Title: Breast Cancer Subtype Prediction Using Protein Sequences and K-Nearest Neighbors (KNN): A Bioinformatics Approach

Abstract: This study presents a comprehensive computational pipeline for the classification of breast cancer subtypes using protein sequence data. By employing Amino Acid Composition (AAC) features and the K-Nearest Neighbors (KNN) algorithm, the model is trained to differentiate between major subtypes: Luminal A, Luminal B, HER2-enriched, Triple-Negative, and Normal-like. The project incorporates dataset parsing, feature extraction, model training, evaluation, and an interactive prediction interface. Designed to be executed within Google Colab and integrated with Google Drive, this tool provides an accessible and effective framework for cancer subtype prediction based on sequence data.

1. Introduction: Breast cancer is a heterogeneous disease comprising multiple subtypes with distinct molecular characteristics and clinical implications. Traditional methods for subtype classification often require laboratory-based molecular testing, which may be time-consuming or inaccessible in some settings. With the rise of computational biology, there is an increasing opportunity to develop models that classify cancer subtypes using omics data, such as protein sequences. This project explores the potential of using Amino Acid Composition (AAC) features derived from protein sequences, combined with the K-Nearest Neighbors (KNN) classifier, to automate and enhance breast cancer subtype prediction.

2. Objectives:

- To develop a machine learning model for predicting breast cancer subtypes using protein sequence data.
- To extract AAC features from. fasta files representing protein sequences.
- To train and evaluate a KNN classifier on labelled datasets.
- To implement an interactive platform that allows prediction of subtypes based on new protein sequences.

3. Materials and Methods:

3.1 Dataset Preparation: Protein sequences in FASTA format were organized into directories by subtype: Luminal A, Luminal B, HER2-enriched, Triple-Negative, and Normal-like. These datasets were split into training, testing, and validation subsets.

3.2 Feature Extraction: Amino Acid Composition (AAC) features were calculated by determining the frequency of each of the 20 standard amino acids in a given sequence. This was implemented using Biopython.

3.3 Model Training: The dataset was normalized and split into training and test sets. A K-Nearest Neighbors (KNN) classifier was trained on the extracted AAC features using scikit-learn. Model evaluation was conducted using classification metrics and confusion matrix visualization.

3.4 Interactive Prediction Interface: The model includes functionality to accept a new protein sequence, extract its AAC features, and predict its cancer subtype. This interactive feature enables real-time prediction and usability in diverse scenarios.

3.5 Platform and Tools: The project was developed in Python using the following libraries: Biopython, Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn. Google Colab was used as the execution environment, with Google Drive integration for data handling.

4. Results: The KNN classifier achieved satisfactory performance in distinguishing between the five breast cancer subtypes. Confusion matrix plots and classification reports demonstrated the model's accuracy, precision, recall, and F1-score across classes. Stratified train-test splitting ensured balanced representation of each subtype in both training and evaluation phases.

5. Discussion: The proposed approach demonstrates the feasibility of predicting breast cancer subtypes using protein sequence data. While traditional diagnostics rely on more complex molecular data, AAC features provide a lightweight alternative for preliminary classification. The performance of the KNN model, though basic, highlights the potential of non-parametric methods in bioinformatics classification problems.

6. Conclusion: This project provides a replicable and accessible framework for classifying breast cancer subtypes based on protein sequences. With further refinement and incorporation of additional sequence features, such models may contribute to low-cost, high-throughput diagnostic pipelines in the future.

7. Future Work:

- Extend to other cancer types and biological datasets.
- Compare performance with other classifiers (SVM, Random Forest, Deep Learning).
- Deploy as a web application for broader access.

8. Acknowledgements: The authors thank the open-source community and academic collaborators for providing datasets and computational tools.

9. References:

- Pedregosa et al., 2011. Scikit-learn: Machine Learning in Python.
- Cock et al., 2009. Biopython: freely available Python tools for computational molecular biology.
- Relevant breast cancer proteomics and bioinformatics literature.