An Framework for Simple Sequence Repeat Reduction with Information Extraction using Machine Learning

Mukesh Perumala
I/c-AKMU-Formerly ARIS Cell
ICAR - Indian Institute of Millets Research (IIMR), GOI,
Rajendranagar, Hyderabad 500030,TS, India
pmukesh29@gmail.com

Vasumathi D
Computer Science and Engineering
Jawaharlal Nehru Technological University (JNTU), Kukutpally
Hyderabad-India in 2011
rachanu@gmail.com

Syeda Sameen Fathima
(Rtd) Professor & Head
Dept Of Computer Science & Engineering, College of Engineering,
Osmania University, Hyderabad
sameenf@gmail.com

Abstract—**Demand for the agricultural improvements using the advanced computer algorithms have increased in the recent years. The primary focus is on the higher crop production rates with least damage to the crops due to various diseases. In the recent times, a good number of research attempts are observed to formulate multiple computerized algorithms to identify the Amino Acid sequence and further protein sequences, which are responsible for diseases to the plants and crops. However, due to the higher complexity of DNA structure and further the complex process for DNA to Amino Acid extraction, these recent researches have produced unsatisfactory outcomes. Henceforth, in order to solve the primary challenge of higher time complexity of the DNA processing methods, this work proposes two algorithms to reduce the DNA sequence length without losing vital information using machine learning. Firstly, the use of clustering method to reduce the size ensures least information loss and best processing time. Secondly, the look up based indexed Amino Acid extraction process ensures higher correctness of the extraction and again in best possible time. The proposed framework produced nearly 98% accuracy in 0.107 sec time frame, which is relatively 5% improvement in accuracy and 10% improvement in time complexity.**

Keywords— Sequence Reduction, K-Means Clustering, Sequence Benchmarking, Amino Acid Information, Indexed LookUp

I. INTRODUCTION

"Simple sequence repeats" (SSRs) are a collection of short and repeated oligonucleotide sequences found mostly in between long DNA sequences, according to S. Mondal et al. [1] These short oligonucleotides differ structurally and functionally from normal DNA sequences. An SSR-based parallel processing paradigm for fundamental DNA sequence operations using k-mers is described in this study. In the initial phase of this effort, k-mer is used to identify SSRs. As a result, this study utilises the MapReduce-based K-mer method in order to benefit from parallel execution and the distributed platform. A pipelined K-mer toolset known as KAnalyze was compared to the results of this study and found a considerable improvement in computing time. MapReduce-based k-mer has the ability to minimise memory footprint and properly identify every SSR of any defined length, according to our research. SSRs (Simple Sequence Repetitions) may be found using our MapReduce-based k-mer approach on any distributed platform.

Microsatellites, which are short sequence repeats, make up a large fraction of genomes. SSRs in organellar genomes, on the other hand, have yet to be fully appreciated for their relevance. With the availability of organelle genome sequences, we can examine how SSRs are organised in the coding and noncoding portions of the genomes of various organisms. We found and classified SSRs in the wheat mitochondrial and chloroplast DNA in the current research. According to the findings of A. K. Mishra et al. [2], the number of SSRs in the non-coding region is higher than in the coding region, and the frequency of mononucleotides is highest in the chloroplast genome of wheat while the frequency of tetranucleotides is highest in the mitochondrial genome of wheat.

SSRs (Simple Sequence Repeats) are common biomarkers in genetic investigations because of their abundance in genomic sequences. C.-P. Sio et al. [3] demonstrated the importance of many SSRs in gene regulation. For example, a disease may be caused by aberrant repetition patterns of these essential SSRs. SSR polymorphism identification was made easier by Next Generation Sequencing methods. Prior techniques to detecting SSR markers were hampered by the need of labour-intensive, manual procedures. SSR polymorphisms at genome scales may be detected using an automated and efficient technique that does not need human curation and examination. De novo or reference mapping techniques to data assembling were both supported in this process. It was then possible to acquire the consensus sequences by assembling the contigs together and aligning them to the specified reference sequences. As a further step, a system for mining SSRs was devised to obtain all possible polymorphic SSRs whenever insertions or deletions happened. The CODIS SSR markers and nine well-known diseases related SSR motifs were used as the testing targets for the 1000 genomes Trio studies. This technique was able to detect known polymorphic SSRs and new SSR markers when there were no sequencing or mapping mistakes in the consensus sequences. NGS technology were used to find SSR polymorphism and speed up related studies in order to enable the identification of new SSR biomarkers and the discovery of regulatory elements

II. FOUNDATIONAL DNA ANALYSIS

After setting the context of the proposed research, in this section of the work, the foundational methods for DNA analysis are furnished.

Assuming that the DSX[] is the set of DNA sequences with m number of samples and each sample with n sequence length. Thus, this can be presented as,

$$DSX[\,] = < A, C, G, T >_n^m \qquad \text{(Eq.1)}$$

Thus, one sample from the set can be represented as,

$$DSX[i] = < A, C, G, T >_n \qquad \text{(Eq.2)}$$

As the extraction of Amino Acids from the DNA sequences, follows a simple 4 step process, the first step involves considering a sample DNA sequence with length 4 as,

$$DSP = < A, C, G, T > \qquad \text{(Eq.3)}$$

Further, matching DNA sequence, $\mathrm{MSeq}_{\mathrm{DNA}}$ must be extracted as,

$$MSeq_{DNA} = DSX[i]\ X\ DSP \qquad \text{(Eq.4)}$$

Once the matching DNA pair is extracted, the messenger RNA or $\mathrm{MSeqm}_{\mathrm{RNA}}$ will be extracted. For the extraction of m-RNA, another matching sequence must be considered as RSP = <A, C, G, U>. Hence, with the help of RSP, the $\mathrm{MSeq}_{\mathrm{m\text{-}RNA}}$ will extracted as,

$$MSeq_{m-RNA} = MSeq_{DNA}\ X_{\{A \to U\}}\ RSP_{G \to C} \quad \text{(Eq.5)}$$

