REPORT - 2

Objective: Extract the features of the given gene sequence dataset.

Solution:

Note:

- Step-1: Open the given Colab file of the gene sequence
- Step-2: Load the AR, DOM, and PR folder (contains gene sequence data) in Colab
- Step-3: Run it
- Step-4: The feature will show in result with the accuracy of the SVM model as well.

Important Keyword:

- **Gene Sequence**: The **sequence** tells scientists the kind of **genetic** information that is carried in a particular **DNA** segment. It contains a huge number of characters in one sequence and there are only four characters A, C, G, and T that represent amino acids.

  Each gene sequence is composed of a series of nucleotides (A, T, G, C), and the distribution of these nucleotides within the sequence can provide insights into the sequence's characteristics.

  For example,
  A small DNA sequence: ATGGCTCCAACTCAGGATCCCAACAGTGT

- **Feature:** A feature refers to an individual measurable property or characteristic of a data sample. Features are typically represented as columns in a dataset, and they provide information that can be used to make predictions or perform analyses.

- **Label:** Set the label to each sample in entire datasets, so we can categorize them.

  Each given gene sequence folder AR, DOM, and PR has its own category.
  For example,

  Feature — class
  AR ---- 0,
  DOM —-- 1,
  PR ---- 2

- **SVM model:** Support Vector Machine (SVM) is a machine learning model that aims to find the best decision boundary or hyperplane to separate data points of different classes, maximizing the margin between the classes.

- **Testing dataset:** The portion of the dataset used to train or fit a machine learning model. It consists of input features and corresponding target labels. The model learns from this data to understand the underlying patterns and relationships in the data.

- **Training dataset:** The portion of the dataset used to evaluate the performance of the trained machine learning model. It is separate from the training data and contains input features, but the corresponding target labels are typically withheld. The model makes predictions on this data, and the predictions are compared to the actual labels to assess the model's accuracy or performance.

For calculating the features of gene sequence we perform the following concept:

1. **Shannon Entropy Feature:** Shannon entropy, also known as information entropy, is a measure of the uncertainty or randomness in a dataset or information source. It quantifies the average amount of information or surprise contained in each data point.
   For Example:
   Original sequence: "ATCGATCGATCGATCG"
   Numeric array: [0, 3, 1, 2, 0, 3, 1, 2, 0, 3, 1, 2, 0, 3, 1, 2]
   Compute_SE_feature(Numeric array): 1.13

   Case 1: Training data = 50% and Testing data = 50%
   Model Accuracy: 38%

2. **Hurst Exponent Feature:** The Hurst exponent is a statistical measure that quantifies the long-term memory or persistence of a time series data. It is named after Harold Edwin Hurst, a British hydrologist who first introduced this concept.
   The Hurst exponent is a measure of the long-term memory or self-similarity of a time series data. It can also be applied to analyze DNA sequences to capture certain characteristics or patterns in the data.

   For Example:

   Original sequence: "ATCGATCGATCGATCG"
   Numeric array: [0, 3, 1, 2, 0, 3, 1, 2, 0, 3, 1, 2, 0, 3, 1, 2]
   Compute_HE_feature(Numeric array): 0.03 (Import hurst module to use this function)

   Case 1: Training data = 50% and Testing data = 50%
   Model Accuracy: 43%

3. **Run Length Encoding Feature:** Run-length encoding (RLE) is a simple compression technique that can also be used as a feature extraction method. It involves representing consecutive repeated elements in a sequence with a count of the element and the element itself.
   Example:

sequence = "AATTTCCGGGGGAA"
Features = [(2, 'A'), (3, 'T'), (2, 'C'), (5, 'G'), (2, 'A')]
Compute_statical_information(Features) = 2.75 (Using Mean of the 2,3,2,5,2)

Case 1: Training data = 50% and Testing data = 50%
Model Accuracy: 41%

Note: