Сдача заданий до 31 октября включительно. Разрешенные форматы отчетов: *.doc, *.docx, *.odt. Отчеты следует отправлять на адрес dravcheev@burnham.org.
Свой вариант задания можно найти в файле **Students.pdf**.

Часть 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.

Введение

Сайты связывания транскрипционных факторов у прокариот достаточно длинны (~15 нуклеотидов), и часто более консервативны, чем их окружение. Также они часто имеют дополнительную внутреннюю структуру: например, являются почти строгими палиндромами. Эти свойства позволяют эффективно находить такие сайты *de novo* методами сравнительной геномики, например, при помощи программы MEME. Для построения профиля сайта связыывания транскрипционного фактора можно использовать промоторные области генов, для которых косвенно показана регуляция этим фактором.
Задача состоит в том, чтобы определить, при каких длинах последовательностей и при каком числе лишних (то есть не содержащих сайта) последовательностей программа способна находить сайты, совпадающие с экспериментальными, и какие параметры нужно для этого использовать.

Исходные данные

1. Набор промоторных областей генов *E.coli*, вида **MEME<номер>.txt** (номер Вашего варианта Вы можете найти в файле **Students.pdf**).
2. Список экспериментально установленных сайты связывания белка PurR. Для всех сайтов указан контекст, но не указаны их координаты

codB
aaaaaatatatttcccc**acgaaaacgattgctt**tttatcttcagatgaatagaatgcggcggatttttgggtttcaaacag
purE
tgatttcacagcc**acgcaaccgttttcct**tgctctctttccgtgctattctctgtgccctctaaagccgagagttgtgcacc
pyrC
agggcgcattcgcgccctttatttttcgtgcaa**aggaaaacgtttccgc**ttatcctttgtgtccggcaaaaacatcccttca
purR
ggcgtaccgcaacacttttgttgtgcgtaaggtgtgtaa**aggcaaacgtttacct**tgcgattttgcaggagctgaagttagg
cvpA
tttattgatgcgcgggaaggaaatccct**acgcaaacgttttctt**tttctgttagaatgcgccccgaacaggatgacagggcg
purM
aaaggttgtgtaaagcagtc**tcgcaaacgtttgctt**tccctgttagaattgcgccgaattttatttttctaccgcaagtaac
guaB
gatagcaagcatttttgcaaaaaggggtag**atgcaatcggttacgc**tctgtataatgccgcggcaatatttattaaccact
glnB
ttcccgacacgagctgg**atgcaaacgatttcaa**ggaatgaattggcgttatgtgttacgtttagcagatcaaaagacaggcg
purL
ttatttcc**acgcaaacggtttcgt**cagcgcatcagattctttataatgacgcccgtttcccccccttgggtacaccgaaagc
purA
tcatttttgagtgcaaaaagtgctgtaactctgaaaaagcgatggtagaatccatttttt**aagcaaacggtgattt**tgaaaaa

Используемая программа

http://meme.nbcr.net/meme/cgi-bin/meme.cgi

Результат выполнения задания

Разметка в исходных последовательностях (файл «MEME<номер>.txt») сайтов связывания PurR, как экспериментальных, так и найденных программой MEME (см. указания). Анализ результатов.

Указания к выполнению задания

- 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 / *.docx / *.odt 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
- В отчете следует указать длины последовательностей (в каждом варианте все последовательности имеют одинаковую длину) и их количество.
- Программу надо будет запустить два раза, задав различное число ожидаемых сайтов в последовательностях (в поле **How do you think the occurrences of a single motif are distributed among the sequences?**).
  1. **One per sequence**
  2. **Zero or one per sequence**
- Остальные параметры всегда остаются одинаковыми:

  Minimum length = 16

  Maximum length = 16

  Maximum number of motifs to find = 1
- Следует дважды указать действующий e-mail, на который придет сообщение о завершении работы.
- Также следует вставить последовательности в поле **the actual sequences here (Sample Protein Input Sequences)** или загрузить соответствующий файл с последовательностями.
- В выдаче программы обратите внимание на p-value, оно должно быть меньше 10─4. Сайты с большим p-value следует игнорировать. Если программа нашла сайт на комплементарной цепи (strand = ─), нужно искать в исходной последовательности обратно-комеплементарный ему сайт (то есть нужно спроецировать этот сайт на прямую цепь).

Часть 2. **Поиск сайтов в эукариотических последовательностях**

Введение

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. После этого следует просканировать полученные выравнивания с помощью матриц из базы данных TFANSFAC, используя ту же программу rVISTA.

Исходные данные

Набор из трёх текстовых файлов, содержащих промоторные области ортологичных генов млекопитающих, экспрессирующихся в мышцах: **rVista<номер>.zip** (номер Вашего варианта Вы можете найти в файле **Students.pdf**).

Используемая программа

http://genome.lbl.gov/vista/rvista/submit.shtml

Результат выполнения задания

Два попарных выравнивания (геном человека — геном другого позвоночного) с разметкой сайтов связывания восьми транскрипционных факторов, найденных программой. Расчёт плотности потенциальных сайтов в геноме.

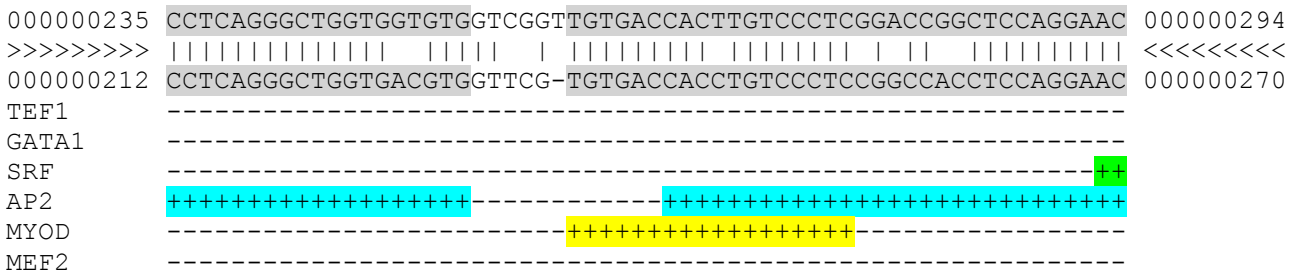
Указания к выполнению задания

- На первом шаге ввода параметров нужно ввести количество последовательностей (3).
- На втором шаге кроме e-mail'а и нуклеотидных последовательностей нужно установить **Alignment program = AVID** и флажок рядом с опцией **Find potential transcription fator binding sites using rVISTA**. Также стоит ввести в качестве названий последовательностей названия организмов.
- На третьем шаге нужно выбрать **Use TRANSFAC matrices** и **vertebrates**.
- После этого программа предлагает выбрать список факторов транскрипции, для которых будет производиться поиск сайтов связывания. Нужно отметить мышечно специфичные факторы:
  AP2
  GATA1
  MEF2
  MEF3
  MYOD
  SRF
  TEF
  TEF1
- После нажатия Submit на Ваш электронный адрес придёт письмо со ссылкой на результат запуска программы. По этой ссылке Вам и следует перейти.
- В открывшемся окне будут приведены сведения о построенных выравниваниях. Всего должно получиться два выравнивания, и с каждым Вы сможете работать по-одтельности. 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 перечисленны 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. 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.

Ответ должен состоять из трех частей:

1. Выравнивание последовательностей, на котором размечены все найденные сайты. Программа выдает выравнивания, на которых отмечен сайт только для одного транскрипционного фактора. Вам же следует на одно выравнивание нанести все найденные сайты.
   Рекомендуемый вариант оформления:



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 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.
3. Ваши выводы