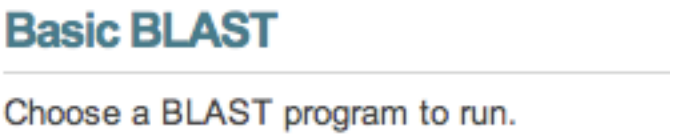
For all of the following, you will have to use this website to determine the answers:
http://blast.ncbi.nlm.nih.gov/Blast.cgi

We are going to be using the programs under this heading:



Answer the following questions. You might want to use the Help function the Blast page.

1. ***You have protein sequence and you wish to know what other proteins look like it.*** Which of the five Basic Blast programs should you use?

Answer: protein blast (also called blastp)

2. ***You a DNA sequence and you wish to search for other DNA sequences to find one that encodes the same or similar protein.*** Which of the five Basic Blast programs should you use?

Answer: tblastx

3. ***You have DNA and you wish to find other DNA sequences that look like it.*** Which of the five Basic Blast programs should you use?

Answer: nucleotide blast or blastn

4. ***You have protein sequence and you wish to search DNA databases to find genes that encode a similar protein.*** Which of the five Basic Blast programs should you use?

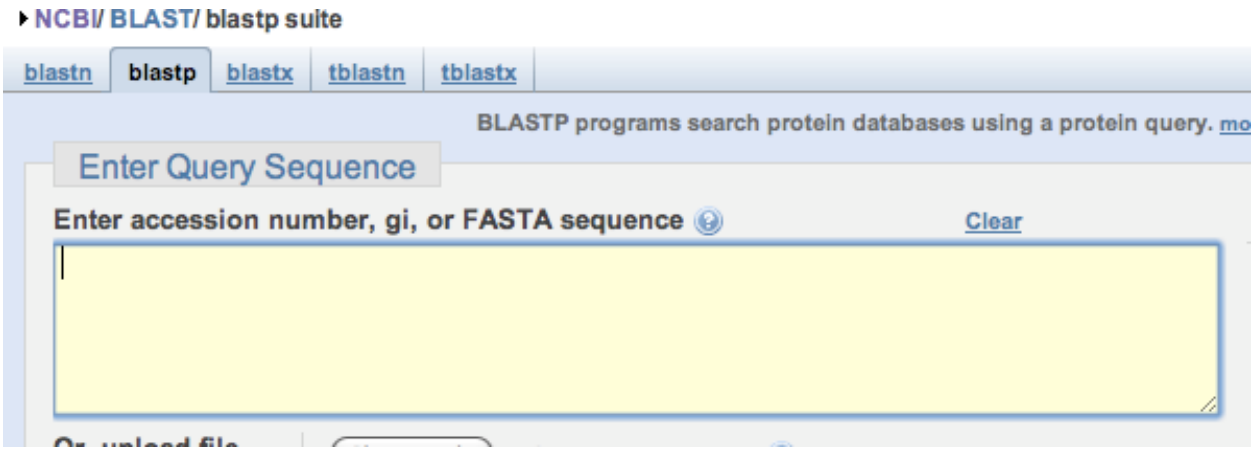
Answer: tblastn

Now we begin an example in which you are going to use blastp.

You have identified a gene that is important for transcription regulation of a collection of genes. You have just obtained some amino acid sequence. You are going to use BLAST to find out the likely identity of this protein.
The protein sequence that you have is called the query.
The query is:

HGTSSGPTVTIVQIPNGNTVQVHGVLQGGQPSVLQSPQVQTVQLSVLGESEDSQESVD

Click the link that says "protein blast".
You will see a box like the one below.



