Does it matter if the code is not concise

No – however simpler is better but point will not be removed

Summary

• Context: you want to develop a DNA identification service

• Identify the closest sequence in the database to the provided sequence

• Input: sequence database (adapted from GEO), test sequence

• Output: the closest sequence, and the difference percentage difference or the alignment showing the differences?

You can choose your own

• Stretch goal 1: Probabilities across database, p-value probability of what ?

• Stretch goal 2: reconstructed phylogeny

Checklist

* Did you choose a project?
* Did you make a plan ? (What will you need to do) ?
* What library will you use?

Plan and summary

Tasks

* Check out this repository it could be used for inspiration [barque/01\_scripts/12\_extract\_multiple\_hit\_sequences.py at master · enormandeau/barque (github.com)](https://github.com/enormandeau/barque/blob/master/01_scripts/12_extract_multiple_hit_sequences.py)
* [bcgsc/unikseq: 🧬unique (& conserved) DNA sequence identification (github.com)](https://github.com/bcgsc/unikseq) – it might be useful to compare only the conserved regions as well as the whole sequence depending on the data provided
  + Unique to individuals sequences might create a false negative if they differ substantially in breed model vs target sequence
* Decide which scoring system to use for alignment
* Use NumPy for probability testing of each alignment
* Create a test folder
* Output
  + Sequence of the closest alignment

Libraries that will be used

* Predominantly Biopython
  + Phylo – to create a phylogenetic tree output file
  + [Phylo - Working with Phylogenetic Trees · Biopython](https://biopython.org/wiki/Phylo)

from Bio import Phylo

# Parse the alignment file (e.g., FASTA format)

alignment\_file = "your\_alignment.fasta"

trees = Phylo.parse(alignment\_file, "fasta")

# Assuming there's only one tree in the file

tree = next(trees)

# Print the tree name

print("Tree name:", tree.name)

# You can also visualize the tree using Matplotlib or other packages

# (requires additional dependencies)

* Numpy for statistical analysis
* Matplotlib to draw the phylogenetic tree graph
* Pytest

Tests

* can a random sequence show similarity with a particular breed ?/ can you get likely alignments with a random sequence
* would a mix of two breeds be recognised?
* Use assert to test the functions as I go
* Pytest library
* Hypothesis testing
  + [Welcome to Hypothesis! — Hypothesis 6.98.2 documentation](https://hypothesis.readthedocs.io/en/latest/)
* Use %timeit function\_name() to optimise functions run time at the end

6 weekends to complete coursework

Week 1 = found the most similar sequence and created the output

Week 2 = test the output and organise/annotate code

Week 3 = stretch goal 1

Week 4 = stretch goal 2

Week 5 = test stretch goals

Week 6 = debugging and presentation

Try displaying a heatmap of a multisequence alignment of all species instead of the alignment of two

To do list

* ~~convert rest of the code to work with the alignment() function~~
* ~~output alignment to a file~~
* ~~output details to a file~~
* convert the file hierarchy to what Tristan suggested
* output better quality alignment to a file
  + [Visualizing Sequence data with Matplotlib - Liam (lamdv.github.io)](https://lamdv.github.io/Scientific%20Journey/Sequence%20visualization%20with%20matplotlib/)
* set up an environment for the project
* make a draft process map from notes

Issues

* dash-bio can not be installed
* ~~top\_alignment\_details not saved to a file~~
* the phylogenetic tree plot does not look presentable