Bioinformatics Analysis on DNA Barcode Sequences for Species Identification: A Review

Huyen-Trang Vu¹,² and Ly Le²*

¹Faculty of Biotechnology, Nguyen Tat Thanh University, 298A-300A Nguyen Tat Thanh Street, Ward 13, Ho Chi Minh City, District 4, 72820, Vietnam.
²Faculty of Biotechnology, International University - Vietnam National University, Vietnam.

Authors’ contributions

This work was carried out in collaboration between both authors. Author HTV designed the study, performed the statistical analysis, wrote the protocol, managed the literature searches and wrote the first draft of the manuscript. Author LL managed the analyses of the study. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/ARRB/2019/v34i130142

Editor(s):
(1) Dr. Xiao-Xin Yan, Professor, Department of Anatomy and Neurobiology, Xiangya School of Medicine (CSU-XYSM), Central South University, China.

Reviewers:
(1) Bhaba Amatya, Tribhuvan University, Nepal.
(2) Bipinchandra B. Kalbande, Rashtrasant Tukadoji Maharaj Nagpur University, India.
Complete Peer review History: http://www.sdiarticle4.com/review-history/53419

Received 12 October 2019
Accepted 22 December 2019
Published 28 December 2019

ABSTRACT

Classification of organisms is the primary step in management of biodiversity, breeding, conservation and development of populations and distinguishing adulterant objects. There are many approaches in taxonomic identification, from morphological, PCR-based to sequence-based techniques. Molecular methods give more accurate results than morphological comparisons and are independent of plant stages. PCR-based methods are low-cost but their limited information gives less reproducibility and can only distinguish samples among determined groups. In contrast, in sequence-based methods each nucleotide site is considered as genetic information hence a sequence of nucleotide represents large data, which is highly specific and more stable than PCR bands. Establishment of worldwide DNA library for barcoding is essential. There were previous reviews on screenings and applications of barcodes in different taxa. In this review we discussed common bioinformatics analyses as well as some new improved techniques relying on barcoding approaches.

*Corresponding author: E-mail: ly.le@hcmiu.edu.vn;
Keywords: Molecular classification; bioinformatics tools; DNA barcoding; sequence analysis; identification technique.

ABBREVIATIONS

AFLP : Amplified Fragment Length Polymorphism;
ARMS : Amplification Refractory Mutation System;
BA : Bayesian;
BLAST : Basic Local Alignment Search Tool;
Cp : Complete;
cpDNA : Chloroplast DNA;
IR : Inverted Repeat;
LSC : Large single-copy region;
MAP : Maximum a posteriori;
ML : Maximum Likelihood;
MP : Maximum Parsimony;
NGS : Next Generation Sequencing;
NJ : Neighbor-joining;
OUT : Operational Taxonomic Units;
PCR : Polymerase Chain Reaction;
RADP : Random Amplified Polymorphic DNA;
RFLP : Restriction Fragment Length Polymorphism;
SCAR : Sequence Characterized Amplified Region;
SNP : Single Nucleotide Polymorphism;
SSC : Small single-copy region;
SSR : Simple Sequence Repeat;
UPGMA : Unweighted Pair Group Method with Arithmetic mean.

1. INTRODUCTION

Collection of genetic information, for looking up the origin of a wide range of organisms linked all over the world, is an advanced and essential idea for the protection of species, phylogenetic inference, management and development of genetic diversity [1,2]. Morphological methods show limitations of accuracy and high reliance on reproductive organs. PCR-based methods can overcome these above problems with just a small piece of sample. However, amplification techniques can only be effectively applied to samples of a defined group using RAPD, RFLPs, AFLPs [3] or to samples containing specific genes using species-specific PCR [4-6]. An unknown taxon cannot be determined using PCR-band techniques.

