**Hands-On Exercise: BLAST**

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.

BLAST can be freely accessed at the NCBI website at:

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

More information on BLAST and the parameters used in the BLAST algorithms can be found at:

<http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs>

Learning Goals:

* Become familiar with BLAST and be able to use it for homology searches
* Be able to interpret BLAST results

***It may help you to open a new Word or Text document to keep all of your findings along with web citations, so that you can keep track of where you’ve been and where you are going.***

**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 <http://blast.ncbi.nlm.nih.gov/Blast.cgi> and click on nucleotide blast. Copy and paste the FASTA sequence above in the “Query Sequence Box”.

Choose “Standard” 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 its own.

**Exclude (optional):** This option allows you to exclude sequences from your search. For example, you might want to exclude sequences with XM (predicted mRNA models). Sequences associated with these identifiers are usually associated with data that come from sequencing only without associated biological information. When doing a protein BLAST, the parallel identifier is XP (predicted protein models). You might also exclude data from incomplete sequencing projects. 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.

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

ii) What is the primary citation for the BLAST program? *Hint:* Look in the BLAST header report.

[Reference](https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Retrieve&list_uids=9254694&dopt=Citation)

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

iii) How many sequences are there in the database that was used for your search? *Hint:*  All this information is there in the header report under Databases

59809978 (07/27)

iv) What does the color code on the graphical interface represent? (Paste a screen shot of the graphical interface)

Results as of 07/27



Color key represents alignment scores – higher score = higher similarity. Red means it is the most similar, followed by pink (not shown here).

v) What is the bit score, E-value and query coverage values of the first ten hits of the BLAST results? Make a table to represent your results.

Results as of 07/27



The bit score or Max score is calculated from the alignment and normalized using the scoring matrix used for the BLAST search (https://www.ncbi.nlm.nih.gov/books/NBK62051/)

vi) From among these ten results, which represents an ortholog in another closely related species? Justify your answer.

*Bacteroides salontronis and Bactoroides xylanisolvens (see encircled in v. above)*

vii) 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.

The first three hits in the table match genes in different strains of *Bacteroides thetaiotamicron*. The fourth hit matches genes in a different species. All these hits have high matching scores, high query coverage and low E-values. These criteria together are used to pick the best matches.

The fifth hit titled *Bacteroides thetaiotomicron neopullulanase* is not a good match due to a low query coverage (20%).

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

The criteria used for identify closely related species is High bit score, low E-value and a high query coverage. In addition, query and subject length should match with few gaps in the sequence alignments