

Задание №1

Поиск регуляторных мотивов транскрипции в бактериальных последовательностях

В первом задании Вам необходимо найти регуляторный мотив (набор сайтов) в полученных последовательностях с помощью программы MEME.

В этом файле после задания №1 и инструкций к программе MEME приведены последовательности перед генами, экспрессия которых регулируется пуриновым репрессором **PurR**. Экспериментально установленные сайты связывания белка **PurR** выделены в них синим цветом. Задача состоит в том, чтобы определить, при каких длинах последовательностей и каком числе лишних (*то есть не содержащих сайта*) последовательностей каждая программа способна находить сайты, совпадающие с экспериментальными. Поэтому с помощью двух упомянутых выше программ Вам надо найти регуляторный мотив длиной 16 нуклеотидов.

Каждому будет выдан текстовый файл с последовательностями в FASTA-формате.

Что представляет собой FASTA-формат:

FASTA-формат – это определенная форма записи последовательностей, с которой работает большая часть программ для анализа геномных последовательностей.

В первой строке должно стоять название последовательности после знака “>”. Начиная со следующей строки приводится сама последовательность. Следующие друг за другом разные последовательности должны быть разделены пустой строкой. Ниже приводится пример записи нескольких последовательностей в FASTA-формате:

```
>guaB
acctgtcccattctcatgctcaagcagcagacgaaccgtttgattcaggcgactaacggtaaaaattgcaggggattgagaa
ggtaacatgtgagcagatcaaattctaaatcagcaggttattcagtcgatagtaacccgcctt

>glnB
gggtgaaaaatacggcgctgccaacctttgttgaggcacgtaatcagtttgaactcaactatttgcgtaagctgctgcaaat
caccaaaaggcaacgtcacccacgcggcgagaatggcgggcgcaaccggacagaa

>purL
attctctgtgtcgtgcgctcccagcttgaaaaacgtaataatagtgaaaggtttactcataaatgagcggcattttgcg
taaacctgcccagatggcaacttattacagccattggcggcacgcgcttgctaattcacga
```

Часть выданных Вам последовательностей не содержит сайтов. Поэтому не удивляйтесь, если сайты будут найдены не во всех последовательностях. Сайт считается совпадающим с экспериментальным, если он пересекается с ним на 8 или более нуклеотидов.

Ответ на задание следует представить в виде файла в формате *.doc с размеченными последовательностями. Для этого

- скопируйте из текстового файла в Word только те последовательности, в которых были найдены сайты. Последовательности должны быть скопированы *полностью*.
- выделите **синим экспериментально установленные** сайты.
- сайты, найденные с помощью программы MEME с параметром «**One per sequence**» (см. в инструкции) должны быть выделены *курсивом*
- сайты, найденные с помощью программы MEME с параметром «**Zero or one per sequence**» (также см. в инструкции) должны быть выделены **жирным шрифтом**
- все сайты (и экспериментальные, и предсказанные) должны быть **на сером фоне**

То есть ответ должен выглядеть так:

>prsA

ttcagcaatgattgcgaggttatc**gcaagaaaacgttttcgcgagg**ttgatgcggtgctttcctggctggt
agaatacgcgcccgctcgcgccctgactgggacaggggctgtgtctttgctgaatttga

В данном случае:

aagaaaacgttttcgc

aaacgttttcgcgagg

- экспериментально установленный сайт связывания **PurR**
- сайт, найденный с помощью программы MEME с параметром «**One per sequence**»
- сайт, найденный с помощью программы MEME с параметром «**Zero or one per sequence**»

gcaagaaaacgttttc

Все выше перечисленные предсказанные сайты считаются совпадающим с экспериментальным, поскольку пересекаются с ним более чем наполовину.

В отчете следует указать длину и количество последовательностей. Желательно также привести Ваши выводы о работе программы MEME и возможности ее применения для предсказания и поиска регуляторных сайтов в бактериальных геномах.

Инструкция по использованию программой MEME

On-line версия программы MEME находится по адресу <http://meme.sdsc.edu/meme/meme.html>

Окно программы содержит следующие поля:

Your e-mail address: Адрес Вашей электронной почты, на который будут высланы результаты.

