Plant Protein Classification Using K-mer Encoding

K. Veningston[1(✉)], P. V. Venkateswara Rao[2], M. Pravallika Devi[1], S. Pranitha Reddy[1], and M. Ronalda[1]

[1] Department of Computer Science and Engineering, National Institute of Technology Srinagar, 190006, Srinagar, Jammu and Kashmir, India
veningstonk@gmail.com

[2] Department of Computer Science and Engineering, GITAM School of Technology, Visakhapatnam, Andhra Pradesh 530045, India

Abstract. Proteins play an important role in the human body and in plants. A lack of expertise in protein labeling in plants can make it extremely difficult to characterize and comprehend the precise roles and activities of different proteins. Furthermore, it restricts development in fields like biotechnology, disease resistance, and crop enhancement. The presented project focuses on plant protein classification, aiming to overcome the challenges arising from limited protein labeling knowledge. Advanced machine learning techniques, including various classification algorithms such as Logistic Regression, Decision Tree, K-nearest neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), Multinomial Naive Bayes (NB), AdaBoost, and XGBoost, are employed to accurately classify protein sequences into their respective families. This classification approach provides valuable insights into the functions and roles of proteins within plants, ultimately advancing our understanding of plant biology. This attempt offers new possibilities for advancement in critical sectors such as agriculture, drug discovery, and genomic research by eliminating the limitations associated with limited protein labeling knowledge.

Keywords: Genome Annotation · Peach (Prunus persica) Genome · Machine Learning · Protein Classification · K-mer Encoding

1 Introduction

Proteins are macromolecules that play crucial roles in the body. They are used to structure the body's organs and help in regulation and body functioning. They are composed of hundreds of amino acids that are covalently connected to each other. There are amino acids of 20 types, and the sequence predicts the protein's 3D structure and corresponding function. Combining four nucleotide bases ('*A*', '*T*', '*G*', and '*C*') results in coded amino acids. Proteins play a crucial role in plant growth and development. They provide a variety of functions, including photosynthesis, biosynthesis, transportation, immunology, etc. A diverse range of proteins, including enzymes, structural proteins, and storage proteins,

R. Muthalagu et al. (Eds.): CINS 2023, CCIS 1978, pp. 90–104, 2024.
https://doi.org/10.1007/978-3-031-48984-6_8

are produced by plants. *Enzymes* act as catalysts and are crucial for plant metabolism, *Structural Proteins* provides support and shape to plant cells, and *Storage proteins* stores amino acids for later use. Deoxyribonucleic acid (DNA) is a macromolecule composed of two polynucleotide chains that form a double helix structure by coiling around each other as shown in Fig. 1. It carries the genetic code, which is essential for all living things, to develop, function, grow, and reproduce. The two DNA strands are made up of nucleotides further stated to as polynucleotides. Each nucleotide is made up of a phosphate group, a deoxyribose sugar, and one of the four nitrogen-containing nucleobases (*A*, *T*, *G*, and *C*). Since more than 98% of human DNA is non-coding, it cannot serve as a guide for the construction of protein sequences. Due to the fact that they move in opposing directions, the two DNA strands are antiparallel. Transcription is the process of converting DNA nucleotides into ribonucleic acid (RNA) strands, with the exception of thymine (T), for which RNA substitutes uracil (U). Translation is the process by which these RNA strands use genetic instructions to direct the arrangement of amino acids in proteins.

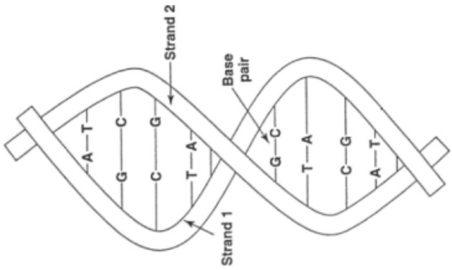

Fig. 1. DNA Double Helix Structure

1.1 Gene Expression

The genetic information of a cell is stored in DNA or RNA chemical form. A *Gene* is a portion or a segment of a DNA molecule and the *Genome* is the entire DNA in a cell. Each cell has a complex and closely controlled process from gene to protein. Transcription and translation are the two main phases. Transcription and translation work together to produce gene expression. The information contained in a gene's DNA is transferred to a similar molecule called RNA in the cell nucleus during the transcription process. Since messenger RNA (mRNA) transmits information from the nucleus to the cytoplasm, it is this type of RNA that contains the instructions needed to make a protein. The cell wall is where translation, i.e., the second step in turning a gene into a protein, takes place. The ribosome and the mRNA interact, and the ribosome "reads" the nucleotide sequence of the mRNA. Three nucleotide sequences make up a codon, which typically codes for one amino acid. One amino acid at a time, transfer RNA (tRNA), a kind of RNA, that assembles the protein. Until the ribosome encounters a "stop" codon (a three-nucleotide sequence that does not code for an amino acid), protein production continues. The transmission of information from DNA through RNA to proteins is one of the fundamental ideas in molecular biology ("central dogma").

