Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Иркутский государственный университет»
Биолого-почвенный факультет

Д. Ю. Щербаков
Р. В. Адельшин
М. В. Коваленкова

АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОВРЕМЕННОЙ ГЕНЕТИКИ

БИОИНФОРМАЦИОННЫЕ МЕТОДЫ АНАЛИЗА БИОРАЗНООБРАЗИЯ

Учебное пособие

ИРКУТСК
ИЗДАТЕЛЬСТВО
ИГУ
2018

УДК 575(075.8)
ББК 28.04я73
Щ61

Печатается по решению
учебно-методической комиссии
биолого-почвенного факультета ИГУ

Авторы:
Д. Ю. Щербаков (предисловие, разд. 1, 3–9, заключение),
Р. В. Адельшин (разд. 2, 10), М. В. Коваленкова (разд. 11)

Рецензенты:
д-р биол. наук, проф. *Ю. М. Константинов,*
канд. биол. наук, доц. *А. А. Приставка*

Щербаков Д. Ю.

Щ61 Актуальные проблемы современной генетики: биоинформационные методы анализа биоразнообразия : учеб. пособие / Д. Ю. Щербаков, Р. В. Адельшин, М. В. Коваленкова. – Иркутск : Изд-во ИГУ, 2018. – 119 с.

ISBN 978-5-9624-1600-7

Рассматриваются современные методы и подходы к анализу генетической информации организмов с целью выяснения их эволюционных историй, геногеографии и популяционных параметров. Приведены примеры применения наиболее распространенных программ для решения конкретных задач.

Издание предназначено для студентов бакалавриата и магистратуры, обучающихся по направлению «Биология», а также аспирантов и научных сотрудников, интересующихся использованием биоинформатики для исследования проблем экологии и эволюции.

УДК 575(075.8)
ББК 28.04я73

ISBN 978-5-9624-1600-7

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Авторы учебного пособия ставили перед собой задачу дать представление о современных подходах к анализу генетической информации организмов с целью выяснения их эволюционных историй, геногеографии и популяционных параметров. Эти сведения необходимы для освоения дисциплины «Биоинформационные методы анализа биоразнообразия», посвященной изучению изменчивости и разнообразия живых организмов. Понимание теоретической основы методов биоинформатики относится к фундаментальным знаниям, которые необходимы для приобретения профессиональных компетенций, необходимых студентам биологических специальностей. Усвоение студентами знаний по дисциплине предполагает предварительное освоение ботаники, зоологии, генетики, высшей математики, биометрии, основ информатики и программирования.

В учебном пособии изложены теоретические и практические аспекты использования биоинформатики для исследования процессов микро- и макроэволюции живых организмов. Рассматриваются разделы генетики популяций, молекулярной генетики и биоинформатики, необходимые для начала самостоятельной практической работы в данной области, дается представление об использовании собственных скриптов.

Для приобретения практических умений и навыков необходимо наличие интернета и компьютера с установленными на нём необходимыми программами.

1. О РОЛИ БИОИНФОРМАТИКИ В ГЕНЕТИЧЕСКИХ ИССЛЕДОВАНИЯХ

На современном этапе развития биологии практически невозможно придумать задачу, которая в том или ином виде ранее не привлекала бы внимания исследователей. Поэтому при постановке любого вопроса необходимо точно представлять себе, что уже известно, какие результаты уже получили ваши предшественники, работающие с помощью самых разнообразных методов в совершенно необычных областях биологии. Если на самом начальном этапе исследования такая работа не будет проделана, то в лучшем случае вы изобретёте довольно уродливый «велосипед», в худшем – станете предметом для недоумённых насмешек. Ещё пару десятилетий назад для знакомства с состоянием дел в любой области биологии достаточно было прочитать несколько современных обзоров и ряд статей в относительно небольшом числе специальных реферируемых журналов. Представление об общем контексте можно было получить из относительно небольшого количества книг. Соответственно, свои результаты следовало излагать в виде научных публикаций, рассчитывая на критику и оценку коллег.

В конце XX в. устройство потока научной информации, с которым приходится иметь дело исследователю, существенно изменилось в сторону усложнения. Эти изменения можно кратко свести к следующим пунктам:

- Число специальных журналов взрывообразно возросло. Раньше существовала единственная модель научного журнала – периодическое издание, на которое подписываются научные библиотеки или индивидуальные исследователи. Опубликовать статью в таком журнале можно было бесплатно (хотя и существовало несколько журналов, которые требовали от авторов денег либо за публикацию статьи, либо – за цветные иллюстра-

ции). В последнее время широко распространилась другая модель, согласно которой доступ к научной публикации бесплатен, но авторы платят довольно много за публикацию. Такие статьи, как правило, доступны в интернете, а бумажные копии играют подчиненную роль. В результате образовалось довольно много мелких журналов, которые, скорее, являются источниками спама, чем достоверной научной информации. Следовательно, надежность публикаций в реферируемых журналах в целом понизилась, а ориентироваться в этом море стало существенно труднее.

- Появились отличные базы данных, в которых содержится колоссальное количество информации. Один из самых впечатляющих примеров – NCBI, которую мы и будем рассматривать далее, хотя есть много и более специализированных серверов вроде flybase (**http://flybase.org**) или treeoflife (**http://tolweb.org/tree/**). Невозможно в одном пособии описать даже схематично это колоссальное разнообразие. Достаточно сказать, что дважды в год выходит специальное издание журнала Nucleic Acids Research, посвященное описанию баз данных и ресурсам, предназначенным для работы с нуклеиновыми кислотами и сопутствующими сведениями онлайн.

- Очень многие задачи на современном этапе развития биологии связаны с анализом очень больших объемов информации. Несмотря на то что начальные этапы любой из таких работ можно выполнить стандартными средствами, которые в основном описаны в настоящем пособии, обязательно в конце концов приходится столкнуться с отсутствием необходимых инструментов, когда придется воспользоваться инструментами программирования. Практически любая операция с очень большим количеством информации в ручном режиме не возможна, и приходится пользоваться простыми программами на интерпретируемых языках программирования – скриптами.

2. ПОИСК НАБОРОВ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

На этапе планирования любой работы, которая связана с анализом генетического разнообразия, как, впрочем, и на всех последующих этапах, необходимо выяснить, какие данные уже получены в интересующей исследователя области другими. Связанные друг с другом общедоступные базы данных о нуклеотидных последовательностях ДНК и последовательностях аминокислот содержат практически все, что было опубликовано, и очень много неопубликованной информации. Помимо последовательностей соответствующих молекул, эти базы данных содержат самые разнообразные сведения, от таксономии до публикаций, из которых можно выяснить, по крайней мере, зачем и кто получил заинтересовавшие исследователя данные.

Поиск нуклеотидных последовательностей отличается от обычного поиска информации в базах данных в первую очередь тем, что часто носит «приблизительный» характер, поскольку происходит в рамках так называемых синонимических множеств, т. е. не совпадающих последовательностей нуклеотидов, которые, тем не менее, выполняют одну и ту же функцию. Определение границ таких множеств представляет собой весьма сложную задачу. Это можно проиллюстрировать на примере полипептидов, одни из которых могут различаться по последовательности более чем на 50 %, но играть одну и ту же биологическую роль (например, казеины – питательные белки молока). Другие, наоборот, у всех многоклеточных животных различаются всего на два аминокислотных остатка, подобно инсулину. В последнем случае картина разнообразия последовательностей осложняется еще и наличием в тех же самых геномах (например, у человека) генов, кодирующих весьма схожие с инсулином полипептиды, выполняющие к тому же похожие функции (соматомедин, или инсулиноподобный фактор роста).

На сложность задачи поиска белковых или нуклеотидных последовательностей влияет и невообразимое количество возможных вариантов. Относительно коротких полипептидов длиной всего 100 аминокислотных остатков (приблизительно 10 кДа) может быть 20^{100}, или примерно 10^{120}, что существенно больше, чем общее число атомов во Вселенной. Перебор такого количества вариантов не в состоянии выполнить за обозримое время ни один из компьютеров. Поэтому приходится использовать либо так называемые логистические стратегии поиска, либо – пользоваться дополнительной информацией вроде контекста, в котором были упомянуты соответствующие фрагменты ДНК или белков, их номенклатурные или тривиальные названия. Естественно, что специализированные базы данных должны быть оборудованы полным набором инструментов, предназначенных для поиска информации и ее первичной обработки. Это инструменты весьма разнообразны и, к сожалению, не всегда просто устроены.

Важной чертой абсолютно всех рассматриваемых ниже баз данных является то, что они общедоступны и бесплатны.

2.1. Раздел GenBank: PopSet

GenBank (Генбанк) – главный ресурс, где представлена вся находящаяся в свободном доступе информация о нуклеотидных последовательностях, находится на веб-странице Американского национального центра биотехнологической информации (NCBI).

На рис. 2.1 показана стартовая страница Генбанка. На 6 сентября 2016 г. в этом банке было зарегистрировано всего 107 045 797 записей, включая 70 427 238 белков, 16 172 490 РНК. Последовательности принадлежат 62 739 организмам. База увеличивается каждый день и содержит просто невообразимое количество самой разнородной информации. Громадный объем приводит к трем основным следствиям.

В Генбанке не может не быть ошибок. Надо быть к этому готовым. Основные ошибки – это неправильная идентификация организмов – источников, ошибочная аннотация последовательностей и банальные ошибки секвенирования.

Рис. 2.1. Стартовая страница портала Генбанка, расположенная по адресу http://www.ncbi.nlm.nih.gov

Выполнение каждого запроса или последовательности операций на результате запроса на сервере – довольно продолжительная процедура. Особенно это верно для больших запросов, в результате выполнения которых можно ожидать появления тысяч записей. Поэтому следует ответственно относиться к экспериментам с запросами, по возможности выполнять большую часть работы локально и пользоваться специализированными инструментами для запросов – скриптами на «Биопитоне» (Biopython) или «Биоперле» (Bioperl). Имеет смысл также скачивать с сайта NCBI локальную копию всей базы данных и тренироваться с нею.

2.1.1. Организация запроса

Приведем пример выяснения вопроса о том, не исследовали ли филогенетические связи петуний с помощью довольно быстро эволюционирующего генетического маркера – находящегося в хлоропластном геноме гена, кодирующего большую субъединицу рибулезо 1,5-бифосфат карбоксилазы – оксигеназы (*rbcL*).

Определение раздела базы данных, к которой обращен запрос: в нашем случае это – целостный набор нуклеотидных последовательностей, опубликованный или просто зарегистрированный в Генбанке как единое целое, как это показано на рис. 2.2. При этом надо помнить, что в качестве PopSet'ов обычно регистрируют наборы последовательностей стандартных маркеров молекулярной эволюции – рибосомной РНК, транскрибируемого спейсера рибосомы РНК (ITS), митохондриальных генов первой субъединицы цитохром b-оксидазы

(COI, cox1), для растений – хлоропластного гена RUBISCO (rbcL, ribulose–1,5–bisphosphate carboxylase/oxygenase large subunit) и др. Наборы могут содержать как данные для популяционного анализа, так и коллекции, предназначенные для филогенетического анализа весьма далеких друг от друга таксонов.

Рис. 2.2. Выбор PopSet в качестве раздела Генбанка

Формирование запроса производится в соответствии с полями записи в соответствующем разделе Генбанка. Запрос формируется в соответствии со структурой записи. Простейшая запись приведена на рис. 2.2. Сама запись представляет собой текстовый файл на английском языке (т. е. не содержит букв кириллицы). Текст состоит из двух колонок, в левой заглавными буквами обозначено поле записи, а в правой содержится значение этого поля.

Интерес представляют поля с обозначением гена и с указанием таксономического положения организма. Для получения записи, из которой был выбран представленный на рис. 2.3. элемент, в строке запроса должно быть указано:

```
Petunia[ORGN]
```

Это значит, что будут показаны все записи, входящие в PopSet и относящиеся к петуниям. В данном случае их оказалось 90. Поэтому запрос лучше уточнить, сформулировав задание, как:

```
Petunia[ORGN] AND rbcL[GENE]
```

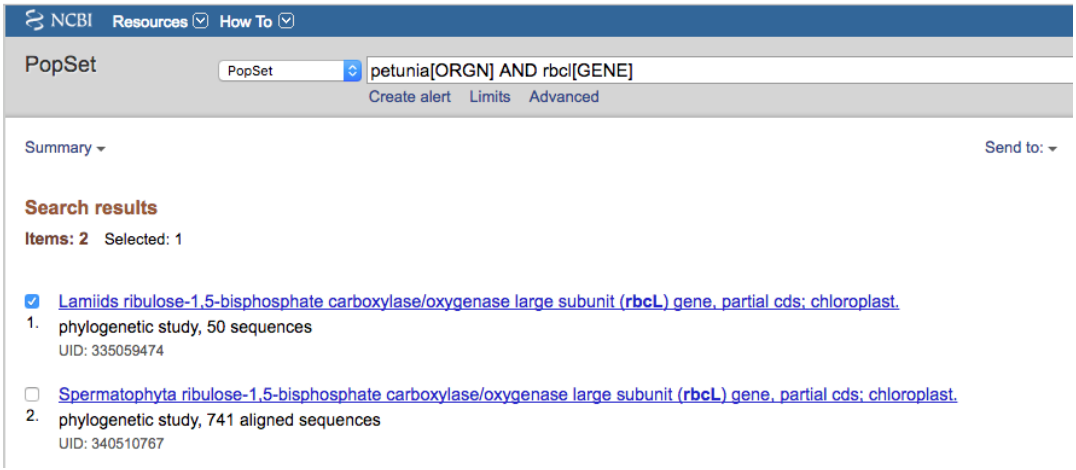


Рис. 2.3. Выбор одного из двух найденных наборов последовательностей rbcL

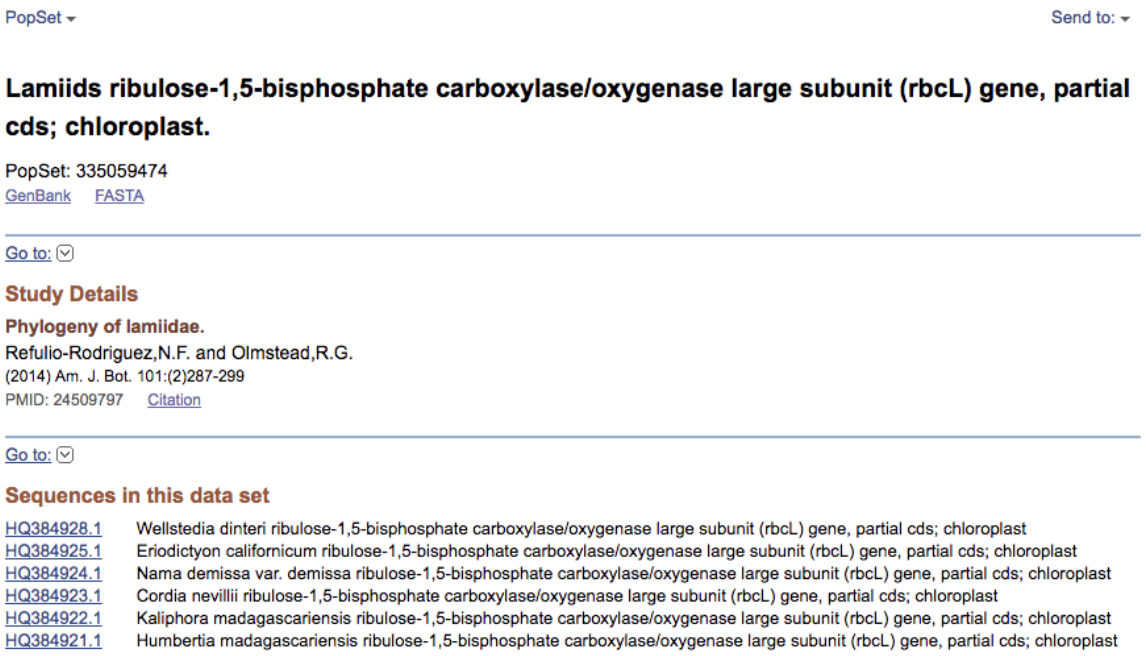
В результате поиск приводит к двум наборам последовательностей (см. рис. 2.3). Этот запрос фактически представляет собой команду серверу выдать ссылки на наборы последовательностей, удовлетворяющие двум требованиям:

- относиться к петуниям;
- содержать последовательности хлоропластного гена, кодирующего большую субъединицу карбоксилазы – оксигеназы.

Здесь `AND` – логический (булев) оператор, означающий, что оба требования должны выполняться одновременно. Запросы можно писать и гораздо более сложные, группируя условия с помощью скобок. Другие логические операторы – `OR` и `NOT`. Они обозначают, соответственно, логические ИЛИ и НЕ. С их помощью мы можем записать:

- `Petunia[ORGN] NOT rbcL[GENE]` обозначает, что вы ищете популяционные наборы с ДНК петунии, все, что угодно, но только не rbcL;
- `Petunia[ORGN] AND (rbcL[GENE] OR ITS[GENE])` обозначает, что требуется найти наборы ДНК петунии, содержащие либо ген rbcL, либо – последовательности транскрибируемого межгенного спейсера (ITS).

Выбор нужного результата поиска. Практически всегда поиск приводит к нескольким результатам (либо – к их отсутствию, но такая ситуация не интересна). Предположим, что выбран первый из наборов, тот, в котором 50 последовательностей. Первая строка представляет собой гиперссылку. Если воспользоваться этой ссылкой, то откроется краткое описание (Short summary) набора последовательностей (рис. 2.4).


PopSet ▾ Send to: ▾

Lamiids ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast.

PopSet: 335059474
GenBank FASTA

Go to:

Study Details

Phylogeny of lamiidae.
Refulio-Rodriguez,N.F. and Olmstead,R.G.
(2014) Am. J. Bot. 101:(2)287-299
PMID: 24509797 Citation

Go to:

Sequences in this data set

HQ384928.1	Wellstedia dinteri ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast
HQ384925.1	Eriodictyon californicum ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast
HQ384924.1	Nama demissa var. demissa ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast
HQ384923.1	Cordia nevillii ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast
HQ384922.1	Kaliphora madagascariensis ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast
HQ384921.1	Humbertia madagascariensis ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast

Рис. 2.4. Краткое описание набора последовательностей rbcL

2.2. Первичный анализ последовательностей

Первичный анализ последовательностей онлайн

Первичный анализ онлайн. После того как выбран для дальнейшей работы определенный набор последовательностей, появляется возможность его скачать на свой компьютер и продолжить их анализ офлайн. Обычно именно этим и кончается поиск. Однако портал NCBI предоставляет целый набор инструментов, позволяющий довольно много узнать о наборе данных. Для этого нам потребуется закладка, обозначенная как 2 на рис. 2.4.

