Heliyon 9 (2023) e20161

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Implementation of machine learning in DNA barcoding for determining the plant family taxonomy

Lala Septem Riza [a,*], Muhammad Iqbal Zain [a], Ahmad Izzuddin [a], Yudi Prasetyo [a], Topik Hidayat [b], Khyrina Airin Fariza Abu Samah [c]

[a] Department of Computer Science Education, Universitas Pendidikan Indonesia, Bandung, Indonesia
[b] Department of Biology Education, Universitas Pendidikan Indonesia, Bandung, Indonesia
[c] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Melaka Kampus Jasin, Melaka, Malaysia

ARTICLE INFO

Keywords:
DNA barcoding
Unsupervised learning
Bioinformatics
Hierarchical clustering
Machine learning
Taxonomy
R programming language

ABSTRACT

The DNA barcoding approach has been used extensively in taxonomy and phylogenetics. The differences in certain DNA sequences are able to differentiate and help classify organisms into taxa. It has been used in cases of taxonomic disputes where morphology by itself is insufficient. This research aimed to utilize hierarchical clustering, an unsupervised machine learning method, to determine and resolve disputes in plant family taxonomy. We take a case study of Leguminosae that historically some classify into three families (Fabaceae, Caesalpiniaceae, and Mimosaceae) but others classify into one family (Leguminosae). This study is divided into several phases, which are: (i) data collection, (ii) data preprocessing, (iii) finding the best distance method, and (iv) determining disputed family. The data used are collected from several sources, including National Center for Biotechnology Information (NCBI), journals, and websites. The data for validation of the methods were collected from NCBI. This was used to determine the best distance method for differentiating families or genera. The data for the case study in the Leguminosae group was collected from journals and a website. From the experiment that we have conducted, we found that the Pearson method is the best distance method to do clustering ITS sequence of plants, both in accuracy and computational cost. We use the Pearson method to determine the disputed family between Leguminosae. We found that the case study of Leguminosae should be grouped into one family based on our research.

1. Introduction

DNA barcoding is the use of *Deoxyribonucleic acid* (DNA) barcodes or specific portions of the DNA [1]. A single gene would ideally be effective in all different groupings of organisms or taxa, however, different portions of the DNA have been found to be more effective in different taxa [2]. For animals, the most effective barcode is a fragment of ~650 base pairs (bp) near the 5′-terminus of the mitochondrial cytochrome *c* oxidase I (COI) gene [3]. In fungi, the more appropriate barcode is the internal transcribed spacer (ITS) nuclear ribosome sequence [4]. In plant species, there are several difficulties with barcoding, one of which is the low nucleotide substitution rate of COI [5]. The Consortium for the Barcode of Life (CBOL) has recommended that the chloroplast ribulose-1,

* Corresponding author.
E-mail addresses: lala.s.riza@upi.edu (L.S. Riza), iqbalzain99@upi.edu (M.I. Zain), ahmadizzuddin@upi.edu (A. Izzuddin), yudiprasetyo@upi.edu (Y. Prasetyo), topikhidayat@upi.edu (T. Hidayat), khyrina783@uitm.edu.my (K.A.F. Abu Samah).

https://doi.org/10.1016/j.heliyon.2023.e20161
Received 9 October 2022; Received in revised form 5 September 2023; Accepted 13 September 2023
Available online 21 September 2023

5-bisphosphate carboxylase large subunit (*rbcL*) genes [6] and Maturase K (*matK*) [7] be used as plant barcode [8]. Another difficulty for plants is the higher identification success rate in animals compared to plants [9].

The method of obtaining DNA barcodes itself has multiple variations depending on the taxa [10]. In general, the process involves collecting a sample, isolation of DNA, matching specific primers, polymerase chain reaction (PCR), analysis of chromatogram, meeting the DNA barcoding standard, The Barcode of Life Data System (BOLD) submission, data analysis and validation, publication and data hosting, and finally the end user [2].

DNA barcoding can be helpful in many real-life applications [11]. It can be used for pest identification for biosecurity purposes to protect from potentially invasive species [12]. Authorities can use DNA barcodes to monitor the illegal trade of animals from protected species [13]. DNA barcoding has been described as being a powerful addition to the identification of wood despite the typical DNA quality of dry tissue being of middling to poor quality [14]. Aside from identification purposes, DNA barcodes can be used for grouping specimens when there is an ambiguity in the morphology, such as due to the lack of descriptions of morphological features [15]. It can also be used as a tool for determining whether unknown species should be grouped with earlier known species or as a new species based on DNA barcodes [16]. It can also be used as a supplement to other taxonomic datasets in the process of delimiting species boundaries [17].

There are several categories of computational approaches for analyzing DNA barcodes: tree-based, similarity-based, and character-based methods [18]. Other approaches include combination and alignment-free [1,19,20]. These approaches each came with their advantages and disadvantages. Similarity and tree-based methods, for example, are dependent on sequence alignment. Diagnostic or character-based methods experience more success than similarity and tree-based approaches, but the accuracies are still less than that of supervised machine learning-based approaches [1]. One such approach mapped barcode sequences into a vector based on *k*-mer frequencies and used a random forest classifier to identify sequences [21]. Several contemporary computational approaches used in DNA barcoding take the form of machine learning [22]. This is due to the complexity and variability of studies involved with genomics.

Heralded as revolutionary for taxonomic discovery, DNA Barcoding was formalized as a broader natural history tool only two decades ago [23]. The formal classification of organisms in Western science dates back to around 1753 with work by Carl Linnaeus [24]. However, the classification of different organisms itself has always presented itself in different human cultures throughout history. The classification or more accurately taxonomy proposed by Linnaeus classified organisms into different ranks with each rank becoming more specific. Since it was first conceived, this design has gone through many revisions and changes [25]. Several sources used by the scientific community define the hierarchy in the following ranks in order of most to the least homogenous: realm, subrealm, kingdom, subkingdom, phylum, subphylum, class, subclass, order, suborder, family, subfamily, genus, and subgenus [26–29]. However, due to the inconsistencies of the ranking system [25] and other factors [30], discrepancies and disputes in taxonomy also arise [31–34] such as in the case of Leguminosae [35,36].

Leguminosae is a large group of agriculturally important flowering plants. The group consists of a variety of species including herbaceous plants, shrubs, and trees [36]. Humans use legumes in various ways, including as a staple food source, animal feed, and fertilizer. Additionally, legumes are also used to synthesize many products including flavorings, drugs, poisons, and colorings. This group of Plantae is also beneficial to other plants by converting atmospheric nitrogen into nitrogen compounds which are useful in biochemical processes. Leguminosae is the third largest group in the flowering plants after Orchidaceae [37] and includes 650 genera with 18,000 species [38]. Dhakad [39] describes this group as holding an important role in biodiversity in the ecosystem and dominating a majority of vegetation types in the world. In addition, Leguminosae also holds an important role in the composition of forests and the management of sustainable goals.

The classification of Leguminosae as one family has become a disputed taxonomic grouping with experts taking several different stances on the issue. The first group of experts agrees that Leguminosae should be classified as a distinct order and subclassified into three distinct families which are Fabaceae (Papilionoid), Caesalpiniaceae, and Mimosaceae [40–42]. This includes the argument that Fabales [40,43] become an order that is subclassified into the three families mentioned above. There are several issues with this perspective. The nomenclature for Fabaceae is ambiguous as it can be used for a family but is also used just for the Papilionoids [39]. Both use cases of Fabaceae are accepted according to articles 18.5 and 18.6 in the International Code of Botanical Nomenclature [26]. Their placement stresses the close relationship between the three aforementioned families under the same order [35]. However, the placement of species and genera of Leguminosae is not systematically consistent [38]. The morphology by itself cannot ascertain phylogenetic relationships. Many species, such as *Acacia* and *Mimosa*, are hard to differentiate based on their morphological characteristics [44,45]. Species in the Mimosaceae and Caesalpiniaceae families are mostly physically similar and are consistently different from Fabaceae which is dominated by herbaceous plants [35].

The second group of experts has the opinion that Leguminosae is a family with three subfamilies of Mimosoideae, Caesalpinioideae, and Papilionoideae [46–48]. The naming changes proposed by several experts are also recorded in the International Code of Botanical Nomenclature, one of which is the change of Fabales into Fabaceae [35]. A few recent studies that covered this dispute by Mondal & Mondal [35] and Patel & Panchal [36] agree that the three groups are distinct. However, Patel & Panchal stresses that the distinction is made as different subfamilies of the same family.

This study aims to assist in clarifying the dispute on the taxonomy of Leguminosae by leveraging machine learning in the form of hierarchical clustering and DNA barcodes. Hierarchical clustering is an unsupervised technique to perform data exploratory analysis. The main aim of the technique is to build a binary merge tree [49]. This technique was the first answer to the limits of similarity-based methods [18]. A dendrogram, the visual drawing of hierarchical clustering, gives rich information for either qualitative or quantitative evaluations [49]. Thus, this visualization created from hierarchical clustering can be used to assess this study. Many studies with DNA Barcoding continue to use hierarchical clustering techniques due to its ubiquity and relative simplicity [50–56]. In this study, hierarchical clustering with DNA Barcodes will be used to achieve two objectives. First, we validate the usability of hierarchical clustering

