MCB 5472 Assignment #5:
RBH Orthologs and PSI-BLAST
February 19, 2014

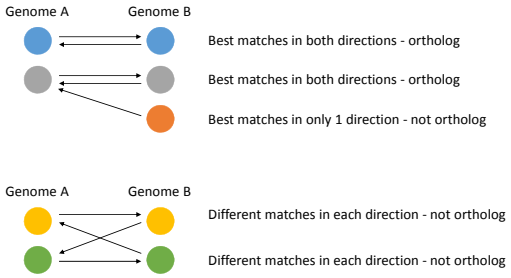
Assignment feedback

- Always check your output files!
- `PRINT` is your friend
- Hash keys and array positions can (and often should) be called directly
  - People seem to want to always loop through – this is unnecessary and seems to be leading people astray
- Protein vs. nucleotide paralogs – why?

This week

1. Count RBH protein orthologs between our complete E.coli genomes
2. Use PSI-BLAST to find divergent homologs of molecular parasites

Recall: RBH orthologs



Discuss: how will this code work?

1. BLAST proteins from each genome vs. themselves
   - What alignment and similarity thresholds to use?
2. Store first set of BLAST hits in a hash
   - Key: query ID Value: best match ID
3. Parse reciprocal BLAST
   - Store in a second hash or evaluate as you go
4. Use first hash to identify RBH hits
   - RBH if: `$first_hash{2nd_BLAST_best_hit} eq 2nd_BLAST_query_id`
5. Tabulate shared and unique proteins
   - Partially do as you go

Submit for part 1:

- Number of shared orthologs, unique sequences in each genome
- Perl scripts
- Short (1-2 sentences) justification of BLAST similarity thresholds

Part 2: Find homologs of molecular parasites

- Make BLAST database of proteins and chromosomes for one or more complete genomes of your choice
  - Can be our model E.coli
- Download the protein sequence for a molecular parasite of your choice

Part 2: BLASTp

- Query reference genomes using BLASTp

```
blastp -db all.faa -query integrase1.fa -evalue 1e-5 -num_threads 2 -out blastp.out -outfmt 6 -comp_based_stats 1
```

`-comp_based_stats`: correction for compositional biases in query; only `0` (off) or `1` can be used with `psiblast` (default is 2)

`-num_threads`: multithreading parameter for computational efficiency

Part 2

`-outfmt 6`: same as `–outfmt 7` but no header lines

i.e., every line is a hit

Can count hits from terminal using `wc –l`

`wc`: word count

`-l` flag: count number of lines

Part 2: PSI-BLAST

- Make pssm of query vs NCBI nr database

```
psiblast -db nr -query integrase1.fa -num_iterations 5 -num_threads 2 -inclusion_ethresh 1e-5 -out blast.out -out_pssm integrase1.pssm -comp_based_stats 1
```

- Search reference genome proteins using that pssm

```
psiblast -db all.faa -in_pssm integrase1.pssm -num_iterations 1 -num_threads 2 -inclusion_ethresh 1e-5 -outfmt 6 -out psiblastp.out -evalue 1e-5 -comp_based_stats 1
```

Part 2: tBLASTn

- Search reference genome itself using `tBLASTn` combined with the pssm

```
tblastn -in_pssm integrase1.pssm -db all.fna -evalue 1e-5 -num_threads 2 -out psitblastn.out -outfmt 6 -comp_based_stats 1
```

Discuss: what are we trying to demonstrate?

Submit for part 2:

- Number of homologs found using each approach
- Short (1-2 sentences) interpretation of these results
- Terminal commands used during each step
  - You should be recording these anyway, just like any experiment you document in a lab book
- Scripts, if you use any