Further, by repeating the Eq. 5 once again with $MSeq_{m-RNA}$ and RSP, the transferable RNA or t-RNA can be extracted as,

$$MSeq_{t-RNA} = MSeq_{m-RNA}\ X_{\{A \to U\}}\ RSP_{G \to C} \quad \text{(Eq.6)}$$

The $MSeq_{t-RNA}$ can further matched with the Amino Acid lookup, LT_{A} to extract the set of matching Amino Acid as,

$$A[\,] \leftarrow MSeq_{t-RNA}\ X\ LT_A \qquad \text{(Eq.7)}$$

Further, in the light of the foundational method of DNA to Amino Acid extraction, the recent research improvements are analyzed in the next section of this work.

III. Parallel Research Outcomes

An artificial life simulation is used in this work to examine the DNA sequence repeat pattern in order to determine if an agent's DNA sequence will be repeated in future generations. This also involves study into identifying patterns in repetitions and the ratio of unique and repeated DNA sequences, as well as the development and lifespan of the agent. A representation of life, 'agents,' was chosen for this piece. The agents are intelligent enough to make judgments about their own existence and to adjust to slight changes in the situation. Clogged corners are avoided on the two-dimensional plane. Reproduction is done using a diploid system and a two-point crossover. They also make use of rudimentary learning algorithms and random selection-based weighting when deciding how they should move. Using simple logic that mimics evolutionary processes improves the efficiency of the agents in their environment. In our approach, DNA sequences are saved and the behaviours of individuals in a population are tracked and recorded. Data from successful sessions is analysed to look for patterns in DNA sequence repeats and abnormalities, as well as how life evolves over time and how many people survive. As established by S. Ismail et al. [4], the agents have evolved to guarantee an ideal ratio of unique and repetitive DNA sequences.

According to the findings of S. V. Tenneti et al. [5], DNA's tandem repeats are periodic segments. DNA testing is critical in forensics, demographic studies, and other areas. This study attempts to solve the challenge of finding them in lengthy DNA sequences. Based on the newly proposed Ramanujan Filter Bank, a novel approach is described (RFB). Many of the older DSP period estimate methods, such as those based on spectral estimation, have been found to have significant shortcomings (STFT etc.). It just requires basic integer operations and uncovers multiple previously undetected repetitions.

An amino acid sequence contains tandemly repeating portions called protein repeats. They have a significant impact on the protein's structure and binding characteristics. However, the most effective methods for detecting such recurrence have relied on costly approaches like dynamic programming, HMMs and so forth. Traditional DSP techniques such as STFT, however, are unable to handle mutations in a meaningful way. S. V. Tenneti et al. [6] offer a unique approach based on the recently built Ramanujan Filter Bank. Using just basic integer calculations, its performance has been shown on various well-known repetition families.

A DNA-specific compression method may be developed based on the features of DNA sequences. Priyanka et al. [7] describe a DNA sequence lossless two-phase compression technique. Prior to compressing the genetic sequences, a modified form of Run Length Encoding (RLE) is used to generate an ASCII-encoded version. For eight-bit ASCII codes, one-fourth compression is guaranteed, no matter whether the sequence is repeated or not, and the modified RLE approach further boosts the compression. Aside from its promising compressibility, the algorithm's straightforward compression method adds to its appeal.

According to Zhang and colleagues [8], the goal of their research was determining the cytochrome B gene (cyt B) features of Zaocys dhumnades. PCR technology was utilised in combination with bioinformatics technologies to build primers, sequence, and blast. The specificity, sensitivity, and stability of a new Zaocys dhumnades DNA test kit was all assessed. China Pharmacopeia results were in line with those obtained using the kit procedure (2012 edition).

The amino acid encoding plays a critical role in the effectiveness of machine-learning based protein structure and function prediction algorithms. It is possible to utilise the amino acid encoding to predict the characteristics of a protein at both the residue level and the sequence level by combining multiple techniques. A lack of attention in the last decades has led to a lack of complete evaluations and evaluations of encoding techniques thus far. However, According to X. Jing et al. [9], a systematic categorization and evaluation of several amino acid encoding techniques are presented in this paper. All of these approaches are divided into five categories depending on how they gather and extract information. These are the five ways listed above: binary code, physical-chemical property code, evolution-based code, structural code, and code learned by machine-learning. Then, 16 typical approaches from five categories were chosen and evaluated using large-scale benchmark datasets for protein secondary structure prediction and protein fold identification. After analysing all of the methods used in this study, PSSM emerged as the best method for encoding amino acid positions, followed by the structure-based and machine learning encoding methods; the neural network based distributed representation of amino acids in particular may shed new light on this area. This study hopes that the examination and evaluation of amino acid encoding experiments will be relevant for future research.

Predicting protein subcellular localization has gotten a lot of interest recently since it's critical for understanding protein function and developing new drugs that are specifically aimed at it. Traditional approaches for determining the subcellular localization of proteins are time-consuming and expensive. More and more machine learning-based protein sub-cellular location predictors have been produced in the previous two decades, and many more are expected in the near future. Most of these predictors, on the other hand, are limited to predicting proteins at a single subcellular site. There have been an increasing number of proteins discovered that may be present in two or more different subcellular sites. Such proteins have a much greater impact on both fundamental biology and bioinformatics research; thus, their study is more important than ever. Prediction accuracy may be improved by extracting much more feature information that can accurately reflect the protein sequence. The multi-label knearest neighbours (ML-KNN) algorithm was used to estimate protein sub-cellular locations in this study by X. Qu et al. [10].

