JOURNAL OF MICROBIOLOGY & BIOLOGY EDUCATION, December 2013, p. 238-243
DOI: http://dx.doi.org/10.1128/jmbe.v14i2.637

JMBE Journal of Microbiology & Biology Education

Supplemental Materials

for

A Small-Group Activity Introducing the Use and Interpretation of BLAST

Peter D. Newell[1], Ashwana D. Fricker[2], Constance Armanda Roco[2], Pete Chandrangsu[2], and Susan M. Merkel[2]*
Departments of [1]Entomology, and [2]Microbiology, Cornell University, Ithaca, NY 14853

Table of Contents

(Total pages 19)

*Corresponding author. Mailing address: Department of Microbiology, 111 Wing Hall, Cornell University, Ithaca, NY 14853. Phone: 607-254-2767. Fax: 607-255-3904. E-mail: smm3@cornell.edu.

Appendix 1: Student handout part 1.

PRINT & COMPLETE THIS BEFORE YOU COME TO CLASS

Learning objectives. *After completing this small group activity, you should be able to ….*

- Label and explain the function of key components in Gram positive and Gram negative bacteria
- Determine the predicted function of a protein sequence using BLAST
- Determine if a gene product is present in a specific organism using BLAST
- Evaluate sequence similarity based on BLAST outputs: E-values, % query cover, and % max identity

This activity introduces BLAST (Basic Local Alignment Search Tool), a valuable tool for analyzing nucleic acid and protein sequence data. In addition, this activity highlights some important differences between the cell envelopes of Gram positive and Gram negative bacteria.

Please note: you need to complete Part I to obtain information for the in-class activity (Part II).

Completion of Part I will be checked at the beginning of class. Please bring laptops to class!

Part 1 (2 points)

A) Cell Envelope Review: Look at the diagrams below of Gram (+) and Gram (-) type cell envelopes.

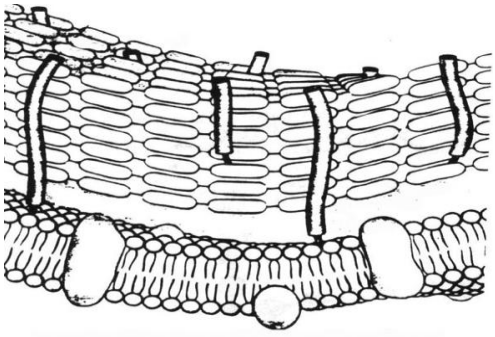
i) Label each cell envelope as being from a Gram (+) or Gram (-) cell type

ii) Label the components of each cell envelope using the list below

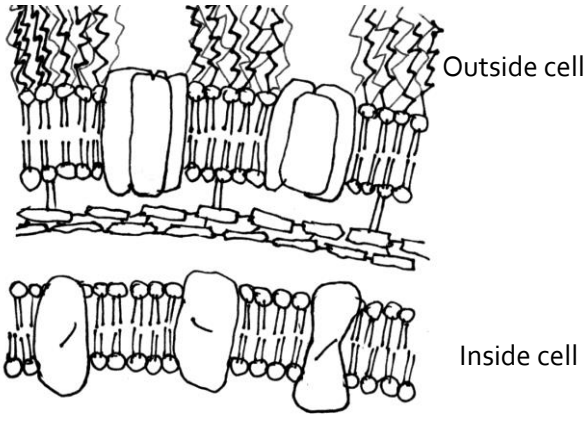
Note: not all cell-types have all the structures.

- cytoplasmic membrane
- outer membrane
- membrane-bound proteins
- peptidoglycan layer
- periplasmic space
- porins

Cell wall type _____



Cell wall type _____



B) BLAST review: One of the bioinformatic tools that you will use is **BLAST** (**Basic Local Alignment Search Tool**), that can be found at the National Center for Biotechnology Information site: http://blast.ncbi.nlm.nih.gov/.

As the name implies, BLAST makes alignments between sequences. Alignment is the process (or result) of matching up the nucleotide or amino acid residues of two or more biological sequences to achieve the best possible match. BLAST identifies sequences similar to your query sequence in the NCBI database by making alignments and assessing how well the sequences match.

Please view the *BLAST Video Tutorial* ((https://www.youtube.com/watch?v=x_dAyY5-VNc) **or the *BLAST PDF Tutorial* to answer the questions below.**

Below is the amino acid sequence of a protein associated with some bacterial cell envelopes. Use a protein BLAST (**BLASTP**) search to obtain information about it, and answer the questions below.

MKLKNTLGVVIGSLVAASAMNAFAQGQNSVEIEAFGKRYFTDSVRNMKNADLYGGSIGYF
LTDDVELALSYGEYHDVRGTYETGNKKVHGNLTSLDAIYHFGTPGVGLRPYVSAGLAHQNI
TNINSDSQGRQQMTMANIGAGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAGLGV
GFNFGGSKAAPAPEPVADVCSDSDNDGVCDNVDKCPDTPANVTVDANGCPAVAEVVRVQ
LDVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANA
VRDVLVNEYGVEGGRVNAVGYGESRPVADNATAEGRAINRRVEAEVEAEAK

Questions:

Identify the Top BLAST hit and fill in the box to answer questions 1-3.

1) What kind of protein does this sequence encode, based on the name given (annotation)?
2) From what organism did it come?
3) What is the BLAST % query cover, E value and % Max Identity for the top hit?

Top BLAST hit for the sequence from your isolate

Protein	Organism	% Query Coverage	E-value	% Max Identity

4) What is the function of this kind of protein?

5) Based on what this protein does and where it is found, do you think this organism is a Gram positive or Gram negative bacterium? Explain your logic.

6) Look at the BLAST tutorial (or look at the glossary section in the BLAST website at http://www.ncbi.nlm.nih.gov/books/NBK62051/) and fill in these definitions:

E-value:

% Max Identity:

7) When running a BLAST search, often times the sequences returned will align with only **part** of your query sequence. NCBI defines query coverage as the percent of the query sequence length that is included in the alignment. This number is significant because it figures into the calculation of the E value— the greater the query coverage, the lower the E value, and the better the match.

