IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 24, 2019, accepted June 12, 2019, date of publication June 17, 2019, date of current version July 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923253

New Exon Prediction Techniques Using Adaptive Signal Processing Algorithms for Genomic Analysis

SRINIVASAREDDY PUTLURI[1], MD. ZIA UR RAHMAN[1], (Senior Member, IEEE), CHANDRA SEKHAR AMARA[2], AND NAGIREDDY PUTLURI[2]

[1]Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Guntur 522502, India
[2]Metabolomics Core, Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA

Corresponding author: Md. Zia Ur Rahman (mdzr55@gmail.com)

ABSTRACT Substantial research and monetary aids to healthcare establishments are provided by cloud computing. A benign position to store and handle vast genome data is offered by cloud services. Labs for gene sequencing send out raw and contingent data over the Internet to multiple sequence collections under conservative flow of gene information. The use of cloud services also reduces the storage costs of deoxyribonucleic acid (DNA) sequencing. Here, an efficient and new bio-informatics genomic system is proposed by the use of cloud services from Amazon to access the stored gene data and process it. A key task in bio-informatics is to locate protein-coding sections in a gene sequence based on three base periodicity (TBP) is for disease diagnosis and design drugs. Here, a novel cloud-based adaptive exon predictor (AEP) using Amazon cloud services is proposed to improve the accuracy in exon finding ability as well as aimed at superior convergence. Noise in the input gene sequence given to the proposed AEPs is pre-processed using normalized LMS filtering. Computational complexity can be reduced using proposed data normalized form of least logarithmic absolute difference (NLLAD) algorithm and its error normalized variants. It was shown that sign regressor NLLAD (SRNLLAD) dependent AEP is efficient in exon forecast applications using different metrics for a performance like sensitivity 0.8037, precision 0.8052 along with specificity 0.8146 by different gene sequences considered from the National Center for Biotechnology Information (NCBI) databank. The proposed AEPs have shown upright performance than typical LMS and other AEPs in terms of exon prediction accuracy, convergence, and computational complexity. Their less computational complexity will be found attractive, and they are suitable to use in bio-informatics nano devices.

INDEX TERMS Amazon cloud services, bio-informatics, convergence, deoxyribonucleic acid, National Center for Biotechnology Information, base three periodicity.

I. INTRODUCTION

Cloud Computing is referred as sharing the data, software and resources thru Internet. Cloud provider manages and controls the data on the physical servers for administration of vast amounts of information. Within genomics research, this offers a path for analyzers in order to increase their ability of sharing in addition to storage of information which saves time and cost [1]. This is an ascendable provision for the DNA sequence data to store, process and manage using data banks of larger scale that are available distantly using multiple platforms over the Internet. The gene informatics is moving to the cloud as the DNA sequence analysis is now becoming cost-effective, quicker compared to storing of data and its computation. Ability of cloud computing thru sequencing of subsequent generations which generates unmatched volumes of information for simplifying results will hasten to develop different tools for diagnosis and treat diseases. A novel genome informatics model based on cloud is presented in current work which is used by many establishments related to healthcare for storage and managing of huge volumes of patient's gene sequence data by using cloud services-based Amazon platform. Genomics is the study that comprises the analysis along with sequencing of genomes. In genomics, the scalable service used to process and store gene sequences using virtual and extensive data

The associate editor coordinating the review of this manuscript and approving it for publication was Sunil Karamchandani.

banks which are remotely accessible thru use of internet is Cloud Computing [2]. In genomics, identifying the regions in gene sequences which codes for proteins is an extensive scope of research. A sequence of DNA is an arrangement of genes along with non-protein coding sections [3]. Learning about the principal structure of coding regions for proteins helps their secondary and the tertiary structure. Once the arrangement of exon segments is completely studied, probability of perceiving whole anomalies helps in preparation of drugs and treat ailments [4], [5]. Hence entire alive beings are classified into prokaryotes as well as eukaryotes. Here, long sequences of data can be processed by using adaptive algorithms in multiple iterations which results in developing an Adaptive Exon Predictor (AEP) for locating gene locations in a gene sequence. Here, gene datasets are accessed by using a typical bioinformatics cloud-oriented system proposed in this work.

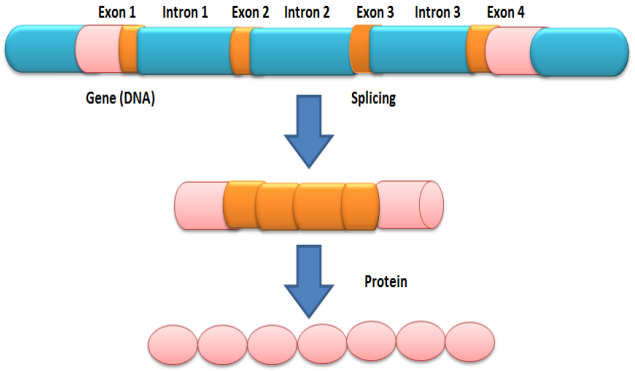

FIGURE 1. Structure of a gene.

To produce beneficial tools at Nano measure using biological blocks, bio-nanotechnology science is used [6], [7]. To treat or prevent diseases, gene therapy is used to directly insert required gene into a specific cell after identifying the gene of interest [8]. For locating the desired gene location in an input genomic sequence, the proposed LLAD based AEPs with low computational complexities can be used in Nano devices. Figure 1 illustrates that exons are the segments that codes for protein in eukaryotes and the segments which codes for non-proteins are introns. In eukaryotes of human, the coded regions are 3% of the progression whereas the residual 97% are non-coded regions. Henceforth, the primary task of DNA sequence is the determination of coding segments. Hence, precisely locating the coding regions is a vital job in bio-informatics.

The property of base three periodicity (TBP) has been exhibited by almost most sequences of DNA. In the PSD plot, at a frequency equal to f1 = 1/3, a sharp peak is clearly depicted [9]. Numerous existing exon locating methods remain several signal processing techniques [10], [11] for locating gene locations in DNA sequences.

In genomic signal processing, to process lengthy sequences, adaptive techniques are used which could alter weight coefficients based on input DNA sequences. A basic adaptive technique which is simple to implement is least mean squares (LMS) technique. Also, the technique undergoes glitches such as drift weight, amplification of gradient noise, and deprived convergence. Hence, to increase the AEP performance, data clipped, error clipped, and data error clipped variants of LMS are used. Adaptive filters based on higher order statistics will perform well in case of LMS algorithm in noise scenarios. In order to examine this, different AEPs are developed to improve exon locating ability and convergence performance using Normalized LLAD (NLLAD) algorithm. The technique which progressively adjusts the conventional cost functions depending on the amount of error is considered. These algorithms are extended by combining with sign algorithms i.e. Sign, Sign Regressor, and Sign-Sign variants are used to reduce the computational complexity. Sign based algorithms needs just half as many multiply operations compared to LMS counter parts which makes them useful in real application point of view.

LMS technique undergoes the difficulty of gradient noise amplification for larger input vector of data. To avoid this problem, normalization of the AEPs can be used for prediction of exons in DNA sequences [13]. With this, adjustment used to the filter weight vector coefficient and its normalization based on input vector squared norm. Here we consider two types of normalizations, namely data normalization and error normalization. They are Normalized LLAD (NLLAD) and Error Normalized LLAD (ENLLAD) algorithms respectively [14], [15]. Here also these algorithms are combined with sign algorithms to ease computational complexity. Normally gene sequences are lengthier in practical applications of bio-informatics and hence it needs lengthy adaptive filters. Thus, LMS technique becomes simple and costly to implement in real case [16]. The computational complexity is reduced largely by using block processing of samples of data [17]. So, maximum variants of adaptive algorithms are considered.

Especially for larger length sequences, the samples overlap each other for larger computational complexities at the input of AEP. This leads to inaccuracy and inter symbol interference (ISI) in prediction. To increase the convergence and stability performance of proposed AEP than LMS algorithm and other AEPs presented in [18], the performance of conventional AEP is improved by using a hybrid version of exon prediction technique in [22]. But the techniques discussed in [22] suffer from a drawback of more complexity in performing the computations. Hence in the contest of development of Nano bio-informatics devices, we intended to use normalized logarithmic based adaptive algorithms.

Difficulties of AEP were prevailed over by the sign depending algorithms and variable normalized adaptive algorithms in practice [20]. Due to normalization, the higher tap length could be minimized to one, by applying an approach named as maximum variable normalization regardless of tap length [21], [22]. In the normalized LMS variant, relation between error and input reference signal is normalized by a value similar to squared norm [23]. The maximum

along with normalized techniques converge quicker compared to conventional LMS technique and also overcome gradient noise application problem [24], [25]. Henceforward error in steady state and rate of convergence of NLLAD are superior compared with LMS [26]. Several AEPs using NLLAD are developed and their performance is measured with standard gene datasets available from National Center for Biotechnology Information (NCBI) gene data base at node 2 with the bio-informatics system based on cloud [27]. We consider sensitivity (Sn), precision (Pr), specificity (Sp), computational complexities and characteristics of convergence as performance metrics to assess the ability of different projected AEPs. Discussion related to performance of several AEPs also theory of adaptive techniques and results of different AEPs is explained in the subsequent sections.

II. TRADITIONAL GENOME BIO-INFORMATICS SYSTEM

Within standard flow of gene data, transmission of raw sequence data thru internet is done by gene sequence laboratories to sequence archives. Casual users either access this data directly or through a website indirectly by value added integrators. Larger sequence datasets are typically downloaded from these archives by power users. Own compute and storage clusters are retained by value added integrators, power users and sequencing archives also their local copies of gene datasets are kept using traditional genomics bioinformatics model shown in Figure 2.

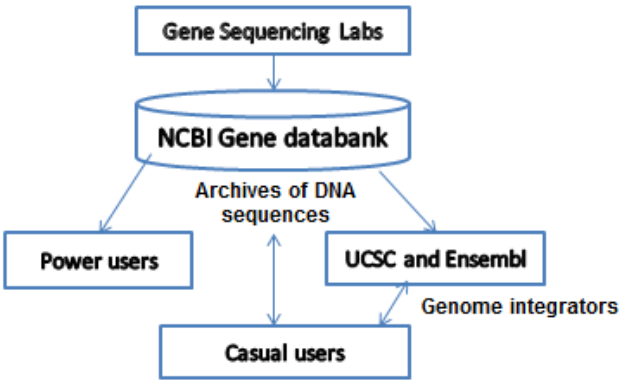

FIGURE 2. Traditional genome bioinformatics system.

From Figure 2, this is clear that the creation and usage of gene information needs an effective and contended eco system for over 25 years [2]. The gene sequencing test centers sends their gene information to large pool of gene databanks like NCBI, DDBJ and EMBL databank of European Bioinformatics Institute which preserves, manage and provides gene data. In traditional genomic informatics system, gene data is accessed by major users either one using value added genome integrators, for instance the Gene Browser University of California on Santa Cruz (UCSC), Ensemble gene databank, casual and power users or through archival databases wherein related websites are produced.

III. PROPOSED GENOME BIO-INFORMATICS SYSTEM

Storage and compute resources of community present in 'cloud' are maintained by large service provider in proposed novel genome bioinformatics system. Challenges of traditional genome informatics such as storing, managing and accessing gene data also cost involved are overcome using the proposed genome bio-informatics cloud-based system shown in Figure 3. Also, the block diagram in Figure 3 presents a detailed analysis of input gene sequences based on dimer nucleotide densities using density plots depicted in Figure 9 followed by mapping of input DNA to binary mapping and pre-processing of DNA sequences is considered. Then the pre-processed sequence is given as input to the AEP for exon prediction.

A new genome bioinformatics cloud-based system with its block diagram is shown in Figure 3. The first step in proposed AEP using bio-informatics cloud-based system in Figure 4 is to get access for cloud services by Researchers also healthcare establishments like hospitals [2]. Secondly the accessed gene sequence from NCBI gene databank from node 2 in the cloud is analyzed based on nucleotide densities and analyzed sequence is converted into digital form. The analyzed sequence is then transformed into binary form and resulting sequence is pre-processed using NLMS algorithm and the output signal after pre-processing is given as AEP input as described in [21]. Here, binary mapping stands as a significant job which is required for denoting alphabetic gene sequence as 4 numeric sequences. The existence of nucleotide is specified as '1' and nonexistence as '0' using conversion process. Accuracy in forecast of genic segments is increased by decreasing noise which is a key in pre-processing [5].

Consider an AEP by new bioinformatics cloud-based system presented using adaptive methods. Consider B(n) as numeric mapped sequence, R(n) be obeyed TBP sequence, K(n) as gene sequence, F(n) signifies a signal for updating the coefficients of weight, O(n) denotes output attained by application of adaptive technique also'' is length of LMS.Succeeding coefficient of weight is anticipated using step size parameter 'Z', the present weight coefficient as v(n), feedback signal F(n) and input gene sequence K(n) at the instance. Analysis along with expression of LMS technique is explained in [14].

The mass update expression of adaptive LMS technique is stated as

$$v(n+1) = v(n) + ZK(n)F(n) \tag{1}$$

In exon identification applications, adaptive algorithms have to possess minimum computational complexity so that they can be attractive for Nano applications. Such reduced value is possible by applying clipping to gene data input otherwise signal of feedback else for both. Techniques for this purpose are demonstrated in [20]. These techniques include three signed variants.

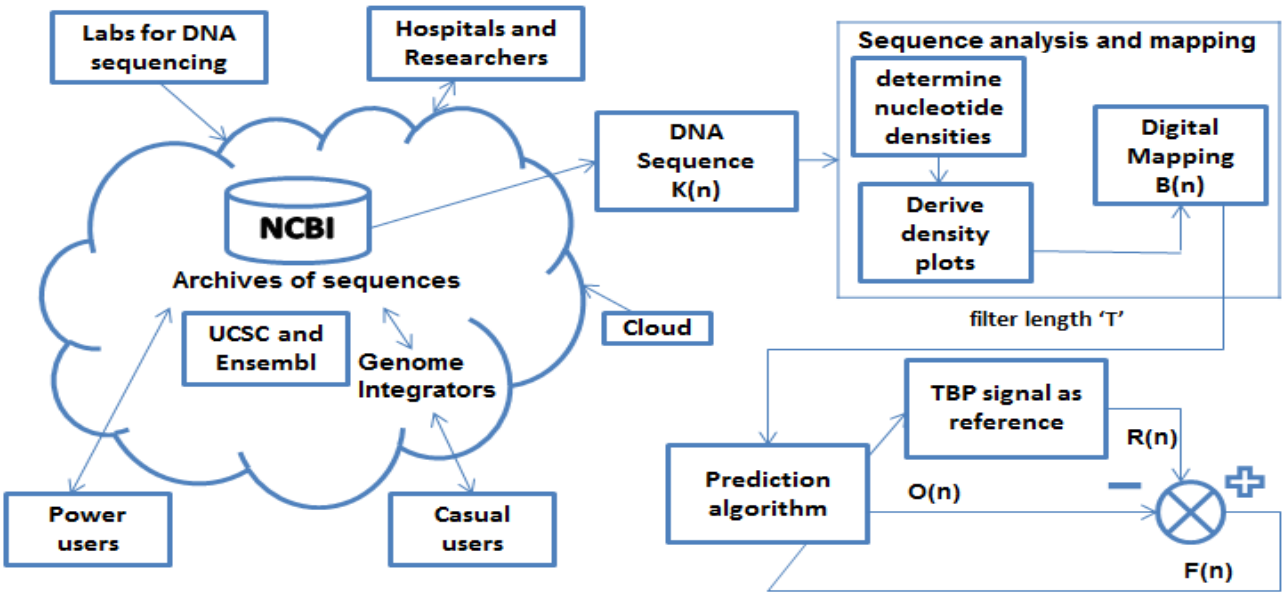

FIGURE 3. Novel cloud-based genome bioinformatics system for exon prediction.

The signum representation is given below: -

$$Q\{K(n)\} = \left\{ \begin{array}{l} 1 : K(n) > 0 \\ 0 : K(n) = 0 \\ -1 : K(n) < 0 \end{array} \right\} \quad (2)$$

These signed variants are preferred to lessen the computational complexity of LMS. LMS algorithm has higher computational complexity compared to these algorithms. Data Clipped LMS (DCLMS) algorithm is derived as recursion of LMS through change of tap input vector.

K(n) is replaced thru means of the vector Q[K(n)], whereas the sign function Q is entered to K(n) upon the element by element basis. Hence, it is referred also as Clipped LMS (CLMS).