1.2 Genome Annotation

Giving the sequence of the genome is never enough we need to know which part of the genome is coding and which part is non-coding as shown in Fig. 2. Here comes Genome Annotation, which is the process of describing the structure and function of a genome. Gene sequencing is the process of determining the nucleic acid sequence, or the arrangement of the nucleotides in DNA. Annotating a genome consists of three major steps; (*1*) Identifying the genome regions that do not contain protein-coding genes, (*2*) Gene Prediction - Process of identifying genome elements, and (*3*) Linking biological information to the elements.

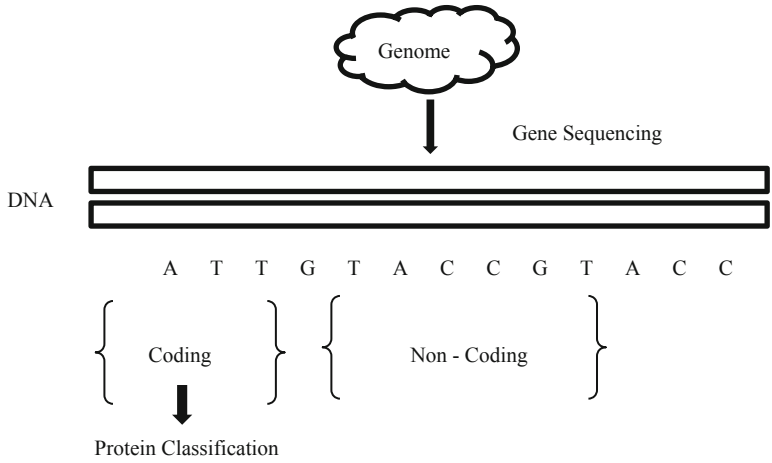

Fig. 2. Genome Annotation

1.3 Contributions

- The classification problem addressed in this paper helps gain insights into the functions and roles of different proteins within plants (specifically the *peach* plant), which has implications in agricultural genomic research.
- It demonstrates the critical role of proteins in *peach* plants and the challenges stemming from limited protein labeling knowledge. By focusing on plant protein classification, it aims to bridge this gap, which is important for understanding plant biology and precision agriculture.
- K-mer encoding represents plant genomic sequences by breaking them into fixed-length subsequences (K-mers), providing a condensed numerical representation for computational analysis of plant genomes.
- By circumventing the constraints of limited protein labeling knowledge, this paper enables informed decisions in agriculture in terms of disease resistance, crop enhancement, and allied fields dependent on a deep understanding of plant proteins.

2 Related Works

There have been several works related to plant protein classification that explore different approaches and techniques. These are a few examples of works related to plant protein classification. The field is evolving, and the efficiency of classification algorithms has been improving for identifying plant proteins.

2.1 Protein Classification Using Genome Neighborhood

Genomic neighborhood information is employed to find photosynthetic proteins and take the UniProt Knowledge Base/Swiss-Prot database [1]. Genome neighborhood network (GeNN) outperformed Random Forest and decision tree methods with an accuracy of 87%. As evidence of the model's potential to increase photosynthetic efficiency, it also showed its capacity to recognize novel photosynthetic proteins. The key outcome of this is the functional relationship between the neighbors of photosynthetic genes.

2.2 Classification of Plant Transcription Factor Proteins

Science still faces difficulties in categorizing amino acid sequences and comprehending the connection between amino acids and protein synthesis. To categorize the amino acid sequences of unidentified species, this study exploits the plant transcription factor database [1]. The model successfully classified transcription factor proteins in the kingdom of plants with a high success rate of 98.23% throughout tests. The hybrid model performs better than traditional long short-term memory – Convolutional Neural Network (LSTM – CNN)-based models due to its lightweight layers and shorter training time. The suggested model is improved by the usage of Word2Vec vectors.

2.3 Plant Allergenic Protein Prediction Based on Sequence

This paper comprehends the allergenic properties of dietary proteins. By combining supervised and unsupervised machine learning approaches, it is possible to predict if plant proteins may cause allergic reactions [2]. The method comprises rating descriptors and evaluating the effectiveness of their categorization. An SVM is utilized for partitioning, while a *k*-nearest neighbor (KNN) classifier is used for classification. For variable selection and final classification, the cross-validation (CV = 5) method is used to validate the KNN classifier. In order to address food allergies, the study emphasizes the necessity for a reliable and practical protein classification system.

