Open Access Research Journal of
Life Sciences
Journals home page: https://oarjpublication/journals/oarjls/
ISSN: 2783-025X (Online)

OARJ OPEN ACCESS RESEARCH JOURNALS

(RESEARCH ARTICLE)

In silico diatoms diversity analysis using 18S rRNA sequence available in NCBI

Manal Khan, Himanshu Sisodia and Aparna Pareek *

Department of Botany, University of Rajasthan, 302004, Jaipur, India.

Open Access Research Journal of Life Sciences, 2026, 11(01), 001-008

Publication history: Received on 09 October 2025; revised on 17 November 2025; accepted on 19 November 2025

Article DOI: https://doi.org/10.53022/oarjls.2026.11.1.0035

Abstract

The diatoms represent a very diverse family of algae that has significant economic and ecological implications. They act as mediators in the global biogeochemical cycles and serve as reliable bioindicators. For this study, diversity analysis was performed using the 18S rRNA sequences of 52 diatom species obtained from the NCBI GenBank database. We checked the sequence quality with AlgaeBase with the intention of verifying the genus and species, it ranged from 1300 to 1600 bp. Multiple sequence alignment was performed in MEGA 12 using the ClustalW algorithm, and Tamura-Nei appeared as the best-fitting nucleotide substitution model. To assess the robustness of the tree, 1,000 bootstrap replicates were run using a Maximum Likelihood construction method. The resultant phylogeny exhibited some divergence and some conserved evolutionary patterns, having well-supported clades corresponding to certain previously identified diatom lineages. In other words, the 18S rRNA gene is very informative about the taxonomy of diatoms and is very useful in the elaboration of evolutionary relationships. The study also provides a molecular infrastructure by which ecological monitoring programs may be furthered and our knowledge of diatom systematics enhanced. Therefore, all these molecular tools, like the 18S rRNA analysis, provide a firm foundation for furthering both the ecological research of diatoms and their biodiversity assessments.

Keywords: Diatoms; Phylogenetic Analysis; Molecular Taxonomy; 18S rRNA Gene; And Bioindicators

1 Introduction

Diatoms are unicellular microscopic algae which are a vital part of phytoplankton communities in aquatic ecosystems including oceans and rivers. Diatoms are recognized throughout the world as mark of water quality evaluation. Regardless of their proven effectiveness, their application in water bodies monitoring in India continues to be constrained. Diatoms based indices can boost ecological assessment usages across Indian water bodies. (Srivastava, *et al*., 2016). In marine ecosystems, which are mainly primary producers, diatoms contribute almost 40 percent of the total primary production, hence the importance of their role in the global carbon cycle. They are valuable assets for industrial and environmental applications, due to their versatility and biotechnological use. (Fu, *et al*., 2022). By promoting nutrient cycling, regulating the climate, and providing useful use in technology, education and medicine. Diatoms benefit important ecosystem services spanning cultural, ecological and technological aspects significantly in human well-being and biodiversity. (B-Béres, *et al.,* 2023).

Diatoms are important biomarkers in freshwater ecosystems in view of their sensitivity to environmental changes and elevated biodiversity. Research indicates that heterogeneity and abundance of Diatoms reveal organic contamination, water quality, and anthropogenic influences. Species like *Gomphonema parvulum* and *Nitzschia palea* are primary pollution indicators. (Parikh *et al.*, 2020). Diatoms are notably valuable for evaluating organic pollution, climatic variations, and water eutrophication. They are effective bioindicators due to their quick development, high biomass, and responsiveness to oxygen concentration, pH value, temperature, and light. (Khudher *et al.,* 2019).

