BRAZILIAN JOURNAL OF BIOLOGY®

ISSN 1519-6984 (Print)
ISSN 1678-4375 (Online)

Original Article

DendroSSR: SSRs and sequence alignment as tools for building phylogeny trees

DendroSSR: SSRs e alinhamento de sequência como ferramentas para a construção de árvores filogenéticas

M. Alhawatema[a]*

[a]Tafila Technical University, Faculty of Science, Department of Applied Biological Science, Tafila, Jordan

Abstract

This study introduces a new method to construct phylogenetic trees by combining both of the Simple Sequence Repeats (SSRs) and sequence alignments. The purpose of this work is to present the DendroSSR program and show it via a case study involving diverse *Aspergillus* species. To show how the DendroSSR program works to resolve complicated species relationships in phylogenetic trees, we employed the *Aspergillus* species as an example of a research case. The DendroSSR employs a technique containing multiple phases beginning with, detecting SSRs, computing SSRs similarities, sequences alignment, building a distance matrix based on SSRs similarity and sequences alignments, and then hierarchical clustering, and presenting the findings in a dendrogram. Sometimes sequence alignments alone may not give adequate information to generate a phylogenetic tree to resolve complicated species relationships. Therefore, establishing a distance matrix that is formed of addition of SSRs similarity across sequences to the traditional sequence alignment helps the process substantially and resolves the connections of complex species on phylogenetic trees. Additionally, it may be hard to distinguish complex relationships across species when studying conserved sequences, which could lead to an incomplete representation of their evolutionary relationships. These limitations are addressed by DendroSSR, which offers a technique to produce phylogenetic trees by incorporating SSRs similarity across species into the approach of generating phylogenetic trees. As it is known, SSRs are extensively scattered across the genomes of species and exhibit a great variation. Therefore, SSRs may support the knowledge gathered from sequence alignments by providing more information on genetic variation and even evolutionary relationships. The use of DendroSSR analysis might be considered for creating phylogenetic trees as a complementary or secondary strategy among the species under examination in circumstances where traditional phylogenetic analysis fails to clarify the species complex phylogenetic relationships.

Keywords: DendroSSR, simple sequence repeats (SSRs), phylogenetic trees, *Aspergillus* species, UPGMA.

Resumo

Este estudo apresenta um novo método para construir árvores filogenéticas, combinando tanto as Sequências Repetitivas Simples (SSRs) quanto os alinhamentos de sequência. O objetivo deste trabalho é apresentar o programa DendroSSR por meio de um estudo de caso envolvendo diversas espécies de *Aspergillus*. Para mostrar como o programa DendroSSR funciona a fim de resolver relações complicadas de espécies em árvores filogenéticas, empregamos a espécie *Aspergillus* como exemplo de caso de pesquisa. O DendroSSR utiliza uma técnica contendo várias fases, começando com detecção de SSRs, computação de similaridades de SSRs, alinhamento de sequências, construção de uma matriz de distância baseada na similaridade de SSRs e alinhamentos de sequências e, em seguida, agrupamento hierárquico e apresentação das descobertas em um dendrograma. Às vezes, os alinhamentos de sequência sozinhos podem não fornecer informações adequadas para gerar uma árvore filogenética a fim de resolver relações complicadas de espécies. Portanto, estabelecer uma matriz de distância, que é formada pela adição de similaridade de SSRs entre sequências ao alinhamento de sequência tradicional, ajuda substancialmente o processo e resolve as conexões de espécies complexas em árvores filogenéticas. Além disso, pode ser difícil distinguir relações complexas entre espécies ao estudar sequências conservadas, o que pode levar a uma representação incompleta de suas relações evolutivas. Essas limitações são abordadas pelo DendroSSR, que oferece uma técnica para produzir árvores filogenéticas ao incorporar a similaridade de SSRs entre as espécies na abordagem de geração de árvores filogenéticas. Como se sabe, os SSRs estão amplamente dispersos nos genomas das espécies e exibem grande variação. Portanto, os SSRs podem apoiar o conhecimento obtido a partir de alinhamentos de sequências, fornecendo mais informações sobre variação genética e até mesmo relações evolutivas. O uso da análise DendroSSR pode ser considerado para a criação de árvores filogenéticas como uma estratégia complementar ou secundária entre as espécies sob exame em circunstâncias em que a análise filogenética tradicional falha em esclarecer as complexas relações filogenéticas das espécies.

