Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Identification Of The Immune Gene Sequences In Plant Species Using Machine Learning Techniques

Mukesh Perumala[1] , Syeda Sameen Fathima[2] , Vasumathi D[3]

[1]I/c-AKMU-Formerly ARIS Cell ICAR - Indian Institute of Millets Research (IIMR), GOI, Rajendranagar, Hyderabad 500030,TS, India

[2](Rtd) Professor & Head Dept Of Computer Science & Engineering, College of Engineering, Osmania University, Hyderabad

[3]Computer Science and Engineering Jawaharlal Nehru Technological University (JNTU), Kukatpally Hyderabad-India in 2011

Abstract—Agriculture has a big influence on nations like India, and current research is responsible for making this subject very cost efficient and profitable. Despite this, a large number of crops are lost each year owing to pathogens such plant diseases. In addition, certain crops have been shown to be particularly resistant to certain illnesses. This study presents a unique framework for the identification of immune gene sequences for different species in order to apply or impute such gene sequences to the particular plant species in order to make them immune to diverse illnesses. Four distinct machine learning techniques are proposed in this paper to improve the process of identifying and extracting gene sequences. With a mean time of 2.89 ns, the study achieves over 90% accuracy in identifying and extracting immune gene sequences.

Keywords— Intelligent Data Separation, Dataset Normalization, Anomaly Removal, Gene Sequence Normalization, Uniqueness Identification, Immune Gene Sequence Identification, Explicit Selection

I. INTRODUCTION

RNA silencing kills plant viruses. As a counter-defensive, they've constructed silencing suppressor proteins. Silence suppressor protein families share little structure and sequence. Sequence-based search methods can't annotate these proteins. Machine learning approaches are more efficient and less time-consuming. Relevant and/or redundant information, class imbalance, and the selection of a suitable learning method all affect machine learning performance. P. Jain et al. [1] proposes a new technique to improve the prediction performance for RNA silencing suppressors. Using SMOTE's synthetic minority over-sampling strategy, a boosted random forest method with a boosted random forest algorithm offers the best results with a sensitivity of 98.90%, specificity of 95.30%, overall accuracy of 96%, and an AUC of 0.93 to 0.993. (SMOTE). Experiment findings reveal that the proposed strategy gives the best results to yet. Using SMOTE, fuzzy rough feature subset selection paired with evolutionary search may achieve these results.

Since sequencing prices have dropped, bulk segregation analysis mapping is prevalent in functional genomics. Processing that much data is difficult. S. Jia et al. [2] developed MSA based on our maize kernel mutants to identify causative genes based on DNA and RNA level variations in mutant pools. Linkage disequilibrium and linkage maxima may be observed in these areas. The method compares mutant and normal F2 populations and their genetic parents. This reduces data noise and identifies linkage peaks.

MicroRNAs (miRNAs) regulate posttranslational gene expression. Non-coding, short (18–26 nt), cell-made. Cross-gene regulation between plants and animals has piqued researchers' interest in regulatory RNAs. This research will first examine the role of plant miRNAs in fighting hepatitis B and C viruses, which attack the liver. Bioinformatics approaches predicted cross-kingdom interactions of plant miRNAs targeting HBV and HCV. This work employed MiR156/157, 166, 169, 172, 390, and 399 to find targets in HBV and HCV GenBank sequences. 12 HBC and HCV genes may be targeted by plant miRNAs. This work documents plant MiRNAs and their HBV and HCV targets. This initial set of data will help clarify the association between plant miRNAs and animal and human pathogens including HBV and HCV, according to M. Y. K. Barozai et al [3].

The rest of the work is furnished such as in the Section – II, the foundational method for gene sequencing are elaborated, in Section – III the current state of arts are elaborated, in Section – IV, the identified research problems are discussed, the proposed solutions are elaborated in the Section – V and Section – VI, the obtained results and the comparative analysis works are discussed in Section – VII and VIII respectively and the research conclusion is presented in the Section – IX.

II. FUNDAMENTALS OF GENE SEQUENCING

After setting the context in the previous section of this work, in this section, the fundamental process for gene sequencing is realized.

