Proposed Automatic Gene Identification and Replacement Framework 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), Kukatpally
Hyderabad-India in 2011
rachanu@gmail.com

Abstract—
Agriculture and the enterprises that support it play an important role in the economy of several countries, such as India. There is little doubt that agricultural research is primarily focused on the fertilizer industry at the moment. The most significant change, on the other hand, can be brought about in the industry by reducing the amount of money lost as a result of the destruction of crops caused by a variety of plant diseases. As a result, the objective of this work is to determine, through the use of computational approaches, the extent to which plant genes have undergone genetic improvement. This article presents a series of algorithms for determining the sequences of immunological genes, causal genes for plant diseases, and improved genetic sequences. All of this is accomplished without changing the species or any other characteristics that are used to distinguish the plants themselves. This work leads to an accuracy rate of 99.6 percent during the genetic enhancements of plants, which in turn leads to a decrease in plant diseases and, by extension, losses due to plant diseases, so enhancing agriculture and making it a field that is more computer-aided.

Keywords— **Gene Identification, Hill Climbing, Genetic Search, Positional Gene Replacement, Gene Replacement**
DOI Number: 10.14704/nq.2022.20.7.NQ33479 **NeuroQuantology 2022; 20(7): 3885-3894**

I. INTRODUCTION

Based on structural controllability of complex networks and a constructed gene network with 9,241 nodes for Arabidopsis thaliana, this work classified nodes into five categories via their roles in control or node deletion, including indispensable, neutral, dispensable, driver, and critical driver nodes. The indispensable nodes can increase the number of drivers after deletion, which are never drivers or critical drivers. About 10 percent of nodes are indispensable. However, more than 60 percent of nodes are neutral ones. More than 62 percent of nodes are drivers, which indicates the gene network is very difficult to be fully controlled. The analysis of enriched Gene Ontology (GO) terms shows that various groups of nodes favor biochemical and physiological processes. Essential genes, genes responsive to drought that do not require abscisic acid (ABA), transcription factors (TFs), core cell cycle genes, and genes related to ABA and Gibberellin (GA) are significantly overrepresented among the nodes that are necessary. Important nodes are enriched in receptor kinase-like genes, while depleted in WRKY TFs and functional genes. Robustness analysis based on node and edge additions; edge rewiring indicates the obtained conclusions are robust to network perturbations. Our investigations clarify control roles of some gene families and provide potential implications for identifying functional genes in other plant species, such as drought responsive genes and TFs as per P. Wang et al. [1].

Transcription factors (TFs) act as master regulators that directly bind to their respective distinct cis-regulatory elements and activate the expression of many downstream target genes (regulon), and thus play a key regulatory role in plant development and stress tolerance. TF families such as AP2/EREBP, AREB/ABF, bHLH, bZIP, C2H2, C3HIS, HB, DREB1/CBF, HSF, MADS, MYB, MYC, NAC, WRKY, etc., were known to regulate stress responses of plants and were relatively well studied in rice and Arabidopsis. Bread wheat (Triticumaestivum L) draught genome is recently released and is available in Ensembl Plants database. By employing previously established rice TFs, this work was able to construct Hidden Markov Model (HMM) profiles for specific protein families within the TF class. These wheat homologs were then searched using the corresponding Profile HMMs. The SMART methodology was applied to the problem of tagging the relevant domains. Based on our data, this work know that the wheat genome contains members of the AP2/EREBP, AREB/ABF, bHLH, bZIP, C2H2, C3HIS, HB, HIS, HSF, MADS, MYB, NAC, and WRKY gene families, in the following numbers: 201, 166, 265, 182, 200, 102, 200, 274, 54, 125, 315, 226 and 199. The identification of 4533 miRNAs from the wheat genome was the result of a genome-wide analysis of miRNAs. Additionally, abiotic stress-responsive TFs that are targeted by miRNAs are determined. Genome-wide and tandem duplication likely contributed to the growth of these gene families in wheat, as postulated by S. Sahu et al. [2], as evidenced by the widespread distribution of TFs and miRNAs that respond to abiotic stress.

To selecting drought-tolerant genotypes and conducting a functional analysis of related genes, it is crucial to detect and characterize physiological processes in crop plants under water-limited conditions. Hyperspectral imaging at close range (HSI) is a promising non-invasive technique for sensing plant traits with the potential to detect plant responses to water deficit stress at an early stage. This paper presents a data analysis strategy for tapping into this potential. It is important to keep in mind that lighting conditions have a significant impact on the reflectance spectra of plants in close-range imaging. To mitigate the effects of linear illumination, standard normal variate (SNV) was used, while non-linear effects were filtered out by omitting the pixels that were affected during a clustering process. Differences in plant spectra were interpreted as

being related to shifts in plant traits after controlling for the effects of illumination. Using a supervised band selection and the direct calculation of a spectral similarity measure against a reference, a spectral analysis procedure was developed to quantify spectral dynamics in response to stress. HSI data of maize plants were acquired in a high-throughput plant phenotyping platform, and the proposed method was applied to them to evaluate their responses to drought stress and their subsequent recovery after being re-watered. According to M. S. M. Asaari et al. [3], spectral analysis consistently revealed recovery effects shortly after the re-watering period, demonstrating its efficacy in detecting drought stress responses at an early stage.