Place an asterisk next to the BLAST hit (A or B) below with the higher query coverage (the two examples are different hits using the same query sequence):

A)

putative outer membrane lipoprotein [Escherichia coli O26:H11 str. CVM9952]

Sequence ID: ref|ZP_14641000.1| Length: 219 Number of Matches: 1

▷ See 1 more title(s)

Range 1: 113 to 213 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
81.3 bits(199)	9e-17	Compositional matrix adjust.	41/102(40%)	62/102(60%)	1/102(0%)

```
Query  241  DVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANAV  300
            +V FD    + +K       +  +A  +K+YP T+  V G+TDS G    N +LS++RA++V
Sbjct  113  NVTFDSSSATLKPAGANTLTGVAMVLKEYPKTAVNVIGYTDSTGGHDLNMRLSQQRADSV  172

Query  301  RDVLVNEYGVEGGRVNAVGYGESRPVADNATAEGRAINRRVE  342
               L+ + GVE  R+   G G + P+A N+TAEG+A NRRVE
Sbjct  173  ASALITQ-GVEASRIRTQGLGPANPIASNSTAEGKAQNRRVE  213
```

B)

Major porin and structural outer membrane porin OprF precursor [Pseudomonas sp. M1]

Sequence ID: ref|ZP_19204629.1| Length: 353 Number of Matches: 1

▷ See 1 more title(s)

Range 1: 1 to 351 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
666 bits(1718)	0.0	Compositional matrix adjust.	317/351(90%)	339/351(96%)	1/351(0%)

```
Query  1    MKLKNTLGVVIGSLVAASAMNAFAQGQNSVEIEAFGKRYFTDSVRNMKNADLYGGSIGYF  60
            MKLKNTLGVVIGS++AASA+NAFAQGQ +VE EAFGKRYFTDS RNMKN DLYGGS+GYF
Sbjct  1    MKLKNTLGVVIGSMIAASAVNAFAQGQGAVEAEAFGKRYFTDSTRNMKNGDLYGGSVGYF  60

Query  61   LTDDVELALSYGEYHDVRGTYETGNKKVHGNLTSLDAIYHFGTPGVGLRPYVSAGLAHQN  120
            LTDDVELALSYGEYHD+RGTYE+GNKKVHGNL SLDAIYHFGTPGVGLRPYVSAG+ HQ+
Sbjct  61   LTDDVELALSYGEYHDIRGTYESGNKKVHGNLASLDAIYHFGTPGVGLRPYVSAGIGHQS  120

Query  121  ITNINSDSQGRQQMTMANIGAGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAGLGV  180
            +TN+NS++ GRQ +TMANIGAGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAG+GV
Sbjct  121  LTNVNSENGGRQNLTMANIGAGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAGVGV  180

Query  181  GFNF-GGSKAAPAPEPVADVCSDSDNDGVCDNVDKCPDTPANVTVDANGCPAVAEVVRVQ  239
            G NF GG+K APAPEPVA+VCSD D+DGVCDNVDKCP+TPANVTVDANGCPAVAEVVRVQ
Sbjct  181  GMNFGGGAKPAPAPEPVAEVCSDEDHDGVCDNVDKCPNTPANVTVDANGCPAVAEVVRVQ  240

Query  240  LDVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANA  299
            LDVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANA
Sbjct  241  LDVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANA  300

Query  300  VRDVLVNEYGVEGGRVNAVGYGESRPVADNATAEGRAINRRVEAEVEAEAK  350
            VRDVLVNEYGVEGGRVNAVGYGESRPVADN+TAEGRA+NRRVEAEVEA+AK
Sbjct  301  VRDVLVNEYGVEGGRVNAVGYGESRPVADNSTAEGRAVNRRVEAEVEAQAK  351
```

Appendix 2: Instructor handout part 1, answers and misconceptions.

Learning objectives. *After completing this small group activity, you should be able to ….*

- Label and explain the function of key components in Gram positive and Gram negative bacteria
- Determine the predicted function of a protein sequence using BLAST
- Determine if a gene product is present in a specific organism using BLAST
- Evaluate sequence similarity based on BLAST outputs: E-values, % query cover, and % max identity

This activity introduces BLAST (Basic Local Alignment Search Tool), a valuable tool for analyzing nucleic acid and protein sequence data. In addition, this activity highlights some important differences between the cell envelopes of Gram positive and Gram negative bacteria.
Note: *students needed to complete Part I to obtain information for the in-class activity (Part II).*
Part I was checked at the beginning of class for completion. Students were asked to bring laptops to class.

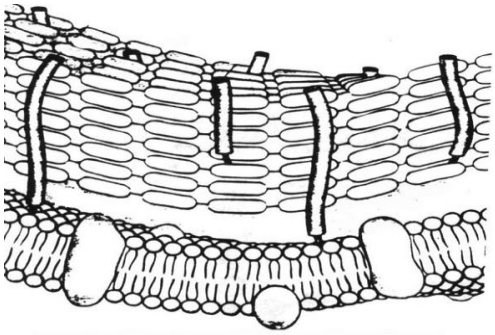
Part 1 (2 points)
NOTE: We did not grade these in detail, but gave students 1 point for labeling the cell, and 1 point for answering the BLAST questions. The 2 points were meant as a small incentive to complete the assignment.

A) Cell Envelope Review: Look at the diagrams below of Gram (+) and Gram (-) type cell envelopes.

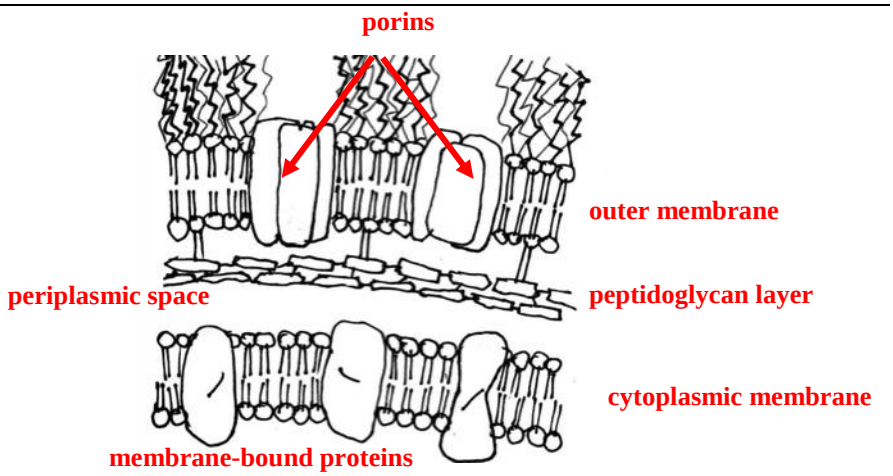
iii) **Label** each cell envelope as being from a Gram (+) or Gram (-) cell type
iv) **Label** the components of each cell envelope using the list below
Note: not all cell-types have all the structures.

- cytoplasmic membrane
- outer membrane
- membrane-bound proteins
- peptidoglycan layer
- periplasmic space
- porins

Cell wall type **GRAM (+)**



Cell wall type **GRAM (-)**



B) BLAST review: One of the bioinformatic tools that you will use is **BLAST** (**Basic Local Alignment Search Tool**), that can be found at the National Center for Biotechnology Information site: http://blast.ncbi.nlm.nih.gov/

As the name implies, BLAST makes alignments between sequences. Alignment is the process (or result) of matching up the nucleotide or amino acid residues of two or more biological sequences to achieve the best possible match. BLAST identifies sequences similar to your query sequence in the NCBI database by making alignments and assessing how well the sequences match.

