Домашнее Задание 1

Были выбраны следующие две бактерии:

Deinococcus radiodurans

Deinococcus radiodurans характеризуется геномной организацией, существенно отличающейся от большинства бактерий. Вместо единственной кольцевой хромосомы, типичной для большинства видов, *D. radiodurans* имеет полиплоидную систему, состоящую из двух хромосом и двух плазмид, присутствующих в клетке в множественных копиях. В частности, геном включает хромосому I (2 648 638 пар оснований), хромосому II (412 348 пар оснований), мега-плазмиду (177 466 пар оснований) и меньшую плазмиду (45 704 пар оснований), GC-содержание порядка 67.1% [Mau+21].

Одной из ключевых характеристик генома *D. radiodurans* является тороидальная (кольцеобразная) морфология ДНК. Такая конфигурация ограничивает диффузию фрагментов ДНК, образующихся при радиационном повреждении, что способствует сохранению пространственной близости между разорванными концами и обеспечивает корректное восстановление цепей даже без наличия непрерывного шаблона [Lev+03].

Deinococcus radiodurans демонстрирует высокую эффективность восстановления генома после многочисленных двойных разрывов ДНК [ZZR24]. Восстановление включает ESDSA (extended synthesis-dependent strand annealing process) с последующей рекомбинацией. [Ben+10].

Clostridium botulinum

Clostridium botulinum существенно отличается от геномной организации *D. radiodurans* и характеризуется наличием четырёх различных групп, способных продуцировать ботулинический нейротоксин — один из самых мощных известных токсинов [Seb+07].

Протеолитический штамм *C. botulinum* Hall A (ATCC 3502), относящийся к Группе I, имеет геном, состоящий из хромосомы длиной 3 886 916 пар оснований и плазмиды (*pBOT3502*) длиной 16 344 пар оснований. Эти генетические элементы содержат 3650 и 19 кодирующих последовательностей соответственно, а GC-содержание плазмиды (26,8%) несколько ниже, чем у хромосомы (28,2%) [Seb+07].

Геном *C. botulinum* демонстрирует адаптацию к специфическим экологическим условиям и патогенному образу жизни. В соответствии с его протеолитическим фенотипом, геном содержит значительное число генов, кодирующих секретируемые протеазы и ферменты, участвующие в метаболизме аминокислот. Заметно отсутствие значительного количества недавно приобретённой ДНК, что указывает на относительную стабильность генома. Такая стабильность характерна для адаптации *C. botulinum* к сапрофитному образу жизни в почве и водных средах с использованием нейротоксина для быстрого разрушения клеточных структур и получения доступа к питательным веществам [Seb+07].

Эволюция геномной архитектуры может обеспечивать поддержку специализированных стратегий жизнедеятельности у бактерий. Структурная организация генома *D. radiodurans*, включая наличие тороидальной структуры, обеспечивает высокую эффективность механизмов восстановления ДНК, в то время как более традиционная, специализированная организация генома *C. botulinum* способствует его адаптации к конкретной экологической нише и производству нейротоксина [CP15].

Загрузка Геномов

Геномные данные были загружены из Европейского архива нуклеотидов (ENA).

Для генома *Deinococcus radiodurans* основная хромосома была загружена по номеру `CP015081.1` (длина: 2 646 742 оснований). Для *Clostridium botulinum* использовался `CP002410.1` (длина: 2 773 157 оснований).

```python
import requests

def download_fasta(ac, output_file):
    url = f"https://www.ebi.ac.uk/ena/browser/api/fasta/{ac}"
    response = requests.get(url)
    if response.status_code == 200:
        with open(output_file, "w") as file:
            file.write(response.text)
```

Код 1: Скачивание Данных

Формирование химерной последовательности

Химерная последовательность создаётся с использованием случайных фрагментов из двух геномов.

