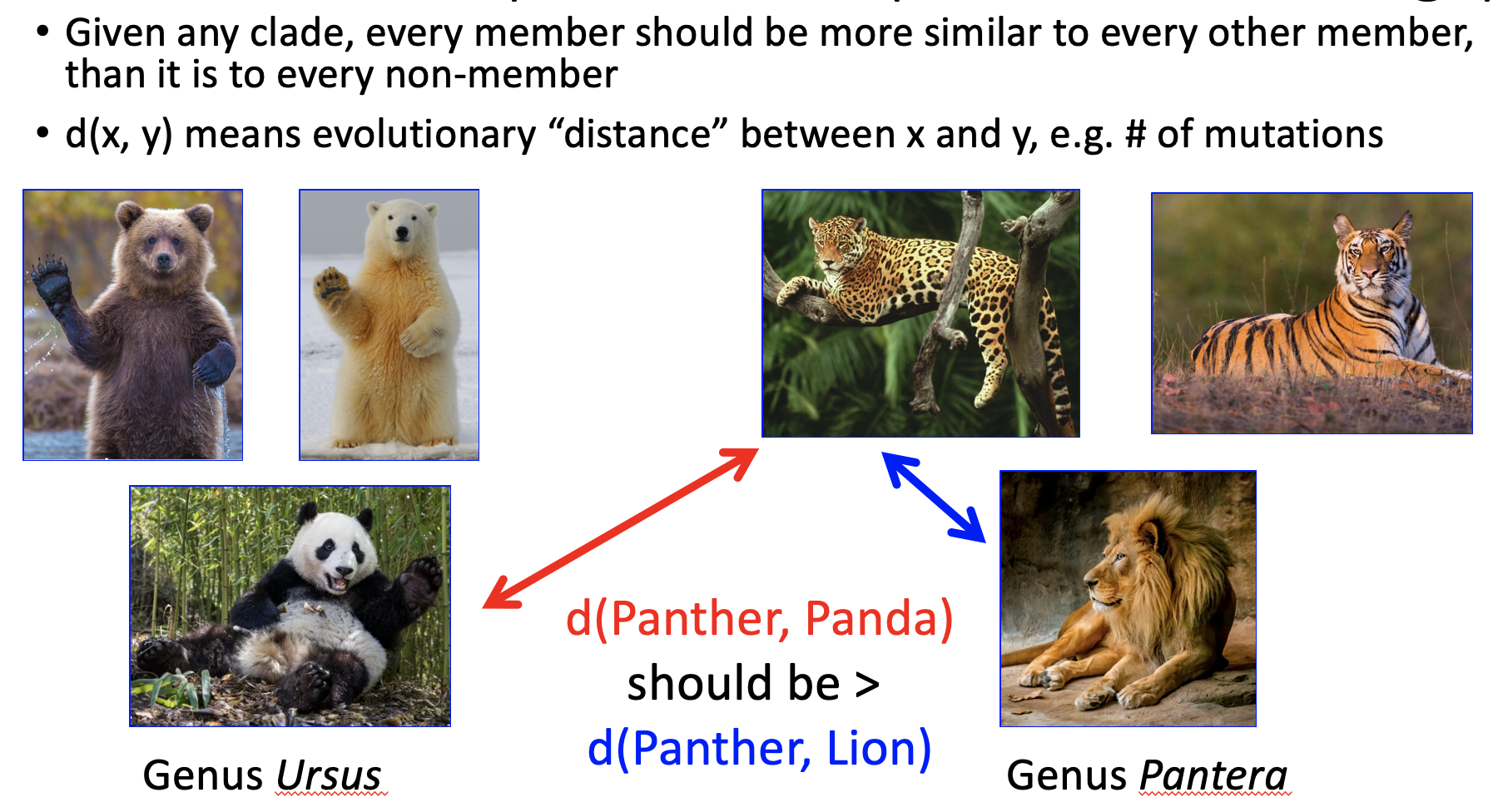
Marine Bioinformatics

Homework 4

Due 11:55 PM Tuesday April 18

Exploring the sponge barcode gap



(Note: these are not sponges, not even the squishy absorbent black-and-white one on the bottom left.)

The barcode gap is the idea that members of any clade should be more similar to one another than they are to anything outside the clade. As you know, the COI (Cytochrome C Oxidase Subunit 1) gene is the de-facto standard barcode for animal life. Unfortunately, it doesn’t always have a clean barcode gap, as we see in sponges.

Suppose you’re a sponge scientist, and you’ve just found a strange-looking sponge that nobody has ever seen before. You sequence its COI gene and blast that against the BOLD database. That’s the very high quality, fully vouchered database that is commonly used for this work. You don’t get a perfect hit with 100% identity. This tells you that nobody has reported this sponge species before. That’s exciting, good for you, but now what? You’re going to have to figure out the taxonomy, and if you’re not careful the barcode gap can misinform you.

It would be reasonable to expect that your best blast hit will be a subject that’s very closely related to your novel sponge. But COI’s lack of a clean barcode gap for sponges means that this won't always happen.

To look closer at this problem, all the sponge sequences in the BOLD database were downloaded and a blastable database was built. Every sequence in the database was blasted against the database. All hits to the species of the query were discarded. This simulates a world where the query was never reported to the database. Strangely, there were many queries where there was a multi-way tie for best hit, rather than one single clearly most similar subject. A software tool identified 991 (out of 5363) queries with suspicious results: at least 1 of the best hits was in the wrong genus. You’re going to look into those results.

Browse to random.org and use the “True Random Number Generator” box to generate 3 random numbers from 1 to 991. What are your numbers?

Look at the file StrangePoriCOIs.txt. It contains the blast results for the 991 suspicious queries. Each block of lines (between “-------------” lines) specifies a query and all the subjects in the multi-way tie for best hit. Individual sequences are listed as i.d. codes followed by taxonomy. The i.d. is *not* the GenBank Accession Number. It’s the i.d. in the BOLD database. That’s really annoying. The taxonomy string gives the phylum, class, order, genus, and species of the sequence. Hopefully the format is obvious.

Make 3 copies of the table below. Fill out one copy for each of your random numbers. The random number is the number of the blast result. Don’t scroll through the big file looking for each number. Use your editor’s search feature. The goal is to think about how similar or dissimilar the subjects are from the query.

|  |  |
| --- | --- |
| Query i.d. |  |
| Query taxonomy |  |
| Number of hits listed |  |
| Number of hits to genus of query |  |
| Number of hits to right family but wrong genus |  |
| Number of hits to right order but wrong family |  |
| Number of hits to right class but wrong order |  |
| Number of hits to right phylum but wrong order |  |

Hmm, I wonder if this weirdness is because the BOLD database is limited. GenBank has lots more sponge COI sequences, though they aren't as reliable as BOLD’s. Pick one of your tables above for queries that have a lot of wrong identifications. Blast that query at GenBank and fill out one more table for those results.

In order to do this blast, you’ll need the query. Here’s where it’s *really irritating* that the BOLD ids can't be converted to GenBank Accession Numbers. You’ll have to do a little extra work. First, try the easy way: browse to <https://www.ncbi.nlm.nih.gov/>, select “Nucleotide” and search for the scientific name of your query organism, followed by “COI”. If you get a hit to the correct organism, you can use that Accession Number as your blast query. If that doesn’t work, open the file SpongeQueries.txt. It’s a fasta file where the deflines contain the BOLD ids, so you can do a text search for your BOLD id, and paste the fasta record (defline and sequence) into the query field.

What is the main take home message from your analysis of the use of CO1 for sponge identification? Why might sponge diversity be difficult to decipher with CO1?