Executive Summary: Protein Domain Periodicity in Eukaryotic Enzymes

Objective: In this study, we aim to replicate and extend Kolker's original findings by applying the SAD methodology to a comprehensive dataset of eukaryotic enzymes. We implemented the exact SAD algorithm as described by Kolker et al. (2002) and applied it to both the full dataset and a non-redundant subset along with focused 10 fold cross-validation for robustness.

Key Finding: Analysis of eukaryotic enzyme sequences reveals a statistically significant periodicity **~126 amino acids** in eukaryotic enzyme lengths, corresponding to fundamental structural domains.

Methodology

- Created a non-redundant dataset (2,199 eukaryotic enzymes) from 18,076 sequences by selecting one representative per eukaryotic enzyme family to prevent statistical bias from overrepresented eukaryotic enzymes
- Applied **Spectral Analysis of Distributions (SAD)** to detect periodic components
- Validated findings with a statistical **mixture model** combining gamma-distributed background with normal distributions.
- Performed **10-fold cross-validatio**n for Robustness:
    - To confirm the reproducibility of the ~126 amino acid periodicity, we performed 10-fold cross-validation using the SAD algorithm. The period search was restricted to the biologically relevant range of 100–150 amino acids.

Why It Matters

Spectral Analysis confirms periodic patterns in non-redundant uniprot enzymes **~126aa**

Mixture Model results suggest fundamental period **~136aa,** p=value 3.35x10^-91

Cross Validation

- Across all folds, the preferred periods consistently clustered between **122–124 aa**, with the full dataset under the same constraint yielding **122 aa.**
- This consistency affirms that the observed periodicity is not an artifact of dataset composition but a stable and biologically meaningful signal.

Figure Highlights Three Approaches:

- **Cosine spectrum**: Dominant peak at **126 aa** with amplitude >1.3, far exceeding any other period
- **Probability density**: Mixture model accurately captures observed distribution **~ 136 with p-value 3.3x10-91**
- **Cross-validation boxplot**: Shows consistent preferred periods between **~122–124 aa** across 10 folds, confirming signal stability

The non-redundant approach was crucial for revealing this pattern, as redundant eukaryotic enzymes would have artificially strengthened certain length frequencies and obscured the true biological signal.