Journal of Theoretical and Applied Information Technology
15th November 2016. Vol.93. No.1

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

SIMPLIFIED ADAPTIVE EXON PREDICTORS FOR EXTRACTING PROTEIN CODING REGIONS IN GENOMIC SEQUENCES

SRINIVASAREDDY PUTLURI and MD ZIA UR RAHMAN
Department of Electronics and Communication Engineering, K. L. University,
Green Fields, Vaddeswaram, Guntur- 522502, Andhra Pradesh, India.
E-mail: sriniputluri@gmail.com, mdzr22@gmail.com

ABSTRACT

In the field of Bio-informatics, exact predicting the regions that code for proteins in a deoxyribonucleic acid (DNA) sequence is a challenging and vital task. Analyzing the exon regions is a major phenomenon which helps in drug design and disease identification. The sections of DNA that contain protein coding information are known as exons. Hence predicting the exons in a DNA sequence is a crucial task in genomics. Nucleotides serve as the basic structural unit of a DNA. Three base periodicity (TBP) has been practical in the protein coding regions of DNA sequences for nucleotides. By applying Signal processing techniques, TBP can be easily predicted. Adaptive signal processing techniques found to be likely due to their distinct capability, with the ability to change weight coefficients depending on the gene sequence. In this paper, we propose an efficient adaptive exon predictor (AEP) based on these considerations using error normalization. To increase the tracking ability of the adaptive algorithm for exon regions, we develop AEPs using ELMS algorithm and its variants. These proposed AEPs prominently reduces computational complexity and offers superior performance in terms of performance measures like sensitivity, specificity, and precision, so that the AEPs are attractive in nano devices. It was shown that maximum error normalized sign regressor LMS (MESRLMS) based AEP is better in exon prediction applications based on performance measures with Sensitivity 0.7198, Specificity 0.7203 and Precision 0.6906 at a threshold of 0.8. Also, this algorithm performs better with respect to convergence because of error normalization. Computational complexity wise also MESRLMS needs only one multiplication operation because of sign regressor operation and using a maximum value in normalization. Finally the ability of various AEPs in prediction of exons is tested using different DNA sequences obtained from National Center for Biotechnology Information (NCBI) database.

Keywords: *adaptive exon predictor, computational complexity, deoxyribonucleic acid, disease identification, exons, three base periodicity*

1. INTRODUCTION

Prediction of exon regions is a substantial area of research in the field of genomics. Essential genes form a subset in organisms which are needed for development, survival or fertility [1]-[2]. Therefore, prediction of exons has practical significance to identify human diseases [3] and discover drug targets in novel pathogens [4]-[5]. Regions which code for proteins and non-coding regions are present in a DNA sequence. The Subarea of genomics that deals with spotting the exon locations in a DNA sequence is known as gene prediction. The study of primary protein region structure helps the secondary and tertiary structure of exon region for detection of all anomalies, design drugs and cure diseases, as soon as the complete structure of protein regions is analyzed. These studies support in knowing the evaluation of phylogenic trees [6] - [7]. Based on the fundamental molecular structure, the living organisms are divided into two classifications termed as eukaryotes and prokaryotes. The protein coding genes are continuous and long in prokaryotes; examples of prokaryotes are bacteria and archaea. The genes are a combination of coding regions separated by long non-coding regions in eukaryotes. These regions which code for proteins are also called as exons, whereas the non-protein coding regions are termed as introns. All living organisms other than archaea and bacteria come under this category. The coding regions present in human eukaryotes are only 3% of the sequence and the residual 97% are non-coding regions. Hence the identification of protein coding regions is a vital 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