The mass update relation of DCLMS algorithm is represented as

$$v(n+1) = v(n) + ZQ\{K(n)\}F(n) \quad (3)$$

The mass renovate expression of ECLMS algorithm is achieved thru replacing F(n) using signed notation as

$$v(n+1) = v(n) + ZK(n)\,Q\{F(n)\} \quad (4)$$

Likewise, mass revise equation of DECLMS is derived via substituting K(n), F(n) using their signed notations as

$$v(n+1) = v(n) + ZQ\{K(n)\}\,Q\{F(n)\} \quad (5)$$

Normalized LMS filter creates a own, namely a small vector of tap input K(n) to overcome the gradient noise amplification of LMS. Numerical problems might arise and hence need to partition little amount of squared norm. In order to overcome this problem, the recursion mentioned above is to be changed by inducing a positive small constant ε. This parameter ε is set to elude divisor as very less also large parameter for step size.

Thus, step size parameter can be expressed as,

$$Z(n) = \frac{Z}{||K(n)||^2} \quad (6)$$

Alternating Z in the expression for LMS with Z(n) tends to Data Normalized LMS (DNLMS) relation expressed as

$$v(n+1) = v(n) + Z(n).K(n).F(n) \quad (7)$$

where Z (n) is a normalized step size.

DNLMS gives minimized error, however in the divisor the squared term increases the MAC operations. Also, complexity along with time for convergence increases. For scaling back amount of calculations, logarithmic based AEPs based on relative logarithmic cost are proposed. A Normalized LLAD algorithm that progressively adjusts the conservative cost functions depending on error amount in its optimization is deliberated for increasing the AEP performance than LMS.

The bound of the step size for mean square convergence of LMS algorithm is:

$$0 < Z < \frac{1}{K^T(n)\,K(n)} \quad (8)$$

Limitations of LMS are overcome by NLLAD also improves convergence speed and exon predicting ability. In order to overcome the weight drift problem connected with LMS, the NLLAD based AEP is presented. The flow chart for NLLAD algorithm used for development of different AEPs is illustrated in Figure 4. Here, B(n) is the desired signal with signal S1 and noise n1 components, reference signal is R(n), filter length as L and weight co-efficient vector w(n) is initialized to 0, I is the iteration parameter and O(i) is

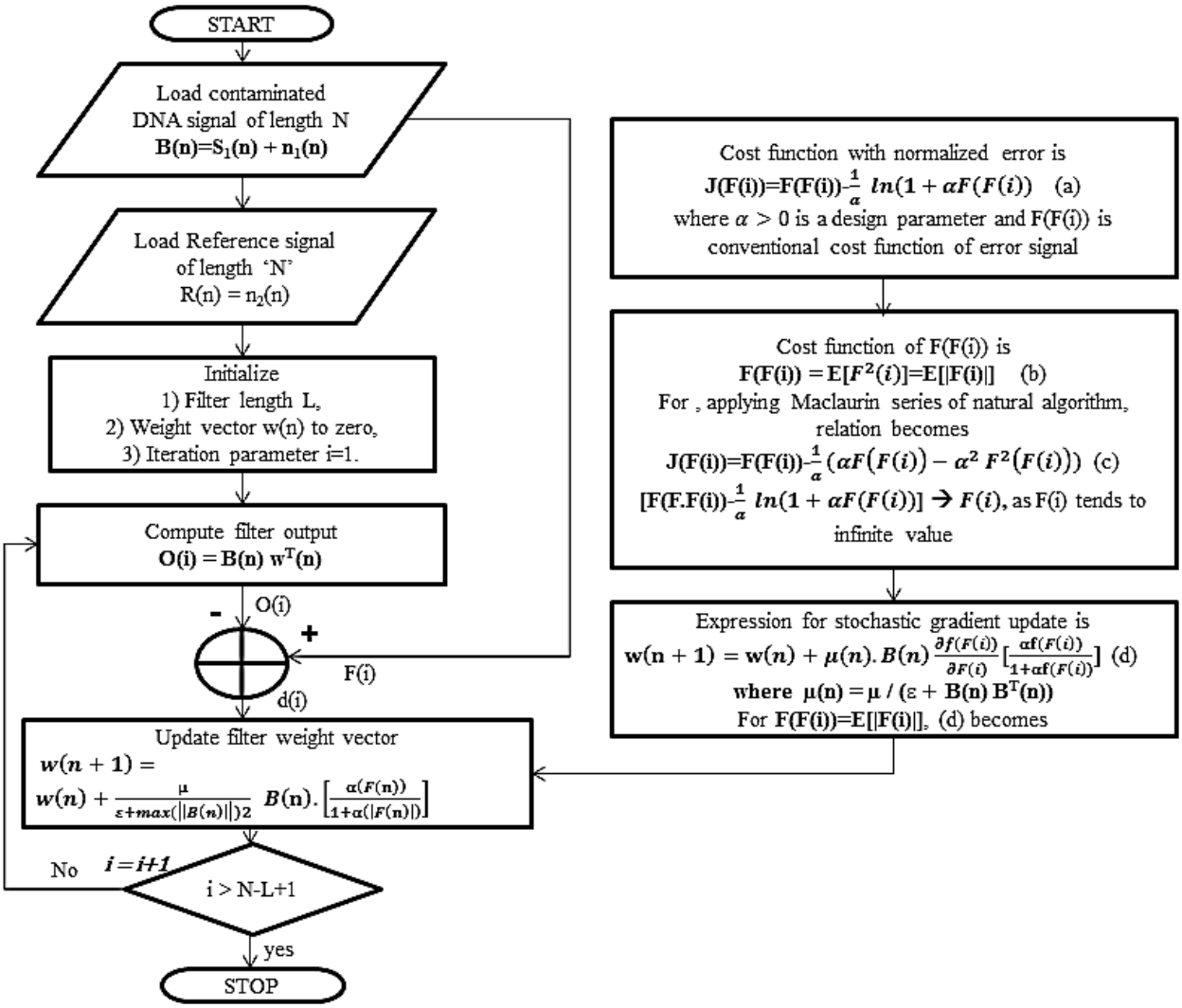

FIGURE 4. Flow chart for proposed NLLAD algorithm.

the Output from the adaptive algorithm. By using the mathematical modeling illustrated in Figure 5, the NLLAD based algorithm was developed for predicting the exon locations in DNA sequence.

The weight update expression of the LLAD technique becomes

$$v(n+1) = v(n) + Z.K(n)\left[\frac{\alpha(F(n))}{1+\alpha(|F(n)|)}\right] \quad (9)$$

Normalized logarithmic variant of LLAD based on relative logarithmic cost is termed as Normalized LLAD algorithm.

Thus, the mass renovate relation of the NLLAD algorithm is given as

$$v(n+1) = v(n) + Z(n).K(n).\left[\frac{\alpha(F(n))}{1+\alpha(|F(n)|)}\right] \quad (10)$$

Replacing $Z(n) = \frac{Z}{\varepsilon+||K(n)||^2}$ in the LLAD weight vector update equation with Z(n) leads to the NLLAD as

$$v(n+1) = v(n) + \frac{Z}{\varepsilon+||K(n)||}^2 K(n).\left[\frac{\alpha(F(n))}{1+\alpha(|D(n)|)}\right] \quad (11)$$

where $Z(n) = \frac{Z}{\varepsilon+||K(n)||}^2$ is a normalized step size with $0 < Z < 2$.

The term α will be either zero or one, based on the value of ε. In case the value of ε is higher compared to the threshold value, then the α will be set to one otherwise it is set to zero, thus reducing the entire numerator to zero and number of calculations reduces. Signed algorithms have minimal convergence speed than DNLMS, but produce a little more error with less complexity in computations. Hence, we combine NLLAD algorithm with sign-based algorithms to lessen complexity involved in computations. The hybrid

versions obtained includes SRNLLAD, SNLLAD and SSNLLAD techniques.

The mass renovate expressions of SRNLLAD, SNLLAD and SSNLLAD algorithms are numerically expressed as,

$$v(n+1) = \frac{Z}{\varepsilon + \max(||K(n)||)^2} \cdot Q[K(n)] \cdot \left[\frac{\alpha(F(n))}{1+\alpha(|F(n)|)}\right] \quad (12)$$

$$v(n+1) = v(n) + \frac{Z}{\varepsilon + \max(||K(n)||)^2} \cdot K(n) \cdot Q\left[\frac{\alpha(F(n))}{1+\alpha(|F(n)|)}\right] \quad (13)$$

$$v(n+1) = v(n) + \frac{Z}{\varepsilon + \max(||K(n)||)^2} Q[K(n)] \cdot Q\left[\frac{\alpha(F(n))}{1+\alpha(|F(n)|)}\right] \quad (14)$$

From expressions (11)-(14), divisor requires 'T' multiplications in normalization. With larger filter length, it requires more amount of multiply operations.

