Teacher Guide:
Introduction to Bioinformatics

Goals

In this lab activity, students will:

- Analyze and manipulate Sanger sequencing chromatograms
- Determine the taxonomic identification of unknown arthropods using NCBI BLAST
- Visualize the evolutionary relatedness of arthropods using phylogenetics

Learning Objectives

Upon completion of this activity, students will (i) compare the identification of arthropods using morphological characterization vs. DNA sequencing; (ii) learn how to convert .ab1 chromatograms to FASTA files; (iii) become familiar with NCBI BLAST; and (iv) build a phylogenetic tree to explore the evolutionary relatedness of arthropods.

Prerequisite Skills

No computer programming skills are necessary to complete this work; prior exposure to personal computers and the Internet is assumed. A review of transcription (DNA → RNA), translation (RNA → protein), and the genetic code (relationship between codons and amino acids) is highly recommended prior to this module.

Group Size

This activity can be performed in small groups (2-4 students) or as an individual project.

Teaching Time

The entire module will take approximately two class periods. Classroom time can be condensed by assigning a few arthropods to each student and/or having students work in groups. Each of the four activities are designed to be standalone units and do not require completion of the previous activity; however, completing the activities in consecutive order ensures comprehension of the stepwise methods.

Supplies

Bioinformatics platforms are highly variable in cost, functionality, and reliability. SnapGene Viewer and NGPhylogeny.fr are recommended for ease of use. Alternatives are listed in Appendix B and C.

- Computer with internet browser, such as Firefox or Chrome
- Project Guide - https://wolbachiaproject.org/bioinformatics/
- Download Example sequences from https://wolbachiaproject.org/bioinformatics/
- Download SnapGene Viewer (free software) - https://www.snapgene.com/snapgene-viewer
    - This software is required for Activity 2 and is not compatible with Chromebooks. Alternative web-based platforms are discussed below.
- Online access to https://ngphylogeny.fr or other phylogenetics programs, discussed below.
- Answer key: Contact info@wolbachiaproject.org to obtain a copy of the Answer Key.

Module Overview

Activity 1: Students will identity arthropods based on morphology. The level of resolution is up to instructor discretion – students may use common knowledge (i.e., spider or beetle) or identity down to *at least* taxonomic order. The goal is to compare morphological (Activity 1) to molecular identification (Activity 3).

Activity 2: Students will assess the quality of Sanger chromatograms, the raw data generated from DNA sequencing, and generate a FASTA file for downstream bioinformatics analyses. This activity is recommended if students will be sequencing their own arthropod samples and need to learn how to analyze .ab1 files.

Activity 3: Students will query DNA sequences against the national sequence archive, NCBI, to putatively identify each arthropod. Results from this activity can be compared against the results from Activity 1 to assess morphological vs. molecular classification. Note that students will be given FASTA files and do not need to complete Activity 2 to participate in Activity 3.

Activity 4: Students will learn how to edit a FASTA file, build a phylogenetic tree, and designate an outgroup. Through participation in this activity, students will develop the skills to add their own taxa to the input file and generate a customized tree of sequenced arthropod DNA.

National Science Standards

This module supports biological evolution components of the national science standards. A description of how the activities address each concept is listed below.

NGSS

HS-LS4 Biological Evolution: Unity and Diversity

Students will learn (i) DNA sequence comparisons of different organisms to infer common ancestry and diversity, and (ii) evolution is a consequence of genetic changes over time due to mutation and sexual reproduction.

AP Biology Framework

Topic 7.6 Evidence of Evolution

Students will learn how to compare DNA nucleotide sequences and use this information as evidence for evolution and common ancestry.

Topic 7.8 Continuing Evolution

Students will analyze nucleotide sequences as a measure of genomic changes over time.

Topic 7.9 Phylogeny

Students will construct phylogenetic trees, based on barcoding genes, to infer evolutionary relationships. They will compare the use of molecular data vs. morphological observations to infer biodiversity and evolution.

Vision and Change

Core Concept 1: Evolution

Students will learn how DNA nucleotide sequences can be analyzed to measure biodiversity and infer evolutionary relationships. They will (i) analyze Sanger sequencing chromatograms to determine DNA sequence, (ii) perform nucleotide alignments to assess substitution rate, (iii) BLAST DNA sequences against the NCBI database to identify closely related organisms, and (iv) build a phylogenetic tree using their DNA sequence.

Selecting a Sanger Chromatogram Viewer

This Project Guide includes three options. All are freely accessible.

