Updated: June 2021

The *Wolbachia* Project

Discover the Microbes Within!

Lab 5: Bioinformatics II

NCBI Sequence Taxonomy & BLAST Searching

Project Guide

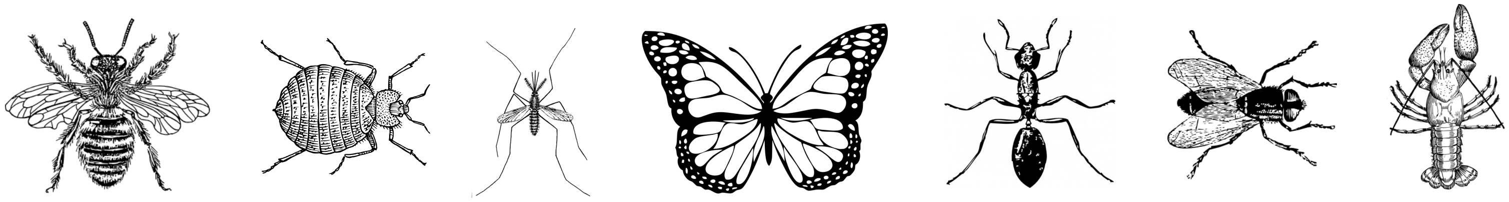


Table of Contents

The *Wolbachia* Project

Discover the Microbes Within!

The *Wolbachia* Project: Discover the Microbes Within! was developed by a collaboration of scientists, educators, and outreach specialists. It is directed by the Bordenstein Lab at Vanderbilt University.

https://www.vanderbilt.edu/wolbachiaproject

Introduction

> *"Understanding nature's mute but elegant language of living cells is the quest of modern molecular biology. From an alphabet of only four letters representing the chemical subunits of DNA, emerges a syntax of life processes whose most complex expression is man... The challenge is in finding new approaches to deal with the volume and complexity of data, and in providing researchers with better access to analysis and computing tools in order to advance understanding of our genetic legacy and its role in health and disease."*
>
> From the National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/

Goals

- *Module 1* (Pages 4-6): To show the ways in which the NCBI online database classifies and organizes information on DNA sequences, evolutionary relationships, and scientific publications.
- *Module 2* (Pages 7-13): To identify an unknown nucleotide sequence from an insect endosymbiont by using the NCBI search tool BLAST

Introduction

This exercise represents two interrelated modules designed to introduce students to modern biological techniques in the area of Bioinformatics. Bioinformatics is the application of computer technology to the management of biological information. The need for Bioinformatics has arisen from the recent explosion of publicly available genomic information, such as that resulting from the Human Genome Project. To address this, the National Center for Biotechnology Information (NCBI) was established in 1988 as a national resource for molecular biology information. The NCBI creates public-access databases, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. The NCBI is a virtual goldmine both in terms of available resources, and treasures yet to be discovered. We will investigate the GenBank DNA sequence database, which is responsible for organizing millions of nucleotide sequence records.

By completing this project, students will be exposed to the tools and databases currently used by researchers in molecular and evolutionary biology, and will gain a better understanding of gene analysis, taxonomy, and evolution.

Prerequisite Skills

While no computer programming skills are necessary to complete the modules in this work, prior exposure to personal computers and the Internet is assumed.

Teaching Time: One class period

Required Resources

- Computer with internet browser, such as Firefox or Chrome
- DNA Sequence Files: https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/

Recommended Background Tutorial

There are a number of online, educational resources devoted to learning bioinformatics. For details that summarize content covered in this exercise and more, see:

- BLAST for beginners: https://digitalworldbiology.com/tutorial/blast-for-beginners

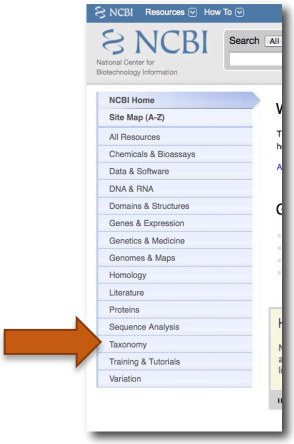
Module 1: Sequence Taxonomy

OBJECTIVE: The goal of this module is to introduce you to the number and diversity of nucleotide sequences in the NCBI database.

