NEW ADAPTIVE EXON PREDICTORS FOR IDENTIFYING PROTEIN CODING REGIONS IN DNA SEQUENCE

Srinivasareddy Putluri and Md Zia Ur Rahman
Department of Electronics and Communication Engineering, K. L. University,Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India
E-Mail: mdzr22@gmail.com

ABSTRACT

Identification of the regions that code for proteins in a deoxyribonucleic acid (DNA) sequence is a vital and challenging task in the area of Bioinformatics. Study of exon regions is a substantial phenomenon in designing drugs and identification of diseases. The fragments of DNA that contain protein coding information are termed as exons. Hence finding the exon locations in a DNA sequence is a crucial job in genomics. Nucleotides aid as the fundamental structural unit of a DNA. Three base periodicity (TBP) has been observed in the regions of DNA sequences which code for proteins in case of nucleotides. By applying signal processing methods, TBP can be easily determined. Adaptive signal processing methods found to be probable in comparison with several other methods. This is due to the distinctive ability of adaptive algorithms to change weight coefficients depending on genomic sequence. We propose a novel adaptive exon predictor (AEP) based on these deliberations using normalization to improve pursuing ability of the adaptive algorithms. We developed AEPs using LMS algorithm with its data clipped; error clipped and signed normalized variants to reduce computational complexity. Hybrid variants of proposed AEPs include DCLMS, ECLMS, ECLMS, DNLMS, DNDCLMS, DNECLMS, and DNDECLMS algorithms. It was shown that DNDCLMS based AEP is better in exon prediction applications based on performance measures with Sensitivity 0.6872, Specificity 0.7043 and precision 0.6722 at a threshold of 0.8. Finally the capability of several AEPs in predicting exon locations is verified using different genomic sequences found from National Center for Biotechnology Information (NCBI) database.

Keywords: adaptive exon predictor, bioinformatics, computational complexity, deoxyribonucleic acid, three base periodicity.

1. INTRODUCTION

A major objective of research in genomics is to know the nature of information along with its role in determining a particular function encoded by the gene. A vital step to achieve this goal is identification of protein coding segments in a DNA sequence [1]. Finding the exon regions is a extensive area of research in the field of genomics. Necessary genes form a subset in organisms which are required for the development, survival or fertility [2] - [3]. Therefore, finding the exons is not only interesting, but also has real importance to find human diseases [4] and discover targets of drugs in new pathogens [5] - [6]. The genic and intergenic segments are present in a DNA sequence. The Subarea of genomics that deals with spotting the protein coding segments in a DNA sequence is known as gene identification. The learning of primary protein region structure helps in analyzing the secondary and tertiary structure of exon regions. Once the whole structure of protein coding regions is analyzed, we can detect all abnormalities, design drugs and cure diseases. These studies help in knowing the assessment of phylogenic trees [7] - [8]. These days, a fast growth of raw data of genomic sequences needs effective biological elucidations, but more cost is involved to conduct biological experiments for predicting gene locations and there is still a practical demand for fast and efficient tools mainly to find genes, to analyze sequences and determine their functions [9] - [10]. Based on the elementary molecular cell structure, all living organisms are divided into two categorizations termed as eukaryotes and prokaryotes. The protein coding regions responsible for synthesis of proteins are continuous and long in prokaryotes; bacteria and archaea are the examples of prokaryotes. The genes are a combination of coding segments separated by long non-coding segments in eukaryotes [11]. These segments which are responsible for protein coding are termed as exons, whereas the non-protein coding fragments are termed as introns. Other than archaea and bacteria, all the living organisms 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 [12] - [13]. Almost in all DNA sequences, a three base periodicity (TBP) is exhibited by the protein coding regions. This is apparent by a sharp peak at a frequency f=1/3 in the power spectral density (PSD) plot [14]. Several exon prediction techniques are presented in literature based on several signal processing techniques [15] - [18]. But, the length of the sequence in real-time gene sequence is extremely long and also the location of the exons varies from sequence to sequence. To process such gene sequences, adaptive algorithms are found to be promising techniques. 3-base periodicity property is useful to find the protein coding segments in a DNA sequence [19]. Adaptive algorithms are able to process very long sequences in multiple iterations and can change weight coefficients in accordance to the statistical behavior of the input sequence. In this paper, we propose to develop an Adaptive Exon Predictor (AEP) using adaptive algorithms. Least mean squares (LMS) algorithm is a fundamental adaptive technique. This algorithm is prevalent because of its simplicity in implementation. But this algorithm undergoes problems like amplification of gradient noise, weight drift and poor convergence. Hence, to increase the performance of AEP, we put forward to use data clipped, error clipped, data error clipped and normalized adaptive algorithms. The three resultant algorithms are Data Clipped LMS (DCLMS), Error Clipped LMS (ECLMS), and Data Error Clipped LMS (DECLMS) algorithms. Data Normalized version of LMS is called as Normalized LMS (DNLMS) algorithm. DNLMS 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 crucial 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 causes inter symbol interference (ISI) and leads to inaccuracy in the prediction. Also, when the AEP is implemented on VLSI circuit or nano device the large computational complexity tends to bigger circuit size and large operations. 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 [20]. 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 normalized LMS 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 [21].The resulting normalized versions are Data Normalized LMS (DNLMS), Data Normalized Data Clipped LMS (DNDCLMS), Data Normalized Error Clipped LMS (DNECLMS) and Data Normalized Data Error Clipped LMS (DNDECLMS) algorithms. In the normalization version of the LMS algorithm the correlation between the error and the reference input is normalized by a factor equal to the squared norm. In normalized algorithms, the gradient noise application problem is minimized and it converges faster than the conventional LMS algorithm. Hence the DNLMS algorithm has a convergence rate and a steady state error better than LMS algorithm. Based on these data clipped, error clipped and normalized adaptive algorithms, we develop various AEPs and the performance is tested using real genomic sequences obtained from the National Center for Biotechnology Information (NCBI) database [22]. We consider convergence characteristics, computational complexity (O), sensitivity (s_n), specificity (s_p) and precision (p_r) as performance characteristics to evaluate the performance of the various AEPs. The theory of the adaptive algorithms, results of AEPs and discussion on the performance of various AEPs is 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 [18]. 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 [19]. Now, we consider an adaptive exon predictor (AEP) to be applied for converting binary sequences. Let G(n) be the DNA sequence, B(n) be 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 ‘T’. 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 [20]. A typical block diagram of proposed AEP is shown in Figure-1.

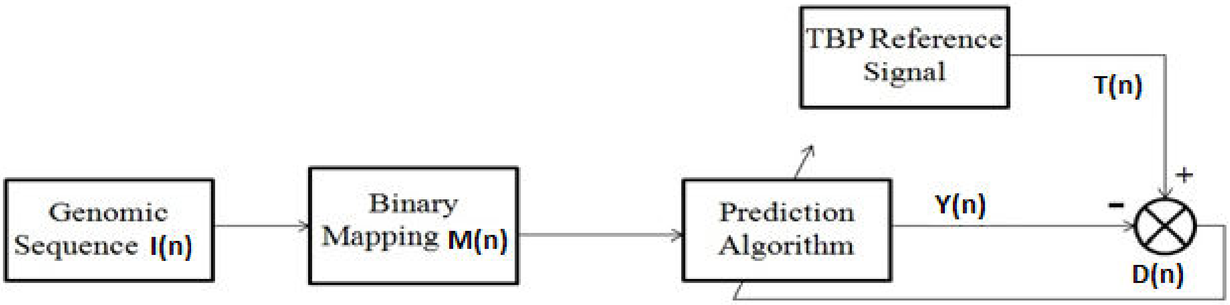

Figure-1. Block diagram of an adaptive exon predictor.

