Редакционно-издательский отдел
Санкт-Петербургского национального
исследовательского университета
информационных технологий, механики
и оптики.
197101, Санкт-Петербург, Кронверкский пр., 49

Порозов Ю.Б.

Биоинформатика

Учебно-методическое пособие

Санкт-Петербург

2012

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

Ю.Б. Порозов

Биоинформатика

Учебно-методическое пособие

Санкт-Петербург

2012

Порозов Ю.Б., Биоинформатика. – СПб: НИУ ИТМО, 2012. – 52 с.

В учебно-методическом пособии приведены лабораторные работы по биоинформатике, описание их пошагового выполнения, вопросы для самоконтроля и последовательности, необходимые при выполнении лабораторных работ.

Для студентов дневного отделения специальности 230201.65 «Информационные системы и технологии», 240700 «биотехнология».

Одобрено и рекомендовано к печати Ученым советом естественно-научного факультета « 18 » сентября 2012. Протокол № 9

В 2009 году Университет стал победителем многоэтапного конкурса, в результате которого определены 12 ведущих университетов России, которым присвоена категория «Национальный исследовательский университет». Министерством образования и науки Российской Федерации была утверждена программа его развития на 2009–2018 годы. В 2011 году Университет получил наименование «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

Содержание.

Лабораторная работа № 1

Целью этой работы является познакомить студента с двумя важнейшими базами данных в Интернете - GenBank (http://www.ncbi.nlm.nih.gov/genbank/index.html) , база данных генов и геномов и Protein Data Bank (PDB) (http://www.rcsb.org/pdb/) – база данных структур биологических макромолекул.

Во время выполнения этой работы студент должен узнать и освоить:

1. Поиск нуклеотидной последовательности по названию определенного белка;
2. Поиск научных статей о конкретном белке;
3. Получать файл, описывающий 3D структуру белка;
4. Изучение третичной структуры белка с использованием 3D-браузера.

ENTREZ машина на NCBI

US National Centre for Biotechnological Information (NCBI http://www.ncbi.nlm.nih.gov/) поддерживает базу данных GenBank для генных и геномных последовательностей. Этот веб-сервер связывает их с другими базами данных и ресурсами, включая National Library of Medicine http://www.nlm.nih.gov/ и предоставляет пакет программ для поиска по GenBank – BLAST

http://blast.ncbi.nlm.nih.gov/Blast.cgi. Этот сервер сделан и поддерживается профессионально. Доступ к нему бесплатный и штат отвечает даже на технические вопросы. Это – лучшее место в Интернете для начала поиска биоинформатических данных.

а) В веб браузере откройте url http://www.ncbi.nlm.nih.gov/Entrez . Несмотря на то, что мы будет использовать только один раздел на сервере (и сможем иметь доступ к нему напрямую) стоит потратить несколько минут для того, чтобы просмотреть домашнюю страницу сайта для того, чтобы оценить его масштаб.

б) Кликните по ссылке Entrez. http://www.ncbi.nlm.nih.gov/Entrez/ Вы перейдете на обновленную страницу Entrez (по французски «enter», произносится как “awntray”). На этой странице можно производить поиск по нескольким базам данных одновременно. Для этого упражнения нас интересует *фермент isocitrate dehydrogenase,* играющий важную роль в генерировании энергии. Мы будем искать ген, кодирующий этот белок, в кишечной бактерии, Escherichia coli. Мы будем сравнивать бактериальный белок и соответствующий белок млекопитающих.

в) Наберите *isocitrate dehydrogenase* в поле поиска и нажмите Go. Результатом будут соответствующие записи из нескольких баз данных. Нас интересуют данные из баз Nucleotide (DNA и RNA), Genome

(целые хромосомы организмов) и Structure (трехмерные структуры белков).

Получение нуклеотидной последовательности

а) Кликните на Nucleotide: вы увидите первую из сотни страниц результатов. Нам нужно как-то сократить этот список до размеров, когда с ним будет удобно работать. Чтобы это сделать:

б) Кликните на ссылке Limits. Вы переместитесь на страницу, на которой можно выбрать среди множества опций для ограничения вашего поиска.

в) Выберите Title из выпадающего меню *Search field tags* и нажмите Search. Вы вернетесь на страницу с результатами и увидите, что общее их количество значительно сократилось. Однако мы хотим ограничить наш поиск последовательностей только Escherichia coli (E. coli). Поэтому:

г) Добавьте «and E coli» в поле поиска после "isocitrate dehydrogenase".

д) Кликните Search. Вы вернетесь на страницу поиска с двумя результатами, второй из которых – то, что нам нужно.

е) Кликните на номере доступа – J02799. Вы перейдете на страницу в стандартном формате GenBank с документацией и ссылками и следующей за ней последовательностью ДНК. Мы ещё вернемся к этим ссылкам, но вначале для того, чтобы убедиться, что файл может использоваться в программах, нам надо убедиться в том, что он содержит правильный текст.

ж) Выберите опцию Text в выпадающем меню, начинающемся с Send To. Это генерирует страницу в формате plain text без ссылок.

з) Выберите Save из меню File браузера и сохраните страницу под именем J02799.gbk. Тонкость состоит в том, что, хотя этот файл содержит только ASCII текст, он в формате Unix. ASCII символ, обозначающий конец строки, различен в Unix, PC и Mac-платформах.

и) Первая строка показывает нам, что длина последовательности 1568 bp (пар оснований) и последовательность есть ДНК. Собственно последовательность начинается со строки ORIGIN. Эту строку можно использовать как индикатор при открытии любых файлов GenBank.

к) Если вы посмотрите на строки с отступами, следующими за FEATURES, вы увидите CDS. Это – расположение кодирующей последовательности и содержит часть гена, транслируемой в белок – нуклеотиды 291-1541. Почему не вся генетическая информация транслируется в аминокислоты белка будет объяснено на лекциях. Это имеет практическую ценность для анализа последовательности ДНК.

Получение научных статей

Практикующим ученым часто нужны работы, связанные с последовательностями, которые они скачали из баз данных. NCBI сервер предоставляет ссылки на аннотации статей в National Library of Medicine (MLM) и часто – pdf версии оригиналов. NLM, разумеется, содержит аннотации статей, описывающих не только последовательнсоти ДНК и белков. Все они доступны через машину PubMed из NCBI.

Вернитесь на страницу последовательнсоти с html ссылками (кнопка Back браузера).

а) Кликните на PUBMED номере (3112144). Это перенесет вас на страницу со ссылкой на оригинальную статью, откуда вы можете попасть на веб сайт журнала, откуда статья может быть скачана в виде Pdf файла.

Получение структурной информации о белке

Файл GenBank J02799 представляет белок isocitrate dehydrogenase (ICDH) как строку символов. Однако белки – это биологические последовательности, и будет очень полезно исследовать этот аспект. 3D структуры белков определяются экспериментально и данные о них депонируются в виде текстовых файлов, описывающих xyz координаты атомов в пространстве. База данных, хранящая эти файлы – Protein Data Bank (бывшая Brookhaven) и формат данных, обычно распознаваемый большинством программ для молекулярной графики называется PDB форматом. GenBank содержит некоторое количество данных из PDB в совей MMDB ((Molecular Modelling Data Base), но также имеет ссылки на Protein Data Bank.

Вернитесь на главную страницу Entrez, где может быть необходимым повторить поиск по многим базам данных по запросу isocitrate dehydrogenase.

а) Кликните на иконке Structure. Это откроет перед вами две страницы ICDH структур. Нам нужно закрыть одну – ту, которая содержит белок со связанным изоцитратом.

б) На второй странице кликните на 5ICD. Вы попадете на страницу этой структуры на MMDB. Используйте RasMol Chime плагин – вы сможете увидеть этот белок. Нам нужно скачать этот файл.

в) Кликните на PDB: 5ICD. Это переместит нас на страницу с этой структурой на сайте PDB http://www.rcsb.org/pdb/.

г) Кликните Download File. Вам нужно ввести PDB ID “5ICD” в главном окне. В опциях сохранения выберите “PDB format” и “No compression”. Сохраните файл на диск как 5ICD.pdb.

Использование RasMol для просмотра и анализа структуры белка

Как уже не раз отмечалось, файлы в формате GenBank представляют белки как строки символов 20-и буквенного алфавита. Хотя эти строки и

несут информацию, как, например, ДНК, эта информация реализуется через трехмерную структуру белка и присущие ему свойства. Мы будем использовать бесплатную кросс-платформенную программу RasMol для визуализации этой структуры.

Запустите RasMol, напечатав rasmol в командной строке unix. Для Windows нужно скачать пакет RasMol с http://www.umass.edu/microbio/rasmol/ или http://rasmol.org/ и установить его. Вы увидите черный графический экран. RasMol – это графическое приложение с GUI. Но полная его мощь раскрывается через использование командной строки и команд, вводимых в текстовом окне, которое обычно скрыто.

Посмотрите на RasMol Reference Card http://www.openrasmol.org/software/RasMol_2.7.3/doc/README.html (Имеется также on-line manual, содержащий полный набор команд и ключей программы). Загрузить RasMol на локальный компьютер можно также с http://www.bernstein-plus-sons.com/software/rasmol/ и http://rasmol.org/

а) Разверните окно командной строки.
б) Если вы хотите, чтобы фон окна командной строки стал белым, введите "set background white" в этом окне.
в) Загрузите файл 5ICD.pdb при помощи команды Open из меню File.
г) Белок появится в виде wire-frame. Это бывает полезно для биологов, поскольку позволяет изучать отдельные аминокислоты, но в этом случае может быть сложно связать их с первичной структурой белка, то есть с его последовательностью.
д) Введите "restrict protein"
е) Выберите подменю Backbone из меню Display. Вы увидите начало белка, однако его конец будет трудно найти на экране.
ж) Выберите Group (или Индекс) в меню Цвет. Теперь вы можете проследить всю цепь благодаря градиенту цвета от голубого до красного, вращая молекулу.
з) Выберите Spacefill (Молекулярная поверхность) из меню Display. Теперь вы можете наблюдать белок как целостный объект, а не набор точек и линий с промежутками между ними. Но такой вариант просмотра затрудняет изучение внутренних структур.
и) Вы имеете возможность вращать (rotate) молекулу белка при помощи левой кнопки мыши или правой кнопки мыши + shift key; перемещать (move) молекулу при помощи правой кнопки мыши; увеличивать или уменьшать масштаб изображения при помощи shift key + левой кнопки мыши.

Теперь давайте взглянем на isocitrate, лиганд, связанный с белком.

а) Выберите Wireframe в меню Display.

б) Выберите CPK в меню Colours.
в) В командной строке наберите «select ligand».
г) Выберите Spacefill в меню Display.

Теперь вы видите isocitrate. Очевидно, что он находится внутри молекулы белка.

а) В командной строке наберите «colour blue».
б) В командной строке наберите «select protein».
в) Выберите Spacefill в меню Display.

Вращая молекулу, вы можете увидеть, что молекула, окрашенная в синий цвет (isocitrate) находится в «кармане» на поверхности белка. Так происходит каталитическая реакция.

Необходимо отметить, что изменения в молекуле белка в области связывания изоцитрата вследствие мутации в ДНК может привести к нарушению взаимодействия молекул. Вероятно, что связывания не произойдёт вообще. Однако изменения в других частях молекулы белка могут не повлиять на связывание.

Лабораторная работа № 2

Цель работы:

1. Ознакомиться с пакетом программ **EMBOSS**;
2. Понять основные принципы декодирования последовательности ДНК в последовательность белка;
3. Понять принципы работы алгоритмов сравнения последовательностей;
4. Получить представление об открытой рамке считывания;
5. Освоить некоторые техники сравнения последовательностей и интерпретации результатов сравнения.

EMBOSS http://emboss.sourceforge.net/ – это "The European Molecular Biology Open Software Suite". EMBOSS является программным пакетом с открытым кодом. Он был специально разработан для нужд молекулярных биологов. Пакет EMBOSS автоматически распознаёт множество биологических форматов данных и позволяет прозрачно для пользователя получать эти данные из сетевых баз. **EMBOSS** включает в себя большое количество утилит для анализа последовательностей, формируя таким образом целостный программный пакет. **EMBOSS** – это хороший бесплатный вариант для начала работы с данными, содержащими последовательности. Для более глубокого анализа существуют коммерческие продукты с соответствующей поддержкой, техническим сервисом и обновлениями.

