SOFT COMPUTING BASED IDENTIFICATION AND ASSESSMENT OF POTENTIAL DNA BARCODES OF SOLANACEOUS SPECIES USING cpDNA SEQUENCES

Bhupinder Pal Singh*

Research Scholar, I.K.Gujral Punjab Technical University, Kapurthala, 144 603, India
System Manager, Centre for I.T. Solutions, Guru Nanak Dev University, Amritsar, 143005, India
bhupinderpalsingh@hotmail.com

Ajay Kumar

Professor, Electronics and Communication Engineering Department,
Beant College of Engineering and Technology, Gurdaspur, 143521, India
ajaykm_20@yahoo.co.in

Harpreet Singh

Assistant Professor, Department of Bioinformatics, Hans Raj Mahila Maha Vidyalaya, Jalandhar, 144008, India
harpreetsingh05@gmail.com

Avinash Kaur Nagpal*

Professor, Department of Botanical and Environmental Sciences,
Guru Nanak Dev University, Amritsar, 143005, India
avnagpal@yahoo.co.in

Abstract

DNA barcoding (a technique that uses short DNA sequences) has become fast, economic and accurate method for discovering and identifying organisms of the three main kingdoms of eukaryotes. In plants, few coding and non coding regions of chloroplast genomes have been tested for their ability to identify species while other regions of genome are still left to be explored for their suitability as DNA barcodes. The present study is about identification of potential DNA barcodes and assessing their potential to discriminate 133 plant species belonging to family Solanaceae from chloroplast DNA (cpDNA) sequences using different machine learning classification algorithms in WEKA and distance based method in SPIDER. Thirty three hyper-variable regions were identified based on nucleotide diversity (π) using sliding window analysis of aligned file of these species. These regions along with well established markers (matK and rbcL) were assessed for their discriminating potential at genus level. Sequence richness regime was followed for six hyper-variable regions 'ycf1', 'cemA, cemA-petA', 'rps12-clpP, clpP / rps12-psbB', 'petA, petA-psbJ, psbJ, psbJ-psbL', 'trnL-trnF, trnF, trnF-ndhJ' and 'ndhF, ndhF-rpl32, rpl32, rpl32-trnL' using BLASTN along with matK and rbcL and were tested for their discrimination potential at genus and species levels. Distance based method SPIDER and machine learning algorithm SMO performed best when compared with other classification methods. It was observed from the study that with increase in number of sequences from particular species, there is increase in percentage correct identification rates. All hyper-variable regions were able to achieve maximum percentage of correct identification rate (100%) at genus level. However region 'ndhF, ndhF-rpl32, rpl32, rpl32-trnL' was able to achieve highest discrimination rate of 69% at species level which was even better than matK and rbcL. The low identification rates at species level as compared to genus level were attributed to ambiguity within species for these regions. This study will provide valuable resource for development of DNA barcodes for Solanaceae family.

Keywords: DNA barcodes; Solanaceae; Machine learning algorithms; SPIDER

1. Introduction

DNA Barcoding is a technique used for species identification with the help of short gene sequences from the standardized region of the genome [Savolainen *et al.* (2005)]. The short DNA sequence or gene sequence which can identify a species is called as DNA barcode. An ideal barcode should have length of 700 bp or less,

simple to sequence and exhibit significant species level genetic variation [Kress *et al.* (2005)]. The traditional species identification techniques rely on morphological characters. However, these methods are time consuming and costly. In addition, the traditional methods often fail to correctly identify closely related species [Pires and Marinoni (2010)]. DNA barcoding can provide fast, low cost and reliable method for discovering new species and identification of existing ones [Hebert *et al.* (2003b)].

DNA barcoding method has been applied for identification of three traditional kingdoms of multicellular eukaryotic life forms like animals, plants and fungi. In animal kingdom, mitochondrial cytochrome c oxidase subunit 1 (COI) gene enabled the discrimination of closely allied species and can be used as DNA barcode [Hebert *et al.* (2003a)]. In Fungi, nuclear ribosomal Internal Transcribed Spacer (ITS) Region can be used as a DNA barcode marker [Schoch *et al.* (2012)]. However in plants, finding a robust and effective barcode is difficult as no single locus is sufficient to discriminate among different species [Hollingsworth (2011)]. Several combinations of coding and intergenic non-coding regions have been identified as DNA Barcodes at family/genus level. These regions include ITS2 and psbA-trnH for Rutaceae [Luo *et al.* (2010)]; atpF-atpH for Lemnaceae [Wang *et al.* (2010)]; ITS2 for Asteraceae [Gao *et al.* (2010)], Rosaceae [Pang *et al.* (2011)], *Uncaria* (Rubiaceae) [Zhang *et al.* (2015)] and *Physalis* (Solanaceae) [Feng *et al.* (2016)]; combination of ITS and trnH-psbA for *Parnassia* (Celastraceae) [Yang *et al.* (2012)], *Ficus* (Moraceae) [Li *et al.* (2012)] and Apiaceae [Liu *et al.* (2014)]; combination of matK and trnH-psbA for *Lamium* (Lamiaceae) [Krawczyk *et al.* (2014)]; combination of atpF-atpH, psbK-psbI and trnH-psbA for Orchidaceae [Kim *et al.* (2014)]; trnH-psbA and ITS region for flowering plants [Kress *et al.* (2005)]. In addition to this, CBOL Plant Working Group *et al.* (2009) have proposed rbcL and matK as standard DNA barcoding regions for land plants.

The DNA barcoding experiments have been confined to only a few specific coding or non-coding regions while many other regions of DNA must be explored to find their suitability as potential DNA barcodes. Although DNA barcoding experimental studies have been simplified over the years, but still these are expensive and require lot of labour for processing large number of samples [de Kerdrel *et al.* (2020)]. With the availability of complete chloroplast genome sequences of large number of plant species, it is now possible to search whole chloroplast genomes of a particular family for identification of hyper-variable regions which could act as potential DNA barcodes.

Solanaceae family is one of the major groups of angiosperms with more than 2500 plant species belonging to 100 genera [Rosario *et al.* (2019)]. Many members of this family have a great agricultural and economical importance. This family is yet to be explored for development of DNA barcodes at family level. In the present study, we have applied soft computing based techniques for identification of hyper-variable regions from all available complete chloroplast genome sequences and to assess them as potential DNA barcodes for identification of various members of Solanaceae family. This study will provide the lead for confirmation of these regions to be used as potential DNA barcodes. To the best of our knowledge, this is perhaps the first study in which an attempt has been made to find potential DNA barcodes for plant species belonging to Solanaceae family using *in silico* methods. The aims of current study are: - (1) *In silico* identification of hyper-variable regions which could act as potential barcodes for discriminating plant species at genus as well species levels for Solanaceae family. (2) Assessment of discriminating potential of selected hyper-variable regions using distance based method as well as machine learning algorithms. (3) Comparison of discriminative potential of the selected hyper-variable regions with two commonly used barcoding regions i.e. matK and rbcL as recommended by CBOL.