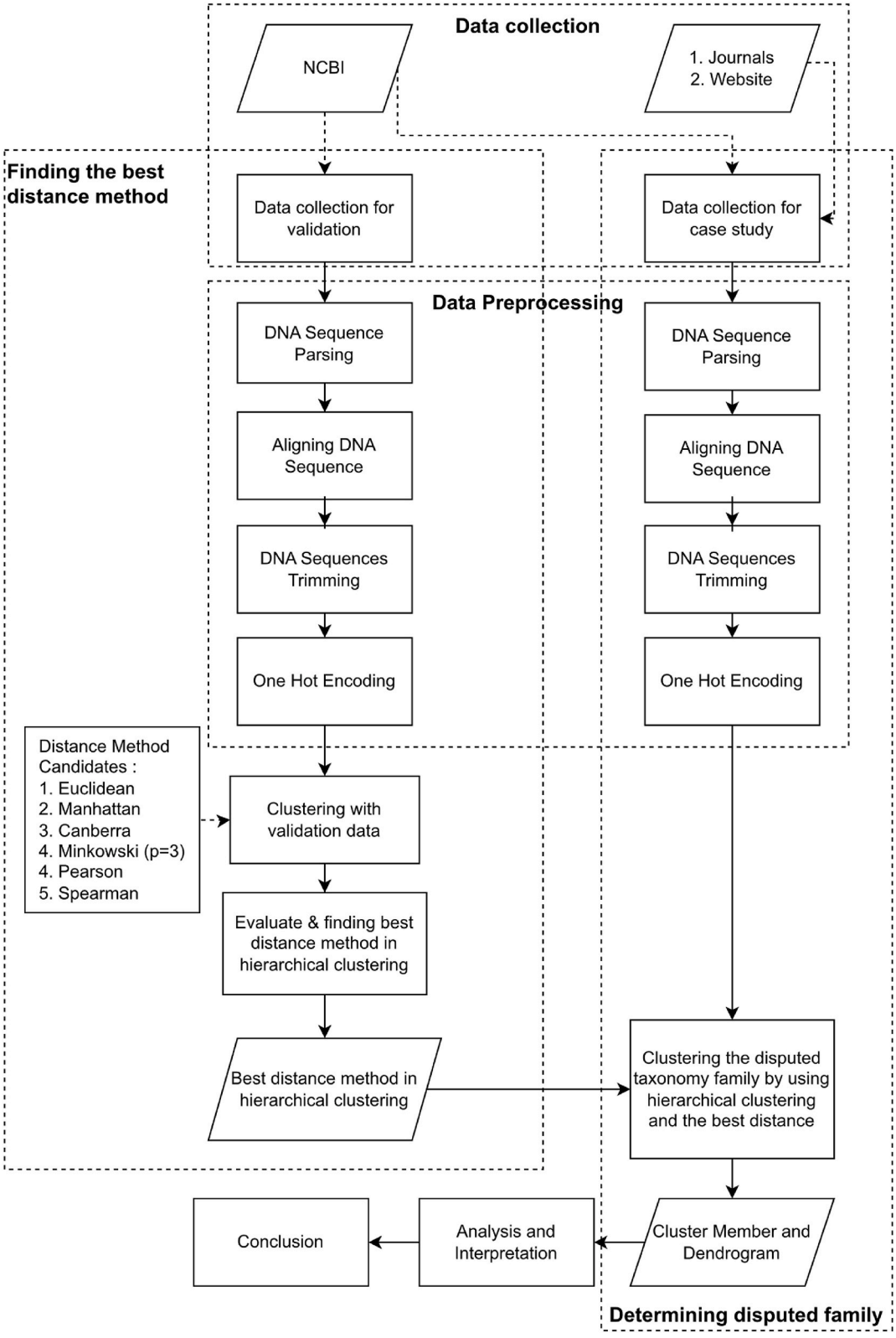

Fig. 1. The proposed computational model.

and different distance methods for the problem. Second, we utilize the validated method to determine the grouping in the taxonomy of Leguminosae. This is done to determine which view on the taxonomy of Leguminosae the results from hierarchical clustering supports. In short, we aim to clarify whether Leguminosae should be classified as three distinct families, subfamilies, or another permutation altogether.

Machine learning is naturally an interdisciplinary field. It draws on insights from a variety of disciplines, including artificial intelligence, probability and statistics, computational complexity theory, control theory, information theory, philosophy, psychology, and neurobiology. In a broad range of domains, machine learning algorithms are proven to be extremely useful, for example, in the domain of speech recognition algorithm-based machine learning outperforms any other approach that has been tried [57]. Machine learning as an approach has the advantage of learning with experience and the lack of need to manually account for the multitude of variations found in genetic data. The category of approaches in the literature does not have a consistent naming scheme. Several articles refer to machine learning as a separate category from tree-based, distance-based, and character-based [20,21,58], despite some approaches in the other categories also being machine learning, albeit unsupervised for the most part such as hierarchical clustering or in other words tree-based approaches [18,22].

2. Material and methods

2.1. Proposed computational model

This section of this paper will describe the computational model that was used in this study. This study uses R version 4.1.2, and the package used in this study is described as follows.

1. *rentrez* [59]: This is an R package used for retrieving data from NCBI. We used version 1.2.3 of this package.
2. *Biostrings* [60]: This package is run in R, the purpose of this package is for data manipulation and for dealing with biological sequences. The version that we use is 2.62.0.
3. *msa* [61]: This package is used for sequence alignment for multiple DNA sequences, the default preset used in this package is ClustalW algorithm. We use version 1.26.0 of this package.
4. *ips* [62]: *ips* is an R package for trimming the beginning and the end of the sequences. We use version 0.0.11 of this package.
5. *factoextra* [63]: This is an R package providing additional distance methods that was used in this study. The version of this package that is used is 1.0.7.

The full flow of the computational model in this study is depicted in Fig. 1. A detailed explanation of each stage is as follows.

1. Data collection: First, we obtain a list of the available families and genera on the National Center for Biotechnology Information (NCBI) Taxonomy Databases [64] (http://www.ncbi.nlm.nih.gov/taxonomy). Subsequently, we identified families and genera with a representation of more than 25 records of ITS marker sequences. From this pool of families and genera that fulfilled the criteria, 3 families or genera were taken randomly over multiple iterations. The data sequences were retrieved with the help of the *rentrez* package [59] version 1.2.3. The sequences that were retrieved are in the *FASTA* format exemplified in Fig. 2.

For the validation phase, we used a query to gather the data on individual organisms for each family and genera. The query also limited the results by the length of the DNA sequences and filtered by gene which in this case is ITS. For the case study data, we gather several references that divide Leguminales into three different groups, in this case, Fabaceae, Mimosaceae, and Caesalpiniaceae. We specifically selected sequences and their species names from websites and other previous research in GenBank to improve the quality of the sequences used in this study. The main references to the data that have been used are [35,36,65], and Plant Specimen Database Program & Publication (https://plantsp-eflora.bnh.gov.bd/family-list). Most of the sources just provide the name of the species. However, we need the corresponding DNA sequence from NCBI.

```
# Load rentrez library
library(rentrez)

# Fetching list of sequences identifier (ids)
fetch_data <- entrez_search(db="nuccore", term="Araceae[Organism] AND (internal
transcribed spacer 1[Title] OR ITS1[Title]) AND (internal transcribed spacer 2[Title] OR
ITS2[Title]) AND 500:850[SLEN] NOT UNVERIFIED",retmax=40)
# Using ids to retrieve sequences
fetchRes <- entrez_fetch(db="nuccore", id=fetch_data$ids, rettype="fasta")
```

Above is an example of retrieving data using the *rentrez* library. In the example, we try to fetch data from the Araceae family and the gene that we want to retrieve is ITS. Since the ITS gene usually ranges from 500 to 850 base pairs, we also filter the sequences based on this range of base pairs. Detailed information about the query in *entrez_search*() can be seen in this article https://www.ncbi.nlm.nih.gov/books/NBK49540/. The *retmax* parameter is to filter the maximum number of records that are retrieved. The function *entrez_fetch*() fetches the sequences from the NCBI Nucleotide database, this function returns a string in the *FASTA* format. The Nucleotide database itself is a collection of sequences from several sources, including GenBank, RefSeq, TPA, and PDB.

```
>MZ686456.1 Crataegus bretschneideri voucher JF1H internal transcribed spacer 1,
partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2,
complete sequence; and large subunit ribosomal RNA gene, partial sequence
TGCGGAAGGATCATTGTCGAACCTGCGCAGCAGAACGACCCGAGAACCAGTTTCAACGTCGGGGCCCCGG
CGGGCCTTCGGGCTCGGCGTCCCCTCGCGTCCCGGGAGCTCGCTCCCGGGCGCACAAACGAACACCGGCG
CGTGTTGCGCCAAGGAACCTGAACGAAAGACCGCGCTCCCGCCGCCCCGGAAACGGTGCGCGCGCGGGCG
CGTCGTCGTCTTCGATATGTCAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAAC
GTAGCGAAATGCGATACTTGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGC
GCCCGAAGCCACTAGGCCGAGGGCACGCCTGCCTGGGCGTCACGCGCCGTTGCCCCCCCTCGCCTCCCTC
GGGAGCGTCGGCGGGCGGAAGATGGCCTCCCGTGCGCCACCCCGCGCGGTTGGCACAAATGCCGAGTCCC
CGGCGGCGAACGCCACGACAATCGGTGGTTGACAAACCTCGGTTGCCTGTTGTGCGCTTTCGCCGCGCTG
CGGGCGGCTCGCGACCCTCGTTTGCTCTGCTTCGGCAGAGTTCTTCAACGCGACCCCAGGTCAGGCGGGG
TTACCCGCTGAATTTAAGC
```

Fig. 2. Example of FASTA format of the *Crataegus bretschneideri* species.

Fig. 2 is an example of *FASTA* formatted data. The main structure of the *FASTA* formatted data the outline of the *FASTA* format consists of two parts. The first part is the header starting with the character ">" and followed by the description of the sequence. The second part is the sequence itself which is a string composed of the characters "A", "C", "T", and "G". This research only considers the second part that is used in the computation. The length of the sequences varies depending on the part or gene that is used.

The example shown in Fig. 2 is from the *Crataegus bretschneideri* DNA sequence, and the sequence contains the internal transcribed spacer 1, 5.8 S, and internal transcribed spacer 2 genes. From the first letter of the header, we can see the "MZ686456.1" as the accession id that is used in NCBI, the detailed information on the sequence can be found at https://www.ncbi.nlm.nih.gov/nuccore/MZ686456.1.

2. Data preprocessing: The initial step of this part is to parse the DNA sequence. DNA sequence parsing is the process of parsing the data into the desired format, in this case, the *DNAStringSet* format or a collection of *DNAString*. The purpose of this change is to allow the data to work with several packages from *Biostrings*, the package allows the manipulation of large biological sequences. The *FASTA* formatted data that were gathered are then converted to *DNAStringSet* format using *Biostrings* [60] version 2.62.0.

DNAStringSet is depicted in Fig. 3. It consists of several columns, including *width* for base pair length, *seq* for sequence, and *names* for the name or label of the sequence. By default, if the *DNAStringSet* data is called to be printed, it will show the first and the last five sequences on the set, indicated by the indexing on the very left of the data.

```
# Load Biostrings library
library(Biostrings)

# Write data file to output.fasta
write(fetchRes, file="output.fasta")
# Read data from output.fasta using readDNAStringSet
dna_data <- readDNAStringSet("output.fasta")
```

To get the *DNAStringSet* data format, we can use the *readDNAStringSet*() function from *Biostrings* library [60]. The "*fetchRes*" variable is a String variable from the previous code block. We need to write the *FASTA* formatted string to some ".fasta" file, in this example, we use "output.fasta" as our *FASTA* formatted file. Afterward, we use the *readDNAStringSet*() function to read the *FASTA* file as the *DNAStringSet* format, and we save it in a variable (named *"dna_data"* in the example). This scenario will return *DNAStringSet* formatted data, saved in a variable called *"dna_data"*.

