# Задания по темам «выравнивание» и «поиск»

Ответы сохраняйте в протоколе. Напишите заголовок: «2***. Выравнивание и поиск».***
Имена всех сохраняемых файлов начинайте с вашей фамилии латинскими буквами – так нам легче будет не запутаться при проверке. Ниже фамилия зашифрована XXXXXXX.

СРАВНЕНИЕ ГЕНОМОВ ВИРУСОВ С ПОМОЩЬЮ КАРТЫ ЛОКАЛЬНОГО СХОДСТВА

**1. Найдите и опишите в протоколе три пары вирусов, геномы которых находятся в следующих отношениях:**

 a) **(первая пара)** гомологичны почти по всей длине: сумма длин гомологичных участков больше половины каждого из геномов;

 b) **(вторая пара)** не имеют гомологичных участков (участки не достоверного сходства не в счёт);

 с) **(третья пара)** имеются сравнительно короткие гомологичные участки: суммарная длина гомологичных участков меньше половины каждого из геномов.

**Описание** должно включать

- семейство, название каждого вируса из пары и его хозяина;

- идентификаторы записей с геномом;

- длины геномов (в п.н.);

- суммарный процент гомологичных участков относительно длины одного из геномов;

- карту локального сходства геномов;

- характеристику сходства геномов – свободный текст (о длинах гомологичных участков, их достоверности – E-value, и типичном проценте совпадающих букв в выравниваниях гомологичных участков).

**Рекомендации**

- откройте страницу геномов: NCBI Genome (<http://www.ncbi.nlm.nih.gov/genome/>) => Viruses => Browse viral genomes by family

- в другой вкладке откройте страницу “BLAST двух последовательностей”, доступную со страницы BLAST на сайте NCBI <http://blast.ncbi.nlm.nih.gov/Blast.cgi> → Nucleotide BLAST → поставьте галочку в чекбоксе «Align two or more sequences».
Чтобы оперативно проверять сходство выбранных геномов:
 в окошках можно указать идентификаторы записей (типа NC\_123..) вместо самих последовательностей;

в меню “Optimize for” выбирайте “ Somewhat similar sequences (blastn)”;

дождавшись результата, смотрите на карту локального сходства (DotMatrix).

- выбирайте вирусы с нефрагментированным геномом длины меньше 20 000 п.н. (соблюдение этих условий не обязательно, но упрощает анализ);

- для a) выбирайте вирусы из одного семейства или даже рода;

- для b) выбирайте вирусы из разных семейств;

- для c) пробуйте разные варианты, используйте свои знания из вирусологии!

Наверное, придется перебрать несколько пар вирусов пока найдете подходящие!

ВЫРАВНИВАНИЕ ГОМОЛОГИЧНЫХ ГЕНОМОВ

**2. Постройте выравнивание полных геномов из задания 1а и укажите координаты гомологичных участков (номера позиций от – до).**

Результат - координаты в отчете и файл с последовательностями и их выравниванием в формате проекта JalView (XXXXXXX\_2genomes.jvp)

**Указания.**

- Сохраните обе последовательности в одном файле в формате fasta. Найдите каждую из последовательностей в GenBank (если не закрыли окно – щелканьем по идентификатору). Сохраните в fasta формате (send => to file, формат fasta); вторую добавьте в тот же файл.

- Откройте файл программой Jalview (File => Input from alignment)
- раскрасьте по нуклеотидам ( color => nucleotides) и

- установите раскраску только при 100% совпадении букв в колонке (color => Above identity threshold)

- выровняйте последовательности (Web service => alignment => Muscle (default) или любая др. программа выравнивания, которая не откажется работать)

- раскрасьте так же выровненные последовательности

- найдите достаточно длинные (более 20 столбцов) участки, на которых нет гомологичности; это либо участки с делециями/вставками, либо такие, на которых идентичных колонок меньше половины; занесите в протокол координаты гомологичных участков; не могу предвидеть, как будет выглядеть ваше выравнивание, поэтому проявляйте здравый смысл по мере необходимости))) Или спрашивайте преподавателей.