Because of its simplicity and vigor, the conventional LMS algorithm may be used in exon prediction applications. For convergence and stability, the LMS filter needs a prior knowledge of the input power level to select the step size parameter. As the input power level is usually one of the statistical unknowns, it is normally estimated from the data before beginning the adaptation process. But 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 problem is the fixed step size. In practice, an algorithm has to be designed such that, it has to handle 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. Normalization has to be applied to avoid this problem. With this, the 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 LMS adaptive algorithm is given by

$$u(n+1) = u(n) + S\,I(n)D(n) \qquad (1)$$

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 [21]. These are sign regressor algorithm (SRA), sign algorithm (SA) and sign sign algorithm (SSA). Among the adaptive algorithms, the SRA, SA and SSA 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.

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

To reduce the computational complexity compared with LMS adaptive algorithms, we use SRA, SA and SA adaptive algorithms.The computational complexity of these algorithms is much less compared to the LMS algorithm. The Data Clipped LMS (DCLMS) algorithm is obtained from the conventional LMS recursion by replacing the tap-input vector I(n) with the vector C[D(n)], where the sign function C is applied to the vector D(n) on an element-by-element basis. This is also called as clipped LMS as we are clipping the input data.

The weight update relation of CLMS algorithm is given by

$$u(n+1) = u(n) + S\,I(n)C[D(n)] \qquad (3)$$

The weight update relation of ECLMS algorithm is obtained by replacing I(n) with its signed form and is given by

$$\mathrm{u(n+1) = u(n) + S\,C[I(n)]D(n)} \qquad (4)$$

Similarly, the weight update relation of DECLMS algorithm is obtained by replacing I(n), D(n) with its signed forms and is given by

$$\mathrm{u(n+1) = u(n) + S\,C[I(n)]C[D(n)]} \qquad (5)$$

In overcoming the gradient noise amplification problem associated with the conventional LMS filter, the normalized LMS filter introduces a problem of its own, namely the tap input vector I(n) is small, numerical difficulties may arise because then we have to divide by a small value for the squared norm. To overcome this problem, we modify the above recursion by adding a small positive constant ε. The parameter ε is set to avoid denominator being too small and step size parameter is too big.

Now the step size parameter is written as,

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

where S(n) is a normalized step size with 0 <S< 2. Replacing S in the LMS weight vector update equation with S(n) leads to the DNLMS, which is given as

$$\mathrm{u(n+1) = u(n)} + \frac{\mathrm{S}}{||\mathrm{I(n)}||2}\mathrm{I(n).\,D(n)} \qquad (7)$$

To further reduce the computational complexity of the sign algorithms and for faster convergence, these three simplified sign algorithms are combined with DNLMS algorithms. The advantage of the DNLMS algorithm is that the step size can be chosen independent of the input signal power and the number of tap weights. Hence the DNLMS algorithm has a convergence rate and a steady state error better than LMS algorithm. On the other hand, some additional computations are required to compute D(n).Further, to reduce the computational complexity of the algorithms we apply data normalization to the LMS adaptive algorithm. In this approach the correlation between the error and the reference input is normalized by a factor equal to the squared norm. This reduces the number of multiplications from a value equal to tap length ‘C’ of the algorithm to only one. In the DNLMS algorithm, step size can be chosen independent of the input signal power and the number of tap weights. This algorithm provides significant improvements in minimizing signal distortion. The advantage of the DNLMS algorithm gives the correlation between the error and the reference input is normalized by a factor equal to the squared norm. Hence the DNLMS algorithm has a convergence rate and a steady state error D(n) better than

LMS algorithm. Compared with LMS algorithm, the DNLMS algorithm requires a small number of computations.

