**ПРАКТИЧЕСКОЕ ЗАНЯТИЕ** В КУРСЕ «Регуляторные геномные последовательности»

## **ТЕМА:** БАЗЫ ДАННЫХ ПО РЕГУЛЯЦИИ ТРАНСКРИПЦИИ и ПРОГРАММЫ РАСПОЗНАВАНИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКИПЦИОННЫХ ФАКТОРОВ в ДНК

(проводила Е.В. Игнатьева)

В ходе практического занятия нужно выполнить приведенные ниже задания, вставить ответы в пустые графы таблиц и выслать эту инструкцию с заполненными таблицами на адрес **eignat@bionet.nsc.ru** в приложении к письму в виде файла в MS word. В письме указать фамилию!! (Бывает, что приходят письма без подписи, а по эл. адресу определить студента бывает трудно, например knowledge75@mail.ru)

**Часть 1**

**База данных Transcription Regulatory Regions Database (TRRD) и поисковая система SRS (Sequence Retrieval System):**

**Цель**: Знакомство с базой данных Transcription Regulatory Regions Database (TRRD <http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/>), которая разработана в ИЦиГ СО РАН и формирование поисковых запросов к базе TRRD с использованием системы SRS (Sequence Retrieval System)

|  |
| --- |
| **Для справки**  SRS – сокращенное название программного средства Sequence Retrieval System, которое обеспечивает хранение, интеграцию и доступ к базам данных. SRS позволяет выполнять простые и сложные запросы к базам данных через Интернет. SRS была использована для доступа к информации, хранящейся в информационных ресурсах портала ExPASy (SIB Bioinformatics Resource Portal, https://www.expasy.org/), где были интегрированы такие базы как UniProtKB/Swiss-Prot, TREMBL, TREMBL\_NEW, и т.д. ([**https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Protein/protein10.html**](https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Protein/protein10.html)).  В настоящее время поисковая система SRS используется как средство доступа к базам данных, разработанным в ИЦиГ СО РАН (<http://srs6.bionet.nsc.ru/srs6/>) |

Посмотрите на рисунок 1 (на стр.6), отображающий доступ к информационной таблице базы данных TRRD через систему SRS. Обратите внимание, что в интерфейсе есть возможность проводить как поиск, так и пролистывание данных.

Перейдите на главную страницу базы TRRD <http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/>

Доступ к поисковой системе **SRS можно получить через гиперссылки: TRRDGENES, TRRDEXP, TRRDSITES, TRRDFACTORS, TRRDBIB, TRRDUNITS, TRRDLCR, TRRDSTARTS. При этом каждая гиперссылка обеспечит доступ к одной конкретной таблице (части = набору полей) базы TRRD**

**Знакомство с функциональной возможностью системы SRS, позволяющей просматривать содержимое полей.**

|  |  |
| --- | --- |
| Задание 1. Просмотрите содержание поля GeneName\_Full (TRRDGENES). Сколько входов базы содержат название генов:  « 2s albumin» ?  « 2'-5' oligoadenylate synthetase» ?? | Ответ: |

Пояснение: Например, чтобы просмотреть содержимое поля GeneName\_Full (то есть пролистать его содержание в алфавитном порядке) необходимо, на странице <http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/> в разделе “SRS access” выбрать опцию TRRDGENES. В таблице, где перечислены все информационные поля из раздела TRRDGENES, кликнуть мышью на название поля «GeneName\_Full». Далее нажать клавишу «List Values» (при этом в окошке для ввода поискового слова должна стоять звездочка).

**Формирование запросов с использованием стандартной формы запроса системы SRS.**

Переход к стандартной форме запроса по каждой конкретной таблице возможен через кнопку “Search”, расположенную справа вверху на каждой из страниц с заголовками: TRRDGENES, TRRDEXP, TRRDSITES, TRRDFACTORS, TRRDBIB, TRRDUNITS, TRRDLCR, TRRDSTARTS (см. Рисунок 1)

Для получения адекватного результата необходимо знать точное название поля, по которому предполагается вести поиск. Определиться с названием поля можно путем просмотра таблиц с названиями полей и их содержания. В связи со сложной эпидемической обстановкой в этом году таблица в задании почти полностью заполнена ☺ . Осталось только заполнить количество входов, которое будет получено в каждом случае.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Задание 2. Сформировать простые запросы к базе TRRD: | | | | | |
| N | смысл поиска | название таблицы для поиска | название поля для поиска | ключевое слово, вводимое в окошко | Сколько входов получено |
|  | Поиск генов, названия которых включают слово «apolipoprotein» | TRRDGENES | GeneName | Apolipoprotein\* |  |
|  | Поиск генов человека (навание вида – homo sapiens, англ. название - human) | TRRDGENES | Species | Human  homo sapiens |  |
|  | Сколько в базе TRRD имеется паттернов экспрессии, описывающих экспрессию в сердце (heart) ? | TRRDEXP | Organ | Heart |  |
|  | Сколько в базе TRRD имеется паттернов экспрессии, в которых упоминается индуктор (или репрессор) глюкокортикоиды (glucocorticoids) ? | TRRDEXP | IndReprName | glucocorticoids |  |
|  | Сколько в базе TRRD имеется сайтов связывания фактора CTCF? | TRRDSITES | FactorName | CTCF |  |
|  | Сколько входов в части базы TRRDUNITS содержат описание энхансеров (enhancer)? | TRRDUNITS | RegUnit | enhancer |  |