На первом этапе есть только одна возможность – выровнять последовательности с помощью программы BLAST. Это далеко не самый совершенный инструмент для выравнивания последовательностей, но на сервере он используется как компромисс между точностью, быстродействием и нагрузкой на сервер. Необходимость выравнивания в качестве первого этапа анализа обусловлена тем, что при депонировании набора последовательностей на самом деле сервер их воспринимает по одной и, соответственно, удаляет все делеции. В итоге последовательности хранятся в сыром виде.

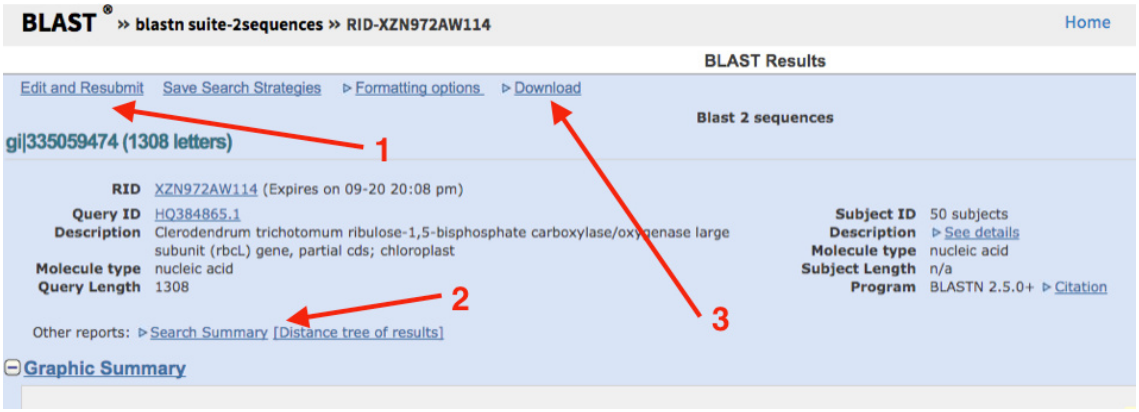


Рис. 2.4. Верхняя панель окна результатов выравнивания: *1* – ссылка на страницу, где можно отредактировать результат и запустить новый поиск с выравниванием; *2* – резюме выравнивания и построение филогенетического дерева; *3* – форматирование и загрузка результата на локальный компьютер

Предварительное выравнивание происходит, если щелкнуть мышью на ссылке 3 с рис. 2.4. Выравнивание не происходит мгновенно. Некоторое время приходится любоваться окном ожидания. Затем возникает окно с результатом, наиболее информативная часть которого – графическое представление выравнивания – приведена на рис. 2.5.

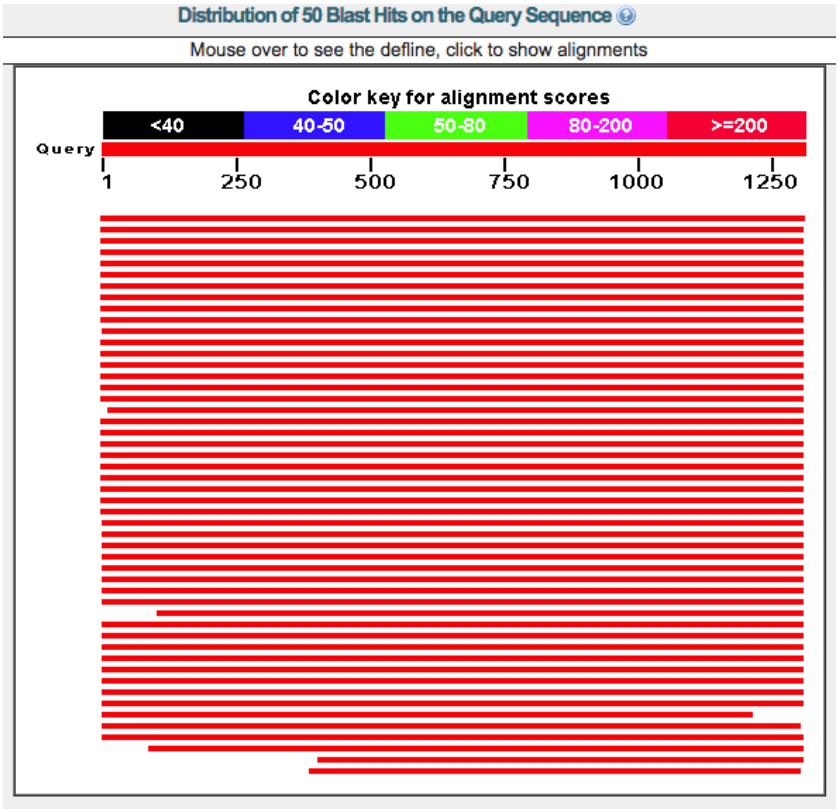


Рис. 2.5. Графическое представление результата выравнивания нуклеотидных последовательностей

Если посмотреть на нижние линии этого набора, то становится ясно, зачем проводилось выравнивание. У некоторых последовательностей оказались «отъеденные» концы.

Выбор продолжения. В этой точке у исследователя есть две возможности продолжения. Можно продолжить анализ данных на сервере. Наиболее интересный путь обозначен как *Б* на рис. 2.6. Он позволяет построить дистантное дерево из найденных последовательностей. Вообще это – далеко не самый точный метод филогенетического анализа, и результат не подходит для публикации, даже если он вам нравится. Но в качестве предварительного этот метод вполне адекватен и главное – быстр.

На рис. 2.6, *А* представлено «черновое» дерево, которое было получено из найденных нами ранее последовательностей методом минимальной эволюции, в котором все нуклеотидные замены считаются равноценными (т. е. транзиции и трансверсии – равновероятными. Ниже мы убедимся, что это не так). В качестве имен отдельных последовательностей были использованы строки, которые сервер Генбанка присваивает своим записям при переводе их в «краткий» формат FASTA (см. ниже). Очевидно, что они содержат повторяющуюся и ненужную информацию. Рисунок 2.6, *Б* иллюстрирует приблизительно то же самое дерево, построенное несколько более точным методом из тех же последовательностей несколько более точным дистантным методом – методом объединения ближайших соседей (*neighbor joining*, обычно сокращается как NJ). Топология деревьев практически не различается, однако соотношение длин ветвей, особенно внутренних – другое.

Выбор метода построения делается в верхнем левом окошке экрана, показанного на рис. 2.6, *А* и *Б*.

Другое важное отличие между рис. 2.6, *А* и 2.6, *Б* состоит в том, что во втором случае в качестве имен OTE (или, если угодно, последовательностей) выведены только наименования вида и рода. В результате читать такое дерево гораздо легче. Выбор информации, которую желательно видеть на дереве, делается в третьем окошке верхнего ряда. В данном случае вместо Show all (показать все) осталось только Taxonomic name.

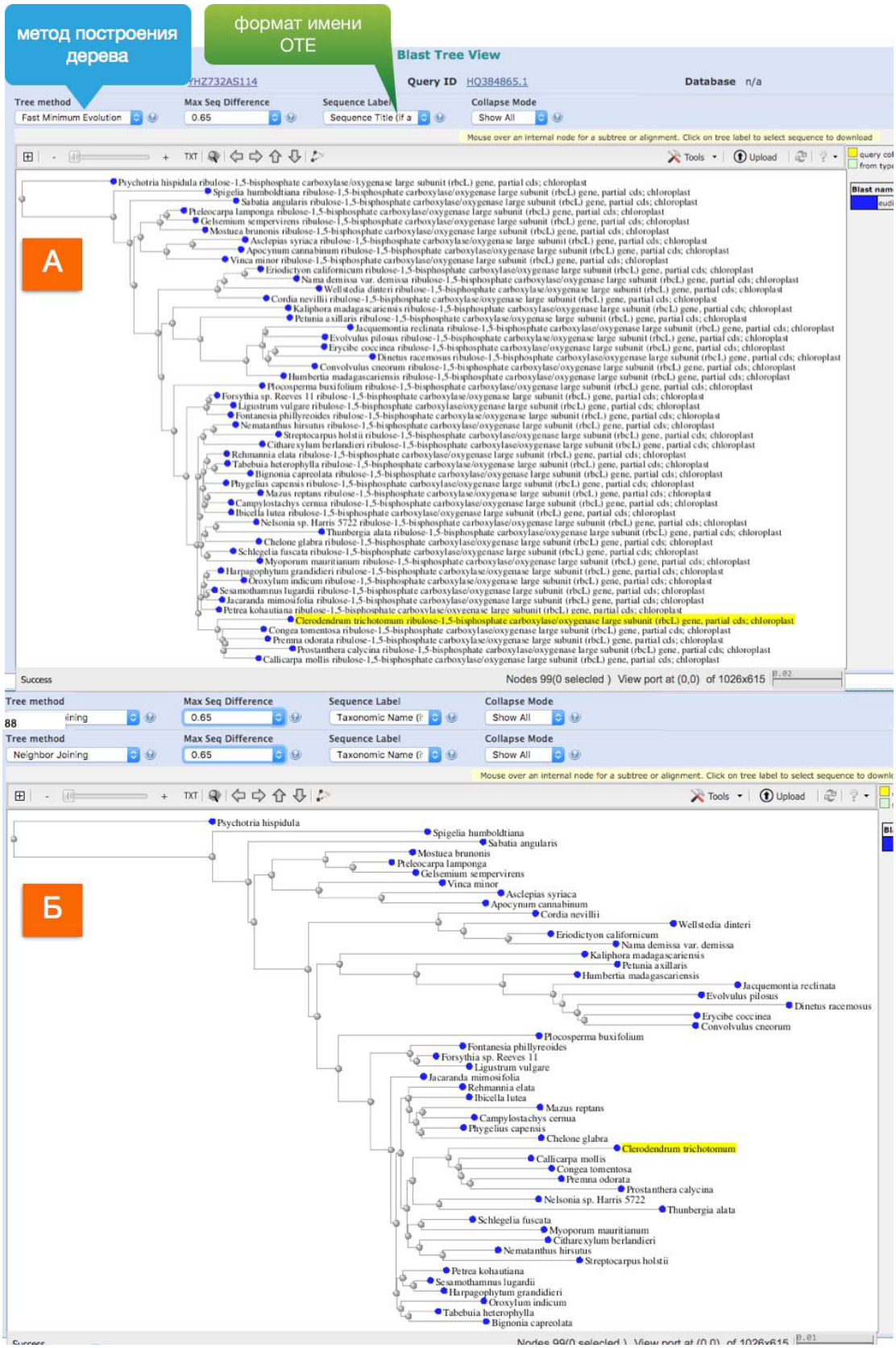


Рис. 2.6. Результаты построения дистантных деревьев. *А* – дерево построено методом минимальной эволюции, *Б* – дерево построено методом объединения ближайших соседей, ветви помечены только с помощью названий видов

Остальные методы работы с последовательностями, без выхода из диалогового окна Генбанка, не пригодны для работы с наборами записей PopSet и будут рассмотрены отдельно.

2.3. Экспорт последовательностей на локальный компьютер

Определить формат, в котором будут загружены данные с сервера Генбанка на локальный компьютер, можно как из окна результатов поиска записей (см. рис. 2.3), так и из окон следующих этапов анализа (например, BLAST – поиска сходных последовательностей или из окна филогенетического анализа). Делать это следует в два этапа:

1. На первом шаге требуется определить формат экспорта. Интерес представляют только две из имеющихся возможностей – это либо формат FASTA, либо внутренний формат, используемый для представления данных в Генбанк – gb. Различия между этими форматами обсуждаются ниже в специальном разделе. При загрузке относительно большого количества последовательностей практически всегда следует выбирать FASTA.

2. Справа от пункта выбора формата есть и следующий выбор, который надлежит сделать. Необходимо указать, куда направить информацию – на экран или в файл. В первом случае файл данных отобразится в окне браузера, во втором – окажется в папке, куда попадают у вас все загружаемые из интернета файлы, например «Загрузки» или Downloads.

Важно отметить, что есть много разных позиций при работе с порталом Генбанка, откуда можно загрузить последовательности, и выше описан только один из возможных способов.

Обзор результатов

В заключение следует отметить, что портал Генбанка дает возможность оценить результаты поиска, выравнивания и первичного филогенетического анализа «с высоты птичьего полета». На рис. 2.7 приведена очень удобная форма текстового представления выровненных последовательностей. Эта форма очень широко распространена. Точки обозначают основания, совпадающие с теми, которые приведены в первой строке. От-

личающиеся обозначены соответствующими символами. Обратите внимание, что не все ряды точек начинаются с одинаковой позиции. Так получается в результате выравнивания, которое по сути максимизирует количество точек, передвигая фрагменты последовательностей друг относительно друга. Это представление очень удобно и полезно для довольно коротких, не длиннее нескольких сотен нуклеотидов длиной, последовательностей.

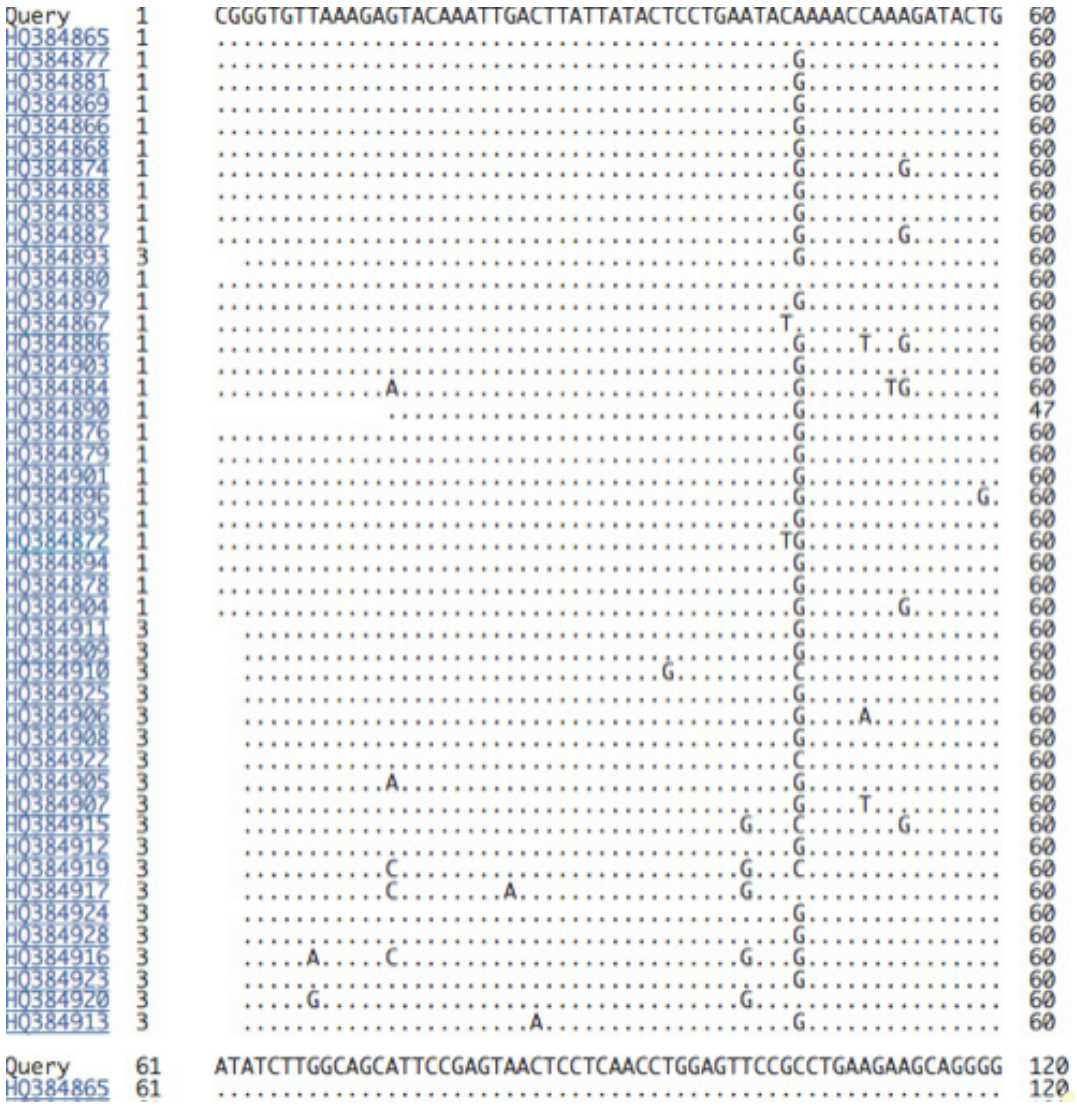


Рис. 2.7. «Точечное» представление результатов выравнивания нуклеотидных последовательностей

Для более длинных лучше подходит графическое представление последовательностей (также на странице результатов выравнивания поиска BLAST), изображенное на рис. 2.8. Оно помогает получить качественное представление как о найденных последовательностях, так и о неравномерности распреде-

ления молекулярного разнообразия по их длине. Для того чтобы получить эту картинку, на экране результатов выравнивания надо выбрать все последовательности (Select: All) в разделе Descriptions окна результатов выравнивания, после чего кликнуть по расположенной справа от (Select: All) ссылке Graphics. Слайдер позволяет изменять масштаб изображения.

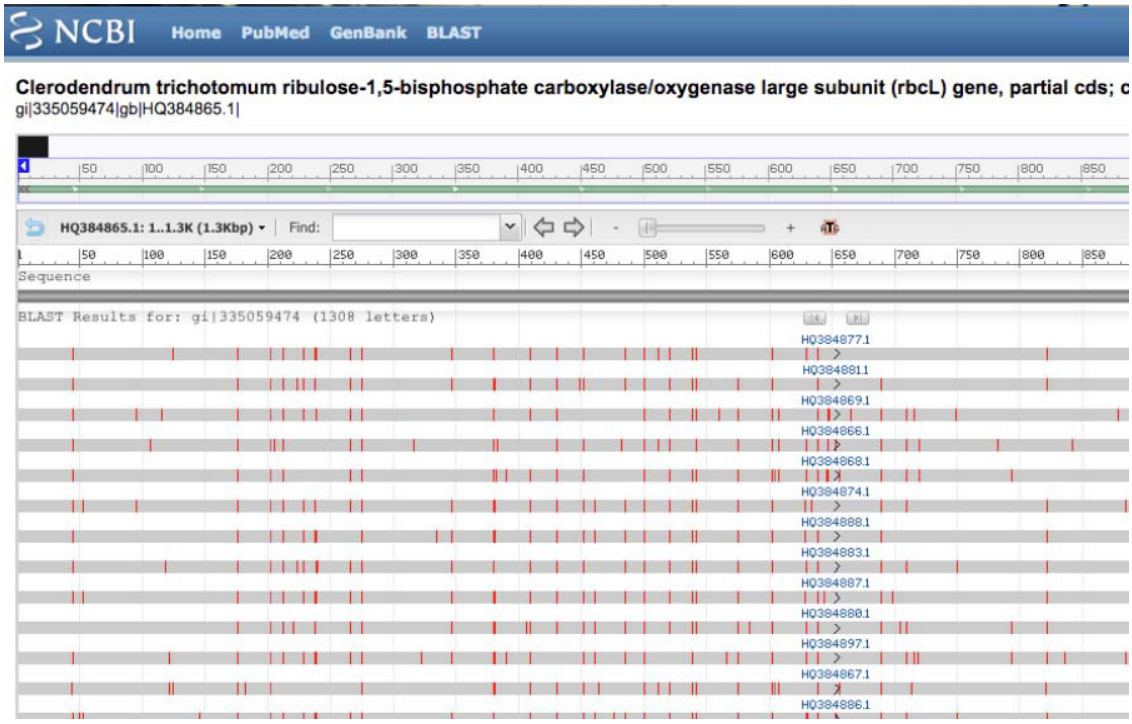


Рис. 2.8. Компактное представление результатов выравнивания нуклеотидных последовательностей

Контрольные вопросы

1. Дайте общую характеристику базы данных NCBI.
2. Перечислите разновидности данных в базе NCBI.

