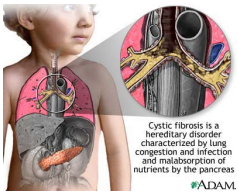
Basic Local Alignment Search Tool

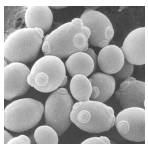
BLAST

Why Use BLAST?

Finding Model Organisms for Study of Disease

Can yeast be used as a model organism to study cystic fibrosis?





Model Organisms

- **Cystic fibrosis is a genetic disorder that affects humans**
  - If **yeast** contain a protein that is related (**homologous**) to the protein involved in cystic fibrosis
  - Then **yeast can be used as a model organism** to study this disease
    - **Study of the protein in yeast will tell us about the function of the protein in humans**

BLAST helps you to find homologous genes and proteins

Homologous Proteins (or genes)

- **Have a common ancestor (they're related)**
- **Have similar structures**
- **Have similar functions**

Criteria for considering two sequences to be homologous

- **Proteins are homologous if**
  - **Their amino acid sequences are at least 25% identical**
- **DNA sequences are homologous if**
  - **they are at least 70% identical**
  - Note that sequences must be over 100 a.a. (or bp) in length

Whenever possible, it is better to compare proteins than to compare genes

What does BLAST do?

BLAST compares sequences

- **BLAST** takes a **query** sequence
- **Compares** it with millions of sequences in the **Genbank databases**
  - By constructing local alignments
- **Lists** those that appear to be similar to the query sequence
  - The **"hit list"**
- **Tells you why** it thinks they are **homologs**
  - **BLAST makes suggestions**
  - **YOU make the conclusions**

How do I input a query into BLAST?

Choose which "flavor" of BLAST to use

- **BLAST comes in many "flavors"**
  - **Protein BLAST (BLASTp)**
    - **Compares a protein query with sequences in GenBank protein database**
  - **Nucleotide BLAST (BLASTn)**
    - **Compare nucleotide query with sequences in GenBank nucleotide database**

Enter your "query" sequence

- **A sequence can be input as a (an)**
  - **FASTA** format sequence
  - **Accession number**
  - Protein blast can only accept amino acid sequences

Choose search set

- **Choose which database to search**
  - **Default is non-redundant protein sequences (nr)**
    - **Searches all databases that contain protein sequences**

Choose organism

- Default is all organisms represented in databases
- Use this to limit your search to one organism (eg. Yeast)

BLAST off!!

- Click on the BLAST button at the bottom of the page!

How do I interpret the results of a BLAST search?

BLAST creates local alignments

- What is a local alignment?
  - BLAST looks for similarities between regions of two sequences

```
Global  FGFTALILLAVKV
        F--TAL-LLA--V

Local   FGFTALILL-AVKAV
        --FTAL-LLAAV---
```


The BLAST output then describes how these aligned regions are similar

- How long are the aligned segments?
- Did BLAST have to introduce gaps in order to align the segments?
- How similar are the aligned segments?

The BLAST Output

The Graphic Display

1. How good is the match?
   - Red = excellent!
   - Pink = pretty good
   - Green = OK, but look at other factors
   - Blue = bad
   - Black = really bad!

2. How long are the matched segments?
   Longer = better

The hit list

- BLAST lists the best matches (hits)
  - For each hit, BLAST provides:
    - Accession number – links to Genbank flatfile
    - Description
    - "G" = genome link
    - E-value
      - An indicator of how good a match to the query sequence
    - Score
      - Link to an alignment

What is an E-value?

- E-value
  - The chance that the match could be random
  - The lower the E-value, the more significant the match
    - $E = 10^{-4}$ is considered the cutoff point
    - E = 0 means that the two sequences are statistically identical

Most people use the E- value as their first indication of similarity!

The Alignment

- Look for:
  - Long regions of alignment
  - With few gaps
  - % identity should be >25% for proteins
    - (>70% for DNA)

BLAST makes suggestions, You draw the conclusions!

- Look at E-value
- Look at graphic display
- If necessary, look at alignment

- Make your best guess!

