

## 3

# Using Bioinformatics to Study Evolutionary Relationships

## Introduction

In this lesson, students learn how to use bioinformatics tools to analyze DNA sequence data and draw conclusions about evolutionary relationships. Students collaborate with their group members by pooling their DNA sequences from *Lesson Two: DNA Barcoding and the Barcode of Life Database (BOLD)* to perform and analyze multiple sequence alignments using the computer programs **ClustalW2** and **JalView**. After comparing relatedness among and between the species within their group, students use their sequence alignment to generate a **phylogenetic tree**, which is a graphical representation of inferred evolutionary relationships. This tree is used to draw conclusions about their research question and hypothesis. In *Lesson Three*, students also learn how **microbiologists** might use bioinformatics tools in their career.

## Learning Objectives

At the end of this lesson, students will know that:

- Bioinformatics tools are used by people in many different career fields, including microbiologists.
- Scientific collaboration and data sharing are vital components of the scientific process.
- The bioinformatics tools ClustalW2 and JalView can be used to analyze long DNA sequences much more quickly and accurately than can be done by hand.
- Phylogenetic trees reflect the similarities and differences among DNA sequences, and these trees are used to infer evolutionary relationships (i.e., similar species are grouped together).

At the end of this lesson, students will be able to:

- Use a group data set with their scientific collaborators to perform a **multiple sequence alignment**, comparing DNA sequences within their data set using ClustalW2 and JalView.
- Analyze multiple sequence alignments to begin to answer the research question and hypothesis they developed in *Lesson Two*.
- Create a phylogenetic tree in **BLAST** based on their multiple sequence alignment.

## Key Concepts

- Bioinformatics tools like ClustalW2, JalView, and BLAST are used by scientists to compare DNA sequences.
- DNA sequences that are more similar are believed to share a more recent common ancestor than DNA sequences that show more differences.

## Class Time

1 class period (50 minutes).

## Prior Knowledge Needed

- DNA contains the genetic information that encodes traits.
- Basic knowledge of taxonomy (specifically the different categories used in taxonomy to classify organisms, and that the study of taxonomy seeks to reflect the evolutionary history of different organisms).
- Binomial nomenclature (i.e., the use of **genus** and **species** to refer to individual species).
- DNA sequence data is needed to answer genetic research questions and evaluate hypotheses (*Lesson One* and *Lesson Two*).
- *COI* is the barcoding gene used for animals (*Lesson Two*).

- The information obtained from multiple sequence alignments can be used to construct phylogenetic trees.
- Phylogenetic trees are a graphical representation of the evolutionary relatedness among the species in the tree.

## Materials

Materials	Quantity
Copies of Student Handout— <i>Careers in the Spotlight</i> (handed out in <i>Lesson One</i> )	1 per student
Copies of Student Handout— <i>The Process of Genetic Research</i> (handed out in <i>Lesson One</i> )	1 per student
Class set of Student Handout— <i>Using Bioinformatics to Study Evolutionary Relationships Instructions</i>	1 per student (class set)
Copies of Student Handout— <i>Using Bioinformatics to Study Evolutionary Relationships Worksheet</i> [Note: This worksheet is for students' answers to lesson questions.]	1 per student
Teacher Answer Key— <i>The Process of Genetic Research</i> (found in <i>Lesson One</i> )	1
Teacher Answer Key— <i>Using Bioinformatics to Study Evolutionary Relationships</i>	1

Computer Equipment, Files, Software, and Media
Computer with internet access and projector to display PowerPoint slides. <b>Alternative:</b> Print PowerPoint slides onto transparency and display with overhead projector.
<i>Lesson Three PowerPoint Slides—Using Bioinformatics to Analyze Evolutionary Relationships.</i> Available for download at: <a href="http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research">http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research</a> .
"Unknown DNA Sequences" 1-30. Available from the Bio-ITEST website under the "Resources" tab at: <a href="http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research">http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research</a> .
A student version of lesson materials (minus Teacher Answer Keys) is available from NWABR's Student Resource Center at: <a href="http://www.nwabr.org/students/student-resource-center/instructional-materials/advanced-bioinformatics-genetic-research">http://www.nwabr.org/students/student-resource-center/instructional-materials/advanced-bioinformatics-genetic-research</a> .
Computer lab with internet access. [Note: Use of Microsoft® Word is not recommended when performing bioinformatics analyses, but can be used to answer homework questions if desired.]

## Teacher Preparation

- Load the classroom computer with the *Lesson Three* PowerPoint slides.
- Make copies of the Student Handout—*Using Bioinformatics to Study Evolutionary Relationships Instructions*, one per student. This handout is designed to be re-used as a class set.
- Make copies of Student Handout—*Using Bioinformatics to Study Evolutionary Relationships Worksheet*, one per student. This worksheet is used for students to write their answers to the lesson questions. Alternatively, answers may also be written in students' lab notebooks or on a separate sheet of paper.

## Procedure

### Warm Up

1. As students enter the classroom, display the PowerPoint slides for *Lesson Three*, beginning with **Slide #1**. This slide highlights microbiologist Lalita Ramakrishnan, PhD.

**Microbiologist**  
 Lalita Ramakrishnan, MD, PhD



**Place of Employment:**  
University of Washington

**Type of Research:**  
Tuberculosis infection

**Model Organism:**  
Zebrafish

Zebrafish are naturally susceptible to tuberculosis. Because their genes are fairly easy to manipulate, we can create some zebrafish that are susceptible to TB and some that are resistant to TB. Zebrafish are also good model organisms because they are transparent, so we can watch the infection process develop.

Bioinformatics & Evolution: **Slide #1**

2. Have students retrieve Student Handout—*Careers in the Spotlight* from *Lesson One*.
3. Students should think about, and write down, the kind of work they think a microbiologist might do (*Microbiologist Question #1*). This will be revisited at the end of the lesson, including how a microbiologist might use bioinformatics in his or her job.
4. Tell students to keep their *Careers in the Spotlight* handout available for future lessons.

### PART I: Multiple Sequence Alignments and Phylogenetic Trees

5. Explain to students the **aims of this lesson**. Some teachers may find it useful to write the aims on the board.
  - a. **Lesson Aim:** Collaborate with other student scientists in your research group.
  - b. **Lesson Aim:** Create **multiple sequence alignments** and **phylogenetic trees** to answer research questions.

Teachers may also wish to discuss the *Learning Objectives* of the lesson, which are listed at the beginning of this lesson plan.

6. Tell students that genetic research often involves comparing DNA sequences to one another. Explain that the students' goal today is to analyze their DNA data from *Lesson Two*, in collaboration with their group members, to answer their research question and evaluate the hypothesis they developed.

#### Multiple sequence alignment:

The process of comparing two or more DNA or protein sequences to one another by aligning the sequences and looking for similarities and differences.

#### Phylogenetic tree:

A branching diagram or "tree" showing the evolutionary relationships among various species, based upon similarities and differences in their physical and/or genetic characteristics. The lengths of the branches on the tree are proportional to the amount of time since the two organisms (or sequences) diverged from one another.

**Reference sequence:** A sequence that has been chosen for the purpose of comparison. At the National Center for Biotechnology Information (NCBI), reference sequences are chosen because they are of high quality and are thought to accurately represent the sequence from the original organism. In genetic testing, a reference sequence is a known and well-studied DNA or protein sequence that does not contain any mutations or changes that are associated with disease. These sequences are used for comparison with patient sequences during genetic testing.

**Mutation:** A change in a DNA or protein sequence.

7. Remind students that the focus of these experiments is phylogenetics, which is the study of the evolutionary relationships among organisms.

8. Describe today's activity: students will compare their DNA sequences by aligning them to other DNA sequences from their collaborators in their group using a process called multiple sequence alignment. Based on these alignments, they will construct a phylogenetic tree to describe the evolutionary relationships among the organisms in their group.

## 9. Alignment to a Reference Sequence:

Show **Slide #2**, which uses the example of breast cancer and genetic testing to illustrate a sequence alignment with a known **reference sequence**. Explain to students that for sequence comparisons such as those used for genetic testing, scientists often compare sequences from healthy people to sequences from families with a history of breast cancer to look for changes or **mutations** that may be associated with cancer. In this case, we refer to the sequence of the *BRCA1* gene found in healthy people as the reference sequence, and the sequences we are searching with (in this case, from patients with a family history of breast cancer) are called the **query sequences** ("query" means "question" or "inquiry"). Scientists look for changes or mutations in the patient sequences by comparing them to the reference sequence. If students have already worked through the Bio-ITEST Introductory bioinformatics lessons, *Using Bioinformatics: Genetic Testing*, remind them of the sequence alignments they performed using DNA and protein sequences from the Lawler family to look for mutations in *BRCA1*.

