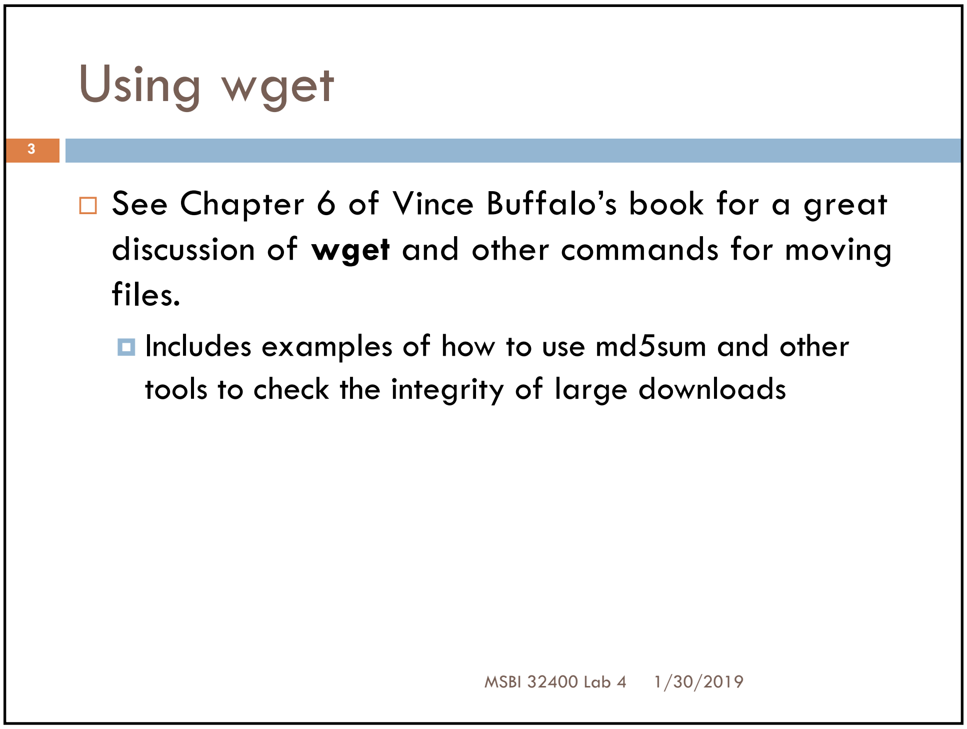
MSBI 32400 – LAB 4
LARRY HELSETH, PHD AND
JASON EDELSTEIN

January 30, 2019

Outline

- □ Install a reference proteome for local BLAST on VM
- □ Configure Linux environment to run BLAST
- □ Run searches
  - ◘ Output to text or console
  - ◘ Output to HTML
- □ Learn to run 2 way BLAST
- □ Identify some unknown sequences

