**BLASTP exercise on susc Protein**

We will be using the protein sequence of susC proteinto query a database.

1. Copy the susC\_protein.txt file.

>gi|29341017|gb|AAO78807.1| SusC [Bacteroides thetaiotaomicron VPI-5482]

MKKGNFMFKVLLMLIAGIFLSIDAFAQQITVKGIVKDTTGEPVIGANVVVKGTTTGTITDFDGNFQLSAK

QGDIIVVSFIGYQPQELPVAAQMNVILKDDTEILDEVVVIGYGQVKKNDMTGSVMAIKPDELSKGITTNA

QDMLSGKIAGVSVISNDGTPGGGAQIRIRGGSSLNASNDPLIVIDGLAIDNEGIKGMANGLSMVNPADIE

TLTVLKDASATAIYGSRASNGVIIITTKKGKNGQAPSVSYNGSVSFSKTQKRYDVLSGDEYRAYANQLWG

DKLPADLGTANTDWQDQIFRTAVSTDHHVSINGGFKNLPYRVSLGYTDDNGIVKTSNFRRFTASVNLAPS

FFEDHLKFNINAKFMNGKNRYADTGAAIGGALAIDPTRPVYSNEDPYQFTGGYWQNINSTTGFSNPDWKY

TSNPNSPQNPLAALELKNDKANSNDFVGNVDVDYKFHFLPDLRLHASIGGEYAEGTQTTIVSPYSFGNNY

YGWNGDVTQYKYNLSYNIYVQYIKSLGANDFDIMVGGEEQHFHRNGFEEGQGWDSYTQEPHDAKLREQTA

YATRNTLVSYFGRLNYSLLNRYLFTFTMRWDGSSRFSKDNRWGTFPSLALGWKIKEENFLKDVNVLSDLK

LRLGWGITGQQNIGDDFAYLPLYVVNNEYAQYPFGDTYYSTSRPKAFNENLKWEKTTTWNAGLDFGFLNG

RITGGIDGYFRKTDDLLNSVKIPVGTNFNAQMTQNIGSLENYGMEFSINAKPIVTKDFTWDLSYNITWNH

NEITKLTGGDDSDYYVEAGDKISRGNNTKVQAHKVGYAANSFYVYQQVYDENGKPIENMFVDRNGNGTID

SGDKYIYKKPAGDVLMGLTSKMQYKNFDFSFSLRASLNNYVYYDFLSNKANVSTSGLFSNNAYSNTSAEA

VALGLSGQGDYMSDYFIHNASFLRCDNITLGYSFQNLWKTQTYKGVGGRVYATVQNPFIISKYKGLDPEV

KSGIDANPYPRAMTFLLGLSLQF

2. Go to <http://blast.ncbi.nlm.nih.gov/Blast.cgi> and click on protein blast under BASIC BLAST.

3. Enter the query sequence: You can either copy and paste the FASTA format of susC protein or upload the susC\_protein.txt file (Note that the sequence is in FASTA format and has the ‘>’ symbol as part of the header).

4. Choose search set:

**Database**: choose **Non-redundant protein sequences (nr).**  This option will query all protein databases including with the exception of metagenomic proteins (env\_nr). You can also choose a specific database e.g. SwissProt or Protein Data Bank (PDB) as your database against which your sequence will be queried.

**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.

5. Program selection:

This option allows you to optimize your search query for

1. Protein-Protein Blast (blastp): blastp the default program of choice for identifying a query amino acid sequence and finding similar sequences in a protein database. This program is designed to find local regions of similarity.
2. Protein-Specific Iterated Blast (PSI-blast): PSI-blast is more sensitive than blastp and is used to find more distantly related proteins or new members of a protein family.
3. Pattern-Hit Initiated Blast (PHI-blast): PHI-blast is designed to search for proteins that contain a pattern specific by the user AND are similar to the query sequence in the vicinity of the pattern.
4. Domain enhanced lookup time accelerated BLAST (DELTA-blast): DELTA-blast

performs a multiple sequence alignment of the query sequence with domains described in a specific database called the Conserved Domain Database (CDD) from NCBI.

More information on each of these programs can be found at the BLAST help pages :

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

For this exercise we suggest that you do the queries using all four programs individually and spot the difference in the results.