One of the most fundamental parts of cell structure and function is SIPs, which are self-interacting proteins (SIPs). Today's molecular biologists are focused on a critical problem: how to find SIPs. SIPs data has been obtained through experimental techniques, however wet laboratory methodologies are both time-consuming and expensive. Because of this, they have a large number of false positives and negatives. Thus, there is a pressing need for in silico algorithms to reliably and effectively anticipate SIPs. A novel sequence-based technique for predicting SIPs has been developed in this work. For the Position-Specific Scoring Matrix (PSSM), evolutionary information is retrieved from proteins with known sequence. As a final step, we feed the characteristics into an ensemble classifier to determine if a protein is self-interacting or not. When compared to the SVM classifier, the approach of J. -Q. Li et al. [11] also performs well. Thus, the suggested technique may be regarded an innovative and promising tool for predicting SIPs.

It is necessary to evaluate protein sequences in order to anticipate their activities because of how the amino acids are arranged in a protein sequence. Due to their speed in comparison to BLAST and FASTA-based techniques, machine learning-based systems fail to perform effectively for large protein sequences (with more than 300 amino acids). Two different feature sets for proteins have been constructed using a bi-directional long short-term memory network that analyses fixed single-sized and multi-sized segments, respectively, in this study by A. Ranjan and colleagues [12]. With the suggested feature set, based on multi-sized segments, paired with the model trained using multilabel linear discriminant analysis (MLDA) features, the accuracy is increased even more.

Secondary and tertiary structure prediction, protein fold prediction, and protein function analysis all benefit from knowledge of protein structural classes. The ability to correctly classify proteins based on their structural classifications is critical. In recent years, a number of computational approaches have been developed to predict protein structure classes with modest sequence similarity (25 percent to 40 percent). There are some discrepancies between the claimed accuracy and the actual accuracy of the predictions. Wei et al. [13] offered three distinct feature extraction approaches in order to increase prediction accuracy even more and created a complete feature set that includes both sequence and structural information. This study further develops a unique approach for predicting structural classes by using a random forest (RF) classifier.

In molecular biology, cell biology, biomedicine, and drug creation, information about protein 3-dimensional (3D) structures is critical. Predicting protein folds is seen as a necessary first step in figuring out the three-dimensional (3D) structures of proteins. In structural bioinformatics, protein fold prediction is a basic challenge. In order to predict protein folds, several new taxonomic approaches have recently been devised. The overall accuracy of current taxonomic approaches is not sufficient, despite the fact that significant progress has been made. A new taxonomy technique termed PFPA, which incorporates an ensemble classifier and a novel feature set, was developed by L. Wei et al. [14] to solve this issue. In particular, a complete feature set is constructed by combining the sequential evolution information from PSI-BLAST profiles with the local and global secondary structure information from PSI-PRED profiles. PFPA outperforms current state-of-the-art predictors in experiments.

Understanding the three-dimensional structure of a protein sequence is an essential and difficult issue in the biological sciences. An intermediate stage in determining a protein's three-dimensional structure is the discovery of protein folds from its basic sequence. For example, a feature extraction approach may be used in conjunction with a good classifier to identify an unknown protein. Several feature extraction methods have been developed in the past, but they have only been able to identify a limited number of objects. It was found that the trigrams derived from Position Specific Scoring Matrices may be used by K. K. Paliwal et al. [15] to extract features. It has been shown on two benchmark datasets that the feature extraction method is successful.

The sequences of gram-positive and gram-negative subcellular localizations were represented in this work by R. Sharma et al. [16] using structural and evolutionary characteristics. A normalising approach was presented in this study to create a normalised Position Specific Scoring Matrix (PSSM) from the original PSSM data. Using normalised PSSM feature vectors and a support vector machine (SVM) and naive Bayes classifier, this paper evaluated the proposed method's performance against previously published findings. This study also calculated and compared the characteristics of the original PSSM and the normalised PSSM. Gram-positive and gram-negative subcellular localizations have been boosted in the archived findings for both benchmarks, we found that using SVM and concatenating features (amino acid composition, Dubchak (physicochemical-based features), auto-covariance normalised PSSM-based features, and bigram normalised PSSM-based features) improved localization accuracy while using nave Bayes classifiers with auto-covariance normalised PSSM-based features increased sensitivity.

In terms of protein structure, disulfide linkage is crucial. When a significant number of proteins are sequenced but not functionally annotated, anticipating disulfide connections purely from protein sequence helps increase our knowledge of protein structure and function. Discriminative features are developed by mixing a novel feature derived from predicted protein 3D structural information with standard features in this work by D. -J. Yu et al. [17]. A random forest regression model is used to predict protein disulfide connections based on the collected characteristics. Cross-validation and independent validation studies on benchmark datasets compare the proposed technique with popular current predictors. Experiment findings show that the suggested technique is better than currently used predictors. The authors of this study feel that the suggested strategy is preferable because of both the new features' capacity to discriminate and the random forest's ability to simulate.

The involvement of DNA-binding proteins in a variety of physiological activities, including gene expression and transcription, is crucial. Although ChIP-sequencing and other experimental approaches are costly and time-consuming, in silico methods, such as machine learning-based methods, are needed to discover these proteins. DNA-binding protein prediction accuracy has improved dramatically in recent years thanks to machine learning techniques. Protein sequence translation into an acceptable discrete model or vector is still a major challenge that must be addressed. X. Fu et al. [18] introduced a new feature creation approach based on a position-specific scoring matrix (PSSM) called K-PSSM-Composition. Evolutionary information about 20 amino acid residues and the local information of each specific sequence may be easily captured by these suggested characteristics. To train the support vector machine model for predicting DNA-binding proteins, this study does a recursive feature reduction. This study evaluates and compares our suggested predictor with existing advanced predictors using two standard benchmark datasets.