Since each site of sequence is considered as a character in bioinformatics analysis, the sequence-based method gives more variable information. This approach allows the read of every nucleotide among the samples. Specific and stable features of monomers are useful in evaluating genetic relationship of a new query sample based on the available sequence library and thus allow to consult origin of the unknown homologous taxon. DNA polymorphism may even provide more information than proteins due to the degradation of genetic code and the presence of large non-coding stretches [7]. DNA fragments represented for the organism can be used as an identifying sequence like the human fingerprint (Fig. 1). This is the most common method in molecular identification strategy today. In animal, the highly conserved sequence of mitochondrial cytochrome c oxidase subunit I (COI), which relates to oxidative phosphorylation for metabolism, was effectively used as barcode in diverse animal species [8,9]. Nevertheless there was no such effective barcode for plants.

![DNA barcoding scheme](https://en.wikipedia.org/wiki/DNA_barcoding)

A numerous studies of barcoding for plants have been conducted. To date, sequencing techniques for determining species using next generation sequencing (NGS) in plants are limited primarily to agricultural crops [10-13] due to their high cost and time. Therefore short sequences are still considered as convenient and effective tools due to the quick and accurate sequencing [14]. A number of studies have reviewed on finding and applying of DNA barcoding in which the efficiency of different biomarkers have been summarized [15-22]. Others discussed the effectiveness of different barcoding techniques, criteria and measurements [7,23,24]. In this review, we discussed some bioinformatics tools in identification analysis of barcoding studies. We also presented some prospects and developed techniques relied on barcoding approaches.

2. COMMON BIOINFORMATICS ANALYSES IN IDENTIFICATION TECHNIQUES USING SEQUENCES

Bioinformatics procedure for species identification using sequences comprises two basic steps: First, the sequence alignment should be conducted on the basis of a
comparative step. This alignment can be performed by two methods. The novel sequences are pairwise aligned against the available sequences in certain databases (similarity search) [25,26] or studied sequences are aligned against each other in a specific set of data (many-against-each other search) [2,27,28]. Following, based on these alignment data, similarity or variation are investigated. This step is performed by evaluating such parameters as GC% content [27,29], genetic distance [30-33], variable sites [32,34], indel appearances [35], or monophyletic clusters [36-38], which can indicate typical characters of the examined sequences. These measurements vary from different studies as they depend on alignment and identification methods (Table 1). Final target of this step is to decide whether the species are different or not. In some studies, species resolution was calculated by counting the number of identified species out of the total number of examined species [2,31,39].

In the alignment step, sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment). For BLAST (Basic Local Alignment Search Tool) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and FASTA [40] programs, each examined sequence is considered a query sequence. A local pairwise alignment is running between two biological sequences: the query sequence and each of database sequences. In contrast to similarity search, alignment process in many-against-each other search is based on two stages, the pairwise alignments followed by multiple alignments. In multiple alignment, whether global alignment or local alignment should be applied depending on similarity level and difference in sequence lengths. ClustalW is a fully automatic program for global multiple alignment of DNA and protein sequences [41]. However, for MAFFT [42] and MUSCLE [43] programs, users can select suitable aligning strategies. Global alignment is effective across their entire length (global alignment) or only in certain regions (local alignment). For BLAST (Basic Local Alignment Search Tool) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and FASTA [40] programs, each examined sequence is considered a query sequence. A local pairwise alignment is running between two biological sequences: the query sequence and each of database sequences. In contrast to similarity search, alignment process in many-against-each other search is based on two stages, the pairwise alignments followed by multiple alignments. In multiple alignment, whether global alignment or local alignment should be applied depending on similarity level and difference in sequence lengths. ClustalW is a fully automatic program for global multiple alignment of DNA and protein sequences [41]. However, for MAFFT [42] and MUSCLE [43] programs, users can select suitable aligning strategies. Global alignment is effective

