

PHYLOGENETIC TREES*

Susan Cates

This work is produced by OpenStax-CNXX and licensed under the Creative Commons Attribution License 1.0[†]

Abstract

This module introduces the student to the concept of cladograms and phylogenetic trees. It explores some online bioinformatics tools that will produce tree diagrams.

A phylogenetic tree is a graphical representation of the evolutionary relationship between taxonomic groups. The term phylogeny refers to the evolution or historical development of a plant or animal species, or even a human tribe or similar group. Taxonomy is the system of classifying plants and animals by grouping them into categories according to their similarities. A phylogenetic tree is a specific type of cladogram where the branch lengths are proportional to the predicted or hypothetical evolutionary time between organisms or sequences. Cladograms are branched diagrams, similar in appearance to family trees, that illustrate patterns of relatedness where the branch lengths are not necessarily proportional to the evolutionary time between related organisms or sequences. Bioinformaticians produce cladograms representing relationships between sequences, either DNA sequences or amino acid sequences. However, cladograms can rely on many types of data to show the relatedness of species. In addition to sequence homology information, comparative embryology, fossil records and comparative anatomy are all examples of the types of data used to classify species into phylogenetic taxa. So, it is important to understand that the cladograms generated by bioinformatics tools are primarily based on sequence data alone. Given that, it is also true that sequence relatedness can be very powerful as a predictor of the relatedness of species.

Cladograms cannot be considered completely true and accurate descriptions of the evolutionary history of organisms, because in any cladogram there are a number of possible evolutionary pathways that could produce the pattern of relatedness illustrated in the cladogram. The cladogram only illustrates the probability that two organisms, or sequences, are more closely related to each other than to a third organism, it does not necessarily clarify the pathway that created the existing relationships. However, the cladogram can be used in the formulation of new hypotheses and to cast new light on existing data. One of the most ambitious cladograms produced to date can be viewed at the Tree of Life¹ website, originated by *David and Wayne Maddison*[2] at the University of Arizona (1). Please take a moment to view the "Root of the Tree" link on the Tree of Life web site. In this phylogenetic tree, the root is at the far left, termed the root of the cladogram because it is at the base of the cladogram, opposite the branches. Return to the home page and click on the link entitled "Popular Pages", then select "Mammals". At the right side of this cladogram are the terminal nodes, located at the tip of the branches in any cladogram. In the Mammalia cladogram illustrated here, there are six terminal nodes, labeled Triconodonts, Monotremata, Multituberculata, Marsupialia, Palaeoryctoids, and Eutheria. An internal node is a hypothetical common ancestor. The branching points between the root and the terminal nodes are internal nodes. Each internal node is also at the base of a clade, which includes the common ancestral node plus all its descendants. Sample a few more links on the Tree of Life. Be sure to

*Version 2.8: Mar 1, 2006 1:01 pm -0600

[†]<http://creativecommons.org/licenses/by/1.0>

¹<http://tolweb.org/tree/phylogeny.html>

read Darwin's quote on the home page and ponder how difficult it would be to get published in a scientific journal today, if it were necessary to write this beautifully in order to succeed.

The Tree of Life is an example of a cladogram illustrating the relationships between taxa, based on the collective evidence from many different fields of biology and bioscience. In contrast, the subject of this tutorial is the construction of cladograms through bioinformatics tools, where the cladograms are based on sequence data. First, use the Biology Workbench² (2)[3] to build a simple unrooted cladogram. The Workbench will require a password (it's free), but it will grant entrance immediately upon registration of a password. Enter the site, and scroll down the page until the five menu buttons are visible. The "Session Tools" button allows the naming of a session, so that different jobs in progress can be saved under distinct sessions. Select "Session Tools", then select "Start New Session" and click on "Run" to change the name of "Default Session" to a new name. Once the workbench has been exited, the session will remain. Subsequently, clicking on the dot to the left of the session name under the "Session Tools" menu, and then selecting "Resume Session", will recall the session. The Workbench policy at the time of this writing is that old jobs are deleted only when an account has not been accessed for 6 months.