Students were asked to view the *BLAST Video Tutorial* ((https://www.youtube.com/watch?v=x_dAyY5-VNc**) or the *BLAST PDF Tutorial* to answer the questions below.**

Below is the amino acid sequence of a protein associated with some bacterial cell envelopes. Use a protein BLAST (**BLASTP**) search to obtain information about it, and answer the questions below.

MKLKNTLGVVIGSLVAASAMNAFAQGQNSVEIEAFGKRYFTDSVRNMKNADLYGGSIGYF
LTDDVELALSYGEYHDVRGTYETGNKKVHGNLTSLDAIYHFGTPGVGLRPYVSAGLAHQNI
TNINSDSQGRQQMTMANIGAGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAGLGV
GFNFGGSKAAPAPEPVADVCSDSDNDGVCDNVDKCPDTPANVTVDANGCPAVAEVVRVQ
LDVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANA
VRDVLVNEYGVEGGRVNAVGYGESRPVADNATAEGRAINRRVEAEVEAEAK

Questions:

Identify the Top BLAST hit and fill in the box to answer questions 1-3.

1) What kind of protein does this sequence encode, based on the name given (annotation)?
2) From what organism did it come?
3) What is the BLAST % query cover, E value and % Max Identity for the top hit?

Top BLAST hit for the sequence from your isolate

Protein	Organism	% Query Coverage	E-value	% Max Identity
PORIN	***PSEUDOMONAS***	**100**	**0.0**	**100**

4) What is the function of this kind of protein?

Porins are membrane proteins that allow for diffusion of molecules in the cell. They can be specific to a specific type of molecule and are found in the outer membrane of GRAM (-) bacteria.

5) Based on what this protein does and where it is found, do you think this organism is a Gram positive or Gram negative bacterium? Explain your logic.

GRAM (-) because only Gram(-) bacteria have porins

6) Look at the BLAST tutorial (or look at the glossary section in the BLAST website (http://www.ncbi.nlm.nih.gov/books/NBK62051/) and fill in these definitions:

E-value:

The E-value represents how well the alignment of your query sequence is to the database sequences. The lower the E-value, or the closer it is to zero, the more significant the alignment and match are.

% Max Identity:

% Maximum identity is the percentage of residues that, after alignment of two sequences are in the same position in the alignment and match up.

7) When running a blast search, often times the sequences returned will align with only **part** of your query sequence. NCBI defines query coverage as the percent of the query sequence length that is included in the alignment. This number is significant because if figures into the calculation of the E value, the greater the query coverage, the lower the E value, the better the match. **Place an asterisk next to the blast hit (A or B) below with the higher query coverage (the two examples are different hits using the same query sequence):**

A)

putative outer membrane lipoprotein [Escherichia coli O26:H11 str. CVM9952]
Sequence ID: ref|ZP_14641000.1| Length: 219 Number of Matches: 1
▷ See 1 more title(s)

Range 1: 113 to 213 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
81.3 bits(199)	9e-17	Compositional matrix adjust.	41/102(40%)	62/102(60%)	1/102(0%)

```
Query  241  DVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANAV  300
            +V FD    + +K       +  +A  +K+YP T+  V G+TDS G    N +LS++RA++V
Sbjct  113  NVTFDSSSATLKPAGANTLTGVAMVLKEYPKTAVNVIGYTDSTGGHDLNMRLSQQRADSV  172

Query  301  RDVLVNEYGVEGGRVNAVGYGESRPVADNATAEGRAINRRVE  342
               L+ + GVE  R+   G G + P+A N+TAEG+A NRRVE
Sbjct  173  ASALITQ-GVEASRIRTQGLGPANPIASNSTAEGKAQNRRVE  213
```

B) ***

Major porin and structural outer membrane porin OprF precursor [Pseudomonas sp. M1]
Sequence ID: ref|ZP_19204629.1| Length: 353 Number of Matches: 1
▷ See 1 more title(s)

Range 1: 1 to 351 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
666 bits(1718)	0.0	Compositional matrix adjust.	317/351(90%)	339/351(96%)	1/351(0%)

```
Query  1    MKLKNTLGVVIGSLVAASAMNAFAQGQNSVEIEAFGKRYFTDSVRNMKNADLYGGSIGYF  60
            MKLKNTLGVVIGS++AASA+NAFAQGQ +VE EAFGKRYFTDS RNMKN DLYGGS+GYF
Sbjct  1    MKLKNTLGVVIGSMIAASAVNAFAQGQGAVEAEAFGKRYFTDSTRNMKNGDLYGGSVGYF  60

Query  61   LTDDVELALSYGEYHDVRGTYETGNKKVHGNLTSLDAIYHFGTPGVGLRPYVSAGLAHQN  120
            LTDDVELALSYGEYHD+RGTYE+GNKKVHGNL SLDAIYHFGTPGVGLRPYVSAG+ HQ+
Sbjct  61   LTDDVELALSYGEYHDIRGTYESGNKKVHGNLASLDAIYHFGTPGVGLRPYVSAGIGHQS  120

Query  121  ITNINSDSQGRQQMTMANIGAGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAGLGV  180
            +TN+NS++ GRQ +TMANIGAGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAG+GV
Sbjct  121  LTNVNSENGGRQNLTMANIGAGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAGVGV  180

Query  181  GFNF-GGSKAAPAPEPVADVCSDSDNDGVCDNVDKCPDTPANVTVDANGCPAVAEVVRVQ  239
            G NF GG+K APAPEPVA+VCSD D+DGVCDNVDKCP+TPANVTVDANGCPAVAEVVRVQ
Sbjct  181  GMNFGGGAKPAPAPEPVAEVCSDEDHDGVCDNVDKCPNTPANVTVDANGCPAVAEVVRVQ  240

Query  240  LDVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANA  299
            LDVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANA
Sbjct  241  LDVKFDFDKSKVKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANA  300

Query  300  VRDVLVNEYGVEGGRVNAVGYGESRPVADNATAEGRAINRRVEAEVEAEAK  350
            VRDVLVNEYGVEGGRVNAVGYGESRPVADN+TAEGRA+NRRVEAEVEA+AK
Sbjct  301  VRDVLVNEYGVEGGRVNAVGYGESRPVADNSTAEGRAVNRRVEAEVEAQAK  351
```

Appendix 3: Student handout part 2.

You should have your completed Part I & a laptop computer with you!

