Домашнее задание №3

Введение

В этом домашнем задании вам предстоит реализовать один из алгоритмов предсказания генов семейства Genemark. Наша реализация алгоритма подразумевает наивный подход предсказания кодирующих участков (используем теорему Байеса):

Вероятность что гипотеза верна

Вероятность возникновения события при условии верности гипотезы

$$P(H|E) = \frac{P(H) \times P(E|H)}{P(E)}$$

Вероятность того, что гипотеза верна, при наших наблюдениях

Вероятность наблюдаемых событий

Вероятность P(E) в знаменателе, которая в нашем случае является P(SEQ), [вычисляется по формуле полной вероятности](#):

Гипотеза/Hypothesis

Наблюдение/Evidence

$$P(NC|SEQ) = \frac{P(NC) \cdot P(SEQ|NC)}{P(SEQ)} = \frac{P(NC) \cdot P(SEQ|NC)}{P(C_1) \cdot P(SEQ|C_1) + P(C_2) \cdot P(SEQ|C_2) + P(C_3) \cdot P(SEQ|C_3) + P(NC) \cdot P(SEQ|NC)}$$

Пример вычисления

В качестве примера вычисления рассмотрим последовательность TATTACTTC.

Таблицы вероятностей:

	pos1	pos2	pos3	nc
T	0.181203	0.296080	0.314970	0.272901
C	0.206748	0.226955	0.233379	0.228714
A	0.256854	0.302191	0.238267	0.276767
G	0.355196	0.174773	0.213384	0.221619

	pos1	pos2	pos3	nc
T|T	0.185541	0.498016	0.316087	0.318294
C|T	0.189713	0.199146	0.180139	0.208396
A|T	0.249076	0.172641	0.221208	0.244180
G|T	0.375670	0.130197	0.282566	0.229130
T|C	0.185250	0.248514	0.269667	0.252617
C|C	0.156196	0.198164	0.244435	0.213797
A|C	0.276236	0.315008	0.244359	0.286844
G|C	0.382318	0.238314	0.241540	0.246742
T|A	0.180048	0.334056	0.307936	0.269508
C|A	0.194071	0.213889	0.193709	0.205264
A|A	0.262876	0.327849	0.330772	0.320696
G|A	0.363005	0.124206	0.167584	0.204531
T|G	0.174164	0.193288	0.384069	0.242311
C|G	0.304287	0.267349	0.377809	0.299724
A|G	0.230045	0.342267	0.099310	0.251184
G|G	0.291504	0.197095	0.138812	0.20678

Для начала посчитаем вероятности того, что последовательность *кодирующая* для трех рамок считывания и вероятность того, что последовательность *некодирующая*:

$$P(SEQ|C_1) = P_C(T_1) \cdot P(A_2|T_1) \cdot P(T_3|A) \cdot P(T_2|T) \cdot P(A_2|T) \cdot P(C_3|A) \cdot P(T_1|C) \cdots = 9.93*10e\text{-}7$$
(0.181 0.172 0.307 0.307 0.172 0.193 0.185)

$$P(SEQ|C_2) = P_C(T_3) \cdot P(A_1|T) \cdot P(T_2|A) \cdot P(T_3|T) \cdot P(A_1|T) \cdot P(C_2|A) \cdot P(T_3|C) \cdot P(T_1|T) \cdot P(C_2|T) = 4.41*10e\text{-}6$$
(0.314 0.249 0.334 0.316 0.249 0.213 0.269 0.185 0.199)

$$P(SEQ|C_3) = P_C(T_1) \cdot P_C(A_2|T) \cdot P(T_1|A) \cdot P(T_2|T) \cdot P(A_3|T) \cdot P(C_1|A) \cdot P(T_2|C) \cdot P(T_3|T) \cdot P(C_1|T) = 1.77*10e\text{-}6$$
(0.181 0.172 0.180 0.498 0.221 0.194 0.248 0.316 0.189)

$$P(SEQ|NC) = P_{NC}(T) \cdot P_{NC}(A|T) \cdot P(T|A) \cdot P(T|T) \cdot P(A|T) \ldots P(T|C) = 4.8 * 10e\text{-}6$$
(0.273 0.244 0.269 0.318 0.244 0.209)