2.4 Evaluating Plant Gene Models

True gene is an ML approach developed for gene model classification and reduces false positive annotations in gene prediction [3]. The Legume Information System (LIS) provided the Pisum annotated gene and protein dataset. The National Center for Biotechnology Information- Non-Redundant (NCBI-NR) database was compared to the annotated genes. It makes use of 41 protein-based traits, including amino acid and nucleotide sequences, and 14 genes. The Pisum genome was used for eXtreme Gradient Boosting (XGBoost) model training, which resulted in optimized models with 87–90% prediction accuracy and F-1 scores of 0.91–0.94.

2.5 Plant Vacuole Protein Prediction Based on Sequence

Prediction of subcellular localization is essential for comprehending gene functions in proteomes. This study addresses this issue by developing various compositions and position-specific scoring matrix (PSSM) based models, resulting in improved accuracy compared to previous methods [4]. They used the UniProt Knowledge Base/SwissProt database. The best model achieved approximately 63% accuracy on a blind dataset, surpassing current tools. To make the models accessible, they developed 'VacPred' [GUI-based software], compatible with Windows and Linux platforms. They reported an accuracy of 86.49%/87.84% and a sensitivity of 90.54%/93.24%.

2.6 DNA Sequence Classification using K-mer Counting

DNA sequences need to be classified in genomic research to determine the applicability of a new protein. This study utilizes machine learning algorithms to identify classes of DNA sequences based on nucleotide sequences. The open DNA sequence dataset was used to obtain the gene sequence dataset. Different datasets representing gene families are analyzed using substrings of defined lengths (determined by the k value) to capture sequence patterns. When they took the human dataset $k = 6$ they obtained the best accuracy 98.4%, and Precision 98.4%. The researchers obtained a gene sequence dataset about chimpanzees, dogs, and humans. The classification of DNA sequences within the framework of the analysis of biomedical data using deep learning (DL) has the ability to extract pertinent features from the input data [12]. They specifically used two different architectures namely CNN-LSTM, and CNN-Bi LSTM (Bidirectional LSTM). Both Label encoding and K-mer encoding strategies [5, 6, 8] were used for representing the DNA sequences. When the models were tested using a variety of classification criteria, the CNN and CNN-Bi LSTM with K-mer encoding both demonstrated good accuracy, scoring 93.16% and 93.13% on the testing data.

2.7 Long Terminal Repeats (LTR) Retrotransposons Classification Using K-mer Method

Although LTR retrotransposons are prevalent repeating sequences in plant genomes, the classification often involves laborious, manual procedures [7]. To overcome the problem, the researchers created a technique for classifying LTR retrotransposons and mapping them into certain families using K-mer-based ML algorithms. They used InpactorDB, which contains 67,241 LTR retrotransposon sequences from 195 plant species that have been systematically categorized into families. This dataset included sequences from Repbase, RepetDB, and Plant Genome and Systems Biology (PGSB). Their approach achieved an impressive 95% F1-Score.

2.8 Plant Protein Classification using Ensemble Classifiers

The method for forecasting the subcellular localization using several classifiers is enhanced in this research. The authors suggested an ensemble ML approach based on average voting [9, 13]. The dataset for testing and training was obtained from PlantmSubP. They gather different features appropriate for each sort of localization, use feature selection to lessen dimensionality, and then train three different models. According to experimental findings, using the testing dataset, the suggested ensemble technique could correctly classify objects in 11 compartments with an accuracy of 84.58%.

3 Proposed Model

Plant protein classification using K-mer encoding involves representing protein sequences as fixed-length feature vectors based on the occurrence frequencies of subsequences called K-mers.

3.1 Problem Statement

Build a robust classification model that correctly maps amino acid sequences to protein families within the PlantGDB Database [11]. In this study, several models were used to solve the above problem. The model analyzes the input sequence data and predicts the protein class it belongs to with high accuracy and precision utilizing Machine Learning algorithms. By proposing an efficient and automated approach for protein family classification, this work aims to improve knowledge of protein structure and function.

3.2 Dataset Description

The PlantGDB is a data repository containing different plant species' genome sequences. In addition to sequence data, it also provides alignments and annotations [10]. '*Prunus persica [peach] genome*' comes with a *peptide file* (Ppersica_139_peptide.fa.gz) and the *annotation file* (Ppersica_139_annotation_info.text.gz). Table 1 shows the dataset statistics.