Bioinformatics & Evolution: **Slide #2**

### Comparing DNA Sequences

Example: Genetic Testing using BLAST

Reference BRCA1 Sequence A T A G C T G

Query Sequence(s): Patient 1 A

Patient 2 A C

Patient 3 \_\_\_\_\_

Look for **mutations** or changes relative to **Reference Sequence**

## 10. Alignment to Other Sequences:

Show **Slide #3**, which uses the example of DNA sequences from different fruits. When comparing the relatedness of papaya, grapes, tomatoes, and watermelon, each of the sequences is compared to the others to evaluate the differences. Scientists use the number of changes between different sequences to understand the evolutionary relatedness of the organisms, as mutations (or differences between sequences) accumulate over time. When sequences from two species are very similar, they are thought to be closely related and share a common ancestor; when sequences from two species are more dissimilar, the species are thought to be more distantly related.

## Comparing DNA Sequences

Example: Genetic Testing using BLAST

Reference BRCA1 Sequence A T A G C T G

Query Sequence(s): Patient 1 A

Patient 2 A C

Patient 3 \_\_\_\_\_

Look for **mutations** or changes relative to **Reference Sequence**

Example: Multiple Sequence Alignments Using JalView and ClustalW

Papaya A T G G T G C

Grape A T G C T G C

Tomato A T G C A G C

Watermelon A T G G A C A

Look for **changes** relative to **each other**

**The amount of change among the sequences reflects the evolutionary relatedness of the organisms.**




Image Source: Wikimedia Commons.

### 11. Predictions:

Ask students to look at the DNA sequence comparisons for the different plant species on **Slide #3** and make a prediction about which species are most closely related. They can write down their prediction on a piece of paper or raise their hands to share their prediction(s) with the class. Based on the short DNA sequences shown, students might predict that papaya and grapes are more closely related to one another than tomatoes and watermelons, as there is one nucleotide difference between the papaya and grape sequences.

### 12. Bioinformatics Tools:

Explain that during today's activity, the students will be making conclusions about the evolutionary relatedness of the organisms in their group based on their multiple sequence alignments using the bioinformatics tools **ClustalW2** and **JalView**. ClustalW2 is the program that will perform the multiple sequence alignments and JalView is the program students will use to view and manipulate the alignments. They will also be using **BLAST** (the program from *Lesson Two*) to make phylogenetic trees.

13. Tell students that the next step in using genetic data to study evolutionary relatedness is to construct a phylogenetic tree. In particular, the phylogenetic tree will be based on pairwise comparisons, as explained in the next slide.

### 14. Pairwise Comparison:

Show students **Slide #4** and explain that bioinformatics tools like ClustalW2 and BLAST compare all possible pairs of sequences to one another. The total number of differences is tallied by the program. Multiple sequence alignment programs like ClustalW2 take these differences into account when making the sequence alignments. This is called a pairwise comparison, as each of the sequences is compared to each other one in pairs. The total number of differences between each pair of sequences is recorded by the program and used to make the best alignment possible between all pairs and the best possible tree. This is similar to the table students made with their canine sequences in *Lesson One*.

**ClustalW2:** A program that performs multiple sequence alignments.

**JalView:** A program provided by the University of Dundee in Scotland that helps users generate multiple sequence alignments by interfacing with the alignment program ClustalW2. The program also allows users to edit their alignments, helping with their genetic analyses.

**BLAST:** Basic Local Alignment Search Tool. A bioinformatics tool used to compare DNA or protein sequences to one or more other sequences, or to compare a DNA or protein sequence to a collection of sequences found in databases, such as the Nucleotide or Protein databases at the NCBI.

## Pairs of Sequences are Compared to Each Other

Papaya: ATGCTGCCG  
Grape: ATGCTGCCG

Grape: ATGCTGCCG  
Watermelon: ATGACACG

Grape: ATGCTGCCG  
Tomato: ATGGTGAAG

Papaya: ATGCTGCCG  
Tomato: ATGACACG

Tomato: ATGACACG  
Watermelon: ATGACACG

Papaya: ATGCTGCCG  
Watermelon: ATGACACG

Number of Nucleotide Differences:





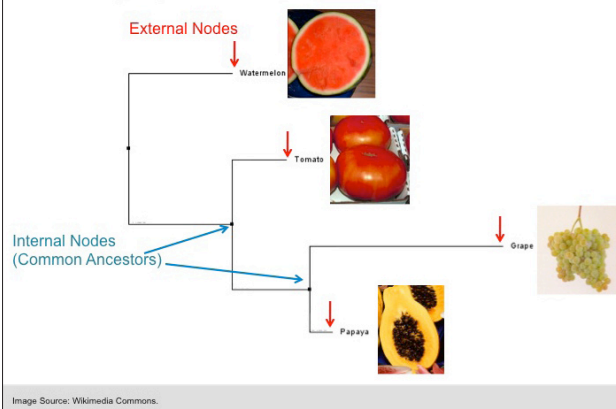
				
Papaya	0	1	2	3
Grape	1	0	2	4
Tomato	2	2	0	3
Watermelon	3	4	3	0

Image Source: Wikimedia Commons.

## 15. Phylogenetic Trees:

Show **Slide #5** and explain that these comparisons are then used to generate a phylogenetic tree, which is a graphical representation of the evolutionary relationships among these organisms.

## Phylogenetic Trees Reflect Evolution



**External node:** On a phylogenetic tree, the external nodes are at the ends of each branch, with each external node representing a different species in the tree.

**Branch:** Horizontal lines on a phylogenetic tree that connect different nodes to one another.

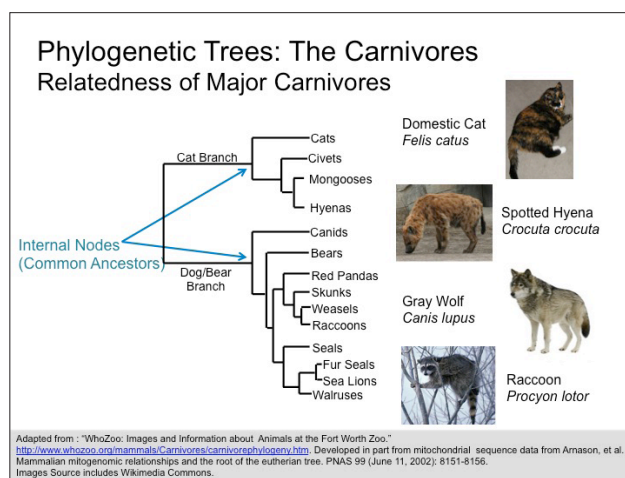
**Internal node:** On a phylogenetic tree, the internal nodes are points where branches split, and represent an inferred evolutionary common ancestor.

- Point out to students that each species is found on a separate **external node** on the tree. Each species has its own **branch**. The internal branching points or **internal nodes** represent common ancestors. In the example shown in this slide, papaya and grape shared a common ancestor long ago, and further back in their evolutionary history that common ancestor shared a common ancestor with tomato, and so forth.
- Ask students to share with the class how their prediction(s) compare to what they see in the phylogenetic tree. If students have already shared their predictions, point out how those predictions relate to what is seen in the tree (that grape and papaya are most closely related, while watermelon is the most distantly related).

18. Slow **Slide #6**. Ask students to discuss the following questions:

- What do they observe in this phylogenetic tree?
- What is the research question this tree is exploring?
- What specific information about evolutionary relationships can be gleaned from this tree?
- What, if anything, surprises them about this tree?

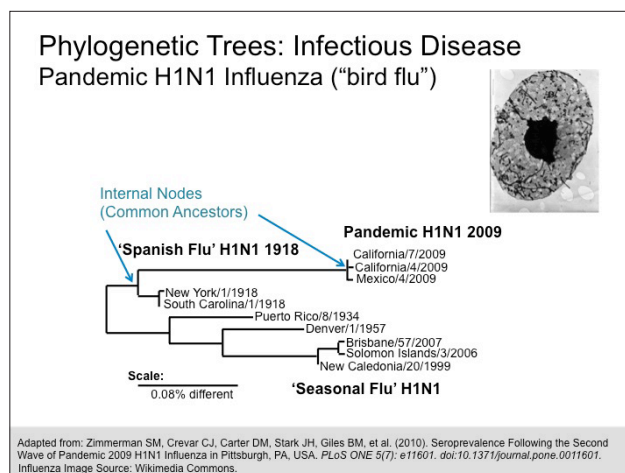
From this DNA analysis, we can see that, contrary to popular belief, hyenas are more closely related to cats than they are to dogs. It is also interesting to note that all of these carnivore species form two distinct clusters – the “cat branch” and the “dog and bear branch.”



Bioinformatics & Evolution: **Slide #6**

19. Show **Slide #7**, which is an example of a phylogenetic tree of influenza (“flu”) virus samples. Ask students to discuss the following questions:

- What do they observe in this phylogenetic tree?
- What is the research question this tree is exploring?
- What specific information about evolutionary relationships can be gleaned from this tree?
- What surprises them about this tree?