Upon parsing the data, it is essential to proceed with the alignment of sequences through DNA sequence alignment. This process involves arranging multiple data sequences in a specific manner with the aim of identifying diagnostic patterns that characterize protein families. Such alignment is instrumental in predicting the secondary and tertiary structures of new sequences and serving as an initial step in molecular evolutionary analysis [66]. The dataset that has been aligned will have the same length. The alignment is

```
DNAStringSet object of length 75:
     width seq                   names
[1]   678   AGGTGAACCT...CTGAGTTTAA MN091569.1 Callit...
[2]   564   TTCATTTGGG...GCTCGTTTTG MW303933.1 Callit...
[3]   680   GTAGGTGAAC...CTGAGTTTAA MN091529.1 Callit...
[4]   681   CGTAGGTGAA...CTGAGTTTAA MN091478.1 Callit...
[5]   685   GGTTTCCGTA...CTGAGTTTAA MN091417.1 Callit...
...   ...   ...
[71] 569    CCCGGGGACG...CTCGGACCGC GU598535.1 Cinnam...
[72] 758    AAGGAGAAGT...AGAAGAAACT MF110041.1 Cinnam...
[73] 612    TCGCGGACAT...ACCCCAGGAA KY271533.1 Cinnam...
[74] 619    ACCACCACCG...GCGCTCGGAC KU139880.1 Cinnam...
[75] 610    GGCGAACCAG...GCGATCGGAC KU139835.1 Cinnam...
```

Fig. 3. Example of *DNAStringSet* format.

performed using the *msa* package [61] version 1.26.0. We run the default preset of the *msa* package which is using the ClustalW alignment algorithm [66].

```
# Load msa library
library(msa)

# Running the sequence alignment
aligned_dna <- msa(dna_data)
```

The input data type of this package is an object of the class *XStringSet* (which includes the class *DNAStringSet*). The data used in this example is in the variable *"dna_data"* which has a *DNAStringSet* data type. The output from this code will be of the class *MsaDNAMultipleAlignment*. In addition to the regular characters representing each base, an additional "-" character is added for alignment purposes.

Afterward, we conducted DNA sequence trimming, a process that involves truncating the initial and final characters of the sequences. The objective of this process is to decrease sequence length, thereby accelerating the clustering process without adversely affecting the model's accuracy. The package that was used for this in this study is the *ips* package [62] version 0.0.11.

```
# Load msa library
library(ips)
# Load dplyr to load pipe
library(dplyr)

# Change the MsaDNAMultipleAlignment to DNAStringSet using unmasked()
# function, and changing to dataframe using as.data.frame() function
df <- unmasked(aligned_dna) %>%
  as.matrix() %>%
  as.DNAbin() %>%
  trimEnds(min.n.seq=75) %>%
  as.character() %>%
  as.data.frame()

# show results:
df <- df %>%
  unite(df,1:ncol(df),sep="")
df$x <- toupper(df$df)
df <- subset(df, select = -df )
```

The example process of DNA trimming is presented in the code block above. The input of the process above is the variable *"aligned_dna"* which is retrieved from the alignment process. The parameter used is "min.n.seq" which is set to 75. From several experiments conducted, the parameter value of 75 speeds up the clustering algorithm, without reducing the accuracy of the algorithm.

To enable us to compute the result of the sequence we do the One Hot Encoding process. One hot encoding is applied to the sequences to allow the data to be processed in the following stages. This process converts characters into numeric representations.

Fig. 4 illustrates the operation of one hot encoding. For each character in the aligned sequence, five columns are established to denote the presence of one of the five unique characters. These columns represent the bases "A", "C", "G", and "T", along with the "-" character originating from the sequence alignment. As an example, if the character being represented is "G", the "G" column will be designated with "1″, while the remaining columns will be marked as "0". This process is iteratively executed for each character in the

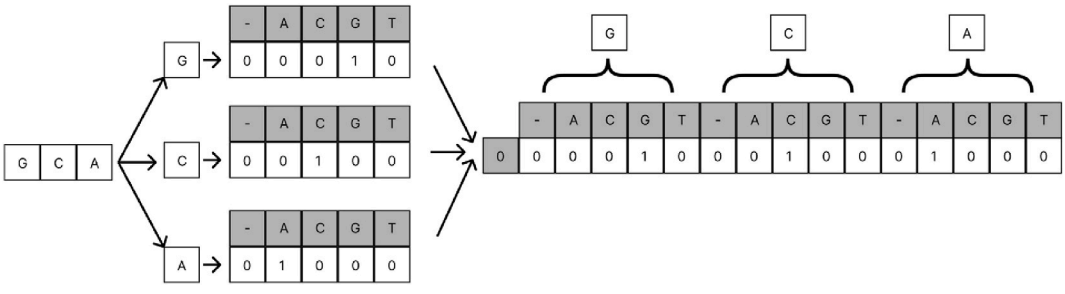

Fig. 4. One hot encoding.

sequence, and the results are then aggregated. Through this procedure, the sequence initially represented by characters is converted into a numeric format, rendering it suitable for processing in the subsequent stages.

The following is the pseudocode of One Hot Encoding:

```
Input: data frame of Multiple DNA Sequences

n <- distinct dna character (A, C, G, T, -)
df <- data frame of multiple dna sequences
seq_col <- aligned dna sequence length
seq_row <- total number of sequences
seq_mat <- matrix of 0 in seq_row x (seq_col * n) dimension
for i=1 to seq_col:
   for j=1 to seq_row:
      position <- position of df[j,i] in n
      seq_mat[j, (length(n)*(i-1))+ position] <- 1

return(seq_mat)

Output: dataframe
```

The following is the implemented code:

```
# OneHotEncoding Function
oneHotDNA <- function(df, n = c("A", "C", "G", "T")) {
  # Construct matrix
  s <- lapply(df$x,toupper)
  seq_col <- nchar(s[1])
  seq_row <- length(df$x)
  seq_mat <- matrix(data = rep(0, seq_col * length(n) * seq_row), nrow = seq_row)
  rownames(seq_mat) <- row.names(df)
  column_names = list()
  for (i in c(1:seq_col)) {
    # Encode
    for (j in c(1:seq_row)) {
      if(substr(s[j],i,i) %in% n) {
        position <- which(n == substr(s[j],i,i))
        seq_mat[j, (length(n)*(i-1))+ position] <- 1
      }
    }
    # Create Column Name
    for (j in n) {
      column_names <- append(column_names, paste(j, i,sep="-"))
    }
  }
  # Set colnames
  colnames(seq_mat) <- column_names
  return(seq_mat)
}
# df is the variable output from trimming
# Running the OneHotEncoding
oneHot_res <- oneHotDNA(df, n = c("-" , "A", "C", "G", "T"))
```

The full functionality of the one hot coding used in this study is illustrated in the code blocks above. The process converts a data frame containing a list of sequences into a *data frame* format based on the results of one-hot encoding. It requires the unmasking of the *MsaDNAMultipleAlignment* datatype back to *DnaStringSet*, followed by the use of the *as. data.frame*() function to transform the *DnaStringSet* into a *data frame*. If the data is already presented in *data frame* format, then the *oneHotDNA*() function is executed as declared above.

3. Finding the best distance method: To determine the most effective distance method for accurate clustering of biological data, we evaluated several distance methods, namely (i) Euclidean [67], (ii) Manhattan [67], (iii) Canberra [68], (iv) Minkowski [69], (v) Pearson [70], and (vi) Spearman [71] distances. We run all of the options with the ITS Marker. To perform hierarchical clustering

after obtaining the distances from the distance method, we used the *hclust*() function provided by the R programming language. The data used for this clustering consisted of established families and genera; we used genera from the same family to assess the ability of the clustering method to differentiate groups with higher similarity:

```
# Load factoextra library
library(factoextra)
# Checking for na in the data
any(is.na(oneHot_res))
# Saving list of distance method
d_met = c("euclidean", "manhattan", "canberra", "minkowski", "pearson", "spearman")
a_met = c("ward.D")
# Set the right margin (allows the text to be seen)
par(mar = par("mar") + c(0,0,0,5))

for (i in d_met){
  start_time <- Sys.time()
  hc <- get_dist(oneHot_res, method = i)
  ward_hc <-  hclust(hc, method = 'ward.D')
  end_time <- Sys.time()

  hc <- as.dendrogram(ward_hc) %>%
    set("labels_cex", 0.50)

  plot(hc, horiz=T, main =  paste(i,'ward.D',sep=" & "))
  print(paste(i,'ward.D time = ',sep=" & "))
  print(end_time - start_time)
}
```

We load the *factoextra* library [63] version 1.0.7 that provides the *get_dist*() function for using several distance methods. We loop through the different distance methods and run the clustering using them. We also capture the speed of each distance method based on the differences between the start and end times.

The following is the pseudocode of hierarchical clustering:

```
Input: Distance matrix

# compute the distance matrix
d = dataset with length n
for i=0 to n:
   for j=0 to i:
      dis_mat[i][j] = distance(d[i], d[j])

each data point represents a single cluster

repeat
   merge the two clusters having minimum distance
   update the distance matrix
until only a single cluster remains

Output: Cluster Membership
```

To validate our approach, multiple experiments were undertaken, after which the performance of each method was assessed. This assessment was centered on the distinctiveness of the groups based on the visualization of the dendrogram that was generated. These procedures were enacted with the intent to identify the most effective method for classifying both the families and the genera of the species under consideration.

We examined three distinct genera after evaluating the outcomes of grouping three distinct family groups. This dual-level validation approach was designed to determine the functioning of the clustering not only in a less homogenous taxonomic rank but also in a more homogeneous one.

Subsequently, we assessed the accuracy of each distance method. Accuracy was gauged by the ability of the method to differentiate between the three groups that were utilized in the validation process. This was performed over a series of experimental runs with varying combinations. Finally, the mean accuracy of each distance method was calculated, and the model with the highest accuracy was studied further.

Following the evaluation step, we decided on the best distance approach to use in our case study to resolve the disputed family classification. The weighted average of the accuracy from the two-tiered validation procedure was used to identify the optimal distance

approach. This selection aims to achieve the best accuracy to solve the disputed problem in the case study, which follows that the result obtained is likely to be more valid and trustworthy.

4. Determining disputed family: Upon identifying the most suitable distance method, we applied it with hierarchical clustering to assess the taxonomic family under dispute. The procedure in this section parallels the steps delineated earlier. Firstly, we initiated data collection for the disputed family. Following data acquisition, the data underwent a data preprocessing stage. The most effective distance method, as determined in the preceding steps, was then used to cluster the disputed family. Subsequent to this, an evaluation was carried out, and conclusions were drawn based on the results of the clustering process. This methodology allowed us to objectively address the taxonomic disputes.

2.2. Experimental setup

In the initial phase of our experiment, the selection of the optimal model for cluster analysis, aimed at addressing the case study issue, was paramount. Subsequently, Internal Transcribed Spacer (ITS) sequence data from individuals of undisputed taxonomic classifications were obtained from the National Center for Biotechnology Information (NCBI). The acquired data was then parsed into the *DNAStringSet* format, which subsequently facilitated string manipulation operations on the sequences. Following this, a sequence alignment was conducted to detect distinctive patterns within the data. Subsequently, One Hot Encoding was implemented to transmute our string data into a numerical format. Once the data conversion process was complete, clustering was performed using an array of selected distance methods for this study. Upon completion of this process, the results were scrutinized in order to identify the most effective distance method for implementation in hierarchical clustering on ITS sequences.

Having determined the most suitable distance method, we proceeded to conduct clustering using data from the Leguminosae family. This data was gleaned from a variety of scholarly journals and a website. The sequences retrieved were parsed into the *DNAStringSet* data type, aligned, and then subjected to the One Hot Encoding process. Finally, clustering was conducted using the selected optimal distance method in order to scrutinize the familial classification dispute within the Leguminosae family.