Particularly during diatom studies, the 18S rRNA gene and more specifically its V4 region have been considered an advantageous tool to gauge microbial diversity. Using a distance-based algorithm method, the V4 region shows high

* Corresponding author: Aparna Pareek

identifying power and separates approximately 97 percent of central diatom species. Species identification improved further through integration of DNA character-based using the 18S rRNA gene. (Luddington *et al.,* 2012). Both taxonomic groups of Diatoms, higher and lower, can be successfully gathered by utilizing the 18S rRNA gene. (Guo, *et al.,* 2015). The 18S rRNA is a reliable tool for diatoms detection due to its broad reference library and accurate amplification, although it is less variable compared to alternatives such as rbcL (Ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit) and COI (Cytochrome c oxidase). Further, to improve its usefulness for diatom categorization, large numbers of databases must be created. (An *et al.,* 2017). Large sequencing datasets and libraries around the world of 18S rRNA sequences of Diatoms are freely accessible, provided by the NCBI. Using biological sequences, annotations, and metadata analytics across many taxa is made effortless by the NCBI dataset's simple interfaces and tools. (O'Leary *et al.,* 2024). GenBank is a primary nucleotide sequence database maintened by the National Center for Biotechnology Information (NCBI). It contains over 557,000 species datasets, with new records being added daily. GenBank allows sharing and accessing data across the globe, in addition to providing a way to retrieve and analyze sequences, through tools such as BLAST and Entrez. (Sayers *et al*.,2024).

2 Methodology

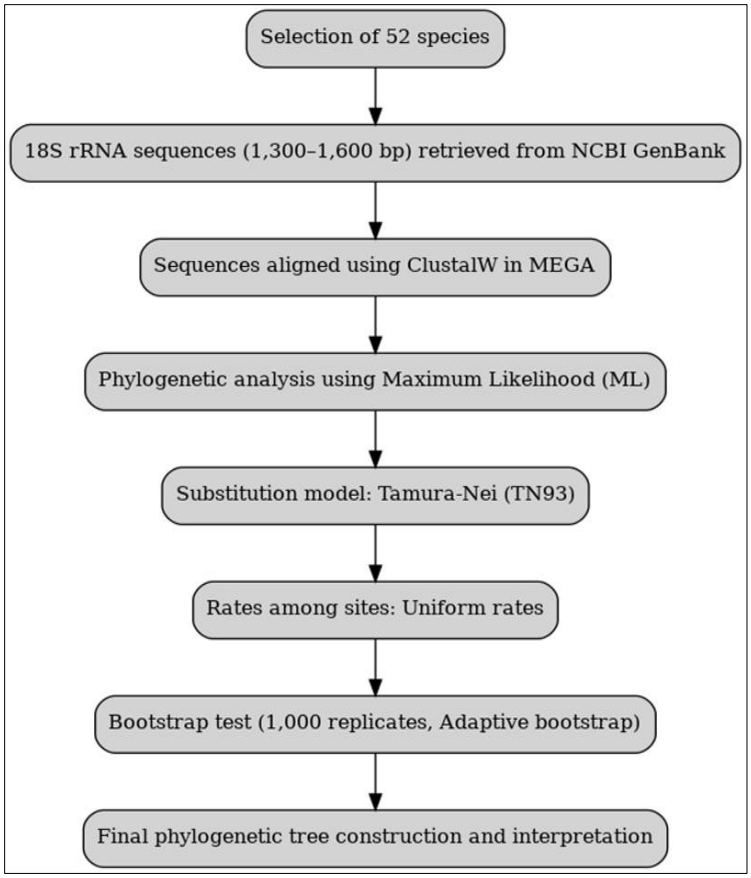

Figure 1 Workflow diagram of the data collection and phylogenetic analysis steps used in this study

A total of 52 diatom species were chosen to represent a wide spectrum of taxonomic and ecological diversity. The taxa were obtained through the Diatoms of North America database, which provides taxonomically verified information on North American diatom taxa and information about their ecological relevance (Spaulding, *et al.*, 2010). For each selected species, 18S rRNA gene sequences were obtained from the NCBI GenBank database. The species names of the diatoms were cross-checked with AlgaeBase, but it was primarily used to check nomenclature and classification (Guiry, *et al.*, 2023). Quality control was strict when selecting sequences for the study. Sequences were retained only if they were between 1300 and 1600 base pairs long, and those with clear annotations that were not ambiguous were preferred. All sequences were downloaded in FASTA format for downstream analyses.