1. Begin by linking to the NCBI homepage at http://www.ncbi.nlm.nih.gov.
If you ever get lost, always return to this page as a starting point.

2. Select **Taxonomy** at the bottom of the left menu bar.

The NCBI Taxonomy database contains the names of more than 160,000 organisms whose sequences have been deposited in the NCBI databases. Only a small fraction of the millions of species estimated to exist on earth is represented!



3. Select the **Taxonomy** link under **DATABASES**.

4. Select the option **Statistics** under Taxonomy Tools.

Q1 For the 'Taxonomy Nodes (all dates)' column, how many Bacterial Species are in the sequence database? ________________________________

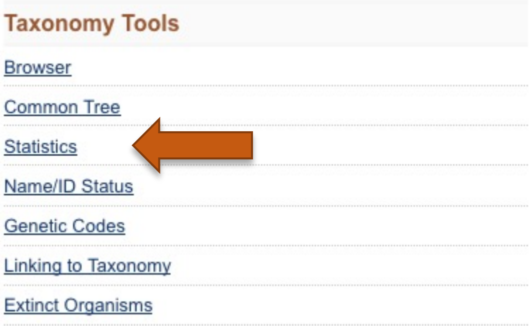
Q2 For the year 2015, how many Bacterial Species were added to the sequence database?

____

Q3 Using the '**Interval**' filter, how many Bacterial Species have been added over the past

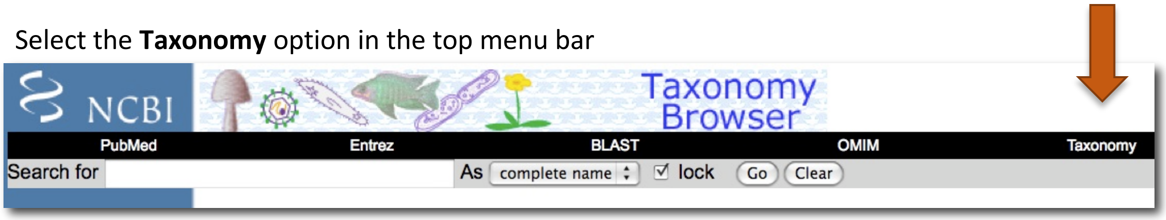
5 years? _________________________________

10 years? _________________________________



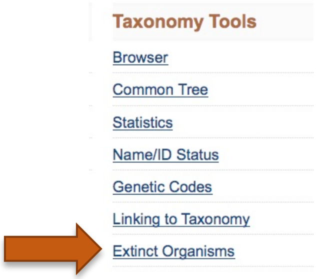
Interestingly, the sequence data from extinct organisms are even listed in the GenBank database. Let's look for a gene sequence from a 120 Mya old insect preserved in amber! To get back to your last webpage,

5. Select the **Taxonomy** option in the top menu bar



Continued on page 5…

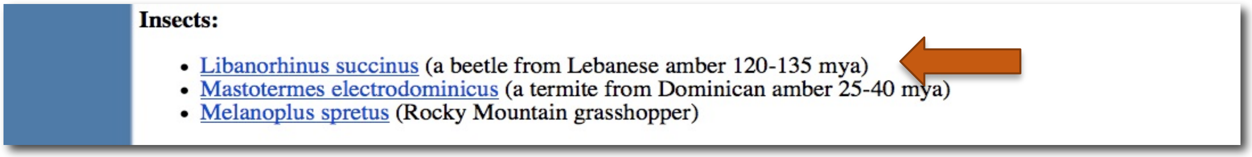
6. Select **Extinct organisms** under Taxonomy Tools



Taxonomy Tools

- Browser
- Common Tree
- Statistics
- Name/ID Status
- Genetic Codes
- Linking to Taxonomy
- Extinct Organisms

7. Scroll down to *Insects* on the main page and select ***Libanorhinus succinus*** **(**a beetle from Lebanese amber 120-135 Mya**)'.**