Assume that, the complete gene dataset, D[], is consisting of species information, SP, gene sequence, GS and the disease class, C. Thus, this collection can be presented as,

$$D[] =< SP, GS[], C > \qquad \text{(Eq.1)}$$

Further, the gene sequence is also a collection and assuming that the length of the gene sequence is n, then is also can be represented as,

$$GS[] = \left| X_1, X_2, X_3 .... X_n \right|_n \qquad \text{(Eq.2)}$$

And, the disease class is also consisting of the number of possible diseases, assuming m, and one label as “No Disease”. Thus, this also can be presented as,

$$C \subset \left| K[] \right|_m \qquad \text{(Eq.3)}$$

Where K[] is the set of diseases.

Further, in order to identify the immune plants, IS[], the no disease labelled records are to be extracted as,

$$IS[] = \prod_{C=No\ Disease} D[] \qquad \text{(Eq.4)}$$

Further, identification of the immune genes, IG[] can be performed as,

$$IG[] = IS[]\ X_{\partial}\ D[] \qquad \text{(Eq.5)}$$

Where, ∂ is the selection factor and can be formulated as,

$$\partial = IS[]\big|_{SP} \in D[]\big|_{SP} \qquad \text{(Eq.6)}$$

As, the base line strategy defines to select the immune genes from the no disease class and further search the same sequences from the diseased classes in order to identify the immune sequence.

Further, based on these understandings, in the next section of this work, the parallel research outcomes are analyzed.

III. LITERATURE REVIEWS

MicroRNAs (miRNAs or miRs) are small (18-25 nt) yet extremely efficient ndRNAs produced by pre-miRNA fragmentation (mRNAs). MiRNAs are increasingly employed as biomarkers for genetic diseases, therefore their discovery and study are crucial in biology. Despite mounting empirical evidence, new ndRNAs derived from longer ncRNAs are likely underestimated, according to recent study. Domain experts use Next Generation Sequencing (NGS) data to find and understand miRNAs. In-silico approaches confront efficiency, effectiveness, and generalizability concerns. Instead of string-based NGS alignment/analysis, our group recommended wavelet-based signal processing to mine ndRNAs. Since this was a novel way for mining RNAs, our initial algorithm concentrated on sdRNAs, tRFs, and miRNAs. Because of their prominence in the literature and the availability of empirically confirmed databases, we picked these RNAs. The prevalence and degree of ndRNA functions from non-miRs, sdRNAs, and tRFs in humans and millions of other organisms is unknown. Given the speed of NGS data output, ndRNA extraction and experiments must be automated. Our method may be used to more than 500 animals and their ncRNA sequences from the NCBI annotation database. Eukaryotic, plants, bacteria, fungus, and protozoa are included. SURFR, a real-time user-friendly application, lets experts and aspiring biomedical scientists examine ndRNAs using RNA-Seq. Our method can recognise ndRNAs from 30 NGS files, analyse, show, and compare them for testing. Users may validate their new findings using NGS files from SRA and ndRNAs from TCGA. This supports our platform's efficacy assessment theoretically [4].

Long non-coding RNAs (lncRNAs) are widely involved in cell and developmental processes, hence several methods have been developed to discover them. Few techniques focus on plant lncRNA identification, whereas most are for animal systems. Plant lncRNAs are different from animal lncRNAs. A reliable computational method should be used to identify plant lncRNAs. ItLnc-BXE was developed by G. Zhang and colleagues [5] as a plant-specific lncRNA identification method. Transcripts are shown by gathering and refining sequence features. Training several base learners and integrating them with ensemble learning creates an ItLnc-BXE model. ItLnc-BXE models outperform other plant lncRNA identification

strategies (AUC>95.91%). Cross-species lncRNA identification tests are also conducted in this work. The results reveal that dicots-based and monocots-based models can accurately identify lncRNAs in mosses and algae.

Sigma factor of RNA polymerase holoenzyme regulates gene expression. Once it finds DNA sites, RNA polymerase delivers its core enzyme to target genes' upstream regions. Understanding functional genomic data requires knowing the promoters of a certain kind of sigma factor. This study created a new method for predicting bacterial sigma-54 promoters. The new method combines motif-finding and machine learning to analyse sigma-54 promoters. In E. coli benchmark tests, our method can distinguish between sigma-54 promoters and surrounding or randomly picked DNA sequences. Based on B. Liu et al. [6] and our own studies, we can infer that our approach is robust and globally applicable.

