Structural Protein Sequences

Introduction

The Protein Data Bank (PDB) is a vital repository for atomic-level structural data of biological macromolecules. Structural biologists rely on advanced techniques such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine and analyze the spatial configurations of proteins and other biomolecules. This study leverages the PDB dataset from 2018, which includes diverse protein structures like ribosomes, oncogenes, drug targets, and viruses, for developing machine learning models aimed at predicting protein structural classes.

Related Literature

1. *A Novel Protein Structural Classes Prediction Method Based on Predicted Secondary Structure*: Discusses innovative approaches for classifying protein structures based on secondary structures.
2. *Prediction of Structural Classes for Protein Sequences and Domains*: Explores the impact of various prediction algorithms and sequence representation methods on accuracy.
3. *Classifier Ensembles for Protein Structural Class Prediction*: Highlights the use of ensemble methods in improving prediction accuracy across different homology levels.
4. *Support Vector Machines for Prediction of Protein Domain Structural Class*: Focuses on the application of SVMs for protein domain classification, emphasizing their effectiveness and limitations.

Methodology and Results

Tools and Technologies:

- Python: Core programming language for data manipulation and model development.
- Pandas & NumPy: Libraries for handling large datasets.
- Scikit-Learn: Implemented traditional machine learning models like Logistic Regression, SVM, and Random Forest.

- TensorFlow & Keras: Used for building and training deep learning models, particularly Sequential Neural Networks.
- Matplotlib & Seaborn: Libraries for visualizing performance metrics.

Evaluation Metrics:

1. Accuracy: Overall correctness of predictions.
2. Precision: Ability to identify true positives without false positives.
3. Recall (Sensitivity): Ability to capture all true positives.
4. F1-Score: Harmonic mean of precision and recall, useful in imbalanced datasets.

Dataset Overview

- Size: 27.29 MB
- Structure: 140,911 rows and 14 columns, representing various biological and structural attributes.

Preprocessing

- Data preprocessing involved standardizing features using StandardScaler from Scikit-Learn.
- The dataset was transformed and exported in CSV format for easier integration during model training and testing.

Model Architecture

The models evaluated for protein structural class prediction include:

1. K-Nearest Neighbors (KNN)
2. Decision Trees
3. Random Forests
4. Naive Bayes
5. Logistic Regression
6. Support Vector Machine (SVM)

Each model was configured to optimize hyperparameters and improve prediction accuracy.

Model Training

- Random Forest achieved the highest accuracy of 93%, demonstrating balanced precision, recall, and F1-scores across all classes, excelling in Class 3.
- Decision Trees followed closely with an accuracy of 91%, though slightly underperforming compared to Random Forest.
- KNN achieved moderate results with 65% accuracy but lagged in recall and precision.
- Naive Bayes (43% accuracy) and Logistic Regression (42% accuracy) were less effective, particularly for certain classes, indicating limited suitability for this dataset.
- SVM achieved 47% accuracy, performing better than Naive Bayes and Logistic Regression but still significantly lower than ensemble methods.

Conclusions

Among the models evaluated, the Random Forest demonstrated the best performance, providing robust and consistent predictions across various structural classes of proteins. Decision Trees also performed well, offering a simpler yet effective alternative. On the other hand, simpler models like Naive Bayes, Logistic Regression, and SVM showed limited capability in capturing the complex patterns within the dataset, highlighting the importance of ensemble methods and deep learning for such tasks.

References

1. A novel protein structural classes prediction method based on predicted secondary structure.
2. Prediction of structural classes for protein sequences and domains—Impact of prediction algorithms, sequence representation, and homology on accuracy.
3. Classifier ensembles for protein structural class prediction with varying homology.
4. Support Vector Machines for Prediction of Protein Domain Structural Class.