BLAST tutorial

Experiential Data science for Undergraduate Cross-disciplinary Education

Dr. Kim Dill-McFarland, U. of British Columbia

Contents

BLAST tutorial

Some of you may be familiar with online BLAST. This online tool is great for BLAST-ing a couple of sequences but what if you have hundreds, thousands, or even more? The best (and quickest) way to accomplish this is using command line BLAST tools.

Learning objectives

- Practice Unix command line
- Complete BLAST of a large amplicon sequencing data set via command line

Setup

If you would like to follow along, please install the NCBI Basic Local Alignment Search Tool (BLAST) as well as download a pre-formatted 16S database and representative data sequences.

Download BLAST

1. Download the BLAST software appropriate for you operating system at ftp://ftp.ncbi.nih.gov/blast/executables/blast+/LATEST.
   - Use the `.exe` file for Windows or the `.dmg` file for Mac.
   - Do not use `tar.gz` files!
2. Open the installer and follow prompts.
3. Check BLAST on your computer by opening your terminal and inputting `blastn`[Enter].
   - If installed correctly, you should see the output: `BLAST query/options error: Either a BLAST database or subject sequence(s) must be specified. Please refer to the BLAST+ user manual.`
   - You may need to restart your computer to fully install.
   - See the 'Download instructions for Unix terminal' if you do not know how to access the terminal on your computer

Download database

When you use BLAST online, it automatically uses the full NCBI sequence database. Since the data we will be using in this tutorial contain only 16S sequences, we do not need this entire database. Instead, we will comparison the data to a curated subset of NCBI containing only 16S reference sequences.

1. Download the 16S microbial database `16SMicrobial.tar.gz` and unzip.

For more information on the databases available for BLAST, see their download page and documentation.

Download data

The sequence data that we will use in this tutorial are 16S amplicon profiles of microbial communities in Saanich Inlet in August 2012. More information on this system can be found here. Specifically, 16S amplicon sequences were processed using mothur to yield representative sequences of each 97% (approximately species-level) operational taxonomic unit (OTU). Details on the data preparation can be found here.

1. Download the representative data sequences and unzip.
    - We will use both `Saanich_OTU_rep.fasta` and `Saanich_OTU_rep_subset.fasta`
    - For ease, you should place the database and sequence files in a single directory on your Desktop like below

```
ls Desktop/BLAST_data
```

```
## 16SMicrobial.nhr
## 16SMicrobial.nin
## 16SMicrobial.nnd
## 16SMicrobial.nni
## 16SMicrobial.nog
## 16SMicrobial.nsd
## 16SMicrobial.nsi
## 16SMicrobial.nsq
## 16SMicrobial.tar.gz
## Saanich_OTU_rep.fasta
## Saanich_OTU_rep_subset.fasta
## taxdb.btd
## taxdb.bti
```

Data format

These data are in the standard `.fasta` format of

Otu#
Sequence of ATCG

```
head Desktop/BLAST_data/Saanich_OTU_rep.fasta
```

```
## >Otu0001
## CTGGGGGGTGCCAGCCGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGTAGGTGGTTTGTTAAGTTAGATGTGA
## >Otu0002
## CTGGGGGGTGCCAGCCGCCGCGGTAATACGGAGGGCGCAAGCGTTACTCGGAATCACTGGGCGTAAAGAGCGTGTAGGCGGGTATTTAAGTTGGAAGTGA
## >Otu0003
## CTGGGGGGTGCCAGCCGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGTAGGCGGTTTATTAAGTCAGATGTGA
## >Otu0004
## TGAGGGGTGCCAGCCGCCGCGGTAATACTGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTAGATCAGTTGGATGTGAA
## >Otu0005
## CTGGGGGGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTGTTCGGAATTACTGGGCGTAAAGGGCGCGCAGGCGGAATAGCAAGTCGGAGGTGA
```

The output in your terminal may look like more than 2 lines, because it wraps the sequence text (ATCG) around to fill multiple lines. If you zoom way out in your terminal, you can see the true '2 line per OTU' fasta format.

