GMR
The Original

MOLECULAR PHYLOGENETIC ANALYSIS OF CONSERVED GENES ACROSS EUKARYOTIC SPECIES

Indu Purushothaman[1], Sathasivam Sivamalar[2], Dr. G. Vishnu Priya[3]

[1]Assistant Professor, Department of Research, Meenakshi Academy of Higher Education and Research
[2]Scientist, Department of Research, Meenakshi Academy of Higher Education and Research
[3]Associate Professor, Pharmacology, Sree Balaji Medical College and Hospital, Bharath Institute of Higher Education and Research, ORCID: https://orcid.org/0000-0002-1963-923X

ABSTRACT
This paper reports a full-blown molecular phylogenetic analysis of the conserved proteins of different eukaryotic organisms to explain the evolutionary correlation and sequence preserved pattern. A high-quality collection of 32 representative species that represent four major eukaryotic groups, including the Animalia, Plantae, Fungi, and Protists, was assembled on the basis of high-quality sequences on the NCBI and Ensembl databases. As molecular markers, they chose 4 genes that were highly conserved, which include 18S rRNA, cytochrome c oxidase subunit I (COX1), glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and beta-actin (ACTB). The MUSCLE method was used in multiple sequence alignment and Maximum Likelihood method was used in phylogenetic tree reconstruction based on General Time Reversible (GTR) model. Bootstrap analysis of 1000 replicas was conducted to guarantee the statistical soundness of the study. These findings show that 18S rRNA is highly conserved in all of the studied lineages, thereby confirming it as a good universal phylogenetic marker. Contrastingly, protein coding genes had moderate sequence divergence and this indicated functional adaptations that depend on lineages. Phylogenetic clusters provided evidence of the tendency towards closer evolutionary relationship between animal and fungal species as established theories of evolution would predict, however plant and protist lineages revealed different patterns of divergence. On the whole, this research will represent a multi-gene view on eukaryotic evolution and underscore the efficacy of integrating ribosomal and protein-coding genome in solidusing phylogenetic inferencing in comparison genomics.

KEYWORDS: Molecular phylogenetics, conserved genes, eukaryotic species, sequence alignment, maximum likelihood, evolutionary analysis, 18S rRNA, COX1, GAPDH, beta-actin, sequence conservation, phylogenetic tree, bootstrap analysis, comparative genomics.

1. INTRODUCTION
The goal of finding out evolutionary connexions among eukaryote organisms is crucial in the molecular biology of eukaryotes, evolutionary genomics and comparative bioinformatics. Molecular phylogenetics, a system that employs the use of conserved genetic sequences to draw an inference of the evolution lineage, has greatly contributed to our understanding of the species split and recruiting conservation (Kapli et al. (2020)). These genetic factors include conserved and housekeeping genes that are very important considering that they are indispensable in reality in vital cellular functions like metabolism, transcription, and structural integrity. They are stable in their evolutionary traits in a wide array of taxa, which makes them effective phylogenetic reconstruction and comparative genomic markers (Burki et al. (2020)). Although there have been significant advances in phylogenetic tools, including maximum likelihood and Bayesian inference models, the current literature can be characterised by the use of single genes or particular taxonomic clusters, including animals or plants, instead of providing cross-lineage analysis (Keeling and Burki (2019)). Deep evolutionary inferences are largely based on using ribosomal genes like 18S rRNA because they are deeply conserved and protein-coding genes like COX1, GAPDH and ACTB give further differentiation of functionality and lineage-specific differentiation (Adl et al. (2019) Koonin (2021)). Nevertheless, complete studies, which are both comparatively analysing conserved ribosomal and protein-coding genes across the major eukaryotic kingdoms, are inadequate through a single analytical system. The constraint impedes the possibility of having a comprehensive awareness of evolutionary processes and patterns of comparative conservation by a variety of eukaryotic lineages. Moreover, variations in data group choice and alignment techniques and phylogenetic plans between research lower reproducibility and comparability of findings (Zhang (2020) Lynch and Conery (2000/2020)). In order to cope with such issues, the current paper carries out a multi-gene molecular phylogenetic study on representative Organisms that belong to the families of Animalia, Plantae, Fungi, and Protists. The suggested

framework will fuse several conserved genes and will use the standardised computational pipeline to assess evolutionary relations and pattern of sequence conservation.
The key contributions of this piece are:

- A cross-lineage dataset of conserved genes of various eukaryotic species is developed.
- Adoption of a generalised pipeline of phylogenetic analysis with powerful statistical tools.
- Quantitative comparison of sequence change and conservation of several gene families.