Pseudomonas aeruginosa is a pathogenic (disease causing) bacterium that can infect a wide variety of animals. *P. aeruginosa* is particularly devastating to patients suffering from Cystic Fibrosis (CF), a genetic disease that causes the buildup of thick mucus in the lungs. The dysfunctional lungs of CF patients are chronically infected with *P. aeruginosa*, which is well adapted to survive in this habitat, in part because it can efficiently utilize amino acids for carbon and energy.

While working in a clinical microbiology lab, you isolate a new strain of *P. aeruginosa* that thrives especially well in CF patients. Comparing its protein expression patterns to previous, non-CF isolates, ***you find that one protein is highly expressed in your isolate relative to other P. aeruginosa strains.***

The amino acid sequence is pasted below. An electronic version can be found in the accompanying file: **Part 2.sequence.doc**

MRTYFERLSAGMALALCTASAAWADEADAKEGFIEGSSLQLLTRNYYFNHDRRHASGHDSKEWA
QGFIATFQSGYTPGVVGFGVDAYGMLGLKLDGGGGTGGTSILPITSPTKDGYESGKAPDEFSSGGAA
LKIRAFDTELKLGDQFLSNPVVAGGESRMLPQTFRGVSLTNNSFEDLTTAGQVSFTKYYNQSGHRRL
GSYYGELPGDRDSHHLSWGGTWGGIEGFTSSLYAAELQNVWKQYYADVDYTYEIDDNWSLNPGA
HYYKTVDSGDSLLGDRIDNNTYSLHFAVGYRQHTVTAVLQKVNGNTPFGDYINQGSIFLDNSQQYS
DFNGPPNEKSWKLQYDYDFVALGLPGLSASASYSRGKLDLTRVDPDSPGYGGWYSADGKAKHWE
RDLDLQYVVQGGPAKDLSLRLRWATHRGTGGYSAVDNDIDEYRVTSSGLKASDTG

To find out what the function of this protein might be, you perform a BLAST (**Basic Local Alignment Search Tool**) search of its amino acid sequence. Go to the **National Center for Biotechnology Information** site (http://blast.ncbi.nlm.nih.gov/) to do a BLASTP search to determine a potential identity of this protein, following the same procedures as Part I and in the BLAST tutorial.

Answer the following questions, adding answers 1-4 to the table for *P. aeruginosa*:

1) What is the name/description of the top BLAST Hit? _____

2) What is the % query coverage? _____

3) What is the E value? _____

4) What is the % maximum identity? _____

Top BLAST hits for the sequence from your isolate

Organism	Protein Name	% Query Coverage	E-value	% Max Identity
P.aeruginosa				

5) Based on the name, what do you think the function of this protein is? (1 point)

6) Why might expressing this protein at a high level help your clinical isolate thrive in patients' lungs? (1 point)

As you scroll down the table giving descriptions of the BLAST hits, you notice that **similar proteins occur in other *Pseudomonas* species besides *aeruginosa*.** You are curious about how widespread this protein may be, so you decide to search the genomes of two well-studied bacteria for similar sequences: *Bacillus subtilis* (Gram positive) and *Escherichia coli* (Gram negative).

Use separate windows or browser tabs for each BLAST search to compare the results.

- Navigate to the BLASTP page and enter the sequence from your *Pseudomonas aeruginosa* isolate as query.
- Under "Choose Search Set" on the same page, find the box where it says "Enter organism name."
- Type in **"Bacillus subtilis (taxid: 1423)." Be sure your text matches this exactly!** This will search the subset of sequences in the NCBI database that come from *B. subtilis*.
- Click the BLAST button
- Repeat these steps in another internet browser window, entering **"Escherichia coli (taxid: 562)"** into the "enter organism name" box, then clicking BLAST. This will search the subset of NCBI sequences from *E. coli*.

7) Fill in the table with your results (add your results from the on the previous page).

Top BLAST hits for the sequence from your isolate

Organism	Protein Name	% Query Coverage	E-value	% Max Identity
P.aeruginosa				
B. subtilis 1423				
E. coli 562				

8) a) Look at your data table. **Based on the E-value data, do you think that either *E. coli* or *B. subtilis* carries the gene for this protein? Explain your answer. (1 point)**

b) Look at your data table. **Based on the % Query Coverage data, do you think that either *E. coli* or *B. subtilis* carries the gene for this protein? Explain your answer. (1 point)**

c) Do these data support what you know about the cell envelope structure of *E. coli* and *B. subtilis*? Explain why or why not. (2 points)

9) How can *Bacillus subtillis* have a higher % Max Identity than *E. coli* but a lower % Query Coverage? Explain this phenomenon. (2 points)

Appendix 4: Instructor handout part 2, answers and misconceptions.

Pseudomonas aeruginosa is a pathogenic (disease causing) bacterium that can infect a wide variety of animals. *P. aeruginosa* is particularly devastating to patients suffering from Cystic Fibrosis (CF), a genetic disease that causes the buildup of thick mucus in the lungs. The dysfunctional lungs of CF patients are chronically infected with *P. aeruginosa*, which is well adapted to survive in this habitat, in part because it can efficiently utilize amino acids for carbon and energy.

While working in a clinical microbiology lab, you isolate a new strain of *P. aeruginosa* that thrives especially well in CF patients. Comparing its protein expression patterns to previous, non-CF isolates, ***you find that one protein is highly expressed in your isolate relative to other P. aeruginosa strains.***

The amino acid sequence is pasted below. An electronic version can be found in the accompanying file: **Part 2.sequence.doc**

MRTYFERLSAGMALALCTASAAWADEADAKEGFIEGSSLQLLTRNYYFNHDRRHASGHDSKEWA
QGFIATFQSGYTPGVVGFGVDAYGMLGLKLDGGGGTGGTSILPITSPTKDGYESGKAPDEFSSGGAA
LKIRAFDTELKLGDQFLSNPVVAGGESRMLPQTFRGVSLTNNSFEDLTTAGQVSFTKYYNQSGHRRL
GSYYGELPGDRDSHHLSWGGTWGGIEGFTSSLYAAELQNVWKQYYADVDYTYEIDDNWSLNPGA
HYYKTVDSGDSLLGDRIDNNTYSLHFAVGYRQHTVTAVLQKVNGNTPFGDYINQGSIFLDNSQQYS
DFNGPPNEKSWKLQYDYDFVALGLPGLSASASYSRGKLDLTRVDPDSPGYGGWYSADGKAKHWE
RDLDLQYVVQGGPAKDLSLRLRWATHRGTGGYSAVDNDIDEYRVTSSGLKASDTG

To find out what the function of this protein might be, you perform a BLAST (**Basic Local Alignment Search Tool**) search of its amino acid sequence. Go to the **National Center for Biotechnology Information** site (http://blast.ncbi.nlm.nih.gov/) to do a BLASTP search to determine a potential identity of this protein, following the same procedures as Part I and in the BLAST tutorial.

