Instructions for the Group Exercise on Fitting Gamma and Log-Normal Distributions to Protein Lengths

Most genes in a genome encode proteins. The distribution of protein lengths is remarkably similar across the tree of life—bacteria, archaea, and eukaryotes. For an overview, see: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02973-2. There are multiple reasons why protein lengths are confined to a relatively narrow interval shared across species. A compelling explanation is that proteins should be short enough to fold and function quickly (activation barriers tend to increase with protein length) yet stable against temperature-induced misfolding. See Eqs. 6–7 and Fig. 4 in: https://www.pnas.org/doi/full/10.1073/pnas.1114477108, which uses data and a model from: https://www.sciencedirect.com/science/article/pii/S0006349511006618.

Gamma and log-normal distributions are often used to describe protein-length distributions in an evolutionary context. Several evolutionary processes can lead to these shapes:

• **Gene duplication and divergence.** Over time, duplicates accumulate mutations, insertions, and deletions, yielding proteins of varying lengths. If insertion/deletion (indel) rates linearly scale with gene length, multiplicative effects can produce an approximately log-normal distribution of lengths.

• **Functional constraints.** Many proteins contain conserved domains under stronger selection, with other regions more variable. This can yield shapes well captured by a Gamma distribution with a peak near conserved domain sizes.

Early modeling ideas along these lines appeared when only a few genomes were available; for examples, see:
https://link.springer.com/article/10.1007/BF00163155
https://www.sciencedirect.com/science/article/pii/S0378437199003702
https://www.sciencedirect.com/science/article/pii/S0168952599019228

In practice, empirical protein-length distributions in a given species are often fit with Gamma or log-normal probability density functions; see Fig. 10 in: https://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-5-85.

Assignment 1 (Escherichia coli K-12 MG1655)

Goal. Determine whether a Gamma or a log-normal distribution provides a better fit to the distribution of protein lengths in *Escherichia coli* str. K-12 substr. MG1655.

1. **Download the gene table.**
   Source (taxon 511145): https://www.ncbi.nlm.nih.gov/datasets/gene/taxon/511145/
   Tip: Use the "Select Columns" option to include all relevant features. Expect roughly 4,600 genes.

2. **Load into MATLAB (or Python).**
   One simple path is to open the downloaded TSV in Excel, save as e_coli_K12_MG1655_genes.xlsx, then import it to MATLAB:

   ```
   a = readmatrix('e_coli_K12_MG1655_genes.xlsx');   % numeric values

   c = readcell('e_coli_K12_MG1655_genes.xlsx');     % text values
   ```

   (Adjust sheet/range options as needed for your file.)

3. **Compute gene lengths (nt).**
   Use the columns "Annotation Genomic Range Start" and "Annotation Genomic Range Stop."
   • Gene length cannot be negative.
   • Check coordinate conventions; if coordinates are inclusive, add 1 when computing length.

4. **Filter to protein-coding genes.**
   Keep rows with "gene type" = PROTEIN_CODING.

5. **Convert to protein lengths (aa).**
   Compute amino-acid length from nucleotide length: three bases per amino acid, and the terminal stop codon is not counted.
   • If any protein lengths come out non-integer, diagnose and correct (e.g., annotation offsets, untranslated regions).
   • Sanity-check a few genes by comparing to NCBI (e.g., dnaA is 467 aa): https://www.ncbi.nlm.nih.gov/datasets/gene/id/948217/products/

6. **Fit distributions in MATLAB.**
   Use the distributionFitter app (or programmatically) to fit both Gamma and log-normal to the protein-length data. Copy the numerical fit summaries to your report. An illustrative output might look like:

   ```
   Distribution:   Gamma
   Log likelihood: -20000.6   % less negative (higher) is better for comparable models
   Domain:         0 < y < Inf
   ```

Mean: 300.0
Variance: 43000.7

In the app, set **Display type → "Probability plot"** and **Distribution → "Lognormal"** to visualize the fit; include a snapshot.

Question. Which distribution fits *E. coli* protein lengths better? Justify your answer using the log-likelihoods.

Assignment 2 (Thermococcus kodakarensis)

Goal. Repeat Assignment 1 for *Thermococcus kodakarensis*, a hyperthermophilic archaeon that inhabits marine hydrothermal vents and terrestrial hot sulfur springs, growing across ~60–100 °C.

1. **Download the gene table.**
   Source (taxon 69014): https://www.ncbi.nlm.nih.gov/datasets/gene/taxon/69014/
2. **Repeat steps for computing protein lengths and fitting both Gamma and log-normal.**
   Report fit statistics (log-likelihoods and, if available, AIC/BIC) and summary measures (mean, median).

Questions.
• Which distribution (Gamma vs. log-normal) better fits *T. kodakarensis* protein lengths?
• Is there a systematic difference between average protein length in *T. kodakarensis* and *E. coli*? Comment on biological plausibility.