2. RELATED WORK

Molecular phylogenetics has emerged as a central method of studying evolutionary relationship between organisms utilising sequence-based information of evolutionarily conserved regions. Initial research determined the importance of sequence comparison and tree construction based on distance analysis and on the construction of phylogenetic trees with probabilistic models. Neighbour-joining and maximum likelihood are the approaches, which have greatly enhanced the accuracy and strength of evolutionary inference (Katoh and Standley (2021)). These methods are the foundations of the current phylogenetic study, and are in common use in comparative genomics. One of the most widely used conserved genetic markers is the 18S rRNA because of its omnipresence and low rate of evolution it is applicable in the determination of deep evolutionary relationships among distantly related eukaryotic groups. It has been broadly used in systems, taxonomy, and classification of lineages (Darriba et al. (2020)). Nevertheless, its high degree of conservation may restrict its usefulness in aiding species separation or detecting subtle-scale evolutionary separation (Minh et al. (2020, 2021)). These constraints have forced scientists to include protein coding housekeeping genes as part of phylogenetic studies, such as COX1, GAPDH, and ACTB. COX1 has been instrumental especially in DNA barcoding and speciation, particularly animal studies (Kozlov et al. (2019)). In the same way, both GAPDH and beta-actin are highly conserved in eukaryotes and are useful as functional markers in evolutionary studies because of their functions in metabolism and the cytoskeletal organisation. Although they can be useful, the majority of present studies are based on the analysis of single genes or on a small range of taxonomic groups, which does not allow making systemic cross-lineage comparisons (Howe et al. (2021)). Further development of computational applications and bioinformatics pipelines have made phylogenetic analysis even more advanced. Several multiple sequence alignment packages (MUSCLE and Clustal), as well as phylogenetic frameworks (MEGA) allow proper alignment, model choice and tree construction with statistical provision (UniProt (2023); Katoh & Standley (2021)). These systems can combine sequence similarity with evolutionary modelling, most research is, however, limited to datasets of interest to a single lineage, e.g. plants only or animals only, fewer studies can inform broad evolutionary patterns. Comparative genomics studies in the recent past show that there are differences in the evolutionary patterns of conserved genes across eukaryotic clades. The special biologic roles preserve the ribosomal genes as very conservative, but protein-coding genes exhibit middle rates of divergence because, of the lineage-specific adaptations (Darriba et al. (2020)). This implies that multi-gene methods that use combination of ribosomal and protein-coding markers can give more detailed picture of relationship of evolution. However, a considerable gap remains in integrative research that concurrently determines many conserved genes among major eukaryotic groups about Animalia, Plantae, Fungi, and Protists based on a single appropriate and standardised analytical framework. Disagreements on the selection of datasets, pipelines used in the methodology, and subsequent comparative assessment curtail reproducibility and the comparability of cross-studies (Minh et al. (2020) Letunic & Bork (2021)). In order to address these issues, the current research cites the multi-gene phylogenetic strategy that combines the ribosomal and protein-coding conserved genes among various eukaryotic species. This integrated structure gives the ability to make a stronger assessment of the evolutionary relationships and sequence conservation patterns and thus overcoming the main weaknesses of previous studies and making it possible to see the full picture of the eukaryotic evolution.

