Introduction to BLAST

# Part 1

## Learning Outcomes:

when you have successfully completed this activity, you should be able to:

* Use BLAST for homology search
* Utilize the BLAST tool to find nucleotide or protein sequences similar to your query
* Interpret BLAST search results, including what makes a significant hit
* Understand how genomic sequence information can be used to find patterns and infer possible function
* Download and upload sequence data in FASTA format
* Find accession and GI numbers that reference a particular sequence

Please keep track of your total time spent in completing the activity. This will be used to provide an average completion time for students in the future, and to assess if there seem to be issues with a particular activity (taking much longer on average for all students than anticipated)

Total time to complete this activity:

|  |
| --- |

#

#

# Activity

## Background

The purpose of this exercise is to get familiar with the Basic Local Alignment Search Tool (BLAST). BLAST finds regions of local similarity between sequences. The program compares a query sequence (a protein or nucleotide sequence) against a sequence database and calculates significance of matches. BLAST divides the query sequence into shorter words and initially looks for matches of these words only. The tool gives a score based on a scoring system e.g. in blastn, it will give +1 for each match and -2 for each mismatch.

Task - Before beginning the tasks of the activity, please review several videos below to review the BLAST tool.

Watch this overview of BLAST video **ONLY to the 6:33 mark**. End before the speaker begins on the Detailed Tasks page. Answer the questions following.

<https://youtu.be/JKD5laNtwSc>

What does BLAST tell you?

|  |
| --- |

What can a Conserved Domain Search (rps-blast) tell you?

|  |
| --- |

To further your understanding of BLAST returns, watch this short videos on E-values by NCBI and answer the questions that follow.

<https://youtu.be/dzRq-5BrGD4>

How are scores and e-values related?

|  |
| --- |

Can e-values be directly compared when searching databases of different sizes? Why or why not? What about the bit scores with the same question?

|  |
| --- |

Is there a magic cutoff for poor e-values?

|  |
| --- |

How might you determine an e-value cutoff?

|  |
| --- |

What is meant by a reported e-value of 0.0?

|  |
| --- |

##

### Part A: Using BLAST to search for closely related organisms:

This exercise demonstrates the different kinds of results you get when you search sequence databases with different algorithms using either a DNA or protein sequence from the same gene.

Task - Retrieve the GenBank sequence for NM\_001026785.1, as follows:

i) Go to the [NCBI web page](http://www.ncbi.nlm.nih.gov)

ii) Enter the ID into the search box, and select “Nucleotide” in the pull-down menu next to the word search box on the left, and then click “Search”. On the preceding page click on the Id to get to the GenBank page.

iii) Retrieve the FASTA formatted sequence for NM\_001026785.1:

Click on the “FASTA” link on the top left of the page. Paste the sequence below in this document and adjust the format using the following rules for FASTA formatted sequence:

The first line is called the “header” and always starts with “>”.

The sequence identifier must immediately follow the “>”. (No space allowed between the “>” and the identifier).

The header may include additional information after the identifier, separated from the identifier by a space. The header must all be on a single line.

The sequence starts on the second line, and can continue on additional lines

The length of the sequence on each line doesn’t matter.

Spaces and non-sequence characters are not allowed within the sequence.

FASTA formatted sequences use single line spacing.

Insert nucleotide sequence in FASTA format below (insert space as needed)

|  |
| --- |

Task - Run a BLASTN search using the NR (Non redundant database) using the sequence associated with NM\_001026785.1 as follows:

iv) Go to the [BLAST homepage](https://blast.ncbi.nlm.nih.gov/Blast.cgi) and select Nucleotide Blast.

v) Enter the query sequence: Paste the FASTA formatted sequence you formatted above into the search box. Be sure that the FASTA formatting rules are maintained. In some cases the header line will wrap around to the second line. In that case, delete the second header line, so that the sequence starts on the second line.

vi) Select the database and blast program: The default should be Standard Databases and “Nucleotide collection (nr/nt)” in the Database pull-down menu. Under “Program Selection”, select “megablast”. Parameters can be selected by looking under “Algorithm Parameters”. Set the “Max Target Sequences” to 500 under “Algorithm Parameters”.

Click BLAST and wait for the results, look at the results and paste a screenshot below.

|  |
| --- |

vii) Interpreting taxonomy report

Click on the “Taxonomy report” link in your BLAST header report.