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

regions are presented in literature based on various signal processing techniques [11] - [14]. 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 algorithms are found to be favorable techniques to process such genomic sequences. 3-base periodicity property is applied to find the protein coding segments accurately in a DNA sequence [15]. Adaptive algorithms are able to process very long sequences in several iterations and can change weight coefficients in accordance to the statistical behavior of the input sequence. In this paper, we propose to develop several Adaptive Exon Predictors (AEPs) using adaptive algorithms for finding protein coding regions. 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 use error normalized and maximum error normalized adaptive algorithms to improve the performance of AEP. Error normalized version of LMS is called as error normalized LMS (ELMS) algorithm. ELMS algorithm overcomes the hitches of LMS and improves tracking ability and convergence speed. 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. Especially 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 prediction and causes inter symbol interference (ISI). Also, the large computational complexity tends to bigger circuit size and large operations, if the AEP is implemented on VLSI circuit or nano device. Hence, to cope up 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 minimizes the number of multiplication operations [16]. The three signum based simplified algorithms are sign regressor algorithm (SRA), sign algorithm (SA) and sign sign algorithm (SSA). Therefore, in order to minimize the computational complexity we combine the three Signum algorithms with the error normalized LMS algorithm. The resulting algorithms are error normalized sign regressor LMS (ESRLMS) algorithm, error normalized sign LMS (ESLMS) algorithm and error normalized sign sign LMS (ESSLMS) algorithm. In these algorithms due to normalization, the denominator of the weight update equation has to compute multiplications equal to the numeric value of tap length of the algorithm. 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. This can be minimized to one, irrespective of tap length by using an approach called maximum normalization [17]. The resultant algorithms are maximum error normalized sign regressor LMS (MESRLMS) algorithm, maximum error normalized sign LMS (MESLMS) algorithm and maximum error normalized sign sign LMS (MESSLMS) algorithm. In error normalized algorithms the step size is normalized with reference to error, the time-varying step-size is inversely proportional to the squared norm of the error vector rather than the input data vector as in the data normalized LMS. ELMS algorithm provides significant improvements in minimizing signal distortion. The advantage of the ELMS algorithm is that the step size can be chosen independent of the input signal power and the number of tap weights. Hence the ELMS algorithm has a convergence rate and a steady state error better than LMS algorithm. Based on these error normalized algorithms, we develop various AEPs and the performance is tested using real genomic sequences obtained from National Center for Biotechnology Information (NCBI) data base [18]. We consider sensitivity (s_n), specificity (s_p), precision (p_r), convergence characteristics, and computational complexity (O) as performance characteristics to evaluate the performance of the various AEPs. 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 proposing AEP, the input genomic sequence is converted into binary representation. This is a vital task in genomic processing since signal processing techniques can be applied only on discrete or digital 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 algorithm. Four binary indicator sequences are used as input to the adaptive filter [15]. Now, we consider

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

an adaptive exon predictor (AEP) to be applied on converted binary sequences. Let G(n) be the DNA sequence, B (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 'M'. In this algorithm, the next weight coefficient can be predicted based on the current weight coefficient, step size parameter 'S', input sequence sample value G(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.

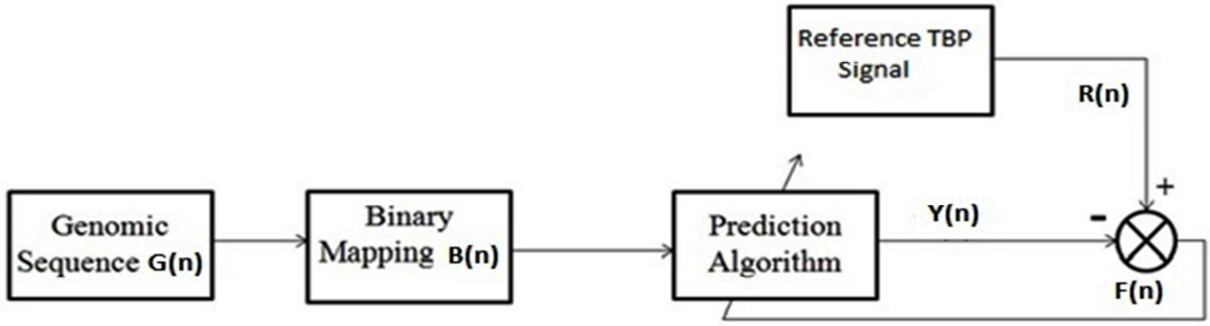

Figure 1. Block diagram of an adaptive exon predictor.