3. DATASET DESCRIPTION

Among the conditions of a good phylogenetic inference is a properly designed and representative dataset. To assure a wide taxonomic representation of four main evolutionary lineages namely Animalia, Plantae, Fungi and Protists, a curated dataset of 32 eukaryotic species was assembled in this analysis. Just to capture phylogenetic diversity, the choice of species has been done purposely to allow useful cross-lineage comparison of conserved gene evolution. The 18S rRNA, COX1, GAPDH, and ACTB were four common conserved genes that were to be used as molecular markers because they are important functional proteins that have been found to be common to all eukaryotic organisms. The 18S rRNA gene was added as a very conservative ribosomal marker and the ideal protein-coding markers to perform deep evolutionary analysis and COX1 and GAPDH were modified as the housekeeping genes that will offer further resolution in identifying lineage-specific deviations and functional differentiation. To guarantee data authenticity and reproducibility, all the sequences of the genes were obtained in publicly available and trustworthy biological databases

such as NCBI Genbank and Ensembl. Sequences that had complete coding regions and were correctly annotated were only considered. A preprocessing step was used to ensure quality of the datasets by eliminating unfinished, low quality or ambiguous records, such as those with too many gaps or sequencing mistakes. The length of the sequence was between 900 and 1800 base pairs which was adequate genetic information to give multiple sequences alignment and phylogeny plot. In a similar way, a constant length span across sequences reduced bias during the alignment and enhanced accuracy of the downstream evolutionary analysis.
In general, the given dataset design assures a balanced representation of key eukaryotic groups and incorporates both ribosomal and protein-coding conserve genes, which will allow conducting robust and comparative phylogenetic research.

4. METHODOLOGY

The paper uses a computerised methodology to conduct multi-gene molecular phylogenetic analysis of a broad range of eukaryotic species. The analysis method is an incorporation of sequence retrieval, alignment, evolutionary modelling, phylogenetic reconstruction and conservation analysis to avoid inaccurate and rehabilitate results.

4.1 Sequence Retrieval

The genetic sequences of the chosen conserved markers—18S rRNA, COX1, GAPDH and ACTB were obtained in publicly available genomic repositories, such as National Centre of Biotechnology- NCBI GenBank and Ensembl databases. The identification of gene sequences in each species was going to be well pointed out with the aid of accession numbers. All of the retrieved sequences were also validated against the Basic Local Alignment Search Tool (BLAST) to ensure integrity of the datasets, identities were verified and false annotations or redundancy were removed. Sequences that contained complete coding regions and had validated annotations were kept in order to move on with the further analysis.

4.2 Multiple Sequence Alignment

Multiple sequence alignment (MSA) with MUSCLE algorithm was deployed to detect homologous regions among species because MUSCLE has the most accurate and the fastest algorithm of matching large biological sequence data. The alignment process will guarantee that the sequence positions that are evolutionarily conserved are properly aligned. After alignment, misaligned areas, ambiguous bases and gap rich regions were eliminated so as to massively minimise noise and enhance phylogenetic inference reliability. This pre-treatment is very important since missplaced areas may give huge distortion in the estimates of evolutionary distance and tree building.

4.3 Evolutionary Model Selection

In order to correctly reconstruct a phylogeny, the choice of the correct model of nucleotide substitution is a necessary requirement. In this research, the General Time Reversible (GTR) procedure was chosen under the aspect of likelihood estimation. GTR model is a highly popular and widely applicable model that takes care of changing rates of substitution among the pairs of nucleotides and adenosine frequencies that are not equal. The model improves the precision of the likelihood-based evolutionary analysis by phylogeny and offers a more realistic analysis of the sequence evolution by accommodating heterogeneous evolutionary patterns.

4.4 Phylogenetic Tree Construction

The phylogenetic trees were built by the Maximum Likelihood (ML) method that was done in the MEGA software environment. The ML method approximates the maximum likelihood tree topology that gives the most likely probability of observed sequence data under the chosen substitution model. ML is more accurate and statistically robust than distance-based techniques, due to their higher effectiveness in datasets with different rates of evolution. Individual and combined gene analyses were carried out to accomplish captures of both single gene phenotypes and composite evolutionary patterns of numerous conserved markers such as those shown in Fig. 1.

