User-Guide

Python Libraries required
Before running the code you'll need a couple libraries and files
so 'pip install beautifulsoup4 requests pandas biopython matplotlib seaborn'
also retrieve 'SearchResults-succinatedehydrogenase.tsv' from blackboard and save it in the current working directory

Step 1: Web Scraping and Data gathering

1. The part of the program scraps the website "https://parasite.wormbase.org/ftp.html" to download save and unpack protein FASTA files.
2. It takes all the files and puts them into a dictionary.
3. 3 random protein.fa.gz files are chosen at random

Step 2: Pfam Data Retrieval

1. This part of the program uses regular expression to extract specific Accession values from 'SearchResults-succinatedehydrogenase.tsv'
2. Pfam files are retrieved using regular expression to and are downloaded, saved and unpacked within the current working directory.

Step 3: SLURM Script and HMMer analysis

1. This part of the program generates a shell script named 'hmmer_script.sh'.
2. Shell script is made executable and is requires 'sbatch ' to run within HPC.
3. If not needed shell script can be executed within the terminal by using './hmmer_script.sh'

Step 4: Data visualisation

1. HMMer analysis produces output files with data
2. First graph produces a bar chart showing the frequency of hit for each species with their respected Pfam domain.
3. The second graph provides information about the distribution of E-values for each Pfam domain, grouped by the different species. The y-axis is set to a logarithmic scale for better visualization.

Important note: Run each part of the script repentantly. As when you get to the hmmer search you'll need to choose to either run the shell script locally within the terminal or send it to the HPC.