а) Скачайте и установите на компьютер последнюю версию EMBOSS (ftp://emboss.open-bio.org/). Документация на пакет доступна на http://emboss.sourceforge.net/apps/#list .

б) Найдите файл GenBank с последовательностью ДНК E.Coli *isocitrate dehydrogenase* J02799, который использовался в лабораторной работе №1.

в) Откройте файл и найдите в нём нуклеотидную и аминокислотную последовательности. Удостоверьтесь, что последовательность нуклеотидов – ДНК – на самом деле кодирует аминокислотную последовательность белка. Вы можете это сделать путем идентификации нуклеотидов (триплетов) в последовательности ДНК, которые точно соответствуют последовательности аминокислот. Будьте внимательны! Рамка считывания, (первый кодон) может начинаться не с первого нуклеотида, а со второго или с третьего!

- Генетический код может быть получен тут: http://helixweb.nih.gov/gcode.html. Аминокислотный код представлен на http://www.mun.ca/biochem/courses/3107/aasymbols.html Вам может потребоваться убрать строку с комментариями, пробелы и иные символы из нуклеотидной последовательности.
- Быстрый, но в то же время грубый способ найти начало рамки считывания (то есть место в последовательности нуклеотидов, соответствующее первой аминокислоте) – это попытаться использовать какие-либо функции поиска подстроки. Мы знаем (см. генетический код), что ATG – M (метионин) и в то же время ATG может являться стартовым кодоном. Также можно попытаться найти и более специфичные (более длинные) подстроки, а не только последовательность, кодирующую M – попробуйте поискать кодоны для как минимум первых шести аминокислот в последовательности нуклеотидов. Можно также осуществить поиск конца последовательности – места, где заканчиваются кодоны, кодирующие аминокислоты белка. Нужно иметь в виду, что существуют три стоп-кодона – TAG, TAA, TGA. Эти кодоны не транслируются в аминокислоты. Вы можете использовать регулярные выражения (если это возможно), поскольку генетический код вырожден, то есть одна аминокислота может кодироваться несколькими кодонами. Например, регулярное выражение для поиска Lysine – AA[A,G] (для UNIX). Упражнение: как записать Leucine регулярным выражением?

г) Вторая проблема кроется в том, что последовательность, кодирующая интересующий нас белок, на самом деле может находиться на комплементарной нити ДНК (а не той, которая приведена в базе данных, так называемые leading strand и lagging strand, strand + и strand-). Как можно получить комплементарную последовательность

известного участка ДНК? Как известно, вторая нить, последовательность ДНК обладает свойством обратной комплементарности к первой:

5' C-A-T-G-T-C-C-A 3'

3' G-T-A-C-A-G-G-T 5'

Комплементарность в молекулярной биологии – это способность нуклеотидов образовывать только строго определенные пары (см. лекции).

Используйте программу **revseq** из пакета **EMBOSS** для:

- построения обратной последовательности;
- построения комплементарной последовательности;
- построения обратно-комплементарной последовательности.

Проделайте эти шаги для последовательности J02799.

д) Сколько возможных трансляций (в последовательность аминокислот) можно получить из одной последовательности нуклеотидов? Мы не будет пытаться получить все эти варианты.

Используйте программу transeq из EMBOSS для того, чтобы транслировать J20799 в возможную последовательность аминокислот. Эта программа может проделать это для любой из трех прямых и трех обратно-комплементарных последовательностей или для всех прямых и\или обратно-комплементарных последовательностей – для этого имеется флаг –frame.

Рамка считывания – это один из трех возможных путей считывания последовательности нуклеотидов в ДНК или РНК как серии неперекрывающихся троек нуклеотидов (триплетов) в зависимости от того, на каком нуклеотиде (первом, втором или третьем) началось чтение последовательности. Например, в TGCTGCTGC имеются следующие три возможные рамки считывания: TGC TGC TGC, GCT GCT и CTG CTG.

е) Проблема, описанная выше, заключается в выборе правильно рамки считывания для трансляции. К счастью, в случае J02799 имеется только один из возможных шести вариантов, который даёт реальную и имеющую смысл ORF (open reading frame).

Используйте программу plotorf из EMBOSS для того, чтобы построить график всех возможных ORF в последовательности J02799 и идентифицировать самый длинный из них. Затем используйте эту информацию для определения начального и конечного нуклеотида в последовательности, которые имеют отношение к трансляции этого гена в соответствующий белок. Что вы можете сказать по поводу выбора ORF – правила, признаки etc? Важно помнить, что plotorf ищет участки, ограниченные старт- и стоп-кодонами. Иными словами,

эукариотические экзоны, не имеющие старт-кодона, будут им пропущены. Plotorf имеет смысл применять на последовательностях ДНК прокариотов или на м-РНК эукариотов. Справочная информация находится в http://emboss.bioinformatics.nl/cgi-bin/emboss/help/plotorf

ж) Теперь давайте посмотрим на получаемые последовательности, которые похожи на J02799. Почему нам может быть это нужно и интересно?

- Эволюция последовательностей – из-за возникающих мутаций и генных перестроек\событий в разных организмах могут оказаться похожие последовательности;
- Поиск консервативных участков;
- Поиск записей с искусственно выполненными мутациями.

Для получения последовательностей, похожих на интересующий нас участок ДНК, мы будем использовать пакет программ BLAST - http://www.ebi.ac.uk/blastall/ или http://www.ncbi.nlm.nih.gov/BLAST/. Вы должны будете сохранять результаты в виде html-страниц – они потребуются нам в дальнейшей работе.

Выберите подходящую базу данных и программу (см. справку http://www.ebi.ac.uk/Tools/sss/ncbiblast/help/index-protein.html). Используйте **blastn** в базе данных **EMBL** для поиска и сравнения нуклеотидных последовательностей и blastp в SwissProt для аминокислотных последовательностей (белков).

Попробуйте найти последовательности, похожие на J02799.

а) Вирус **H5N1 avian** (птичий грипп) является очень вирулентным и опасным не только для птиц, но и для человека. Одним из его свойств является высокая мутационная активность.
Получите любую из последовательностей H5N1 в **GenBank** и затем при помощи **BLAST** найдите похожие последовательности. Каковы характеристики и локализация найденных последовательностей?

б) Даны три последовательности неизвестного происхождения и неизвестных функций:

- мРНК:

acauuugcuu cugacacaac uguguucacu agcaaccuca aacagacacc auggugcacc ugacuccuga ggagaagucu gcgguuacug cccugugggg caaggugaac guggaugaag uuggugguga ggcccugggc aggcugcugg uggucuaccc uuggacccag agguucuuug aguccuuugg ggaucugucc acuccugaug caguuauggg caacccuaagg ugaaggcuc auggcaagaa agugcucggu gccuuuagug auggccuggc ucaccuggac aaccucaagg gcaccuuugc cacacugagu gagcugcacu gugacaagcu gcacguggau ccugagaacu ucaggcuccu gggcaacgug cuggucugug ugcuggccca ucacuuuggc aaagaauuca ccccaccagu gcaggcugcc uaucagaaag ugguggcugg uguggcuaau gcccuggccc acaaguauca cuaagcucgc uuucuugcug uccaauuucu auuaaagguu ccuuuguucc

cuaaguccaa cuacuaaacu gggggauauu augaagggcc uugagcaucu ggauucugcc uaauaaaaaa cauuuauuuu cauugc

– сДНК:

acatttgctt ctgacacaac tgtgttcact agcaacctca aacagacacc atggtgcacc tgactcctga ggagaagtct gcggttactg ccctgtgggg caaggtgaac gtggatgaag ttggtggtga ggccctgggc aggctgctgg tggtctaccc ttggacccag aggttctttg agtcctttgg ggatctgtcc actcctgatg cagttatggg caaccctaag gtgaaggctc atggcaagaa agtgctcggt gcctttagtg atggcctggc tcacctggac aacctcaagg gcacctttgc cacactgagt gagctgcact gtgacaagct gcacgtggat cctgagaact tcaggctcct gggcaacgtg ctggtctgtg tgctggccca tcactttggc aaagaattca ccccaccagt gcaggctgcc tatcagaaag tggtggctgg tgtggctaat gccctggccc acaagtatca ctaagctcgc tttcttgctg tccaatttct attaaaggtt cctttgttcc ctaagtccaa ctactaaact gggggatatt atgaagggcc ttgagcatct ggattctgcc taataaaaaa catttatttt cattgc

– Трансляция в аминокислотную последовательность:

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESF
GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSEL
HCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGV
ANALAHKYH

Используйте каждую из этих последовательностей для поиска при помощи BLAST. Оцените результаты поиска. В чём смысл полей *Score*, *E-value* и *Identities* (для ответа на вопрос нужно прокрутить страницу с результатами вниз, до строк, начинающихся с >)?

- Каким организмам принадлежат эти последовательности?
- Какова функция белка, закодированного в этом гене?
- Какие последовательности довольно близки к исследуемому участку ДНК? Каким организмам они принадлежат и какова их функция?
- Одинаковые ли результаты получились в трех поисковых запросах?
- Аккуратно удалите 21 нуклеотид из последовательностей мРНК и сДНК и *соответствующие* аминокислоты из последовательности белка. Изменило ли это результаты поиска BLAST?
- Аккуратно удалите некоторые нуклеотиды\аминокислоты в конце последовательностей и попробуйте провести BLAST снова.
    - Получили ли Вы те же результаты? Каковы значения *scores* и *identities*?
    - Сколько нуклеотидов\аминокислот нужно удалить для того, чтобы поиск перестал выдавать прежние результаты?
    - Что произойдёт, если Вы удалите нуклеотиды\аминокислоты из середины последовательностей?
    - Что произойдёт с трансляцией ДНК в аминокислотную последовательность, если Вы удалите 1\2\3 нуклеотида из середины последовательности?

Лабораторная работа № 3
Парное выравнивание

Лабораторная работа № 2 была посвящена поиску по базам данный при помощи сервиса **BLAST**, который производит парное сравнение последовательностей и возвращает оценку выравнивания, E-value и собственно выровненные последовательности. BLAST может производить парное очень быстро. Например, поиск похожих последовательностей с использованием участка в 1568 нуклеотидов или 416 аминокислот из J02799 по базам последовательностей GenBank+EMBL+DDBJ+PDB (без EST, STS, GSS), то есть среди более чем 3,7 миллионов последовательностей (~16,5 млрд. нуклеотидов или аминокислот) производится за вполне приемлемое время. BLAST выполняет это, разумеется, не производя глобальное выравнивание, но при помощи определенных эвристических алгоритмов. Однако нужно понимать, что иные подходы могут быть иногда более эффективными, например, точные (не эвристические) методы, подходы, базирующиеся на глобальном выравнивании, глобальное и локальное выравнивание и даже точечные диаграммы – эффективные способы визуального сравнения двух последовательностей.

Мы будем производить сравнение последовательностей при помощи пакета EMBOSS:

а) глобальные выравнивания (см. лекции);

б) локальные выравнивания;

в) работа с точечным графиком – визуализация выравнивания.

Для выполнения лабораторной работы вам понадобится пакет EMBOSS, установленный на локальном компьютере. Документация пакета доступна в сети Интернет.

Данные, необходимые для работы: последовательность белка hemoglobin http://en.wikipedia.org/wiki/Hemoglobin (его β-chain, доступна по адресу http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_HUMAN]).

1. ***needle*** – программа, выполняющая глобальное выравнивание http://en.wikipedia.org/wiki/Sequence_alignment#Global_alignment по алгоритму Needleman-Wunsch http://en.wikipedia.org/wiki/Needleman-Wunsch_algorithm.

а) Вы можете использовать *needle* с параметрами для выполнения выравнивания и оценки его результатов, а также для генерации таблицы с оценками выравнивания, идентичности и похожести последовательностей и количеством вставленных промежутков для:

– человеческих последовательностей (по умолчанию);

– горилла;

– кролик;
– свинья.

б) Попробуем сконструировать филогенетическое дерево для человека, гориллы, кролика и свиньи на основе оценок глобального выравнивания (наподобие приведенного на рисунке 1). Для построения дерева нам понадобится создать таблицу 4x4 (таблица 1) и заполнить её оценками выравнивания (являющимися, в сущности, показателями эволюционного расстояния между последовательностями).