RNAi is a sequence-specific post-transcriptional gene silencing process caused by double-stranded RNA. MicroRNAs govern various biological processes in insect development and metamorphosis (miRNAs). MicroRNAs and their target genes are being utilised to better understand developmental processes. This study sought to identify miRNAs that target Bombyx mori's juvenile hormone epoxide hydrolase (JHEH). Juvenile Hormone (JH) is a hydrolytic enzyme that affects development physiology and reproductive maturation in Lepidoptera. NCBI utilised JHEH's genomic sequence to identify miRNA (NW004582036.1). Only the strongest miRNAs targeting JHEH were selected for gene silencing. RNAi may be employed to extend the larval stage of silkworms, resulting in increased silk output, because JH degradation signals pupation. In vitro and in vivo studies are underway to employ miRNAs to block JH degrading enzyme [7].

MicroRNA (miRNA) prediction techniques relying on annotations may miss functioning miRNAs. Updated miRNA annotation criteria may improve plant miRNA prediction. Alzahrani and colleagues [8] anticipate Arabidopsis thaliana miRNA. They provide a degradome-aided method for finding functional miRNAs. This research tested how a new criterion and a more lenient criterion affects miRNA prediction systems. This work used degradome sequencing to predict miRNAs. Degradome-assisted miRNA prediction exceeds unassisted prediction in this research. This research compared projected miRNAs to several parameters and found a previously undiscovered candidate in Arabidopsis thaliana. This article introduces PAREfirst, a freeware degradome-aided application. Some miRNAs may have been missed due to the strictness of the prior annotation criteria. A degradome-assisted approach with more lenient miRNA criteria may improve miRNA predictions.

MicroRNAs affect post-transcriptional gene regulation (miRNAs). Many machine learning-based studies identify miRNAs using miRNA properties. Since plant pre-miRNAs are more variable than animal pre-miRNAs, it's harder to tell them apart. This study identifies authentic and fake plant pre-miRNAs. P. Ihalagedara et al. [9] presented a machine learning model using 280 compositional, sequence-based, and thermodynamic features.

Predicting lncRNA-protein connections is crucial for understanding fundamental biological processes and plant and animal sickness. In recent years, LncRNAs have proliferated (lncRNAs). Little effort has been done to predict lncRNA-protein interactions (LPI) to describe plant lncRNAs. LPI-DL predicts plant lncRNA-protein interactions using deep learning. We use the optimal blend of k-nucleotide frequency and codon-based encoding for the model's input. Recurrent neural networks develop discriminative long-term dependencies (RNN). J. S. Wekesa et al. [10] employ RFE-SVM to determine the optimum features and connection pruning to sparsely project input sequences' hidden states. Two plant datasets demonstrate LPI-

superiority over other techniques. The proposed strategy performs best in comparison tests. This study improves the accuracy of interaction prediction for future research into lncRNA functions.

The Smith-Waterman (SW) method permits quick alignment of a small query sequence with a large reference sequence "db" in DNA, RNA, and protein sequences. BLAST features a heuristic step of seed indexing and an extension phase using Smith-Waterman (SW) sequence comparison. This paper recommends using a two-dimensional matrix instead of a sparse one to hold BLAST's seed index. It uses our GPU to boost planting performance, lowering processing time by 11.24 percent compared to sequential and threaded multi-CPU implementations. We utilised an O-time technique to acquire pattern-key-matching seeds (1). Hash key length enhances efficiency [11].

G. Leitao et al. [12] introduced a real-time warning processing system to forecast abnormal operational situations. The recommended method uses a database of critical occurrences. Using rules and projected alert sequences, crucial situations are modelled. These situations have occurrence indices to estimate their likelihood. The most probable critical scenario is determined by the expected situation's similarity index. Simulated oil refinery used to test concept.

Computational prediction of novel microRNAs in a complete genome requires identifying miRNA precursor sequences (pre-miRNA). These sequences are miRNA candidates. The number of well-known pre-miRNAs is tiny compared to the hundreds of thousands of probable miRNA candidates, making this task a classification problem with a high class-imbalance level. Classical training approaches employed well-known pre-miRNAs as positive classes and arbitrarily constructed negative classes. Negative examples are much harder to find than positive ones, making it difficult to find adequate training samples for unsupervised labelling. G. Stegmayer et al. [13] use machine learning to avoid defining negative instances. Clustering unlabeled genome sequences with known miRNA precursors allows quick discovery of the best miRNA candidates. Too few positive class labels are addressed by a deep model. Deeper layers screen out fewer likely pre-miRNA sequences. Our method correctly predicts new pre-miRNAs in many mammals. Our approach is less time-consuming to learn and enables greater visualisation and comprehension of results.