The Error Normalized version of LLAD is called as Error Normalized LLAD (ENLLAD) technique. Here, instead via data instantaneous vector, the normalized error vector with its squared norm is considered. ENLLAD technique performs well in comparison to LMS based on less inaccuracy in its steady state and its rate of convergence.

The mass update expression of ENLLAD algorithm becomes

$$v(n+1) = v(n) + \frac{Z}{\varepsilon + \max(\|F(n)\|)^2} G(n) \cdot \left[\frac{\alpha(F(n))}{1+\alpha(|F(n)|)}\right] \quad (15)$$

where $\|F(n)\|^2 = \sum_{k=0}^{N-1} |F(n-k)|^2$ denotes squared norm of the error F(n), estimated over its complete updated length.

As in (11), ε is added in (15) to prevent instability of the algorithm if $\|F(n)\|^2$ is too small. For reducing the excess multiplications, normalization of step size is done in regards to error amount in ENLLAD. Using the sign regressor based error normalized approach only four multiplications are needed instead of 'T' multiplications. This version is called as error normalized LLAD (ENLLAD) algorithm. The hybrid variants of ENLLAD using sign versions give SRENLLAD, SENLLAD also SSENLLAD techniques.

The mass renovate relations of signed variants of ENLLAD algorithm are given as,

$$v(n+1) = v(n) + \frac{Z}{\varepsilon + \max(\|F(n)\|)^2} Q[K(n)] \cdot \left[\frac{\alpha(F(n))}{1+\alpha(|F(n)|)}\right] \quad (16)$$

$$v(n+1) = v(n) + \frac{Z}{\varepsilon + \max(\|F(n)\|)^2} K(n) \cdot Q\left[\frac{\alpha(F(n))}{1+\alpha(|F(n)|)}\right] \quad (17)$$

$$v(n+1) = v(n) + \frac{Z}{\varepsilon + \max(\|F(n)\|)^2} Q[K(n)] \cdot Q\left[\frac{\alpha(F(n))}{1+\alpha(|F(n)|)}\right] \quad (18)$$

There are certain factors of which the error normalized algorithms for continuing with minimum complexity. They are algorithm's sign capability of good filtering due to the presence of normalized term. Finally, the proposed NLLAD based techniques are successfully applied to actual gene datasets derived as of NCBI databank and shown that they are more precise for gene prediction.

IV. COMPUTATIONAL COMPLEXITIES AND CONVERGENCE ISSUES

For estimating and comparison of complexity of algorithm in general, the number of multiplications required is chosen as a measure. Most DSPs uses hardware to perform MAC computations. Typically, this operation is performed in one instruction cycle besides subtraction or addition. Here, focus is on assessment of several adaptive techniques, rather using exact analysis for complexity to perform computations. Further, these sign dependent techniques are without multiplications, which are required for exon identification applications. In lieu of example, LMS algorithm requires T + 1 MACs for computing the mass update equation. Only one MAC operation is needed for computation of 'S. F(n)' for signed regressor algorithm.

Whereas multiply operations are not needed, in case of other two signed algorithms. NLLAD technique is more difficult with respect to complexity in computations; as it needs T + 4 multiplication operations to implement the mass renovation equation for NLLAD. In case of the SRNLLAD, less complexity in performing computations is presented with 4 multiplications compared to other normalized techniques. Nevertheless, with implementation of maximum normalized techniques, multiplications in divisor are minimized beginning 'T' to '1'. The Computational complexities of the NLLAD algorithm and its error normalized variants are shown in Table 1.

TABLE 1. Multiplication operations required for implementation of various logarithmic AEPs.

S.No.	Algorithm	Multiplications
1	LMS	T+1
2	NLLAD	T+4
3	SRNLLAD	4
4	SNLLAD	T+3
5	SSNLLAD	3
6	ENLLAD	T+4
7	SRENLLAD	4
8	SENLLAD	T+3
9	SSENLLAD	3

When comparing to the remaining normalized algorithms, SRNLLAD algorithm needs minimum number

of computations. In order to manage mutually with the issues like computational complexity and convergence without any restrictive trade-off, the resultant error and sign based normalized logarithmic variants includes Normalized LLAD (NLLAD), Sign Regressor Normalized LLAD (SRNLLAD), Sign Normalized LLAD (SNLLAD), Sign-Sign Normalized LLAD (SSNLLAD), Error Normalized LLAD (ENLLAD), Sign Regressor Error Normalized LLAD (SRENLLAD), Sign Error Normalized LLAD, (SENLLAD), and Sign-Sign Error Normalized LLAD (SSENLLAD) algorithms. The computational complexities of the data normalized LLAD algorithm and its error normalized variants are shown in Table 1.

All these proposed algorithms provide minimum amount of computational complexity owing to use of sign in algorithms with superior capacity of filtering due to the normalized factor. These data normalized LLAD algorithm and its error normalized variants provides upright filtering capability also less complex compared with LMS. These leads to its application at Nano scale such as architecture which is streamlined and aimed at system on chip (SOC) otherwise lab on a chip (LOC). For fabrication of tools required at the Nano scale, Bio nanotechnology can be used to determine the structural elements of cell [7]. Gene therapy has grown colossal interest for researchers due to its ability to replace a gene of interest with a healthy gene and can be very useful for surgery and treating drugs. For locating the desired gene location in an input genomic sequence, the proposed LLAD based AEPs with low computational complexities can be used in Nano devices [8].

The convergence characteristics of proposed data and error normalized LLAD based variants are shown in Figure 5 and Figure 6 respectively. Thus, is evident that all proposed data normalized LLAD adaptive algorithms have a faster convergence rate than LMS and other AEPs.

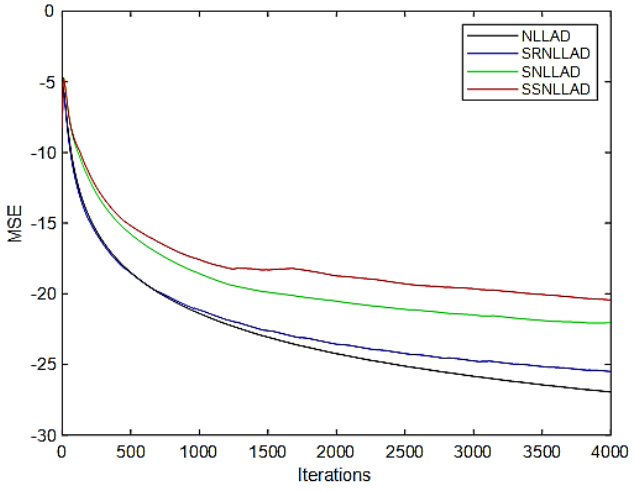

FIGURE 5. Characteristics of convergence for NLLAD with its signed versions.

Hence, among the algorithms considered for the implementation of AEPs, the SRNLLAD adaptive algorithm is considered to be the better, based on the convergence characteristics and computation complexity compared to other normalized algorithms. From convergence characteristics, it was evident that SRNLLAD convergences faster when compared to SNLLAD, and SSNLLAD based AEPs.

V. RESULTS AND DISCUSSION

A. PLATFORM AND INPUT DATA

The proposed model with use of cloud adopts three nodes for computing. Every node needs to be equipped using three core CPU's along with 64 GB of RAM using Xeon X-5550 by Intel of frequency 2.67GHz. The above-mentioned nodes include the gene databases taken from Ensemble, these contains the elucidated genome of homo sapiens also 50 alternative species (annotations of 150 gigabytes added to 100 gigabytes of gene sequence) considered as node 1, virtual machine images along with a comprehensive imprint of NCBI data pairs (100 gigabytes) in the act of node 2, and data sets obtained as of the 1000 Genomes assignment (500 gigabytes) designed as node 3. The genome data from various gene databases are represented at different nodes is as presented in [22]. Every node is associated along with 1 Gigabit Ethernet.

In our proposed work, node 2 generates the input genomic sequence with the help of Amazon cloud services in FASTA (fast A) file pattern. In our work, a list of ten gene sequences obtained from National Centre for Biotechnology Information is considered in [21]. Due to space constraint the experimental results for the sequence AF009962 are shown in this paper.