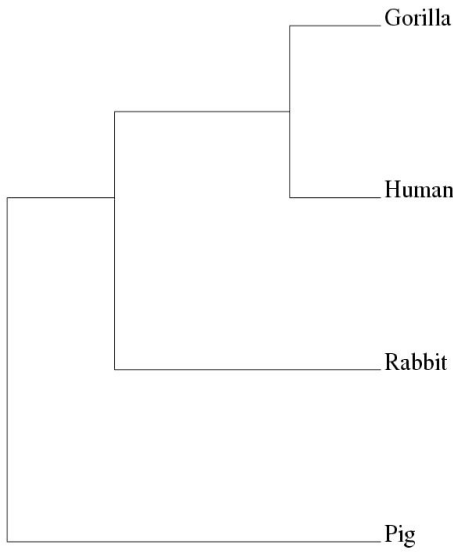

Рисунок 1. Пример укорененного филогенетического дерева

Таблица 1.

	Human	Gorilla	Rabbit	Pig
Human	x	x	x	x
Gorilla		x	x	x
Rabbit			x	x
Pig				x

Вы также можете попробовать использовать значения *sequence similarity* (для белков), заполнив предлагаемую таблицу еще раз. Будет ли построено такое же дерево, как при использовании *sequence identity*?

2. Используя навыки, полученные в результате выполнения заданий лабораторной работы № 2, скачайте аминокислотные и нуклеотидные последовательности *human isocitrate dehydrogenase*. Используя *needle*,

проведите сравнение этих последовательностей и уже имеющейся *isocitrate dehydrogenase* от E.coli (J02799).

3. Повторите п.2, используя программу *stretcher* (алгоритм Myers-Miller, работает быстрее, но менее точно, чем классический алгоритм). Отметьте различия в:
   а) score, значениях *sequence identity, sequence similarity,*количестве вставленных промежутков, иные отличия;
   б) времени выполнения (имеется ли достаточно сильная разница)?
4. Точечный график, по всей видимости, самый старый метод сравнения последовательностей (рисунок 2). Такой график – это визуальное представление похожих или идентичных участков в двух последовательностях. В таком представлении длина окна может оказаться фиксированной, также, как и длины двух последовательностей (но не всегда). Всякий раз, когда символ (или серия последовательно идущих символов – окно) одной последовательности идентичен\похож (в биологическом смысле) на символ из другой последовательности, на график наносится точка или короткая диагональ в соответствующей позиции.

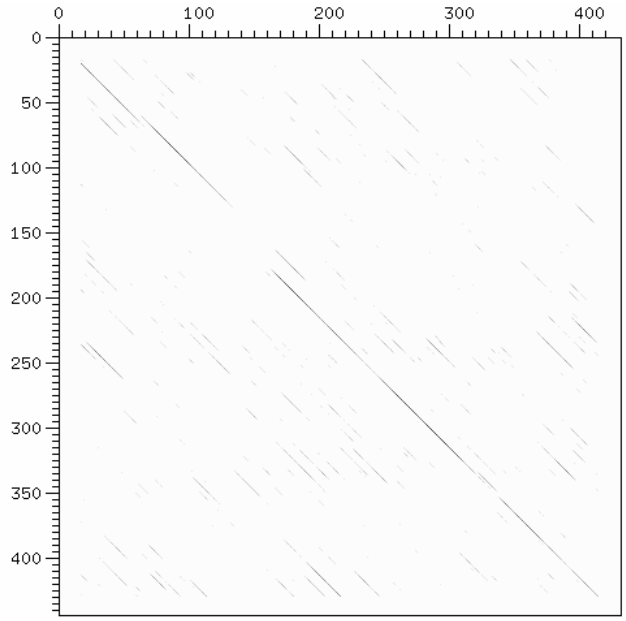

Рисунок 2. Точечный график для оценки идентичности двух последовательностей

Таким образом, когда две последовательности имеют похожие\идентичные участки по своей длине, диагональные линии, идущие от одного угла графика к противоположному, показывают эти участки. Это может быть главная диагональ (из района нуля и до конца обеих последовательностей), возможно, с перерывами, или отдельные

линии в разных местах графика. Попробуйте объяснить физический и биологический смысл коротких линий, рассеянных по графику.

Программа **dotmatcher** отображает точечный график выравнивания двух последовательностей. Позволяет менять длину слова – минимальное количество подряд совпадающих символов.
Программа **Dotter** – более интерактивна по сравнению с предыдущей. Программа и документация: http://bioinformatist.org/index.php/Dotter и http://210.86.230.110/bioinfo/material/20070822_seq/Dotter%20binaries/Dotter%20info.pdf. Программы, строящие точечные графики, позволяют быстро найти консервативные и изменчивые участки в двух последовательностях.

Используйте одну из программ, строящих точечные графики, для визуального сравнения:

а) *Isocitrate dehydrogenase* человека и E.coli;
б) *Haemoglobin* от двух разных организмов;
в) любой последовательности с ней же;
г) последовательность с её обратной копией. Вам потребуется программа ***revseq*** (см. лабораторную работу № 2);
д) соедините в текстовом редакторе две очень разных последовательности в различном порядке (вопрос по ходу: можно ли соединить нуклеотидную и аминокислотную последовательности? Например, соедините последовательности *haemoglobin* (H) и *isocitrate dehydrogenase* (I) в порядке HI и IH. Сравните получившиеся синтетические последовательности (называемые в биологии химерами) с использованием метода глобального и локального выравнивания (в соответствующих программах);
е) попробуйте сделать это для:
   – нуклеотидной последовательности;
   – аминокислотной последовательности.
ж) в чём разница выравниваний?
з) Попробуйте изменить параметр *thershold*. Получится ли получить лучшее представление совпадающих регионов.

5. Программа ***water*** **и** ***matcher*** рассчитывают локальное выравнивание (http://en.wikipedia.org/wiki/Sequence_alignment#Local_alignment) – поиск похожих\идентичных регионов (локальных участков) в двух (в случае парного выравнивания) последовательностях по всей их длине. Методы локального выравнивания бывают очень полезны для поиска по базам данных и для решения других задач, например, для поиска небольших идентичных или похожих участков у белков – доменов.

Используйте эти программы для выравнивания некоторых из последовательностей, с которыми сегодня работали. Получилось ли обнаружить качественное локальное выравнивание?

Приложения к лабораторной работе № 3

Globin Sequences (Beta-Chain)

Human HEMOGLOBIN BETA CHAIN.(HBB_HUMAN) SEQUENCE 146 AA; 15867 MW;

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAH
KYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_HUMAN]

Human sickle beta-hemoglobin SEQUENCE 147 AA

MVHLTPVEKSAVTAXWGKVNVDEVGGEALGRLLVVYPWTQRFFESFG
DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCD
KLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANAL
AHKYH

http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=protein&dopt=GenPept&list_uids=183945

Gorilla gorilla gorilla (Lowland gorilla).HEMOGLOBIN BETA CHAIN. (HBB_GORGO) SEQUENCE 146 AA; 15839 MW;

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
HVDPENFKLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAH
KYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_GORGO]

Hylobates lar (Common gibbon).HEMOGLOBIN BETA CHAIN.(HBB_HYLLA) SEQUENCE 146 AA; 15925 MW;

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDK
LHVDPENFRLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALA
HKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_HYLLA]

Presbytis entellus (Hanuman langur).HEMOGLOBIN BETA CHAIN.(HBB_PREEN) SEQUENCE 146 AA; 15895 MW;

VHLTPEEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGD
LSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCD
KLHVDPENFRLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_SEMEN]

Colobus badius (Red colobus).HEMOGLOBIN BETA CHAIN.(HBB_COLBA) SEQUENCE 146 AA; 15870 MW;

VHLTPDEKNAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGD
LSTADAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCD
KLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_COLBA]

Oryctolagus cuniculus (Rabbit).HEMOGLOBIN BETA-1 AND BETA-2 CHAINS.(HBB_RABIT) SEQUENCE 146 AA; 16001 MW;

VHLSSEEKSAVTALWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDL
SSANAVMNNPKVKAHGKKVLAAFSEGLSHLDNLKGTFAKLSELHCDKL
HVDPENFRLLGNVLVIVLSHHFGKEFTPQVQAAYQKVVAGVANALAHK
YH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_RABIT]

Bison bonasus (European bison).HEMOGLOBIN BETA CHAIN.(HBB_BISBO) SEQUENCE 145 AA; 15976 MW;

MLTAEEKAAVTAFWGKVHVDEVGGEALGRLLVVYPWTQRFFESFGDL
SSADAVMNNAKVKAHGKKVLDSFSNGMKHLDDLKGTFAALSELHCDK
LHVDPENFKLLGNVLVVVLARHFGKEFTPVLQADFQKVVTGVANALAH
RYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_BISBO]

Sus scrofa (Pig).HEMOGLOBIN BETA CHAIN.(HBB_PIG) SEQUENCE 146 AA; 16034 MW;

VHLSAEEKEAVLGLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGD
LSNADAVMGNPKVKAHGKKVLQSFSDGLKHLDNLKGTFAKLSELHCD
QLHVDPENFRLLGNVIVVVLARRLGHDFNPNVQAAFQKVVAGVANAL
AHKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_PIG]

Lutra lutra (European river otter).HEMOGLOBIN BETA CHAIN.(HBB_LUTLU) SEQUENCE 146 AA; 15950 MW;

VHLTGEEKAAVTSLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGD
LSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDK
LHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALA
HKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_LUTLU]

Theropithecus gelada (Gelada baboon).HEMOGLOBIN BETA CHAIN.(HBB_THEGE) SEQUENCE 146 AA; 15925 MW;

VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGD
LSSPAAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCD

KLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_THEGE]

Лабораторная работа № 4

Множественное выравнивание

В этой работе вам предстоит выполнить множественные выравнивания последовательностей. Программа, в которой обычно выполняется выравнивание – ClustalW. Документацию по этому инструменту можно найти на http://www.ebi.ac.uk/2can/tutorials/nucleotide/clustalw.html .

Прошлые лабораторные работы познакомили нас с методиками сравнения последовательностей, базирующимися на локальном и глобальном выравнивании. На этот раз мы будем использовать различные инструменты и различные матрицы оценок для этого.

Для чего нужно Multiple sequence alignment (MSA)? http://en.wikipedia.org/wiki/Multiple_Sequence_Alignment

Множественное выравнивание может быть использовано при изучении эволюционных отношений между белками или последовательностями ДНК. Так как изменения в нуклеотидных последовательностях в процессе эволюции накапливаются, то можно проанализировать гомологичные гены (гены, имеющие одно эволюционное начало) от разных организмов и затем сравнить их при помощи MSA. Результаты такого выравнивания могут быть для определения консервативных, сохранившихся, общих областей, то есть областей, изменение которых может повлиять на функцию. Это бывает очень полезно для планирования экспериментов определения и модификации функции определенных белков, для предсказания структуры и функции белка и для определения новых членов известных белковых семейств (то есть для классификации).

MSA бывает полезна для представления свойств набора (семейства) последовательностей. Эти свойства (дескрипторы) могут в дальнейшем использоваться в качестве шаблонов для поиска по базам данных последовательностей, похожих на имеющийся набор (bootstrapping).

MSA программы и техники

Прогрессивные стратегии для MSA:

Общий подход (эвристический) для множественного выравнивания – это последовательное выравнивания пар последовательностей. Все техники этого подхода подразумевают выполнение двух этапов: попарное выравнивание всех последовательностей с последующей кластеризацией, с построением guide tree и выравнивание результатов парных выравниваний в соответствии с guide tree.

Одна из самых популярных программ для MSA – ClustalW (и ее модификации). Она использует вышеописанный подход для выполнения MSA последовательностей ДНК или белков. Она определяет лучшее парное выравнивание и помещает его на первые позиции, выделяет цветом и т.д. Программа также генерирует филограмму (филогенетическое дерево) и кладограмму (граф сестринского соответствия между известными группами без учёта фактора времени в эволюции). Это может быть полезно для изучения эволюционных взаимосвязей в наборе последовательностей.

1. Для выполнения MSA используйте **ClustalW** на сайте EBI http://www.ebi.ac.uk/Tools/msa/clustalw2/. Самый простой формат последовательностей при использовании **ClustalW - FASTA** http://www.ebi.ac.uk/Tools/sss/fasta/:

>SEQUENCE_NAME1 PLUS ANY OTHER COMMENTS
SEQUENCE1 (может быть в несколько строк)
>SEQUENCE_NAME2 PLUS ANY OTHER COMMENTS
SEQUENCE2 (может быть в несколько строк)
...
>SEQUENCE_NAMEn PLUS ANY OTHER COMMENTS
SEQUENCEn (может быть в несколько строк)

Можно также использовать **ClustalW** локально в виде командной строки (на UNIX) или многочисленные реализации для Windows.