Re-enter e-mail address: То же самое, еще раз (*это сделано для того, чтобы предотвратить введение адреса с ошибками*)

Ввести последовательности, в которых будет производиться поиск сайтов, возможно двумя способами:

- **Enter the name of a file containing the sequences here:** нужно сослаться на файл, содержащий последовательности в FASTA-формате. Удобнее воспользоваться кнопкой “Обзор...” (*в некоторых браузерах эта кнопка называется “Browse...”*)
- **The actual sequences here (Sample Input Sequences):** в окно вводятся непосредственно сами последовательности, тоже в FASTA-формате.

Description of your sequences: описание вводимых последовательностей, данное поле не обязательно для заполнения.

How do you think the occurrences of a single motif are distributed among the sequences?

– необходимо пометить, сколько сайтов одного типа Вы рассчитываете найти в каждой последовательности.

- **One per sequence** – по одному сайту в каждой последовательности;
- **Zero or one per sequence** – найти в каждой последовательности по одному сайту или вообще не найти в ней сайтов;
- **Any number of repetitions** – позволяет найти в каждой последовательности несколько сайтов одного типа.

*При выполнении задания необходимо произвести поиск два раза: в первый раз установив **One per sequence**, во второй – **Zero or one per sequence**.*

MEME will find the optimum width of each motif within the limits you specify here:

длина искомого сайта, необходимо задать минимальную и максимальную длину в располагающихся ниже окнах.

*При выполнении задания надо установить и минимальную, и максимальную длину сайта **16** нуклеотидов.*

Maximum number of motifs to find: количество различных типов мотив, которые предполагается найти.

При выполнении задания необходимо указать значение 1, то есть искать мотивы только одного типа.

MEME will find the optimum number of sites for each motif within the limits you specify here:

- количество сайтов каждого типа, которое предполагается найти во всей обучающей выборке. Данная функция имеет смысл лишь в том случае, если предполагается найти более 1 сайта в каждой последовательности. Поэтому, при выполнении задания поля **Minimum sites** (≥ 2) и **Maximum sites** (≤ 300) следует оставить пустыми.

Следующие четыре функции не потребуются для выполнения задания и поэтому возле них **не должно** стоять галочек:

- **Text output format** – формат, в котором будут представлены результаты: по умолчанию результаты будут оформлены в виде гипертекста (HTML формат).
- **Shuffle sequence letters** – осуществляет перетасовку букв в последовательности.
- **Search given strand only** – поиск сайта осуществляется только в приведенной последовательности, по умолчанию программа ищет сайт как в приведенной последовательности, так и в комплементарной.
- **Look for palindromes only** – осуществляется поиск только палиндромных сайтов.

После того, как все необходимые поля заполнены, нажимайте “**Start search**”
(в некоторых браузерах эта кнопка называется “**Submit Query**”)

В обновленном окне браузера при этом откроется страница **PENDING**, откуда по ссылке после следует перейти, на страницу **Job Outputs** (ссылка на эту страницу также высылается Вам по почте на адрес, указанный в окне **Your e-mail address**). На данной странице находятся ссылки на файлы с результатами работы программы. Результаты поиска сайтов программы MEME содержатся в файле [meme.html](#).

Работа с результатами программы MEME.

Файл с результатами содержит несколько разделов:

[MEME - Motif discovery tool](#) – информация об используемой версии программы.

[REFERENCE](#) – ссылка на статью о программе.

[TRAINING SET](#) – сведения о введенных последовательностях.

[COMMAND LINE SUMMARY](#) - информация обо всех параметрах, часть из которых программа сама и устанавливает.

Далее следует описание каждого найденного мотива. В первой строке сообщаются сведения о длине сайта (**width**), количестве найденных сайтов (**sites**) и приводятся различные критерии оценки статистической значимости сайта (**llr** и **E-value**).