Noncoding RNAs and post-translational changes boost plant growth (PTM). Wu et al. [14] employed miRNA-seq and RNA-seq to explore PTM-associated circRNAs in Populus euphratica Oliver heteromorphic leaves. Sequencing data and the ceRNA hypothesis were used to build circRNA-miRNA-mRNA regulatory links in P. euphratica heteromorphic leaves. GO analysis was improved based on circRNAs' targets. Antagonizing 51 miRNAs revealed 17 circular RNAs in P.euphratica that may co-control the development of heteromorphic leaves by protein modification and panning regulation.

Further, in the next section of this work, the identified problems are presented.

IV. PROBLEM FORMULATION

After analysing the recent research outcomes on gene sequencing, the following problems can be identified as bottlenecks for further improvements.

Firstly, the time complexity of the present systems is significantly high due to the nature of the algorithms used. Also, added to that, the length of the gene sequences is naturally lengthy, which increases the time complexity further. In order to prove the same fact, continuing from the Eq. 2 and 3, assuming that the length of the gene sequence is n and the number of diseases are m. These two factors can be formulated as,

$$n = \lambda(GS[]) \qquad \text{(Eq.7)}$$

And,

$$m = \lambda(C[]) \qquad \text{(Eq.8)}$$

Where, λ is the arbitrary function to calculate the length.

Henceforth, in Eq. 5, the length of IG[] can be calculated as,

$$\lambda(C[]) = n.m \qquad \text{(Eq.9)}$$

Or,

$$\lambda(C[]) = n^2 \qquad \text{(Eq.10)}$$

Naturally, this implies the time complexity as, $O(n^2)$

Secondly, the few of the samples in each dataset can incorrect as, they can have the same gene sequences as the immune gene sequences and still labelled as diseased plant samples. These samples must be removed from the dataset initially in order to reduce the complexity and increase the accuracy of the proposed models.

Further, in the next section of this work, the proposed solutions are furnished using the mathematical models.

V. PROPOSED SOLUTIONS

After the detailed analysis of the existing systems and research bottlenecks in the previous section of this work, in this section the proposed solutions are furnished using the mathematical model.

Continuing from the Eq. 1, the dataset D[] is separated into two parts as K1[] and K2[] with disease infected plants and plants with no diseases respectively. Thus, these two can be formulated as,

$$K1[] = \prod_{C <> No\ Disease} D[] \qquad \text{(Eq.11)}$$

And,

$$K2[] = \prod_{C = No\ Disease} D[] \qquad \text{(Eq.12)}$$

Further, the reduction of the no disease dataset must be carried out and is identified as UK2[] as,

$$UK2[] = K2[0..\frac{n}{2}] \notin K2[\frac{n}{2}+1..k] \quad \text{(Eq.13)}$$

Where the items available only once are compared with the complete no disease dataset to identify the unique sequences.

Further, the disease dataset is compared with the no disease datasets to identify unique matches of the gene sequences to build a new set, RK1[], without any data items where the gene sequence matches with the immune gene sequence but identified as diseased. The RK1[] can be formulated as,

$$RK1[] = K1[] - \prod_{C=Disease \,\&\&\, K1[GS]=UK2[GS]} K1[] \quad \text{(Eq.14)}$$

Further, the immune gene sequences can be identified as,

$$IG[] = \prod_{UK2[GS] \subset RK1[GS]} RK1[] \qquad \text{(Eq.15)}$$

Thus, finally, the immune gene sequences can be identified as IG[GS].

Further, based on the proposed mathematical models, in the next section of this work, the proposed algorithms are furnished and explained.

VI. PROPOSED ALGORITHMS AND FRAMEWORK

Further based on the proposed model, the algorithms are furnished here.

Firstly, the Intelligent Data Separation Process using Clustering (IDSPC) Algorithm is furnished here.

