Homology and Clustering

Stephen A. Smith[1] and James B. Pease[2]

[1]*Department of Ecology and Evolutionary Biology, University of Michigan;*
[2]*Department of Biology, Wake Forest University*

EMBO Computational Molecular Evolution 2018

1. How to use this tutorial

1.1. In the tutorial below, individual command lines are present in boxes using `monospaced font`.

1.2. Comments and commands are numbered to help with asking questions.

1.3. Lines in the command boxes ending in "\" indicate a single command syntax is continued the next line of text.

1.4. While you *could* copy-paste the command lines, I **strongly** encourage you to practice typing them in manually, as this will help you remember and practice terminal command technique.

2. Getting started

2.1. Download the gzip-compressed TAR archive http://jbpease.github.io/crete2018/exercise_2.tar.gz.

2.2. Decompress the file using this command:

2.2.1.
```
tar -xzf exercise_2.tar.gz
```

2.2.2. This should create a directory called `exercise_2`.

2.3. Change directory to this folder and be sure to run all commands below within this folder.

2.3.1.
```
cd exercise_2
```

3. Phylogeny and BLAST

3.1. First we will look at `cary.matK.fasta` and see what happens when we BLAST a sequence into this set and compare it to a tree we have already built.

3.2. We want to build a database and then BLAST a sequence from that database against itself.

3.3. We are going to search with `Suessenguthiella_scleranthoides_FN825756`.

3.3.1. Look at this in the extracted portion of the tree

3.3.2.
```
figtree cary.matK.rr.sub.tre
```

3.4. Now we will do a blast of this sequence against these sequences and many others.

3.4.1.
```
makeblastdb -in cary.matK.fasta -dbtype nucl
```

3.4.2.
```
blastn -query cary.test -db cary.matK.fasta \
-num_alignments 0 -num_descriptions 10
```

3.4.3. How many need to be listed (change `num_descriptions` to be higher) before we get `Coelanthum_parviflorum_FN825759` in the list?

3.4.4. What does this mean for analyses that interpret the order of the hit?

3.5. Pulling out sequences for alignment with BLAST (Baited BLAST) We can use a file that contains all the sequences from GenBank for the plant group Caryophyllalles for this example called cary.renamed.fasta. We will use another file (some bait) as the database to pull out all of the rbcL sequences from cary.renamed.fasta (maybe backwards from what you would think).

3.5.1.
```
makeblastdb -in cary.atpB.fasta -dbtype nucl
```

3.5.2.
```
blastn -db cary.atpB.fasta -query cary.fasta -evalue 10e-10 -num_threads 2 \
-max_target_seqs 2 -out cary.fasta.raw.blastn \
-outfmt '6 qseqid qlen sseqid slen frames pident nident length mismatch \
gapopen qstart qend sstart send evalue bitscore'
```

Output is cary.fasta.rawblastn in case things don't finish (14 secs for my laptop). We are going to filter the results with some basic filters for length of hit using a provided script

3.5.3.
```
python filter_results_bait.py cary.fasta.rawblastn \
cary.fasta cary.fasta.table > atp.results
```

3.6. We can now conduct a quick multiple sequence alignment and phylogeny

3.6.1.
```
mafft --adjustdirection atp.results > atp.aln
```

3.6.2.
```
python fasta2phylip.py atp.aln atp.phy
```

3.6.3.
```
raxmlHPC-PTHREADS-AVX2 -T 2 -s atp.phy -n ATP -p 12345 -m GTRCAT
```

If we want to combine the bait with the results we can do:

3.6.4.
```
cat cary.atpB.fasta atp.results > cary_and_atp.results
```

3.6.5.
```
mafft --adjustdirection cary_and_atp.results > cary_and_atp.aln
```

3.6.6.
```
python fasta2phylip.py cary_and_atp.aln cary_and_atp.phy
```

3.6.7.
```
raxmlHPC-PTHREADS-AVX2 -T 2 -s cary_and_atp.phy -n CARY_AND_ATP \
-p 12345 -m GTRCAT
```

3.7. Creating all by all alignments with BLAST and SWIPE

3.8. First, we will look at a file ITS.gbget that is one gene region (ITS) and a bunch of species. I know that these can be aligned together, but results can vary given clustering analyses.