Table 1. Dataset statistics

Data characteristics	Count
# Samples	3824
# Features (length of the sequence)	1577562
# Classes	34

3.3 Dataset Preparation

The Peptide file considered for preparing data is in FASTA format (text-based) and it needs to be changed to CSV format for applying Machine Learning algorithms. For reading the FASTA file SeqIO.parse() function from *BioPython* Library is used. Extracted required attributes, stored them in a pandas data frame, and finally saved them in a file named 'peptide_data.csv' as shown in Fig. 3.

```
>ppa010779m|PACid:17640292
MAASNSGTVSLKLLIDTTRNKVLFAEAGKDFVDFLFTLLSLPAGTIIRLLSKDAMVGSLGKLYESVETLNDEYLQPNLNK
DTLLKPKEPVAAGPNLLGLLTDVKSDAPKTIYRCSNGSCIYRISYVADDPKAICPGCKHSMTTAVTYVASPTSEVQATYS
GAGKVGYVKGLVTYMIMDDLEVKPLSTISCITLLSRFSVKDVGVLEEKVVDLGMDEGVKLLKASLQTKSVLTQVFLR*
>ppa004704m|PACid:17640293
MSSHLYGYGATQSAAAAAAATAGLSSVYTSRTLTDPTLRYLSGSDPFASATDHHSRSSSMYLATSHLMSQSSWPAPDVEP
GVPGVKRPSEALYHQSFMGAYNTIGQEAWYSALAKRPRYESASNLPIYPQRPGEKDCAHYMLTRTCKFGELCKFDHPIWV
PEGGIPDWKEVPLVAPSESLPERPGEPDCPYFIKTQRCKFGMRCKFNHPKEKLAAAVASENADVFALPERPSEPPCAFYM
KTGQCKFGATCKFHHPKDIQIPSAEQENKIGETGTTIQPEGTGFAVKLPVSFSPALLYNSKELPVRPGEPDCPFYLKTGS
CKYGATCRYNHPDRYAINPPIGAISHPIVAPPAAGLNIGVINPAASIYQTLAQPTVGGGQTVYPQRFGQIECDYYMKTGE
CRFGEQCKYHHPIDRSAVTLSTTKPVQQQNVKLTLAGLPRREGVAICVYYLKTGTCKYGATCKFDHPPPGEVMGMAASQG
ASGGEANDFTHEQQQ*
```

	ID	PACId	Sequence
0	ppa010779m	17640292	MAASNSGTVSLKLLIDTTRNKVLFAEAGKDFVDFLFTLLSLPAGTI...
1	ppa004704m	17640293	MSSHLYGYGATQSAAAAAAATAGLSSVYTSRTLTDPTLRYLSGSDP...
2	ppa025459m	17640294	MDQQYSRDFPILSYILSRLDPESNPPLSPQLQETLLTQLPHLNHPK...
3	ppa025141m	17640295	MAMKQPHVIIFPFPLQGHMKPLLCLAELLCHAGLHVTYVNTHHNHQ...
4	ppa001905m	17640296	MCSSTNVPRWTPSPSPTRSLLASAVEGPSSKESHVLDDDHAEMKRQ...

Fig. 3. Peptide File Conversion (.fa to.csv)

The annotation file was in text format(tab-separated). Read the file, convert it into a *pandas* data frame, and finally save it to a CSV file named 'annotation_info.csv' as shown in Fig. 4. The problem statement lies in classifying amino acid sequences into protein families. Extract relevant features such as Sequence and ID columns from 'peptide_file.csv' and Protein families (*Defline*) and Transcript name columns from 'annotation_file.csv'. Merge these two files to make a single dataset named 'filtered_data.csv' shown in Fig. 5 that is based on common identifiers present in both files (*ID* and *Transcript name*) respectively. Protein class labels were extracted based on their occurrence count of at least 50 counts (count $>=50$), the required condition to choose labels, and their corresponding sequence and id. The final dataset named 'filtered_data.csv' consists of 3824 sequence samples over 34 different protein classes as shown in Table 2.

Convert peptide data from FASTA (.fa) format to comma separated values (.csv) format by extracting gene sequences information that ensures compatibility with various data processing tools.

Table 2. Sample Class Labels and its Description

S. No.	Class Label	Significance
1	'F-box family protein'	Responsible for vegetative and reproductive growth
2	'Ankyrin repeat family protein'	Involved in mediating protein-protein interactions
3	'Disease resistance protein'	Pathogen Recognition
4	'Leucine-rich repeat family protein'	Regulates shoot and root growth
5	'Leucine-rich repeat transmembrane protein'	Innate immunity in plants
6	'Protein kinase family protein'	Respond to environmental stresses
7	'Protein kinase, putative'	Regulate biological activity
8	'Short-chain dehydrogenase/reductase family protein'	Metabolic regulation
9	'Zinc finger protein-related'	Regulate growth and Stress adaptation
10	'Zinc knuckle family protein'	Nucleic acid and zinc ion binding

```
ppa000001m      PF07728,PF07726 PTHR22908       KOG1808                 AT1G67120.1
        ATP binding / ATPase/ nucleoside-triphosphatase/ nucleotide binding / transcription
factor binding
ppa000002m      PF02207,PF00569 PTHR21725       KOG1776                 AT3G02260.1     BIG
        BIG (BIG); binding / ubiquitin-protein ligase/ zinc ion binding
ppa000003m      PF00097 PTHR12183                                       AT5G23110.1             zinc
finger (C3HC4-type RING finger) family protein
ppa000004m      PF00169 PTHR16166       KOG1809                 AT4G17140.2
        unknown protein
ppa000005m      PF00169 PTHR16166       KOG1809                 AT1G48090.1
        phosphoinositide binding
```

