Часть 0. Шпаргалка

Программы

Используют внутренние статистические свойства последовательности:

ORF Finder http://www.ncbi.nlm.nih.gov/gorf/gorf.html

GeneMark http://exon.gatech.edu/genemark/gmhmm2_prok.cgi

GENSCAN http://genes.mit.edu/GENSCAN.html

Ищут известные гомологи:

BLAST http://blast.ncbi.nlm.nih.gov/Blast.cgi

Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway
(при помощи программы **BLAT**, очень похожей на BLAST)

Итог работы

Задание состоит из двух частей, прокариотической и эукариотической, они сдаются и проверяются одновременно. Отчёт, файл в формате Word, нужно послать по электронной почте pseudoobscura@gmail.com Екатерине Олеговне Ермаковой, последний день сдачи отчётов — 30 ноября 2011 года. Будут приниматься и проверяться только отчёты о выполнении всех пунктов задания (и прокариотической, и эукариотической части). Файл с отчётом должен быть назван так: genes_<фамилия латиницей.docx (или .doc), например, genes_Petrov.docx.

Правила оформления

Выравнивания должны быть набраны моноширинным шрифтом.

Если нужно выделить цветом ячейку таблицы или фрагмент выравнивания, используйте цветной фон, а не цветной текст.

Часть 1. Прокариоты

Вам дан фрагмент генома бактерии *Abiotrophia defectiva* (см. файл **a<номер>.txt**). Ваша задача — при помощи программ **ORF Finder**, **GeneMark** и **blastp** проаннотировать этот фрагмент — определить границы генов, по возможности предсказать их функцию — и обосновать Ваше решение.

С помощью программы ORF Finder найдите открытые рамки считывания в геномной ДНК. Скопируйте последовательность ДНК (о цифрах и пробелах можно не беспокоиться, программа их автоматически пропустит) в текстовое поле формы, в окошке выбора генетического кода выберите **11 Bacterial Code** и нажмите кнопку **OrfFind**. Вы получите список обнаруженных открытых рамок. Шесть белых полосок в выдаче ORF Finder изображают исходную последовательность в шести возможных рамках считывания, в порядке +1, +2, +3 (прямая цепь), –1, –2, –3 (обратная цепь), а бирюзовые полоски — найденные в данной рамке ORFs. Мы рассматриваем рамки считывания не короче 60 кодонов, из существенно перекрывающихся выбираем самую длинную (перекрывание рамок будем считать существенным, если оно затрагивает хотя бы половину одной из них). Длина ORF должна делиться на 3.

Занесите данные о рамках считывания, прошедших фильтрацию по длине и перекрыванию, в таблицу:

Abiotrophia defectiva: ORF Finder				
начало	конец	длина	цепь	описание
7877	8314	438	+	Транскрипционный регулятор, принадлежащий к семейству MarR

Если щёлкнуть на одну из рамок считывания на первичной странице результатов **ORF Finder**, откроется её подробное описание. Если нажать кнопку **BLAST** на страничке с подробным описанием рамки, можно найти программой blastp предположительные гомологи предсказанного гена, при помощи которых можно предсказать функцию белка. На открывшейся странице форматирования результатов в строке **Show** отметьте **Advanced view**, в строке **Limit results** выставите во всех трёх полях **50**; (для начала) нажмите кнопку **View report**. Все белковые хиты будут расположены в порядке возрастания e-value. Чем меньше e-value, тем лучше. Значимыми можно считать хиты с e-value $<10^{-4}$. Если из двух рамок со значительным перекрытием (>20 нуклеотидов) одна подтверждается blast'ом, а другая нет, последняя, скорее всего, не является геном. Сходство запроса и белка из базы должно быть хорошим, белок должен покрывать запрос по возможности полностью. Быть может, старт последовательности придётся сместить.

Для аннотации нужно использовать не hypotetical и не predicted белки, даже если они немного хуже выравниваются с запросом (скажем, белки с идентификаторами NP_ лучше, чем белки с идентификаторами XP_). В базе данных лежат сырые предсказания белков *A. defectiva*, но их функция не предсказана, просто у всех написано hypothetical. Такие белки просто пропускаем.

Предположительная функция белка должна быть описана по-русски.

Обязательной вставьте в отчёт данные о хитах, при помощи которых Вы предсказали функцию гена (выравнивание, идентификатор, организм, длина белка, длина выравнивания, сходство, e-value). Если Вы сместили старт гена, нужно привести такое же обоснование.

