Abstract

In a world where evolution is an evergoing and rapid process, phylogenetic trees serve as a method to gain insight into the evolutionary mechanism. While phylogenetic trees can be built, they require a complicated and intensive coding process. By introducing PhyloME, an application that builds and generates phylogenetic trees, users are given the ability to create and understand evolutionary relationships between species. This application offers a multitude of Multiple Sequence Alignment methods, clustering algorithms, the capability to bootstrap, and visualize phylogenetic trees and tanglegrams. As an emerging software in the field of bioinformatics, PhyloME builds gene trees to highlight evolutionary relationships not just between species, but how mutations in the gene can dramatically alter the composition of the phylogenetic tree.

Introduction

As the natural process of change in which descendants of organisms differ from their ancestors, evolution is the driving force of life[1]. Through the preservation, replication, and expression of hereditary information in cells, deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and amino acids (Figure 1) contribute to sustaining this force[2]. DNA is the chemical composition of the genes which encode and replicate the genetic information passed through generations[2]. The theory of the central dogma states genetic information flows unidirectionally from DNA, where it is transcribed to a RNA intermediate, and translated to a protein[3]. As the precursor to proteins, amino acids contribute to the composition of peptides and proteins[4]. Upon completion of the translation process (Figure 2), an enzymatic-catalyzed change of attaching a biochemical functional group or

removing a functional group or an amino acid to a protein can occur resulting in a post-translational modification (PTM)[5]. Regardless of whether a sequence is composed of DNA, RNA, or amino acids, a gene shared by two closely related organisms should have homologous, or even identical, sequences. The sequences can be used to rank the degree of relation between organisms and can depict evolutionary relationships in a treelike pattern[6].

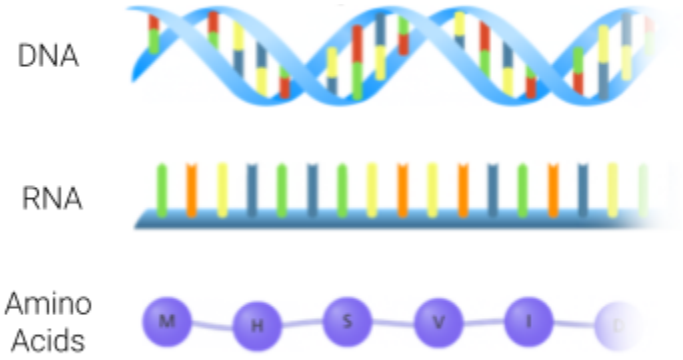

Figure 1: Showcases DNA, RNA, and amino acid molecules respectively

[Illustration of DNA, RNA, and Protein]. The LGMD2i Protein Fund. https://www.lgmd2ifund.org/science-basics/from-gene-to-protein

