Выравнивание белков BLAST

Выравнивание белков

Задание 1.

Откройте файл `CoV_task.fa` для выравнивания, перейдите на вкладку Translated Protein Sequence, чтобы получить последовательности белков.

- Align Protein

Запустите выравнивание с настройками по умолчанию и сохраните результат. Какая длина у получившегося выравнивания?

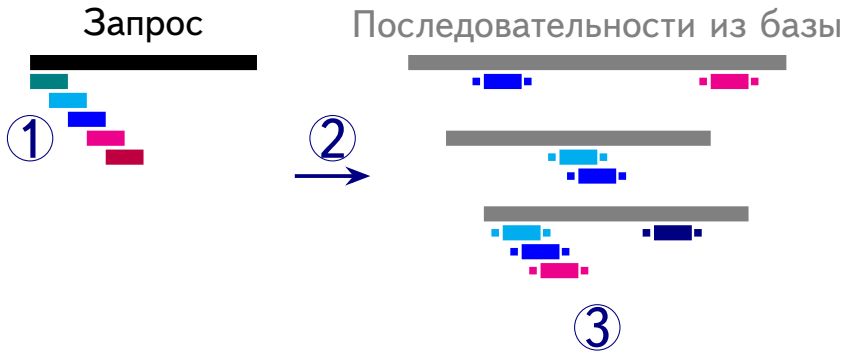
Настройки выравнивания

Option		Setting
GAP PENALTIES		
Gap Open	☑	*-2.90*
Gap Extend	☑	*0.00*
Hydrophobicity Multiplier	☑	*1.20*
MEMORY/ITERATIONS		
Max Memory in MB	☑	*2048*
Max Iterations	☑	*16*
ADVANCED OPTIONS		
Cluster Method (Iterations 1,2)	☑	*UPGMA*
Cluster Method (Other Iterations)	☑	*UPGMA*
Min Diag Length (Lambda)	☑	*24*

Задание 2. Сделайте обрезку полученного выравнивания с помощью Gblocks, какая длина у полученного выравнивания?

BLAST

Basic Local Alignment Search Tool



1. Последовательность запроса разбивается на «слова»
2. Поиск слов в индексированной базе
3. Расширение участков совпадений

Проблемы оценки качества

1. Вероятность случайно найти совпадение зависит от длин последовательностей.
Например, вероятность случайной встретить последовательность `ATGC`:

$$P = f_A \times f_T \times f_G \times f_C = 0.25^4$$

Длина сайта (п.н.)	Вероятность	*E.coli* 4.6 млн п.н.	*S.cerevisiae* 9 млн п.н.	*H.sapiens* 3.3 млрд п.н.
6	0.25^6	1123	2197	805664
8	0.25^8	70	137	50354
10	0.25^{10}	4	8	3147

Проблемы оценки качества

1. Вероятность случайно найти совпадение зависит от длин последовательностей.
Например, вероятность случайной встретить последовательность `ATGC`:

$$P = f_A \times f_T \times f_G \times f_C = 0.25^4$$

Длина сайта (п.н.)	Вероятность	*E.coli* 4.6 млн п.н.	*S.cerevisiae* 9 млн п.н.	*H.sapiens* 3.3 млрд п.н.
6	0.25^6	1123	2197	805664
8	0.25^8	70	137	50354
10	0.25^{10}	4	8	3147

2. Итоговая оценка качества выравнивания зависит от выбранной системы оценки.

Оценка выравниваний BLAST

1. Bit-score, S' — нормированная оценка качества выравнивания

$$S' = \frac{\lambda \times S - ln(K)}{ln(2)}$$

S — оценка полученного выравнивания; K, λ — параметры функции распределения вероятности получить конкретное значение S для заданных m и n (длины последовательностей)