- (\*) для интереса найдите рамку считывания какого-нибудь гена по чередованию: две консервативные позиции, одна неконсервативная и т.д. (ясно, в чем дело?)

- coхраните результат как «проект JalView» в виде файла XXXXXXX\_2genomes.jvp (в самом большом окне File → Save Project → находите свою папку, в нижнем меню выбираете JalView, в верхнем окошке пишете имя файла).

ВЫРАВНИВАНИЕ КОДИРУЮЩИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ: ВЫРАВНИВАТЬ ГЕНЫ ИЛИ БЕЛКИ?

**3. Для пары энтеровирусов сравните процент гомологичных участков, определяемых выравниванием нуклеотидных последовательностей геномов (программа blastn) и формальных трансляций этих геномов в шести рамках (программа tblastx)**

**Указания.**

- используйте ту же страницу “BLAST двух последовательностей”, запомните процент гомологичных участков и сохраните карту в протоколе;

- вернитесь (Edit and resubmit) и во второй раз выберите в верхнем меню программу tblastx вместо blastn

- оформите протокол (две карты, два процента “покрытия” и комментарий)

МНОЖЕСТВЕННОЕ ВЫРАВНИВАНИЕ. РЕКОНСТРУКЦИЯ ФИЛОГЕНИИ.

**4. Постройте выравнивание последовательностей белков, отобранных вами на предыдущем занятии (задание II.5) и сохраните филогенетическое дерево.**

Результат — файл с выравниванием в формате проекта JalView (XXXXXXX\_homologs.jvp), файл с филогенетическим деревом в формате Newick (XXXXXXX.tre) и графический файл с изображением дерева (XXXXXXX.png).

**Указания.**

- используйте Jalview для выравнивания;
- раскрасьте по методу Clustalx и подберите порог “by conservation” так, чтобы лучше было видно сходство и различия между последовательностями;

- проверьте выравнивание! В нем могут оказаться последовательности негомологичных белков (что видно по малому числу совпадений) или короткие фрагменты белков. Удалите “плохие” последовательности из выравнивания;

– постройте филогенетическое дерево (Calculate → Calculate Tree → Average Distance Using % Identity).

 – сохраните дерево в формате Newick (имя файла XXXXXXX.tre где XXXXXXX – ваша фамилия латинскими буквами)

 – сохраните «Проект JalView» в виде файла XXXXXXX\_homologs.jvp (в самом большом окне File → Save Project → находите свою папку, в нижнем меню выбираете JalView, в верхнем окошке пишете имя файла).

 – Откройте дерево (файл с расширением tre) программой Mega. Придайте ему какую-нибудь красивую форму и сохраните изображение.

АННОТАЦИЯ ПО СХОДСТВУ

**5. Предскажите функцию белка g17, полученного в результате секвенирования и аннотации (поиска генов) паразита водорослей *Amoeboaphelidium protococcarum*.**

**Указания.**

- используйте поиск blastp в базе данных Refseq Proteins;

- выберите лучшую находку с хорошо аннотированным белком;

- опишите результат в протоколе (E-value, coverage %, identity %, функция, таксономия выбранной находки).

ПОИСК ОТДАЛЕННЫХ ГОМОЛОГОВ

**6. Найдите по четыре-пять гомологов белка integrase.fasta среди a) вирусов, b) бактерий, c) архей, d) эукариот.** Верно ли, что представители таксонов находятся на отдельных кладах филогенетического дерева, построенного по выравниванию последовательностей?

**Указания.**

- используйте поиск blastp в базе данных Refseq Proteins с ограничением таксона; каждый раз отметьте несколько находок галочкой и сохраните их в формате fasta (Download => Fasta (complete sequences) ),

- добавляйте скачанные последовательности в один файл, используя редактор (WordPad, например);

- откройте файл в JalView, выровняйте последовательности и постройте филогенетическое дерево;

- ответ на вопрос включите в протокол.

=================

(\*) ДОПОЛНИТЕЛЬНОЕ ЗАДАНИЕ

Определите доменную архитектуру какого-либо белка.

**Указание.**

Используйте БД Pfam и в ней поиск по последовательности.