Henceforth, after setting the initial research context, this work formulates the foundational method for gene replacement and based on the foundational methods, recent improves are also analyzed in Section – II and Section – III. The current and long persistent research problems are identified in the Section – IV. The proposed methods are analyzed, furnished and discussed in Section – V and VI, and the obtained results are analyzed in Section – VII. The comparative analysis and the research conclusion after comparative analysis are furnished in Section – VIII and Section – IX respectively.

II. Foundational Method for Gene Replacement

After setting the initial research context in the previous section of this work, in this section the base line method for gene replacement is analyzed.

Assuming that the complete dataset, DLS[] is collection of gene sequences with n number of attributes in the collection. This can be formulated as,

$$DLS[] = < GS[] >_0^n \quad \text{(Eq.1)}$$

Further, each and every component of the gene sequences are collection of the actual sequence, S, plant class or species, C and the type of the diseases, D. Thus, this can be formulated as,

$$GS[] = \{S, C, D\}[] \quad \text{(Eq.2)}$$

The base line method indicates to identify two tuples in the same plant species with, GS[i], and without, GS[j], the diseases.

$$GS[i] = \{S, C, D\}[i]_{i.D=Disease} \quad \text{(Eq.3)}$$

And,

$$GS[j] = \{S, C, D\}[]_{j.D=No\_Disease\ and\ i.C=j.C} \quad \text{(Eq.4)}$$

Further, as per the baseline method, the gene sequences of the diseased plant must be replaced with the gene sequence of the plant without disease and a completely new plant characteristics, GS'[i] must be developed as,

$$GS'[j] = \{j.S \leftarrow i.S, C', D'\} \quad \text{(Eq.5)}$$

Based on the baseline method, the recent improvements to this strategy are analyzed in the next section of this work.

III. Recent Research Reviews

Further, based on the foundational method, the recent research outcomes are analyzed here in this section of the work.

Senescence, or plant ageing, is a complicated and strictly controlled process. The increased risk of illness that comes with advancing age makes studying the ageing process all the more important. Knowledge how genes and proteins interact is crucial to our understanding of ageing, but analysis of gene expression or sequence data prevents this from happening. Understanding the molecular processes underpinning ageing has motivated this research to model Gene Regulatory Networks (GRNs) for the senescence process in the leaves of Arabidopsis thaliana. Upregulated and downregulated gene expression levels during leaf senescence were taken into account to generate these networks using a Dynamic Bayesian Network model. Six Dynamic Bayesian Networks for studying leaf behaviour across time were produced in this study. We have used dynamic GRNs to anticipate many gene-gene interactions that may have an effect on leaf senescence in Arabidopsis. Next, we examined the gene networks of differentially expressed genes throughout time. Multiple topology-sensitive metrics were used to the derived dynamic Senescence-specific GRNs. The findings demonstrate that changes in local topologies of networks are more strongly related with the loss of connectivity in genes during the activation of visual senescence of a leaf than are changes in global features. Genes implicated in signalling pathways that regulate leaf senescence are uncovered by GO enrichment analysis. Many Senescence Associated Genes (SAGs) were confirmed by our research, and numerous predicted gene-gene interactions were discovered during the network inference phase, as reported by H. M. S. D. Herath et al. [4].

Powdery mildew is the most expensive disease to affect commercial grapevines across the world. The agricultural community has a critical need for a deeper knowledge of the complex genetic underpinnings of powdery mildew (PM) resistance via the discovery of possible gene biomarkers implicated in plant defence systems. The hunt for therapeutically beneficial genes may be aided by machine learning analyses of gene expression data. In order to identify putative gene biomarkers linked with resistance to powdery mildew disease in grapevines, this work used a data-driven computational model in its analysis of gene expression data. Node-Based Resilience Clustering (NBR-

Clust) is a clustering method that utilises a graph-based approach. This study used two distinct graph formats to compare average differences in gene expression levels across 6 time periods between PM inoculation and mock inoculated Cabernet and Norton (PM disease resistant) species (geometric and kNN). By taking an opposite tack, this study postulated that genes located in smaller clusters would exhibit time- and space-specific differences in PM-induced transcript expression, suggesting possible biological importance. This study contrasted the smaller clusters seen in Norton to the bigger clusters observed in Cabernet, analysing the genes in common between the two (by the intersection of sets). These findings demonstrate the superiority of geometric graphs over kNN graphs in this context. Researchers J. Dale et al. [5] found that between the Norton and Cabernet species, there was time-to-time variation in the expression of genes involved in physiologically important processes.

In spite of the widespread adoption of mapping-by-sequencing due to falling sequencing prices, bulk segregation analysis mapping is still routinely used in the field of functional genomics. However, processing such a massive amount of data is not easy. In this study, we used a collection of maize kernel mutants to create a novel mapping-by-sequencing analysis software tool (MSA) for locating disease-causing genes in a population by analysing DNA and RNA expression differences. In order to pinpoint where on the chromosomes linkage peaks are located, it must first locate the association areas that are in linkage disequilibrium with the causal mutations. The technique includes contrasting pools of F2 individuals, either mutant or normal, and may also make use of the parents of the F2s. In order to produce linkage peaks with less background noise, S. Jia et al. [6] report evaluating numerous mutants and normal samples simultaneously.