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. 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 data vector is large. To avoid this problem normalization has to be applied. With this, the adjustment applied to filter weight vector coefficient is normalized with respect to a squared Euclidian norm of the input vector at each iteration. Due to normalization, the step size 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.

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

$$u(n+1) = u(n) + S\,F(n)B(n) \qquad (1)$$

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 [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.

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

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 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).

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

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

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

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

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

Further, to reduce the computational complexity of the algorithms we apply data error normalization and maximum error normalization. In this approach, instead of using instantaneous data vector for normalization squared norm of the error vector can be used. The length of the error vector is the instantaneous number of iterations. Because the step size is normalized with reference to error, the resulting adaptive algorithm is called as ELMS algorithm. In the ELMS algorithm, the time-varying step-size is inversely proportional to the squared norm of the error vector rather than the input data vector as in the normalized LMS algorithm. This algorithm provides significant improvements in minimizing signal distortion. The advantage of the ELMS algorithm is that the step size can be chosen independent of the input signal power and the number of tap weights. Hence the ELMS algorithm has a convergence rate and a steady state error better than LMS algorithm. Compared with other normalized algorithms, the ELMS algorithm requires a small number of computations.

Thus, the weight update equation of the ELMS algorithm becomes

Normalized Sign-Sign LMS (ESSLMS) algorithm. In this paper, we have also considered maximum error normalized variants using ELMS which include Maximum Error Normalized Sign Regressor LMS (MESRLMS) algorithm, Maximum Error Normalized Sign LMS (MESLMS) algorithm and Maximum Error Normalized Sign-Sign LMS (MESSLMS) algorithm. These algorithms enjoy less computational complexity because of the sign present in the algorithm and good filtering capability because of the normalized term. The less computational complexity leads to simplified architecture for system on chip (SOC) or lab on chip (LOC).

Here, we have considered error normalized ELMS algorithm and its signed versions in this paper. Due to consideration of the maximum normalization for ELMS algorithm, the weight update relation of Maximum Error Normalized LMS (MELMS) algorithm for $e_{Li} \neq 0$ is written as,

$$h(n+1) = h(n) + \frac{S}{e_{Li}^2}F(n)B(n) \quad (10)$$

$$h(n+1) = h(n) + \frac{S}{\varepsilon+\|e(n)\|^2}F(n)B(n) \quad (6)$$

The weight update relation of ESRLMS is obtained by clipping the input data in ELMS. Now, the weight recursion of ESRLMS becomes

$$h(n+1) = h(n) + \frac{S}{\varepsilon+\|e(n)\|^2}C[F(n)]B(n) \quad (7)$$

The weight update relation of ESLMS is obtained by clipping the error with variable step size and is given by

$$h(n+1) = h(n) + \frac{S}{\varepsilon+\|e(n)\|^2}F(n)C[B(n)] \quad (8)$$

Similarly, the weight update relation of ESSLMS is obtained by clipping both the data and error and is free from multiplications which is given by

$$h(n+1) = h(n) + \frac{S}{\varepsilon+\|e(n)\|^2}C[F(n)]C[B(n)] \quad (9)$$

In order to cope up with both the complexity and convergence issues without any restrictive tradeoff, we propose various sign based adaptive algorithms using ELMS algorithm and their maximum error normalized variants in this paper. The corresponding Signum based adaptive algorithms using ELMS are Error Normalized Sign Regressor LMS (ESRLMS) algorithm, Error Normalized Sign LMS (ESLMS) algorithm and Error

$$where, e_{Li} = max\{|e_k|, k \in Z_i'\}, Z_i' = \{iM,\ iM+1, \dots, iM+M-1\}, \mathrm{i} \in \mathrm{Z}$$