Palavras-chave: DendroSSR, sequências repetitivas simples (SSRs), árvores filogenéticas, espécies de *Aspergillus*, UPGMA.

*e-mail: m.hawatmeh@ttu.edu.jo

Received: June 4, 2023 – Accepted: August 6, 2023

1. Introduction

Phylogenetic trees are utilized in various ways across multiple subfields of biological research, including but not limited to taxonomy, molecular biology, and ecology (Townsend et al., 2012). The presence of these trees is essential for understanding the ancestral relationships among different genera species. Computational phylogeny approaches are used to build phylogenetic trees from a large number of DNA sequences. Distance-matrix approaches like neighbor-joining or UPGMA (Unweighted Pair Group Method with Arithmetic Mean), that utilize multi-sequences alignment to build distance matrix, are the easiest to use but don't use an evolutionary model (Felsenstein, 2004). A number of approaches for sequences alignment, like ClustalW, also build phylogenetic trees by applying the easier techniques that are based on distance (Felsenstein, 2004). The Maximum parsimony is another simple method to estimate evolutionary trees, but it assumes a model of evolution. When estimating an evolutionary tree, more sophisticated approaches use the measure of maximum likelihood such as a Bayesian approach system (Felsenstein, 2004).

Despite the fact that phylogenetic trees made from mapped genes or genomic sequences from various species can infer valuable information about evolution, these studies aren't perfect and need to be improved. Furthermore, the trees they make are not always right. They do not always show how the groups they include have changed over time. As can be with any science result, they can be proven wrong with more research (for example, by getting more data or studying the data we already have with better tools). The data that they depend on may be unclear; studies can be messed up by genetic recombination (Townsend et al., 2012), horizontal gene transfer(Arenas and Posada, 2010), and hybridization among species that were not closest to one another on the tree before hybridization, convergent evolution, and conserved sequences (Woese, 2002).

Regarding the process of generating phylogenetic trees, conventional computational methodologies primarily depend on sequence alignments to compute estimated genetic distances and concluded evolutionary relationships among diverse organisms. Alternatively, these methodologies may exhibit certain limitations is due to factors such as genetic recombination, horizontal gene transfer, hybridization, convergent evolution, or conserved sequences (Townsend et al., 2012; Arenas and Posada, 2010; Woese, 2002; Parhi et al., 2019; Felsenstein, 2004).

This study presents the software program DendroSSR (2023) as a tool for constructing phylogenetic trees. DendroSSR relies on both sequence alignments and Simple Sequence Repeats (SSRs) in its methodology. The SSRs have proven to be of significant value in the study of genetic variety and the processes of evolution (Ellegren, 2004), due to the high level of variability they exhibit as well as the large quantity of which they are composed (Townsend et al., 2012; Arenas and Posada, 2010; Woese, 2002; Parhi et al., 2019; Felsenstein, 2004; Geneious, 2022). The DendroSSR performs a new way to infer phylogenetic species relationships in comparison to the most common software such as MEGA or PAUP* which mainly uses sequence alignments only (Kumar et al., 2018; Swofford, 2002). Here where DendroSSR program comes to add an importance to the traditional phylogenetic analysis by incorporating SSRs' similarities that might help resolving complex species relationships on phylogenetic trees when traditional phylogenetic analyses that are based on sequence alignments fail.

This is accomplished by including SSRs into phylogenetic analysis in addition to traditional sequences alignments. The SSR identification, computation of SSR similarity, sequence alignment, generation of distance matrix based on SSR similarity and alignment distances, hierarchical clustering, and dendrogram visualization are all included in the DendroSSR program as an implementation of a systematic approach (Kofler et al., 2007; Python Software Foundation, 2021; McKinney, 2017; Needleman and Wunsch, 1970; Real and Vargas, 1996; Cock et al., 2009; Virtanen et al., 2020; Hunter, 2007; Ward Junior, 1963). This strategy offers a beneficial method to standard methods of building trees, which are dependent primarily on sequence alignments in order to generate phylogenetic trees.

