M1 - BIM Project

Roberto Netti, Giovanni Peinetti - Sorbonne Université

January 14, 2025

Exploration of Nitroreductase Protein Sequence Datasets for Generative Modeling

Objective

This project is designed to provide students with expertise in bioinformatics and machine learning techniques for analyzing biological sequence data. The focus is on constructing and analyzing a dataset of the Nitroreductase protein family [1], enzymes involved in the reduction of nitro-containing compounds and significant in various biological and biotechnological contexts, uncovering patterns and relationships within the sequence space. Generative modeling will be considered as a final step, contingent on the progress of the analytical phase.

Key Activities

1. Dataset Construction and Feature Extraction

Explore public protein/RNA databases (e.g., UniProt, Rfam) to identify and retrieve relevant sequence data. Perform multiple sequence alignments (MSA) using tools like BLAST and HMMER to curate and preprocess the dataset. Extract features from aligned sequences, identify conserved regions, and prepare data representations for computational analysis.

2. Dimensionality Reduction and Visualization of Sequence Space

Apply dimensionality reduction techniques, such as Principal Component Analysis (PCA), to visualize the sequence space and reveal latent patterns.

3. Clustering and Classification of Sequences

Use clustering methods (e.g., k-means, hierarchical clustering) to group sequences based on their extracted features and explore interrelationships within the dataset. Perform classification tasks, if applicable, to annotate or predict the functional properties of the sequences.

4. Evaluation and Interpretation

Interpret and validate results from clustering, classification, and dimensionality reduction for their biological relevance.

5. Exploration of Generative Models (Optional)

Based on the insights gained during the analysis phase, explore the development of generative models (e.g., Variational Autoencoders, Gaussian Mixture Models, or Potts Models) to generate

synthetic sequences with properties similar to the original dataset. Assess the quality and utility of synthetic sequences in comparison to the original dataset.

Expected Outcomes

Students will gain proficiency in handling biological sequence data, applying bioinformatics tools, and leveraging machine learning techniques for exploratory data analysis. The project will focus on analysis-driven insights, with generative modeling providing an optional extension of the learned concepts.

Conclusion

This internship provides a unique opportunity for students interested in bioinformatics, computational biology, and data science to work on real-world problems at the intersection of these fields.

References

[1] Janine N Copp et al. "Exploring the sequence, function, and evolutionary space of protein superfamilies using sequence similarity networks and phylogenetic reconstructions". In: *Methods in Enzymology* 620 (2019), pp. 315–347.