To analyse omics data, gene co-expression networks (GCN) are becoming more useful. The sample size is far less than the number of genes, which presents a significant problem for GCN creation. Massive samples are required by conventional procedures. As a result, it is likely to be challenging to identify association signals among thousands of candidates since they are weak, nonlinear, and stochastic. This study provides a new method for constructing gene differential co-expression networks (GDCNs) by using the grey correlation coefficient (GCC). In order to use the GDCNs effectively for gene discovery, three metrics are recommended. Overcoming the scarcity of GCNs that can assess the changes in co-expression relationships that may be generated by treatments, the suggested strategy makes optimal use of the data supplied by a small number of samples. Brassica napus RNA-seq data is used to build and study GDCNs in a variety of experimental settings. The GCC-based approach is shown to be particularly resistant to processing errors. Understanding the functions of specific genes and isolating the factors that influence a system's response to stress are both made simpler by the existence of GDCNs. P. Wang et al. [7] note that the GDCN-based techniques combine the 'guilt by association' and 'guilt by rewiring' criteria to provide new methods for analysing omics data.

In terms of fatality, hepatocellular carcinoma (HCC) is the most frequent form of liver cancer. Diagnosis, prognosis, and therapies are all subpar because scientists don't understand the underlying biological process of illness development. This paper evaluated mRNA expression levels in 115 samples of HCC tumour tissue retrieved from the Gene Expression Omnibus (GEO) database with the goal of better comprehending the molecular foundation for HCC's rapid development. Understanding the subset of DEGs that are co-expressed in HCC samples but not in controls is the goal of this study. After that, a PPI network was constructed for the set of genes that had been shown to be significantly co-expressed. The PPI analysis in this study revealed a total of just six genes to be interconnected (MSH3, DMC1, ALPP, IL10, ZNF223, and HSD17B7). Only MSH3, DMC1, HSD17B7, and IL10 were enriched in GO Terms & Pathway enrichment study; these genes are mostly engaged in cellular process, metabolic, and catalytic activities that promote the formation & progression of HCC. Finally, the driver miRNAs and TFs connected with our important genes are revealed using a composite 3-node FFL, as shown by S. Bhatt et al. [8].

The purpose of any clustering technique is to find groupings within a dataset that share similarities among themselves but are distinct from those shared by other clusters. This article uses three case studies using actual gene expression data to demonstrate that common algorithms (including k-means and Markov clustering) do not always accomplish the purpose of clustering as articulated in the literature. This becomes more of a problem when dealing with multi-dataset analyses or multi-dimensional data. This paper suggests using Clust, an automated consensus clustering technique, to solve this problem. Clust can process several datasets simultaneously and produces clusters with better within-cluster similarity and lower intra-cluster similarity than prior approaches. For a more solid and accurate definition of clustering, see B. Abu-Jamous et al[9] .'s work on Clust.

Bacteria of the genus Rhizobium create root nodules on the roots of leguminous plants in order to aid the plant in its ability to fix atmospheric nitrogen. This means they may serve as a supplementary nitrogen source in fertilisers. Research into how different Rhizobium species use their respective codon resources is gaining traction. This paper compared the codon use of three NCBI-accessible Rhizobium strains: Sinorhizobium meliloti 1021, Bradyrhizobium japonicum USDA110, and Rhizobium tropici CIAT899. Analysis of the codon use patterns in the rhizobium genome demonstrates a general preference for G and C codons over A and T codons. The ENc figure reveals a bias in codon use as a result of compositional restrictions and translational selection. According to the results of the correspondence analysis (COA), the first two axes explain the majority of the observed codon use variance. Using a Pearson correlation analysis, we found that the first axis of the COA is significantly correlated with the Codon adaptation index (CAI) and other determinants of codon use bias. N. Rai et al. [10] report that amongst these three strains, there is conservation of 17 optimum codons.

One of the most important crops in the world is the durum type of wheat (Triticum turgidum L.). Durum breeding aims to improve upon existing strains by developing new ones that produce bumper crops in a broad variety of climates. Quantitative trait loci (QTL) have a significant role in determining this personality feature (QTL). In order to pinpoint the QTL associated with grain yield, it is necessary to generate a genomic map with a dense set of markers. The goal of this study was to identify potential genes that are spatially adjacent to quantitative trait loci (QTL) linked with grain yield. Using the Lahn/Chaml map population, this study annotated and discovered QTLs for grain production under varying environmental circumstances. Quantitative trait locus (QTL)-detection areas were selected, and the sequences of 583 SNP markers were analysed bioinformatically. There are 122 identified sequences, and 53 percent of these are putative genes involved in stress tolerance. Tolerance to biotic pressures accounts for 44.6%, tolerance to abiotic stresses for 32.3%, and 23% show tolerance to both. In addition, 29.5% has to do with growth and reproduction in plants. To intracellular transport, 3.3% of the total. And another 2.4% were discovered to serve no discernible purpose or be retro/transposants. Furthermore, 9.8% were determined to be involved in activities inside cells that have not yet been characterised. Approximately 66.7% of the potential genes for stress tolerance were found to be located on chromosome 4B. The remaining 78.1 percent of candidate genes are likewise found on the short arm of 4B chromosome and play important roles in plant development and reproduction. These results highlight the use of the 4B short arm in abiotic stress breeding strategies for durum. This paper details how to use SNPs, which I. Farouk et al. [11] claim to be a powerful strategy for discovering candidate genes related with the grain yield characteristic in durum wheat.

