**Using DNA Subway to Analyze Sequence Relationships**

**Teacher Instructions:** Page 1

**Handout and Lab:** Page 2-15

**Description**: These are instructions for a 120-minute bioinformatics lab in which students will learn how to use the DNA Subway Blue Line workflow to determine the species of an organism using a DNA sequence.

**Resources and Materials:**

* **DNA Subway:** Online Bioinformatics Tool. Registration (free) required to save work. May also be used without registration as guest. https://dnasubway.cyverse.org/
* **DNA Subway manual:** Additional help for using DNA Subway

https://cyverse-dnasubway-guide.readthedocs-hosted.com/en/latest/

**Learning Objectives:**

Completing this lesson students should be able to:

* Understand that DNA Barcode sequences can be used to determine the identity of an organism
* Interpret a DNA sequence (trace file), determine if it provides usable data, and use software tools to clean DNA sequence reads.
* Understand the fundamental principles of BLAST including:
	+ where BLAST matches are retrieved from
	+ how BLAST matches are ranked
	+ what the limitations of BLAST are
* Create a multiple alignment using software tools
* Understand the relationship between a multiple sequence alignment and a phylogenetic tree
* Create a Neighbor-joining phylogenetic tree using software tools
* Understand that a phylogenetic tree shows relationships between DNA sequences and inferred relationships between organisms.

**Prerequisites and background**

The laboratory assumes that students have an understanding of the premise of DNA barcoding. While students may have done the wet lab protocols to generate their own DNA sequences, we will use sample sequences provided in DNA Subway. We will not cover every DNA subway functionality; see the DNA Subway manual at the link above for more information. A CyVerse account is suggested (in order to save work), but DNA Subway may be used without registration as a Guest.

* **Laboratory’s source material:** This resource hosts the complete wet lab protocols for generating DNA sequences from collected specimens http://www.dnabarcoding101.org/lab/index.html
* **Multimedia resources:** including the animation “What is DNA Barcoding”(See Animations & Videos tab) http://www.dnabarcoding101.org/resources/

**INTRODUCTION**

DNA Subway is an online educational bioinformatics platform. It bundles research-grade bioinformatics tools, high-performance computing, and databases into workflows with an easy-to-use interface.

In this exercise, we will show you how to identify a species based on DNA sequence, using a collection of DNA sequences generated by students at Cold Spring Harbor High School in 2010.

To generate these samples, students collected plant materials from their homes and from school grounds. DNA was extracted from these samples, and PCR was used to amplify a region of DNA sequence called a barcode region. Finally, DNA sequencing was used to read the order of the DNA nucleotides. From this information, we will perform the following steps to generate an identification:

* **Clean and assemble sequence results:** We will examine the quality of the DNA sequence and use two DNA sequences taken from the same sample to verify accurate sequencing results
* **Identify similar sequences:** Using BLAST, we will search for similar (related) DNA sequences from a database of known sequences
* **Determine sequence relationships:** We will compare our DNA sequence and use a phylogenetic tree to place our sequence in context with the tree of life

**LABORATORY**

1. **Create a DNA Subway Blue Line Project**

In this step, we will setup a new barcoding project. Since the Cold Spring Harbor samples are all plant material, we chose *rbcL* barcoding project type. *rbcL* (RuBisCo large subunit) is the gene sequence commonly used for plant DNA barcoding.

1. Log-in to [DNA Subway](https://dnasubway.cyverse.org/); unregistered users may ‘Enter as Guest’ but your work will not be saved.
2. Click the blue square (“Determine Sequence Relationships”) on the left side of the screen to start a Blue Line project.
3. Under “Select Project Type.” Select **rbcL** from the “Barcoding” column.
4. Under “Select Sequence Source” choose **CSH HS Oct 2010** from “Select a set of sample sequences”
5. Give the project a tile and, if desired a description. Click ‘Continue’ to complete the project setup.
6. **Clean and assemble sequence results**

Results obtained from DNA sequencing are not perfect. If they were, we could skip this step and use DNA sequencing data directly in our analysis. In reality, DNA sequencing can include error at several steps:

* **Human error**: Samples we send to the sequencing facility may have been mislabeled, or they may be mishandled or mixed at the sequencing facility.
* **Systemic error**: The cycle DNA sequencing used will often be unreliable at the very beginning or end of the sequence read.
* **Signal to Noise**: At any given position in the read the amount of signal (fluorescently labeled nucleotide) may be too low to distinguish from background, leading to one or more “miscalled” bases.