3. ФОРМАТЫ ДАННЫХ

Наборы последовательностей, которые являются результатом поиска в Генбанке, можно загрузить на локальный компьютер в двух форматах – FASTA и собственном формате gb (GenBank). Оба имеют свои преимущества и недостатки и служат для разных целей. Скорее всего (в зависимости от опций, выбранных при загрузке) набор последовательностей потеряет выравнивание. Даже если этого и не произойдет, следует помнить, что быстрое выравнивание последовательностей на сервере в общем случае следует рассматривать как предварительное и начать работу с последовательностями следует с более точного выравнивания, как описано в специальном разделе ниже.

Файлы с последовательностями не должны содержать непечатных символов (тех, которых не видно в текстовом окне, например TAB), а также букв кириллицы, греческих и пр. ТОЛЬКО ASCII! Названия файлов не должны содержать пробелов и знаков нелатинских алфавитов (даже если на вид они похожи вроде C и C), а также они должны быть короткими. Придется их печатать много раз, и опечатки будут сильно отвлекать.

3.1. Формат FASTA

Последовательности в формате FASTA начинаются с однострочного описания, за которым следуют строки с данными последовательности. Описание отмечается при помощи символа > (больше) в первом столбце. Текст после этого знака до конца строки за ним является идентификатором последовательности, далее, через пробел, следует описание, которое что-то говорит о том, что же это за последовательность. Описания может и не быть. Обычно строки нуклеотидных или аминокислотных последовательностей в формате FASTA ограничены длиной в 80 символов. Данные последовательности располагаются до следующего описания.

Несмотря на то что в официальной документации по формату FASTA это не упомянуто, суммарная длина заголовка обычно не должна превышать 35 знаков. Дело в том, что для наборов последовательностей в этом формате пишут программы многие исследователи, которым не очень хочется читать и тем более исполнять придуманные другими правила. В случае превышения этого ограничения пользователи часто сталкиваются с ошибками вроде «имена некоторых последовательностей дублируют друг друга» или программа просто перестает работать, ни на что не «жалуясь». Пример одной последовательности в формате FASTA:

```
>gi|335059490|gb|HQ384877.1|:1-1306 Petrea kohautiana ribulose-1,5-
bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial
cds; chloroplast
CGGGTGTTAAAGAGTACAAATTGACTTATTATACTCCTGAATACGAAACCAAAGATACTGATATCTTGGC
AGCATTCCGAGTAACTCCTCAACCTGGAGTTCCGCCTGAAGAAGCAGGGGCCGCAGTAGCTGCCGAATCT
TCTACTGGTACATGGACAACTGTGTGGACCGATGGACTTACCAGCCTTGATCGTTACAAAGGGCGATGCT
ACCACATCGAGCCCGTTCCTGGAGAAGCAGATCAATATATCTGTTATGTAGCTTACCCTTTAGACCTTTT
TGAAGAAGGTTCTGTTACTAACATGTTTACTTCCATTGTAGGAAATGTATTTGGATTCAAAGCCCTGCGT
GCTCTACGTCTGGAAGATCTGCGAATCCCTACTGCTTATATTAAAACTTTCCAAGGCCCGCCTCATGGGA
TCCAAGTTGAGAGAGATAAATTGAACAAGTATGGTCGTCCTCTGTTGGGATGTACTATTAAACCTAAATT
GGGGTTATCTGCTAAAAACTACGGTAGAGCAGTTTATGAATGTCTTCGCGGTGGACTTGATTTTACCAAA
GATGATGAGAACGTAAACTCCCAGCCATTTATGCGTTGGAGAGATCGTTTCTTATTTTGTGCCGAAGCAC
TTTATAAAGCACAGGCTGAAACAGGTGAAATCAAAGGGCATTACTTGAATGCTACTGCAGGTACATGCGA
AGAAATGATCAAAAGAGCTGTATTTGCTAGAGAATTGGGAGTTCCTATCGTAATGCATGACTACTTAACA
GGAGGATTCACTGCAAATACTAGCTTGGCTCATTATTGCCGAGATAATGGCCTACTTCTTCACATTCACC
GTGCAATGCATGCAGTTATTGATAGACAGAAGAATCATGGTATGCACTTCCGTGTACTAGCTAAAGCGTT
ACGTATGTCTGGTGGAGATCATATTCACGCTGGTACCGTAGTAGGTAAACTTGAAGGAGAAAGAGACATC
ACTTTGGGCTTTGTTGATTTACTGCGTGATGATTTTATTGAAAAAGATCGAAGTCGCGGTATTTATTTCA
CTCAAGATTGGGTCTCTCTACCAGGTGTTATTCCCGTGGCTTCAGGGGGTATTCACGTTTGGCATATGCC
TGCTCTGACCGAGATCTTTGGGGATGATGCCGTACTACAGTTCGGTGGAGGAACTTTAGGACACCCTTGG
GGTAATGCGCCAGGTGCCGTAGCTAACCGAGTAGCTCTAGAAGCATGTGTAAAAGCTCGTAATGAAGGAC
GTGATCTTGCTGCTGAGGGTAATACAATTATCCGTGAGGCTAGCAA
```

Очевидно, что сам GenBank не обращает внимания на ограничение в 35 знаков никакого внимания, пытаясь сообщить в строке заголовка максимум информации. Эта информация часто оказывается очень избыточной, более того, она просто мешает на последующих этапах. Например, в примере о петуниях, каждая строка заголовка содержит текст «ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast». От него желательно избавиться. Часто полезно удалить индикаторы доступа к записи и т. п. Для простых случаев это можно сделать с помощью текстового редактора, однако для очень большого количества длинных последо-

вательностей лучше пользоваться простыми скриптами (наборами автоматически следующих друг за другом команд на интерпретируемом языке программирования).

Расширения (идентификаторы) названий файлов, содержащих последовательности в формате FASTA, не стандартизированы, хотя попытки этого и предпринимали разные международные организации. Наиболее распространенные: .FAS, .FA и .FST.

3.2. Внутренний формат записей GenBank (gb)

Каждая запись в Генбанке хранится в виде простого файла, содержащего всю информацию о последовательности, которую сообщили авторы. Файл состоит из двух колонок, на левую отведено 12 символов. Она содержит название полей (разделов) записи. Подробное описание значений этих полей дано на специальной странице, и приводить его здесь не имеет смысла, тем более, что NCBI иногда меняет этот формат. Текущий стандарт расположен по адресу: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html. Пример из этого документа приведен ниже.

```
LOCUS       SCU49845    5028 bp    DNA       PLN    21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds,
           and Axl2p(AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION  U49845
VERSION    U49845.1 GI:1293613
KEYWORDS   .
SOURCE     Saccharomyces cerevisiae (baker's yeast)
ORGANISM   Saccharomyces cerevisiae
           Eukaryota; Fungi; Ascomycota; Saccharomycotina;
           Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE  1 (bases 1 to 5028)
AUTHORS    Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE      Cloning and sequence of REV7, a gene whose function is
           required for DNA damage-induced mutagenesis in Saccharo-
           myces cerevisiae
JOURNAL    Yeast 10 (11), 1503-1509 (1994)
PUBMED     7871890
REFERENCE  2 (bases 1 to 5028)
  AUTHORS  Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE    Selection of axial growth sites in yeast requires Axl2p,
           a novel plasma membrane glycoprotein
  JOURNAL  Genes Dev. 10 (7), 777-793 (1996)
  PUBMED   8846915
REFERENCE  3 (bases 1 to 5028)
```

```
  AUTHORS  Roemer,T.
  TITLE    Direct Submission
  JOURNAL  Submitted (22-FEB-1996) Terry Roemer, Biology, Yale
           University, New Haven, CT, USA
FEATURES         Location/Qualifiers
  source         1..5028
                 /organism="Saccharomyces cerevisiae"
                 /db_xref="taxon:4932"
                 /chromosome="IX"
                 /map="9"
   CDS           <1..206
                 /codon_start=3
                 /product="TCP1-beta"
                 /protein_id="AAA98665.1"
                 /db_xref="GI:1293614"
/translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
AEVLLRVDNIIRARPRTANRQHM"
   gene          687..3158
                 /gene="AXL2"
   CDS           687..3158
                 /gene="AXL2"
                 /note="plasma membrane glycoprotein"
                 /codon_start=1
                 /function="required for axial budding pattern of S.
                 cerevisiae"
                 /product="Axl2p"
                 /protein_id="AAA98666.1"
                 /db_xref="GI:1293615"
/translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
VDFSNKSNVNVGQVKDIHGRIPEML"
   gene          complement(3300..4037)
                 /gene="REV7"
   CDS           complement(3300..4037)
                 /gene="REV7"
                 /codon_start=1
                 /product="Rev7p"
                 /protein_id="AAA98667.1"
                 /db_xref="GI:1293616"
/translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
```

```
LISGDDKILNGVYSQYEEGESIFGSLF”
ORIGIN
1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaa-
gaccaa
181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taa-
taaaccg
241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta ta-
taattcaa
301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg
tcacaaataa
361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca
agaatgtaat
421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tcctt-
gccga
481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat
tttcttattc
541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaa-
gaacaga
   . . .
1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca
gacattgaag
1381 gattttctgc cgttgaggta gaattcgaat ta
. . .
//
```

Для наших целей достаточно отметить следующие свойства этого формата.

Содержит уникальный буквенно-цифровой идентификатор записи. Многие программы и в локальном, и в удаленном режиме могут его использовать для того, чтобы получить всю запись; эти идентификаторы можно использовать на промежуточных этапах анализа для того, чтобы избежать длинных и содержащих пробелы, точки, скобки и прочие подписи. Часто при анализе больших объемов данных это оказывается критично для успеха.

Есть краткое описание последовательности, и указаны некоторые ее свойства, в частности длина. В случае, если последовательность содержит белок кодирующие фрагменты, правильные аминокислотные последовательности также содержатся в записи.

Важнейшая особенность – запись Генбанка по идее содержит ссылку на статью (если она, конечно, существует), где данная последовательность была впервые описана. Поскольку депонирование последовательностей в Генбанке является обяза-

тельным условием редакций всех ведущих профессиональных журналов, часто оказывается, что последовательность в банке есть, а статью так и не приняли к печати. Тогда она все равно рано или поздно появляется в открытом доступе и на нее можно с разрешения авторов ссылаться как на «личное сообщение».

Запись содержит полное родовое и видовое название организма, из которого была получена ДНК. Это название однозначно соответствует записи в сопутствующем разделе Генбанка Taxonomy (таксономия).

Пример таксономической записи для вида петуний из примера FASTA-формата:

```
     Taxonomy ID: 222881
     Inherited blast name: eudicots
     Rank: species
     Genetic code: Translation table 1 (Standard)
     Mitochondrial genetic code: Translation table 1 (Standard)
     Other names:
     authority: Sesamothamnus lugardii N.E.Br. ex Stapf
     Lineage( full )
     cellular organisms; Eukaryota; Viridiplantae; Streptophyta;
Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Sperma-
tophyta; Magnoliophyta; Mesangiospermae; eudicotyledons; Gunner-
idae; Pentapetalae; asterids; lamiids; Lamiales; Pedaliaceae;
Sesamothamnus
```

Таким образом, приведена вся иерархия, начиная с многоклеточных организмов. Необходимо отметить, что идентификация вида целиком лежит на совести авторов, и поэтому именно эта часть базы данных содержит много ошибок.

В заключение следует отметить, что поскольку генбанковский формат – самый богатый информацией, то из него можно получить файлы всех других форматов, но наоборот – вряд ли. Не хватит сведений о последовательностях. Поэтому программы-браузеры наборов последовательностей могут открывать файлы .gb, редактировать их и сохранять в других форматах, но транслировать данные из других форматов в .gb они не могут.

3.3. Форматы NEXUS и PHYLIP

NEXUS

Как и другие форматы, это простой текстовый файл, содержащий «команды» для программы, которая его обрабатывает. В этом смысле NEXUS напоминает скорее не формат файла данных, а программу на интерпретируемом языке (т. е. на языке, команды которого выполняются строчка за строчкой, по современной терминологии – скрипт).

Правила синтаксиса. Первая строчка любого файла в этом формате должна начинаться со слова #nexus. Именно это проверяют все программы, работающие с этим форматом. Квадратные скобки ограничивают комментарии. На эту часть файла программа не обращает внимания. Файл разбит на блоки, каждый из которых начинается командой BEGIN block_name;[1] и заканчивается командой END;. Обратите внимание на точку с запятой. Основная ошибка при приготовлении nexus – забытые знаки препинания. Пример простого NEXUS-файла:

```
#NEXUS
Begin data;
Dimensions ntax=4 nchar=15;
Format datatype=dna missing=? gap=-;
Matrix
Species1  atgctagctagctcg
Species2  atgcta??tag-tag
Species3  atgttagctag-tgg
Species4  atgttagctag-tag
;
End;
```

Основные блоки:

1. Блок TAXA

Содержит информацию о таксонах. Другими словами, содержит список наименований последовательностей.

2. Блок DATA

DATA содержит матрицу данных (например, стопку выровненных последовательностей).

---

[1] block_name может быть название любого *разрешенного данной программой* блока, например data. За названием блока должна обязательно быть точка с запятой.

3. Блок TREES

TREES содержит филогенетическое дерево, в формате, напоминающем ньюикский (см. ниже подробное описание): ((A,B),C)[2]; например:

```
\#NEXUS
BEGIN TAXA;
TAXLABELS A B C;
END;
BEGIN TREES;
TREE tree1 = ((A,B),C);
END;
```

Основное отличие этого формата от других состоит в том, что он предназначен для того, чтобы комбинировать в одном файле различные данные, относящиеся к одним и тем же объектам. Это могут быть молекулярные данные, как нуклеотидные, так и аминокислотные последовательности, данные о морфологических признаках, даже координаты сбора проб. В последней комбинации данных используются для филогенетического анализа по total evidence (всем сведениям), что приводит к получению более обоснованных выводов и о взаимоотношении современных видов, и об их эволюционной истории.

Формат PHYLIP

Этот формат был введен Джо Фельзенштейном, автором пакета программ для филогенетического анализа PHYLIP. Этот формат затем подвергался некоторым модификациям, которые, однако, совместимы с исходным вариантом, который мы здесь и приводим.

Первая строка как минимум содержит два числа: первое – число последовательностей, второе – их длина. Имя последовательности занимает не более 10 начальных символов строки. Далее идет последовательность. Здесь возможны два варианта, и не все программы умеют обходиться с ними правильно и автоматически. Приведем оба:

---

[2] В отличие от ньюикского формата, в nexus-файлах деревья всегда имеют имена. В данном случае это tree1.

```
10 705
Cow    ATGGCATATCCCATACAACTAGGATTCCAAGATGCAACATCACCAATCATAGAAGAACTA
Carp   TGGCACACCCAACGCAACTAGGTTTCAAGGACGCGGCCATACCCGTTATAGAGGAACTT
Chicken ATGGCCAACCACTCCCAACTAGGCTTTCAAGACGCCTCATCCCCCATCATAGAAGAGCTC
Human   ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTACTTCCCCTATCATAGAAGAGCTT
Loach   ATGGCACATCCCACACAATTAGGATTCCAAGACGCGGCCTCACCCGTAATAGAAGAACTT
Mouse   ATGGCCTACCCATTCCAACTTGGTCTACAAGACGCCACATCCCCTATTATAGAAGAGCTA
Rat     ATGGCTTACCCATTTCAACTTGGCTTACAAGACGCTACATCACCTATCATAGAAGAACTT
Seal    ATGGCATACCCCCTACAAATAGGCCTACAAGATGCAACCTCTCCCATTATAGAGGAGTTA
Whale   ATGGCATATCCATTCCAACTAGGTTTCCAAGATGCAGCATCACCCATCATAGAAGAGCTC
Frog    ATGGCACACCCATCACAATTAGGTTTTCAAGACGCAGCCTCTCCAATTATAGAAGAATTA

     CTTCACTTTCATGACCACACGCTAATAATTGTCTTCTTAATTAGCTCATTAGTACTTTAC
     CTTCACTTCCACGACCACGCATTAATAATTGTGCTCCTAATTAGCACTTTAGTTTTATAT
     GTTGAATTCCACGACCACGCCCTGATAGTCGCACTAGCAATTTGCAGCTTAGTACTCTAC
     ATCACCTTTCATGATCACGCCCTCATAATCATTTTCCTTATCTGCTTCCTAGTCCTGTAT
     CTTCACTTCCATGACCATGCCCTAATAATTGTATTTTTGATTAGCGCCCTAGTACTTTAT
     ATAAATTTCCATGATCACACACTAATAATTGTTTTCCTAATTAGCTCCTTAGTCCTCTAT
     ACAAACTTTCATGACCACACCCTAATAATTGTATTCCTCATCAGCTCCCTAGTACTTTAT
     CTACACTTCCATGACCACACATTAATAATTGTGTTCCTAATTAGCTCATTAGTACTCTAC
     CTACACTTTCACGATCATACACTAATAATCGTTTTTCTAATTAGCTCTTTAGTTCTCTAC
     CTTCACTTCCACGACCATACCCTCATAGCCGTTTTTCTTATTAGTACGCTAGTTCTTTAC
```

Interleaved-формат позволяет «свернуть» строки, как показано выше. Более простой формой того же формата является sequential, когда вся последовательность без переносов располагается за идентификатором:

```
10 60
Cow     ATGGCATATCCCATACAACTAGGATTCCAAGATGCAACATCACCAATCATAGAAGAACTA
Carp    ATGGCACACCCAACGCAACTAGGTTTCAAGGACGCGGCCATACCCGTTATAGAGGAACTT
Chicken ATGGCCAACCACTCCCAACTAGGCTTTCAAGACGCCTCATCCCCCATCATAGAAGAGCTC
Human   ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTACTTCCCCTATCATAGAAGAGCTT
Loach   ATGGCACATCCCACACAATTAGGATTCCAAGACGCGGCCTCACCCGTAATAGAAGAACTT
Mouse   ATGGCCTACCCATTCCAACTTGGTCTACAAGACGCCACATCCCCTATTATAGAAGAGCTA
Rat     ATGGCTTACCCATTTCAACTTGGCTTACAAGACGCTACATCACCTATCATAGAAGAACTT
Seal    ATGGCATACCCCCTACAAATAGGCCTACAAGATGCAACCTCTCCCATTATAGAGGAGTTA
Whale   ATGGCATATCCATTCCAACTAGGTTTCCAAGATGCAGCATCACCCATCATAGAAGAGCTC
Frog    ATGGCACACCCATCACAATTAGGTTTTCAAGACGCAGCCTCTCCAATTATAGAAGAATTA
```

Обратите внимание на то, что во втором случае длина последовательности – всего 60 пар нуклеотидов вместо 705 в первом. Главным препятствием при использовании последовательной записи (sequential) является то, что у многих текстовых редакторов есть ограничение на длину строки, что не позволяет редактировать последовательности длиной в десятки тысяч символов.