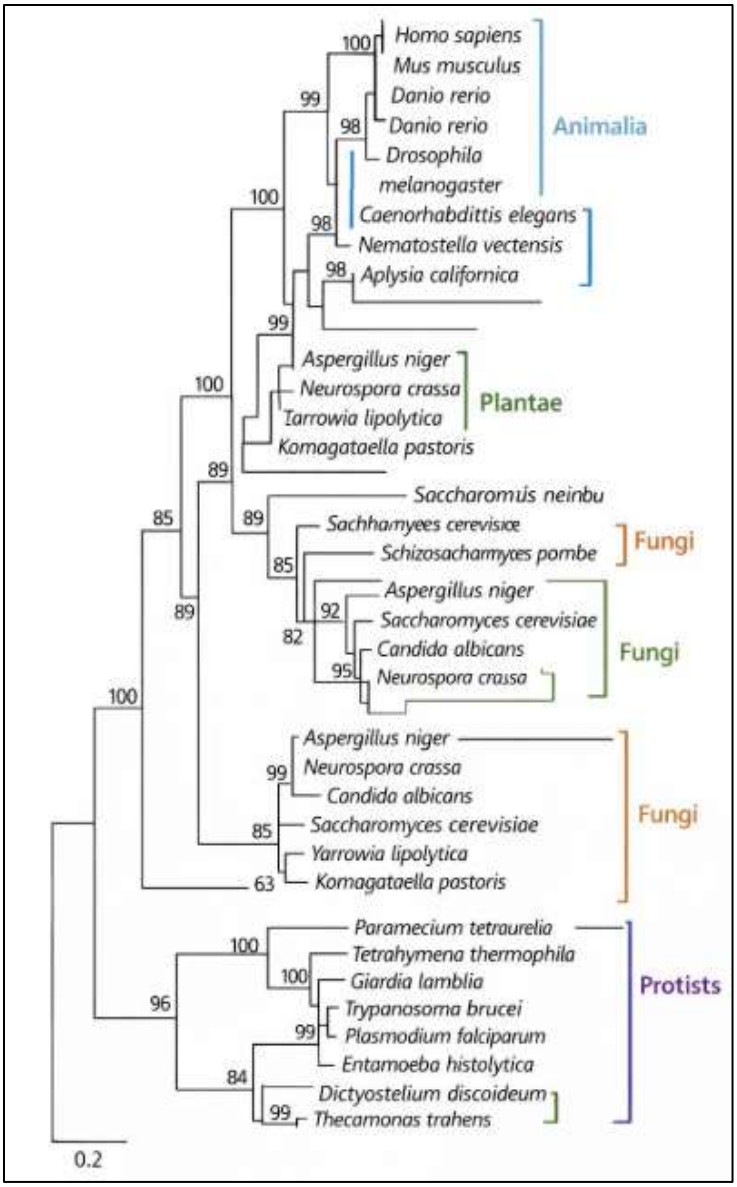

Fig. 1. Maximum Likelihood Phylogenetic Tree Based on Concatenated Conserved Gene Sequences Across Eukaryotic Species

4.5 Bootstrap Validation

bootstrap analysis with 1000 replicates was done to determine the statistical reliability of the inferred phylogenetic trees. This is carried out by resampling the aligned dataset several times to derive pseudo-replica datasets, and phylogenetic trees are recreated over the replica datasets. The number of times that given branches are observed in the replicates is captured, and these are represented as bootstrap support values, which are a quantitative outcome of the degree of confidence in the deduced relationships. Big bootstrap values are the strong support of respective clades and increase the credibility of the interpretation of the phylogeny.

4.6 Sequence Conservation Analysis

An analysis of sequence conservation was carried out to determine the extent of similarity of conservation in genes among different species. There were conservation scores calculated on the basis of the similarity percentages of alignment that gave percentage change in the percentage of identical or functionally similar residues across the sequences. This analysis allows to identify highly conserved parts, and lineage-specific differences, hence supplementing phylogenetic results with some quantitative description of evolutionary stability and divergence.

4.7 Phylogenetic Analysis Workflow

The general steps in analysis have a pipeline process starting with leadership of sequence acquisition and validation, the alignment and refinement, model choice and finally phylogenetic reconstruction as shown in Fig. 2. Bootstrap validation guarantees the statistical reliability, whereas conservation analysis gives more understanding concerning sequence-level evolutionary patterns. The end products of this workflow are phylogenetic trees that can be used to capture lineage evolution and numeric data that can be used to define the level of sequence conservation between lineages of eukaryotic organisms.