All experimental observations contain error. In this section, we will determine the quality of the sequence data we obtained. Additionally, we have sequenced each sample twice – once in the forward direction (capturing the 5’-3’ strand) and in the reverse direction (capturing the 3’-5’ strand). These complementary sequences should be identical, and differences between the two reveal sequencing error.

1. **View Sequences**

In this section, we will view the DNA sequences, both in text format (i.e. a sequence of nucleotides) and as an electropherogram.

1. Click "Sequence Viewer" to display the sequences you have input in the project creation section. As you scroll (left to right) you will see the DNA sequence obtained for each student sample.
2. Click on a sequence name (e.g. Student01\_PlantB\_F). This is the electropherogram for that DNA sequence.
	* The DNA sequencing software measures the fluorescence emitted in each of four channels – A, T, C, G – and records these as a trace, or electropherogram. In a good sequencing reaction, the nucleotide at a given position will be fluorescently labeled far in excess of background (random) labeling of the other three nucleotides, producing a "peak" at that position in the trace. Thus, peaks in the electropherogram correlate to nucleotide positions in the DNA sequence.
	* A software program called Phred analyzes the sequence file and "calls" a nucleotide (A, T, C, G) for each peak. If two or more nucleotides have relatively strong signals at the same position, the software calls an "N" for an undetermined nucleotide.
	* Phred also examines the peaks around each call and assigns a quality score for each nucleotide. The quality scores correspond to a logarithmic error probability that the nucleotide call is wrong, or, conversely, to the accuracy of the call.

|  |  |  |
| --- | --- | --- |
| **Phred Score** | **Error (bases miscalled)** | **Accuracy** |
| 10 | 1 in 10  | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |



**DNA Subway electropherogram**. DNA Sequence (as measured by fluorescence) is shown as colored peaks at the bottom of the readout. The quality (corresponding to Phred score) is shown as a histogram (blue bars at the top). The thin blue line in the histogram represents a Phred score of 20.

* The electropherogram viewer represents each Phred score as a blue bar. The horizontal line equals a Phred score of 20, which is generally the cut-off for high-quality sequence. Thus, any bar at or above the line is considered a high-quality read.
* Every sequence read begins with nucleotides (A, T, C, G) interspersed with Ns. In "clean" sequences, where experimental conditions were near optimal, the initial Ns will end within the first 25 nucleotides. The remaining sequence will have very few, if any, internal Ns. Then, at the end of the read the sequence will abruptly change over to Ns.
* Large numbers of Ns scattered throughout the sequence indicate poor quality sequence. Sequences with average Phred scores below 20 will be flagged with a "Low Quality Score Alert." You will need to be careful when drawing conclusions from analyses made with poor quality sequence.
* **Note:** The exclamation icon (!) indicates poor quality sequence.
1. Use the “X” and “Y” buttons to adjust the level of zoom. You can undo zooming by pressing the “Reset” button.

Examine the quality of the sequence(s). Any sequence for which the forward or reverse has the warning icon indicating a low-quality score in not of good enough quality to publish and any determination of novelty will be tentative as sequencing errors could appear to be novel polymorphisms.

After you have examined your sequences. Close the “Sequence Viewer” window.

**Section Questions**

* **What is the error rate and accuracy associated with a Phred score of 20?**
* **What do you notice about the electropherogram peaks and quality scores at nucleotide positions labeled "N"?**
* **What sequence(s) in the Cold Spring Harbor data set should not be used in further analyses?**
1. **Sequence Trimmer and Pair Builder**

In this section, we will remove leading and trailing “N” calls from the sequence data. We will also pair complementary sequences taken from the same sample.

1. Click “Sequence Trimmer”; when the run is completed, a green “V” will appear. Click the “Sequence Trimmer” again to view the results as in section A above.
2. Click “Pair Builder.’
3. Select the check boxes next to the sequences that represent bidirectional reads of the same sequence set (e.g. Student01\_PlantB\_**f** corresponds to the forward read of Student01\_PlantB and Student01\_PlantB\_**r** corresponds to the forward read of Student01\_PlantB). Alternatively select the ‘Auto Pair’ function and verify the pairs generated.
	* By default, *DNA Subway* assumes that all reads are in the forward orientation, and displays an "F" to the right of the sequence. If any sequence is not in that orientation, click the "F" to reverse compliment the sequence. The sequence will display an "R" to indicate the change.
	* Reverse complementing involves reversing the order the reverse read and then changing the bases to their complement bases. In this way, the two sequences should be identical, and should (mostly) overlap.
4. As necessary, reverse compliment sequences that were sequenced in the reverse orientation by clicking the ‘F’ next to the sequence name. The ‘F’ will become an ‘R’ to indicate the sequence has been reverse complimented.
5. Click “Save” to save the created pairs.