Bioinformatics & Evolution: **Slide #7**



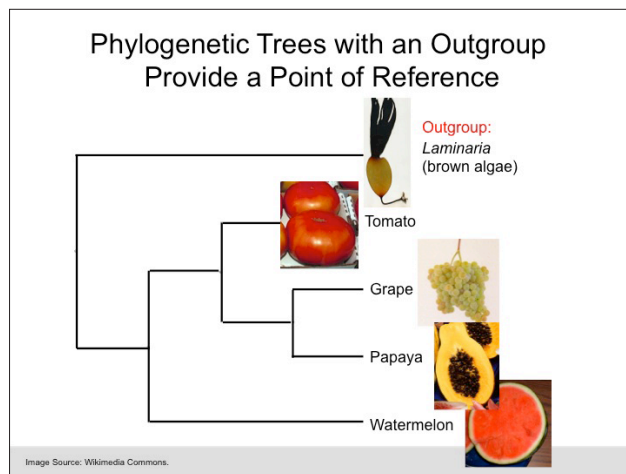
**Outgroup:** A sequence included in phylogenetic trees from an older or more distantly related species, used as a point of reference and to mark the evolutionary time. Inclusion of an outgroup is said to **root** the tree.

**Rooted tree:** Phylogenetic trees that contain an outgroup are said to be “rooted.”

Bioinformatics & Evolution: **Slide #8**

## 20. Outgroups:

Show **Slide #8**. This is very similar to the tree from **Slide #5**, except that this tree includes an **outgroup**. An outgroup is a distantly related species to the ones you are studying that helps provide a point of reference for your tree. An outgroup is said to **root** the tree – it anchors it and makes comparisons within the tree more meaningful. In this tree, we can still see that grapes and papayas are more closely related to **each other**, followed by tomatoes and watermelons, but we can also see that all of these plants are much more closely related to **each other** than they are to the outgroup, *Laminaria setchellii*, which is a kind of brown algae found in the ocean.



21. Pass out Student Handout–*Using Bioinformatics to Study Evolutionary Relationships* and have students work through the handout in their groups at the computer.

## PART II: Putting it all Together

22. Students should complete their handout and work with their collaborators (group members) to answer their research questions and to support or reject their hypotheses. Next, ask students to work within their groups to prepare a short summary of what they have learned.

23. Write questions on the board for each group of students to answer. The answers to these questions could be shared in a class discussion, or written answers could be turned in to the teacher at the end of class. Some questions to ask include:

- What taxonomic group did you study?
- Which species did you study?
- Within your group, which two or three species were mostly related to one another?



## Closure

24. Summarize today's lesson:

- Students have learned how to compare DNA sequences to one another by performing multiple sequence alignments.
- They have used those alignments to construct phylogenetic trees to infer the evolutionary relatedness among the organisms in their group.
- They have also used these analyses to answer their research questions and to support or reject their hypotheses.
- They also learned the importance of scientific collaboration by pooling or sharing their DNA data with other group members.

In the next lessons, they will learn about other types of analyses they can perform with their DNA sequences, including translating their DNA sequences into protein sequences *in silico* using the tools of bioinformatics.

***In silico*:** An expression used to mean “performed on computer or via computer simulation.”

25. Ask students to fill out the section about *Lesson Three* in Student Handout—*The Process of Genetic Research*, which was handed out during *Lesson One*. Students could also answer these questions in their lab notebooks:

- What did you do in this lesson?
- **Methods:** What bioinformatics tool(s) and/or database(s) did you use?
- **Results & Conclusions:** What did you find? What could you conclude from your analysis?
- What **skills** did you learn or practice?

26. Show **Slide #9**, which returns to the picture of the microbiologist from **Slide #1**.

**Microbiologist**  
 Lalita Ramakrishnan, MD, PhD



**Place of Employment:**  
University of Washington

**Type of Research:**  
Tuberculosis infection

**Model Organism:**  
Zebrafish

Zebrafish are naturally susceptible to tuberculosis. Because their genes are fairly easy to manipulate, we can create some zebrafish that are susceptible to TB and some that are resistant to TB. Zebrafish are also good model organisms because they are transparent, so we can watch the infection process develop.

Bioinformatics & Evolution: **Slide #9**

27. Show **Slide #10**, which provides job information for a microbiologist. Review this information with students.

CAREERS IN SPOTLIGHT:  
**Microbiologist**

**What do they do?**  
Microbiologists study microbes: bacteria, viruses, fungi, and protists. Dr. Ramakrishnan is an expert in tuberculosis, a type of bacteria that infects almost a third of humanity worldwide. She also studies immunology, including the body's reaction to or defense from microbes.

**What kind of training is involved?**  
Most Microbiologists who run their own lab have a Bachelor's degree and a PhD (which is usually 5–6 years of research training). However, each lab often employs scientists with diverse backgrounds, including people with Associate's, Bachelor's, and Master's degrees.

**What is a typical salary for a Microbiologist?**  
Associate's degree: \$35,000/year (\$17.50/hour)  
PhD, Full Professor: \$100,000/year or more (\$48.00/hour)

Source: Bureau of Labor and Statistics

28. Ask students, "What more do we know about microbiologists after today's lesson?" Point out that microbiologists perform genetic research to help them identify the causes of infectious diseases like influenza, tuberculosis, and HIV. They also measure genetic changes in these organisms over time, as seen in the example of influenza at the beginning of this lesson. Dr. Ramakrishnan studies tuberculosis and uses genetic research to manipulate the tuberculosis bacteria in an effort to understand more about how the bacteria grow and reproduce, how the immune system responds to tuberculosis infection, and to develop effective treatments.
29. Ask students to answer *Microbiologist Question #2* on their *Careers in the Spotlight* handout, which has students explain how this lesson has changed their understanding of the kind of work a microbiologist does.
30. Ask students to also answer *Microbiologist Question #3* on their *Careers in the Spotlight* handout, which has students explain how a microbiologist might use bioinformatics in his or her work.
31. Tell students to keep their *Careers in the Spotlight* handout available for future lessons.

## Homework

The following are suggested homework activities to follow this lesson:

- A. As homework, ask students to write about the things they learned in *Lesson Three* in their lab notebooks, on another sheet of paper, or in a word processing program like Microsoft® Notepad or Word which they then provide to the teacher as a printout or via email. This can serve as an entry ticket for the following class. Have them complete these prompts:
  - a. Today I learned that...
  - b. An important idea to think about is...
  - c. Something that I don't completely understand yet is...
  - d. Something that I'm really confident that I understand is....

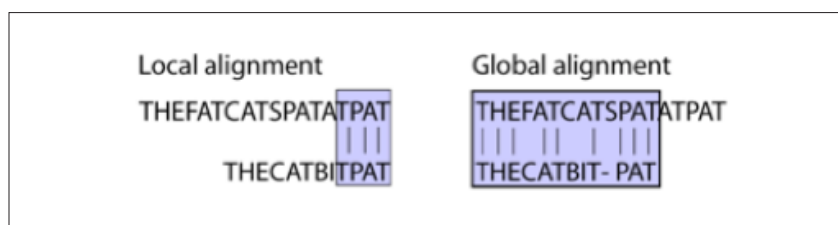
B. The *Lesson Three* Section of Student Handout—*The Process of Genetic Research* could also be assigned as homework.

### Extension

- Ask for volunteers from each student group to present their phylogenetic tree results to the class to support discussion in *Part II*.
- Students could pool DNA sequences across groups to construct larger multiple sequence alignments and phylogenetic trees. A file containing all DNA sequences used in this lesson can be found on the NWABR website. Butterfly (*Outgroup #5*) is a good outgroup to use for this larger analysis.
- Using the groups' phylogenetic trees, or a tree using DNA sequences from all of the groups, the teacher could remove some of the identifying names from the tree and have students propose positions for specimens.

### Teacher Background: Multiple Sequence Alignments

As described in *Part I*, the bioinformatics tool BLAST is used to compare sequences. It can compare a query sequence to a database of sequences, such as when searching the Nucleotide database at the NCBI (see *Lesson Two*). BLAST can also compare a query sequence to another specific sequence or a group of specified sequences. ClustalW2, on the other hand, compares each sequence in a collection of sequences to one another (i.e., because no particular reference sequence is specified). When performing sequence alignments, there are two ways in which the sequences are compared to one another: **local alignments** and **global alignments** (see **Figure 1**). Local alignments, like the **algorithms** in BLAST (the Basic Local Alignment Search Tool), break both the reference sequence and the database sequences into shorter strings of text called “words,” and then compare all of these words to one another. Global alignment algorithms, on the other hand, try to align every residue in every sequence, and are useful when comparing sequences that are more similar and are of roughly the same size. ClustalW2, the program used in this exercise, uses the global alignment approach.



**Figure 1:** Local vs. Global Alignments.

For more information, see the **Appendix** section “Understanding BLAST” and the article “Phylogeny for the faint of heart: A tutorial” cited in the *Resources* section at the end of this lesson.

**Local alignment:** A method to generate DNA or protein sequence alignments by breaking both the query or reference sequence, and all the sequences in the searched database, into short “words,” and then comparing all of these words to one another. This method is the one used by BLAST (Basic Local Alignment Search Tool) and is useful when trying to match sequences that are quite different from one another. This method is in contrast to the global alignment approach.

