Instructions for the group exercise on gamma distribution

Most genes in a genome encode proteins. The distribution of protein lengths is remarkably across the entire tree of life including bacteria, archaea and eukaryotes (see e.g., https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02973-2 for a recent publication on this topic). There are multiple reasons why protein lengths are confined to a relatively narrow interval that is the same for all species. In my opinion the most compelling reason is that proteins should be sufficiently short to fold and function fast (activation barriers increase with protein length) and yet be stable against temperature-induced misfolding (see Eqs. 6,7 and Fig. 4 in https://www.pnas.org/doi/full/10.1073/pnas.1114477108 using data and model from (https://www.sciencedirect.com/science/article/pii/S0006349511006618).

Gamma distribution is sometimes used to describe the distribution of protein lengths in an evolutionary context due to certain assumptions and processes that can lead to such a distribution. Here are some evolutionary processes that could potentially give rise to a Gamma distribution of protein lengths:

Gene Duplication and Divergence: Over evolutionary time, gene duplication events can occur, leading to the presence of multiple copies of a gene. These duplicate genes can accumulate mutations, insertions, or deletions, which can result in proteins of varying lengths. The diversification of gene copies through mutations and deletions can follow a stochastic process that, under certain conditions, can lead to a Gamma-like distribution of protein lengths.

Insertions and Deletions: Evolutionary events involving the insertion or deletion of genetic material can lead to changes in protein length. If these events occur randomly and accumulate over time, they may contribute to a distribution of protein lengths that resembles a Gamma distribution.

Neutral Evolution: In some cases, random genetic changes that do not have a strong impact on the fitness of an organism can accumulate over time due to genetic drift. This neutral evolution can result in a wide range of protein lengths, with the distribution potentially approximating a Gamma distribution.

Functional Constraints: Proteins often have specific functional domains or regions. Evolution may act to preserve the functional integrity of these domains while allowing variability in other regions. This differential selection pressure can contribute to a distribution of protein lengths that approximates a Gamma distribution, with a peak at a particular length representing the conserved functional domain.

Some of these processes were already used in the very early models when one had only a handful of genomes to analyze (see e.g., https://link.springer.com/article/10.1007/BF00163155 , https://www.sciencedirect.com/science/article/pii/S0378437199003702, and https://www.sciencedirect.com/science/article/pii/S0168952599019228)

It's important to note that the Gamma distribution is just one of many statistical models that researchers may use to describe protein length distributions in an evolutionary context. The actual distribution of protein lengths can vary significantly between different organisms and gene families, and the choice of distribution depends on the specific biological context and the data being analyzed. Empirically observed distribution of protein lengths in a given species is fitted by either Gamma or Log-normal PDFs (Fig. 10 in https://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-5-85).

Assignment 1:

Find whether Gamma or Log-normal is a better fit to the distribution of protein coding genes in Escherichia coli str. K-12 substr. MG1655 – perhaps the best studied model organism

1. Download all 4,639 genes in this strain of E. coli from the NCBI website: https://www.ncbi.nlm.nih.gov/datasets/gene/taxon/511145/ . Tip: put checkmarks to all features provided by the "Select Columns" option.
2. Import this file into Matlab (or Python if you prefer). I opened .tsv file in Excel, saved it again as 'e_coli_K12_MG1655_genes.xlsx' file and imported it to Matlab using a=readmatrix('e_coli_K12_MG1655_genes.xlsx'); c=readcell('e_coli_K12_MG1655_genes.xlsx'); for tables of numeric values in a and text values in c
3. Calculate the length of each gene as the difference between "Annotation Genomic Range Stop" and "Annotation Genomic Range Start" columns. Hint 1: gene length cannot be negative. Hint 2: if a gene starts at position 21 and ends also at position 21 what is its length? Does one need to add 1 to the length?
4. Filter the list of genes to only include genes marked as "PROTEIN_CODING" in "gene type" column.
5. Calculate lengths of all proteins (in number of amino acids) from lengths of the corresponding genes. Hint: 3 bases code for one amino acid. The last three bases of a gene are a stop codon which is not counted in the protein length. Are there any proteins whose length is not an integer? Find analyse why and correct. Check for a couple of genes that the protein length you calculated agrees with what is published by the NCBI (e.g. the protein dnaA has 467 aa: https://www.ncbi.nlm.nih.gov/datasets/gene/id/948217/products/)
6. Use distributionFitter tool in Matlab to fit the Gamma distribution and separately a Log-normal distribution to the data on protein lengths in E. coli str. K-12 substr. MG1655. Copy the results of both fits to your report. The results of the fit generated by distributionFitter look like this (I modified all the numbers on purpose):
   Distribution: Gamma
   Log likelihood: -20000.6 (<u>the closer is this number to zero - the better is the fit</u>)
   Domain: 0 < y < Inf
   Mean: 300.003
   Variance: 43000.7

Question: which of the two distributions provided a better fit for E. coli? Illustrate it by quoting Log likelihood value and selecting Display type -> "Probability plot", Distribution -> "Log normal". Include a snapshot of this figure in your report.

Assignment 2:

Repeat the analysis in Assignment 1 using genes of Thermococcus kodakarensis (https://www.ncbi.nlm.nih.gov/datasets/gene/taxon/69014/), a hyperthermophilic archaeon that lives in marine hydrothermal vents and terrestrial hot sulfur springs). It can grow in a range of temperatures from 60 to 100°C.

Which distribution, Gamma or Log-normal better fits gene lengths in Thermococcus kodakarensis? Is there any systematic difference between average gene length in Thermococcus kodakarensis and E. coli ?