	Transcript Name	PFAM	Panther	KOG	KEGG ec	KEGG Orthology	Name	Symbol	Defline
0	ppa000001m	PF07728,PF07726	PTHR22908	KOG1808	NaN	NaN	AT1G67120.1	NaN	ATP binding / ATPase/ nucleoside-triphosphatas...
1	ppa000002m	PF02207,PF00569	PTHR21725	KOG1776	NaN	NaN	AT3G02260.1	BIG	BIG (BIG); binding / ubiquitin-protein ligase/...
2	ppa000003m	PF00097	PTHR12183	NaN	NaN	NaN	AT5G23110.1	NaN	zinc finger (C3HC4-type RING finger) family pr...
3	ppa000004m	PF00169	PTHR16166	KOG1809	NaN	NaN	AT4G17140.2	NaN	unknown protein
4	ppa000005m	PF00169	PTHR16166	KOG1809	NaN	NaN	AT1G48090.1	NaN	phosphoinositide binding

Fig. 4. Annotation File Conversion (.txt to.csv)

	ID	Sequence	Defline
0	ppa000003m	MESPVATPESIFLEDFGQKVYLTRRIREVLVNYPEGTTVLKELIQN...	zinc finger (C3HC4-type RING finger) family pr...
1	ppa000015m	MATLSQAQAVKSLNKSPGRRRFVFKSFSQRLEEVEIDVFRSLDKVK...	binding
2	ppa000021m	MPDVDLPSAVDPTHHLPLQFLPSDPTPPAPTRSDPPGCTLDWLPDF...	transducin family protein / WD-40 repeat famil...
3	ppa000041m	MGTEPALLLVDIIFKTLYIYDDRGSRKAVDDIITKGLQEVAFMKSF...	binding
4	ppa000048m	MKAGSAAKLIVDALLQRFLPLARRRIETAQAQDGQYLRPSDPAYEQ...	binding

Fig. 5. Final Dataset (filtered_data.csv)

3.4 K-mer Encoding

When processing the Genome sequence, conversion from string format to numerical value is necessary, to form a matrix input for model training. The features of the existing sequence encoding methods are shown in Table 3.

Table 3. Sequence Encoding Methods

Encoding Method	Features
Sequential encoding	Encodes each character into a numeric value
One-hot encoding	Represents the categorical to binary value mapping in a binary vector format
K-mer encoding	Divide the sequence into *K*-length overlapping short sequences or segments

Table 4. Details of the Dataset Split

Dataset	Class labels included	# Search Queries
D1	'F-box family protein', 'nucleic acid binding/ribonuclease H', and 'pentatricopeptide (PPR) repeat-containing protein'	169
D2	All remaining 31 class labels	595

The problem that exists in the methods (other than *k-mer* encoding) mentioned does not result in vectors of uniform length, which is the required condition to feed data to an algorithm (classification or regression). For the other two methods, we have to curtail or fill with "*n*" or "*0*" to meet the requirement.

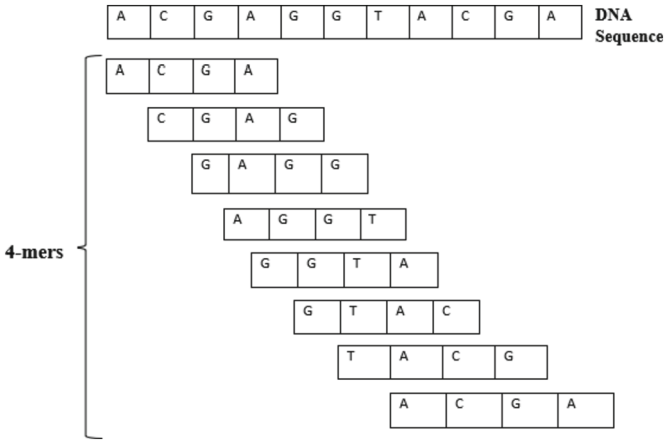

Fig. 6. *K*-mer representation [$K = 4$]

