Часть 1. Поиск регуляторного мотива в заданной выборке фрагментов

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 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

Набор промоторных областей генов *E. coli*: MEME<номер>.txt.

Список экспериментально установленных сайтов связывания транскрипционного фактора PurR, включающий контекст, но не включающий координаты сайта:

Экспериментально установленные сайты связывания PurR

codB
aaaaaatatatttcccc**acgaaaacgattgctt**tttatcttcagatgaatagaatgcggcggatttttgggtttcaaacagcaa
purE
tgatttcacagcc**acgcaaccgttttcct**tgctctctttccgtgctattctctgtgccctctaaagccgagagttgtgcaccaca
pyrC
agggcgcattcgcgccctttatttttcgtgcaa**aggaaaacgtttccgc**ttatcctttgtgtccggcaaaaacatcccttcagcc
purR
ggcgtaccgcaacacttttgttgtgcgtaaggtgtgtaa**aggcaaacgtttacct**tgcgattttgcaggagctgaagttagggtc
cvpA
tttattgatgcgcgggaaggaaatccct**acgcaaacgttttctt**tttctgttagaatgcgccccgaacaggatgacagggcgtaa
purM
aaaggttgtgtaaagcagtc**tcgcaaacgtttgctt**tccctgttagaattgcgccgaattttatttttctaccgcaagtaacgcg
guaB
gatagcaagcatttttgcaaaaaggggtag**atgcaatcggttacgc**tctgtataatgccgcggcaatatttattaaccactctg
glnB
ttcccgacacgagctgg**atgcaaacgatttcaa**ggaatgaattggcgttatgtgttacgtttagcagatcaaaagacaggcgacc
purL
ttatttcc**acgcaaacggtttcgt**cagcgcatcagattctttataatgacgcccgtttcccccccttgggtacaccgaaagctta
purA
aggtcatttttgagtgcaaaaagtgctgtaactctgaaaaagcgatggtagaatccatttttt**aagcaaacggtgattt**tgaaaaa

Используемая программа

MEME http://meme.sdsc.edu/meme4_6_1/cgi-bin/meme.cgi

Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

Результат выполнения задания

Разметка в исходных апстримах сайтов связывания PurR, как экспериментальных, так и найденных программой MEME (см. указания). Анализ результатов.

Указания

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 размечеными 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» должны быть выделены курсивом.

Сайты, найденные с помощью программы MEME с параметром «Zero or one per sequence» должны быть выделены жирным шрифтом.

Все сайты (и экспериментальные, и предсказанные) должны быть на сером фоне.

В отчете следует указать длину и количество последовательностей.

Программу нужно запустить 2 раза, задав различное число ожидаемых сайтов в последовательностях:

1. How do you think the occurrences of a single motif are distributed among the sequences?
=One per sequence

2. How do you think the occurrences of a single motif are distributed among the sequences?
= Zero or one per sequence

Установка остальных параметров одинакова:

Minimum length = 16
Maximum length = 16
Maximum number of motifs to find = 1

Также нужно подгрузить файл с последовательностями или скопировать их в окно формы и указать e-mail.

Все поля раздела Options нужно оставить пустыми.

В выдаче программы обратите внимание на p-value, оно должно быть меньше 10^{-4}. 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. Описание всех потенциальных сайтов связывания факторов транскрипции в заданном фрагменте генома человека, оценка плотности потенциальных сайтов

Введение

Если профиль (позиционная матрица весов) сайта связывания транскрипционного фактора известен, можно сканировать промоторные области генов и размечать в них сайты связывания транскрипционных факторов различными программами. Программа rVISTA позволяет проводить такое сканирование одновременно для любого количества транскрипционных факторов из обширной базы данных TRANSFAC.

Исходные данные

Набор из трёх текстовых файлов, содержащих промоторные области ортологичных генов позвоночных, экспрессирующихся в мышцах: rVista<номер>.

Используемая программа

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

Loots, G., Ovcharenko,I., Pachter,L., Dubchak,I., Rubin, E. rVISTA for comparative sequence-based discovery of functional transcription factor binding sites. (2002) Genome. Res. 12:832-839

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

На первом шаге ввода параметров нужно ввести количество последовательностей (3). На втором кроме e-mail 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 произфодиться 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 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-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 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 перечислены все факторы транскрипции, сайты для которых Вы пытаетесь найти. В настоящей работе необходимо найти все выравненные (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.

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 способ оформления:

```
000000235   CCTCAGGGCTGGTGGTGTGGTCGGTTGTGACCACTTGTCCCTCGGACCGGCTCCAGGAAC  000000294
>>>>>>>>>   ||||||||||||||  |||||  | |||||||||| ||||||||| | ||  ||||||||||  <<<<<<<<<
000000212   CCTCAGGGCTGGTGACGTGGTTCG-TGTGACCACCTGTCCCTCCGGCCACCTCCAGGAAC  000000270
TEF1        ------------------------------------------------------------
GATA1       ------------------------------------------------------------
SRF         ------------------------------------------------------------++
AP2         ++++++++++++++++++----------++++++++++++++++++++++++++++++++
MYOD        ----------------------++++++++++++++++++----------------------
MEF2        ------------------------------------------------------------
```

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

1) Поделите среднюю длину выравнивания одной пары последовательностей (той пары, которую Вы выравнивали) на суммарное число всех найденных сайтов. Получится довольно точная оценка числа 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