2. Methods

2.1. *Sequence downloading and alignment*

Complete chloroplast genome sequences of 133 plant species belonging to 19 genera of Solanaceae family were downloaded in FASTA format from CpGDB (http://www.gndu.ac.in/CpGDB/) database [Singh *et al.* (2020)] (Supplementary Table 1). Above sequences were aligned with MAFFT (Multiple Alignment using Fast Fourier Transform) [Katoh and Standley (2013)] online web server (https://mafft.cbrc.jp/alignment/server/) using the default parameters. Fig. 1 shows the workflow adopted for identification and validation of hyper-variable regions for development of potential DNA barcodes to identify plant species belonging to Solanaceae family.

2.2. *Identification of hyper-variable regions*

Identification of hyper-variable regions of the chloroplast genomes can lead to development of DNA barcodes to discriminate plant species which was done in two phases in the present study. In the first phase, a snapshot of overall variability in the aligned genome sequences was obtained using in house written PERL script which calculates the variable, singleton and parsimony informative sites. A variable site is a position in the alignment which contains at least two types of nucleotides. Some of the variable sites can be singleton or parsimony-informative sites (https://www.megasoftware.net/web_help_7/helpfile.htm) [Kumar *et al.* (2016)]. A singleton site in the alignment contains at least two types of nucleotides (or amino acids) with, at most, one occurring multiple times. A site is parsimony-informative if it contains at least two types of nucleotides (or amino acids), and at least two of them occur with a minimum frequency of two.

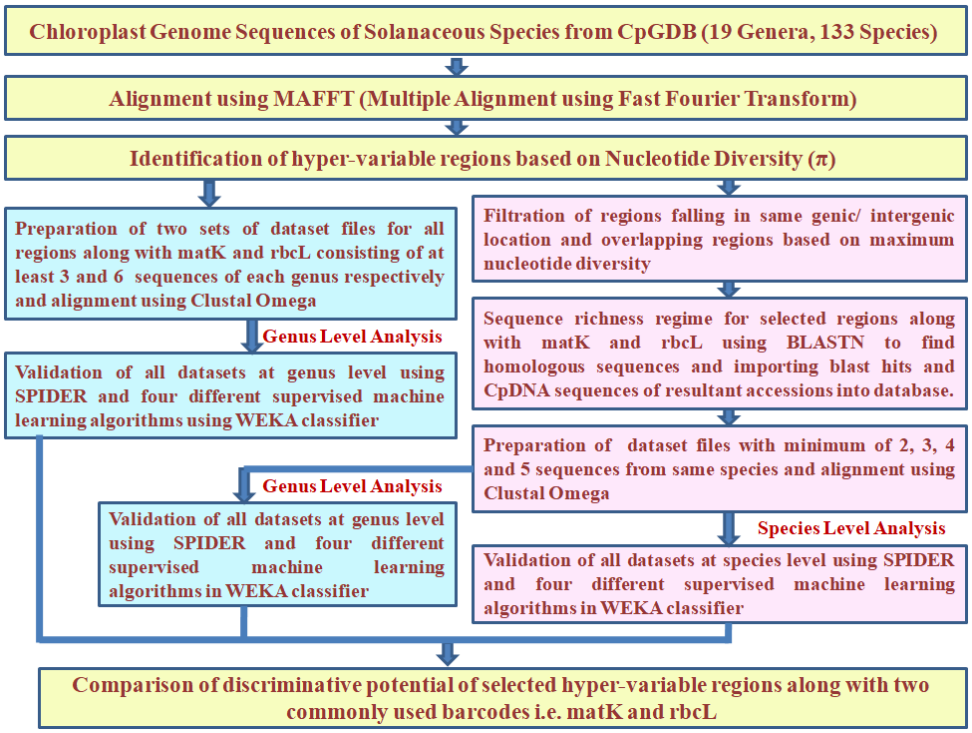

Fig. 1. Complete workflow to identify and assess potential DNA barcodes for plant species belonging to Solanaceae family

In the second phase, the aligned file was converted to MEGA format using script written in PERL language. The converted file in MEGA format was used to perform sliding window analysis using DnaSP ver 5.10 with window length as 600 bp and step size as 200 bp [Librado and Rozas (2009)]. In order to have uniform number of net nucleotides in all the windows, sites with alignment gaps are not considered. Nucleotide diversity (π) i.e. average number of nucleotide differences per site for each window was computed. Windows having π greater than mean+2SD were considered as hyper-variable regions [Bi *et al.* (2018)]. A program was written using vb.net to find corresponding genic/intergenic locations of these regions from the annotations available in their GenBank files and stored in the SQL server 2012 database. The actual locations of the hyper-variable regions corresponding to each gene in the alignment were computed using complete window as well as its mid-point of the window. This was done to double check the location of the regions which sometimes is ambiguous due to incomplete annotation of genes and alternate gene names. In addition, regions corresponding to the matK and rbcL genes, the well established DNA barcodes for plant species were also included as a positive control.

2.3. *Preparation of datasets and data cleaning*

The sequences of all hyper-variable regions along with matK and rbcL genes were extracted using program written in vb.net for those genera which have representation of three or more species (list of genera along with number of species is given as Supplementary Table 2). This collection of sequences was used to create two distinct datasets with one dataset consisting of minimum 3 sequences from each genus and the second dataset consisting of minimum of 6 sequences from each genus. The sequences not fulfilling above criteria were filtered out. The datasets thus obtained were aligned using Clustal Omega command line version 1.2.2 (http://www.clustal.org/omega/). The alignment generated was also used to compute variable, singleton and parsimony informative sites using PERL script. These datasets were used for genus level identification as each species in the dataset was represented by only one sequence in the dataset.

In order to validate hyper-variable regions in terms of their potential to be used as DNA barcoding regions at species level using machine learning approaches, more than one sequence of the same region from same species are required. Therefore a sequence richness regime was followed using web based BLASTN (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to find homologous sequences related to same species for each region. Before proceeding for sequence richness regime, the regions falling in the same genic/intergenic location and overlapping regions were filtered out based on the maximum nucleotide diversity. For each selected hyper-variable region, the sequence with shortest length i.e. the sequence with least gaps in the aligned window among all the 133 species was selected. Apart from the selected hyper-variable regions, two more regions, matK and rbcL genes were used as control.

The selected sequences were used one by one as query sequence in the web based nucleotide blast program. Restricted search was performed in nucleotide collection database by using the subset “Solanaceae (taxid:4070)” in organism field. The other parameters were set to default values except maximum target sequences

as 5000. The resulted blast hits were filtered based on 90% or more query coverage and imported to SQL server 2012 database. Also complete cpDNA sequences of resultant accessions were downloaded and stored in the database. A program was written in vb.net to extract matched sequences from the database based on starting and ending position of matching record. The species which have at least two or more accessions were considered for further analysis. Separate dataset files i.e. files with minimum 2, 3, 4 and 5 sequences from same species were prepared in FASTA format to perform analysis at genus as well as species levels for each region. This was done to further evaluate the effect of number of sequences of a particular species on the efficiency of the barcoding region.

2.4. *Validation of hyper-variable regions*

Identification success rate of different hyper-variable regions was calculated both at genus and species levels using traditional distance based method as well as supervised machine learning algorithms.