Data analysis was conducted using MEGA version 12 (Tamura, *et al.,* 2021). A multiple sequence alignment was performed with the CLUSTALW algorithm set to the default options to identify homologous regions across the dataset (Thompson *et al.*, 1994). The "Find Best DNA/Protein Model" function in MEGA 12 was used to select the most

appropriate nucleotide substitution model, and the best-fitted model based on statistical criteria was the Tamura-Nei model (Tamura *et al.,* 1993). Phylogenetic reconstruction was carried out using the Maximum Likelihood (ML) approach under the Tamura-Nei model. Analyses were conducted with the set of sites having uniform rates, deleting gaps and gaps between sites paired, and 1,000 bootstrap replicates with the Adaptive Bootstrap method. The reliability of each clade was assessed using the bootstrap values after the resulting phylogenetic tree was analyzed in MEGA 12.

3 Results

3.1 Dataset Summary

We analyzed fifty-two 18S rRNA sequences (one per species), each of which was between 1300-1600 bp. Following alignment in MEGA 12, the dataset included a total of 1,808 nucleotide positions. In total, there were 925 constant sites (51.2% constant sites), 523 (29.4%) parsimony informative sites, and 348 singleton sites (19.4%) visualized in Figure 2. Together, this represents a good phylogenetic signal for likelihood analysis.

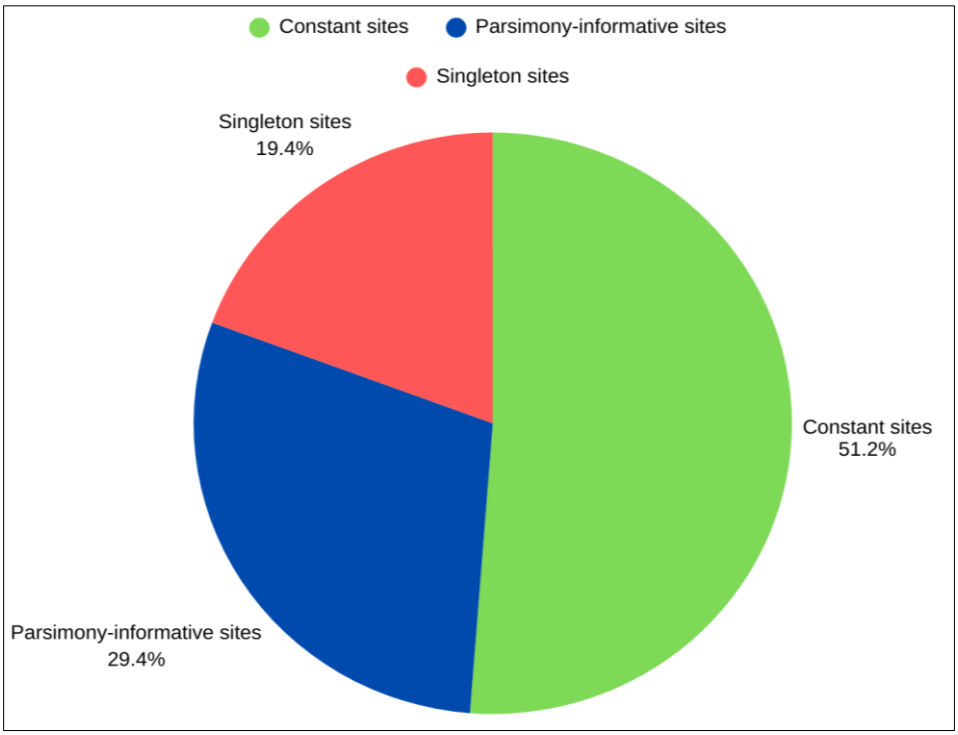

Figure 2 Proportion of constant (51.2 %), parsimony-informative (29.4 %), and singleton (19.4 %) sites in the 18S rRNA alignment of 52 diatom species

3.2 Phylogenetic Tree Construction

The most appropriate substitution model was the Tamura-Nei (TN93) model of uniform rates. Six primary clades (A-F) were recovered with different support values from the bootstrap analysis by maximum likelihood analysis (Figure 3). Overall, most nodes had significant support (≥70%), and most of the vital clusters are >90–100% support, reflecting confidence in the topology inferred.