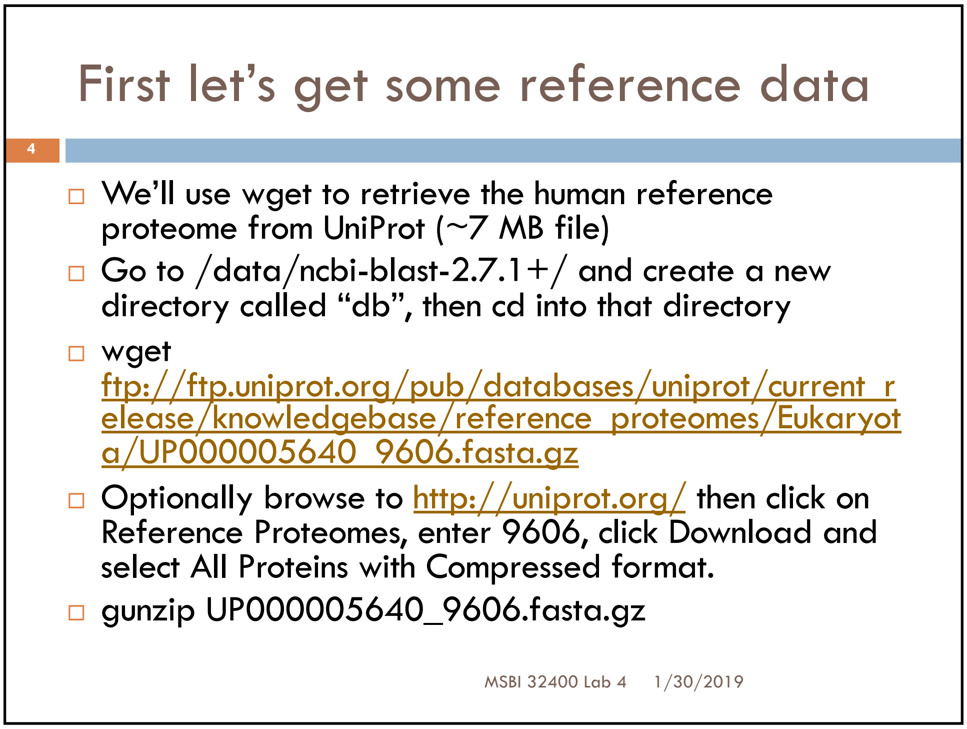


Using wget

3

- See Chapter 6 of Vince Buffalo's book for a great discussion of **wget** and other commands for moving files.
  - Includes examples of how to use md5sum and other tools to check the integrity of large downloads

MSBI 32400 Lab 4 1/30/2019



First let's get some reference data

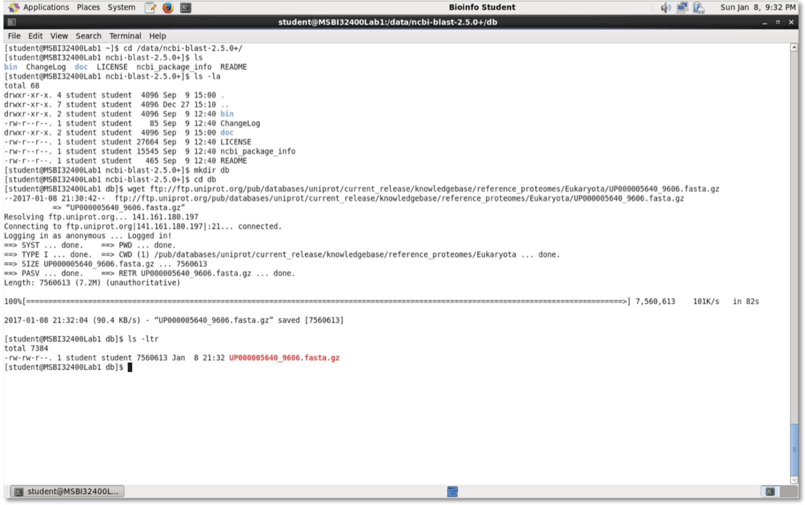
4

- We'll use wget to retrieve the human reference proteome from UniProt (~7 MB file)
- Go to /data/ncbi-blast-2.7.1+/ and create a new directory called "db", then cd into that directory
- wget ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640_9606.fasta.gz
- Optionally browse to http://uniprot.org/ then click on Reference Proteomes, enter 9606, click Download and select All Proteins with Compressed format.
- gunzip UP000005640_9606.fasta.gz

