PLANT SCIENCE TODAY
Vol x(x): xx–xx
https://doi.org/10.14719/pst.6418

eISSN 2348-1900

REVIEW ARTICLE

Bioinformatics and computational tools for post-sequencing data analysis in DNA barcoding studies - A review

Adot Vivek[1] & Vadakkemuriyil Divya Nair[2*]

[1]Department of Plant Sciences, Central University of Kerala, Tejaswini Hills, Kasargod 671 320, Kerala, India
[2]Department of Plant Sciences, Central University of Himachal Pradesh, Shahpur Campus, Kangra 176 206, Himachal Pradesh, India

*Correspondence email - divyanair013@hpcu.ac.in

Received: 30 November 2024; Accepted: 07 April 2025; Available online: Version 1.0: 25 June 2025

Cite this article: Adot V, Vadakkemuriyil DN. Bioinformatics and computational tools for post-sequencing data analysis in DNA barcoding studies - A review. Plant Science Today (Early Access). https:/doi.org/10.14719/pst.6418

Abstract

DNA barcoding is a significant and valuable method for identifying species and it is one of the key fields of biodiversity and evolutionary research. It changed biodiversity studies with computational techniques and next-generation sequencing. Post-sequencing data analysis is an essential stage encompassing many critical procedures to ensure precise identification of organisms and their classification. Processing generated mass-sequenced information is a significant problem in any barcoding studies. Various analysis methods are employed for inferring organismal taxonomy, such as tree-based, similarity-based, composition-based and hybrid methods. This study will review the diversity of these computational methods for post-sequence data analysis in DNA barcoding studies. Tree-based techniques (e.g.: MrBayes and RAxML) illustrate evolutionary relationships among species, similarity-based techniques (e.g.: SOrt-ITEMS and BLAST) assist in identifying species by sequence similarities and composition-based techniques (e.g.: Phymm and NBC) categorize species according to their nucleotide composition. These methods are combined in hybrid approaches (e.g.: PhyScimm and RITA) to provide an in-depth investigation. Computational tools for post-sequence analysis use graphical user, command line or web-based interfaces with supervised, unsupervised, or semi-supervised machine learning approaches. Operating systems such as Linux, UNIX, Windows and macOS are used to analyze DNA barcoding data, while Java, R, Python, C/C++ and Perl are the most widely used programming languages. This review emphasizes how crucial it is to incorporate such bioinformatics and computational techniques to improve the robustness and consistency of DNA barcoding studies and provide an adequate set of tools for advanced biodiversity research.

Keywords: BLAST; DNA barcoding; PhyloPythiaS; post-sequencing; SOrt-ITEMS; SPHINX

Introduction

DNA barcoding is one of the critical areas in biodiversity and evolutionary research and is an emerging and effective tool for species identification (1, 2). This procedure makes it possible to identify and distinguish between different species by using specific DNA segments known as barcode regions. These areas contain distinct genetic information that makes them ideal for species differentiation (3). Undoubtedly, DNA barcoding has changed biodiversity studies with computational techniques and next-generation sequencing (NGS) (4). The significance of DNA barcoding in ecological and evolutionary research has risen because of recent advancements in technology and cost reduction (5). Sophisticated bioinformatics and computational techniques used by the sequenced data from DNA barcoding studies enable extracting relevant information on genetics, such as identifying species, phylogenetic analyses, or habitat monitoring (6). However, these tools allow the quick and easy processing of huge sequenced data (4). With novel data processing methods and a suitable NGS platform, a large-scale DNA barcoding study will undoubtedly contribute to creating a reference library of DNA barcodes from every species, preventing misidentification and misinterpretation (7). Advanced bioinformatics and computational tools are essential for accurate and practical analysis of given data because NGS technologies have made it possible to generate large-sequenced data quickly and efficiently (4).

Various analysis methods are employed for efficient data analysis, such as tree-based, similarity-based, composition-based and hybrid methods (8). A commonly used computational technique is BLAST (similarity-based), which facilitates the comparison of DNA sequences and aids in the identification of comparable sequences from the reference database (9). Software MEGA (tree-based) allows post-sequence analysis, which uses phylogenetic methods such as maximum likelihood (ML), neighbour-joining (NJ), Bayesian inference (BI) (in MEGA12) and maximum parsimony (MP) (10). Similarly, a wide range of computational tools is available for efficient post-evaluation of sequenced data in composition-based methods (e.g., PhyloPythia) and hybrid approaches (e.g., SPHINX). Data processing efficiency and quality in DNA barcoding studies might significantly improve by developing specific algorithms and pipelines suited to experimental settings (11). This study aimed to discuss the significant computational and bioinformatic tools used in post-sequencing data analysis during DNA barcoding studies.

DNA barcoding

DNA barcoding is one of the effective taxonomic approaches for identifying and discovering species (3). It identifies taxa using one or more short, standard DNA segments (7). DNA barcoding is an economically viable, time-saving, objective approach and a potent tool for species identification when cryptic species and phenotypic plasticity are problematic and morphological keys are not accessible (12). The concept of a barcoding gap, which describes the variation in mean intra- and interspecific genetic distances, is the backbone of DNA barcoding such that the reliability of species identification increases with the length of the barcoding gap (13). DNA barcoding differs from other molecular approaches because it uses standard markers viz., RbcL, rpoB, COI, MatK, ITS, PsbA-trnH, rpoC1, psbK-psbI and atpF-atpH in plants, COI and COII in animals, 16S rRNA in bacteria and archaea and ITS in fungi (7, 14). The standard marker gene must meet parameters like conserved flanking regions to construct universal PCR primers, distinct barcoding gaps and short sequence lengths (3). The steps followed during DNA barcoding are as follows: a sample is taken from the field, the DNA is extracted, amplification of the barcoding gene with a universal primer, the amplified DNA molecule is sequenced using Sanger sequencing or high-throughput sequencing to ascertain its diversity and data analysis software is used after the data is sequenced (15). Fig. 1 represents the sequential steps involved in the DNA barcoding process.

Post-sequencing data analysis in DNA barcoding

In DNA barcoding, post-sequencing data analysis is an essential stage encompassing many critical procedures to guarantee precise identification of organisms and their classification (16). The processing of generated mass-sequenced information is a significant problem in many barcoding studies (17) because high throughput sequencing creates so much data that post-processing, analysis and interpretation require quick and efficient computing and bioinformatics tools. Following the sequencing of DNA barcode regions, a thorough post-sequencing data analysis must be performed to provide highly accurate and reliable species identifications. It involves several steps like quality control, taxonomy assignment, sequence alignment, error correction and more, which are essential for precise data analysis. Additionally, comparing new sequences with existing reference databases can provide valuable insights into the diversity and distribution of species (18, 19). Various analysis methods are employed for inferring organismal taxonomy, such as tree-based, similarity-based, composition-based and hybrid methods (8). The required phylogenetic and machine-learning assessments, information from reference databases and possible taxonomic range can vary amongst the abovementioned approaches (8). There are various computational and bioinformatic tools/software in each of these approaches, which are briefly described below and in Table 1, the computational features of these tools are given in Fig. 2, which presents the significant importance of computational and bioinformatic tools in post-sequence data analysis during DNA barcoding studies.

Tree-based approaches

In the tree-based approach, a querys' taxonomy of an operational taxonomic unit (OTU) is determined by positioning the OTU within a preformed reference phylogenetic tree (8). This approach, which allocates query sequences to species based on their cluster membership in a barcode tree, is frequently used to determine phylogenetic relationships. Hierarchical clustering techniques like NJ, ML, MP and BI are used to construct phylogenetic trees and combining these algorithms improves classification accuracy compared to single-method approaches by leveraging their strengths and mitigating individual biases (49). The combination of multiple algorithms ensures robust, error-checked results and improved resolution of phylogenetic trees by cross-validating results (50). The major tree-based programs were covered briefly below.

MEGA

MEGA (latest version MEGA12) is a computer program created to compare homologous genome sequences from multi-gene families or various species, with a focus on identifying variations in DNA and protein evolution as well as relationships in evolution (10, 20). MEGA offers several useful services for assembling sets of sequencing data from files or web-based sources and it has resources for displaying the outcomes visually, such as evolutionary distance matrices and interactive phylogenetic trees (20). A wide range of applications are available in MEGA to do alignment of sequences, to determine phylogenetic trees, to calculate time-trees, to measure diversities and genetic distances and to check selection (51, 52). Three distinct parts make up MEGA: a processing engine, a report engine and a graphical user interface and it is compatible with several phylogenetic techniques, including MP, BI (in MEGA12), NJ and ML (10, 20). Furthermore, MEGA calculates statistical data analysis like codon frequencies, transition/transversion biases and the frequency of variable sites in certain portions of nucleotide and amino acid strands (53).