Testing and improving evolutionary ideas is crucial in the fields of biology and medicine, but this can only be accomplished with a comprehensive knowledge of evolutionary connections throughout the whole genome. Theory and observation both indicate that phylogenetic trees representing distinct genes (loci) should not have entirely consistent topologies. The principal source of such phylogenetic incongruence is meiotic sexual recombination in eukaryotes or horizontal transfers of genetic material in prokaryotes, both of which occur often during the evolutionary history of most species. Despite this, a large number of genes should exhibit topologically connected phylogenies, and therefore should cluster together in polydimensional "tree space" into one or more (for genetic hybrids). "Outlier" genes are those whose phylogenies are not clustered in the same region of tree space as the majority of other genes. This may be due to their distinctive evolutionary histories or the impacts of selection. To better detect phylogenetic outliers in a given collection of ortholog groups collected from several genomes, this study offers a unique phylogenomic technique termed CURatio, which uses ratios of total branch lengths in gene trees. CURatio has an advantage over other approaches since it can account for genes that are absent from and/or duplicated in certain genomes. Given a sufficiently enough species depth and topological difference, as proposed by Q. Kang et al. [12], this work used a simulation analysis inside the coalescent model to show that.

The Institute for Plant Genetic Resources in Sadovo uses the GenBank System for gene bank management, and this page will provide an overview of that system (IPGR). Details on the system's make-up and architecture are laid here. Incorporating operational assistants into the design, as outlined by A. Stoyanova-Doycheva et al. [13], is examined with the use of an ontology for data storage and processing related to plant genetic resources.

MicroRNAs (miRNAs) are tiny, naturally occurring non-coding RNAs that play a significant role in post-transcriptional gene regulation. There have been several research using machine learning to identify miRNA characteristics. Classifying plant pre-miRNAs as genuine or faux is more difficult because of their greater diversity compared to animal pre-miRNAs. Therefore, this study aims to classify plant pre-miRNAs as either genuine or fake. In this paper, we provide a 280-feature machine learning model that draws on compositional, sequence-based, and thermodynamic data. Here, we compare the performance of four different classifiers using a wide range of features. The Random Forest classifier outperforms all other methods on the testing dataset, with an accuracy of 97%, as reported by P. Ihalagedara et al. [14].

Medicine development depends significantly on bioassays, which measure the biological activity of a chemical like a hormone or drug using either live animals or plants (in vivo) or tissue or cells (in vitro). The primary objective of this study was to develop methods for optimising the utilisation of the biological data generated by the BioAssay, with the end aim of systematically identifying previously unknown biological correlations. Since drug misuse is one of the top 10 main causes of mortality, understanding its molecular underpinnings is crucial for preventing and controlling the epidemic. Data from PubChem BioAssay on chemicals, genes, and diseases were used to build a biological network with the goal of better understanding the connections between drug misuse and these factors. In this study, we utilised network analysis to identify influential nodes in the drug addiction network using a suite of centrality metrics and then used perturbation to rank the resulting collection of unique relationships.

To a large extent, microRNAs (miRNAs) are responsible for modulating almost every biological function in multicellular eukaryotes. Wheat (Triticum aestivum) is, without a doubt, one of the most commonly farmed cereal crops on the planet. However, the function of the few conserved miRNAs discovered so far in wheat has not been fully elucidated. Accordingly, the purpose of this work was to make some informed predictions regarding the roles of various miRNAs in various wheat metabolic pathways. We used the 8496 mature miRNAs and their precursor sequences from the Viridiplantae to predict 39 mature miRNAs in wheat. In all, 11380 target genes spanning all 21 chromosomes of the human genome were discovered using meta-analysis and then characterised for their biological roles. Thus, the impacts of different circumstances on the regulatory pathways and network topologies of miRNA-targets were investigated. In this study, we present the identification of a miRNA community that functions in nitrogen metabolism and propose the name "miRNA community" to characterise it. Stem loop pulsed RT-PCR

was used to confirm the expression of the top two candidate miRNAs in this community in root and leaf tissue when the plants were nitrogen deficient. D. Nigam et al. [16] state that a deeper knowledge of metabolic processes, which may lead to improved genotypes, may be attained by examining the communities to which these miRNAs belong.

After realizing the recent research outcomes, in the next section of this work, the persistent research problems are furnished.

IV. Problem Formulation – Mathematical Model

After analyzing the research problems in the previous section of this work, it is natural to realize that the gene replacement methods have come a long way from the baseline method. Nevertheless, due to the change in the complete gene sequences, the plants tend to complete change the original form of the gene structure.

The identified problem is furnished here using mathematical models for better clarifications.

Continuing from the Eq. 5, the $GS'[j]$ is the plant, for which the genetic replacement must happened. If the gene sequence, $j.S_1$ contains the immune gene and $j.S_2$ contains the non-immune gene from the total gene sequence $j.S$. This can be formulated as,

$$j.S_1 = \prod j.S \qquad \text{(Eq.6)}$$

And,

$$j.S_2 = \prod j.S \qquad \text{(Eq.7)}$$

Nonetheless, during the gene replacement process, both $j.S_1$ and $j.S_2$ must be replaced as $i.S$ and clearly $j.S \neq i.S$. Thus, it is natural to realize that,

$$j.C \neq i.C \qquad \text{(Eq.8)}$$

And,

$$GS'[j] \neq GS[j] \qquad \text{(Eq.9)}$$

Thus, $GS'[j]$ is a completely new plant sequence, which is not the intension of making the plants immune.

Henceforth, in the next section of this work, the proposed solutions are furnished.

V. Proposed Solutions

After the identification of the research problem in the previous section of this work, in this section, the proposed solutions are furnished.