MSBI 32400 Lab 4 1/30/2019

Getting the reference data

5


Applications Places System
Bioinfo Student
Sun Jan 8, 9:32 PM
student@MSBI32400Lab1:/data/ncbi-blast-2.5.0+/db
File Edit View Search Terminal Help
student@MSBI32400L...

```
[student@MSBI32400Lab1 ~]$ cd /data/ncbi-blast-2.5.0+/
[student@MSBI32400Lab1 ncbi-blast-2.5.0+]$ ls
bin  ChangeLog  doc  LICENSE  ncbi_package_info  README
[student@MSBI32400Lab1 ncbi-blast-2.5.0+]$ ls -la
total 68
drwxr-xr-x. 4 student student  4096 Sep  9 15:00 .
drwxr-xr-x. 7 student student  4096 Dec 27 15:10 ..
drwxr-xr-x. 2 student student  4096 Sep  9 12:40 bin
-rw-r--r--. 1 student student    85 Sep  9 12:40 ChangeLog
drwxr-xr-x. 2 student student  4096 Sep  9 15:00 doc
-rw-r--r--. 1 student student 27664 Sep  9 12:40 LICENSE
-rw-r--r--. 1 student student 15545 Sep  9 12:40 ncbi_package_info
-rw-r--r--. 1 student student   465 Sep  9 12:40 README
[student@MSBI32400Lab1 ncbi-blast-2.5.0+]$ mkdir db
[student@MSBI32400Lab1 ncbi-blast-2.5.0+]$ cd db
[student@MSBI32400Lab1 db]$ wget ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640_9606.fasta.gz
--2017-01-08 21:30:42--  ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640_9606.fasta.gz
           => “UP000005640_9606.fasta.gz”
Resolving ftp.uniprot.org... 141.161.180.197
Connecting to ftp.uniprot.org|141.161.180.197|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota ... done.
==> SIZE UP000005640_9606.fasta.gz ... 7560613
==> PASV ... done.    ==> RETR UP000005640_9606.fasta.gz ... done.
Length: 7560613 (7.2M) (unauthoritative)

100%[==================================================================================================================>] 7,560,613    101K/s   in 82s

2017-01-08 21:32:04 (90.4 KB/s) - “UP000005640_9606.fasta.gz” saved [7560613]

[student@MSBI32400Lab1 db]$ ls -ltr
total 7384
-rw-rw-r--. 1 student student 7560613 Jan  8 21:32 UP000005640_9606.fasta.gz
[student@MSBI32400Lab1 db]$ 
```

MSBI 32400 Lab 4 1/30/2019

How many proteins?

6

- □ Extract the fasta file using gunzip UP000005640_9606.fasta.gz
- □ grep the new fasta file for “^>” (lines that start with the greater than symbol, meaning they have FASTA annotation) and **pipe** that to the word count program **wc** with the -l (just show the number of lines)
- □ Include that information (and the wget command) in your Lab 4 README file

MSBI 32400 Lab 4 1/30/2019

Let’s add ncbi binaries to PATH

7

- Tell your server where to find the BLAST binaries:
  - PATH=$PATH:/data/ncbi-blast-2.7.1+/bin
  - export PATH

“Hidden” files in Linux

8

- If you type “ls -l” in your home directory you’ll see some files. If you type “ls -la” you will start seeing things like “.bash_profile”, “.bashrc”, etc.
  - Your terminal history is kept in “.bash_history”
- The preceding “.” hides the file or directory from ls unless you ask it to show all
- Most are program or user profile files & directories
- .bash_profile and .bashrc contains your environmental variables (loaded when you sign in)

BLAST lab

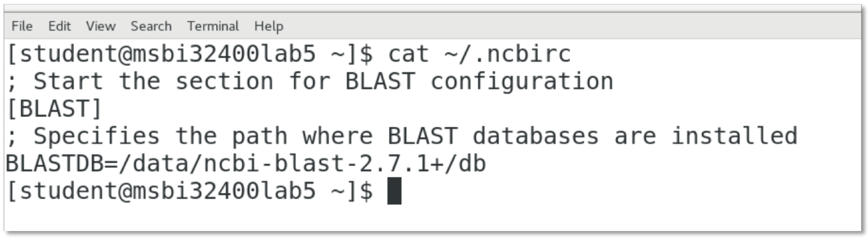
9

- Need a hidden file for BLAST to work correctly
- See **Recipe 11: Running Blast** from Safari on-line version of “**Managing Your Biological Data with Python**” by Allegra Via, Kristian Rother & Anna Tramontano, CRC Press 2014 (pp 431-435 print)
  - (Files/Lab4)
- Create a .ncbirc file in your <u>home</u> directory (/home/student)
  - vi ~/.ncbirc (or nano or gedit)

