КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ФУНДАМЕНТАЛЬНОЙ МЕДИЦИНЫ И БИОЛОГИ
Кафедра биохимии и биотехнологии

Н.И.АКБЕРОВА

Методы молекулярной филогении

Учебно-методическое пособие

Казань – 2014

Настоящее пособие посвящено знакомству с методами молекулярной филогении и их программными реализациями. Предназначено для получения навыков практической работы с биоинформатическими банками данных, программами множественного выравнивания полинуклеотидных и полипептидных последовательностей и определению их филогенетического родства.

Филогения

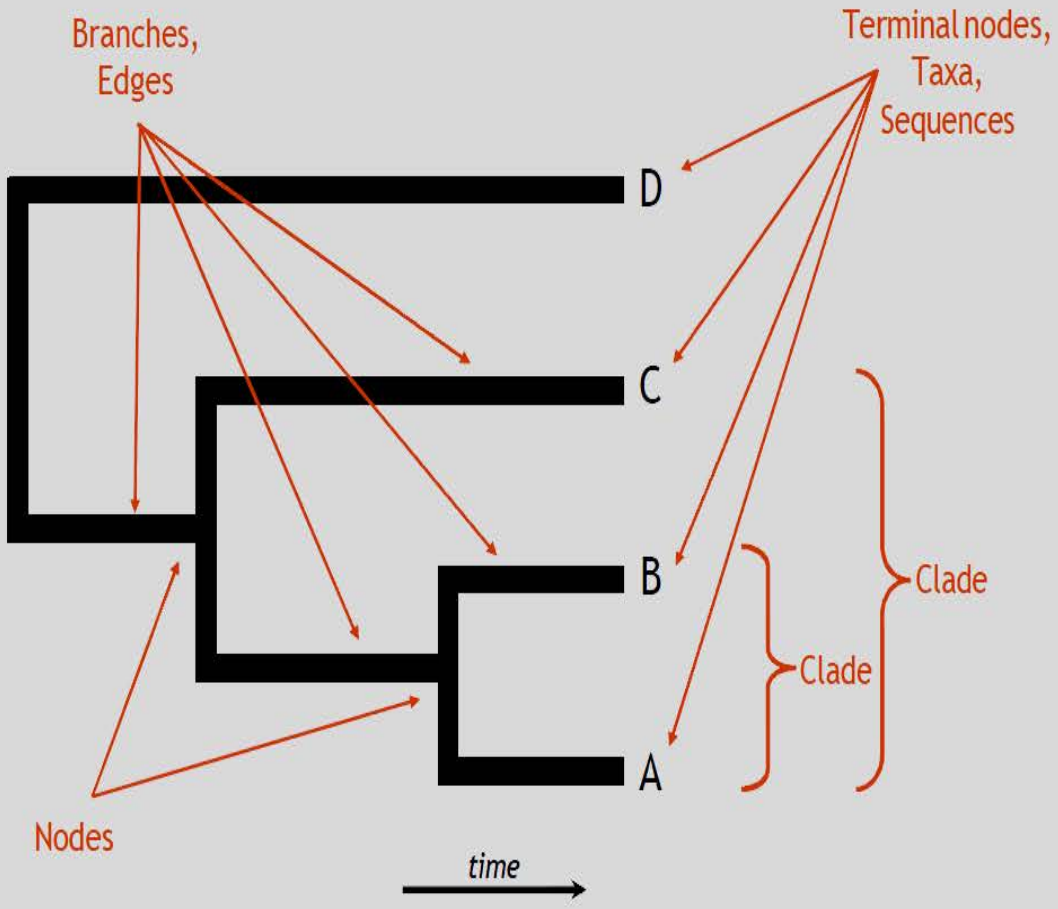
[Необходимый софт: MEGA 5 и выход в Интернет]

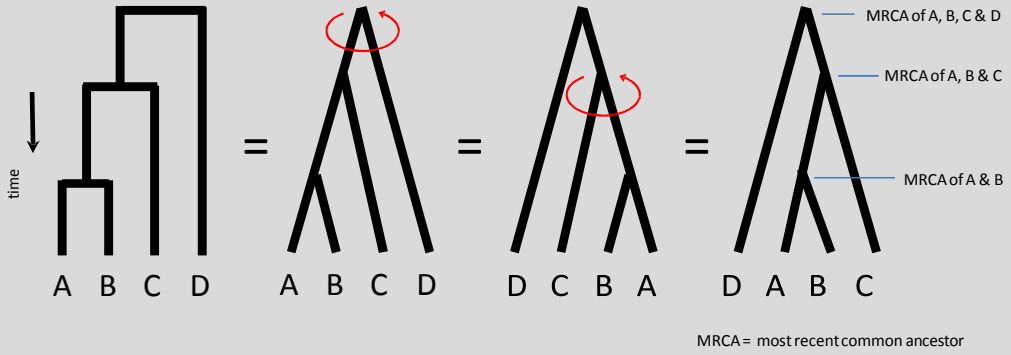
В этой работе мы будем использовать множественное выравнивание последовательностей (MSA). Филогенетический анализ производит ветвящиеся диаграммы, они могут четко иллюстрировать отношения между последовательностями, которые не очевидны из BLAST или MSA. Филогенетические деревьяполезны для эволюционных и сравнительных исследований, ориентированных на выяснение эволюционных взаимоотношений и моделей дивергенции, но они также становятся все более важными длягенерации гипотез относительно функции гена или белка для молекулярных и биохимических исследований.

Филогения - обширная область и часто сама по себе занимает весь курс. Мы только коснемся этой темы, введем некоторые основные понятия и идеи, познакомимся с основными инструментами. Есть две основных категории филогенетических инструментов: методы, рассматривающие расстояния между последовательностями (distance-basedmethods) и методы, рассматривающие признаки (character-basedmethods). Мы будем использовать оба этих подхода. Цель этих занятий не только получение представления о том, как построить

филогенетическое древо, но и понимание того, что принцип «вырезать и вставить», который слишком часто предполагается для биоинформатических подходов, не работает, что главное – это необходимость выбора адекватного метода. Часто разные подходы дают различные выводы, это подразумевает, что вы должны использовать несколько методов и разумно изменять значения их параметров, и в конечном итоге использовать биологическую интуицию, чтобы генерировать лучшее дерево и сделать качественный хороший анализ и правильно его интерпретировать.