Thus, the weight update equation of the DNLMS algorithm becomes

$$\mathrm{u(n+1) = u(n)} + \frac{\mathrm{S}}{\varepsilon+\|\mathrm{I(n)}\|^2}\mathrm{I(n)}D(\mathrm{n}) \tag{8}$$

Similarly, the weight update relations of DNDCLMS, DNECLMS and DNDECLMS becomes

$$\mathrm{u(n+1) = u(n)} + \frac{\mathrm{S}}{\varepsilon+\|\mathrm{I(n)}\|^2}\mathrm{I(n)}C[D(\mathrm{n})] \tag{9}$$

$$\mathrm{u(n+1) = u(n)} + \frac{\mathrm{S}}{\varepsilon+\|\mathrm{I(n)}\|^2\mathrm{C}[}\mathrm{C[I(n)]}D(\mathrm{n}) \tag{10}$$

$$\mathrm{u(n+1) = u(n)} + \frac{\mathrm{S}}{\varepsilon+\|\mathrm{I(n)}\|^2}\mathrm{C[I(n)]}C[D(\mathrm{n})] \tag{11}$$

In order to cope up with both the complexity and convergence issues without any restrictive tradeoff, we propose various data clipped, error clipped, and normalized adaptive variants of LMS in this paper. The corresponding adaptive algorithms using LMS and DNLMS are Data Clipped LMS (DCLMS), Error Clipped LMS (ECLMS), Data Error Clipped LMS (DECLMS), Data Normalized LMS (DNLMS), Data Normalized Data Clipped LMS (DNDCLMS), Data Normalized Error Clipped LMS (DNECLMS) and Data Normalized Data Error Clipped LMS (DNDECLMS) algorithms. The normalized algorithms enjoy less computational complexity because of the sign present in the algorithm and good filtering capability because of the normalized term.

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 thesis, we are not trying to provide a precise analysis of a computational complexity; rather we concentrate on presenting a comparison between different adaptive algorithms. The computational complexity figures required to compute various algorithms considered are 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 requires T+1 MACs to compute the weight update equation. In case of signed regressor algorithm only one multiplication and accumulate operation is required to compute 'S.D(n)'. Whereas other two signed LMS algorithms does not require multiplication 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 DNLMS algorithm is more complex; as it requires 2T+1 MACs and 1 division operations implement the weight updating equation (8) on a DSP processor. In case of the DNDCLMS adaptive algorithm, computational complexity is less compared with other normalized algorithms with 1 MAC and 1 Division operations. Note that ASC and shift operations require less logic circuitry when compared to MAC operations. However, by using a maximum normalization approach, we can minimize multiplications in the denominator from 'T' to '1'.

Compared with other normalized algorithms, the DNDCLMS 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	T+1	Nil	Nil	Nil
2	DCLMS	1	Nil	Nil	Nil
3	ECLMS	T	Nil	Nil	Nil
4	DECLMS	Nil	T	Nil	Nil
5	DNLMS	2T+1	Nil	1	Nil
6	DNDCLMS	1	Nil	1	Nil
7	DNECLMS	T	Nil	1	Nil
8	DNDECLMS	Nil	T	1	T

In order to cope up with both the complexity and convergence issues without any restrictive tradeoff, the corresponding signum based normalized adaptive algorithms considered using LMS are Data Normalized LMS (DNLMS), Data Normalized Data Clipped LMS (DNDCLMS), Data Normalized Error Clipped LMS (DNECLMS) and Data Normalized Data Error Clipped LMS (DNDECLMS) algorithms. These algorithms provide less computational complexity because of the sign present in the algorithm and good filtering capability because of the normalized term. These normalized adaptive algorithms offers low computational complexity and good filtering capability compared to converntional LMS adaptive algorithm. The less computational complexity of these adaptive algorithms leads to simplified architecture for system on chip (SOC) or lab on chip (LOC).

The convergence characteristics of the Data Clipped LMS (DCLMS), Data Error Clipped LMS (DECLMS) and its data normalized adaptive algorithms are shown in Figure-2. From these characteristics, it is clear that normalized adaptive algorithms have a faster convergence rate than LMS. Hence, among the algorithms considered for the implementation of AEPs, the DNDCLMS adaptive algorithm is found to be better with reference to computational complexity and convergence characteristics than other normalized algorithms.