We give in this study a detailed case study, and this study used several *Aspergillus* species for demonstrating the capabilities of DendroSSR in resolving species complex relationships in phylogenetic studies. The study highlights the ability of this software to resolve confusing places within phylogenetic trees and reveal the evolutionary relationships across taxa.

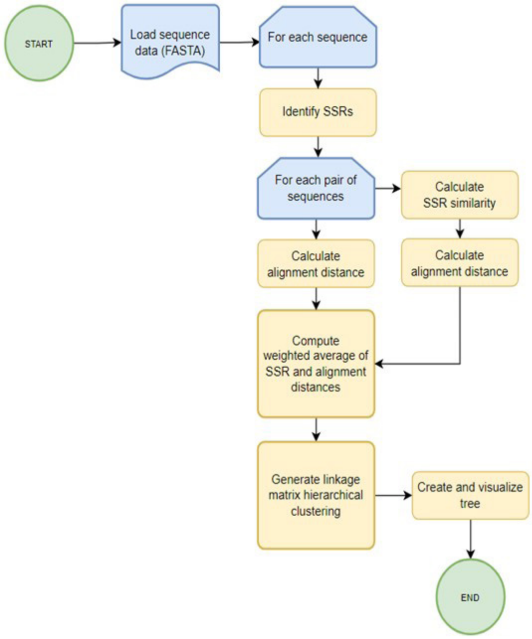
Therefore, the goal of this study is to introduce the tool **DendroSSR** which uses a new approach to build phylogenetic trees and demonstrate it by using this program in a case study including several *Aspergillus* species.

2. Materials and Methods

In this section, we provide a methodology and workflow of DendroSSR (Flowchart 1), which includes the input of sequence data, SSR identification, the calculation of SSR similarity, sequence alignment, distance matrix computation, hierarchical clustering, and the visualization of dendrograms. We applied the DendroSSR program to analyze the phylogenetic relationships among 12 *Aspergillus* species, and the phylogenetic analysis obtained by DendroSSR were then compared to results to those obtained from a traditional sequence alignment analysis.

2.1. Sequences retrieval from the GenBank

The 13 sequences were retrieved from the GenBank database (NCBI, 2023). The selected sequences for the phylogenetic analysis were collected from the GeneBank database, including the internal transcribed spacer (*ITS*) sequences of *Aspergillus* species and their corresponding accession numbers as follows: *A. fumigatus* (AB185273.1), *A. niger* (AJ853742.1), *A. terreus* (FR837963.1), *A. fischeri* (NR_137479.1), *A. nidulans* (EF567977.1), *A. oryzae* (JX878609.1), *A. flavus* (KM977885.1), *A. clavatus* (KX022485.1), *A. costaricensis* (KU945902.1), *A. luchuensis* (NR_135449.1), *A. steynii* (KP281455.1), and *A. campestris* (NR_135396.1). Additionally, *Pythium sp*. was included as an outgroup with the accession number AB095042.1. We chose 12 sequences from 42 *Aspergillus* species recorded to simplify our study. Using all 42 would make study


START
Load sequence data (FASTA)
For each sequence
Identify SSRs
For each pair of sequences
Calculate SSR similarity
Calculate alignment distance
Calculate alignment distance
Compute weighted average of SSR and alignment distances
Generate linkage matrix hierarchical clustering
Create and visualize tree
END

Flowchart 1. Workflow for DendroSSR.

confusing. In addition, these 12 sequences demonstrate how well DendroSSR handles DNA sequence variation and species evolution. These 12 sequences explain DendroSSR compared with UPGMA.

2.2. UPGMA tree construction

After that, the sequences were aligned with the help of a standardized set default that was already set into the Geneious program (Geneious, 2022). The UPGMA in the Geneious software was used to create a phylogenetic tree of *Aspergillus* species (Geneious, 2022). This tree was compared with the tree that was generated by DendoSSR program.