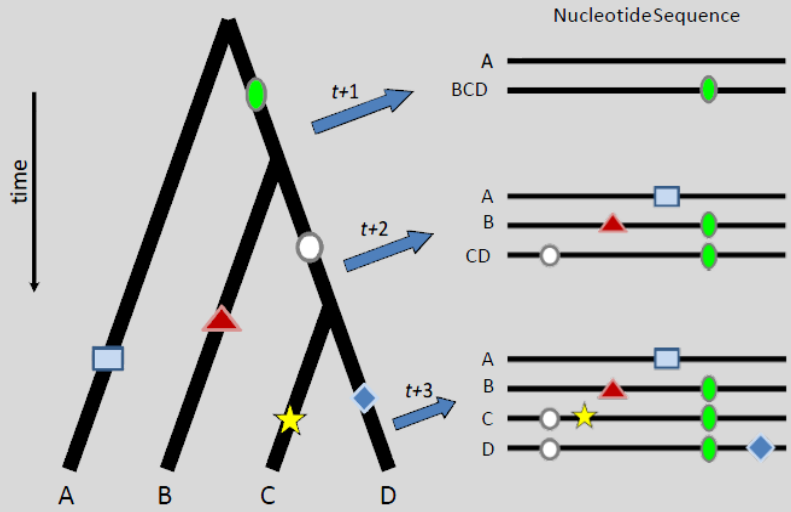
B. 1 Филогенетическое дерево



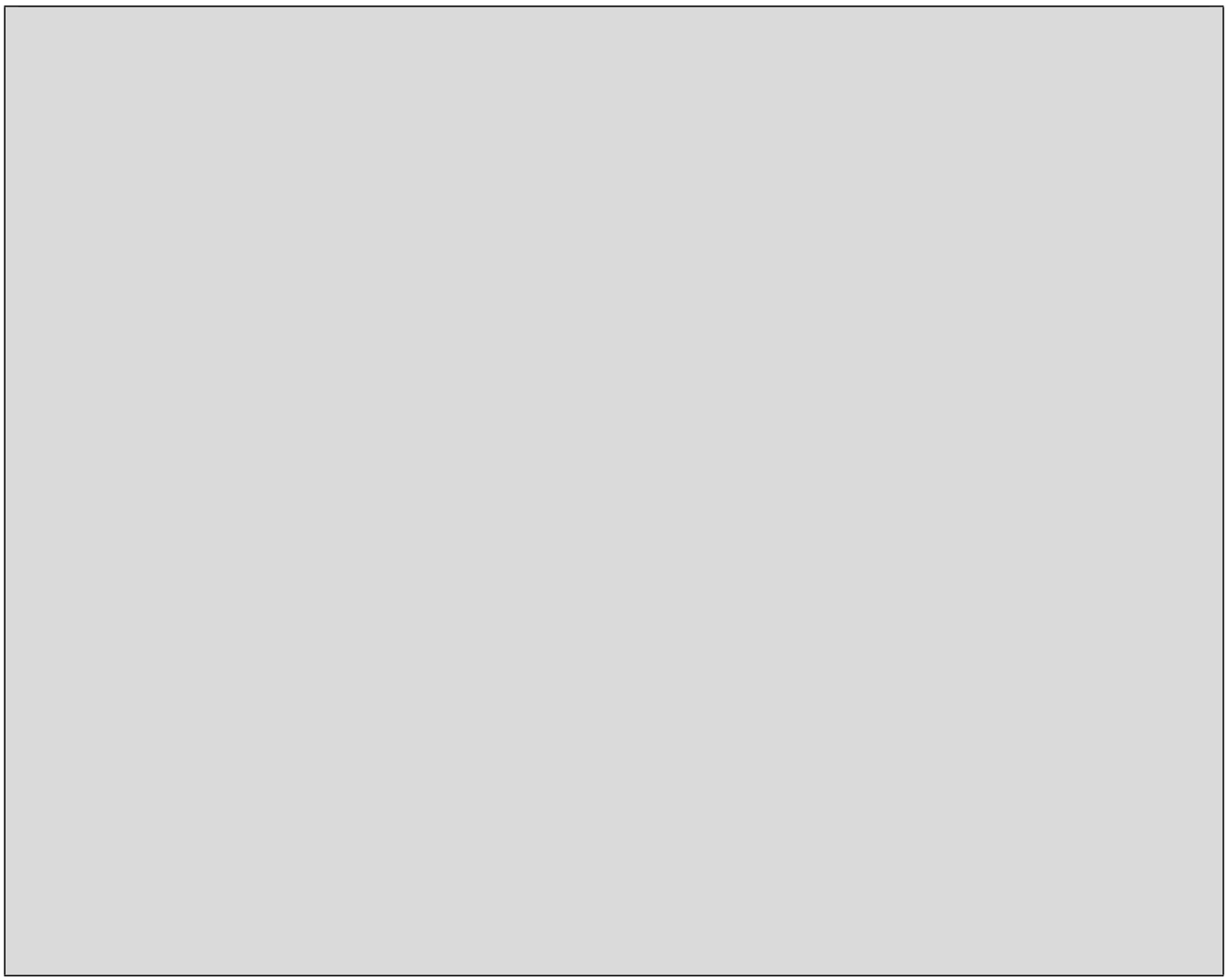


Филогенетические деревья могут быть представлены в различных формах и ориентациях. что Важно, что единственный способ определить эволюционное расстояние между двумя последовательностями – это определить, как далеко назад во времени вы должны пойти прежде, чем найти общий предок. Так, на дереве справа, хотя последовательность A ближе к D, чем к C физически на странице, на самом деле A более тесно связана с C, так как они имеют больше общих предков. Эволюционные отношения между A-D также могут быть представлены с использованием формата Newick следующим образом (((A, B), C), D): вложенность в круглые скобки соответствует разделению на деревях выше.

B 2. Рост филогенетического дерева



В ходе эволюции организмы изменяются, накапливают мутации (цветные фигуры). Эти мутации будут передаваться в потомстве всем дочерним линиям. Мутация, которая происходит очень рано в истории группы, например, зеленый овал, которая произошла у предка последовательностей B, C & D, будет найдена во всех трех линиях потомков. Мутация, что происходит позже (например, синий квадрат), вероятно, можно найти в меньшем количестве линий. Положение мутаций на нуклеотидных последовательностях совершенно произвольно и предназначено только, чтобы показать, как много уникальных последовательностей есть в каждый момент времени, распределение мутации среди этих последовательностей



Начнем с построения дерева с помощью neighbour-joining метода в программе MEGA 5. Вы уже знакомы с этой программой, делали в ней множественное выравнивание. MEGA (Molecular Evolutionary Genetic Analysis) довольно простое в использовании, хотя и очень мощное приложение для проведения филогенетического анализа. Метод Neighbour-joining – быстрый и достаточно надежный, поэтому в нем делают большинство стартовых деревьев для построения гипотез о общем деревм топологии / расстояния, прежде чем перейти к более строгим программам.

1. Откройте программу MEGA 5 и конвертируйте выравненные и сохраненные полинуклеотидные последовательности *DNA_aligned.fas* из FASTA формата в MEGA формат.
   - Нажмите на **File/Convert File Format to MEGA…**
   - Чтобы найти файл, вам необходимо нажать на значок очень небольшой папки на правой стороне **Data file to convert** бокса.
   - Найдите ваш файл с выровненными нуклеотидными последовательностями в FASTA формате (***aligned nucleotide sequence file***). Выбирите **Data Format (FASTA**).
   - Нажмите **OK**

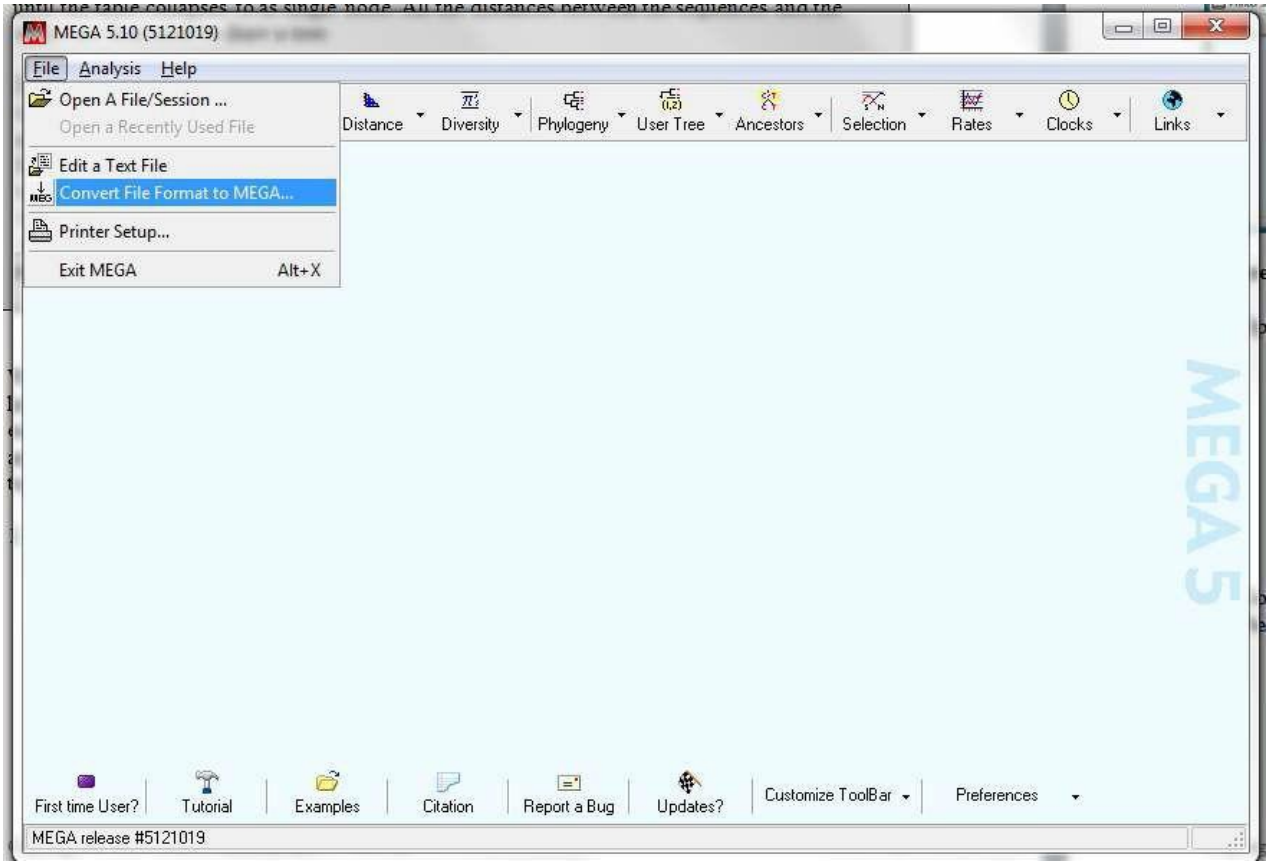


Рис. 1. MEGA 5 – интерфейс пользователя

- Предполагая, что файл преобразован правильно, сохраните его. Вы заметите, что к имени файла добавляется расширение **.meg**. MEGA и FASTA форматы очень похожи для этих простых файлов, основное отличие, что MEGA сохраняет больше информации в разных полях
  - MEGA формат может быть несколько придирчивым. Вот правила:
  - Входные файлы не могут иметь имена последовательностей с пробелами, или любым из следующих символов, ; : ‘ “ ! ? > < [ ] ~ @ # ^ &
  - Фигурные скобки можно использовать только тогда, когда они в паре.
  - Первая строчка всегда: **#MEGA**
  - Вторая строка всегда: **!Title: xxx** , где xxx что угодно

2. Откройте новый файл
   - Нажмите **File > Open a File/Session**
   - Найдите и откройте новый файл с расширением **.meg**
   - Отметьте что этот файл содержит **Nucleotide Sequences**, и любую другую необходимую информацию (protein coding sequence = Y, select genetic code = standard).
   - Нажмите на иконку “**TA**” и откройте эксплорер данных- Data Explorer. Это полезно для визуализации и выбора данных и областей для анализа

Примечание: если есть недопустимые символы в файле, МЕГА покажет вам, на каких линиях они находится.