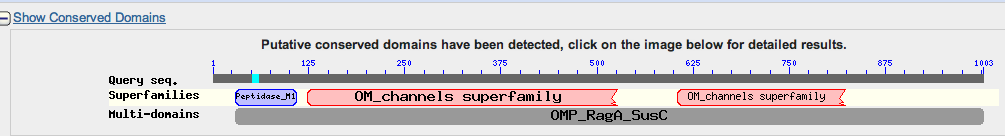
6. Algorithm parameters:

Please look at the BLAST manual for more details.

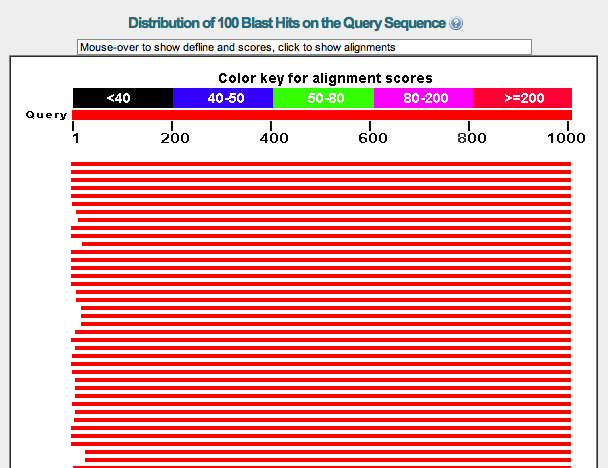
<http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs>

BLASTP results:

With protein searches, blastp automatically searches for conserved domains that are associated with known functional or structural regions and displays them while the search is being done. The same information is also available under the “Show Conserved domains” tab.



The graphical output of blastp looks like

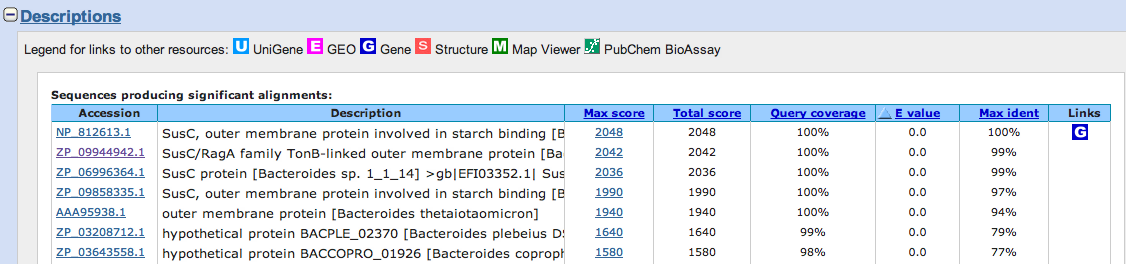


1. The graphical overview aligns hits (database sequences retrieved during

BLAST search) with the query sequence.

1. The thick red bar at the top represents the query sequence, and the numbers below the red bar correspond to those of nucleotides.
2. As shown in the color key above the thick red bar, the bar color for a hit refers to an alignment score, a mathematically derived value that reflects the degree of similarity between the hit and query sequences. The higher the score, the more similar the two sequences are.
3. The Color Key at the top of the graphical display gives the range of alignment scores assigned to each color. For example, red hits are most similar, with alignment scores greater than or equal to 200, while black hits are least similar, with alignment scores lower than 40

Below the graphical results are descriptions of statistically significant alignments (top 100) with accession numbers, scores, E-values and links to other resources, e.g. Gene resources (G) in the example below. The most significant alignments are at the top.



**Accession No**: is the unique identifier for the entire sequence record.

**Max Score**: Score of the single best-aligned sequence. This score is calculated from the number of gaps and substitutions associated with each alignment. **The higher the score more significant is the alignment.**  Each score refers to a pairwise alignment between the query and the hit sequence (subject sequence).

**Total score**: Score all the aligned sequences.

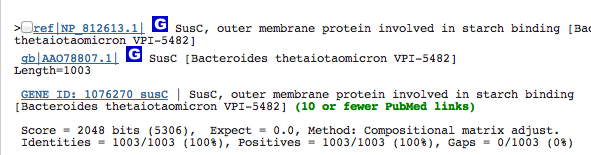
**Query coverage**: What percentage of the query sequence was aligned.

**Max Ident**: Represents the percent identity of the match

**E-value**: is a parameter that describes the number of hits one can “expect” to see by chance when searching a database of a particular size. **Lower the E-value, the higher is the significance of the match.** In general an E-value less than one is necessary to be a valid match, with smaller E-values being better. While picking a sequence it is good to correlate the query coverage and E-value. An E-value of 0 is an exact or nearly exact match.