The proposed model works in three phases. Firstly, the identification of the non-immune gene must be realized as,

$$GS[i] = \{S, C, D\}[i]_{i.D=Disease} \qquad \text{(Eq.10)}$$

Further, the extraction of the gene sequence, which is responsible for non-immunity must be extracted as,

$$S_X[] = \prod_{GS[i].C \in GS[].C \;\&\&\; GS[].D=Disease} GS[].S \qquad \text{(Eq.11)}$$

Naturally, the collection $S_X[]$ will contain multiple gene sequence matching with the requirements. Hence, the proposed system applies, φ function to find the first and the best match as,

$$S_X = \varphi\{S_X[]\} \qquad \text{(Eq.12)}$$

Henceforth, the non-immune gene sequence is identified as S_X.

Secondly, to replace S_X with an immune gene, the immune gene identification process is carried out as,

$$GS[j] = \{S, C, D\}[j]_{j.D=No\_Disease} \qquad \text{(Eq.13)}$$

And, subsequently the immune gene is extracted with a similar approach as Eq. 13,

$$S_Y = \varphi\{S_Y[]\} = \prod_{GS[j].C \in GS[].C \;\&\&\; GS[].D=No\_Disease} GS[].S \qquad \text{(Eq.14)}$$

Finally, the gene replacement process takes place. As highlighted in the problem formulation section at Eq. 6 and Eq. 7, the non-immune and immune gene must be separated.

For this process, the overlapping of the immune and non-immune genes must be identified. This proposed system, lays down a method to identify the positions, P[], of the non-immune gene, where it is not matching with the immune gene as,

$$P[] \leftarrow \phi\{S_X \notin S_Y\} \qquad \text{(Eq.15)}$$

ϕ is the function to identify the non-matching positions.

Further, with the help of the positions, only those non-matching points are replaced instead of the complete gene sequences as,

$$S_X' = \{S_X \Leftarrow S_Y\}_0^{P[]} \quad \text{(Eq.16)}$$

Finally, the modified immune plant is furnished as,

$$GS'[X] = \{X.S \leftarrow S_X', C, D' = No\ Disease\} \quad \text{(Eq.17)}$$

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

VI. Proposed Algorithms and Frameworks

This section of the work furnishes and explains the proposed algorithms and the proposed integrated framework.

Firstly, the Non-Immune Gene Identification using Hill Climbing (NIGI-HC) Algorithm is furnished.

Algorithm - I: Non-Immune Gene Identification using Hill Climbing (**NIGI-HC**) Algorithm
Input: Plant dataset as GDS[] Species as C
Output: Non Immune gene for each species as {S[], C}
Process: Step - 1. Accept the live dataset as GDS[] Step - 2. For each element in the GDS[] as GDS[i] a. If GDS[i].D <> "No Disease" and GDS[i].C==C b. Then, S[][j] = GDS[i].S, C[][j] = GDS[i].C Step - 3. For each element in S[j] // Hill Climbing Phase // a. If Length{S[][j]} > Length{S[][j+1]} and (C[][j]) = Closed_To{C[][j+1]} b. Then, S[] = S[][j], C = C[][j] Step - 4. Return {S[], C}

Hill climbing is a local search-based mathematical optimization method used in numerical analysis. It's an iterative method that takes a starting point of an arbitrary solution and iteratively tries to improve it. If the adjustment improves the solution, then another adjustment is made to the adjusted solution, and so on, until no more enhancements can be made.

Secondly, the Immune Gene Identification using Genetic Search (IGI-GS) Algorithm is furnished.

Algorithm - II: Immune Gene Identification using Genetic Search (**IGI-GS**) Algorithm
Input: Plant dataset as GDS[] Species as C
Output: Immune gene for each species as {IS[], C}
Process: Step - 1. Load the live dataset as GDS[] Step - 2. For each element in the GDS[] as GDS[i] a. If GDS[i].D == "No Disease" and GDS[i].C==C b. Then, S[][j] = GDS[i].S, C[][j] = GDS[i].C Step - 3. For each element in S[j] // Genetic Search Phase // a. If {S[][j]} = Close_To{S[][j+1]} and (C[][j]) = Higher_Than{C} b. Then, IS[] = S[][j], C = C[][j] Step - 4. Return {IS[], C}

Since evolution is an iterative process, each iteration's population is referred to as a "generation," and the process often begins with a population of randomly created individuals. Every member of the population is scored on their fitness every generation, where fitness is the value of the objective function in the optimization issue at hand. The most fit members of the population are randomly chosen, and their genomes are recombined and maybe randomly altered to create a new generation. The algorithm iterates by using the newly generated set of potential solutions. Once a certain threshold has been achieved in terms of population fitness or the number of generations the algorithm is allowed to run, it typically stops.

Finally, the Positional Gene Replacement (PGR) Algorithm is furnished.