After the detailed analysis of the existing methods, the identified problems are discussed in the next section of this work.

IV. Problem Formulation

In the recent time, the analysis for DNA to Amino Acid extraction have improved a lot. However, the persistent research problems are furnished here:

A. *Research Challenge 1: Higher Processing Complexity*

Firstly, the time complexity for processing a single DNA sequence is analyzed here.

Assuming that the length of a DNA sequence is n, thus n can be calculated using a function for length extraction as, λ

$$n = \lambda\{DSX[i]\} \qquad \text{(Eq.8)}$$

Further, in order to extraction the matching DNA sequence, each element in the DSX[i] will be compared with 4 DNA base elements. Thus, the time complexity, T, will be calculated as,

$$T = 4.n \qquad \text{(Eq.9)}$$

Further converting the same matched DNA sequence to m-RNA, each base element will be again compared with 4 elements each. Thus, the updated time complexity will be,

$$T = 4*4.n \qquad \text{(Eq.10)}$$

Or,

$$T = 16.n \qquad \text{(Eq.11)}$$

Or assuming k as constant,

$$T = k^2.n \qquad \text{(Eq.12)}$$

Further, during the look process for finding the Amino Acid sequence, each element in t-RNA base will be compared with 20 known Amino bases. Thus, the updated complexity can be calculated as,

$$T = 16 * 20.n \qquad \text{(Eq.13)}$$

Or,

$$T = 320.n \qquad \text{(Eq.14)}$$

Or,

$$T = \log(k^{k^k}).n \qquad \text{(Eq.15)}$$

Thus, it is natural to realize that, for a higher order of n, the total complexity can be significantly higher.

The proposed solution to this problem, which is explain in the next section of this work, is to cluster the sequences in similar parts and further reduce by reduction of the length n.

B. *Research Challenge 2: Inconsistent Amino Acid Sequence*

Secondly, the due to the higher repetition of the sequences in the DNA, the final extracted Amino Acid sequences are also compromised.

Assuming that, the primary DNA sequence is consisting of two independent sequences as, DSX[i] and DSX[j].

$$DSX \subset [DSX[i], DSX[j]] \qquad \text{(Eq.16)}$$

Here, both the sequences are nearly similar or actually similar as,

$$DSX[i] \approx DSX[j] \qquad \text{(Eq.17)}$$

Now, due to the variation of the positions of DSX[i] and DSX[j], the extracted Amino Acid sequences can be different, which are presented as A[i] and A[j].

$$DSX[i] \rightarrow A[i] \qquad \text{(Eq.18)}$$

Or,

$$DSX[j] \rightarrow A[j] \qquad \text{(Eq.19)}$$

Hence, it is natural to realize that,

$$A[i] \neq A[j] \qquad \text{(Eq.20)}$$

But, in the reality, the following relation should be established as,

$$A[i] = A[j] \qquad \text{(Eq.21)}$$

This problem can also be solved by reducing the DNA sequences, which is furnished in the next section of this work, and by achieving the unique Amino Acid sequences.

C. *Research Challenge 3: Linear Search for Amino Acid Extraction Process*

Finally, the extraction of the Amino Acid sequences from the t-RNA sequences is the final step for further processing towards protein sequences.

The lookup process is a linear search as explain using Eq. 7. Thus, the linear search process can lead to a very higher time complexity.

Thus, during this lookup process, the time complexity, T.LT, can be formulated as,

$$T.LT = O(n.t) \qquad \text{(Eq.22)}$$

Assuming that, t is the length of the look up.

The solution to this problem, which is furnished in the next section of this work, is achieved using the indexed look up process.

V. Proposed Solutions

After the detailed analysis of the research problems, in this section of the work, the proposed solutions are furnished using mathematical models.

Firstly, the reduction of the DNA sequences is carried out. Assuming that, the complete DNA sequence, DSX[i], is divided in to P[] sets of parts with length k each. As,

$$DSX[i] =< \text{ATGATGATGGTTTCTAAGTAAGTAGTCTTCTA....}>$$

(Eq.23)

Thus, dividing the same DNA sequence into P[] sets as,

$$P[] = \{\text{ATGA,TGAT,GGTT,TCTA,AGTA,AGTA,GTCT,TCTA,GGTT,CAGT,GGTT,ATGA......}\}$$

(Eq.24)

Further, by applying the K-Means clustering algorithm on P[] with x-number of clusters into K[] set as,

$$K[] = \{\text{(TCTA), (AGTA), (GTCT), (CAGT), (ATGA), (TGAT), (GGTT)}\} \quad \text{(Eq.25)}$$

Here, after the clustering process, only the unique sequences are kept and the naturally the length is shortened. Now, in the light of the Eq. 15, assuming that the new time complexity is T1, then T1 can be calculated as,

$$T1 = \log(k^{k^k}).Y \quad \text{(Eq.26)}$$

Where, Y is the new length.

As Y<<n, thus, T1<<T.

Secondly, the finalization of the extraction of the Amino Acids are carried out.

As the extraction of the Amino Acid is solely relying on the DNA sequence, thus the new Amino Acid sequence, A1[], can be formulated as,

$$K[] \rightarrow A1[] \quad \text{(Eq.27)}$$

Now, in the view of the Eq. 21, only one static sequence of Amino Acids is extracted thus,

$$A1[] = A[i] = A[j] \quad \text{(Eq.28)}$$

Finally, the index-based lookup for the Amino Acids is formulated here. Assuming that, the standard lookup table LT[] is now reconstructed using the index, I[], as, D.LT[] ,

$$D.LT[] \Leftarrow \{LT[], I[]\} \quad \text{(Eq.29)}$$

Thus, in the view of the Eq. 22, the new time complexity, T.D_LT can be formulated as,

$$T.D\_LT = O(Y.\frac{t}{m}) \quad \text{(Eq.30)}$$

Here, Y is the length of the new sequence and m is the number of iterations due to the indexing process. Thus, it is natural to state that,

$$T.D\_LT << T.LT \quad \text{(Eq.31)}$$

This reduces the overall time complexity.