List the taxonomic groups below that have homologs. Remember to check that the E-value is significant before considering a “hit” to be a homolog.

|  |
| --- |

Repeat the search, except under “Program Selection”, select “blastn” (you chose (megablast) on the prior search)

List below the major taxonomic groups with homologs (be sure that the E-value is significant).

|  |
| --- |

Were there any homologs in any plant species? Support your answer with E-values.

|  |
| --- |

How and why are the results different from the first blast (megablast)?

|  |
| --- |

### Part B: Picking the best match for your query sequence

###

This set of questions demonstrates the different matches you get when you run a BLAST search and how to pick the best matches for your search.

The FASTA sequence for this part of your problem set is given below:

>Bacteroides thetaiotaomicron\_SusC

atgaaaaaaggaaactttatgttcaaggtcctgcttatgcttatagctgg

aatattcttgtccattgacgcatttgctcagcaaattactgtcaaaggaa

tagtgaaagacacaacgggtgaaccggttatcggtgccaatgttgtggtg

aaaggcactaccaccggaacgattaccgatttcgacggcaacttccagtt

gtctgccaagcaaggtgacataattgttgtttcattcatcggataccagc

cacaggaacttcccgtcgccgcacaaatgaatgtaatactgaaagacgat

acggaaatactggacgaagtagtagtcatcggttacggtcaggtgaaaaa

gaacgatatgaccggttcggtaatggctatcaagcccgatgaactaagta

aaggtattacgacgaatgctcaggatatgttatccggtaaaatagccggt

gtcagcgtgatctccaatgacggtacaccgggtggtggcgctcaaatccg

tattcgtggcggttcttcattgaatgcaagcaatgacccgctgatcgtta

ttgacggtctggctattgacaatgaaggtatcaaagggatggcaaacggt

ttgtcaatggtcaaccctgcggatatcgaaacccttactgtactgaaaga

tgcctctgcaactgccatttacggttcgcgtgcatccaacggtgttatta

ttatcaccaccaagaaaggaaagaacggacaagctcccagcgtaagctat

aacggttctgtatccttctccaaaactcaaaagcgctatgatgtattgag

cggagatgaatatcgcgcttacgccaatcagttatggggtgacaaattac

cggcagatttaggaaccgccaatacagactggcaggatcagatattccgt

actgctgtcagcaccgaccatcatgtttctatcaacggaggattcaagaa

cctgccttaccgtgtatctttaggttatacagacgacaatggtattgtga

aaacatccaacttccgacgcttcactgcttccgtgaacctggctccttcc

ttctttgaagatcatctgaagttcaacattaatgccaaattcatgaacgg

taaaaaccgctatgccgacacaggtgccgctattggcggggcattggcta

tcgaccctacccgtccggtttattctaacgaagacccttaccagtttaca

ggcggctactggcagaatataaattctaccacaggtttcagcaatccgga

ctggaaatacacgtccaatccgaactctccccaaaatccgctggctgcac

tggaactcaaaaatgacaaggcgaacagcaacgactttgttggaaatgta

gacgttgactataaattccatttcctgcctgacctccgtctgcacgcaag

cataggtggcgaatatgcggaaggtacacagactacgattgtttctccat

actcattcggcaataattactatggttggaatggcgacgttacccaatat

aaatacaacctttcgtacaacatatacgtacagtatatcaagtctttggg

tgcaaacgactttgacatcatggtcggtggtgaagaacaacacttccatc

gcaacggatttgaagaaggccagggctgggattcctatacgcaagaaccc

catgacgccaaattgcgcgaacagacagcttatgcaaccagaaatacact

ggtctcttacttcggccgtctgaattactccctgctgaaccgttacttgt

ttacctttaccatgcgttgggatggctcgtcacgtttctccaaagacaac

cgctggggtacattcccgtcattggcactgggatggaagattaaagaaga

aaacttcctgaaagatgtaaatgtcctgtctgatctgaaattgcgtttag

gctggggtattaccggtcagcaaaacataggtgatgattttgcttatctt

cctctgtatgtagtcaataacgagtatgcccagtatccttttggcgatac

ctattactctacttcccgcccgaaggctttcaatgaaaatctgaaatggg

aaaaaacgaccacatggaatgccggactggacttcggattcctgaatgga

agaatcacaggcggtatcgacggatacttccgtaaaacggatgacctgct

gaacagcgttaagatccccgtaggaactaacttcaatgcccagatgacac

agaatatcggttcactggaaaactacggtatggaattttccatcaacgcc

aaaccaattgtgactaaggacttcacctgggacctcagctataacattac

atggaaccacaatgaaatcaccaagttgacaggtggcgacgacagcgatt

attacgtagaagcaggcgataagatttccagaggtaacaataccaaggta

caggcgcataaagtaggttacgcagccaactctttctacgtttaccagca

ggtatacgacgaaaatggcaaacctattgaaaatatgtttgttgaccgta

acggaaacggaacaatagacagcggagacaaatatatctacaaaaaaccg

gcaggcgatgttttgatgggactgacctccaaaatgcagtataagaactt

tgacttcagcttctccttacgtgccagtctgaataactacgtgtactatg

acttcctgagcaacaaagccaacgtcagcacttcgggactgttctccaat

aatgcatatagcaacaccagtgccgaagccgtcgcactcggtctcagcgg

acaaggtgattacatgagcgactattttatacataacgcatcattcttac

gttgtgataacatcacgttaggttattctttccagaatctgtggaagact

caaacctacaaaggtgttggcgggcgtgtatatgctacagtacagaatcc