<table>
<thead>
<tr>
<th>Alignment methods</th>
<th>Species Identification methods</th>
<th>Identification criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity search</td>
<td>Blast</td>
<td>Correct Identification</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ambiguous Identification</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incorrect Identification</td>
</tr>
<tr>
<td>Many-against-each other</td>
<td>Genetic distance-based</td>
<td>Nearest distance</td>
</tr>
<tr>
<td>search</td>
<td></td>
<td>Correct Identification</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ambiguous Identification</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incorrect Identification</td>
</tr>
<tr>
<td>Many-against-each other</td>
<td>Best match</td>
<td>Correct (match)</td>
</tr>
<tr>
<td>search</td>
<td></td>
<td>Ambiguous</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incorrect (mismatch)</td>
</tr>
<tr>
<td></td>
<td>Best close match</td>
<td>Correct (match)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ambiguous</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incorrect (mismatch)</td>
</tr>
<tr>
<td></td>
<td>All species barcodes</td>
<td>Correct (match)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ambiguous</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incorrect (mismatch)</td>
</tr>
<tr>
<td>Many-against-each other</td>
<td>Barcoding gap</td>
<td>Intra-specific genetic distance</td>
</tr>
<tr>
<td>search</td>
<td></td>
<td>Inter-specific genetic distance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Barcoding gap</td>
</tr>
<tr>
<td>Many-against-each other</td>
<td>Tree-based</td>
<td>Monophyletic</td>
</tr>
<tr>
<td>search</td>
<td></td>
<td>Paraphyletic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polyphylectic</td>
</tr>
<tr>
<td>Many-against-each other</td>
<td>Character-based (nucleotide</td>
<td>Variable sites</td>
</tr>
<tr>
<td>search</td>
<td>polymorphism)</td>
<td>Indels</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sequence lengths</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GC% contents</td>
</tr>
</tbody>
</table>
when the input sequences share global homology, and the similarity level is high. On the other hand, with fragmentary and divergent sequences, local alignment would be the better option [44]. For "BLAST" method, the target is searching for the best homologous sequences (best hits) from the available databases (GenBank, BOLD, others). Based on this background, "Correct Identification" means that: the best hit is the sequence with species name as expected. "Ambiguous Identification": the best hits are some sequences belong to different species including the expected species. "Incorrect Identification": the best hit does not match with expected species [25,26,28,45]. However, because "BLAST" depends on available sequences reported in databases, query species must be already included in a database otherwise the result will be a failure, therefore a negative "Incorrect Identification" will occur.

For genetic distance-based methods, the "best hit" feedback of "Correct", "Ambiguous" and "Incorrect" criteria are similar to that in "BLAST", but based on the smallest genetic distances ("Nearest Distance") [28] or the most similarity ("Best match", "Best close match") [46-49]. The query sequence is compared with each other in the given data set. "Best close match" differs from "Best match". A similarity threshold value (e.g. 95%) of all intraspecific distances is established to determine how similar a barcode match needs to be and the results under this similar value (No match) would be removed before identification step.

For "All species barcodes" methods, a list of sequences sorted by similarity to each query using the same threshold as for "Best close match" are assembled. If the query is followed by all conspecific barcodes with at least two ones then the species is identified. If the query is followed by only one or some of conspecific barcodes then the identification is ambiguous. If the query is followed by none of conspecific barcodes but other species then misidentification occurs [47].

Barcoding gap method analyses the divergence between intra-specific and inter-specific genetic distances of each query versus other conspecific and hetero-specific sequences in a data set [25-28, 30-33,50,51]. However this method may be not precise in case of using mean instead of smallest inter-specific distances versus largest intra-specific distances [52]. If barcoding gap exists, the species is successfully identified.

Another method is the use of nucleotide polymorphism features, such as variable sites, sequence length variation, indel information, GC% content [4,19,32,53,54] as tools of identification. This approach is known as character-based method in which each nucleotide is consider as the fifth character beside four traditional characters A, T, C and G.

Regardless of which method is used, the core property of the identification strategy is that all conspecific sequences should be grouped together without blending with any other species. To address this issue, the tree-based method is a simple and visualized approach that is most common in such classification studies [2,25-27,29,34,36-39,45,46,48,49,55]. Two species are completely separated when all sequences of one species are clustered in a monophyletic branch [56]. There are some problems that should be noticed to avoid errors in identification process. Small sample size [4,26,29,36,51,57-59] or a wide range but less con-level of taxonomic samples [27,60,61] may lead to over-fitting [56].