Similarly the weight update relations of MESRLMS, MESLMS, and MESSLMS adaptive algorithms becomes

$$h(n+1) = h(n) + \frac{S}{e_{Li}^2}C[F(n)]B(n) \quad (11)$$

$$h(n+1) = h(n) + \frac{S}{e_{Li}^2}F(n)C[B(n)] \quad (12)$$

$$h(n+1) = h(n) + \frac{S}{e_{Li}^2}C[F(n)]C[B(n)] \quad (13)$$

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

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

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 M+1 MAC operations are required to compute the weight update equation. In case of error normalized signed regressor algorithm only one multiplication is required to compute 'S.F(n)'. Whereas other ESSLMS, MESLMS and MESSLMS based algorithms does not require multiplications if we choose 'S' value a power of 2. In these cases 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 ELMS adaptive algorithm is more complex, as they require 2M+1 MACs and 1 division operations to implement the weight updating equation (6) on a DSP processor. Among the proposed AEPs, ESRLMS and MESRLMS algorithms provide less computational complexity with 1 MAC and 1 division operations. However, by using a maximum normalization approach, we can minimize multiplications in the denominator from 'M' to '1'.

Compared with other normalized algorithms, the ELMS algorithm requires a small number of computations. To compute the variable step 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	ASC	Divisions	Shifts
1	LMS	M+1	Nil	Nil	Nil
2	ELMS	2M+1	Nil	1	Nil
3	ESRLMS	M	Nil	1	Nil
4	ESLMS	2M	Nil	1	Nil
5	ESSLMS	Nil	M	1	2M
6	MELMS	M+1	Nil	1	Nil
7	MESRLMS	1	Nil	1	Nil
8	MESLMS	Nil	M	1	M
9	MESSLMS	Nil	M	1	M

The ELMS algorithm provides significant improvements in minimizing signal distortion. The advantage of this algorithm is that the step size can be chosen independent of the input signal power and the number of tap weights. Hence the ELMS algorithm has a convergence rate and a steady state error better than LMS algorithm. In order to cope up with both the complexity and convergence issues without any restrictive tradeoff, the corresponding signum based adaptive algorithms using ELMS is Error Normalized Sign Regressor LMS (ESRLMS) algorithm, Error Normalized Sign LMS (ESLMS) algorithm and Error Normalized Sign-Sign LMS (ESSLMS) algorithm. These algorithms provide less computational complexity because of the sign present in the algorithm and good filtering capability because of the normalized term. By applying maximum normalization, we have also considered maximum error normalized variants using ELMS which include Maximum Error Normalized Sign Regressor LMS (MESRLMS) algorithm, Maximum Error Normalized Sign LMS (MESLMS) algorithm and Maximum Error Normalized Sign-Sign LMS (MESSLMS) algorithm for low computational complexity and good filtering capability. 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 ESRLMS is just inferior to its non-sign regressor version. Hence, among the algorithms considered for the implementation of the AEPs MESRLMS algorithm is found to be better with reference to computational complexity and convergence characteristics.

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

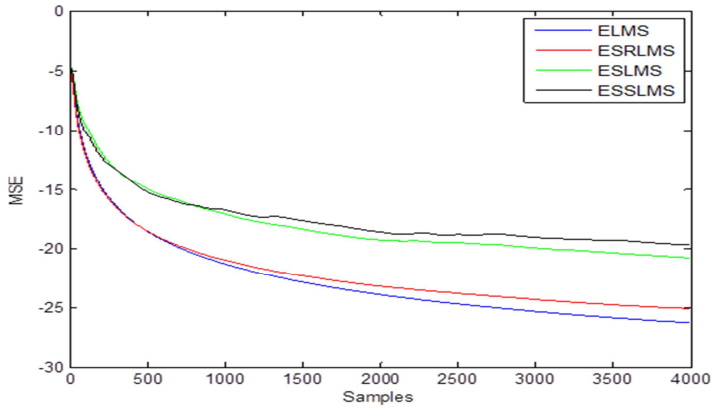

Figure 2: Convergence characteristics of error normalized LMS with its signed based variants.