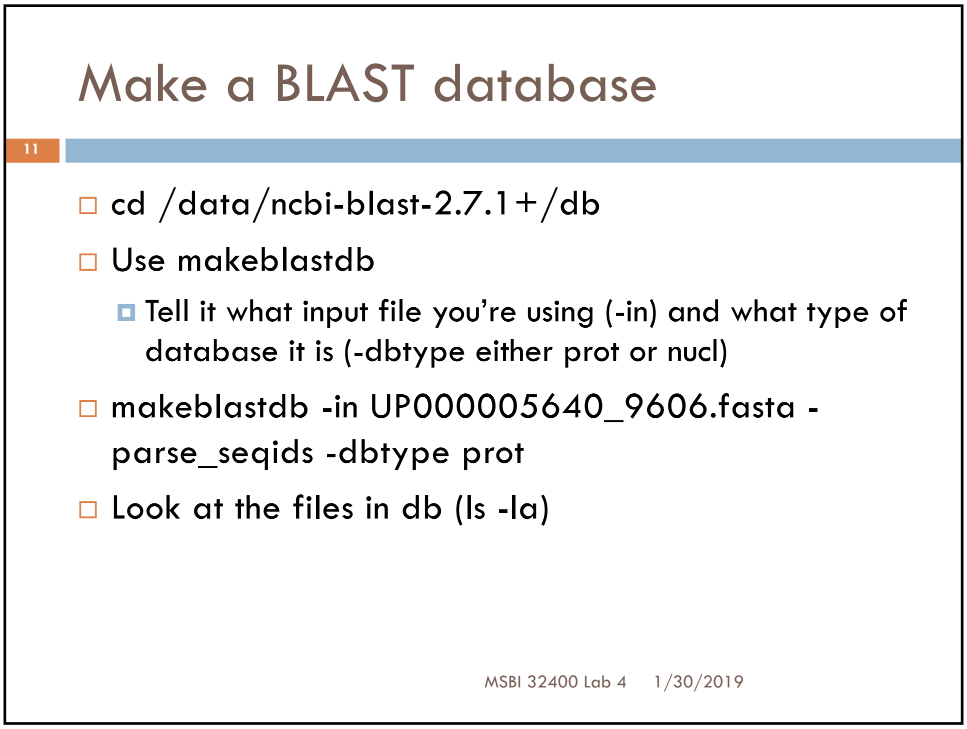
Syntax for .ncbirc

10

- ; Start the section for BLAST configuration
  [BLAST]
  ; Specifies the path where BLAST databases are installed
  BLASTDB=/data/ncbi-blast-2.7.1+/db



```
[student@msbi32400lab5 ~]$ cat ~/.ncbirc
; Start the section for BLAST configuration
[BLAST]
; Specifies the path where BLAST databases are installed
BLASTDB=/data/ncbi-blast-2.7.1+/db
[student@msbi32400lab5 ~]$ 
```