MRBAYES

The application MRBAYES does Bayesian inference of phylogeny utilizing a variation of Markov chain Monte Carlo (MCMC) to estimate posterior probability of trees. It employs Metropolis coupled Markov chain Monte Carlo (variation of MCMC) ((MC)^{3}for short) in addition to the traditional MCMC technique and the programme reads a typical NEXUS-formatted aligned matrix of amino acid or DNA sequences (54). The users can modify the assumptions of the substitution model, the prior and the specifics of the (MC)^3 analysis any time during the process and also the users has the ability to add, remove and recover characters and taxa in the analysis (54). The updated versions include convergence diagnostics and let users perform various studies simultaneously while keeping track of the convergence process. Additionally, it has substantially quicker likelihood calculations because it provides support for the BEAGLE library and streaming single-instruction-multiple-data extensions (SSE), which enable graphics processing units (GPUs) on suitable hardware to do likelihood calculations (21, 55).

RAxML

RAxML is a command line tool that offers extra features like phylogenetic placement of a short sequence of environmental reads (56). RAxML provides various methods for post-analyzing the trees using consensus tree algorithms (22). It does post-

Table 1. Features of standard computational and bioinformatic tools which are used in post-sequencing data analysis in DNA barcoding studies

Name of the software	Software type	Operating system	Programming language	Input data	Algorithm	Interphase	Machine learning	Link	Reference
MEGA	Package	Windows Linux macOS	Java	Contigs	Neighbour-joining Minimum evolution method UPGMA Maximum parsimony	Graphical user interface Command line interface	-	https://www.megasoftware.net/	(20)
MrBayes	Package	Macintos Windows UNIX	-	-	Markov chain Monte Carlo model	Command line interface	-	http://www.mrbayes.net/	(21)
RAxML	Package	Linux macOS	C++	Contigs	Maximum likelihood	Command line interface	-	http://https//github.com/stamatak/standard-RAxML	(22)
IQTREE	Package	Windows Linux	Python C++	Reads or contigs	NNI algorithm	Command line interface	Supervised	http://www.cibiv.at/software/iqtree	(23)
BEAST	Package	Linux	Java	Reads or contigs	Metropolis-Hastings MCMC	Command line interface	-	http://beast-mcmc.googlecode.com/	(24)
MLTreeMap	Pipeline	Linux MacOS Windows	Perl	Reads or contigs	Hidden Markov models Maximum-likelihood BLAST	Command line interface, Web-based interface	Supervised	http://mltreemap.org/	(25, 26)
pplacer	Package	Linux	-	Reads	Maximum-likelihood, Bayesian	Command line interface	-	https://github.com/matsen/pplacer/releases	(27)
SOrt-ITEMS	Pipeline	Linux	-	Reads	Lowest common ancestor	Command line interface	-	http://meta-genomics.atc.tcs.com/binning/SOrt-ITEMS	(25, 28)
BLAST	Standalone	Windows Linux, UNIX Mac OS	C/C++	Reads or contigs	-	Web-based interface	Unsupervised	http://www.ncbi.nlm.nih.gov/BLAST/	(29)
MARTA	Pipeline	Linux	Java	-	Lowest common ancestor	Command line interface	-	http://bergelson.uchicago.edu/software/marta	(25, 30)
PhyloPythiaS	Pipeline	Linux	PHP and Java	Reads or contigs	Support Vector Machines (SVMs) approaches	Command line interface, Web-based interface	Supervised	http://binning.bioinf.mpi-inf.mpg.de/	(25, 31)
TACOA	Standalone	Linux, MacOS Windows	-	Reads or contigs	k-means/k-nearest neighbor	Command line interface	Supervised	http://www.cebitec.uni-bielefeld.de/brf/tacoa/tacoa.html	(25, 32)
RDP Classifier	Package	Linux, MacOS Windows	Java	Read	k-means/k-nearest-neighbour Naive Bayes classifier	Web-based interface Command line interface	Supervised	https://github.com/rdpstaff/classifier	(25, 33)
Phymm	Package	Linux	Perl	Reads	Interpolated Markov Models	Command line interface	-	http://www.cbcb.umd.edu/software/phymm/	(34)

NBC	-	Linux macOS	-	Reads	Bayess' theorem/naive Bayes classifier	Command line interface, Web-based interface	Supervised	http://nbc.ece.drexel.edu/	(25, 35)
RAIphy	Package	Linux MacOS Windows	C++	Reads or contigs	Relative Abundance Index	Graphical user interface	Semi-supervised	http://bioinfo.unl.edu/raiphy.php	(25, 36)
CARMA	Pipeline	Linux	C/C++	Reads	Hidden Markov models	Command line interface, Web-based interface	-	http://www.cebitec.uni-bielefeld.de/brf/	(25, 37)
MEGAN	Pipeline	Linux/ UNIX MacOS Windows	Java	Rawreads or contigs	Lowest common ancestor	Graphical user interface	-	http://www-ab.informatik.uni-tuebingen.de/software/megan	(25, 38)
Scimm	Pipeline	Linux	Python	Rawreads or contigs	Interpolated Markov models	Command line interface	Unsupervised	http://www.cbcb.umd.edu/software/scimm	(25, 39)
Treephyler	-	Linux/ UNIX MacOS Windows	Perl	Reads	Hidden Markov model	Command line interface	-	http://www.gobics.de/fabian/treephyler.php	(25, 40)
DiScRIBinATE	Pipeline	Linux	Python	Raw reads or contigs	-	Command line interface	Supervised	Not available	(41, 42)
INDUS	Package	Linux	-	Reads	k-means	Command line interface	-	http://meta-genomics.atc.tcs.com/INDUS/	(42, 43)
PhymmBL	Package	Linux	Perl	Reads	Interpolated Markov models	Command line interface	Supervised	http://www.cbcb.umd.edu/software/phymmbl/	(25, 34)
SPHINX	-	-	-	Reads	k-means/kNN	Web-based interface	Supervised	http://meta-genomics.atc.tcs.com/SPHINX/	(25, 44)
NB-BL	-	Linux MacOS Windows	Java	Reads or contigs	Naive Bayes classifier	Command line interface	-	https://projects.cs.dal.ca/Software/FCP	(45)
PhyScimm	Pipeline	Linux	Python	Reads or contigs	Interpolated Markov models k-means	-	Supervised and unsupervised	http://www.cbcb.umd.edu/software/scimm	(46)
RITA	Standalone	Linux MacOS Windows	Python	Reads	NBC BLAST	Web-based interface	-	http://kiwi.cs.dal.ca/Software/RITA	(47)
TWARIT	-	Linux MacOS Windows	Python Java	Reads	Hit-pair based assignment Signature sorting-based assignment	-	Supervised	http://meta-genomics.atc.tcs.com/Twarit/	(48)