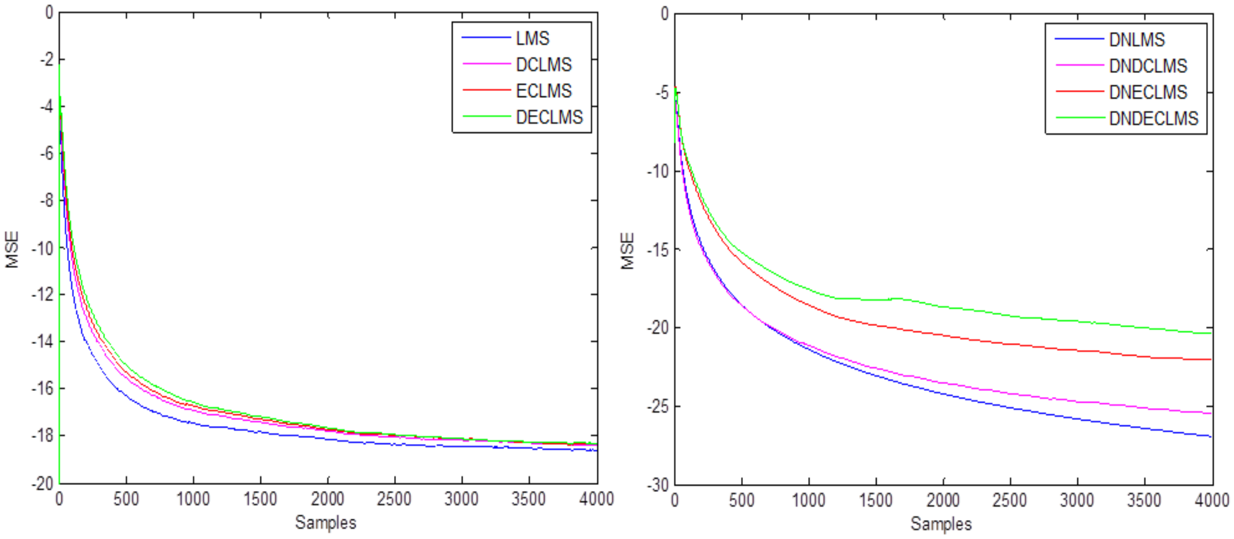

Figure-2. Convergence characteristics of data clipped, error clipped LMS and its normalized variants.

4. RESULTS AND DISCUSSIONS

In this section, performances of various AEPs are compared. 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 [22]. 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 [18][23]. 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 is 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:

a) Input DNA sequences are chosen from NCBI database. Using binary mapping technique convert DNA sequence to binary data. Provide obtained binary data as input to AEP structure shown in Figure-1.
b) A biological sequence obeying three base periodicity is given as reference to the AEP.
c) As shown in Figure-1, a feedback signal that is generated is used to update filter coefficients.
d) The feedback signal when becomes minimum, the location of the coding region sequence is predicted accurately.
e) With the help of power spectral density, location of the predicted exon region is plotted. The performance measures like Sn, Sp and Pr are measured.