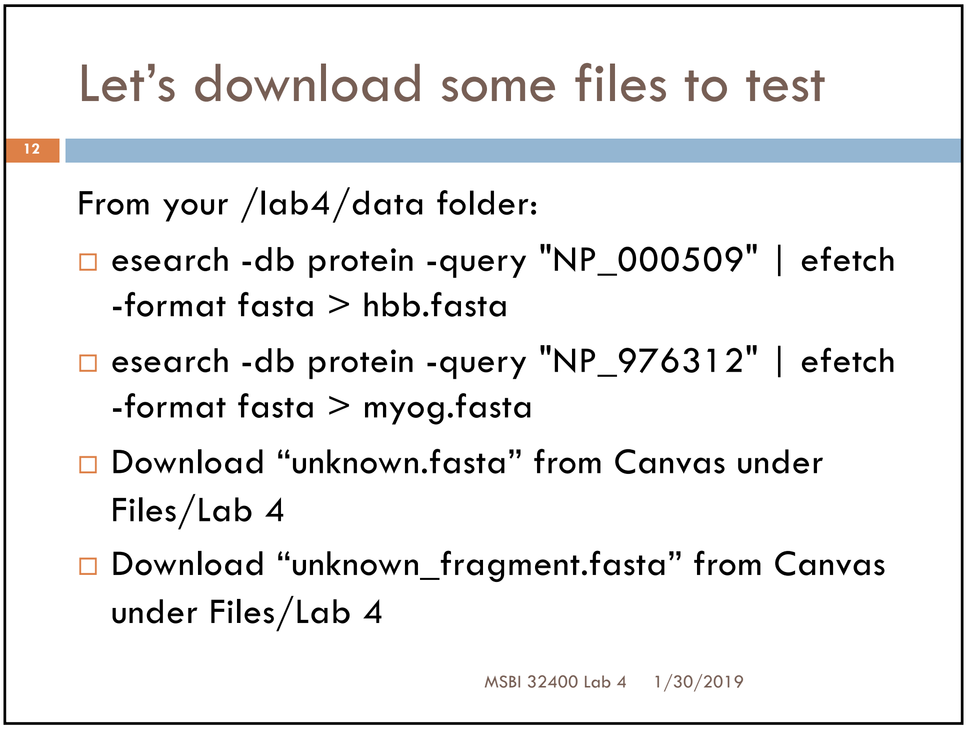


Make a BLAST database

11

- cd /data/ncbi-blast-2.7.1+/db
- Use makeblastdb
  - Tell it what input file you're using (-in) and what type of database it is (-dbtype either prot or nucl)
- makeblastdb -in UP000005640_9606.fasta -parse_seqids -dbtype prot
- Look at the files in db (ls -la)

MSBI 32400 Lab 4 1/30/2019



Let's download some files to test

12

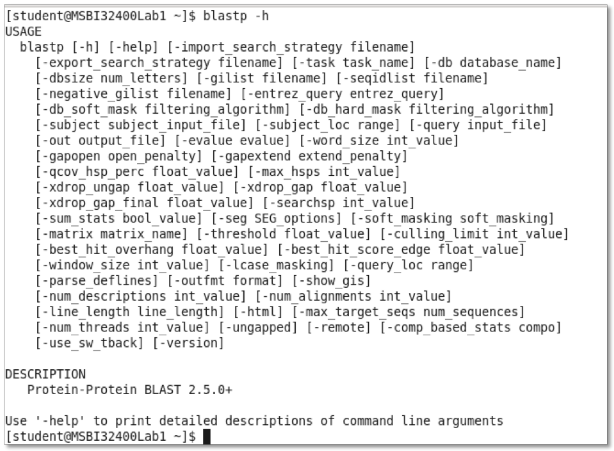
From your /lab4/data folder:

- esearch -db protein -query "NP_000509" | efetch -format fasta > hbb.fasta
- esearch -db protein -query "NP_976312" | efetch -format fasta > myog.fasta
- Download "unknown.fasta" from Canvas under Files/Lab 4
- Download "unknown_fragment.fasta" from Canvas under Files/Lab 4

MSBI 32400 Lab 4 1/30/2019

Running blastp (comparing proteins)

Check your PATH if you don't get this:



```
[student@MSBI32400Lab1 ~]$ blastp -h
USAGE
  blastp [-h] [-help] [-import_search_strategy filename]
    [-export_search_strategy filename] [-task task_name] [-db database_name]
    [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
    [-negative_gilist filename] [-entrez_query entrez_query]
    [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
    [-subject subject_input_file] [-subject_loc range] [-query input_file]
    [-out output_file] [-evalue evalue] [-word_size int_value]
    [-gapopen open_penalty] [-gapextend extend_penalty]
    [-qcov_hsp_perc float_value] [-max_hsps int_value]
    [-xdrop_ungap float_value] [-xdrop_gap float_value]
    [-xdrop_gap_final float_value] [-searchsp int_value]
    [-sum_stats bool_value] [-seg SEG_options] [-soft_masking soft_masking]
    [-matrix matrix_name] [-threshold float_value] [-culling_limit int_value]
    [-best_hit_overhang float_value] [-best_hit_score_edge float_value]
    [-window_size int_value] [-lcase_masking] [-query_loc range]
    [-parse_deflines] [-outfmt format] [-show_gis]
    [-num_descriptions int_value] [-num_alignments int_value]
    [-line_length line_length] [-html] [-max_target_seqs num_sequences]
    [-num_threads int_value] [-ungapped] [-remote] [-comp_based_stats compo]
    [-use_sw_tback] [-version]

DESCRIPTION
   Protein-Protein BLAST 2.5.0+

Use '-help' to print detailed descriptions of command line arguments
[student@MSBI32400Lab1 ~]$ █
```

Let's BLAST some proteins

- blastp -query hbb.fasta -db /data/ncbi-blast-2.7.1+/db/UP000005640_9606.fasta
  - This sends the output to console. You can either pipe this to more or redirect to a file then view that with less or more
- blastp -query hbb.fasta -db /data/ncbi-blast-2.7.1+/db/UP000005640_9606.fasta -out mysearch.html –html
  - firefox mysearch.html to view
  - What is the E-value for the top hit? (record in README)

What is the unknown sequence?