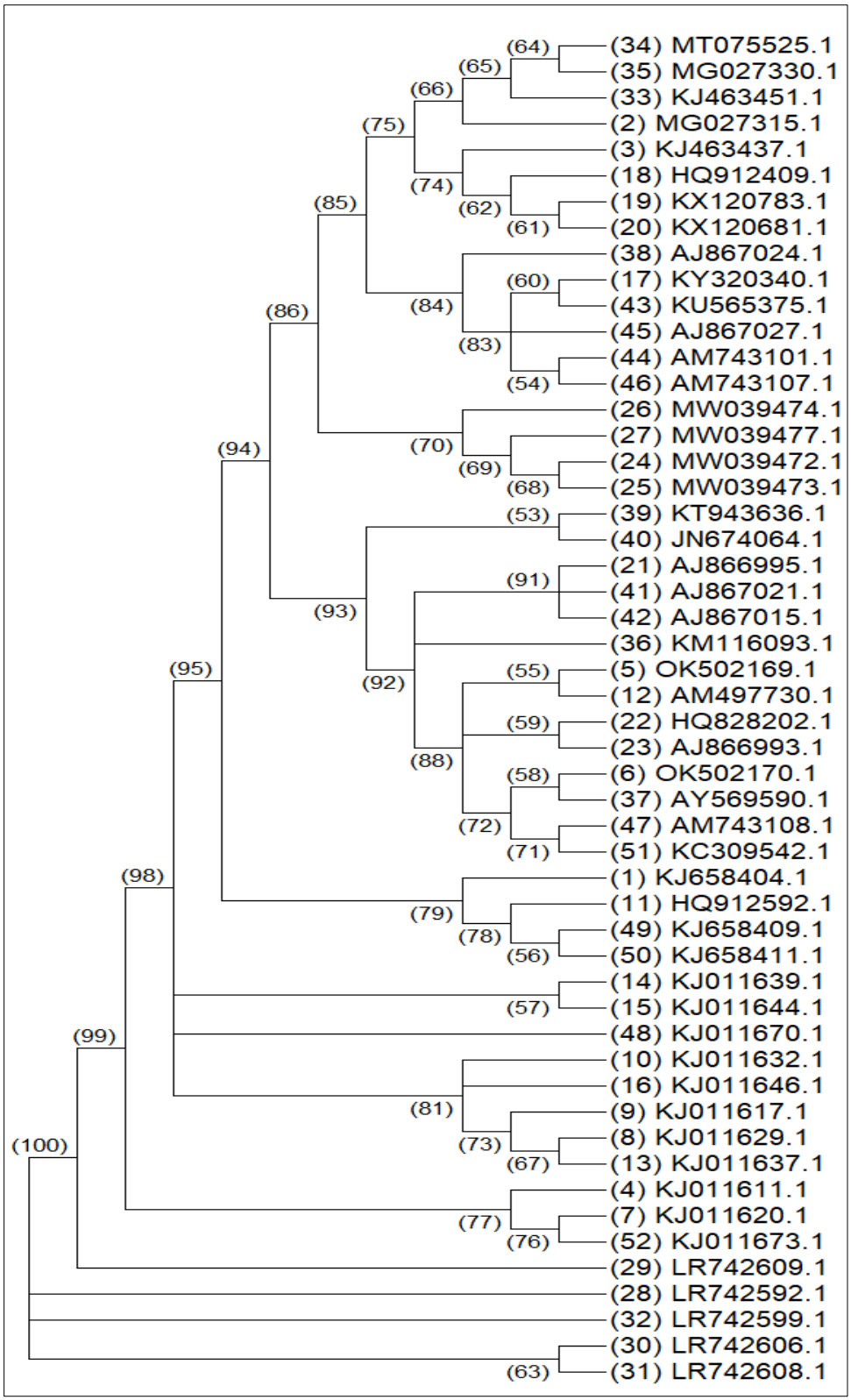

Figure 3 52 diatom species' maximum likelihood phylogeny based on 1,808 bp of 18S rRNA gene. The Tamura–Nei model with uniform rates was used to build the tree in MEGA. Only values ≥ 50% are presented at the nodes, which display bootstrap values (1,000 repeats, adaptive approach). The majority of the six major clades (A–F) have high statistical support.

