BioE 131/231
Fall 2018

Lab Assignment #2

Overview and references

You will download some sequences, import them into your iPython notebook, analyze them, and plot the results. You should be able to complete this task using the lab slides as a reference, along with the resources listed below. Going forward, we will provide you with the tools to solve problems in lab and expect you to apply them in your own way.

You are expected to keep a thorough record of everything you did in your notebook. Create a folder in your home directory for each lab, and keep all your files there. Try to create a directory hierarchy that makes sense, like the one we went over in Lab 1. Copy and paste any terminal commands you used into a Markdown section and explain what the input was, what the tool did, and what the output was. Plot any results in-line and explain them.

An iPython notebook containing your analysis is due at the beginning of lab section next week. Please upload your .ipynb file to bCourses.

Matplotlib
https://matplotlib.org/users/pyplot_tutorial.html
https://matplotlib.org/examples/statistics/boxplot_demo.html
https://matplotlib.org/examples/statistics/histogram_demo_features.html

BioPython
https://biopython.org/wiki/SeqIO
https://biopython.org/wiki/Phylo
https://biopython.org/DIST/docs/api/Bio.Align.MultipleSeqAlignment-class.html

Background

You PI has given you a set of DNA sequences and told you they are related. It's not clear whether they are from the same taxon (paralogs), different taxa (orthologs), or perhaps a combination of the two. Perhaps they have some, but not all, domains in common (partial homology). What are some ways we can analyze them and figure out what their relationship is? What are some properties of DNA sequences, and how can we compare them to each other?

Getting the data into iPython

Head on over to bCourses > Files > Labs > Lab 2 and download seqs.fa. You'll then need to upload this file to iPython. From the main iPython landing page (after logging in), click "Upload." The location of data and procedure for uploading will be the same in future labs.

Generating a phylogenetic tree

Sequence alignment is the bread-and-butter of comparative genomics. When two sequences are aligned, we assume that each column corresponds to the same position in some ancestral sequence. Any indels or point mutations that occurred over evolutionary time then become apparent. Columns where the sequence has not changed show us which positions are conserved. Columns with differences show us where and how sequences have evolved since the most recent common ancestor.

We will cover this in greater detail later in the semester, but for now, it's safe to consider a sequence aligner as a "black box." In its simplest form, two unaligned sequences go in, and two aligned sequences come out. All the aligner really does is insert gaps (dashes, "-") such that some arbitrary scoring function is maximized. Consider the following example:

```
Seq1 ATGATTAAG
     ||     |
Seq2 ATTGGCTAAG
```

The sequences, with no gaps and beginning at the first column, match at only three positions. By eye, it looks like those two `TAAG` blocks at the end ought to match up, right? Recall that we're trying to infer which parts of the sequence were present in the common ancestor of both Seq1 and Seq2, so we can infer which parts of the sequence have changed.

```
Seq1 AT-GATTAAG
     || |  ||||
Seq2 ATTGGCTAAG
```

By inserting a gap in Seq1, we now have seven matching positions. We can now infer that there was an indel in column 3 (i.e., a T was either inserted in Seq2 or deleted from Seq1—note that we can't distinguish between these two possibilities without more data). It also looks like there may have been two transitions (i.e., A↔G or C↔T) in columns 5 and 6—note that we don't know the directionality of these changes, again, without more data.

What if we have many sequences, rather than just two? Can we align those to each other? Yes, of course! The result is called a multiple sequence alignment (MSA). When presented with a bunch of related sequences to analyze, generating an MSA is a great place to start.

There's a fantastic tool available to perform this very task: MUSCLE. Written by Bob Edgar, it has been cited over 22,000 times. If you're ever wondering how to get your citation count up, writing some simple tool that everyone will use is definitely the way to go.

Open up your terminal and use this command:

```
muscle –in seqs.fa –out seqs.aligned.fa
```

If you haven't already, examine the contents of the input and output of this command. What do you notice?