**Global alignment:** A method to generate DNA or protein sequence alignments by comparing one entire reference or query sequence to another sequence. Local alignments, in contrast, work by looking for short regions where sequences are highly related.

**Algorithm:** A detailed, unambiguous set of instructions to carry out a given task.

**Neighbor joining tree:** A method for making phylogenetic trees with DNA or protein sequences that involves calculating the percent difference between each pair of sequences, and using those percent differences to construct the phylogenetic tree.

**Maximum likelihood analysis:**

A more complex method than neighbor joining for making phylogenetic trees that takes into account various statistical properties of the data, including the probability of particular mutations occurring.

**Teacher Background: Phylogenetic Trees**

**Neighbor joining trees**, which are the type of phylogenetic trees used in this lesson, are a common method used in phylogenetics. Other methods, such as **maximum likelihood analysis**, are more robust, but require high levels of computational power and are more intensive than required for this lesson. Neighbor joining takes two sequences, counts the number of differences between them, divides it by the total sequence length, and calculates a percent difference. This method assumes the best answer is the one that requires the smallest amount of change and hence discards all other variations. Likelihood methods are the most powerful ones simply because they work with models of nucleotide substitution that take into account as much variation as theoretically possible, and this is what makes them computationally intensive. It is important to note that many times the kind of algorithm used (such as neighbor joining or maximum likelihood) can give slightly different trees.

## Glossary

**Algorithm:** A detailed, unambiguous set of instructions to carry out a given task.

**BLAST:** Basic Local Alignment Search Tool. A bioinformatics tool used to compare DNA or protein sequences to one or more other sequences, or to compare a DNA or protein sequence to a collection of sequences found in databases, such as the Nucleotide or Protein databases at the NCBI.

**Branch:** Horizontal lines on a phylogenetic tree that connect different nodes to one another.

**ClustalW2:** A program that performs multiple sequence alignments.

**Cluster:** Grouping together of two or more sequences in a phylogenetic tree that indicates the two sequences are closely related to one another.

**Consensus:** Consensus means “agreement” and is used in bioinformatics to describe a case in which two or more aligned DNA or protein sequences have the same amino acid or nucleotide in a given position. In the words “CAT” and “BAT,” the “AT” in each word would be the consensus.

**Consensus sequence:** A way of representing the results of a multiple sequence alignment by showing the amino acid or nucleotide found most often at each position in the alignment. In the alignment of the words “CAT” and “BAT,” the consensus sequence would be “+AT.” The plus indicates that there is no consensus at the position of the first letter.

**External node:** On a phylogenetic tree, the external nodes are at the ends of each branch, with each external node representing a different species in the tree.

**Global alignment:** A method to generate DNA or protein sequence alignments by comparing one entire reference or query sequence to another sequence. Local alignments, in contrast, work by looking for short regions where sequences are highly related.

**Internal node:** On a phylogenetic tree, the internal nodes are points where branches split, and represent an inferred evolutionary common ancestor.

**In silico:** An expression used to mean “performed on computer or via computer simulation.”

**JalView:** A program provided by the University of Dundee in Scotland that helps users generate multiple sequence alignments by interfacing with the alignment program ClustalW2. The program also allows users to edit their alignments, helping with their genetic analyses.

**Local alignment:** A method to generate DNA or protein sequence alignments by breaking both the query or reference sequence, and all the sequences in the searched database, into short “words,” and then comparing all of these words to one another. This method is the one used by BLAST (Basic Local Alignment Search Tool) and is useful when trying to match sequences that are quite different from one another. This method is in contrast to the global alignment approach.

**Maximum likelihood analysis:** A more complex method than neighbor joining for making phylogenetic trees that takes into account various statistical properties of the data, including the probability of particular mutations occurring.

**Multiple sequence alignment:** The process of comparing two or more DNA or protein sequences to one another by aligning the sequences and looking for similarities and differences.

**Mutation:** A change in a DNA or protein sequence.

**Neighbor joining tree:** A method for making phylogenetic trees with DNA or protein sequences that involves calculating the percent difference between each pair of sequences, and using those percent differences to construct the phylogenetic tree.

**Outgroup:** A sequence included in phylogenetic trees from an older or more distantly related species, used as a point of reference and to mark the evolutionary time. Inclusion of an outgroup is said to **root** the tree.

**Phylogenetic tree:** A branching diagram or “tree” showing the evolutionary relationships among various species, based upon similarities and differences in their physical and/or genetic characteristics. The lengths of the branches on the tree are proportional to the amount of time since the two organisms (or sequences) diverged from one another.

**Reference sequence:** A sequence that has been chosen for the purpose of comparison. At the National Center for Biotechnology Information (NCBI), reference sequences are chosen because they are of high quality and are thought to accurately represent the sequence from the original organism. In genetic testing, a reference sequence is a known and well-studied DNA or protein sequence that does not contain any mutations or changes that are associated with disease. These sequences are used for comparison with patient sequences during genetic testing.

**Rooted tree:** Phylogenetic trees that contain an outgroup are said to be “rooted.”

## Resources

For an online tutorial on making phylogenetic trees, visit “A Beginner’s Guide to Phylogenetic Trees” in the “Tutorials” section at Digital World Biology: <http://www.digitalworldbiology.com/dwb/Home.html>

Baldauf, Sandra. Phylogeny for the faint of heart: A tutorial. *Trends in Genetics*. 2003; 19: 345-351.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T., Higgins, D., and J. Thompson. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*. 2004; 31: 3497-3500.

Clamp, M., Cuff, J., Searle, S., and G. Barton. Jalview Java alignment editor. *Bioinformatics Applications Notes*. 2004; 20: 426-427.

## Credit

The authors wish to thank Wikimedia Commons, ClustalW2, JalView, and NCBI BLAST for many of the images found in the PowerPoint slides and handouts associated with this lesson.

Ramakrishnan, Lalita. Personal Interview. 13 July 2010.



# 3 Using Bioinformatics to Study Evolutionary Relationships Instructions

## Student Researcher Background:

### Making and Using Multiple Sequence Alignments

One of the primary tasks of genetic researchers is comparing DNA sequences to one another. In this exercise, you will collaborate with your other group members and use the DNA sequence data you obtained in *Lesson Two* to address your research question and support or reject your research hypothesis. Analyzing small pieces of DNA by hand is possible, but for complex analyses involving multiple long DNA sequences, scientists use bioinformatics tools like ClustalW2 and JalView. This not only makes the process faster, but allows scientists to quantify their results.

**Aim:** Today, your job as a researcher is to:

1. Generate a **multiple sequence alignment** to compare the DNA data within your group.
2. Create a **phylogenetic tree** from your sequence alignment to study the evolutionary relatedness of the organisms within your group.

### Multiple sequence alignment:

The process of comparing two or more DNA or protein sequences to one another by aligning the sequences and looking for similarities and differences.

### Phylogenetic tree:

Phylogenetics is the study of evolutionary relationships among organisms. A phylogenetic tree is a branching diagram or "tree" showing the evolutionary relationships among various species, based upon similarities and differences in their physical and/or genetic characteristics.



**Instructions:** Write the answers to your questions on Student Handout—*Using Bioinformatics to Study Evolutionary Relationships Worksheet*, in your lab notebook, or on a separate sheet of paper, as instructed by your teacher.

## PART I: Generating Your Multiple Sequence Alignment in JalView



1. At the top of your answer sheet or in your lab notebook, re-write your research hypothesis from *Lesson Two* (you can find this on Student Handout—*The Process of Genetic Research* or Student Handout—*Using BLAST and BOLD for Genetic Research, Question #23*).
2. Go to the Bio-ITEST website and click on the **Resources** tab (black box, **Figure 1**): <http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research>.
3. Obtain the **Group DNA** sequence file for your group by clicking on the **DNA** link beside your Group number (black arrow, **Figure 1**).
4. Copy all of your group's DNA sequences, including the species names and the carets ">". Be sure to select **all** of the sequences.

Using Bioinformatics: Advanced: Genetic Research

bio-itest

This is the second of a two-part series in NWABR's bioinformatics curriculum, funded by a grant called Bio-ITEST. ITTEST grants are for Innovative Technology Experiences for Students and Teachers, from the National Science Foundation (NSF). The three-year grant provides funding for education outreach programs that help secondary school teachers and their students learn about how information technology is used in biological research.

Major collaborators include Digital World Biology, EdLab Group, and Shoreline Community College. The program also draws on NWABR's strong relationships with school districts, community colleges, and NWABR member research institutions.