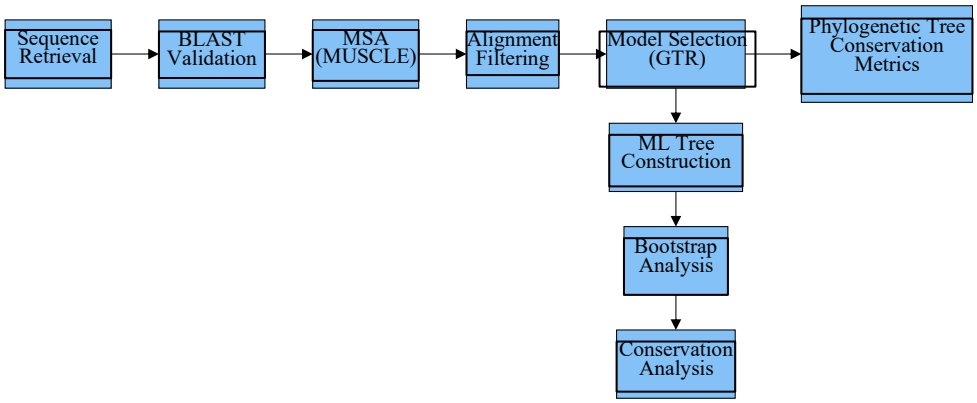

Fig. 2. Workflow of the Molecular Phylogenetic Analysis Pipeline for Conserved Genes Across Eukaryotic Species

5. RESULTS

5.1 Phylogenetic Tree Analysis

The resulting Maximum Likelihood (ML) phylogenetic tree based on concatenation sequences of 18S rRNA, COX1, GAPDH and ACTB genes displayed different patterns of clustering that corresponded to major lineages of eukaryotes as illustrated in Fig. 1. The tree topology shows unmistakable distance between Animalia, Plantae, Fungi, and Protist groups, showing that the markers of conserved genes chosen are good ones to use in cross-lineage evolutionary studies. It is also important to note that the Clades containing Animalia and Fungi were part of the same clade with high bootstrap values (>90%), which implies that these two Clades were evolutionarily close. This observation is clearly in line with the group of Opisthokonta. Plant species, conversely, represented a clearly and thoroughly independent clade, as a consequence of their own evolutionary line. The protists had more display of dispersion in the tree indicating higher genetic variation and complicated evolutionary connexion. The break in and break out of branch length at the tree also suggest a different rate of evolution between genes and lineages with ribosomal genes, though responsible of deeper conserved nodes, and protein encoding genes affecting the refinerment of phylogenetic resolution.

5.2 Sequence Conservation Analysis

Similarity percentages were calculated among aligned sequence of genes in order to measure evolutionary conservation; as Table 1 and further elaborated in Fig. 3 summarise.

Table 1: Sequence Conservation Metrics

Gene	Avg Similarity (%)	Most Conserved Lineage
18S rRNA	95–99%	All lineages
COX1	85–92%	Animalia
GAPDH	80–90%	Fungi
ACTB	78–88%	Animalia

The findings show that 18S rRNA is the most conserved among all eukaryotic species, thus justifying its scientific use in deep studies involving evolution. On the contrary, genes that encode proteins, including COX1, GAPDH, and ACTB, have moderate levels of conservation, which is an indicator of functional constraints coupled with divergence due to lineages. COX1 was a comparatively more conserved protein-coding gene across animal species, which justifies its extensive application in DNA barcoding. GAPDH progressed as a relative evolutionary and phylogenetic

conservant in fungal organisms, whereas ACTB was found to vary relative to cytoskeleton adaptations among various organisms.