2.4.1. *Distance based analysis using SPIDER*

SPecies IDentity and Evolution in R (SPIDER) is a package which is being used for DNA barcoding studies based on distance based analysis [Brown *et al.* (2012)]. A script was written in R (https://www.r-project.org/) language to evaluate sequence identification success using 'nearNeighbour' and 'bestCloseMatch' functions for various datasets at genus and species levels. These functions were used to test the barcoding efficiency of each region. To use these functions, all sequences in the dataset must be identified before analysis. During analysis, each sequence, one by one, was considered unidentified and used as query sequence. Remaining sequences in the dataset were used to identify the query sequence based on distance. The function 'nearNeighbour' returns true if the closest individual matches with the species of query sequence otherwise returns false. The function 'bestCloseMatch' returns the closest individual within given threshold. If no individual found within given threshold then it returns 'no identification' whereas if more than one individual is found then it returns 'ambiguous'.

2.4.2. *Machine learning based analysis using WEKA*

Apart from distance based analysis, machine learning algorithms were also used to test the efficiency of DNA barcoding regions for identification of plant species [Weitschek *et al.* (2014); Hartvig *et al.* (2015)]. DNA barcoding was considered as a classification problem and analysis was performed using Waikato Environment for Knowledge Analysis (WEKA) classifier [Hall *et al.* (2009)]. Efficiency of each region was tested using different supervised machine learning algorithms like Decision Tree based C4.5 (J48) [Salzberg (1994)], Function based Support Vector Machine using Sequential Minimal Optimization (SMO) [Platt (1999)], Naïve Bayes classifier [John and Langley (1995)] and propositional rule based learner RIPPER (Jrip) [Cohen (1995)].

J48 is a supervised classifier and is an open source java implementation of C4.5 decision tree algorithm in WEKA. Decision tree algorithm builds a tree like structure based on various attributes. The decision tree is human readable classification model which consists of set of logic rules based on nucleotide positions in the sequence. SMO is Support Vector Machine supervised learning algorithm which is implemented in WEKA classifier. It converts the input dataset into multi dimensional vectors and defines optimal hyperplane which separates different output classes. However there is no human readable model provided with this algorithm.

Naïve Bayes is a Bayesian-based classifier which is implemented in WEKA and is often used when a large dataset is available. In this algorithm, prior probabilities are calculated for each class from the input dataset and posterior probabilities of query sequence are calculated for each class and prediction is made based on the highest probability. This algorithm also does not provide a readable model. Jrip is a WEKA classifier which implements propositional rule based Repeated Incremental Pruning to produce Error Reduction (RIPPER). This algorithm first generates initial set of rules for each class and then optimizes the rule set k times and provides classification model which is composed of logic rules for each class. This method can have advantage in DNA barcoding as the set of rules for each species can be applied manually for identification.

To run machine learning algorithms, all dataset files in FASTA format were converted to Attribute-Relation File Format (ARFF) using FASTA to WEKA converter [Weitschek *et al.* (2014)]. WEKA experimenter was set up to run four machine learning algorithms on each dataset at genus and species levels. The percentage identification was evaluated using 10 fold cross-validation and each experiment was repeated 10 times. The final percentage identification for each dataset was taken as average of all repetitions.

3. Results and Discussion

The alignment of 133 complete chloroplast genome sequences with length ranging from 154,289 bp to 157,390 bp was analyzed using the PERL script. The information obtained from this analysis revealed that the aligned length of 173,636 bp harbours 21498 variable, 9721 singleton and 11602 parsimony informative sites.

3.1. *Identification of hyper-variable regions*

Nucleotide diversity (π) was used to find the hyper-variable regions in the solanaceous chloroplast genomes. The value of π was computed using sliding window analysis (window size: 600 bp and step size: 200 bp) for different regions. Fig. 2 shows the variation of π across the complete chloroplast genome sequences of 133 species of Solanaceae family. The value of π ranges from 0.00005 to 0.06278 with average value as 0.01148 for 697 window regions.

Out of total 697 window regions in the analysis, 33 were having π greater than mean+2SD (i.e. 0.03121) and therefore considered as hyper-variable regions [Bi *et al.* (2018)]. The actual locations of these regions in the chloroplast genome of respective species were computed to pinpoint genic/intergenic regions to which they belong (Supplementary Table 3). From the location of these regions in the genome, it was found that some of the regions are falling in the same genic/intergenic location. In these cases, region with maximum nucleotide diversity were

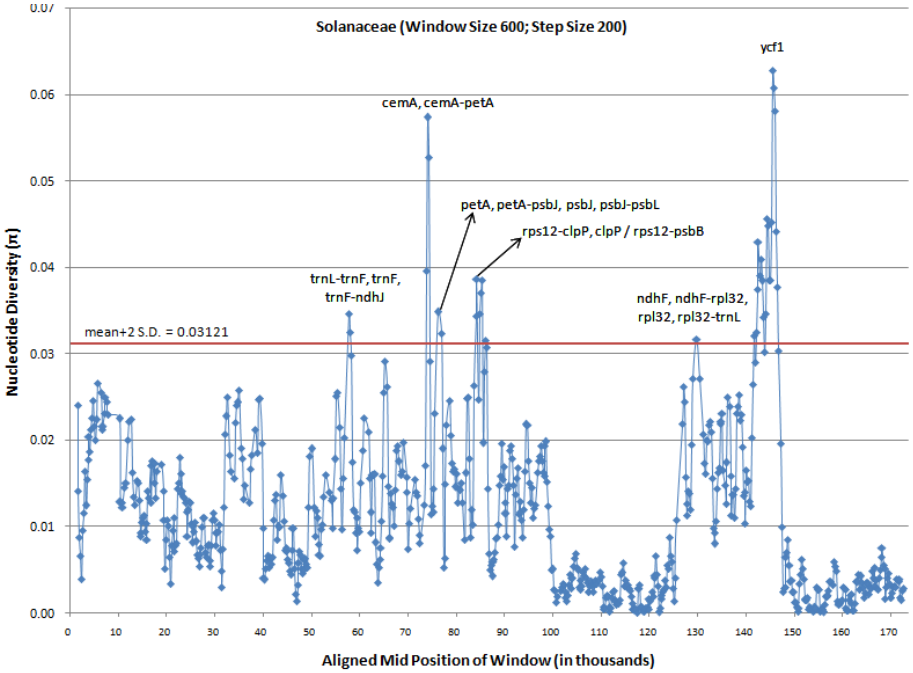

Fig. 2. Variation of nucleotide diversity (π) across complete chloroplast genome sequences of 133 species of Solanaceae family calculated using sliding window analysis (window length: 600 bp, step size: 200 bp). Y-axes: nucleotide diversity (π) of each window; X-axes: position of the midpoint of window in the aligned file.

selected for further analysis at species level (Fig. 2).

It was seen from the Fig. 2 that ycf1 is the most variable region which was also considered as the most promising DNA barcode in another study by Dong *et al.* (2015) for land plants due to its variability. The second most variable region observed in the present study i.e. cemA was also reported in six species of *Pyropia* (Bangiaceae) genus based on nucleotide diversity [Choi *et al.* (2019)]. In another study based on sliding window analysis of whole chloroplast genomes belonging to nine species of *Diospyros* (Ebenaceae) genus, ycf1, ndhF and ndhF-rpl32-trnL were among the eight hyper-variable regions [Li *et al.* (2018)]. Six highly variable regions rps15/ycf1, ndhF/rpl32, pasbE-petL, petA/psbJ, trnL/trnF and trnK/rps16 were reported in 72 species of Brassicaceae family based on nucleotide diversity using sliding window analysis [de Santana Lopes *et al.* (2018)]. Ten highly variable regions ycf1a, ycf1b, psbM-trnD, rpl31-trnL, rpl32-trnL, rps4-trnT-trnL, ycf4-cemA, petA-psbJ, rps11-rpl36-rps8 and trnK-rps16 were reported in *Fritillaria* (Liliaceae) genus based on nucleotide diversity [Bi *et al.* (2018)]. It can be seen that hyper-variable regions which were reported in the earlier studies on few genera/families are common to hyper-variable regions observed in the present study on Solanaceae family.