3. Вернитесь в главное окно и войдите в **Analysis > Phylogeny > Construct/Test Neighbor- Joining Tree**. Мы будем придерживаться параметров по умолчанию на этот раз, но только не забудьте выбрать "метод Bootstrap" под **Phylogeny Test/Test of Phylogeny**. Нажмите **Compute.**

Рис. 2. MEGA 5 - окно настройки Neighbour-Joining анализа

4. Результатом должно быть красиво отформатированое филогенетическое дерево в новом окне. Обратите внимание на значения, указанные над каждым узлом или ветвью дерева. Они называются загрузочными значениями являются мерой статистической достоверности для каждого узла. Любая загрузочная оценка> 70, как правило, рассматривается как достаточно надежная.

- В окне TreeExplorer выбирите **View > Options > Branch**
- Выбирите **Hide values lower than** и поставьте **70%**
    a. ***Сколько надежных узлов в дереве?***
    b. ***Более надежные узлы ближе к основанию найдены или на терминальных концах дерева?***
    c. ***Можете ли вы назвать возможные причины этого?***
    d. ***Есть ли у вас большое доверие к этому дереву?***

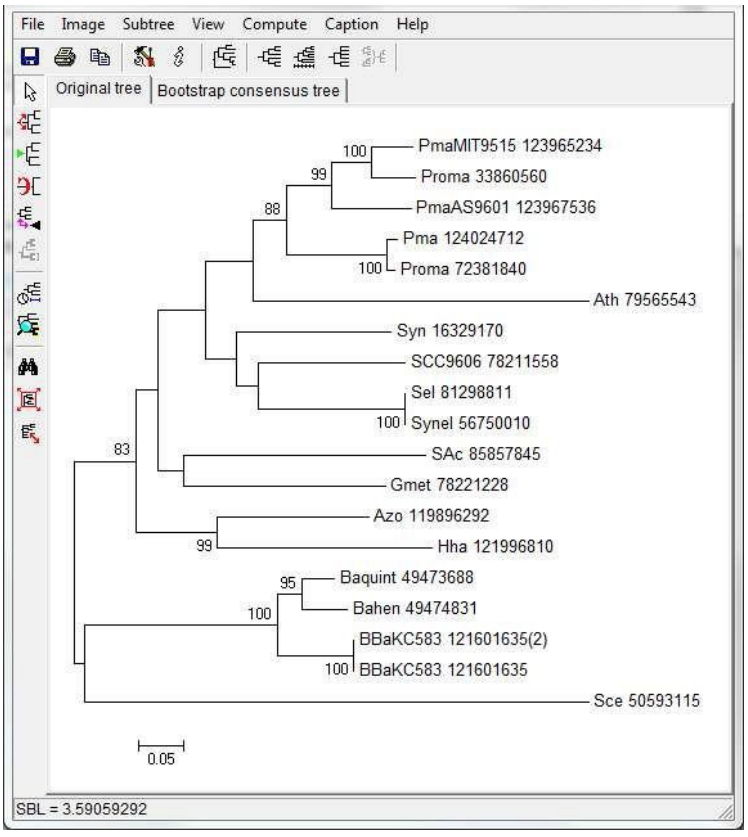


Рис. 3. Дерево, построенное методом ближайшего соседа с бутстрап-поддержкой (Bootstrapped Neighbour-Joining Tree), показаны лишь бутстрап-оценки ≥70%

5. Вы очень просто можете изменить в MEGA способ представления дерева.
    - В окне TreeExplorer идите **View > Tree Branch Style** и выбирите один из форматов: circular, radial, traditional straight…
    - Проверьте относительный порядок ветвления из последовательностей с этими различных форматах

Изменились ли взаимосвязи последовательностей?

6. MEGA TreeExplorer очень мощный. Вы можете манипулировать своим деревом бесконечно посредством **Options** (View > Options) и меню **Subtree**. Верните дерево в традиционный формат (traditional / rectangular) и попробуйте поиграть с различными опциями. Большинство опций меню **Subtree** доступны также с иконок в окне слева. Обратите внимание, что все эти изменения являются обратимыми, поэтому не стесняйтесь «играть» "с ними!

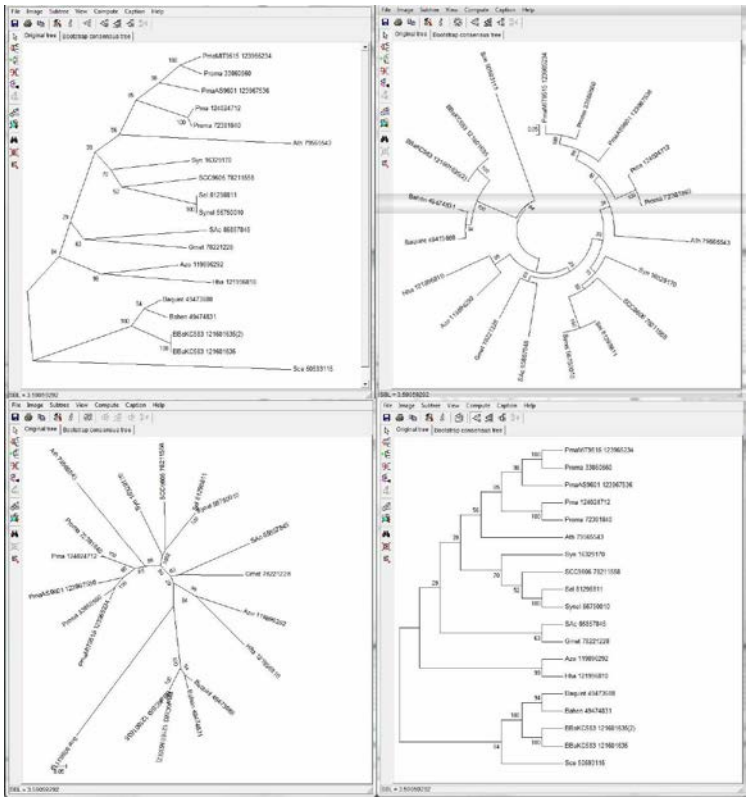


Рис.4. Четыре формата одного и того же дерева.

В 4. Укоренение филогенетических деревьев

Корень филогенетического дерева - это точка на основании дерева, которая уходит дальше всех в прошлое. Хотя это простая идея, определения корня дерева на самом деле может быть очень трудным. Есть два основных способа укоренения деревьев:

1. Средняя точка укоренения предполагает размещение корня в самом центре тяжести дерева. Это является методом по умолчанию используется в MEGA и многих других программах. Midpoint корень очень легко сделать (идите по **View > Root on Midpoint**), но при этом предполагается, что все последовательности эволюционируют приблизительно с одной и той же скоростью. Если это не так, то срединное укоренение может быть неуместным

2. При использовании внешней группы для укоренения (Outgroup rooting) вы включаете в том числе последовательность, для которой известно, что она более отлична (не родственна), чем остальная часть последовательностей в анализе, и убеждаетесь, что эта последовательность ветвится в самом основании дерева. Такое укоренение является очень надежным, если у вас есть надежная предварительная информация, которая позволяет вам выбрать хорошую внешнюю группу. К сожалению, во многих исследованиях .эта информация не известна

Если вы не уверены в укоренения дерева, вы всегда можете представить неукорененные дерево, излучающееся из центральной точки, без направления, представляющего время Многие считают, чтобы эти деревья труднее интерпретировать, но они представляют ту же самую информацию, без дополнительных предположении о том, какие последовательности разветвляются первыми.

7. Давайте посмотрим, как изменение укоренения дерева влияет на наши выводы. Щелкните левой кнопкой мыши непосредственно на одной из внутренних ребер дерева, а затем щелкните правой кнопкой мыши, выбирите **Place Root**. Дерево будет переделано так, чтобы выбранный вами узел будет у основания дерева

- Повторите несколько раз на различных ответвлениях и посмотрите взаимосвязи последовательностей.

  a. Есть ли изменения во взаимосвязях?