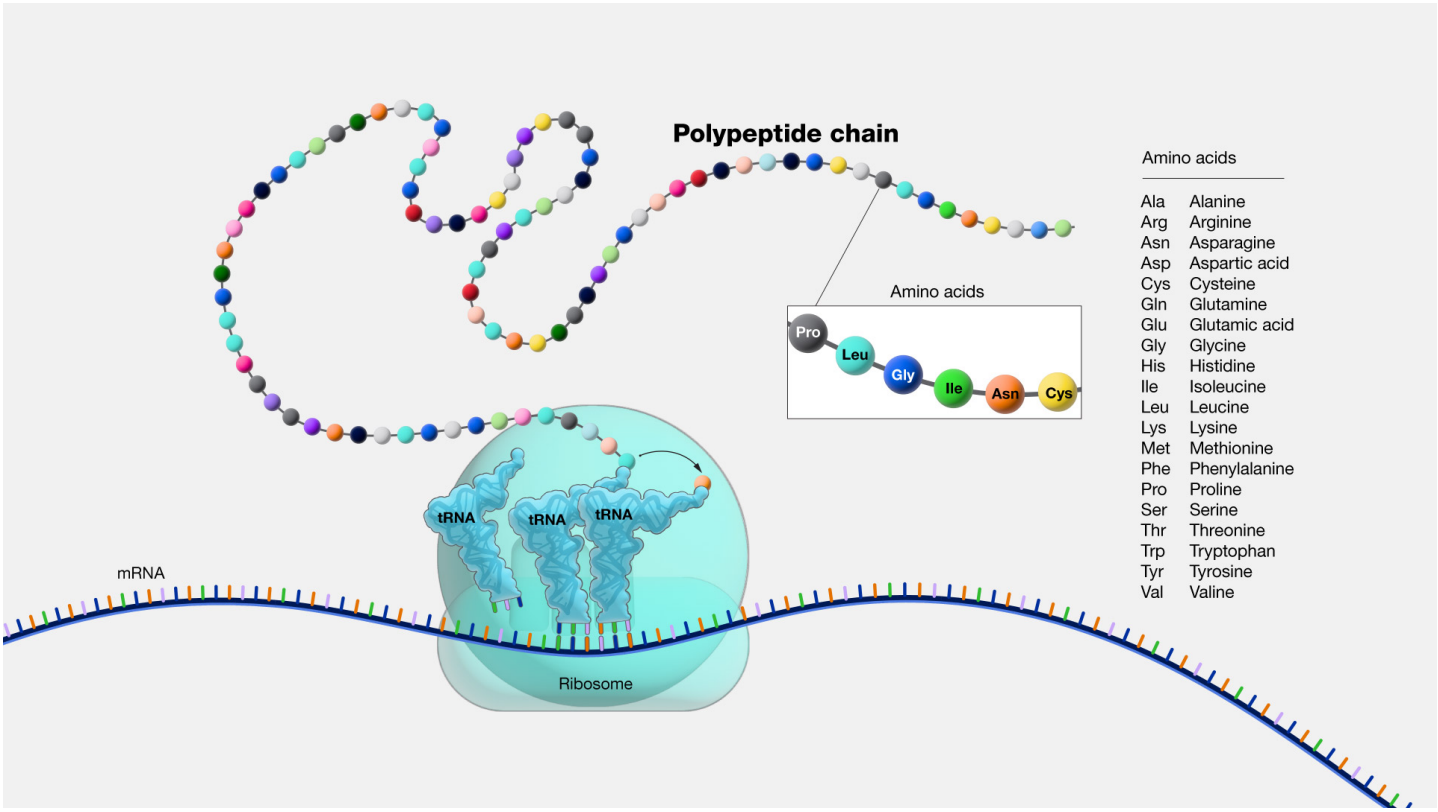

Figure 2: Exhibits the translation process from (m)RNA to amino acids to a polypeptide chain.

[Illustration of mRNA, tRNA, Ribosomes, a Polypeptide Chain, and all 20 Amino Acids]. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Peptide

Structured with dichotomously branching lines, a phylogenetic tree depicts the evolutionary relationships and distances between species differentiating ancestors and descendants proportional to the branch length[6]. While there are many branches of types of phylogenetic trees, an unrooted (Figure 3A) and rooted tree (Figure 3B) are among the simplest forms. An unrooted tree does not depict relationships by evolutionary direction due to the lack of evolutionary specification[7]. However, a rooted tree specifies the oldest common ancestor and evolutionary splits are represented chronologically. By specifying DNA, RNA, or amino acid sequences for various proteins in a FASTA format, FASTA files can be used to compose phylogenetic trees[8]. Each file is organized with a single line description of a sequence, followed by lines of sequence data[8].

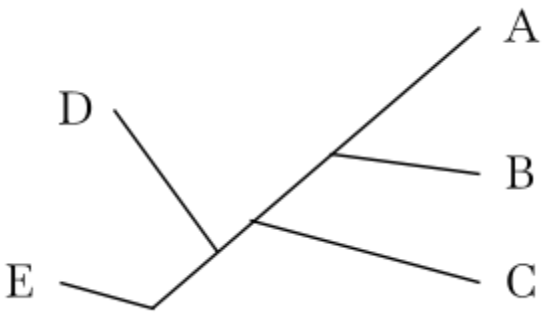

Figure 3A: An example of a rooted phylogenetic tree.