[simplified pos.-specific probability matrix](#) – построенная на основании найденных сайтов матрица вероятности нуклеотидов. По вертикали указаны нуклеотиды, а по горизонтали – позиции в последовательности сайта. Вероятность данного нуклеотида в данной позиции указывается в десятых долях, то есть, если в матрице стоит число 7, то частота этого нуклеотида в данной позиции равна 0,7. Также используются следующие обозначения: “:” –

данный нуклеотид не встречается в этой позиции; “а” – данный нуклеотид встречается в этой позиции со 100%-ной вероятностью.

Information Content Diagram – диаграмма, показывающая информационное содержание каждой позиции.

Multilevel consensus sequence – консенсусная последовательность для найденного сайта.

Далее приводится таблица, включающая сведения о найденных сайтах:

NAME	STRAND	START	P-VALUE	SITES
gapA	+	74	1.68e-08	GCTGCACCTA AAATCGTGATGAAAATCACATTT TTATCGTAAT
mtlA	+	21	6.74e-08	ATCAAAACAA AAATGTGACACTACTCACATTT AAATGCCATT
tnaL	+	206	9.82e-08	CTCCCGAAC GATTGTGATTCGATTCACATTT AAACAATTTTC
caiT	-	143	9.82e-08	ATAAGCTGTA TTCTGTGATTGGTATCACATTT TTGTTTCGGG
exuT	-	148	1.40e-07	TACAACSTTA AAAGGTGAGAGCCATCACAAAT GTGGGAATAT

NAME – имя последовательности

STRAND – Цепь ДНК, в которой найден сайт: “+” - введенная последовательность, “-” - комплементарная ей

START – положение сайта (то есть положение **первой** позиции сайта относительно начала последовательности)

P-VALUE – критерий статистической значимости сайта, чем он ниже, тем сайт имеет большую значимость. Сайты в списке перечислены именно по возрастанию **p-value**.

SITES – выравненные последовательности найденных сайтов. Приводятся последовательности сайтов (раскрашены) плюс по 10 нуклеотидов с каждой стороны.

В случае, если сайт найден не во введенной цепи, а в комплементарной ей, необходимо искать обратнo-комплементарный ему.

Например, в результатах приведен следующий сайт:

codB - 17 9.57e-09 TGAAGATAAA **AAGCAATCGTTTTCTG** GGGAAATATA

Вам следует искать **обратно-комплементарный** ему, то есть CACGAAAACGATTGCTT

В рассматриваемой последовательности такой сайт будет располагаться следующим образом:

aaaaaatatatttccc**cacgaaaacgattgctt**tttatcttcagatgaatagaatgcggcgatttttt

16 нуклеотидов

сайт, обратно-комплементарный тому, который выдала программа.

Будьте внимательны!

Указывается не количество нуклеотидов до начала сайта, но **положение первой позиции** сайта.

Block diagrams – графическое отображение расположения сайтов по последовательностям, “+” и “-” обозначают цепи ДНК, в которых найден сайт (так же, как и в предыдущем случае).

Экспериментально установленные сайты связывания PurR

codB

aaaaaatatatttcccc**acgaaaacgattgctt**tttatcttcagatgaatagaatgcggcggattttttgggtttcaaacagcaa

purE

tgatttcacagcc**acgcaaccgttttcct**tgctctctttccgtgctattctctgtgccctctaagccgagagttgtgcaccaca

purC

agggcgcatttcgcgccctttatttttcgtgcaa**aggaaaacgtttccgc**ttatcctttgtgtccggcaaaaacatcccttcagcc

purR

ggcgtaccgcaacacttttgttggtgcgtaaggtgtgtaa**aggcaaacgtttacct**tgcgattttgcaggagctgaagttagggtc

cvpA

ttatttgatgcgcgggaaggaaatccct**acgcaaacgttttcct**tttctgttagaatgcgccccgaacaggatgacagggcgtaa

purM

aaaggttggtgtaaagcagtc**tcgcaaacgtttgctt**tcctgttagaattgcgccgaattttatttttctaccgcaagtaacgcg

guaB

gatagcaagcattttttgcaaaaagggtag**atgcaatcggttacgc**tctgtataatgccgcggcaatatttattaaccactctg

glnB

ttcccgcacagagctgg**atgcaaacgatttcaa**ggaatgaattggcgttatgtgttacgtttagcagatcaaagacagggcgacc

purL

ttattttcc**acgcaaacggtttcgt**cagcgcacagattctttataatgacgccggtttcccccttggttacaccgaaagctta

purA

aggtcatttttgagtgcaaaaagtgctgtaactctgaaaaagcgatggtagaatccattttt**aagcaaacggtgattt**tgaaaaa

