



**ACTIVITY 2: Using BLAST (one bioinformatics tool)**

To know more about what all those notations mean that are given on the results page of a BLAST search, there is an excellent tutorial to take you through the steps and terminology of a BLAST search – go to the page titled “Blast for Beginners” located at <http://www.geospiza.com/outreach/BLAST/index.html>. Follow the green arrows to complete the 12 slide tutorial.

As you work through this tutorial, you should be able to answer the following questions.

1. What does BLAST stand for?
2. The BLAST site is maintained by what agency?
3. How long is the first sequence that the tutorial pasted in the BLAST database box?
4. How many sequences are in the database for comparison?
5. What organism is the source of the sequence?
6. Note the blue letters (hyperlinks) that are given to the left of the sequence description. In general, what are they used for?

7. What is the definition of the E value?

Is a higher or lower E value better? \_\_\_\_\_ Why?

8. Even though the tutorial searched 4183 bits, how many bits from the query matched the sequence stored in the database? \_\_\_\_\_
9. In this same tutorial example, what % of the query matched exactly with the database?
10. Using the tutorial page that lists the accession number at the top left of the screen find out:
  - (a) the taxonomy of this organism (just list the first 3)
  - (b) name the journal this sequence was first published in and the year
  - (c) the authors of the journal are?

**ACTIVITY 3:**

How would you like to have access to all known genes at your fingertips? This activity uses the BLAST (Basic Logical Alignment Search Tool) search engine to locate previously explored proteins from partial sequence data. To make sure we understand the role of DNA in protein synthesis, answer the following questions:

- 1.) What are the four nucleotides that make up a DNA code? What are their structures?
  
- 2.) What does DNA code for?
  
- 3.) What is a gene? Where / how do we get genes?
  
- 4.) Where are genes located?
  
- 5.) Explain how DNA determines the traits of an organism.
  
- 6.) What is the cause of a genetic disease?
  
- 7.) What will happen to an organism's homeostasis if a gene for an important protein becomes defective?
  
- 8.) What would happen if an organism inherited a gene that codes for a defective protein?

Directions:

In the following exercise you will be given nucleotide sequences found in real DNA that are associated with genetic diseases when mutated. Your job is to compare the sequences you are given with the nucleotide sequences of most known genes. Using a search engine (BLAST) for genetic databases, you should be able to:

- a.) Name the gene (or abbreviation) that contains the sequence you are investigating
- b.) Locate the gene on its corresponding chromosome (chromosome number)
- c.) State the genetic disease associated with the defective gene
- d.) Describe the effects/symptoms that results when the gene is defective

PROTOCOL:

1. Connect to the Internet.

2. Find the home page for NCBI (National Center for Biotechnology Information) at [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

3. Click on the word "**BLAST**" located on the blue bar at the top of the page.

4. Scroll down the screen until you find the heading, "**Nucleotide Blast.**" Click on the link, "**Standard Nucleotide-Nucleotide Blast.**"

5. Type the **exact** nucleotide sequence you were given into the large empty box. **Accuracy counts!!** Hint: It is easier to read in threes to your partner, but copy and paste is more accurate.

6. When you have finished entering your sequence, click on BLAST! On the next screen, click on the "**Format**" bar. You should then see a screen asking you to wait 10-20 seconds for the search. Be patient while formatting takes place.

7. After the search has ended, scroll down the screen until you find the words "**Sequences Producing Significant Alignments.**" Listed in order are the closest matches with your DNA sequence. You may notice that the first few listings are all identical matches. **Click on the blue reference number preceding the first listing or the first few listings (the closest matches).** This will tell you the name of the gene and its abbreviation (if available).

8. Once you have identified the gene, **return to the NCBI Home Page** by repeated clicks on the **BACK** button in the upper left corner of the screen. ***In order to do so, you may have to close one of the BLAST search pages.***

9. In the right hand column, find and click on the link, "**Genes and Diseases.**"

10. Across the top of the page, you will notice the numbers 1-22 XY. These numbers represent the chromosomes found in humans. **Click on each number to locate the position of your gene on a chromosome.**

11. Click on the name or abbreviation of the gene on the chromosome in order to find out the effects or symptoms produced by the defective gene.