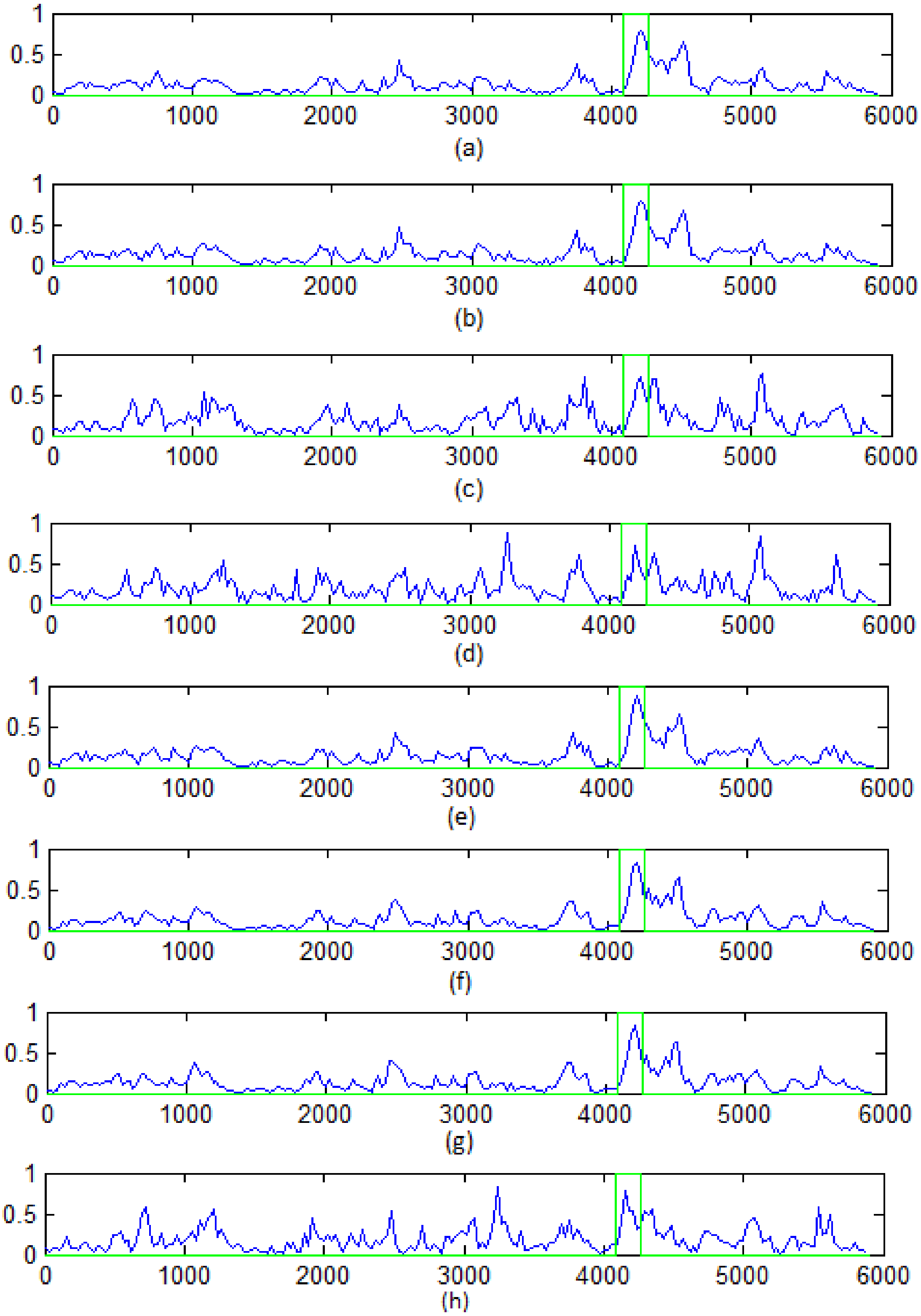

Figure-3. Locations of exons predicted using various adaptive algorithms, (a). LMS based AEP, (b). DCLMS based AEP, (c). ECLMS based AEP, (d). DECLMS based AEP, (e). DNLMS based AEP, (f). DNDCLMS based AEP (g).DNECLMS based AEP, and (h). DNDECLMS based AEP

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 is 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. At the same time the actual exon location 4084-4268 is not predicted. Prediction measures such as sensitivity, specificity and precision of DCLMS, ECLMS and DECLMS algorithms are observed a bit inferior than LMS adaptive algorithm where these are much better in case of normalized algorithms. In the case of normalized versions, the DNLMS, DNDCLMS, DNECLMS and DNDECLMS algorithms exactly predicted the exon locations at 4084-4268 with good intensity of PSD are observed. These PSDs are shown in Figures 3 (b), (c) and (d).

Because of the normalization involved in these algorithms the tracking capability of these algorithms is better than LMS algorithm. Among these three algorithms DNDCLMS is found to be better with reference to its computational complexity and convergence characteristics. 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 DNDCLMS are better than other normalized algorithms. In the case of DNDECLMS, 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 DNDCLMS based AEP is found to be the better candidate in practical applications.