OVERVIEW

LESSONS

RESOURCES

LINKS

EVENTS

Resource Materials

Lesson Two: DNA Barcoding and the Barcode of Life Database

Unknown DNA Sequences:

1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 10

11 - 12 - 13 - 14 - 15 - 16 - 17 - 18 - 19 - 20

21 - 22 - 23 - 24 - 25 - 26 - 27 - 28 - 29 - 30

Lesson Three: Using Bioinformatics to Study Evolutionary Relationships AND

Lesson Four: Using Bioinformatics to Analyze Protein Sequences

Group Sequences:

Group	DNA	Protein	Outgroup
1	DNA	Protein	Outgroup
2	DNA	Protein	Outgroup
3	DNA	Protein	Outgroup
4	DNA	Protein	Outgroup
5	DNA	Protein	Outgroup

**Figure 1:** The Bio-ITEST **Resources** Page for Advanced Bioinformatics. Source: NWABR.



EMBL-EBI  Enter Text Here  [Help](#) | [Feedback](#)

[Databases](#) [Tools](#) [Research](#) [Training](#) [Industry](#) [About Us](#) [Help](#) [Site Index](#)

- Help
- FAQ
- Clustal website
- Jalview
- Programmatic Access
- Download

Related Applications

- Pairwise Sequence Alignment
- Multiple Sequence Alignment
- Phylogeny

Clustal related literature

Search for Clustal related literature in Medline... [more](#)

EBI > Tools > Multiple Sequence Alignment > ClustalW2

### ClustalW2 - Multiple Sequence Alignment

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

New version! Clustal Omega is now available - give it a try!

Use this tool

**STEP 1 - Enter your input sequences**

Enter or paste a set of  sequences in any supported format:

```
ATGTTGCGCGACGCTGCTCTACAAACCACAAAGATATTGGAACACTATACCTACTATTGCGCG
CATGAGCTGGAGTCTCTGGGCACAGCCCTAAGTCTCCTTATTGCGGCTGAACTAGGCCAACCGCAACCT
TCTAGGTAATGACCACATCTACAATGTCATGTCACAGCCCATGCAITCGTAATAATCTTCTCATAGTA
ATGCCTATTATAATCGGAGGCTTTGGCAACTGGCTAGTTCCTTGATAATTGGTGCCCGGACATGGCAT
TCCCCCGCATAAACACATAAGCTTCTGGCTCCTGCCCCCTTCTCCTACTCCTACTTGCATCTGCCAT
AGTAGAAGCCGCGCGCGGAACAGGTGGAACAGTCTACCCTCCCTTAGCGGGAACTACTCGCATCCTGGA
GCCTCCGTAGACCTAACCATCTTCTCCTTGCATCTGGCAGGCGTCTCCTCTATCCTAGGAGCCATTAACT
```

Or, upload a file:

**STEP 2 - Set your Pairwise Alignment Options**

Alignment Type: ☒ Slow ☐ Fast

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

**STEP 3 - Set your Multiple Sequence Alignment Options**

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

**STEP 4 - Submit your job**

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Figure 2: Entering Sequences in ClustalW2. Source: EBI ClustalW2.

EMBL-EBI  Enter Text Here  [Help](#) | [Feedback](#)

[Databases](#) [Tools](#) [Research](#) [Training](#) [Industry](#) [About Us](#) [Help](#) [Site Index](#)

- Help
- FAQ
- Jalview

Related Applications

- Multiple Sequence Alignment
- Phylogeny

EBI > Tools > Multiple Sequence Alignment > ClustalW2

### ClustalW2 Results

[Alignment](#) [Result Summary](#) [Guide Tree](#) [Submission Details](#) [Submit Another Job](#)

[Download Alignment File](#) [Show Colors](#)

CLUSTAL 2.1 multiple sequence alignment

Chimpanzee	ATGTTGCGCGACGCTGACTATTCTCTACAAACCACAAAGATATTGGAACACTATACTTA 60
Gorilla	ATGTTGCGCGACGCTGATTATTCTCTACAAACCACAAAGATATTGGAACACTATACTTA 60
Orangutang	ATGTTGCGCGACGCTGGCTATTCTCCAGAACCAAAAGATATTGGAACGCTATACTTG 60
SpiderMonkey	ATGTTGATAACTGCTGATTATTCTCAACCAACCAAAAGACATCGGAACACTATACTTA 60
LeafMonkey	-----GGTTATTCTCTACAAACCACAAAGATATTGGAACCTTATACTTA 44
RhesusMacaque	ATGCTCAITAAATGCTGACTCTTTTCAACAAATCACAAAGACATTGGAACCTGTATTTA 60

Figure 3: ClustalW2 Alignment. Source: EBI ClustalW2.

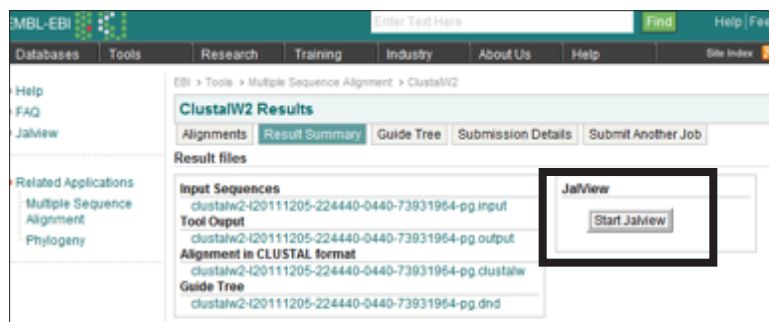


Figure 4: ClustalW2 Alignment Results Summary. Source: EBI ClustalW2.

- Open a new tab in your browser or a new browser window and go to ClustalW2 at the European Bioinformatics Institute (EBI) at: <http://www.ebi.ac.uk/Tools/msa/clustalw2/>.
- Paste your **Group DNA** sequences into the sequence box in STEP 1 (black arrow, **Figure 2**).
- Select **DNA** from the drop down menu (black box, **Figure 2**).
- Click **Submit** (gray arrow, **Figure 2**).

**Consensus:** Consensus means “agreement” and is used in bioinformatics to describe a case in which two or more aligned DNA or protein sequences have the same amino acid or nucleotide in a given position. In the words “CAT” and “BAT,” the “AT” in each word would be the consensus.

**Consensus sequence:** A way of representing the results of a multiple sequence alignment by showing the amino acid or nucleotide found most often at each position in the alignment. In the alignment of the words “CAT” and “BAT,” the consensus sequence would be “+AT.” The plus indicates that there is no consensus at the position of the first letter.

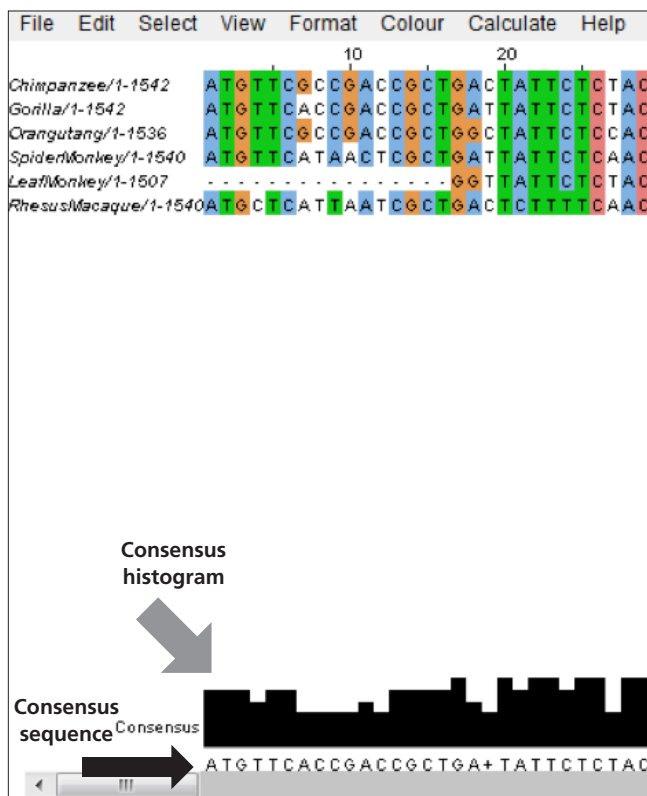
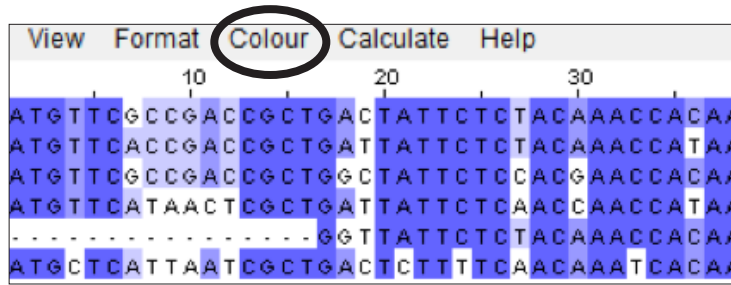


Figure 5: Multiple Sequence Alignment in JalView. Source: JalView.