Затем, по теореме Байеса, посчитаем *финальные* вероятности, для *первой*, *второй*, *третьей* рамок считывания соответственно:

TATTACTTC

$$P(C_1|SEQ) = \frac{P(C_1) \cdot P(SEQ|C_1)}{P(C_1) \cdot P(SEQ|C_1) + P(C_2) \cdot P(SEQ|C_2) + P(C_3) \cdot P(SEQ|C_3) + P(NC) \cdot P(SEQ|NC)} = 8\%$$
(0.5 9.93*10e-7 0.5 4.41*10e-6 0.5 1.77*10e-6 0.5 4.8*10e-6)

TATTACTTC

$$P(C_2|SEQ) = \frac{P(C_2) \cdot P(SEQ|C_2)}{P(C_1) \cdot P(SEQ|C_1) + P(C_2) \cdot P(SEQ|C_2) + P(C_3) \cdot P(SEQ|C_3) + P(NC) \cdot P(SEQ|NC)} = 36\%$$
(0.5 9.93*10e-7 0.5 4.41*10e-6 0.5 1.77*10e-6 0.5 4.8*10e-6)

TATTACTTC

$$P(C_3|SEQ) = \frac{P(C_3) \cdot P(SEQ|C_3)}{P(C_1) \cdot P(SEQ|C_1) + P(C_2) \cdot P(SEQ|C_2) + P(C_3) \cdot P(SEQ|C_3) + P(NC) \cdot P(SEQ|NC)} = 16\%$$
(0.5 9.93*10e-7 0.5 4.41*10e-6 0.5 1.77*10e-6 0.5 4.8*10e-6)

И, наконец, *финальная* вероятность для *некодирующей* последовательности:

TATTACTTC

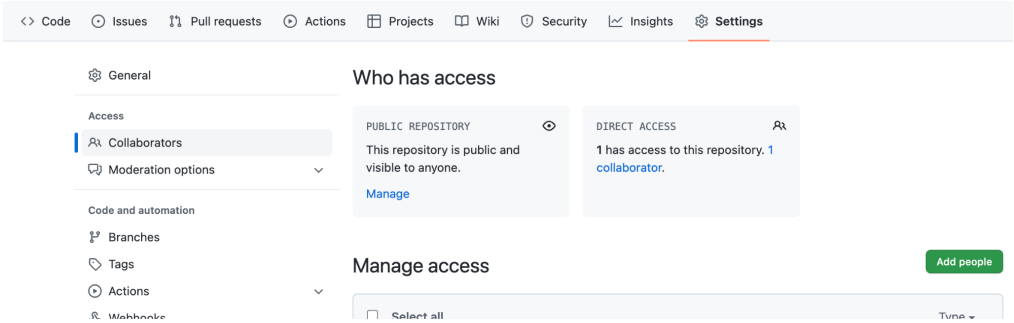
$$P(NC|SEQ) = \frac{P(NC) \cdot P(SEQ|NC)}{P(C_1) \cdot P(SEQ|C_1) + P(C_2) \cdot P(SEQ|C_2) + P(C_3) \cdot P(SEQ|C_3) + P(NC) \cdot P(SEQ|NC)} = 40\%$$

