

Задание №1

Поиск сигналов регуляции транскрипции в бактериальных последовательностях

В первом задании Вам необходимо найти сигнал (набор сайтов) в полученных последовательностях с помощью программ AlignACE 3.0 и MEME.

В этом файле после задания №1 и инструкций к программам AlignACE 3.0 и 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 сигнал 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
acctgtcccatctcatgctcaagcagcagacgaaccgtttgattcaggcgactaacggtaaaaattgcaggggattg
agaaggtaacatgtgagcagatcaaattctaaatcagcaggttattcagtcgatagtaacccgcctt

>glnB
gggtgaaaaatacggcgctgccaacctttgttgaggcacgtaatcagtttgaactcaactatttgcgtaagctgctgc
aaatcaccaaaaggcaacgtcacccacgcggcgagaatggcggggcgcaaccggacagaa

>purL
attctctgtgtcgtgcgctcccagcttgaaaaaacgtaataatagtgaaaggtttactcataaatgagcggcattt
tgcgtaaacctgcgccagatggcaacttattacagccattggcggcacgcgcttgctaattcacga
```

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.
- пометьте подчеркиванием те сайты, которые были найдены с помощью программы AlignACE 3.0
- 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

ttcagcaatgattgcgagggttatcgcaagaaaacgttttcgcgaggttgatgcggtgctttcct
ggctgttagaatacgcgcccgtcgcgctgactgggacaggggcctgtgtctttgctgaatttga

В данном случае:

aagaaaacgttttcgc
ggttatcgcaagaaaacgt
aaacgttttcgcgagg

gcaagaaaacgttttc

- экспериментально установленный сайт связывания **PurR**
- сайт, найденный с помощью программы AlignACE 3.0
- сайт, найденный с помощью программы MEME с параметром «**One per sequence**»
- сайт, найденный с помощью программы MEME с параметром «**Zero or one per sequence**»

Все выше перечисленные предсказанные сайты считаются совпадающим с экспериментальным, поскольку пересекаются с ним более чем на 8 нуклеотидов.

- gggacaggggcctgtgtc - сайт, найденный с помощью программы AlignACE 3.0, но **не совпадающий** с экспериментальным.

Инструкция по использованию программы AlignACE 3.0

Программа AlignACE предназначена для поиска сходных сайтов в нескольких последовательностях. **On-line** версия программы [AlignACE 3.0](http://atlas.med.harvard.edu/cgi-bin/alignace.pl) находится по адресу <http://atlas.med.harvard.edu/cgi-bin/alignace.pl>.

Окно программы содержит следующие поля:


Enter sequence description – описание вводимых последовательностей (заполнять не обязательно).

Number of columns to align – минимальная длина для искомого слова (*при выполнении задания следует искать сигнал длиной не менее 16 нуклеотидов*).

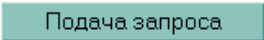
Number of sites to expect – суммарное количество сайтов, которое Вы желаете найти **во всех** введенных последовательностях.

Fractional background GC content – содержание GC-пар в искомых сайтах; указывается в виде десятичной дроби. Этот параметр необходим для того, чтобы программа не принимала за сайты достаточно часто встречающиеся AT-богатые последовательности. По умолчанию программа выставляет значение 0.38; его и рекомендуется использовать при выполнении задачи.

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

- Внести последовательности в FASTA-формате в окно **Enter FASTA-formatted sequence below**
- Указать размещение файла, содержащего последовательности, в окне **Or select file** - можно вбивать руками, а можно воспользоваться кнопкой  (в некоторых браузерах эта кнопка называется “Browse...”)

*Этой опцией пользоваться не рекомендуется, так как она ведет к неудобствам в использовании результатов (см. ниже). Поэтому при выполнении задания последовательности должны быть представлены в FASTA-формате, то есть внесены в окно **Enter FASTA-formatted sequence below***

После заполнения всех необходимых полей, нажимайте  (в некоторых браузерах эта кнопка называется “Submit Query”)

После этого перезагружается та же самая страница, но уже с полученными результатами внизу.

Результаты.

Результаты представляют собой сообщения о параметрах, с которыми был запущен запрос, список введенных последовательностей и все найденные сайты, выравненные между собой.

В первых двух строках сообщается информация о версии программы (в данном случае - **3.0**) и исходных параметрах.

Далее приводится информация уже обо всех параметрах, часть из которых программа сама и устанавливает, и нумерованный список последовательностей, например:

```
Input sequences:
#0 caiT
#1 araB
#2 nmpC
#3 nagB
#4 dadA
#5 gapA
#6 flhD
#7 dsdX
```

Последовательности нумеруются в соответствии с порядком, в котором они приводятся в окне **Enter FASTA-formatted sequence below**.

Будьте внимательны! Нумерация начинается не с 1, а с 0.

Сведения о каждом наборе найденных сайтов представлены в виде следующей таблицы:

Motif	Sequence	Start	End	Chain
Motif 1	ATTGGTGATCCATAAAAACAATA	0	22	0
	TTCTGTGATTGGTATCACAATTT	0	142	0
	ATTTGTGAAGTAGATCTCTATT	2	231	0
	TTTGGTGACAAAACSTCACAATA	3	122	0
	AGATGTGAGCCAGCTCACCAATA	4	183	1
	AATCGTGATGAAAATCACAATTT	5	73	1
	ATGCGTGATGCAGATCACACAA	6	29	0
	TAAAGTGAACCATATCTCAATT	7	145	1
	TGGAGTGATTTACATCTAAAAA	7	175	0
	** **** * *****			
MAP Score: 16.8184				

1. Выравненные последовательности найденных сайтов

2. Номер введенной последовательности ¹⁾

3. Положение сайта в последовательности ²⁾

4. Цепь ДНК, в которой найден сайт:
1 - Введенная последовательность
0 - Комплементарная ей

Звездочками помечены консервативные позиции

Вес сигнала

- 1) при использовании функции **Or select file** для ввода последовательностей **on-line** версия программы автоматически присваивает всем последовательностям номер **#0**, поэтому приходится вносить последовательности в окно **Enter FASTA-formatted sequence below**
- 2) положение сайта указывается в виде числа нуклеотидов от начала последовательности до начала сайта, то есть, если во второй колонке стоит "22", значит, сайт начинается с 23-го нуклеотида. То же самое верно, если данный мотив найден в комплементарной цепи. В этом случае в имеющейся последовательности следует искать сайт, обратнo-комплементарный приведенному.

Например, для одного из сайтов программа выдала следующее:

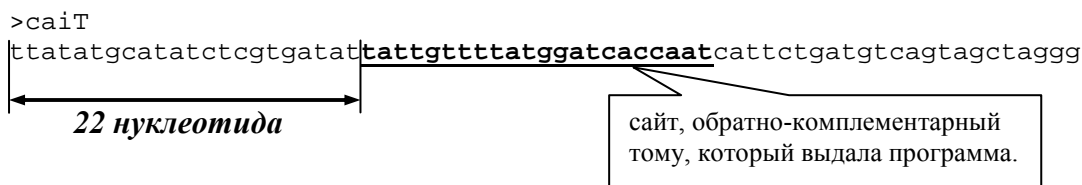
```
ATTGGTGATCCATAAAAACAATA 0 22 0
```

Это значит, что сайт найден в последовательности, которой программа присвоила #0. Во-первых, выясните название этой последовательности (см. выше). То, что в последней колонке стоит "0", означает, что сайт найден в цепи, комплементарной введенной.

Следовательно, необходимо искать не тот сайт, который выдала программа, а **обратно-комплементарный** ему. Например, для приведенного выше сайта обратно-комплементарная последовательность будет иметь вид:

TATTGTTTTATGGATCACCAAT

Допустим, номер #0 был присвоен последовательности с именем `caiT`. Тогда найденный в ней сайт будет располагаться вот так:



Наборы сайтов располагаются в порядке уменьшения веса сигнала. Понятно, что наибольший вес будет иметь первый набор сайтов, который и является лучшим.

При выполнении задания необходимо изменять параметр **Number of sites to expect**, пока не будет найден **наилучший набор сайтов**. Вначале следует в соответствующем поле указать число, равное количеству последовательностей в Вашем варианте задания (то есть попытаться найти в каждой последовательности по одному сайту). Далее необходимо при каждой следующей попытке уменьшать **Number of sites to expect** на единицу. **Наилучшим набором сайтов** будет считаться тот первый набор из списка, вес сигнала для которого будет превышать вес второго набора сайтов **как минимум в 1,5 раза**. Эти сайты и следует пометить в последовательностях.

В случае, если такового набора сайта не удастся обнаружить ни при одном значении параметра **Number of sites to expect**, то лучшим набором сайтов будет считаться тот, вес которого будет максимальным для всех лучших наборов сайтов, найденных при всех значениях **Number of sites to expect**.