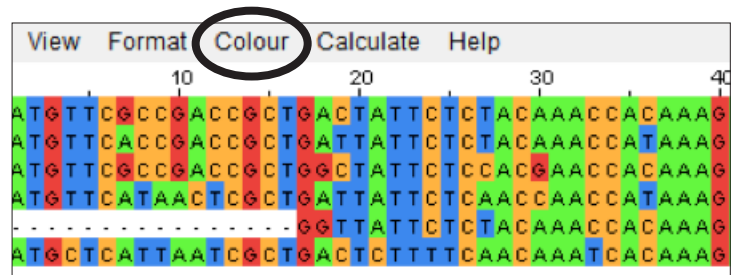
- When the multiple sequence alignment is complete, you will see all of your group’s DNA sequences aligned with one another (**Figure 3**). Next, you will use the JalView program to help you analyze your multiple sequence alignment.
- Click **Result Summary** from the top menu bar (black box, **Figure 3**). This will take you to a page that includes a summary of your multiple sequence alignment results (**Figure 4**).
- Click **Start JalView** (black box, **Figure 4**).
- When JalView opens, you will see all of your sequences aligned with one another, and highlighted with different colors. The regions of black boxes below your sequence alignment is the “**Consensus Histogram**,” which is a graphical representation of the number of sequences that have the same nucleotide at a given position (gray arrow, **Figure 5**). The **consensus sequence** is shown below the consensus histogram (black arrow), which is the “average” or most frequent nucleotide at each position.

13. From the **Colour** menu select **Percent Identity** [also called **Percentage Identity**]. This highlights similarities and differences based on sequence identity, as shown in **Figure 6**. **Dark blue** highlights areas of **consensus** (where all the sequences are the same). **Lighter blue** is used to color-code regions where some, but not all, of the sequences are the same. Regions of difference remain **white**.



**Figure 6:** Color Coding Regions Using the Percent Identity Selection. Source: JalView.

14. From the **Colour** menu select **Nucleotide**. This is another color coding approach which highlights each base as a different color, as shown in **Figure 7**.



**Figure 7:** Color Coding Regions Using the Nucleotide Selection. Source: JalView.



15. Which format do you prefer for analyzing the similarities and differences among your sequences: Percent Identity or Nucleotide? Explain your answer.



16. Select your preferred format (either "Percent Identity" or "Nucleotide").



17. Look at the **consensus sequence** below your multiple sequence alignment (black arrow in **Figure 5**). In some positions in the consensus sequence, there is a plus sign "+" instead of a base. What do you think this "+" sign represents?



18. What do all of the "+" base positions have in common? [Hint: How many sequences in your alignment have each base at this position? Be sure to select and copy **all** of the sequences.]



19. Scroll through your multiple sequence alignment using the scroll bar on the right. What can you tell from the color coded sequences? Do some sequences appear to be more similar than others?



20. Based on your answer to the question above, can you use this information about sequence similarities to draw any conclusions about relatedness among the species in your group?

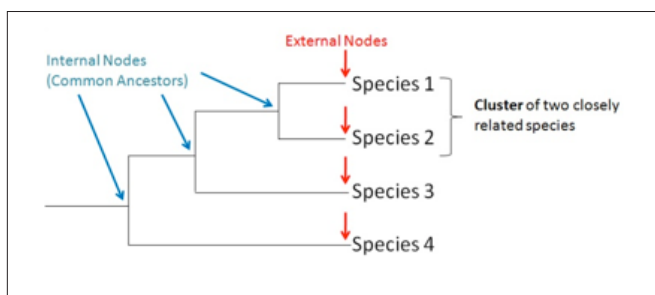
21. Compare your answer to the previous question with the answers of your collaborators in your group. Did you all come to the same conclusion? If not, why do you think that is?

### PART II: Using Phylogenetic Trees to Evaluate Evolutionary Relationships

In the previous exercise, you estimated species relatedness based on the similarities and differences you observed in the multiple sequence alignment. Multiple sequence alignment programs like ClustalW2 take these similarities and differences into account when generating the sequence alignments. Another useful way to present the number of differences (or the percent difference) among sequences is with a graphical representation called a **phylogenetic tree**.

Sequences that are most closely related to one another will be grouped together in a **cluster**. In **Figure 8**, Sequences 1 and 2 form one cluster. Each horizontal line represents a **branch**, and each site where two branches split is called a node. **Internal nodes** represent common ancestors, and each species is located on its own **external node**. By including an **outgroup** in your analysis (in this case, Species 4, from the brown algae *Laminaria setchellii*), you will have a point of reference within your tree, and can make conclusions about evolutionary relatedness.

We will use the program BLAST (Basis Alignment Search Tool) to make our phylogenetic tree with our pre-selected outgroups.



**Figure 8:** Anatomy of a Phylogenetic Tree. Source: NCBI BLAST.

**Lesson Three: Using Bioinformatics to Study Evolutionary Relationships**

AND

**Lesson Four: Using Bioinformatics to Analyze Protein Sequences**

**Group Sequences:**

Group	DNA	Protein	Outgroup
1	DNA	Protein	Outgroup
2	DNA	Protein	Outgroup
3	DNA	Protein	Outgroup
4	DNA	Protein	Outgroup
5	DNA	Protein	Outgroup

**Figure 9:** Click the **Outgroup** sequence for your group. Source: NWABR.

**Phylogenetic tree:** A branching diagram or “tree” showing the evolutionary relationships among various species, based upon similarities and differences in their physical and/or genetic characteristics. The lengths of the branches on the tree are proportional to the amount of time since the two organisms (or sequences) diverged from one another.

**Cluster:** Grouping together of two or more sequences in a phylogenetic tree that indicates the two sequences are closely related to one another.

**Branch:** Horizontal lines on a phylogenetic tree that connect different nodes to one another.

**Internal node:** On a phylogenetic tree, the internal nodes are points where branches split, and represent an inferred evolutionary common ancestor.

**External node:** On a phylogenetic tree, the external nodes are at the ends of each branch, with each external node representing a different species in the tree.

- Go to the BLAST homepage: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>.
- Select **nucleotide blast**.
- Open a new browser window or tab, and obtain the **Outgroup** sequence file **for your group** from the Bio-ITEST website **Resources** tab (**Figure 9**). The address for the Bio-ITEST website is: <http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research>.
- On the BLAST site, check the box that says **Align two or more sequences** (black circle in **Figure 10**).
- Paste the **Outgroup** sequence in the top text box (Query Sequence text box, gray arrow in **Figure 10**).

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

ATTTGGGTAAACCTTCTTCCCAACACTTCTTGGCTATCTGGGATGCCCCGACGTTACT  
CGGACTACCCGATGCATATACCATGAATGTCTATCATCCGTAGGCTCATTATCTCCCTGAC  
AGCAGTAATATTAAATTTTCATGATTGAGAGCCTTTGCTTCAAAACGAAAAGTCTTAATAGTA  
GAAGAGCCCTCCACAACCTGGAGTGACTATATGGATGCCCTCCACCTACCAACATTGAAAGAC  
CGTATACATAAAATCTAGA

Clear Query subrange

From

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

☒ Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence

GTTAATCTAACCTTCTTCCCAACACTTCTTGGCTATCTGGAATACCCGACGTTACTCGGACT  
ACCCGATGCATATACCATGAATATCTGTCATCCGTGGGCTCATTCAITTCCTTAACAGCAGT  
AATATTAAATTTTATAATCTGAGAGCCTTCCGCTCAAAACGAAAAGTCTTAATATCGAAGAA  
CCCTCCACAATCTGGAGTGACTGTATGGATGCCCTCCACCTATCATACATTTGAAGAGCCTGTAT  
ATATAAGTCTAAA

Clear Subject subrange

From

Or, upload file Browse...

Figure 10: Entering Sequences into BLAST to Make a Phylogenetic Tree. Source: NCBI BLAST.

Program Selection

Optimize for

Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

☒ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search nucleotide sequence using Blastn (Optimize for somewhat similar sequences)

☒ Show results in a new window

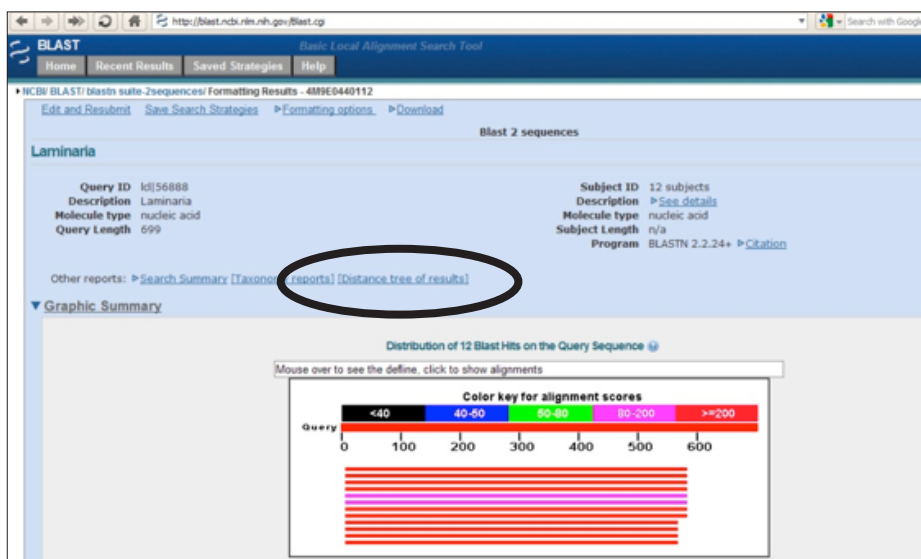
+ Algorithm parameters

Figure 11: Use the *Somewhat similar sequences (blastn)* Option to Compare Sequences. Source: NCBI BLAST.