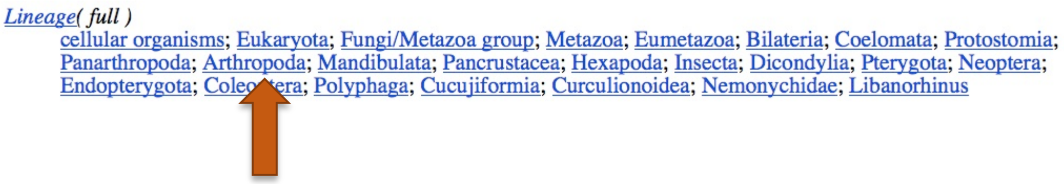


Insects:

- Libanorhinus succinus (a beetle from Lebanese amber 120-135 mya)
- Mastotermes electrodominicus (a termite from Dominican amber 25-40 mya)
- Melanoplus spretus (Rocky Mountain grasshopper)

This page gives you very specific information about the ancestry of this organism.

8. Select the option **Arthropoda** under *Lineage.*



Lineage(full)
cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Protostomia; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Coleoptera; Polyphaga; Cucujiformia; Curculionoidea; Nemonychidae; Libanorhinus

Q4 What are some other organisms that belong to this phylum of animals?

____

Q5 Can you think of any body traits that these organisms have in common?

____

____

Q6 Go back one page. How many 'Nucleotide' sequences have been deposited into the Entrez Records from this organism? (*Hint:* Look at the box on the top right labeled 'Entrez records')

____

Q7 What is the name of the gene that was sequenced for this organism (to find out, click on the number 1 next to nucleotide)?____

Q8 How does this relate to 16S rRNA? ____

Q9 How many nucleotide base pairs (bp) does this DNA entry contain? (*Hint:* the answer is in the first line just before DNA) ____

Continued on page 6…

9. Scroll through the complete reference report on this sequence.

A lot of information may seem confusing, but it is all there to provide scientists with as much information as possible about this sequence. This data is formatted into what is called a "flatfile". At the bottom of the screen, you will find the nucleotide sequence (all of the A,T,G,C base pairs in this gene) of this gene.

10. Click on the **PUBMED '8505978**' to directly link to the title, authors, and abstract of the published paper!

Amazing, now you can read the research article that discovered this nucleotide sequence.

Q10 What is the title of the research article that published this gene sequence?

____

____

11. Go back and select the **NCBI** link in the top left corner of the screen (next to the DNA symbol) to return to the NCBI home page.

Congratulations! You have completed Module 1.

Module 2: BLAST Searching

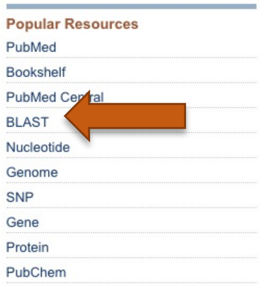
OBJECTIVE: The goal of this module is to retrieve genetic sequence data from the NCBI database that identifies the '*Wolbachia* sequence' you generated. The Basic Local Alignment Search Tool (BLAST) is an essential tool for comparing a DNA or protein sequence to other sequences in various organisms. Two of the most common uses are to a) determine the identity of a particular sequence and b) identify closely related organisms that also contain this particular DNA sequence.

A slide show introduction (optional): Begin by linking to a BLAST for beginners slide show that is simple and easy to follow (**https://digitalworldbiology.com/tutorial/blast-for-beginners**). Let the slide show guide your learning by clicking on the bright green arrow to proceed through the pages. It is meant to give a general feel for using BLAST and it is not necessary to complete the whole slide show.

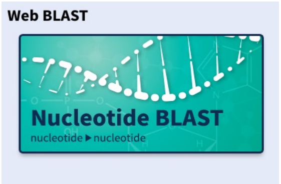
1. Begin by linking to the **NCBI** homepage.