**Section Questions**

* **Why is it important to remove excess Ns from the ends of the sequences?**
* **Will a sequence and its reverse complement sequence completely overlap?**
1. **Consensus Builder**

In this section, we will combine the “cleaned” data from the forward and reverse reads. We will also remove poor quality areas at the 5’ and/or 3’ ends of the consensus sequence.

1. Click “Consensus Builder”; when the run is completed, a green “V” will appear. Click the “Consensus Builder” again to view the results. One sequence will now be displayed for each of the pairs.
2. Click on a sequence name to view the two sequences that were used to generate the consensus sequence.
	* The consensus sequence is the “sum” of both the forward and reverse read. Where the reads overlap, the highest quality read is selected.



**DNA Subway Consensus Editor**. Where reads conflict – the nucleotides will be highlighted in yellow; clicking on a mismatched nucleotide will also display the electropherogram at that location.

* + A large number of mismatches in properly paired and reverse complemented sequences indicate that one or both sequences is of poor quality. Often, one of the sequencing reactions produces a high quality read that can be used on its own.
1. A consensus sequence should be trimmed so that it includes only the region of overlap between the forward and reverse read. To trim, click the “Trim Consensus” link beneath the sequence names. On the consensus sequence, click on the last nucleotide you wish to remove from the beginning of the sequence. Repeat this procedure on the end of the sequence.



**DNA Subway Consensus Editor in Trim Mode**. A red line will indicate the sequences to be trimmed. Lower quality sequences will appear in black font. Although they may not appear as a mismatch, you may wish to trim these as well, especially since “N” bases will not be flagged as mismatches.

1. After you have selected the regions to trim, click the “Trim” button. Repeat this procedure for all your sequences.

**Section Questions**

* **For each of the consensus sequences, complete the following chart to indicate how you trimmed the sequence:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sequence name | Left trim bp | Right trim bp | Consensus length (after trimming) | Notes |
| *Student01\_PlantB* |  |  |  |  |
| *Student02\_PlantA* |  |  |  |  |
| *Student03\_PlantA* |  |  |  |  |
| *Student04\_PlantA* |  |  |  |  |
| *Student05\_PlantA* |  |  |  |  |
| *Student06\_PlantA* |  |  |  |  |
| *Student07\_PlantA* |  |  |  |  |
| *Student08\_PlantB* |  |  |  |  |
| *Student09\_PlantB* |  |  |  |  |
| *Student10\_PlantB* |  |  |  |  |
| *Student11\_PlantA* |  |  |  |  |
| *Student12\_PlantA* |  |  |  |  |
| *Student13\_PlantA* |  |  |  |  |

1. **BLAST DNA Sequences**

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.

1. Click “BLASTN”; next to the first DNA sequence, click the BLAST link. When the BLASTN search is complete (this may take up to a minute) you will be presented with a results table.
2. The BLASTN results table contains detailed information about sequences (matches) that are similar to your DNA sequence (query). Click on the species name for the first match in the list to learn more information about the species. Links to Wikipedia/Encyclopedia of Life may also contain photos of this species.
3. In the Accession column next to each match, there is a checkbox; click to select the top match, the click the “Add BLAST hits to project” button.
4. Repeat the BLASTN procedure to for each DNA sequence and answer the questions below:

**Section Questions**

* **For each of the consensus sequences, complete the following chart**

|  |  |  |  |
| --- | --- | --- | --- |
| Sequence name | Species name of top hit | Alignment (ALN) length (bp) | No. of Mismatches |
| *Student01\_PlantB* |  |  |  |
| *Student02\_PlantA* |  |  |  |
| *Student03\_PlantA* |  |  |  |
| *Student04\_PlantA* |  |  |  |
| *Student05\_PlantA* |  |  |  |
| *Student06\_PlantA* |  |  |  |
| *Student07\_PlantA* |  |  |  |
| *Student08\_PlantB* |  |  |  |
| *Student09\_PlantB* |  |  |  |
| *Student10\_PlantB* |  |  |  |
| *Student11\_PlantA* |  |  |  |
| *Student12\_PlantA* |  |  |  |
| *Student13\_PlantA* |  |  |  |

* + **Do you notice hits that have the same statistics (e.g. alignment length, bit score, e-value, mismatches) but which originate from different species? What does this tell you about the DNA Barcoding technique?**

 **More about BLAST**

**BLAST** (Basic Local Alignment Search Tool) is an algorithm to search for similarities between DNA or Protein Sequences. BLASTN specifically is used to compare a query nucleotide sequence and search for close matches in a database of known sequences. When BLAST finds a similar sequence, it returns that sequence along with statistical information that help you to judge the significance of the match (i.e. is there a match because these sequences share a biological relationship, or is the match more likely due to random chance).