- Вы можете вернуться к среднему укоренению, выбрав **View / Root on Midpoint.**

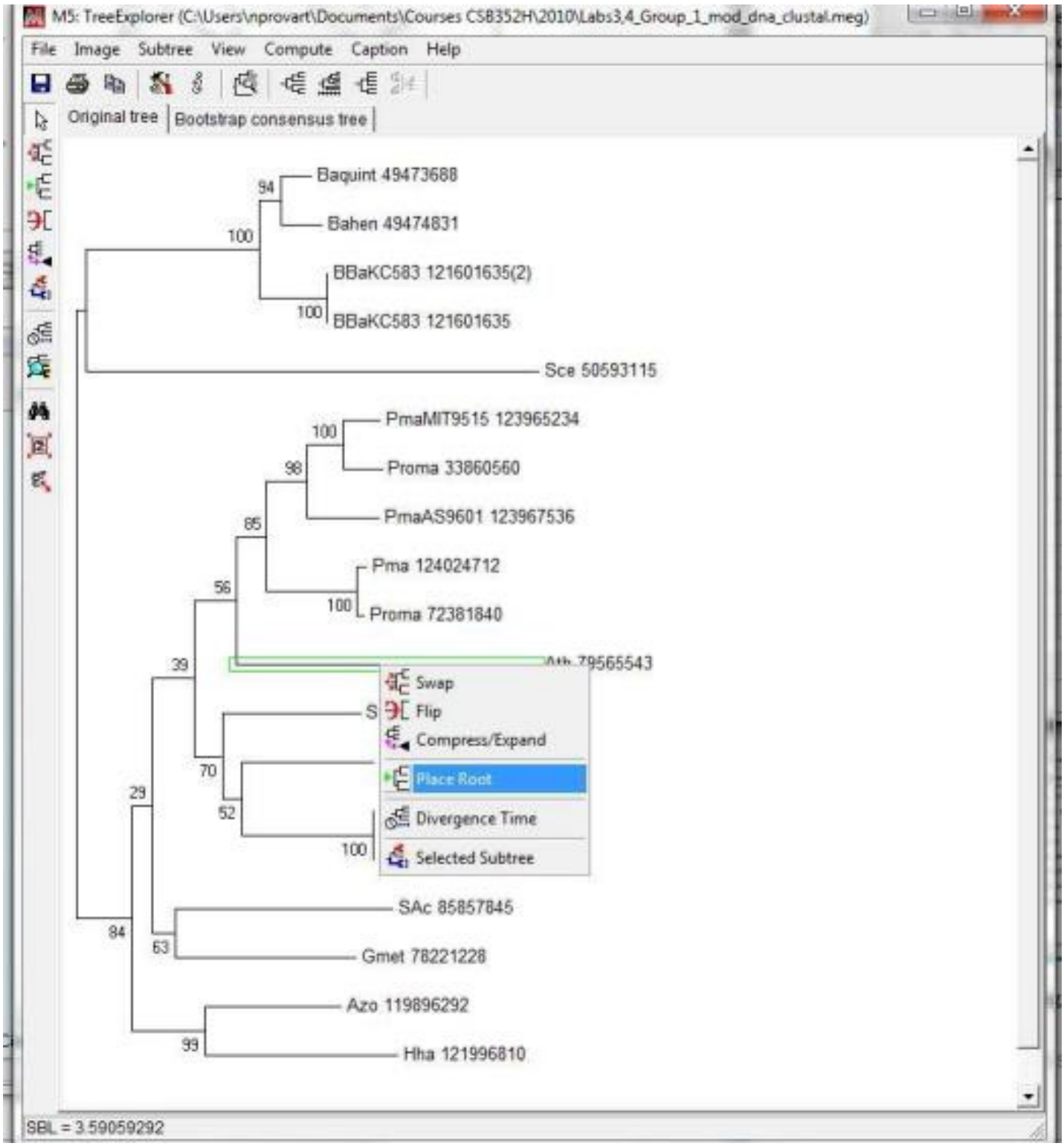


Рис 5. Переукоренение в MEGA

8. Нарисуйте неукорененное дерево, выбрав **View > Tree Branch Style > Radiation**.

   a. ***Попробуйте совместить это дерево с рассматриваемыми ранее.***

Контрольное задание 1

Одинаковы ли группы P. marinus 124024712 и P. marinus 72381840 в средне-укорененном (mid-point- rooted) и неукорененном (radiation-style unrooted) деревьях?

9. MEGA 5 производит автоматическую подпись для каждого анализа. Нажмите **Caption,** чтобы увидеть детальное описание проведенного филогенетического анализа, причем это делаете в формате, необходимом для публикации со всеми ссылками. Это очень удобная опция, как в случае включения анализа в статью, так и для сохранения и четкого представления деталей вашего анализа

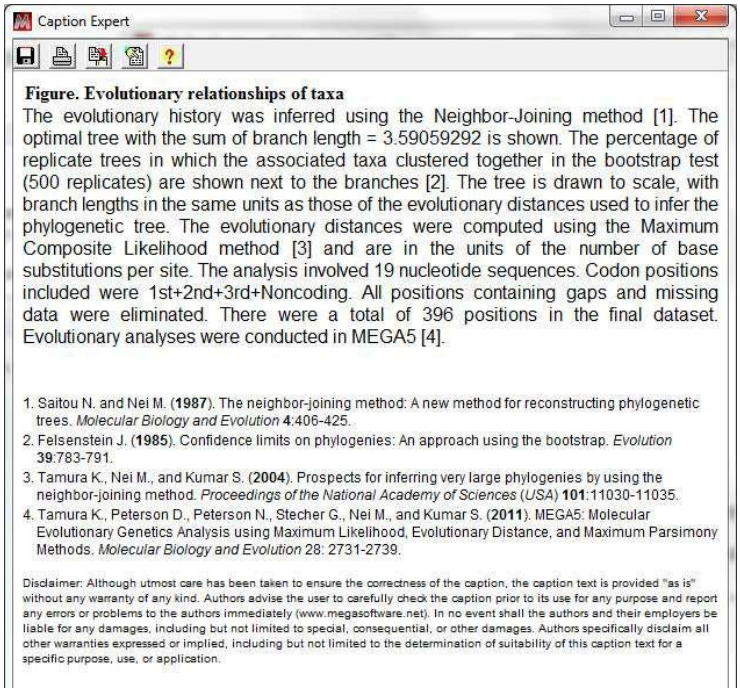


Рис. 6. Функция автозаголовок в MEGA 5. Обратите внимание, как замечательно цитируются биоинформатические методы и софт.

10. Теперь давайте сделаем дерево белковых последовательностей. Вы можете оставить окно TreeExplorer открытым, но закрыть текущий файл данных, перейдя по значку **Close Data** (в главном окне). Теперь нужно открыть соответствующий файл с выровненными белковыми последовательностями (скачать "Lab3,4_sequences_prot_aligned.fas"). Сначала вам придется конвертировать ваши выровненные белковые последовательности в формат MEGA (Смотрите пункт 1)

11. Сделайте дерево методом ближайших соседей neighbour-joining tree, так же, как пункте 1, уберите слишком низкие бутстрапы так же, как делали ранее. Сравните нуклеотидные и белковые деревья.

a. ***Есть ли различия между белковым и нуклеотидным деревьями?***
b. ***Если да, как это можно объяснить?***
c. ***Какое дерево вызывает большее доверие?***

Контрольное задание 2

Сколько хорошо поддерживаемых узлов (bootstrap scores >= 70%) в полученном вами дереве на основе филогенетического анализа белковых последовательностей?

12. Теперь давайте поиграем с некоторыми параметрами, чтобы посмотреть, как они влияют на анализ. Откройте снова файл выровненных нуклеотидных последовательностей (обратите внимание, что вы можете сделать это, просто нажав на иконку **Close Data**, и затем выбрав **File > Open a Recently Used File**), а затем выберите **Phylogeny > Bootstrap Test of Phylogeny > Neighbor-joining** (см рис. 2).

- Помните, что вы ничего не сломаете, поэтому попробуйте как можно больше вариантов. Один очень хороший способ определить, какое дерево лучше, это посмотреть на бутстрап-оценки. Попробуйте оптимизировать параметры таким

образом, чтобы у вас как можно больше узлов с наибольшей поддержкой (самые высокие баллы начальной загрузки). Обратите особое внимание на следующее:

- **Gaps-Missing Data / Pairwise Deletion**

• Этот параметр полезен, если у вас есть последовательности с большим количеством вставок и удалений (indels). Complete deletion (настройка по умолчанию) удаляет все столбцы выравнивания с indels из любой последовательности. Pairwise deletion удаляет indels только в парных сравнениях. Полное удаление является наиболее консервативныым подходом, иногда количество вставок и удалений будет настолько высоким, что вы теряете важную информацию при таком подходе

- **Model / Nucleotide**

• Эти модели замещения такие же, как замещения матриц, рассмотренные ранее. Попробуйте использовать различные модели, чтобы увидеть,как они влияют на ваши выводы. См. В 5 для описания некоторых из моделей

- **Rates among sites / Different (Gamma Distributed)**
- **Gamma Parameter /** изменяйте в пределах 0.1 – 2.0

• Этот параметр позволяет контролировать изменения скорости эволюции через последовательности. Например, может быть, одна область высоко консервативна (эволюционирует очень медленно), а другая вовсе не имеет консервативных остатков (имеет гораздо более высокую скорость эволюции).