In Solanaceae family, few studies have been conducted at genus level for development of DNA barcode e.g. ITS2 region of nuclear ribosomal DNA was found to be potential DNA barcode for *Physalis* (Solanaceae) [Feng *et al.* (2016)]. However to the best of our knowledge, this is the first study in which hyper-variable regions based on chloroplast genome sequences are reported in solanaceous species at family level. These regions are required to be validated as potential DNA barcodes for identification of solanaceous species.

3.2. *Validation of hyper-variable regions*

3.2.1. *Validation at genus level*

Out of 19 genera, 7 genera have been represented by three or more species (Supplementary Table 2). Out of these 7 genera, three genera viz. *Physalis*, *Lycium* and *Dunalia* have representation of 3 plant species each while other four genera viz. *Solanum*, *Capsicum*, *Iochroma* and *Nicotiana* have representation of 6 or more species, therefore two sets of dataset files were prepared for all the 33 hyper-variable regions along with two standard barcoding regions matK and rbcL genes as recommended by CBOL for land plants. The first set consists of 121 plant species from 7 genera with at least three representatives of each genus. The second set consists of 112 plant species from 4 genera with at least six representatives of each genus. To compute variable, singleton and parsimony informative sites and to find percentage of correct species identification using distance based analysis in SPIDER and Machine learning algorithms in WEKA, each dataset file was stored in three formats and aligned separately using Clustal Omega command line version 1.2.2. A PERL script was executed on each aligned dataset file to compute variable, singleton and parsimony informative sites and the results are shown in Table 1. Each parameter contains two sub-columns, first column shows results from first set of files in which each dataset file consists of minimum 3 plant species for each genus and second column shows results from second set of files in which each dataset file consists of minimum 6 plant species for each genus.

Table 1. Characteristics of 33 hyper-variable window regions along with matK and rbcL

S. No.	Position of hyper-variable windows in MSA (Nucleotide Diversity)	No. of Sequences/ Genera		Alignment Length		Variable Sites		Singleton Sites		Parsimony Informative Sites	
		Set 1	*Set 2*	*Set 1*	*Set 2*	*Set 1*	*Set 2*	*Set 1*	*Set 2*	*Set 1*	*Set 2*
1.	145095-145812 (0.06278)	121/7	112/4	692	668	216	202	46	48	170	154
2.	145298-146091 (0.06081)	121/7	112/4	762	741	249	233	64	65	185	168
3.	145519-146321 (0.05806)	121/7	112/4	759	759	254	241	75	76	179	165
4.	73724-74394 (0.05733)	121/7	112/4	633	633	127	126	33	34	94	92
5.	73924-74624 (0.05268)	121/7	112/4	648	648	139	135	33	35	106	100
6.	143895-144679 (0.04555)	121/7	112/4	768	750	236	205	89	74	147	131
7.	144895-145518 (0.04525)	121/7	112/4	625	600	148	136	40	44	108	92
8.	144178-144894 (0.04479)	121/7	112/4	688	679	209	191	57	59	152	132
9.	145813-146527 (0.04418)	121/7	112/4	689	689	204	194	83	85	121	109
10.	142218-142949 (0.04298)	121/7	112/4	700	685	217	170	50	50	166	120
11.	142657-143447 (0.04085)	121/7	112/4	753	759	241	197	57	64	184	133
12.	73524-74123 (0.03956)	121/7	112/4	600	600	90	87	17	18	73	69
13.	142418-143193 (0.03897)	121/7	112/4	738	723	234	171	64	62	169	109
14.	83824-84704 (0.03861)	121/7	112/4	844	836	241	212	64	67	176	144
15.	144432-145094 (0.03845)	121/7	112/4	634	625	159	139	44	45	115	94
16.	84916-85605 (0.03844)	121/7	112/4	682	682	129	118	30	28	99	90
17.	144680-145297 (0.03844)	121/7	112/4	619	616	139	124	40	43	99	81
18.	142950-143688 (0.03843)	121/7	112/4	734	734	193	170	58	63	135	107
19.	146092-146780 (0.03772)	121/7	112/4	669	669	165	155	59	58	106	97
20.	141954-142656 (0.03737)	121/7	112/4	704	683	177	141	50	45	126	96
21.	84705-85376 (0.03705)	121/7	112/4	664	663	128	118	43	45	85	73
22.	75425-77015 (0.03481)	121/7	112/4	1473	1452	572	515	111	202	461	313
23.	143689-144431 (0.03462)	121/7	112/4	744	738	193	179	72	72	121	107
24.	84375-85115 (0.0346)	121/7	112/4	732	730	136	125	33	33	103	92
25.	56673-58057 (0.03459)	121/7	112/4	1214	1185	557	461	112	115	443	346
26.	83609-84374 (0.03428)	121/7	112/4	719	711	178	163	28	26	150	137
27.	143194-143894 (0.03424)	121/7	112/4	702	702	173	165	58	60	115	105
28.	56978-58330 (0.03251)	121/7	112/4	1178	1179	456	390	91	93	365	297
29.	141742-142417 (0.0325)	121/7	112/4	671	671	142	125	42	40	100	85
30.	76040-77222 (0.03236)	121/7	112/4	1118	1104	393	379	112	162	281	217
31.	141207-141953 (0.03206)	121/7	112/4	718	718	159	152	61	58	97	93
32.	128753-131079 (0.03166)	121/7	112/4	2022	2010	1003	902	218	225	783	674
33.	85606-86435 (0.03147)	121/7	112/4	784	782	227	207	63	69	164	138
34.	matK	121/7	112/4	1675	1675	239	223	113	107	126	116
35.	rbcL	121/7	112/4	1456	1435	99	90	33	35	66	55

Set 1 consists of minimum of three sequences (species) per genus; **Set2** consists of minimum of six sequences (species) per genus

Discrimination ability of all hyper-variable regions was evaluated using SPIDER and four different machine learning classification methods in WEKA. The percentages of correct identification rates for each region

along with minimum, maximum and average rates across all regions are shown in the Table 2. It was seen that with increase in minimum number of plant species for a particular genus, there is improvement in the correct identification rates. Distance based analysis in SPIDER and all the machine algorithms except Naïve Bayes gave correct identification rates of more than 93% at genus level for all the regions. Average correct identification rates across all regions ranges from 95% to 99% for these methods. Whereas Naïve Bayes gave average identification rate of 76% in case of datasets with minimum 3 species from each genus and 82% in case of datasets with minimum 6 species from each genus. The high identification rates signify that there is enough variability at genus level across hyper-variable regions under study.