Further, based on the proposed mathematical model solutions, in the next section of the work, the proposed algorithms are furnished.

VI. PROPOSED ALGORITHMS AND FRAMEWORK

After the satisfactory confirmation using the mathematical model in the previous section of this work, in this section the proposed algorithms and the framework is furnished.

Firstly, the Sequence Reduction using K-Means Clustering with Sequence Benchmarking (SR-KMeans-SB) Algorithm is furnished and discussed.

Algorithm - I: Sequence Reduction using K-Means Clustering with Sequence Benchmarking (**SR-KMeans-SB**) Algorithm

Input:

DSX[] as DNA Sequence

Output:

K[] set of unique sequences

Process:

```
Step - 1. Assume K[] as emptry set
Step - 2. Accept the DNA sequence as DSX[]
Step - 3. For each element in DSX[i]
    a. count++
    b. If count == 4
    c. Then,
        i. DSXX[] = DSX[start..count]
        ii. Count = 0
    d. Else, Continue
Step - 4. For each elements set of 4 in DSXX[] as DSXX[j]
    a. If DSXX[j] Des not Belong to K[]
    b. Then, Add DSXX[j] to K[]
    c. Else, Continue
Step - 5. Return K[]
```

Outcomes: This algorithm will:

- Uniquely identify the DNA sequences.
- Reduce the length without losing vital information for

further Amino Acid translations.

Primers can be created manually to search for microsatellite markers in specific genomic areas, such as an intron. This includes looking for microsatellite repeats in the genome's DNA sequence, which can be done manually or automatically using programmed like repeat masker. The flanking sequences can be utilized to construct oligonucleotide primers that will amplify a specific microsatellite repeat in a PCR reaction once the possibly useful microsatellites have been identified.

Secondly, the Extraction of Amino Acid Information using Indexed LookUp (EA-AI-IL) Algorithm is furnished and discussed.

Algorithm - II: Extraction of Amino Acid Information using Indexed LookUp (**EA-AI-IL**) Algorithm

Input:

K[] set of DNA unique sequences

AA[] as set of Amino Acids

Output:

A[] as set of Amino Acids

Process:

```
Step - 1. Build the set of Amino Acids as AA[]
Step - 2. Sort the elements as AA[]
Step - 3. Set the order of AA[] into index I[]
Step - 4. Read the DNA sequence as K[]
Step - 5. For each element in K[] with multiple of 3
          elements as K[i]
        a. If K[i] matches with AA[] / I[]
        b. Then, A[] = AA[]
        c. Else If K[i] matches with {AA[] / I[]} to the left
        d. Then, AA[] = Move to the left and start from 5.A
        e. Else If K[i] matches with {AA[] / I[]} to the right
        f. Then, AA[] = Move to the right and start from 5.A
Step - 6. Return A[]
```

Outcomes: This algorithm will:

- Reduce the time complexity to find the Amino Acid sequences.
- Reduce the chance of repetitive Amino Acid sequences.

All of the genetic information is included in both strands of double-stranded DNA. When the two strands split, this information is re-created in the process. More than 98 percent of DNA in humans is non-coding, which means that these regions do not serve as templates for protein sequences. They are antiparallel because the two strands of DNA are running in opposing directions. One of four different nucleobases is attached to each sugar in the genome (or bases). Genetic information is encoded in the DNA sequence of these four nucleobases. A process known as transcription uses DNA strands as templates to make RNA strands, with the exception of thymine (T), for which RNA replaces uracil (U). A process called translation is used to determine the sequence of amino acids in proteins.

Finally, the framework is furnished.

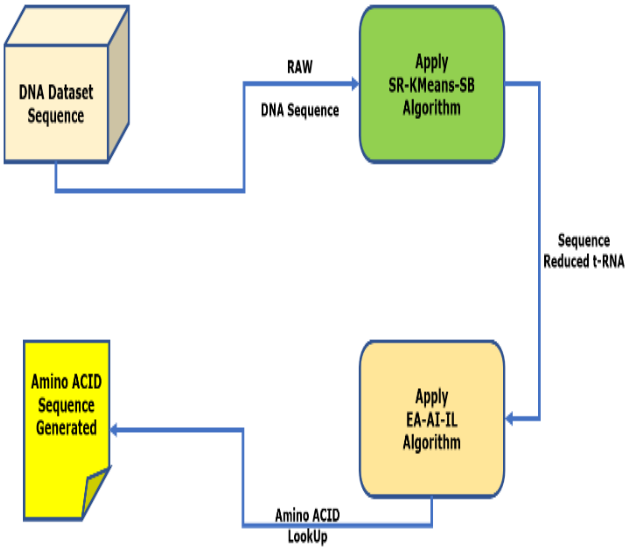

Fig. 1. Proposed Amino ACID Extraction Framework

Further the proposed framework is tested on benchmarked dataset and the results are furnished in the next section of the work.

VII. Results and Discussions

The obtained results are highly satisfactory and are discussed in this section of the work.

Firstly, the dataset characteristics are explained. The proposed algorithms and the framework are tested on PRGdb 4.0 dataset [19] and the initial characteristics are furnished here [Table – 1].

