Зачетное задание по BLAST

**Дано: неаннотированный фрагмент генома бактерии *Paenibacillus larvae***

**Задача: определить, где в данном фрагменте закодированы белки, похожие на известные белки родственной бактерии (сенной палочки *Bacillus subtilis*).**

**Задание по этапам**

1. Напишите короткий план работы в файле XXXXXXX\_credit.doc: что, в каком порядке и какими программами воспользуетесь для решения поставленной задачи.
2. Создайте индексные файлы для поиска программами пакета BLAST (предварительно проверьте свою [квоту](http://kodomo.fbb.msu.ru/FBB/year_09/term3/quota.html) и удалите, если надо, ненужные файлы).
3. Создайте файл с трансляциями всех открытых рамок считывания длиной не менее 200 нуклеотидов из вашего фрагмента. Используйте программы пакета EMBOSS.
4. Создайте Excel файл с находками BLAST. Используйте табличную выдачу. Не забудьте включить названия всех столбцов. (Страница blast\_hits)
5. Создайте страницу orfs в файле Excel, включающую информацию обо всех открытых рамках считывания (>200 п.н.) в вашем фрагменте генома. Для каждой рамки должно быть указано: начало во фрагменте, конец во фрагменте, направление (прямое или обратное), число сходных последовательностей, найденных программой BLAST в протеоме *B. subtilis* при условии E-value<0,001. начало CDS, конец CDS предполагаемого белка.
6. В отчете XXXXXXX\_credt.doc приведите таблицу, содержащую эту информацию только для тех открытых рамок, для которых нашлась хотя бы одна сходная последовательность. Вместо границ открытой рамки приведите предполагаемые границы кодирующей последовательности (начало CDS может не совпадать с началом открытой рамки). Кроме того, в отчётной таблице должны присутствовать два дополнительных столбца, в которых приведены: идентификатор самого близкого из найденных белков *B. subtilis*, E-value находки, предполагаемая функция белка *Paenibacillus larvae*
7. Изобразите схематически положение на фрагменте тех открытых рамок, для которых нашлись сходные последовательности в *B. subtilis*.
8. \* Если в полученном наборе предполагаемых генов имеются аномалии (перекрывания генов), постарайтесь их объяснить. Предложите наиболее правдоподобную, по вашему мнению, структуру генов на данном участке генома.
9. \* Приведите выравнивания трансляций открытых рамок считывания и сходных последовательностей из протеома сенной палочки.

Ход работы:

Contig\_ID: AARF01000048(11500-18300)

1)Вырезать данный мне неаннотированный фрагмент генома *Paenibacillus larvae*

команда seqret -sask

julia270692@kodomo:~/Term3/Block2/Credit$ seqret -sask

Reads and writes (returns) sequences

Input (gapped) sequence(s): pla\_contigs.fasta:AARF01000048

 Begin at position [start]: 11500

 End at position [end]: 18300

 Reverse strand [N]: Y

output sequence(s) [aarf01000048.fasta]::

Полученный файл:aarf01000048.fasta

2)Необходимо создать индексные файлы для поиска программами пакета BLAST для поиска по протеому *Bacillus subtilis*

formatdb (-i, -p и –n)

formatdb -i bsu\_proteom.fasta -p T -n bsu

3) Для создания файла с трансляциями всех открытых рамок считывания длиной не менее 200 нуклеотидов из данного фрагмента, можно воспользоваться командой getorf( при этом использовать опции minsize,find,table.)

4)Для создания Exсel файла с находками BLAST необходимо воспользоваться командой blastall(конкретно программой blastp)

|  |  |  |  |
| --- | --- | --- | --- |
| BLASTP | Белок | Белки | Поиск гомологов |

(+параметры -d (базовое имя индексных файлов), -i (входной файл), -o (выходной файл))

blastall -p blastp -d bsu -i naraykina\_orf\_transl.fasta -o naraykina\_blast\_hits.xls -m 9

5) Создание страницы orfs c информацией обо всех открытых рамках считывания в моем фрагменте генома:

* Всего рамок 14
* Число рамок, для которых нашлась хотя бы одна находка:3
* Число сходных последовательностей, найденных программой BLAST в протеоме *B. Subtilis :*

blastall -p blastp -d bsu -i naraykina\_orf\_transl.fasta -m 9 -e 0.001 -o alig.txt

Количество находок: скрипт blastp\_grep.scr( активация -chmod +x blastp\_grep.scr

 ./blastp\_grep.scr)

Данные о функциях белков из UniProt.

6) Положение на фрагменте тех открытых рамок, для которых нашлись сходные последовательности в *B. Subtilis*:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Оттранслированная рамка | CDS, (направление у всех прямое) | наилучшая находка в *B. subtilis* | E-value находки |  Предполагаемая функция белка *Paenibacillus larvae* |
| >AARF01000048\_1 | [33-611]  | sp|P24012|COX3\_BACSU | 6e-54 | Оксидаза, окисление цитохрома с |
| >AARF01000048\_3 | [1075-1875] | sp|O34329|CTAG\_BACSU | 3e-38 | Мембранный белок |
| >AARF01000048\_6 | [3501-4391] | sp|C0SPB4|YHAQ\_BACSU | 1e-83 | АТФ-связывающий белок |
| >AARF01000048\_7 | [4387-5640] | sp|O07523|YHAP\_BACSU | 3e-61  | Мембранный белок |

# Гипотетические гены во фрагменте

# AARF01000048(11500-18300)

####  (генома бактерии *Paenibacillus larvae)*

 **[4378-5652,YHAP-BACSU]**

**3'--[3-617,СOX3\_BACSU]--[1066-1929,CTAG\_BACSU]---[3465-4391,YHAQ\_BACSU]----5'**

**5'-----------------------------------------------------------------------3'**

Как видно,два последних гена перекрываются, но область перекрывания очень маленькая. Поэтому,можно предоложить,что границы кодирующей последовательности белка YHAP\_BACSU немного другие, так как он мало изучен и неизвестна его точная функция.