2.3. Data input of DendroSSR

The FASTA file format is the only format accepted and it can be used to submit sequence data to DendroSSR program. With the graphical user interface, users have the ability to download sequence files, and the software will then read and make processing for the sequences that file has. The software will read the labels that relate to sequences and consider them. For this study, we collected ITS sequences of 12 *Aspergillus* species in FASTA format and downloaded them to the program.

2.4. SSR identification

Regular expressions are utilized by DendroSSR in order to determine the SSRs that are present in each sequence. Identified SSRs are saved in a list together with the positions where they began and ended as well as their length were determined (Kofler et al., 2007; Python Software Foundation, 2021; McKinney, 2017). In this study, we identified the SSRs for each of the 12 *Aspergillus* species using DendroSSR software.

2.5. SSR similarity

In order for the software to determine the degree of similarity between two sequences on the basis of their SSRs, it must first compute the intersection between the sets of SSRs contained in each sequence. The DendroSSR used, The Jaccard index, which is the size of the intersection divided by the size of the union of the SSR sets, is then used to compute the SSR similarity. This is done after the Jaccard index has been calculated. We calculated the SSR similarity for all pairs of *Aspergillus* species (Needleman and Wunsch, 1970; Real and Vargas, 1996; Python Software Foundation, 2021).

2.6. Sequence alignment

Using the BioPython pairwise 2 modules, the DendroSSR program computes an alignment score for each pair of sequences that are being compared. We performed global sequence alignments for all *Aspergillus* species pairs (Needleman and Wunsch, 1970; Real and Vargas, 1996; Cock et al., 2009: Python Software Foundation, 2021).

2.7. Distance matrix

The DendroSSR computes a distance matrix for all sequence pairs based on a weighted average of SSR distance (1 - SSR similarity) and traditional alignment distance.

For the 12 *Aspergillus* species, we first calculated the SSR similarity between each pair of species, as described in the SSR Similarity section. Then, DendroSSR program computed the alignment distances for each pair of species using the global sequence alignment performed with the Needleman-Wunsch algorithm (Python Software Foundation, 2021; McKinney, 2017; Needleman and Wunsch, 1970; Real and Vargas, 1996; Cock et al., 2009). After both SSR similarities and traditional alignment distances calculated. This program then created a distance matrix for sequences analyzed. As a demonstration, we built a distance matrix based on both calculated SSR similarities and traditional alignment distances for *Aspergillus* species sequences (Python Software Foundation, 2021; McKinney, 2017; Needleman and Wunsch, 1970; Real and Vargas, 1996; Cock et al., 2009)

2.8. Hierarchical clustering

The hierarchical clustering found in the SciPy library of python is used side by side with the Ward method in order to get the distance matrix into a linkage matrix. After that, the linkage matrix is used to construct a dendrogram, the matplotlib library of python is implemented to show the tree (Virtanen et al., 2020; Hunter, 2007). In order to construct the phylogenetic tree, we applied hierarchical clustering to the dataset of ITS sequences of Aspergillus species (Ward Junior, 1963; Python Software Foundation, 2021).

2.9. Dendrogram visualization

The resulted tree of high resolution has been showed using the matplotlib library. The resulted tree can be saved as an image file on personal computer by users (Virtanen et al., 2020; Hunter, 2007; Python Software Foundation, 2021). The generated DendroSSR tree of *Aspergillus* species sequences was generated and saved.

2.10. Comparison with traditional sequence alignment analysis

To evaluate the performance of DendroSSR, we compared its results with those obtained from a traditional sequence alignment analysis of UPGMA method using the same dataset of 12 *Aspergillus* species.

3. Results

3.1. The UPGMA phylogenetic tree

The UPGMA phylogenetic tree (Figure 1) was created after a set multiple sequence alignment in the Geneious software had performed (Geneious, 2022), The results grouped the *Aspergillus* species as follows:

Group 1: (*A. nidulans* as a separate branch)
Group 2: (*A. steynii* as a separate branch*)*
Group 3: (*A. oryzae*, *A. flavus*)
Group 4: (*A. niger*, *A. costaricensis*, *A. luchuensis*, *A. terreus*. Where *A. luchuensis*, *A. costaricensis* as a sister clade)
Group 5: (*A. fumigatus, A. fischeri, A. clavatus, and A. campestris*. Where *A. fumigatus, A. fischeri* as a sister clade)
Outgroup: (Pythium)

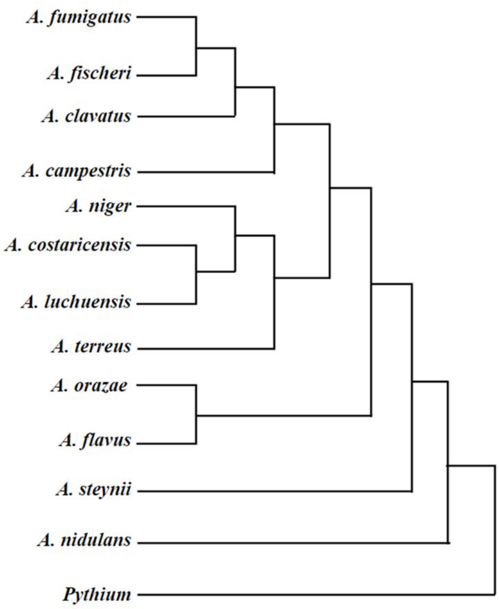


Figure 1. Phylogenetic tree of ITS sequences of *Aspergillus* species constructed using UPGMA method: traditional sequence alignment approach.

3.2. The DendroSSR phylogenetic tree

The DendroSSR program, based on sequence alignments and SSRs (Simple Sequence Repeats), was used for analysis and to build a phylogenetic tree (Figure 2). The results grouped the *Aspergillus* species as follows:

Group 1: (*A. fumigatus*, *A. fischeri*)
Group 2: (*A. terreus*, *A. oryzae*, *A. flavus*. Where *A. oryzae, A. flavus as* a sister clade)
Group 3: (*A. steynii*, *A. campestris*)
Group 4: (*A. niger*, *A. luchuensis*, *A. costaricensis*. Where *A. niger, A. luchuensis* as a sister clade)
Group 5: (*A. nidulans*, *A. clavatus*)
Outgroup: (*Pythium*)

3.3. Comparison of the DendroSSR tree with the UPGMA tree

Comparison of the DendroSSR tree (Figure 2) with the UPGMA tree (Figure 1) are summarized as following:

Both trees place *A. fumigatus* and *A. fischeri* in the same group. Unlike the UPGMA tree, the DendroSSR tree classifies *A. steynii* and *A. campestris* as sister species. Therefore, the DendroSSR method provides a more faithful portrayal of the true connections between species.

DendroSSR's greater resolution over the UPGMA approach is shown by its ability to be more clearly depict the relationships among *A. terreus*, *A. oryzae*, and *A. flavus*. With this addition, the DendroSSR may show more resolution to solve complex connections across species.

Both trees agree on how to classify *A. niger*, *A. luchuensis*, and *A. costaricensis*, hence their relationships are consistent, and they are strongly related. The UPGMA approach identifies genetic similarities and common evolutionary

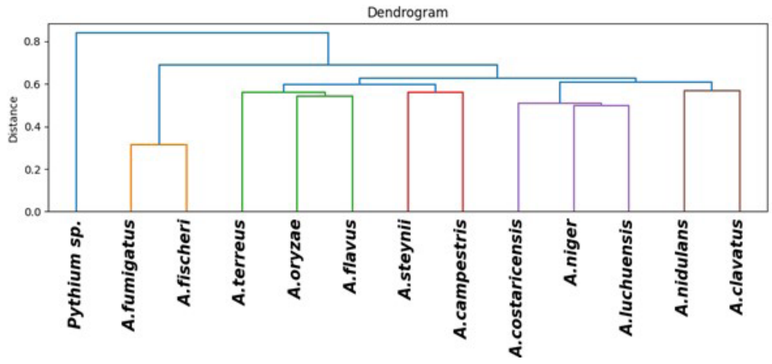


Figure 2. Phylogenetic Analysis: DendroSSR-generated Tree of ITS sequences of *Aspergillus* Species with Integrated SSRs and Sequence Alignments.

history, and the DendroSSR tree maintains this information, demonstrating its reliability and consistency.