**A word of caution on E-values:** Since BLAST works by dividing the query sequence into shorter words (strings of sequence), a low E-value may not always be a good parameter to look for homology. For example a shorter word sequence may have a low E-value, but may not be a perfect match, since it is part of the query sequence. In this case, look out for the query coverage and make sure it matches the original query sequence.

Clicking on the Max score will direct you to a pairwise alignment, which looks like this:

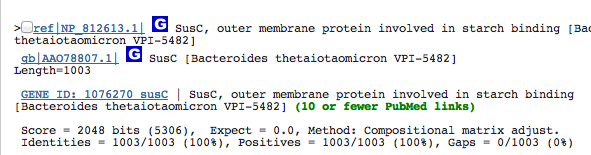


The header line contains a sequence identifier, a full definition line and the length of the sequence. Following this is the score (in bits and the raw score) as well as the statistical significance of the match (E-value), followed by the number of identities and positive matches according to the scoring system (e.g., Compositional matrix) and, if applicable, the number of gaps in the alignment.

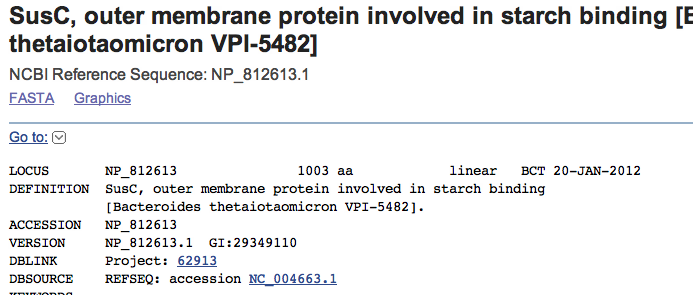
In the sequence alignment, the identical matches are marked by letter code with "homologous" substitutions (determined by the scoring matrix used) marked by "+" symbol in a line between the query and the database sequence. The numbers on either side of the sequence are the sequence co-ordinates.



Click on the ID associated with your pairwise alignment:



You will be directed to a new page with a flat file, which looks like:



**Questions to answer:**

1. How long is the protein in terms of amino acids?
2. What is the academic citation (Title, Journal and Authors) associated with this sequence?
3. What region of the sequence is associated with Peptidase like activity (give co-ordinates)?
4. Can you find a sub-sequence within the protein sequence? (Hint look at the options on the right-hand side of the page)

**Problem Set 2: BLAST exercises**

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 search
* Become familiar with the BLAST output and 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.***

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

This set of questions 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.

*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 into a word document and adjust the format using the following rules for FASTA formatted sequence:

1. The first line is called the “header” and always starts with “>”.
2. The sequence identifier must immediately follow the “>”. (No space allowed between the “>” and the identifier).
3. 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.
4. The sequence starts on the second line, and can continue on additional lines
5. The length of the sequence on each line doesn’t matter.
6. Spaces and non-sequence characters are not allowed within the sequence.
7. FASTA formatted sequences use single line spacing.

*Run a BLASTN search using the NR (Non redundant database) using the sequence associated with NM\_001026785.1*

iv) Go to <http://blast.ncbi.nlm.nih.gov/Blast.cgi> and click on **nucleotide blast** under BASIC BLAST.

v) Enter the query sequence:

Paste the FASTA formatted sequence 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

Choose “Others” for database and select the “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 screen shot in your word document.

vii) Interpreting taxonomy report

Click on the “Taxonomy report” link in your BLAST header report. (More information on how to interpret the results from a “Taxonomy report” can be found at: <http://www.ncbi.nlm.nih.gov/blast/taxblasthelp.shtml>)

a) List the taxonomic groups 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”*

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

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

x) How and why are the results different from Part B?

**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 <http://blast.ncbi.nlm.nih.gov/Blast.cgi> and click on nucleotide blast under BASIC 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.

*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

iii) 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.

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

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.

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

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.

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

**Part C: Using blastp to identify repeats**

The third 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.

*Retrieve the FASTA formatted sequence for the* ***protein*** *CAC40682.1 from NCBI and paste it here.*

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

ii) 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”.

iii) 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.

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

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

vi) 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” in under “Algorithm Parameters”. Did you find a homolog in Drosophila melanogaster?

vii)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.