The list of species names, GenBank accession numbers, and details of the sequences analyzed in this study can be found in Supplementary Table S1. A–D clades consistently presented high support values, which reflects strong taxonomic groupings. Clade B and Clade D also presented support bootstraps of more than >90–100% and could also reflect safe evolutionary lineages. Clade E with <50% support probably presents these patterns because 18S rRNA is conserved across fine taxonomic scales. Clade F, with mixed values, possessed certain subclusters with strong support, whereas the deeper nodes possessed moderate support. The Maximum Likelihood tree had a log-likelihood of –18,669.00, which indicated an acceptable model fit to the data. Clustered species relationships were generally consistent with known diatom taxonomic groups; however, low bootstrap support for some clusters indicates that either incomplete lineage sorting (ILS) or un-resolved phylogenetic signal may have occurred. This reinforces the need for additional complementary markers for better resolution of the species group potential for use at finer scales of taxonomic placement. Clade F showed more mixed values with strong support in several subclusters, but moderate support at deeper nodes.

3.3 Clade Wise Analysis

3.3.1 Clade A: Cymbellales–Gomphonemataceae Cluster

- **Species:** *Anomoeoneis sculpta, Cymbella sp. (cistula, tumida, aspera), Cymbopleura naviculiformis, Didymosphenia geminata, Encyonema minutum and prostratum, Encyonopsis sp., Rhoicosphenia abbreviata,* multiple *Gomphonema spp.*
- Support: Notable bootstrap values (76-100%).
- Interpretation: This significant, consistently supported clade represents a core lineage of pennate diatoms in Cymbellales and Gomphonemataceae. Support was present from Tarcz, *et al.,* (2013) when they put forth that Gomphonema groups in a monophyletic fashion in Gomphonemataceae, and is sister to other cymbelloid forms. (Thomas,*et al.,* 2016*). Morphologically*, all of the taxa have asymmetric valves and a benthic mode of life, which also confirms the phylogenetic relationship. The placement of Rhoicosphenia (Zhu, *et al*.,2016) here is also consistent with some evidence that the association with Cymbellales is a likely one, but there continues to be debate about a connection to either Cocconeis or Achnanthidium.(Kociolek, *et al.,* 1986)

3.4 Clade B: *Achnanthidiaceae* Group

- **Species:** *Achnanthidium minutissimum, Planothidium frequentissimum, Planothidium lanceolatum.*
- Support: Modest bootstrap (56-79%).
- Interpretation: This compact clustering represents monoraphid diatoms that fit in Achnanthidiaceae (Znachor, *et al.,* 2015). Achnanthidium has also been recognized to create biofilms in freshwater habitats, where they also have species with unique extracellular polymeric capsules concomitant to bacteria to form biofilm (Windler, *et al.,* 2015). The close genetic match to Planothidium also supports taxonomic beliefs that the genera share a family boundary. (Round, *et al.,* 1996)

3.5 Clade C: Naviculaceae–Stauroneidaceae Complex

- **Species:** *Frustulia gaertnerae, F. septentrionalis, F. vulgaris, Frustulia sp., Navicula pseudacceptata, Navicula hippodontofallax***.**
- Support: High (68–100%).
- Interpretation: The clade brings together a group of acidophilic Frustulia representatives and a more cosmopolitan group of Navicula taxa—good valve shapes are essentially doing the same ecological thing, with Frustulia spending time in acidic peat bogs and acidic streams (Bouchard, *et al.,* 2021). Molecular work shows that 18S rRNA is a trustworthy marker for the resolution of such high-level relationship, while more variable markers would resolve species-level differences more effectively. (Urbánková, *et al.,*2016)