3. COMMON BIOINFORMATICS TOOLS IN IDENTIFICATION TECHNIQUES USING SEQUENCES

For various measurements, different bioinformatics tools could be used also. The BLAST method is presented by BLASTn program from NCBI. Intra- and inter-specific genetic distances, matching sequences and clustering sequences based on pairwise distances can be calculated in "Best match", "Best close match" and "All species barcodes" methods by Species Identifier tool using TaxonDNA software. When multiple regions are compared for selection of optimal biomarkers, TaxonGap program is used to infer intra- and inter-specific distance for "Nearest" method in high-throughput sequencing researches [62].

For tree-based reconstruction, probability model should be estimated. Neighbor-Joining (NJ), Maximum Likelihood (ML), Maximum Parsimony (MP), Bayesian (BA) or Unweighted Pair Group Method with Arithmetic mean (UPGMA) are common phylogenetic algorithms inferred along with suitable models. Which method should be used in different studies is still a problem that need to be noticed for obtaining the most accurate results. Some studies performed comparisons on different methods [27,45]. Although algorithms are inferred from suppositions, having a thorough understanding of
our algorithms and data is the best way to achieve the highest efficiency. The standard principal of tree building is to examine all possible topologies or certain topologies that represent the true structure. Neighbor-joining performance [39,46] is a time-consuming method in reconstructing phylogenetic trees. It finds pairs of operational taxonomic units (OTUs) that minimize the total branch length at each stage of OTUs clustering and starts with a star-like tree [63]. However the reliability of Neighbor-joining tree is a problematic issue [47] as it quickly generates only one phylogenetic tree while others may be more fitting. In contrast, Maximum Likelihood (ML) accounts the probability for all events that can happen simultaneously and the best tree is supported at a higher probability. Hence ML become a powerful and professional method in phylogenetic algorithm [64,65] although it requires significant running time for optimal tree especially with large data [66,67].

The Maximum parsimony method is based on the least character state changes required to infer a tree. In case of heterogeneous evolution, maximum parsimony (MP) is strongly biased towards recovering an incorrect tree. However this method outperforms ML over a wide range of conditions, including low and moderate heterogeneity [68].

While both ML and MP use the probabilities called likelihood, the Bayesian (BA) technique represents the posterior probability (Bayes' rule). A known theory called prior is implemented. The posterior is in direct proportion with the product of likelihood and prior [67]. Bayesian posterior probability gives more generous estimates of subtree reliability than the maximum likelihood analysis, particularly when using the gamma distribution modeling [65,67]. When the size of data is small, the probabilities inferred from ML may be over-fitting. Maximum a posteriori (MAP) can be taken accounts to solve this problem.

Regarding genetic distance and tree-based methods, MEGA is the most popular software used due to its friendly interface and optimal analysis time. Since genetic variations have to be inferred from evolutionary distance matrices, MEGA versions have integrated these evolutionary models into their program along with different algorithms, i.e. Neighbor-joining, Maximum Likelihood, Maximum Parsimony and UPGMA [69]. However, the number of models in MEGA is pruned for its convenience. For more accurate and reliable results, PAUP* and MRBAYES are often used although it takes much more time. Maximum Likelihood and Maximum Parsimony can also be calculated using PAUP* [70]. Data was estimated for revolutionary models using software jModelTest [71] before running in PAUP*. The software MRBAYES analyses the Bayesian Inference [72]. However, the use of PAUP* and MRBAYES needs bioinformatics skill to run the program, which is quite complex for some researchers.