Algorithm - III: Positional Gene Replacement (**PGR**) Algorithm
Input: Non Immune gene for each species as {S[], C} Immune gene for each species as {IS[], C}
Output: Modified Gene Sequence as MS[]
Process: Step - 1. Accept the Non Immune gene as S[] Step - 2. Accept the Immune gene as IS[] Step - 3. For each element in S[] a. If S[i] == IS[j] b. Then, i++, j++, P[k]=i c. Else, j=0, P[] <- 0 Step - 4. For each element in P[] as P[k] a. MS[] = Replace S[k] with IS[k] for each position of P[k] Step - 5. Return MS[]

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

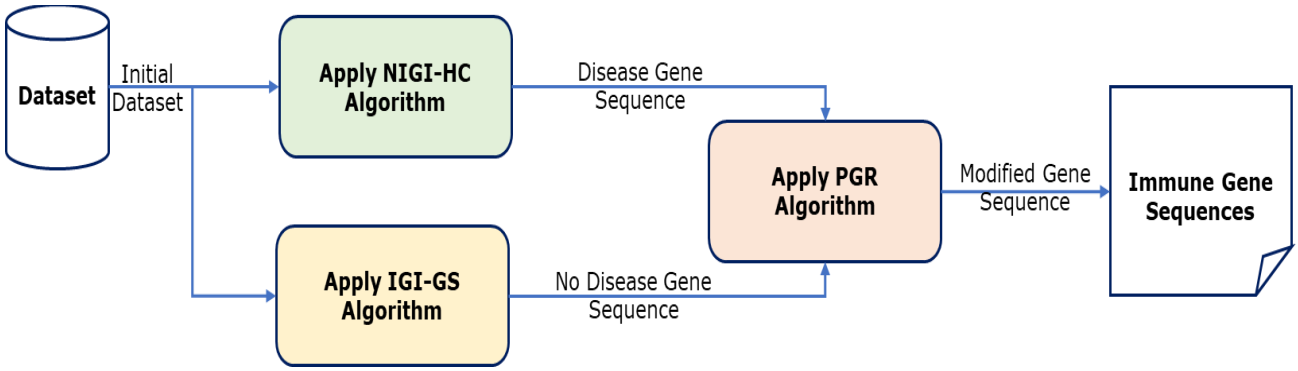

Fig. 1. Proposed Automatic Gene Identification and Replacement Framework using Machine Learning

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

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 graphically here [Fig – 2].

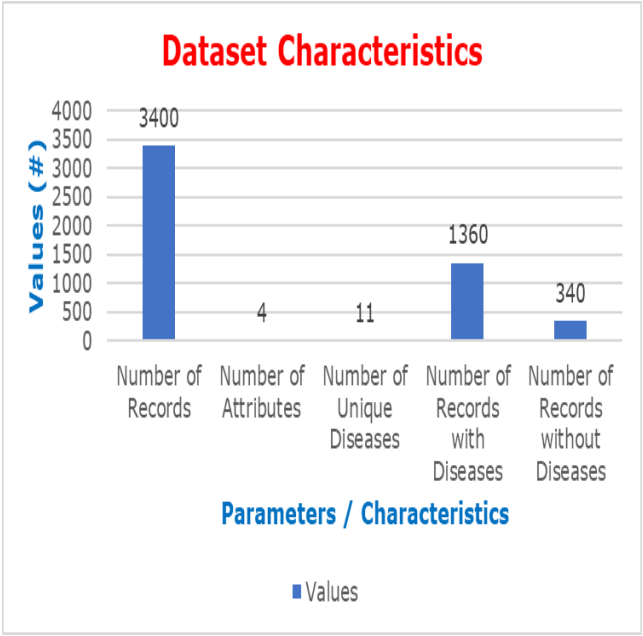

Fig. 2. Initial Dataset Characteristics

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

Secondly, the immune gene sequences are analyzed here [Table – 2].

TABLE II. Immune Gene Sequence

Disease Type	*Actual Gene Sequence*	*Immune Gene Sequence*	*Detection Accuracy (%)*
Cucumoviruses	CTGGAAATCTAAGATGGCTTGCAATCAAAAAACTGGACATTATGCGGA	CTGGAAATCTAAGATGGCTTGCAA	97
late blight	CATTTGCTTCGACTGAGGCAACCCTCTTGAAATGGAAAGTCAAGAACCATAATT	CATTTGCTTCGACTGAGGCAACCCTCT	98
Speck	CTGGAAATCTAAGATGGCTTGCAATCAAAAAACTGGACATTATGCGGA	CTGGAAATCTAAGATGGCTTGCAA	98
Canker	CTTTTTGGCTTCATGGATTCCAAGTAATGCCAAGGACTGGTATGGAGTTGT	CTTTTTGGCTTCATGGATTCCAAGT	97
Mosaic	CATTTGCTTCGACTGAGGCAACCCTCTTGAAATGGAAAGTCAAGAACCATAATT	CATTTGCTTCGACTGAGGCAACCCTCT	98
Tobacco Streak	ATGGTTTCTAGAAAAGTAGTCTCACTTCAGTTTTTCACTTACCTCACA	ATGGTTTCTAGAAAAGTAGTCTCA	97

eISSN 1303-5150

Disease Type	*Actual Gene Sequence*	*Immune Gene Sequence*	*Detection Accuracy (%)*
Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	TTTTGATATGCAGAACAAACTTTCTGG	97
Fusarium	ATTCATATGAAGGTAGATTACGTGATCCAGTTTCAAGTTGCACTGTGT	ATTCATATGAAGGTAGATTACGTG	98
Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	TTTTGATATGCAGAACAAACTTTCTGG	98
Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	TTTTGATATGCAGAACAAACTTTCTGG	98

The mean accuracy for the testing dataset with 3400 records are 97.6%.

Further, the results are visualized graphically here [Fig – 3].