Задание №2

Поиск сайтов в эукариотических последовательностях

Во втором задании Вам необходимо будет найти сайты для известных сигналов в полученных последовательностях с помощью программы rVISTA. Для выполнения этого задания необходимо будет сделать попарные выравнивания последовательности из генома человека с последовательностями из геномов других млекопитающих. Всего Вам будет выдано три последовательности: одна – для человека, две другие – для других млекопитающих (*например*, Human_Aldolase.txt, Mouse_Aldolase.txt и Dog_Aldolase.txt).

Примечание: Некоторым из вас достанется последовательность Guinea pig. Хочу обратить ваше внимание, что это не порода свиней, а морская свинка (*Cavia porcellus*), которая, как известно, относится к отряду грызунов.

Все последовательности, с которыми Вы будете работать, представляют собой промоторные области генов, экспрессирующихся в мышечной ткани. Поэтому в них Вы будете пытаться найти сайты связывания мышечно-специфичных факторов транскрипции.

Вам потребуется найти сайты для восьми таких факторов:

AP2	GATA1	MEF2	MEF3
MYOD	SRF	TEF1	TEF

Список этих факторов приведен также в инструкции по использованию программы.

Ответ должен состоять из трех частей:

1. Выравнивание последовательностей, на котором размечены **все** найденные сайты (*должно быть представлено в формате *.doc*)

Программа выдает выравнивания, на которых отмечен *сайт только для одного* транскрипционного фактора. Вам же следует на одно выравнивание нанести **все** найденные сайты.

Например, программа выдала:

Для фактора MEF2 :

```

60          70          80          90          100          110
seq1      TTGCCCCTCTAGCCCCCTGTCGGTACCGAGAAGCCCCAAGAGGAGCAGGACCCCAAGCGGA
          | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      TGGCTCCSCTGACCCCTGTCCSTACCAAAGAGTGTTCAGGAGAAGCAGGGCACCACCAAGCAGA
          50          60          70          80          90          100

seq1      120          130          140          150          160          170
          GCCCCTTGCCATCCTGCTGCGTGGTTCTCAGGGTTATCTGAGCTCTGGCAGGCTTGGAGG
          | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      GCCCCTTGCCATCCTGATGTCTGGTTCTCA-GGTTATCTGAGCTTTGACAAGCTT-GAGG
          110          120          130          140          150          160
    
```

Для фактора MYOD :

```

seq1      TTGCCCCTCTAGCCCCCTGTCGGTACCGAGAAGCCCCAAGAGGAGCAGGACCCCAAGCGGA
          | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      TGGCTCCSCTGACCCCTGTCCSTACCAAAGAGTGTTCAGGAGAAGCAGGGCACCACCAAGCAGA
          50          60          70          80          90          100

seq1      120          130          140          150          160          170
          GCCCCTTGCCATCCTGCTGCGTGGTTCTCAGGGTTATCTGAGCTCTGGCAGGCTTGGAGG
          | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      GCCCCTTGCCATCCTGATGTCTGGTTCTCA-GGTTATCTGAGCTTTGACAAGCTT-GAGG
          110          120          130          140          150          160
    
```

Рекомендуется оформлять ответ следующим образом:

```

seq1      TTGCCCTCTAGCCCTGTCCGTACCGAGAAGCCCAAGAGGAGCAGGACCCCAAGCGGA
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      TGGCTCCCCTAGCCCTGTCCCTACCAAAGAGTGTGTCAGGAGAAGCAGGGCACCAAGCAGA
MYOD     -----+-----
MEF2     -----

seq1      GCCCTTGCCATCCTGCTGCGTGGTTCTCAGGTTATTCTGAGCTCTGGCAGGCTTGGAGG
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      GCCCTTGCCATCCTGATGTCTGGTTCTCA-GTTATTCTGAGCTTTGACAAGCTT-GAGG
MYOD     -----+-----
MEF2     -----

```

2. Результаты расчетов, на сколько нуклеотидов приходится один сайт (для каждого выравнивания).

