Bioinformatics Lab

ACTIVITY AT A GLANCE

"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/

Goal:

- Module 1: To show the ways in which the NCBI online database classifies and organizes information on DNA sequences, evolutionary relationships, and scientific publications.

- Module 2: To identify an unknown nucleotide sequence from an insect endosymbiont by using the NCBI search tool BLAST

Teaching Time:
45 minutes

Introduction:
This exercise represents two interrelated modules designed to introduce the student 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.

Online Resources: There are a number of online, educational resources devoted to learning bioinformatics. For details that summarize what we will cover in this exercise and more, see:

- BLAST for beginners (Helps the learner with a slide show; we will use this one!): http://digitalworldbiology.com/dwb/BLAST

Significance and Supplies Needed: By completing this project, you will be exposed to the tools and databases currently used by researchers in molecular and evolutionary biology, and you will gain a better understanding of gene analysis, taxonomy, and evolution. While no computer programming skills are necessary to complete the modules in this work, prior exposure to personal computers and the Internet will be assumed. The main program that you will need is an Internet browser, such as Firefox or Internet Explorer.

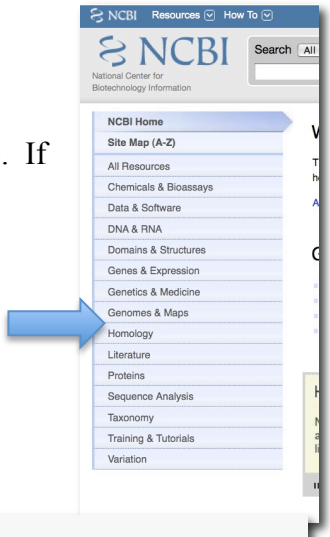
Student Activity Sheet Name:_____

Bioinformatics Lab

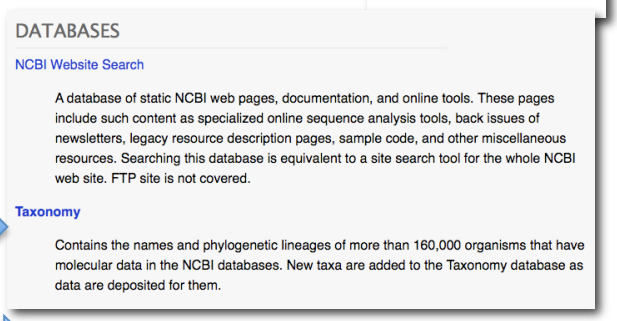
MODULE 1: Sequence Taxonomy

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

Begin by linking to the NCBI homepage (www.ncbi.nlm.nih.gov). If you ever get lost, always return to this page as a starting point. 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! Select the '**Taxonomy**' link under DATABASES.



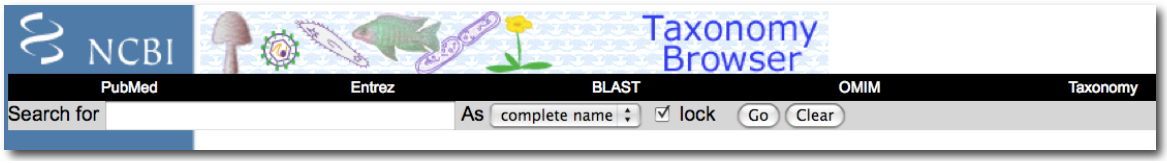
Then select the option '**Statistics**' under Tools.

1. For the 'Taxonomy Nodes (all dates)' column, how many Bacterial Species were in the sequence database?_____

2. For the year 2004, how many Bacterial Species were added in the sequence database?_____ Wow, what a difference a decade makes!

