**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)

|  |
| --- |
| **Часть 1**  **Знакомство с базой данных Transcription Regulatory Regions Database (TRRD) и поисковой системой SRS (Sequence Retrieval System):**  1) Знакомство с базой данных Transcription Regulatory Regions Database (TRRD <http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/>), которая разработана в ИЦиГ Со РАН.  2) Знакомство с принципами работы поисковой системы SRS (Sequence Retrieval System) на примере базы данных TRRD.  3) Знакомство с браузерами базы TRRD.  4) Выполнение самостоятельной работы по составлению простых и сложных запросов к базе TRRD через систему SRS. |

|  |
| --- |
| **Пояснение - Что такое SRS ?**  **https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Protein/protein10.html**  SRS (Sequence Retrieval System) – программное средство хранения, интеграции и доступа к базам данных. SRS позволяет выполнять простые и сложные запросы к базам данных через Интернет. SRS используется для доступа к информации хранящейся в информационных ресурсах портала ExPASy (SIB Bioinformatics Resource Portal, https://www.expasy.org/), где интегрированы такие базы как UniProtKB/Swiss-Prot, TREMBL, TREMBL\_NEW, и т.д.  Поисковая система SRS используется как средство доступа к базам данных, разработанным в ИЦиГ СО РАН ( <http://srs6.bionet.nsc.ru/srs6/> ) |

**Задание**

**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).

**Сравнение способа индексации в информационных полях базы.**

Пролистайте содержимое полей “StageOrgDev” и «InductionTime» в таблице базы TRRDEXP4. Обратите внимание на то, что в поле “StageOrgDev” вся запись воспринимается поисковой программой, как единое целое (например, 1 day-old), а в поле «InductionTime» числа (например, 1, 2, 10) представлены в отдельных строчках, а единицы измерения «days», «months» (и т.д.) в других отдельных строчках. Исключения составляют случаи, когда запись была сделана без пробела (например, 10-11days). Способ индексации необходимо учитывать при поиске, поскольку добавление звездочки перед (либо после) поискового термина может изменить количество данных, полученных на выходе. Добавление звездочки осуществляется с помощью команды «append wildcards to words» либо вручную.

|  |  |
| --- | --- |
| Задание 2. Сколько записей в таблице TRRDEXP4, содержащих ключевые слова:  А) С помощью опции «List Values» в информационном поле «Organ» (Таблица TRRDEXP4 ) определите, сколько входов(записей) в базе содержат ключевые слова:  “ adrenal cortex ”  “ adrenal gland ”  “ adrenal medulla ”  “ adrenal primordium ”  **Для выполнения следующих заданий нужно пользоваться стандартной поисковой формой системы SRS , доступ к которой осуществляется из таблицы TRRDEXP4 через кнопочку справа с надписью “Search”**  Б) Сколько входов получиться при выполнении поиска по полю Organ (Таблица TRRDEXP4 ) и ключевому слову “ adrenal ” со звездочкой и без звездочки после слова.  В) Добавление звездочки в задании «Б)» повлияло на результат. Такой эффект наблюдается при определенном способе индексирования.  Проверьте, наблюдаются ли различия в количестве входов  - при поиске по полю StageOrgDev (Таблица TRRDEXP4 ) по ключевому слову  “ 11 ” и «11 со звездочкой» (11\*)  - при поиске по полю InductionTime (Таблица TRRDEXP4 ) по ключевому слову  “ 11 ” и «11 со звездочкой» (11\*) | А) Ответ:  Б) Ответ:  В) Ответ:  Количество входов, полученных при запросе по информационному полю StageOrgDev  « 11 » - \_\_\_\_\_\_\_\_\_\_  «11 со звездочкой» - \_\_\_\_\_\_\_\_\_\_\_  Количество входов, полученных при запросе по информационному полю InductionTime  « 11 » - \_\_\_\_\_\_\_\_\_\_  «11 со звездочкой» - \_\_\_\_\_\_\_\_\_\_\_  Вывод: результат зависит от добавления звездочки при поиске по полю \_\_\_\_\_\_\_\_\_\_\_\_ и не зависит при поиске по полю \_\_\_\_\_\_\_\_\_\_\_\_ |