Next, select "Protein Tools" from the menu buttons, highlight "Ndjinn Multiple Database Search", and click "Run". In the query box to the right of the term "Contains", type HSP70, for the molecular chaperone, heat shock protein 70 kDa. Scroll down the database list and check the box to the left of the database entitled "PDBFINDER" before hitting the "Search" button. Among the results, find 2BUP, chaperone, and check the box to the left. Then select the menu button entitled "Import sequence(s)". This will import the sequence in fastA format into the open session. Now, under the box of session options, there should be a listing for the 2BUP sequence, with a small box to the left. Notice that the main menu under "Protein Tools" allows more options such as "Delete Protein Sequence", "Copy Protein Sequence" and "Add New Protein Sequences". For now, select the "Ndjinn Multiple Database Search" again. Search the PDBFINDER Database again by scrolling down the page and selecting it, but this time, just search using the PDB ID codes 1HKB, 1ATN and 1DKG for hexokinase, actin and the molecular chaperone DnaK (use the OR operator between each PDB ID code to search for all three in the same search). Import all three sequences simultaneously by checking the box to the left of the PDB ID codes used in the query and clicking on "Import sequence(s)". 1DKG will return three chains, A, B and D. Only chain D is the molecular chaperone, chains A and B are nucleotide exchange factors that co-crystallized with DnaK. Delete chains A and B by checking the box to the left of 1DKG_A and 1DKG_B, highlighting "Delete Protein Sequence", and clicking on "Run". Actin (1ATN) returns two chains, but chain A is the actin, chain D should be deleted in the same manner. Hexokinase (also called phosphotransferase) will return two chains as well. They are both hexokinase, but two identical sequences are not desirable in the cladogram, so delete chain B. Four sequences should remain, 1DKG_D, 1ATN_A, 1HKB_A, and 2BUP_A; check the boxes to the left of each of these. Scroll down the protein tools menu and highlight "CLUSTALW - Multiple Sequence Alignment", then click "Run". The default parameters will be sufficient for our purposes, just select "Submit". When the sequence alignment is returned, scroll down the page and view the multiple alignment. The Workbench automatically returns an unrooted tree with the alignment. Look at the unrooted tree.

Exercise 1

Which two sequences appear to be most closely related by viewing the unrooted tree?

Look at the multiple sequence alignment scoring section. Notice the sequence list that assigns numbers to each sequence. The alignment scores are labeled by the assigned sequence numbers, so this list is necessary to interpret the scores.

Exercise 2

According to the pairwise scores, which two sequences are most similar?

Exercise 3

What is the score of the best pairwise alignment?

²<http://workbench.sdsc.edu/>

Return to the top of the page and select "Import Alignment(s)". This will save this sequence alignment under the "Alignment Tools" menu of this session. Select the alignment by clicking in the small box to the left of the listing. Choose "DRAWGRAM" from the options box, to view this alignment in a rooted tree. Accept all the default values on the drawgram page and click the submit button at the bottom of the page. Drawgram will return a rooted tree using the program "*PHYLIP*"^[1], a Phylogeny Inference Package. The PHYLIP suite includes packages that can infer phylogenies by parsimony, compatibility, distance matrix methods, and maximum likelihood methods. PHYLIP can also draw several types of tree diagrams, and allows editing of trees. The suite accepts an impressive number of input formats, including nucleotide and protein sequences in fastA format. Detailed information on the PHYLIP suite can be viewed at the PHYLIP website³.

Exercise 4

Do the same two sequences appear to be the most closely related by viewing the rooted tree in comparison to the unrooted tree?

Click on the "Return" button and this will yield the "Alignment Tools" menu. It will be necessary to return to "Protein Tools" to complete this tutorial, so select the "Protein Tools" icon with the mouse. Now, to put together a little more complex cladogram, search the PDBFINDER in the Ndjinn Database for the following sequences by copying and pasting the following into the query box:

```
1ECL OR 1BGW OR 1DUB OR 1AUX OR 1KAN OR 1BPE OR 2AAC
```

Import all of these sequences into this session. Once again, there are some duplicated sequences from multimeric proteins, so delete 2AAC_B, 1KAN_B, 1AUX_B and 1DUB_B, C, D, E, and F. This should leave 11 sequences, including the first four sequences imported in the example above. On the scroll down menu under "Protein Tools", highlight "Select all sequences" and click "Run". Next, select "ClustalW" and click "Run". On the ClustalW input page, change the "unrooted" tree option to "rooted and unrooted trees", then submit.

Exercise 5

Once the results are returned, click on the option "Download a PostScript version of the output" for the rooted tree. Send the postscript file as an attachment to your lab assignment.

Exercise 6

Which two sequences have the highest number of ancestral nodes as represented in the cladogram?

Exercise 7

Which sequence has the longest branch between the terminal node and the closest ancestral node, as represented in the cladogram?

Exercise 8

In this cladogram, what is the relationship between the two sequences that scored the highest pairwise alignment in the first example?

Attempt to construct another type of tree using sequences of personal interest, without explicit instructions.

Exercise 9

Find at least six related nucleotide sequences (e.g., download the sequences for superoxide dismutase genes from six different species) and construct rooted and unrooted trees containing these sequences using the Biology Workbench⁴. Send the postscript files as attachments to your lab, and list the 6 (or more) PDB IDs for the chosen sequences, with brief descriptions (protein name).

³<http://evolution.genetics.washington.edu/phylip.html>

⁴<http://workbench.sdsc.edu/>

References

- [1] J. Felsenstein. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.*, pages 266:418–27, 1996.
- [2] Maddison DR Maddison WP. Interactive analysis of phylogeny and character evolution using the computer program macclade. *Folia Primatol (Basel)*, pages 53(1–4):190–202, 1989.
- [3] Subramaniam S. The biology workbench—a seamless database and analysis environment for the biologist. *Proteins*, pages 32(1):1–2, 1998.