|  |  |
| --- | --- |
| Задание 3  При поиске по полю Species в таблице TRRDGENES4 запрос по ключевому слову «human» и «human со звездочкой» дает разное количество результатов: 763 и 785 (соответственно). Запрос по ключевому слову «Homo sapiens» дает возможность получить 763 входа. Так все-таки, какое количество генов человека содержится в базе?  Пролистайте записи в поле *Species* **(кнопка “List Values”)** и приведите пример входов, содержащих слово «human», но к человеку не относящихся? |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Задание 4. Сформировать сложные запросы к базе TRRD (соединены союзом) | | | | | |
| N | смысл поиска | название таблицы для поиска | название поля для поиска | ключевое слово, вводимое в окошко | Сколько входов получено |
|  | Поиск сайтов с названием ТАТА box  (учтите возможные варианты написания «TATA-box», «TATA box», «TATAbox») | TRRDSITES | SiteName | TATA-box | TATA box | TATAbox |  |
|  | Поиск сайтов связывания транскрипционного фактора SF1  (учтите возможные варианты написания «SF-1», «SF1», «SF 1») | TRRDSITES | SiteName | SF-1 | SF1 | SF 1 |  |
|  | Сколько записей в таблице TRRDEXP говорят о том, что экспрессия гена индуцирована инсулином? | TRRDEXP | IndReprName  AND  Influence | insulin  Induction |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Задание 5 Сформировать запросы к базе TRRD с помощью линковки между таблицами.  Поиск выполняется в два этапа. Первый запрос идет к таблице TRRDSITES. Затем необходимо выполнить переход между таблицами (линковку), как это показано на Рис.2 | | | | | | | |
| N | Смысл поиска | название таблицы для поиска | название поля для поиска | ключевое слово, вводимое в окошко | С какой таблицей базы TRRD осуществляется линковка | Вариант линковки | Сколько входов получено |
|  | Поиск **генов**, в регуляторных районах которых имеется сайт связывания CTCF | TRRDSITES | SiteName | CTCF | TRRDGENES | Первый вариант: «in the selected databanks which are linked to the current query» |  |
|  | Поиск **генов**, в регуляторных районах которых имеется сайт связывания SF1 | TRRDSITES | SiteName | SF-1 | SF1 | SF 1 | TRRDGENES | Первый вариант: «in the selected databanks which are linked to the current query» |  |

|  |
| --- |
|  |
| Рисунок 1. Доступ к информационной таблице базы данных TRRD через систему SRS. В нижней части экрана видно начало таблицы, включающей названия информационных полей (первый и второй столбцы таблицы) и количества записей в каждом поле ( четвертый и пятый столбцы таблицы) . |

|  |
| --- |
|  |
| Рисунок 2. Один из этапов реализации сложных запросов через систему SRS с использованием связей между таблицами. В рассматриваемом примере осуществляется поиск генов, в регуляторных районах которых имеется сайт связывания CTCF. Слева отображен результат поиска с использованием ключевого слова «CTCF» по полю «FactorName» таблицы TRRSITES. После выбора опции «Link» осуществляется переход на страницу (на рисунке справа), где пользователь выбирает информационную таблицу, с которой необходимо определить связь (в данном случае выбрана таблица TRRDGENES), а также способ линковки с выбранной таблицей. |

**Часть 2**

**Сравнение информационного содержания двух баз**

В ходе выполнения предыдущего задания (по базе TRRD) было определено количество генов, в регуляторных районах которых имеются сайты связывания транскрипционных факторов CTCF

Перейдите на страницу базы TRRUST (<https://www.grnpedia.org/trrust/>). Эта база данных, представленная в журнале Nucleic Acids Research в 2017 году. TRRUST создана на основе поиска данных в публикациях с помощью Text mining (автоматический анализ текстов). Перейдите на страницу “SEARCH” (<https://www.grnpedia.org/trrust/Network_search_form.php>) и сделайте запрос по названию транскрипционного фактора CTCF. Определите количество генов человека и мыши, регулируемых этим фактором. Результат занесите в таблицу.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | База TRRD (цифра из предыдущего задания) | База TRRUST | | |
| Человек | Мышь | Всего в базе TRRUST |
| Количество генов, регулируемых фактором CTCF |  |  |  |  |