3.4. Редактирование файлов с последовательностями

Только что загруженные из интернета файлы нуждаются в некоторой редакции перед тем, как с ними можно будет что-то сделать. Для относительно небольшого (несколько десятков) количества последовательностей это можно сделать с помощью текстовых редакторов ***(не путать с текстовыми процессорами! Word и иже с ним ни в коем случае использовать нельзя!).***

Редакторы, пригодные для работы с файлами последовательностей, должны отвечать следующим требованиям:

- не вносить невидимых знаков форматирования текста;
- поддерживать возможность работы с очень длинными строками;
- искать и убирать невидимые знаки;
- менять кодировку конца строки (в разных операционных системах она разная, а файл данных придется пересылать между компьютерами с разными ОС);
- не заниматься самодеятельностью вроде автоматической замены серии пробелов одним знаком табулятора.

Желательные возможности редакторов:

- способность выделять и работать с вертикальными блоками текста;
- иметь простой макроязык для автоматизации редактирования;
- уметь самостоятельно загружать, редактировать и сохранять файлы на удаленных компьютерах с помощью защищенного соединения.

Типичная сессия редактирования файла последовательностей представлена на рис. 3.1 (в данном случае – с помощью редактора Editra, http://editra.org).

Очевидно, что следует отредактировать названия последовательностей и проследить за тем, чтобы были «правильные» концы. Для дальнейшей работы также желательно приготовить файл – словарь, содержащий в первой колонке номера доступа (Accession numbers) последовательностей, а во второй – таксономические имена соответствующих организмов.

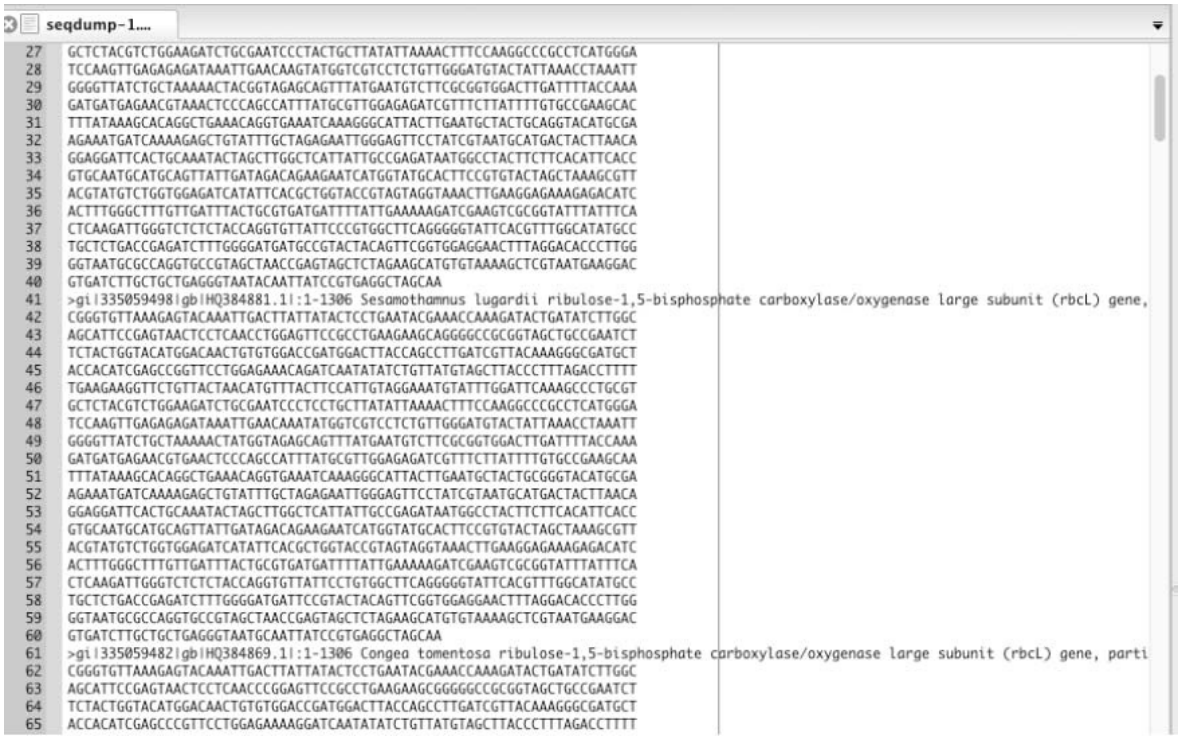


Рис. 3.1. Редактирование набора последовательностей гена rbcL петуний, загруженного с сервера NCBI

В нашем примере – 50 последовательностей. Поэтому все необходимые простейшие операции можно за обозримое время и при определенной аккуратности проделать с помощью текстового редактора. Но с ростом объемов данных необходимость автоматизации любых, даже самых простых действий становится все острее. К счастью, в наше время часто уже не требуется писать специальные программы для любого шага. Можно обойтись короткими наборами инструкций на специальных интерпретируемых языках.

Для «переработки» текстовых файлов очень хорошо приспособлен язык Perl, который доступен для всех операционных систем, включая даже Android и iOS. Скачать и установить его можно с сайта http://perl.org. В интернете можно найти много бесплатных и хороших руководств, а также активных и доброжелательных групп, которые помогут новичку. Следует отметить, что этот язык обязательно входит в минимальный набор, который устанавливается вместе с операционной системой, и пользователи Linux или Mac OS X имеют его в своем распоряжении по умолчанию.

Рассмотрим короткий скрипт, который можно набрать в любом из текстовых редакторов и сохранить, например, под именем namez.pl в той же директории (папке), что и только что скачанный файл с последовательностями. Для дальнейшей работы следует его сделать исполняемым:

```
chmod +x namez.pl
```

После выполнения этой команды компьютер сможет отличить файл скрипта от обычного текстового и станет относиться к нему как к программе.

```
#!/usr/bin/perl
while(<>){
    if($_ =~ /^>/){
        @line = split;
        print «>$line[1]_$line[2]\n";
    }
    else {
        print "$_";
    }
}
```

Для того чтобы воспользоваться этим скриптом, следует дать команду

```
./namez.pl <petunia.fas >p.fas
```

«./» обозначает, что исполняемый файл находится в той же директории, что и файл данных, которая открыта в настоящее время в окне терминала. Затем следует наименование скрипта. Значок < направляет файл, наименование которого следует за ним, для обработки скриптом. Значок > указывает, куда направить информацию, распечатываемую скриптом. В данном случае это файл p.fas. После выполнения этой команды **все** заголовки в исходном файле изменятся так же, как и название первой последовательности:

в исходном файле petunia.fas:

```
>gi|335059474|gb|HQ384865.1|:1-1308 Clerodendrum trichotomum
ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit
(rbcL)...
CGGGTGTTAAAGAGTACAAATTGACTTATTATACTCCTGAATACAAAACCAAAGA-
TACTGATATCTTGGC
. . .
```

в обработанном файле p.fas это превращается в:

```
>Clerodendrum_trichotomum
CGGGTGTTAAAGAGTACAAATTGACTTATTATACTCCTGAATACAAAACCAAAGA-
TACTGATATCTTGGC
. . .
```

Эта форма гораздо компактнее и лучше читается. Знак «_» добавлен между родовым и видовым наименованиями для того, чтобы избежать пробелов, что требуется для корректной работы некоторых программ.

Полезно также иметь словарь соответствия номеров доступа и таксономических имен. Для приготовления такого файла модифицируем скрипт namez.pl и сохраним его с именем table.pl.

```
1  #!/usr/bin/perl
2  while(<>){
3      if($_=~'>'){
4          @line = split;
5          @ID = split /\|/,$line[0];
6          print "$ID[3]\t$line[1]_$line[2]\n";
7      }
8  }
9
```

Его следует сделать исполняемым, как описано выше, а затем:

```
./table.pl <petunia.fas >dict.txt
```

Результат деятельности этого коротенького скрипта сильно отличается от входного файла:

```
HQ384865.1 Clerodendrum_trichotomum
HQ384877.1 Petrea_kohautiana
HQ384881.1 Sesamothamnus_lugardii
HQ384869.1 Congea_tomentosa
HQ384866.1 Premna_odorata
HQ384868.1 Callicarpa_mollis
HQ384874.1 Rehmannia_elata
HQ384888.1 Jacaranda_mimosifolia
HQ384883.1 Harpagophytum_grandidieri
HQ384887.1 Tabebuia_heterophylla
HQ384893.1 Phygelius_capensis
HQ384880.1 Schlegelia_fuscata
HQ384897.1 Nematanthus_hirsutus
HQ384867.1 Prostanthera_calycina
HQ384886.1 Oroxylum_indicum
HQ384903.1 Forsythia_sp.
HQ384884.1 Bignonia_capreolata
HQ384890.1 Ibicella_lutea
. . .
```

То есть мы получили искомую таблицу, которую сможем в дальнейшем использовать в других скриптах.

3.5. Простейшая анатомия скрипта на Perl

В рамках настоящего пособия не стоит задачи обучить читателя программированию на языке Perl или каком-либо другом языке программирования, который нам потребуется в дальнейшем. Вместе с тем необходимо хотя бы приблизительно понимать, что означают команды, тем более, что очевидность не является сильной стороной Perl'а.

Рассмотрим построчно скрипт table.pl:

первая строчка

```
#!/usr/bin/perl
```

означает, что в качестве интерпретатора последующих команд будет работать программа Perl, и указывает, где она расположена. Расположение Perl может изменяться от одной операционной системы к другой, и часто в скриптах, скачанных из интернета, приходится её редактировать. Для того чтобы найти свой вариант, следует дать команду:

```
bash-3.2$which perl
```

и увидеть, что на данном компьютере адрес оказывается:

```
/opt/local/bin/perl
```

и соответственно надо изменить первую строчку.

Вторая строка открывает цикл while, ограниченный фигурными скобками:

```
while(<>){
```

Это означает, что команды цикла будут выполняться, пока верно условие, которое является аргументом (т. е. то, которое находится в простых скобках) функции while. В нашем случае использована специфическая конструкция «ромб» (<>), означающая: «пока не закончится поступающий сюда *построчно* входной файл, который скрипту надлежит читать строка за строкой». Когда файл данных закончится, это условие перестанет выполняться.

Далее по правилам языка считываемая строка становится значением специализированной переменной $_ (все символы переменных в Perl начинаются с символа доллара $), и скрипт проверяет истинность предположения о том, что первый знак этой строки равен '>', т. е. что эта строка – имя последовательности в соответствии с правилами формата FASTA:

```
If($_=~'>'){
```

если это утверждение верно, то открывается следующий блок команд, выделенный фигурной скобкой.

Первая из этих команд

```
@line=split;
```

заводит массив @line (его имя начинается с @), расщепляет строку, хранящуюся в переменной $_ по пробелам, и заключенные между пробелами фрагменты становятся элементами массива @line, нумерация которых начинается с нуля.

Ниже название первой последовательности из файла petunia.fas по-разному окрашено в соответствии с блоками, получившимися после его расщепления по пробелам.

```
>gi|335059474|gb|HQ384865.1|:1-1308 Clerodendrum trichotomum ribulose-1,5-bisphos-
phate carboxylase/oxygenase large subunit (rbcL) gene, partial cds; chloroplast
```

Видно, что родовое название стало элементом № 1, а видовое – элементом № 2. Номер доступа содержится в окрашенном белым элементе № 0. Поэтому следующей командой

```
@ID = split /\|/,\$line[0];
```

скрипт берет нулевой элемент из @line, $line[0] (символ @ при этом меняется на $ потому, что это – просто переменная, а не массив, состоящий из единственного элемента), и расщепляет его по знаку «|» (вертикальная черточка). Результат записывается в массив @ID, который, если его раскрасить подобно предыдущему случаю (символы, по которым происходит разрезание, в состав элементов не входят!), превращается в:

```
>gi|335059474|gb|HQ384865.1|:1-1308
```

Нужный нам белый фрагмент – номер доступа, оказывается третьим в массиве, если считать с нуля. Разбор строки – имени последовательности закончен, и командой

```
print "\$ID[3]\t$line[1]\t$line[2]\n";
```

скрипт составляет строку выходного файла и распечатывает её.

В случае первой последовательности файла petunia.fas, на примере которой мы рассматривали работу скрипта, результат будет

```
HQ384865.1 Clerodendrum_trichotomum
```

Такие же действия скрипт будет проводить со всеми заголовками, пока не закончится входной файл. Если же строка не является именем последовательности, этот скрипт с ней не будет делать ничего. В скрипте закрывающиеся фигурные скобки завершают соответствующие блоки команд, завершается и работа скрипта.

Написание такого рода небольших скриптов – сначала весьма трудоемкое занятие. Которое, тем не менее, быстро начинает окупаться за счет того, что раз написанный и отлаженный скрипт применим много раз, и со временем у каждого исследователя формируется собственная библиотека небольших, но сильно ускоряющих работу инструментов. Единственное условие, которое следует соблюдать при создании библиотеки скриптов, – их нужно тщательно комментировать прямо в файле с кодом. Иначе легко просто забыть, о чем он, и начать изобретать велосипед. Как и набор «физических» инструментов, собственную библиотеку скриптов следует поддерживать в порядке.

Контрольные вопросы

1. Дайте основные характеристики FASTA-формата.
2. Приведите примеры различных типов данных, используемых при анализе биоразнообразия.
3. Охарактеризуйте форматы данных NEXUS и GenBank, опишите их принципиальные отличия от других форматов.

4. ТОЧНОЕ МНОЖЕСТВЕННОЕ ВЫРАВНИВАНИЕ

Выравнивание последовательностей – этап любого анализа наборов последовательностей аминокислот, который тем более важен, чем больше эволюционное расстояние между организмами, из которых эти последовательности происходят. Несмотря на полувековую историю исследований, на сегодняшний день не существует идеального подхода, который бы позволил точно утверждать, что набор последовательностей выровнен оптимальным образом. Критерием оптимальности, или критерием качества, является количество инделей (инсерций или делеций нуклеотидов или аминокислот, внесенных в набор данных в результате выравнивания). Иногда важное значение играет также число непрерывных отрезков – инделей.

Как уже отмечалось выше, довольно часто наборы последовательностей, полученные из интернета, теряют выравнивание. Те же, что его сохраняют, оказываются выровнены довольно приблизительно, что может негативно сказаться на результатах дальнейшего анализа. Поэтому выравнивание является первым шагом в любой работе.

Выравнивание последовательностей – метод, основанный на сдвигании двух или более последовательностей ДНК, РНК или белков друг относительно друга таким образом, чтобы сделать максимальным сходство этих последовательностей. Сходство первичных структур двух молекул может отражать их функциональные, структурные или эволюционные взаимосвязи.

Выровненные последовательности нуклеотидов или аминокислот обычно представляют в виде строк матрицы. В процессе выравнивания в эти последовательности вносятся разрывы между мономерами таким образом, чтобы одинаковые или похожие элементы были расположены в следующих друг за другом столбцах матрицы.

Множественное выравнивание – это выравнивание трёх и более последовательностей. В большинстве случаев построение множественного выравнивания – необходимый этап реконструкции филогенетических деревьев и других анализов. Нахождение оптимального множественного выравнивания методом динамического программирования слишком сложно, поэтому для множественного выравнивания применяют различные эвристические алгоритмы.

Наиболее распространенные программы, осуществляющие множественное выравнивание:

- Clustal (http://www.clustal.org/);
- T-COFFEE (http://www.tcoffee.org);
- MUSCLE (http://www.drive5.com/muscle/);
- mafft (http://mafft.cbrc.jp/alignment/software/).

Следует помнить, что очень во многих случаях не существует единственного выравнивания, которое было бы лучше остальных. Более того, методы сравнения различных вариантов субъективны и сильно зависят от относительного веса, который придают разным мутациям.

Иногда полезным оказывается привлечение дополнительной информации, например, о консервативной пространственной структуре нуклеиновых кислот. В последнем случае говорят о структурном выравнивании. Оно применяется к белкам и рибонуклеиновым кислотам (РНК), для которых известна пространственная структура. Целью является нахождение и сопоставление участков, одинаково уложенных в пространстве. Структурное выравнивание обычно сопровождается наложением структур, т. е. нахождением движений пространства, применение которых к заданным молекулам наилучшим образом совмещает их.

Рассмотрим подробнее, пожалуй, наиболее точный из инструментов множественного выравнивания на время написания настоящего пособия – mafft[3]. Обычно удобнее всего пользо-

[3] MAFFT version 5: improvement in accuracy of multiple sequence alignment / K. Katoh, K. Kuma, H. Toh, T. Miyata // Nucleic Acids Res. 2005. Vol. 33. P. 511–518; MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform / K. Katoh, K. Misawa, K. Kuma, T. Miyata // Nucleic Acids Res. 2002. Vol. 30. P. 3059–3066.

ваться серверами, на которых установлены интерфейсы к mafft. Таких серверов много. Часть из них перечислены ниже. Версии программы на различных серверах могут различаться, соответственно, может различаться и результат выравнивания. Основные серверы, которыми можно воспользоваться, перечислены ниже:

- http://www.ebi.ac.uk/Tools/msa/mafft/;
- http://toolkit.tuebingen.mpg.de/mafft/;
- http://www.genome.jp/tools/mafft/;
- http://myhits.isb-sib.ch/cgi-bin/mafft;
- http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::mafft;
- https://www.ddbj.nig.ac.jp/search/help/wabi/wabi_mafft_help-j.html.

В рамках одной программы в mafft реализовано несколько методов, каждый из которых имеет свой набор опций и применим в различных ситуациях. Различия в результатах можно оценить по рис. 4.1 [4].

Ниже приведены варианты консольных команд, запускающих mafft в его различных инкарнациях.

mafft [options] input [> output]
linsi input [> output]
ginsi input [> output]
einsi input [> output]
fftnsi input [> output]
ftns input [> output]
nwns input [> output]
nwnsi input [> output]
mafft-profile group1 group2 [> output]

Рассмотрим, чем эти методы отличаются друг от друга.