4. RELATION BETWEEN MOLECULAR IDENTIFICATION AND APPLICATION FOR PHYLOGENETIC STUDY

For discrimination of species, tree-based method seems to be the most common technique in many different studies on variety taxa. Phylogenetic tree can be presented for both phylogenetic and barcoding studies. However phylogenetic analysis represents the measurement and estimation of evolutionary past. Whereas barcoding analysis is used to identify taxa in certain taxonomic groups [2,39,48] or to determine a new taxon [60,73] by DNA sequence comparison. In barcode-phylogenetic tree, the differences between nucleotide characters are more important than the way they form through the evolutionary time. This means that taxa of the same species should be clustered in a monophyletic branch and the different ones should be distributed in separated clades [45,49]. The length of branches and the members of clusters are not strictly evaluated. Therefore, barcode-phylogenetic tree is not really a phylogenetic tree. However, as the trees are built based on specific DNA sequences, information from them can be used for phylogenetic investigations. Barcoding and phylogenetic relationships of species have also been studied in combination previously [28,29,36,74]. For phylogenetic relationship analyses, homologous variations at each alignment site are considered. In this case, Maximum Likelihood or Maximum Parsimony is used instead of Neighbor-joining [27].

Yang et al. (2013) used six data sets including sequences of whole complete (cp) genomes, protein-coding exons, large single-copy region, small single-copy region, inverted repeat region, introns and spacers for phylogenetic tree establishments to indicate congruent among different data partitions (Fig. 2). Only cp genome tree and the intron and intergenic spacer tree gave genetic similarity. The relationships between seven species C. danzaiensis,
Fig. 2. Phylogenetic trees constructed from different data partitions from whole chloroplast genome, with all clades were absolutely separated by high genetic variation [75]

(CP Genome: complete genome; LSC: large single-copy regions; SSC: small single-copy regions; IR Region: inverted repeat region; numbers above the lines on the left indicate the maximum parsimony bootstrap of each clade >50%; numbers above the lines on the right indicate the Bayesian posterior probabilities; numbers below each branch are the maximum likelihood bootstrap of each clade >50%)

C. pitardii, C. impressinervis, C. cuspidata, C. taliensis 7, C. taliensis 8 and C. yunnanensis were not consistent in other trees [75]. This result posed a question: whether a partial DNA sequence region is sufficient to represent species and the relationship between them. The great data of whole genome might be more reliable for estimating the evolution compared to shorter barcodes.

5. THE USE OF COMPLETE CHLOROPLAST GENOME AS ULTRA BARCODE (SUPER-BARCODE)

Short DNA sequences can solve most questions in identification at species and above-species levels but still have some limitations with closely related taxa. Some researchers have suggested the way of serving complete chloroplast genome as a single barcode in plants [31,76]. This method was called ultra-barcoding by Kane et al. when they compared nine complete plastid genotypes of Theobroma cacao and one related species T. grandiflorum with a control GenBank accession [77]. The complete chloroplast DNA (cpDNA) could absolutely separate all examined intra-species in the study. This approach promise new effective applications in identification at species and below-species level [78]. Fritillaria, a popular herbal medicine genus in China, experienced some efforts in characterizing closely related species using universal
molecular markers (ITS, trnL-trnF etc.) but could not be distinguished entirely [79-81]. Using complete chloroplast sequences, phylogenetic analyses of eight *Fritillaria* species were well resolved in the study of Bi et al. [82]. In other studies, even no potential regions in cp genome were proposed but the complete genome itself had a capability in distinguishing samples as a single barcode [83]. The interesting thing is that you can also use this meta-data to develop potential mini-barcodes which are high variability for quick authentication of certain taxa [84-88].

This genomic strategy not only allows the use of complete cpDNA as a single barcode, but also facilitate the utilization of other data partitions to identify plants. Protein-coding exons, large single-copy region, small single-copy region, inverted repeat region, introns and intergenic spacers (Fig. 2) [75], pseudogenes [89] or the differences between size, number of annotated genes [86] could be taken into accounts in phylogenetic analyses. Length-variations which based on the differences of indels (insertions and deletions) at certain locations of cpDNA genome were also considered as effective barcodes [35,90].

Genetic information of this super-data is definitely great, enough to avoid the analytical limitation by different bioinformatics methods such as Maximum Likelihood (ML), Maximum Parsimony (MP) or Bayesian (BA). This means that the topologies of phylogenetic trees based on these three algorithms, which be usually congruent using short DNA sequences, are highly similar in the case [75].