Here are some important definitions about the information generated by the DNA Subway BLASTN search:

* **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 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 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 0 means that a search with your sequence would be expected to turn up no matches by chance. Blast results are ranked so that matches with the highest alignment length and bit score as well as the lowest e-value and mismatch number are at the top of the list (the best matches). In some cases, multiple species will have identical BLAST results – it is in these cases we see that BLAST is limited in differentiating very closely related sequences (and by extension, species).
* Examine the last column in the report called “**Mismatches**.” For barcodes, this is the informative column, with the best hits being those with the lowest number of mismatches. Note that hits with low numbers of mismatches can sometimes be lower on the list, as the bit scores are used to arrange the hits in the table. High bit scores can occur when the alignment length is longer, even when there are more mismatches than for other hits.

**More about novel sequences and the limitations of BLAST**

If there are zero mismatches between your sequence and a BLAST result, it is unlikely that your sequence is unique. Instead, the identical sequences probably match because they are in the same taxonomic group as your sample. Check to see if the matching sequences are from species that seem reasonable for your sample. If your best matches include some mismatches, you may have identified a novel barcode. The more mismatches you find, the more likely that your sequence is unique, especially in regions of the sequence with high quality scores. However, sequencing errors could explain the difference, so it will be important to reexamine the trace files at any sites with mismatches to ensure that the consensus at those locations is of high quality.

It is also important to note that BLAST results alone cannot conclusively determine the identity of your species. Using a single DNA Barcoding locus (See references here: http://www.dnabarcoding101.org/lab/index.html) **we can only expect to identify the genus of an unknown organism**. In other words, in closely related species, a single DNA barcode may not reveal any mutations that differentiate these species. To overcome this limitation, it may be necessary to use more than one DNA barcoding locus.

1. **Compare sequences using a multiple alignment and phylogenetic tree**

In this section, we will assemble and align all of our collected sequences. From this information, we will construct a phylogenetic tree, a hypothesis about how these sequences are related.

1. **Add reference sequences and create a multiple sequence alignment**

First, we will add some additional species to our analysis so that we can use these as a point of comparison. These common plants contain representatives from many of the large plant families. A multiple sequence alignment allows you to compare several DNA sequences at once, visualizing similarities and differences between the sequences.

1. Click “Reference Data” and select “Common plants.” Next click “Add ref data”.
2. Click “Select Data” and select sequences to add to your project. Be sure to exclude any sequences you previously identified as being of low quality. Select all Blast hits and the list of common plants. Click “Save Selections” to add these plants to your analysis.
3. Click “MUSCLE” to generate the multiple sequence alignment. When the “V” appears, click MUSCLE again to view the alignment results.



**DNA Subway Alignment Viewer**. *Sequence Conservation* displays a histogram across the displayed sequences. At positions were most nucleotides are the same, the histogram approaches 100%, dips in the histogram are more variable regions. *Sequence Variation* displays the nucleotides that occur at that position relative to the consensus sequence; the colors of the bars (Green = A, Red = T, Black =G, Blue =C) display what alternative nucleotide(s) appear at that position. Missing sequence is indicated in each row by a grey block on either end of the sequence.

**How to interpret the multiple alignment**

A consensus sequence shows the most common nucleotide at a given position along the length of the match. Using the (+) button or the (ATCG) button to zoom down to the nucleotide level, the entire consensus sequence is shown, and nucleotides in the rows of sequences are only shown if they differ from the consensus. At this point, it should be possible to see clear patterns of sequences that share mutations (polymorphisms).

* 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).
* 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 "MUSCLE" to recalculate.
1. Click the “TRIM ALIGNMENT” button to trim the alignment to a region where all the selected DNA sequences overlap. Without this trimming step, those missing regions of sequences would be interpreted by the phylogenetic tree building algorithm as true deletions in the sequence, rather than missing data.