• Параметры Lower gamma используются для последовательностей с высокой степенью изменчивости (вариабельности).

a. ***Опишите, как изменение параметров влияет на реконструкцию филогении.***

b. ***Можете ли вы сделать обоснованное предположение, почему вы можете (или не можете)включить или исключить определенные сайты, такие как 3-й позиции в кодонах или в некодирующих участках?***

B 5. Модели замещений

Как обсуждалось ранее, модели замещения используются для моделирования, насколько последовательности ДНК или белок изменились в течение эволюционного времени. В то время как матрицы, как PAM и BLOSUM полезны для моделирования эволюции последовательности белка, другие модели используются для ДНК-последовательностей. Вот небольшая подборка моделей и их допущений:

- Jukes-Cantor
  - Равные частоты для всех нуклеотидов
  - Нет смещения по частоте, когда один нуклеотид мутирует в другой (равные стоимости замен)
- Felsenstein-81
  - Неравные частоты для нуклеотидов
  - Нет смещения при заменах
- Kimura 2-Parameter
  - Равные частоты для всех нуклеотидов
  - Различные скорости замещения для транзиций (пуриновые основания - в пуриновые или пиримидиновые- в пиримидиновые) и трансверсий (пуриновых –в пиримидиновые или наоборот). Обычно транзиции происходят примерно в два раза чаще, чем трансверсии.
- Tajima-Nei
  - Неравные частоты для всех нуклеотидов
  - Равные частоты трансверсий
  - Разные частоты транзиций
- Tamura 3-Parameter
  - Равные частоты для всех нуклеотидов
  - Различные скорости для транзиций и трансверсий
  - Смещение за G+C содержание
- Hasegawa-Kishino-Yano(HKY)
  - Неравные частоты для всех нуклеотидов
  - Различные скорости для транзиций и трансверсий

- Как вы выбрать, какую модель использовать? Самое главное, вы должны исходить из ваших данных. Это должно дать вам представление о том, одинаковы ли частоты для нуклеотидов Если вы хотите делать это правильно, то лучше использовать программу jModelTest 2, которая использует метод правдоподобия, чтобы помочь вам определить наилучшую модель замен и гамма-параметр (доступна в качестве приложения Java на https://code.google.com/p/jmodeltest2/ (Darriba et al., 2012).

В 6. Методы, основанные на признаках (Character-Based Methods)

Существует множество таких методов. Все они на самом деле сравнивают состояние каждого остатка (нуклеотидов или аминокислот) в каждом столбце выравнивания в MSA. Они пытаются определить наиболее вероятное или простейшее объяснение, необходимое для объяснения отношений, наблюдаемых в данных. Как правило, они делают это, рассматривая все возможные объяснения (другими словами, все возможные деревья), и определяют дерево или набор деревьев, что лучше всего объясняет данные на основе конкретных критериев, используемых в методе.

Метод максимального правдоподобия описывает статистические рамки, применительные к филогенетической реконструкции в данном случае. Он проходит через все возможные структуры дерева и спрашивает, насколько вероятно, что вашему набору данных присваивается конкретное дерево. Так, например, гораздо более вероятно, что очень сходные последовательности должны находиться очень близко друг к другу (в терминальных узлах дерева), а не сходные – далеко друг от друга (около ствола дерева).

Такие методы, в т.ч. и метод максимального правдоподобия, как правило, очень сложные подходы, которые позволяют реалистично моделировать эволюционные изменения на статистической основе. К сожалению, с ними труднее работать (или по крайней мере работать должным образом), и, возможно, самое главное, они требуют большого объема вычислений, так как они должны эффективно исследовать все возможные структуры дерева. Это практически означает, что они не могут быть применены к очень большим наборам данных.

Теперь построим дерево методом максимального правдоподобия (ML). Как уже говорилось выше, ML является одним из самых мощных филогенетических методов, но, к сожалению, он не так прост для выполнения, как метод объединения ближайших соседей. Есть ряд хороших приложений для ML-анализа в свободном доступе. Мы будем использовать реализацию в MEGA, но вы можете ознакомиться с инструментами ML, доступными через PHYLIP (Phylogeny Inference Package), который является мощным и всеобъемлющим набором свободно доступных филогенетических приложений. PHYLIP работает на большинстве компьютерных платформ через командный интерфейс, он также доступен через ряд общедоступных веб-интерфейсов (см " Полезные ссылки " ниже).

1. Запустите MEGA и загрузите выровненные последовательности ДНК в MEGA формате как описано в пункте 2 (**File > Open a Recently Used File** и выбрать файл в MEGA формате).

2. В **Analysis > Phylogeny > Construct/Test Maximum Likelihood Tree** установить в **Phylogeny Test** / Test of Phylogeny to Bootstrap - 500 реплик , остальное оставьте по умолчанию (однако увеличение количества потоков может ускорить анализ). Нажмите **Compute** для запуска анализа.

3. Подождите немного дольше, чем вы делали в Neighbour Joining analysis, т.к. в алгоритме максимального правдоподобия используется намного больше вычислений. Индикатор выполнения покажет, в какой степени ваш анализ завершен. После того, как анализ завершится, откроется TreeViewer и отобразит полученное дерево.
4. Теперь вы можете совершать все те же манипуляции с деревом, которые делали раньше.
   a. ***Отличается ли полученное дерево от предыдущих?***
   b. ***Сравните порядок ветвления полученного дерева с другими, есть ли разница?***

Контрольный вопрос 3
Какова бутстрап - оценка для филогенетической ветви (клады) P. Marinus?

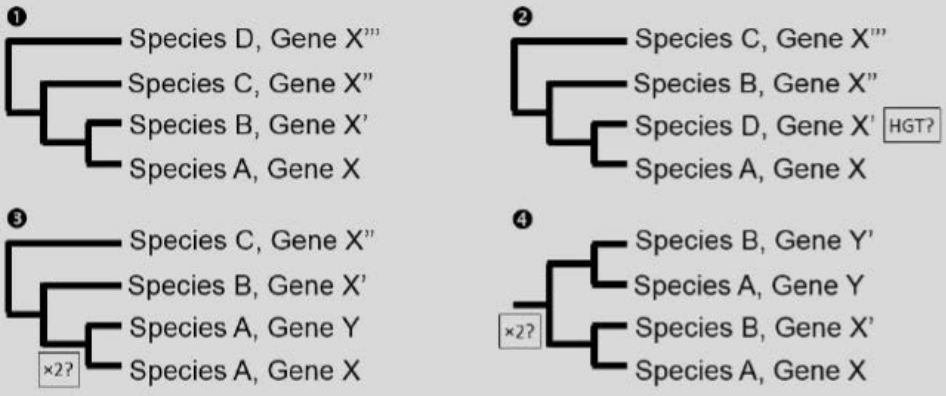
- Одним из сильных способов поддержки филогенетического анализа является выполнение его с использованием не менее двух независимых методов. Если вы получаете ту же базовую топологию, используя оба метода, neighbour-joining и ML, например, то у вас есть очень веские

основания полагать, что ваш анализ корректен. Почти каждый журнал по молекулярной эволюции требует, чтобы филогенетический анализ производился с использованием нескольких подходов.

5. Вернитесь назад и попробуйте поиграть опциями Maximum Likelihood analysis в MEGA для того, чтобы увидеть как они влияют на топологию дерева.

В 7. Интерпретация результатов филогенетического анализа

Филогенетический анализ весьма мощный для определения эволюционной истории интересующего нас гена. Рассмотрим следующие четыре сценария, и предположим, что все узлы имеют хорошие (высокие уровни) бутстрап- поддержки. Гены X, X ', X', X " ' являются ортологами , ген Y и Y' являются паралогичными к этим генами. Предположим виды A, B, C и D будут более отдаленно связаны друг с другом , чем дальше по алфавиту находятся. В сценарии 1, ген падает в ожидаемом клады, то есть дерево для гена аналогично дереву видов. Здесь нет сюрпризов. В сценарии 2, ген не выпадает, где мы ожидаем, на основании дерева видов. Этот ген может быть приобретен в случае горизонтального переноса HGT? от видов (или близкородственных видов), с которыми он группируется . В сценарии 3, существует паралог у вид A, но нет с никаких паралогов в других видах. Предполагая, что геномы видов были отсеквинированы и порог E-value не был слишком жестким, это может свидетельствовать о частичной дупликации или дупликации всего генома, обозначается × 2 ?. В сценарии 4 имеются паралоги гена также у гомологов других видов. Опять же, с теми же оговорками хорошего покрытия генома и надлежашей отсечки E-value (E-value cutoff), это может означать, что событие дупликации произошло у предков обоих видов, в точке, обозначенной × 2 ?.