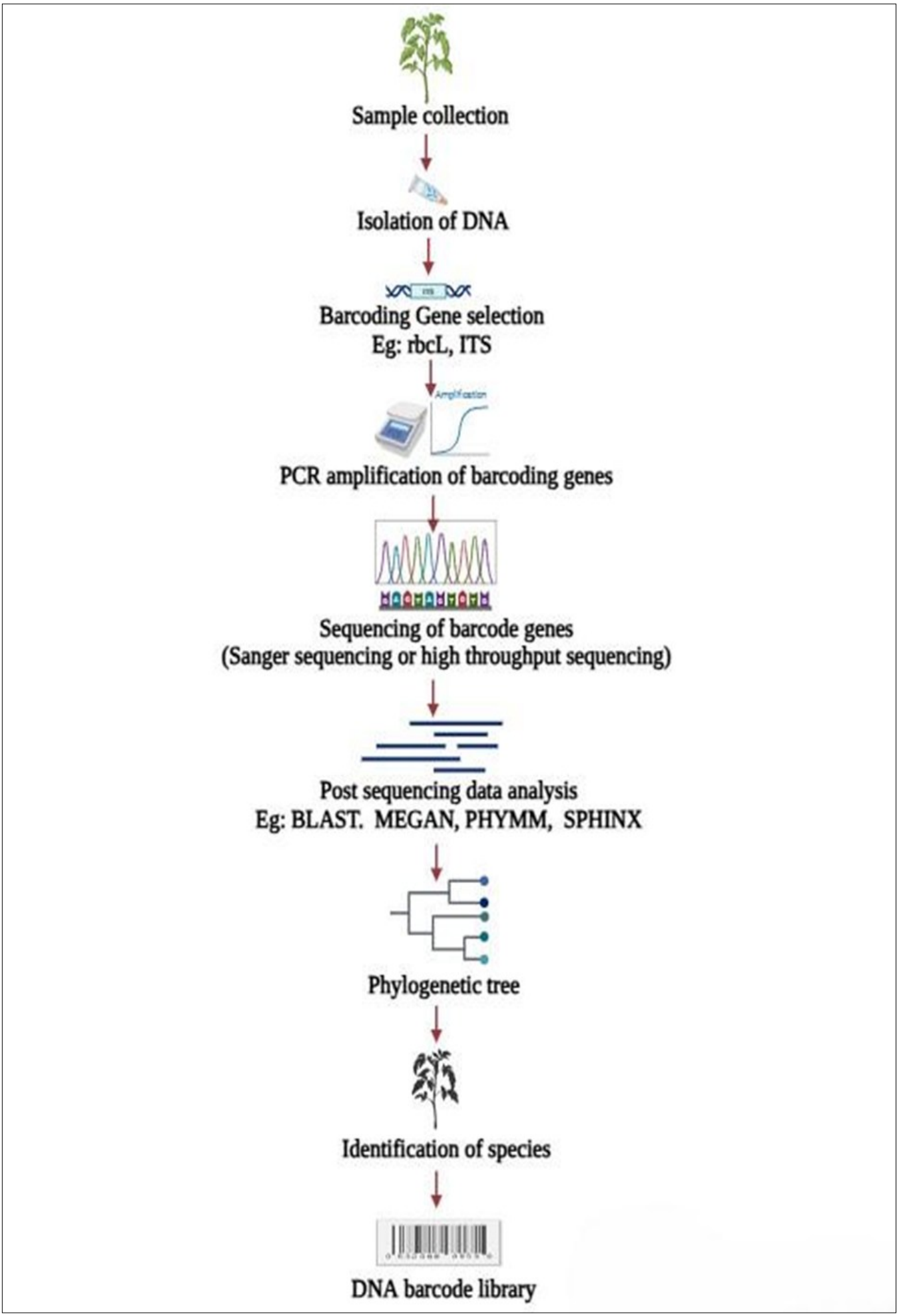

Fig. 1. Sequential steps involved in DNA barcoding studies.

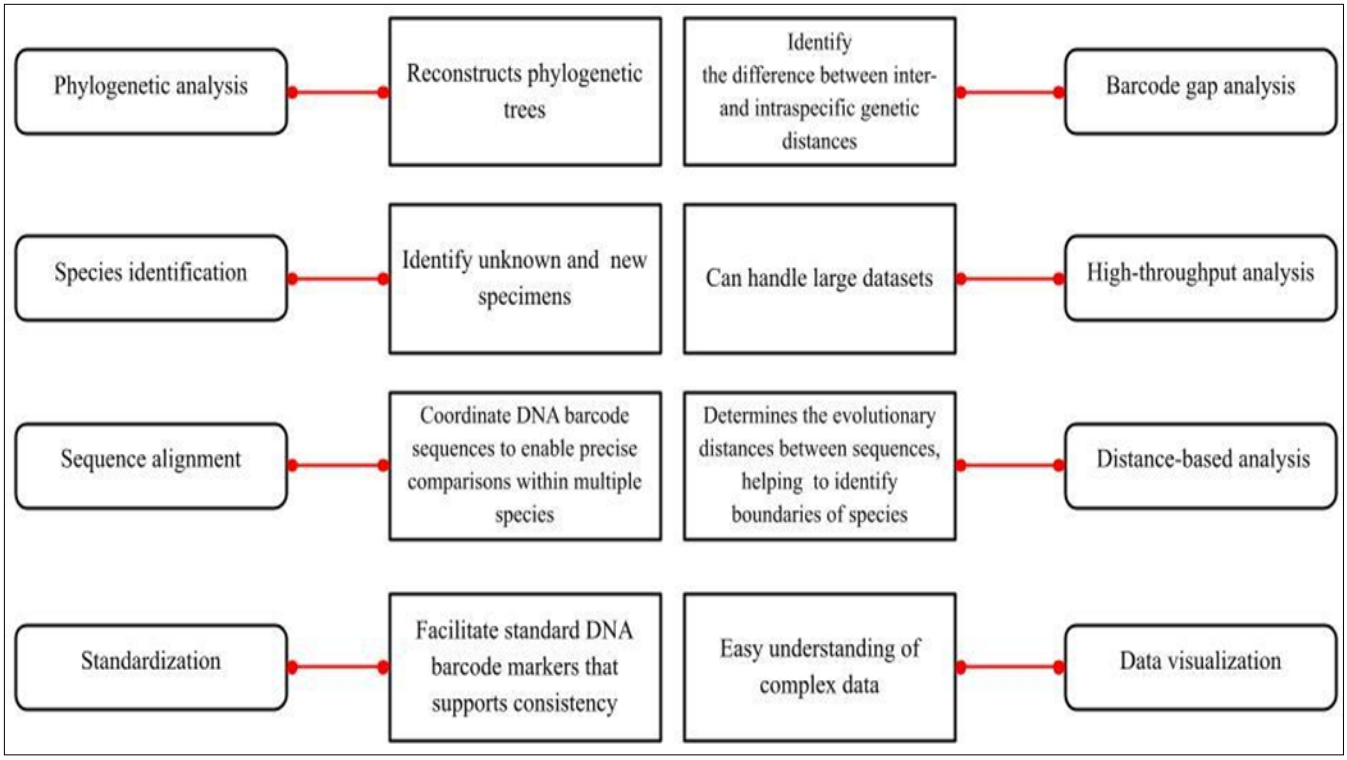

Fig. 2. Application of computational and bioinformatic tools in post sequence data analysis of DNA barcoding studies.

analysis under ML and provides techniques for phylogeny inference using binary, multi-state, RNA secondary structure, protein and DNA data. RAxML also offers a variety of methods, such as the bootstrap operation and statistical tests, like the approximate likelihood-ratio test, for determining support values on phylogenies (57-60).

IQ-TREE

IQ-TREE is software for reconstructing ML trees that saves time and searches well. It performs superior to RAxML in terms of ML search and enhances accessible ML programs (61). The three crucial processes in phylogenetic analysis-quick model selection using Model Finder, a successful tree likely search algorithm and a unique ultrafast bootstrap approximation-are made better by the clever integration of novel phylogenetic methodologies, which leads to IQ-TREEs' excellent performance (62).

BEAST

It is a quick and adaptable software framework for the Bayesian examination of DNA sequences connected by a tree of life (24) and Metropolis-Hastings MCMC is the core algorithm (63). It was the first program that supported flexible molecular clock models and allowed for the deduction of the actual phylogenetic tree (64). BEAST stands out because of its emphasis on calibrated phylogenies and genealogies, that is, rooted trees integrated on a time scale. This is made possible by precisely modelling the rate of molecular evolution on each branch of the tree (24). The component-based nature of BEASTs' model definition makes it challenging to summarize the vast array of potential evolutionary models (24).

Similarity-based approaches

It needs raw reference sequences accessible in public nucleotide databases such as NCBI, GenBank, EMBL and DDBJ and they include taxonomic details (8, 65). Here, the algorithm arranges and classifies the raw reads close to neighbours in the phylogeny and hence relieson closely comparable genomes being available in the database. Such approaches fail to detect homologs for reads from new species because most existing databases are significantly biased in depicting actual diversity (65). This technique can function even with read lengths as small as 80 base pairs (66). The major similarity-based programs were covered briefly below.

MLTreeMap

The MLTreeMap aims to provide insights into the phylogenetic structure and functional characteristics of lifeforms based on ML and manually selected protein-coding marker gene sets as well as 16S and 18S rRNA information, where the query sequences are placed within pre-existing phylogenies (67). To decrease false positive findings, deep paralogs of these genes should be further inspected and removed if found after the BLAST search for the marker genes from the query (26). Subsequently, the Genewise software extracts marker genes based on Hidden Markov models (HMMs), aligns them with referencing proteins (using hmmalign and Gblocks), concatenates them if multiple marker genes are on a fragment and performs mild gap removal. Further, the sequences are processed with RAxML for phylogenetic classification using ML (26).

pplacer

pplacer is intended to carry out phylogenetic placements and sequence analysis, which makes it easier to place reads accurately using ML and posterior probability (27, 68, 69). It aims to be user-friendly for single-runs and pipeline applications and may be used whenever a reference alignment and phylogenetic tree are available. Its robust and fully integrated visualization capabilities may be used to view both the placements and the accompanying uncertainty on a single tree through branch thickness and colour (27). The method can put 20000 short reads on a reference tree of 1000 taxa per hour per processor, is simple to execute in parallel and has roughly linear time and memory complexity in the number of reference taxa. It might inform the user about positional uncertainty for

query sequences by calculating the estimated distance between locations of placements. This is crucial for estimating the uncertainty when using a well-sampled reference tree (27, 70).