Algorithm - I: Intelligent Data Separation Process using Clustering (**IDSPC**) Algorithm	
Input:	
Dataset as D[]	
Output:	
K1[] as Plants with Disease Dataset	
K2[] as Plants with No Disease Dataset	
Process:	
Step - 1.	Load the initial dataset as D[]
Step - 2.	For each element in D[] as D[i]
Step - 3.	If D[i].C == "No Disease"

Step - 4.	Then, K1[j]=D[i] using Eq. 11
Step - 5.	Else, K2[k]=D[i] using Eq. 12
Step - 6.	Return K1[] and K2[]

Secondly, Dataset Normalization by Anomaly Removal (DNAR) Algorithm is furnished here.

Algorithm - II: Dataset Normalization by Annomaly Removal (**DNAR**) Algorithm
Input: Disease Dataset as K2[] No Disease Dataset as K1[]
Output: Processed Dataset as KK2[]
Process: Step - 1. Load the dataset with disease as K2[] Step - 2. Load the dataset without disease as K1[] Step - 3. For each element in K2[] as K2[i] a. If K1[j].Gene <> K2[i].Gene and K2[i].C <> "No Disease" b. Then, KK2[] = K2[i] Step - 4. Return K2[]

Thirdly, Gene Sequence Normalization by Uniqueness Identification (GSNUI) Algorithm is furnished here.

Algorithm - III: Gene Sequence Normalization by Uniqueness Identification (**GSNUI**) Algorithm
Input: No Disease Dataset as K1[]
Output: Processed Dataset as KK1[]
Process:

Step - 1. Load the No Disease dataset as K1[] Step - 2. For each element in K1[] as K1[i] a. If K1[i] <> K1[i+1] b. Then, KK1[j] = K1[i] using Eq. 13 Step - 3. Return KK1[]

Finally, the Immune Gene Sequence Identification using Explicit Selection (IGSIES) Algorithm is furnished here.

Algorithm - IV: Immune Gene Sequence Identification using Explicit Selection (**IGSIES**) Algorithm
Input: Processed Dataset as KK1[] Processed Dataset as KK2[]
Output: Immune Gene Sequence as GS[]
Process: Step - 1. Load the dataset as KK1[] and KK2[] Step - 2. For each element in KK1[] as KK1[i] a. If KK1[i].Gene Contains KK2[j].Gene using Eq. 14 b. Then GS[k] = Gene using Eq. 15 Step - 3. Return GS[]

Further the proposed framework is furnished here [Fig – 1].

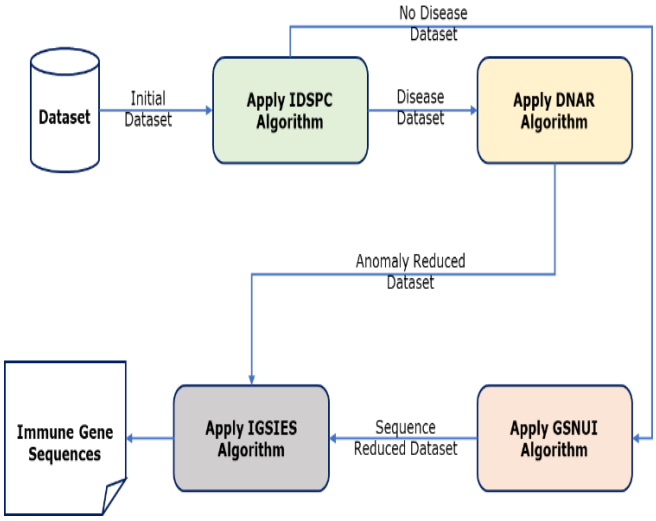

Fig. 1. Proposed Immune Gene Sequence Identification Framework

Further, the obtained results are discussed in the next section of this work.

VII. RESULTS AND DISCUSSIONS

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

Firstly, the initial dataset characteristics are discussed [Table – 1].

TABLE I. DATASET CHARACTERISITCS

Characteristics	Values
Number of Records	3400
Number of Attributes	4
Number of Unique Diseases	11
Number of Records with Diseases	2420
Number of Records without Diseases	680

The synthetic dataset is a good distribution of nearly 80% of the data with diseased gene sequences and 20% complete immune gene sequences, which makes it perfect for this analysis.

Further, the analysis is visualized fraphically here [Fig – 2].