Answer the following questions, adding answers 1-4 to the table for *P. aeruginosa*:

1) What is the name/description of the top BLAST Hit? histidine porin OpdC

2) What is the % query coverage? 97%

3) What is the E value? 0%

4) What is the % maximum identity? 96%

Top BLAST hits for the sequence from your isolate

Organism	Protein Name	% Query Coverage	E-value	% Max Identity
P.aeruginosa	**histidine porin OpdC**	**97**	**0**	**96**

5) Based on the name, what do you think the function of this protein is? (1 point)

Answer: As a porin, it transports histidine through the outer membrane into the cell.

We gave ½ point for mentioning histidine transport and ½ point for noting the outer membrane

Misconceptions: A few students thought the porin was called a his porin because it was made of histidine

6) Why might expressing this protein at a high level help your clinical isolate thrive in patients' lungs? (1 point)

Answer: It helps the bacterium to efficiently transport amino acids (histidine) for biosynthesis and energy, thus giving cells a growth advantage.

We gave 1 point for mentioning more efficient transport

As you scroll down the table giving descriptions of the BLAST hits, you notice that **similar proteins occur in other *Pseudomonas* species besides *aeruginosa*.** You are curious about how widespread this protein may be, so you decide to search the genomes of two well-studied bacteria for similar sequences: *Bacillus subtilis* (Gram positive) and *Escherichia coli* (Gram negative).

Use separate windows or browser tabs for each BLAST search to compare the results.

- Navigate to the BLASTP page and enter the sequence from your *Pseudomonas aeruginosa* isolate as query.
- Under "Choose Search Set" on the same page, find the box where it says "Enter organism name."
- Type in **"Bacillus subtilis (taxid: 1423)." Be sure your text matches this exactly!** This will search the subset of sequences in the NCBI database that come from *B. subtilis*.
- Click the BLAST button
- Repeat these steps in another internet browser window, entering **"Escherichia coli (taxid: 562)"** into the "enter organism name" box, then clicking BLAST. This will search the subset of NCBI sequences from *E. coli*.

7) Fill in the table with your results (add your results from the on the previous page).

Top BLAST hits for the sequence from your isolate

Organism	Protein Name	% Query Coverage	E-value	% Max Identity
P.aeruginosa	**histidine porin OpdC**	**97**	**0**	**96**
B. subtilis 1423	**Hypothetical protein**	**7**	**8.8**	**34**
E. coli 562	**Outer membrane porin**	**93**	**6×10^{-4}**	**22**

NOTE: Due to the dynamic nature of the BLAST database, the actual values could vary. Facilitators could ask: what do you think "hypothetical protein" means?

8) a) Look at your data table. Based on the E-value data, do you think that either *E. coli* or *B. subtilis* carries the gene for this protein? Explain your answer. (1 point)

Answer: E.coli should carry this gene because it has an E-value closer to 0, and the lower the E-value the more significant the score and the alignment. Therefore, the Ecoli gene sequence has a better alignment to the porin sequence of Pseudomonas, while the Bacillus E-value is much higher and therefore the sequence is not as similar.

(1 point was given for answers that reflected an understanding of how to interpret the E-value.)

b) Look at your data table. Based on the % Query Coverage data, do you think that either *E. coli* or *B. subtilis* carries the gene for this protein? Explain your answer. (1 point)

Answers: E.coli more likely carries this gene because it has a higher % Query Coverage. . This large number means that 93% of the Ecoli sequence matches the query sequence. The Bacillus % Query Coverage is much lower. This small number means that only 7% of the Bacillus sequence matches the query sequence. Therefore, the Ecoli gene sequence has a better alignment to the porin sequence of Pseudomonas.

(1 point was given for answers that reflected an understanding of how to interpret the % Query Coverage.)

NOTE: Students were referred to the graphical output on page 2 of Part 1 that shows Bacillus with a tiny bit of coverage and Ecoli with almost the entire coverage.

c) Do these data support what you know about the cell envelope structure of *E. coli* and *B. subtilis*? Explain why or why not. (2 points)

Answer: Yes, E coli is Gram negative and therefore would have an outer membrane that would support porins, which are outer membrane proteins. Bacillus shouldn't have porins because it is Gram positive and therefore lacks an outer membrane.

(1 point was given for noting that *E.coli* (a Gram (-) bacterium) has porins and
1 point was given for noting that *Bacillus* (a Gram (+) bacterium) does not.)

Misconceptions:

The presence of the histidine porin protein suggests E.coli is a Gram negative organism (rather than saying E.coli is Gram negative, therefore has a porin)

The data are consistent since E coli is Gram negative. [Some students did not state that Gram negative bacteria have an outer membrane that and will therefore have a porin, while Gram positive lack an outer membrane and therefore will lack a porin.]

9) How can *Bacillus subtillis* have a higher % Max Identity than *E. coli* but a lower % Query Coverage? Explain this phenomenon. (2 points)

Answer: The lower % Query Coverage of the Bacillus sequence indicates that there is not a lot of overlap (only 7%) with the Pseudomonas gene. However, higher % Max Identity (96%) indicates that what did overlap matched very well. So the Bacillus gene contained a small segment that matched well.

The higher % Query Coverage of Ecoli shows the sequence had much more overlap with the Pseudomonas gene, (93%). but the relatively lower %Max Identity (22%) shows that a fair number of the sequences did not match exactly. Still, the % query coverage indicates much more overall similarity. One should look at the % query coverage first, before considering % max identity.

(1 point was given for answers that reflected an understanding of how to interpret the % Query Coverage; anther point was given for answers that reflected an understanding of how to interpret the % Max Identity.)

NOTE: Students were referred to the graphical output on page 2 of Part 1 that shows Bacillus with a tiny bit of coverage and Ecoli with almost the entire coverage.

Misconceptions:

-Bacillus has one big query that matches the sequence we have, which might lead to the high % max identity. E.coli found more match all across the query, but couldn't find the one big area where the sequences exactly matches with each other, leading to lower % max identity. [Did not realize that Bacillus has a short alignment with the query sequence and that E.coli overall has a higher number of amino acids that match the query].

-It has a higher % identity due to chance, as the E value is relatively high. Chance results like that account for this phenomenon.[Does not understand how BLAST works]

-The queries for B. subtilis have a greater number of matching residues to the protein sequence we entered, however, E coli results in a greater number of overall queries across the target sequences. This explains why B subtillis, which it has a higher % max identity, has only one query and therefore a less significant E value and lower % query [Does not understand that it is amino acid matches and that B subtilis has a small number of total amino acids that match]

Appendix 5: Part 2 amino acid sequence.