B. TASK DISTRIBUTION AND PERFORMANCE

In current work, gene data input datasets from gene databank of NCBI at node 2 are used. The distribution of task is done depending on location of gene sequence input at one of three presented nodes. Whole three nodes can be accessed as virtual machine images. Hence, an account needs to setup by use of web services by Amazon, then unveil an occurrence from existing three images focused on bio-informatics also one image among the available three is to be attached for genomic data processing. Distribution of tasks is done from the node where data input sequence is to be chosen. The NCBI databank as of node 2 is considered and access the input genomic sequence using Amazon cloud services consuming minimum time when compared to the original method of data access. The location of exon is predicted by giving the pre-processed genomic sequence as an input to AEPs.

C. PRE-PROCESSING OF GENOMIC SEQUENCE

Because of period-3 behavior exhibited by exon segments and distinct behavior of DNA made it suitable to use signal processing techniques for analysis. The output gene signal after the mapping consists of noise as presented in [6]. This signal has been pre-processed using NLMS and the pre-processed output is given as input to the proposed AEPs.

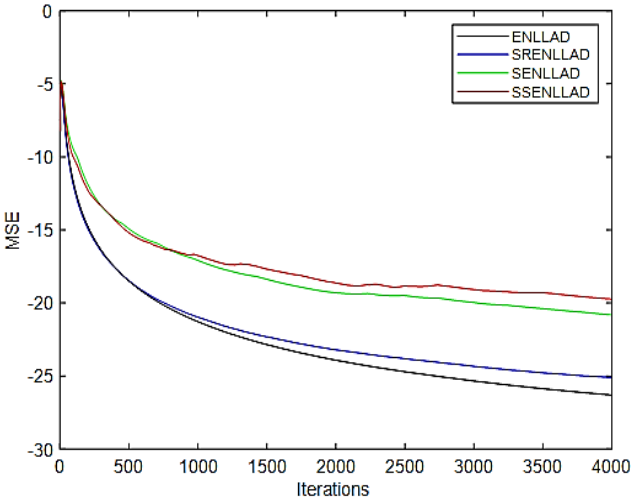

FIGURE 6. Characteristics of convergence for ENLLAD with its signed versions.

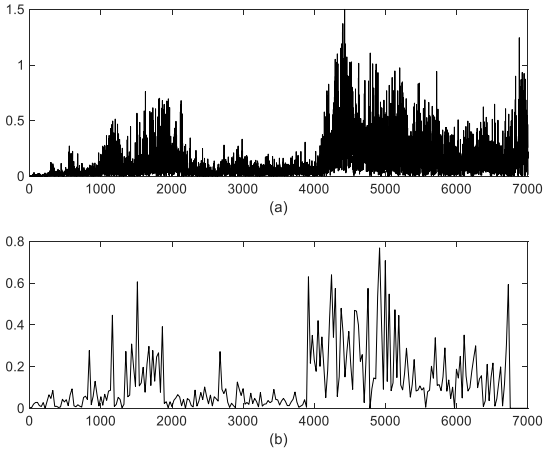

FIGURE 7. Pre-processing of digital genomic sequence using NLMS. a) Pattern of genomic sequence with noise. b) Enhanced genomic sequence using NLMS based adaptive noise canceller.

Figure 7 shows the pre-processing of digital gene sequence using NLMS algorithm. We have used the frame work of the adaptive noise cancellation presented in [24]. The experimental results of the gene sequence with Accession AF009962 are shown in this paper. To evaluate the performance of the enhancement process we have considered three basic performance measures, namely signal to noise ratio improvement (SNRI), mis-adjustment (MSD) and excess mean square error (EMSE) were shown in Table 2.

D. EXON PREDICTION RESULTS AND DISCUSSION

The discussion is regarding the comparison of performances of variety of AEPs. Figure 4 shows the outline of AEP. The maximum data normalized LMS algorithms along with sign variants are used to derive several AEPs. In order to compare, we have also developed an LMS depending AEP. For the purpose of evaluation, ten DNA sequence datasets from the database of NCBI are obtained [25]. It is carried out in order to provide a consistency in results and increase the output of variety of algorithms under consideration as shown in Table 3. The output performance is calculated by considering parameters such as Specificity (Sp), Sensitivity (Sn) and Precision (Pr) in [18], [24] demonstrates the expressions and the theory of the above-mentioned parameters. Part of locating the exon segments with use of DSP methods, few measures are definite based on change of level of threshold in output spectrum which are used for comparison in this paper. Number of nucleotides appropriately located as introns in exon identification step is defined as true negative (TN), while those properly identified as exons is indicated as true positive (TP). Likewise, a total of exon segments located as intron nucleotides is defined as false negative (FN), whereas those quantity of introns truly predicted as exon nucleotide is measured as false positive (FP). Thus, expressions for performance measures such as the specificity, sensitivity, as well as precision were written as: -

$$\begin{aligned} \mathrm{Sn} &= \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \\ \mathrm{Sp} &= \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \\ \mathrm{Pr} &= (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}) \end{aligned}$$

Quantity of exons located which are actually present in exon sections is known as Specificity (Sp), whereas the amount of exon regions those are appropriately forecasted as exons is measured as Sensitivity (Sn). The results of exon prediction of sequence 5 using data normalized LLAD algorithms are shown in Figure 10, whereas for error normalized LLAD algorithms are in Figure 11.

The Threshold values are taken between 0.4 and 0.9 at an interval of 0.05. By using these values, the performance of Pr, Sn, and Sp are measured. The prediction of exon is precise at the threshold value of 0.8. Therefore, Table 3 illustrates the performance measures at the value of threshold 0.8. Sections of a DNA sequence with a high percent of A + T nucleotides usually indicate intergenic parts of the sequence, while low A + T and higher G + C nucleotide percentages indicate possible genes. Many times, high CG dinucleotide content is located before a gene.

The sequence statistics functions are useful to determine if input gene sequence has the characteristics of a protein-coding region. For instance, Figure 9 depicts the typical plot for dimer nucleotide densities for the nucleotide sequence with accession AF009962.The dimers distribution in a gene sequence with accession AF009962 is depicted in the form of bar illustration using MATLAB software. From Figure 3, it was shown that dimers using T-T base pairs are more when compared with all other dimers in the gene sequence with accession AF009962. There are 680 dimers of T-T base pairs in the considered gene sequence. In this gene sequence, there are 527 A-T dimers and 70 G-C dimers. In the considered gene sequence, G + C content is less compared to A + T dimers indicates that it has a smaller number of genes.

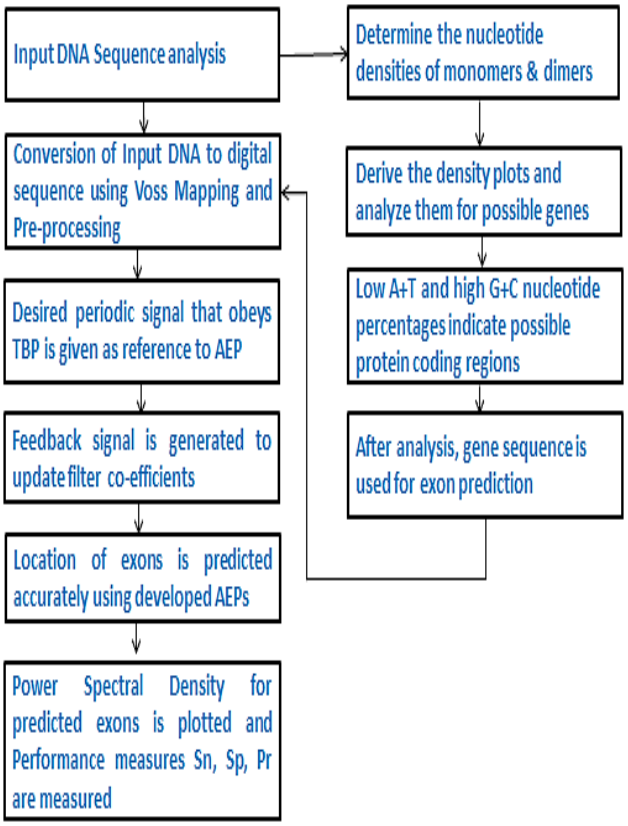

FIGURE 8. Flow diagram for adaptive exon prediction in an input DNA sequence using the proposed model.