а) Запустите **ClustalW** (в случае локального использования – без параметров)
б) Выберите тип последовательности Protein.
в) Загрузите последовательности в форму в формате Fasta передайте на обработку fasta-файл с диска (поле Upload).
г) Установите переключатель Slow/Fast в Slow – полный точный MSA.
д) Вы можете получить результаты по e-mail, отметив соответствующий checkbox или увидеть их на открывшейся веб-странице.

Используйте последовательности белков из семейства глобинов, представленные в конце лабораторной работы (как список или fasta-файл).
Символы консенсуса:
Полученное выравнивание будет содержать следующие символы под каждым блоком последовательностей, показывающие степень консервации символа в каждой колонке:
* - все аминокислоты/нуклеотиды в этой колонке идентичны во всех последовательностях;
: - имеются консервативные замены (см. таблицу цветов на странице);
. – имеются полуконсервативные замены.
Вам нужно обратить внимание на:
а) Файл выравнивания (лучше с цветами);

б) Филограмму и кладограмму (отметьте разницу между ними);
в) Выравнивание в редакторе Jalview (может быть запущен на странице результатов ClustalW http://www.jalview.org/).

Вы также можете построить выравнивание других интересующих вас последовательностей, найденных в базе NCBI (http://www.ncbi.nlm.nih.gov/) и сохраненных в fasta-формате. Например:

- если вас интересуют последовательности вируса птичьего гриппа, выполните поиск по H5N1. В случае ВИЧ – поиск по HIV-2;
- вам потребуется выбрать последовательности (базы данных), с которыми вы хотите поработать – это либо нуклеотидные либо аминокислотные последовательности и соответствующие базы данных;
- установите режим отображения FASTA;
- сохраните данные в текстовом файле;
- отредактируйте сохраненный файл: удалите все пробелы из первой строки с комментариями и иные символы, которые вы не хотите видеть в результатах выравнивания и убедитесь в том, что строки комментариев для каждой последовательности в файле уникальна, при необходимости отредактируйте эти строки для того, чтобы последовательности можно было различать.

В конце лабораторной работы – последовательности H5N1 с их локализацией, видовой принадлежностью, местом обнаружения вируса и прочими данными.

Вы также можете использовать одну интересующую вас последовательность (например, глобин) и затем использовать BLAST для поиска похожих последовательностей. После этого вы можете сохранить все или некоторые из них (например, от различных видов) в формате fasta и использовать полученный файл для MSA.

Самостоятельно изучите **2can ClustalW** tutorial (http://www.ebi.ac.uk/2can/tutorials/nucleotide/clustalw.html).

Сейчас мы рассмотрим выравнивание нескольких нуклеотидных последовательностей гена белка тропомиозина (http://www.biology-online.org/dictionary/Tropomyosin), номера доступа которого приведены ниже:

BF056441 BE8487196 BF022813 BF452255 BG089808 BG147728 BI817778 AF186109 AF186110 AF310722 AF362886 AF362887 AF087679 SSAJ803 SSAJ804

Эти последовательности (в конце лабораторной работы) имеют больше различий, чем исследованные нами глобины.

Попробуйте использовать другие программы для MSA, выполнив выравнивания на ваших данных.

- **T-coffee** (http://www.igs.cnrsmrs.fr/Tcoffee/tcoffee_cgi/index.cgi) – эта программа выполняет более точное выравнивание по сравнению с ClustalW для последовательностей с идентичностью < 30%. Работает медленнее;
- **Muscle** (http://phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py) – результаты по e-mail.

Если вы хотите использовать пакет **EMBOSS:**

Документация по **EMBOSS** доступна на http://emboss.sourceforge.net/;

Пакет EMBOSS доступен как online, так и для использования локально.

2. Исследование из лабораторной работы № 3.

Программы ***water*** **и** ***matcher*** из пакета **EMBOSS** выполняют локальное выравнивание двух последовательностей и находят в них похожие участки. (http://en.wikipedia.org/wiki/Sequence_alignment#Local_alignment)

Выравниваемые последовательности могут быть разной длины. Методы локального выравнивания могут быть очень полезны при поиске в базах данных и других случаях, например, при поиске наиболее совпадающих небольших участков совпадений, например, доменов в белках.

Используя эти программы, проведите выравнивание последовательностей, приведенных ниже. Удалось ли выявить хорошее локальное выравнивание?

3. Программы, полезные для поиска последовательностей (таблица 2), можно найти на EBI tools page (http://www.ebi.ac.uk/Tools/index.html). Попробуйте поработать с некоторыми из них. Сравните результаты их работы. Одинаковы ли они при одном и том же запросе\последовательности? Оцените время работы различных приложений для поиска последовательностей и результаты\подробности в отчетах. Подробная online помощь по вопросам гомологии и сравнению последовательностей - http://www.ebi.ac.uk/Tools/sss/ .

Таблица 2.

Fasta3 http://www.ebi.ac.uk/Tools/sss/fasta/	Sequence similarity and homology searching against nucleotide and protein database using Fasta3
WU-Blast2 http://www.ebi.ac.uk/Tools/ss	Washington University

s/wublast/	blast2 (blast 2.0 with gaps)
NCBI-Blast2 http://www.ebi.ac.uk/Tools/sss/ncbiblast/	NCBI blast2 (blastall) program
MPsrch http://www.ebi.ac.uk/Tools/MPsrch/	Edinburgh University's new implementation of the Smith and Waterman algorithm
Scanps2.3 http://www.ebi.ac.uk/Tools/scanps/	Version 2.3 of Scanps.Fast implementation of the true Smith & Waterman algorithm for protein database searches

4. Попробуйте провести различные поиски и выравнивания последовательностей по базам данных, используя при этом различные матрицы расстояний. Online помощь по выбору матриц – http://www.ebi.ac.uk/2can/tutorials/matrices.html .
5. Повторите поиск, используя BLAST с различным expected thresholds. Это дает возможность уменьшить или увеличить список совпадающих последовательностей\участков. Online помощь по использованию BLAST - http://blast.ncbi.nlm.nih.gov/Blast.cgi .
6. Вы можете производить поиск раздельно по целым геномам или протеомам, используя сервер Proteomes & Genomes Fasta3 http://www.ebi.ac.uk/Tools/sss/fasta/.
7. Попробуйте поэкспериментировать с ним.

Приложение к лабораторной работе № 4

Последовательности и материалы, необходимые для выполнения лабораторной работы.

Globin Sequences (Beta-Chain)

Human HEMOGLOBIN BETA CHAIN.(HBB_HUMAN) SEQUENCE 146 AA; 15867 MW;
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAH
KYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_HUMAN]

Human sickle beta-hemoglobin SEQUENCE 147 AA

MVHLTPVEKSAVTAXWGKVNVDEVGGEALGRLLVVYPWTQRFFESFG
DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCD
KLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANAL
AHKYH

http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=protein&dopt=GenPept&list_uids=183945

Gorilla gorilla gorilla (Lowland gorilla).HEMOGLOBIN BETA CHAIN. (HBB_GORGO) SEQUENCE 146 AA; 15839 MW;

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
HVDPENFKLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAH
KYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_GORGO]

Hylobates lar (Common gibbon).HEMOGLOBIN BETA CHAIN.(HBB_HYLLA) SEQUENCE 146 AA; 15925 MW;

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDK
LHVDPENFRLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALA
HKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_HYLLA]

Presbytis entellus (Hanuman langur).HEMOGLOBIN BETA CHAIN.(HBB_PREEN) SEQUENCE 146 AA; 15895 MW;

VHLTPEEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGD
LSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCD

KLHVDPENFRLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_SEMEN]

Colobus badius (Red colobus).HEMOGLOBIN BETA CHAIN.(HBB_COLBA) SEQUENCE 146 AA; 15870 MW;

VHLTPDEKNAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGD
LSTADAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCD
KLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_COLBA]

Oryctolagus cuniculus (Rabbit).HEMOGLOBIN BETA-1 AND BETA-2 CHAINS.(HBB_RABIT) SEQUENCE 146 AA; 16001 MW;

VHLSSEEKSAVTALWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDL
SSANAVMNNPKVKAHGKKVLAAFSEGLSHLDNLKGTFAKLSELHCDKL
HVDPENFRLLGNVLVIVLSHHFGKEFTPQVQAAYQKVVAGVANALAHK
YH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_RABIT]

Bison bonasus (European bison).HEMOGLOBIN BETA CHAIN.(HBB_BISBO) SEQUENCE 145 AA; 15976 MW;

MLTAEEKAAVTAFWGKVHVDEVGGEALGRLLVVYPWTQRFFESFGDL
SSADAVMNNAKVKAHGKKVLDSFSNGMKHLDDLKGTFAALSELHCDK
LHVDPENFKLLGNVLVVVLARHFGKEFTPVLQADFQKVVTGVANALAH
RYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_BISBO]

Sus scrofa (Pig).HEMOGLOBIN BETA CHAIN.(HBB_PIG) SEQUENCE 146 AA; 16034 MW;

VHLSAEEKEAVLGLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGD
LSNADAVMGNPKVKAHGKKVLQSFSDGLKHLDNLKGTFAKLSELHCD
QLHVDPENFRLLGNVIVVVLARRLGHDFNPNVQAAFQKVVAGVANAL
AHKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_PIG]

Lutra lutra (European river otter).HEMOGLOBIN BETA CHAIN.(HBB_LUTLU) SEQUENCE 146 AA; 15950 MW;

VHLTGEEKAAVTSLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGD
LSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDK

LHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALA
HKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_LUTLU]

Theropithecus gelada (Gelada baboon).HEMOGLOBIN BETA CHAIN.(HBB_THEGE) SEQUENCE 146 AA; 15925 MW;

VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGD
LSSPAAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCD
KLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH

http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[swissprot-id:HBB_THEGE]

Глобины в формате FASTA:
>Human
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAH
KYH
>Human_sickle
MVHLTPVEKSAVTAXWGKVNVDEVGGEALGRLLVVYPWTQRFFESFG
DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCD
KLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANAL
AHKYH
>Gorilla
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
HVDPENFKLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAH
KYH
>Common_gibbon
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDK
LHVDPENFRLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALA
HKYH
>Hanuman_langur
VHLTPEEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGD
LSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCD
KLHVDPENFRLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH
>Red_colobus
VHLTPDEKNAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGD
LSTADAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCD

KLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH

>Rabbit

VHLSSEEKSAVTALWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDL
SSANAVMNNPKVKAHGKKVLAAFSEGLSHLDNLKGTFAKLSELHCDKL
HVDPENFRLLGNVLVIVLSHHFGKEFTPQVQAAYQKVVAGVANALAHK
YH

>European_bison

MLTAEEKAAVTAFWGKVHVDEVGGEALGRLLVVYPWTQRFFESFGDL
SSADAVMNNAKVKAHGKKVLDSFSNGMKHLDDLKGTFAALSELHCDK
LHVDPENFKLLGNVLVVVLARHFGKEFTPVLQADFQKVVTGVANALAH
RYH

>Pig

VHLSAEEKEAVLGLWGKVNV DEVGGEALGR LLVVYPWTQR
FFESFGDLSN ADAVMGNPKV KAHGKKVLQS
FSDGLKHLDNLKGTFAKLSELHCDQLHVDPENFRLLGNVIVVVLARRLG
HDFNPNVQAAFQKVVAGVANALAHKYH

>European_river_otter

VHLTGEEKAAVTSLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGD
LSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDK
LHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALA
HKYH

>Gelada_baboon

VHLTPEEKNAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGD
LSSPAAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCD
KLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANAL
AHKYH

H5N1 (вирус птичьего гриппа) в формате Fasta:

>1A_Cygnus_olor_Astrakhan_Ast05_2_3_2005

>2A_Goose_Guangdong_1_96

>3A_Hong_Kong_1073_99

>4A_chicken_Hebei_326_2005

>5A_chicken_Hebei_718_2001

>6A_chicken_Hebei_326_2005

>7A_chicken_Hebei_718_2001

>8A_chicken_Hebei_326_2005

>9A_chicken_Hebei_108_02

Последовательности тропомиозинов.

>embl:BF056441 BF056441; 7k05a04.x1 NCI_CGAP_GC6 Homo sapiens cDNA clone IMAGE:3443238 3' similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN, FIBROBLAST NON-MUSCLE TYPE; mRNA sequence.