SOrt-ITEMS

SOrt-ITEMS is an upgraded similarity-based binning technique. The process initially determines a suitable taxonomic level using alignment characteristics other than the bit score where the read can be assigned, then employs an orthology-based methodology for the final assignment (71). The orthology technique aims to find hits that exhibit a remarkable orthology that is reciprocal in similarity with the query read sequence. The final read assignment is then based on the hits found to be orthologs of the query sequence. First, reads with negligible alignments are found and classified as "Unassigned". Secondly, if they have significant alignment parameters, it determines a taxonomic rank at which the reads can be assigned (28). Eliminating irrelevant hits based on the bit-score cut-off and determining the LCA are the sole processes that SOrt-ITEMS and MEGAN share. Compared to MEGAN, assigning a read takes almost twice as long since it completes many more steps beyond these common steps. SOrt-ITEMSs' remarkably low false-positive rate justifies this extra-time investment (28, 72).

BLAST

BLAST locates areas of local similarity between sequences using a query sequence and a huge database (NCBI, National Center for Biotechnology Information). This software compares sequences of nucleotides or proteins to the constantly expanding sequence databases and estimates the statistical significance of hits with NCBIs' sequence search engine (9, 15). It lists BLAST search types, including translated, nucleotide, protein and genome (73). It looks for short similarities between two sequences and tries to initiate alignments from these hot spots. Besides facilitating alignments, BLAST offers statistical data that aids in interpreting the biological importance of the alignment; this is known as the false positive rate or expected value (74). The BLAST server provides a Request Identifier (RID) once the query is given there, either as a sequence identification or in FASTA format. After an RID is given, the query and results are kept in a structured manner for a maximum of twenty-four hours. The query is identified by the RID, which enables many forms to view the results (ASN.1, BLAST report and XML) (73).

MARTA

MARTA was developed to use DNA sequence data to describe the taxonomic diversity of communities. This program coordinates BLAST with the taxonomy database to ascertain if taxonomic consensus exists among the top hits that BLAST returns. It uses resources of NCBI for taxonomic classification from the sequenced data. It necessitates a reduced consensus requirement (66 %) at the genus and species level and a strict percentage consensus (100 %) among ties at the six higher taxonomic ranks (domain to family), also it can be modified at the command line easily. MARTA allows either the best-score technique (above) or the slip-score (percentile-based) strategy. And it also has a revote option that enables users to consider extra voting strategies during BLAST which avoids the bottleneck experienced (30).

Composition-based approaches

This method uses pattern recognition of the k-mer length word composition to give a sequence to a taxonomic unit. Also, before carrying out a querys' taxonomic assignment, the algorithms must learn the word composition of reference sequences (8). This program is quicker and uses less computational power, which compares reads to sequences or models already available in reference databases using compositional features such as oligonucleotide usage patterns, codon use and GC percentage (71). It creates models from the genomes of the reference organisms, then groups the input sequence reads according to the model that best matches the read (42). The degree of compositional similarity, expressed either in absolute or relative terms, determines the final taxonomic assignments and these approaches also need sufficiently long query sequences. Composition-based approaches can vary depending on how they reflect, measure and differentiate compositional features (71). Most approaches begin with a training phase wherein "genome-specific" reference models or classifiers are constructed using one or more compositional characteristics in known genomes (like Support Vector Machines, Naive-Bayesian approach, kernelized -Nearest Neighbor (k-NN) approach, Relative Abundance Index and Interpolated Markov Models) (71). The major composition-based programs were covered briefly below.

PhyloPythiaS

The successor of PhyloPythia, PhyloPythiaS, is a quick and precise classifier based on sequence composition that uses the hierarchical links between clades (31, 75). Utilizing Support Vector Machines (SVMs) techniques, Phylophithia creates genome or clade-specific classifiers that freeze and describe oligonucleotide using patterns found in recognized taxonomic clades (19). High prediction accuracy is demonstrated by PhyloPythiaS, which enables quick analysis of datasets containing hundreds of megabases or gigabases (76, 77). It performs well in prediction and was tested on simulated and actual data sets. Its processing times are significantly less than those of MEGAN and PhymmBL. There are two modes of application for PhyloPythiaS, such as general and sample-specific (31). The expert-trained PhyloPythiaS package is one of the best methods for recovering species bins from a single sample. Machine learning relies on matching the training and test data sets to achieve high classification accuracy. A human expert uses marker genes and contig coverage information to identify the training sequences from the sample and then determines which taxa to include in the composition-based taxonomic model based on the availability of data (78). The successor to PhyloPythiaS is called PhyloPythiaS+. In addition to carrying out the tasks that the human expert had previously completed, it incorporates a novel k-mer counting method that amplifies k-mer counting by 100 (78).

TACOA

For input query DNA fragments, this composition-based technique can yield a significant taxonomic level and taxonomic group and the modified kernel-NN (nearest neighbour) theory is used to create this classifier (79, 80). The k-NN technique in TACOA makes it possible to create an accurate multi-class classifier (19, 32). Until the rank class, TACOA can

accurately categorize genomic segments between 800 bp and 1000 bp in length. The taxonomic origin of genomic segments as short as 800 bp may be reliably predicted by it. Additionally, it can yield accurate solutions when a fragments' taxonomic origin is not recorded in the reference set, categorizing it as unknown (32). Additionally, adding recently sequenced genomes to the reference set that the proposed classifier utilizes is simple. Researchers may quickly integrate sample-specific models from certain species into the TACOA framework, just like with PhyloPythia (32).

RDP classifier

This method is composition-based (8) and it is dependent upon the naive Bayes classifier (NBC) that uses the 8-mer decomposition of a sequence (81). It is a standard sequence classifier that exhibits low error and excellent sensitivity at the family and order levels (82). The RDP Classifier might be used to analyze libraries of thousands of sequences and individual rRNA sequences. It can quickly offer summary information and taxonomic placement, including how many input sequences belong to each taxon (83). It has been established that reference libraries for taxonomy categorization using RDP classifiers utilizing plant rbcL sequences have been developed (84).

Phymm

The rank-specific Phymm characterizes variable-length oligonucleotides (where n-mers up to n=10) using interpolated Markov models (IMMs) (19). Using details from many oligonucleotides and integrating the information is one benefit of IMMs over other approaches, especially those that rely on oligonucleotide counts. Therefore, Phymm may employ both 5 -mers and 6-mers for categorization at the class/phylum levels rather than having to choose between them (34, 85). The software creates an IMM within a genome for every chromosome and plasmid (45). Any read can be classified using Phymm and its genus-level accuracy was 32.8 % for 100 bp reads (compared to 6 % for CARMA) and 71.1 % for 1000 bp reads (compared to 7.1 % for PhyloPythia) (34). Phymms' log-likelihood output, ease of use and speed of training and testing make them attractive, simple but sophisticated designs for extensive classification and comparison (86).

NBC

Applying Bayes' theorem and supposing that every feature in the classification is independent of every other feature is the base of a naive Bayes classifier (NBC) (87). Using Naive-Bayesian techniques, NBC classifiers create genome or clade-specific classifiers that freeze and describe oligonucleotide use patterns found in known taxonomic clades (19). In contrast to other tools, NBC classifies every read, is simple to use, completes a dataset in a reasonable amount of time and produces competitive results. Moreover, NBC can classify next-generation sequencing reads according to their taxonomic classification and identify significant genera populations that other classifiers might ignore (35). NBC can handle data anywhere in the genome, making it competitive with different classifiers. It can determine the species by comparing the type of repeat in the query to the databases' repetition (88). Usually, it employs motifs that are longer than 12-mers. NBC uses the likelihood that a sequenced read originates from a certain reference genome to determine scores. Every read has been assigned to the genome with the highest score (89).