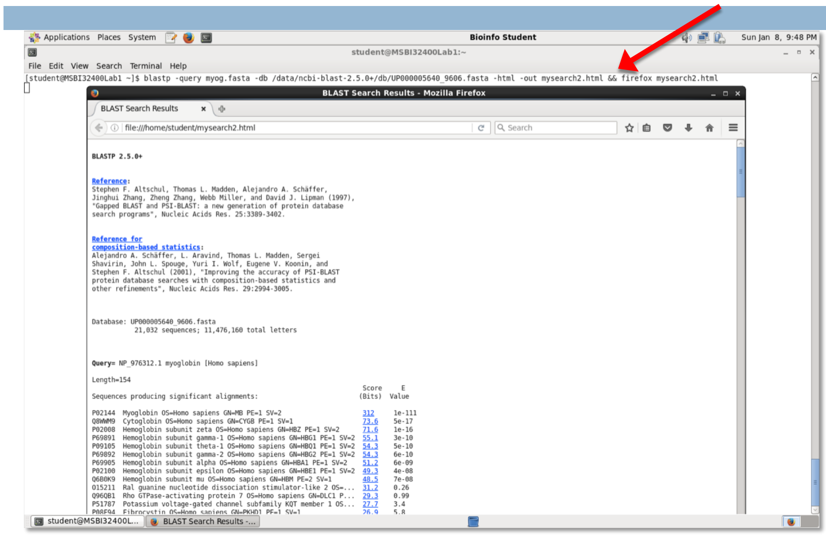
15

- □ Assuming it’s human, run blastp with unknown.fasta as your query against the UniProt database
- □ Record the top hit in the README you send to Jason