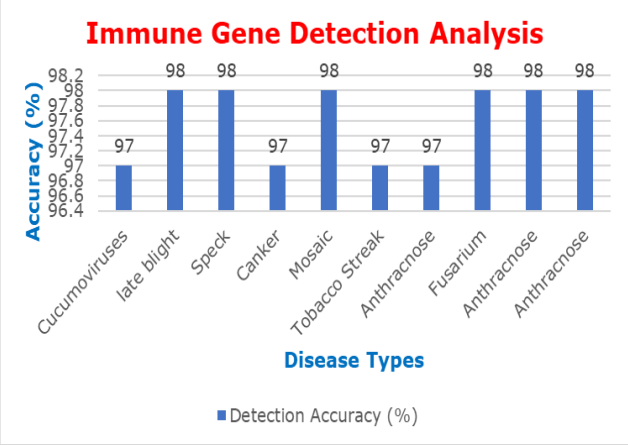

Fig. 3. Immune Gene Detection Accuracy Analysis

Further, the replaceable gene sequence identification accuracy is analyzed [Table – 3].

TABLE III. REPLACEABLE GENE SEQUENCE IDENTIFICATION

Disease Type	*Actual Gene Sequence*	*Gene Sequence for Replacement*	*Detection Accuracy (%)*
Cucumoviruses	CTGGAAATCTAAGATGGCTTGCAATCAAAAAACTGGACATTATGCGGA	TCAAAAAACTGGACATTATGCGGA	100
late blight	CATTTGCTTCGACTGAGGCAACCCTCTTGAAATGGAAAGTCAAGAACCATAATT	TGAAATGGAAAGTCAAGAACCATAATT	100
Speck	CTGGAAATCTAAGATGGCTTGCAATCAAAAAACTGGACATTATGCGGA	TCAAAAAACTGGACATTATGCGGA	100
Canker	CTTTTTGGCTTCATGGATTCCAAGTAATGCCAAGGACTGGTATGGAGTTGT	AATGCCAAGGACTGGTATGGAGTTGT	100
Mosaic	CATTTGCTTCGACTGAGGCAACCCTCTTGAAATGGAAAGTCAAGAACCATAATT	TGAAATGGAAAGTCAAGAACCATAATT	99
Tobacco Streak	ATGGTTTCTAGAAAAGTAGTCTCACTTCAGTTTTTCACTTACCTCACA	CTTCAGTTTTTCACTTACCTCACA	99
Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	GACTCTTCCAACAAATAGCATATGGAT	99
Fusarium	ATTCATATGAAGGTAGATTACGTGATCCAGTTTCAAGTTGCACTGTGT	ATCCAGTTTCAAGTTGCACTGTGT	99
Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	GACTCTTCCAACAAATAGCATATGGAT	99
Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	GACTCTTCCAACAAATAGCATATGGAT	100

The mean accuracy for the testing dataset with 3400 records are 99.5%. Further, the results are visualized graphically here [Fig – 4].

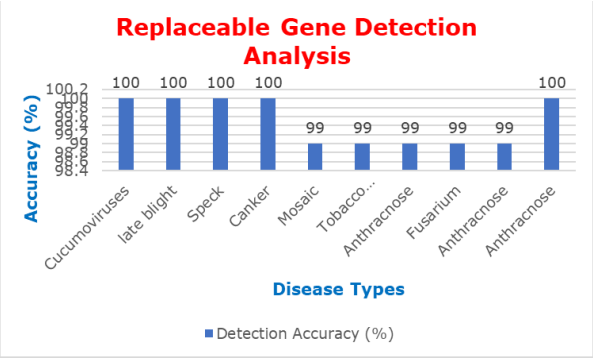

Fig. 4. Replaceable Gene Detection Accuracy Analysis

Further, the replaced and new gene sequence results are furnished here [Table – 4].

TABLE IV. IMPROVED GENE SEQUENCE

Disease Type	*Actual Gene Sequence*	*New Gene Sequence after Replacement*
Cucumoviruses	CTGGAAATCTAAGATGGCTTGCAATCAAAAAACTGGACATTATGCGGA	CTGGAAATCTAAGATGGCTTGCAACTGGAAATCTAAGATGGCTTGCAA
late blight	CATTTGCTTCGACTGAGGCAACCCTCTTGAAATGGAAAGTCAAGAACCATAATT	CATTTGCTTCGACTGAGGCAACCCTCTCATTTGCTTCGACTGAGGCAACCCTCT
Speck	CTGGAAATCTAAGATGGCTTGCAATCAAAAAACTGGACATTATGCGGA	CTGGAAATCTAAGATGGCTTGCAACTGGAAATCTAAGATGGCTTGCAA
Canker	CTTTTTGGCTTCATGGATTCCAAGTAATGCCAAGGACTGGTATGGAGTTGT	CTTTTTGGCTTCATGGATTCCAAGTCTTTTTGGCTTCATGGATTCCAAGT
Mosaic	CATTTGCTTCGACTGAGGCAACCCTCTTGAAATGGAAAGTCAAGAACCATAATT	CATTTGCTTCGACTGAGGCAACCCTCTCATTTGCTTCGACTGAGGCAACCCTCT
Tobacco Streak	ATGGTTTCTAGAAAAGTAGTCTCACTTCAGTTTTTCACTTACCTCACA	ATGGTTTCTAGAAAAGTAGTCTCAATGGTTTCTAGAAAAGTAGTCTCA
Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	TTTTGATATGCAGAACAAACTTTCTGGTTTTGATATGCAGAACAAACTTTCTGG
Fusarium	ATTCATATGAAGGTAGATTACGTGATCCAGTTTCAAGTTGCACTGTGT	ATTCATATGAAGGTAGATTACGTGATTCATATGAAGGTAGATTACGTG
Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	TTTTGATATGCAGAACAAACTTTCTGGTTTTGATATGCAGAACAAACTTTCTGG
Anthracnose	TTTTGATATGCAGAACAAACTTTCTGGGACTCTTCCAACAAATAGCATATGGAT	TTTTGATATGCAGAACAAACTTTCTGGTTTTGATATGCAGAACAAACTTTCTGG

Further, the obtained results are compared with the parallel and recent research outcomes, in the next section of this work.

