Домашнее задание 1

Срок выполнения — утро 16 апреля 2025. Отчёты присылайте на sspirin@hse.ru

1. Выберите две бактерии с заметно разным GC-составом геномов
2. Скачайте последовательности самых больших хромосом каждого из геномов
3. Создайте химерную последовательность из нескольких тысяч букв, состоящую из случайных кусков этих геномов средней длины 300. Лучше всего разыгрывать длину куска исходя из экспоненциального распределения со средним 300, затем начало куска равномерно по длине хромосомы
4. Запрограммируйте алгоритм Витерби для декодирования входной последовательности произвольной длины в соответствии со скрытой марковской моделью с двумя эмиссионными состояниями. Эмиссионные вероятности сделайте соответствующими частотам букв в двух геномах, переходные вероятности — так, чтобы среднее число шагов в каждом из состояний было равно 300. Программа должна выдавать файл со строкой (или столбцом) символов “1” и “2”, соответствующим двум состояниям.
5. Примените свою программу к химерной последовательности, а также к каким-нибудь участкам исходных геномов
6. Посчитайте процент ошибок в каждом случае, сделайте выводы

Отчёт должен быть в формате pdf и включать достаточно подробное описание каждого шага (что за бактерии, что за хромосомы, из каких участков составлена химерная последовательность, параметры HMM и т.д.). К отчёту приложите код (файл *.py, *.c и т.п.) и инструкцию по запуску программы из командной строки на произвольной входной последовательности, можно в виде ссылки на репозиторий из текста отчёта. Не присылайте ссылку на Colab или файл Jupiter Notebook — я должен убедиться, что программа корректно работает на любой последовательности.

Несколько замечаний

По п. 1: здесь есть (довольно произвольный) список бактерий с указанием процента букв G+C в них.

По п.2: если EMBL AC хромосомы, например, CP093938.1, то добыть последовательность в формате fasta можно по адресу https://www.ebi.ac.uk/ena/browser/api/fasta/CP093938.1

По п. 3: в геномах иногда встречаются буквы, отличные от A, T, G, C (чаще всего N), что означает неточно прочитанные нуклеотиды. При выборе фрагментов геномов следите, чтобы в этих фрагментах не было таких букв.

По п. 4: программа может принимать на вход либо fasta-файл (строка с названием, затем последовательность нуклеотидов), либо файл, где кроме собственно последовательности нуклеотидов ничего нет. Fasta-файл, наверное, удобнее всего читать средствами BioPython; если же программа будет читать файл с “чистой” последовательностью, то нужно предусмотреть игнорирование пробелов и переносов строк (результат не должен

зависеть от их расстановки). Что касается модели, то разумно для обоих состояний сделать эмиссионные вероятности для G и C одинаковыми (и равными половине частоты G+C в соответствующем геноме), то же для A и T.