TABLE I. DATASET CHARACTERISTICS ANALYSIS

Characteristics	*Values*
Number of Records	1700
Number of DataSet Items	1700
Number of Unique Species	41

Characteristics	Values
Number of RNA Sequences	47

The dataset is uniquely distributed proportion of possibilities with 1700 unique plant DNA sequences and additionally 47 RNA sequences, which makes the dataset to be utilized for further cross verification on the obtained results.

The proposed algorithms are deployed on the 1700 records, but here only 5 samples are showcased for the representation purposes.

Secondly, the DNA to DNA Pair sequence identification is carried out and the sample 5 sequences are furnished.

TABLE II. DNA TO DNA PAIR EXTRACTION ANALYSIS

Dataset Item ID	*DNA Sequence*	*Reduced DNA Sequence*	*Time (sec)*
PRGDB135	ATGATGATGGTTTCTAGAAAAGTAGTCTCTTCACTTCAGTTTTTCACTCTTTTCTACCTCTTTACAGTT	ATG,GTT,TCT,AGA,AAA,GTA,GTC,TCA,CTT,CAG,TTT,TTC,ACT,TAC,CTC,ACA	0.019
PRGDB136	CATTTGCTTCGACTGAGGAGGCAACTGCCCTCTTGAAATGGAAAGCAACTTTCAAGAACCAGAATAATT	CAT,TTG,CTT,CGA,CTG,AGG,CAA,CCC,TCT,TGA,AAT,GGA,AAG,TCA,AGA,ACC,ATA,ATT	0.013
PRGDB149	CTTTTTGGCTTCATGGATTCCAAGTTCTAATGCATGCAAGGACTGGTATGGAGTTGTATGCTTTAATGG	CTT,TTT,GGC,TTC,ATG,GAT,TCC,AAG,TAA,TGC,CAA,GGA,CTG,GTA,TGG,AGT,TGT	0.015
PRGDB259	AGGGTAAACACGTTGAATATTACAAATGCTAGTGTCATTGGTACACTCTATGCTTTTCCATTTTCATCC	AGG,GTA,AAC,ACG,TTG,AAT,ATT,ACA,GCT,AGT,GTC,GGT,CTC,TAT,TTT,CCA,TCA,TCC	0.013
PRGDB459	TCCCTTCTCTTGAAAATCTTGATCTTAGCAAGAACAATATCTATGGTACCATTCCACCTGAGATTGGTA	TCC,CTT,CTC,TTG,AAA,ATC,TTA,GCA,AGA,ACA,ATA,TCT,ATG,GTA,CCA,TTC,CAC,CTG	0.011

Here during the sample analysis of 5 elements, the average time is 0.0142 sec and for the complete dataset the reduction time is 0.0156 sec. The results are visualized graphically here [Fig – 2].

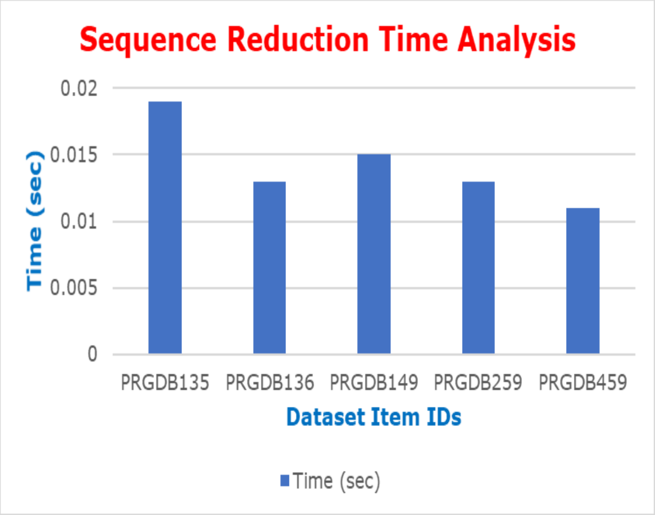

Fig. 2. DNA Seqeunce Reduction Time Compelxity Analysis

Further, the Amino ACIDs are extracted from the original DNA sequence and from the reduced DNA sequence. The comparisons are furnished here [Table – 3].

TABLE III. ACTUAL AMINO ACID AND REDUCED AMINO ACID EXTRACTION COMPARISONS

Dataset Item ID	*Original Extracted Amino ACID Sequence*	*Extracted Amino ACID Sequence from Reduced DNA*	*Accuracy (%)*
PRGDB135	Tyrosine, Stop, Lysine, Cysteine, Glutamine, Serine, Valine, Tyrosine, Glutamic acid, Phenylalanine, Histidine, Methionine	Tyrosine, Stop, Lysine, Cysteine, Glutamine, Serine, Valine, Tyrosine, Glutamic acid, Phenylalanine, Histidine, Methionine	99
PRGDB136	Leucine, Asparagine, Aspartic acid, Tryptophan, Stop, Serine, Valine, Proline, Glycine, Tyrosine,Valine, Threonine, Glutamic acid, Alanine, Phenylalanine	Leucine, Asparagine, Aspartic acid, Tryptophan, Stop, Serine, Valine, Proline, Glycine, Tyrosine,Valine, Threonine, Glutamic acid, Alanine, Phenylalanine	96
PRGDB149	Glutamic acid, Leucine, Aspartic acid, Lysine, Proline, Valine, Serine, Tyrosine, Isoleucine, Threonine, Glutamic acid, Arginine, Phenylalanine, Histidine	Glutamic acid, Leucine, Aspartic acid, Lysine, Proline, Valine, Serine, Tyrosine, Isoleucine, Threonine, Glutamic acid, Arginine,	98