Пример.

Вы получили вариант, где содержится 12 последовательностей. При первой попытке поиска сигнала Вы устанавливаете **Number of sites to expect** равным 12, и, если не удалось сразу же получить разницу весов между первым и вторым набором сайтов в 1,5 раза, то производятся повторные попытки поиска сигнала, но уже с меньшими значениями **Number of sites to expect**. Однако, использовать значения данного параметра, меньшие 6, не имеет смысла, поскольку при таких значениях статистическая значимость обнаруженного сигнала будет недостаточно высока.

Если же ни при одном значении **Number of sites to expect** не удалось обнаружить такого наилучшего набора сайтов, то за наилучший выбирается тот набор, вес которого максимален среди всех первых наборов сайтов для различных значений **Number of sites to expect**.

Инструкция по использованию программы MEME

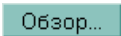
On-line версия программы MEME находится по адресу <http://meme.sdsc.edu/meme/website/meme.html>.

Окно программы содержит следующие поля:

Your e-mail address: Ваш электронный адрес, на который будут посланы результаты.

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

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 

(в некоторых браузерах эта кнопка называется “Browse...”)

- or the **actual sequences** here: в окно вводятся непосредственно сами последовательности, тоже в FASTA-формате.

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.*

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

*При выполнении задания необходимо указать значение 1, то есть искать **только один** тип сигнала.*

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**.

optimum width of each motif : 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.

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”)

После этого результаты будут высланы Вам по почте на адрес, указанный в окне **Your e-mail address**. В обновленном окне браузера при этом появляется информация о входных данных:

- e-mail, на который высланы результаты
- имя файла, содержащего последовательности
- количество искомых сайтов (в каждой последовательности)
- количество искомых сигналов (разных типов сайтов)
- минимальное и максимальное число искомых сайтов в каждой последовательности (если предполагается найти более 1 сайта в каждой последовательности)
- минимальная и максимальная длина искомого сайта
- таблица, отражающая статистику ваших данных

type of sequence	dna	- тип последовательностей
number of sequences	20	- количество последовательностей
shortest sequence (residues)	400	- длина самой короткой последовательности
longest sequence (residues)	400	- длина самой длинной последовательности
average sequence length (residues)	400.0	- средняя длина последовательности
total dataset size (residues)	8000	- суммарная длина всех последовательностей

Работа с результатами.

В ответ на запрос на почтовый ящик высылаются следующие письма:

- 1) Подтверждение о том, что запрос принят и обрабатывается
- 2) Результаты обработки последовательностей программой MEME
- 3) Результаты обработки последовательностей программой MAST

Программа MAST, используя полученный с помощью программы MEME профиль для распознавания сайта, картирует сайты в тех последовательностях, которые были использованы для поиска сигнала.