**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 необходимо путем просмотра таблиц с названиями полей и их содержания.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Задание 3. 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» |  |  |  |  |
|  | Поиск генов семейства глобинов «globin» |  |  |  |  |
|  | Поиск генов человека (навание вида – homo sapiens, англ. название - human) |  |  |  |  |
|  | Сколько в базе TRRD имеется паттернов экспрессии, описывающих экспрессию в сердце (heart) |  |  |  |  |
|  | Сколько в базе TRRD имеется паттернов экспрессии, в которых упоминается индуктор (или репрессор) глюкокортикоиды (glucocorticoids) |  |  |  |  |
|  | Сколько входов в части базы TRRDUNITS содержат описание энхансеров (enhancer)? |  |  |  |  |

|  |  |
| --- | --- |
| Задание 4  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. Какое количество генов человека содержится в базе? Пролистайте записи в поле *Species* и приведите пример входов, содержащих слово «human», но к человеку не относящихся? |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Задание 5. 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 «ТАТА-box», «ТАТА box», «ТАТАbox») |  |  |  |  |
|  | Поиск сайтов связывания транскрипционного фактора HNF4  (учтите возможные варианты написания «HNF-4», «HNF 4», «HNF4») |  |  |  |  |
|  | Поиск генов человека из семейства глобинов (globin) |  |  |  |  |
|  | Сколько записей в таблице TRRDEXP содержат в поле IndReprName ключевые слова «insulin» либо «glucocorticoids» |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Задание 6 Сформировать сложные запросы к базе TRRD (линковка между таблицами) | | | | | |
| N | Смысл поиска | название таблицы для поиска | название поля для поиска | ключевое слово, вводимое в окошко | Сколько входов получено |
|  | Поиск **генов**, экспрессия которых описана в сердце (heart) |  |  |  |  |
|  | Поиск генов, у которых есть энхансер (enhancer) |  |  |  |  |

Для выполнения сложного запроса с использованием линковки между таблицами необходимо сделать следующее (см. Рисунок 2):

1) после получения ответа на запрос по первой таблице воспользоваться опцией «Link» (слева на желтом фоне);

2) раскрыть список таблиц базы TRRD (знак + слева от названия базы);

3) выбрать нужную таблицу, с которой будет осуществлена линковка (например TRRDGENES);

4) выбрать способ линковки (in the selected databanks which are linked to the current query);

5) запустить поиск командой “Submit Link”

|  |
| --- |
|  |
| 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, экспрессия которых описана в легких (lung) . Слева отображен результат поиска с использованием ключевого слова «lung» по полю «Organ» таблицы TRRDEXP. После выбора опции «Link» осуществляется переход на страницу (на рисунке справа), где пользователь задает информационную таблицу, с которой необходимо определить связь (в данном случае выбрана таблица TRRDGENES), а также способ линковки с выбранной таблицей. |

|  |
| --- |
| **Часть 2**  **Анализ последовательности ДНК программой распознавания сайтов базы CIS-BP**  (http://cisbp.ccbr.utoronto.ca/TFTools.php)   1. Подготовка последовательности ДНК в нужном формате; 2. Выполнение анализа программой базы CIS-BP; 3. Сопоставление результатов, полученных при использовании разных пороговых значениях функции распознавания |

**ЭТАП 1**

Взять из базы GenBank карточку нуклеотидной последовательности цыпленка «Gallus gallus HS4 insulator, partial sequence» (ACCESSION=U78775)

https://www.ncbi.nlm.nih.gov/nuccore/U78775.2

Как обозначено в карточке (поле TITLE), функция этой нуклеотидной последовательности: «A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila» То есть, это последовательность локус-контролирующего района. Указывается, что последовательность имеет инсуляторную функцию.