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

Seq. no.	Parameter	LMS	DCLMS	ECLMS	DECLMS	DNLMS	DNDCLMS	DNECLMS	DNDECLMS
1	Sn	0.6286	0.5813	0.5313	0.4413	0.7085	0.6872	0.6687	0.6436
	Sp	0.6435	0.6261	0.5774	0.5171	0.7267	0.7043	0.6802	0.6642
	Pr	0.5922	0.5694	0.5334	0.5278	0.6954	0.6722	0.6545	0.6115
2	Sn	0.6384	0.6023	0.5802	0.4486	0.7137	0.6996	0.6741	0.6582
	Sp	0.6628	0.6054	0.5745	0.5171	0.7458	0.7263	0.7057	0.6876
	Pr	0.5894	0.5727	0.5583	0.5429	0.7027	0.6638	0.6336	0.6195
3	Sn	0.6437	0.6236	0.5937	0.4835	0.7227	0.7027	0.6838	0.6612
	Sp	0.6587	0.6084	0.5716	0.5581	0.7321	0.7137	0.6946	0.6736
	Pr	0.5902	0.5694	0.567	0.5587	0.6987	0.6602	0.6435	0.6295
4	Sn	0.6273	0.5473	0.5136	0.4831	0.7086	0.6862	0.6694	0.6462
	Sp	0.6405	0.6315	0.5756	0.5257	0.7278	0.7036	0.6851	0.6674
	Pr	0.5858	0.5634	0.5586	0.5489	0.7096	0.6734	0.6553	0.6156
5	Sn	0.6481	0.5849	0.4762	0.4514	0.724	0.7026	0.6852	0.6614
	Sp	0.6518	0.6105	0.5799	0.5684	0.7378	0.7114	0.6902	0.6707
	Pr	0.5904	0.5751	0.571	0.5704	0.6927	0.6672	0.6433	0.6225
6	Sn	0.6162	0.6072	0.5827	0.4528	0.7162	0.6814	0.6524	0.6372
	Sp	0.6324	0.6151	0.5633	0.4765	0.7284	0.7035	0.6727	0.6582
	Pr	0.5786	0.5686	0.5463	0.5329	0.6857	0.6526	0.6295	0.6084
7	Sn	0.6193	0.5929	0.5364	0.4927	0.7192	0.6894	0.6602	0.6492
	Sp	0.6529	0.6145	0.5746	0.5049	0.7396	0.7112	0.6994	0.6776
	Pr	0.5896	0.5764	0.5345	0.5132	0.6904	0.6793	0.6484	0.6175
8	Sn	0.6241	0.5915	0.5705	0.4862	0.7162	0.6814	0.6524	0.6372
	Sp	0.6289	0.6438	0.5726	0.4686	0.7284	0.7035	0.6727	0.6582
	Pr	0.5856	0.5753	0.5539	0.5191	0.6857	0.6526	0.6295	0.6084
9	Sn	0.6268	0.5941	0.534	0.4562	0.7285	0.7074	0.6786	0.6512
	Sp	0.6452	0.565	0.5481	0.432	0.7393	0.7068	0.6874	0.6642
	Pr	0.5814	0.5726	0.5362	0.5052	0.6896	0.6554	0.6285	0.5893
10	Sn	0.6202	0.5848	0.564	0.4321	0.7286	0.7046	0.6772	0.6598
	Sp	0.5965	0.5457	0.5243	0.5152	0.6976	0.6738	0.6546	0.6353
	Pr	0.5761	0.5525	0.5372	0.5151	0.6825	0.6688	0.6458	0.6156