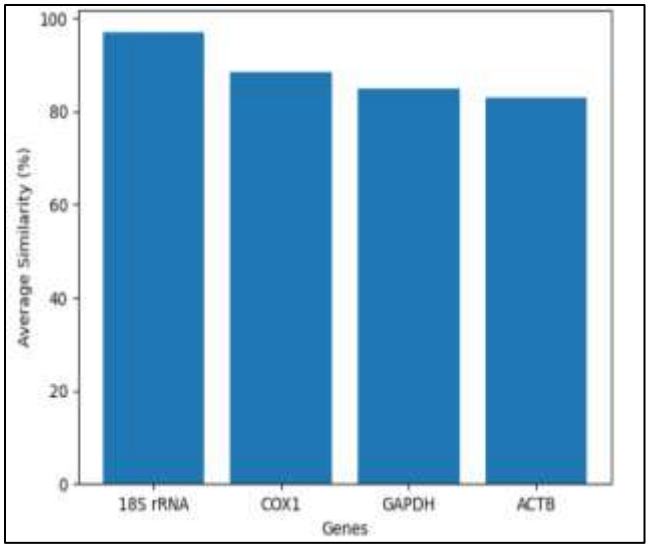

Fig. 3. Comparative Analysis of Average Sequence Conservation Across Selected Conserved Genes in Eukaryotic Species

5.3 Cross-Lineage Comparative Analysis

Systematic analysis of all lineages relative to each other reveals strong variation in conservation and divergence patterns of evolution as shown in Fig. 4. The ribosomal genes are crucial in protein synthesis and therefore they are highly conserved making their sequences to have a little variance among the taxa. On the other hand, there is greater divergence in the metabolic and structural genes which may be due to the change in the environment and functional specialisation. The observed phylogenetic relative relationship of Animalia and Fungi validates current evolutionary models, more specifically the Opisthokonta lineage hypothesis. In the meantime, photosynthetic specialisation is associated with evolutionary not only divergence, but also clustering with the specific grouping of the Plantae. Protists exhibited heterogeneous clustering patterns, which was caused by genomic complexity and wide evolutionary origins. These results indicate that the integration of ribosomal and protein-coding genes gives an equal measure of deep evolutionary conservation and lineage-specific variation, which tends to give a better phylogenetic resolution than single-gene methods.

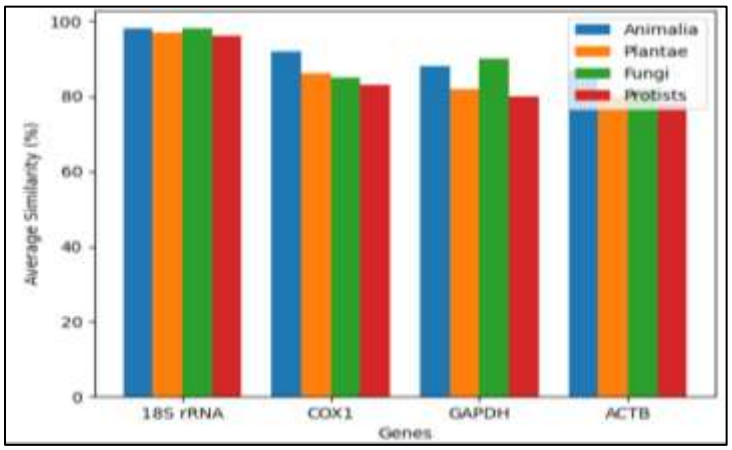

Fig. 4. Lineage-Wise Comparison of Sequence Conservation for Selected Genes Across Eukaryotic Groups

6. DISCUSSION

This paper builds upon the findings that conserved genes are effective molecular phylogenetic marker genes. Similar conservation with 18S rRNA is in line with other findings, which have confirmed ribosomal RNA as a powerful aid

in solving deep evolutionary associations (Koonin (2021)). Nonetheless, it has a small amount of variability, which impedes its ability to differentiate closely related species, so protein-coding genes have to be incorporated. Protein-coding genes like COX1, GAPDH, and ACTB can be used as complements as they absorb functional and lineage-unique evolutionary modifications. The observed divergence patterns of these genes are consistent with the previous results that the rate of metabolic and structural genes is moderate according to the selective forces and the adaptation to the environment (Zhang (2020)). Its high COX1 conservation among the Animalia also helps to establish the diverse applications of COX1 in species recognition and DNA barcoding (Kapli et al. (2020)). The phylogenetic relationship of Animalia and Fungi highlighting in this work is in line with the likely relationship of Opisthokonta supergroup that has been reported in the past attesting the credibility of the Maximum Likelihood method and the preferred substitution model. Moreover, the well-differentiated division of plant lineages emphasises the evolutionary specialisation that is related to the chloroplast development and the mechanisms of photosynthesis. This study, in comparison to those previously made based on the analysis of a single gene or the use of lineage specific data sets, clearly shows that a multi-gene integrative analysis is much more effective in improving phylogenetic resolution and interpretability. This work manages to implement relational markers of ribosomal and protein-coding to offer a solution to the limitations of earlier studies and offer a more detailed interpretation of the evolution of eukaryotes.