Interesting new clustering: *A. nidulans* and *A. clavatus* are now grouped together in the DendroSSR tree, although they were previously separated in the UPGMA tree. This indicates that DendroSSR may be better able to uncover unexpected relationships between species.

Consistent Representation of the Outgroup: The *Pythium* outgroup is consistently represented in both trees, providing a stable baseline against which to evaluate the divergence of the Aspergillus species. Accurately interpreting species connections relies on maintaining a stable outgroup with constant representation.

These contrasts show that the DendroSSR approach is preferable than the UPGMA method for several studies because it provides a more comprehensive, precise, and consistent representation of species relationships.

4. Discussion

The present investigation employed DendroSSR to analyze a dataset of DNA sequences obtained from various *Aspergillus* species. The efficacy of DendroSSR program was tested through its proficient execution of SSR identification, SSR distance matrix computation, and alignment distance matrix generation. The software DendroSSR was utilized to carry out hierarchical clustering, leading to the production of a dendrogram that visually represents the clustering of sequences (Kofler et al., 2007; Python Software Foundation, 2021; McKinney, 2017; Needleman and Wunsch, 1970; Real and Vargas, 1996; Cock et al., 2009; Virtanen et al., 2020; Hunter, 2007; Ward Junior, 1963). The dendrogram presented a thorough depiction of the phylogenetic relationships among sequences, by utilizing both SSR content and alignment distances, thereby leading to generate phylogenetic relationships of the *Aspergillus* species. The utilization of DendroSSR analysis (as depicted in Figure 2) has enhanced the level of resolution for certain *Aspergillus* species in contrast to the UPGMA Method in this study (as illustrated in Figure 1). The taxonomic classification of certain *Aspergillus* species has been refined, resulting in a clearer grouping of *A. terries, A. oryza*e, and *A. flavus*. Additionally, *A. steynii* has been reclassified and is now more definitively grouped with *A. campestris*, indicating a more precise relationship between these species. Furthermore, *A. nidulans* and *A. clavatus* are classified as a cluster, which contrasts with the UPGMA dendrogram. The enhanced resolution observed in the DendroSSR analysis can be ascribed to the utilization of both sequence alignments and SSRs, which can show more resolution into the phylogenetic associations among species. However, the differences in the underlying methodologies employed to construct the trees may account for the alterations in the tree topology between the UPGMA Method and DendroSSR trees (Townsend et al., 2012; Arenas and Posada, 2010; Woese, 2002; Parhi et al., 2019; Felsenstein, 2004; Geneious, 2022). Phylogenetic trees are considered as the most reliable depictions of evolutionary relationships, based on the DNA and protein sequences. In order to enhance our understanding of phylogenetic species relations, we need to advance our methods in analysis of phylogenetic trees by incorporating more informative characters. The DendroSSR uses informative characters based on both traditional sequence alignment and SSRs similarity between genera species to come out with the best tree topology to solve complex species relationships that are not resolved by trees that are generated based on traditional sequence alignment.