0.5 4.8*10e-6

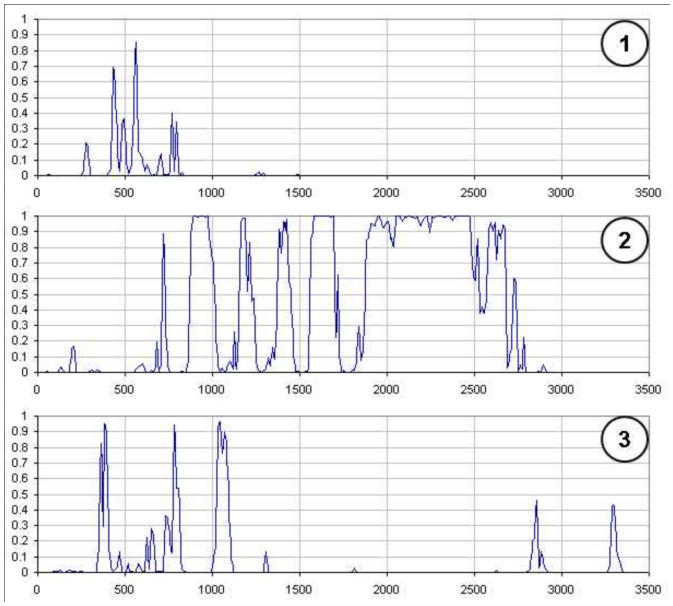
0.5 9.93*10e-7 0.5 4.41*10e-6 0.5 1.77*10e-6 0.5 4.8*10e-6

Обязательная часть задания (8 баллов)

- На сайте github.com создаем **приватный** репозиторий и приводим ссылку на этот репозиторий в общей гугл-таблице (**вкладка HW3**) https://docs.google.com/spreadsheets/d/1bdKBlDMFvhX8xStZUL2YHRTGLpmdiLNU5JcVmy8K-Yw/edit?usp=sharing
    - Также необходимо дать доступ ассистентам (**dRabbit-ab, efrsw**) к репозиторию для будущей проверки (Settings => Collaborators => Add people):

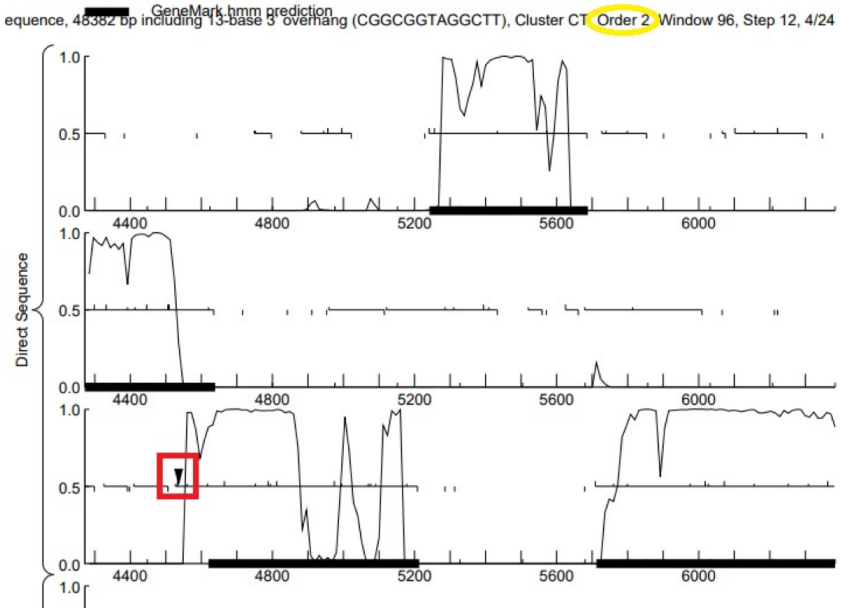


- Рекомендуется выполнять работу в Google Collab ноутбуках.
    - Если вы будете выполнять работу на сервере или на своем компьютере, необходимо будет также загрузить написанный код на Github
- Подготавливаем исходные данные для начала имплементации алгоритма GeneMark
    - Пример Google Colab ноутбука:
        - https://colab.research.google.com/drive/1bJnuCnWStcupKug_l1N9yaVxfmqtIORn?usp=sharing#scrollTo=P1AqeOAbn61s
- Итогом работы будет построение графика вероятностей кодирования в трех рамках для выбранного вами участка ДНК длиной около 3кб (важно, чтобы в этом участке был хотя бы один ген и все гены были расположены на "+" цепи, размер окна 96 п.о. с шагом 12), например:



Бонусная часть задания 1 (2 балла)

- Добавить на графики метки для всех start и stop кодонов, а также горизонтальными линиями соединить те пары start-stop, которые образуют открытые рамки считывания (ORF):
    - Подсказка -- для поиска открытых рамок считывания разрешается использовать сторонние библиотеки, такие как Biopython
- Сопоставить информацию про открытые рамки считывания с посчитанными вероятностями кодирования в разных рамках и выделить те ORF-ы, которые мы классифицируем как гены. Отметить эти предсказанные гены на графике также:



Бонусная часть задания 2 (5 баллов)

- Сделать возможным предсказание генов как на прямой, так и на обратной цепи:
    - Для этого улучшить функцию подсчета вероятностей для данной последовательности (текущего окна), чтобы она возвращала вероятности для 7 моделей, а именно 4 модели, которые были ранее (NC, COD1, COD2, COD3), плюс три возможные рамки на обратной цепи (NEG_COD1, NEG_COD2, NEG_COD3)
- Обновить функцию построения графика с вероятностями кодирования, чтобы он включал 6 кодирующих панелей -- 3 для "+" цепи и 3 для "-" цепи, как это сделано в реальной выдаче GeneMark.
    - Получить график для выбранного участка ДНК, чтобы в нем был 1 ген на "+" цепи и 1 ген на "-" цепи

Бонусная часть задания 3 (5 баллов)

- Создать полноценный *ab initio* алгоритм GeneMark:
    - Входные данные -- файл .fasta с одним прокариотическим геномом, состоящим из одной или несколько последовательностей ДНК
    - Инициализация (первая итерация)
        - Во всех последователях ДНК найти все ORF-ы (во всех 6-ти рамках)
        - Выбрать участки ДНК, которые не содержат длинных ORF-ов (например не длиннее 90 нт) ни в одной из 6-ти рамок. Для этого можно отметить все ORF-ы длиннее 90 нт (в любой из 6-ти рамок) и участки между ними считать некодирующими. Из полученных некодирующих последовательностей, посчитать вероятности нуклеотидов нулевого и первого порядка
        - Выбрать все длинные ORF-ы (длиной более 900 нт) и по ним получать частоты первой, второй и третьей позиций кодона (COD1, COD2 и COD3).
    - Итерации
        - С полученными параметрами для моделей NC, COD1, COD2 и COD3 осуществить предсказание генов на всем геноме как на +, так и на минус цепи
        - По полученным предсказаниями генов обновить параметры моделей NC, COD1, COD2 и COD3
    - Остановка итераций и выдача результатов
        - когда ни один из генов (или большинство генов -- например 95%) не меняют свои координаты
- Для вашей реализации алгоритма GeneMark получить полный список генов и сравнить его с тем списком, который выдал оригинальный алгоритм GeneMark из прошлого ДЗ (этот список генов мы считаем золотым стандартом)
    - Интересно оценить, какой % генов успешно предсказывается нашей версией GeneMark
    - Имейте в виду, что координата начала гена может немного варьировать, так как есть альтернативные старт-кодоны. Поэтому если у гена, предсказанного нашим GeneMark-ом, немного отличается старт-координата от оригинального GeneMark, то мы этим пренебрегаем и считаем, что ген предсказан верно

Список файлов для сдачи

- В репозитории в файле *README*.md
    - Ссылки на google colab ноутбуки
    - График с профилями вероятностей кодирования в трех рамках участка ДНК длиной около 3кб
    - Результаты выполнения бонусной части задания
- В репозитории в папке src:
    - любой другой код, который был использован для выполнения задания

Форма отчетности

Github репозиторий, содержащий все полученные результаты.

Последний срок сдачи: 13 ноября до 23:59 (будет отслеживаться по последнему коммиту в репозиторий). Штраф -0.5 балла за каждый день просрочки.

В случае возникновения вопросов обращаться по каналам связи:
Telegram ассистентов:

- @dbushnev
- @efrsw