The flow diagram with steps in adaptive exon prediction shown in Figure 8 is explained below:

a) Novel method of cloud dependent genome bio-informatics is used to pick an input DNA sequence obtained from Node 2 in database of NCBI and to load the virtual image present in Node. We have analyzed the input DNA based on nucleotide densities of A + T and G + C base pairs using density plots shown in Figure 6 to determine the presence of gene locations. This sequence after analysis is then transformed to binary form using binary mapping methodology. Where the AEP input is the binary data obtained as illustrated in Figure 5.

b) The resulting signal is now pre-processed by applying Data Normalized LMS (NLMS) algorithm to remove the noise before giving it as input to the proposed AEPs as shown in Figure 3. Three base periodicity (TBP) obeyed biological sequence is given as reference signal to the proposed NLLAD based AEPs.

c) From Figure 3, a signal for feedback signal F(n) derived was used for updating coefficients of filter.

d) When this signal becomes minimum, genic regions are accurately located from the DNA sequence using PSD plot.

e) The plots for predicted coding regions remain shown with help of PSD. Also, measures Sn, Pr as well as Sp are derived and plotted.

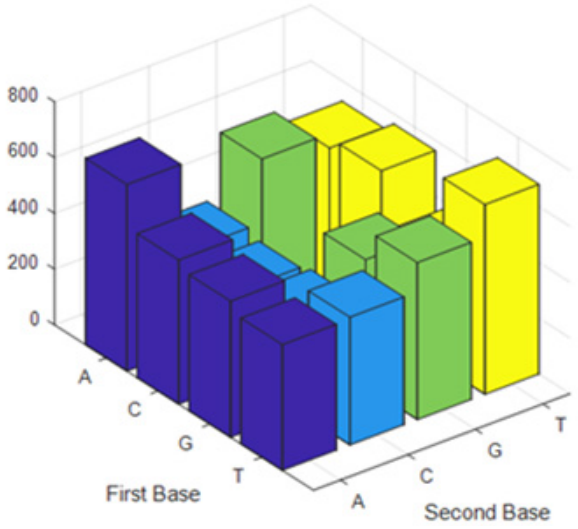

FIGURE 9. Nucleotide density plot of dimers for gene sequence with accession AF009962.

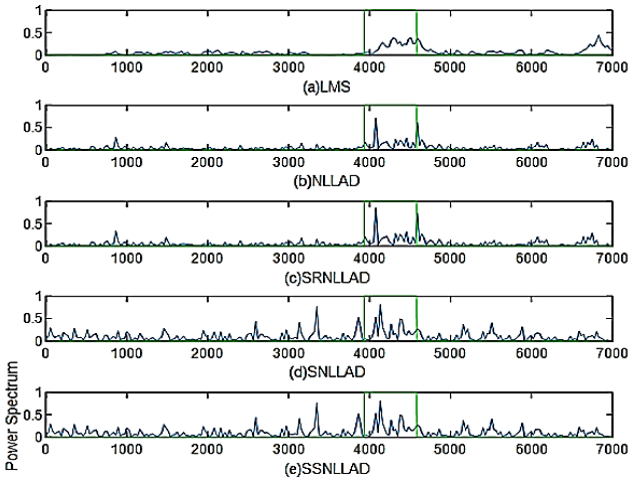

FIGURE 10. PSD with the exon position (3934-4581) predicted for gene sequence with accession AF009962 with several AEPs, (a). AEP with LMS, (b). AEP with NLLAD, (c). AEP with SRNLLAD, (d). AEP with SNLLAD, (e). AEP with SSNLLAD.

Figure 10 and Figure 11 show the predicted exon locations of sequence 5 by applying various adaptive algorithms. Table 2 deliberates the dataset of genome taken from the database of NCBI gene database [28]. It is clear from the table that there is no accurate prediction of coded regions by the LMS dependent AEP. Here, some non-coded regions are also predicted which leads to uncertainty in exon location prediction.

In Figure 10 (a) have certain undesirable peaks were recognized at 1200th, 2300th and 3200th sample values with LMS. On the identical occasion, the exact exon location 3934-4581 is not identified. In case of data normalized LLAD versions, it was observed that the NLLAD, SRNLLAD, SNLLAD and SSNLLAD algorithms precisely identified the location of exon at 3934-4581 with higher

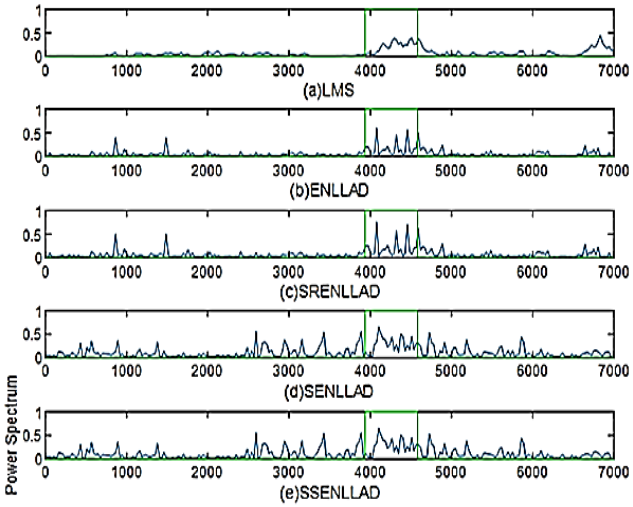

FIGURE 11. PSD with the exon position (3934-4581) predicted for gene sequence with accession AF009962 with several AEPs, (a). AEP with LMS, (b). AEP with ENLLAD, (c). AEP with SRENLLAD, (d). AEP with SENLLAD, (e). AEP with SSENLLAD.

TABLE 2. Performance measures obtained in Pre-processing using NLMS.

S.No.	Seq. No.	SNRI (dbs)	EMSE (dbs)	MSD
1	E15270.1	10.1864	-23.2863	0.284
2	X77471.1	10.2932	-24.5825	0.279
3	AB035346.2	10.2857	-23.3862	0.268
4	AJ225085.1	10.2786	-23.4623	0.273
5	AF009962	10.2875	-24.5254	0.272
6	X59065.1	10.3864	-22.9437	0.284
7	AJ223321.1	10.2972	-24.2561	0.277
8	X92412.1	10.2881	-22.9235	0.287
9	U01317.1	10.3154	-23.1353	0.279
10	X51502.1	10.2968	-24.4618	0.27

intensity as depicted in the Power Spectral Density (PSD). Figures 10 (b), (c), (d) and (e) show their PSD plots. Similarly, in case of error normalized LLAD versions, it was observed that the ENLLAD, SRENLLAD, SENLLAD and SSENLLAD algorithms precisely identified the location of exon at 3934-4581 with higher intensity as depicted in the PSD. Figures 11 (b), (c), (d) and (e) show their PSD plots. The plot for performance measures like Sensitivity, Specificity and Precision use data and error normalized LLAD algorithms is shown in Figure 12. It was shown that data and error normalized LLAD versions are far superior to existing LMS method in terms of performance measures shown in Figure 12. Prediction measurements like specificity, precision and sensitivity of NLLAD, SRNLLAD, SNLLAD

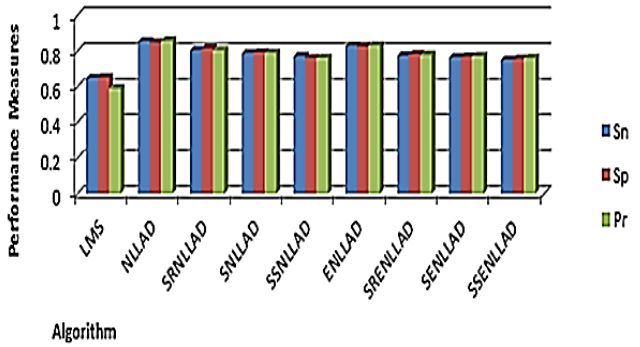

FIGURE 12. Performance Measures of gene sequence with accession AF009962 using proposed AEPs (i). AEP built on LMS, (ii). AEP built on NLLAD, (iii). AEP built on SRNLLAD, (iv). AEP built on SNLLAD, (v). AEP built on SSNLLAD, (vi). AEP built on ENLLAD, (vii). AEP built on SRENLLAD, (viii). AEP built on SENLLAD, (ix). AEP built on SSENLLAD.

and SSNLLAD techniques viewed better compared to LMS whereas much superior in SRNLLAD algorithm.

The tracking abilities of proposed normalized techniques are higher in comparison with LMS algorithm and also better compared with signed regressor form of VNLMS algorithm in [22] which performed better among normalized VSLMS variants. In Table 3, the results obtained using VNSRLMS are compared with all the proposed AEPs including LMS based AEP. Higher the values of Sn, Sp and Pr, the better the accuracy in prediction of exon locations in an input DNA sequence. However, it also depends on the computational complexity and convergence performance of adaptive algorithm.