We will go into more detail on generating trees later in the semester, but for now, you can use Morgan Price's excellent tool FastTree to take your multiple alignment and turn it into a Newick-formatted tree.

Back in your terminal:

```
fasttree –nt < seqs.aligned.fa > tree.nwk
```

Finally, head into iPython to load and visualize your tree using Bio.Phylo. Are there any obvious clusters of sequences? Are there any pairs or groups of sequences that seem very closely related? Any that are any sequences that seem far more distantly related than the others?

Identifying sequences by BLAST

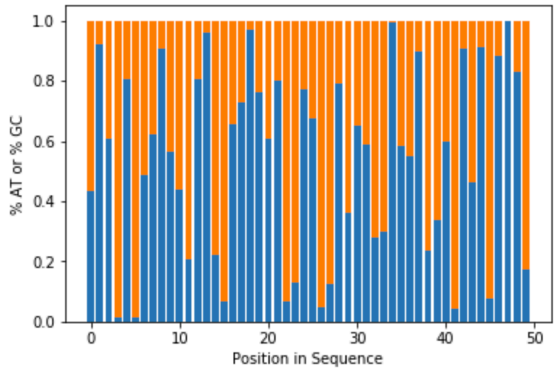
Comparing sequences in your dataset to themselves can be highly informative. There's also lots gain by comparing them to one of many large databases currently available. In this lab, we haven't been told where these sequences are from—are they even from a genome? Are they coding (i.e., do they encode a protein) or non-coding?

We could pop our seqs.fa file into BLAST and see what comes up, but that's going to take a while. We have a lot of sequences, and going through the results for all of them is slow. Using the tree you generated above, pick a single specimen from each obvious cluster. Now, load the sequences in iPython using Bio.SeqIO and print the sequence of each chosen specimen to the screen. Point your web browser at https://blast.ncbi.nlm.nih.gov/Blast.cgi and BLAST each representative against `nr/nt` using default parameters.

What are some of the most common matches for each cluster? Write them down in your lab notebook. Do they all match the same gene? Are there any that definitely do not match the others? Why might we not trust the annotations for sequences that come up in our BLAST?

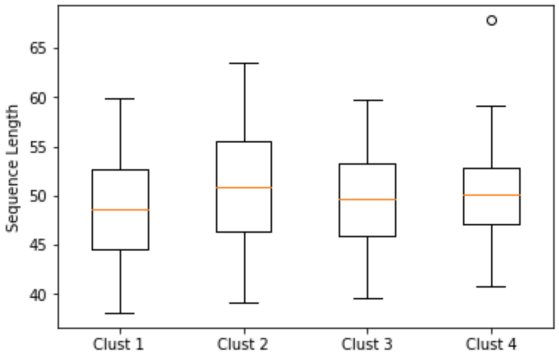
Calculating sequences statistics for each cluster

Now that you have your sequences broken out into clusters and a reasonable annotation for each cluster, let's calculate some simple statistics and plot them. For each cluster, generate a plot that looks like the one below. This lets us identify any regions that are skewed toward GC or AT, which might be useful for your final project later in the semester.



For each position in the MSA of each cluster, calculate the fraction of A or T, and the fraction of G or C. Note that the sum of each quantity will sum to 1.0. You'll plot the results using **plt.bar()**. Check out the "bottom" keyword argument to get the two bars stacked on top of each other, and the **plt.xlabel()** and **plt.ylabel()** commands to label the axes.

Next, find the length of each sequence in each cluster. Generate a box plot that shows the distribution of sequence lengths within each cluster. It will look something like this:



To do this part of the assignment, you'll need to have loaded the **seqs.aligned.fa** file you obtained above into your iPython notebook using Bio.SeqIO. You might also try **Bio.AlignIO** which will return a MultipleSequenceAlignment object that can be sliced by column.

See: https://biopython.org/DIST/docs/api/Bio.Align.MultipleSeqAlignment-class.html