Dataset Item ID	*Original Extracted Amino ACID Sequence*	*Extracted Amino ACID Sequence from Reduced DNA*	*Accuracy (%)*
		Phenylalanine, Histidine	
PRGDB259	Leucine, Asparagine,Serine, Stop, Cysteine, Lysine, Glutamine, Serine, Proline, Glycine, Isoleucine, Glutamic acid, Arginine, Histidine	Leucine, Asparagine,Serine, Stop, Cysteine, Lysine, Glutamine, Serine, Proline, Glycine, Isoleucine, Glutamic acid, Arginine, Histidine	89
PRGDB459	Asparagine, Aspartic acid, Stop, Cysteine, Lysine, Serine, Valine, Tyrosine, Glycine,Arginine, Glutamic acid, Arginine, Phenylalanine, Histidine	Asparagine, Aspartic acid, Stop, Cysteine, Lysine, Serine, Valine, Tyrosine, Glycine,Arginine, Glutamic acid, Arginine, Phenylalanine, Histidine	97

During the sample accuracy analysis, the framework demonstrated nearly 95.8% accuracy and for the complete dataset with 1700 records, the accuracy is 98.8%. The result is also visualized graphically here [Fig – 3].

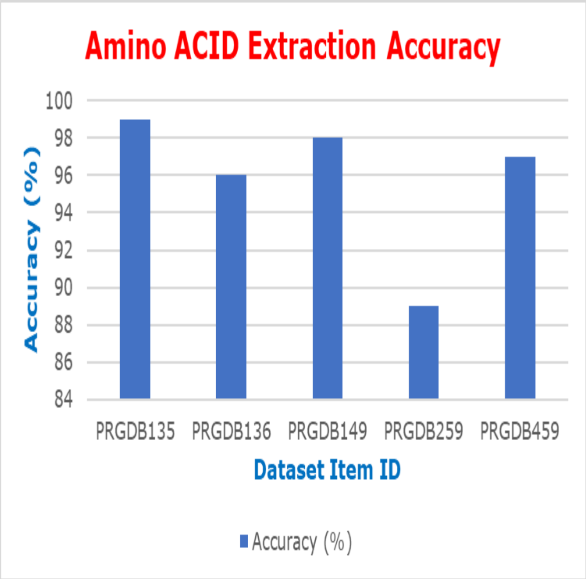

Fig. 3. AMINO ACID Extraction Accuracy Comparison

Further, the final time complexity of the process is analyzed here [Table – 4].

TABLE IV. TIME COMPLEXITY ANALYSIS

Dataset Item ID	*Amino Acid Exaction Time from Original Sequence (Sec)*	*Amino Acid Exaction Time from Reduced Sequence (Sec)*
PRGDB135	0.991	0.124
PRGDB136	0.98	0.089
PRGDB149	0.953	0.087
PRGDB259	0.961	0.137
PRGDB459	0.792	0.099

Hence, the improvement in the time complexity is clearly visible. The analysis is also visualized graphically here [Fig – 4].

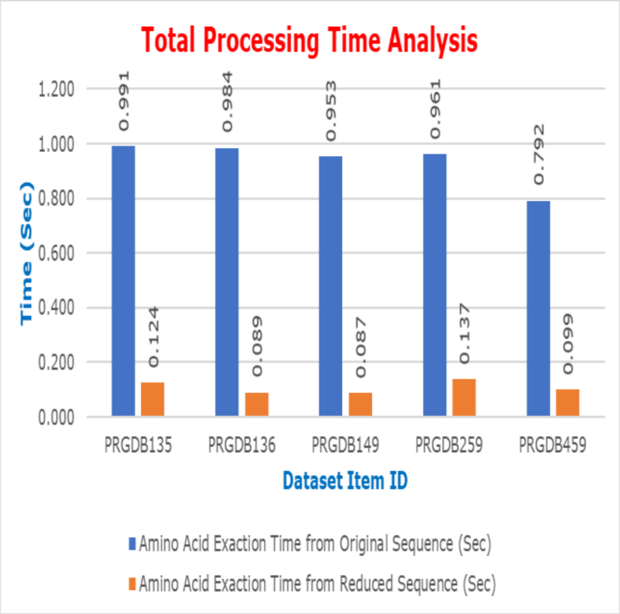

Fig. 4. Total Processing Time Comparison

Henceforth, with the clear understanding of the improvements by the proposed system, the work is compared with the parallel benchmark works in the next section of the work.

VIII. COMPARATIVE ANALYSIS

The outcomes from the proposed algorithms in the framework are compared with the parallel research outcome in order to analyse the improvements [Table – 5].

TABLE V. COMPARATIVE ANALYSIS

Author, Year	*Framework Complexity*	*Mean Time (sec)*	*Accuracy (%)*
J. -Q. Li et al [11], 2017	O(n^3)	0.997	94
S. Mondal et al. [1], 2018	O(n^3)	0.758	96
X. Jing et al. [9],	O(n^3)	0.339	97

Author, Year	*Framework Complexity*	*Mean Time (sec)*	*Accuracy (%)*
2020			
A. Ranjan et al. [12], 2020	$O(n^3)$	0.389	97
Proposed Framework	$O(\log n^2)$	0.107	98.8

Henceforth, with the clear understanding of the improvements over the parallel research outcomes, in the next section of this work, the research conclusion is furnished.

IX. Conclusion and Future Scope

As realized from the parallel research outcomes, during the research surveys, many research efforts have been made in recent years to develop numerous computerised algorithms that may be used to determine the amino acid sequence and, in some cases, protein sequences, that are responsible for the transmission of plant and agricultural diseases. However, owing to the increased complexity of DNA structure, as well as the more difficult procedure for DNA to Amino Acid extraction, the results of these current studies have been disappointing. As a result, in order to address the major obstacle of increased time complexity of DNA processing techniques, this study provides two algorithms that use machine learning to minimise the length of DNA sequences without sacrificing essential information. First and foremost, the use of the clustering approach, in Sequence Reduction using K-Means Clustering with Sequence Benchmarking (SR-KMeans-SB) Algorithm, to minimise the size guarantees that the least amount of information is lost and that the processing time is optimised. Second, using the Extraction of Amino Acid Information using Indexed LookUp (EA-AI-IL) Algorithm, the look-up-based indexed Amino Acid extraction procedure provides improved accuracy of the extraction while also completing the extraction in the shortest amount of time. The suggested framework achieved approximately 98 percent accuracy in a time frame of 0.107 seconds, representing a 5 percent gain in accuracy and a 10 percent improvement in time complexity over the previous framework.