Полезные ссылки:

MEGA 5 http://www.megasoftware.net/
PHYLIP http://evolution.genetics.washington.edu/phylip.html
Он-лайн ML анализ http://bar.utoronto.ca/webphylip/
http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::fastdnaml

Приобретенные умения и навыки

- знание терминологии, построением дендрограмм и быть в состоянии идентифицироватьсамого последнего общего предка любых двух терминальных узлов (таксонов) на дереве;

- знание основных элементов и терминологии филогении и возможные эволюционные пути к данному полученному состоянию (как могут возникать сходные формы (homoplasy);

- быть в состоянии идентифицировать корень дерева и знать разницу между укорененными и некорневых деревьями;

- иметь представление о методах филогенетического анализа, понимать , как они работают, плюсы и минусы каждого из них;

- познакомиться с различными моделями замен;

- знакомство с бутстрапом, понимание, о чем свидетельствует бутстрап- оценка на узле;

- практическое умение строить деревья методами объединения ближайших соседей и максимального правдоподобия, используя MEGA

Дополнительная литература

Chapter 7 “Recovering Evolutionary History” in *Understanding Bioinformatics* by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. pp 223-264.

Chapter 8 “Building Phylogenetic Trees” in *Understanding Bioinformatics* by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. pp 267-311.

K Tamura, J Dudley, M Nei, S Kumar (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24(8):1596-9.

WF Doolittle (1999) Phylogenetic classification and the universal tree. Science 284: 2124-2128.

RDM Page and MA Charleston (1997) From gene to organismal phylogeny: reconcidle trees and the gene / species tree problem. Mol. Phylogenet. Evol. 7:231-240.

N Saitou and M Nei (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.

S Guidon and O Gascuel (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52: 696-704.

T Sitnikova (1996) Bootstrap method of interior-branch test for phylogenetic trees. Mol. Biol. Evol. 13: 605-611.

Appendix 1: Форматы файлов

Fasta Format

```
>A_thaliana_79565543    ATCAGCGATATCCCAAGAAGAACAAAGTTTCAGAAACATCATCGAGGAAGAATTAATAAA
GGAGTATCTTCTCAGGGGTATATTTGTAGTAGATATGCTCTTCAAACACTTGAACCAGCT
TGGATCACTTCTAGACAAATAGAAGCAGGACGACGAGCAATGAC
>Azoarcus_sp_119896292    ATGCTGCAGCCGTCGAGAAGGAAATACCGCAAGGAGCAGAAAGGTCGCAACACCGGCCTG
GCGACGCGCGGCACCAAGGTCAGCTTCGGTGATTTCGGTCTGAAGGCGATCGCCCGCGGT
CGTCTGACTGCCCGTCAGATTGAATCCGCGCGTCGCGCGATGAC
>P_marinus_124024712    ATGCTTAGCCCAAAAAGAACCAAATTTCGTAAACAACAAAGAGGCCGTATGCGCGGTGTT
GCTACTAGAGGCAACAAAATCGCTTTTGGTCAGTTTGCATTGCAAGCTCAAGACTGTGGA
TGGGTCACTTCAAGGCAAATCGAGGCAAGTCGACGAGCAATGAC
>H_halophila_121996810    ATGTTACAGCCGAAACGGACCAAGTACCGCAAGAAGCAAAAGGGCCGCTGCTCGGGCCTC
GCGACCCGCGGTGATCGCGTGAGCTTCGGCGAGTTCGGCCTCAAGGCAACCACCCGCGGG
CCGATCACCTCGCGGCAGATCGAGGCGGCGCGGCGTGCCATCAA
>B_bacilliformis_121601635
ATGTTGCAGCCAAAGCGCACAAAGTTCCGTAAGCAATTCAAAGGTCGTATTCACGGTGCT
TCGAAAGGTGGTACGGATTTGAATTTTGGTGCTTACGGCCTGAAGGTTGTCGAGCCAGAG
CGTATTACTGCGCGTCAAATTGAAGCAGCTCGTCGTGCAATTAC
>S_aciditrophicus_85857845
ATGTTAATGCCAAAAAGGGTGAAATATAGGAAGTTGCAAAGGGGTCGAAGGACAGGAACC
GCCACAAGAGGAAGTAAAATATCTTTTGGGGAATATGGACTTCAAGCAGAAGAATGTGGC
TGGATAACCGCAAGGCAGATTGAGGCAGCGAGAATTGCCATTAC
>S_elongatus_81298811    ATGCTCAGTCCACGTCGTACCAAATTCCGGAAGCAGCAACGTGGCCGCATGACCGGCAAA
GCGACGCGCGGGAATACTCTCGCCTTCGGTAACTTCGGTCTGCAGGCGCTGGAATGCTCC
TGGATCACGGCTCGCCAAATTGAGGCTAGCCGTCGTGCCATGAC
>G_metallireducens_78221228
ATGTTGATGCCCAAAAGAGTTAAGTATAGAAAGCAAATGAAGGGGCGCATGACGGGCGCT
GCAATGCGCGGGGCCACACTGTCGTACGGTGATTTCGGTCTCCAGGCAACGGAGTGTGGA
TGGGTTGATTCCCGTCAGATAGAGGCTGCTCGTATTGCAATGAC
>Synechococcus_sp_78211558
ATGCTGAGTCCAAAACGCGTCAAATTCCGTAAGCAGCAGCGAGGCCGCATGCGCGGCGTC
GCCACCCGGGGCAACACCATTGCCTTCGGACAATTCGCGCTGCAGGCACAGGAATGTGGC
TGGATCACCTCGCGCCAGATCGAGGCCAGCCGTCGTGCCATGAC
>B_henselae_49474831    ATGTTGCAGCCAAAGCGCACAAGGTTCCGTAAACAGTTCAAAGGTCGTATTCATGGTGTT
TCGAAAGGTGGTACGGATTTGAATTTCGGTGCTTATGGTTTAAAAGCTGTTGAACCGGAG
CGGATTACTGCCCGCCAAATTGAGGCGGCGCGTCGTGCGATTAC
>B_quintana_49473688    ATGTTGCAGCCAAAGCGCACAAGGTTCCGTAAACAATTCAAAGGTCGTATTCACGGTGTT
TCGAAAGGTGGTACGGATCTAAATTTCGGTGCTTATGGTTTGAAAGCTGTTGAGCCGGAG
CGGATTACTGCCCGCCAAATTGAAGCGGCGCGTCGTGCGATTAC
```