BLAST via command line

The full data set `Saanich_OTU_rep.fasta` contains 4,368 OTUs (*e.g.* species), which can take quite some time to run through BLAST. To start, we will run BLAST on just the first 10 OTUs, which are contained in `Saanich_OTU_rep_subset.fasta`.

blastn

There are several BLAST functions; we will use `blastn` for nucleotide BLAST since our data are nucleotide sequences. Similar to other command line functions, you provide the function name and then various parameters. `blastn` requires a:

- query - sequences we want to know the taxonomy of (`.fasta`)
- db - database of sequences and taxonomies to compare the query to (`16SMicrobial` files)

We will also add the following parameters to customize BLAST for these data.

- out - what to name the output (results) file
- num_descriptions - how many BLAST hit descriptions to include in the output
- num_alignments - how many BLAST hit alignments to include in the outputs

So, to BLAST the data, we use

```
blastn -query Desktop/BLAST_data/Saanich_OTU_rep_subset.fasta \
-db Desktop/BLAST_data/16SMicrobial \
-out Desktop/BLAST_data/blast_results.txt \
-num_descriptions 3 -num_alignments 3
```

Note that as our command gets longer, we can split it across multiple lines using `\`.

BLAST results

We can look at the first 50 lines of the output file with `head -50`.

```
head -50 Desktop/BLAST_data/blast_results.txt
```

```
## BLASTN 2.7.1+
##
##
## Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb
## Miller (2000), "A greedy algorithm for aligning DNA sequences", J
## Comput Biol 2000; 7(1-2):203-14.
##
##
##
## Database: 16S Microbial Sequences
##            20,767 sequences; 30,197,079 total letters
##
##
##
## Query= Otu0001
```

```
##
## Length=302
##                                                                      Score     E
## Sequences producing significant alignments:                         (Bits)  Value
##
## NR_025016.1  Methylobacter psychrophilus strain Z-0021 16S riboso...  414     1e-115
## NR_126264.1  Bacterioplanes sanyensis strain GYP-2 16S ribosomal ...  409     5e-114
## NR_112914.1  Arenicella xantha strain KMM 3895 16S ribosomal RNA,...  409     5e-114
##
##
## >NR_025016.1 Methylobacter psychrophilus strain Z-0021 16S ribosomal RNA,
## partial sequence
## Length=1527
##
##  Score = 414 bits (224),  Expect = 1e-115
##  Identities = 270/293 (92%), Gaps = 0/293 (0%)
##  Strand=Plus/Plus
##
## Query  8    GTGCCAGCCGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAA  67
##             |||||||| ||||||||||||||||||||| ||||||||||||||||||||||||||||
## Sbjct  499  GTGCCAGCAGCCGCGGTAATACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAA  558
##
## Query  68   GCGTGCGTAGGTGGTTTGTTAAGTTAGATGTGAAAGCCCTGGGCTCAACCTAGGAACTGC  127
##             ||||||||||||||||| ||||||| ||||||||||||||||||||||| ||||||||
## Sbjct  559  GCGTGCGTAGGTGGTTCGTTAAGTCAGATGTGAAAGCCCTGGGCTCAACCTGGGAACTGC  618
##
## Query  128  ATTTAAAACTGGCAAACTAGAGTATAGGAGAGGAAAGTGGAATTTCAGGTGTAGCGGTGA  187
##             |||| ||||||||  |||||||| ||| |||||  ||||||||||||||||||||||||
## Sbjct  619  ATTTGAAACTGGCGGACTAGAGTTTAGTAGAGGGGAGTGGAATTTCAGGTGTAGCGGTGA  678
##
## Query  188  AATGCGTAGATATCTGAAGGAACATCAATGGCGAAGGCAGCTTTCTGGACTAATACTGAC  247
##             ||||||||||||||| |||||||| || ||||||||||  ||  ||||||||| ||||||
## Sbjct  679  AATGCGTAGATATCTAAAGGAACACCAGTGGCGAAGGCGACTCCCTGGACTAAAACTGAC  738
##
## Query  248  ACTGAGGTACGAAAGCGTGGGTAGCAAACAGGATTAGAAACCCCGGTAGTCCA  300
```

These first lines show you just the top BLAST hits for OTU1. Let's go through the parts.

1. BLASTN version and citation.
2. Database name and how many sequences it contains.
3. First sequence query name and length.
4. Top **BLAST hits**. These are the sequences in the database which our query is significantly similar to. For each hit, there is a(n):
   - NR ID which is unique for each sequence in the database
   - name of the organism from which the database sequence was obtained
   - BLAST score
   - BLAST E value
5. The alignment showing the direct base pair comparison of the query and database sequence.
