An Exploration of Command-Line BLAST (Basic Local Alignment Sequence Tool)

Using BLAST to Search Watermelon Sequence Data: Hands-On Exercise

Searching sequence data for similarities is one of the most common tasks in bioinformatics. In fact, identifying and quantifying sequence similarity (either nucleotide or amino acid) underlies *many* other types of sequence analyses. Sequence identity (as a percent) is straightforward, but other metrics of sequence similarity are not trivial to calculate and slight differences in parameter choices can change the score for matches (or "hits") between your sequence of interest ("query") and those sequences you are searching against in the database ("subject[s]"). In addition, the "best" hit might change over time as researchers add new sequences to public databases.

BLAST is a sequence-searching algorithm that finds statistically significant similarities between sequences by evaluating pairwise alignments. These alignments can be global (across the entire length) or local (the best subsequence). There are many flavors of BLAST and a variety of ways to execute this program—today we will perform a series of searches to acquaint ourselves with the nuances of this powerful and ubiquitous tool. Learning to perform command-line blast is especially useful for researchers, as it can be customized and is faster than using the web interface.

Overview: This exercise does not ask or address a specific scientific question. However, you will learn to manipulate real data, and will orient yourselves to the tools and possibilities of command-line BLAST. You will be performing command-line BLASTN searches using nucleotide sequence from a watermelon gene (NAD4L) against six *different* databases of nucleotide sequences using default (Highly similar) parameters. Then you will re-run the same 6 searches, altering the parameters (Somewhat similar), so you can compare the outputs. Lastly, you will use the whole watermelon mtDNA genome to search against all watermelon mtDNA proteins.

BLAST should already be downloaded on your computer. If however it is not, download **ncbi-blast** to your computer. Check for ncbi-blast and find where it is on your computer. Because command-line **ncbi-blast** does not have a GUI, no icon will appear on your Desktop, but you can search for where it has been downloaded on your computer:

If you are on a **Mac**, it should download to: **usr/local/ncbi/blast/bin/**
(To search this manually on your computer: **Go > Go to Folder… > type "/usr"**)

`If you are on a` **`PC,`** `it should download to:` **`C:\Program Files\NCBI\`**
`(To search this manually on your computer:` **`Computer > Program Files > NCBI)`**

IMPORTANT!!!: Throughout the module, if instructions differ between Mac and PC, the PC versions will appear `in courier font`. Also, the slashes are often reversed in Mac versus PC paths.

Navigate to the link on the web page for this module called '**watermelon_files'** and click on it. You will be directed to a folder of files on Google drive. Go **File→ Download→ Save File**. Once this file has saved to your computer, move it to your Desktop and unzip it. It should contain the following folders and files (make sure they are all there):

watermelon_nt
watermelon_aa
Plant_mt_genomes
example_output_files
watermelonmt.fsa
watermelonmt.gff

Within these files, you have several key pieces of data that you will need for this exercise, including the nucleotide sequence of the nad4L gene, the nucleotide sequence of all watermelon (Citrullus) mtDNA genes (**watermelon_nt**), a copy of the whole watermelon mtDNA genome (**watermelonmt.fsa**), and a file of many plant whole mtDNA genomes (**Plant_mt_genomes**). You are also going to need to record data, so download the **watermelon_blast_statistics** Google spreadsheet from the web page to your Desktop folder as well (download in **.xlsx** format).

Part I: BLASTN of the nad4L Gene from Watermelon

Mitochondria are organelles found in almost all eukaryotes, and are thought to have resulted from an early endosymbiotic event whereby one prokaryote engulfed another prokaryote. Thereby leading to the evolution of the respiratory organelle (the mitochondrion) in the ancestral eukaryote. As a result of the endosymbiotic formation of the mitochondrion, it contains its own genome (the mtDNA). In animals, mtDNA is highly streamlined—in most species it ranges from 13,000 - 16,000 bp long, with few gene duplication events, little intergenic material, and almost no introns. In plants, however, mtDNA sizes can be big (and range widely). For example, the primrose mtDNA genome is 195 kb long, while the cantaloupe mtDNA

genome is 2400 kb long. Many scientists are interested in what factors drive the diversity of mtDNA genome size in plants, and/or understanding how plant mtDNA produce energy under different conditions in order to optimize energy production in plants. (You can read more about the various theories explaining the origin of mitochondria here: http://www.nature.com/scitable/topicpage/the-origin-of-mitochondria-14232356.)
The gene for the mitochondrial protein NADH dehydrogenase subunit 4L (nad4L) is a key component of a respiratory-chain enzyme that catalyzes electron transfer from NADH to ubiquinone. It is part of a larger group of NAD genes, which is one of several gene families encoded in the mtDNA that are related to energy production—the main function of mitochondria in plant growth and development.
You will perform a **BLASTN** search of the nad4L *nucleotide* sequence against the following databases that you will need to make on your computer using the files you downloaded:

Citrullus nad4L gene
Citrullus nad genes
Citrullus mt genes
Citrullus mt genome
Various Plant mt genomes
GenBank (nr)*

*This is the non-redundant database of all nucleotide genomic sequences that are publicly available. Rather than download that to our computers, this search we will conduct online.