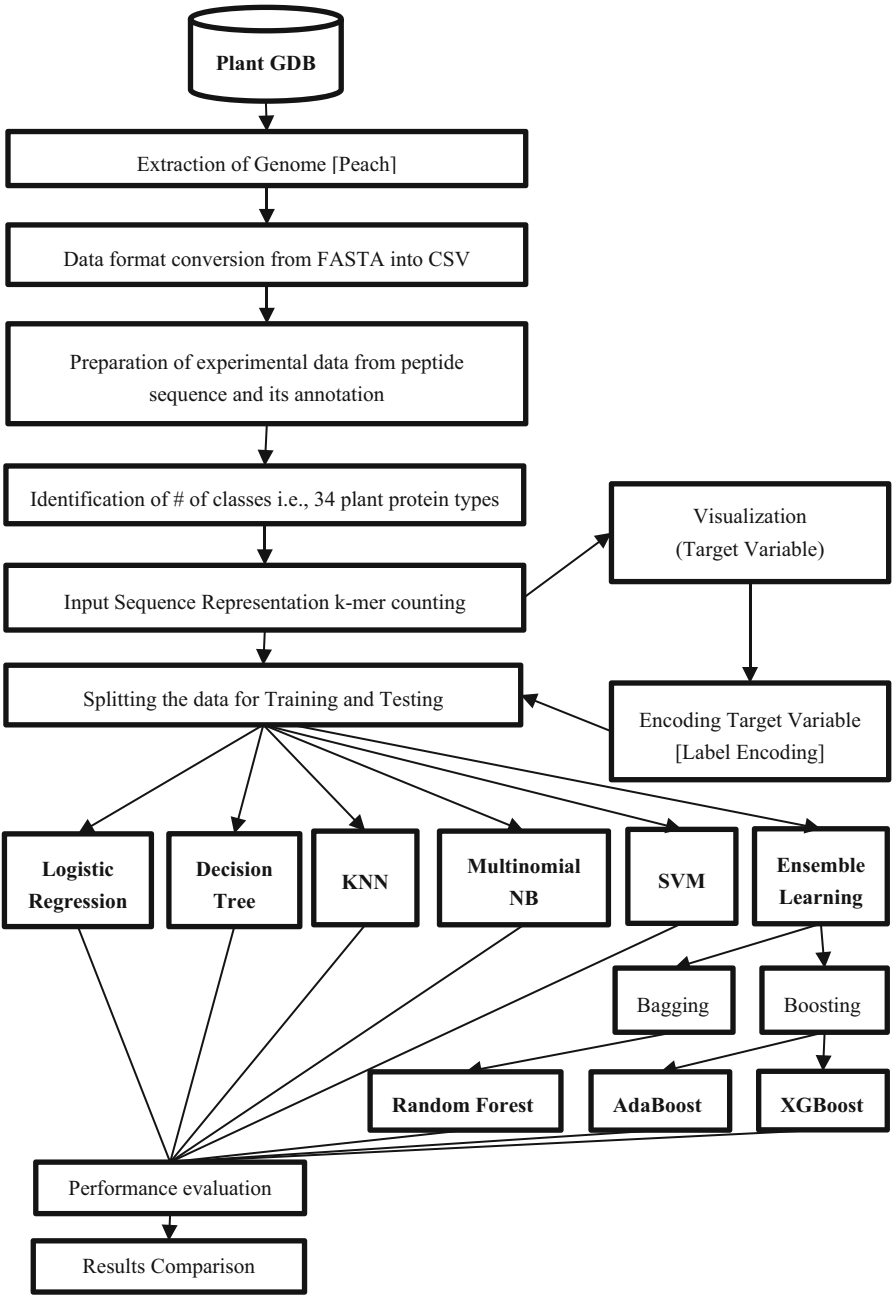

Fig. 7. Detailed Workflow

Choose the value of K, which represents the length of the subsequences (K-mers) used for encoding as shown in Fig. 6. Typically, K is set to a small value, such as 4 or

6. It is considered a hyper-parameter while training machine learning models on this representation. Using Algorithm 1, all possible 6-mers were generated from the protein sequences of peach fruit plants. Figure 7 shows the detailed workflow of the contributions made in this paper.

4 Empirical Evaluation

4.1 Algorithms

ALGORITHM 1: Generate *K*-mer representation

Input:

- A set of amino acid sequences, S = $\{s_1, s_2, \ldots, s_m\}$ where each string is represented as $s_i, 1 \leq i \leq n$ and $n = |s_i|$

Output:

- Numeric *N*-gram representation (*N*=4)

Method:

for each s_i **do**
    Divide s_i into overlapping segments of length *K* to generate *K*-mers.
    Construct a bag-of-words model using the frequency of *K*-mers.
    for each *K*-mer **do**
        count the frequency
    end for
end for
Return the *K*-mers in *document-term matrix* representation.