- Фрагменты для объединения выбирались случайно из обоих геномов. Длины фрагментов подбирались из экспоненциального распределения со средней длиной в 300 нуклеотидов. Для получения такого распределения можно была использована одна из функций библиотеки `numpy`:

  ```python
      piece_length = max(1, int(np.random.exponential(avg_length)))
  ```

  Код 2: Химерная Последовательность

  После чего проверить, состоит ли полученный фрагмент только из `ACTG`, и добавить его в последовательность.

- Для каждого сегмента фиксировался источник (состояние 1 для *Deinococcus radiodurans* и 2 для *Clostridium botulinum*), что позволяет установить истинную аннотацию при декодировании.

- Итоговая длина последовательности — 50 000 нуклеотидов.

Расчёт эмиссионных вероятностей

На основе чистых участков геномов вычислялись нуклеотидные частоты:

Геном 1 — *Deinococcus radiodurans*:
`A: 0.16463316, T: 0.16463316, G: 0.33536684, C: 0.33536684`.

Геном 2 — *Clostridium botulinum*:
`A: 0.35755170, T: 0.35755170, G: 0.14244830, C: 0.14244830`.

Эти значения отражают соотношение нуклеотидов, где вероятности для `A` и `T` заданы равными (половина суммарной AT-частоты), аналогично для `G` и `C`.

Параметры HMM и алгоритм Витерби

Построенная HMM включает два состояния:

Состояние 1: соответствует участкам из *Deinococcus radiodurans*.

Состояние 2: соответствует участкам из *Clostridium botulinum*.

Основные параметры модели:

Начальные вероятности:
Для обоих состояний установлены равные шансы $\frac{1}{2}$.

Переходные вероятности:
При условии, что средняя длина фрагмента составляет 300 нуклеотидов, вероятность продолжения того же состояния вычисляется как $\frac{299}{300}$, а вероятность смены состояния — $\frac{1}{300}$.

Эмиссионные вероятности:
заданы в соответствии с вычисленными частотами нуклеотидов для каждого генома.

```
    total = len(sequence)
    count_A = sequence.count("A")
    count_T = sequence.count("T")
    count_G = sequence.count("G")
    count_C = sequence.count("C")
    at_freq = (count_A + count_T) / total
    gc_freq = (count_G + count_C) / total
    probs = {
        "A": at_freq / 2,
        "T": at_freq / 2,
        "G": gc_freq / 2,
        "C": gc_freq / 2,
    }
```

Код 3: Вероятности для каждой последовательности

Алгоритм Витерби реализован через логарифмы для повышения точности — замены операций умножения на сложение. Он применяется для декодирования как химерной последовательности, так и сегментов исходных геномов.

Результаты

- Для химерной последовательности алгоритм Витерби показал **2.85%** ошибок. Итоговая последовательность предсказанных состояний сохранена в файле `predicted_statesa_chimeric.txt`.