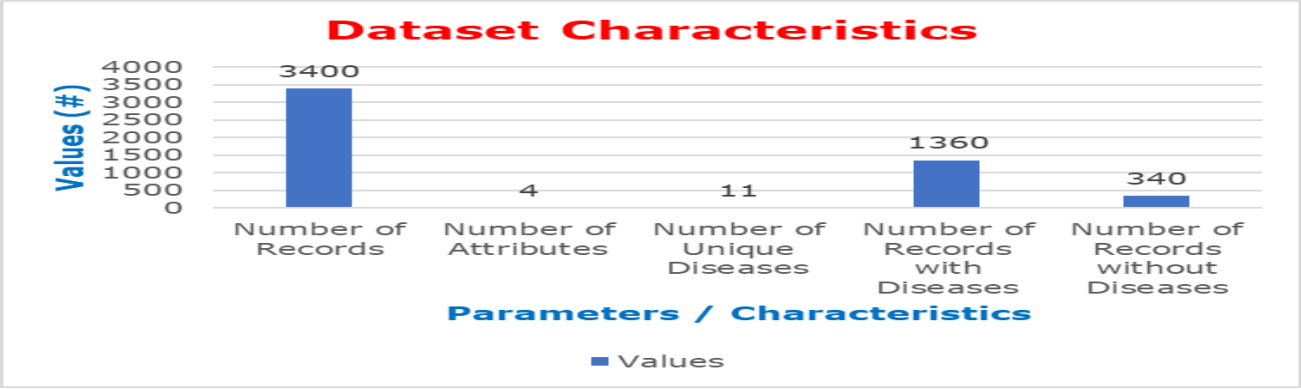

Fig. 2. Initial Dataset Characteristics

The actual analysis is carried on more than 3400 sequences, however only 10 are listed here for the representation purpose.

Secondly, the disease type, actual gene sequence and the immune gene sequence are furnished [Table – 2].

TABLE II. GENE SEUQNECE IDENTIFICATION

Data set Sequence ID	Disease Type	Actual Gene Sequence	Immune Gene Sequence
Seq – 1	Cucumoviruses	CTGGAAATCTAAGATGGCTTGCAATCAAAAAACTGGACATTATGCGGA	CTGGAAATCTAAGATGGCTTGCAA
Seq – 2	late blight	CATTTGCTTCGACTGAGGCAACCCTCTTGAAATGGAAAGTCAAGAACCATAATT	CATTTGCTTCGACTGAGGCAACCCTCT
Seq – 3	Speck	CTGGAAATCTAAGATGGCTTGCAATCAAAAAACTGGACATTATGCGGA	CTGGAAATCTAAGATGGCTTGCAA
Seq – 4	Canker	CTTTTTGGCTTCATGGATTCCAAGTAATGCCAAGGACTGGTATGGAGTTGT	CTTTTTGGCTTCATGGATTCCAAGT
Seq – 5	Mosaic	CATTTGCTTCGACTGAGGCAACCCTCTTGAAATGGAAAGTCAAGAACCATAATT	CATTTGCTTCGACTGAGGCAACCCTCT
Seq – 6	Tobacco Streak	ATGGTTTCTAGAAAAGTAGTCTCACTTCAGTTTTTCACTTACCTCACA	ATGGTTTCTAGAAAAGTAGTCTCA
Seq – 7	Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	TTTTGATATGCAGAACAAACTTTCTGG
Seq – 8	Fusarium	ATTCATATGAAGGTAGATTACGTGATCCAGTTTCAAGTTGCACTGTGT	ATTCATATGAAGGTAGATTACGTG
Seq – 9	late blight	CTGGAAATCTAAGATGGCTTGCAATCAAAAAACTGGACATTATGCGGA	CTGGAAATCTAAGATGGCTTGCAA
Seq – 10	Anthracnose	ATGGTTTCTAGAAAAGTAGTCTCACTTCAGTTTTTCACTTACCTCACA	ATGGTTTCTAGAAAAGTAGTCTCA

During the testing process for all the 3400 data items for nearly 89% of the gene sequences the immune gene sequences are identified.

Thirdly, the identified immune gene sequence positions in the actual gene sequence are identified [Table – 3].

