



Выравнивание белков BLAST

Выравнивание белков

Задание 1.

Откройте файл `CoV_task.fasta` 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?

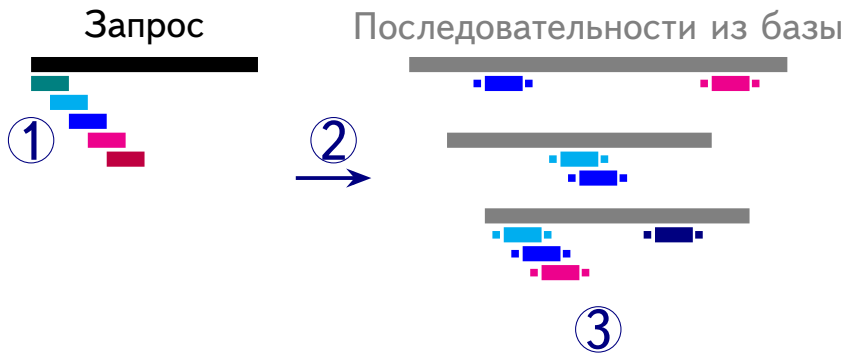
Настройки выравнивания

Option	Setting
GAP PENALTIES	
Gap Open	<input checked="" type="checkbox"/> -2.90
Gap Extend	<input checked="" type="checkbox"/> 0.00
Hydrophobicity Multiplier	<input checked="" type="checkbox"/> 1.20
MEMORY/ITERATIONS	
Max Memory in MB	<input checked="" type="checkbox"/> 2048
Max Iterations	<input checked="" type="checkbox"/> 16
ADVANCED OPTIONS	
Cluster Method (Iterations 1,2)	<input checked="" type="checkbox"/> UPGMA
Cluster Method (Other Iterations)	<input checked="" type="checkbox"/> UPGMA
Min Diag Length (Lambda)	<input checked="" type="checkbox"/> 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 индексируемой 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	<i>E.coli</i> 4.6 млн п.н.	<i>S.cerevisiae</i> 9 млн п.н.	<i>H.sapiens</i> 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	<i>E.coli</i> 4.6 млн п.н.	<i>S.cerevisiae</i> 9 млн п.н.	<i>H.sapiens</i> 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

The screenshot shows the BLASTn search interface with four numbered annotations in purple circles:

- 1**: Points to the large text input field for the query sequence, labeled "Enter accession number(s), gi(s), or FASTA sequence(s)".
- 2**: Points to the "Query subrange" section, which includes "From" and "To" input fields.
- 3**: Points to the "Job Title" input field.
- 4**: Points to the checkbox labeled "Align two or more sequences".

Other visible elements include a "Clear" button, a "Choose File" button, and a "No file chosen" status.

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

Настройки BLASTn. База для поиска

The screenshot shows the 'Choose Search Set' section of the BLASTn interface. It is divided into four numbered regions by dashed purple lines:

- 1. Database:** Includes radio buttons for 'Standard databases (nr etc.)', 'rRNA/ITS databases', 'Genomic + transcript databases', 'Betacoronavirus', and 'Experimental databases'. A red 'New' tag is next to 'Experimental databases'. Below is a button 'Try experimental taxonomic nt databases' with a 'Download' link and a link to 'What are taxonomic nt databases?'. A dropdown menu shows 'Nucleotide collection (nr/nt)'.
- 2. Organism:** Labeled 'Optional'. Includes a text input field 'Enter organism name or id--completions will be suggested', an 'exclude' checkbox, and an 'Add organism' button. Below is a note: 'Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown'.
- 3. Exclude:** Labeled 'Optional'. Includes checkboxes for 'Models (XM/XP)', 'Uncultured/environmental sample sequences', and 'Sequences from type material'.
- 4. Entrez Query:** Labeled 'Optional'. Includes a text input field for the query.

1. Выбор базы данных для поиска
2. Выбор организма

3. Исключить из поиска
4. Искать по запросу

Настройки BLASTn. Программа, алгоритм

- ▶ megablast
- ▶ discontinuous 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?

Organism
Optional

Gammacoronavirus (taxid:694013) ☐ exclude [Add organism](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Infectious bronchitis virus isolate ahysx-1, complete genome	Infectiou...	60.8	108	8%	5e-08	80.95%	27718	MK142676.1
<input checked="" type="checkbox"/>	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.fasta`. 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. Выпишите ACCESSION белка, который демонстрирует наилучшее совпадение в этом случае?