II. Identify Species and Phylogenetic Relationships Using *DNA Subway*

The following directions explain how to use the Blue Line of *DNA Subway* to analyze novel DNA sequences generated by a DNA sequencing experiment. If you did not sequence your own DNA sample, you can follow these directions to use DNA sequences produced for other students. You can find supplementary instructions by clicking on the "manual" link on the *DNA Subway* homepage.

DNA Subway is an intuitive interface for analyzing DNA barcodes. Generally, you progress in a stepwise fashion through the button "stops" on each "branch line." An R indicates that analysis is available. A blinking *R* indicates an analysis is in process. A V means that results are ready to view.

1. Create a *DNA Subway* Project and Upload DNA Sequences
   a. Log into *DNA Subway* at www.dnasubway.org. If you do not have an account, you will need to register first to save and share your work.
   b. Select "Determine Sequence Relationships" (Blue Line) to begin a project.
   c. Select "*rbc*L" or "COI" from the "Select Project Type" section. (*rbc*L (plant) sequences must be analyzed separately from COI (animal) sequences.)
   d. "Select Sequence Source" provides several ways to obtain sequences for barcode analysis:
      - *Upload sequence(s) in ab1* (files ending with .ab1) *or FASTA format.* Click "Browse" to navigate to a folder on your desktop or drive containing your sequence(s). Select a sequence by clicking on its file name. Select more than one sequence by holding down the ctrl key while clicking file names. Once you have selected the sequences you want, click "Open".
      - *Enter a sequence in FASTA format.* Below is an example of this format. The

good quality.

- If one of the sequences seems of good quality, return to Pair Builder, and click the red x to undo the pairing.
- Continue on to Step 4.

i. Few or no internal mismatches indicate good quality sequence from forward and reverse reads. If you like, you can check the consensus sequence at yellow mismatches and override the judgment made by the software:

- Click on a highlighted mismatch to see the electropherograms and graphic summarizing Pfred scores for each read. Remember that the horizontal line equals a Phred score of 20, the cut-off for high-quality sequence.
- Click on the desired nucleotide in the black rectangle to change the consensus sequence at that position. You should only change the consensus if you have a strong reason to believe the consensus is wrong.

Changing the consensus sequence arbitrarily is likely to create a change in the sequence that does not represent the sequence in the organism.

- Click the button to "Save Change(s)."

4. BLAST Your Sequence

A BLAST search can quickly identify any close matches to your sequence in sequence databases. In this way, you can often quickly identify an unknown sample to the genus or species level. It also provides a means to add samples for a phylogenetic analysis.

a. On the *Add Sequences* branch, click on "BLASTN". Then, click on the "BLAST" button next to the sequence you want to query against DNA databases.

b. The returned list has information about the 20 most significant alignments (hits):

- Accession number, a unique identifier given to each sequence submitted to a database. Prefixes indicate the database name – including gb (GenBank), emb (European Molecular Biology Laboratory), and dbj (DNA Databank of Japan).
- Organism and sequence description or gene name of the hit. Click on the genus and species name for a link to an image of the organism, with additional links to detailed descriptions at Wikipedia and Encyclopedia of Life (EOL).
- Several statistics shown in the window allow comparison of hits across different searches. The number of mismatches over the length of the alignment gives a rough idea of how closely two sequences match. The bit score formula takes into account gaps in the sequence; the higher the score the better the alignment. The Expectation or E value is the number of alignments with the query sequence that would be expected to occur by chance in the database. The lower the E value, the higher the probability that the hit is related to the query. For example, an E value of 1 means that a search with your sequence would be expected to turn up 1 match by chance. Why do the most significant hits typically have E values of 0? (This is not the case with BLAST searches with primers.) What does it mean when there

are multiple BLAST hits with similar E values?

- Add BLAST sequence data to your phylogenetic analysis by checking the box(es) above any accession number(s), then clicking on "Add BLAST hits to project" at the bottom of the BLAST results window.

5. Add Sequences to Your Analysis

   a. Click on "Upload Data" to include additional data. Either upload data in abl or FASTA format or import data from other sources.

   b. Click on "Reference Data" to select data that will let you compare your barcode sequence in an appropriate phylogenetic context.

If you have a good idea of the taxonomy of your sample, you may want to select Reference Data from a narrow range of plants or animals including the putative family your sample is from. If you have little idea of the taxonomy of your sample, include a very broad selection of Reference Data.

6. Analyze Sequences: Select and Align

Many unknown species can be rapidly identified by a BLAST search. In this case, a phylogentic analysis adds depth to your understanding by showing how your sequence fits into a broader taxonomy of living things. If your BLAST search fails to identify your sequence, phylogenetic analysis can usually identify it to at least the family level.

   a. Click on "Select Data" on the "Analyze Sequences" branch. Then check boxes to select any or all of the sequences you have uploaded from your own sequencing projects, from BLAST searches, and from reference data sets. Click on Save.