**Section Questions**

* **Looking at the sequence conservation histogram, are some regions in the barcode more conserved or lesson conserved than other regions?**
* **What visual patterns can you see from looking at a multiple alignment?**
1. **Create a Neighbor-Joining phylogenetic tree**

In this final section, we will use an algorithm to generate a phylogenetic tree. A Phylogenetic tree is hypothesis about how sequences are related. Although a phylogenetic tree may look like a family tree, the analogy is incomplete, since in a family tree we (usually) know for certain what the relationships between members are. In a phylogenetic tree, we are using only some of the available evidence – in this case DNA sequence from a single barcoding region – to make an inference about how species may be related.

1. Click “PHYLIP NJ” to construct your phylogenetic tree. When the “V” appears, click PHYLIP NJ again to view the alignment results.

**More on Phylogenetic Trees**

Phylogenetic trees can display inferred relationships between organisms by comparing changes in DNA or protein sequences and grouping collections of DNA or protein sequences by similarity.  Often, changes in DNA or protein sequences are used to infer the overall relationships between organisms.

Trees consist of branch tips (also called leaves), which are labeled with the name of the sequence and/or organism, and nodes, which represent the point at which descendants from an inferred common ancestor diverged into different lineages.

An analysis of 10 sequences (perhaps from 10 different organisms) will result in a tree with 10 leaves. The number of nodes depends on how many hypothetical common ancestors can be inferred to exist between the sequences. Just how many common ancestors (nodes) may have existed, and what justifies hypothesizing a common ancestor (i.e. creating a node) varies according to the tree building algorithm, and underlying evolutionary assumptions.

Relatedness between any two sequences decreases as the number of nodes increases; sequences sharing only one node are more closely related than sequences that can only be joined by two nodes.  If your analysis consists of sequences from different organisms, nodes may represent distinct taxonomical units (such as families or species).

The NJ method is a “distance-based” method of tree building. First a matrix of the sequences is created and the distance between each sequence is calculated. To construct a tree, pairs of sequences are connected by nodes starting with the most closely related sequences. Inside the MUSCLE window you can click the “Sequence Similarity %” button in the upper right-hand corner of the alignment window to see a similar matrix.

To find the most likely tree and determine the reliability of the branches in this tree, NJ in DNA Subway uses bootstrapping. In bootstrapping, the columns in the sequence alignment are sampled randomly to make many new alignments – 100 for NJ in DNA Subway – and these alignments are used to construct NJ trees. The final tree represents the “most likely” tree, and shows the confidence of relationships with bootstrap levels. Each bootstrap value is the number of times that relationship appears in the 100 resampled trees. The values do not represent the distance between sequences. Instead, a higher value indicates that a branch of the tree is well supported, while low values indicate that the relationships are less certain. In general bootstrap values above 70 might be considered as plausible given the data, and above 95 can be considered “correct.”

For more on interpreting phylogenetic trees see:

* + Teaching the Process of Molecular Phylogeny and Systematics: A Multi-Part Inquiry-Based Exercise <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995769/pdf/cbe513.pdf>
	+ Teaching Tree-Thinking to Undergraduate Biology Students

<http://evolution.berkeley.edu/UToL/meisel10.pdf>

**Section Questions**

* **What assumptions are made when one infers evolutionary relationships from sequence differences?**
* **How should you view relationships with low bootstrap support (i.e. bootstrap value < 70)?**
* **How well do your conclusions about possible species identifications agree (or disagree) with BLASTN search results?**

**Conclusion**

In this exercise, the Cold Spring Harbor High school sequences were analyzed, first by cleaning these sequence data, using BLASTN to search a known database of sequences for matches, and by placing these sequences in the phylogenetic contexts of known organisms. DNA Barcoding is a powerful technique for identifying organisms. However, DNA Barcoding alone may only be once piece of evidence needed to establish the identity of an unknown sequence. In general, barcoding at a single locus (such as rbcL) will not yield a conclusive identification for the species, but should establish the genus the sample belongs to. With additional barcode sequences, it may be possible to have enough evidence for a conclusive identification. Finally, traditional taxonomic characters such as the shape/size/structure of the organism and its various components are used for definitive identification.

**Further work**

Visit the DNA Barcoding 101 website (http://www.dnabarcoding101.org/) for more information on DNA Barcoding, including how to use DNA Subway to analyze your own DNA sequences and even contribute them to the global online databases for DNA sequencing and barcoding.