Predictive model for DNA sequencing and Explaining BIAS

Neelanjan Mitra mitra.ne@northeastern.edu

Neehar Satti satti.n@northeastern.edu

Reetikesh Patel patel.reet@northeastern.edu

ABSTRACT

The aim of the model running on K-mer analysisis used to provide information on the frequency and distribution of k-mers in a DNA sequence. This information can be used to make inferences about various aspects of the DNA sequence, such as its composition, complexity, and potential functions. The data obtained from this model is utilized to compare various sets of sequences and make predictions about their classification within a particular group of organisms. However, We encountered a few Biases that were predicted by our model.

1.0 Methedology

1.1 String to array:

Label Encoder class is imported, and an array of characters, including 'a', 'c', 'g', 't', and 'n' are used to train the label_encoder object. This produces a relationship between each character and a corresponding integer label, which can subsequently be utilized for machine learning tasks or other analytical purposes.

1.2 Ordinal Encoder:

The ordinal_encoder function is applied by initially converting the DNA sequence to an array of individual characters utilizing the string_to_array function. This array is then passed to the ordinal_encoder function. The resulting output in the array consisting of float values representing the encoded DNA sequence to exemplify how to encode a DNA sequence by adopting an ordinal encoding method, where each nucleotide is uniquely represented by a float value.

1.3 One Hot Encoder:

One hot encoder accepts a DNA sequence string as input. It utilizes the label_encoder to convert each character in the sequence to an integer value. One Hot Encoder to convert the integer-encoded sequence into a one-hot encoded matrix.The function returns the resulting one-hot encoded matrix as output One Hot Encoder convert the integer-encoded sequence into a one-hot encoded matrix. Finally, the function returns the resulting one-hot encoded matrix as output.

1.4 Kmers Function:

K-mer analysis is a powerful method used in DNA sequence prediction to identify patterns of nucleotides (the building blocks of DNA) in a sequence. The analysis involves dividing a DNA sequence into overlapping k-length subsequences (k-mers) and counting the frequency of each k-mer in the sequence. These counts can then be used to infer certain characteristics of the DNA sequence, such as its composition, complexity, and potential functions.

The provided code entails a function named Kmers_funct that accepts two arguments, a DNA sequence string, and an integer value called size.The function returns a list comprising all k-mers of the size definied that can be extracted from the input DNA sequence.

2.0 Model

Multinomial Naive Bayes (MultinomialNB) is a classification algorithm based on Bayes' theorem. It is primarily used for text classification and is particularly effective for tasks such as spam filtering, sentiment analysis, and document categorization.

In MultinomialNB, each document is represented as a bag-of-words. The algorithm assumes that the frequency of each word is independent of the frequency of all other words.

The model is used to classify DNA sequences into different classes based on their features. Naive Bayes classifier is used to predict the class of DNA sequences based on their k-mer features. The k-mers are generated by breaking the DNA sequences into short overlapping subsequences of given length, and then converting each subsequence into a feature vector of counts of each possible k-mer. The resulting feature vectors are then used to train a Naive Bayes classifier to predict the class of

each DNA sequence. This type of analysis can be used to identify patterns in DNA sequences that are associated with different types of organisms.

3.0 Results

```
Confusion matrix for predictions on human test DNA sequence

Predicted   0    1   2    3    4   5    6
Actual
0          99    0   0    0    1   0    2
1           0  104   0    0    0   0    2
2           0    0  78    0    0   0    0
3           0    0   0  124    0   0    1
4           1    0   0    0  143   0    5
5           0    0   0    0    0  51    0
6           1    0   0    1    0   0  263
accuracy = 0.984
precision = 0.984
recall = 0.984
f1 = 0.984
```

```
Confusion matrix for predictions on Chimpanzee test DNA sequence

Predicted    0    1    2    3    4    5    6
Actual
0          232    0    0    0    0    0    2
1            0  184    0    0    0    0    1
2            0    0  144    0    0    0    0
3            0    0    0  227    0    0    1
4            2    0    0    0  254    0    5
5            0    0    0    0    0  109    0
6            0    0    0    0    0    0  521
accuracy = 0.993
precision = 0.994
recall = 0.993
f1 = 0.993
```

```
Confusion matrix for predictions on Dog test DNA sequence

Predicted    0   1   2   3    4   5    6
Actual
0          127   0   0   0    0   0    4
1            0  63   0   0    1   0   11
2            0   0  49   0    1   0   14
3            1   0   0  81    2   0   11
4            4   0   0   1  126   0    4
5            4   0   0   0    1  53    2
6            0   0   0   0    0   0  260
accuracy = 0.926
precision = 0.934
recall = 0.926
f1 = 0.925
```

4.0 BIAS

The frequency of certain k-mers can be influenced by the overall GC content of the genome, genome size, and repetitive sequences. Additionally, the frequency of k-mers can vary between different species or strains of organisms, which can make it more difficult to compare and analyze their genomes. This is particularly true for highly divergent organisms or those with unusual genomic characteristics. Two organisms can also have similar DNA sequences for a protein, which may result in only a few distinct k-mers. This lack of information could prevent the k-mer analysis model from predicting the correct organism classification and lead to false positive results.

5.0 Conclusion

Based on the analysis and implementation of the Naive Bayes Classifier model on DNA sequences, we can conclude that it can be a useful tool for DNA sequence classification. The accuracy of the model is high and it can effectively differentiate between different species based on their DNA sequences.

However, it is important to note that the model has some limitations and biases. For example, it still uses k-mers analysis which can be sensitive to certain biases in the data. Furthermore, the model could fail when dealing with larger and more complex DNA sequences. Therefore, it is important to carefully consider the limitations and biases of the model when using it for DNA sequence analysis and classification.