2.2.1. Data collection for validating the best distance method

In the initial phase, the data used in this study were retrieved from GenBank [72] (accessed August 2022). Datasets contain Internal Transcribed Spacer (ITS) from the ribosomal RNA gene of the plant. Detailed information about the data that was used in this study is explained in Table 1. All of the information in Tables 1 and 2 can be accessed at https://www.ncbi.nlm.nih.gov/nuccore using the filter.

- "species_name" [Organism] filter for filtering the family or genera of the organism.
- (internal transcribed spacer 1 [Title] OR ITS1 [Title]) AND (internal transcribed spacer 2 [Title] OR ITS2 [Title]) filter to obtain the ITS gene sequences.
- 500:850 [SLEN] filter to refine the result to the ITS gene which is generally 500 to 850 bp in length.
- NOT UNVERIFIED filter to exclude the unverified data on NCBI.

Table 1
Detailed information on validation data at the family level.

No	Family	Total Record	Average Base Pair Length	No	Family	Total Record	Average Base Pair Length
1	Cannabaceae	25	640.72	9	Alismataceae	25	707.8
	Cucurbitaceae	25	671.24		Arecaceae	25	712.8
	Zosteraceae	25	597.24		Burseraceae	25	680.96
2	Cleomaceae	25	676.88	10	Chrysobalanaceae	25	692.44
	Dilleniaceae	25	614.72		Hypericaceae	25	675.68
	Typhaceae	25	709.72		Iridaceae	25	670.52
3	Brassicaceae	25	621.08	11	Boraginaceae	25	637.88
	Hydrangeaceae	25	645.68		Convolvulaceae	25	685.16
	Linaceae	25	616.76		Haloragaceae	25	688.04
4	Buxaceae	25	650.84	12	Clusiaceae	25	693.44
	Cactaceae	25	634.24		Menispermaceae	25	596.24
	Haloragaceae	25	693.48		Zosteraceae	25	585.52
5	Iridaceae	25	666.32	13	Brassicaceae	25	639.12
	Linaceae	25	617.36		Ceratophyllaceae	25	658.08
	Malvaceae	25	697.64		Haloragaceae	25	661.2
6	Amaryllidaceae	25	645.8	14	Araliaceae	25	623.84
	Eriocaulaceae	25	743.04		Elaeocarpaceae	25	634.12
	Urticaceae	25	643.72		Meliaceae	25	676.92
7	Cymodoceaceae	25	614.64	15	Buxaceae	25	664.16
	Ericaceae	25	672.2		Ceratophyllaceae	25	655.72
	Urticaceae	25	657		Chrysobalanaceae	25	712.88
8	Ceratophyllaceae	25	644.04	**Average**		**25**	**658.6702**
	Haloragaceae	25	700.44				
	Juncaceae	25	612.84				

Table 2
Detailed information on validation data at the genera level.

No	Family	Genera	Average Base Pair Length	Total Sequences	No	Family	Genera	Average Base Pair Length	Total Sequences
1	Orchidaceae	*Anoectochilus*	693.52	25	6	Poaceae	*Brachypodium*	596.56	25
		Bulbophyllum	662.24	25			*Briza*	641.6	25
		Coelogyne	639.96	25			*Chusquea*	657.64	25
2	Orchidaceae	*Anoectochilus*	696.04	25	7	Rosaceae	*Acaena*	684.64	25
		Calopogon	675.96	25			*Alchemilla*	638.84	25
		Cypripedium	709.8	25			*Crataegus*	633.2	25
3	Orchidaceae	*Aerides*	648.2	25	8	Rosaceae	*Acaena*	677.88	25
		Anoectochilus	700.08	25			*Alchemilla*	640.64	25
		Coelogyne	631.92	25			*Amelanchier*	603.16	25
4	Asteraceae	*Ambrosia*	613.44	25	9	Sapindaceae	*Acer*	699.92	25
		Chrysanthemum	699.08	25			*Aesculus*	625.56	25
		Coreopsis	651	25			*Cardiospermum*	603.6	25
5	Brassicaceae	*Aethionema*	655.48	25	10	Ranunculaceae	*Adonis*	608.08	25
		Alyssum	668.96	25			*Caltha*	614.32	25
		Brassica	633.84	25			*Coptis*	647.88	25
Average								**651.768**	**25**

The example of the filter will look like this:

```
"(Campanulaceae[Organism]) AND (internal transcribed spacer 1[Title] OR ITS1[Title]) AND
(internal transcribed spacer 2[Title] OR ITS2[Title]) AND 500:850[SLEN] NOT UNVERIFIED".
```

After we get the result from the *rentrez* library, we take the random sample of the ids to be used.

The validation dataset contains a total of 1875 sequences of ITS markers from various families and genera among the Plantae Kingdom. The validation dataset was retrieved by using the family and the genera on the NCBI that were not disputed and contained more than 25 records. The family and the genera that were used were collected randomly from the list of eligible families. We used the data on different genera to make sure that the model that we have developed can cluster groups with higher levels of similarity or in other words are more homogenous.

2.2.2. Case study data

The case study data were collected from GenBank [72]. The dataset used in this study contains Internal Transcribed Spacer (ITS) from the ribosomal RNA. The brief information about the data is explained in Table 3 and the detailed information can be accessed in Appendix 1.

The case study family data consist of 63 data on Caesalpiniaceae, 95 data on Fabaceae, and 41 data on Mimosaceae, 199 data in total. The length of the sequences is varying from 510 bp to 785 bp and has an average length of the sequences of 672,105.

2.2.3. Hardware specification

All of the programs in this study run on an 8-core CPU computer, with a RAM capacity of 52 GB, and storage using Solid State Disk (SSD). The programming language used in this study is R version 4.1.2 which runs in **RStudio**. The libraries used can be found at **CRAN** and **Bioconductor III**.

3. Results and discussion

3.1. Validation phase

The aim of the validation phase is to get the best distance method that can cluster the DNA sequences data clearly, the method that was retrieved and then used in the case study to assess the dispute between the Leguminosae group. The distance method that is examined in this phase is Euclidean [67], Manhattan [67], Canberra [68], Minkowski [69], Pearson [70], and Spearman [71] distances. The test was run 25 times using random data from the family and genera that are not in dispute, the list of the family is shown in Table 1 and 2.

Table 3
Detailed information on case study data.

Case study data		
Family	Average bp length	Total records
Caesalpiniaceae	700.35	63
Fabaceae	668.33	95
Mimosaceae	637.46	41

Table 4
Result of the Validation phase.

	Family Validation (15 Experiments)		Genera Validation (10 Experiments)		Weighted Average	
Method	Accuracy	Time (s)	Accuracy	Time (s)	Accuracy	Time (s)
Canberra	98.22%	0.0377	99.33%	0.0328	98.67%	0.0357
Euclidean	98.22%	0.0165	**99.47%**	0.0143	98.72%	0.0156
Manhattan	98.76%	0.0164	**99.47%**	0.0146	**99.04%**	0.0157
Minkowski (p = 3)	**98.93%**	0.0566	98.53%	0.0453	98.77%	0.0521
Pearson	98.76%	**0.0160**	**99.47%**	**0.0136**	**99.04%**	**0.0150**
Spearman	98.76%	0.0645	**99.47%**	0.0528	**99.04%**	0.0598
Average	**98.61%**	**0.0346**	**99.29%**	**0.0289**	**98.88%**	**0.0323**

The summary of the Validation phase is mapped in Table 4, The table shows a summary of the accuracy and computational time of each method in each experimental run. The family validation phase consists of 15 experiments (Table 1), each experiment using 3 different families with 25 sequences each. Whereas the genera validation phase consists of 10 experiments (Table 2), each experiment using 3 different genera from the same family with 25 sequences each. The weighted average column is the weighted average of family validation and genera validation phase, calculated by:

$$waa = \frac{(fva \times 15) + (gva \times 10)}{25} \tag{1}$$

$$wat = \frac{(fvt \times 15) + (gvt \times 10)}{25} \tag{2}$$

where:

waa = Weighted average accuracy (%)
fva = Family validation accuracy (%)
gfa = Genera validation accuracy (%)
wat = Weighted average time (s).
fvt = Family validation time (s).
gft = Genera validation time (s).

The first equation (1) is used to calculate the weighted average accuracy that is used in Table 4, and the second equation (2) is used to calculate the weighted average computational time that is used in Table 4.

During the validation phase, the aggregated results indicated that the Pearson correlation emerged as the best distance method, resulting in an overall accuracy of 99.04% and an execution time of 0.0150 s. The highest accuracy was observed in Pearson, Manhattan, and Spearman methods, each achieving 99.04% accuracy, correctly classifying nearly all the cases used in this study. Any instances of misclassification could be due to anomalies or imbalances inherent in the data. However, in terms of execution time, the Pearson method proved to be the fastest, completing classification in 0.0150 s, followed closely by the Euclidean and Manhattan methods, requiring 0.0156 and 0.0157 s, respectively.

In the family validation phase, where we assessed three different families to identify the best distance method, a total of 15 test scenarios were implemented. The specifics of these tests are elaborated in Table 1. In this phase, the Minkowski method (p = 3) proved

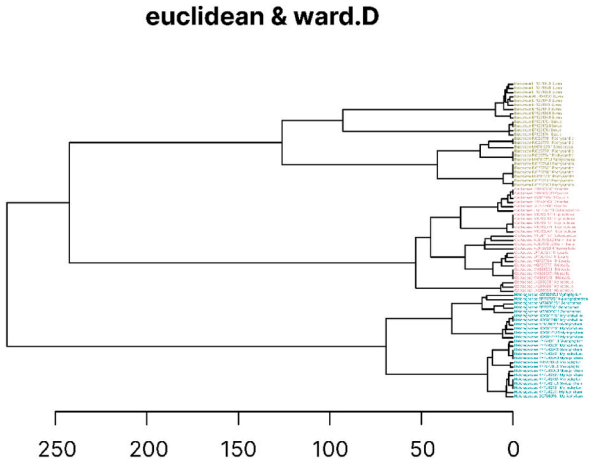

Fig. 5. Example of the well-separated families. The color represents the family of the sequence. Family: Haloragaceae (Turquoise), Cactaceae (Pink), and Buxaceae (Green).

to be the most effective for classifying the three different families, with an accuracy rate of 98.93%. Nevertheless, this method attained the lowest accuracy in the genera validation phase, scoring 98.53%. The Pearson correlation, alongside the Spearman and Manhattan distances, earned the second-highest accuracy of 98.76%. Regarding speed, the Pearson correlation method was the fastest, averaging 0.0160 s to classify different families, followed by the Manhattan and Euclidean methods, which required 0.0164 and 0.0165 s, respectively.

Genera validation is the phase that uses 3 different genera from the same family, the purpose of this phase is to examine whether the methods can classify the sequence with higher similarity, the detailed information about this phase is explained in Table 2. The overall accuracy in this phase shows a higher average accuracy than the family validation phase with 99.29% accuracy compared to the family validation phase with 98.61% accuracy. In a term of computational speed, Pearson correlation gains the fastest run time with 0.0136 s on average to cluster the different genera, followed by Euclidean and Manhattan methods at 0.0143 and 0.0146 s respectively. In contrast, although the Spearman method has high accuracy at 99.47%, it takes the longest run time to cluster the genera at 0.0528 s.

Based on the result, the ITS marker, despite being recommended for fungi, can be used to distinguish the families and the genera of the data in the experiment, and all of the distance methods give the consistently good result with more than 98.22% percent accuracy for all distance methods as long as the data used is not disputed or has problems with their taxa. We can also see that the genera validation phase is more accurate and less time-consuming than the family validation phase. Fig. 5 shows how genera of the same family are well separated using the Euclidean distance method.

The distribution of the data that we used in this validation experiment in Fig. 6 explains that most of the sequences fall into 600–675 bp (Fig. 6a) on the family validation experiments (Table 1) and 575–675 bp (Fig. 6b) on the genera validation experiments (Table 2). Fig. 5 shows us the example of visualization of the hierarchical clustering that can cluster each of the families clearly, the result is from the experiment on 3 different families: Haloragaceae (Turquoise), Cactaceae (Pink), and Buxaceae (Green).

3.2. Case study: Leguminosae

This phase aimed to resolve the disputed classification of the Leguminosae family, specifically whether it should be categorized as one or three distinct families. The data used in this study consisted of the ITS sequences of Fabaceae (Papilionoid), Caesalpiniaceae, and Mimosaceae, procured from various journals and a website featuring species from this group. If a source did not provide the NCBI accession id for the data, researchers located the corresponding sequence for that species and included it in this study. The Pearson method, due to its best accuracy and computational speed as demonstrated in the validation phase, was employed to ascertain the familial placement of the Leguminosae group.

Following the validation phase, we applied the most effective distance method to perform clustering on the case study data, which resulted in a dendrogram (Fig. 7) that advocated for the consolidation of the three families into a single family termed Leguminosae. The arrangement of each data sample on the dendrogram was determined by the similarity of the DNA sequences; the more closely the

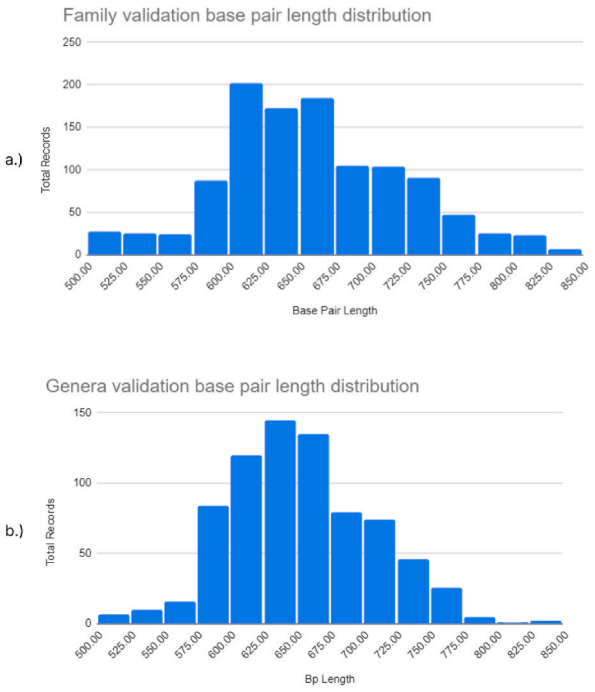

Fig. 6. Base pair distribution that is used in the validation phase of this study. a.) Family validation base pair length distribution from Table 1, b.) Genera validation base pair distribution from Table 2.

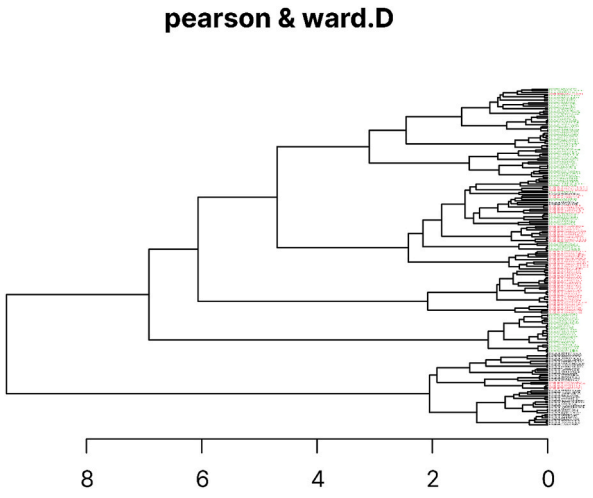

Fig. 7. Clustering result from case study.

two samples resembled each other, the more closely they were positioned on the dendrogram.

Fig. 7 shows the dendrogram visualization of the members of Fabaceae indicated with green labels on the dendrogram, Caesalpiniaceae with red labels, and Mimosaceae with black labels. We can see that some of the groups are clustered correctly, like at the top branch of the visualization, the group of Fabaceae (green) gathered in one place. However, overall, most were mixed and were not gathered in the same branch with the other group members. This is different compared to the visualization presented in Fig. 5, where each of the groups gathered on the same branch of the dendrogram. Samples from three families did not converge to produce clusters for their own families. This indicates that the three families are not different enough to be grouped into separate families. Thus, we can conclude that the group of Fabaceae, Caesalpiniaceae, and Mimosaceae should be grouped into one family. The morphological similarities among these three families are further reinforced by the resemblance in the shape of their fruits. The fruit from the Fabaceae family, illustrated in Fig. 8 a, resembles the fruit from the Mimosaceae family, shown in Fig. 8 b, as well as the fruit from the Caesalpiniaceae family, depicted in Fig. 8 c.

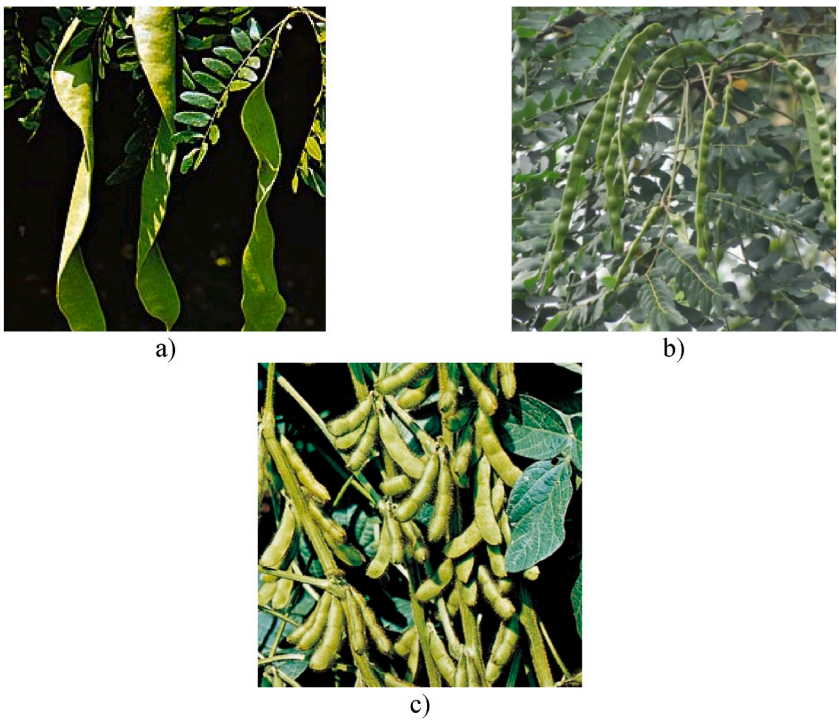

Fig. 8. a.) Fabaceae fruit, adapted from Ref. [73], b.) Mimosaceae fruit, adapted from Ref. [74], c.) Caesalpiniaceae fruit, adapted from Ref. [75].

The purposed method in this study uses common mathematical distance measures such as Euclidean, Manhattan, Canberra, Minkowski (P = 3), Pearson, and Spearman and does not use pairwise distance methods like Kimura 2-parameter (K2P) distance and Jukes and Cantor distance [76]. We also did not compare the result with the biological approach like electrophoretic analysis to cluster the species [35]. The machine learning approach may also be inconsistent if the libraries used in this approach receive an update or adjustment in their parameters, which is not a significant concern in traditional methods. This inconsistency directly ties into another limitation of this research which is that the dendrogram result needs to be validated by an expert to interpret the result. This human interpretation is limited to the bigger picture homogeneity of clusters. A computational method for interpreting the dendrogram may be able to parse out further details in the finer structures of the dendrogram [77], but this approach may be debatable. Finally, this research only uses a hierarchical clustering algorithm, any other algorithms like K-means [78], DBSCAN [79], Gaussian Mixture [80], etc. Can be used for this purpose and may give a different result.

The results from our experiment support several previous works that classify legumes as one family. Lewis [47] argues that the argument for three separate families is untenable because of two reasons. First, apparently, Mimosoideae and Papilionoideae are unique and distinct lineages arising in the Caesalpinioid alliance and are not comparable to it on the same taxonomic level. Second, Caesalpinioideae are under scrutiny and once further detailed studies are concluded it seems inevitable for divisions into more definable groups comparable in rank to the other two subfamilies. Hsuan [46], while not providing any arguments for the one-family classification, address the three groups as subfamilies in describing their morphology. Takhtajan [48] and Patel & Panchal [36] both refer to Leguminosae as one family.

On the other hand, a number of works argue against the one-family classification and instead classify legumes as three families. One such work by Cronquist [40] describes the author's preference for this classification because it is more in harmony with the customary classifications of families within angiosperms. Other works refer to a specific group in the legumes as a family such as Hou [41] for Caesalpiniaceae and Nielsen [42] with Mimosaceae. The argument for three families is also supported by other works that refer to the whole group as Fabales as an order, such as the works by Cronquist [40] and Dahlgren [43].

IV. Conclusion.

In this study, we validated our proposed machine learning, namely hierarchical clustering, for the objective of clustering a disputed group of Plantae–the Leguminosae. There are four main steps in this research, as follows: (i) data collection, (ii) data preprocessing, (iii) finding the best distance method, and (iv) determining the disputed family. According to the third step, our study shows that the Pearson correlation method is the best distance method to cluster different groups of families and genera. Through the application of the Pearson correlation approach within our hierarchical clustering experiments, the case study of the Leguminosae family, we ascertained that the Fabaceae, Mimosaceae, and Caesalpiniaceae are appropriately clustered into a single family. This conclusion is supported by the classification used or referred to by a number of previous works [36,46–48].

Author contribution statement

Lala Septem Riza: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. Muhammad Iqbal Zain; Ahmad Izzuddin; Yudi Prasetyo: Performed the experiments; Wrote the paper. Topik Hidayat: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data. Khyrina Airin Fariza Abu Samah: Analyzed and interpreted the data; Wrote the paper.

Data availability statement

Data associated with this study has been deposited at http://www.ncbi.nlm.nih.gov/taxonomy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix 1

To get access to sequence link in NCBI you can access through http://www.ncbi.nlm.nih.gov/nuccore/[Accession Number].