RAIphy

RAIphy, a semi-supervised classification method, uses the Relative Abundance Index (RAI) to show patterns of over- or under-abundance of k-mers in sequences from different known taxonomic clades. After that, this index is employed to link a specific taxon to a query sequence (19). The taxonomic designations of genomic fragments of uncertain origin are determined using membership scores that assess the genomic fragments' fitness to each taxon-specific index pattern (90). Compared to similarity-search-based applications, RAIphy is quick and easy to use as a standalone desktop program. The approach exhibits accurate performance for longer environmental contigs while maintaining competitive binning accuracies for read lengths (100 to 1000 bp) (36). Since RAIphy uses models that have been estimated using genomes that are now accessible in the RefSeq database, it falls under the category of semi-supervised methods. At its runtime, RAIphy consumes considerable memory (304 MB) when training at the species level and 47 MB at the genus level (36).

CARMA

CARMA predicts the taxonomic origins of DNA fragments. Firstly, Pfam profile hidden Markov models identify protein family and domain fragments in a samples' unassembled reads. The second step is reconstructing a phylogenetic tree for every Pfam family that matches (91, 92). The readings are then categorized into a taxonomically higher-order stage based on their evolutionary links to sequences in the database that are known to have taxonomic origins (93). Because CARMA uses complete protein/nucleotide domains and tries to categorize any given genome fragment, it may be used for binning and is commonly used for phylotyping. Even for short sequences (80–400 bp), these techniques have been demonstrated to be accurate despite their high computing expense (36). CARMA uses a reciprocal BLAST search to further filter the findings rather than utilizing all hits in a particular region to create the LCA of a single read. CARMA refines the taxonomic categorization to a lower level in the taxonomic tree using the bit scores of the reciprocal search. Although this reciprocal search produces a thorough categorization, it is computationally expensive (94).

MEGAN

MEGAN uses BLAST to map query sequences to the NCBI-NR database, then classifies them taxonomically based on the top databases' lowest common ancestor (LCA) (36, 93). Each read is assigned to the LCA of the set of taxa that the program hits throughout the comparison using a simple algorithm. As a result, high-order taxa at the roots of the NCBI tree are assigned widely conserved sequences. In contrast, taxa close to the leaves are assigned species-specific sequences, enabling easy usage and the study of big data sets. Fragments as small as 35 bp are appropriately assigned by MEGAN (version 1) analysis, which also enables the integration of several taxonomic systems and offers filters to subsequently modify the stringency level to a suitable level (38). Version 2 of the application made it possible to compare the taxonomy of several datasets, while version 3 sought to offer a functional analysis based on the GO ontology as well (95) and two new functional analysis techniques (KEGG and SEED classification)

have taken the role of the GO analyzer in version 4 (96, 97). Additionally, versions five and six introduced better speed and a better user interface (39, 98). To use MEGAN, compare reads against BLASTX against NCBI-NR, import reads and BLAST files, calculate taxonomic and functional classifications and a single comparative document can simultaneously open multiple datasets, offering comparative views of different classifications (99).

Scimm

Scimm is an IMM-based unsupervised sequence clustering technique (46). Scimm employs the same general technique as CEM (Classification Expectation Maximization), in which IMMs are cluster models and data points are read sequences (46). For IMM modelling, Scimm employs entirely autonomously determined bins, which improves precision and accuracy. Before using IMM for the data, initial bins must be formed. This may be accomplished with a different binning technique (likelyBin and compostBin initialized) or k-means clustering, which requires a predicted number of clusters as an input (39).

Treephyler

Treephyler is involved in a quick and precise taxonomic profile of big data sets (40). During analysis, Treephyler may use several processing cores to make assignments at taxonomic ranks higher than the genus level. It can also provide measures of assignment confidence (25).

DiScRIBinATE

DiScRIBinATE is the successor to SOrt-ITEMS (42). It substitutes the SOrt-ITEMS' orthology approach with a faster alignment-free approach. By using a new reclassifying technique, DiScRIBinATE lowers the overall misclassification rate to about 3 to 7 %. Compared to SOrt-ITEMS, this misclassification rate is 1.5 to 3 times lower and compared to MEGAN, it is 3-30 times lower. The suggested algorithms' enormous applicability with high specificity and accuracy is demonstrated by the notable binning time reduction combined with excellent assignment accuracy (41). DiScRIBinATE utilizes the procedures used in the initial stage of SOrt ITEMS to keep assignment accuracy. In order to guarantee assignment specificity, it uses a faster alternative approach based on the ratio of bit-score to distance information derived from the hits corresponding to a read, avoiding the tedious orthology stage of SOrt-ITEMS (41). Additional techniques used by DiScRIBinATE, such as alignment parameter thresholds and a reciprocal BLAST search stage, have been demonstrated to increase the precision of taxonomic designations in specific situations (42).

INDUS

Instead of assuming a "one genome-one composition model", the INDUS algorithm demonstrates every genome as several vectors. The tetranucleotide frequency pattern of distinct (non-overlapping) 1kb segments produced by slicing the corresponding genome is shown by each vector. INDUS determines the proper taxonomic level of assignment for the query by utilizing the compositional distance between the identified set of reference segments and the query read. The nearest reference segments at or above the designated taxonomic level are assigned to a consensus taxon at the end (43).

Hybrid approaches

This approach utilizes either tree-based and similarity-based methods or composition-based and similarity-based approaches together (8). The hybrid strategy is here to put effort into improving classification or speed. In the case of enhancing classification, for every prediction, scores from the composition and similarity areas might be combined and to improve speed, utilize the composition method to reduce the number of candidate species, allowing the similarity search to be conducted against a smaller subset of the original database (34, 44). The major hybrid programs were covered briefly below.

PhymmBL

It is an approach that combines Phymm and BLAST and performs better than either of the two approaches alone. PhymmBL performs better than both Phymm and BLAST for all read lengths and clade levels, with a 6 % improvement over BLAST alone for the 1000 bp query set for all taxonomic levels. PhymmBL produces extremely reliable and repeatable results, in every instance; the accuracy standard deviation was less than 1 % (34). To get more precision, PhymmBL combines log-scores from Phymm and BLASTN linearly. Either the entire set of PhymmBL predictions or just the subset based on composition and homology can be included (47). It is composed of two components: one is Phymm-dependent composition-directed taxonomic predictions and the second one is BLAST-based homology. These are combined by PhymmBL, which then assigns each input sequence its best estimate of the source organisms' taxonomy. It phylogenetically classifies input sequences as short as 100 bp more accurately and predicts phylum to species for each read (100).

SPHINX

The hybrid binning approach SPHINX, uses both composition and alignment-based binning algorithm concepts to achieve higher binning efficiency and it targets to reduce the total time requirement of alignment-based binning techniques by an order of magnitude (101). It depends on the sequence alignment and generated alignment parameter (43). There are three steps in the sequence binning by SPHINX. In the first stage, small sequence subsets (in the reference database) that are similar in composition to the query are found using the compositional properties of the query sequence. In the second stage, BLAST similarly searches the query sequence against this small subset of database sequences. The last phase involves analyzing the results of these BLAST searches and assigning the query to a taxon or clade that produced significant hits (44). Compared to pure BLAST-based approaches, compositional characteristics are employed to limit the initial space of search for BLAST, resulting in a 4-to 8-fold decrease in total binning time (48).

NB-BL

NB-BL was developed using Naive Bayes and BLAST algorithms as its foundations (45). Both NB-BL classifier and PhymmBL perform almost identically and produce higher average sensitivity when compared to BLASTN (45). Using a vector of nucleotide frequencies, it might be used to detect genomic fragments of varying length (80).

PhyScimm

A combination of the SCIMM and PHYMM (initial partitioning of the sequences) strategies is called PhyScimm. When dealing with very complicated data, such as a combination of several species, it performs poorly and cannot differentiate between species with lesser abundance. Additionally, noise contributed by species with lower abundance impairs the clustering accuracy of high-abundance species (102). Initially, a subset of the sequences was chosen at random to classify and cluster the sequences at a particular taxonomic level. In most cases, Phymm returns too many clusters because of misclassification noise. A helpful recommendation for removing misclassification noise is to retain clusters that comprise more than 20 % of the sequenced bases (where the genome number in the mixture is denoted by k). Phymms' ability to return k clusters depends on the strictness of the filtering, which the user must specify in a new environment. All uncluttered sequences should be moved to a different cluster once clusters have been filtered; otherwise, SCIMM tends to force these sequences into the generally high-quality clusters from Phymm classifications. At the end, IMM clustering was continued as in the SCIMM algorithm (46).