MUSCLE is a multiple sequence alignment program, like CLUSTALW, which aligns two or more sequences in a manner that produces the fewest gaps. Jalview is a Java utility for viewing and editing the alignments produced by Muscle. Jalview also calculates and displays phylogenetic trees.

   b. Click on "MUSCLE" to align your sequences. When the program is finished, click again to view the alignment in Jalview.

      - Scroll through your alignments to see similarities between sequences. Nucleotides are color coded, and each row of nucleotides is the sequence of a single organism or sequencing reaction. Columns are matches (or mismatches) at a single nucleotide position across all sequences. Dashes (-) are gaps in sequence, where nucleotides in one sequence are not represented in other sequences.

      - Note that the 5' (leftmost) and 3' (rightmost) ends of the sequences are usually misaligned, due to gaps (-) or undetermined nucleotides (Ns). What causes these problems?

      - Note any sequence that introduces large, internal gaps (-----) in the alignment. This is either poor quality or unrelated sequence that should be excluded from the analysis. To remove it, return to "Select Data," uncheck that sequence, and save your change. Then click on "MUSCLE" to recalculate.

   c. Trim Unaligned Ends of the Sequences

      - Identify the leftmost point at which all or most sequences show corresponding nucleotide color bars. (There should be few or no gaps in the vertical column of nucleotides at this point.)

      - Click in the nucleotide coordinate bar directly above this nucleotide in the first sequence. This will activate a red cursor and a pop-up menu.

      - Click on "Remove left" to trim the leftmost sequences to this nucleotide position.

- Repeat first two steps of 6.c. above, and click "Remove right" to trim the rightmost sequences.
- You can return to "Select Data" (in step b. above) to remove any sequence that has large sequence gaps. Why is it important to remove sequence gaps and unaligned ends?
- Click "Submit trimmed alignment."

7. Analyze Sequences: Create a Phylogenetic Tree

Tree-building algorithms attempt to reconstruct the order in which sequence mutations accumulated as different lineages diverged from a common ancestor. A number of plausible trees can be constructed from any set of sequences, so an algorithm presents what it determines to be the optimal one. The maximum likelihood algorithm evaluates possible trees and determines which is mostly likely to have been produced by the observed data. Because it fits mutations to a tree, the maximum likelihood method produces the most parsimonious tree – one that accounts for the data with the shortest branch lengths.

The tree visualization software may assign a numerical value to each branch, which is proportional to its length.

a. Click on "PHYLIP ML" to generate a phylogenetic tree using the maximum likelihood method. A tree will open in a new window; and the MUSCLE alignment used to produce it will open in another window.

b. A phylogenetic tree is a graphical representation of relationships between taxonomic groups. In this experiment, a *gene tree* is determined by analyzing the similarities and differences in DNA sequence.

c. Look at your tree.
   - The branch tips are the DNA sequences of individual species or samples you analyzed. Any two branches are connected to each other by a node (□), which represents the common ancestor of the two sequences.
   - The length of each branch is a measure of the evolutionary distance from the ancestral sequence at the node. Species or sequences with short branches from a node are closely related, those with longer branches are more distantly related.
   - A group formed by a common ancestor and its descendants is called a clade. Related clades, in turn, are connected by nodes to make larger, clades.
   - Click on a node (□) to highlight sequences in that clade. Click the node again to deselect the clade. What assumptions are made when one infers evolutionary relationships from sequence differences?
   - Generally, the clades will follow established phylogenetic relationships ascending from genus > family > order > class > phylum. However, gene and phylogenetic trees do disagree on some placements, and much research is focused on "reconciling" these differences. Why do gene and phylogenetic trees sometimes disagree?

d. Find and evaluate your sequence's position in the tree.
   - If your sequence is closely related to any of the reference or uploaded sequences, it will share a single node with those species.
   - If your sequence is identical to another sequence, the two will diverge directly from the node *without branches.*
   - If your sequence is distantly related to all of the species in your tree, your sequence will sit on a branch by itself – with the other sequences grouping together as a clade.
   - To identify the smallest clade that includes your sequence, click on the node that is directly connected to your sequence. The sequences that are

highlighted are the closest relatives of your sequence in the tree.

- Look at the scientific names of sequences within the most closely associated clade. If all members share the same genus name, you have identified your sequence as belonging to that genus. If different genus names are represented, check and see if they belong to the same family or order.

e. Return to the menu, and click on "PHYLIP NJ" to generate a phylogenetic tree using the neighbor joining method. How does it compare to the maximum likelihood tree? What does this tell you?

f. If neither tree places your sequence within an identifiable clade -- or if that clade is only at order level – you will need to add more sequences that may increase the resolution of your analysis. Return to Step 5, and add more reference sequences or obtain sequences within the order or family clade that contained your sequence. Then repeat Steps 6-7 to select, align, and generate trees from your refined data set.

The neighbor-joining algorithm builds a tree from the bottom up by comparing the evolutionary distance between pairs of DNA sequences. Sequences with best matching sequences are linked as "neighbors" that share common nodes in the tree. Because the branch distances are produced in a pairwise manner, neighbor joining does not optimize branch length and tree parsimony. The chief advantage of neighbor joining – that it is less computational intensive than maximum likelihood – has become less important as the processing power of computers has increased.