MRTYFERLSAGMALALCTASAAWADEADAKEGFIEGSSLQLLTRNYYFNHDRRHASGHDSKEWAQG
FIATFQSGYTPGVVGFGVDAYGMLGLKLDGGGGTGGTSILPITSPTKDGYESGKAPDEFSSGGAALKIR
AFDTELKLGDQFLSNPVVAGGESRMLPQTFRGVSLTNNSFEDLTTAGQVSFTKYYNQSGHRRLGSYY
GELPGDRDSHHLSWGGTWGGIEGFTSSLYAAELQNVWKQYYADVDYTYEIDDNWSLNPGAHYYKT
VDSGDSLLGDRIDNNTYSLHFAVGYRQHTVTAVLQKVNGNTPFGDYINQGSIFLDNSQQYSDFNGPPN
EKSWKLQYDYDFVALGLPGLSASASYSRGKLDLTRVDPDSPGYGGWYSADGKAKHWERDLDLQYV
VQGGPAKDLSLRLRWATHRGTGGYSAVDNDIDEYRVTSSGLKASDTG

Appendix 6: Pre- and post-test questions, answers and misconceptions.

For our analysis, answers to these questions were marked as being either correct (+1) or incorrect (0)

PRE-TEST

_____ 1) Are you familiar with **BLAST**?

A. I know what it is and how to use it.
B. I have some idea of what it is, but I don't know how to do it.
C. I have heard of it, but I do not know what it is.
D. I have never heard of it.

If you chose option A or B, explain for what BLAST is used.

Answer: *BLAST finds similar sequences in a sequence database by making alignments and assessing how well the sequences match; used to find genes or proteins in a genome*

Misconceptions: *Helps to visualize protein structures, used to study structure of biological molecules*

__A__ 2) Which **e-value** would indicate a very good match for a **protein sequence BLAST**?

A. 0.0
B. 0.5
C. 1.0
D. I don't know

POST-TEST

__C__ 1) BLAST is a tool that…

A. applies sound energy to disrupt cell envelopes.
B. builds a phylogenetic tree.
C. finds regions of local similarity between sequences.
D. removes gaps from sequence alignments.
E. I don't know.

__A__2) Which e-value would indicate a very good match for a protein sequence BLAST?

A. 0.0
B. 0.5
C. 1.0
D. I don't know

3) The results below are from a BLAST search using the FunE protein of *E. comica.* Based on these data, which bacterium MOST likely contains a FunE protein? Pick one and explain your answer

Bacterium	% query coverage	% max identity
Y.Gabbagabbaea	20	80
H. simpsonius	30	30
P. Griffinia	80	50

Answer: *P. griffinia* has the best alignment with a higher percentage of the sequence (%QC)

Misconceptions: *Y.Gabbagabbaea: has the highest % max identity; has the highest combined values, has the best alignment with a small part of the sequence.*

Appendix 7: BLAST tutorial.

You are interested in finding out if this sequence codes for a protein with an interesting function!

To do this, you will use a Basic Local Alignment Search Tool on the National Center for Biotechnology Information.

Part 1A

1. Go to the BLAST website at NCBI: http://blast.ncbi.nlm.nih.gov/.

2. Click on "protein blast".

3. Enter the protein sequence into the box labeled "Enter accession number(s), gi(s), or FASTA sequence(s)". You can copy + paste the sequence.

4. Under database, make sure "non-redundant protein sequence (nr)" is selected.

5. Click on "BLAST".

6. Wait until sequence has been completely processed.

7. Scroll down to "Descriptions".

8. Fill in the required information on your Small Groups sheet.

Part 1B

To find information on how BLAST works and definitions of the term.

1. Go to http://www.ncbi.nlm.nih.gov/Web/Newsltr/V15N2/BLView.html.

2. Find the required information on your Small Groups sheet.

Part II- You will need to know this for class.

1. Go to the BLAST website at NCBI: http://blast.ncbi.nlm.nih.gov/ .

2. Click on "protein blast".

3. Enter the protein sequence into the box labeled "Enter accession number(s), gi(s), or FASTA sequence(s)". You can copy + paste the sequence.

4. Under "Database", make sure "non-redundant protein sequence (nr)" is selected.

5. Under "Organism", type in the name indicated in your Small Groups packet.

6. Click on "BLAST".

7. Wait until sequence has been completely processed.

8. Scroll down to "Descriptions".

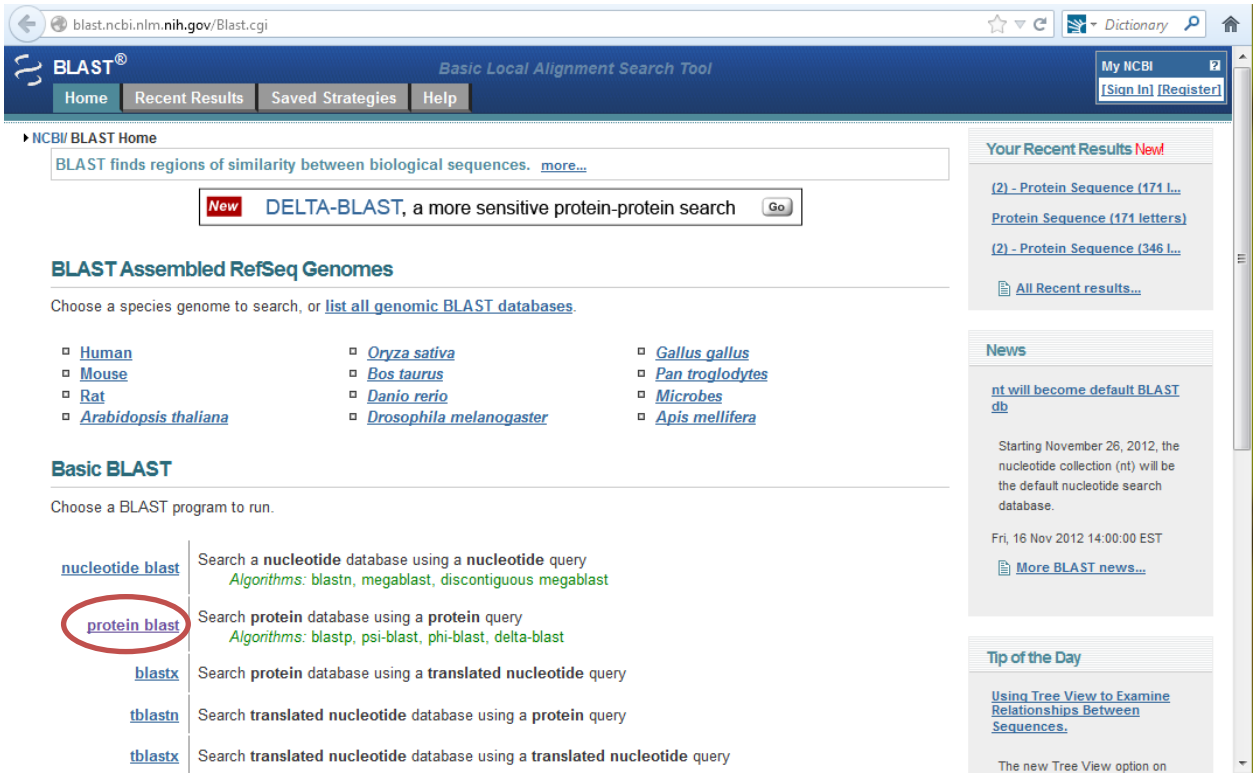
9. Fill in the required information on your Small Groups sheet.