TABLE III. GENE SEUQNECE POSITION IDENTIFICATION

Dataset Sequence ID	Disease Type	Gene Sequence Position
Seq – 1	Cucumoviruses	24
Seq – 2	late blight	14
Seq – 3	Speck	17
Seq – 4	Canker	24
Seq – 5	Mosaic	23
Seq – 6	Tobacco Streak	15
Seq – 7	Anthracnose	16
Seq – 8	Fusarium	17
Seq – 9	late blight	11
Seq – 10	Anthracnose	5

These identified positions are the proof that, the immune gene sequences are available in the actual gene sequences and further can be imputed in multiple places to make the plant species highly immune to such diseases. Further, the results are visualized graphically here [Fig – 3].

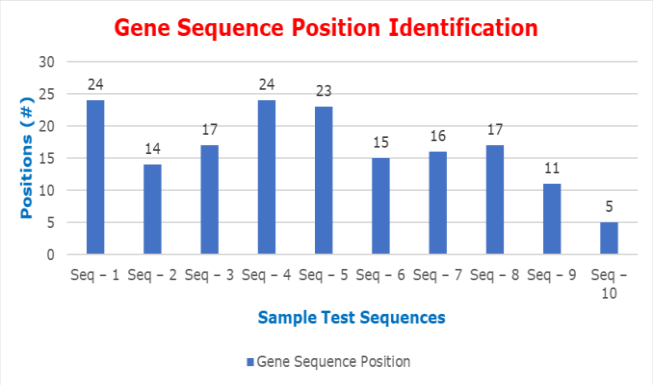

Fig. 3. Immune Gene Sequence Positions

Fourthly, the time complexity for each iteration is carried out [Table – 4].

TABLE IV. GENE SEUQNECE POSITION IDENTIFICATION TIME COMPLEXITY

Dataset Sequence ID	Disease Type	Time to Match (ns)
Seq – 1	Cucumoviruses	1.273
Seq – 2	late blight	1.537
Seq – 3	Speck	3.312
Seq – 4	Canker	3.899
Seq – 5	Mosaic	4.265
Seq – 6	Tobacco Streak	2.864
Seq – 7	Anthracnose	3.707
Seq – 8	Fusarium	2.221

Dataset Sequence ID	Disease Type	Time to Match (ns)
Seq – 9	late blight	1.653
Seq – 10	Anthracnose	4.146

The average time complexity is nearly 2.89 ns. Further the results are visualized graphically here [Fig – 4].

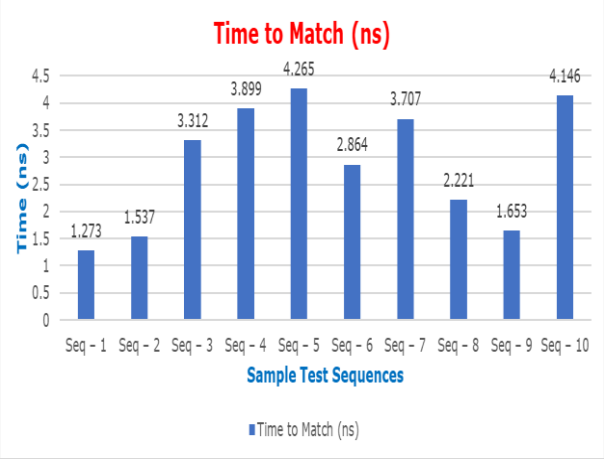

Fig. 4. Immune Gene Sequence Positions matching Time Analysis

Further, the obtained results and the proposed framework are compared with the other benchmarked research works in the next section of this work.

VIII. COMPARATIVE ANALYSIS

The obtained results are also benchmarked against most popular parallel research outcomes and are furnished here [Table – 5].

TABLE V. COMPARATIVE ANALYSIS

Author, Year	Model Complexity	Mean Time (ns)	Extraction Accuracy (%)
P. Jain et al. [1], 2019	$O(n^2)$	3.96	70.011
M. V. Kasukurthi et al. [4], 2021	$O(n^2)$	4.12	13.500
P. Ihalagedara et al. [9], 2020	$O(n^2)$	8.22	59.658
Proposed Framework, 2022	O(n)	2.89	89.641

Hence, it is natural to realize that the proposed framework has outperformed most the parallel research outcomes.

Further, in the next section, the research conclusion is presented.