IMPORTANT: DO *NOT* COPY/PASTE THE FOLLOWING COMMANDS, TYPE THEM!!!!!!!

REMEMBER!!!!!!!!!!!!! Mac instructions are in Times font and `PC instructions are in courier`. Things that you will need to change based on your individual computer are highlighted in grey.

To begin:
Mac: Open the **Terminal** window: **Finder > Go > Utilities > Terminal**
`PC: Start button > Run > cmd`
When you are working, there are two different prompt symbols
Mac: $
`PC: >`

change to the directory where you installed BLAST to make sure it is there
cd /usr/local/ncbi/blast/bin
PC users first need to move up one directory
`cd.. to go to C:\>`
PC users then can change to the directory where you installed BLAST to make sure it is there
`cd \Program Files\NCBI\`

list contents of directory to make sure it is there
ls
`dir`
PC users need to move up two directories
`cd.. (NOTE:you may have to perform this command more than once to move back to C:\>)`

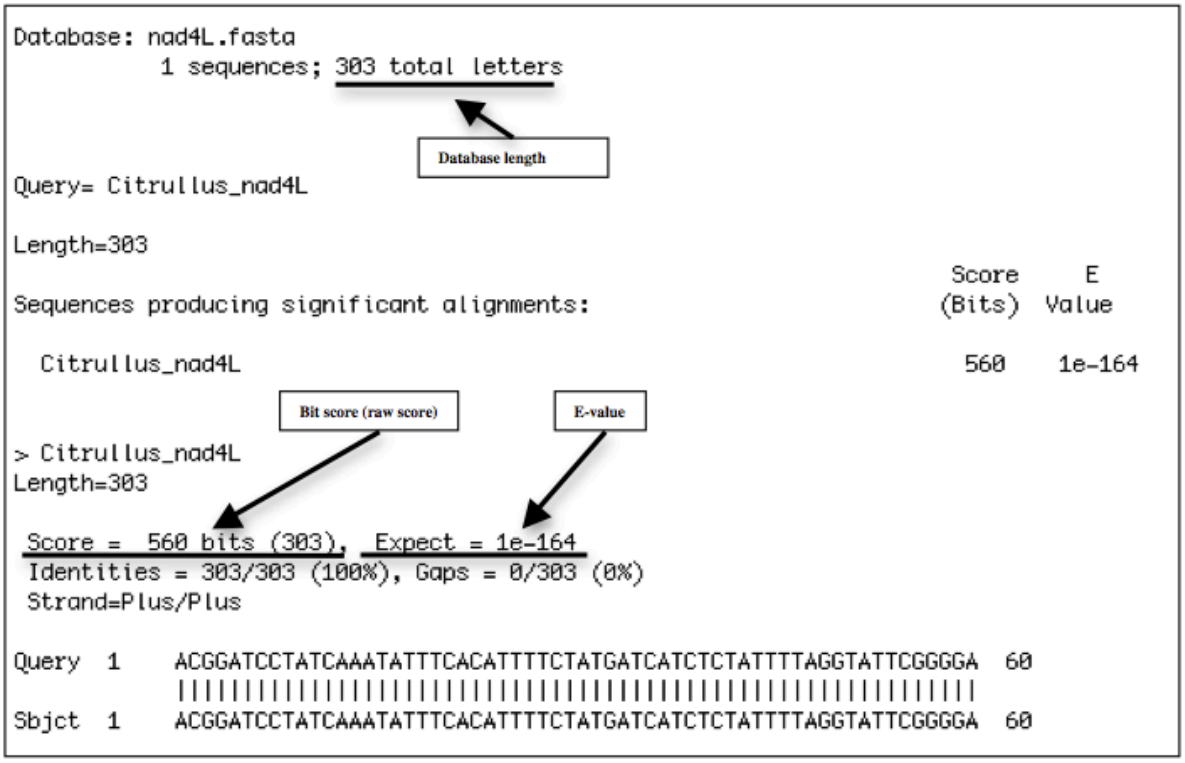
change directory to move to path where your data files are located
cd /Users/your computer's name/Desktop/watermelon_files/watermelon_nt
`cd \Users\your computer's name\Desktop\watermelon_files\watermelon_nt`

list contents of directory to make sure they are there
ls
`dir`

make a database from **nad4L.fasta**
note: making a blast database will also result in the production of associated other files (.nhr, .nin, .nsq)
makeblastdb –in nad4L.fasta –dbtype nucl
[same]

