Project Objective: **Deciphering the functional role of a bacterial gene by colinearly conserved genomic neighborhood context.**

Part 1 (20% of final grade).

a. Design an efficient algorithm and write a program that does the following: Given a set of COG-spelled genomes S, where each genome is segmented into segments such that each segment could contain one or more operons. Also given parameters q, ℓ and an "unknown" COG X: find all strings c (over the COG alphabet) of length ℓ that are conserved in $\geq q$ of the genomes in the database S , such that COG X appears at least once in c. Implement the algorithm in python and document your code very clearly. Provide a "high level" description of your method\algorithm in the report.
b. Group your results (all COG strings c complying by the requirements stated in a) by their length (number of COGs in the string) and sort each group by decreasing quorum (the number of genomes in which the string was found).
c. Answer: How did you enforce the pre-specified constraints $q, \ell, and\ X$ on the sought strings c? How did you handle the fact that q refers to the minimal number of genomes in which the string c should appear, while in the input data set (the "cog words" file) the genomes are segmented into segments that could contain one or more operons?
d. Answer: What is the time and space complexity of your algorithm? Explain clearly and in detail.
e. Select an "unknown" COG X from the file "COG_INFO_TABLE" implementation assists in deciphering the function of a COG that is "uncharacterized", "poorly characterized", or "function uknown". Run your program with various parameter settings for q and ℓ, find conserved strings that include X and examine the results. Write a few sentences describing the functional context of X according to your results. Note that each group should chose a different COG X to analyze.
f. Your submission should include your documented code (preferably in a notebook), and a report of length 2-4 pages addressing the previous items a-e. The report can be in Hebrew, but you need to run "spellcheck" on it before submitting. Do not send me incremental versions of your report by email and ask me to give you comments again and again… instead, you are invited to schedule office hour meetings and ask me questions directly.

Code (documented notebook) and report due date: 15\11\2021

Behatzlaha!