References

[1] S. Mondal and S. Khatua, "Finding simple sequence repeats (SSRs) within human genome using MapReduce based K-mer algorithm," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, India, 2018, pp. 340-345.

[2] A. K. Mishra, S. Chaudhary, A. Kumar and H. Chandrasekharan, "Identification of Simple Sequence Repeats in chloroplast and mitochondrial genome of wheat," 2014 International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2014, pp. 265-270.

[3] C. -P. Sio, Y. -L. Lu, C. -M. Chen, T. -W. Pai and H. -T. Chang, "Mining Polymorphic SSRs from Individual Genome Sequences," 2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, Taichung, Taiwan, 2013, pp. 570-575.

[4] S. Ismail, M. I. Bhuiyan, S. A. Osmani and M. A. Alim Mukul, "An Artificial Life Simulation to Observe DNA Sequence Repeat Pattern," 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2018, pp. 1-10.

[5] S. V. Tenneti and P. P. Vaidyanathan, "Detecting tandem repeats in DNA using Ramanujan Filter Bank," 2016 IEEE International Symposium on Circuits and Systems (ISCAS), Montreal, QC, Canada, 2016, pp. 21-24.

[6] S. V. Tenneti and P. P. Vaidyanathan, "Detection of protein repeats using the Ramanujan Filter Bank," 2016 50th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 2016, pp. 343-348.

[7] Priyanka and S. Goel, "A compression algorithm for DNA that uses ASCII values," 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, India, 2014, pp. 739-743.

[8] X. Zhang, L. Song, W. Yu, T. Zhou and M. Li, "Development and evaluation of a DNA detection kit on authentication of Zaocys dhumnades based on a bioinformatics method and PCR technology," 2014 7th International Conference on Biomedical Engineering and Informatics, Dalian, China, 2014, pp. 649-653.

[9] X. Jing, Q. Dong, D. Hong and R. Lu, "Amino Acid Encoding Methods for Protein Sequences: A Comprehensive Review and Assessment," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 6, pp. 1918-1931, 1 Nov.-Dec. 2020.

[10] X. Qu, D. Wang, Y. Chen, S. Qiao and Q. Zhao, "Predicting the Subcellular Localization of Proteins with Multiple Sites Based on Multiple Features Fusion," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 1, pp. 36-42, 1 Jan.-Feb. 2016.

[11] J. -Q. Li, Z. -H. You, X. Li, Z. Ming and X. Chen, "PSPEL: In Silico Prediction of Self-Interacting Proteins from Amino Acids Sequences Using Ensemble Learning," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 5, pp. 1165-1172, 1 Sept.-Oct. 2017.

[12] A. Ranjan, M. S. Fahad, D. Fernández-Baca, A. Deepak and S. Tripathi, "Deep Robust Framework for Protein Function Prediction Using Variable-Length Protein Sequences," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 5, pp. 1648-1659, 1 Sept.-Oct. 2020.

[13] L. Wei, M. Liao*, X. Gao and Q. Zou, "An Improved Protein Structural Classes Prediction Method by Incorporating Both Sequence and Structure Information," in IEEE Transactions on NanoBioscience, vol. 14, no. 4, pp. 339-349, June 2015.

[14] L. Wei, M. Liao*, X. Gao and Q. Zou, "Enhanced Protein Fold Prediction Method Through a Novel Feature Extraction Technique," in IEEE Transactions on NanoBioscience, vol. 14, no. 6, pp. 649-659, Sept. 2015.

[15] K. K. Paliwal, A. Sharma, J. Lyons and A. Dehzangi*, "A Tri-Gram Based Feature Extraction Technique Using Linear Probabilities of Position Specific Scoring Matrix for Protein Fold Recognition," in IEEE Transactions on NanoBioscience, vol. 13, no. 1, pp. 44-50, March 2014.

[16] R. Sharma, A. Dehzangi, J. Lyons, K. Paliwal, T. Tsunoda and A. Sharma, "Predict Gram-Positive and Gram-Negative Subcellular Localization via Incorporating Evolutionary Information and Physicochemical Features Into Chou's General PseAAC," in IEEE Transactions on NanoBioscience, vol. 14, no. 8, pp. 915-926, Dec. 2015.

[17] D. -J. Yu, Y. Li, J. Hu, X. Yang, J. -Y. Yang and H. -B. Shen, "Disulfide Connectivity Prediction Based on Modelled Protein 3D Structural Information and Random Forest Regression," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, no. 3, pp. 611-621, 1 May-June 2015.

[18] X. Fu, W. Zhu, B. Liao, L. Cai, L. Peng and J. Yang, "Improved DNA-Binding Protein Identification by Incorporating Evolutionary Information Into the Chou's PseAAC," in IEEE Access, vol. 6, pp. 66545-66556, 2018.

[19] Joan Calle García, Anna Guadagno, Andreu Paytuvi-Gallart, Alfonso Saera-Vila, Ciro Gianmaria Amoroso, Daniela D'Esposito, Giuseppe Andolfo, Riccardo Aiese Cigliano, Walter Sanseverino, Maria Raffaella Ercolano, PRGdb 4.0: an updated database dedicated to genes involved in plant disease resistance process, Nucleic Acids Research, 2021.