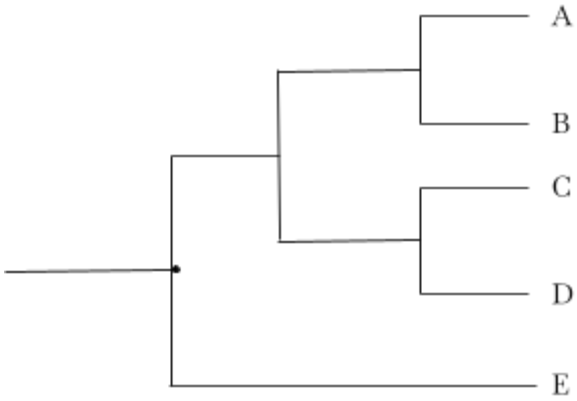

Figure 3B: An example of a rooted phylogenetic tree.

As evolution occurs and genes become more diverse, PhyloME serves to identify the evolutionary relationships between organisms relative to the gene. The initial proposal described PhyloME as an application that builds phylogenetic trees. However, as the development process began, conversations regarding the design and functionalities of

the application took precedent. It was decided that the app would have functionalities that would be considered most useful for the user and streamline the process of building the phylogenetic tree. By taking this into account, the application was built to automatically detect the sequence type and provide a multiple sequence alignment algorithm accordingly, allow users to select their clustering algorithm and rooting style, decide whether users would like to confirm the accuracy of the tree through bootstrapping, and finally select how they would like their tree to display and download their file of choice.

Prior to testing the application, research was conducted regarding which gene to use. The tumor suppressor gene *TP53* was selected as it is known for being one of the most frequently somatically altered genes in human cancer. The alteration in the *TP53* genes result in mutated forms of the p53 protein thus impacting the function of these proteins[9]. Because *TP53* is known for its mutations, we hypothesized the unmodified form and the PTM form would generate completely different trees.

An individual immersing themselves into the world of bioinformatics for the first time would greatly benefit from PhyloME. Its step-by-step process guides the user through the behind-the-scenes construction of the phylogenetic tree, while displaying highly advanced results regarding evolutionary relationships. In academia, the application can be used to explain how the relationships between organisms change as their genes evolve. PhyloME can be used to revolutionize the drug development industry, especially considering its ability to generate and compare two trees. For example, if a gene and a modified version of the gene are available, PhyloME can generate and compare the two

phylogenetic trees. These trees will depict how the evolutionary relationships change as the gene is modified and how drug testing can occur in organisms more closely related to humans. This provides additional insight into drug development as the organisms the drugs are tested on will have effects similar to those in humans.

Methodology

As previously mentioned, the data used to test PhyloME was the *TP53* gene due to its highly mutated variations. The unmodified *TP53* gene was downloaded from the Uniprot[10] database while information about the phosphorylated *TP53* gene was downloaded from the PhosphoSitePlus[11] database. The Uniprot and PhosphoSitePlus databases provide an immense amount of data on proteins. Uniprot contains protein sequence and functional information in addition to large amounts of information regarding the biological function of proteins derived from research[10]. Meanwhile, PhosphositePlus provides information and tools for the study of protein PTMs including phosphorylation, acetylation, and others[11].

