Bioinformatics: project 4

The purpose of this project is to build a Random Forest classifier that predict of ‘being an enhancer’ (https://en.wikipedia.org/wiki/Enhancer_(genetics))

1. Prepare all combinations with the nucleotides "A", "T", "C", "G" with a length of k = 4, but without reverse complement sequences (you can use https://biopython.org/DIST/docs/api/Bio.Seq-module.html#reverse_complement)
2. *example of 4-mer:* AAAA, AAAT, ... - without the 4-mer reverse you should have 136 sequences. These will be your features

3. Prepare a function that will count the frequency of each 4-mer in the DNA sequence (number_of_4mer / sequence_length, here sequence_length = 1500).
   Do it with a step = 1. You should have (number_of_this_4mer / length_ sequence) in each column.

4. Treat both the original sequence and reverse_complement as the same feature.
   example: both "AAAT" and "ATTT" (which is reverse_complement to ‘AAAT’) refer to the same feature.

5. Convert file with positive (vista1500) and negative (random1500) data into a 4-mer frequency format (hint: concatenate positive and negative data to create training data, hint2: create a list of '0' and '1' to create a label for training data)

6. Build a Random Forest classifier based on this data using the sklearn package

7. Divide the entire chromosome 21 sequence (chr21.fa) on frames=1500 with step=750. Convert each frame into ‘4-mers frequency’ format (with same order as in Random Forest classifier !)

8. Use RF classifier on each frame

9. Count the average prediction for entire chromosome - set this value for each frame with ‘N’ in the frame sequence

10. Save results in WIG format (https://genome.ucsc.edu/goldenPath/help/wiggle.html)

    Points:

- 2p. for converter sequence > ‘4-mers frequency’
- 2p. for trained Random Forest classifier
- 2p. for compute chr21 using RF
- 1p. for preparation results in WIG format