Lab 1

1. Скачать данные - Plasmodium falciparum chromosome 13: ftp://ftp.ncbi.nlm.nih.gov/genomes/Protozoa/Plasmodium_falciparum/NC_004317.gbk или скачиваем chtomosome 13 и ее анатацию с ncbi. http://www.ncbi.nlm.nih.gov/genome/?term=Plasmodium+falciparum
2. Прочесть позиции начал кодирующих регионов из gbk файла (CDS FEATURES). Позиции, в которых встречаются “<” или “>”, игнорировать(обозначают, что точные позиции начала или конца региона не известны).
3. Вывести в файл последовательности нуклеотидов для начал сайтов трансляции(+- 10 нуклеотидов от первой позиции трансляции, т.е. всего 21 нуклеотид). Неизвестные нуклеотиды ‘N’ не надо учитывать. Если кодирующий регион идет “не последовательно”, то надо считывать только кодирующую область. (Например, если кодирующая область join(1000..1008,1200..1500), то начальная позиция +10 будет соответствовать координате 1201).
4. Делим хромосому на 2 выборки (сайты трансляции, которые начинаются в первой половине хромосомы - обучающий набор, остальные - тестовый). Далее работаем с обучающей выборкой. Посчитайте частоту встречаемости нуклеотидов для каждой позиции сайта трансляции(PFM).
5. Используя найденные последовательности в пункте 2 постройте лого с помощью weblogo.
6. Посчитать матрицу весов(position specic scoring matrix), используя результаты, полученные в пункте 4.
7. Используя матрицу весов, постройте гистограмму scores для
    a. известных начал CDS в обучающем наборе (из пунктов 1-2)
    b. всех подпоследовательностей во всей хромосоме. Приблизить это распределение нормальным.
8. По данным, полученным в предыдущем пункте, постройте график FP, FN в зависимости от отсечки для score, по которой определяется, что последовательность в хромосомы - это CDS.
9. С использованием значением отсечки, которое кажется вам разумным, проведите поиск CDS во второй половине хромосомы и проанализируйте результат (FP, FN, sensitivity, specificity, and PPV).