RITA

RITA generates predictions more precisely than a one-step BLASTN search by combining homology-based predictions with the Naive Bayes technique for compositional grouping (103). It uses three BLAST algorithms to reduce and save the computational time spent (47). RITA assigns sequences as short as 50 nt to several classification groups with differing levels of confidence based on the concordance between composition and homology and thus, comparable homology search techniques make it significantly quicker than the PhymmBL (47). RITA can classify sequences more precisely than the best rank-flexible classifier currently in use by applying prior information about taxonomic distributions to improve assignment accuracy in data sets with different levels of taxonomic novelty and it could use paired-end details that might enhance the accuracy of short reads (47). There are three key differences between PhymmBL and RITA. Firstly, RITA employs NBC rather than Phymms' IMM. This improves speed by more than ten times without losing accuracy. Secondly, RITA proposes a methodology that increases the importance of homology-based predictions by first examining the agreement between homology and compositional predictions and then determining if homology results in them significantly preferring one taxonomic label over all others. Thirdly, in order to limit the range of forecasts that RITA can make, a rank-flexible version of the program enables the user to supply a list of predicted taxonomic categories (47).

TWARIT

TWARIT offers expected improvements in binning precision (101). Its binning method combines composition-based signature sorting techniques with short-read alignment to attain high binning rates without reducing binning specificity and accuracy (48). TWARIT is a major step forward in the creation of computationally inexpensive tools (48). In two separate stages, TWARIT assigns input reads phylogenetically. Firstly, using a special hit-pair-based assignment (HPBA) process, query sequences that come from known genomes are found in the first phase. Secondly, a novel signature sorting-based assignment (SSBA) process is used to bin the remaining query sequences (48).

Conclusion

DNA barcoding is a significant and valuable method for identifying species and it is one of the key fields of biodiversity and evolutionary research. It allows the analysis of large amounts of DNA sequence data effectively. The appropriate use of reference databases, visualization tools and advanced algorithms improves the speed and precision of post-sequence analysis. Depending on the study environment, tree-based, similarity-based and composition-based approaches each have unique benefits. Similarity-based techniques assist in identifying species by sequence similarities, composition-based techniques categorize species according to their nucleotide composition and tree-based techniques illustrate evolutionary relationships. These various methods are combined in hybrid approaches to provide an in-depth investigation. These techniques are widely applied to multiple types of biological research, including species identification. Through direct comparisons of DNA, RNA and proteins of various species, similarity-based methods can be used to identify homologous regions, predict the functions of proteins and find genetic elements that are conserved across species. Composition-based methods are valuable in meta-genomic classifications and genome assembly studies. Tree-based methods support species classification, reconstruction of evolutionary relationships and analysis of protein family divergence. Computational tools for post-sequence analysis use graphical user, command line or web-based interfaces with supervised, unsupervised or semi-supervised machine learning.

Operating systems such as Linux, UNIX, Windows and macOS are used to analyze DNA barcoding data. At the same time, Java, R, Python, C/C++ and Perl are the most widely used programming languages for data analysis.. For evolutionary precision, tree-based algorithms (RAxML, IQ-TREE, BEAST) work best if there are reference trees; they are, however, slower and need high-quality data. Similarity-based methods (BLAST, MLTreeMap) are very sensitive for taxonomic classification; however, they struggle with new species and are computationally expensive. Composition-based algorithms (Phymm, NBC, PhyloPythiaS) are fast and flexible; they are less accurate for closely related taxa. For large or complex datasets, hybrid approaches (SPHINX, PhyScimm, RITA) integrate strengths for accuracy and higher efficiency. Both RAxML and IQ-TREE are fast tools for phylogenetics, but IQ-TREE offers better model selection.

MEGA is simple to use but has limitations when handling large datasets. The RDP Classifier and NBC are fast but may lack accuracy for taxonomy classification compared to BLAST, which is slow but sensitive. MEGAN and Scimm maintain a balance between accuracy and speed. The pplacer is a rapid phylogenetic placement program but requires reference trees. Phymm is better for short-read classification, while PhyloPythiaSs' oligomer-based taxonomic assignment

makes it more useful in novel environments. Hybrid strategies often work better in general than single approaches. Combining tree-based, sequence similarity-based and composition-based techniques, hybrid barcoding data analysis improves accuracy, flexibility and robustness by mitigating their weaknesses-particularly for complicated, noisy, or incomplete datasets. They utilize fast initial filters (such as k-mers) (that help to balance speed and sensitivity) before implementing more computationally intensive phylogenetic refinements. They also manage database biases more effectively than single-method technologies. There are still issues with improving database accuracy, managing massive datasets and developing user-friendly interfaces. Future algorithms and interdisciplinary cooperation developments will improve these methods even more, allowing for a more thorough and accurate understanding of biodiversity.

Acknowledgements

DN acknowledge the funding agency DST-SERB, New Delhi, India, for supporting this work (File SRG/2021/00114 dated 30/12/21) as Start-Up Research Grant awarded to the corresponding author. DN wish to thank authorities of the Central University of Himachal Pradesh, Kangra, Himachal Pradesh, India who facilitated the smooth operation of her research activities.

Authors' contributions

VDN was responsible for the conception and design of the study, performed material preparation, data collection and analysis. AV performed material preparation, data collection & analysis, compiled the first draft of the manuscript and both the authors commented on previous versions. Both authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest: The authors state that they have no known competing financial interests or personal relationships that could have influenced the work presented in this study.

Ethical issues: None

References

1. Trivedi S, Aloufi AA, Ansari AA, Ghosh SK. Role of DNA barcoding in marine biodiversity assessment and conservation: an update. Saudi J Biol Sci. 2016;23(2):161–71. https://doi.org/10.1016/j.sjbs.2015.01.001
2. 2. Nair VD, Aseema P, Saini KC. DNA barcoding: an effective molecular tool for species identification, molecular authentication and phylogeny studies in plant science research. Plant Sci Today. 2024;11(4):204–18. https://doi.org/10.14719/pst.3183
3. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc R Soc Lond B Biol Sci. 2003;270(1512):313–21. https://doi.org/10.1098/rspb.2002.2218
4. Abdi G, Singh S, Selvakumar S, Dhar SK, Mudgal G, Swaminathan P, et al. DNA barcoding and its applications. In: Singh V (ed.) Advances in Genomics. Singapore: Springer Nature Singapore; 2024. p. 91–117. https://doi.org/10.1007/978-981-97-3169-5_5
5. Lahaye R, van ber bank M, Bogarin D, Warner J, Pupulin F, Gigot G, et al. DNA barcoding the floras of biodiversity hotspots. Proc Natl Acad Sci. 2008;105(8):2923–28. https://doi.org/10.1073/pnas.0709936105
6. Tan MP, Wong LL, Razali SA, Afiqah-Aleng N, Mohd Nor SA, Sung YY, et al. Applications of next-generation sequencing technologies and computational tools in molecular evolution and aquatic animals conservation studies: a short review. Evol Bioinforma. 2019;15:1176934319892284. https://doi.org/10.1177/1176934319892284
7. Antil S, Abraham JS, Sripoorna S, Maurya S, Dagar J, Makhija S, et al. DNA barcoding, an effective tool for species identification: a review. Mol Biol Rep. 2023;50(1):761–75. https://doi.org/10.1007/s11033-022-08015-7
8. Tanabe AS, Toju H. Two new computational methods for universal DNA barcoding: a benchmark using barcode sequences of bacteria, archaea, animals, fungi and land plants. Fontaneto D (ed) PLoS ONE. 2013;8(10):e76910. https://doi.org/10.1371/journal.pone.0076910
9. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. Blast+: architecture and applications. BMC Bioinformatics. 2009;10(1):421. https://doi.org/10.1186/1471-2105-10-421
10. Kumar S, Stecher G, Suleski M, Sanderford M, Sharma S, Tamura K. Mega12: Molecular Evolutionary genetic analysis version 12 for adaptive and green computing. Mol Biol Evol. 2024;41(12):msae263. https://doi.org/10.1093/molbev/msae263
11. Hakimzadeh A, Abdala Asbun A, Albanese D, Bernard M, Buchner D, Callahan B, et al. A pile of pipelines: an overview of the bioinformatics software for metabarcoding data analyses. Mol Ecol Resour. 2024;24(5):e13847. https://doi.org/10.1111/1755-0998.13847
12. Yao H, Song J, Liu C, Luo K, Han J, Li Y, et al. Use of the ITS2 region as the universal DNA barcode for plants and animals. PLoS ONE. 2010;5(10):e13102. https://doi.org/10.1371/journal.pone.0013102
13. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci. 2012;109(16):6241–46. https://doi.org/10.1073/pnas.1117018109.
14. Comtet T, Sandionigi A, Viard F, Casiraghi M. DNA (meta) barcoding of biological invasions: a powerful tool to elucidate invasion processes and help manage aliens. Biol Invasions. 2015;17(3):905–22. https://doi.org/10.1007/s10530-015-0854-y
15. DeSalle R, Goldstein P. Review and interpretation of trends in DNA barcoding. Front Ecol Evol. 2019;7:302. https://doi.org/10.3389/fevo.2019.00302
16. Letsiou S, Madesis P, Vasdekis E, Montemurro C, Grigoriou ME, Skavdis G, et al. DNA barcoding as a plant identification method. Appl Sci. 2024;14(4):1415. https://doi.org/10.3390/app14041415
17. Thielecke L, Aranyossy T, Dahl A, Tiwari R, Roeder I, Geiger H, et al. Limitations and challenges of genetic barcode quantification. Sci Rep. 2017;7(1):43249. https://doi.org/10.1038/srep43249
18. Thomas T, Gilbert J, Meyer F. Metagenomics - A guide from sampling to data analysis. Microb Inform Exp. 2012;2(1):3. https://doi.org/10.1186/2042-5783-2-3
19. Tripathi LK, Nailwal TK. Metagenomics: applications of functional and structural approaches and meta-omics. In: Recent Advancements in Microbial Diversity. Elsevier; 2020. p. 471–505. https://doi.org/10.1016/B978-0-12-821265-3.00020-7
20. Sohpal VK, Dey A, Singh A. MEGA biocentric software for sequence and phylogenetic analysis:a review. Int J Bioinforma Res Appl. 2010;6(3):230. https://doi.org/10.1504/IJBRA.2010.034072

21. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539–42. https://doi.org/10.1093/sysbio/sys029

22. Stamatakis A. Using RAxML to infer phylogenies. Curr Protoc Bioinforma. 2015;51(1). https://doi.org/10.1002/0471250953.bi0614s51

23. 2Minh BQ, Lanfear R, Ly-Trong N, Trifinopoulos J, Schrempf D, Schmidt HA. IQ-TREE version 2.2.0: Tutorials and Manual Phylogenomic software by maximum likelihood. 2022. http://www.iqtree.org/

24. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Bio. 2007;7(1):214. https://doi.org/10.1186/1471-2148-7-214

25. Bazinet AL, Cummings MP. A comparative evaluation of sequence classification programs. BMC Bioinformatics. 2012;13(1):92. https://doi.org/10.1186/1471-2105-13-92

26. 26. Stark M, Berger SA, Stamatakis A, Von Mering C. MLTreeMap - accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. BMC Genomics. 2010;11(1):461. https://doi.org/10.1186/1471-2164-11-461

27. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics. 2010;11(1):538. https://doi.org/10.1186/1471-2105-11-538

28. Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS. SORT-ITEMS: Sequence orthology-based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics. 2009;25(14):1722–30. https://doi.org/10.1093/bioinformatics/btp317

29. Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. Nucleic Acids Res. 2006;34(Web Server issue):W6–W9. https://doi.org/10.1093/nar/gkl164

30. Horton M, Bodenhausen N, Bergelson J. MARTA: a suite of Java-based tools for assigning taxonomic status to DNA sequences. Bioinformatics. 2010;26(4):568–69. https://doi.org/10.1093/bioinformatics/btp682

31. Patil KR, Roune L, McHardy AC. The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. PLoS ONE. 2012;7(6):e38581. https://doi.org/10.1371/journal.pone.0038581

32. Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. TACOA - Taxonomic classification of environmental genomic fragments using a kernelised nearest neighbour approach. BMC Bioinformatics. 2009;10(1):56. https://doi.org/10.1186/1471-2105-10-56

33. Lan Y, Wang Q, Cole JR, Rosen GL. Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. PLoS ONE. 2012;7(3):e32491. https://doi.org/10.1371/journal.pone.0032491

34. Brady A, Salzberg SL. Phymm and Phymmbl: metagenomic phylogenetic classification with interpolated Markov models. Nat. Methods. 2009;6(9):673–76. https://doi.org/10.1038/nmeth.1358

35. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the Naïve Bayes Classification tool web server for taxonomic classification of metagenomic reads. Bioinformatics. 2011;27(1):127–29. https://doi.org/10.1093/bioinformatics/btq619

36. Nalbantoglu OU, Way SF, Hinrichs SH, Sayood K. RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics. 2011;12(1):41. https://doi.org/10.1186/1471-2105-12-41

37. Gerlach W, Stoye J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Res. 2011;39(14):e91–e91. https://doi.org/10.1093/nar/gkr225

38. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86. https://doi.org/10.1101/gr.5969107

39. Sedlar K, Kupkova K, Provaznik I. Bioinformatics strategies for taxonomy-independent binning and visualization of sequences in shotgun metagenomics. Comput Struct Biotechnol J. 2017;15:48–55. https://doi.org/10.1016/j.csbj.2016.11.005

40. Schreiber F, Gumrich P, Daniel R, Meinicke P. Treephyler: fast taxonomic profiling of metagenomes. Bioinformatics. 2010;26(7):960–61. https://doi.org/10.1093/bioinformatics/btq070

41. Ghosh TS, Haque MM, Mande SS. DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. BMC Bioinformatics. 2010;11(S7):S14. https://doi.org/10.1186/1471-2105-11-S7-S14

42. Peabody MA, Van Rossum T, Lo R, Brinkman FSL. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinformatics. 2015;16(1):362. https://doi.org/10.1186/s12859-015-0788-5

43. Mohammed MH, Ghosh TS, Reddy RM, Reddy CVSK, Singh NK, Mande SS. INDUS - a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. BMC Genomics. 2011;12(S3):S4. https://doi.org/10.1186/1471-2164-12-S3-S4

44. Mohammed MH, Ghosh TS, Singh NK, Mande SS. SPHINX-an algorithm for taxonomic binning of metagenomic sequences. Bioinformatics. 2011;27(1):22–30. https://doi.org/10.1093/bioinformatics/btq608

45. Parks DH, MacDonald NJ, Beiko RG. Classifying short genomic fragments from novel lineages using composition and homology. BMC Bioinformatics. 2011;12(1):328. https://doi.org/10.1186/1471-2105-12-328

46. Kelley DR, Salzberg SL. Clustering metagenomic sequences with interpolated Markov models. BMC Bioinformatics. 2010;11(1):544. https://doi.org/10.1186/1471-2105-11-544

47. MacDonald NJ, Parks DH, Beiko RG. Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Res. 2012;40(14):e111. https://doi.org/10.1093/nar/gks335

48. Reddy RM, Mohammed MH, Mande SS. TWARIT: an extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences. Gene. 2012;505(2):259–65. https://doi.org/10.1016/j.gene.2012.06.014

49. Alam M, Abbas K, Usmani N, Mustafa M, Husain A. A comprehensive review on DNA barcoding for species identification across diverse taxa. Munis Entomol Zool J. 2024;19.

50. Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020;21(7):428–44. https://doi.org/10.1038/s41576-020-0233-0

51. Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2016;33(7):1870–74. https://doi.org/10.1093/molbev/msw054

52. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547–49. https://doi.org/10.1093/molbev/msy096

53. Kumar S, Tamura K, Nei M. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Bioinformatics. 1994;10(2):189–91. https://doi.org/10.1093/bioinformatics/10.2.189

54. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17(8):754–55. https://doi.org/10.1093/bioinformatics/17.8.754

55. Heritage S. MBASR: Workflow-simplified ancestral state reconstruction of discrete traits with MrBayes in the R environment. 2021. https://doi.org/10.1101/2021.01.10.426107

56. Berger SA, Krompass D, Stamatakis A. Performance, accuracy and web server for evolutionary placement of short sequence reads under maximum likelihood. Syst Biol. 2011;60(3):291–302. https://doi.org/10.1093/sysbio/syr010

57. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: a fast, accurate and powerful alternative. Syst Biol. 2006;55(4):539–52. https://doi.org/10.1080/10635150600755453

58. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76. https://doi.org/10.1007/BF01734359

59. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 1985;39(4):783. https://doi.org/10.2307/2408678

60. Stamatakis A. RAxML version 8:a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30 (9):1312–13. https://doi.org/10.1093/bioinformatics/btu033

61. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74. https://doi.org/10.1093/molbev/msu300

62. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE 2:New models and efficient methods for phylogenetic inference in the genomic era. Teeling E (ed.) Mol Biol Evol. 2020;37(5):1530–34. https://doi.org/10.1093/molbev/msaa015

63. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys. 1953;21(6):1087–92. https://doi.org/10.1063/1.1699114

64. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. Penny D (ed.) PLoS Biol. 2006;4(5):e88. https://doi.org/10.1371/journal.pbio.0040088