- **SnapGene Viewer** (https://www.snapgene.com/snapgene-viewer): This software is the easiest to use, displays quality scores, and allows students to generate FASTA files directly from the chromatogram sequence. The software must be downloaded and is not compatible with Chromebooks.
- **Benchling** (https://www.benchling.com/educators): This software is web-based and displays quality scores. It requires an email address for login and students will need to generate their own FASTA files using a text editor.
- **Teal** (https://www.gear-genomics.com/teal): This software is web-based and does not require an email address for login. Quality scores are not included on the chromatograms. Students will need to generate their own FASTA files using a text editor.

	SnapGene Viewer (Lab Activity 2)	**Benchling** (Appendix B.1)	**Teal** (Appendix B.2)
Software download; compatible with MacOS and Windows	✓		
Web-based; compatible with MacOS, Windows, and Chrome		✓	✓
Email address required	✓	✓	
View chromatogram	✓	✓	✓
View quality scores (QS)	✓	✓	
Trim/edit chromatogram	✓	✓	
Directly export sequence as FASTA file	✓		

Selecting a Phylogenetics Workflow

This Project Guide includes four options. All are freely accessible and web-based; therefore, they should be compatible with all operating systems (Mac, Windows, Linux, and Chrome). Because they are hosted on external servers, we recommend briefly testing the program(s) prior to class to make sure the web servers are online and responsive. They are listed below in order of ease of use to complexity (most customizable).

- **NGPhylogeny.fr** (https://ngphylogeny.fr/): This is the default program featured in Lab 4 and is recommended for introductory classes and non-specialists. It features a "One-Click Workflow" to provide a seamless experience for students. Each of the steps in the table below are automatically processed. Students may observe the workflow in real-time and click on associated links to investigate each step. Occasionally, we receive feedback that the server is unresponsive (it should not take longer than a few minutes to build the tree). Appendix C lists alternative options.
- **Phylogeny.fr** (http://www.phylogeny.fr/): This program predates NGPhylogeny.fr and features a similar workspace. It is often online and accessible when the previous server is unavailable. Use caution that the site connection is "not secure".
- **MAFFT version 7** (https://mafft.cbrc.jp/alignment/server/): This option is the most customizable and requires users to manually progress through each of the steps listed in the table below. It is recommended for more experienced users or for a more comprehensive learning experience, such as independent research projects. The MAFFT server is a multiple alignment program that aligns the FASTA sequences and generates a tree file for downstream phylogenetic analysis. Users must select a separate tree building software to visualize the phylogeny.
    - **Phylo.io**: This option is offered through the MAFFT environment. It is quick and easy, but not customizable.

- **Interactive Tree of Life (iTol)**: iTol is highly customizable and recommended for experienced users. The program generates publication-quality trees. Users will need to download the tree file from MAFFT and navigate to the external website. The countless options may be overwhelming to beginners. We recommend exploring the tools. Reset and Undo buttons are available to reverse unwanted changes.

In general, each program utilizes the same task workflow. Just as we can select a preferred word processing software to best suite our needs – such as Microsoft Word, Google Docs, or Apple Pages – bioinformaticians have multiple software options for each type of analysis. The workflows offered in this Project Guide are reliable, quick, and easy. Research labs employ these programs, as well as more expensive and robust platforms to work with larger datasets.

PHYLOGENETICS WORKFLOW		
STEP	**TASK**	**DESCRIPTION / UNDERLYING SOFTWARE**
1	**Input Data**	Users input a FASTA file that lists all sequences to be included in the phylogenetic tree.
2	**Multiple Sequence Alignment (MSA)**	Sequences from the FASTA file are aligned based on similarity. NGPhylogeny.fr and MAFFT use MAFFT; Phylogeny.fr uses MUSCLE.
3	**Alignment Curation**	The most informative regions (or nucleotides) for phylogenetic analysis are selected. NGPhylogeny.fr uses BMGE; Phylogeny.fr uses Gblocks; MAFFT is highly customizable. In general, indels and sequence ends are trimmed because they do not contain genetic information across all sequences. When comparing a sequence with 500 nucleotides to a sequence with 1500 nucleotides, for example, the alignment is curated to include only the shared region of 500.
4	**Phylogenetic Tree Inference**	Information from the curated alignment is converted to a phylogenetic tree structure. NGPhylogeny.fr uses FastMe; Phylogeny.fr uses PhyML; MAFFT uses NJ. Because each program uses a different underlying algorithm to infer the tree, different trees may be generated with each analysis. More robust platforms allow the users to select modes of evolution and other customizable variables to best analyze the dataset.
5	**Tree Visualization**	Information from the tree file is converted to a visual display. NGPhylogeny.fr uses Newick; Phylogeny.fr uses TreeDyn; MAFFT is customizable. Users can select from Phylo.io, Archaeopteryx, or download various tree files for external tree viewers, such as iTol.