Using Blast To Ask Questions About Evolutionary Relationships

One of the tools used to determine how recently two species share a common organism is to compare their molecular sequences. Species that share a relatively recent common ancestor will have fewer differences than species that diverged in the more distant past. By comparing sequences for the same proteins found in different species, a phylogeny can be created.

In this activity, we will first examine the Coding DNA Sequence (CDS) of a specific protein for a group of organisms. When selecting a protein, it is important to consider what question you want to ask and then, what protein may be found within that group of organisms. Generally speaking, you want to select a protein that will be found in a large group of organisms, is fairly conserved within a group, and one where you can research the underlying biology. Sources like the Protein Data Bank are perfect for this research.

In this activity, we will ask which organism is most closely related within a group (Our Group will be: Raccoon, Opossum, Mouse, Rat, Guinea Pig, Rabbit). Depending on the group of organisms you select, you can also ask, what two organisms are least related (identify the outgroup), or are organisms that share a common name (such as Giant Panda and Lesser Panda or Atlantic, Chinook, and Sockeye Salmon) more closely related than other species (in this case of Panda's compare to racoons or compare the salmon to arctic char). You can also select a group of organisms, build the tree and then come up with a follow up research question.

The protein we will examine is Cytochrome C Oxidase subunit I. Cytochrome C Oxidase is a large transport protein complex associated with the inner membrane of the mitochondria. It is an important component of the Electron Transport Chain as it receives electrons from cytochrome c and uses these electrons to convert molecular oxygen to water. It is also essential in establishing the concentration gradient of hydrogens that powers ATP synthesis. While the complex can have anywhere from three to thirteen subunits, we will focus on subunit I. Cytochrome C Oxidase I is the main subunit of the complex, it is fairly conserved in aerobic organisms. The essential nature of the protein and it highly conserved nature are properties make it a great choice for cladistic analysis. More information can be found by reading up on this protein in the Protein Data Bank.

Following this Investigation, you will be able BLAST your own genes of interest and investigate it in different species. You will compile those sequences, upload them to a supercomputer to have them aligned, and then create a phylogenetic tree based on the alignment.

Part I: Using BLAST to Compile Gene Sequences

1. Pick a gene that you want to investigate. Our example, you can try cytochrome c. To locate a gene, you will go to the **Entrez Gene** (https://www.ncbi.nlm.nih.gov/gene) website. Search for "Cytochrome C Oxidase I Raccoon."
2. Scroll down and click on the "COX1" link. This link describes the Cytochrome C Oxidase I found in Racoons.
3. Now scroll down under "mRNA and Proteins," click on the first file name. It will be named "**NC_009126.1** " or something similar. These standardized numbers make cataloging sequence files easier. Do not worry about the file number for now.
4. Just below the gene title click on "FASTA.: This is the name for a particular format for displaying sequences.
5. The nucleotide sequence displayed is that of the COX1 gene in raccoons.
6. Once you have found the gene on the website, you can copy the gene sequence. Paste the sequence into a text file. Change the header of the sequence to only the species name.
7. Now that you have the sequence, you can find the sequences for other species by doing one of the following strategies:
    a. Return to Entrez Gene and search for the same protein coupled with other species (such as "opossum Cytochrome C Oxidase ").
    b. Copy the entire gene sequence, and then go the to the BLAST homepage (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and select one of the species (Such as Cow). Paste the sequence into the box where it says "Enter Query Sequence." Under "Program Selection," and choose somewhat similar sequences. Click BLAST. In the results you can find the FASTA sequences for your protein in other species. Copy these FASTA Sequences to the text document that you have already started.
    c. Copy the entire gene sequence, and then go the to the BLAST homepage. Click on "nucleotide blast" under the Basic BLAST menu. Paste the sequence into the box where it says "Enter Query Sequence." Under "Choose Search Set," select all genomes available. Under "Program Selection," and choose somewhat similar sequences. Click BLAST. In the results you can find the FASTA sequences for your protein in other species. Copy these FASTA Sequences to the text document that you have already started.
8. Once you have compiled your sequences, placed them into a single .txt file, and edit the header of each sequence to be the name of the species for that sequence, you are ready to align your sequence.

Part II: Aligning your Sequences with ClustalOmega

1. Go to Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/).
2. Change the sequence type to "DNA".
3. Copy and paste all of your FASTA sequences from your Part I document into the box.
4. Click submit.