GeneMark

Скопируйте последовательность ДНК в поле **Sequence Text**. Включите опции **Use RBS model, if avaliable**, **Print GeneMark 2.4 predictions in addition to GeneMark.hmm predictions** и **Generate PDF graphics (screen)**. Запустите программу кнопкой **Start GeneMark.hmm**. Вы получите таблицы генов, предсказанных двумя программами: GeneMark и GeneMark 2.4. В отчете используйте результаты работы GeneMark 2.4.

Посмотрите, как распределен кодирующий потенциал по ДНК, для этого нажмите гиперссылку **View PDF Graphical Output** на странице с предсказаниями GeneMark. На открывшемся рисунке в формате PDF Вы увидите графики распределения кодирующего потенциала для каждой рамки считывания на обеих цепях ДНК. Вставьте графики в отчёт.

Занесите результаты работы программы GeneMark 2.4 в такую же таблицу, как для ORF Finder.
Знак < или > возле границы предсказанного гена означает, что программа предполагает продолжение гена за пределами данной ей последовательности.

В обеих таблицах выделите зелёным гены, одинаково предсказанные ORF Finder и GeneMark, жёлтым — пересекающиеся, но не совпадающие полностью, оранжевым — гены, предсказанные ровно одной программой. ORF, которые Вы не считаете настоящими генами по результатам BLAST'а, оставьте незакрашенными.

Часть 2.1 Эукариоты: GENSCAN, Genome Browser

Вам дан фрагмент ДНК из генома человека (h<номер>.txt). Ваша задача — определить экзон-интронную структуру гена и описать его альтернативный сплайсинг, используя программы **GENSCAN**, **BLAT** и **Genome Browser**.

GENSCAN

Скопируйте последовательность ДНК в текстовое поле формы и нажмите кнопку Run GENSCAN. GENSCAN представляет результаты в виде таблицы экзонов. Вам будут нужны её колонки Type, S (Strand), Begin и End. Обязательно прочтите расшифровку обозначений в разделе Explanation после таблицы. Занесите в свою таблицу начало, конец, цепь и тип всех предсказанных программой экзонов, заведя отдельную таблицу на каждый предполагаемый ген (тип экзона: Init — initial (начальный), Intr — internal (внутренний), Term — terminal (конечный); PlyA — это не экзон, а сайт полиаденилирования). Пример таблицы:

Homo sapiens: GENSCAN			
начало	конец	цепь	тип
315	490	+	начальный
1009	1300	+	внутренний
2000	2101	+	внутренний

Genome Browser

База **Genome Browser (ранее Human Genome Browser, HGB)** (http://genome.ucsc.edu/cgi-bin/hgGateway) содержит гены, белки, мРНК и другие объекты, картированные на различные аннотированные геномы. Браузер позволяет просмотреть разнообразную информацию, относящуюся к заданному фрагменту ДНК. Программа **BLAT** аналогично BLAST позволяет искать последовательности в геноме с учетом возможной фрагментированности генома. Доступ к программе можно получить по ссылке Blat с основной страницы портала (на синей полосе сверху).

Поместите последовательность ДНК в текстовое поле формы, выберите поиск в геноме человека, сборка (Assembly) Feb. 2009, и нажмите кнопку Submit. Вы получите список найденных фрагментов генома. Если в этом списке больше одной строки, выберите ту строку, которая имеет максимальное сходство с вашей последовательностью по SCORE и максимальную длину выравнивания. Если Ваша последовательность длиннее 25000 нуклеотидов, придётся поделить её на части, найти их по отдельности, записать координаты находок и убедиться, что они нашлись в геноме рядом. После этого выставить в окне просмотра координаты целого фрагмента.

Перейдите к просмотру найденного фрагмента генома человека: нажмите гиперссылку browser. Может быть, чтобы увидеть весь ген, нужно будет расширить область просмотра или уменьшить масштаб. Поэкспериментируйте с кнопками! Под картинкой находятся выпадающие меню для выбора отображаемых объектов. Поставьте на pack переключатель Blat Sequence в группе Mapping and Sequencing Tracks, а также переключатели Human mRNAs и Spliced ESTs в группе mRNA and EST Tracks, остальные переключатели поставьте на hide. Нажмите кнопку refresh, она находится в самом низу страницы. Теперь Вы видите, как выравниваются с геномной ДНК Ваша последовательность, а также сплайсированные EST и мРНК из базы. Быть может, с Вашим запросом выравняется только часть гена человека, тогда нужно будет настроить браузер так, чтобы был виден ген целиком.