2. Select **BLAST** in the right menu bar under "Popular Resources."

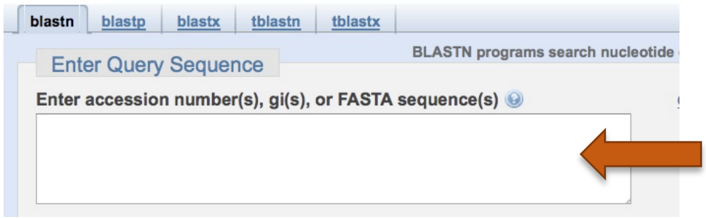
With your new knowledge of Sequence Searching and BLAST, let's begin with a sequence you make up and then your *Wolbachia* sequence.



3. Select **Nucleotide BLAST** under the Web BLAST category.

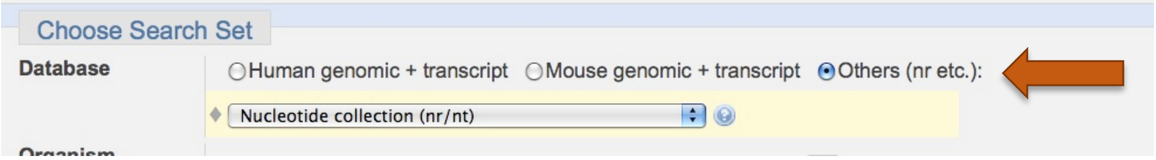


4. Input your own, random nucleotides (A,T,G,C) that fill one complete line in the blank box at the top under "Enter Query Sequence". Your sequence is referred to as the query sequence. (Make sure that the "Align two or more sequences box is unchecked.)


blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTN programs search nucleotide
Enter accession number(s), gi(s), or FASTA sequence(s)

Continued on page 8…

5. VERY IMPORTANT – Click on the circle for **Others (nr etc)** under “Choose Search Set.”



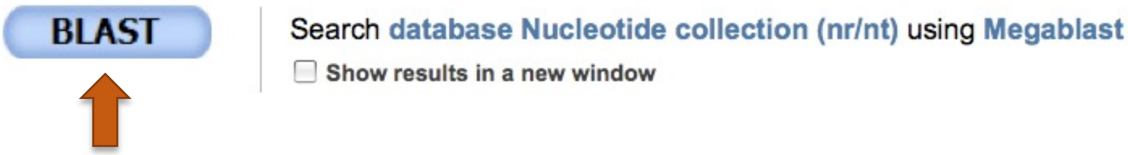
Choose Search Set

Database ○Human genomic + transcript ○Mouse genomic + transcript ●Others (nr etc.):

Nucleotide collection (nr/nt)

Organism

6. Page down and click **BLAST** at end of page. A new window appears.



BLAST

Search database Nucleotide collection (nr/nt) using Megablast

□ Show results in a new window

Wait for the results page to automatically launch. The wait time depends on the type of search you are doing and how many other researchers are using the NCBI website at the same time you are. Look at **“Search Summary”** once the run is finished.

Q1 Did your sequence have any significant similarity to anything in the NCBI databases? How do you determine significance? (Hint: A significant hit has an **E-value** below E-5 or E raised to the negative 5, a very small number). If there was no significant similarity, can your offer an explanation why?