5. CONCLUSIONS

In this paper, the problem of identifying exons in a DNA sequence is illustrated. The concept of predicting the exact location of exons has several applications in current health care technology. At this point, we considered adaptive exon identification technique. To fulfill this we considered data clipped, error clipped, data error clipped adaptive LMS algorithms to minimize the number of computations. In order to further reduce computational complexity of the proposed implementation, we introduced the concept of normalization in addition to conventional LMS. To further minimize the computational complexity the proposed DNLMS algorithm is combined with its sign based and normalized algorithms. As a result seven new hybrid algorithms come into the scenario of exon prediction. The hybrid variants are DCLMS, ECLMS, DECLMS, DNLMS, DNDCLMS, DNECLMS, and DNDECLMS are considered for the current implementation. Different AEPs are developed and tested using these seven algorithms on real DNA sequences obtained from NCBI database. It is evident that DNDCLMS based AEP is better in exon prediction applications, based on the convergence characteristics shown in Figure-2 and computational complexities shown in Table 1. This is also clear from the performance measures tabulated in Table-3 and PSD of exon locations shown in Figure 3. Proposed AEPs exactly predicted the exon locations at 4084-4268 with good intensity as shown in PSD plot. The proposed DNDCLMS based AEP based realization provides superior performance in terms of computational complexity based on performance measures with Sensitivity 0.6872, Specificity 0.7043 and precision 0.6722 obtained at a threshold value of 0.8. Therefore, the proposed normalized based AEPs are suitable for practical genomic applications for the development of nano devices, LOCs, and SOCs.

REFERENCES

[1] Sitanshu Sekhar Sahu and Ganapathi Panda. 2011. Identification of Protein-Coding Regions in DNA Sequences Using A Time-Frequency Filtering Approach. Genomics Proteomics & Bioinformatics. 9(1-2): 45-55.

[2] Itaya M. 1995. An estimation of minimal genome size required for life.Federation of European Biochemical Societies (FEBS) letters. 362(3): 257-260.

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

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

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

[6] Cole S. 2002. Comparative myco bacterial genomics as a tool for drug target and antigen discovery. The European Respiratory Journal. 20(36 suppl):78s-86s.

[7] S.Nemati, M. E. Basiri, N.Ghasem-Aghaee and M.H. Aghdam. 2009. A novel ACO-GA hybrid algorithm for feature selection in protein function prediction.Expert Systems with Applications. 36(10): 12086-12094.

[8] S. A. Marhon and S. C. Kremer. 2011. Gene prediction based on DNA spectral analysis: a literature review. Journal of Computational Biology. 18(4): 639-676.

[9] C. Mathe, M. Sagot, T. Schiex and P. Rouze. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research. 30(19): 4103-4117.

[10] N. Y. Song and H. Yan. 2011. Short exon detection in DNA sequences based on multifeature spectral analysis. EURASIP Journal on Advances in Signal Processing. 2011(Article ID 780794): 1-8.

[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 Applied Anaysis. 2014(2014): 1-14.

[12] S. Maji and D. Garg. 2013. Progress in gene prediction: principles and challenges.Current Bioinformatics. 8(2): 226- 243.

[13] N. Goel, S. Singh, and T. C. Aseri. 2013. A review of soft computing techniques for gene prediction.ISRN Genomics. 2013(Article ID 191206): 1-8.

[14] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya and R. Ramaswamy. 1997. Prediction of probable genes by Fourier analysis of genomic sequences.Computer Applications in the Biosciences. 13(3): 263-270.

[15] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis. 2004. Autoregressive modeling and feature analysis of DNA sequences.EURASIP Journal on Applied Signal Processing. 2004(1): 13-28.

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

[17] N. Rao, X. Lei, J. Guo, H. Huang and Z. Ren. 2009. An efficient sliding window strategy for accurate location of eukaryotic protein coding regions.Computers in Biology and Medicine. 39(4): 392-395.

[18] Parameswaran Ramachandran, Wu-Sheng Lu, Andreas Antoniou. 2012. Filter-Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA. IEEE Transactions on Biomedical Engineering. 59(6).

[19] Mohammed Abo-Zahhad, Sabah M. Ahmed, Shimaa A. Abd-Elrahman. 2012. Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques. International Journal of Information Technology and Computer Science. 8(2012): 22-36.

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

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

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

[23] Yusuke Azuma and Shuichi Onami. 2014.Automatic Cell Identification in the Unique System of Invariant Embryogenesis in Caenorhabditis elegans. Biomedical Engineering Letters. 4(2014): 328-337.