Table 2. Position of 33 hyper-variable windows along with matK and rbcL, their nucleotide diversity and percentage correct identification of two datasets at genus level using different classification methods.

S. No.	Position of hyper-variable windows in MSA (Nucleotide Diversity)	%age correct identification at genus level									
		SPIDER (NN)		J48		Jrip		SMO		Naïve Bayes	
		Set 1	*Set 2*	*Set 1*	*Set 2*	*Set 1*	*Set 2*	*Set 1*	*Set 2*	*Set 1*	*Set 2*
1.	145095-145812 (0.06278)	95.04	99.11	96.04	98.85	95.29	99.11	95.37	99.11	84.07	95.35
2.	145298-146091 (0.06081)	95.87	100	94.63	100	93.22	99.64	96.85	100	91.73	98.24
3.	145519-146321 (0.05806)	97.52	100	94.97	98.24	95.79	99.11	97.53	100	87.62	94.67
4.	73724-74394 (0.05733)	95.87	99.11	96.70	99.11	96.52	100	96.70	99.12	71.09	76.82
5.	73924-74624 (0.05268)	96.69	99.11	96.70	99.10	96.21	99.10	96.70	99.10	72.42	78.18
6.	143895-144679 (0.04555)	99.17	100	96.88	99.11	93.74	98.57	98.35	100	71.18	76.82
7.	144895-145518 (0.04525)	95.04	99.11	94.63	98.05	93.72	97.61	95.47	99.12	71.10	76.82
8.	144178-144894 (0.04479)	97.52	99.11	98.34	99.11	95.88	99.56	97.51	99.11	78.53	84.87
9.	145813-146527 (0.04418)	95.87	100	96.72	98.91	94.82	97.19	96.79	100	71.09	76.82
10.	142218-142949 (0.04298)	97.52	100	96.69	99.10	95.62	99.10	97.52	100	78.53	85.76
11.	142657-143447 (0.04085)	96.69	100	97.50	100	96.75	100	96.77	100	79.36	84.88
12.	73524-74123 (0.03956)	96.69	100	96.71	100	95.64	99.03	97.54	100	71.09	76.82
13.	142418-143193 (0.03897)	97.52	100	97.52	100	96.52	100	96.12	100	71.09	84.86
14.	83824-84704 (0.03861)	98.35	100	94.49	98.11	96.16	99.11	96.38	99.91	77.71	83.96
15.	144432-145094 (0.03845)	93.39	99.11	97.44	99.83	95.94	98.77	96.69	99.12	71.09	76.82
16.	84916-85605 (0.03844)	97.52	100	96.71	99.30	96.71	100	97.53	100	71.11	76.82
17.	144680-145297 (0.03844)	95.04	99.11	96.53	98.59	93.80	98.23	96.12	99.11	76.72	82.80
18.	142950-143688 (0.03843)	96.69	99.11	96.69	99.11	96.19	98.74	96.03	99.11	78.53	84.86
19.	146092-146780 (0.03772)	94.21	99.11	95.71	98.24	95.60	98.24	95.77	99.11	71.09	76.82
20.	141954-142656 (0.03737)	97.52	100	95.12	98.39	94.72	97.30	96.76	100	71.10	76.82
21.	84705-85376 (0.03705)	97.52	100	97.12	100	96.87	100	97.53	100	71.10	76.82
22.	75425-77015 (0.03481)	95.87	97.32	96.20	98.03	93.89	98.48	94.72	98.20	86.81	88.06
23.	143689-144431 (0.03462)	99.17	100	98.35	99.09	94.88	99.09	99.17	100	71.09	76.82
24.	84375-85115 (0.0346)	95.04	100	95.87	99.11	96.36	100	97.52	100	71.09	76.82
25.	56673-58057 (0.03459)	95.87	99.11	95.87	98.20	95.21	98.57	95.21	99.11	92.58	100
26.	83609-84374 (0.03428)	96.69	99.11	97.51	99.11	97.27	100	96.69	100	71.10	76.82
27.	143194-143894 (0.03424)	96.69	99.11	97.53	100	93.89	99.09	96.69	99.09	77.71	83.95
28.	56978-58330 (0.03251)	95.04	100	97.51	100	95.12	97.26	95.46	100	92.56	100
29.	141742-142417 (0.0325)	98.35	100	94.55	98.09	95.54	98.82	95.37	100	71.09	76.82
30.	76040-77222 (0.03236)	95.04	100	95.12	97.36	94.62	96.80	95.94	98.22	79.79	85.73
31.	141207-141953 (0.03206)	97.52	100	96.13	100	95.63	99.45	97.54	100	71.09	76.82
32.	128753-131079 (0.03166)	95.87	98.21	94.08	98.06	95.14	98.68	96.72	99.12	86.81	93.82
33.	85606-86435 (0.03147)	97.52	100	96.92	100	95.67	99.92	96.69	100	71.09	76.82
34.	matK	97.52	100	96.71	99.11	96.04	98.94	97.54	100	71.09	76.82
35.	rbcL	96.69	99.11	97.54	100	96.71	99.57	96.71	99.12	71.09	76.82
Minimum %age		93.39	97.32	94.08	97.36	93.22	96.80	94.72	98.20	71.09	76.82
Maximum %age		99.17	100	98.35	100	97.27	100	99.17	100	92.58	100
Average %age		96.58	99.54	96.39	99.07	95.48	98.95	96.68	99.57	76.35	82.65

Set 1 consists of minimum of three sequences (species) per genus, **Set2** consists of minimum of six sequences (species) per genus, **SPIDER (NN)** = 'NearNeighbour' function in SPIDER, **J48** = Decision tree based classifier, **Jrip** = Rules based classifier, **SMO** = Support vector machine based classifier, **Naïve Bayes** = Bayesian-based classifier. Results are averaged over 10 fold cross-validation repeated 10 times for all four machine learning classification methods

3.2.2. *Validation at species level*

From location of 33 hyper-variable regions given in Supplementary Table 3, it was found that some of the regions are falling in the same genic/intergenic location and often overlap. In such cases, the region with maximum nucleotide diversity was selected for further analysis. This approach helped us to narrow down our analysis to 6 hyper-variable regions (Supplementary Table 4). In order to fulfil requirement of more than one sequence of a particular plant species, sequence richness regime was followed using BLASTN to find homologous

sequences related to same species for each of the above six regions. Furthermore, out of 133 sequences in the aligned window, the sequence with least gaps i.e. sequence with shortest length was used as query in the BLASTN search. A separate BLASTN search was also carried out using the standard matK and rbcL sequences as queries. The results were filtered based on 90% or more query coverage and stored in the local database. From this database, 4 separate dataset files i.e. files with minimum 2, 3, 4 and 5 sequences from same species were prepared in FASTA format for each region. In this way, a total of 32 dataset files belonging to 8 regions were created. The sequence names and description lines in these dataset files were further fine tuned to meet specific requirements of WEKA, SPIDER and in house written PERL script. All these datasets were then aligned separately using Clustal Omega command line version 1.2.2. Variable, singleton and parsimony informative sites were computed for each aligned dataset and the results are shown in Table 3.

Table 3. Characteristics of 6 hyper-variable windows along with matK and rbcL selected after sequence richness regime.