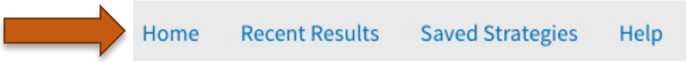
____

____

____

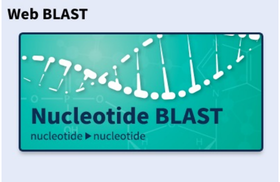
Q2 What was your E-value? ____

7. Select **Home** at the top right of the BLAST page to return.



Home Recent Results Saved Strategies Help

8. On the BLAST home page, select **Nucleotide BLAST** under the Web BLAST category.



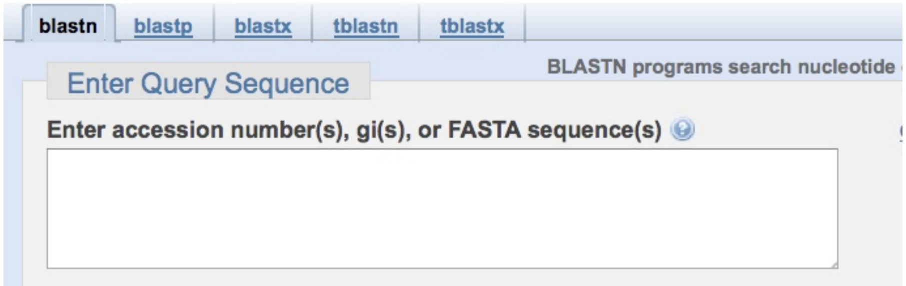
Web BLAST

Nucleotide BLAST

nucleotide ▸ nucleotide

Continued on page 9…

9. Download (or copy) the *Wolbachia* sequence below and enter in the Search box. (Make sure "Align two or more sequences" is unchecked.)


blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTN programs search nucleotide
Enter accession number(s), gi(s), or FASTA sequence(s)

https://www.vanderbilt.edu/wolbachiaproject/lab-5-dna-sequences/

```
GTTGCAGCAATGGTAGACTCAACGGTAGCAATAACTGCAGGACCTAGAGGAAAAACAGTAGGG
ATTAATAAGCCCTATGGAGCACCAGAAATTACAAAAGATGGTTATAAGGTGATGAAGGGTATCA
AGCCTGAAAAACCATTAAACGCTGCGATAGCAAGCATCTTTGCACAGAGTTGTTCTCAATGTAAC
GATAAAGTTGGTGATGGTACAACAACGTGCTCAATACTAACTAGCAACATGATAATGGAAGCTTC
AAAATCAATTGCTGCTGGAAACGATCGTGTTGGTATTAAAAACGGAATACAGAAGGCAAAAGAT
GTAATATTAAAGGAAATTGCGTCAATGTCTCGTACAATTTCTCTAGAGAAAATAGACGAAGTGGC
ACAAGTTGCAATAATCTCTGCAAATGGTGATAAGGATATAGGTAACAGTATCGCTGATTCCGTGA
AAAAAGTTGGAAAAGAGGGTGTAATAACTGTTGAAGAGAGTAAAGGTTCAAAAGAGTTAGAAG
TTGAGCTGACTACTGGCATGCAATTTGATCGCGGTTATCTCTCTCCGTATTTTATTACAAATAATGA
AAAAATGATCGTGGAGCTTGATAATCCTTATCTATTAATTACAGAGAAAAAATTAAATATTATTCA
ACCTTTACTTCCTATTCTTGAAGCTATTGTTAAATCTGGTAAACCTTTGGTTATTATTGCAGAGGAT
ATCGAAGGTGAAGCATTAAGCACTTTAGTTATCAATAAATTGCGTGGTGGTTTAAAAGTTGCTGC
AGTAAAAGCTCCAGGTTTTGGTGACAGAAGAAAGGAGATGCTCGAAGACATAGCAACTTTAACT
GGTGCTAAGTACGTCATAAAAGATGAACTT
```

10. Select **BLAST** – A new window appears. Using the "Manage Columns" drop-down at the top of the table, select all fields.

Q3 How long (query length) is the Wolbachia sequence that you used to search the database?

____

Q4 What is the E-value and Maximum Identity (%) of the best hit (in this case, the first matching sequence)? ___________________________________ and ___________________________________

Q5 What is the most likely identity of this sequence? (click on the blue 'Accession' link to the right of the top hit, AY714811.1) ____

Q6 What is the title of the scientific publication that reported this sequence, if applicable? (Click on the PUBMED link) ____

Continued on page 10…

11. Return to the **NCBI BLAST Results** window.

12. Select **Distance tree of results** under **Other Reports** (above the four results tabs). This will open a separate page with a phylogenetic tree that includes your sequences (highlighted in yellow).

13. Print the phylogenetic tree (if you have access to a printer) and discuss what the tree tells you about the evolutionary relatedness of your *Wolbachia* strain to other strains in the database.

> The class might want to create a portfolio of their specific *Wolbachia* trees along with a picture and general information on their insects. In particular, what insects are the closely related *Wolbachia* from and are they the same as yours or different? What does this tell you about horizontal transmission of *Wolbachia*?