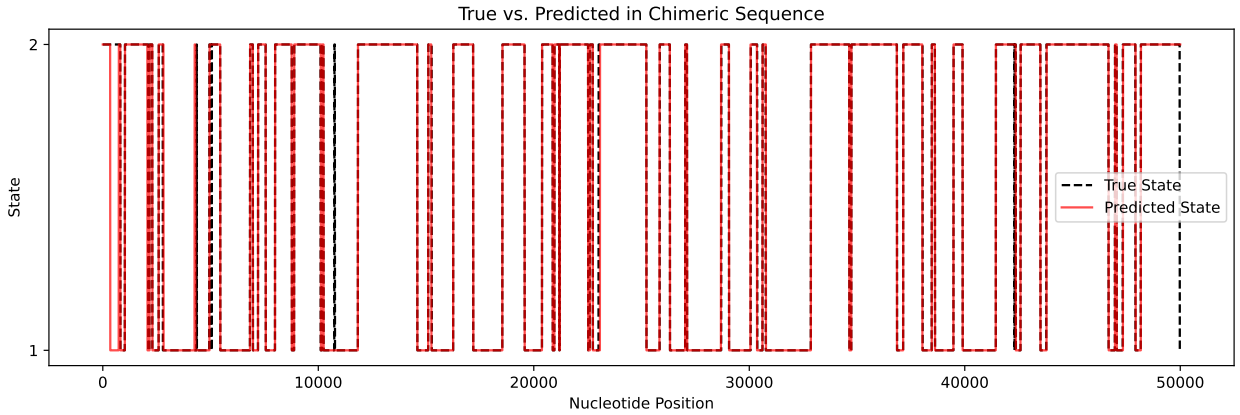

Рис. 1: Предсказания и Ошибки

- Для однородных сегментов: ошибка декодирования для *Deinococcus radiodurans* составила **0.29%**, а для *Clostridium botulinum* — **1.52%**.

Можно запустить программу несколько раз, чтобы получить лучшее представление об ошибках декодирования. Чтобы результаты не повторялись, каждый раз мы будем брать новые случайные сегменты двух изначальных последовательностей.

Таким образом, получим следующие значения
(в формате среднее $\pm$ среднеквадратичное отклонение): 2.05 ± 0.93 для химерной последовательности, 0.64 ± 0.84 для *Deinococcus radiodurans* и 0.94 ± 2.59 для *Clostridium botulinum*.

Заметим, что у *Clostridium botulinum* для некоторых участков встретились значения, заметно превышающие среднее, что говорит о неравномерном распределении `GC`-состава в геноме. У химеры среднее значение заметно выше значений исходных геномов, что говорит о том что алгоритм может заметить переход не сразу, или же, что можно заметить на Рис. 1, может и вовсе его пропустить если участок был довольно коротким.