It's always better to choose an AEP developed using adaptive algorithm which offers low computational complexity and better convergence, so that they can be used in real applications. For instance, in table 3, consider values of performance measures Sn, Sp and Pr for results analysis and discussion. Here, the values obtained using VNSRLMS are Sn as 0.7884 (78.84%), Sp as 0.7812 (78.12%) and Pr as 0.7936 (79.36%), which are less accurate in prediction accuracy when compared to the proposed SRNLLAD based AEP. Using SRNLLAD based AEP, the measures obtained are Sn as 0.8037 (80.37%), Sp as 0.8146 (81.46%) and Pr as 0.8052 (80.52%). These values are just inferior to the values obtained using NLLAD based AEP. But, due to a smaller number of computations required for SRNLLAD based AEP, it is a better candidate for exon prediction applications. Also, the values obtained using SNLLAD and SSNLLAD based AEPs are inferior in comparison with NLLAD and SRNLLAD based variants. Similarly, the values of ENLLAD variants are derived and analyzed for their performance are presented in Table 3.

From these, the values obtained using SRENLLAD are Sn as 0.7734 (77.34%), Sp as 0.7825 (78.25%) and Pr as 0.7783 (77.83%), which are less accurate in prediction accuracy when compared to the proposed SRNLLAD based AEP. These values are just inferior compared to ENLLAD based AEP, whereas the values for SENLLAD and SSENLLAD

TABLE 3. Metrics for performance of various adaptive exon prediction techniques in terms of sensitivity, specificity and precision.

Sq. No.	Metric	LMS	VNSRLMS	NLLAD	SRNLLAD	SNLLAD	SSNLLAD	ENLLAD	SRENLLAD	SENLLAD	SSENLLAD
1	Sn	0.6292	0.7972	0.8523	0.8012	0.7862	0.7694	0.8296	0.7732	0.7654	0.7503
	Sp	0.6467	0.7836	0.8424	0.8124	0.7902	0.7586	0.8235	0.7811	0.7695	0.7538
	Pr	0.5984	0.7783	0.8506	0.8096	0.7896	0.7612	0.8311	0.7796	0.7705	0.7592
2	Sn	0.6373	0.7835	0.8553	0.8102	0.7911	0.7708	0.8292	0.7724	0.7642	0.7552
	Sp	0.6586	0.7841	0.8462	0.8164	0.7895	0.7611	0.8273	0.7863	0.7645	0.7567
	Pr	0.5912	0.7924	0.8531	0.8196	0.7934	0.7586	0.8328	0.7785	0.7783	0.7525
3	Sn	0.6494	0.7882	0.8534	0.8079	0.7891	0.7686	0.8285	0.7795	0.7688	0.7595
	Sp	0.6608	0.7936	0.8464	0.8114	0.7936	0.7592	0.8274	0.7835	0.7644	0.7576
	Pr	0.5985	0.7823	0.8561	0.8097	0.7892	0.7618	0.8335	0.7737	0.7736	0.7593
4	Sn	0.6315	0.7936	0.8618	0.8094	0.7868	0.7697	0.8238	0.7736	0.7685	0.7545
	Sp	0.6428	0.7835	0.8514	0.8138	0.7995	0.7535	0.8297	0.7876	0.7697	0.7564
	Pr	0.5793	0.7941	0.8538	0.8068	0.7894	0.7697	0.8378	0.7778	0.7736	0.7582
5	Sn	0.6487	0.7884	0.8537	0.8037	0.7878	0.7705	0.8282	0.7734	0.7651	0.7522
	Sp	0.6546	0.7812	0.8468	0.8146	0.7931	0.7587	0.8246	0.7825	0.7685	0.7554
	Pr	0.5932	0.7936	0.8578	0.8052	0.7892	0.7612	0.8311	0.7783	0.7718	0.7608
6	Sn	0.6186	0.7974	0.8445	0.8125	0.7954	0.7596	0.8263	0.7846	0.7658	0.7582
	Sp	0.6378	0.7872	0.8575	0.8086	0.7878	0.7632	0.8323	0.7752	0.7746	0.7598
	Pr	0.5821	0.7793	0.8643	0.8102	0.7872	0.7704	0.8245	0.7765	0.7682	0.7562
7	Sn	0.6216	0.7892	0.8532	0.8142	0.7988	0.7546	0.8288	0.7882	0.7702	0.7572
	Sp	0.6575	0.7894	0.8542	0.8074	0.7874	0.7632	0.8365	0.7792	0.7754	0.7602
	Pr	0.5911	0.7966	0.8531	0.8029	0.7868	0.7714	0.8302	0.7746	0.7696	0.7714
8	Sn	0.6305	0.7845	0.8582	0.8054	0.7882	0.7732	0.8265	0.7751	0.7672	0.7528
	Sp	0.6296	0.7902	0.8476	0.8106	0.7976	0.7594	0.8252	0.7848	0.7696	0.7562
	Pr	0.5902	0.7864	0.8586	0.8124	0.7862	0.7676	0.8347	0.7783	0.7724	0.7624
9	Sn	0.6294	0.7982	0.8548	0.8075	0.7882	0.7715	0.8294	0.7756	0.7656	0.7542
	Sp	0.6485	0.7876	0.8472	0.8154	0.7954	0.7592	0.8262	0.7875	0.7684	0.7584
	Pr	0.5692	0.7838	0.8565	0.8086	0.7895	0.7635	0.8378	0.7792	0.7735	0.7627
10	Sn	0.6251	0.7947	0.8572	0.8054	0.7892	0.7718	0.8294	0.7758	0.7698	0.7545
	Sp	0.6493	0.7836	0.8486	0.8187	0.7964	0.7592	0.8265	0.7864	0.7686	0.7568
	Pr	0.5822	0.7978	0.8593	0.8076	0.7914	0.7684	0.8342	0.7808	0.7712	0.7674

*'Sq. No.' is the Gene Sequence Serial Number

based variants are inferior than both the ENLLAD and SRENLLAD based AEPs. The simulated values of performance measures using data and error normalized LLAD based AEPs are tabulated in Table 3. Among ENLLAD based variants, SRENLLAD based AEP is a better AEP in terms of exon locating ability due to its low computational complexity and convergence performance compared to other ENLLAD based AEPs. Overall, among these eight proposed NLLAD based AEPs, SRNLLAD possess higher convergence capabilities and low computational complexity. This SRNLLAD based AEP requires less amount of multiply operations independent of its length of the tap vector. The characteristics of convergence for SRNLLAD is significant compared to its other variants, however metrics considered are a little inferior than NLLAD, SNLLAD, also SSNLLAD algorithms. The simulated values of performance measures using data and

error normalized LLAD algorithms are tabulated in Table 3. The exon prediction output is greater when compared to LMS along with remaining standard signed versions. Hence, with respect to complexity in computations, rate of convergence, exon plots, calculation of Sp, Sn, and Pr measures, SRNLLAD based AEP was shown to be superior in practice for locating exon segments in an input gene sequence. Finally, the entire AEPs in proposed methodology are more efficient compared to the existing LMS technique to find the regions of exons in genomic sequences.

VI. CONCLUSION

In proposed paper, the problem based on exon location identification in a gene sequence is demonstrated. Here, a new adaptive exon identification methodology is proposed. In order to fulfill the above-mentioned problem, ''Virtual Machines'' with use of cloud services by Amazon with Virtual Hard Disks that are custom designed are proposed. In order to store and access the information of genome database from NCBI node using windows platform, logarithmic based normalized adaptive algorithms are taken into consideration to process various DNA sequences. This is clear from metrics in Table 3 and PSD plots for exon positions as shown in Figure 10 and Figure 11. The presented AEPs precisely predicted the position of exon at 3934-4581 thru great intensity from plot for PSD. Overall SRNLLAD delivers greater performance pertaining to computational complexity and the measures obtained are Sn as 0.8037 (80.37%), Sp as 0.8146 (81.46%) and Pr as 0.8052 (80.52%) attained using gene sequence 5 with accession as AF009962 with a value of threshold equal to 0.8. These values are just inferior to values obtained using NLLAD based AEP. But, due to its less computational complexity and better convergence performance, it is a better candidate for exon prediction applications. Therefore, the AEP based on SRNLLAD seems to be better than the other counter algorithms in this family. Hence, AEPs based on SRNLLAD could be used in the development of Nano bioinformatics devices in SOC and LOC applications.

REFERENCES

[1] K. Charlebois, N. Palmour, and B. M. Knoppers, ''The adoption of cloud computing in the field of genomics research: The influence of ethical and legal issues,'' *PLoS ONE*, vol. 11, no. 10, 2016, Art. no. e0164347.