- 1) Поделите среднюю длину одной пары последовательностей (то есть той пары, которую Вы выравнивали) на суммарное число всех найденных сайтов. Это и будет число нуклеотидов, на которое в среднем приходится один сайт.
- 2) Вы искали сайты для 8 мышечно-специфичных факторов. А теперь представьте себе, что Вам необходимо найти сайты для всех 455 факторов, имеющихся в арсенале программы rVISTA. На какое число нуклеотидов в среднем тогда приходился бы один сайт? Чтобы узнать это, разделите полученное значение на 56.

3. Ваши выводы

Инструкция по использованию программы rVISTA.

On-line версия программы находится по адресу <http://genome.lbl.gov/vista/rvista/submit.shtml>

Перейдя по данному адресу необходимо в окне Total number of sequences набрать цифру “3” (поскольку у Вас три последовательности) и нажать “Submit”

Таким образом Вы перейдете в окно программы.

Окно программы содержит следующие поля:

Your email address: Ваш электронный адрес, на который будут высланы результаты.

[Sequence #1](#): последовательность из генома человека;

[Sequence #2](#), [Sequence #3](#): последовательности из других геномов.

Обязательно проследите, чтобы первой была именно последовательность из генома человека, иначе можете запутаться в результатах.

Для заполнения этих полей требуется воспользоваться кнопкой “Обзор” (в некоторых браузерах эта кнопка называется “Browse...”)

Также желательно, чтобы Вы написали названия организмов в разделе **Additional options**, в окнах “Name” – чтобы впоследствии не запутаться в результатах.

Проследите, чтобы стояли следующие пометки:

- в **Alignment program** должна стоять пометка возле [AVID](#)
- возле поля “Find potential transcription factor binding sites using [rVISTA](#)” должна стоять галочка,

Если все установлено правильно, нажмите “Submit”

В новом окне содержатся функции, требуемые для поиска сайтов связывания.

Пометки должны стоять возле надписей “Use TRANSFAC matrices” и “vertebrates”. Если все правильно, можно нажимать “Submit”

После этого загрузится окно со списком известных для позвоночных факторов транскрипции. Здесь Вам необходимо будет поставить галочки возле следующих названий:

AP2	GATA1	MEF2	MEF3
MYOD	SRF	TEF1	TEF

и нажать “Submit”

После этого в новом окне появится сообщение

Your sequences were successfully submitted.
An email will be sent to you when your request is processed.

Это значит, что результаты уже высланы на адрес, указанный в поле Your email address

В полученном письме будет содержаться ссылка на страницу с результатами. По этой ссылке Вам и следует перейти.

В нижней части открывшегося окна будут приведена таблица со ссылками. Вам необходимо пройти по ссылке [rVISTA](#), находящейся напротив последовательности из генома человека. В новом окне приводятся сведения о построенных парных выравниваниях. Всего должно получиться два выравнивания, и с каждым Вы сможете работать по-отдельности. Для того, чтобы приступить к работе с выравниванием, пройдите по соответствующей **rVISTA** (*внизу, в правой части экрана, например, rVISTA: [Human-Cow](#)*)

По этой ссылке Вы перейдете в окно **Choose matrices to visualize**, в котором будут перечислены все факторы транскрипции, отмеченные вами ранее. Возле каждого из названий поставьте галочку и нажмите “Submit”

После этого Вы попадете на страницу **Visualization Options**, где в средней колонке увидите перечислены все факторы транскрипции, сайты для которых Вы пытаетесь найти. В настоящей работе необходимо найти все выровненные (aligned) сайты. Для этого в правой колонке необходимо поставить галочку возле “aligned”, убрать галочки возле “conserved” и “all” и нажать “Submit”. Возле каждого имени фактора находится надпись [view in alignment](#), кликнув по которой Вы перейдете на страницу с выравниванием.

В каждом выравнивании найденные сайты связывания *данного* фактора будут показаны **на розовом фоне**.