VIII. COMPARATIVE ANALYSIS

After the detailed analysis of the obtained results, in this section of this work, the results are compared with the parallel and significant research outcomes.

TABLE V. COMPARATIVE ANALYSIS

Author, Year	*Model Complexity*	*Immune Gene Identification Accuracy (%) - Mean*	*Gene Replacement Accuracy (%) – Mean*
S. Bhatt et al. [8], 2022	$O(n^2)$	98.56	98.60
I. Farouk et al. [11], 2020	$O(n^2)$	97.58	96.35
P. Ihalagedara et al. [14], 2020	$O(n^2)$	96.32	97.83
Proposed Method	O(n)	97.6	99.6

It is natural to realize that the proposed method has outperformed the parallel and recent research outcomes.

Further, the final research conclusions are presented in the next section of this work.

IX. CONCLUSION

India, like many other developing nations, relies heavily on agriculture and related businesses. However, the fertilization sector is where agricultural research now stands. However, the industry's considerable disruption may be mitigated by lessening the costs associated with crop failures brought on by a wide range of plant diseases. So, the purpose of this study is to compute the genetic improvement of plant genes. Firstly, the Non-Immune Gene Identification using Hill Climbing (NIGI-HC) Algorithm identifies the non-immune genes for any given plant with 97.6% accuracy, secondly, the Immune Gene Identification using Genetic Search (IGI-GS) Algorithm identifies immune genes to replace with the non-immune genes with 99.6% accuracy and finally, the Positional Gene Replacement (PGR) Algorithm finalized the gene sequences after replacements.

REFERENCES

[1] P. Wang, D. Wang and J. Lü, "Controllability Analysis of a Gene Network for Arabidopsis thaliana Reveals Characteristics of Functional Gene Families," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 3, pp. 912-924, 1 May-June 2019.

[2] S. Sahu, A. R. Rao, K. C. Bansal, S. K. Muthusamy and V. Chinnusamy, "Genome-wide analysis and identification of abiotic stress responsive transcription factor family genes and miRNAs in bread wheat (Triticumaestivum L.): Genomic study of bread wheat," 2016 International Conference on Bioinformatics and Systems Biology (BSB), Allahabad, India, 2016, pp. 1-4.

[3] M. S. M. Asaari, S. Mertens, S. Dhondt, N. Wuyts and P. Scheunders, "Detection of Plant Responses to Drought using Close-Range Hyperspectral Imaging in a High-Throughput Phenotyping Platform," 2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Amsterdam, Netherlands, 2018, pp. 1-5.

[4] H. M. S. D. Herath, A. R. Weerasinghe and C. R. Wijesinghe, "Constructing and analyzing gene regulatory networks in leaf senescence of arabidopsis thalina," 2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2017, pp. 1-8.

[5] J. Dale, J. Matta, S. Howard, G. Ercal, W. Qiu and T. Obafemi-Ajayi, "Analysis of grapevine gene expression data using node-based resilience clustering," 2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), St. Louis, MO, USA, 2018, pp. 1-8.

[6] 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.

[7] P. Wang and D. Wang, "Gene differential co-expression networks based on RNA-seq: construction and its applications," in IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[8] S. Bhatt et al., "Deciphering Key Genes and miRNAs Associated With Hepatocellular Carcinoma via Network-Based Approach," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 2, pp. 843-853, 1 March-April 2022.

[9] B. Abu-Jamous and S. Kelly, "Simultaneous clustering of multiple heterogeneous gene expression datasets," 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Aqaba, Jordan, 2017, pp. 1-6.

[10] N. Rai et al., "Genome analysis of Rhizobium species using codon usage bias tools," 2016 International Conference on Bioinformatics and Systems Biology (BSB), Allahabad, India, 2016, pp. 1-4.

[11] I. Farouk et al., "Dissection of QTL linked to grain yield and identification of candidate genes involved in grain yield formation using comparative SNP sequences analysis," 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, 2020, pp. 1-6.

[12] Q. Kang, N. Moore, C. L. Schardl and R. Yoshida, "CURatio: Genome-Wide Phylogenomic Analysis Method Using Ratios of Total Branch Lengths," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 3, pp. 981-989, 1 May-June 2020.

[13] A. Stoyanova-Doycheva, E. Doychev, S. Stoyanov and A. Toskova, "An Intelligent Gene Bank Management System," 2020 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 2020, pp. 1-5.

[14] 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.

[15] W. Qin and Q. Zhu, "Network analysis with PubChem BioAssay: Preliminary study on drug abuse," 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV, USA, 2016, pp. 248-251.

[16] D. Nigam et al., "Meta-analysis of potential miRNA in Triticum astivum reveals their genome biased association with different metabolisms EST based potential miRNA identification in wheat," 2016 International Conference on Bioinformatics and Systems Biology (BSB), Allahabad, India, 2016, pp. 1-5.