Для выполнения задания Вам потребуются только результаты 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 сигнала. 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 position-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.

“.” – данный нуклеотид не встречается в этой позиции

“a” – данный нуклеотид встречается в этой позиции со 100%-ной вероятностью

[Information Content Diagram](#) – диаграмма, показывающая информационное содержание каждой позиции. Является аналогом Logo.

[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	AAATGTGACACTACATTT	AAATGCCATT
tnaL	+	206	9.82e-08	CTCCCCGAAC	GATTGTGATTGGATTCACATTT	AAACAATTTTC
caiT	-	143	9.82e-08	ATAAGCTGTA	TTCTGTGATTGGTATCACATTT	TTGTTTCGGG
exuT	-	148	1.40e-07	TACAACCTTTA	AAAGGTGAGAGCCATCACAAT	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*.

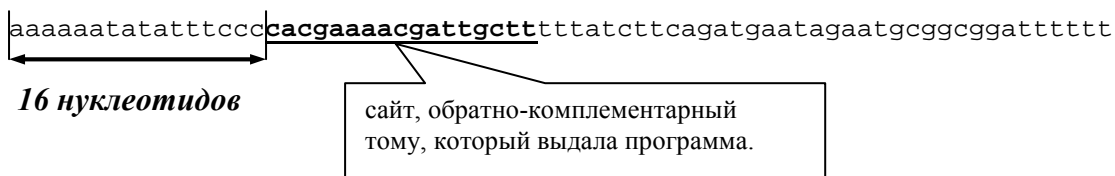
SITES – выравненные последовательности найденных сайтов. Приводятся последовательности сайтов (раскрашены) плюс по 10 нуклеотидов с каждой стороны.

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

Например, в результатах приведен следующий сайт:

codB + 17 9.57e-09 TGAAGATAAA **AAGCAATCGTTTTCGTG** GGGAAATATA

Вам следует искать **обратно-комплементарный** ему, то есть CACGAAAACGATTGCTT
В рассматриваемой последовательности такой сайт будет располагаться следующим образом:



Будьте внимательны!

В отличие от программы [AlignACE 3.0](#), здесь 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).

Сайты связывания белка PurR

codB

aaaaaatatatttccccacgaaaacgattgctttttatcttcagatgaatagaatgcggcggat
tttttgggtttcaaacagcaaaaagggggaatttcg

purE

tgatttcacagccacgcaaccggttttccttgctctctttccgtgctattctctgtgccctctaa
agccgagagttgtgcaccacaggagttttaagacgc

purC

agggcgcattcgcgcctttatTTTTcgtgcaaaggaaaacggtttccgc ttatcctttgtgtcc
ggcaaaaacatcccttcagccggagcatagagatta

purR

ggcgtaccgcaacacttttggtgtgcgtaagggtgtgtaaaggcaaacggtttaccttgcgatttt
gcaggagctgaagttagggctctggagtgaaatggaa

cvpA

tttattgatgcgcggaaggaaatccctacgcaaacggttttcctttttctggttagaatgcgcccc
gaacaggatgacagggcgtaaaatcgtgggacacat

purM

aaaggttgtgtaaagcagtcctcgcaaacggttgctttccctggttagaattgcgccgaattttat
ttttctaccgcaagtaacgcgtggggacccaagcag

guaB

gatagcaagcattttttgcaaaaaggggtagatgcaatcggttacgctctgtataatgccgcgg
caatatttattaaccactctggtcgagatattgcc

glnB

ttcccgcacagagctggatgcaaacgatttcaaggaaatgaattggcggttatgtgttacgtttag
cagatcaaaagacaggcgaccttttcaaggaaatagc

purL

ttattttccacgcaaacggtttcgtcagcgcacagattctttataatgacgcccggtttcccccc
cttgggtacaccgaaagcttagaagacgagagactt

purA

aaaacagactgatcgaggtcatttttgagtgcaaaaagtgctgtaactctgaaaaagcgatggt
agaatccatttttaagcaaacggtgatttgaaaaa

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 сравнить две пары последовательностей. Одна последовательность из каждой пары взята из генома человека, другая – из генома мыши. При выполнении задания следует сравнивать последовательности из файлов с одинаковыми названиями (*Например, из файлов B1_Human.txt и B1_Mouse.txt. Если же вы возьмете, допустим, B1_Human.txt и A1_Mouse.txt, у Вас ничего не получится – эти последовательности относятся к разным генам и выравниваться не будут*).

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 двух 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 и *только по одной цепи*. Вам же следует на одно выравнивание нанести **все** найденные сайты, отметив их **по обеим цепям**.

Например, программа выдала:

Для фактора MEF2 :