Приведите примеры альтернативного сплайсинга в найденном гене человека, указав тип альтернативы (это могут быть, например, кассетные экзоны, чередующиеся экзоны, альтернативные донорные и акцепторные сайты сплайсинга, удержанные интроны) и идентификаторы мРНК или EST, подтверждающих альтернативный сплайсинг (минимум 2 на каждую альтернативу, например, для кассетного экзона нужно указать транскрипт, пропускающий экзон и транскрипт, включающий его). Обязательно вставьте в отчёт картинку (размер окна по длине гена). Обведите на ней найденные альтернативы. **Внимание! Начало первого и конец последнего экзона EST использовать при аннотации нельзя, они обрываются в произвольном месте!**

Часть 2.2 Эукариоты: blastx

Вам дан фрагмент ДНК из генома голубиного гороха *Cajanus cajan* (c<номер>.txt). Ваша задача — при помощи программы **blastx** проаннотировать этот фрагмент —разметить экзон-инронную структуру генов и предсказать их функцию. Если можно предсказать несколько изоформ гена, их нужно описать.

Исключите (Exclude) поиск по моделям и природным последовательностям (поставьте галочки возле Models (XM/XP) и Uncultured/environmental sample sequences).

Как и ранее, нужно стараться использовать не предсказанные белки. Обратите особое внимание на белки из SWISSPROT.

Стоит запустить поиск несколько раз. В меню Database можно оставить предложенный по умолчанию банк **nr,** а можно ограничить поиск базой **SWISSPROT**. Можно ограничить поиск только белками сои (другое бобовое растение, пожалуй, самый близкий к голубиному гороху вид из ранее секвенированных) или резуховидки Таля (растение, изученное лучше других) в окне Organism. Если включить Advanced view результатов, можно будет сортировать предсказанные экзоны по порядку их следования в запросе или последовательности из базы (последнее наиболее удобно).

Занесите аннотацию экзонов в таблицу (отдельно для каждого гена, если Вы найдёте их несколько), учитывая что blastx определяет границы экзонов не точно и их нужно уточнять вручную:

- «Экзоны», размеченные BLAST'ом, могут перекрываться как по ДНК, так и по белку. Вам нужно посмотреть на выравнивания таких «экзонов» и уточнить их границы на ДНК. Для этого посмотрите, какой «экзон» лучше выравнивается в области перекрытия. Считайте, что перекрытие принадлежит «экзону» с наилучшим выравниванием.
- Длинная вставка в последовательности ДНК по сравнению с белком, скорее всего, является интроном. Если вставка ДНК содержит стоп-кодон (отмечается знаком * на выравнивании), это прямое указание на интрон. Такой «экзон» нужно разбить на два «экзона».
- Иногда «экзоны» нужно, наоборот, объединить.
- Blast выводит «экзоны» в случайном порядке. Нужно расположить «экзоны» в порядке возрастания координат **по белку**. Белковая координата конца предыдущего экзона должна быть на единицу меньше белковой координаты начала следующего экзона (или перекрываться на одну-три аминокислоты).

В отчете для каждого гена приведите **выравнивание** и отдельную таблицу координат «экзонов» на белке и ДНК:

Идентификатор и название белка из базы, организм, цепь	
координаты по белку	координаты по ДНК
начало экзона 1	начало экзона 1
конец экзона 1	конец экзона 1

начало экзона 2	начало экзона 2
конец экзона 2	конец экзона 2

При подсчёте координат помните о том, что на каждую аминокислоту должно приходиться по три нуклеотида! Пример таблицы:

>gi|9621790|gb|AAF89534.1| serine protease [Mus musculus], +	
координаты по белку	координаты по ДНК
1	603
169	1109
169	1211
237	1417

Укажите отличия уточнённых экзонов от предсказания blastx. Для этого добавьте колонку QQ в таблицу уточнённых экзонов. Перекрытие QQ — мера близости двух систем отрезков, вычисляется как отношение длины пересечения к длине объединения отрезков из этих систем. Скажем, если у сырого экзона координаты 1-98, а у уточнённого 3-100, то QQ=96/100=0,96. Вам нужно вычислить QQ для уточнённого экзона и исходного blastx-экзона, его породившего. Так же, кстати, можно описывать сходство аннотаций, полученных разными программами.

Если Вы заносите результат blast'а в таблицу, нужно привести в тексте соответствующее выравнивание.

Экзоны, полученные blastx, обязательно должны быть отсортированы, а уточнённые экзоны отмечены на выравниваниях цветом (используйте цветной фон текста).