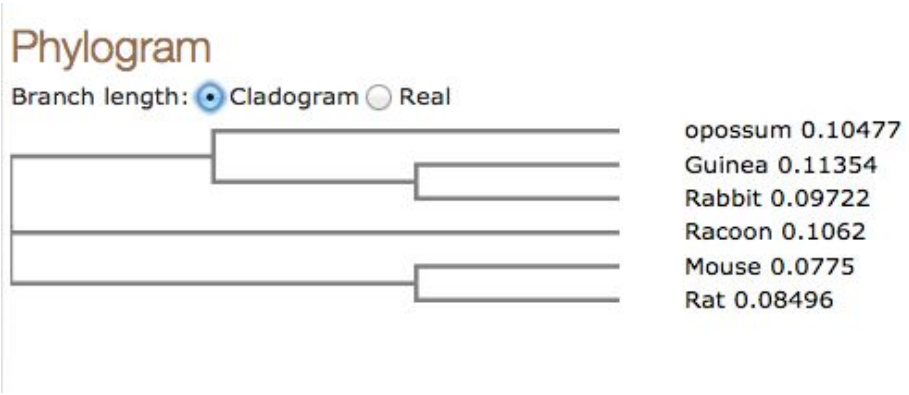
5. Once the alignments have been complete, the “Alignments” tab will show you the bases that are shared and different in the group. Here is what the first line of the comparison should look like:

```
CLUSTAL O(1.2.1) multiple sequence alignment


opossum     ATGTTCATCAATCGTTGACTTTTTTCAACTAACCACAAAGACATCGGAACACTATACTTA
Guinea      ATGTTAATTAATCGTTGATTATTTTCTACCAATCATAAAGACATTGGTACCCTATACCTC
Rabbit      ATGTTCGTCAATCGTTGACTTTTCTCTACCAACCACAAAGACATCGGCACTCTTTATCTC
Racoon      ATGTTCATAACCCGATGGCTATTTTCCACAAATCACAAGGATATTGGCACTCTCTACCTT
Mouse       ATGTTCATTAATCGTTGATTATTCTCAACCAATCACAAAGATATCGGAACCCTCTATCTA
Rat         ATGCTCGTAAACCGTTGACTCTTTTCAACTAACCACAAAGATATCGGAACCCTCTACCTA
            *** *  * *  ** **  * ** ** ** ** ** ** ** ** ** ** ** **  *
```

You can use these alignments to count up the differences between individual animals and hand draw a cladogram (but you don’t have to because the program can do this for you!)

6. Next, click on “Phylogenetic Tree” tab from the menu. It should look something like this:



7. There are a variety of other tools you can use within this program. One additional example is the “Percent Identity Matrix” found under the “Result Summary” Tab. This allows you to get a computational comparison of the sequences. It should look something like this:

```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#


     1: opossum     100.00   77.24   79.31   78.47   79.25   78.73
     2: Guinea       77.24  100.00   78.92   76.91   77.63   77.56
     3: Rabbit       79.31   78.92  100.00   77.89   79.96   78.92
     4: Racoon       78.47   76.91   77.89  100.00   80.26   78.90
     5: Mouse        79.25   77.63   79.96   80.26  100.00   83.75
     6: Rat          78.73   77.56   78.92   78.90   83.75  100.00
```

Questions and Follow Up Research:

1. What are the scientific and common names of the organisms you chose for this activity?
2. Examine the Alignment that you generated. What is the significance of the blue and black lettering shown? What generated those differences?
3. Based on your Phylogeny, which species are most closely related? Justify your answer.

4. Based on your Phylogeny, are there any aspects of your cladogram that appear to be slightly off? (such as organisms that should be closely related, but they do not share a recent common ancestor).
5. In addition to molecular data, what other forms of evidence would you use to determine the evolutionary relationships of organisms?