3.6 Clade D: Amphora–Epithemia–Pinnularia Assemblage

- **Species:** *Halamphora spp. (siqueirosii, subtropica, oligotraphenta), Amphora spp., Epithemia spp., Entomoneis paludosa, Pinnularia spp. (mesolepta, rupestris, catenaborealis, termitina).*
- Support: High (70–98%).
- Interpretation: This mixed assemblage includes benthic lineages from at least three families: Amphipleuraceae, Epithemataceae (which are known to have nitrogen-fixing symbionts), and Pinnulariaceae. The highlighted characteristics of Epithemia's nitrogen-fixing ecology have been widely reported as part of previous studies, in both oligotrophic habitats; the clustering suggests they share an ancestor even if a shared ecology has diverged. (Moulin, *et al*.,2023)

3.7 Clade E: Eunotiales–Bacillariales Link

- **Species:** *Eunotia bilunaris, Nitzschia pusilla, Nitzschia inconspicua, Hantzschia amphioxys.*
- Support: Moderate (33–91%).
- Interpretation: This small clade connects taxa traditionally present in Eunotiales (Eunotia) with Bacillariales (Nitzschia, Hantzschia), suggesting an evolutionary connection between araphid and raphid lineages. Similar arrangements have been demonstrated in mitochondrial phylogenetic trees, displaying a complex lineage history for monoraphid diatoms. (Thomas, *et al.,*2016).

3.8 Clade F: Centric and Araphid Lineage

- Species: *Asterionella formosa, Aulacoseira pusilla, Diatoma tenuis, Fragilariforma sp., Fragilaria ulna, Melosira varians, Pinnularia viridiformis, Pseudosolenia calcar-avis.*
- Support: Robust (74–100%) in subclusters.
- Interpretation: This clade groups centric (e.g., *Aulacoseira, Melosira)* and araphid (e.g., *Asterionella, Fragilaria*) diatoms together; connecting lineages that have clear morphological differences, but likely conserved deeper molecular signals. *Pseudosolenia,* a marine taxon, falls inside this group, suggesting a common evolutionary foundation across marine and freshwater lineages. These results provide support for molecular phylogenetic schemes that align araphid lineages near centric ancestors. (Sims, *et al.,* 2006*).*

The Maximum Likelihood phylogenetic tree demonstrated six clades (A–F), with varying levels of bootstrap support. Clade B and Clade D had the highest support values (>90–100%), suggesting they are well-defined evolutionary lineages of the diatom taxa analyzed. Clade A and Clade C also had relatively high support (65–95%) and thus we believe the sub-groupings to be stable. Clade F internally had moderate support (42–95%), with a reasonable mixture of well-supported and less well-supported branches. While Clade E had low bootstrap values (<50%), indicating unresolved phylogenetic relationships, it may also reflect the existence of some cryptic diversity or taxonomic ambiguities. By and large, the bootstrap analysis suggests that most of the nodes are supported with values > 70%, which reinforces the reliability of the inferred evolutionary relationships in the reconstructed tree and confidence in the relationships between the 52 diatom species presented in the study.

4 Discussions

This study helps substantiate phylogenetic relationships among 52 diatom species employing 18S rRNA sequences. The maximum Likelihood analysis identified six overall clades with varying degrees of bootstrap support. These results are generally in line with earlier research (Medlin, *et al.*, 2004; Alverson, *et al.*, 2008). The strong values in Clades A–D indicate better support among stable evolutionarily related lineages, whereas Clade E showed much less support (<50%); this is likely due to the conservative nature of 18S and hence greatly reduced divergence at the level at which we are working (Guo, *et al.*, 2015). We acknowledge this reduction in our original selection of markers because while it causes relatively little error in higher-level phylogenetic inference, 18S is often insufficient for species-level resolution (Theriot *et al.,* 2009). Our curated dataset eliminated many errors that are present in misidentified, poorly sampled, and low insertion quality entries, thereby improving the phylogenetic integrity of our analyses. In future studies investigating the same taxa, additional markers like ITS (Internal Transcribed Spacer), COI, or hypervariable regions of 18S rRNA (Luddington, *et al*., 2012) would improve resolution. The results from this study validate some strengths of 18S for diatom systematics but also describe the limitations with regards to levels of resolution for more taxonomically recent divergences.