>embl:BE848719 BE848719; uw40c07.y1 Soares_thymus_2NbMT Mus musculus cDNA clone IMAGE:3419148 5' similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN, FIBROBLAST NON-MUSCLE TYPE ; mRNA sequence.
>embl:BF022813 BF022813; uw40c07.x1 Soares_thymus_2NbMT Mus musculus cDNA clone IMAGE:3419148 3' similar to SW:TPM4_RAT P09495 TROPOMYOSIN 4, EMBRYONIC FIBROBLAST ISOFORM ; mRNA sequence.
>embl:BF452255 BF452255; uz86d11.y1 NCI_CGAP_Lu29 Mus musculus cDNA clone IMAGE:3675957 5' similar to SW:TPM4_RAT P09495 TROPOMYOSIN 4, EMBRYONIC FIBROBLAST ISOFORM ; mRNA sequence.
>embl:BG089808 BG089808; mab82b11.x1 NCI_CGAP_BC3 Mus musculus cDNA clone IMAGE:3976676 3' similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN, FIBROBLAST NON-MUSCLE TYPE ; mRNA sequence.
>embl:BG147728 BG147728; mab53f06.x1 Soares_NMEBA_branchial_arch Mus musculus cDNA clone IMAGE:3974147 3' similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN, FIBROBLAST NON-MUSCLE TYPE; mRNA sequence.
>embl:BI817778 BI817778; G3-F20 Axolotl Lambda Zap Library Ambystoma mexicanum cDNA similar to Homo sapiens gb|AAG17014.1|AF186109_1 (2.0e-40), TPM4-ALK fusion oncoprotein type 2, mRNA sequence.
>embl:AF186109 AF186109; Homo sapiens TPM4-ALK fusion oncoprotein type 2 (TPM4-ALK fusion) mRNA, partial cds.
>embl:AF186110 AF186110; Homo sapiens TPM4-ALK fusion oncoprotein type 1 (TPM4-ALK fusion) mRNA, partial cds.
>embl:AF310722 AF310722; Homo sapiens tropomyosin 4-anaplastic lymphoma kinase fusion protein (TPM4-ALK) mRNA, partial cds.
>embl:AF362886 AF362886; Homo sapiens tropomyosin 4-anaplastic lymphoma kinase fusion protein major isoform mRNA, partial cds.
>embl:AF362887 AF362887; Homo sapiens tropomyosin 4-anaplastic lymphoma kinase fusion protein minor isoform mRNA, partial cds.
>embl:AF087679 AF087679; Sus scrofa tropomyosin 4 (TPM4) mRNA, complete cds.

Лабораторная работа № 5

Изучение шаблонов последовательностей

Цель этой лабораторной работы – получение опыта исследований и поиска шаблонов (паттернов) последовательностей. Потребуются материалы лекции о множественном выравнивании, шаблоны, мотивы (http://en.wikipedia.org/wiki/Sequence_motif) и профили.

Базы данных и инструменты поиска:

а) **PROSITE** (http://ca.expasy.org/prosite/) – база данных доменов и семейств белков. Она содержит биологически важные сайты, мотивы и профили, которые помогают довольно четко определить семейство (если таковое имеется), к которому принадлежит новая последовательность белка.

б) Приложение **ScanProsite** (http://ca.expasy.org/tools/scanprosite/) позволяет пользователю сканировать последовательности белков (в базе данных UniProt-SwissProt/TrEMBL, PDB или пользовательские последовательности) на наличие мотивов, профилей или шаблонов (паттернов), имеющихся в базе данных **PROSITE** или производить поиск по белковым базам данных в поисках определенных мотивов.

Большинство программ, используемых в этой лабораторной работе, находится на http://ca.expasy.org/tools/#pattern.

Некоторые программы, которые могут использоваться для поиска и изучения шаблонов\мотивов (http://bioweb.pasteur.fr/seqanal/motif/intro-uk.html):

а) **PRATT** на **EBI**, http://ca.expasy.org/tools/pratt/ или **PrattWWW** (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::pratt): инструмент для поиска мотивов (I. Jonassen). Руководство пользователя на Pratt находится по адресу http://www.ebi.ac.uk/Tools/pratt/. Pratt ищет повторяющиеся мотивы, может генерировать шаблоны - регулярные выражения в синтаксисе ProSite и может находить шаблоны и паттерны, являющимися важными, определяющими для введенных последовательностей. Кроме того, программа может улучшать, оптимизировать паттерны и шаблоны, которые находит.

б) **MEME** http://meme.sdsc.edu/meme/meme-input.html (Multiple EM for Motif Elicitation): поиск мотивов (высококонсервативных обрастей) в группах (наборах) последовательностей гомологичных ДНК или белков (T. Bailey, C. Elkan, B. Grundy).

в) **PFTOOLS** (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::pftools): PROFILE tools (P. Bucher).

г) **SMILE** (http://www-igm.univ-mlv.fr/~marsan/smile_english.html): Structured Motif Inference and Evaluation (L. Marsan, J. Allali).

Простые, структурные и распеделенные мотивы http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::smile;

а) **Prophet** (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::prophet) веб-сервер: выравнивание профилей, возможно с промежутками. **EMBOSS-prophet** (http://emboss.bioinformatics.nl/cgi-bin/emboss/prophet) – как и весь EMBOSS, может быть запущен локально;

б) **Consensus** (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::consensus): Определение консенсусных шаблонов в невыровненных последовательностях ДНК и белков. Развитая статистическая база для штрафов за промежутки (G.Z.Hertz и G.D.Stormo).

в) **PatternHunter -** http://www.bioinformaticssolutions.com/all-products/ph

Дополнительные ресурсы для поиска мотивов и шаблонов: http://bioweb2.pasteur.fr/motif/

Цитохромы P450 – семейство наиболее мощных детоксицирующих ферментов организма. Известно более 60-и ключевых форм с сотнями возможных генетических вариаций, выливающихся в огромный спектр восприимчивости к различным токсинам. Дополнительную информацию по P450 можно найти по ссылке http://www.mall-net.com/mcs/p450.html.

Биохимики высказали предположение (благодаря интуиции и опыту), что последовательности белка P450 может быть описана (характеризована) следующим мотивом: **FMFEGHDTTA.**

Этот мотив был обнаружен в наборе последовательностей Цитохрома P450 (последовательности в конце лабораторной работы).

Однако, возник вопрос, нет ли других мотивов\паттернов, которые могут характеризовать и выделять последовательности цитохромов P450 из других семейств.

В этом реальном исследовании нужно:

а) Доказать, что мотив **FMFEGHDTTA** действительно является уникальным и может идентифицировать семейство P450 – это значит, что можно использовать указанный мотив для поиска по базе ProSite для идентификации;

б) Использовать любую из найденных последовательностей для того, чтобы попытаться извлечь какую-либо дополнительную информацию о возможных мотивах\шаблонах, имеющихся в найденном сете последовательностей. Эти шаблоны будут регулярными выражениями, генерированными при помощи программы Pratt.

в) Оценить эти сгенерированные шаблоны\паттерны при помощи поиска по сету последовательностей цитохрома P450 и сету последовательностей не-P450. Оценить их описательную и дискриминантную силу.

1. Проверьте, что вышеуказанный мотив **(FMFEGHDTTA)** присутствует во всех последовательностях набора изученных цитохромов P450 (в конце лабораторной работы). Вы можете использовать приложение *patmatdb* локально (предварительно установив пакет EMBOSS) или в любой из многочисленных её online реализаций – например http://bips.u-strasbg.fr/EMBOSS/.

Помощь, документация на *patmatdb* – http://emboss.sourceforge.net/apps/release/5.0/emboss/apps/patmatdb.html.

2. Используя ScanProsite – http://ca.expasy.org/tools/scanprosite/ – (**Поиск «Motif(s) to scan for:»** – правое поле), проведите поиск мотива **FMFEGHDTTA** среди последовательностей базы SwissProt (галочка – только около UniProtKB/Swiss-Prot).
3. Загрузите найденные последовательности в формате Fasta – последовательно откройте каждую найденную последовательность и нажмите “fasta”. Создайте затем файл multi-fasta – все последовательности в одном файле.

4. Проверьте, что найденные последовательности специфичны в отношении искомого мотива – поместите их в поле *Sequences to be scanned*.

   Теперь нам нужно попытаться автоматически генерировать мотивы, определяющие (характеризующие) набор полученных последовательностей белков P450, а затем протестировать качество этих шаблонов:
5. Используйте **PRATT** (http://www.ebi.ac.uk/Tools/pratt/) – это приложение, которое автоматически составляет регулярные выражения, описывающие эти последовательности. Пример работы PRATT – http://www.ii.uib.no/~inge/papers/mdl/small.html.
   - Произведите обратный поиск по базе SwissProt с использованием сгенерированных паттернов. Какие последовательности были найдены?

     Эти поиски вернули исходные последовательности?

     Какие иные последовательности (если есть) были найдены по введенным шаблонам? Являются ли они последовательностями P450?
   - Вы можете также использовать приложение IBM **TEIRESIAS** http://cbcsrv.watson.ibm.com/Tspd.html для генерации паттернов\шаблонов.
6. Используйте набор последовательностей P450, который вы сумели идентифицировать при помощи **FMFEGHDTTA** в качестве набора ***positive examples.***
7. Случайным образом разделите этот набор на обучающую и тестовую выборки (в соотношении обучающая\тестовая между 1:1 и 1:3).
8. Используйте **PRATT** для обучения (получения) надёжного паттерна по обучающей выборке.

9. Произведите поиск в тестовом наборе (или в полной базе данных) по найденному паттерну\шаблону. Вы можете использовать приложение **EMBOSS *patmatdb*** (для поиска с использованием паттернов, предварительно найденных в обучающем наборе. Документация на *patmatdb* – http://emboss.bioinformatics.nl/cgi-bin/emboss/.
10. Создайте новый тестовый набор с *negative examples* в fasta-формате – возьмите последовательности глобинов от разных организмов (см. предыдущие лаб. работы или Приложение). Тут мы предполагаем, что последовательности глобинов **не содержат** мотив **FMFEGHDTTA.** Является ли это предположение верным? Вы можете проверить это предположение, поместив эти последовательности в поле поиска *Sequences to be scanned* на странице поиска ScanProsite, а затем произведите поиск в базе данных Prosite для того, чтобы увидеть мотивы, соответствующие последовательностям глобинов.
11. Теперь вы имеете два тестовых набора (позитивный тест-набор из п.6 и негативный тест-набор с данными о глобинах).
12. Используйте приложение EMBOSS *patmatdb* для поиска некоторых паттернов\мотивов, полученных ранее благодаря PRATT, в позитивном и негативном тест-наборах последовательностей.
13. Вы должны знать, что в каждом поиске True Positives (TP), True Negatives (TN), False Positives (FP) и False Negatives (FN) значат следующее:
    - TP – шаблон найден в позитивном наборе;
    - FN – шаблон не найден в позитивном наборе;
    - FP – шаблон найден в негативном наборе;
    - TN – шаблон не найден в негативном наборе.
14. Используйте ваши результаты для подсчета различных показателей качества для шаблона (см. rating_patterns.pdf). Убедитесь в том, что когда вы выбираете слабый шаблон, предложенный PRATT, то и показатели качества будут плохими. Но что является интуитивным показателем слабого шаблона? Можете ли вы сказать что-либо, оценить силу паттерна только глядя на него вне связи с целевыми последовательностями? Всегда ли короткие паттерны являются слабыми?
15. Проведите обратную проверку – создайте шаблоны-паттерны для глобинов (теперь они выступают в качестве позитивного набора). Используйте набор P450 в качестве негативного. Удалось ли вам обнаружить дискриминирующие, определяющие шаблоны для глобинов?

Другие программы и ресурсы

а) **PPSEARCH** (http://www.ebi.ac.uk/Tools/ppsearch/): производит поиск мотивов в исследуемой последовательности, быстро сравнивает введенную последовательность с шаблонами\паттернами, накопленными в базе данных шаблонов Prosite и пытается предсказать функцию неизвестного белка. Приложение запрашивает на вход последовательность белка. ДНК\РНК могут быть транслированы в последовательность белка при помощи *transeq* (http://www.ebi.ac.uk/Tools/emboss/).

