**Building a Profile Hidden Markov Model for the Kunitz-type Protease Inhibitor Domain**

**Overview**

This repository contains the code and data for building a **Profile Hidden Markov Model (HMM)** designed to identify the Kunitz-type protease inhibitor domain in protein sequences. The Kunitz domain is a conserved protein domain critical for inhibiting serine proteases, with significant implications in biological regulation and drug development.

The model was constructed using **experimentally validated structural data** and **multiple sequence alignments (MSA)** to accurately detect Kunitz domains across a variety of protein sequences. The performance of the model was evaluated using metrics such as **accuracy (ACC)**, **Matthews Correlation Coefficient (MCC)**, and **Receiver Operating Characteristic (ROC)** analysis.

**Project Steps**

**1. Data Collection and Preparation**

The first step involves collecting a comprehensive dataset of experimentally determined protein structures containing the Kunitz domain. The dataset was sourced from the **Protein Data Bank (PDB)**, using the following criteria:

* Data resolution of less than 3 Ångströms.
* Sequence lengths between 50 and 80 amino acids.
* Exclusion of proteins with mutations.

Once the data was collected, it underwent cleaning and processing to ensure consistency and reliability.

**2. Multiple Sequence Alignment (MSA)**

The next step involved performing a **Multiple Sequence Alignment (MSA)** to identify conserved regions and structural similarities across Kunitz domain-containing proteins. The alignment was performed using the **PDBeFold** server, and the results were visualized with **Jalview** to identify conserved regions critical for domain detection.

**3. Model Generation**

The **HMM** was generated using the **HMMER 3.4.0.2** package. The aligned sequences from the MSA were used as input, and the model was trained to capture both variability and conservation across the Kunitz domain sequences.

For model visualization, the **SkyAlign** platform was used, offering intuitive tools to explore the structural and evolutionary features of the model.

**4. Model Evaluation**

To evaluate the model’s performance, a two-fold **cross-validation** approach was employed. The dataset was split into two subsets, with the model being trained on one and tested on the other. **Accuracy (ACC)**, **Matthews Correlation Coefficient (MCC)**, and **ROC curves** were used to measure the model's performance.

Key findings from the evaluation:

* The model reached peak performance at an **e-value threshold of 1e-6**, with an accuracy of approximately **0.99995** and an MCC of **0.9995**.
* The confusion matrix showed minimal misclassifications with only **one false negative** and no **false positives** at this threshold.
* ROC curve analysis confirmed the model’s robust discriminatory capacity.

**5. Results**

The model effectively identifies the Kunitz domain with high accuracy and precision. The results show:

* **Accuracy**: Ranging from approximately **0.6362** to **0.9999** across different thresholds.
* **MCC**: Fluctuating from **0.0471** to **0.9995**.
* The optimal threshold for classification was determined to be **1e-6**, balancing sensitivity and specificity.

**6. Conclusion**

The constructed Profile HMM successfully identifies the Kunitz-type protease inhibitor domain, with high classification accuracy and robustness across various thresholds. The model is an effective tool for protein domain annotation, with potential applications in **functional genomics** and **drug discovery**.

Further improvements could include:

* Expanding the dataset for enhanced model generalization.
* Integrating additional methods for detecting divergent sequences.