Generating *K-mer* representations for plant protein sequences using Algorithm 1is significant as it transforms intricate biological data into a format that can be efficiently processed by machine learning algorithms. K-mers capture sequence patterns and motifs, preserving crucial information about plant protein structure and function. This encoding method aids in feature extraction, making it easier to classify plant proteins accurately and understand their roles in plants growth, resistance against pests, and enhance crop yields. This *K-mer* representations helps in bridging the gap between biological knowledge and computational analysis, and thus opens avenues for agricultural disease resistance and genomic analytics.

4.2 Evaluation Metrics

When evaluating the performance of a protein classification task, several evaluation metrics can be used to assess the accuracy and effectiveness of the classification model. The proposed method has been evaluated using the following metrics.

***Confusion Matrix*:** It can be used to derive other evaluation metrics such as accuracy, precision, and recall.

***Accuracy*:** Accuracy measures the proportion of correctly classified protein instances over the total number of instances in the dataset. It provides an overall measure of the model's correctness.

$$Accuracy = TP + TN/(TP + TN + FP + FN) \tag{1}$$

***Precision*:** Precision represents the proportion of true positive predictions (correctly classified positives) over the total positive predictions. It indicates the model's ability to avoid false positives.

$$Precision = TP/(TP + FP) \tag{2}$$

Recall (*a.k.a.* Sensitivity or True Positive Rate): Recall measures the proportion of true positive predictions over the total actual positive instances in the dataset. It reflects the model's ability to correctly identify positive instances.

$$Recall = TP/(TP + FN) \tag{3}$$

***F1-score*:** It is the harmonic mean of *precision* and *recall*. It provides a balanced measure, which is useful when the dataset has imbalanced classes.

$$F1 - score = 2 * (precision * recall)/(precision + recall) \tag{4}$$

4.3 Experimental Evaluation Using Entire Dataset (with 34 classes)

The dataset shown in Fig. 5 is used to train the models. The detailed dataset preparation is mentioned in Sect. 3.3. The dataset consists of 764 sequence samples over 34 different protein classes as sample classes. Different machine learning algorithms were considered for building the classification model.

It shows that the Multinomial Naïve Bayes model outperformed other models while observing the results shown in Fig. 8. Analyzing various classifiers, it is observed that three class labels namely '*F-box family protein', 'nucleic acid binding/ribonuclease H', and 'pentatricopeptide (PPR) repeat-containing protein'* were assigned to most of the instances. Therefore, the dataset is split into two parts and has been trained separately such that one dataset comprises 3 classes whereas the other dataset comprises of remaining 31 classes. Due to the rise in classification errors of those 3 classes, it was decided to divide the dataset into two parts called **D1** and **D2** (Table 4).

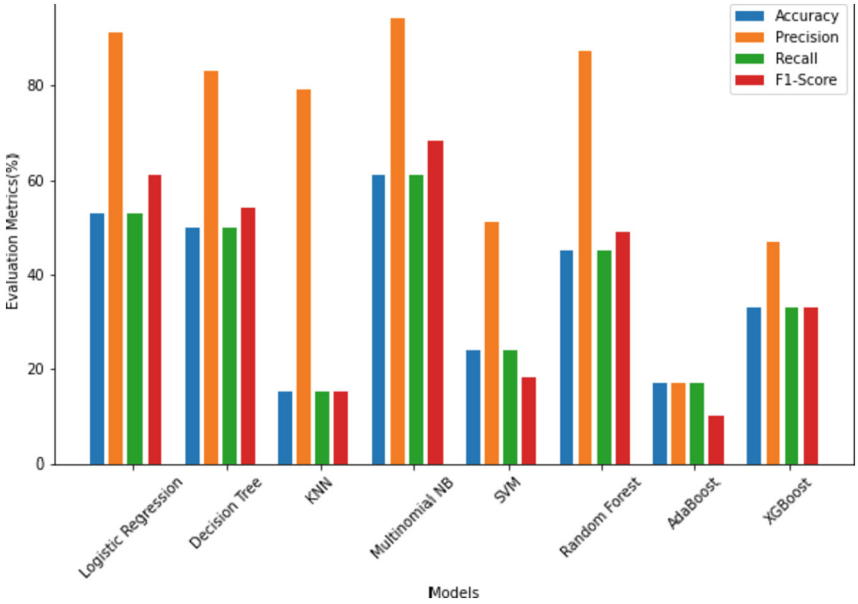

Fig. 8. Evaluation scores of 7 ML models considered for experimentation

4.4 Performance Evaluation on D1 and D2

It is noted that the Support Vector Machine (SVM) outperformed on dataset **D1** that addresses 3-class classification task, while Multinomial NB outperformed on dataset **D2** that addresses 31-class complex classification task from Tables 5 and 6 respectively. Assessing complex protein classes entails rigorous analysis of their structure, function, and interactions. This process involves data integration, and computational modeling that enables deeper insights into plant biological systems.

