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."


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://www.geospiza.com/outreach/BLAST/index.html
Discover the Microbes Within: The *Wolbachia* Project

- Similarity search (Summarizes the basic concepts and vocabulary of BLAST)
- NCBI Education (Provides educational tutorials, software, and mini-courses):

**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, Internet Explorer or Safari.
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 ‘TaxBrowser’ at the top right. The NCBI Taxonomy database contains the names of those organisms whose sequences have been deposited. Only a small fraction of the millions of species estimated to exist on earth is represented! Select the option ‘Taxonomy Statistics’ in the middle of the left-side navigation bar.

1. For the ‘all dates’ column, how many Bacterial Species were in the sequence database?_____________

2. For the year 1999, how many new Bacterial Species were added to the sequence database?_____________ Wow, what a difference a few years 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!

From your last website,

• Select the ‘Taxonomy’ option in the right of the top menu bar
• Select ‘Taxonomy home’ in the left-hand navigation menu
• Select ‘Extinct organisms’ in the bottom of the left-hand navigation menu to see the organism list
• 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’.

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

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

4. Go back one page. How many ‘Nucleotide’ sequences have been deposited into the Entrez Records from this organism?______________________________________________________
5. What is the name of the gene that was sequenced for this organism (to find out, click on the number 1 next to nucleotide)?

6. How many nucleotide base pairs does this DNA entry contain? (the answer is in the first line of the flatfile after you select the Identification link)

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. At the bottom of the screen, you will find the nucleotide sequence (all of the A,T,G,C base pairs) 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 describes the discovery of this nucleotide sequence.

7. 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.
Discover the Microbes Within: The *Wolbachia* Project

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://www.geospiza.com/outreach/BLAST/index.html](http://www.geospiza.com/outreach/BLAST/index.html)). Let the slide show guide your learning by clicking on the bright green arrow to proceed through the pages. Note that this slideshow is not updated and based on the old BLAST format. 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 top menu bar. 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 nucleotides (A,T,G,C) that fill one complete line** into the Search Box. This is referred to as the query sequence.
- **VERY IMPORTANT - Click on the circle** for ‘Others (nr etc.) under Choose Search Set
- **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 produce a significant match (also known as ‘hit’ to another sequence in the genetic database (a significant hit is an E value below E-10)_____________ If yes, how many________________**

2. **How many sequences did it search in the database?_____________**

3. **How many nucleotide letters did it search in the database?_____________**

- **Select Home** at the top 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, provided by the Bordenstein Lab, or use the one provided below if you don’t have them yet)

GTTGCAGCAATGGTAGACTCAACCGGTAGCAATAACTGCAGGACCTAG
AGGAAAAACAGTGGGATTAATAAAGCCCTATGGAGCACCAGAAATTA
CAAAAGATGGTTATAAGTGATGAAAGGATCAAGACCTGAAACCA
TTAAACGCTGAGTACCAAGCATATCTTGGCACCAGATGTTGTCTCAATGT
AAGTATAAAGTTGCTGATGAAACCAACGTGCTCAATACTAAACTAGC
AACATGATAATGGAAGCCTTCAAAATCAATTTGCTGCTGAAACGATCGT
GTTGGATATTAAACACGAATACAGGAAGCAGAAGGCAAAGATGTAATTTAA
GGGATTTCGCTAAATCTCCTGAATAATTCTCTAGAGAAAAATAGCGTA
AGTGGCACAAGTTGCAATAATCTCCTGCAAATGTTGATAAGGATAAGG
TAAACGATATCCTGATTGGTCCTCGTGAAGAAAAAGTTGGAAAAAGAGGGTGTA
AACATGTTGAAGAGAGTAAAGGTTCAAAAGAGTTAGAAGTTGAGCTG
ACTACTGGCATGCAATTGCTGCTGCTGATTATTCTCCTGCAAATTAAA
CAAAATGAAAAATGTGAGTCGTAGCTGTTATAATCATATATTAA
TTACAGAGAAAAATTAAATATTTCACTTTTACCTCATTCTCTGG
AGCTATTGTAAATCTGTTAATACCTTGGATCAATCTGTTAGATGT
GAAGGTGAACTCAATTTTGTATATCAATTTTTTTGCCTATCTGGGTGTT
TTAAAAAGATTGCTGAGTAAAGAGCCTCAGTTTGTGGTGACAAGAAGAA
AGATGCTGAGCAGACATAGCAACTTTAACTGGCTAAGTGACGTCTAC

• 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 Max score 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)

What is the title of the scientific publication that reported this sequence (click on the PUBMED 16267140 link)

________________________________________________________________________

• Go back twice when you’re done.
• Select Distance tree of results NEW just above the list of the sequences producing significant alignments. This will open a separate page with a
Discover the Microbes Within: The *Wolbachia* Project

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.

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

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

9. What is the E-value and Max score of the best hit (the first matching sequence)?_______ and _________. Is the E-value more or less significant than you BLASTED the longer *Wolbachia* sequence in step 3?

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

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

---

Bioinformatics Lab-CIBT Version
• **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/Education/BLASTinfo/information3.html](http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html)