Q7 What does a phylogenetic tree show? For instance, what does the length and order of the branches tell you about evolutionary relatedness?

____

____

____

Q8 What is this strain most closely related to in the phylogenetic tree?

____

14. Return to the **NCBI BLAST Results** window.

15. Select **Home** at the top of the BLAST page.

16. Select **Nucleotide BLAST** under the Web BLAST category.

17. Now enter only the first 25 base pairs of the *Wolbachia* sequence below into the Search box.

```
GTTGCAGCAATGGTAGACTCAACGG
```

18. As you did before, select **BLAST**. A new window appears.

Continued on page 11…

Q9 What is the E-value and Maximum Identity (%) of the best hit (the first matching sequence)?

_______________________________________ and ____

Q10 Is the E-value more or less significant than when you BLASTED the longer *Wolbachia* sequence in question 3? ____

Q11 Is the identity of the best hit different from when you used the complete nucleotide sequence?

____

Q12 From the two BLAST searches you performed, what can you deduce about how the length of a query sequence affects your confidence in the sequence search?

____

____

____

STOP HERE if you do not have your own sequences.

Wolbachia BLAST

19. Return to the **NCBI BLAST Results** window.
20. Select **Home** at the top of the BLAST page.
21. Select **Nucleotide BLAST** under the Web BLAST category.
22. Enter your first *Wolbachia* Sequence into the Search box.
23. Select **BLAST.**
24. Repeat for additional *Wolbachia* sequences and record information below.

Sample ID	Arthropod Host	Best BLAST Hit	Query Coverage	E-value	% Identity

STOP HERE if you do not have arthropod sequences.

Continued on page 12…

Arthropod 1 BLAST

25. Return to the **NCBI BLAST Results** window.

26. Select **Home** at the top of the BLAST page.

27. Select **Nucleotide BLAST** under the Web BLAST category.

28. Enter your first Arthropod Sequence into the Search box.

29. Select **BLAST** and complete info below.

Sample ID	Best BLAST Hit	Query Coverage	E-value	% Identity

Q13 Based on DNA homology, what is the most likely identity of your arthropod?

____

Q14 Does it match the initial classification from Lab 1? __________________________________

- If not, explain: ____
- Which identification do you believe is most reliable? Explain.

____

____

____

30. Select **Taxonomy reports** from the top of the page.

Q15 Which organisms are included in the BLAST report? (See ‘Blast name’ under Lineage Report)

____

____

Q16 What does this phylogenetic tree tell you about the evolutionary relatedness of your arthropod to others in the database?

- What is the most closely related organism? ________________________________
- What is a more distantly related organism? ________________________________
- Do they share the same genus? ____

Continued on page 13…

Arthropod 2 BLAST

31. Return to the **NCBI BLAST Results** window.

32. Select **Home** at the top of the BLAST page.

33. Select **Nucleotide BLAST** under the Web BLAST category.

34. Enter your second Arthropod Sequence into the Search box.

35. Select **BLAST** and complete info below.

Sample ID	Best BLAST Hit	Query Coverage	E-value	% Identity

Q17 Based on DNA homology, what is the most likely identity of your arthropod?

____

Q18 Does it match the initial classification from Lab 1? ______________________________

- If not, explain: ____
- Which identification do you believe is most reliable?

____

36. Select **Taxonomy reports** from the top of the page.

Q19 Which organisms are included in the BLAST report? (See 'Blast name' under Lineage Report)

____

____

Q20 What does this phylogenetic tree tell you about the evolutionary relatedness of your arthropod to others in the database?

- What is the most closely related organism? _____________________________
- What is a more distantly related organism? _____________________________
- Do they share the same genus? ______________________________________

Close all web windows. This exercise is now complete. You successfully mastered one of the state-of-the-art tools used by most molecular and evolutionary biology researchers today. There is a lot of information on the NCBI website. Feel free to explore the website and you can find more tutorials at: http://www.ncbi.nlm.nih.gov/home/tutorials.shtml