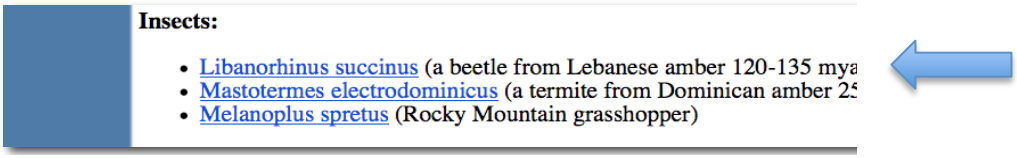
Interestingly, the sequence data from extinct organisms are even listed in the GenBank database. Let's look for a gene sequence from a 120 Mya old insect preserved in amber! To get back to your last webpage,

- Select the '**Taxonomy**' option in the top menu bar

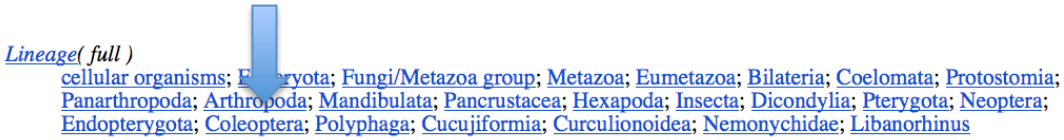


- Select '**Extinct organisms**' under Taxonomy Tools

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



- This page gives you very specific information about the ancestry of this organism. Select the option '**Arthropoda**' under *Lineage.*



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

   _____

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

   _____

   _____

5. Go back one page. How many 'Nucleotide' sequences have been deposited into the Entrez Records from this organism? Look at the box on the top right labeled 'Entrez records' _____

6. What is the name of the gene that was sequenced for this organism (to find out, click on the number 1 next to nucleotide)?_____(how does this relate to 16S rRNA) _____

7. How many nucleotide base pairs (bp) does this DNA entry contain? (the answer is in the first line just before DNA)

   _____

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. 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.

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

   _____

   _____

9. Select the '**NCBI link**' in the top left corner of the screen (next to the DNA symbol) to return to the NCBI home page. Great! That's where we started with Module 1.

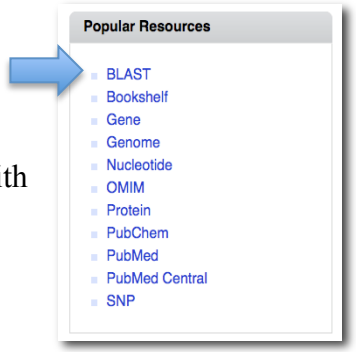
Bioinformatics Lab

MODULE 2: Sequence Searching and BLAST

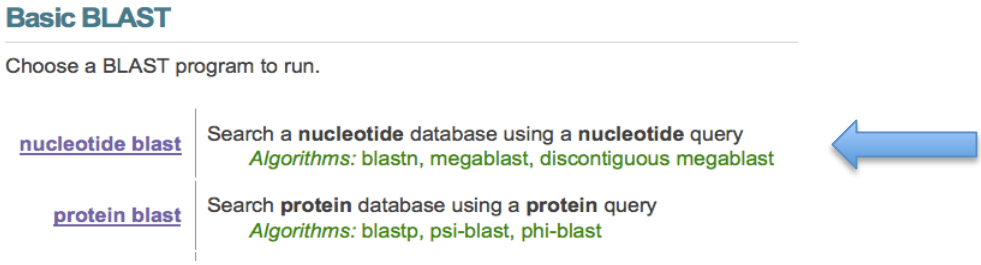
Objective: The goal of this module is to retrieve genetic sequence data from the NCBI database that identifies the '*Wolbachia* sequence' you generated. The Basic Local Alignment Search Tool (BLAST) is an essential tool for comparing a DNA or protein sequence to other sequences in various organisms. 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 (**http://digitalworldbiology.com/dwb/BLAST**). 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.

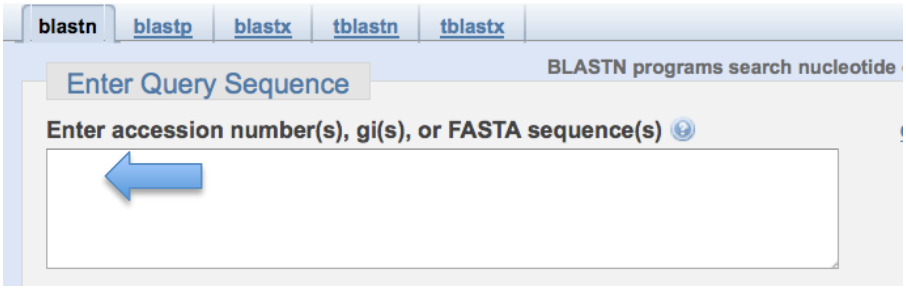
Using BLAST to identify a fake sequence and your '*Wolbachia* Sequence': Begin by linking to the NCBI homepage (**www.ncbi.nlm.nih.gov**/). **Select** '**BLAST**' in right menu bar under "Popular Resources". With your new knowledge of Sequence Searching and BLAST, let's begin with a sequence you make up and then your *Wolbachia* sequence.