[4] Рисунок 2 из статьи: Kazutaka Katoh, Daron M. Standley. A simple method to control over-alignment in the MAFFT multiple sequence alignment program // Bioinformatics. 2016. Vol. 32, N 13. P. 1933–1942. doi: 10.1093/bioinformatics/btw108

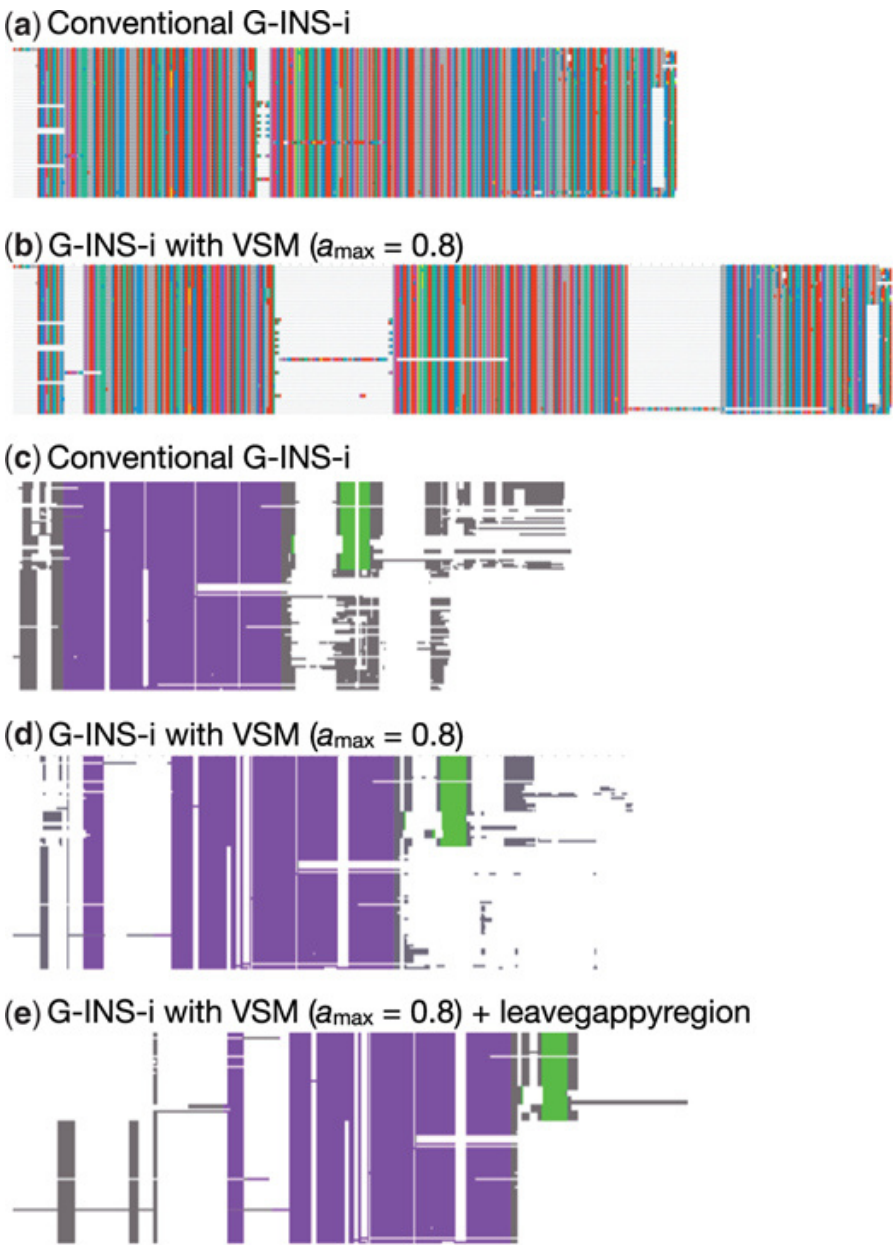


Рис. 4.1. Графическое представление результатов применения различных методов mafft к одному и тому же набору последовательностей

Точные методы:

- L-INS-i

Возможно, самый точный из реализованных в mafft методов. Применим к наборам менее 200 последовательностей. Итеративный метод, опирающийся на локальное выравнивание.

Варианты командной строки:

```
mafft –localpair –maxiterate 1000 input [> output]
linsi input [> output]
G-INS-i
```

Применим к наборам данных менее 200 последовательностей примерно одинаковой длины. Опирается на результаты глобального выравнивания.

Варианты командной строки:

```
mafft – globalpair –maxiterate 1000 input [> output]
ginsi input [> output]
E-INS-i
```

Предназначен для наборов из менее чем 200 последовательностей, содержащих длинные невыравниваемые куски.

Варианты командной строки:

```
mafft –ep 0 –genafpair –maxiterate 1000 input [> output]
einsi input [> output]
```

Методы, предназначенные для быстрого, но менее точного выравнивания:

- FFT-NS-i

Метод итеративного выравнивания, включает только два цикла

Варианты командной строки:

```
mafft –retree 2 –maxiterate 2 input [> output]
fftnsi input [> output]
```

- FFT-NS-2

Быстрый прогрессивный метод

Варианты командной строки:

```
mafft –retree 2 –maxiterate 0 input [> output]
fftns input [> output]
```

- NW-NS-i

Быстрый итеративный метод без быстрого Фурье-преобразования

Варианты командной строки:

```
mafft –retree 2 –maxiterate 2 –nofft input [> output]
nwnsi input [> output]
```

- NW-NS-PartTree-1

Очень быстрый метод, рекомендуется для 10 000–50 000 последовательностей.

```
    mafft –retree 1 –maxiterate 0 –nofft –parttree
input [> output]
```

Mafft можно использовать и для выравнивания двух наборов данных, один из которых используется в качестве матрицы или «профиля»:

```
    mafft-profile group1 group2 [> output]
    mafft –maxiterate 1000 –seed group1 –seed group2
/dev/null [> output]
```

Контрольный вопрос

Перечислите наиболее распространенные программы для множественного выравнивания нуклеотидных и аминокислотных последовательностей.

5. КЛАДИСТИКА

Основание современной молекулярной филогенетики в значительной степени было заложено в работах, посвященных формализации использования морфологических признаков при эволюционных построениях. Многие термины, предложенные тогда, сохранились и используются до сих пор, несмотря на то, что породившая их теория в значительной степени потеряла свою актуальность. Именно поэтому имеет смысл начать изложение с краткого введения в наиболее логичный раздел филогенетической систематики кладистику и определить её основные понятия.

Кладистика (от др.-греч. κλαδωσ – ветвь) – раздел филогенетической систематики, использующий в качестве метода кладистический анализ, т. е. набор строго формализованных методов построения схемы филогенетических взаимоотношений между видами, а также имеющий свой набор понятий, используемых для их описания.

В рамках кладистического подхода предполагается вначале использовать набор признаков организмов для выяснения эволюционных взаимоотношений этих организмов, а затем на основании филогенетической схемы построить классификацию этих организмов. Иерархичность классификации (виды, роды, семейства и т. д.) при этом оказывается вторичной по отношению к филогенетической схеме. Филогенетическую схему принято представлять в виде эволюционного дерева. Вместе с тем важным элементом кладистики является требование взаимно однозначного соответствия между реконструированной филогенией и иерархической классификацией. Кладистический анализ – основа большинства принятых в настоящее время биологических классификаций, учитывающих родственные отношения между организмами. Другое важное требование кладистики – монофилия, т.е. происхождение от единственного общего предка всех классифицируемых организмов.

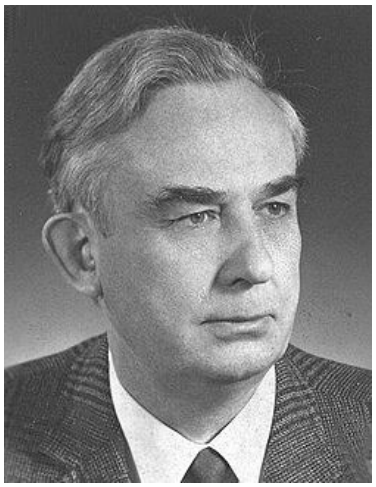


Рис. 5.1. Вилли Хенниг – создатель кладистики

Смежные направления филогенетической систематики, например фенетика или эволюционная таксономия, отличаются от кладистики в основном по их отношению к этим требованиям.

Основы кладистики были заложены в работах немецкого энтомолога и систематика Вилли Хеннига (1913–1976, рис. 5.1), написанных в 1950–1960-х гг.

Само название «кладистика» было впервые использовано только в конце 1960-х гг. критиками этого направления, а современные методы разработаны только в 1970-е гг. и предполагают довольно большой объём вычислений, которые практически невозможно проделать вручную.

5.1. Предположения в основе кладистики

Одним из главных мотивов, побудивших Хеннига к началу работы над формализацией методов систематики и использованием филогенетической систематики, был царивший в то время полный произвол в этой дисциплине, вызванный, в частности, отсутствием строгих критериев сравнения различных систем живых организмов и тем, что в одни и те же термины разные исследователи вкладывали разный смысл[5].

Во-первых, предполагается, что эволюционную историю группы видов (популяций, других таксономических единиц, далее – ОТЕ от **О**перационная **Т**аксономическая **Е**диница, англ. OTU) корректно описывает эволюционное дерево. Соответственно формулируются и основные постулаты кладистики:

[5] Блестящий пример ухода от таких проблем продемонстрировал Чарльз Дарвин, которому удалось в своей книге **«Происхождение видов»** ни разу не пояснить, а что же он обозначает словом «вид». До сих пор друг с другом конкурируют множество часто не вполне совместимых определений этого фундаментального понятия.

1. Для любой рассматриваемой группы ОТЕ предполагается, что она монофилетична, т. е. происходит от единственного общего предка.

2. Эволюционное дерево бинарно (состоит исключительно из бифуркаций). Это означает, что от одного общего предка могут происходить только два ОТЕ-потомка.

3. Изменение состояний признаков происходит с постоянной скоростью и менее вероятно, чем бифуркация. В идеале на всем дереве изменение состояния каждого из признаков происходит только один раз.

Первое предположение по своей сути является общим для любого эволюционного исследования. Оно в пределе означает, что на Земле жизнь возникла только один раз и все населяющие и населявшие её организмы в той или иной степени родственны друг другу. Следовательно, можно, обладая достаточной информацией, установить эволюционную историю любого набора ОТЕ. Другое дело, что обычно выбор группы исследуемых организмов должен быть достаточно оправдан и представлять самостоятельный интерес. Подбор рассматриваемых видов весьма сильно может сказаться на результатах анализа эволюционной истории организмов и поэтому обычно требует отдельного и подробного обоснования.

Второе из этих предположений, пожалуй, наиболее спорно. Оно означает, что образование новых видов возможно *исключительно* в результате того, что группа организмов предков разделяется на две и только две группы потомков. Кстати, именно в результате принятия этого предположения в рамках кладистики невозможно говорить о том, что один вид является предком другого: вид-предок при разделении на группы прекращает свое существование, превращаясь в два «потомка». Известно довольно много примеров, когда наиболее естественным объяснением наблюдаемой картины была бы гипотеза о том, что одновременно произошло разделение предковой линии на несколько линий потомков (жесткая политомия). Либо имела место серия быстрых бифуркаций, и у исследователя недостаточно информации для того, чтобы заметить промежутки между этими бифуркациями (мягкая политомия). Это серьёзная пробле-

ма в кладистике, однако следует заметить, что и в рамках других подходов к эволюционной систематике в настоящее время нет эффективных методов анализа быстрой видовой радиации.

Последнее из этих предположений наиболее важно с методологической точки зрения, поскольку оно определяет свойства алгоритмов поиска эволюционных историй видов в рамках кладистики, а также определяет критерии, с помощью которых можно оценивать доказательную силу исследований в рамках этой дисциплины.

Естественно, что, как и многие другие естественнонаучные дисциплины, кладистика содержит целый набор серьёзных упрощений, которые трудно считать реалистическими и которые приняты только для того, чтобы облегчить анализ. К таким упрощениям относится, во-первых, предположение о том, что время, в течение которого образуется новая ОТЕ, пренебрежимо мало по сравнению со временем её существования. В результате точки бифуркации на дереве являются точками и во времени. В реальности же в зависимости от механизма видообразования силы, его вызывающие, должны действовать в течение достаточно долгого времени, и только в крайне редких случаях начало видообразования удается привязать к какой-то определенной дате. В качестве примера можно привести случаи внезапно наступившей географической изоляции в результате, например, тектонических явлений, как это произошло при формировании Панамского перешейка 12 млн лет назад.

Другое предположение, адекватность которого требуется всегда доказывать, – независимость эволюции признаков. Доказательство коэволюции разных признаков, так же как и их независимость, очень часто нетривиальная задача, требующая серьёзных генетических исследований, далеко не всегда возможных, и часто касается того, что же исследователь считает отдельным признаком.

Помимо независимости, к признакам предъявляются довольно жесткие формальные требования: они должны в рамках исследуемой группы принимать только два состояния и менять его на дереве единственный раз. Иначе такие признаки порождают плезиоморфии и не полностью разрешенные деревья.

Кладистика не зависит от признания ни одной из теорий эволюции. По сути это набор приемов и понятий, предназначенный для совместного анализа большого количества признаков. Благодаря этому кладистические методы можно использовать в небиологических дисциплинах, включая историческую лингвистику и задачи по определению авторства текстов.

5.2. Основные понятия и термины кладистики

Кладистика обладает весьма разработанным набором терминов, каждый из которых имеет своё определение. Эти термины можно применять и вне кладистики, поскольку их смысл зависит только от предположения о том, что дерево адекватно описывает эволюционную историю, а алгоритмы его вычисления не влияют на смысл понятий, перечисленных ниже (рис. 5.2, 5.3).

Монофилия – происхождение от единственного общего предка.

Плезиоморфия – состояние признака как у корневого (для данного дерева) вида и его потомков («близкая форма», «предшествующая форма» или «примитивная форма»).

Апоморфия – состояние признака у позднейших видов, отличающееся от того, которое присутствует у корневой формы.

Аутоапоморфия – состояние признака у одного из позднейших видов, отличающееся от того, которое присутствует у корневой формы.

Синапоморфия – апоморфия, присущая монофилетичной (происходящей от единственного общего предка) группе видов. Использование синапоморфий и является основой кладистического подхода.

Синплезиоморфия – случай плезиоморфии, общей для группы напрямую связанных с корнем видов.

Гомоплазия – сборный термин для обозначения всех ситуаций, когда эволюцию признака невозможно объяснить с помощью предположения о том, что он менял свое состояние в рассматриваемой группе видов единственный раз. То есть носители состояний этого признака не монофилетичны.

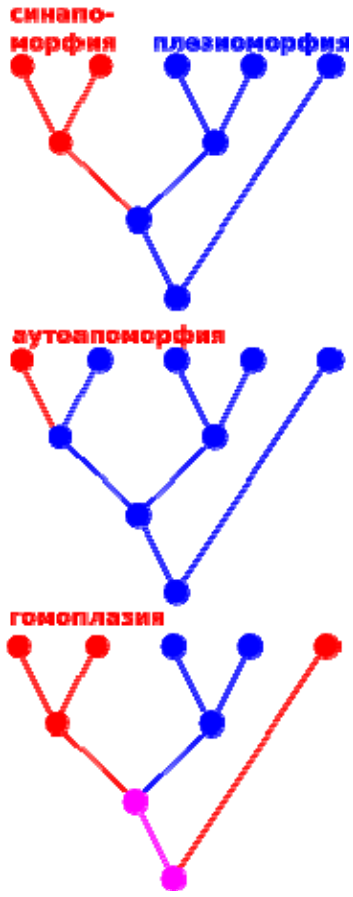


Рис. 5.2. Термины, описывающие взаимное расположение ОТЕ на дереве

Сестринская группа – ближайшая клада к исследуемой – ближайшая внешняя группа.

Эти термины ввели для того, чтобы избежать имеющих оценочный характер прилагательных «примитивный» и «передовой» (или «продвинутый»). Во-первых, состояние, плезиоморфное в рамках рассматриваемой группы видов, может оказаться гомоплазией относительно бóльшей группы. Во-вторых, оба состояния признака могут соответствовать структурам, крайне полезным при некоторых обстоятельствах и вредным – при других. Нередко набор плезиоморфных форм называется «базисом» для соответствующих клад.

Само название кладистики происходит, как отмечалось выше, от слова «клада». Поэтому естественно определить это понятие. Определения клады операциональны, т. е. кладой называется отличающаяся от остальных ОТЕ монофилетическая группа, выбранная одним из трёх способов.

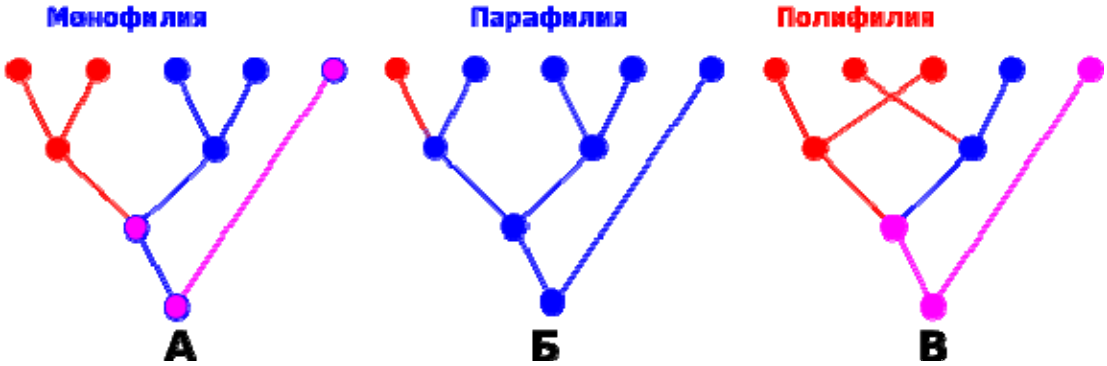


Рис. 5.3. Различные варианты эволюционных взаимоотношений ОТЕ монофилия, парафилия и полифилия (обратите внимание на положение общего предка в последних двух случаях)

На рис. 5.3 показаны различные варианты отношений, в которых могут находиться ОТЕ по отношению к своему происхождению.

Способы определения клады в кладистической таксономии:

1) *с помощью базального узла* – всю кладу представляет ближайший общий предок всех составляющих её ОТЕ;

2) *с помощью базальной ветви* – аналогичен предыдущему, только клада обозначается ветвью, ведущей к ближайшему общему предку клады;

3) *апоморфный* – через характерную для данной клады апоморфию.

В идеале морфологические, молекулярно-биологические и другие (поведенческие, экологические, палеонтологические и т. д.) филогенетические данные должны обобщаться при разработке итогового заключения, при этом никакой из методов не является более доказательным, чем другие, но все они имеют разные внутренние источники ошибок. Например, конвергенция развития признаков (гомоплазия) гораздо чаще появляется при анализе морфологических данных, чем в данных, полученных при молекулярном секвенировании, но реверсии признаков у них обоих встречаются с примерно одинаковой частотой; обычно морфологические гомоплазии могут быть вскрыты при достаточно внимательном и детальном анализе характерных признаков.