5 Conclusion

This study illustrates the potential of 18S rRNA as the preferred marker for documenting phylogenetic relationships in diatoms. Phylogenetic analysis of the 52 species showed six main clades, with most supported with high bootstrap support, and with phylogenetic relationships that aligned with previous work performed on Bacillariophyta. The analysis has shown the discriminatory potential of using 18S rRNA to resolve relationships at monophyletic higher taxonomic groups, while acknowledging that 18S rRNA is not very useful to resolve species-level relationships. The study utilized a curated portion of the dataset, and a researcher reproducible analysis workflow, which improves common molecular phylogenetics sources of error, and offers an effective method for conducting relevant biodiversity assessments and facilitating taxon's validation. Future work should integrate markers that evolve faster than 18S, which can improve species-level resolution.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Alverson, A. J. (2008). Molecular systematics and the diatom species. Protist, 159(3), 339–353. https://doi.org/10.1016/j.protis.2008.04.001

[2] An, S. M., Choi, D. H., Lee, J. H., Lee, H., and Noh, J. H. (2017). Identification of benthic diatoms isolated from the eastern tidal flats of the Yellow Sea: Comparison between morphological and molecular approaches. PLOS ONE, 12(6), e0179422. https://doi.org/10.1371/journal.pone.0179422

[3] B-Béres, V., Stenger-Kovács, C., Buczkó, K., et al. (2023). Ecosystem services provided by freshwater and marine diatoms. Hydrobiologia, 850, 2707–2733. https://doi.org/10.1007/s10750-022-04984-9

[4] Bouchard, A. (2021). Incorporating molecular data in the taxonomic study of diatoms: an example using two well-known genera, Frustulia and Navicula S.S. (Bacillariophyceae, Naviculales). https://doi.org/10.20381/ruor-26609

[5] Eric, W. S., Cavanaugh, M., Clark, K., Pruitt, K. D., Sherry, S. T., Yankie, L., Karsch-Mizrachi, I., et al. (2024). GenBank 2024 update. Nucleic Acids Research, 52(D1), D134–D137. https://doi.org/10.1093/nar/gkad903

[6] Fu, W., Shu, Y., Yi, Z., Su, Y., Pan, Y., Zhang, F., and Brynjolfsson, S. (2022). Diatom morphology and adaptation: Current progress and potentials for sustainable development. Sustainable Horizons, 2, 100015. https://doi.org/10.1016/J.HORIZ.2022.100015

[7] Gaonkar, C. C., Piredda, R., Minucci, C., Mann, D. G., Montresor, M., Sarno, D., and Kooistra, W. H. C. F. (2018). Annotated 18S and 28S rDNA reference sequences of taxa in the planktonic diatom family Chaetocerotaceae. PLOS ONE, 13(12), e0208929. https://doi.org/10.1371/journal.pone.0208929

[8] Guiry, M. D., and Guiry, G. M. (2023). AlgaeBase. World-wide electronic publication, National University of Ireland, Galway.

[9] Guo, L., Sui, Z., Zhang, S., Ren, Y., and Liu, Y. (2015). Comparison of potential diatom 'barcode' genes (the 18S rRNA gene and ITS, COI, rbcL) and their effectiveness in discriminating and determining species taxonomy in the Bacillariophyta. International Journal of Systematic and Evolutionary Microbiology, 65(Pt 4), 1369–1380. https://doi.org/10.1099/ijs.0.000076

[10] Khudher, E. K., and Al-Jasimee, A. S. (2019). Diatoms (Bacillariophyta) as bio-indicators. International Journal of Research in Pharmaceutical Sciences.