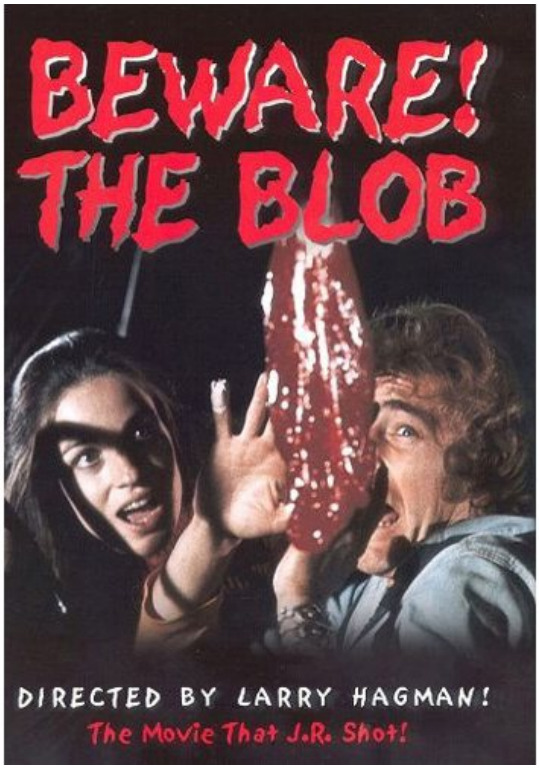
Extend By Doing Your Own Research:
Now that you have completed the sample phylogeny create your own tree. To do this you will want to select a specific protein and a group of organisms you wish to you may want to BLAST. Possible proteins you may want to consider may be:

- Actin, Catalase, Keratin, PAX1, Ubiquitin, or search the PDB Molecule of the Month Archive for additional ideas.

Some possible groups of organisms you may want to select may include:

- Bears (giant panda, lesser panda, koala, polar bear, black bear, brown bear, racoon)
- Fish (bluefin tuna, Atlantic salmon, shark, sturgeon, king salmon, rainbow trout)
- More diverse groups such as: Birds, crocks, turtles, lizards, snakes, amphibians

What on Earth is “The Chilean Blob”?



Directions for the Identification of the Chilean Blob

a. The nucleotide sequence, in FASTA, of a segment of the blob's DNA is on the previous page.
b. On the internet, go to the NCBI website (http://www.ncbi.nlm.nih.gov/).
c. Choose BLAST from the menu on the right.
d. Select nucleotide BLAST from the menu, under Basic BLAST.
e. In the box under "Enter Query Sequence" insert the blob's DNA sequence.
f. Near the bottom of the page, under "Choose Search Set" and under "Database", click on "Others".
g. Click on the" BLAST" button at the bottom of the page.
h. Wait for your results.
i. When **BLAST** is done with its search, you can scroll down and see a colorized diagram indicating the degree of similarity of the BLAST hits to blob's DNA sequence. Red and pink/purple mean a good match, while green, blue and black indicate a poor match. If the colored line spans the entire length of the window, then the "hit" sequence matches the query sequence along its entire length. Since, the first match will be that of the blob's won species, we expect to see a high quality match along the entire length of the query sequence.
j. Below the colorized diagram is a "hit list" of your results. The first "hit" will identify the blob, since it represents the closest match to the query sequence.

Nucleotide sequence of DNA isolated from Chilean Blob.

TAATACTAACTATATCCCTACTCTCCATTCTCATCGGGGGTTGAGGAGGACTAAACCAGACTCAACTCCG
AAAAATTATAGCTTACTCATCAATCGCCCACATAGGATGAATAACCACAATCCTACCCTACAATACAACC
ATAACCCTACTAAACCTACTAATCTATGTCACAATAACCTTCACCATATTCATACTATTTATCCAAAACT
CAACCACAACCACACTATCTCTGTCCCAGACATGAAACAAAACACCCATTACCACAACCCTTACCATACT
TACCCTACTTTCCATAGGGGGCCTCCCACCACTCTCGGGCTTTATCCCCAAATGAATAATTATTCAAGAA
CTAACAAAAAACGAAACCCTCATCATACCAACCTTCATAGCCACCACAGCATTACTCAACCTCTACTTCT
ATATACGCCTCACCTACTCAACAGCACTAACCCTATTCCCCTCCACAAATAACATAAAAATAAAATGACA
ATTCTACCCCACAAAACGAATAACCCTCCTGCCAACAGCAATTGTAATATCAACAATACTCCTACCCCTT
ACACCAATACTCTCCACCCTATTATAG