12. Fill in the attached data sheet for the unknown gene, using the information from your search.



Teacher Sequence Key:

Sequence A: Huntington's Disease

Sequence B: William's Syndrome

Sequence C: Alzheimer's

Sequence D: Cystic Fibrosis

Sequence E: Marfan Syndrome

Sequence F: Retinoblastoma

Sequence G: Human Menkes Disease

Sequence H: Muscular Dystrophy



---

Student Data Sheet

Name \_\_\_\_\_

Nucleotide Sequence:

Abbreviation of gene (if applicable):

Genetic disease associated with defective gene:

Chromosome number (location of gene)

Describe the effects/symptoms (phenotype) of the genetic disease:



### Alternative Activity 3 (use same sequences)

#### Worksheet: Trying Your Hand at Bioinformatics

**\*\*hint:** once you get your search results, find and use the Blue “L” box (LocusLink) (located to the right of the E-value under “sequences producing significant alignments” to lead you to further information, particularly to OMIM site for this gene. Another approach is to go back to the NCBI Home Page, click on “Genes and Disease” on menu on right side, and then click on the correct chromosome number to locate your disease gene on the chromosome.

Sequence	# bases	Human Disease	Name of Gene	Other
#1				What chromosome? (blue locuslink box) – Click on O box –get to OMIM What is a triplet repeat? How many in this disease?
2				What chromosome?
3				What chromosome?
4				What chromosome?
5				What chromosome?
6				What chromosome?
7				What chromosome?
8				What chromosome?

Follow up questions:

1. What does OMIM stand for?
2. Which diseases were inherited on the X chromosome?
3. How would inheritance of the disease be different in males than in females for X-linked disorders?

## ACTIVITY 4

### Try your hand at bioinformatics: "BLASTing through the Kingdoms of Life"

Here is a simple exercise in using a national database to identify a DNA sequence. It's as easy as cutting and pasting! See the worksheet for questions. Your teacher will decide whether you need to do all the sequences, or whether you will have one sequence assigned to you.

*\*\* HINT: it works easier if you have 2 browser windows open at the same time \*\*.*

1. Open this page <http://www.geospiza.com/outreach/BLAST/62000sequences.html>. Read the instructions listed before continuing.
2. In another window open <http://www.ncbi.nlm.nih.gov>. Next select "BLAST" from the top navigation bar. Next under the heading **nucleotide** select the link "Nucleotide-nucleotide BLAST (blastn)".
3. You are now ready to simply "copy" each of the 16 sequences listed on the Geospiza BLAST practice page and "paste" them in the BLAST search box to find the database information available.
4. As you find each of the 16 sequences, fill out the worksheet from that exercise.

### Teacher BLAST sequences key

1. Halobacterium DNA polymerase
2. Chinese radish pre-protein for fungicide (seeds and leaves)
3. Zebra fish receptor
4. Nurse shark immunoglobulin cDNA expression (antibodies)
5. American toad transcription factor 3A in oocytes
6. Purple Sea Urchin mRNA for myosin 6
7. Scorpoin precursor of cDNA for neurotoxin
8. part of kanamycin resistance used with E. coli
9. Rice chlorophyll a/b binding protein
10. Mouse amylase gene
11. Yeast RNA polymerase
12. Lactobacillus amylase gene
13. Streptococcus streptokinase (TPA)
14. Feline leukemia virus --destroys protein on T lymphocyte
15. Herpes simplex DNA polymerase
16. rRNA sequence from dirt bacteria

---

BLAST Tutorial follow-up questions  
Sequence number used \_\_\_\_\_

Name(s) \_\_\_\_\_

1. How long is the sequence that you used to search the database?
2. What is the most likely identity of this sequence? What data supports this conclusion?
3. What organism is the source of the sequence?
4. What is the common name for this organism?
5. What phylum contains this organism?
6. What is the accession number for this sequence?
7. Is this sequence expressed? How do you know?
8. If your sequence is expressed, *where* (tissue) and *when* is it expressed?
9. Is anything known about factors that cause your sequence to be expressed?
10. How many sequences can you find with an E-value less than 0.05?
11. What organisms did those sequences come from?
12. Look at the first matching sequence, how long is the extent of the alignment and what fraction of the nucleotides match?
13. Use PubMed to find the possible function of the protein specified by your DNA sequence. Describe what's known about the role of this protein in the organism that provided the DNA.

**To Summarize –**

1. Now that you have used a computer and the Internet to obtain identities of DNA sequences with relative ease, imagine doing this task – taking a DNA sequence and searching a database of sequences – without a computer. In your own words, describe why bioinformatics is a part of today's biology.

2. For a description of the role of bioinformatics in the Human Genome Project, see [www.accessexcellence.com/WN/SUA14/genome2.html](http://www.accessexcellence.com/WN/SUA14/genome2.html).