б) База данных **EMOTIF** (http://motif.stanford.edu/ или https://bioinformatics.cs.vt.edu/emotif/) – это коллекция более чем 170000 высокоспецифичных и чувствительных мотивов последовательностей белка, представляющих консервативные свойства и биологические функции. Эти белковые мотивы были выделены при помощи более чем 7600 выравниваний последовательностей в базе данных BLOCKS+ (23 июня 2000) и всех (8244) выравниваний белковых последовательностей в базе PRINTS. Для построения использовался также алгоритм emotif-maker, разработанный Nevill-Manning с соавторами. Благодаря тому, что аминокислоты и их группы в найденных мотивах отражают критические позиции в белках, закрепленных эволюционно, поисковые алгоритмы, примененные в EMOTIF, могут идентифицировать и классифицировать весьма большую группу различных белковых последовательностей – большую, чем методы, основанные на глобальном выравнивании последовательностей. База данных мотивов\паттернов EMOTIF доступна по адресу http://motif.stanford.edu/ и https://bioinformatics.cs.vt.edu/emotif/.

– **Emotif Maker** (https://bioinformatics.cs.vt.edu/emotif/emotif-maker.html) – программа для генерации мотива\шаблона (регулярных выражений для последовательностей, соответствующих функционально важным областям) на основании результатов выравнивания;

– **Emotif Scan** (https://bioinformatics.cs.vt.edu/emotif/emotif-scan.html) – программа для поиска последовательности, соответствующей созданному регулярному выражению, в базе данных последовательностей.

– **Emotif Search** (https://bioinformatics.cs.vt.edu/emotif/emotif-search.html) – программа для сопоставления последовательности определенным участкам по базе данных блоков (*blocks* (http://blocks.fhcrc.org/blocks/help/about_blocks.html)) и отпечатков (*prints* (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php)). Блоки – это сегменты без блоков в множественном выравнивании

последовательностей, соответствующие высоко консервативным регионам (highly conserved regions) в белках. Отпечатки – это резюме белковых «отпечатков пальцев» или fingerprints – группа консервативных мотивов, используемая для описания, аннотирования, характеризации семейств белков.

в) PHI-BLAST (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome или http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::phiblast) – это BLAST, использующий паттерны определенного формата (формат похож на Prosite, см. сайт PHI-BLAST) для инициации поиска с использованием BLAST. Попробуйте провести поиски с использованием PHI-BLAST.

Приложение к лабораторной работе № 5

Последовательности и материалы, необходимые для выполнения лабораторной работы

Цитохромы P450:

>1cpt00
>1bvyA0
>1eupA0
>1akd00
>1cmnA0

Globin Sequences (Beta-Chain)

Последовательности глобинов для лабораторной работы № 5 аналогичны таковым для лабораторной работы № 4.

Лабораторная работа № 6
Филогенетические деревья

Цель лабораторной работы – получение практических навыков и понимания филогенетических деревьев и их построения.

Программное обеспечение

Polytree: Вы можете использовать web-версию программы по адресу http://www.dina.dk/~sestoft/bsa/Match7Applet.html

Phylip: Онлайн доступ к сервису Phylip: http://bioweb2.pasteur.fr/phylogeny/intro-en.html. Список программ, входящих в пакет Phylip, находится по адресу http://evolution.genetics.washington.edu/phylip/programs.html. С сервера Phylip можно скачать (http://evolution.genetics.washington.edu/phylip/getme.html) и установить этот пакет для использования на локальном компьютере. Там же находятся

и исходные коды пакета, инструкции и примеры и ссылки на другие программы для филогенетических исследований.

BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi и его разновидность PSI-Blast (меню distance tree или results).

Упражнения

А. Построение филогении в семействе глобинов (бета-цепи)

Используйте **ClustalW/Phylip** для вычисления дистанций и матриц замен.

а) Последовательности глобинов в формате fasta, необходимые для проведения работы, приведены в приложении.

б) Откройте любую из реализаций ClustalW (онлайн версии, локальные реализации или в EMBOSS (программа *emma*, которая является оболочкой ClustalW)).
- Загрузите файл fasta, содержащий последовательности глобинов, при помощи кнопки Browse.
- Оставьте настройки, установленные по умолчанию. Формат вывода должен быть установлен в PHYLIP.
- Нажмите Submit – алгоритм выравнивания будет запущен.
- Скопируйте выравнивание в новый текстовый файл и сохраните его (это будет служить входными данными для программ Phylip).

в) Откройте каталог, в котором вы сохранили файл выравнивания в формате Phylip.

г) С помощью программы protdist из пакета Phylip (или онлайн версии: http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::protdist и http://www.cbib.u-bordeaux2.fr/pise/protdist.html) рассчитайте матрицу расстояний, соответствующую построенному выравниванию глобинов. Введите имя сохранённого выравнивания в формате Phylip в окно protdist.

д) Сохраните результаты работы protdist в файле matrix.

е) Теперь мы можем построить деревья по алгоритму neighbor-joining, используя программу neighbor из пакета Phylip или на сервере http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::neighbor. Введите матрицу расстояний в окно ввода. Заметьте, что на данный сервис позволяет выбрать метод построения дерева также и по алгоритму UPGMA. Вы можете попробовать построить деревья по обоим алгоритмам и сравнить их.

ж) Neighbor выводит два различных файла – outfile (визуализацию дерева) и treefile (расстояния).

з) Вы можете построить неукорененное дерево для neighbor joining , используя программу drawtree или укорененное (для UPGMA) при помощи drawgram. Пригодится созданный treefile.

и) Вы также можете построить филогенетические деревья, используя программы fitch или kitsch из пакета Phylip (или онлайн версии http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::fitch и http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::kitsch). Для визуализации используйте уже известные drawtree или drawdram.

Можно пройти более простым автоматизированным путем, поместив последовательности в поле ввода на сервере http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::protein_distance_phylogeny

B. Начало и эволюция ВИЧ

Данные:

Нуклеотидные последовательности ВИЧ в формате fasta (в конце лаб. работы). Эти последовательности ВИЧ подтипа C из Южной Африки (последовательности C.ZA) и последовательности из Индии (C.IN) или

Fasta-файлы, содержащие аминокислотные последовательности для envelope, gag и pol белков изолятов, приведенных ниже (см. последовательности в *fasta* в приложении и http://artedi.ebc.uu.se/course/UGSBR/hiv/):

Таблица 3.

no	**Isolate**	**Accession no (GenBank, ENTREZ)**	**Subtype/animal**
1	HIV-1_{ELI}	K03454 http://www.ncbi.nlm.nih.gov/nuccore/K03454	D
2	HIV-1_{LAI}	K02013 http://www.ncbi.nlm.nih.gov/nuccore/K02013	B
3	HIV-1_{NDK}	M27323 http://www.ncbi.nlm.nih.gov/nuccore/M27323	D
4	HIV-2_{D205}	X61240 http://www.ncbi.nlm.nih.gov/nuccore/X61240	B
5	HIV-2_{ROD}	M15390 http://www.ncbi.nlm.nih.gov/nuccore/M15390	A
6	HIV-2_{ST}	M31113 http://www.ncbi.nlm.nih.gov/nuccore/M31113	A
7	HIV-2_{UCI}	L07625 http://www.ncbi.nlm.nih.gov/nuccore/L07625	B
8	SIV_{mac}	M19499 http://www.ncbi.nlm.nih.gov/nuccore/M19499	macaque

9	SIV_{cpz}	X52154 http://www.ncbi.nlm.nih.gov/nuccore/X52154	chimpanzee
10	SIV_{agm}	M58410 http://www.ncbi.nlm.nih.gov/nuccore/M58410	African green monkey
11	SIV_{man}	X14307 http://www.ncbi.nlm.nih.gov/nuccore/X14307	mangabey

а) Проведите филогенетическое исследование этих последовательностей (MSA по ClustalW с выводом в Phylip), аналогичное проделанному в п. А.

б) Используя подходящие приложения из пакета Phylip:
- Постройте матрицу расстояний с помощью *dnadist* (почему?);
- Постройте дерево;
- Визуализируйте построенное в п.b филогенетическое дерево.

C. Начало и эволюция H5N1

Самостоятельно постройте филогенетические деревья по алгоритму UPGMA и Neighbor Joining для различных последовательностей ДНК птичьего гриппа (см. приложение). Объясните различия в построенных деревьях.

D. Метод максимальной бережливости Maximum Parsimony (MP)

В этом упражнении вы будете реконструировать филогенетическое дерево, используя программу ***Dnapars*** из пакета **Phylip.** Как и в предыдущих упражнениях, вы можете установить **Phylip** на компьютер и использовать это ПО локально (http://evolution.genetics.washington.edu/phylip/install.html) или воспользоваться онлайн сервисом, предоставляющим функционал **Phylip** (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::dnapars)**.**

Справочная информация по ***Dnapars*** – http://evolution.genetics.washington.edu/phylip/doc/dnapars.html.

Важно: ***Dnapars*** используется только для реконструкции филогенетических деревьев по методу MP для последовательностей ДНК. Для работы с последовательностями белков нужно использовать программу *Protpars* (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::protpars).

а) Установите Phylip на компьютер или откройте http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::dnapars (дальнейшие шаги приведены для случая использования онлайн-сервиса).

б) Постройте выравнивание последовательностей из файла cytb.txt (см. приложение) и сохраните его в формате, поддерживаемом Phylip.

в) Введите выровненные последовательности в окно ввода.

г) Оставьте параметры по умолчанию и нажмите **RUN.** Введите действующий адрес электронной почты.

д) Посмотрите на построенное дерево. Какие данные вывел сервис? Проанализируйте и объясните их. Сохраните дерево и сопутствующие результаты.

е) Постройте дерево последовательностей из cytb.txt (приложение), используя **Phylip** и алгоритм NJ. Сохраните дерево. Сравните филогенетические деревья.

ж) В поле *Tree file* выберите *view with archaeopteryx*. Сравните построенные деревья. Похоже ли это дерево на полученное при помощи ClustalW?

з) Bootstrap анализ.

- Откройте ClustalW (http://clustalw.ddbj.nig.ac.jp/)
- Поместите в окно ввода последовательности из cytb.txt (приложение)
- В поле Bootstrap отметьте «ON».
- Нажмите «Submit».
- Сохраните на диске файл выравнивания *.aln.
- Выберите в пакете **Phylip** раздел *Phylogeny – tree analyser* и проанализируйте построенное выравнивание и bootstrap значения.

Приложения к лабораторной работе № 6:
Globin Sequences (Beta-Chain)

Для выполнения лабораторной работы используйте последовательности глобинов из лабораторной работы № 4.

Последовательности HIV (к п. В):

>C.ZA_AY047297
>C.ZA_AY047296
>C.ZA_AY047295
>C.ZA_AY047299
>C.ZA_AY047298
>C.IN_Y17891
>C.IN_Y17892
>C.IN_Y17893
>C.IN_Y18199

HIV envelope proteins (к п. В):

>H1_ELI_D
>H1_LAI_B
>H1_MAL
>H1_NDK_D
>H2_D205_B
>H2_ROD_A
>H2_ST_A
>H2_UCI_B

>S_MAC
>S_CP2
>S_AGM
>S_MAN
HIV gag proteins (к п. B):
>H1_ELI_D
>H1_LAI_B
>H1_MAL
>H1_NDK_D
>H2_D205_B
>H2_ROD_A
>H2_ST_A
>H2_UCI_B
>S_MAC
>S_CP2
>S_AGM
>S_MAN
HIV pol proteins (к п. B):
>H1_ELI_D
>H1_LAI_B
>H1_MAL
>H1_NDK_D
>H2_D205_B
>H2_ROD_A
>H2_ST_A
>H2_UCI_B
>S_MAC
>S_CPZ
>S_AGM
>S_MAN
ДНК вирусов H5N1 (к п. C):
Для выполнения лабораторной работы используйте последовательности ДНК вирусов птичьего гриппа из лабораторной работы № 4.
Последовательности к п D.
>Papio_hamadryas (baboon)
>Pan_troglodytes (chimpanzee)
>Homo_sapiens (human)
>Pongo_pygmaeus (orangutan)
>Pan_paniscus (pygmy chimpanzee)
>Phoca_vitulina (harbor seal)
>Gorilla_gorilla (gorila)

Лабораторная работа № 7

Текстовые методы в биоинформатике

Цель работы: продемонстрировать потребности и пользу от методов text mining (TM) в системной биологии и биоинформатике.