BLAST myog.fasta

16

Use ‘&&’ to string together commands

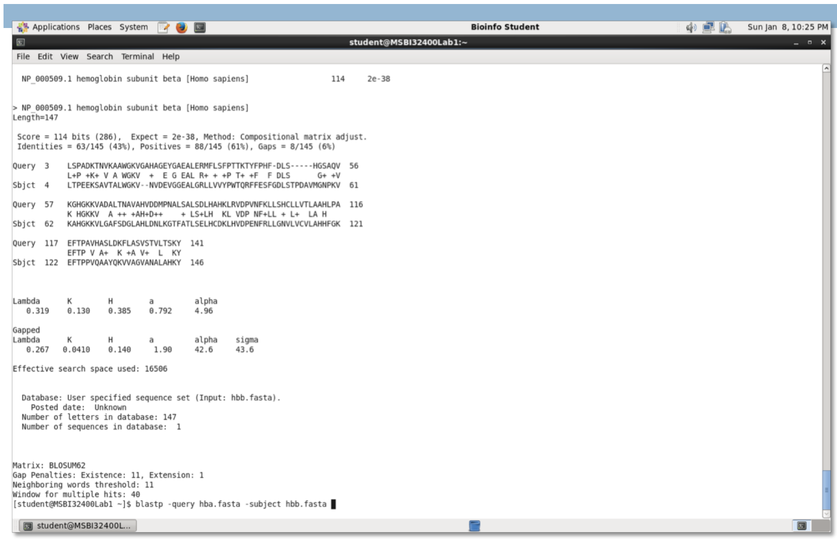

Applications Places System
Bioinfo Student
Sun Jan 8, 9:48 PM
student@MSBI32400Lab1:~
File Edit View Search Terminal Help
BLAST Search Results - Mozilla Firefox
BLAST Search Results
file:///home/student/mysearch2.html
Search
student@MSBI32400L...
BLAST Search Results -...

```
[student@MSBI32400Lab1 ~]$ blastp -query myog.fasta -db /data/ncbi-blast-2.5.0+/db/UP000005640_9606.fasta -html -out mysearch2.html && firefox mysearch2.html
```

```
BLASTP 2.5.0+


Reference:
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.


Reference for
composition-based statistics:
Alejandro A. Schäffer, L. Aravind, Thomas L. Madden, Sergei
Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and
Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST
protein database searches with composition-based statistics and
other refinements", Nucleic Acids Res. 29:2994-3005.



Database: UP000005640_9606.fasta
           21,032 sequences; 11,476,160 total letters



Query= NP_976312.1 myoglobin [Homo sapiens]

Length=154
                                                                     Score     E
Sequences producing significant alignments:                         (Bits)  Value

P02144  Myoglobin OS=Homo sapiens GN=MB PE=1 SV=2                    312     1e-111
Q8WWM9  Cytoglobin OS=Homo sapiens GN=CYGB PE=1 SV=1                 73.6    5e-17
P02008  Hemoglobin subunit zeta OS=Homo sapiens GN=HBZ PE=1 SV=2     71.6    1e-16
P69891  Hemoglobin subunit gamma-1 OS=Homo sapiens GN=HBG1 PE=1 SV=2 55.1    3e-10
P09105  Hemoglobin subunit theta-1 OS=Homo sapiens GN=HBQ1 PE=1 SV=2 54.3    5e-10
P69892  Hemoglobin subunit gamma-2 OS=Homo sapiens GN=HBG2 PE=1 SV=2 54.3    6e-10
P69905  Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2   51.2    6e-09
P02100  Hemoglobin subunit epsilon OS=Homo sapiens GN=HBE1 PE=1 SV=2 49.3    4e-08
Q6B0K9  Hemoglobin subunit mu OS=Homo sapiens GN=HBM PE=2 SV=1       48.5    7e-08
O15211  Ral guanine nucleotide dissociation stimulator-like 2 OS=... 31.2    0.26
Q96QB1  Rho GTPase-activating protein 7 OS=Homo sapiens GN=DLC1 P... 29.3    0.99
P51787  Potassium voltage-gated channel subfamily KQT member 1 OS... 27.7    3.4
P08F94  Fibrocystin OS=Homo sapiens GN=PKHD1 PE=1 SV=1               26.9    5.8
```

BLAST two files

17



```
 NP_000509.1 hemoglobin subunit beta [Homo sapiens]                 114     2e-38


> NP_000509.1 hemoglobin subunit beta [Homo sapiens]
Length=147

 Score = 114 bits (286),  Expect = 2e-38, Method: Compositional matrix adjust.
 Identities = 63/145 (43%), Positives = 88/145 (61%), Gaps = 8/145 (6%)

Query  3    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQV  56
            L+P +K+ V A WGKV  +  E G EAL R+ + +P T+ +F  F DLS      G+ +V
Sbjct  4    LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61

Query  57   KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  116
            K HGKKV  A ++ +AH+D++    + LS+LH  KL VDP NF+LL + L+  LA H
Sbjct  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121

Query  117  EFTPAVHASLDKFLASVSTVLTSKY  141
            EFTP V A+  K +A V+  L  KY
Sbjct  122  EFTPPVQAAYQKVVAGVANALAHKY  146



Lambda      K        H        a         alpha
   0.319    0.130    0.385    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 16506


  Database: User specified sequence set (Input: hbb.fasta).
    Posted date:  Unknown
  Number of letters in database: 147
  Number of sequences in database:  1



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 11
Window for multiple hits: 40
[student@MSBI32400Lab1 ~]$ blastp -query hba.fasta -subject hbb.fasta
```

MSBI 32400 Lab 4 1/30/2019

BLAST shows mismatches

18

- □ Download unknown2.fasta from Files/Lab 4 on Canvas
- □ Run blastp with this file as query
- □ Look at the first hit in the HSP alignment. Did it match all amino acids? What is the base in the query and what is the base in the reference sequence? Ignoring the first M in the sequence, what position is any potential mismatch? (Hint- You've seen it before!) Record what you found in the README file you send to Jason.

MSBI 32400 Lab 4 1/30/2019

Repeat some of the searches on-line

- Upload unknown.fasta to https://blast.ncbi.nlm.nih.gov/Blast.cgi
- Run 16S ribosomal search against NR_119358.1
- Download and BLAST unknown_fragment.fasta
  - Look at the file by **head** to see if it's protein/nucleotide and use appropriate search engine
  - Assume it's human (9606)
  - What gene/protein is this from? What is the E-value?

Homework

- Submit to Canvas or e-mail Jason (jasone@uchicago.edu) the README with the file information requested above before next class (2/13) with "**Lab #4**" in the subject line