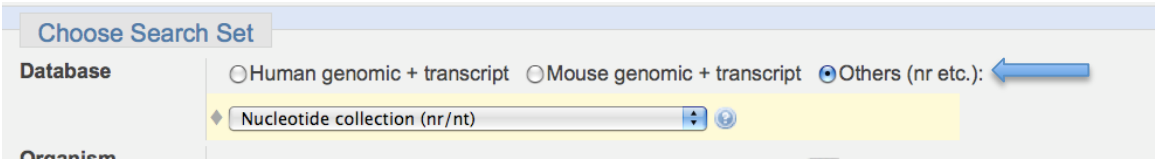
- **Select 'nucleotide blast'** under the Basic BLAST category



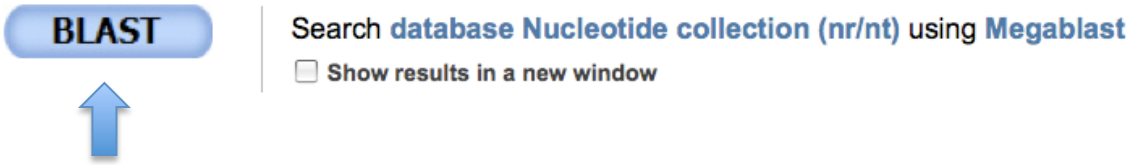
- **Input your own, random nucleotides (A,T,G,C) that fill one complete line** in the blank box at the top under "Enter Query Sequence". Your sequence is referred to as the query sequence.



- **VERY IMPORTANT - Click on the circle** for 'Others (nr etc.) under "Choose Search Set".



- **Page Down and Select 'BLAST!'** at end of page. A new window appears.



- 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.

1. 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?

   _____

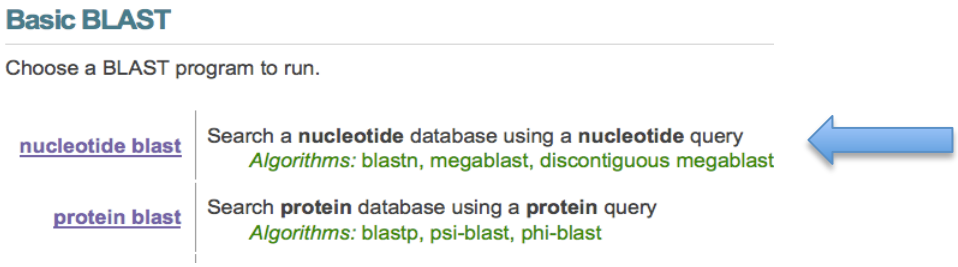
   _____

   _____

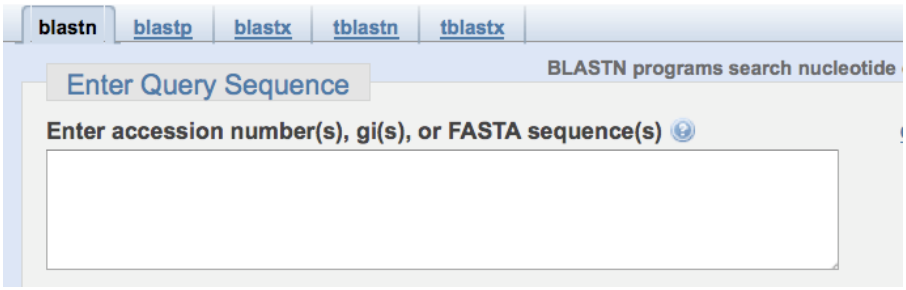
2. What was your E-Value?_____

- **Select Home** at the top left of the BLAST page.