TM – это метод извлечения информации из больших объемов текста. Очевидно, что эта информация зависит, в том числе, и от потребностей пользователя, от правильности составления им запросов. В системной биологии одной из целей является представление организма в виде схем и диаграмм. Иными словами организм должен быть представлен как система взаимосвязей их базовых компонентов, таких, как гены и белки. Многие из таких взаимосвязей могут наблюдаться исследователем независимо от других исследовательских групп в мире. Некоторые взаимосвязи могут являться побочным продуктом исследований, проводимых с другой целью, а не для установления случайно выявленной (или вообще какой либо) взаимосвязи.

На данный момент мы будем рассматривать белок-белковые взаимодействия. Белки являются продуктами экспрессии генов. Каждый белок выполняет специфическую, присущую ему функцию в клетке. Их можно рассматривать как «разумные» частицы, выполняющие свои задачи в клетках. Белки могут влиять друг на друга при связывании и, возможно, происходящем при этом ингибировании или активации. Это значит, что взаимодействующие белки могут образовывать сети или пути. Такие пути могут быть классифицированы, разделены на метаболические, сигнальные, транскипционные и другие.

В первой части работы мы проведём исследование одного сигнального пути. Этот путь является небольшой частью большей сети путей, вовлеченной в регуляцию деления клетки. Когда этот сигнальный путь ломается, распадается, то клетка начинает делиться бесконтрольно, что, в свою очередь, приводит к раку. Поэтому этот путь является исключительно важным для исследователей механизмов канцерогенеза. Мы будем использовать доступные инструменты (медицинские базы данных) для изучения различных аспектов этого пути.

Во второй части работы мы уделим внимание известному инструменту TM – GATE (http://gate.ac.uk/). Мы будем использовать его (расчет F-score) для оценки распознавания определенных терминов в биологических текстах. GATE является довольно мощным инструментом, однако он может быть неточен при работе с биологическими текстами. Мы проверим насколько эффективно инструмент общего предназначения работает с медико-биологическими текстами и насколько сильно биологические тексты отличаются от, скажем, новостных текстов.

Инструменты и ПО, используемое в лабораторной работе:

The Gene Ontology (http://www.geneontology.org/) - иерархический курируемый словарь, описывающий «продукты экспрессии генов в терминах их ассоциаций с биологическими процессами, клеточными компонентами и молекулярными функциями без учета видоспецифичности». Этот инструмент поиска может быть полезен при изучении компонентов различных путей.

KEGG (http://www.genome.jp/dbget-bin/show_pathway?hsa04010+5894) – сайт предназначен для представления и визуализации различных метаболических, сигнальных и иных путей. Сайт содержит ссылки на аннотации и описания на все белки, вовлеченные в метаболические пути и сети, а также перекрестные ссылки на пути, содержащие общие белки и ссылки на белки в других базах данных.

Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=gene) – база данных, содержащая информацию о белках и генах. Она также имеет ссылки на оригинальные статьи, имеющие отношение к имеющимся записям о последовательностях. В Entrez Gene хранится информация о синонимах для всех имеющихся в ней белках – это помогает составлять более точные и полные запросы к базе.

PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=pubmed) – главный портал для доступа к базам данных Medline. Эта база данных содержит цитаты на статьи в рецензируемых журналах (статьи, публикуемые в журналах с расширенным рецензированием). Цитаты, хранящиеся в PubMed, имеют полную библиографическую информацию и аннотации для большинства научных статей.

Web of Knowledge (http://portal.isiknowledge.com/?DestApp=WOS) – поисковая машина для медико-биологических текстов и статей. Она находится в собственности и может быть доступна только из университетских сетей или по подписке.

EBIMed (http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp) – поисковая машина, которая представляет результаты поиска в виде таблицы корреляций, терминов и гиперссылок, встречающихся в статьях. Мы будем использовать ее для проверок связей, описанных в путях. Эта поисковая машина имеет в своем составе элементы ТМ, такие, как распознавание именованных сущностей с использованием лексикона.

ChiliBot (http://www.chilibot.net/) – еще одна поисковая машина, которая несет в себе элементы ТМ. Это позволяет пользователям проводить поиск по нескольким ключевым словам и затем выводить результаты по каждому из них или по всем сразу и строить сети одновременного появления этих ключевых слов в предложениях – сети взаимодействия и взаимосвязей терминов.

GATE (http://gate.ac.uk/) – мощное средство ТМ, которое позволяет вам использовать предварительно заготовленные компоненты (такие, как блоки) для построения вашей собственной ТМ машины. GATE была разработана для работы с текстами общего назначения. Инструмент недостаточно пригоден для поиска по биологическим текстам, так как для их обработки требуются дополнительные компоненты. В этих случаях нужно использовать инструмент Annotation Diff для оценки работы on-line версии программы на выборке из аннотированной вручную базы GENIA.

Lingpipe (http://alias-i.com/lingpipe/web/demos.html) – онлайн демоверсия ТМ Java библиотеки, которая располагает компонентами, обученными на биологических текстах. Однако в ней отсутствуют хороший интерфейс пользователя и некоторые другие средства и инструменты для оценки.

Сигнальный путь RKIP, показанный на рисунке 3, очень хорошо описан в статье Cho et al (2003)[1]. Понимание механизмов, заложенных в этом сигнальном пути исключительно важно при изучении рака. Исследователи проводят десятки экспериментов на клетках, в которых имеется RKIP, удаляя или блокируя части пути, стимулируя рецепторы-белки, которые инициализируют каскады реакций или добавляя различные соединения (лекарства). Структура и функция RKIP понемногу стали проявляться благодаря многочисленным экспериментам. Путь RKIP состоит из части MAPK-пути, который содержит белок Raf-1, взаимодействующий с ингибитором Raf-киназой (RKIP).

Представлены белки Raf-1, RKIP, ERK-PP и другие, образующие этот путь. Raf-1/RKIP представляют их взаимодействие, тогда как «PP», находящийся после MEK или ERK (MEK-PP и ERK-PP соответственно) обозначает активированную форму белка. Это, в свою очередь, означает, что ERK-PP не идентичен ERK, однако не существует стандартной формы записи этих названий, и поэтому многие ученые употребляют в написании pERK или ERKPP.

Как было сказано выше, в первой части лабораторной работы мы будем использовать онлайн-ресурсы для ознакомления с различными аспектами этого сигнального пути и для более полного понимания тех трудностей, с которыми сталкиваются биологи при попытке получения более специфичной биологической информации при помощи стандартных инструментов.

---

[1] Cho, K.-H., et al., *Mathematical Modeling of the Influence of RKIP on the ERK Signaling Pathway*, in *Proceedings of the First International Workshop on Computational Methods in Systems Biology*. 2003, Springer-Verlag. p. 127-141.

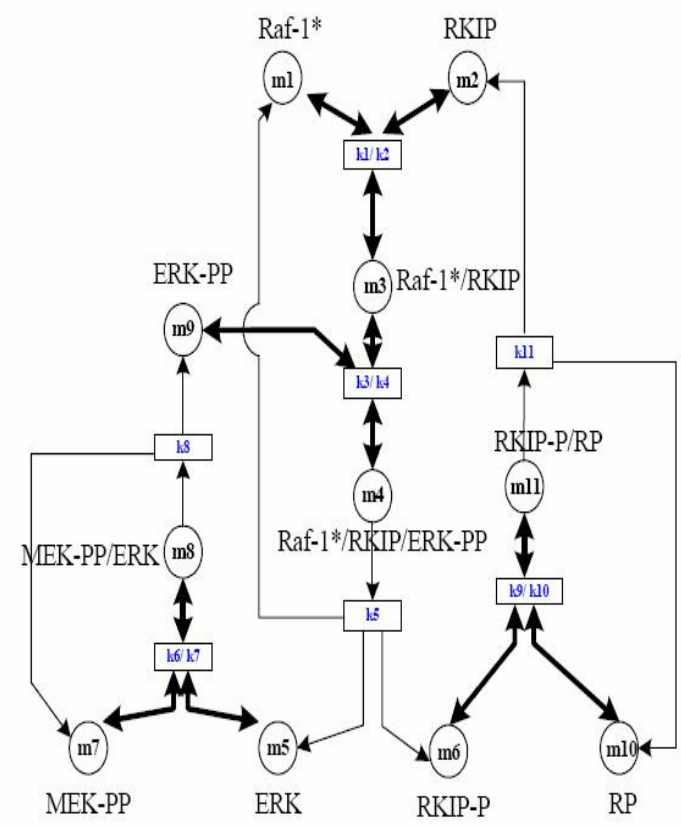

Рисунок 3. Сигнальный путь RKIP.

Часть 1

Некоторые из этих ресурсов не обладают большой скоростью, в особенности Gene Ontology AmiGO, поэтому стоит выполнять задания лабораторной работы параллельно в разных окнах браузера.

1. Знакомство с сигнальным путём RKIP

В этом упражнении вы будете находить различные компоненты пути и определять их синонимы. Вы попытаетесь научиться понимать функции путей при помощи анализа составляющих их компонентов и проходам по сигнальным путям и соеинениям, связывающим узлы путей.

а) Откройте сайт **KEGG** http://www.genome.jp/dbget-bin/show_pathway?hsa04010+5894. Взгляните на полный MAPK путь. Сможете ли вы сказать, где конкретно путь RKIP вписывается в большую сеть MAPK? Какие компоненты RKIP являются также компонентами MAPK? Кликните на любом из компонентов MAPK чтобы увидеть дополнительную информацию по выбранному компоненту. Какие ещё сигнальные пути имеют в своем составе выбранный вами компонент\соединение?

б) Откройте в другом окне браузера для поиска RKIP сайт Entrez Gene http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=gene. В списке «Search» выберите «Protein». Что вам удалось узнать об этом белке? Выпишите эту информацию и его синонимы.

в) Более подробный поиск данных об этом белке можно провести на сайте The Gene Ontology (работа с ним иногда занимает много времени) http://www.geneontology.org/.

2. Изучение сигнального пути

Используйте базы данных PubMed http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pubmed или Web of Knowledge http://portal.isiknowledge.com/portal.cgi?DestApp=WOS&Func=Frame для поиска информации о RKIP пути. Попробуйте провести такой поиск с использованием различных ключевых слов, которые, по вашему мнению, могут помочь вам получить больше информации о компонентах сигнального пути из литературных источников.

3. Использование специализированных веб ресурсов

Откройте EBIMed (http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp) и попытайтесь поискать компоненты сигнального пути. Какие результаты возвратил вам сервер?

Выберите пять компонентов и произведите по ним поиск с использованием ChiliBot (http://www.chilibot.net/). Насколько полученная вами карта соответствует исходному сигнальному пути? Нажмите на кнопку «Edit Synonyms» (слева). Добавьте выбранные вами в п.1 синонимы. Произведите повторный поиск.

Часть 2

В этой части мы попытаемся произвести распознавание автоматичеки именованных биологических сущностей с использованием среда Lingpipe и GATE. Вы можете получить дополнительную информацию и навыки работы с GATE, просмотрев видео на странице инструмента (http://gate.ac.uk/demos/gate-tutorial-all/part-one-all-with-annic-and-ontology.html).

1. Для оценки эффективности различных программных средств очень важно иметь эталонный тестовый набор для тестирования. Этот набор обычно называется «Золотой стандарт». В этом упражнении мы будем использовать очень маленькую часть GENIA corpus (http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/), дополненную некоторыми деталями в целях выполнения этой работы.

2. Скачайте и установите GATE на ваши компьютеры http://gate.ac.uk/download/. На странице есть дистрибутивы для *NIX, Windows и MacOS систем.

3. Запустите **GATE**. Параметры окон и шрифтов настраиваются в **Options→Configuration**.

4. Откройте меню **File→New Language Resource→GATE document.** В поле «имя» введите «genia» и затем нажмите на кнопку «browse» для

открытия файла (текст из приложения 1 сохраните в отдельном файле .xml).

5. Вы увидите, что файл с Genia corpus откроется в *Language Resourses* в левой панели. Двойной щелчок на ней открывает вторую закладку в центральном окне. Кликните на *Annotation Sets,* разверните исходный список и попробуйте выбирать различные закладки. Вы увидите, что текст подсвечивается по-разному. Наведение указателя мыши на подсвеченные части текста вызывает всплывающие окна. Проведите этот опыт. Что отображается в всплывающих окнах при наведении указателя мыли на подсвеченный текст?