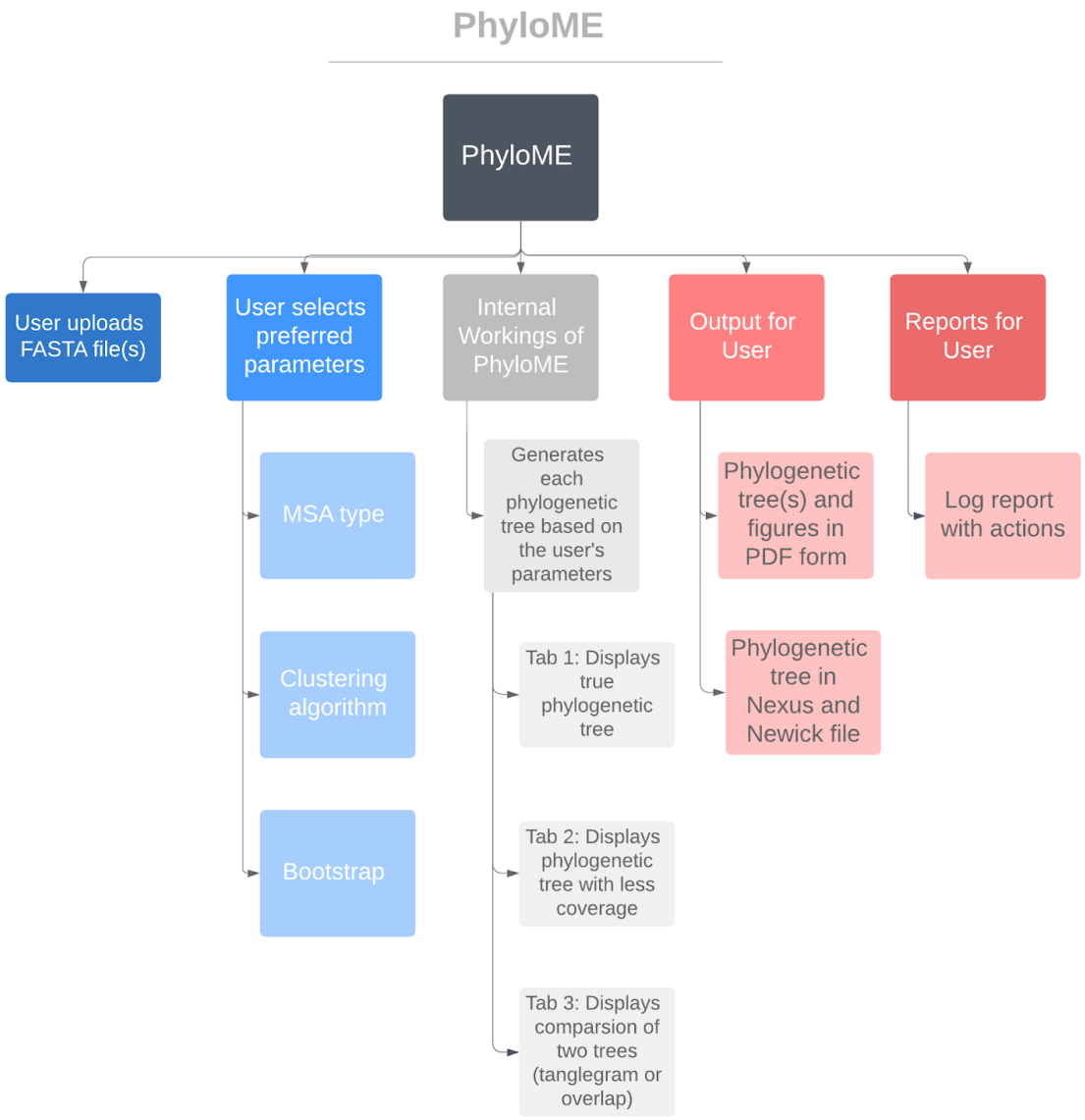

Figure 4: Displays each step in the PhyloME pipeline.

The application uses a series of algorithms to generate its output with its overall process shown in Figure 4. After uploading the FASTA file, the user will choose their Multiple Sequence Alignment method. PhyloME is designed to detect the sequence types

automatically and will present the designated methods accordingly. For DNA/RNA sequences, users will choose one of Muscle, Clustal W - gonnet, Clustal W - id, and Clustal Omega - Gonnet. For amino acid sequences, users will be presented with Muscle, Clustal W - blosum, Clustal W - pam, Clustal Omega - BLOSUM30, Clustal Omega - BLOSUM40, Clustal Omega - BLOSUM50, Clustal Omega - BLOSUM65, Clustal Omega - BLOSUM80, and Clustal Omega - Gonnet.

Upon completion of the Multiple Sequence Alignment, the tree is constructed after users choose their preferred clustering algorithm. Users have the option of creating a Neighbour Joining or Maximum Likelihood. Once the tree is generated, users can choose to root their tree, with either midpoint or outgroup rooting, or leave the tree unrooted. In the final step, users can choose to confirm the accuracy of their tree by bootstrapping it.

Users have the option of creating a second tree following the same outline described previously. Each tree is created and displayed in an individual tab. The third tab is used for visualizing and comparing the generated tree(s). Users have the option of displaying their trees individually or as a tanglegram.

It is important to note all softwares and packages required for this project are open source. The list of R packages includes ape, Biostrings, dendextend, dipsaus, dplyr, forcats, ggplot2, ggtree, grid, lubridate, maps, msa, phangorn, phylogram, phytools, purrr, readr, seqinr, shiny, shinythemes, shinyWidgets, spiralize, stringr, taxize, tidyr, tidyverse. These packages are cited in the references section.