4. RESULTS AND DISCUSSIONS

In this section, the performances are compared for various AEPs. The structure of AEP is shown in Figure 1. The maximum data normalized LMS algorithm and its sign based versions are used to develop various AEPs. For comparison purpose, we also develop an LMS based AEP. For evaluation purpose, we obtained ten DNA sequences from NCBI database [14]. 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 1 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. At threshold 0.8 the exon prediction seems to be better. Hence at threshold 0.8 the values are shown 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 for 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 obeying 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

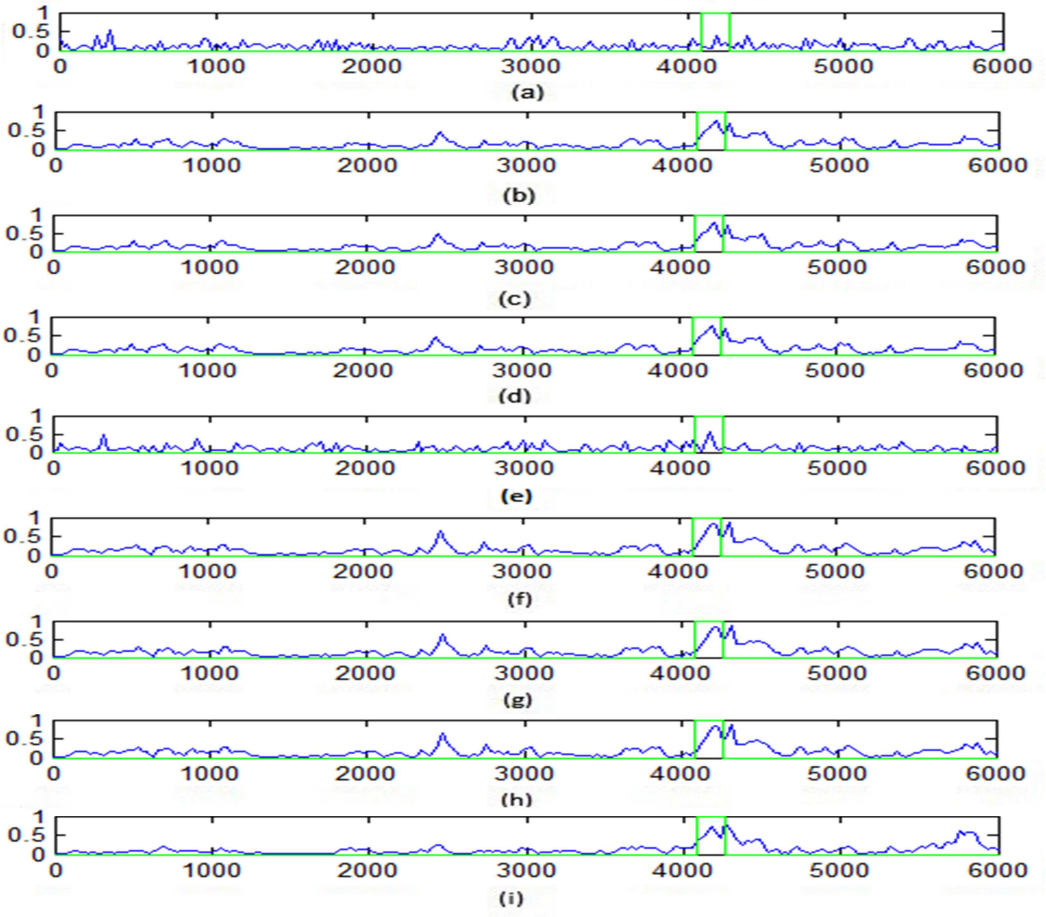

Figure 3: Locations of exons predicted using various adaptive algorithms (a). LMS based AEP, (b). ELMS based AEP, (c). ESRLMS based AEP, (d). ESLMS based AEP, (e). ESSLMS based AEP, (f). MELMS based AEP (g). MESRLMS based AEP, (h). MESLMS based AEP and (i). MESSLMS based AEP