5.3. Задачи кладистики

Для практически любого таксона можно построить не одну, а множество различных кладограмм, основываясь на различных корневых видах и наборах характерных признаков; но из них выбирают одну-единственную, руководствуясь принципом парсимонии: компактную систему, которая, в сочетании с наименьшими возможными изменениями характерных признаков (синапоморф), дает непротиворечивую картину происхождения клады (в общем, это вариант соображения по бритве Оккама). Хотя в начале такой анализ проводился «вручную», впоследствии для него стали применяться компьютеры со специальным программным обеспечением, которое позволяет опери-

ровать на порядки большими наборами данных и количеством признаков. Такие программы (вроде PAUP и других подобных) позволяют делать статистическую оценку вероятности нод (узлов и разветвлений) построенной кладограммы.

Важно также заметить, что ноды кладограммы не обязательно отражают различия эволюционных ветвей, а лишь различия постоянных признаков, которые наблюдаются между этими ветвями. Признаки, которые заключаются в разнице последовательностей ДНК, способны расходиться после того, как генный дрейф между популяциями редуцируется к некоторой пороговой величине, в то время как заметные морфологические изменения, обычно будучи эпистатическими (т. е. результатом взаимодействия нескольких генов), выявляются только после того, как разошедшиеся таксоны отдельно эволюционно развивались в течение некоторого (обычно довольно значительного) времени; так, биологические подвиды зачастую могут быть разрознены генетически, но не морфологически (по строению тела или внутренней анатомии).

Кладистическая классификация

Во второй половине XX в. в биологии возникла тенденция под названием «кладизм» или «кладистическая таксономия». Эта тенденция состоит в том, чтобы считать таксоны кладами. Следовательно, биологическую классификацию предполагалось реформировать таким образом, чтобы прекратить считать таксонами все группы, не удовлетворяющие определению клады. Это отличается от общепринятого подхода практических систематиков, которые стремятся к тому, чтобы каждая таксономическая группа живых существ отражала её филогенетическую историю. Кладистический подход при этом используется часто, но также допускается в формировании классификационных деревьев использование как монофилетических (что совпадает с кладистичным подходом), так и парафилетических таксонов. В результате с начала XX в. роды и таксоны низшего уровня формировали, основываясь на монофилетическом подходе, в то время как таксоны высокого ранга могут быть (а такие как класс и выше – обычно и являются) парафилетическими. В последнее время приходится все чаще отходить от прин-

ципа монофилетичности на уровне вида. Это вызвано всё бóльшей разрешающей способностью методов генетики. С точки зрения систематики это означает резкое увеличение количества признаков. Что, в свою очередь, делает попытки эти признаки согласовать друг с другом всё более сложной задачей.

В кладистической систематике монофилетической группой считается клада, состоящая из (предполагаемого) предшественника и всех его потомков, которые формируют одну и только одну эволюционную группу. Парафилетической называется группа, лишённая группы своих современных членов (например, вымерших или подвергшихся существенным преобразованиям, в результате которых сама их принадлежность к этой группе становится сомнительной)[6]. Например, традиционный класс Пресмыкающиеся не включает птиц, хотя птицы произошли от пресмыкающихся.

Группа, состоящая из членов, которые происходят из разных эволюционных линий, называется полифилетической. Например, сформированная ранее таксономическая группа «толстокожие» (Pachydermata) была затем признана полифилетической, поскольку включенные в нее слоны, носороги и бегемоты произошли от разных, не родственных между собой предшественников. Эволюционные таксономисты считают полифилетические группы ошибками классификации, вызванными конвергенцией и другими видами гомоплазии, которые ошибочно интерпретируются как гомологии. В случае молекулярных признаков, относительно простые правила их эволюции и большое количество позволяют минимизировать риск ложной гомологии. Однако во многих случаях невозможно выяснить вопрос об исходном состоянии признака, более того, из-за ненулевой вероятности инсерций или делеций зачастую невозможно доказать или опровергнуть гомологии, казалось бы, одной и той же позиции в последовательности. Именно по этой

---

[6] Например, для выделения группы был использован набор признаков, которые есть у её представителей и отсутствуют у не принадлежащих к этой группе организмов. Резкое изменение части этих признаков даст основания выделить некое подмножество в отдельную группу, тем самым превратив исходную в парафилетическую.

причине в молекулярной филогенетике широко распространён термин «ортологи», который во многих случаях можно перевести как «может быть, даже гомологи».

Следуя Хеннигу, кладисты считают, что парафилия является не менее неприемлемой для классификации, чем полифилия. В рамках кладистики принимается, что монофилетические группы могут быть объективно определены с помощью общих предшественников, или синапоморфов. В отличие от них, парафилетические и полифилетические группы выделяются на основе анализа ключевых характеристик, при этом оценка того, ключевой или не очень конкретный признак, очень субъективна.

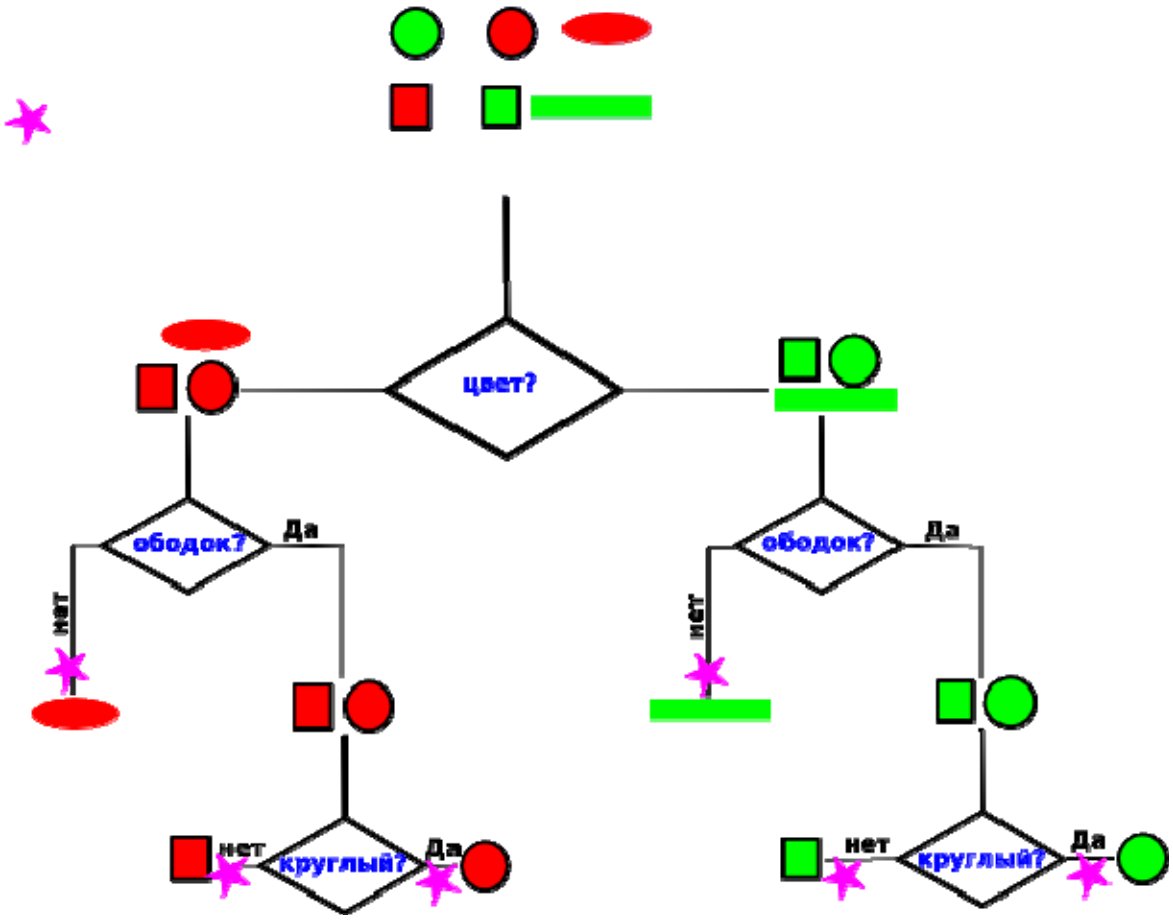


Рис. 5.4. Пример определителя, построенного для шести объектов с использованием трех независимо варьирующих признаков, каждый из которых принимает всего два состояния. Пурпурная звездочка обозначает, что объект идентифицирован. Для однозначного определения объекта достаточно максимум три шага. Порядок использования признаков не важен

Некоторые последователи кладистики утверждают, что выделение таксонов высшего по отношению к виду ранга является слишком субъективным, и поэтому их выделение не представляет никакой содержательной информации. На основе этого, по их мнению, эти таксоны не должны использоваться в систематике вообще. Таким образом, биологическая систематика, в их понимании, должна полностью отойти от линнеевской таксономии и стать простой иерархией клад.

В эволюционной систематике также существует точка зрения, согласно которой все таксоны по своей сути субъективны, даже если они отражают эволюционные взаимоотношения, поскольку живые организмы формируют непрерывное эволюционное дерево. В таком случае любая линия раздела является искусственной и создает монофилетическое выделение над парафилетическим. Парафилетические таксоны являются необходимыми для классификации более ранних секций дерева – например ранние позвоночные, которые через некоторое время развились в семью Гоминиды, не могут быть помещены более в одну монофилетическую семью. Также в этой системе аргументов приводится то, что парафилетические таксоны предоставляют информацию о существенных изменениях в морфологии, экологии и эволюционной истории организмов – короче говоря, и таксоны, и клады являются ценными для построения истинной картины систематики живой природы, но и те и те – с некоторыми ограничениями.

Формальный сборник филогенетической номенклатуры (ФилоКод) находится в перманентной разработке. Целью этой разработки является приспособление его к кладистической таксономии. Его планируют использовать и те, кто пытается полностью избегать линнеевской номенклатуры, и те, кто использует в систематике таксоны вместе с кладами. Таким образом, комбинация этих двух систем, возможно, позволит сформировать таксономическую картину, которая разместит группу живых существ на эволюционном дереве и при этом непротиворечивым образом учтет всю имеющуюся научную информацию.

Несколько других распространенных терминов кладистики введены для описания кладограмм и позиций таксонов внутри них. Вид, или клада, является базальным относительно дру-

гой клады, если первая имеет больше плезиоморфных черт, чем вторая. Базальная группа насчитывает меньшее количество видов по сравнению с развитыми группами; наличие в кладограмме базальной группы не является обязательным. Например, при совместной кладистичной классификации птиц и млекопитающих одна из этих групп не является базальной для другой.

Клада, или вид, находящаяся в кладограмме внутри другой клады, называется ингруппой, внутренней группой, «вложенной» в эту кладу.

Методы кладистики

Массив информации, который может быть подвергнут кладистическому анализу, должен быть организован специальным образом. Для этого, прежде всего, необходимо провести разграничение между признаками (или характеристиками) и их состояниями (характеристическими состояниями). Например, цвет перьев может быть голубым у одного вида и красным у другого. Тогда «голубые перья» и «красные перья» будут двумя состояниями одного признака – «цвета перьев».

Исследователь должен определить, какой признак (или признаки) присутствовал до появления последнего общего предшественника (плезиоморфия), а какой имелся у последнего общего предшественника (синапоморфия) путем выделения одного или нескольких корневых видов. Корневой вид – это организм, который не относится к исследуемой группе, но является ему близкородственным. Это делает выбор корневого вида важной задачей, поскольку такой выбор способен серьезно повлиять на структуру кладистического дерева. Надо заметить, что при характеристике клад используются только синапоморфии.

Следующей стадией является составление различных возможных кладограмм и их проверка. В идеале клады имеют много «согласованных» синапоморфий; в таком идеальном случае ожидают наличие достаточно большого количества настоящих синапоморф, которые не могут быть скрытыми гомоплазиями, появление которых вызывает конвергентная эволюция (т. е. качествами, которые воспроизводят друг друга благодаря влиянию окружающей среды или общего функционального использования, а не всеобщему происхождению). Известным примером гомоплазии, возникшей благодаря конвергентной эволю-

ции, служит признак «наличие крыльев». Хотя крылья птиц, летучих мышей и насекомых выполняют одинаковые функции, каждое из этих крыльев эволюционировало независимо, что может быть прослежено благодаря их анатомии. Если же птица, мышь и крылатое насекомое будут объединены признаком «наличие крыльев», в массив данных будет искусственно внесена гомоплазия, что разрушит анализ и скорее всего приведет в результате к неверному построению эволюционной картины.

Применение гомоплазии в морфологических наборах данных часто можно избежать путем точнейшего определения характерных признаков и увеличения их количества: в предыдущем примере, используя в качестве признаков «крылья с перьями», «крылья с хитиновым экзоскелетом» и «кожистые крылья» как характерные признаки, можно избежать эволюционно ложного объединения трех перечисленных групп животных на основе гомоплазии. При анализе «супердеревьев» (баз данных, включающих большее число таксонов исследуемой клады) применение неточных признаков может стать неизбежным, поскольку в обратном случае признаки могут стать неприменимыми для всех многочисленных таксонов. Скажем, такой признак, как наличие крыльев, не может быть применен для анализа филогении настоящих многоклеточных животных (Eumetazoa), так как большинству видов этого таксона данный признак не присущ. Таким образом, осторожный выбор и определение характерных признаков является другим важным элементом кладистического анализа. При ошибочном определении корневого вида и набора признаков, никакие методы построения кладограмм не смогут дать в результате филогенетической системы, которая соответствует эволюционной реальности.

Контрольные вопросы

1. Информативны ли аутоапоморфии и почему?
2. Почему в рамках одного популяционного набора последовательности называются паралогичными, а не гомологичными?
3. Перечислите критерии гомологии объектов.
4. Перечислите предположения о свойствах признаков, на которых основана кладистика.
5. Что такое гомоплазия?

6. ОСНОВНЫЕ ПОНЯТИЯ МОЛЕКУЛЯРНО-ФИЛОГЕНЕТИЧЕСКОГО АНАЛИЗА

6.1. Филогенетические деревья

Вначале определим основные термины, используемые при описании молекулярной эволюции. Итогом или обязательным промежуточным этапом любого молекулярно-филогенетического анализа является эволюционное дерево – граф, отражающий историю разделения эволюционных линий. Те объекты, которые подвергаются анализу, называются **о**перационными **т**аксономическими **е**диницами (**ОТЕ**).

Существуют две формы графического представления дерева, показанные на рис. 6.1, *А* и *Б*. Первая и наиболее традиционная из них корневое дерево. В этом случае сама форма дерева в явном или неявном виде предполагает, что процесс ветвления происходил на протяжении некоторого периода времени и расстояние точек ветвления от общего корня отражает срок, когда случилось соответствующее эволюционное событие. Бескорневой способ представления дерева (рис. 6.1, *Б*) используется главным образом в тех случаях, когда нет достаточных оснований указать на тот или иной узел дерева как на общий корень, соответствующий наиболее древнему эволюционному событию. Можно, конечно, такой узел произвольным образом выбрать. В обоих случаях соблюдается важнейшее правило, общее для всех деревьев, называющихся эволюционными. Это правило заключается в том, что всегда, во всех случаях на концах ветвей в качестве ОТЕ появляются только современные объекты (виды, организмы и т. п.). С другой стороны, во внутренних узлах дерева могут быть только ископаемые объекты. Таким образом, любая ветвь дерева, идущая «изнутри наружу» независимо от формы представления дерева, отражает направление течения

времени. Это правило не соблюдается только для так называемых простирающихся деревьев. Простирающиеся деревья (в англоязычной литературе *spanning trees*) представляют собой специальный класс деревьев, в которых современные объекты могут находиться во внутренних узлах.

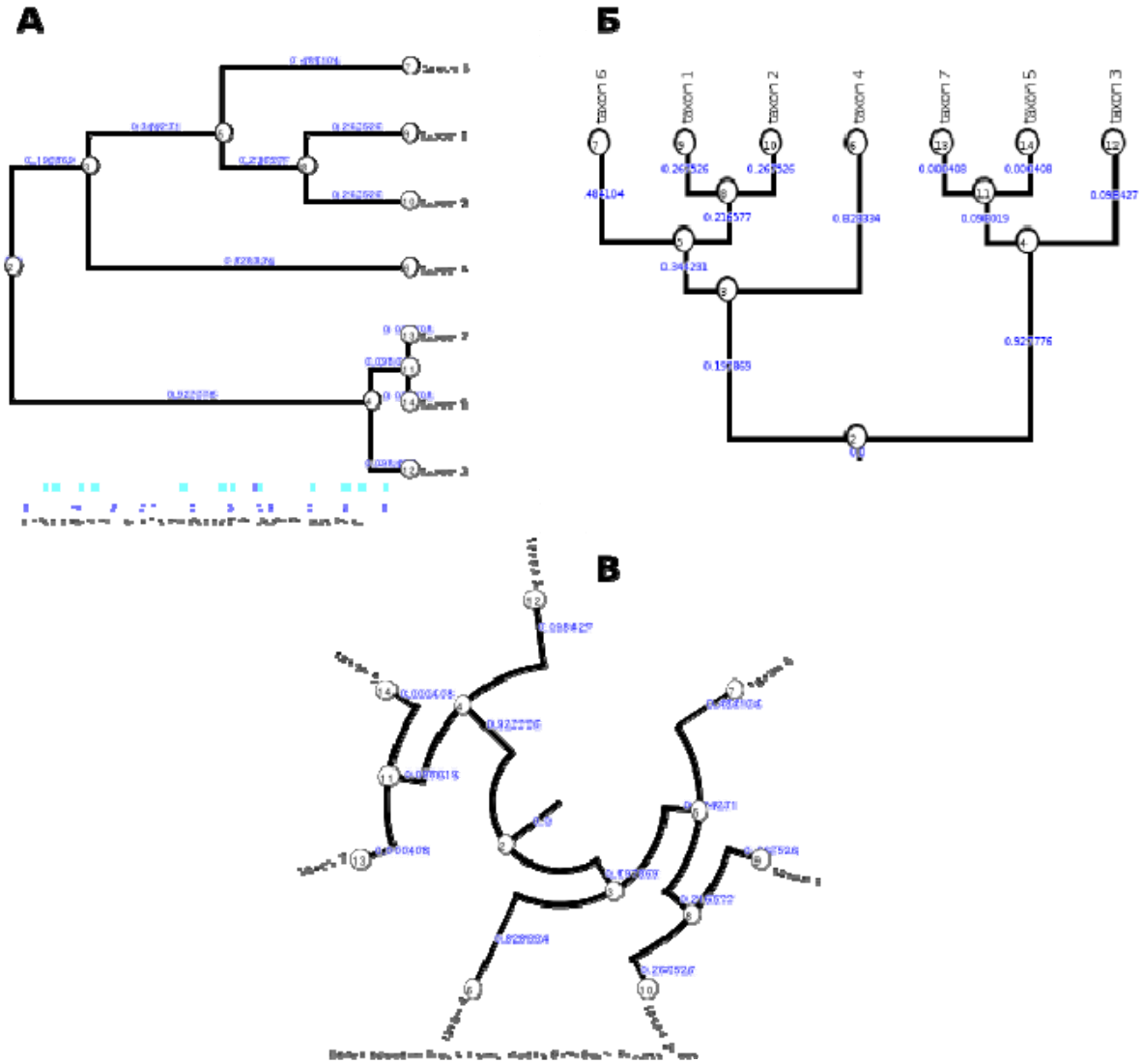


Рис. 6.1. Формы представления эволюционных деревьев. **А** – дерево представлено в традиционной или укорененной форме, длины ветвей пропорциональны возрасту, разделяющему соответствующие узлы. **Б** – то же самое дерево, представленное в виде кладограммы. Длины ветвей не имеют значения. **В** – бескорневая форма того же дерева

Такая форма используется для представления взаимосвязи между сосуществующими объектами, которые дивергировали относительно недавно, т. е. в основном в популяционных ис-

следованиях. Пример такого дерева приведен на рис. 6.2. Длины некоторых ветвей установили равными 0. Соответствующие узлы и ОТЕ на рисунке выделены рамками. Это, кстати, подчеркивает тот факт, что простирающиеся деревья на самом деле можно рассматривать как специальный класс эволюционных деревьев, у которых длины некоторых внешних ветвей могут быть равны нулю.

