Final project

Background: Mutations arise in the genome of individual organisms. When within a gene, these mutations can be classified into 2 categories: synonymous and non-synonymous substitutions. Synonymous substitutions are single base changes that modify a codon without a modification in its encoded amino acid. For instance, AGU and AGC are 2 different codons that code for the same amino acid (serine) and only differ by one base. On ther other hand, non-synonymous mutations are those which modify both the codon and the amino acid. For example, mutating a C to an A in the third position of the UAC codon will change a tyrosine to a stop codon (UAA).

Goal: To estimate the number of non-synonymous, synonymous and intergenic SNPs present in a sampled population.

Input files:

1. SNP file: a list of single nucleotide polymorphisms (SNPs) from a bacterial population. The information in this file was obtained by mapping the raw sequencing reads of individual bacterial isolates to a complete reference genome of the same species. The contig of origin, position in the contig and nucleotide change are displayed for every SNP.
2. Annotation file: contains the geneID, gene description, contig of origin and genome coordinates (gene start and gene end) for every gene encoded in the reference genome.
3. Genome file: a .fasta file containing the complete genome of the reference sample.

Requirements:

1. A text file with 9 columns separated by TAB: (1) contig of origin, (2) SNP position, (3) substitution type, (4) gene description, (5) gene ID, (6) the reference codon, (7) the modified codon, (8) the reference amino acid (single letter) and (9) the modified amino acid (single letter). Please, check the example output file attached.
2. A barplot showing the number of non-synonymous, synonymous and intergenic SNPs present in each contig. Please, check the example output file attached.
3. The above files should be produced with the execution of a single python script.
4. Your script must be in .py format and executable in Python3.

Hints:

For each SNP:

1. Check whether that SNP is present in any gene in the assigned contig;
2. Use the gene coordinates to obtain the gene sequence from the original genome;
3. Identify the codon containing the SNP position of interest;
4. Store the amino acid encoded by the reference;
5. Store the amino acid encoded by the modified codon;
6. Compare the 2 amino acids to identify the type of substitution resulting from the SNP of interest.

7. Remember that, if the gene where the SNP falls in is on the reverse strand (the start position of the gene is greater than the end position), you must account for that on your code. For instance, if the SNP file has a C for position 35 and the gene is on the reverse strand, you must reverse translate the gene in order to get the correct codon (read it backwards and change the bases to the reverse complement) and change the SNP to a G.