Results

PhyloME was used to demonstrate the differences between a gene tree with the unmodified *TP53* protein (figure 5) and PTM *TP53* protein (figure 6). Despite being the same gene, there is a significant distinction between the modified and PTM gene tree. The unmodified gene tree exhibited a close relationship between a mouse and human and showed the African Green Monkey, European rabbit, and rat had a shared ancestry. On the other hand, the PTM gene tree grouped the African Green Monkey, human, and European rabbit together with African Green Monkey and human being more recently related. The mouse and rat are also grouped together. The PTM gene tree also provides a more detailed insight into the organisms as some organisms (human and mouse) show isoforms.

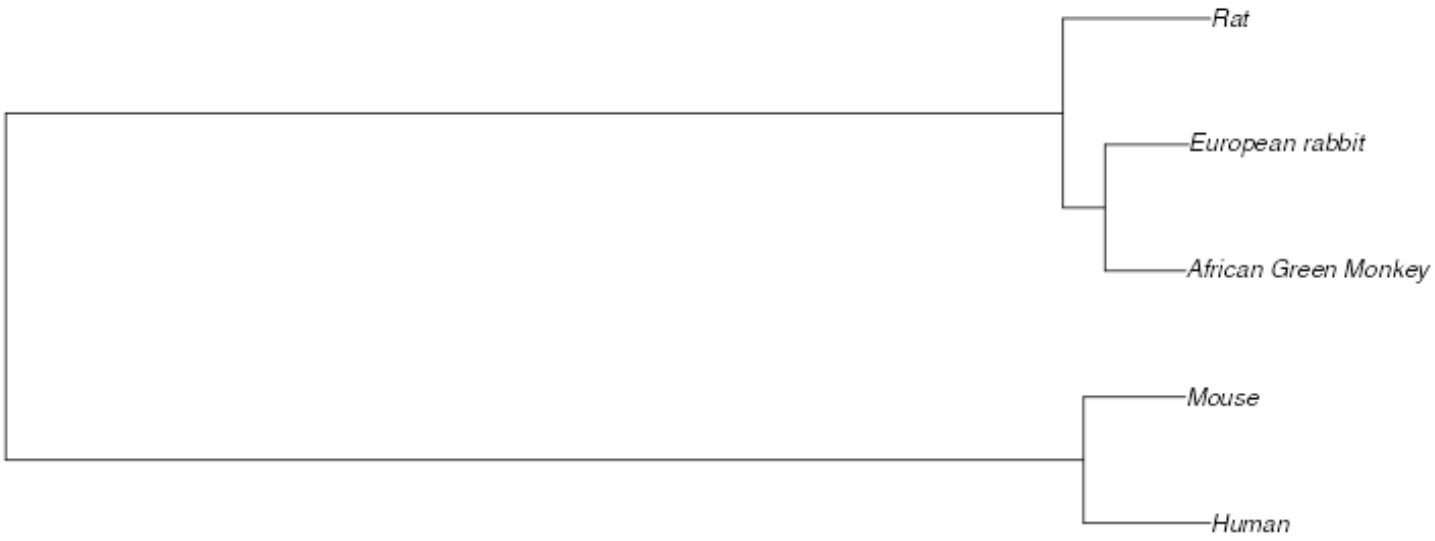

Figure 5: Unmodified *TP53* gene tree generated by PhyloME.

