Input-output task demonstrating your analytical abilities

In this challenge, we would like you to analyse the TumE superfamily of proteins from sequence, taxonomic and functional perspectives. Support your conclusions with figures, plots or calculations that you obtained or carried out during the analysis.

1. How many relevant sequence clusters, or protein families, are present within the TumE superfamily? Examine the similarities and differences among the provided sequences of the superfamily members to address this question.

2. Integrate your sequence cluster analysis with the taxonomy and predicted functions data. Do the taxonomy and function distributions within sequence clusters match what you would expect?

3. What is the importance of applying computational approaches in biomedical research, particularly in the investigation of gut microbiota? Support your argument with a detailed exploration 1-2 examples from the TumE superfamily dataset. You can use the type II toxin-antitoxin system discovery in Durairaj et al. 2023 for inspiration.

The input dataset is available to download using the following link:

https://drive.google.com/drive/folders/1I-n-jYsBOlcKsJcIKQ2Jyz0i9NkA-rKB?usp=sharing

'*TumE.fasta*' file contains 860 protein sequences from the TumE superfamily, in FASTA format. Each sequence's header contains a UniProt Accession ID and a title, separated by a `|`.

'*TumE_annotations.csv*' file provides annotations for each of the 860 proteins, listed in the `uniprotAC` column. The annotations include taxonomy (Superkingdom, Phylum, Taxa, Class, Family, Genus, Order and Species) and two gene ontology (GO) term predictions 1 and 2. The GO terms represent function-related predictions generated by the DeepFRI tool. For most proteins, the two top scoring predicted terms are listed, however some may only have one; these cases will have an empty value in the `deepfri prediction 2` column.

The dataset is ready to use, does not require any manipulation, and should be easy to work with using a basic laptop. Feel free to use any tools for your analysis, and move beyond the provided annotations. Just make sure to clearly state what you use so that we can replicate your analysis.

For convenience, please present your code, results (e.g. figures, tables) and conclusions in a single document that can be shared with us and also clearly seen on screen. A Jupyter Notebook would be ideal for us.

Reference

Durairaj J, Waterhouse AM, Mets T, Brodiazhenko T, Abdullah M, Studer G, Tauriello G, Akdel M, Andreeva A, Bateman A, Tenson T, Hauryliuk V, Schwede T, Pereira J. Uncovering new families and folds in the natural protein universe. Nature. 2023 Oct;622(7983):646-653.