No	Full Name	Accession Number	Family Name
1	*Adenanthera pavonina*	KP092694.1	Mimosaceae
2	*Mimosa diplotricha*	MH768250.1	Mimosaceae
3	*Prosopis glandulosa*	AF174630.1	Mimosaceae
4	*Prosopis juliflora*	JX139107.1	Mimosaceae
5	*Mimosa pudica*	KX057889.1	Mimosaceae
6	*Leucaena leucocephala*	MH070604.1	Mimosaceae

(*continued on next page*)

(*continued*)

No	Full Name	Accession Number	Family Name
7	*Desmanthus pumilus*	AF458845.1	Mimosaceae
8	*Desmanthus virgatus*	AF458843.1	Mimosaceae
9	*Neptunia oleracea*	KX057891.1	Mimosaceae
10	*Entada abyssinica*	KX057869.1	Mimosaceae
11	*Albizia julibrissin*	FJ572041.1	Mimosaceae
12	*Samanea saman*	JX870770.1	Mimosaceae
13	*Calliandra surinamensis*	JX870747.1	Mimosaceae
14	*Acacia lycopodiifolia*	AF360716.1	Mimosaceae
15	*Dichrostachys paucifoliolata*	AF458812.1	Mimosaceae
16	*Leucaena lanceolata*	JF339948.1	Mimosaceae
17	*Archidendron utile*	KT767599.1	Mimosaceae
18	*Archidendron lucidum*	KT321363.1	Mimosaceae
19	*Acacia victoriae*	DQ029281.1	Mimosaceae
20	*Gleditsia triacanthos*	AF509980.1	Caesalpiniaceae
21	*Gleditsia microphylla*	AF510029.1	Caesalpiniaceae
22	*Caesalpinia pulcherrima*	JX856420.1	Caesalpiniaceae
23	*Haematoxylum campechianum*	KX372832.1	Caesalpiniaceae
24	*Haematoxylum brasiletto*	KX372834.1	Caesalpiniaceae
25	*Haematoxylum dinteri*	KX372830.1	Caesalpiniaceae
26	*Cassia fistula*	JX856430.1	Caesalpiniaceae
27	*Senna odorata*	HM116996.1	Caesalpiniaceae
28	*Senna siamea*	KJ638423.1	Caesalpiniaceae
29	*Chamaecrista choriophylla*	KR134122.1	Caesalpiniaceae
30	*Chamaecrista potentilla*	KR134123.1	Caesalpiniaceae
31	*Maniltoa grandiflora*	MG949352.1	Caesalpiniaceae
32	*Bauhinia purpurea*	JX856406.1	Caesalpiniaceae
33	*Bauhinia syringifolia*	AY258398.1	Caesalpiniaceae
34	*Cynometra letestui*	MG949304.1	Caesalpiniaceae
35	*Maniltoa gemmipara*	KY306626.1	Caesalpiniaceae
36	*Crudia papuana*	MH535137.1	Caesalpiniaceae
37	*Tamarindus indica*	MG949357.1	Caesalpiniaceae
38	*Flemingia macrophylla*	MN165994.1	Fabaceae
39	*Flemingia mengpengensis*	MN177611.1	Fabaceae
40	*Phaseolus sinuatus*	AF115194.1	Fabaceae
41	*Glycine pindanica*	AY433933.1	Fabaceae
42	*Pisum sativum*	AY143482.1	Fabaceae
43	*Phaseolus amblysepalus*	AF115218.1	Fabaceae
44	*Glycine max*	FJ609734.1	Fabaceae
45	*Cicer arietinum*	DQ312219.1	Fabaceae
46	*Medicago sativa*	AF053142.1	Fabaceae
47	*Cicer microphyllum*	KP338131.1	Fabaceae
48	*Glycyrrhiza pallidiflora*	EU591998.1	Fabaceae
49	*Glycyrrhiza astragalina*	GQ246134.1	Fabaceae
50	*Pueraria montana*	AF338215.1	Fabaceae
51	*Lupinus albus*	AF007481.1	Fabaceae
52	*Ulex parviflorus*	AF007470.1	Fabaceae
53	*Trifolium buckwestiorum*	AF053148.1	Fabaceae
54	*Lathyrus aphaca*	AY839345.1	Fabaceae
55	*Vicia sativa*	MH808491.1	Fabaceae
56	*Pongamia pinnata*	AF467493.1	Fabaceae
57	*Melilotus indicus*	MK918730.1	Fabaceae
58	*Acacia nilotica*	JX139101.1	Mimosaceae
59	*Acacia auriculiformis*	KC955519.1	Mimosaceae
60	*Acacia farnesiana*	AF360728.1	Mimosaceae
61	*Albizia lebbeck*	MN181375.1	Mimosaceae
62	*Leucaena leucocephala*	MH050230.1	Mimosaceae
63	*Senna alata*	MH050234.1	Caesalpiniaceae
64	*Celtis occidentalis*	DQ499147.1	Caesalpiniaceae
65	*Delonix regia*	KY321088.1	Caesalpiniaceae
66	*Phaseolus vulgaris*	MW843824.1	Fabaceae
67	*Sesbania grandiflora*	AF536354.1	Fabaceae
68	*Tephrosia purpurea*	MH768297.1	Fabaceae
69	*Abrus precatorius*	MF440357.1	Fabaceae
70	*Butea monosperma*	KJ436384.1	Fabaceae
71	*Cicer arietinum*	MW424513.1	Fabaceae
72	*Clitoria ternatea*	MH260279.1	Fabaceae
73	*Crotalaria pallida*	MH050227.1	Fabaceae
74	*Crotalaria retusa*	KP698625.1	Fabaceae
75	*Dalbergia sissoo*	JX856444.1	Fabaceae
76	*Erythrina variegata*	MT023961.1	Fabaceae

(*continued on next page*)

(*continued*)

No	Full Name	Accession Number	Family Name
77	*Glycyrrhiza glabra*	MT350378.1	Fabaceae
78	*Indigofera tinctoria*	MN879515.1	Fabaceae
79	*Melilotus albus*	MN560612.1	Fabaceae
80	*Melilotus indicus*	MW241661.1	Fabaceae
81	*Pisum sativum*	AY839340.1	Fabaceae
82	*Phaseolus vulgaris*	MW843825.1	Fabaceae
83	*Vigna mungo*	MF467912.1	Fabaceae
84	*Canavalia lineata*	KT751442.1	Fabaceae
85	*Lathyrus odoratus*	AY839377.1	Fabaceae
86	*Trifolium repens*	MT481887.1	Fabaceae
87	*Sesbania sesban*	MW560073.1	Fabaceae
88	*Sesbania bispinosa*	MH768288.1	Fabaceae
89	*Tephrosia purpurea*	MH768296.1	Fabaceae
90	*Desmodium gangeticum*	KP092721.1	Fabaceae
91	*Guilandina bonduc*	MH768079.1	Caesalpiniaceae
92	*Senna alata*	MH050233.1	Caesalpiniaceae
93	*Cassia fistula*	MW367522.1	Caesalpiniaceae
94	*Senna occidentalis*	MH558633.1	Caesalpiniaceae
95	*Senna siamea*	KJ638421.1	Caesalpiniaceae
96	*Senna sophera*	HQ833042.1	Caesalpiniaceae
97	*Senna tora*	KP092708.1	Caesalpiniaceae
98	*Tamarindus indica*	KF055236.1	Caesalpiniaceae
99	*Saraca asoca*	MW301610.1	Caesalpiniaceae
100	*Delonix regia*	KX057862.1	Caesalpiniaceae
101	*Acacia mangium*	KC955551.1	Mimosaceae
102	*Acacia catechu*	KC952019.1	Mimosaceae
103	*Pithecellobium dulce*	JX856483.1	Mimosaceae
104	*Adenanthera pavonina*	KP092695.1	Mimosaceae
105	*Leucaena leucocephala*	MG755502.1	Mimosaceae
106	*Bauhinia purpurea*	MH548397.1	Caesalpiniaceae
107	*Bauhinia tomentosa*	KX057838.1	Caesalpiniaceae
108	*Libidibia coriaria*	JX856416.1	Caesalpiniaceae
109	*Caesalpinia pulcherrima*	KX057841.1	Caesalpiniaceae
110	*Chamaecrista absus*	KT279729.1	Caesalpiniaceae
111	*Cassia fistula*	MW367497.1	Caesalpiniaceae
112	*Senna italica*	KY576676.1	Caesalpiniaceae
113	*Cassia javanica*	MW386314.1	Caesalpiniaceae
114	*Chamaecrista mimosoides*	KX057847.1	Caesalpiniaceae
115	*Senna obtusifolia*	KX057900.1	Caesalpiniaceae
116	*Senna occidentalis*	MW326931.1	Caesalpiniaceae
117	*Cassia roxburghii*	MW326753.1	Caesalpiniaceae
118	*Senna siamea*	KC984644.1	Caesalpiniaceae
119	*Senna surattensis*	MW367670.1	Caesalpiniaceae
120	*Senna tora*	MH712712.1	Caesalpiniaceae
121	*Delonix elata*	KY321105.1	Caesalpiniaceae
122	*Delonix regia*	KY321089.1	Caesalpiniaceae
123	*Parkinsonia aculeata*	KF379226.1	Caesalpiniaceae
124	*Tamarindus indica*	JX856519.1	Caesalpiniaceae
125	*Vachellia farnesiana*	KF532059.1	Mimosaceae
126	*Prosopis juliflora*	OK184559.1	Mimosaceae
127	*Mimosa pudica*	MN081594.1	Mimosaceae
128	*Leucaena leucocephala*	KF048811.1	Mimosaceae
129	*Senegalia senegal*	KY688828.1	Mimosaceae
130	*Pithecellobium dulce*	KX057895.1	Mimosaceae
131	*Albizia amara*	MW699936.1	Mimosaceae
132	*Albizia lebbeck*	MW699948.1	Mimosaceae
133	*Albizia procera*	MW699953.1	Mimosaceae
134	*Samanea saman*	EF638210.1	Mimosaceae
135	*Medicago lupulina*	MW241681.1	Fabaceae
136	*Medicago polymorpha*	OK036671.1	Fabaceae
137	*Trifolium repens*	MT481888.1	Fabaceae
138	*Melilotus albus*	MW241669.1	Fabaceae
139	*Lathyrus odoratus*	JN115031.1	Fabaceae
140	*Vicia hirsuta*	MH808488.1	Fabaceae
141	*Vicia sativa*	MW540820.1	Fabaceae
142	*Lupinus albus*	MK532380.1	Fabaceae
143	*Aeschynomene indica*	MN718416.1	Fabaceae
144	*Arachis hypogaea*	MT230611.1	Fabaceae
145	*Gliricidia sepium*	AF398816.1	Fabaceae
146	*Sesbania sesban*	KY968926.1	Fabaceae

(*continued on next page*)

(*continued*)

