Sex Determination Model Using K-Means Clustering

1. Introduction

In our study, we utilized a combination of sequencing data from **692 samples** derived from a thermo sequencer and **190 samples** from nanopore sequencing. The primary goal is to determine the sex of samples based on the read counts of the X and Y chromosomes.

2. Approach

We based our sex determination logic on the following principles:

- **Female Samples**: Display an XX pattern, indicating the presence of two X chromosomes.
- **Male Samples**: Exhibit an XY pattern, characterized by one X and one Y chromosome.

The fraction of X chromosome reads tends to indicate female samples, while a near 1:1 ratio of X to Y chromosome reads indicates male samples. However, variations in read counts can lead to false positives in gender determination. To improve the accuracy of our predictions, we employed a **K-means clustering** machine learning model.

3. K-Means Clustering Algorithm

K-means clustering is an unsupervised machine learning algorithm used for partitioning a dataset into **K distinct clusters**. The algorithm works as follows:

1. **Initialization**: Randomly select K initial centroids from the data.
2. **Assignment Step**: Assign each data point to the nearest centroid, forming K clusters.
3. **Update Step**: Calculate the new centroids as the mean of all points assigned to each cluster.
4. **Iteration**: Repeat the assignment and update steps until convergence, where centroids no longer change significantly.

The output of the algorithm consists of cluster labels for each data point, allowing for the classification of samples into distinct groups, such as **male** and **female** in our case.

4. Feature Generation

To train the K-means clustering model, we generated features based on:

- **Read Counts of Chromosomes X and Y**: Essential for determining the sex ratio.
- **SRY Gene Presence**: The SRY gene, located at **chrY: 2,654,896 - 2,655,295**, is exclusively found on the Y chromosome. Its presence is indicative of male samples.
- **Fraction of Chromosomes X and Y**: Used to understand the distribution of reads.

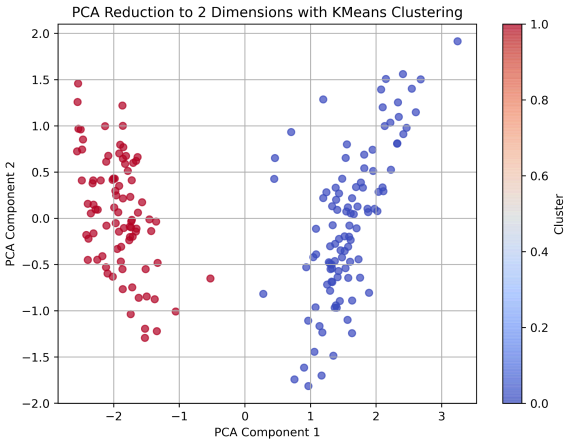
5. Data Scaling

To enhance the clustering process, we applied **StandardScaler** from **sklearn** to standardize the features. This scaling method ensures that each feature contributes equally to the distance calculations in the clustering algorithm.

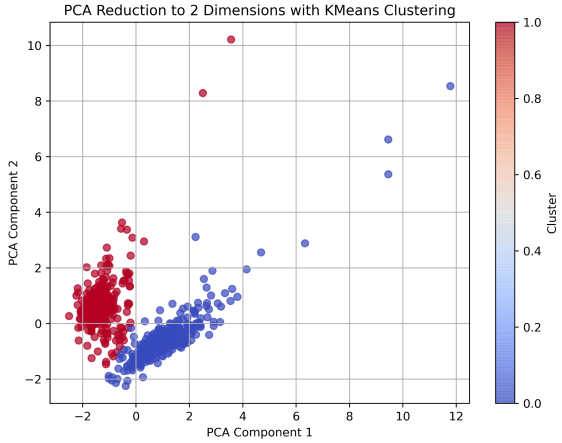
6. Model Training and Validation

After generating the feature set and scaling the data, we trained the K-means clustering model. We validated the model using a test dataset, achieving an impressive **99.5% accuracy** in sex determination.

The plots below demonstrate that male and female samples are **linearly separable** based on the generated features, showcasing the effectiveness of our model.



1 . Nanopore Samples



2. Thermo Samples

7. Future Improvements

To further enhance the model's accuracy and robustness, we plan to:

- Incorporate more samples into the training dataset.
- Explore additional features that could improve gender classification.