3.9. There are a number of ways that we can construct the all-by-all comparisons that are necessary for the MCL analyses. We have already learned some about the tools SWIPE and BLAST and with the help of some scripts we can use those tools to construct the information we need for MCL

3.10. First we will conduct the all-by-all comparison

3.11. We make the database

3.11.1.
```
makeblastdb -in ITS.gbget -parse_seqids -dbtype nucl -out ITS.gbget.db
```

3.12. Now I will conduct the all-by-all using swipe (if you have swipe you can also do this, if not, watch or skip to the blast section below)

3.12.1.
```
swipe -d ITS.gbget.db -i ITS.gbget -a 4 -c 50 -p blastn \
-o ITS.gbget.swipe -m 8 -e 1000000
```

3.13. Look at the results in a plain text editor (they have been provided in the folder in case you do not have skype as the file ITS.gbget.swipe). The results are like Query id, Subject id, s. start, s. end, e-value, bit score. Now we convert the file to be read by mcl which is seq1 seq2 score

3.13.1.
```
python swipe_to_mcl.py ITS.gbget.swipe
```

3.13.2. `ITS.gbget.swipe.filtered` will be created Now we are going to run all-by-all blast on the same database

3.13.3.
```
blastn -db ITS.gbget.db -query ITS.gbget -evalue 0.00001 -num_threads 4 \
-max_target_seqs 100 -out ITS.gbget.rawblastn \
-outfmt '6 qseqid qlen sseqid slen frames pident nident length mismatch \
gapopen qstart qend sstart send evalue bitscore'
```

Check out these results in a plain text editor. How do the evalues differ for the same comparisons? Remember we are comparing Smith-Waterman and the BLAST heuristic. Now we are going convert the file so that it can be run by MCL

3.13.4.
```
awk -F '\t' '{if($6>0.7 && $7>200 && $7/$2>0.2) print $1,$3,$15}' \
ITS.gbget.rawblastn >bl.evalue
```

3.13.5.
```
sed -i 's/ / /g' bl.evalue
```

3.13.6.
```
awk '{if($3==0.0)print $1,$2,180;else print $1,$2,-log($3)/log(10)}' \
bl.evalue >bl.evalue-log
```

3.14. Remember that we want log E-values. These commands change the file into what we need with $\log_{10}$ E-values.

4. Constructing clusters with MCL

4.1. To run a basic MCL analysis, we just run

4.1.1.
```
mcl bl.evalue-log --abc -I 2
```

The -I is the inflation value. If you open the results file in a text editor, you will notice that on each line is a cluster. Then, separated by spaces, there is the name of each sequence in the cluster.Change the inflation value from 1.1 to 10.1 incrementing by 1. Look at the resulting number of clusters. How does increasing the value change the number of clusters? Do you notice how the runtime and iterations change?

4.2. You can extract the sequences so that alignments can be run using the command:

4.2.1.
```
python write_fasta_files_from_clusters_simple.py \
ITS.gbget out.bl.evalue-log.I20 .
```

In addition to adjusting the inflation value, we can, on the fly, filter with evalues. Because we did log10 evalues, we are looking at the values in reverse, so large is better. So we can include only relationships (blast results) with greater than or equal to (gq) 30 like this

4.2.2.
```
mcl all.evalue-log --abc -I 2 -tf 'gq(30)'
```

4.2.3. Try different values. How does this change the results?

4.3. Now let's look at another file `all.unaln`. This includes a number of different gene regions that shouldn't go together.

4.4. Using the procedures outlined above, repeat the all-by-all BLAST followed by mcl.

4.4.1. How do different inflation values change the clusters? (The sequence names are prepended with the gene.)

4.4.2. Are the clusters reflecting this well?

4.4.3. What is the "best" inflation value, if we are going by the sequence name?

4.5. Are there potential problematic sequences?

4.5.1. BLAST them on webBLAST!

5. Extended Exercise 1: Running orthology and paralogy analyses

5.1. We do not have time to do more intensive analyses, in the exercises and so you will need to do them on your own.

5.2. If you want to do orthology and paralogy analyses you can check out this repository and instructions https://bitbucket.org/yangya/phylogenomic_dataset_construction

5.3. The scripts used in the exercises are associated with the publication:
Yang, Y. and S.A. Smith. 2014. "Orthology inference in non-model organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics." *Molecular Biology and Evolution.* doi: 10.1093/molbev/msu245

5.4. If you want something more general, email me and I can send you more general procedures.

5.5. For baited blast analyses or other more general procedures, you can use PHLAWD http://phlawd.net or send me an email and I can send you more general scripts.