Лабораторная работа 3.

Первое знакомство с биоинформатикой.

Платформа:

<https://tio.run/>

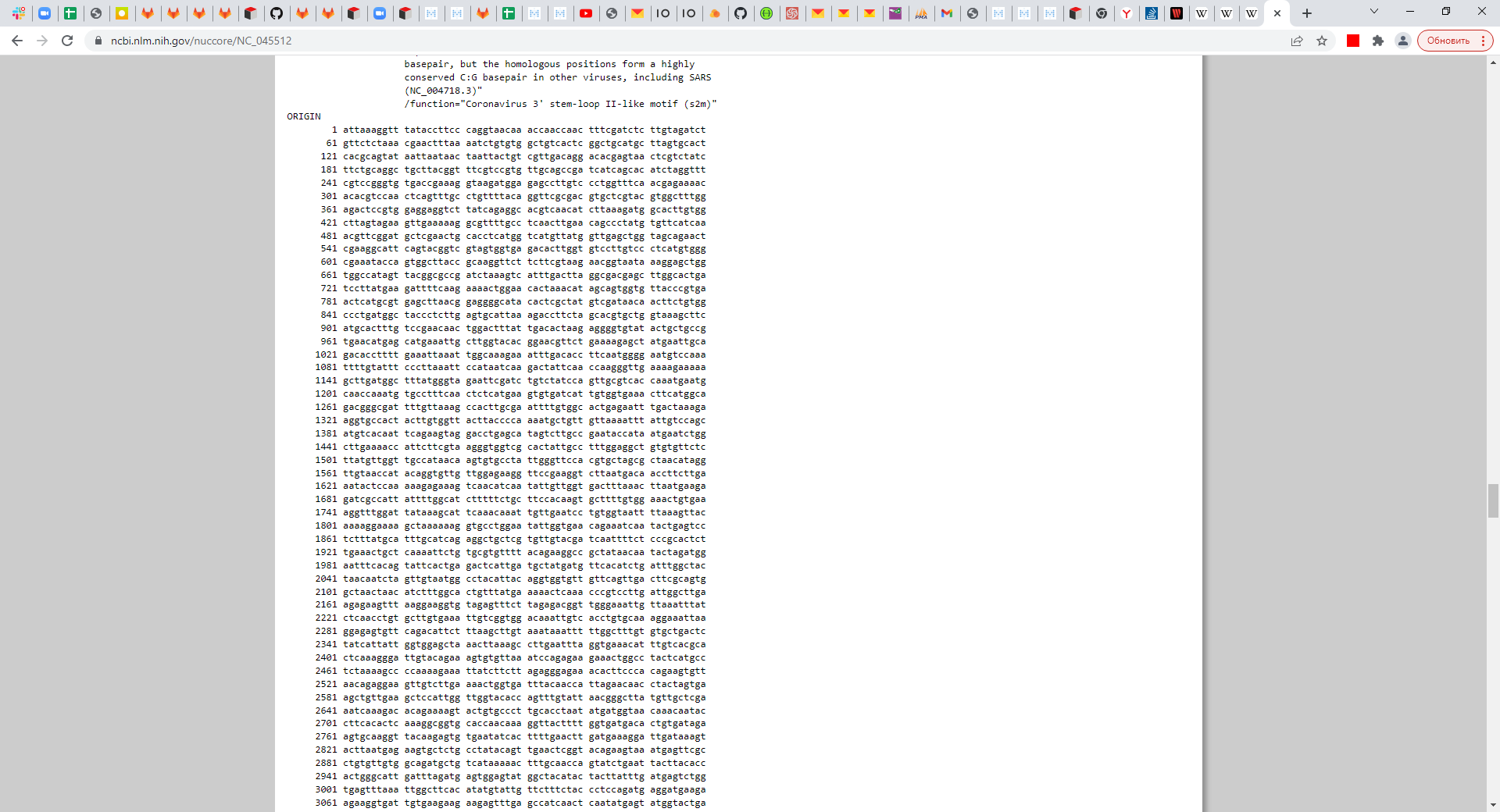
Цели:

* Знакомимся с некоторыми задачами биоинформатики
* Работаем со строками
* Работаем с сырыми данными из источников

Синопсис, тезисно:

* Биоинформатика содержит класс задач, плотно связанный с работой со строками – как правило, поисх общих подпоследовательностей максимальной длины, поиск самых коротких уникальных последовательностей для пары строк.
* Геном – (если предельно упростить) это последовательность пар оснований, которую часто можно представить строкой с алфавитом {A,T,C,G}
* Есть сайты, где выложены дампы таких последовательностей, например, для вируса гриппа, или SARS-Cov-2, с ними мы будет работать

Задание:

1. Пример ниже рассматривается с языком Python (версия 3).
2. Берем дамп вируса Ковид (Ухань)   
   **(1)** <https://www.ncbi.nlm.nih.gov/nuccore/NC_045512>  
   Вот его начало  
   
3. В Python можно использовать строку, записанную в несколько строк. Пример -   
     
   a = “””многострочная строка  
   берется в тройные двойные кавычки  
   “””  
   print(a)
4. Для начала работы нам нужно строки дампа вируса преобразовать в строку без переносов (“\n”), пробелов и цифр  
   a = "abc123"  
   a = a.replace("1", "")  
   print(a) # Выведет abc23
5. Можно сделать отдельную программу для преобразования, либо выполнять это прямо в коде решения, оставив в коде оригинальные дампы вируса.
6. Возьмем вирус гриппа H5N1  
   **(2)** <https://www.ncbi.nlm.nih.gov/nuccore/DQ150426.2>
7. Найти минимальную специфичную последовательность для (1), отсутствующую в (2). Такая последовательность – необходимые данные для теста ПЦР. Синтезированная молекула с такой последовательностью при наличии в исследуемом материале такой же последовательности, начинает активно размножаться (реплицироваться), что позволяет понять, что исследуемый материал содержит такой вирус. В то же время из-за сильно растущей стоимости синтеза от длины последовательности и возникает требование ее минимальности.
8. Найти минимальную специфичную последовательность для (2), отсутствующую в (1). Такая последовательность может использоваться для диагностики этого штамма гриппа методом ПЦР.
9. Найти минимальную общую последовательность для (1) и (2). Такой тест позволяет выявлять, не разделяя, наличие любого из этих двух вирусов в исследуемом материале.
10. Появился новый штамм ковид - Ковид (Дельта)  
    **(3)** <https://www.ncbi.nlm.nih.gov/nuccore/OK091006.1>
11. В этот момент возникает неопределённость – работает ли наша последовательность из п.7 для нового штамма? Как вы думаете, что придётся делать, если найденная последовательность отсутствует в новом штамме COVID?
12. Не срабатывает ли тест на грипп (2) – п. 8 ложно на штамм Дельта?
13. Найти минимальную специфичную последовательность для (1), отсутствующую в (3) и (2)
14. Найти минимальную специфичную последовательность для (3), отсутствующую в (1) и (2)
15. Найти максимальную общую подпоследовательность для (1) и (3), чтобы оценить отличие этих двух штаммов COVID по длине такой подпоследовательности к общей длине генома вируса.

Ожидаемый результат работы программы:

Вывести минимальную специфичную последовательность для (1), отсутствующую в (2);

Вывести минимальную специфичную последовательность для (2), отсутствующую в (1);

Вывести минимальную общую последовательность для (1) и (2);

Вывести минимальную специфичную последовательность для (1), отсутствующую в (3) и (2);

Вывести минимальную специфичную последовательность для (3), отсутствующую в (1) и (2);

Вывести максимальную общую подпоследовательность для (1) и (3) и соотношение длины этой последовательности к общей длине генома.

Варианты:

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | Lua | 6 | Java |
| 2 | JavaScript | 7 | Go |
| 3 | Python | 8 | Ruby |
| 4 | C(любой) | 9 | Haskell |
| 5 | PHP |