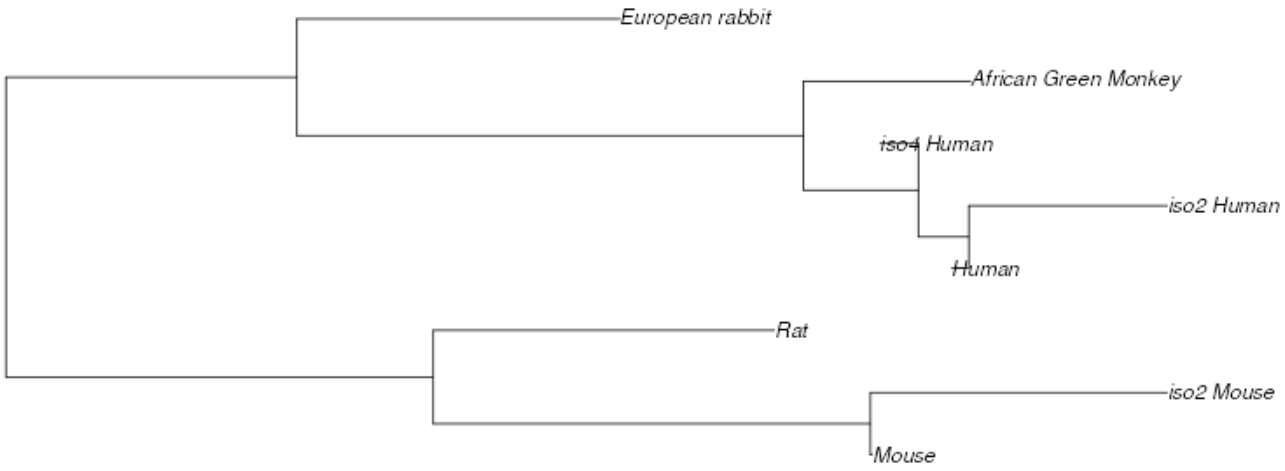

Figure 6: PTM *TP53* gene tree generated by PhyloME.

Discussion

PhyloME serves as an application that depicts phylogenetic trees from genomic, transcriptomic, and proteomic level data. While the genomic and transcriptomic level provide a significant degree of analysis, the proteomic level of data offers an incomparable level of insight through PTMs. At times, there is a significant disparity between the evolutionary relationships demonstrated on an unmodified and PTM gene tree. The unmodified and PTM gene trees of the *TP53* gene showed this vast range of difference. First of all, every organism present was more closely related to a different organism on the PTM tree. This analysis shows the *TP53* gene evolved differently and even gave rise to certain isoforms. The significance of this result will be of use especially in the drug development aspect of industry. If a drug for the PTM *TP53* gene is tested

without taking the PTM tree into consideration, the developed drug would be tested in mice (mice are more closely related according to the unmodified *TP53* gene). However, the PTM tree reveals the *TP53* gene in humans is more closely related to the one in the African Green Monkey. Therefore, it would be best to test the drug on the African Green Monkey. Similarly, PhyloME can be used to display the evolutionary relationships using other genes and provide the basis for which organisms should be used when drug testing.

While PhyloME can generate two phylogenetic trees concurrently, the application is capable of generating just one phylogenetic tree as well. This tool is useful when studying diseases, such as viruses, and understanding how they have evolved. A prime example of this is regarding the spike protein of the SARS-CoV-2 virus. Due to the rapid evolution rate of the virus, the spike protein changes rapidly as well. By using PhyloME, the user will be able to understand how closely related each strain is and gain an understanding of how the spike protein is evolving.

Future Prospects

The future of PhyloME revolves around introducing features to make the application completely unique from competing resources. The first and most important feature will be the ability to connect to the Uniprot[10] and PhosphositePlus[11] databases directly. With this feature, users will not have to independently download their FASTA files and instead will be able to search directly in the application. This will be a completely unique feature, not yet available in similar applications, and will increase the accessibility for users. The next feature to be introduced will be the ability to edit the trees. With this

feature, users will be able to edit the names of organisms to refer to the organism with the local name. This will be especially useful when sharing the generated tree with individuals who are not from a scientific background. Lastly, a molecular clock function will be added to show when a divergence between two species occurred. The molecular clock would be measured in millions of years to accurately display the evolutionary relationships. This feature is another that is not available on any other application and will provide users the ability to analyze when divergences occurred.

Conclusion

Revolutionizing phylogenetic trees means revolutionizing the industry. PhyloME has made that revolution possible by providing access to visualizing evolutionary relationships through a multitude of methods. Whether the user prefers an unrooted or rooted tree, wants to verify the accuracy of their tree through bootstrapping, or to simply understand what a phylogenetic tree is, PhyloME has the capability to handle whichever features the user chooses. Through the various options available to generate a unique phylogenetic tree with just one FASTA file. The additional option to compare phylogenetic trees serves the user beyond what a traditional application can achieve, allowing for the immediate analysis of results. With the introduction of PhyloME, the world of phylogenetics can serve as a catalyst for drug development, immunology and virology, and many more. PhyloME is an exceptional tool transforming the understanding of evolution one tree at a time.