27. Open your **Group DNA** sequences from the Bio-ITEST website, copy them, and paste all of your Group DNA sequences in the bottom text box (Subject Sequence text box, black arrow in **Figure 10**). [Note: Be sure that each sequence contains the information at the beginning of each FASTA-formatted sequence (>name\_of\_species) **with no spaces in the name**. Be sure to select and copy **all** of the sequences.]
28. Choose **Somewhat similar sequences (blastn)** from the **Program Selection** menu (black arrow in **Figure 11**).
29. Click the **blue triangle** or **plus sign** beside the **Algorithm Parameters** menu to adjust the BLAST settings (green box in **Figure 11**).

30. From the **Algorithm Parameters** menu, make the following changes as seen in **Figure 12**:
  - a. Uncheck the **Short queries** box.
  - b. Change the **Word size** to 7 in the drop down menu.
  - c. Uncheck the Filter box for **Low complexity regions**.
  - d. Uncheck the Mask box for **Mask for lookup table only**.
31. Click **BLAST** (blue button at the bottom of the page as shown by the black arrow in **Figure 12**).
32. When BLAST has completed the comparisons, you will see a Results window similar to the one you saw in *Lesson Two* when you identified your unknown DNA sequence. Near the top of the page, click the link for the **Distance tree of results** (black circle in **Figure 13**).

**Figure 12:** Click the **Algorithm Parameters** Menu to Change the Default Settings for Your BLAST Analysis. Source: NCBI BLAST.



**Figure 13:** Select **Distance tree of results** to View Your Phylogenetic Tree. Source: NCBI BLAST.



# LESSON 3

## CLASS SET



33. On your answer sheet or in your lab notebook, draw your phylogenetic tree. Be sure to write the names of the species beside the appropriate branches.



34. Based on your phylogenetic tree, which species appear to be most closely related to one another (which species cluster closest together)?



35. What species is/are most distantly related (other than your outgroup)?



36. Is this the same conclusion you reached in your analysis of the multiple sequence alignment? Why or why not?



37. Do these data support or refute your hypothesis, or is your analysis inconclusive?



38. Discuss your results with the collaborators in your group. Did you all reach the same conclusions about species relatedness? Why or why not?



Name \_\_\_\_\_ Date \_\_\_\_\_ Period \_\_\_\_\_

# 3

## Using Bioinformatics to Study Evolutionary Relationships Worksheet

**Aim:** Today, your job as a researcher is to:

1. Generate a **multiple sequence alignment** to compare the DNA data within your group.
2. Create a **phylogenetic tree** from your sequence alignment to study the evolutionary relatedness of the organisms within your group.



**Instructions:** Use Student Handout—*Using Bioinformatics to Study Evolutionary Relationships* Instructions to complete this worksheet.

### PART I: Generating Your Multiple Sequence Alignment in JalView

1. Hypothesis:
15. Which format do you prefer for analyzing the similarities and differences among your sequences: Percent Identity or Nucleotide? Explain your answer.
17. In some positions in the consensus sequence, there is a plus sign "+" instead of a base. What do you think this "+" sign represents?
18. What do all of the "+" base positions have in common? [**Hint:** How many sequences in your alignment have each base at this position?]
19. Scroll through your multiple sequence alignment using the scroll bar on the right. What can you tell from the color coded sequences? Do some sequences appear to be more similar than others?
20. Based on your answer to the question above, can you use this information about sequence similarities to draw any conclusions about relatedness among the species in your group?

21. Compare your answer to the previous question with the answers of your collaborators in your group. Did you all come to the same conclusion? If not, why do you think that is?

### PART II: Using Phylogenetic Trees to Evaluate Evolutionary Relationships

33. Draw your phylogenetic tree. Be sure to write the names of the species beside the appropriate branches.

34. Based on your phylogenetic tree, which species appear to be most closely related to one another (which species cluster closest together)?

35. What species is/are most distantly related (other than your outgroup)?

36. Is this the same conclusion you reached in your analysis of the multiple sequence alignment? Why or why not?

37. Do these data support or refute your hypothesis, or is your analysis inconclusive?

38. Discuss your results with the collaborators in your group. Did you all reach the same conclusions about species relatedness? Why or why not?

# 3 Using BLAST and BOLD for Genetic Research

## Teacher Answer Key

[**Note:** The suggested total point value for this worksheet is **25 points**, or 1 point for restating their hypothesis and 2 points per question thereafter.]

### PART I: Generating Your Multiple Sequence Alignment in JalView

1. Hypothesis:

Students should restate their hypothesis from *Lesson Two*.

15. Which format do you prefer for analyzing the similarities and differences among your sequences? Explain your answer.

This is largely a personal preference for the students; for some students, the representation of bases with different colors (Nucleotide coloring) will allow for easier visualization of similarities and differences, while others will prefer the blue shading of the Percent Identity option. However, the question should encourage them to reflect on the fact that the DNA data is being represented or interpreted in a graphical way that makes it possible to make generalizations or conclusions about the similarity or relatedness of the sequences.

17. In some positions in the consensus sequence, there is a plus sign “+” instead of a base. What do you think this “+” sign represents?

Students may notice that the plus sign indicates cases in which a consensus sequence cannot be determined. However, they may require the “hint” provided in the following question.

18. What do all of the “+” base positions have in common? [**Hint:** How many sequences in your alignment have each base at this position?]

The plus signs indicate cases in which two sequences have one base in common while the other two sequences have a different base in common. Therefore, a consensus sequence cannot be determined at this position because half of the sequences share one base and half of the sequences share another.

19. Scroll through your multiple sequence alignment using the scroll bar on the right. What can you tell from the color coded sequences? Do some sequences appear to be more similar than others?

Students should be able to determine that some of the sequences are more similar to one another than others.

20. Based on your answer to the question above, can you use this information about sequence similarities to draw any conclusions about relatedness among the species in your group?

This will vary by Group as shown below, but it may be difficult for students to make any meaningful conclusions based solely on the multiple sequence alignment. That is one of the reasons genetic researchers make phylogenetic trees, instead of looking at multiple sequence alignments alone. If students can make conclusions, these answers may vary. Conclusions based on the phylogenetic trees (*Part III*) will be easier for students to draw.

**Group 1:** Gorillas and Chimpanzees are the most closely related, followed by Orangutans, and then the Monkeys.

**Group 2:** Chickens and Partridges appear closely related, as do Nutcrackers and Ravens. Herons and Penguins are the most distantly related.

**Group 3:** Chilepepper and Snapper appear to be the most closely related, followed by Soldierfish, Alfonsino, and Mackerels, with Coho Salmon being the most distantly related.

**Group 4:** Both the Dusky Shark and Whitecheek Sharks are closely related, as are the Great White and Mackerel Sharks. The two Stingrays are closely related, but appear more distantly related to the Sharks.

**Group 5:** The Nile Crocodile and American Alligator are closely related, followed by the Iguana and Komodo Dragon. Chameleons and Agamas appear to be the most distantly related.

21. Compare your answer to the previous question with the answers of your collaborators in your group. Did you all come to the same conclusion? If not, why do you think that is?

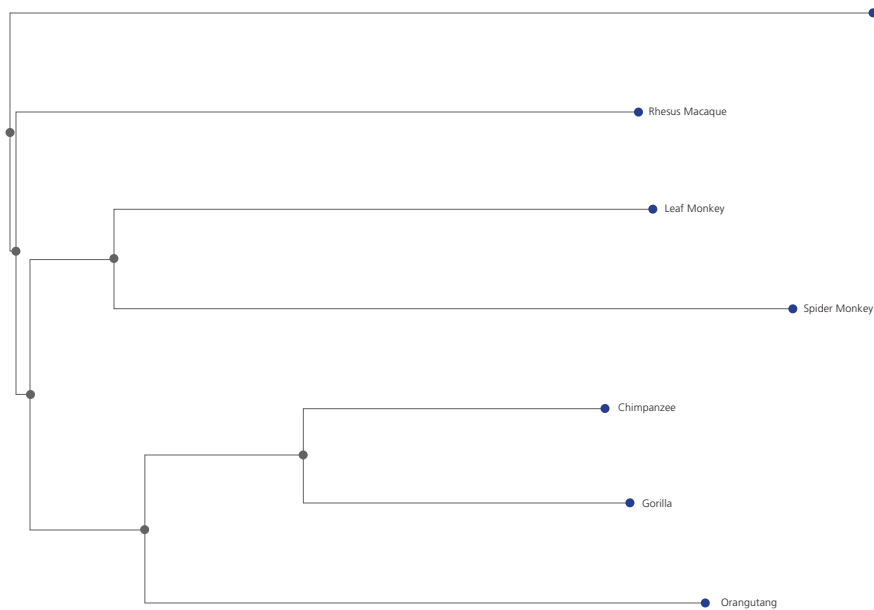
While each group was formed around the same general research question, each student formed his or her own hypothesis. Other group members may reach the same or different conclusions, based on their original hypotheses.