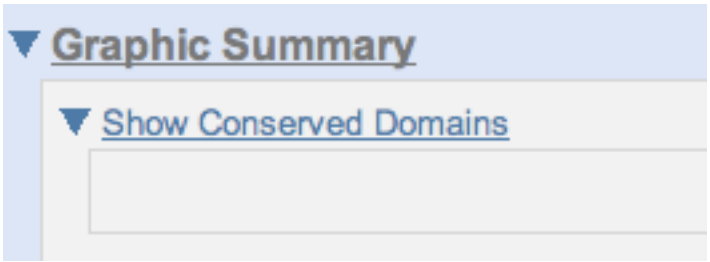
The accession number and gi numbers(see above) mean that if the sequence is contained in the Blast database then one could type in the ID number of the sequence. Your sequence is not this database so we can't use these ID Your query (above)is not exactly in fasta format but it is close enough. The program will accept it.

Copy and past the query into the the box labelled "Enter Query Sequence".

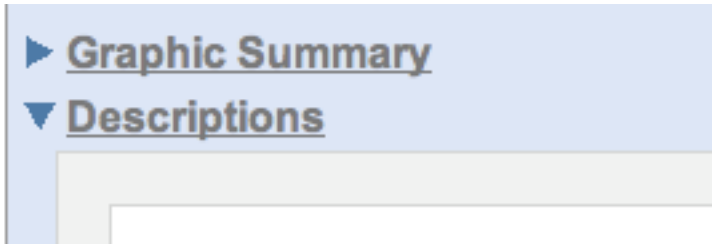
Next you click the Blast button and wait a bit.

After a while you get a graphic summary.

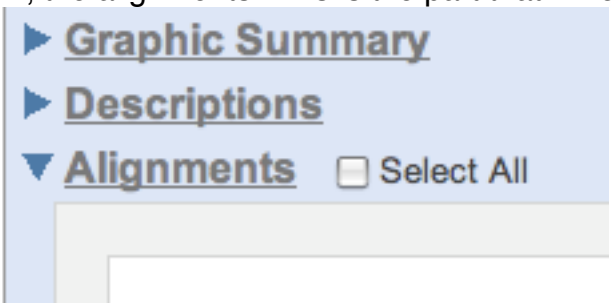
The graphic summary is convenient and like most conveniences—somewhat useless. Use it only if it shows conserved domains. Take a look at them and vanish the Graphic Summary. Click the triangle and it will vanish.



The descriptions are more useful. Look at them and then click the triangle to vanish them



Ah, the alignments. This is the part that I like.



What is the top hit? The first one is the most statistically improbable to be a random hit. This means that it is likely to be one that we want to look at.

You should a line that says something like this:
`Identities = 45/58 (77%), Positives = 51/58 (87%), Gaps = 0/58 (0%)`

5. ***What do you think the difference between Identities and Positives is?***

****
Answer: Conservative substitutions

****

Do you understand how to interpret the alignment?
The top sequence is your query. The bottom sequence is the hit that was found in the database. The middle line shows which sequence are the same.

6. ***What do the pluses mean?***
****
Answer: CHANGES THAT ARE MORE FAVORABLE - MORE CONSERVATIVE THAN THE BLANKS.

****

Now give your protein a name.
Don't make one up.

7. ***What does Blast tell you it's name should be?***

Answer: CREB

****

8. ***In blastp page what is the default scoring matrix? Look under Algorithm parameters. Tell me something about this matrix (look in lecture notes.)***

Answer:

Answer: BLOSUM62

Blocks of conserved amino acid patterns (aka blocks or motifs) were identified. This is a comparison of motifs that define gene families. About 2000 aligned and the changes recorded.

****

9. ***In blastp page what is the default word size? Look under Algorithm parameters. What does word size mean?***

Answer: _____

Answer: 3

****

10. ***In the blastp page what are the default Gap insertion and extension penalties? They are different than the ones used in class. Look under Algorithm parameters.***

Answer: 11 and 1

Alignment used in the above example.