Для выполнения анализа, нужно подготовить эту последовательность в фаста-формате:

>U78775.2

gagctcacggggacagcccccccccaaagcccccagggatgtaattacgtccctcccccgctagggggcagcagcgagccgcccggggctccgctccggtccggcgctccccccgcatccccgagccggcagcgtgcggggacagcccgggcacggggaaggtggcacgggatcgctttcctctgaacgcttctcgctgctctttgagcctgcagacacctggggggatacggggaaaaagctttaggctgaaagagagatttagaatgacagaatcatagaacggcctgggttgcaaaggagcacagtgctcatccagatccaaccccctgctatgtgcagggtcatcaaccagcagcccaggctgcccagagccacatccagcctggccttgaatgcctgcagggatggggcatccacagcctccttgggcaacctgttcagtgcgtcaccaccctctgggggaaaaactgcctcctcatatccaacccaaacctcccctgtctcagtgtaaagccattcccccttgtcctatcaagggggagtttgctgtgacattgttggtctggggtgacacatgtttgccaattcagtgcatcacggagaggcagatcttggggataaggaagtgcaggacagcatggacgtgggacatgcaggtgttgagggctctgggacactctccaagtcacagcgttcagaacagccttaaggataagaagataggatagaaggacaaagagcaagttaaaacccagcatggagaggagcacaaaaaggccacagacactgctggtccctgtgtctgagcctgcatgtttgatggtgtctggatgcaagcagaaggggtggaagagcttgcctggagagatacagctgggtcagtaggactgggacaggcagctggagaattgccatgtagatgttcatacaatcgtcaaatcatgaaggctggaaaagccctccaagatccccaagaccaaccccaacccacccaccgtgcccactggccatgtccctcagtgccacatccccacagttcttcatcacctccagggacggtgacccccccacctccgtgggcagctgtgccactgcagcaccgctctttggagaaggtaaatcttgctaaatccagcccgaccctcccctggcacaacgtaaggccattatctctcatccaactccaggacggagtcagtgagaatatt

**ЭТАП 2**

Обратиться к программе распознавания сайтов базы CIS-BP

Адрес базы CIS-BP: http://cisbp.ccbr.utoronto.ca/index.php

Адрес программы распознавания сайтов (вкладка TOOLS): http://cisbp.ccbr.utoronto.ca/TFTools.php

В верхнее окошко вставить нуклеотидную последовательность в фаста-формате. Выбрать режим работы программы, как это показано на рисунке ниже[[1]](#footnote-1).

Выбрать вид организма (Gallus\_gallus).

Выбрать модель мотива (PWM – LogOdds)[[2]](#footnote-2). Выбрать порог распознавания (Threshold) =16

|  |
| --- |
|  |
| Рисунок 3. Интерфейс программы распознавания сайтов базы CIS-BP |

Запустить программу (красная кнопка внизу, на рисунке не показана).

Дождаться результата - программа считает долго :-((( .

Занести результат в таблицу (см. ниже)

Повторить анализ с более мягкими порогами распознавания (Threshold=15 и Threshold=12)

Результаты также внести в таблицу.

|  |  |  |  |
| --- | --- | --- | --- |
| Порог распознавания (Threshold) | Количество предсказанных сайтов связывания | Названия транскрипционных факторов | Скоры предсказанных сайтов (Колонка Score) |
| 16 |  |  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| 15 |  | Здесь ничего писать не надо | Здесь ничего писать не надо |
| 12 |  | Здесь ничего писать не надо | Здесь ничего писать не надо |

**ВОПРОС:** Был сайт CTCF[[3]](#footnote-3) среди предсказанных сайтов связывания ????

ДА НЕТ

(нужное подчеркнуть)

1. В работе предложено выбрать именно эти параметры в качестве примера. Для получения публикабельных результатов необходимо протестировать все режимы работы и выбрать оптимальный для конкретного исследования. [↑](#footnote-ref-1)
2. PWM - position weight matrix [↑](#footnote-ref-2)
3. CTCF -(CCCTC-binding factor), является маркером участков ДНК, разделяющих активный и репрессированный хроматин. Функция инсуляторов реализуется благодаря их взаимодействию с этим фактором. [↑](#footnote-ref-3)