```

60          70          80          90          100         110
seq1      TTGCCCCTCTAGCCCTGTCCGTACCGAGAAGCCCAAGAGGAGCAGGACCCCAAGCGGA
          | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      TGGCTCCCCTGACCCCTGTCCSTACCAAAGAGTGTCCAGGAGAAGCAGGGCACCAAGCAGA
          50          60          70          80          90          100

120        130        140        150        160        170
seq1      GCCCTTGCCATCCTGCTGCGTGGTTCTCAGGGTATTCTGAGCTCTGGCAGGCTTGGAGG
          | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      GCCCCTGCCATCCTGATGTCTGGTTCTCA-GGTTATTCTGAGCTTTGACAAGCTT-GAGG
          110        120        130        140        150        160
    
```

Для фактора MYOD :

```

seq1      TTGCCCCTCTAGCCCTGTCCGTACCGAGAAGCCCAAGAGGAGCAGGACCCCAAGCGGA
          | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      TGGCTCCCCTGACCCCTGTCCSTACCAAAGAGTGTCCAGGAGAAGCAGGGCACCAAGCAGA
          50          60          70          80          90          100

120        130        140        150        160        170
seq1      GCCCTTGCCATCCTGCTGCGTGGTTCTCAGGGTATTCTGAGCTCTGGCAGGCTTGGAGG
          | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2      GCCCCTGCCATCCTGATGTCTGGTTCTCA-GGTTATTCTGAGCTTTGACAAGCTT-GAGG
          110        120        130        140        150        160
    
```

Ваш ответ в данном случае должен выглядеть так:

```

                MYOD
seq1  TTGCCCCTC TAGCCCTGTCC GTACCGAGAAGCCCAAGAGGAGCAGGACCCCAAGCGGA
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2  TGGCTCCCC TGACCCTGTCC CTACCAAAGAGTGT CAGGAGAAGCAGGGCACCAAGCAGA

                MEF2
seq1  GCCCTTGCCATCCTGCTGCGTGGTTCTCAGGGTTATTCTGAGCTCTGGCAGGCTTGGAGG
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2  GCCCTTGCCATCCTGATGTCTGTTCTCA -GGTTATTCTGAGCTTTGACAAGCTT-GAGG
  
```

Если сайты разных факторов будут перекрываться между собой, то их необходимо выделять так, чтобы было видно, какие именно нуклеотиды принадлежат к данному фактору, например

```

                MYOD      SPI
seq1  TTGCCCCTC TAGCCCTGTCC GTACCGAGAAGCCCAAGAGGAGCAGGACCCCAAGCGGA
      | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
seq2  TGGCTCCCC TGACCCTGTCC CTACCAAAGAGTGT CAGGAGAAGCAGGGCACCAAGCAGA
  
```

2. Результаты расчета, на сколько нуклеотидов приходится один сайт.

- 1) Поделите среднюю длину одной пары последовательностей (то есть той пары, которую Вы выравняли) 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 407 факторов, имеющихся в арсенале программы rVISTA. На какое число нуклеотидов в среднем тогда приходился бы один сайт? Чтобы узнать это, разделите полученное значение на 50.

В ответе Вы должны привести выравнения и результаты вычислений для обеих пар последовательностей.

Инструкция по использованию программы rVISTA.

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

Перейдя по данному адресу необходимо в окне Total number of sequences набрать цифру “2” (поскольку Вам необходимо будет выровнять между собой две последовательности) и нажать

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: последовательность из генома мыши.

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

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

Далее, оставив остальные окна нетронутыми, необходимо поставить галочку возле поля **Find potential transcription factor binding sites using rVISTA**, чтобы программа нашла в выравненных последовательностях потенциальные сайты связывания транскрипционных факторов, и нажать

В новом окне содержатся функции, требуемые для поиска сайтов связывания. Следует поставить пометки возле надписей Use TRANSFAC matrices и vertebrates, после чего нажать

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

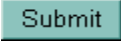
После этого в новом окне появится сообщение

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 открывшемся окне вам будут предложено несколько возможных форматов для полученных результатов. Наиболее удобным для выполнения данной задачи является формат программы **rVISTA**. Поэтому Вам необходимо кликнуть на любую из ссылок [rVISTA](#) (в правой части экрана).

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 

После этого Вы попадете на страницу **Visualization Options**, где в средней колонке увидите перечислены все факторы транскрипции, сайты для которых Вы пытаетесь найти. Возле каждого имени фактора находится надпись [view in alignment](#), кликнув по которой Вы перейдете на страницу с выравниванием.

Найденные сайты связывания *данного* фактора будут показаны  на розовом фоне.