[2] L. D. Stein, ''The case for cloud computing in genome informatics,'' *Genome Biol.*, vol. 11, p. 207, May 2010.

[3] L. W. Ning, H. Lin, H. Ding, J. Huang, N. Rao, and F. B. Guo, ''Predicting bacterial essential genes using only sequence composition information,'' *Genet. Mol. Res.*, vol. 13, pp. 4564–4572, Jan. 2014.

[4] J. E. Dickerson, A. Zhu, D. L. Robertson, and K. E. Hentges, ''Defining the role of essential genes in human disease,'' *PLoS ONE*, vol. 6, no. 11, Nov. 2011, Art. no. e27368.

[5] V. Tomar, D. Gandhi, and C. V. Kumar, ''Advanced filters for genomic signal processing,'' in *Proc. IEEE TENCON*, Hyderabad, India, Nov. 2008. doi: 10.1109/TENCON.2008.4766648.

[6] O. Abbasi, A. Rostami, and G. Karimian, ''Identification of exonic regions in DNA sequences using cross-correlation and noise suppression by discrete wavelet transform,'' *BMC Bioinf.*, vol. 12, p. 430, Nov. 2011.

[7] T. M. Sun, Y. C. Wang, F. Wang, J. Z. Du, C. Q. Mao, C. Y. Sun, R. Z. Tang, Y. Liu, J. Zhu, Y. H. Zhu, X. Z. Yang, and J. Wang, ''Cancer stem cell therapy using doxorubicin conjugated to gold nanoparticles via hydrazone bonds,'' *Biomaterials*, vol. 35, no. 2, pp. 836–845, Jan. 2014.

[8] S. Massadeh, M. Al-Aamery, S. Bawazeer, O. AlAhmad, R. AlSubai, S. Barker, and D. Craig, ''Nano-materials for gene therapy: An efficient way in overcoming challenges of gene delivery,'' *J. Biosensors Bioelectron.*, vol. 7, no. 5, p. 94, May 2017.

[9] M. Li, Q. Li, G. U. Ganegoda, J. Wang, F. Wu, Y. Pan ''Prioritization of orphan disease-causing genes using topological feature and GO similarity between proteins in interaction networks,'' *Sci. China Life Sci.*, vol. 57, no. 11, pp. 1064–1071, Nov. 2014.

[10] S. T. Cole, ''Comparative mycobacterial genomics as a tool for drug target and antigen discovery,'' *Eur. Respiratory J.*, vol. 20, pp. 78s–86s, Jan. 2002.

[11] S. Maji and D. Garg, ''Progress in gene prediction: Principles and challenges,'' *Current Bioinf.*, vol. 8, no. 2, pp. 226–243, 2013.

[12] H. Saberkari, M. Shamsi, H. Heravi, and M. H. Sedaaghi, ''A novel fast algorithm for exon prediction in eukaryotic genes using linear predictive coding model and goertzel algorithm based on the Z-curve,'' *Int. J. Comput. Appl.*, vol. 67, no. 17, pp. 25–38, Jan. 2013.

[13] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, ''Prediction of probable genes by Fourier analysis of genomic sequences,'' *Comput. Appl. Biosci.*, vol. 13, no. 3, pp. 263–270, Jun. 1997.

[14] W. M. Ismail, Y. Ye, and H. Tang, ''Gene finding in metatranscriptomic sequences,'' *BMC Bioinf.*, vol. 15, p. 8, Sep. 2014.

[15] R. F. Voss, ''Evolution of long-range fractal correlations and 1/f noise in DNA base sequences,'' *Phys. Rev. Lett.*, vol. 68, no. 25, pp. 3805–3808, Jun. 1992.

[16] G. Liu and Y. Luan, ''Identification of protein coding regions in the eukaryotic DNA sequences based on marple algorithm and wavelet packets transform,'' *Abstract Appl. Anal.*, vol. 2014, Jul. 2014, Art. no. 402567.

[17] G. Mahin and K. Hamed, ''Bioinformatics approaches for gene finding,'' *Int. J. Sci. Res. Sci. Technol.*, vol. 1, pp. 12–15, Sep. 2015.

[18] S. R. Putluri and M. Z. Rahman, ''Identification of protein coding region in dna sequence using novel adaptive exon predictor,'' *J. Sci. Ind. Res.*, vol. 77, pp. 87–91, Feb. 2018.

[19] Y. Azuma and S. Onami, ''Automatic cell identification in the unique system of invariant embryogenesis in caenorhabditis elegans,'' *Biomed. Eng. Lett.*, vol. 4, no. 4, pp. 328–337, Dec. 2014.

[20] P. Srinivasareddy, M. Z. U. Rahman, ''Computer based genomic sequences analysis using least mean forth adaptive algorithms,'' *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 9, pp. 2006–2014, May 2017.

[21] S. Putluri and M. Z. U. Rahman, ''New adaptive exon predictors for identifying protein coding regions In DNA sequence,'' *ARPN J. Eng. Appl. Sci.*, vol. 11, no. 23, pp. 13540–13549, Dec. 2016.

[22] S. Putluri, M. Z. U. Rahman, and S. Y. Fathima, ''Cloud-based adaptive exon prediction for DNA analysis,'' *IET Healthcare Technol. Lett.*, vol. 5, no. 1, pp. 25–30, Feb. 2018.

[23] M. Z. U. Rahman, G. V. S. Karthik, S. Y. Fathima, A. Lay-Ekukaille, ''An efficient cardiac signal enhancement using time–frequency realization of leaky adaptive noise cancelers for remote health monitoring systems,'' *Measurements*, vol. 46, no. 10, pp. 3815–3835, Dec. 2013.

[24] M. Z. U. Rahman, R. A. Shaik, and D. V. R. K. Reddy, ''Efficient and simplified adaptive noise cancelers for ECG sensor based remote health monitoring,'' *IEEE Sensors J.*, vol. 12, no. 3, pp. 566–573, Mar. 2012.

[25] S. O. Haykin, *Adaptive Filter Theory*, 5th ed. London, U.K.: Pearson, 2014.

[26] M. O. Sayin, N. D. Vanli, and S. S. Kozat, ''A novel family of adaptive filtering algorithms based on the logarithmic cost,'' *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4411–4424, Sep. 2014.

[27] *National Center for Biotechnology Information*. Accessed: Jan. 10, 2019. [Online]. Available: www.ncbi.nlm.nih.gov/

SRINIVASAREDDY PUTLURI is currently an Engineer with Tata Consultancy Services Ltd., Hyderabad. He is currently pursuing the Ph.D. degree with the Department of Electronics and Communication Engineering, Koneru Lakshmaiah Educational Foundation named as K. L. University, Guntur, India. His research interests include genomic signal processing and adaptive signal processing.

MD. ZIA UR RAHMAN (M'09–SM'16) received the M.Tech. and Ph.D. degrees from Andhra University, Visakhapatnam, India. He is currently a Professor with the Department of Electronics and Communication Engineering, Koneru Lakshmaiah Educational Foundation (K. L. University), Guntur, India. His current research interests include adaptive signal processing, biomedical signal processing, array signal processing, MEMS, and nano photonics. He has published over 100 research papers in various journals and proceedings. He is currently serving on various editorial boards as the Editor-in-Chief, an Associate Editor, and a Reviewer of publishers like the IEEE, Elsevier, Springer, IGI, American Scientific Publishers, and Hindawai.

CHANDRA SEKHAR AMARA is a Postdoctoral Associate with the Department of Molecular and Cellular Biology, Baylor College of Medicine, TX, USA. He is an active Researcher, and his work is focused in the area of metabolomics profiling of tumors to obtain biochemical insights into cancer progression and develop non-invasive metabolic markers for the detection and the prognosis of tumors. His research interests include proteomics and metabolomics.

NAGIREDDY PUTLURI received the Ph.D. degree in analytical chemistry-mass spectrometry, metabolomics, and proteomics from the Indian Institute of Chemical Technology, Hyderabad, in 2007. He is currently the Director of the Metabolomics Core, Baylor College of Medicine, TX, USA. He has received a Postdoctoral Fellowship on metabolomics with Medical College of Georgia, in 2009. He has published over 60 publications in journals pertinent to metabolomics and metabolomics flux, and in high impact papers such as *Cell*, in 2014, *Science*, in 2018, *Nature*, in 2017 and 2018, the *Journal of Clinical investigations* in 2014, 2015, and 2017, *Cell Metabolism*, in 2017, and *Cancer Research*, in 2011 and 2016. His research interests include mass spectrometry, metabolomics, and proteomics.