No	Full Name	Accession Number	Family Name
147	*Indigofera tinctoria*	MH595834.1	Fabaceae
148	*Dalbergia lanceolaria*	JX856439.1	Fabaceae
149	*Erythrina suberosa*	MT023956.1	Fabaceae
150	*Clitoria ternatea*	KT876054.1	Fabaceae
151	*Cajanus cajan*	MK253074.1	Fabaceae
152	*Rhynchosia minima*	MH768286.1	Fabaceae
153	*Butea monosperma*	MN700631.1	Fabaceae
154	*Pueraria montana*	JN407470.1	Fabaceae
155	*Glycine max*	MW391260.1	Fabaceae
156	*Lablab purpureus*	MH518283.1	Fabaceae
157	*Phaseolus vulgaris*	MW843826.1	Fabaceae
158	*Vigna aconitifolia*	JN008333.1	Fabaceae
159	*Abrus precatorius*	MN091943.1	Fabaceae
160	*Tephrosia candida*	HE681571.1	Fabaceae
161	*Tephrosia villosa*	MN173946.1	Fabaceae
162	*Smithia sensitiva*	MF281645.1	Fabaceae
163	*Alysicarpus vaginalis*	MH768274.1	Fabaceae
164	*Lathyrus aphaca*	KJ864924.1	Fabaceae
165	*Vigna radiata*	MW366905.1	Fabaceae
166	*Vigna unguiculata*	JN008290.1	Fabaceae
167	*Mimosa pigra*	KT364060.1	Mimosaceae
168	*Albizia procera*	JX856397.1	Mimosaceae
169	*Cynometra ramiflora*	MG949301.1	Caesalpiniaceae
170	*Senna hirsuta*	KJ638428.1	Caesalpiniaceae
171	*Senna occidentalis*	MZ505523.1	Caesalpiniaceae
172	*Senna siamea*	KJ638422.1	Caesalpiniaceae
173	*Senna alata*	MH915657.1	Caesalpiniaceae
174	*Mezoneuron hymenocarpum*	KX372820.1	Caesalpiniaceae
175	*Moullava digyna*	KX372803.1	Caesalpiniaceae
176	*Brownea coccinea*	MH535219.1	Caesalpiniaceae
177	*Senna tora*	MH050240.1	Caesalpiniaceae
178	*Caesalpinia* sp	KP003675.1	Caesalpiniaceae
179	*Cassia javanica*	MW386313.1	Caesalpiniaceae
180	*Bauhinia acuminata*	JX856404.1	Caesalpiniaceae
181	*Guilandina crista*	KX372808.1	Caesalpiniaceae
182	*Crotalaria juncea*	KP698651.1	Fabaceae
183	*Crotalaria calycina*	KP698617.1	Fabaceae
184	*Crotalaria pallida*	MH050226.1	Fabaceae
185	*Crotalaria verrucosa*	KP698648.1	Fabaceae
186	*Crotalaria saltiana*	KX371754.1	Fabaceae
187	*Desmodium triflorum*	LC377412.1	Fabaceae
188	*Uraria crinita*	JN407474.1	Fabaceae
189	*Aeschynomene americana*	MT902905.1	Fabaceae
190	*Mucuna bracteata*	LC494604.1	Fabaceae
191	*Millettia pinnata*	KF848293.1	Fabaceae
192	*Dalbergia volubilis*	KM276224.1	Fabaceae
193	*Trigonella foenum-graecum*	MH645773.1	Fabaceae
194	*Derris trifoliata*	MT312808.1	Fabaceae
195	*Grona heterocarpos*	MK933480.1	Fabaceae
196	*Vigna marina*	MH768299.1	Fabaceae
197	*Pongamia pinnata*	MN076243.1	Fabaceae
198	*Flemingia strobilifera*	MW732036.1	Fabaceae
199	*Derris scandens*	JX506450.1	Fabaceae

References

[1] E. Weitschek, G. Fiscon, G. Felici, Supervised DNA Barcodes species classification: analysis, comparisons and results, Dec, BioData Min. 7 (1) (2014) 4, https://doi.org/10.1186/1756-0381-7-4.

[2] R. Purty, S. Chatterjee, DNA barcoding: an effective technique in molecular taxonomy, Austin J. Biotechnol. Bioeng. 3 (2016) 1059, 1.

[3] A.A. Saddhe, K. Kumar, DNA barcoding of plants: selection of core markers for taxonomic groups, Dec, Plant Sci. Today 5 (1) (2017) 9–13, https://doi.org/10.14719/pst.2018.5.1.356.

[4] C.L. Schoch, et al., Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi, Apr, Proc. Natl. Acad. Sci. USA 109 (16) (2012) 6241–6246, https://doi.org/10.1073/pnas.1117018109.

[5] A.J. Fazekas, et al., Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well, Jul, PLoS One 3 (7) (2008) e2802, https://doi.org/10.1371/journal.pone.0002802.

[6] L. Gielly, P. Taberlet, The use of chloroplast DNA to resolve plant phylogenies: noncoding versus rbcL sequences, Sep, Mol. Biol. Evol. 11 (5) (1994) 769–777, https://doi.org/10.1093/oxfordjournals.molbev.a040157.

[7] K.W. Hilu, gping Liang, The matK gene: sequence variation and application in plant systematics, Jun, Am. J. Bot. 84 (6) (1997) 830–839, https://doi.org/10.2307/2445819.

[8] China Plant Bol Group, et al., Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants, Dec, Proc. Natl. Acad. Sci. USA 108 (49) (2011) 19641–19646, https://doi.org/10.1073/pnas.1104551108.

[9] A.J. Fazekas, et al., Are plant species inherently harder to discriminate than animal species using DNA barcoding markers?, May, Mol. Ecol. Resour. 9 (2009) 130–139, https://doi.org/10.1111/j.1755-0998.2009.02652.x.

[10] W.H. Wong, Y.C. Tay, J. Puniamoorthy, M. Balke, P.S. Cranston, R. Meier, 'Direct PCR' optimization yields a rapid, cost-effective, nondestructive and efficient method for obtaining DNA barcodes without DNA extraction, Nov, Mol. Ecol. Resour 14 (6) (2014) 1271–1280, https://doi.org/10.1111/1755-0998.12275.

[11] Ž. Fišer, Pečnikar, E.V. Buzan, 20 years since the introduction of DNA barcoding: from theory to application, J. Appl. Genet. 55 (1) (Feb. 2014) 43–52, https://doi.org/10.1007/s13353-013-0180-y.

[12] M.J.L. Madden, R.G. Young, J.W. Brown, S.E. Miller, A.J. Frewin, R.H. Hanner, Using DNA barcoding to improve invasive pest identification at U.S. ports-of-entry, Sep, PLoS One 14 (9) (2019), e0222291, https://doi.org/10.1371/journal.pone.0222291.

[13] P.F.M. Gonçalves, A.R. Oliveira-Marques, T.E. Matsumoto, C.Y. Miyaki, DNA barcoding identifies illegal parrot trade, J. Hered. 106 (S1) (2015) 560–564, https://doi.org/10.1093/jhered/esv035.

[14] L. Jiao, Y. Lu, T. He, J. Guo, Y. Yin, DNA barcoding for wood identification: global review of the last decade and future perspective, Oct, IAWA J. 41 (4) (2020) 620–643, https://doi.org/10.1163/22941932-bja10041.

[15] R. Tänzler, K. Sagata, S. Surbakti, M. Balke, A. Riedel, DNA barcoding for community ecology - how to tackle a hyperdiverse, mostly undescribed melanesian fauna, Jan, PLoS One 7 (2012), e28832, https://doi.org/10.1371/journal.pone.0028832, 1.

[16] B.C. Rossini, et al., Highlighting Astyanax species diversity through DNA barcoding, Dec, PLoS One 11 (12) (2016), e0167203, https://doi.org/10.1371/journal.pone.0167203.

[17] V.A. Lukhtanov, A. Sourakov, E.V. Zakharov, DNA barcodes as a tool in biodiversity research: testing pre-existing taxonomic hypotheses in Delphic Apollo butterflies (Lepidoptera, Papilionidae), Nov, Syst. Biodivers. 14 (6) (2016) 599–613, https://doi.org/10.1080/14772000.2016.1203371.

[18] A. Sandionigi, et al., Analytical approaches for DNA barcoding data – how to find a way for plants?, Dec, Plant Biosyst. - Int. J. Deal. Asp. Plant Biol. 146 (4) (2012) 805–813, https://doi.org/10.1080/11263504.2012.740084.

[19] D.P. Little, D. Wm Stevenson, A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms, Feb, Cladistics 23 (1) (2007) 1–21, https://doi.org/10.1111/j.1096-0031.2006.00126.x.

[20] C.-H. Yang, K.-C. Wu, L.-Y. Chuang, H.-W. Chang, DeepBarcoding: deep learning for species classification using DNA barcoding, Jul, IEEE ACM Trans. Comput. Biol. Bioinf 19 (4) (2022) 2158–2165, https://doi.org/10.1109/TCBB.2021.3056570.

[21] P.K. Meher, T.K. Sahu, A.R. Rao, Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier, Nov, Gene 592 (2) (2016) 316–324, https://doi.org/10.1016/j.gene.2016.07.010.

[22] H. Soueidan, M. Nikolski, Machine learning for metagenomics: methods and tools (2015), https://doi.org/10.48550/ARXIV.1510.06621.

[23] R. DeSalle, P. Goldstein, Review and interpretation of trends in DNA barcoding, Sep, Front. Ecol. Evol. 7 (2019) 302, https://doi.org/10.3389/fevo.2019.00302.

[24] P.F. Stevens, History of taxonomy, in: eLS, first ed., John Wiley & Sons, Ltd, Ed. Wiley, 2003 https://doi.org/10.1038/npg.els.0003093.

[25] J.C. Avise, J.-X. Liu, On the temporal inconsistencies of Linnean taxonomic ranks, Apr, Biol. J. Linn. Soc. 102 (4) (2011) 707–714, https://doi.org/10.1111/j.1095-8312.2011.01624.x.

[26] W. Greuter, et al. (Eds.), International Code of Botanical Nomenclature (Saint Louis Code): Adopted by the Sixteenth International Botanical Congress, St Louis, Missouri, July-August 1999, Koeltz scientific books, Königstein, 2000.

[27] C.T. Parker, B.J. Tindall, G.M. Garrity, International code of nomenclature of prokaryotes: prokaryotic code (2008 revision), Jan, Int. J. Syst. Evol. Microbiol. 69 (1A) (2019) S1–S111, https://doi.org/10.1099/ijsem.0.000778.

[28] C.L. Schoch, et al., NCBI Taxonomy: a comprehensive update on curation, resources and tools, Database 2020 (2020), https://doi.org/10.1093/database/baaa062 baaa062, Jan.

[29] P.J. Walker, et al., Changes to virus taxonomy and the international code of virus classification and nomenclature ratified by the international committee on taxonomy of viruses (2019), Sep, Arch. Virol. 164 (9) (2019) 2417–2429, https://doi.org/10.1007/s00705-019-04306-w.

[30] J.D. Sigwart, M.D. Sutton, K.D. Bennett, How big is a genus? Towards a nomothetic systematics, Jun, Zool. J. Linn. Soc. 183 (2) (2018) 237–252, https://doi.org/10.1093/zoolinnean/zlx059.

[31] S. Das, Domestication, phylogeny and taxonomic delimitation in underutilized grain Amaranthus (Amaranthaceae) – a status review, Feb, Feddes Repert. 123 (4) (2012) 273–282, https://doi.org/10.1002/fedr.201200017.