S. No.	Position of hyper-variable windows in MSA (Midpoint) Location of the Region	Minimum number of sequences for each species	Number of Sequences/ Species / Genera	Alignment Length	Variable Sites	Singleton Sites	Parsimony Informative Sites
1.	145095-145812 (145418) ycf1	2	368/95/11	692	247	14	233
		3	276/50/5	653	163	11	152
		4	228/35/2	629	124	5	119
		5	196/27/2	629	118	4	114
2.	73724-74394 (74023) cemA, cemA-petA	2	372/96/11	600	107	3	104
		3	280/51/6	600	91	4	87
		4	232/36/3	600	82	3	79
		5	200/28/3	600	76	1	75
3.	83824-84704 (84218) rps12-clpP, clpP / rps12-psbB	2	367/95/11	842	293	13	280
		3	275/50/5	830	243	28	215
		4	227/35/2	808	167	22	145
		5	195/27/2	805	175	18	157
4.	75425-77015 (76152) petA, petA-psbJ, psbJ, psbJ-psbL	2	360/92/11	1493	654	21	630
		3	274/50/5	1489	446	115	331
		4	226/35/2	1453	268	22	246
		5	194/27/2	1446	245	30	215
5.	56673-58057 (57623) trnL-trnF, trnF, trnF-ndhJ	2	351/92/11	1206	786	199	587
		3	257/46/5	1107	653	244	409
		4	215/33/2	1033	282	8	274
		5	187/26/2	1034	280	8	272
6.	128753-131079 (129801) ndhF, ndhF-rpl32, rpl32, rpl32-trnL	2	336/83/7	1972	848	17	831
		3	260/46/2	1888	532	42	490
		4	227/35/2	1875	484	39	445
		5	195/27/2	1867	456	40	416
7.	matK	2	481/136/16	1680	963	60	903
		3	337/65/10	1668	881	31	850
		4	259/40/5	1660	837	29	808
		5	219/31/5	1660	818	27	791
8.	rbcL	2	469/118/17	1439	266	110	156
		3	361/66/14	1439	200	65	135
		4	295/45/7	1440	166	58	108
		5	255/35/7	1440	162	56	106

Using program written in R language, "BestCloseMatch" function of SPIDER was executed on each dataset to find percentage of correct/ incorrect/ ambiguous/ not identified sequences. The aligned files were further analysed using NearNeighbour function of SPIDER and four different machine learning classification methods in WEKA to compute correct identification rates both at genus and species levels and the results are shown in Table 4. From the results, it was seen that for these 8 different regions, correct identification rates at genus level are almost similar with identification rates predicted earlier using sequences without enrichment as shown in Table 2. However at species level, the correct identification rates dropped significantly as compared to those obtained at genus level. The presence of large number of ambiguous sequences within threshold distance of 1% may explain the reason of less identification rates at species level.

3.3. *Comparisons of different methods and hyper-variable regions*

Machine learning algorithm SMO and near neighbour function in SPIDER performed best with highest correct identification rates both at genus and species levels. In another study by Hartvig *et al.* (2015), distance

based method 'Taxon DNA' and SMO also gave highest correct identification rates for 21 species of genus *Dalbergia* based on various DNA barcode regions. At genus level before sequence richness regime, 100% correct identification rate was achieved for few hyper-variable regions with all the methods in case of datasets with minimum 6 species for each genus. Naïve Bayes was less efficient than other machine learning algorithms due to requirement of large dataset for analysis. Weitschek *et al.* (2014) reported less efficiency of J48 and Jrip in comparison to other methods for empirical dataset of fungi and algae. However efficiency of these algorithms was comparable with SMO for datasets at genus level in the present study. Similar results were observed from the datasets obtained after sequence richness regime as shown in Table 4. At species level, maximum of 69% correct identification rate was achieved both in case of SMO and SPIDER. The average correct identification rate was highest in case of SMO, followed by SPIDER, Jrip, J48 and Naïve Bayes.

Table 4. Location of six hyper-variable windows along with matK and rbcL selected after sequence richness regime, Correct/Incorrect/ Ambiguous/Not Identified sequences using 'BestCloseMatch' function of SPIDER, percentage correct identification at genus and species levels using different classification methods.

S. No.	Location of hyper-variable windows	Min. No.	Genus Level Analysis						Species Level Analysis					
			SPIDER (BCM) Correct/ Incorrect/ Ambiguous / Not Identified	%age correct identification					SPIDER (BCM) Correct/ Incorrect/ Ambiguous/ Not Identified	%age correct identification				
				SPIDER (NN)	J48	Jrip	SMO	Naïve Bayes		SPIDER (NN)	J48	Jrip	SMO	Naïve Bayes
1.	ycf1	2	365/0/0/3	100	97.72	97.42	100	94.30	117/20/228/3	40.22	28.93	16.07	42.87	8.59
		3	274/0/0/2	99.64	99.28	99.28	99.64	97.86	78/17/179/2	37.68	37.63	27.77	42.20	11.07
		4	228/0/0/0	100	100	100	100	100	55/11/162/0	30.70	35.26	30.83	35.08	14.40
		5	196/0/0/0	100	100	100	100	100	54/7/135/0	34.18	37.31	32.51	38.35	15.92
2.	cemA, cemA-petA	2	372/0/0/0	100	97.38	96.76	100	88.72	64/8/300/0	26.08	24.33	12.42	31.00	7.74
		3	279/1/0/0	99.64	98.93	98.36	99.64	86.93	58/7/215/0	27.14	32.86	24.54	33.29	10.64
		4	232/0/0/0	100	100	100	100	95.78	51/6/175/0	26.72	33.61	27.70	33.87	12.48
		5	200/0/0/0	100	100	100	100	93.40	48/2/150/0	26.00	36.00	30.00	34.45	14.70
3.	rps12-clpP, clpP / rps12-psbB	2	363/0/0/4	99.73	96.65	96.49	99.92	89.95	103/18/242/4	40.05	28.88	16.57	43.93	9.50
		3	272/0/0/3	99.64	99.57	99.64	99.64	94.57	70/18/184/3	41.09	42.56	26.34	44.11	13.93
		4	225/0/0/2	100	100	100	100	100	47/15/163/2	37.89	41.19	27.18	43.26	15.17
		5	193/0/0/2	100	100	100	100	100	47/9/137/2	34.36	40.28	29.40	38.73	19.14
4.	petA, petA-psbJ, psbJ, psbJ-psbL	2	359/0/0/1	100	97.47	96.69	99.72	95.28	181/30/148/1	55.56	38.11	25.39	57.92	9.56
		3	273/0/0/1	99.64	99.64	99.64	99.64	94.55	125/26/122/1	49.64	48.69	39.94	53.52	11.54
		4	226/0/0/0	100	100	100	100	100	92/23/111/0	46.02	44.97	39.14	50.71	14.23
		5	194/0/0/0	100	100	100	100	100	78/13/103/0	42.78	42.88	40.95	49.21	16.95
5.	trnL-trnF, trnF, trnF-ndhJ	2	342/0/0/9	99.72	98.01	96.87	99.72	95.73	110/26/206/9	45.01	27.26	22.62	48.80	12.36
		3	252/0/0/5	99.61	99.00	99.23	99.62	98.54	76/17/159/5	42.80	36.92	35.04	44.74	19.30
		4	214/0/0/1	100	100	100	100	100	59/14/141/1	40.00	34.58	33.42	39.52	21.25
		5	186/0/0/1	100	100	100	100	100	52/11/123/1	42.78	38.41	37.34	41.96	25.29
6.	ndhF, ndhF-rpl32, rpl32, rpl32-trnL	2	335/0/0/1	100	97.87	97.57	100	96.43	194/41/100/1	69.05	51.38	35.43	69.29	13.98
		3	259/0/0/1	100	100	100	100	100	129/40/90/1	63.08	60.04	47.62	64.62	13.73
		4	226/0/0/1	100	100	100	100	100	112/29/85/1	63.88	61.48	48.88	65.40	16.70
		5	194/0/0/1	100	100	100	100	100	95/22/77/1	65.64	62.72	53.08	67.36	19.09
7.	matK	2	458/13/6/4	96.47	93.76	92.81	96.11	67.99	186/76/215/4	49.06	36.45	20.50	51.91	3.95
		3	331/3/0/3	98.81	96.00	95.67	98.81	79.59	156/32/146/3	57.57	51.70	38.00	60.78	5.64
		4	257/0/0/2	100	100	99.84	100	84.18	111/17/129/2	54.44	49.99	44.04	58.81	7.33
		5	216/0/0/3	99.54	99.50	99.50	99.54	84.94	92/11/113/3	52.97	53.26	47.27	58.27	8.67
8.	rbcL	2	462/3/0/4	99.15	97.66	95.78	99.62	74.84	120/32/313/4	47.33	42.47	21.68	49.81	4.50
		3	358/1/0/2	99.72	98.06	96.38	99.72	75.07	141/16/202/2	56.23	52.74	32.36	56.45	5.93
		4	293/0/0/2	100	97.96	98.00	99.93	81.72	119/11/163/2	59.66	55.23	39.77	59.16	7.26
		5	253/0/0/2	100	97.65	97.53	100	80.40	116/6/131/2	62.35	59.77	42.90	61.38	8.35