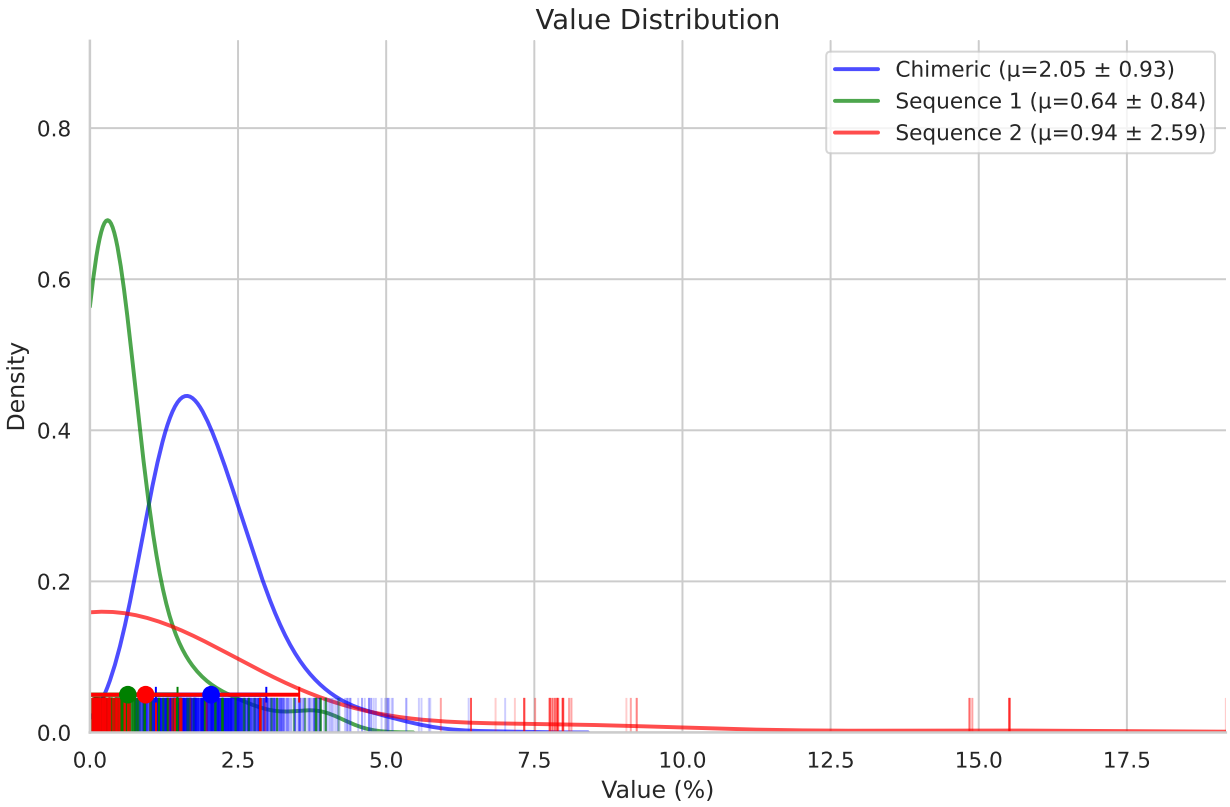

Рис. 2: Ошибки Декодирования

Инструкция по запуску программы

Программа написана на Python и использует файлы в формате FASTA для чтения геномных последовательностей. Для запуска программы из командной строки используйте следующую команду:

```
python BioAlg_HW1.py file1.fasta file2.fasta
```

Код 4: Запуск программы

Где `BioAlg_HW1.py` — исходный код, который можно скачать тут, а `file1.fasta` и `file2.fasta` — выбранные вами последовательности в формате `.fasta`.

Требования:

- Python 3.x
- `sys`
- `numpy`
- `random`
- `math`

- `matplotlib`
- `BioPython` (для работы с FASTA-файлами, т.е. `SeqIO`)
- `io` (для `StringIO`)

Программа выведет в консоль длины исходных геномов, а также информацию о длине химерной последовательности, эмиссионных вероятностях, процент ошибок для химерной и изначальных последовательностей.

```
Genome 1 length: 2646742
Genome 2 length: 2773157
Chimeric length: 50000
Chimeric error rate: 1.88%
Genome 1 error rate: 0.76%
Genome 2 error rate: 0.05%
```

Код 5: Пример Вывода

Предсказанная строка состояний будет сохранена в файл `predicted_states_chimeric.txt`.

Библиография

[Lev+03] S. Levin-Zaidman и др. «Ringlike structure of the Deinococcus radiodurans genome: a key to radioresistance?» в: *Science* 299.5604 (янв. 2003). PMID: 12522252, с. 254—256. DOI: 10.1126/science.1077865.

[Seb+07] M. Sebaihia и др. «Genome sequence of a proteolytic (Group I) Clostridium botulinum strain Hall A and comparative analysis of the clostridial genomes». в: *Genome Res.* 17.7 (июль 2007). Epub 2007 May 22; PMID: 17519437; PMCID: PMC1899119, с. 1082—1092. DOI: 10.1101/gr.6282807.

[Ben+10] Esma Bentchikou и др. «A major role of the RecFOR pathway in DNA double-strand-break repair through ESDSA in Deinococcus radiodurans». в: *PLoS Genetics* 6.1 (2010), e1000774. DOI: 10.1371/journal.pgen.1000774.

[CP15] A. T. Carter и M. W. Peck. «Genomes, neurotoxins and biology of Clostridium botulinum Group I and Group II». в: *Res Microbiol* 166.4 (май 2015). Epub 2014 Nov 4; PMID: 25445012; PMCID: PMC4430135, с. 303—317. DOI: 10.1016/j.resmic.2014.10.010.

[Mau+21] G. K. Maurya и др. «Molecular insights into replication initiation in a multipartite genome harboring bacterium Deinococcus radiodurans». в: *J Biol Chem* 296 (июнь 2021). Epub 2021 Feb 21; PMID: 33626388; PMCID: PMC7988490, с. 100451. DOI: 10.1016/j.jbc.2021.100451.

[ZZR24] K. Zahradka, D. Zahradka и J. Repar. «Structural Differences between the Genomes of Deinococcus radiodurans Strains from Different Laboratories». в: *Genes (Basel)* 15.7 (июнь 2024). PMID: 39062626; PMCID: PMC11276467, с. 847. DOI: 10.3390/genes15070847.