[32] M.F. Docker (Ed.), Lampreys: Biology, Conservation and Control, Volume 1, first ed., Dordrecht: Springer Netherlands, Imprint: Springer, 2015 https://doi.org/10.1007/978-94-017-9306-3, 2015.

[33] Y. Ji, C. Liu, J. Yang, L. Jin, Z. Yang, J.-B. Yang, Ultra-barcoding discovers a cryptic species in Paris yunnanensis (melanthiaceae), a medicinally important plant, Apr, Front. Plant Sci. 11 (2020) 411, https://doi.org/10.3389/fpls.2020.00411.

[34] J.B. Ristaino, The importance of mycological and plant herbaria in tracking plant killers, Front. Ecol. Evol. 7 (2020), https://doi.org/10.3389/fevo.2019.00521.

[35] A.K. Mondal, S. Mondal, Circumscription of the families within Leguminales as determined by cladistic analysis based on seed protein, Apr, Afr. J. Biotechnol. 10 (15) (2011) 2850–2856, https://doi.org/10.5897/AJB10.206.

[36] S. Patel, H. Panchal, Evolutionary studies of few species belonging to Leguminosae family based on RBCL gene, Discovery 9 (22) (Jan. 2014) 38–50.

[37] J.J. Doyle, M.A. Luckow, The rest of the iceberg. Legume diversity and evolution in a phylogenetic context, Mar, Plant Physiol. 131 (3) (2003) 900–910, https://doi.org/10.1104/pp.102.018150.

[38] R.M. Polhill, P.H. Raven, Advances in Legume Systematics, Kew/Surrey: Royal botanic gardens, 1981.

[39] A.K. Dhakad, Molecular Phylogeny of Selected Tree Species of Families Fabaceae Caesalpiniaceae and Mimosaceae of Uttarakhand [Online]. Available:, Forest Research Institute University, 2018 http://hdl.handle.net/10603/203120.

[40] A. Cronquist, An Integrated System of Classification of Flowering Plants, Columbia University Press, New York, 1981.

[41] D. Hou, K. Larsen, S.S. Larsen, Caesalpiniaceae (Leguminosae-Caesalpinioideae), Jan, Flora Malesiana 12 (2) (1996) 409–730.

[42] I.C. Nielsen, Mimosaceae (Leguminosae-Mimosoideae), Flora Malesiana 11 (1) (1992) 1–226.

[43] R. Dahlgren, "General aspects of angiosperm evolution and macrosystematics," Nord, J. Bot., Le 3 (1) (1983) 119–149, https://doi.org/10.1111/j.1756-1051.1983.tb01448.x.

[44] G. Bentham, "Notes on Mimoseae, with a synopsis of species," Lond. J. Bot., vol. 1, pp. 318–528, 1842..

[45] T.J. Wardill, G.C. Graham, M. Zalucki, W.A. Palmer, J. Playford, K.D. Scott, The importance of species identity in the biocontrol process: identifying the subspecies of Acacia nilotica (Leguminosae: Mimosoideae) by genetic distance and the implications for biological control, Dec, J. Biogeogr. 32 (12) (2005) 2145–2159, https://doi.org/10.1111/j.1365-2699.2005.01348.x.

[46] K. Hsuan, Orders and Families of Malayan Seed Plants, Singapore University Press, Singapore, 1983.

[47] G.P. Lewis (Ed.), Legumes of the World, Royal Botanic Gardens, Kew, Richmond, UK, 2005.

[48] A.L. Takhtajan, Outline of the classification of flowering plants (magnoliophyta), Jul, Bot. Rev. 46 (3) (1980) 225–359, https://doi.org/10.1007/BF02861558.

[49] F. Nielsen, Introduction to HPC with MPI for Data Science, Springer International Publishing, Cham, 2016, https://doi.org/10.1007/978-3-319-21903-5.

[50] Q. An, et al., Predicting medicinal resources in Ranunculaceae family by a combined approach using DNA barcodes and chemical metabolites, Aug, Phytochem. Lett. 50 (2022) 67–76, https://doi.org/10.1016/j.phytol.2022.04.009.

[51] C. Lucas, T. Thangaradjou, J. Papenbrock, Development of a DNA barcoding system for seagrasses: successful but not simple, Jan, PLoS One 7 (1) (2012), e29987, https://doi.org/10.1371/journal.pone.0029987.

[52] E.V. Nikitina, F.I. Karimov, N.V. Savina, S.V. Kubrak, A.V. Kilchevsky, Inventory of some Tulipa species from Uzbekistan using DNA barcoding, BIO Web Conf 38 (2021), https://doi.org/10.1051/bioconf/20213800086, 00086.

[53] E.V. Nikitina, N. Yu Beshko, S.A. Omarov, Assessment of plant species diversity (Lamiaceae Lindle.) in Uzbekistan based on DNA barcoding, Jul, IOP Conf. Ser. Earth Environ. Sci. 1068 (1) (2022), 012042, https://doi.org/10.1088/1755-1315/1068/1/012042.
[54] Y. Papa, P. Le Bail, R. Covain, Genetic landscape clustering of a large DNA barcoding data set reveals shared patterns of genetic divergence among freshwater fishes of the Maroni Basin, Aug, Mol. Ecol. Resour. 21 (6) (2021) 2109–2124, https://doi.org/10.1111/1755-0998.13402.
[55] H. Xu, P. Li, G. Ren, Y. Wang, D. Jiang, C. Liu, Authentication of three source spices of arnebiae radix using DNA barcoding and HPLC, Front. Pharmacol. 12 (Jul. 2021), 677014, https://doi.org/10.3389/fphar.2021.677014.
[56] L. Zhao, X. Yu, J. Shen, X. Xu, Identification of three kinds of Plumeria flowers by DNA barcoding and HPLC specific chromatogram, J. Pharm. Anal. 8 (3) (Jun. 2018) 176–180, https://doi.org/10.1016/j.jpha.2018.02.002.
[57] T.M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.
[58] T. He, L. Jiao, A.C. Wiedenhoeft, Y. Yin, Machine learning approaches outperform distance- and tree-based methods for DNA barcoding of Pterocarpus wood, Planta 249 (5) (May 2019) 1617–1625, https://doi.org/10.1007/s00425-019-03116-3.
[59] D.J. Winter, Rentrez: an R Package for the NCBI eUtils API, PeerJ Preprints, preprint, Aug. 2017, https://doi.org/10.7287/peerj.preprints.3179v2.
[60] P.A.H. Pagès, "Biostrings." Bioconductor (2017), https://doi.org/10.18129/B9.BIOC.BIOSTRINGS.
[61] C.H.-K. Enrico Bonatesta, "msa." Bioconductor, MSA, 2017, https://doi.org/10.18129/B9.BIOC.
[62] C. Heibl, PHYLOCH: R Language Tree Plotting Tools and Interfaces to Diverse Phylogenetic Software Packages, Jan[Online]. Available:, 2008 http://www.christophheibl.de/Rpackages.html.
[63] A. Kassambara, F. Mundt, Factoextra: Extract and Visualize the Results of Multivariate Data Analyses [Online]. Available:, 2020 https://CRAN.R-project.org/package=factoextra.
[64] S. Federhen, The NCBI Taxonomy database, Jan, Nucleic Acids Res. 40 (D1) (2012) D136–D143, https://doi.org/10.1093/nar/gkr1178.
[65] S. Tripathi, A.K. Mondal, Taxonomic diversity in epidermal cells (stomata) of some selected Anthophyta under the order Leguminales (Caeselpiniaceae, Mimosaceae and Fabaceae) based on numerical analysis: a systematic approach, Dec, Int. J. Sci. Nat. 3 (4) (2012) 778–798.
[66] J.D. Thompson, D.G. Higgins, T.J. Gibson, Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (22) (1994) 4673–4680, https://doi.org/10.1093/nar/22.22.4673.
[67] M.D. Malkauthekar, Analysis of euclidean distance and manhattan distance measure in face recognition, in: Third International Conference on Computational Intelligence and Information Technology, CIIT 2013), Mumbai, India, 2013, pp. 503–507, https://doi.org/10.1049/cp.2013.2636.
[68] M. Faisal, E.M. Zamzami, Sutarman, Comparative analysis of inter-centroid K-means performance using euclidean distance, Canberra distance and manhattan distance, Jun, J. Phys. Conf. Ser. 1566 (1) (2020), 012112, https://doi.org/10.1088/1742-6596/1566/1/012112.
[69] J.M. Merigó, M. Casanovas, A new Minkowski distance based on induced aggregation operators, Apr, Int. J. Comput. Intell. Syst. 4 (2) (2011) 123–133, https://doi.org/10.1080/18756891.2011.9727769.
[70] J.H. Weber, K.A. Schouhamer Immink, S.R. Blackburn, Pearson codes, Jan, IEEE Trans. Inf. Theor. 62 (1) (2016) 131–135, https://doi.org/10.1109/TIT.2015.2490219.
[71] Y. Xie, Y. Wang, A. Nallanathan, L. Wang, An improved K-Nearest-Neighbor indoor localization method based on spearman distance, Mar, IEEE Signal Process. Lett. 23 (3) (2016) 351–355, https://doi.org/10.1109/LSP.2016.2519607.
[72] D.A. Benson, et al., Nov, "GenBank," Nucleic Acids Res. 41 (D1) (2012) D36–D42, https://doi.org/10.1093/nar/gks1195.
[73] J.H. Gerard, Common honey locust (gleditsia triacanthos) [Online]. Available: https://www.britannica.com/plant/Fabales/Classification-of-Fabaceae#/media/1/199654/115993. (Accessed 5 September 2023).
[74] Vijay, Adenanthera pavonina L [Online]. Available: https://indiabiodiversity.org/species/show/245164. (Accessed 5 September 2023).
[75] J.C. Allen, Son. Soybeans (Glycine max) [Online]. Available: https://www.britannica.com/plant/common-bean#/media/1/199651/7490. (Accessed 5 September 2023).
[76] T. Nishimaki, K. Sato, An extension of the Kimura two-parameter model to the natural evolutionary process, J. Mol. Evol. 87 (1) (Jan. 2019) 60–67, https://doi.org/10.1007/s00239-018-9885-1.
[77] P. Lee, S.T. Yang, J.D. West, B. Howe, PhyloParser: a hybrid algorithm for extracting phylogenies from dendrograms, in: 2017 *14*th IAPR International Conference On Document Analysis And Recognition (ICDAR), Kyoto, IEEE, Nov, 2017, pp. 1087–1094, https://doi.org/10.1109/ICDAR.2017.180.
[78] J.A. Hartigan, M.A. Wong, Algorithm as 136: a K-means clustering algorithm, Applied Statistics 28 (1) (1979) 100, https://doi.org/10.2307/2346830.
[79] M. Hahsler, M. Piekenbrock, D. Doran, Dbscan : fast density-based clustering with R, J. Stat. Software 91 (1) (2019), https://doi.org/10.18637/jss.v091.i01.
[80] M. Ouyang, W.J. Welsh, P. Georgopoulos, Gaussian mixture clustering and imputation of microarray data, Bioinformatics 20 (6) (Apr. 2004) 917–923, https://doi.org/10.1093/bioinformatics/bth007.