[11] Kociolek, J. P., and Stoermer, E. F. (1986). Phylogenetic relationships and classification of monoraphid diatoms based on phenetic and cladistic methodologies. Phycologia, 25(3), 297–303. https://doi.org/10.2216/i0031-8884-25-3-297.1

[12] Luddington, I. A., Kaczmarska, I., and Lovejoy, C. (2012). Distance and character-based evaluation of the V4 region of the 18S rRNA gene for the identification of diatoms (Bacillariophyceae). PLOS ONE, 7(9), e45664. https://doi.org/10.1371/journal.pone.0045664

[13] Medlin, L., and Kaczmarska, I. (2004). Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. Phycologia, 43(3), 245–270. https://doi.org/10.2216/i0031-8884-43-3-245.1

[14] Moulin, S. L. Y., Frail, S., Doenier, J., Braukmann, T., and Yeh, E. (2023). The endosymbiont of Epithemia clementina is specialized for nitrogen fixation within a photosynthetic eukaryote. bioRxiv [Preprint]. https://doi.org/10.1101/2023.03.08.531752

[15] O'Leary, N. A., Cox, E., Holmes, J. B., Anderson, W. R., Falk, R., Hem, V., Tsuchiya, M. T. N., et al. (2024). Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. Scientific Data, 11, 732. https://doi.org/10.1038/s41597-024-03571-y

[16] Round, F. E., and Bukhtiyarova, L. (1996). Four new genera based on Achnanthes (Achnanthidium) together with a re-definition of Achnanthidium. Diatom Research, 11(2), 345–361. https://doi.org/10.1080/0269249x.1996.9705389

[17] Sims, P. A., Mann, D. G., and Medlin, L. K. (2006). Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia, 45(4), 361–402. https://doi.org/10.2216/05-22.1

[18] Singh, M., and Parikh, P. (2020). Freshwater diatoms as bio-indicators in urban wetlands of central Gujarat, India. Indian Journal of Ecology, 47(1), 7–11.

[19] Spaulding, S. A., Lubinski, D. J., and Potapova, M. (2010). Diatoms of North America: The freshwater flora. Academy of Natural Sciences of Philadelphia. https://diatoms.org

[20] Srivastava, P., Verma, J., Grover, S., and Khan, A. (2016). On the importance of diatoms as ecological indicators in river ecosystems: A review. Indian Journal of Plant Sciences, 5, 70–86.

[21] Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution, 10(3), 512–526.

[22] Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis version 11. Molecular Biology and Evolution, 38(7), 3022–3027. https://doi.org/10.1093/molbev/msab120

[23] Tarcz, S., et al. (2013). Molecular phylogeny of Gomphonema (Bacillariophyceae) inferred from nuclear and plastid genes. Journal of Phycology. https://pmc.ncbi.nlm.nih.gov/articles/PMC11680788

[24] Theriot, E. C., Cannone, J. J., Gutell, R. R., and Alverson, A. J. (2009). The limits of nuclear-encoded SSU rDNA for resolving the diatom phylogeny. European Journal of Phycology, 44(3), 277–290. https://doi.org/10.1080/09670260902749159

[25] Thomas, E. W., Stepanek, J. G., and Kociolek, J. P. (2016). Historical and current perspectives on the systematics of the 'enigmatic' diatom genus Rhoicosphenia (Bacillariophyta), with single and multi-molecular marker and morphological analyses and discussion on the monophyly of 'monoraphid' diatoms. PLOS ONE, 11(4), e0152797. https://doi.org/10.1371/journal.pone.0152797

[26] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22(22), 4673–4680. https://doi.org/10.1093/nar/22.22.467

[27] Urbánková, P., Scharfen, V., and Kulichová, J. (2016). Molecular and automated identification of the diatom genus Frustulia in northern Europe. Diatom Research, 31, 217–229. https://doi.org/10.1080/0269249X.2016.1224780

[28] Windler, M., Leinweber, K., Bartulos, C. R., Philipp, B., and Kroth, P. G. (2015). Biofilm and capsule formation of the diatom Achnanthidium minutissimum are affected by a bacterium. Journal of Phycology, 51(2), 343–355. https://doi.org/10.1111/jpy.12280

[29] Zhu, H., et al. (2016). The diatom genus Rhoicosphenia: Phylogenetic placement and biogeographic distribution. Protist, 167(2), 178–191. https://pmc.ncbi.nlm.nih.gov/articles/PMC4821588

[30] Znachor, P., et al. (2015). Extracellular polymeric substances of Achnanthidium minutissimum provide attachment sites for bacteria. FEMS Microbiology Ecology, 91(5). https://pmc.ncbi.nlm.nih.gov/articles/PMC4380156