Рис. 6.2. Бескорневое дерево группы эндемичных гастропод. Выделены виды, расположенные на ветвях нулевой длины и политомии

Те эволюционные деревья, у которых все внутренние ветви имеют ненулевую длину, называются дихотомическими или бинарными. Другими словами это означает, что у объекта в каждом узле есть только один предок и (если узел внутренний) обязательно два потомка. Если же длина внутренней ветви оказывается равной нулю, то на дереве это выглядит, как будто у объекта, соответствовавшего такому узлу, было более двух потомков (как на рис. 6.2). Такие случаи называются политомиями. Количество дихотомических бескорневых деревьев N_n для n ОТЕ определяется формулой

$$N_n = (2n - 5)(2n - 7) \cdot \ldots \cdot 2 \cdot 1.$$

Для десяти видов, следовательно, можно построить 20 270 225 различных неукоренённых деревьев. Укоренение дерева эквивалентно добавлению еще одной ОТЕ – корня. Соответственно, число укорененных деревьев для того же числа ОТЕ будет ещё больше.

В филогенетическом анализе различают два вида политомий: мягкие и жесткие. Разница между ними состоит в том, что мягкие возникают тогда, когда не хватает информации для построения полностью разрешенного дерева. Жесткие политомии соответствуют реальной последовательности эволюционных событий, и их появление не зависит от того, сколько информации имеется в распоряжении исследователя. Если, конечно, он пользуется адекватными методами анализа.

Теперь рассмотрим, как же можно преобразовывать деревья. На рис. 6.3 исходное дерево, представленное на левой панели, подвергается последовательно двум преобразованиям, которые практически не приводят к изменению топологии дерева, хоть в результате получаются очевидно отличающиеся картинки. Вначале изменили внешнюю группу (или аутгруппу). Внешней группой называется ОТЕ, или ветвь, объединяющая несколько ОТЕ и соединяющаяся непосредственно с его общим корнем. Таким образом, операция изменения внешней группы эквивалентна объявлению другого узла корневым. Несмотря на то что при этом изменяется направление течения времени вдоль внутренних ветвей, соединяющих старый и новый корни, общая схема, характеризующая относительную близость ОТЕ и внутренних узлов друг к другу, не изменяется. Многие подходы к вычислению топологий филогенетических деревьев (см. ниже) на самом деле приводят к бескорневым деревам, и определение внешней группы требует привлечения дополнительных соображений. Следующая синонимическая операция, приводящая к еще более заметным изменениям дерева, однако совершенно не меняющая его смысла, – вращение вокруг предковых ветвей. Эта операция очень часто требуется для повышения «читаемости» деревьев, облегчения сравнения деревьев, полученных для сходных наборов организмов разными методами, особенно ес-

ли, как это часто бывает, они друг от друга отличаются. Вообще рисунок эволюционного дерева может нести различную смысловую нагрузку. Часто в зависимости от предположений, сделанных при вычислении топологий и методов, возникают специальные классы деревьев. Поскольку для их обозначения используются специальные термины, то давайте их и рассмотрим в данном разделе.

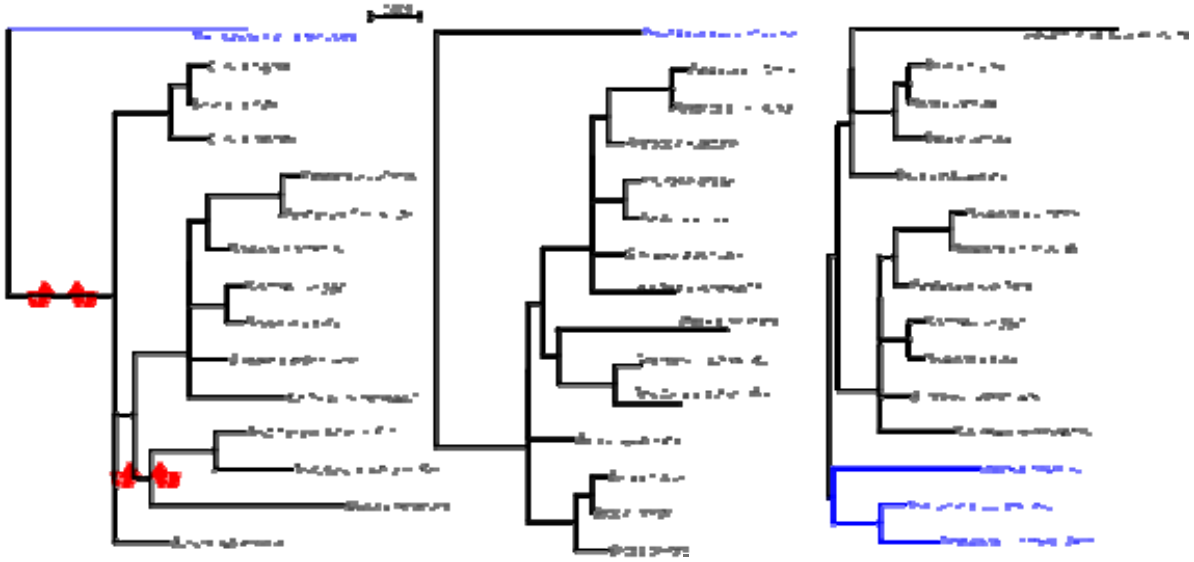


Рис. 6.3. Синонимичные преобразования эволюционных деревьев. Левое дерево описывает филогенетические взаимоотношения нескольких видов байкальских эндемичных гастропод и укоренено с помощью «центра тяжести». Второе получено с помощью вращения двух внутренних ветвей первого дерева (показаны красными стрелками), а третье – из второго путем укоренения с помощью внешней группы (выделена синим)

Дерево, представленное на рис. 6.1, *А,* было получено в предположении существования молекулярных часов. Эта гипотеза состоит в следующем: предполагается, что скорость возникновения и фиксации нуклеотидных (или аминокислотных) замен строго постоянна и одинакова во всех эволюционных линиях. Насколько оправдано это предположение и насколько часто оно соблюдается в реальности, будет обсуждаться ниже. Здесь же важно, что если оно сделано в процессе молекулярно-филогенетического анализа, то в результате положение корня дерева однозначно определяется и расстояния от ОТЕ до корня обязательно равны друг другу. Такие деревья называются ультраметрическими. Почти не отличимы на вид от ультраметри-

ческих деревьев так называемые кладограммы. Кладограммами называют такие деревья, которые отражают только порядок ветвления, или, иначе говоря, последовательность эволюционных событий, но длины ветвей никакого смысла не имеют. Как правило, они появляются либо в результате филогенетического анализа с использованием морфологических признаков, что выходит за рамки нашего пособия, либо появляются в результате применения статистических методов исследования достоверности полученной топологии. Бывают и другие случаи, когда длины ветвей не имеют смысла. В результате появляются рисунки, выглядящие как ультраметрические деревья только потому, что такой стиль представления сложился исторически задолго до того, как появились методы оценки длины ветвей.

И наконец, в противоположность кладограммам деревья, несущие информацию и о порядке ветвления, и о длинах ветвей, но не ультраметрические, называют филограммами (см. рис. 6.1).

Метод разбиений

Филогенетическую информацию можно представить не только в форме графа (дерева). Полезен оказался и другой способ, основанный на том, что филогенетический анализ обязательно включает определение порядка попарной близости ОТЕ друг к другу. Этот способ получил название **метода разбиений** (splits decomposition). Метод основан на том, что если для набора ОТЕ существует филогенетическое дерево, то каждая ветвь разбивает все множество ОТЕ на две неперекрывающиеся части, или “раздела”. Те разбиения, которые оставляют в одной из частей только одну ОТЕ, не представляют интереса. Поэтому обычно рассматриваются только те разбиения, которые содержат более одной ОТЕ в обеих частях. Очевидно, что для любого набора последовательностей данные будут удовлетворять только некоторому набору разбиений. Набор таких разбиений и будет описывать филогенетические взаимоотношения ОТЕ. Представляют разбиения в виде таблиц, подобных приведенной ниже для дерева из пяти видов, когда возможно только два информативных варианта.

Два непротиворечивых разбиения для дерева из 5 ОТЕ.

ОТЕ	**разбиение 1**	**разбиение 2**
s1	A	A
s2	A	A
s3	B	A
s4	B	B
s5	B	B

Эта таблица однозначно задаёт бескорневое дерево, которое в ньюикской нотации записывается как

(((s1,s2),s3),(s4,s5));

Метод разбиений, несомненно, существенно уступает в наглядности традиционному представлению эволюционных историй в виде деревьев, однако в ряде случаев оказывается очень полезным. Особенно интересен этот метод, если не удается однозначно построить филогенетическое дерево. Этот метод реализован в программе Splitstree.

Ниже мы встретимся еще с одним исключительно важным понятием «филогенетическая гипотеза». Так называется эволюционное дерево, отражающее всю полученную из анализа экспериментальных данных информацию плюс набор предположений, на которых этот анализ основан.

Контрольные вопросы

1. Перечислите предположения, лежащие в основе филогенетического анализа.
2. Формы представления филогенетических деревьев.
3. Сравните основные группы методов филогенетического анализа и перечислите программы, в которых они реализованы.

7. МОДЕЛИ МОЛЕКУЛЯРНОЙ ЭВОЛЮЦИИ

7.1. Нуклеотидные последовательности

Сравнение паралогичных[7] последовательностей нуклеотидов часто выявляет большое количество замен одного нуклеотида на другой. Типичная картина представлена на рис. 7.1.

Каждому нуклеотиду соответствует вертикальная черточка, при этом нуклеотидные замены выделены красным цветом там, где основания отличаются от самой первой последовательности.

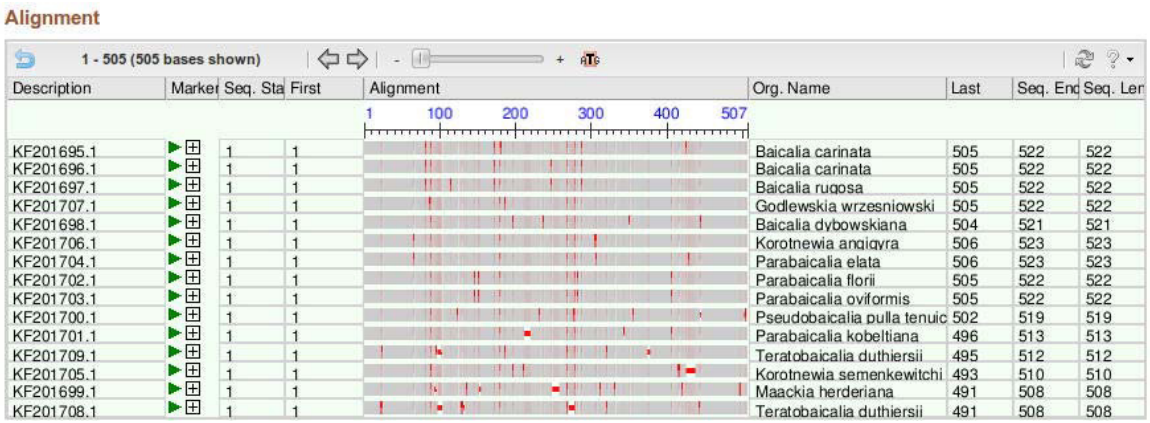


Рис. 7.1. Пример набора («стопки») нуклеотидных последовательностей в сжатой форме, которая используется в базе данных NCBI (в данном случае – http://www.ncbi.nlm.nih.gov/popset/544206152)

Как видно на этом рисунке, набор последовательностей обычно содержит довольно много замен. Представленный набор и дальше будет использован для построения различных деревьев, а получить его для самостоятельных упражнений можно по ссылке, приведенной в подписи к рис. 7.1.

[7] То есть наборов нуклеотидных последовательностей, кодирующих один и тот же белок, но происходящих из организмов, принадлежащих к различным таксонам. При этом последовательности располагают друг относительно друга таким образом, чтобы сходство между ними было максимальным. А вот происходят ли они от единственного общего предка – не определено. Отсюда и термин.

Задача построения филогенетических деревьев на основании наборов последовательностей обычно сводится к попытке выяснить максимум о механизмах возникновения и закрепления в ряду поколений нуклеотидных замен, а затем – в поиске наиболее подходящего сценария, который бы соответствовал наблюдаемой картине генетического разнообразия организмов. К счастью, законы, определяющие эволюцию на молекулярном уровне, гораздо проще, чем законы преобразования морфологических признаков. Настоящий раздел посвящен первому этапу молекулярно-филогенетического анализа – описанию правил преобразования нуклеотидов или аминокислотных остатков в эволюции конкретной группы организмов.

Решив эту задачу для своего набора данных, исследователь может как минимум ответить на вопрос, какие пары последовательностей ближе друг к другу по сравнению с другими парами, и насколько. То есть определить, сколько и каких мутационных событий произошло с момента дивергенции любой пары ОТЕ. Для этой цели используются модели молекулярной эволюции, которых в настоящее время используется порядка сотни только лишь для полинуклеотидов.

Модели с обратимым временем

Самая простая из моделей – Джукса – Кантора (обычно сокращается как JC). Она не учитывает отличия между различными типами замен и соответствует следующей матрице скоростей (или вероятностей возникновения и фиксации мутации):

$$Q = \left[\begin{array}{c|cccc} & A & T & C & G \\ \hline A & - & \alpha & \alpha & \alpha \\ T & \alpha & - & \alpha & \alpha \\ C & \alpha & \alpha & - & \alpha \\ G & \alpha & \alpha & \alpha & - \end{array}\right]$$

В строках и столбцах этой таблицы (или матрицы) стоят вероятности (скорости) появления *и закрепления в ряду поколений* соответствующих мутаций. Таким образом, в рамках этой модели принимаются предположения:

1) единицей мутации является один нуклеотид;

2) мутации в разных положениях ДНК происходят независимо друг от друга;

3) в неявном виде предполагается, что все мутации примерно нейтральны;

4) замены любого нуклеотида на любой равновероятны (нет различия между транзициями и трансверсиями);

5) различия частот нуклеотидов не имеют значения;

6) возможны повторные замены.

От прямого подсчета замен между всеми парами ОТЕ[8] эту модель отличает наличие последнего предположения – учёт возможности повторных замен. Соответственно, ожидаемую генетическую дистанцию, исходя из наблюдаемой, можно вычислить, используя уравнение

$$d = -\frac{3}{4}\ln(1-\frac{4}{3}p) \qquad (7.1)$$

Результат применения этой коррекции показан на рис. 7.2.

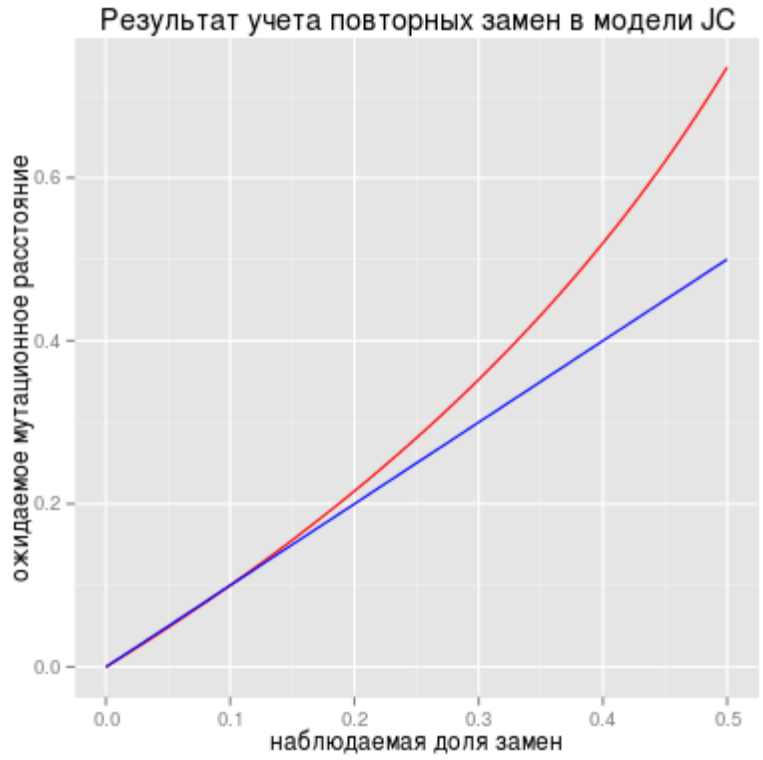


Рис. 7.2. Сравнение модели JC (красная кривая) с прямым подсчетом числа замен (синяя прямая). Возможность повторных мутаций приводит к тому, что для достижения наблюдаемой частоты замен их в реальности должно произойти больше

[8] Число положений, в которых 2 последовательности отличаются друг от друга, называется *P*-дистанцией.

Красная кривая показывает, что, объясняя *наблюдаемое* число мутаций, разделяющих две нуклеотидные последовательности, следует предполагать, что на самом деле их скорее всего происходило больше, но некоторые из них в силу случайности процесса приходятся более одного раза на одно и то же положение. Синяя прямая соответствует случаю, когда возможность повторных мутаций не учитывается[9]. Необходимо отметить, что на рис. 7.1 по осям отложены не числа замен, а их доли: $d = \frac{n}{N}$, где n – число положений, в которых две последовательности различаются, а N – длина последовательности. Очевидно, что при относительно низкой частоте замен, примерно до 10 % – разницы между этими кривыми практически нет, что эквивалентно утверждению, что в этих пределах вероятность повторных замен практически нулевая.

Расхождение кривых ожидаемой и наблюдаемой доли замен называется **эффектом насыщения**.