Figure 3 shows the predicted exon locations of sequence 3 applying various adaptive algorithms. From this plots it is clear that the LMS based AEP not predicted the coding regions accurately. This algorithm causes some ambiguities in location prediction by identifying some non-coding regions. 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 have been presented in the literature [11]–[14]. But, using proposed error normalized and maximum error normalized based AEP versions, the ELMS, ESRLMS, ESSLMS, MELMS, MESRLMS and MESSLMS algorithms 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 algorithm. Among these three algorithms ESRLMS 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 this algorithm are independent of tap length of AEP. The convergence characteristics of ESRLMS are just inferior to ELMS, but due to a large number of reduced multiplications this inferior behavior in convergence can be tolerable. In case of MESSLMS, due to clipped input sequence and clipped feedback signal the performance of exon perdition is inferior to other signed versions. Therefore, based on computational complexity, convergence characteristics, exon prediction plots, Sn, Sp and Pr calculations, it is found that MESRLMS based AEP is found to be the better candidate in practical applications for the development of SOCs, LOCs and nano devices in future research.

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

Table 3: Performance measures of various AEPs with respect to Sn, Sp and Pr calculations.

Seq. No.	Parameter	LMS	ELMS	ESRLMS	ESLMS	ESSLMS	MELMS	MESRLMS	MESLMS	MESSLMS
1	Sn	0.6286	0.7017	0.6813	0.6625	0.6323	0.7337	0.7198	0.6931	0.6714
	Sp	0.6435	0.7106	0.6924	0.6897	0.6764	0.7456	0.7203	0.6922	0.6859
	Pr	0.5922	0.6797	0.6529	0.6307	0.6176	0.7131	0.6906	0.6767	0.6547
2	Sn	0.6384	0.6927	0.6743	0.6512	0.6363	0.7365	0.7131	0.6898	0.6622
	Sp	0.6628	0.7234	0.7106	0.6951	0.6732	0.7496	0.7203	0.7098	0.6819
	Pr	0.5894	0.6832	0.6565	0.6316	0.6024	0.7023	0.6694	0.6427	0.6225
3	Sn	0.6437	0.7027	0.6873	0.6635	0.6353	0.7295	0.7053	0.6896	0.6624
	Sp	0.6587	0.7224	0.7006	0.6824	0.6689	0.7456	0.7205	0.6912	0.6717
	Pr	0.5902	0.6734	0.6514	0.6322	0.6117	0.7031	0.6716	0.6574	0.6217
4	Sn	0.6273	0.7013	0.6837	0.6685	0.6396	0.7235	0.7021	0.6897	0.6618
	Sp	0.6405	0.7102	0.6923	0.6717	0.6575	0.7436	0.7202	0.7194	0.6827
	Pr	0.5858	0.6724	0.6578	0.6314	0.6127	0.7131	0.6912	0.6792	0.6524
5	Sn	0.6481	0.7038	0.6849	0.6645	0.6371	0.7365	0.7180	0.6847	0.6640
	Sp	0.6518	0.7110	0.6925	0.6669	0.6563	0.7435	0.7221	0.6942	0.6732
	Pr	0.5904	0.6722	0.6441	0.6246	0.6037	0.7045	0.6711	0.6544	0.6254
6	Sn	0.6162	0.7035	0.6897	0.6613	0.6475	0.7315	0.7163	0.6912	0.6716
	Sp	0.6324	0.7194	0.6912	0.6651	0.6483	0.7418	0.7103	0.6936	0.6643
	Pr	0.5786	0.6702	0.6559	0.6314	0.6136	0.7111	0.6926	0.6727	0.6537
7	Sn	0.6193	0.7027	0.6823	0.6615	0.6423	0.7327	0.7131	0.6894	0.6614
	Sp	0.6529	0.7214	0.7018	0.6841	0.6628	0.7446	0.7203	0.7014	0.6779
	Pr	0.5896	0.6734	0.6557	0.6332	0.6186	0.7121	0.6994	0.6747	0.6525
8	Sn	0.6241	0.7095	0.6823	0.6643	0.6435	0.7343	0.7125	0.6934	0.6718
	Sp	0.6289	0.7054	0.6918	0.6894	0.6342	0.7438	0.7242	0.7012	0.6617
	Pr	0.5856	0.6736	0.6539	0.6328	0.6186	0.7137	0.6902	0.6724	0.6515
9	Sn	0.6268	0.7019	0.6827	0.6647	0.6463	0.7397	0.7191	0.6997	0.6724
	Sp	0.6452	0.7207	0.6984	0.6807	0.6928	0.7428	0.7203	0.6927	0.6619
	Pr	0.5814	0.6722	0.6577	0.6324	0.6184	0.7113	0.6902	0.6756	0.6513
10	Sn	0.6202	0.7087	0.6853	0.6643	0.6423	0.7347	0.7131	0.6982	0.6702
	Sp	0.5965	0.6824	0.6526	0.6331	0.6213	0.7324	0.7013	0.6796	0.6443
	Pr	0.5761	0.6716	0.6569	0.6314	0.6176	0.7111	0.6906	0.6738	0.6515

