***Executive Summary: Eukaryotic Enzyme Length Distribution Analysis***

Isaac Oyediran, Rishav Eliyahu Sofer Dasgupta, Chien-Hung Lu, Jameson Born

***Research Question and Background***

Do eukaryotic enzymes exhibit preferred length distributions? Inspired by the foundational work of Berman et al. (1994) and Kolker et al. (2002), we explored whether these proteins show modular size patterns suggestive of evolutionary constraints. Using a prompt-driven AI framework, we revisited the hypothesis of ~125 amino acid periodicity with modern statistical modeling and spectrum analysis.

***Methodology***

We retrieved 41,705 reviewed eukaryotic enzymes from UniProt and filtered to 18,076 validated entries (50–600 amino acids). To remove redundancy, we applied two manual filtering strategies:

**Either filter the protein sequences by protein names or by sequences, and apply either Rule 1 or Rule 2:**

1. Rule 1: For any two protein entries with the same name or sequence, if the length difference is <20%, remove the shorter; otherwise, keep both.
2. Rule 2: If the shortest and second shortest differ by <20%, remove the shortest only; otherwise, keep all.

**Filtered dataset sizes:**

1. By protein name → Rule 1: 2,237 | Rule 2: 16,197
2. By sequence → Rule 1: 3,440 | Rule 2: 15,103

We then conducted exploratory data analysis, applied Spectral Analysis of Distributions (SAD), Cosine Fourier Transform, and developed a Gamma + Normal mixture model to characterize periodicity.

**Key Findings:**

SAD and Fourier analyses revealed consistent peak periodicities across datasets, ranging from 121–182 amino acids, with species-specific signatures (e.g., human: 180 aa, maize: 182 aa).

Our statistical model identified:

* Base periodic unit: μ = 121.4 ± 1.1 aa
* Gamma background: α = 4.17, β = 61.5
* Weighted peaks at 1×, 2×, 3×, 4× μ (p1–p4 all significant)

Model fit superiority: Likelihood ratio test (LRT = 2207.13, p < 0.00001) confirmed periodicity is statistically significant.

AIC/BIC favored a full periodic model over background-only.

**Conclusions and Implications:**

Our study confirms that eukaryotic enzymes exhibit a modular length architecture centered around ~121–127 amino acid units. These periodic patterns likely reflect structural or evolutionary constraints such as protein folding efficiency or domain organization. By combining manual curation with AI-driven statistical tools, we successfully replicated and extended prior work—offering fresh insight into enzyme structure and bioinformatic discovery workflows.