Module 1 Capstone Assignment (Bioinformatics 2019)

When comparing genomes it is necessary to compare “apples to apples”. That is, you want to know that gene A in organism 1 is related by descent to gene A in organism 2 (i.e., the genes are orthologous, or orthologs). The two genes are the same gene, but find themselves in different organisms because the lineages of the two organisms diverged in the past (but share a common ancestry).

Orthology relationships are important to determine for many reasons, but primarily because they give a basis for comparison between species. For example, a candidate gene for autism is found in flies - what is the ortholog in humans and is it associated with autism? Known orthology is also necessary for studying gene, gene family, and genome evolution. If you are to make any comparison between genomes it is first necessary to make sure you are examining the same features across the species, regardless of the biological nature of your question.

So, how is this done? There are many ways, such as through tracing gene evolution using phylogenetics (OrthoDB uses this approach, for example). One of the simplest ways is via BLAST and is called “reciprocal best hit BLAST” or rbhBLAST. The idea is pretty simple, are the genes their own best hits when BLASTed into the comparison genomes? If yes, let’s call them 1:1 orthologs, if not, maybe they are closely related, and even true orthologs, but we are not certain (this gets increasingly confusing when you have gene duplications and gene family expansions).

Your assignment:

- Awhile ago I became interested in a particular, and rather enigmatic, gene family - Osiris. Last year we published a paper where we, among other things, suggested that this family is likely confined to insects. The problem was, though, that there are not many genomes available that are near the origin of insects, but not insects themselves (i.e., good outgroups). But that was then...and genomes are being sequenced all of the time.
- The protein fasta of the D. melanogaster Osiris protein sequences are available for download at: https://cluster.earlham.edu/~crsmith/DmelOsiProt.faa
- There are now several Collembola genomes available. They share a common ancestor with insects that existed 75 my before the first lineages of modern insects began to diverge (~475 mya).
- Choose a collembolan species that has an available proteome (some of these genomes are not fully annotated).
- Write a script that is capable of doing rbhBLAST for the Osiris proteins of the fruit fly (*Drosophila melanogaster*) and your organism of choice.
- The output of your script should be two tables, one with the BLAST results from fly to X, and the other from X to fly.
- And then interpret your results.

- One of the most interesting things about the Osiris genes is that tend to occur in a cluster (or in clusters - some genomes are not well-enough put together to say that they are always in just one cluster). Do your best hits tend to occur near each other on the genome?
- Ultimately, I want you to try and answer the question (and note, I don't know the answer) of whether the Collembola have Osiris genes - is the Osiris cluster really unique to insects?
- To really interpret well and understand what you are doing, it is assumed you are doing some outside reading...among others, I would suggest our paper (HERE).

What to turn in:

1. Your commented script (as a .sh file)
2. The table that was generated by the script.
3. A 1-2 page (maximum) interpretation of your results. *If you use information from other sources, cite them (format does not matter, but you need to have sufficient information and a consistent format - typical formats such as MLA, Harvard, or the ECBioCite, are all sufficient).*

Your “packet” will be due in one week, Feb. 25.