IX. CONCLUSION AND FUTURE SCOPES

The primary intension of this research is to build a robust framework for identification of the immune gene sequences to make certain plant species immune to specific diseases. During the propose this work firstly proposes the Intelligent Data Separation Process using Clustering (IDSPC) Algorithm to cluster the dataset into various categories, second proposes the Dataset Normalization by Anomaly Removal (DNAR) Algorithm to reduce the anomalies from the diseased sets in the original dataset, thirdly proposes the Gene Sequence Normalization by Uniqueness Identification (GSNUI) Algorithm to identify the infected species with immune gene and finally, proposes the Immune Gene Sequence Identification using Explicit Selection (IGSIES) Algorithm to identify the immune gene sequences for various specifies for various plant diseases. The final outcome of this work is a robust framework with 89.64% accuracy to extract the immune genes.

The work intend to highlight that, the identified immune genes can be imputed to the plants for making the plant specifies immune to various diseases.

REFERENCES

[1] P. Jain, A. K. Tiwari and T. Som, "Enhanced Prediction of plant virus-encoded RNA silencing suppressors by incorporating Reduced Set of Sequence Features using SMOTE followed by Fuzzy-Rough Feature Selection Technique," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-7.

[2] S. Jia, D. Holding and C. Zhang, "A mapping-by-sequencing tool for searching causative genes in mutants," 2017 IEEE International Conference on Electro Information Technology (EIT), Lincoln, NE, USA, 2017, pp. 338-340.

[3] M. Y. K. Barozai and M. Din, "Initial screening of plant most conserved MicroRNAs targeting infectious viruses: HBV and HCV," 2017 14th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 2017, pp. 192-196.

[4] M. V. Kasukurthi et al., "SURFR: A Real-Time Platform for Non-Coding RNA Fragmentation Analysis Using Wavelets," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 2021, pp. 2720-2727.

[5] G. Zhang, Z. Liu, J. Dai, Z. Yu, S. Liu and W. Zhang, "ItLnc-BXE: A Bagging-XGBoost-Ensemble Method With Comprehensive Sequence Features for Identification of Plant lncRNAs," in IEEE Access, vol. 8, pp. 68811-68819, 2020.

[6] B. Liu, L. Han, X. Liu, J. Wu and Q. Ma, "Computational Prediction of Sigma-54 Promoters in Bacterial Genomes by Integrating Motif Finding and Machine Learning Strategies," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 4, pp. 1211-1218, 1 July-Aug. 2019.

[7] D. M. Mamatha, S. Jyothi, V. V. Satyavathi and K. S. Kumari, "RNA interference (RNAi) technology of microRNAs targetingjuvenile hormone epoxide hydrolase (JHEH) gene for increased silk productivity in Bombyx mori," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2016, pp. 1636-1643.

[8] S. Alzahrani, C. Applegate, D. Swarbreck, T. Dalmay, L. Folkes and V. Moulton, "Degradome Assisted Plant MicroRNA Prediction under Alternative Annotation Criteria," in IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[9] P. Ihalagedara, S. Lokuge, S. Jayasundara, D. Herath and I. Kahanda, "miRNAFinder: A pre-microRNA classifier for plants and analysis of feature impact," 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Via del Mar, Chile, 2020, pp. 1-7.

[10] J. S. Wekesa, Y. Luan and J. Meng, "LPI-DL: A recurrent deep learning model for plant lncRNA-protein interaction and function prediction with feature optimization," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 499-502.

[11] F. L. A. Cruz-Gamero and J. C. Gutiérrez Cáceres, "Optimization of BLAST Seed Indexing in the Alignment of DNA Sequences with GPU using CUDA," 2018 XLIV Latin American Computer Conference (CLEI), Sao Paulo, Brazil, 2018, pp. 527-532.

[12] G. Leitao and L. Affonso Guedes, "Real Time Alarm Processing for Predictive Failure Diagnosis in Petrochemical Plants," in IEEE Latin America Transactions, vol. 14, no. 7, pp. 3481-3489, July 2016.

[13] G. Stegmayer, C. Yones, L. Kamenetzky and D. H. Milone, "High Class-Imbalance in pre-miRNA Prediction: A Novel Approach Based on deepSOM," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 6, pp. 1316-1326, 1 Nov.-Dec. 2017.

[14] H. Wu, C. Li, S. Qin and L. Zhao, "Regulatory networks of circRNAs associated with protein modification in the morphogenesis of Populus euphratica Oliv. leaf shape," 2021 11th International Conference on Information Technology in Medicine and Education (ITME), Wuyishan, Fujian, China, 2021, pp. 347-351.