Since the cost for whole genome sequencing has significantly decreased, from $2.7 billion in 2003 for the first human genome to $300,000 in 2006 and from there to $1,000 in 2016, a series of studies on plant barcoding was published in the next two years 2017 and 2018. Along with sequencing and assembling techniques, whole-plastome barcode may offer more informative sites and is considered as accurate and effective single barcode for identification in plants.

6. APPLICATION OF BARCODES TO DEVELOP OTHER IDENTIFICATION TECHNIQUES

Nucleotide information can be used to develop some species-specific amplification markers for quick and cheap identification of specific subjects. PCR primers derived from these techniques only react upon annealing to specific DNA sequences and more specific and reproducible results than random amplifications. SCAR (sequence characterized amplified regions) technique succeeded in authentication three species of *Paphiopedilum armeniacum*, *Paphiopedilum micranthum*, *Paphiopedilum delenatii* and their hybrids by developing three species-specific primer pairs from ITS sequences [6]. Some studies focused on developing and comparing simple sequence repeat (SSR) system among screened taxa [86,91,92]. Kim et al. designed new primers for ARMS (amplification refractory mutation system) technique based on specific insertion in sequence of *Cypripedium macranthos* and SNPs (single nucleotide polymorphisms) in sequences of *Cypripedium japonicum* and *Cypripedium formosanum* located inside atpF-atpH barcode. These three primer pairs can be used in combination to distinguish four *Cypripedium* species with different-size bands on the electrophoresis gel [4]. RFLP (restriction fragment length polymorphism) technique is a type of random PCR-based method in identification of subjects. However, RFLP based on the combination of ITS and some chloroplast sequences gave more reproducible and successful identification of native *Dendrobium* species in Thailand by Peyachoknagula et al. [93]. Therefore using known barcodes to develop other time- and cost-saving methods can also support molecular identification.

Furthermore, metabarcoding using high-through put sequencing can help identify a variety of species in multiple samples. A detection of species was performed simultaneously in 55 commercial salep products based on ITS barcode. Each sample was found to contain 1-55 species [94]. *RbcL*, *matK* and ITS were also used in combination to determine 16 orchid species as components of a common food named *Chikanda* in Zambia [95]. Based on this foundation, the authors alerted the over-harvesting condition and called for the conservation of these rare orchids. This technique also opened a new trend in application of DNA barcodes.

7. CONCLUSIONS

Barcoding technique had been shown to be useful in many practical applications and classification studies. This method promises more optimal results than the PCR method especially when the price of the sequencing is
decreasing. In barcoding technique using sequences, the chosen loci and algorithms are directly related to the identification results.

Whole genome comparison based on next generation sequencing and high-throughput sequencing has shown to be more convenient and contain greater data than traditional barcoding. However, in terms of price, it is about $200 per sample up to now, which is 20 times more expensive than mini-barcodes ($11-13). In terms of technology, a powerful computer, which is not available in small laboratories, is needed to manage large genome data. Besides, it also takes a considerable time from the sequencing to the analyzing. For those reasons, traditional barcoding methods are still popular and effective in many cases. Hence these two barcoding trends can be performed according to the demands and conditions of different laboratories. In parallel, the sequencing techniques and analyzing tools should be improved to make it simpler and more convenient for researchers. The next orientation of identification should be a species identification gadget that is portable and requires no amplification step.

ACKNOWLEDGEMENT

The study was supported by Air Force Office of Scientific Research and Asian Office of Aerospace Research and Development (grant number FA23861514119). All authors declared no other competing financial interests.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES


14. Hebert PD, Cywinska A, Ball SL, deWaaard JR. Biological identifications through DNA barcodes. Proceedings of the Royal


31. Singh HK, Parveen I, Raghuvanshi S, Babbar SB. The loci recommended as universal barcodes for plants on the basis of floristic studies may not work with congeneric species as exemplified by DNA barcoding of Dendrobium species. BMC Research Notes. 2012;5:42.


79. Day PD, Berger M, Hill L, Fay MF, Leitch AR, Leitch IJ, Kelly LJ. Evolutionary relationships in the medicinally important genus Fritillaria L. (Liliaceae). Molecular