NOTE: if you are not IN the folder with the executable (in this case, makeblastdb) or with the input file (in this case nad4L.fasta), you need to add a path to the file name so that the computer can find it (e.g., /Users/schaack/Desktop/watermelon_files/watermelon_nt/nad4L.fasta). Similarly, if you are IN the folder, and therefore you have an unnecessary path, sometimes it won't work, and you will need to remove the path.
Now that you have your databases made, you can run a BLASTN. To run a **Highly similar** BLASTN search, you do not need to change any parameters (by default, they are: **-word_size 28 -reward 1 -penalty -2 -gapopen 0 -gapextend 2.5**).

run BLASTN to compare nad4L nucleotide sequence to the nad4L database that you have made using default ("highly similar") parameters
blastn –query nad4L.fasta –db nad4L.fasta
[same]
Record the **Database size (nt), Top hit: Raw score, bitscore, and E-value** in your spreadsheet**.**
The top hit has the lowest e-value (remember: 7e-164 is lower than 3e-16, and therefore there is a lower probability of getting this hit by chance, and it is a better hit!) Hits will be ordered from best to worst starting at the top of the BLAST output**.** To find specific values (raw score, bitscore, etc.), refer to the following sample output:



re-run BLASTN to compare nad4L nucleotide sequence to this database, but CHANGE the parameters to run a **Somewhat similar** BLASTN search. Use these (Somewhat similar) parameter values (below) as an example to see how changing values changes your output:
blastn –query nad4L.fasta –db nad4L.fasta -word_size 11 -reward 2 -penalty -3 -gapopen 5 -gapextend 2
[same]
Record the **Database size (nt), Top hit: Raw score, bitscore, and E-value** in worksheet**.**

Now, you will make your other databases. In some cases, you will need to concatenate files *prior to* making your database and running your BLASTs.

change directory and path to the watermelon_files folder (PC users will just move up one directory using **cd..**)
cd /Users/username/Desktop/watermelon_files
cd.. to go to \Users\Username\Desktop\watermelon_files
You are going to need to repeat the commands **cat**, **makeblastdb** and **blastn -query** for your other databases (**A: *Citrullus* nad genes, B: *Citrullus* mt genes, C: *Citrullus* mt genome, D: Plant mt genomes)**, until all 4 are done and the data are recorded in your spreadsheet. REMEMBER: for the following BLAST commands you need to perform BLASTN searches using both **Highly Similar** and **Somewhat Similar** parameters (and continue to fill out the excel spreadsheet).

for example, for A (to BLASTN of your nad4L gene against all *Citrullus* nad genes), use these commands:

cat watermelon_nt/nad*.fasta > watermelonnadgenes.fasta
type watermelon_nt\nad*.fasta>watermelonnadgenes.fasta

makeblastdb –in watermelonnadgenes.fasta –dbtype nucl
[same]

blastn –query watermelon_nt/nad4L.fasta –db watermelonnadgenes.fasta
[same, except slashes]

to do B, use the commands below (to BLASTN of your nad4L gene against all *Citrullus* mt genes)

cat watermelon_nt/*.fasta > watermelonmtgenes.fasta
makeblastdb –in watermelonmtgenes.fasta –dbtype nucl
blastn –query watermelon_nt/nad4L.fasta –db watermelonmtgenes.fasta
[same, except 'type' instead of 'cat' and slashes]

for C, the file already contains multiple sequences, so you don't need to con**cat**enate it first. Use the commands below (to BLASTN of your nad4L gene against all *Citrullus* mt genome)

makeblastdb –in watermelonmt.fsa –dbtype nucl
blastn –query watermelon_nt/nad4L.fasta –db watermelonmt.fsa
[same, except slashes]

to do D, use the commands below
this blastn output file is large, so you will generate an output text file (instructions below) named '**results.txt'**, this file will be saved automatically to your **watermelon_files** folder

cat Plant_mt_genomes/*.fasta > plantmtgenomes.fasta
makeblastdb –in plantmtgenomes.fasta –dbtype nucl
blastn –query watermelon_nt/nad4L.fasta –db plantmtgenomes.fasta –out results.txt