Answer: This is a modified version of human creb

HGTSSGPTVTIVQIPNGNTVQVHGVLQGGQPSVLQSPQVQTVQLSVLGESEDSQESVD

>ref|XP_001139429.1| PREDICTED: similar to CREB isoform 2 [Pan troglodytes]
Length=226

GENE ID: 459901 CREB1 | cAMP responsive element binding protein 1 [Pan troglodytes]

Score = 97.4 bits (241), Expect = 3e-19, Method: Compositional matrix adjust.
Identities = 45/58 (77%), Positives = 51/58 (87%), Gaps = 0/58 (0%)

```
Query  1   HGTSSGPTVTIVQIPNGNTVQVHGVLQGGQPSVLQSPQVQTVQLSVLGESEDSQESVD  58
           H TSS PTVT+VQ+PNG TVQVHGV+Q  QPSV+QSPQVQTVQ+S + ESEDSQESVD
Sbjct  5   HATSSAPTVTLVQLPNGQTVQVHGVIQAAQPSVIQSPQVQTVQISTIAESEDSQESVD  62
```

NEW EXAMPLE.

This DNA is from a real cDNA. There are no sequencing errors in it.

```
gctggtccagaaggctaaactggccgagcagtcagaacgttacgatgatatggcccaggccatgaagtc
cgtcacagagactggcgttgagctctcaaatgaggaaagaaatctgctctccgttgcctacaaaaatgt
ggtcggtgcccgcaggtcatcgtggcgtgtcatctcctccattgagcagaaaaccgaagcatccgctag
aaaacagcagctcgcccgtgagtacagagagc
```

You are to search the DNA databases of blast looking for other DNA molecules that encode similar or identical proteins.

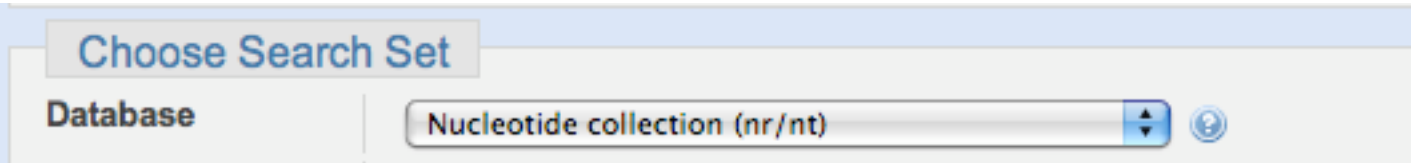
11. ***Which basic blast program should you use?*** It is a different blast than in the last question.

****

Answer: tblastx

****

Paste the sequence into the query window.
Double check that the correct database is selected. Yu should be using this one:



Now do the alignment—click the BLAST button.

In your alignment window you will see multiple possible alignments to the same chunk of DNA. Actually, if you are on the right track, all of these should be aligned amino acid sequeences.

All of the most favorable alignments are listed under a link that has a format like this:
>gb|BBXXXXX.X|
This link identifies a file that contains the sequence from the gene that matched your query. Lower down the list you will find other genes that match too.
Click the link.

12. ***What is the name of this gene?*** (hint scroll down and look under FEATURES for the term called "gene". Next to it will an entry that says
/gene = "Name of this darned gene".

Answer: Name of gene: _____14-3-3zeta-RD_____

13. ***Some of the alignments have asterisks in them. These are the product of STOP CODONS!!! Why are there stop codons?***

Answer: Different translation frames.

14. ***Knowing this, find the first one that does not have a stop codon,What is the expect value for this one?***

Answer: The expect value is _____Expect = 9e-41_____

Here is the match. They don't have to include this.
The match.
Answer:

```
 Score =  171 bits (381),  Expect = 9e-41
 Identities = 79/79 (100%), Positives = 79/79 (100%), Gaps = 0/79 (0%)
 Frame = +2/+1

Query  2    LVQKAKLAEQSERYDDMAQAMKSVTETGVELSNEERNLLSVAYKNVVGARRSSWRVISSI
181
            LVQKAKLAEQSERYDDMAQAMKSVTETGVELSNEERNLLSVAYKNVVGARRSSWRVISSI
Sbjct  145  LVQKAKLAEQSERYDDMAQAMKSVTETGVELSNEERNLLSVAYKNVVGARRSSWRVISSI
324

Query  182  EQKTEASARKQQLAREYRE  238
            EQKTEASARKQQLAREYRE
Sbjct  325  EQKTEASARKQQLAREYRE  381
```