SPIDER (BCM) = 'BestCloseMatch' function in SPIDER, **SPIDER (NN) =** 'NearNeighbour' function in SPIDER, **J48** = Decision tree based classifier, **Jrip** = Rules based classifier, **SMO** = Support vector machine based classifier, **Naïve Bayes** = Bayesian-based classifier. Results are averaged over 10 fold cross-validation repeated 10 times for all four machine learning classification methods

Efficiency of methods also depends on the minimum number of sequences for a particular genus or species. It was observed that for genus as well as species specific datasets, the efficiency of methods increases with increase in minimum number of sequences representing each genus or species. At genus level, increase in number of sequences from each species from 2 to 5 resulted in 100% correct identification for all the methods. In fact, for most of the methods, 100% correct identification rate was obtained with even four samples per species. However at species level, the maximum correct percentage identification obtained was 69% using SMO and SPIDER. This lower level of accuracy is attributed to high percentage of ambiguous sequences (Table 4). Due to

increase in ambiguous sequences at species level, significant drop in the percentage of correct identification rates was observed for all the methods. The ambiguity itself depends upon the variability in the sequences as well as the number of samples representing each species. All the hyper-variable regions detected in this study were found to be ambiguous when considered for species level identification. The region 'ndhF, ndhF-rpl32, rpl32, rpl32-trnL' resulted in high (69%) correct identification using SMO and SPIDER even when using minimum two samples from each species. High identification rate was attributed to less ambiguity in this region as compared to other regions. The percent identification rate of above region was found to be higher as compared to well established markers for identification of plant species (matK and rbcL). The accuracy for these regions may further increase by including more rich datasets with a very high number of samples from each species as well as using more diverse species. Therefore, it is reflected from this investigation that SMO and SPIDER can be the best methods for identification at genus as well as at species levels. All the hyper-variable regions detected in this study can be used for identification of genera but 'ndhF, ndhF-rpl32, rpl32, rpl32-trnL' region emerged as the best region when it comes to identify plant species of Solanaceae family.

4. Conclusion

The present study involves utilization of *in silico* methods to identify potential DNA barcodes for solanaceous species and assess their discriminating potential using distance based method SPIDER and machine learning classification algorithms J48, Jrip, SMO and Naïve Bayes. Out of these methods, SMO and SPIDER were the best methods for species identification both at genus and species levels while Naïve Bayes was the poorest among all methods for our datasets. All the identified DNA barcodes performed well at genus level while DNA barcode 'ndhF, ndhF-rpl32, rpl32, rpl32-trnL' gave highest correct identification rate at species level. These barcodes can be validated experimentally by using DNA extraction, amplification and sequencing from more plant species belonging to Solanaceae family. With the availability of more sequences, the correct identification rates will also increase at species level. This study will provide a lead for the scientists to develop DNA barcodes for Solanaceous species which are yet to be explored for DNA barcoding studies at family level.

References

[1] Bi, Y. *et al.* (2018): Chloroplast genomic resources for phylogeny and DNA barcoding: a case study on Fritillaria. Sci Rep, **8**, Article 1184.
[2] Brown, S.D.J. *et al.* (2012): Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Mol Ecol Resour, **12**, pp. 562-565.
[3] CBOL Plant Working Group *et al.* (2009): A DNA barcode for land plants. Proc Natl Acad Sci U S A, **106**, pp. 12794–7.
[4] Choi, S.J.; Kim, Y. and Choi, C. (2019): Chloroplast genome-based hypervariable markers for rapid authentication of six Korean pyropia species. Diversity, **11**, Article 220.
[5] Cohen, W.W. (1995): Fast Effective Rule Induction. In *Machine Learning Proceedings*, pp. 115-123.
[6] de Kerdrel, G.A. *et al.* (2020): Rapid and cost-effective generation of single specimen multilocus barcoding data from whole arthropod communities by multiple levels of multiplexing. Sci Rep, **10,** Article 78.
[7] de Santana Lopes, A. *et al.* (2018): The Crambe abyssinica plastome: Brassicaceae phylogenomic analysis, evolution of RNA editing sites, hotspot and microsatellite characterization of the tribe Brassiceae. Gene, **671**, pp. 36-49.
[8] Dong, W. *et al.* (2015): ycf1, the most promising plastid DNA barcode of land plants. Sci Rep, **5**, Article 8348.
[9] Feng, S. *et al.* (2016): Application of the ribosomal DNA ITS2 region of physalis (Solanaceae): DNA barcoding and phylogenetic study. Front Plant Sci, **7,** Article 1047.
[10] Gao, T. *et al.* (2010): Evaluating the feasibility of using candidate DNA barcodes in discriminating species of the large Asteraceae family. BMC Evol Biol, **10**, pp. 1–7.
[11] Hall, M. *et al.* (2009): The WEKA data mining software: An update. ACM SIGKDD Explor Newsl, **11**, pp. 10-18.
[12] Hartvig, I. *et al.* (2015): The use of DNA barcoding in identification and conservation of rosewood (Dalbergia spp.). PLoS One, **10**, Article e0138231.
[13] Hebert, P.; Ratnasingham, S.; DeWaard, J.R. (2003): Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc London, **270**, pp. 96–99.
[14] Hebert, P.D.N. *et al.* (2003): Biological identifications through DNA barcodes. Proc Biol Sci, **270**, pp. 313–21.
[15] Hollingsworth, P.M. (2011): Refining the DNA barcode for land plants. Proc Natl Acad Sci, **108**, pp. 19451–19452.
[16] John, G.H. and Langley, P. (1995): Estimating Continuous Distributions in Bayesian Classifiers. Proc Elev Conf Uncertain Artif Intell, pp. 339-345.
[17] Katoh, K. and Standley, D.M. (2013): MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol, **30**, pp. 772-80.
[18] Kim, H.M. *et al.* (2014): DNA barcoding of orchidaceae in Korea. Mol Ecol Resour, **14**, pp. 499–507.
[19] Krawczyk, K.; Szczecinska, M.; Sawicki, J. (2014): Evaluation of 11 single-locus and seven multilocus DNA barcodes in Lamium L. (Lamiaceae). Mol Ecol Resour, **14**, pp. 272–285.
[20] Kress, W.J. *et al.* (2005): Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci U S A, **102**, pp. 8369–74.
[21] Kumar, S.; Stecher, G. and Tamura, K. (2016): MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol, **33**, pp. 1870-1874.
[22] Li, H.Q. *et al.* (2012): Evaluation of six candidate DNA barcoding loci in Ficus (Moraceae) of China. Mol Ecol Resour, **12**, pp. 783–790.
[23] Li, W. *et al.* (2018): Interspecific chloroplast genome sequence diversity and genomic resources in Diospyros. BMC Plant Biol, **18**, Article 210.
[24] Librado, P. and Rozas, J. (2009): DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics, **25**, pp. 1451-1452.
[25] Liu, J. *et al.* (2014): Identification of species in the angiosperm family Apiaceae using DNA barcodes. Mol Ecol Resour, **14**, pp. 1231–1238.