References

1. Futumya, J.D. Evolution. *AccessScience*. Available at: https://doi.org/10.1036/1097- 8542.475150 (2019).
2. Johnson, E., & Pardue, M. L. Nucleic acid. *AccessScience*. https://doi-org.ezproxy.langara.ca/10.1036/1097-8542.46060. (2019).
3. Central Dogma. *National Human Genome Research Institute*. https://www.genome.gov/ genetics-glossary/Central-Dogma . (2023).
4. Cummings, M. R., Klug, W. S., Killian, D. J., Palladino, M. A., & Spencer, C. A. Genomic Analysis. *Concepts of Genetics.* 12. Pearson Education Inc. Hoboken, New Jersey. (2018).
5. Yu. F., Wu, Y., & Xie, Q. Precise protein post-translational modifications modulate ABI5 activity. *Trends in Plant Science. https://doi.org/10.1016/j.tplants.2015.05.004*. 20. (2015).
6. Hine, R (Ed.). Evolutionary Tree. *A Dictionary of Biology*. Oxford University Press. (2019).
7. Hine, R (Ed.). Phylogram. *A Dictionary of Biology*. Oxford University Press. (2019).
8. FASTA. *Integrative Genomics Viewer.* Preprint at https://software.broadinstitute.org/ software/igv/FASTA (2021).
9. Liacine Bouaoun. *TP53* Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Human Mutation Variation, Informatics, and Disease*. Human Genome Variation Society. https://doi-org.ezproxy.langara.ca/10.1002/humu.23035. 37. (2016).

10. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research.* https://doi.org/10.1093/nar/gkac1052. 51. 2022.
11. Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., & Skrzypek, E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* 2015.

References for R Packages

1. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
2. Paradis E. & Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35: 526-528.
3. Pagès H, Aboyoun P, Gentleman R, DebRoy S (2022). _Biostrings: Efficient manipulation of biological strings_. R package version 2.64.1, <https://bioconductor.org/packages/ Biostrings>.
4. Guangchuang Yu. (2022). Data Integration, Manipulation and Visualization of Phylogenetic Trees (1st edition). Chapman and Hall/CRC.
5. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
6. U. Bodenhofer, E. Bonatesta, C. Horejs-Kainrath, and S. Hochreiter (2015) msa: an R package for multiple sequence alignment. Bioinformatics 31(24):3997-9999. DOI: 10.1093/bioinformatics/btv176.

7. Schliep, K., Potts, A. J., Morrison, D. A., Grimm, G. W. (2017), Intertwining phylogenetic trees and networks. Methods in Ecology and Evolution, 8: 1212--1220. DOI:10.1111/2041 -210X.12760
8. Wilkinson SP, Davy SK (2018) phylogram: an R package for phylogenetic analysis with nested lists. Journal of Open SourceSoftware 3:790. DOI: 10.21105/joss.00790
9. Revell, L. J. (2012) phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3 217-223. doi:10.1111/j.2041-210X.2011.00169.x
10. Charif, D. and Lobry, J.R. (2007). Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis.
11. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2022). _shiny: Web Application Framework for R_. R package version 1.7.2, <https://CRAN.R-project.org/package=shiny>.
12. Perrier V, Meyer F, Granjon D (2022). _shinyWidgets: Custom Inputs Widgets for Shiny_. R package version 0.7.4, <https://CRAN.R-project.org/package=shinyWidgets>.
13. Chang W (2021). _shinythemes: Themes for Shiny_. R package version 1.2.0, <https://CRAN.R-project.org/package=shinythemes>.
14. Gu, Z. (2021) spiralize: an R package for Visualizing Data on Spirals. Bioinformatics.

15. Wickham H (2022). _stringr: Simple, Consistent Wrappers for Common String Operations_. R package version 1.4.1, <https://CRAN.R-project.org/package=stringr>.
16. Scott Chamberlain and Eduard Szocs (2013). taxize - taxonomic search and retrieval in R. F1000Research, 2:191. URL: https://f1000research.com/articles/2-191/v2
17. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” _Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.