5. 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 error normalized adaptive algorithms. In order to reduce computational complexity of the proposed implementation, we introduced the concept of error adaptive normalization instead of data normalization. To further minimize the computational complexity, the proposed ELMS algorithm is combined with its sign based and maximum normalized algorithms. As a result seven new hybrid algorithms come into the scenario of exon prediction. The hybrid variants are ESRLMS, ESLMS, ESSLMS, MELMS, MESRLMS, MESLMS and MESSLMS are considered for present implementation. Different AEPs are developed and tested using these seven algorithms on real DNA sequences obtained from NCBI database. It is evident that MESRLMS 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 0.7198, Specificity 0.7203 and precision 0.6906 obtained at a threshold value of 0.8. 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. The limitation of presented work is that proposed AEPs may not be so accurate for finding very short length exons. Therefore, proposed AEP realizations are suitable 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] Itaya M, "An estimation of minimal genome size required for life," *Federation of European Biochemical Societies (FEBS) letters*, 362(3), 1995, pp. 257–260.

[2] Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen K, Arnaud M, Asai K,Ashikaga S, Aymerich S, Bessieres P, "Essential Bacillus subtilis genes," *Proceedings of the National Academy of Sciences of the United States of America*, 100(8), 2003, pp. 4678–4683.

[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] Chalker AF, Lunsford RD, "Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach," *Pharmacology & Therapeutics*, 95(1), 2002, pp. 1–20.

[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.Nemati, M. E. Basiri, N.Ghasem-Aghaee and M.H. Aghdam, "A novel ACO-GA hybrid algorithm for feature selection in protein function prediction," *Expert Systems with Applications*, vol. 36, no. 10, 2009, pp. 12086–12094.

[7] S. A. Marhon and S. C. Kremer, "Gene prediction based on DNA spectral analysis: a literature review," *Journal of Computational Biology*, vol. 18, no. 4, 2011, pp. 639–676.

[8] S. Maji and D. Garg, "Progress in gene prediction: principles and challenges," *Current Bioinformatics*, vol. 8, no. 2, 2013, pp. 226–243.

[9] N. Goel, S. Singh, and T. C. Aseri, "A review of soft computing techniques for gene prediction," ISRN Genomics, vol. 2013, Article ID 191206, pp. 1-8.

[10] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Computer Applications in the Biosciences*, vol. 13, no. 3, 1997, pp. 263–270

[11] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 1, 2002, pp. 13–28.

[12] Fox, T.W. and Alex Carreira, "A digital signal processing method for gene prediction with improved noise suppression," *EURASIP Journal on Applied Signal Processing*, vol. 1, 2004, pp. 108-114.

[13] N. Rao, X. Lei, J. Guo, H. Huang, and Z. Ren, "An efficient sliding window strategy for accurate location of eukaryotic protein coding regions," *Computers in Biology and Medicine*, vol. 39, no. 4, 2009, pp. 392–395.

[14] Parameswaran Ramachandran, Wu-Sheng Lu, Andreas Antoniou, "Filter-Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA," *IEEE Transactions on Biomedical Engineering*, vol. 59, no.6., 2012

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

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