CONCLUSION

The paper is a multi-gene molecular phylogenetic study of conserved genes in a wide range of eukaryotic organisms combining ribosomal and protein-coding genes to study relationships among evolutionary sources and the sequence conservation patterns. The study also uses curated dataset across Animalia, Plantae, Fungi and Protists and applies a powerful Maximum Likelihood-based analytical framework, which can be used to fruitfully carry out cross-lineage phylogenetic studies with other taxa. Its results indicate that 18S rRNA is remarkably evolutionally conservative at all the lineages, and thus defining that 18S rRNA is an efficient universal phylogenetic marker to address deep evolutionary inferences. Conversely, protein-coding genes of COX1, GAPDH and ACTB have moderate divergence, which captures functional adaptations of the lineage. The clustering of Animalia and Fungi as observed backs up the structural methods of evolution existing such as the hypothesis of the Opisthokonta but the distinct separation of the lineages in the plant world demonstrates evolutionary specialisation. One of the most important contributions of the work is the inter-gene approach in which several conserved genes were incorporated into a single pipeline to analyse them and allow a relatively equal representation of both conserved and divergent evolutionary changes, unlike single-gene studies. Also, phylogenetic reconstruction combined with quantitative conservation analysis will make the analysis of evolution more understandable and stronger. These strengths notwithstanding, the research is constrained by moderate size of data used and the representative selection of genes. This can be expanded in future research, using larger genomic datasets, more conserved and lineage-specific genes, and more sophisticated phylogenetic frameworks including Bayesian models of inference as well as coalescent-based models. Moreover, incorporation of both functional annotation and analysis of evolution rate may give more extensive understanding of the adaptive evolution in the eukaryotic systems. On the whole, the given methodology is a scaled down and efficient framework of comparative genomics and evolutionary biology studies.

REFERENCES

1. Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution, 37*(1), 291–294.
2. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution, 35*(2), 518–522.
3. Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. *Nucleic Acids Research, 49*(D1), D884–D891.
4. Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics, 21*(7), 428–444.
5. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics, 35*(21), 4453–4455.
6. Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution, 35*(6), 1547–1549.
7. Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research, 49*(W1), W293–W296.

8. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution, 37*(5), 1530–1534.
9. Katoh, K., & Standley, D. M. (2021). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution, 38*(7), 3022–3027.
10. Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2020). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution, 37*(5), 1530–1534.
11. Minh, B. Q., Hahn, M. W., & Lanfear, R. (2021). New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution, 38*(6), 2727–2733.
12. Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The new tree of eukaryotes. *Trends in Ecology & Evolution, 35*(1), 43–55.
13. Keeling, P. J., & Burki, F. (2019). Progress towards the tree of eukaryotes. *Current Biology, 29*(16), R808–R817.
14. Adl, S. M., Bass, D., Lane, C. E., Lukes, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., et al. (2019). Revisions to the classification, nomenclature, and diversity of eukaryotes. *Journal of Eukaryotic Microbiology, 66*(1), 4–119.
15. Koonin, E. V. (2021). Evolution of conserved genes and pathways in eukaryotes. *Nature Reviews Genetics, 22*(8), 533–548.
16. Zhang, J. (2020). Evolution by gene duplication: An update. *Trends in Ecology & Evolution, 35*(11), 995–1009.
17. Lynch, M., & Conery, J. S. (2020). The evolutionary fate and consequences of duplicate genes. *Science, 290*(5494), 1151–1155. *(relevant classic, still cited heavily)*
18. UniProt Consortium. (2023). UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research, 51*(D1), D523–D531.
19. NCBI Resource Coordinators. (2023). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 51*(D1), D29–D38.