Table 5. Results obtained on dataset D1

Models	Accuracy	Precision	Recall	F1-score
Logistic Regression	63	81	63	60
Decision Tree	66	80	66	62
KNN	12	26	12	11
Multinomial NB	21	92	21	22
SVM*	**76**	**70**	**76**	**73**
Random Forest	56	77	56	46
AdaBoost	53	35	53	38
XGBoost	64	79	64	59

Table 6. Results obtained on dataset D2

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	50	90	50	55
Decision Tree	50	88	50	55
KNN	15	81	15	21
Multinomial NB*	66	90	66	72
SVM	27	57	27	25
Random Forest	42	83	42	47
AdaBoost	16	21	16	12
XGBoost	34	51	34	34

4.5 Tools and Libraries Used

A range of Python libraries and toolkits were harnessed for the implementation of the presented *peach* plant classification models. Particularly, the *sci-kit-learn* library was employed for modeling algorithms including Logistic Regression, Decision Tree, KNN, multinomial NB, SVM, Random Forest, AdaBoost, and XGBoost on the prepared dataset. *BioPython* library was employed in handling biological data and gene sequences from the peach genome. It provided functionalities for gene sequence manipulation, feature extraction, and data preprocessing, ensuring the effective integration of biological data and knowledge into the machine-learning pipeline.

5 Conclusions

Plant protein classification using *K*-mer encoding involves representing protein sequences as fixed-length feature vectors based on the frequencies of subsequences. This encoding method provides a compact and informative representation of protein sequences, enabling the application of ML algorithms for classification. Additionally, experimenting with different *K*-mer lengths would improve the classification accuracy of the models. It is observed that the *multinomial NB* outperforms with an F-score of 72% when evaluated on the dataset containing 31 features while *SVM* reports 73% on the dataset containing 3 features. Achieving a remarkable F-score of 72% in a complex 31-class classification problem demonstrates robust model performance, effective feature engineering, and rigorous model tuning. These developments contribute to the improvement of classification accuracy and the understanding of plant protein functions and interactions, which have significant implications in the precision agriculture domain. Further attempts could be made to interpret the model's prediction output through explainable artificial intelligence (EAI) capabilities. Explaining the output of the ML models is still an open area for research as it is considered challenging to interpret the output of an extremely complex model.

References

1. Öncül, A.B., Çelik, Y.: A hybrid deep learning model for classification of plant transcription factor proteins. SIViP **17**, 2055–2061 (2023)
2. Nedyalkova, M., Vasighi, M., Azmoon, A., Naneva, L., Simeonov, V.: Sequence-based prediction of plant allergenic proteins: machine learning classification approach. ACS Omega **8**(4), 3698–3704 (2023). https://doi.org/10.1021/acsomega.2c02842
3. Upadhyaya, S.R., et al.: Evaluating Plant Gene Models Using Machine Learning. Plants **11**(12), 1619 (2022). https://doi.org/10.3390/plants11121619
4. Yadav, A.K., Singla, D.: VacPred: Sequence-based prediction of plant vacuole proteins using machine-learning techniques. J. Biosci. **45**, 1–9 (2020)
5. Simon, O.A., et al.: K-mer-based machine learning method to classify LTR-retrotransposons in plant genomes. PeerJ **9**, e11456 (2021)
6. Warin, W., et al. Ensemble of multiple classifiers for multilabel classification of plant protein subcellular localization. Life **11.4**, 293 (2021)
7. Guo, Y., Hou, L., Zhu, W., Wang, P.: Prediction of Hormone-Binding Proteins Based on K-mer Feature Representation and Naive Bayes. Front. Genet. **12**, 797641 (2021)
8. Juneja, S., Dhankhar, A., Juneja, A., Bali, S.: An approach to DNA sequence classification through machine learning: DNA sequencing, K-Mer counting, thresholding, sequence analysis. Int. J. Reliable Qual. E-Healthcare (IJRQEH) **11**(2), 1–15 (2022)
9. Sangphukieo, A., Laomettachit, T., Ruengjitchatchawalya, M.: Photosynthetic protein classification using genome neighborhood-based machine learning feature. Sci. Rep. **10**(1), 7108 (2020)
10. Gotoh, O., Morita, M., Nelson, D.R.: Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. BMC Bioinform. **15**(1), 1–13 (2014)
11. http://plantgdb.org/PeGDB/ - Prunus persica [Peach] Genome. Accessed 14 June 2023
12. Pan, J., et al.: DWPPI: a deep learning approach for predicting protein-protein interactions in plants based on multi-source information with a large-scale biological network. Front. Bioeng. Biotechnol. **10**, 807522 (2022)
13. Li, L.-P., Zhang, B., Cheng, L.: CPIELA: computational prediction of plant protein-protein interactions by ensemble learning approach from protein sequences and evolutionary information. Front. Genet. **13**, 857839 (2022)