gttcattatcagtaaatacaaaggccttgatccggaagtaaaaagcggta

tcgacgccaatccatatcccagagctatgactttcttattaggtttaagt

ctgcaattctaa

i) Go to [BLAST](https://blast.ncbi.nlm.nih.gov/Blast.cgi) and select nucleotide blast and copy and paste the FASTA sequence above in the “Query Sequence Box”.

Choose “Others” for database, under “Choose search set”.

Since the rest of the search sets are optional, you need not check them for this exercise, but their functionality is explained below:

Organism (optional): You could enter the name of the organism, however the program identifies the organism on it’s own.

Exclude (optional): This option allows you to exclude identifiers with XM (predicted mRNA models) and XP (predicted protein models) as well as other samples from your query. Sequences associated with these identifiers are usually associated with incomplete data or data that come from sequencing only without associated biological information. Excluding these identifiers will make the query search quicker.

Entrez query (optional): This option allows you to restrict your search to a subset of entries.

Task - Run a blastn and wait for the results in the preceding page and answer the following questions:

What is the primary citation for the BLAST program? Hint: Look in the BLAST header report

|  |
| --- |

What database was used to query your sequence? How many sequences are there in the database? Hint: All this information is there in the header report. Look at the different reports in the same area where the “Taxonomy report” is linked.

|  |
| --- |

What does the color code on the graphical interface represent? (Paste a screenshot of the graphical interface below)

|  |
| --- |

What is the bit (max) score, E-value and query coverage values of the first ten hits of the BLAST results? Insert a screenshot below of these results.

|  |
| --- |

Among these ten results, which represents a paralog in another close species? Justify your answer.

|  |
| --- |

There are some entries with significant E-values (low E-values).

Are these all best matches to the entire query sequence?

|  |
| --- |

How would you pick the best match between the query and subject? For this it would help you to give an example of the best match and one that may not be a good match, in spite of a low E-value. Justify your answers.

|  |
| --- |

How would you identify a closely related species to your query sequence?

|  |
| --- |

### Part C: Using BLASTP to identify repeats

The next set of questions demonstrates how repeating characters in a sequence (such as repeating patterns of amino acids) can affect the results of a database search.

Task - Retrieve the FASTA formatted sequence for the protein CAC40682.1 from NCBI and paste it below.

What organism is this sequence from? How long is the sequence?

|  |
| --- |

Does anything stand out about this sequence when you look at it?

|  |
| --- |

Run a BLASTP search of CAC40682.1 using the “Reference Proteins” database and set organism to “Drosophila melanogaster”.

Can you identify the conserved domains?

|  |
| --- |

What database is associated with the annotations? Hint: Look at the conserved domains graphic and click on it to identify the database from where the annotation is imported.

|  |
| --- |

What is the best hit (the first match listed) and what is the E-value?

|  |
| --- |

Look at the alignment (scroll down). Which amino acid is aligned most frequently? (You can answer this by just looking at the alignment).

|  |
| --- |

Using the same query sequence for BLASTP, set the database to “Reference Proteins” and the organism to “Drosophila melanogaster”. This time, turn on “Filter Low Complexity Regions” under “Algorithm Parameters”.

Did you find a homolog in Drosophila melanogaster?

|  |
| --- |

Based on results from the previous question (i-v)s, do you think this protein has a homolog in Drosophila melanogaster? Explain why or why not.

|  |
| --- |