MEGA Format

```
#mega
TITLE: Written by EMBOSS 27/01/09
#A_thaliana_79565543     ATCAGCGATATCCCAAGAAGAACAAAGTTTCAGAAACATCATCGAGGAAG
#Azoarcus_sp_11989629      ATGCTGCAGCCGTCGAGAAGGAAATACCGCAAGGAGCAGAAAGGTCGCAA
#P_marinus_124024712     ATGCTTAGCCCAAAAAGAACCAAATTTCGTAAACAACAAAGAGGCCGTAT
#H_halophila_12199681      ATGTTACAGCCGAAACGGACCAAGTACCGCAAGAAGCAAAAGGGCCGCTG
#B_bacilliformis_1216      ATGTTGCAGCCAAAGCGCACAAAGTTCCGTAAGCAATTCAAAGGTCGTAT
#S_aciditrophicus_858      ATGTTAATGCCAAAAAGGGTGAAATATAGGAAGTTGCAAAGGGGTCGAAG
#S_elongatus_81298811      ATGCTCAGTCCACGTCGTACCAAATTCCGGAAGCAGCAACGTGGCCGCAT
#G_metallireducens_78      ATGTTGATGCCCAAAAGAGTTAAGTATAGAAAGCAAATGAAGGGGCGCAT
#Synechococcus_sp_782      ATGCTGAGTCCAAAACGCGTCAAATTCCGTAAGCAGCAGCGAGGCCGCAT
#B_henselae_49474831     ATGTTGCAGCCAAAGCGCACAAGGTTCCGTAAACAGTTCAAAGGTCGTAT
#B_quintana_49473688     ATGTTGCAGCCAAAGCGCACAAGGTTCCGTAAACAATTCAAAGGTCGTAT
#A_thaliana_79565543     AATTAATAAAGGAGTATCTTCTCAGGGGTATATTTGTAGTAGATATGCTC
#Azoarcus_sp_11989629      CACCGGCCTGGCGACGCGCGGCACCAAGGTCAGCTTCGGTGATTTCGGTC
#P_marinus_124024712     GCGCGGTGTTGCTACTAGAGGCAACAAAATCGCTTTTGGTCAGTTTGCAT
#H_halophila_12199681      CTCGGGCCTCGCGACCCGCGGTGATCGCGTGAGCTTCGGCGAGTTCGGCC
#B_bacilliformis_1216      TCACGGTGCTTCGAAAGGTGGTACGGATTTGAATTTTGGTGCTTACGGCC
#S_aciditrophicus_858      GACAGGAACCGCCACAAGAGGAAGTAAAATATCTTTTGGGGAATATGGAC
#S_elongatus_81298811      GACCGGCAAAGCGACGCGCGGGAATACTCTCGCCTTCGGTAACTTCGGTC
#G_metallireducens_78      GACGGGCGCTGCAATGCGCGGGGCCACACTGTCGTACGGTGATTTCGGTC
#Synechococcus_sp_782      GCGCGGCGTCGCCACCCGGGGCAACACCATTGCCTTCGGACAATTCGCGC
#B_henselae_49474831     TCATGGTGTTTCGAAAGGTGGTACGGATTTGAATTTCGGTGCTTATGGTT
#B_quintana_49473688     TCACGGTGTTTCGAAAGGTGGTACGGATCTAAATTTCGGTGCTTATGGTT
#A_thaliana_79565543     TTCAAACACTTGAACCAGCTTGGATCACTTCTAGACAAATAGAAGCAGGA
#Azoarcus_sp_11989629      TGAAGGCGATCGCCCGCGGTCGTCTGACTGCCCGTCAGATTGAATCCGCG
#P_marinus_124024712     TGCAAGCTCAAGACTGTGGATGGGTCACTTCAAGGCAAATCGAGGCAAGT
#H_halophila_12199681      TCAAGGCAACCACCCGCGGGCCGATCACCTCGCGGCAGATCGAGGCGGCG
#B_bacilliformis_1216      TGAAGGTTGTCGAGCCAGAGCGTATTACTGCGCGTCAAATTGAAGCAGCT
#S_aciditrophicus_858      TTCAAGCAGAAGAATGTGGCTGGATAACCGCAAGGCAGATTGAGGCAGCG
#S_elongatus_81298811      TGCAGGCGCTGGAATGCTCCTGGATCACGGCTCGCCAAATTGAGGCTAGC
#G_metallireducens_78      TCCAGGCAACGGAGTGTGGATGGGTTGATTCCCGTCAGATAGAGGCTGCT
#Synechococcus_sp_782      TGCAGGCACAGGAATGTGGCTGGATCACCTCGCGCCAGATCGAGGCCAGC
#B_henselae_49474831     TAAAAGCTGTTGAACCGGAGCGGATTACTGCCCGCCAAATTGAGGCGGCG
#B_quintana_49473688     TGAAAGCTGTTGAGCCGGAGCGGATTACTGCCCGCCAAATTGAAGCGGCG
#A_thaliana_79565543     CGACGAGCAATGAC
#Azoarcus_sp_11989629    CGTCGCGCGATGAC
#P_marinus_124024712     CGACGAGCAATGAC
#H_halophila_12199681    CGGCGTGCCATCAA
#B_bacilliformis_1216    CGTCGTGCAATTAC
#S_aciditrophicus_858    AGAATTGCCATTAC
#S_elongatus_81298811    CGTCGTGCCATGAC
#G_metallireducens_78    CGTATTGCAATGAC
#Synechococcus_sp_782    CGTCGTGCCATGAC
#B_henselae_49474831     CGTCGTGCGATTAC
#B_quintana_49473688     CGTCGTGCGATTAC
```

Clustal Format

```
A_thaliana_7956     ATCAGCGATATCCCAAGAAGAACAAAGTTTCAGAAACATCATCGAGGAAGAATTAATAAA
Azoarcus_sp_119     ATGCTGCAGCCGTCGAGAAGGAAATACCGCAAGGAGCAGAAAGGTCGCAACACCGGCCTG
P_marinus_12402     ATGCTTAGCCCAAAAAGAACCAAATTTCGTAAACAACAAAGAGGCCGTATGCGCGGTGTT
H_halophila_121     ATGTTACAGCCGAAACGGACCAAGTACCGCAAGAAGCAAAAGGGCCGCTGCTCGGGCCTC
B_bacilliformis     ATGTTGCAGCCAAAGCGCACAAAGTTCCGTAAGCAATTCAAAGGTCGTATTCACGGTGCT
S_aciditrophicu     ATGTTAATGCCAAAAAGGGTGAAATATAGGAAGTTGCAAAGGGGTCGAAGGACAGGAACC
S_elongatus_812     ATGCTCAGTCCACGTCGTACCAAATTCCGGAAGCAGCAACGTGGCCGCATGACCGGCAAA
G_metallireduce     ATGTTGATGCCCAAAAGAGTTAAGTATAGAAAGCAAATGAAGGGGCGCATGACGGGCGCT
Synechococcus_s     ATGCTGAGTCCAAAACGCGTCAAATTCCGTAAGCAGCAGCGAGGCCGCATGCGCGGCGTC
B_henselae_4947     ATGTTGCAGCCAAAGCGCACAAGGTTCCGTAAACAGTTCAAAGGTCGTATTCATGGTGTT
B_quintana_4947   ATGTTGCAGCCAAAGCGCACAAGGTTCCGTAAACAATTCAAAGGTCGTATTCACGGTGTT

A_thaliana_7956     GGAGTATCTTCTCAGGGGTATATTTGTAGTAGATATGCTCTTCAAACACTTGAACCAGCT
Azoarcus_sp_119     GCGACGCGCGGCACCAAGGTCAGCTTCGGTGATTTCGGTCTGAAGGCGATCGCCCGCGGT
P_marinus_12402     GCTACTAGAGGCAACAAAATCGCTTTTGGTCAGTTTGCATTGCAAGCTCAAGACTGTGGA
H_halophila_121     GCGACCCGCGGTGATCGCGTGAGCTTCGGCGAGTTCGGCCTCAAGGCAACCACCCGCGGG
B_bacilliformis     TCGAAAGGTGGTACGGATTTGAATTTTGGTGCTTACGGCCTGAAGGTTGTCGAGCCAGAG
S_aciditrophicu     GCCACAAGAGGAAGTAAAATATCTTTTGGGGAATATGGACTTCAAGCAGAAGAATGTGGC
S_elongatus_812     GCGACGCGCGGGAATACTCTCGCCTTCGGTAACTTCGGTCTGCAGGCGCTGGAATGCTCC
G_metallireduce     GCAATGCGCGGGGCCACACTGTCGTACGGTGATTTCGGTCTCCAGGCAACGGAGTGTGGA
Synechococcus_s     GCCACCCGGGGCAACACCATTGCCTTCGGACAATTCGCGCTGCAGGCACAGGAATGTGGC
B_henselae_4947     TCGAAAGGTGGTACGGATTTGAATTTCGGTGCTTATGGTTTAAAAGCTGTTGAACCGGAG
B_quintana_4947   TCGAAAGGTGGTACGGATCTAAATTTCGGTGCTTATGGTTTGAAAGCTGTTGAGCCGGAG

A_thaliana_7956     TGGATCACTTCTAGACAAATAGAAGCAGGACGACGAGCAATGAC
Azoarcus_sp_119     CGTCTGACTGCCCGTCAGATTGAATCCGCGCGTCGCGCGATGAC
P_marinus_12402     TGGGTCACTTCAAGGCAAATCGAGGCAAGTCGACGAGCAATGAC
H_halophila_121     CCGATCACCTCGCGGCAGATCGAGGCGGCGCGGCGTGCCATCAA
B_bacilliformis     CGTATTACTGCGCGTCAAATTGAAGCAGCTCGTCGTGCAATTAC
S_aciditrophicu     TGGATAACCGCAAGGCAGATTGAGGCAGCGAGAATTGCCATTAC
S_elongatus_812     TGGATCACGGCTCGCCAAATTGAGGCTAGCCGTCGTGCCATGAC
G_metallireduce     TGGGTTGATTCCCGTCAGATAGAGGCTGCTCGTATTGCAATGAC
Synechococcus_s     TGGATCACCTCGCGCCAGATCGAGGCCAGCCGTCGTGCCATGAC
B_henselae_4947     CGGATTACTGCCCGCCAAATTGAGGCGGCGCGTCGTGCGATTAC
B_quintana_4947   CGGATTACTGCCCGCCAAATTGAAGCGGCGCGTCGTGCGATTAC
```

PHYLIP Interleaved Format

```
11 164
A_thalianaATCAGCGATA   TCCCAAGAAG  AACAAAGTTT  CAGAAACATC  ATCGAGGAAG
Azoarcus_sATGCTGCAGC   CGTCGAGAAG  GAAATACCGC  AAGGAGCAGA  AAGGTCGCAA
P_marinus_ATGCTTAGCC   CAAAAAGAAC  CAAATTTCGT  AAACAACAAA  GAGGCCGTAT
H_halophilATGTTACAGC   CGAAACGGAC  CAAGTACCGC  AAGAAGCAAA  AGGGCCGCTG
B_bacillifATGTTGCAGC   CAAAGCGCAC  AAAGTTCCGT  AAGCAATTCA  AAGGTCGTAT
S_aciditroATGTTAATGC   CAAAAAGGGT  GAAATATAGG  AAGTTGCAAA  GGGGTCGAAG
S_elongatuATGCTCAGTC   CACGTCGTAC  CAAATTCCGG  AAGCAGCAAC  GTGGCCGCAT
G_metallirATGTTGATGC   CCAAAAGAGT  TAAGTATAGA  AAGCAAATGA  AGGGGCGCAT
SynechococATGCTGAGTC   CAAAACGCGT  CAAATTCCGT  AAGCAGCAGC  GAGGCCGCAT
B_henselaeATGTTGCAGC   CAAAGCGCAC  AAGGTTCCGT  AAACAGTTCA  AAGGTCGTAT
B_quintanaATGTTGCAGC   CAAAGCGCAC  AAGGTTCCGT  AAACAATTCA  AAGGTCGTAT

           AATTAATAAA  GGAGTATCTT  CTCAGGGGTA  TATTTGTAGT  AGATATGCTC
           CACCGGCCTG  GCGACGCGCG  GCACCAAGGT  CAGCTTCGGT  GATTTCGGTC
           GCGCGGTGTT  GCTACTAGAG  GCAACAAAAT  CGCTTTTGGT  CAGTTTGCAT
           CTCGGGCCTC  GCGACCCGCG  GTGATCGCGT  GAGCTTCGGC  GAGTTCGGCC
           TCACGGTGCT  TCGAAAGGTG  GTACGGATTT  GAATTTTGGT  GCTTACGGCC
           GACAGGAACC  GCCACAAGAG  GAAGTAAAAT  ATCTTTTGGG  GAATATGGAC
           GACCGGCAAA  GCGACGCGCG  GGAATACTCT  CGCCTTCGGT  AACTTCGGTC
           GACGGGCGCT  GCAATGCGCG  GGGCCACACT  GTCGTACGGT  GATTTCGGTC
           GCGCGGCGTC  GCCACCCGGG  GCAACACCAT  TGCCTTCGGA  CAATTCGCGC
           TCATGGTGTT  TCGAAAGGTG  GTACGGATTT  GAATTTCGGT  GCTTATGGTT
           TCACGGTGTT  TCGAAAGGTG  GTACGGATCT  AAATTTCGGT  GCTTATGGTT

           TTCAAACACT  TGAACCAGCT  TGGATCACTT  CTAGACAAAT  AGAAGCAGGA
           TGAAGGCGAT  CGCCCGCGGT  CGTCTGACTG  CCCGTCAGAT  TGAATCCGCG
           TGCAAGCTCA  AGACTGTGGA  TGGGTCACTT  CAAGGCAAAT  CGAGGCAAGT
           TCAAGGCAAC  CACCCGCGGG  CCGATCACCT  CGCGGCAGAT  CGAGGCGGCG
           TGAAGGTTGT  CGAGCCAGAG  CGTATTACTG  CGCGTCAAAT  TGAAGCAGCT
           TTCAAGCAGA  AGAATGTGGC  TGGATAACCG  CAAGGCAGAT  TGAGGCAGCG
           TGCAGGCGCT  GGAATGCTCC  TGGATCACGG  CTCGCCAAAT  TGAGGCTAGC
           TCCAGGCAAC  GGAGTGTGGA  TGGGTTGATT  CCCGTCAGAT  AGAGGCTGCT
           TGCAGGCACA  GGAATGTGGC  TGGATCACCT  CGCGCCAGAT  CGAGGCCAGC
           TAAAAGCTGT  TGAACCGGAG  CGGATTACTG  CCCGCCAAAT  TGAGGCGGCG
           TGAAAGCTGT  TGAGCCGGAG  CGGATTACTG  CCCGCCAAAT  TGAAGCGGCG

           CGACGAGCAA  TGAC
           CGTCGCGCGA  TGAC
           CGACGAGCAA  TGAC
           CGGCGTGCCA  TCAA
           CGTCGTGCAA  TTAC
           AGAATTGCCA  TTAC
           CGTCGTGCCA  TGAC
           CGTATTGCAA  TGAC
           CGTCGTGCCA  TGAC
           CGTCGTGCGA  TTAC
           CGTCGTGCGA  TTAC
```

PHYLIP Non-interleaved Format

```
11 164
A_thalianaATCAGCGATA TCCCAAGAAG AACAAAGTTT CAGAAACATC ATCGAGGAAG
          AATTAATAAA GGAGTATCTT CTCAGGGGTA TATTTGTAGT AGATATGCTC
          TTCAAACACT TGAACCAGCT TGGATCACTT CTAGACAAAT AGAAGCAGGA
          CGACGAGCAA TGAC
Azoarcus_sATGCTGCAGC CGTCGAGAAG GAAATACCGC AAGGAGCAGA AAGGTCGCAA
          CACCGGCCTG GCGACGCGCG GCACCAAGGT CAGCTTCGGT GATTTCGGTC
          TGAAGGCGAT CGCCCGCGGT CGTCTGACTG CCCGTCAGAT TGAATCCGCG
          CGTCGCGCGA TGAC
P_marinus_ATGCTTAGCC CAAAAAGAAC CAAATTTCGT AAACAACAAA GAGGCCGTAT
          GCGCGGTGTT GCTACTAGAG GCAACAAAAT CGCTTTTGGT CAGTTTGCAT
          TGCAAGCTCA AGACTGTGGA TGGGTCACTT CAAGGCAAAT CGAGGCAAGT
          CGACGAGCAA TGAC
H_halophilATGTTACAGC CGAAACGGAC CAAGTACCGC AAGAAGCAAA AGGGCCGCTG
          CTCGGGCCTC GCGACCCGCG GTGATCGCGT GAGCTTCGGC GAGTTCGGCC
          TCAAGGCAAC CACCCGCGGG CCGATCACCT CGCGGCAGAT CGAGGCGGCG
          CGGCGTGCCA TCAA
B_bacillifATGTTGCAGC CAAAGCGCAC AAAGTTCCGT AAGCAATTCA AAGGTCGTAT
          TCACGGTGCT TCGAAAGGTG GTACGGATTT GAATTTTGGT GCTTACGGCC
          TGAAGGTTGT CGAGCCAGAG CGTATTACTG CGCGTCAAAT TGAAGCAGCT
          CGTCGTGCAA TTAC
S_aciditroATGTTAATGC CAAAAAGGGT GAAATATAGG AAGTTGCAAA GGGGTCGAAG
          GACAGGAACC GCCACAAGAG GAAGTAAAAT ATCTTTTGGG GAATATGGAC
          TTCAAGCAGA AGAATGTGGC TGGATAACCG CAAGGCAGAT TGAGGCAGCG
          AGAATTGCCA TTAC
S_elongatuATGCTCAGTC CACGTCGTAC CAAATTCCGG AAGCAGCAAC GTGGCCGCAT
          GACCGGCAAA GCGACGCGCG GGAATACTCT CGCCTTCGGT AACTTCGGTC
          TGCAGGCGCT GGAATGCTCC TGGATCACGG CTCGCCAAAT TGAGGCTAGC
          CGTCGTGCCA TGAC
G_metallirATGTTGATGC CCAAAAGAGT TAAGTATAGA AAGCAAATGA AGGGGCGCAT
          GACGGGCGCT GCAATGCGCG GGGCCACACT GTCGTACGGT GATTTCGGTC
          TCCAGGCAAC GGAGTGTGGA TGGGTTGATT CCCGTCAGAT AGAGGCTGCT
          CGTATTGCAA TGAC
SynechococATGCTGAGTC CAAAACGCGT CAAATTCCGT AAGCAGCAGC GAGGCCGCAT
          GCGCGGCGTC GCCACCCGGG GCAACACCAT TGCCTTCGGA CAATTCGCGC
          TGCAGGCACA GGAATGTGGC TGGATCACCT CGCGCCAGAT CGAGGCCAGC
          CGTCGTGCCA TGAC
B_henselaeATGTTGCAGC CAAAGCGCAC AAGGTTCCGT AAACAGTTCA AAGGTCGTAT
          TCATGGTGTT TCGAAAGGTG GTACGGATTT GAATTTCGGT GCTTATGGTT
          TAAAAGCTGT TGAACCGGAG CGGATTACTG CCCGCCAAAT TGAGGCGGCG
          CGTCGTGCGA TTAC
B_quintanaATGTTGCAGC CAAAGCGCAC AAGGTTCCGT AAACAATTCA AAGGTCGTAT
          TCACGGTGTT TCGAAAGGTG GTACGGATCT AAATTTCGGT GCTTATGGTT
          TGAAAGCTGT TGAGCCGGAG CGGATTACTG CCCGCCAAAT TGAAGCGGCG
          CGTCGTGCGA TTAC
```