**Часть 3**

**Анализ последовательности ДНК программой распознавания сайтов MoLoTool** (http://molotool.autosome.ru/)

1. Подготовка последовательности ДНК в фаста формате;
2. Поиск в данной последовательности потенциальных сайтов связывания транскрипционных факторов CTCF с помощью программы **MoLoTool**; (http://molotool.autosome.ru/)
3. Сопоставление результатов, полученных при использовании программы MoLoTool для сайта CTCF при разных параметрах распознавания

**ЭТАП 1**

Нужно обратиться к базе EPD (<https://epd.epfl.ch//index.php>), выбрать вид организма H. Sapiens. В окне для поиска ввести название гена APP (amyloid beta precursor protein). Получится таблица, верхние три строчки которой содержат промоторы гена APP. Нужно получить последовательность ДНК первого промотора с идентификатором APP\_1. Для этого надо воспользоваться командой Download as: FASTA from -499 to 100 (эти позиции выставлены по умолчанию)

Для контроля Ваших действий, нужная ДНК-последовательность приведена ниже:

>FP027464 APP\_1 :+U EU:NC; range -499 to 100.

CCTTTAAGCTTCACTCGTTCTCATTCTCTTCCAGAAACGCCTGCCCCACCTCTCCAAACC

GAGAGAAAAAACGAAATGCGGATAAAAACGCACCCTAGCAGCAGTCCTTTATACGACACC

CCCGGGAGGCCTGCGGGGTCGGATGATTCAAGCTCACGGGGACGAGCAGGAGCGCTCTCG

ACTTTTCTAGAGCCTCAGCGTCCTAGGACTCACCTTTCCCTGATCCTGCACCGTCCCTCT

CCTGGCCCCAGACTCTCCCTCCCACTGTTCACGAAGCCCAGGTGGGCCGTCGGCCGGGGA

GCGGAGGGGGCGCGTGGGGTGCAGGCGGCGCCAAGGGCGCGTGCACCTGTGGGCGCGGGG

CGCGAGGGCCCCTCCCGGCGCGAGCGGGCGCAGTTCCCCGGCGGCGCCGCTAGGGGTCTC

TCTCGGGTGCCGAGCGGGGTGGGCCGGATCAGCTGACTCGCCTGGCTCTGAGCCCCGCCG

CCGCGCTCGGGCTCCGTCAGTTTCCTCGGCAGCGGTAGGCGAGAGCACGCGGAGGAGCGT

GCGCGGGGGCCCCGGGAGACGGCGGCGGTGGCGGCGCGGGCAGAGCAAGGACGCGGCGGA

**ЭТАП 2**

Необходимо обратиться к программе распознавания сайтов связывания транскрипционных факторов MoLoTool (http://molotool.autosome.ru/)

В окошко с надписью «Insert nucleotide sequence in fasta format here» вставить подготовленную ранее нуклеотидную последовательность промотора APP\_1 в фаста-формате.

Выбрать режим работы программы, как это показано на рисунке 2:

(1) выбрать тип сайта связывания, который будет предсказываться (CTCF\_HUMAN H11MO.0.A);

(2) выбрать параметр распознавания **P-value =5.00e-4**

|  |  |
| --- | --- |
|  | Рисунок 2. Интерфейс программы MoLoTool . Введена последовательность ДНК, выбран конкретный тип сайта связывания, который будет предсказываться (CTCF\_HUMAN H11MO.0.A). Выбран параметр распознавания P-value =5.00e-4[[1]](#footnote-1) |

Нажать кнопку «Submit», ознакомиться с результатом распознавания.

Нажать кнопку «Exel», результат будет получен в виде таблицы.

**Заполните таблицу:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Идентификатор проанализированной последовательности ДНК | ДНК-последовательность предсказанного сайта связывания ТФ | P-value предсказанного сайта | Позиция первого нуклеотида предсказанного сайта | Позиция последнего нуклеотида предсказанного сайта | Цепь ДНК | Название ТФ (Идентификатор базы Uniprot) |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Проведите анализ этой же последовательности повторно, но теперь используйте параметр **P-value =1.00e-4**.

**ЭТАП 3**

З**аполните таблицу:**

|  |  |
| --- | --- |
| Параметр распознавания (P-value) | Количество распознанных сайтов связывания фактора CTCF |
| 5.00e-4 |  |
| 1.00e-4 |  |

1. Параметр 5.00e-4 рекомендован разработчиками программы (см. вкладку “Guide” . Там написано «The recommended permissive P-value = 0.0005 (approximately 1 expected prediction in a stretch of random 1000 nucleotides of double-stranded DNA).» Это означает, что при таком параметре в случайной двухцепочечной последовательности ДНК длиной 1000 нуклеотидов будет распознан примерно 1 сайт связывания данного типа) [↑](#footnote-ref-1)