Phylogeny using bioinformatics

How do scientists investigate the evolutionary relationships among plants?

A unifying characteristic of all organisms is the genetic code. Genomes are information dense structures that provide historical evidence of the relationships among living species. The field of **phylogenetics** uses genetic information (e.g., DNA sequences) to identify these evolutionary relationships. This technique is useful to classify new or unknown species as well as provide measures of genetic distance among species. A common output of phylogenetic analysis is a **phylogeny**, or **phylogenetic tree**, which provides an illustrative explanation of the relationships.

A simple phylogenetic analysis can be performed with basic **bioinformatic tools**. These tools use computers and apply statistical analyses to help compare large or complex biological data, such as genomes or multiple DNA sequences. Certain genomic regions have been found to differ across many species. Genetic differences within the genome that help distinguish species are called **barcodes** or barcoding regions.

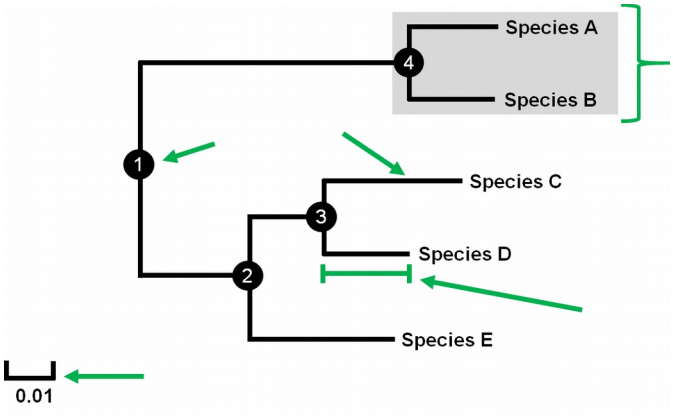
For classifying land plants, the ***rbcL* gene** is the primary barcode used in phylogenetic studies. This gene codes for Ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCo). RuBisCo is an enzyme in plants that catalyzes the critical chemical reaction of capturing carbon from CO2 for photosynthesis and downstream glucose production. RuBisCo is thought to be the most abundant protein on the planet.

This activity explores the basic tools to create and interpret phylogenetic trees from the barcode gene *rbcL*. The *rbcL* gene tree grows a new branch with each new DNA sequence, resulting in an ever-expanding database of life.

Activity 1: What are the parts of the phylogenetic trees?

Label and/or define the components of the phylogenetic tree.

Branch, Branch length, Node, Clade, and Distance Scale Bar.



Evolution and phylogeny

Questions

1. What Species is the most closely related to Species C?
    1. Species D
2. Approximately how much distance is between Species A and Species B
    1. Using the distance scale bar of 0.01, it is the sum of the branch lengths. So, Species A branch + Species B branches or 0.02 + 0.02 = 0.04 distance units.
3. Is Species C, D, and E in the same clade? Why or why not?
    1. Yes, they are connected through the common ancestor or node 3
4. Given the tree above, what two species do NOT represent a clade?
    1. Species A (or B) and Species C
    2. Species A (or B) and Species D
    3. Species A (or B) and Species E

Activity 2. Build a Phylogenetic Tree

1. Follow this link to a collection (also known as a library) of *rbcL* genes for many common crops.

https://drive.google.com/file/d/1g8oWamxyTAbLrxe4d7pN9RmcbUEHEP36/view?usp=sharing

The *rbcL* sequences are formatted in a standard form that is recognized across a broad range of bioinformatic tools on the internet. The format is called FASTA (pronounced Fast-A), which stands for Fast-All. Essentially, FASTA sequences have names and other descriptions about the sequence in the first line, beginning with a greater than (>) sign. The second line contains the nucleotide sequence. Two samples are recognized by programs with the addition of another > within the text file. For example:

> Sample1

AAACGCTTACGCG

>Sample 2

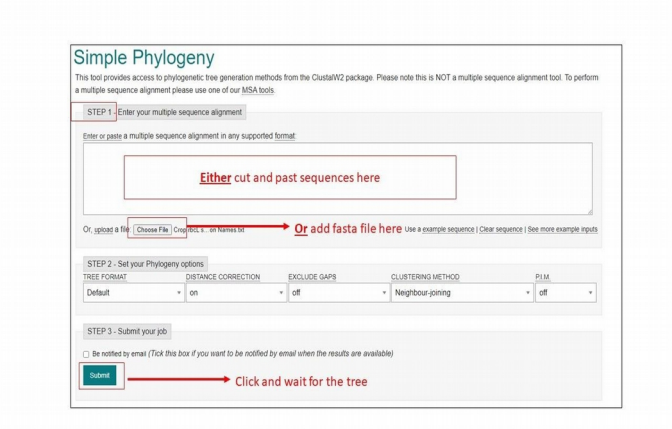
GGGACGTTACGGA

*Note the FASTA files can be read by using Word or Google Docs, but must be saved as a text (.txt) file to be used in various bioinformatic programs.

2. Go to: https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/ and follow the two steps below in the image. Load the sequences by either cutting and pasting into the textbox or uploading the file directly using the choose file option. Set the Phylogeny options (Step 2) as Default, Off, Off, Neighbour-Joining, Off.

    Interpretation note: The output will be a phylogenetic tree that shows the relationships of the plants with gene tree branches of equal lengths. The decimals in the parentheses after the plant name are the genetic distances between sequences within each node or clade.

Evolution and phylogeny



Additional (optional) tree view

- Once the tree appears in the Simple Phylogeny program, press “View phylogenetic tree file” under the tree. A new window will appear with the tree provided in the Newick format. The raw output will look strange, but this format can be recognized by other tree building programs.
- For a more interactive tree open a new window and go to: https://itol.embl.de/.
- On the front page of the Interactive Tree of Life tree building program click ‘Upload’.
- Provide a name for the tree (eg., Crop Phylogeny).
- Paste the Newick format tree into the tree text box and click the Upload button at the bottom.
- A tree will appear with a control box to explore the different ways to present the phylogenetic tree. Specifically, this analysis will apply the branch lengths calculated in the original analysis.

Questions

1. How many major clades are there in this list of sequences?
    1. 3
2. Which crop is most related to Canola?
    1. Papaya
3. What is a specialized feature shared between the clade containing lentil, peanut, and soybeans?
    1. Legumes - nodules for N2 fixation.
4. Teosinte and corn are closely related according to the phylogenetic tree. What about their history explains this relationship?
    1. Modern corn or maize was selectively bred through domesticating ancestral teosinte plants.
5. What are the branch lengths for soybean and wild soybean? What does this suggest about their *rbcL* gene sequences?
    1. Branch lengths 0; Sequences are identical
6. Can fungi be classified in the same manner as the crops above? Why?
    1. Yes, but using a different barcode gene. Anything with DNA can be analyzed in this manner.