### Part II: Using Phylogenetic Trees to Evaluate Evolutionary Relationships

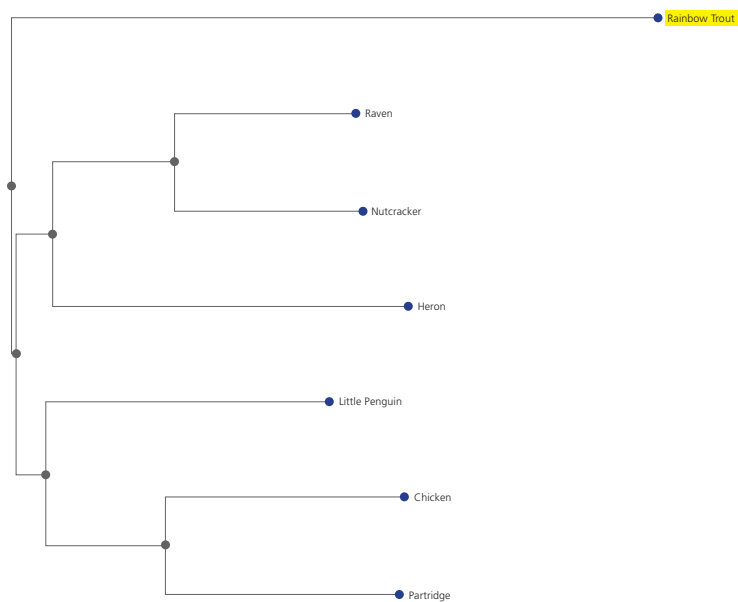
33. Draw your phylogenetic tree. Be sure to write the names of the species beside the appropriate branches.

Phylogenetic trees will differ for each group as shown below.

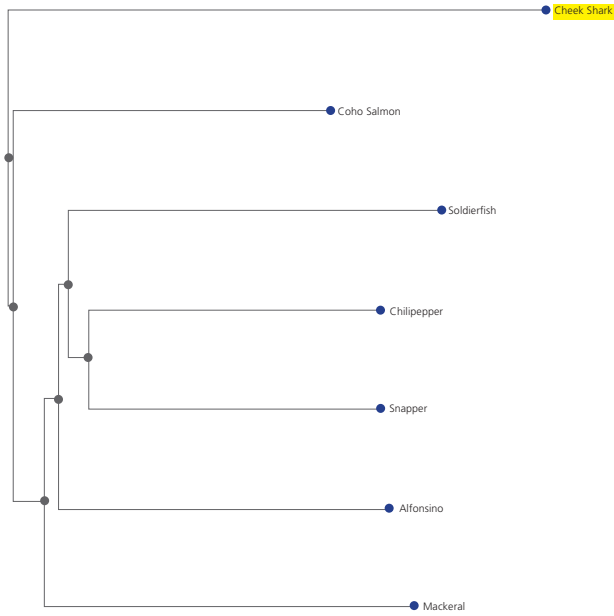
**Group 1: Class Mammalia (Primates).** Outgroup = Raven (*Corvus corax*).



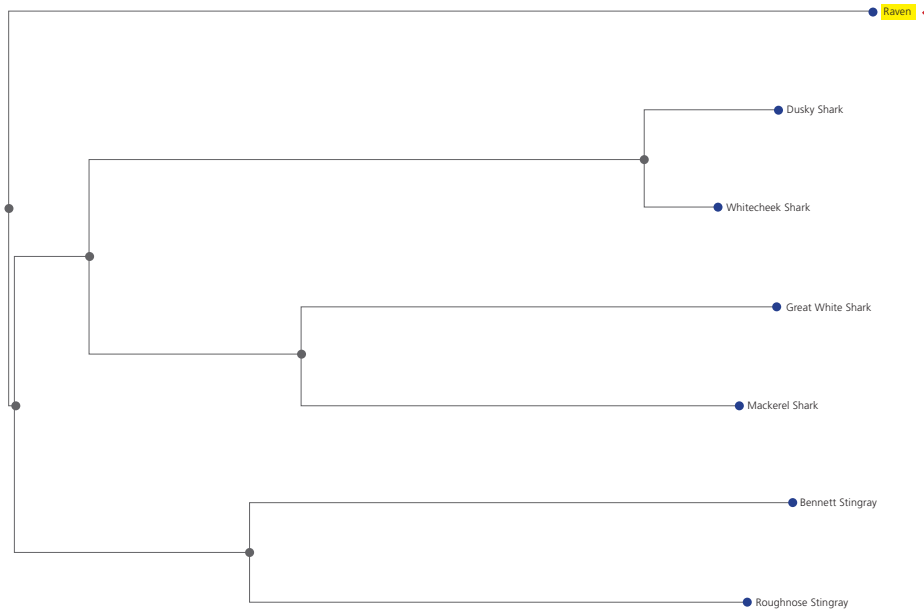
**Group 2: Class Aves (Birds).** Outgroup = Rainbow Trout (*Oncorhynchus mykiss*)



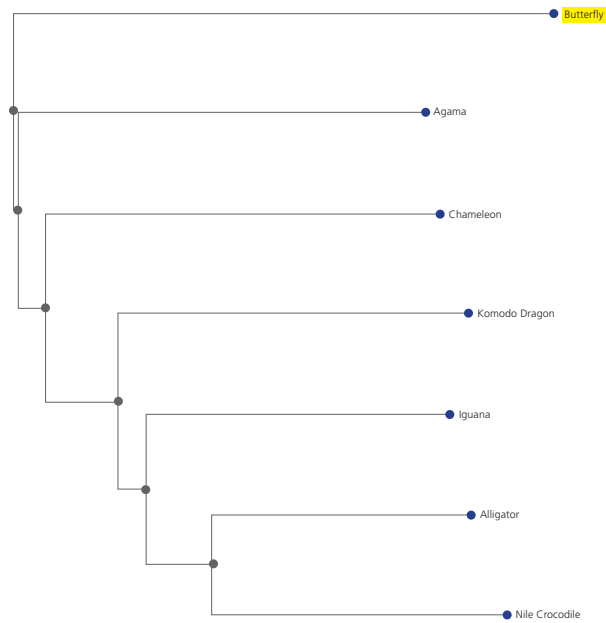
**Group 3:** Class Osteichthyes. (Bony Fishes). Outgroup = Whitecheek Shark (*Carcharinus dussumieri*)



**Group 4:** Class Chondrichthyes (or Class Elasmobranchii) (Cartilaginous Fishes). Outgroup = Raven (*Corvus corax*).



**Group 5:** Class Reptilia (Reptiles). Outgroup = Butterfly (*Lepidoptera*)



34. Based on your phylogenetic tree, which species are most closely related to one another (which species cluster closest together)?

This will vary by group as shown below:

**Group 1:** Gorillas and Chimpanzees are the most closely related, followed by Orangutans.

**Group 2:** Chickens and Partridges appear closely related, as do Nutcrackers and Ravens.

**Group 3:** Chilipepper and Snapper appear to be the most closely related, followed by Soldierfish, Alfonsino, and Mackerels.

**Group 4:** Both the Dusky Shark and Whitecheek Sharks are closely related, as are the Great White and Mackerel Sharks. The two Stingrays are closely related, but appear more distantly related to the Sharks.

**Group 5:** The Nile Crocodile and American Alligator are closely related, followed by the Iguana and Komodo Dragon.

35. What species is/are most distantly related (other than your outgroup)?

This will vary by group as shown below:

**Group 1:** The Rhesus Macaque, the Leaf Monkey, and the Spider Monkey are most distantly related.

**Group 2:** Herons and Penguins are the most distantly related.

**Group 3:** Coho Salmon are the most distantly related.

**Group 4:** The two Stingrays are closely related, but appear more distantly related to the Sharks.

**Group 5:** Chameleons and Agamas appear to be the most distantly related.



# LESSON 3

## KEY

36. Is this the same conclusion you reached in your analysis of the multiple sequence alignment? Why or why not?

Students should compare their answers above in *Part I Question #13* (analysis of Multiple Sequence Alignment) and *Part II Question #26* (Phylogenetic Tree). Some students may have been unable to draw conclusions in *Part I Question #13*.

37. Do these data support or refute your hypothesis, or is your analysis inconclusive?

In this question, students refer back to their original research question and hypothesis, and then evaluate their hypothesis in the context of their data analysis above. For instance, if their hypothesis was that Chimpanzees and Spider Monkeys were more closely related than Chimpanzees and Gorillas, they would reject their hypothesis.

38. Discuss your results with your collaborators in your group. Did you all reach the same conclusions about species relatedness? Why or why not?

Students within each group should reach the same general conclusions to *Questions #34* and *#35* above.