Detailed tutorial using the following protein sequence:

MKKIACLSALAAVLAFTAGTSVAATSTVTGGYAQSDAQGQMNKMGGFNLKYRYEEDNSPLGVIGS
FTYTEKSRTASSGDYNKNQYYGITAGPAYRINDWASIYGVVGVGYGKFQTTEYPTYKHDTSDYGFS
YGAGLQFNPMENVALDFSYEQSRIRSVDVGTWIAGVGYRF

1. Go to the BLAST website at NCBI: http://blast.ncbi.nlm.nih.gov/.

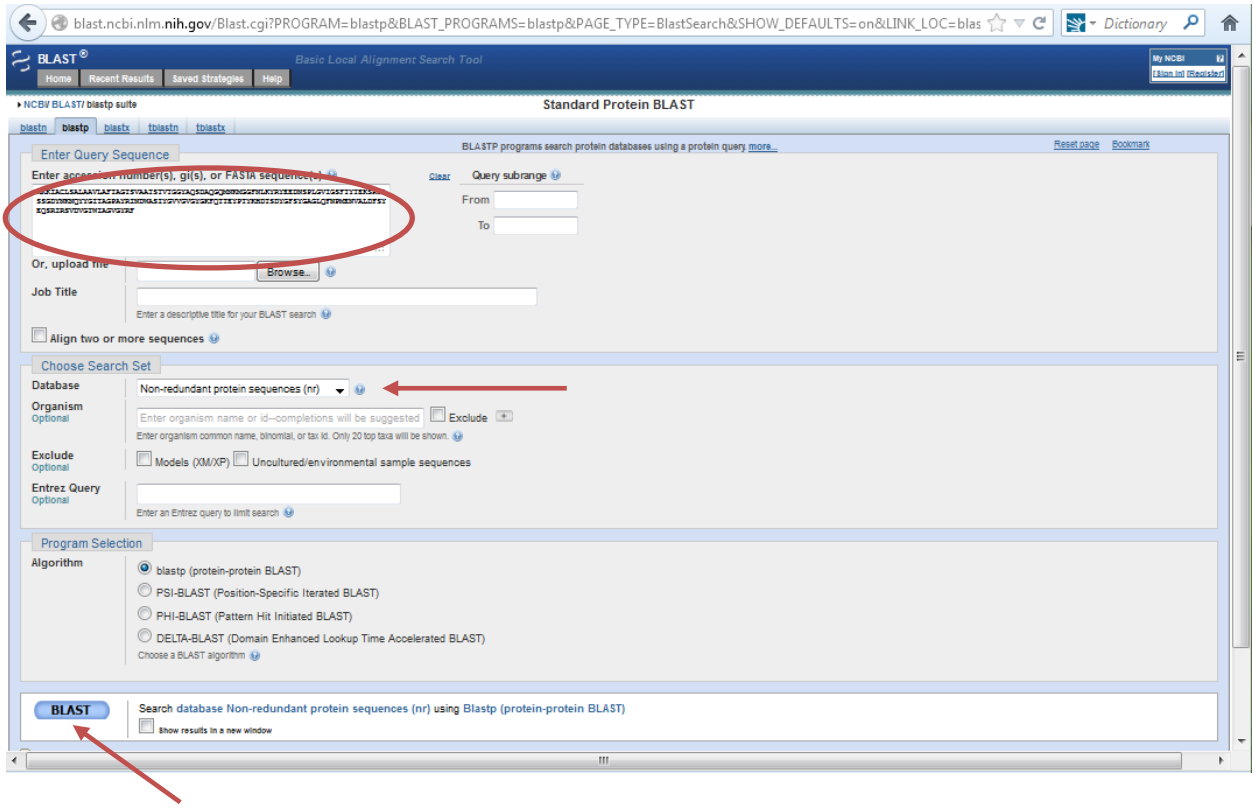
2. On that page, click on protein blast.



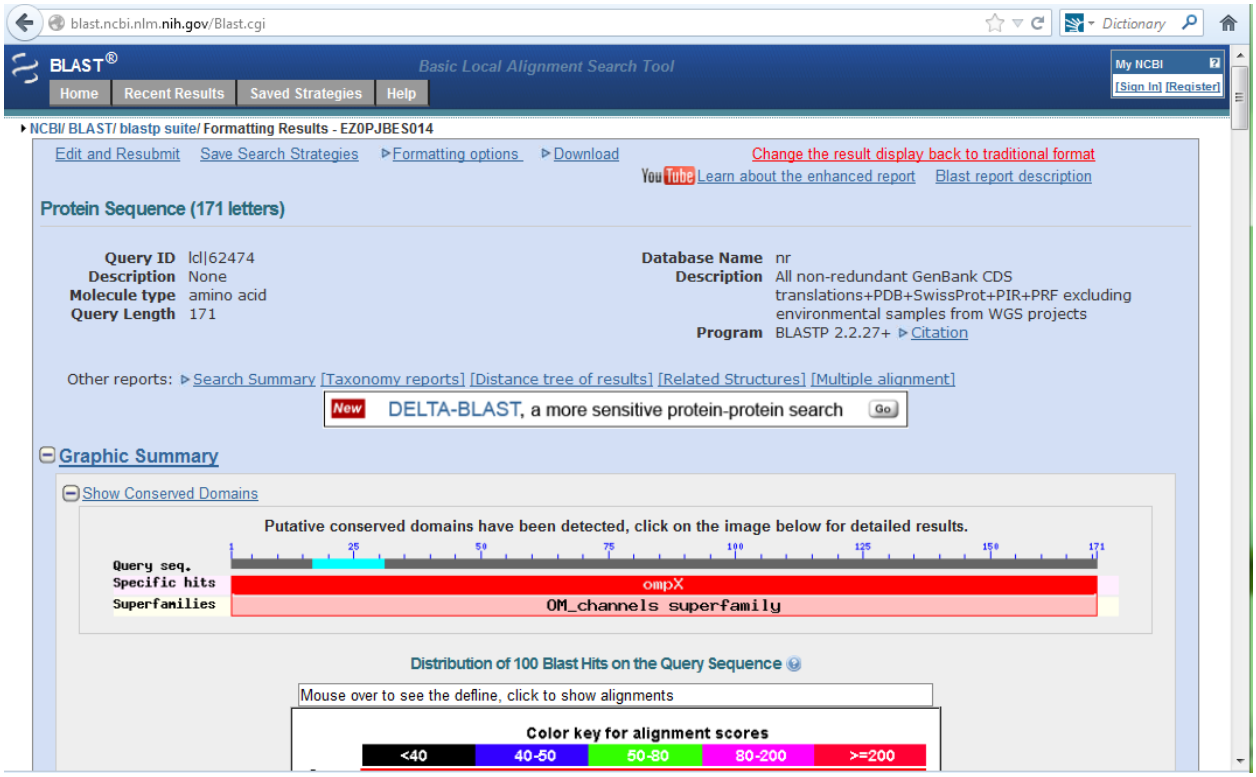
3. Enter the protein sequence into the box labeled "**Enter accession number(s), gi(s), or FASTA sequence(s)**". You can copy + paste the sequence.