In this study, we used 12 species from the genus *Aspergillus*. The genus *Aspergillus* species may have undergone rapid species diversification within a short time period, creating challenges in detecting genetic or morphological differences among various species. The challenge may be intensified in cases where the method employed for constructing the tree is not optimal, particularly if specific assumptions are not valid for the given species leading to an unclear phylogenetic relationships among species on a tree (Townsend et al., 2012; Arenas and Posada, 2010; Woese, 2002; Parhi et al., 2019; Felsenstein, 2004). And this was the case with the phylogenetic analysis of *Aspergillus* species in this study (Figure 1). In order to enhance the understanding of the interrelationships among *A. steynii*, *A. nidulans*, and other *Aspergillus* species groups, it is recommended that this study explore the utilization of other phylogenetic methodologies in addition to a traditional phylogenetic analysis. DendroSSR presents a unique methodology for conducting phylogenetic analysis of DNA sequences, which places significant emphasis on the SSR content in addition to traditional sequence alignments (Kofler et al., 2007; Python Software Foundation, 2021; McKinney, 2017; Needleman and Wunsch, 1970; Real and Vargas, 1996; Cock et al., 2009; Virtanen et al., 2020; Hunter, 2007; Ward Junior, 1963). When comparing highly divergent sequences, alignment-based methods may not provide meaningful results (Townsend et al., 2012; Arenas and Posada, 2010; Woese, 2002; Parhi et al., 2019; Felsenstein, 2004; Geneious, 2022). The SSRs in phylogenetic analysis allow the study of how these repeating elements affect genome evolution, resulting in a full tree demonstrating inter-species relationships. This Python-based DendroSSR builds phylogenetic trees where SSRs and sequence alignments are used to do this (Geneious, 2022; Kofler et al., 2007; Python Software Foundation, 2021; McKinney, 2017; Needleman and Wunsch, 1970; Real and Vargas, 1996; Cock et al., 2009; Virtanen et al., 2020; Hunter, 2007; Ward Junior, 1963). This approach uses SSRs and sequence alignments to show species-related sequences' evolutionary connections. DendroSSR is particularly useful when standard phylogenetic analysis cannot distinguish species connections on the tree. Thus, DendroSSR analysis may be used to build phylogenetic trees for species when standard methods fail to clarify their complicated connections.

References

ARENAS, M. and POSADA, D., 2010. The effect of recombination on the reconstruction of ancestral sequences. *Genetics*, vol. 184, no. 4, pp. 1133-1139. http://dx.doi.org/10.1534/genetics.109.113423. PMid:20124027.

COCK, P.J., ANTAO, T., CHANG, J.T., CHAPMAN, B.A., COX, C.J., DALKE, A., FRIEDBERG, I., HAMELRYCK, T., KAUFF, F., WILCZYNSKI, B. and DE HOON, M.J., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, vol. 25, no. 11, pp. 1422-1423. http://dx.doi.org/10.1093/bioinformatics/btp163. PMid:19304878.

DENDROSSR, 2023 [viewed 4 June 2023]. *DendroSSR for PC download* [online]. Available from: https://drive.google.com/file/d/1cX6As3c_ZKg2eWTpx85pBPc-mmqer_FI/view

ELLEGREN, H., 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews. Genetics*, vol. 5, no. 6, pp. 435-445. http://dx.doi.org/10.1038/nrg1348. PMid:15153996.

FELSENSTEIN, J., 2004. *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.

GENEIOUS, 2022 [viewed 4 June 2023]. *Geneious software (version 6.1)* [online]. Available from: https://www.geneious.com

HUNTER, J.D., 2007. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90-95. http://dx.doi.org/10.1109/MCSE.2007.55.

KOFLER, R., SCHLÖTTERER, C. and LELLEY, T., 2007. SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics*, vol. 23, no. 13, pp. 1683-1685. http://dx.doi.org/10.1093/bioinformatics/btm157. PMid:17463017.

KUMAR, S., STECHER, G., LI, M., KNYAZ, C. and TAMURA, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, vol. 35, no. 6, pp. 1547-1549. http://dx.doi.org/10.1093/molbev/msy096. PMid:29722887.

MCKINNEY, W., 2017. *Python for data analysis: data wrangling with Pandas, NumPy, and IPython*. 2nd ed. Sebastopol, CA: O'Reilly Media.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION – NCBI, 2023 [viewed 4 June 2023]. *GenBank database* [online]. Available from: https://www.ncbi.nlm.nih.gov/genbank/

NEEDLEMAN, S.B. and WUNSCH, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443-453. http://dx.doi.org/10.1016/0022-2836(70)90057-4. PMid:5420325.