this next BLASTN search uses **GenBank (nr)** as the database, which would take a long time to download and takes a long time to run remotely. Thus, instead of running command-line blast, let's use the BLAST webpage (http://blast.ncbi.nlm.nih.gov/Blast.cgi) for this one. Navigate there and paste your nad4L sequence, click on the option to search for "Highly similar sequences" and record the data in your spreadsheet. Because the alternative parameters we used are actually identical to the parameters for searching for "Somewhat similar sequences", you can simply click on this option and get the output you need to fill in the last cell in your table. ***NOTE**: to find the raw score and bit score in this online output scroll down to the alignment of the best hit; the size of the GenBank (nr) database is not displayed in the online output, do some quick research to find out how large the database is.*

#if you wanted to run this last search using the command-line, this is what you would type:

blastn –query watermelon_nt/nad4L.fasta –remote –db nr –num_descriptions 10
[same]

Once you have compiled all of your results in the spreadsheet, answer the following questions. It may be helpful to click on the column headers on the output table on web version of BLAST (which are linked to explanation of the various metrics) or it may be helpful to discuss the results of your various BLASTs with your partner.

Consider the different databases you searched against.
Do the raw scores for your best hits change depending on the database you used? Why or why not?
Does the E-value associated with your best hit change depending on the database you search against? Why or why not?
What was the database that resulted in the BEST (i.e., lowest) E-value? _____
Did your raw scores change depending on the parameters you used to search your database?_____
Did they get better or worse in your second set of searches?_____
Did your bit scores do the same (i.e., also get better or also get worse)?_____
Did your E-values get better or worse in the second set of searches?_____
What does "word size" mean with respect to a BLAST search?

Other than word size, what was the biggest parameter change you made between your 1st set and 2nd set of searches? Why do you think this might be important for changing the scores and E-values of your BLASTN?

On your own….

Part II: BLASTX of mt Genome against mt Proteins for Watermelon

Imagine you sequenced the whole mtDNA genome for a strain of watermelon that produced especially large melons and you were interested in studying a few of the most important ATP-related genes more carefully. You might want to use BLAST to sift through your large dataset (the watermelon mtDNA genome is over 300,000 bp long!) to focus on the protein-coding genes (which are coded for by only < 35,000 bp subset!)

Perform a **BLASTX** search of the whole watermelon mt genome against a *Citrullus* (watermelon) mtDNA protein database (containing amino acid sequence data).
to do this, you need to concatenate the amino acid files, make a database, and perform a command-line BLASTX using the following commands:
this blastx output file is large, so you will again generate an output text file (instructions below) named '**protalign.txt'**, this file will be saved automatically to your **watermelon_files** folder

cat watermelon_aa/*.fasta > watermelonaa.fasta
makeblastdb –in watermelonaa.fasta –dbtype prot
blastx –query watermelonmt.fsa –db watermelonaa.fasta –out protalign.txt

Look carefully at the output. Note that the **Query** line is in terms of nucleotide numbers in the mt genome sequence, while the Subject (**Sbjct**) line is in terms of the amino acid numbers in the amino acid database sequence. Also, the beginning and ending nucleotide numbers can be ordered for all exons to find their order along the chromosome. First, only consider hits that have reasonably high **Identities** (~95-100%). Note that the reading frame (**Frame = +1, +2, +3, -1, -2, or -3**) may change from one exon to the next. *(Why?)*
Answer the following questions on the spreadsheet.
*How many exons are there in **nad5**? (HINT: you can use Command+F or `Control+F` to search)*
What comes between exons in the nucleotide sequence that does not affect the amino acid sequence?
*How many exons are there in **cox1**?*
Reading frame may change from one exon to another, why is this?
How do the BLASTX results compare with the BLASTN results?
Why might some sequences be identical according to BLASTX but differ according to BLASTN?

when you are all done

exit
`[same]`

Further Reading:

*mtDNA: Alverson, A. J. X. Wei, D. W. Rice, D. B. Stern, K. Barry, and J. D. Palmer 2010. Insights into the evolution of mitochondrial genome size from complete sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae). Mol. Biol. Evol. **27**: 1436-1448. Mackenzie, S. A. (2010) The Influence of Mitochondrial Genetics on Crop Breeding Strategies, in Plant Breeding Reviews, Volume 25 (ed J. Janick), John Wiley & Sons, Inc., Oxford, UK. BLAST algorithm: Zhang, Z. S. Schwartz, L. Wagner, and W. Miller 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7:203-214.*