2. E — оценка количества случайных выравниваний с индексом S'

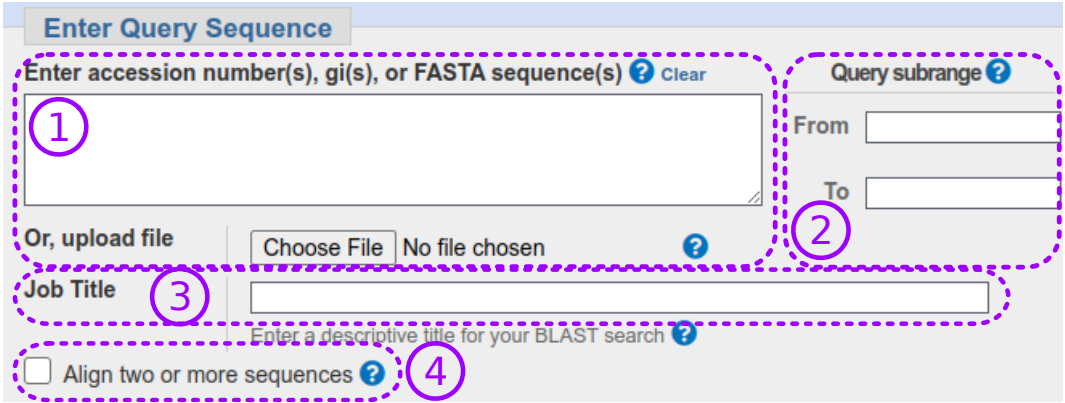
$$E = m \times n \times 2^{-S'}$$

Варианты BLAST

Web приложение

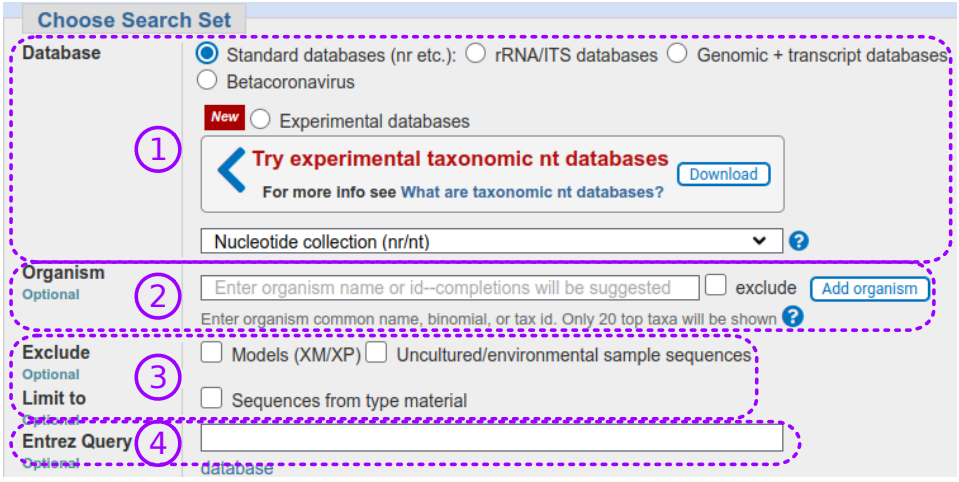
Вариант BLAST	Последовательность запроса	Последовательность в базе
blastn	Нуклеотидная	Нуклеотидная
blastp	Белковая	Белковая
blastx	Транслированная	Белковая
tblastn	Белковая	Транслированная
tblastx	Транслированная	Транслированная

Настройки BLASTn. Запрос



1. Поле для ввода последовательности или идентификатора
2. Ограничение фрагмента для запроса
3. Название
4. Выравнивание двух последовательностей

Настройки BLASTn. База для поиска



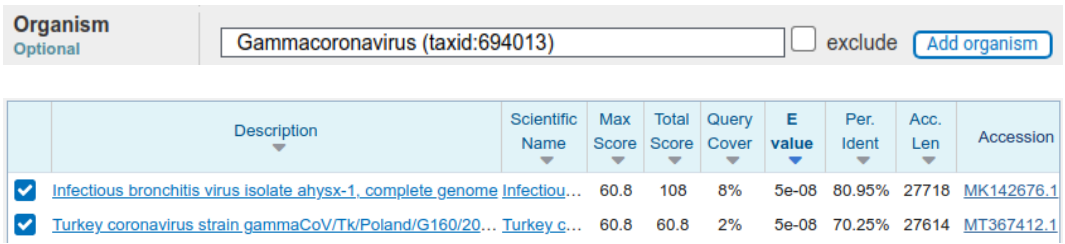
1. Выбор базы данных для поиска
2. Выбор организма
3. Исключить из поиска
4. Искать по запросу

Настройки BLASTn. Программа, алгоритм

- megablast
- discontigous megablast
- blastn

- Algorithm parameters
  - Максимальное количество результатов поиска
  - Пороговое значение для E-value
  - Размер слова
  - Максимальное количество совпадений на одном участке
  - Матрица
  - Поправка на состав
  - Фильтры

Задание 3. Проведите поиск последовательностей, сходных с CDS для белка S (первый в файле `CoV_task.fa`), среди рода Gammacoronavirus. Какой максимальный процент идентичности последовательностей?



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Infectious bronchitis virus isolate ahysx-1, complete genome	Infectiou…	60.8	108	8%	5e-08	80.95%	27718	MK142676.1
☑	Turkey coronavirus strain gammaCoV/Tk/Poland/G160/20…	Turkey c…	60.8	60.8	2%	5e-08	70.25%	27614	MT367412.1

Задание 4. Найдите потенциальные гомологи белка β-синуклеина в референсном протеоме человека. Последовательность белка можно найти в файле `b_syn.fa`. Какой вариант BLAST нужно использовать? Исключите из анализа «предсказанные» последовательности. Сколько потенциальных кандидатов вы получили в результате? Не закрывайте вкладку.

Задание 5. Найдите потенциальные гомологи белка β-синуклеина у других позвоночных. Исключите из анализа «предсказанные» последовательности. Сколько потенциальных кандидатов вы получили в результате?

Задание 6. Для белка кабана (*Sus scrofa*) с наилучшим совпадением с β-синуклеина человека проведите реципроктный BLAST против протеома человека. Выпишете ACESSION белка, который демонстрирует наилучшее совпадение в этом случае?