Homework Exercise 7

1)

A)

Download from Pfam's FTP site all the occurrences of Pfam domains/HMM profiles in the entire human proteome. Use version **31.0** (not the newer ones). It should be a compressed .tsv file with ~100k rows and 14 columns. Each row in this file indicates an occurrence of a single domain/HMM profile in a single protein (each profile can occur multiple times in different proteins, and even in the same protein, and each protein can have multiple profiles).

Among the fields in this file are:

- **seq id**: the UniProt ID of the protein where the profile occurs.
- **alignment start** & **alignment end**: 1-based coordinates indicating the position of the profile within the protein.
- **hmm acc** & **hmm name**: The accession ID and the name of the profile occurring in the protein.
- **type**: the type of the occurring HMM profile (most commonly **Domain**).

How many occurrences are there of each type?

B)

Filter only the records of type **Domain**. What is the average length of Pfam domains? What are the 5 longest and 5 shortest domains (on average)?

Note that the field **hmm_length** indicates the theoretical length of the Pfam profile, not the actual length of its occurrence within a given protein.

C)

Compare the lengths of the **G-alpha** and **tyrosine kinase** domains. Is there a large difference in their average lengths?

Is this difference significant? Consider using Scipy's t-test (see this video for a refresher on t-tests). Other appropriate statistical tests will also be accepted.

D)

Download the required protein sequences from UniProt and add a new column to the table for the entire protein sequences of the records. Add another column for the sequence of the HMM profile within the protein (i.e. the latter is a subsequence of the former).

E)

What is the biological role of the **PDZ** domain?

Save the sequences of all the occurrences of the PDZ domain as a FASTA file. Submit it to an online Multiple Sequence Alignment tool (e.g. https://www.ebi.ac.uk/Tools/msa/clustalo/) and create a logo from it (e.g. using http://www.cbs.dtu.dk/biotools/Seq2Logo-2.1/).

Compare your logo to Pfam's HMM logo for this domain. Identify similarities and differences between the two logos, and suggest plausible explanations for the differences.

F)

There's one protein with 9 occurrences of the **PDZ** domain. Find it.

What is known about the function of this protein?

G)

[**Optional**, **recommended**]

What percentage of the exon junctions in the coding region of the transcript of the protein you have found occur within a **PDZ** domain? What percentage would you expect to see at random?

H)

[**Optional**, **recommended**]

At which locations inside the **PDZ** domain do exon junctions occur (across the entire human proteome)? How are they distributed compared to random uniform distribution?

I)

[**Optional**, **recommended**]

Do **PDZ** occurrences within the same protein more similar to each other than occurrences in different proteins?

Bonus Questions

1)

What is “time series” data and why is it important in biological and medical research?

Demonstrate Panda’s time-series capabilities with an example biological/medical dataset (e.g. you may use the BioTIME dataset).