6. Сохраните неразмеченный текст из приложения 2 в файле .txt. Зайдите на страницу Lingpipe http://alias-i.com/lingpipe/web/demo-ne.html и попытайтесь открыть страницу для поиска по тексту **Named Entity Demo on the Web→English Biomedical Text: GENIA Corpus (TokenShapeChunker).** Введите в открывшееся текстовое поле текст из приложения 2 (неразмеченный текст Genia Corpus). Нажмите кнопку «Submit Text».

7. Сохраните результаты на диск и затем откройте сохраненный файл как документ GATE.

8. Внутри окна GATE нажмите **Tools→Annotation Diff.** Там, где написано *Document* (слева вверху в окне *Annotation Diff*) выберите ваш документ *genia* в первом выпадающем меню и ваш результат *lingpipe* во втором. Установите *Original Markup* для обоих в *Annotation Set.* Кликните на *Do Diff.*

9. Обратите внимание на F-score и метки, присвоенные программой и имеющиеся в «Золотом стандарте». Посмотрите на частичные совпадения и то, как они соответствуют друг другу – вероятно, будут лишь незначительные различия в расстановке проделов в двух документах.

10. Сохраните текст из приложения 3 в файле .html.

11. Откройте **File→New Language Resource→GATE corpus.** Задайте приемлемое имя и нажмите на иконку со списком справа. Добавьте *genia* документ и документ с новостями.

12. Откройте **File→Load ANNIE System→With defaults.**

13. Сделайте двойной щелчок по новому приложению *ANNIE*, которое отображается под *Applications* на левой панели. Убедитесь в том, что ваш *Corpus* документ выбран в выпадающем меню в центральной панели и нажмите *Run.*

14. Сделайте двойной щелчок на документе с новостями и разверните *Annotation Sets* в верхнем списке. Затем попробуйте выбирать различные метки, наблюдая за тем, какие аннотации выбирает GATE в этом простом новостном тексте.

15. Посмотрите на ваш документ *genia* и определите, какие аннотации были выбраны в нем GATE.

Приложения к лабораторной работе № 7

Приложение 1. GENIA corpus .xml

Приложение 2

Genia corpus неразмеченный

MEDLINE:95369245

IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase .

Activation of the CD28 surface receptor provides a major costimulatory signal for T cell activation resulting in enhanced production of interleukin-2 (IL-2) and cell proliferation.

In primary T lymphocytes we show that CD28 ligation leads to the rapid intracellular formation of reactive oxygen intermediates (ROIs) which are required for CD28 -mediated activation of the NF-kappa B / CD28 -responsive complex and IL-2 expression.

Delineation of the CD28 signaling cascade was found to involve protein tyrosine kinase activity , followed by the activation of phospholipase A2 and 5-lipoxygenase.

Our data suggest that lipoxygenase metabolites activate ROI formation which then induce IL-2 expression via NF-kappa B activation .

These findings should be useful for therapeutic strategies and the development of immunosuppressants targeting the CD28 costimulatory pathway.

MEDLINE:95333264

The peri-kappa B site mediates human immunodeficiency virus type 2 enhancer activation in monocytes but not in T cells .

Human immunodeficiency virus type 2 (HIV-2) , like HIV-1 , causes AIDS and is associated with AIDS cases primarily in West Africa.

HIV-1 and HIV-2 display significant differences in nucleic acid sequence and in the natural history of clinical disease.

Consistent with these differences , we have previously demonstrated that the enhancer/promoter region of HIV-2 functions quite differently from that of HIV-1.

Whereas activation of the HIV-1 enhancer following T-cell stimulation is mediated largely through binding of the transcription factor NF-kappa B to two adjacent kappa B sites in the HIV-1 long terminal repeat, activation of the HIV-2 enhancer in monocytes and T cells is dependent on four cis-acting elements : a single kappa B site, two purine-rich binding sites, PuB1 and PuB2, and a pets site.

We have now identified a novel cis-acting element within the HIV-2 enhancer, immediately upstream of the kappa B site , designated peri-kappa B.

This site is conserved among isolates of HIV-2 and the closely related simian immunodeficiency virus , and transfection assays show this site to mediate HIV-2 enhancer activation following stimulation of monocytic but not T-cell lines .

This is the first description of an HIV-2 enhancer elementwhich displays such monocyte specificity, and no comparable enhancer element has been clearly defined for HIV-1.

While a nuclear factor(s) from both peripheral blood monocytes and T cells binds the peri-kappa B site, electrophoretic mobility shift assays suggest that either a different protein binds to this site in monocytes versus T cells or that the protein recognizing this enhancer element undergoes differential modification in monocytes and T cells, thus supporting the transfection data.

Further , while specific constitutive binding to the peri-kappa B site is seen in monocytes, stimulation with phorbol esters induces additional, specific binding.

Understanding the monocyte-specific function of the peri-kappa B factor may ultimately provide insight into the different role monocytes and T cells play in HIV pathogenesis.

Лабораторная работа № 8

Цель работы – изучение возможностей горизонтального переноса генов у бактерий на примере т-РНК синтетазы (рисунок 4).

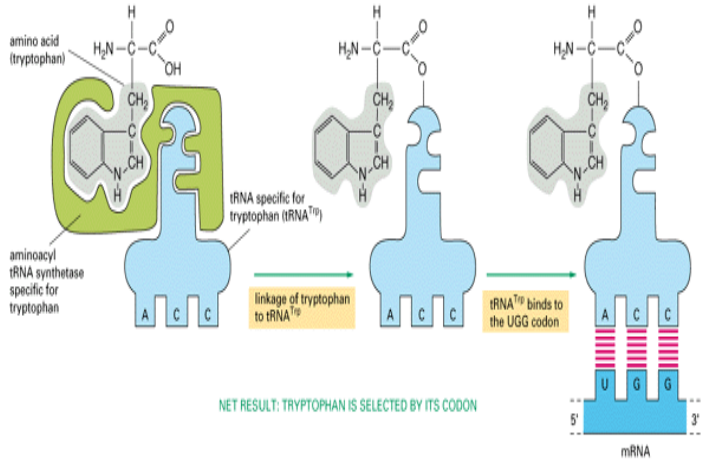

Рисунок 4. Работа т-РНК синтетазы.

Роль т-РНК синтетазы.

При помощи BLAST выберите последовательности бактериальных т-РНК. Постройте филогенетическое дерево, используя выбранные последовательности.

Постройте филогенетическое дерево видов, используя последовательность ssu-16s (последовательность рибосомальной РНК 16s субъединицы).

Сравните полученные деревья и объясните их сходство и различия.

Вопросы для самоконтроля

1. Парное выравнивание. Виды, авторы алгоритмов, цели, значение. Глобальное выравнивание.
2. Вторичные структуры белков, их характеристики и предсказание. ПО и сервисы.
3. Локальное выравнивание. Цели, значение. Алгоритм локального выравнивания.
4. Биоинформатика. Объекты биоинформатики. Задачи, решаемые этой наукой. Методы биоинформатики.
5. Матрицы сравнения последовательностей. PAM, BLOSUM.
6. По приведенной матрице расстояний постройте филогенетическое дерево (Neighbor Joining method, UPGMA). Опишите процесс построения.

	1	**2**	**3**	**4**
1	**0**	**0.3**	**0.5**	**0.6**
2		**0**	**0.6**	**0.5**
3			**0**	**0.9**
4				**0**

7. Основные алгоритмы построения филогенетических деревьев – их достоинства и недостатки. UPGMA и NJ (их отличия), максимальной бережливости (maximal parsimony), максимального правдоподобия, минимальной эволюции.
8. Группы аминокислот. Группировка аминокислот с эволюционной и структурной точек зрения.
9. Открытая рамка считывания. Её нахождение и транскрибция. Поиск открытых рамок считывания – методы.

10. Биоинформатика и филогенез. Молекулярные часы. Клада, OTU, ветвь, лист, корень. Ультраметрическое и неультраметрическое дерево. Ортологи, паралоги, гомологи, ксенологи.
11. Определение филы, таксона, вида. Филогенетические деревья видов и генов, их различия.
12. Какой тест может быть проведен на филогенетическом дереве, построенном по NJ алгоритму, чтобы высказать предположение о его корректности.
13. Редакционное расстояние между двумя последовательностями. Сложность наивного алгоритма его определения.
14. Дано: последовательности WATER и WINE. Скоринг: match- 5, mismatch- -5, вставка промежутка (gap insertion)- -1. Построить таблицу выравнивания и найти по ней путь для него.
15. Локальное выравнивание, задачи, примеры.
16. Множественное выравнивание.
17. Третичная структура белка. Фолдинг.
18. Предсказание третичной структуры белка. Моделирование гомологов. Методы, ПО и сервисы
19. Предсказание третичной структуры белка. Распознавание фолда. ПО, сервисы.
20. Динамическое программирование и выравнивание последовательностей. Способы оптимизации поиска – FASTA, BLAST
21. Классификации белков. Базы данных Pfam, SCOPE, CATH
22. NCBI, ENTREZ и BLAST – назначение, инструменты, задачи.
23. Штрафы за вставку промежутка, схемы, различия.
24. Профиль и консенсус. Сходство и различия.
25. Выравнивание и его статистическая достоверность. Bootstrap.
26. Докинг – цель и задачи. Трудности.
27. Жёсткий докинг. Методы, применение.
28. Гибкий докинг.
29. Экспериментальное определение структур белка. Рентгено-структурный анализ.
30. Экспериментальное определение структур белка. Ядерно-магнитный резонанс.

31. Экспериментальное определение структуры белка. Оценка качества полученной структуры.

32. Hamming distance и Edit distance – отличия.

33. Метод GOR и Chou-Fasman. Их применение.

34. Дана следующая матрица скоринга ДНК:

	A	C	G	T
A	10	2	5	2
C	2	10	2	5
G	5	2	10	2
T	2	5	2	10

Какова максимально возможная оценка выравнивания AATAAT и AAGG, при условии цены промежутка -5?

35. Допустим, нам даны 4 последовательности: S1=act, S2=agct, S3=aact, and S4=acct. Парные выравнивания этих последовательностей следующие:

a-ct	a-ct	a-ct
agct	aact	acct

По ним были построены 2 варианта MSA (в зависимости от параметров\применения алгоритма). Какой из этих вариантов Вы предпочтёте и почему?

a-ct	a---ct
agct	ag--ct
aact	a-a-ct
acct	a--cct

36. MSA последовательностей с параметрами по умолчанию (gap penalty -10) выглядит так:

AGCT

ACCT

– ACT

AACT

В полученном выравнивании имеется 2 полных совпадения и один промежуток. Очевидно, что есть лучшее выравнивание с тремя полными совпадениями и одним промежутком. Объясните логику построения этого

выравнивания. Как нужно изменить параметры выравнивания для того, чтобы его улучшить?

37. Интерпретация результатов выравнивания – Score, E-value. Проверка достоверности полученного выравнивания.

38. BLAST. Алгоритм поиска оптимального выравнивания. Состав BLAST на NCBI.

39. Почему динамическое программирование используется для парного выравнивания и практически не применяется для выравнивания нескольких последовательностей (MSA)?

40. Штрафы за вставку промежутков. Аффинные и линейные штрафы, их сравнение и влияние на выравнивание.

41. Что может сказать об эволюционном процессе организмов свойство ультраметричности построенного для них дерева (при условии его истинности)?

42. Свойство аддитивности филогенетического дерева. Его определение.

В 2009 году Университет стал победителем многоэтапного конкурса, в результате которого определены 12 ведущих университетов России, которым присвоена категория «Национальный исследовательский университет». Министерством образования и науки Российской Федерации была утверждена программа его развития на 2009–2018 годы. В 2011 году Университет получил наименование «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».

Естественнонаучный факультет был создан в НИУ ИТМО в 1994 году объединением кафедр, обеспечивающих базовую физико-математическую подготовку студентов первого и второго курсов практически по всем направлениям и специальностям Университета.

На факультете обучается более 650 студентов — будущих специалистов по применению информационных систем и технологий в различных областях производства науки и образования (информатике, математике, физике, экологии, компьютерной графике).

Лаборатория биоинформатики была создана в НИУ ИТМО в 2011 году. Основным направлением исследований лаборатории являются проблемы структурной биологии белка.

Порозов Юрий Борисович
"Биоинформатика".
Учебно-методическое пособие

В авторской редакции
Редакционно-издательский отдел НИУ ИТМО
Зав. РИО Н.Ф. Гусарова
Лицензия ИД № 00408 от 05.11.99
Подписано к печати
Заказ №
Тираж
Отпечатано на ризографе