Introduction to BLAST and protein homology searches

Frank Aylward

Abstract

Citation: Frank Aylward Introduction to BLAST and protein homology searches. **protocols.io**
dx.doi.org/10.17504/protocols.io.piedkbe
Published: 17 Apr 2018

Protocol

Make sure the right tools are installed first

Step 1.

We'll be using BLASTP

Download the data

Step 2.

wget -O prochlorococcus_phage_PSSM2.faa ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/859/585/GCF_000859585.1_ViralProj15135/GCF_000859585.1_ViralProj15135_protein.faa.gz

wget -O prochlorococcus_phage_PSSM3.faa ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/907/775/GCF_000907775.1_ViralProj209210/GCF_000907775.1_ViralProj209210_protein.faa.gz

Get some basic stats about the files and what's inside

Step 3.

```
grep "^>" prochlorococcus_phage_PSSM2.faa | wc
# or
grep "^>" prochlorococcus_phage_PSSM3.faa | wc
```

Choose one protein file to be the reference, and make the appropriate BLAST databases

Step 4.

```
makeblastdb -in prochlorococcus_phage_PSSM2.faa -dbtype prot
```

Now use the other protein file as the query, and the reference file as the database

Step 5.

```
blastp -query prochlorococcus_phage_PSSM3.faa -db prochlorococcus_phage_PSSM2.faa | head -n 100
```

Vary the parameters a bit and see what the different results look like

Step 6.

Do the same thing, but use a tab-delimited output and only look at the top 10 hits.
blastp -query prochlorococcus_phage_PSSM3.faa -db prochlorococcus_phage_PSSM2.faa -outfmt 6 | head

Now we can play around with the output parameters to ensure that only "good hits" are reported, and only the best hit for each query protein is given.
blastp -query prochlorococcus_phage_PSSM3.faa -db prochlorococcus_phage_PSSM2.faa -outfmt 6 -max_target_seqs 1 -evalue 0.00001 -max_hsps 1 -qcov_hsp_perc 50 | head

breakdown of the flags:
-query: this is the input file, so the file with all of the protein sequences that we want to search
-db: this is the database, so the file we just indexed with the makeblastdb command above. Note that makeblastdb creates multiple reference files and that only the root name needs to be given here (so if the database was called refdb, then refdb would be given here even though the index files are called refdb.pin, refdb.phr, etc.)
-max_target_seqs: This flat specifies that we only want the best hit for each query protein. Otherwise all hits are provided.
-outfmt: This specifies that we want the tab-delimited output format rather than the full alignment output. If you forget what the columns are you can use -outfmt 7.
-evalue: This indicates that we want to exclude all hits with evalues above this threshold. A good value is about 0.00001, or 1e-5.
max_hsps: HSPs are "high-scoring segment pairs". A query protein can make several separate alignments to a single reference, so this tells the program we want only the best-scoring alignment.
-qcov_hsp_perc: This is the "query coverage high-scoring sequence pair percent", or the percent of the query protein that has to form an alignment against the reference to be retained. Higher values prevent spurious alignments of only a short portion of the query to a reference.

Now let's calculate the one-way amino acid identity (AAI) of the two genomes

Step 7.

What is the average amino acid identify of PSSM2 and PSSM3, using PSSM2 as the query?
To help with this we can install a package for simple math called "datamash"
"sudo apt install datamash" OR "brew install datamash"
Datamash will allow for quick calculation of averages straight from the command line. Once this package is installed you can run:
blastp -query prochlorococcus_phage_PSSM3.faa -db prochlorococcus_phage_PSSM2.faa -outfmt 6 -max_target_seqs 1 -evalue 0.00001 -max_hsps 1 -qcov_hsp_perc 50 | datamash mean 3
And the output should be a single number, which is the average of all of the % identity scores from the blast output