PARHI, J., TRIPATHY, P.S., PRIYADARSHI, H., MANDAL, S.C. and PANDEY, P.K., 2019. Diagnosis of mitogenome for robust phylogeny: a case of Cypriniformes fish group. *Gene*, vol. 713, pp. 143967. http://dx.doi.org/10.1016/j.gene.2019.143967. PMid:31279710.

PYTHON SOFTWARE FOUNDATION, 2021 [viewed 4 June 2023]. *Python language reference (version 3.9)* [online]. Available from: https://docs.python.org/3.9/reference/index.html

REAL, R. and VARGAS, J.M., 1996. The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, vol. 45, no. 3, pp. 380-385. http://dx.doi.org/10.1093/sysbio/45.3.380.

SWOFFORD, D.L., 2002. _PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4_. Sunderland: Sinauer Associates.

TOWNSEND, J.P., SU, Z. and TEKLE, Y., 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. *Systematic Biology*, vol. 61, no. 5, pp. 835-849. http://dx.doi.org/10.1093/sysbio/sys036. PMid:22389443.

VIRTANEN, P., GOMMERS, R., OLIPHANT, T.E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W., BRIGHT, J., VAN DER WALT, S.J., BRETT, M., WILSON, J., MILLMAN, K.J., MAYOROV, N., NELSON, A.R.J., JONES, E., KERN, R., LARSON, E., CAREY, C.J., POLAT, İ., FENG, Y., MOORE, E.W., VANDERPLAS, J., LAXALDE, D., PERKTOLD, J., CIMRMAN, R., HENRIKSEN, I., QUINTERO, E.A., HARRIS, C.R., ARCHIBALD, A.M., RIBEIRO, A.H., PEDREGOSA, F., VAN MULBREGT, P., VIJAYKUMAR, A., BARDELLI, A.P., ROTHBERG, A., HILBOLL, A., KLOECKNER, A., SCOPATZ, A., LEE, A., ROKEM, A., WOODS, C.N., FULTON, C., MASSON, C., HÄGGSTRÖM, C., FITZGERALD, C., NICHOLSON, D.A., HAGEN, D.R., PASECHNIK, D.V., OLIVETTI, E., MARTIN, E., WIESER, E., SILVA, F., LENDERS, F., WILHELM, F., YOUNG, G., PRICE, G.A., INGOLD, G.-L., ALLEN, G.E., LEE, G.R., AUDREN, H., PROBST, I., DIETRICH, J.P., SILTERRA, J., WEBBER, J.T., SLAVIČ, J., NOTHMAN, J., BUCHNER, J., KULICK, J., SCHÖNBERGER, J.L., DE MIRANDA CARDOSO, J.V., REIMER, J., HARRINGTON, J., RODRÍGUEZ, J.L.C., NUNEZ-IGLESIAS, J., KUCZYNSKI, J., TRITZ, K., THOMA, M., NEWVILLE, M., KÜMMERER, M., BOLINGBROKE, M., TARTRE, M., PAK, M., SMITH, N.J., NOWACZYK, N., SHEBANOV, N., PAVLYK, O., BRODTKORB, P.A., LEE, P., MCGIBBON, R.T., FELDBAUER, R., LEWIS, S., TYGIER, S., SIEVERT, S., VIGNA, S., PETERSON, S., MORE, S., PUDLIK, T., OSHIMA, T., PINGEL, T.J., ROBITAILLE, T.P., SPURA, T., JONES, T.R., CERA, T., LESLIE, T., ZITO, T., KRAUSS, T., UPADHYAY, U., HALCHENKO, Y.O. and VÁZQUEZ-BAEZA, Y., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, vol. 17, no. 3, pp. 261-272. http://dx.doi.org/10.1038/s41592-019-0686-2. PMid:32015543.

WARD JUNIOR, J.H., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236-244. http://dx.doi.org/10.1080/01621459.1963.10500845.

WOESE, C., 2002. On the evolution of cells. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 13, pp. 8742-8747. http://dx.doi.org/10.1073/pnas.132266999. PMid:12077305.