[26] Luo, K. *et al.* (2010): Assessment of candidate plant DNA barcodes using the Rutaceae family. Sci China Life Sci, **53**, pp. 701–708.
[27] Pang, X. *et al.* (2011): Applying plant DNA barcodes for Rosaceae species identification. Cladistics, **27**, pp. 165–170.
[28] Pires, A.C.; Marinoni, L. (2010): DNA barcoding and traditional taxonomy unified through Integrative Taxonomy: a view that challenges the debate questioning both methodologies. Biota Neotrop, **10**, pp. 339–346.
[29] Platt, J. (1999): Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In *Advances in Kernel Methods, Support Vector Learning* ed. B. Schölkopf, C.J.C. Burges and A.J. Smola. The MIT Press, Cambridge, pp.185–208.
[30] Rosario, L.H. *et al.* (2019): DNA barcoding of the solanaceae family in puerto rico including endangered and endemic species. J Am Soc Hortic Sci, **144,** pp. 363–374.
[31] Salzberg, S.L. (1994): C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach Learn, **16**, pp. 235-240.
[32] Savolainen, V. *et al.* (2005): Towards writing the encyclopaedia of life: An introduction to DNA barcoding. Philos Trans R Soc B Biol Sci, **360**, pp. 1805–1811.
[33] Schoch, C.L. *et al.* (2012): Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A, **109**, pp. 1–6.
[34] Singh, B.P. *et al.* (2020): CpGDB : A Comprehensive Database of Chloroplast Genomes. Bioinformation, **16**, pp. 171–175.
[35] Wang, W. *et al.* (2010): DNA barcoding of the Lemnaceae, a family of aquatic monocots. BMC Plant Biol, **10**, Article 205.
[36] Weitschek, E.; Fiscon, G. and Felici, G. (2014): Supervised DNA Barcodes species classification: Analysis, comparisons and results. BioData Min, **7**, Article 4.
[37] Yang, J.B. *et al.* (2012): Applying plant DNA barcodes to identify species of Parnassia (Parnassiaceae). Mol Ecol Resour, **12**, pp. 267–275.
[38] Zhang, Z.L. *et al.* (2015): DNA barcoding in medicinal plants: Testing the potential of a proposed barcoding marker for identification of Uncaria species from China. Biochem Syst Ecol, **60**, pp. 8–14.

Authors Profile

Bhupinder Pal Singh received B.E. (E&EC) degree from Punjab Engineering College, Chandigarh and M.Tech. (IT) from Punjab Technical University, Jalandhar. He is pursuing Ph.D. from I.K. Gujral Punjab Technical University, Jalandhar. Presently he is working as System Manager, Centre for IT Solutions, Guru Nanak Dev University, Amritsar. His areas of interest include development of web based applications using vb.net, machine learning algorithms, bioinformatics, DNA barcoding, database designing and statistical analysis of research data. He has more than 15 years of experience in teaching post graduate students and software development for the University.

Dr. Ajay Kumar received his Master's degree in Electrical Engineering from Punjab Technical University, Jalandhar in 2003. He did his Ph.D. from Punjab Technical University, Jalandhar in 2010 in the field of Design and Analysis of Electromagnetic Devices. Presently he is working as Professor in Department of Electronics and Communication Engineering, Beant College of Engineering and Technology, Gurdaspur. He has a vast experience of guiding M.Tech. and Ph.D. students. His research interests include design of electromagnetic devices and digital signal processing. He has more than 40 research publications to his credit.

Dr. Harpreet Singh has completed Advanced PG Diploma in Bioinformatics from JNU, New Delhi. He did his Ph.D. from Guru Nanak Dev University, Amritsar. Presently he is working as Head, PG Department of Bioinformatics, Hans Raj Mahila Maha Vidyalaya, Jalandhar. He is also appointed as Secretary, APBioNET; Finance Secretary, Bioclues.org and Representative GALAXY India. His research interest covers the areas of Structural Bioinformatics, Molecular Graphics, Data Analytics, and Machine learning. He has published 19 research papers and 2 book chapters in journals of International repute. He has 15 years of experience in the corporate and academic sector in the field of Bioinformatics. As a member of International organizations, he is involved in collaborative research with teammates from various countries.

Dr. (Mrs.) Avinash Kaur Nagpal is working as Professor in the Department of Botanical and Environmental Sciences, Guru Nanak Dev University, Amritsar, Punjab, India. She completed her Ph.D. in Biology in the year 1988 from the same University and joined as Faculty (Lecturer) in the year 1989. She has teaching experience of 33 years and research experience of 39 years. Her research interests include Environmental monitoring and genotoxicity, antimutagenic/ anticarcinogic potential of medicinal plants, Plant tissue culture, Plant databases, Bioinformatics etc. She has supervised 18 Ph.D. students and 9 are presently pursuing their Ph.D. under her supervision. She has published more than 150 research/ review articles in journals of International repute, completed 6 major research projects and had been the coordinator of major research grants of University ('Environmental Management' component of University with Potential for Excellence (UPE)) and Department (Departmental Research Support (DRS)) levels. She has held number of administrative positions in the University. She has been conferred with number of Awards including Biotechnology National Associateship in 1995-96 to have training in India and 3 months component for training in a Foreign Lab (AIST, Tsukuba, Japan) in 1998, travel grants by CSIR and DBT to present her research abroad (Berlin, Germany; Manchester, UK, Tsukuba, Japan). She has presented her work in more than 60 conferences (both national and international) and also chaired scientific sessions in many of them. She had been an active member of organizing committees of more than 10 conferences.