4. Under database, make sure "non-redundant protein sequence (nr)" is selected.

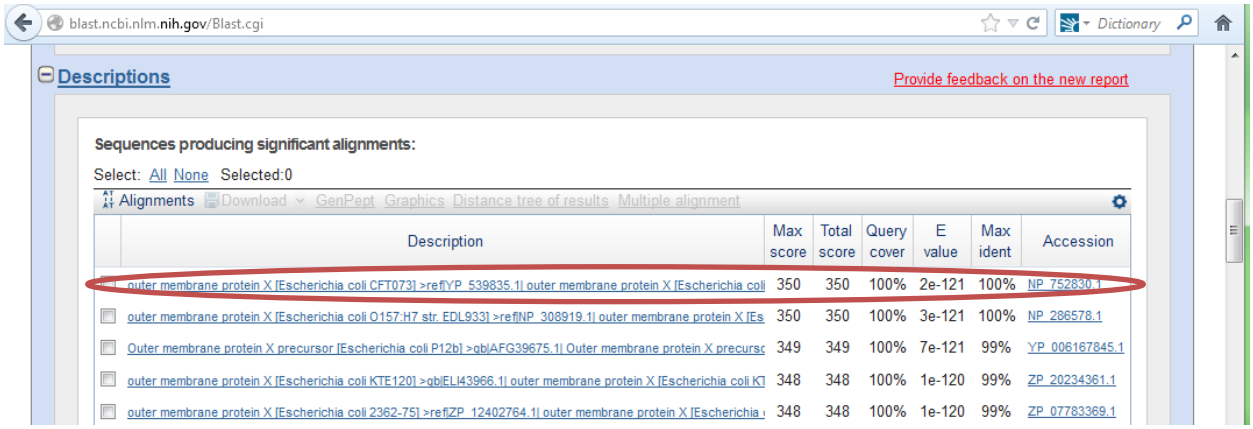
5. Click on "BLAST".


blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blas
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
Standard Protein BLAST
Enter Query Sequence
Query subrange
From
To
Or, upload file
Browse...
Job Title
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Exclude
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

…wait…


blast.ncbi.nlm.nih.gov/Blast.cgi
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite/ Formatting Results - EZ0PJBES014
Edit and Resubmit
Save Search Strategies
Formatting options
Download
Change the result display back to traditional format
Learn about the enhanced report
Blast report description
Protein Sequence (171 letters)
Query ID lcl|62474
Description None
Molecule type amino acid
Query Length 171
Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.2.27+ Citation
Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment]
New DELTA-BLAST, a more sensitive protein-protein search Go
Graphic Summary
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
Query seq.
Specific hits
Superfamilies
ompX
OM_channels superfamily
Distribution of 100 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200

6. Scroll down to “Descriptions”.



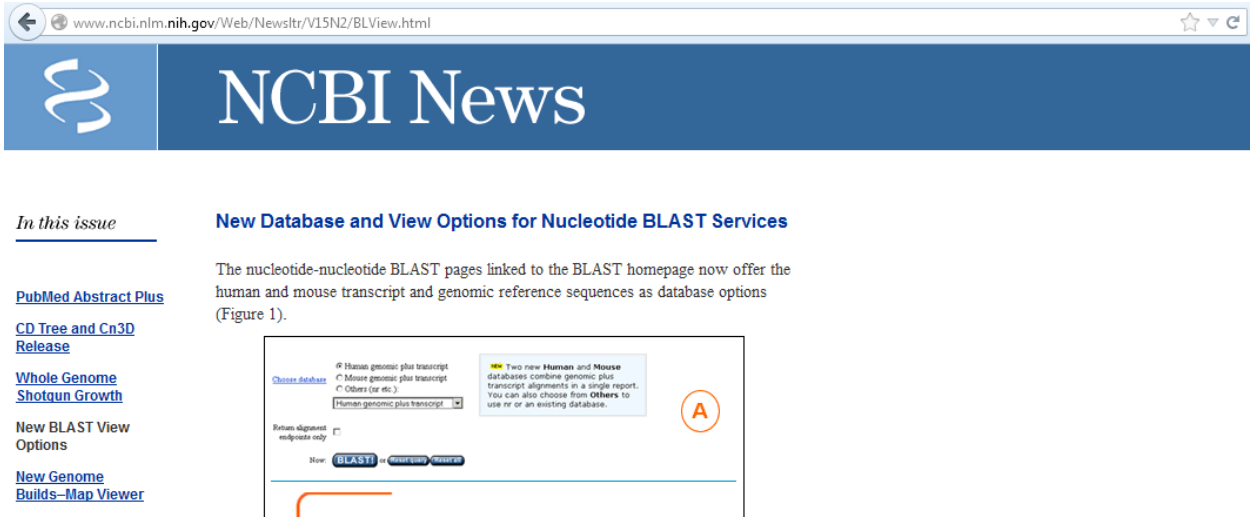
7. Fill in the required information on your Small Groups sheet.

Part 1B

To find information on how BLAST works and definitions of the term.

1. Go to http://www.ncbi.nlm.nih.gov/Web/Newsltr/V15N2/BLView.html.

2. Find the required information on your Small Groups sheet.



Part II

5. Under "Organism", type in the name indicated in your Small Groups packet.