65. Prabhakara S, Acharya R. A two-way Bayesian mixture model for clustering in metagenomics. In: Loog M, Wessels L, Reinders MJT, De Ridder D (eds.) Pattern Recognition in Bioinformatics. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 25–36. https://doi.org/10.1007/978-3-642-24855-9_3

66. Ander C, Schulz-Trieglaff OB, Stoye J, Cox AJ. metabeetl: high-throughput analysis of heterogeneous microbial populations from shotgun DNA sequences. BMC Bioinformatics. 2013;14 (S5):S2. https://doi.org/10.1186/1471-2105-14-S5-S2

67. Stark M. MLTreeMap - maximum likelihood placement of environmental DNA sequence reads into curated reference phylogenies. PhD [Dissertation]; Zurich: University of Zurich. 2011. https://doi.org/10.5167/UZH-59539

68. Land TA, Fizzano P, Kodner RB. Measuring cluster stability in a large-scale phylogenetic analysis of functional genes in metagenomes using PPLACER. IEEE/ACM Trans Comput Biol Bioinform. 2016;13(2):341–49. https://doi.org/10.1109/TCBB.2015.2446470

69. Wedell E, Cai Y, Warnow T. Scalable and accurate phylogenetic placement using pplacer-XR. In: Martín-Vide C, Vega-Rodríguez MA, Wheeler T (eds.) Algorithms for Computational Biology. Cham: Springer International Publishing; 2021. p. 94–105. https://doi.org/10.1007/978-3-030-74432-8_7

70. Koning E, Phillips M, Warnow T. ppIacerDC: a new scalable phylogenetic placement method. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology and Health Informatics. Gainesville, Florida: ACM; 2021. p. 1–9. https://doi.org/10.1145/3459930.3469516

71. Mande SS, Mohammed MH, Ghosh TS. Classification of metagenomic sequences: methods and challenges. Brief Bioinform. 2012;13(6):669–81. https://doi.org/10.1093/bib/bbs054

72. Wang J, McLenachan PA, Biggs PJ, Winder LH, Schoenfeld BIK, Narayan VV, et al. Environmental bio-monitoring with high-throughput sequencing. Brief Bioinform. 2013;14(5):575–88. https://doi.org/10.1093/bib/bbt032

73. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32:W20–W25. https://doi.org/10.1093/nar/gkh435

74. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2

75. Abdelsalam NA, Elshora H, El-Hadidi M. Interactive web-based services for metagenomic data analysis and comparisons. In: Mitra S (ed.) Metagenomic Data Analysis. New York, NY: Springer US; 2023. p. 133–74. https://doi.org/10.1007/978-1-0716-3072-3_7

76. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods. 2007;4(1):63–72. https://doi.org/10.1038/nmeth976

77. Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T, et al. Taxonomic metagenome sequence assignment with structured output models. Nat Methods. 2011;8(3):191–92. https://doi.org/10.1038/nmeth0311-191

78. Gregor I, Dröge J, Schirmer M, Quince C, McHardy AC. PhyloPythiaS: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. PeerJ. 2016;4:e1603. https://doi.org/10.7717/peerj.1603

79. Hou T, Liu Y, Xue J, Li M, Liu F. Taxonomic classification of DNA fragments of metagenome with a novel model. In:2016 35th Chinese Control Conference (CCC). Chengdu, China:IEEE; 2016. p. 9325–30. https://doi.org/10.1109/ChiCC.2016.7554840

80. Tao H, Yun L, Fu L, Ke W, Jian X. Binning DNA fragment of metagenome using a novel model. In:The 27th Chinese Control and Decision Conference (2015 CCDC). Qingdao, China: IEEE; 2015. p. 4760–65. https://doi.org/10.1109/CCDC.2015.7162767

81. Somervuo P, Koskela S, Pennanen J, Henrik Nilsson R, Ovaskainen O. Unbiased probabilistic taxonomic classification for DNA barcoding. Bioinformatics. 2016;32(19):2920–27. https://doi.org/10.1093/bioinformatics/btw346

82. Richardson RT, Bengtsson-Palme J, Johnson RM. Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Mol Ecol Resour. 2017;17(4):760–69. https://doi.org/10.1111/1755-0998.12628

83. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–67. https://doi.org/10.1128/AEM.00062-07

84. Bell KL, Loeffler VM, Brosi BJ. An rbcL reference library to aid in the identification of plant species mixtures by DNA metabarcoding. Appl Plant Sci. 2017;5(3):1600110. https://doi.org/10.3732/apps.1600110

85. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9 (8):811–14. https://doi.org/10.1038/nmeth.2066

86. Rosen GL, Polikar R, Caseiro DA, Essinger SD, Sokhansanj BA. Discovering the unknown: Improving detection of novel species and genera from short reads. BioMed Res Int. 2011;2011 (1):495849. https://doi.org/10.1155/2011/495849

87. Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B. Metagenome fragment classification using N -mer frequency profiles. Adv Bioinforma. 2008;2008(1):205969. https://doi.org/10.1155/2008/205969

88. Rosen GL, Lim TY. NBC update:The addition of viral and fungal databases to the Naïve Bayes classification tool. BMC Res Notes. 2012;5(1):81. https://doi.org/10.1186/1756-0500-5-81

89. Matsushita N, Seno S, Takenaka Y, Matsuda H. Metagenome fragment classification based on multiple motif-occurrence profiles. PeerJ. 2014;2:e559. https://doi.org/10.7717/peerj.559

90. Karagöz MA, Nalbantoglu OU. Taxonomic classification of metagenomic sequences from Relative Abundance Index profiles using deep learning. Biomed Signal Process Control. 2021;67:102539. https://doi.org/10.1016/j.bspc.2021.102539

91. Gerlach W, Jünemann S, Tille F, Goesmann A, Stoye J. Webcarma: a web application for the functional and taxonomic classification of unassembled metagenomic reads. BMC Bioinformatics. 2009;10(1):430. https://doi.org/10.1186/1471-2105-10-430

92. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, et al. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 2008;36(7):2230–39. https://doi.org/10.1093/nar/gkn038

93. Liu B, Gibbons T, Ghodsi M, Pop M. MetaPhyler: Taxonomic profiling for metagenomic sequences. In:2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Hong Kong, China: IEEE; 2010. p. 95–100. https://doi.org/10.1109/BIBM.2010.5706544

94. Ander C. Bioinformatic methods for the analysis and comparison of metagenomes and metatranscriptomes. PhD [dissertation]. Bielefeld University, Germany; 2014. Available from: https://core.ac.uk/download/pdf/211842477.pdf

95. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: a tool for the unification of biology. Nat Genet. 2000;25(1):25–29. https://doi.org/10.1038/75556

96. Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27

97. Overbeek R. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33(17):5691–5702. https://doi.org/10.1093/nar/gki866

98. Beier S, Tappu R, Huson DH. Functional analysis in metagenomics using MEGAN 6. In: Charles TC, Liles MR, Sessitsch A (eds.) Functional Metagenomics: Tools and Applications. Cham: Springer International Publishing. 2017. p. 65–74. https://doi.org/10.1007/978-3-319-61510-3_4

99. Huson DH, Mitra S. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. In: Anisimova M (ed) Evolutionary Genomics. Totowa, NJ:Humana Press; 2012. p. 415–29. https://doi.org/10.1007/978-1-61779-585-5_17

100. Brady A, Salzberg S. Phymmbl expanded: confidence scores, custom databases, parallelisation and more. Nat Methods. 2011;8(5):367–367. https://doi.org/10.1038/nmeth0511-367

101. Dutta A, Tandon D, Mh M, Bose T, Mande SS. Binpairs: utilization of Illumina paired-end information for improving efficiency of taxonomic binning of metagenomic sequences. Melcher U (ed) PLoS ONE. 2014;9(12):e114814. https://doi.org/10.1371/journal.pone.0114814

102. Gadhe UG. Metagenomic sequence analysis using a hybrid approach. College of engineering, pune-5; 2013. Available from: https://www.coep.org.in/page_assets/341/Metagenomic_Sequence_Analysis_using_Hybrid_Approach.pdf

103. Hug LA, Beiko RG, Rowe AR, Richardson RE, Edwards EA. Comparative metagenomics of three Dehalococcoides-containing enrichment cultures: the role of the non-dechlorinating community. BMC Genomics. 2012;13(1):327. https://doi.org/10.1186/1471-2164-13-327

Additional information

Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

Publisher information: Plant Science Today is published by HORIZON e-Publishing Group with support from Empirion Publishers Private Limited, Thiruvananthapuram, India.