- **Select 'nucleotide BLAST'** under the Basic BLAST category



- **Enter your Wolbachia sequence below into the Search box.** (At this point in the lab, if students generated their own *Wolbachia* sequences, they could BLAST their own sequence. Here everyone will BLAST the same sequence provided to you below)



>Your *Wolbachia* Sequence (either use the new *Wolbachia* sequences from the insects that the class discovered infections in, or if you have no sequences yet, its OK! Use the one provided online (shown below) so you can copy and paste it in the BLAST Box. Just go to this website to get the sequence - http://serc.carleton.edu/microbelife/k12/bioinformatics/module2.html

GTTGCAGCAATGGTAGACTCAACGGTAGCAATAACTGCAGGACCTAG
AGGAAAAACAGTAGGGATTAATAAGCCCTATGGAGCACCAGAAATTA
CAAAAGATGGTTATAAGGTGATGAAGGGTATCAAGCCTGAAAAACCA
TTAAACGCTGCGATAGCAAGCATCTTTGCACAGAGTTGTTCTCAATGT
AACGATAAAGTTGGTGATGGTACAACAACGTGCTCAATACTAACTAGC
AACATGATAATGGAAGCTTCAAAATCAATTGCTGCTGGAAACGATCGT
GTTGGTATTAAAAACGGAATACAGAAGGCAAAAGATGTAATATTAAA
GGAAATTGCGTCAATGTCTCGTACAATTTCTCTAGAGAAAATAGACGA
AGTGGCACAAGTTGCAATAATCTCTGCAAATGGTGATAAGGATATAGG
TAACAGTATCGCTGATTCCGTGAAAAAAGTTGGAAAAGAGGGTGTAA
TAACTGTTGAAGAGAGTAAAGGTTCAAAAGAGTTAGAAGTTGAGCTG
ACTACTGGCATGCAATTTGATCGCGGTTATCTCTCTCCGTATTTTATTA
CAAATAATGAAAAAATGATCGTGGAGCTTGATAATCCTTATCTATTAA
TTACAGAGAAAAAATTAAATATTATTCAACCTTTACTTCCTATTCTTGA
AGCTATTGTTAAATCTGGTAAACCTTTGGTTATTATTGCAGAGGATATC
GAAGGTGAAGCATTAAGCACTTTAGTTATCAATAAATTGCGTGGTGGT
TTAAAAGTTGCTGCAGTAAAAGCTCCAGGTTTTGGTGACAGAAGAAAG
GAGATGCTCGAAGACATAGCAACTTTAACTGGTGCTAAGTACGTCATA
AAAGATGAACTT

- **Select 'BLAST!'** A new window appears

4. How long (query length) is the *Wolbachia* sequence that you used to search the database?_____

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

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

7. What is the title of the scientific publication that reported this sequence (click on the PUBMED 16267140 link). Briefly describe in one sentence what this article reported_____
_____

- **Go back twice** when you're done.
- **Select Distance tree of results** in the light blue box at the top of the page. This will open a separate page with a phylogenetic tree that includes your sequences (highlighted in yellow with blue dot)
- **Print the phylogenetic tree** (if you can print) **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 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*?

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

9. What is your strain most closely related to in the phylogenetic tree?
_____
_____
_____

- **Return** to the page with your BLAST results that should still be open in another tab.

- **Select Home** at the top of the BLAST page.
- **Select 'nucleotide BLAST'** under the Basic BLAST category
- **Now enter only the first 25 base pairs of your *Wolbachia* sequence below into the Search box.**

>Your *Wolbachia* Sequence
GTTGCAGCAATGGTAGACTCAACGG

- As you did before, select **'BLAST!'** A new window appears

10. What is the E-value and Maximum Identity (%) of the best hit (the first matching sequence)?_____ and _____. Is the E-value more or less significant than when you BLASTED the longer *Wolbachia* sequence in step 3? _____

11. Is the identity of the best hit different from when you used the complete nucleotide sequence?_____

12. 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? _____
_____
_____

- **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