Различные вероятности транзиций и трансверсий в следующей по сложности модели – двухпараметровой модели Кимуры (K2P). Необходимость её появления обусловлена тем, что уже первые результаты сравнения наборов последовательностей ДНК из разных видов организмов показали, что транзиции накапливаются существенно (в разы) быстрее, чем трансверсии, и, следовательно, быстрее достигают насыщения. Матрица скоростей накопления замен в этом случае будет:

$$Q = \left[\begin{array}{c|cccc} & A & T & C & G \\ \hline A & - & \beta & \beta & \alpha \\ T & \beta & - & \alpha & \beta \\ C & \beta & \alpha & - & \beta \\ G & \alpha & \beta & \beta & - \end{array}\right]$$

Соответственно, уравнение 7.1 превращается в

[9] Разница стирается, если считать, что длина фрагмента ДНК стремится к бесконечности.

$$d = \frac{1}{2}\ln\frac{1}{1-2P-Q} + \frac{1}{2}\ln\frac{1}{1-2Q},$$

которое содержит уже два параметра: P (частота транзиций) и Q (частота трансверсий) в отличие от модели JC, где используется только один параметр – частота мутаций p. При анализе экспериментальных данных обычно определяют отношение частот транзиций к трансверсий:

$$K = \frac{T_s}{T_v},$$

и именно его используют для характеристики наборов последовательностей. Эта величина характеризует скорость, с которой начинает проявляться и искажать картину эффект насыщения. На рис. 7.3 видно, что чем больше сдвиг в сторону транзиций, тем быстрее возрастает кривая. То есть для того, чтобы образовались наблюдаемые 10 % замен, при $K = 4$ должно произойти гораздо больше мутационных событий, чем при $K = 2$, и больше, чем при эволюции согласно модели JC. Этот пример показывает, насколько важно выбрать адекватную модель молекулярной эволюции и доказать правильность своего выбора.

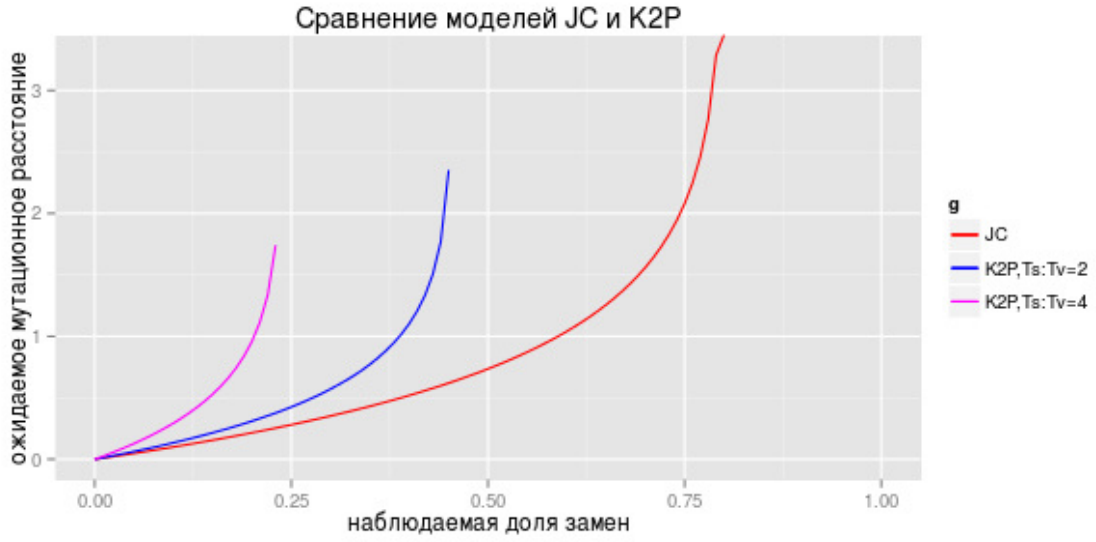


Рис. 7.3. Зависимость ожидаемой частоты мутаций от наблюдаемой в соответствии с разными моделями молекулярной эволюции

Сравнение нуклеотидного состава ДНК из разных организмов показало, что наблюдаемые частоты встречаемости нуклеотидов очень далеки от ожидаемых 0,25. У бактерий, например, доля (G+C) может варьировать от 75 % у термофилов до 25 % у психрофилов. Соответственно, даже если не учитывать различные скорости возникновения разных типов замен, то очевидно, что у организма, у которого 35 % составляет C и только 11 % A и эта особенность поддерживается отбором, вероятности замен C→T, T→C и G→A различны. Это обстоятельство учитывает модель, предложенная в 1981 г. Фельзенштейном (сокращённо F81). По сути это – та же самая модель JC, но – учитывающая различные частоты нуклеотидов.

$$Q = \left[\begin{array}{c|cccc} & A & T & C & G \\ \hline A & - & \mu\pi_T & \mu\pi_C & \mu\pi_G \\ T & \mu\pi_A & - & \mu\pi_C & \mu\pi_G \\ C & \mu\pi_A & \mu\pi_T & - & \mu\pi_G \\ G & \mu\pi_A & \mu\pi_T & \mu\pi_C & - \end{array}\right], \sum_{i\in(A,T,G,C)} \pi_i = 1$$

Согласно этой модели, чем дальше отклоняются частоты нуклеотидов от 25 %, тем быстрее сказываются эффекты насыщения. Если же все частоты равны 25 %, то F81 превращается в JC.

Модель, которая получается, если аналогичным образом учитывать частоты нуклеотидов в K2P, заслуживает отдельного упоминания. Она была предложена Хасегавой, Кишино и Яно. Сокращенно она называется HKY95 и в настоящее время чаще всего применяется по умолчанию при анализе экспериментальных данных с помощью большинства специализированных программ. Матрица скоростей в случае HKY85 выглядит так:

$$Q = \left[\begin{array}{cccc} - & \beta\pi_T & \beta\pi_C & \alpha\pi_G \\ \beta\pi_A & - & \alpha\pi_C & \beta\pi_G \\ \beta\pi_A & \alpha\pi_T & - & \beta\pi_G \\ \alpha\pi_A & \beta\pi_T & \beta\pi_C & - \end{array}\right], \sum_{i\in(A,T,G,C)} \pi_i = 1$$

И, наконец, наиболее общая модель насчитывает шесть параметров, т. е. для всех шести возможных пар нуклеотидов частоты различаются ([GTR]). Это – самая сложная модель, у кото-

рой матрица скоростей остается симметричной. Именно поэтому она называется «общей моделью с обратимым временем»[10]:

$$Q = \begin{bmatrix} - & \alpha\pi_T & \beta\pi_C & \gamma\pi_G \\ \alpha\pi_A & - & \delta\pi_C & \varepsilon\pi_G \\ \beta\pi_A & \delta\pi_T & - & \zeta\pi_G \\ \gamma\pi_A & \varepsilon\pi_T & \zeta\pi_C & - \end{bmatrix}, \sum_{i\in(A,T,G,C)} \pi_i = 1$$

LogDet-дистанция

Три последние модели позволяют учитывать нуклеотидный состав ДНК сравниваемых организмов, однако ***они предполагают, что состав ДНК этих организмов достоверно не различается***. Для близких видов обычно так оно и есть. Однако это требует подтверждения в каждом конкретном случае. Известно, что отношение (A+T)/(G+C) отражает температурный диапазон обитания. То есть даже у близких эволюционно таксонов, обитающих в резко контрастных условиях, это соотношение будет существенно различаться. Показано, что такие различия могут приводить к серьёзным искажениям результатов филогенетического анализа и ОТЕ могут сгруппироваться не в соответствии с эволюционным расстоянием между ними, а по сходству отношения (A+T)/(G+C). Именно для таких случаев предназначен метод LogDet .

Предположим, что имеются 2 последовательности *x* и *y*:

```
ACAGCAAAAAAACCGAAACGTCAGG
ACAGCAATAGAATCGAAACGTCAGG
```

На первом этапе составляем таблицу, в которой указываем число всех возможных пар нуклеотидов, один из которых принадлежит последовательности *x*, другой – *y*:

	x			
$y\downarrow$	A	C	G	T
A	11	0	0	0
C	0	5	0	0
G	1	0	4	0
T	1	1	1	1

[10] General time reversible model, или GTR.

Эту таблицу можно превратить в матрицу частот (в нашем примере длина полинуклеотидов – 25):

$$\mathbf{F}_{xy} = \begin{bmatrix} 0.44 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 \\ 0.04 & 0 & 0.16 & 0 \\ 0.04 & 0.04 & 0.04 & 0.04 \end{bmatrix}$$

Генетическое расстояние между последовательностями нуклеотидов вычисляют как минус логарифм определителя матрицы частот замен:

$$d_{xy} = -\ln(\det(\mathbf{F}_{xy}))$$

Неравномерность скорости накопления замен

При анализе реальных достаточно далеко дивергировавших нуклеотидных последовательностей практически всегда приходится принимать во внимание неравномерность скорости накопления нуклеотидных замен. В случае белок-кодирующих генов подавляющее большинство транзиций и в меньшей степени трансверсий в третьем положении кодона – синонимичны, в первом положении возможны синонимичные транзиции, а во втором – все замены приводят к заменам аминокислот. Следовательно, наибольшее количество замен будет наблюдаться в третьем положении кодона, а наименьшее – во втором. Возникают три класса позиций с разными скоростями эволюции. Кстати, зачастую если из одного набора белок-кодирующих фрагментов составить три отдельных, объединив в один набор все нуклеотиды, находящиеся в первом положении кодона, во второй – вторые и в третий – третьи, то для описания эволюции этих наборов будут подходить разные модели с очень различающимися коэффициентами. Аналогичным образом дело обстоит и в случае многих некодирующих последовательностей, например рибосомальных РНК. Известно, что у рРНК различные позиции существенно различаются по своей функциональной нагрузке или по важности для поддержания её третичной структуры. Как следствие, скорости замен в этой молекуле могут различаться на четыре порядка.

Наиболее распространённой моделью неравномерности скорости накопления нуклеотидных замен в настоящее время

является гамма-распределение Г. Форма этого распределения задается единственным «параметром формы» α. Если значение этого параметра меньше либо равно единице, то форма функции напоминает гиперболу (рис. 7.4). Это соответствует ситуации, когда в большинстве положений совсем не наблюдается никакой вариации, в вариабельных же положениях скорости накопления замен изменяются в очень широких пределах. Если же параметр формы больше единицы, то кривая приобретает максимум и начинает отдаленно напоминать колокол. Степень разброса скоростей уменьшается, а доля константных позиций – стремится к нулю. Исследование большого количества наборов реальных нуклеотидных последовательностей показало, что значения параметра формы колеблются в пределах от 0,1 до 1,4. Для относительно близкородственных последовательностей этот параметр обычно меньше единицы.

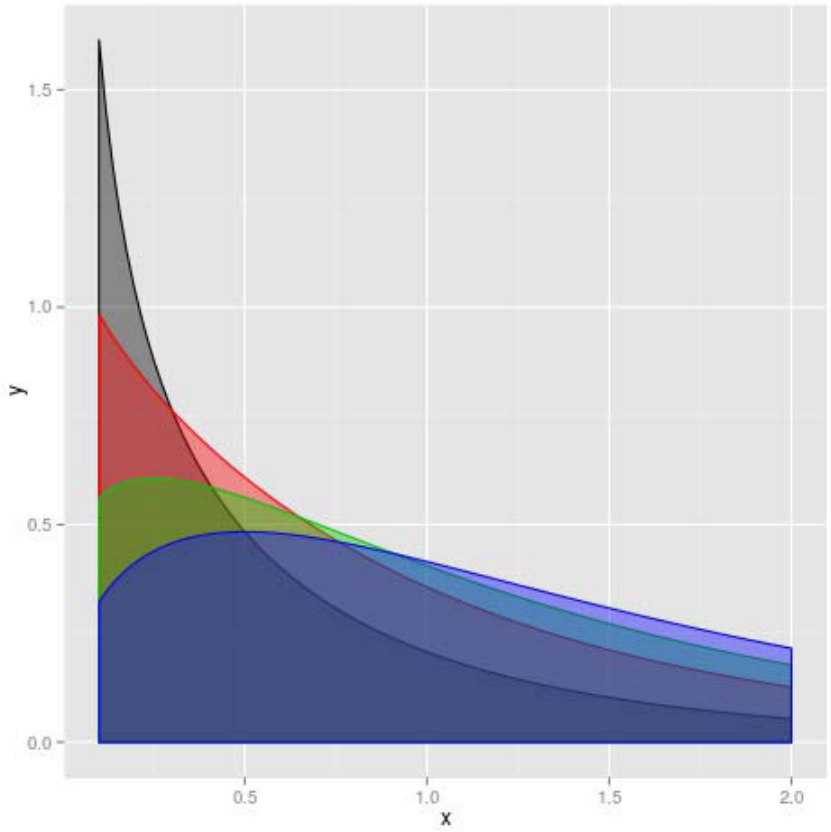


Рис. 7.4. Гамма-распределение при различных значениях параметра формы

Все «простые» модели молекулярной эволюции нуклеиновых кислот можно модифицировать так, чтобы они учитывали неравномерность скоростей в разных позициях. Делается это двумя дополняющими друг друга способами: во-первых, все положения делят на два класса – переменные и константные, и оценивают размеры этих классов. Такое разбиение не отражается на соотношении дистанций между ОТЕ, однако влияет на их абсолютное значение. Следовательно, топология дерева в результате применения этого приёма не изменится, а вот длины ветвей пропорционально станут другими. Это может оказаться важно, например, при оценке возрастов эволюционных событий. Во-вторых, для вариабельных позиций используется описанное выше Г-распределение. Использование состоит в предположении о том, что замены в вариабельных участках не равновероятны, а их вероятность распределена по закону, представленному на рис. 7.4. Поскольку выбор оптимального значения параметра формы для непрерывного Г-распределения – довольно трудоемкая задача, которая к тому же для реальных данных очень редко оправдана, распределение аппроксимируют дискретным, считая, что существует несколько классов позиций, в которых эволюция идёт с разной скоростью. Для белок-кодирующих последовательностей очевидно, что таких классов должно быть три – по числу положений кодона.

В результате любую из моделей с обратимым временем (см. выше) можно превратить в составную. Самая общая из таких моделей будет обозначаться как GTR+I+Г. Для того чтобы её применить к набору данных, требуется оценить:

1) частоты нуклеотидов (4 параметра);

2) скорости каждого типа нуклеотидных замен (6 параметров);

3) долю неизменных положений;

4) параметр формы Г-распределения,

т. е. – всего 12 параметров.

На первый взгляд, наилучших результатов можно достигнуть, используя эту – наиболее сложную модель, которая учитывает всё – и неравновесные частоты нуклеотидов, и консервативность части положений, и неравномерность эволюции. На самом деле это не так по двум причинам. Во-первых, использо-

вание такой модели сильно увеличивает объем вычислений. Даже при существенно возросшей в последнее время скорости компьютеров, для больших наборов данных эта проблема может оказаться критической. Другая причина состоит в том, что если последовательностей относительно немного (несколько десятков) и они не очень отличаются друг от друга, некоторые редкие типы мутаций могут не встречаться совсем либо наблюдаться всего в нескольких случаях. В результате оценка их вероятности будет содержать большую ошибку, что, скорее всего, исказит все результаты последующего анализа.

7.2. Аминокислотные последовательности

В отличие от генетических дистанций, применяемых при сравнении последовательностей ДНК, модели, описывающие скорости накопления аминокислотных замен, основаны не на теоретических выкладках, а на эмпирических наблюдениях. Первая такая модель была разработана Дэйхофф с соавторами в семидесятых годах. Эта модель послужила базовой для разработки целого семейства более точных и специализированных моделей. Эмпирический подход состоял в том, что авторы собрали наборы всех известных на то время аминокислотных последовательностей, которые отличались бы друг от друга не более чем на 25 %. Смысл этого ограничения состоял в том, что для любой пары таких последовательностей, если мы наблюдаем в одной из них, например, Leu, а в другой – Ile, то вероятностью того, что это произошло не в результате просто замены Leu→Ile, а в результате Leu→*x*→Ile или еще более длинной цепочки событий, можно пренебречь. В результате подсчета статистики замен были сконструированы матрицы скоростей замен *P(T)*, где *T* – время, аналогичные тем, что обсуждались в предыдущем разделе. Серия матриц P(0.5), P(1) и P(2.5), известных обычно как PAM50, PAM11 и PAM250, были предназначены главным образом для поиска по банкам данных белковых последовательностей и оценки достоверности неслучайного сходства обнаруженных последовательностей. Однако эти матрицы использовались и при филогенетическом анализе для вычисления мутационного (генетического) расстояния, однако без большого успеха.

В начале девяностых годов Джонс с соавторами применили ту же самую методику для гораздо большего количества белковых последовательностей, содержавшихся в базах данных (модель ЛТТ). Модель аминокислотных замен, сконструированная в результате этого исследования, привела уже к более интересным результатам. Важно отметить, что эти авторы вычислили матрицу скоростей аминокислотных замен отдельно для большого количества трансмембранных сегментов. Коэффициенты этой матрицы существенно отличаются от модели Дэйхофф, которая была получена в основном для глобулярных водорастворимых белков.

Другая специализированная модель сконструирована Адачи и Хасегавой для митохондриальных белков млекопитающих. В настоящее время считается, что применение этой модели для филогенетического анализа аминокислотных последовательностей митохондриальных белков приводит к наилучшим результатам. Супруги Хеникофф для построения модели эволюции аминокислотных последовательностей применили иной подход. Они использовали короткие сходные без делеций и вставок участки очень дальнеродственных белков и получили серию матриц скоростей, известную как BLOSUM. Матрицы этой серии отличаются друг от друга номерами (например, 50), которые обозначают минимальный процент сходства фрагментов, которые использовались для подсчета частот замен. Показано, что эти матрицы более эффективны при поиске гомологов в банке данных, однако они относительно редко используются в молекулярно-филогенетическом анализе.

В последнее время появились модели, описывающие замены аминокислот и основанные на том же принципе, что и LogDet, предназначенные для тех же целей, что и в случае нуклеотидных замен – для коррекции резко различающихся аминокислотных составов у относительно далеко дивергировавших белков.

Контрольные вопросы

1. Приведите причины, по которым необходим подбор подходящей модели молекулярной эволюции.

2. В чем заключается гипотеза молекулярных часов, насколько широко она применима?

8. ГЕНЕТИЧЕСКАЯ СТРУКТУРА ПОПУЛЯЦИЙ

В этом разделе мы рассмотрим, как использовать данные о разнообразии нуклеотидных последовательностей в рамках одной или нескольких популяций для определения её основных популяционно-генетических параметров. В популяционной генетике используют самые различные признаки, более того, решение реальных задач в настоящее время предполагает смешение разных типов признаков – морфологических, физиологических, молекулярных самой различной природы, и к тому же обычно требуется принимать во внимание географическую привязку образцов. Здесь мы, естественно, уделим основное внимание молекулярным признакам, а из них – последовательностям нуклеиновых кислот и SNP[11], в меньшей степени – полиморфизму длин микросателлитных повторов.

Разнообразие объектов, задач и подходов в популяционной генетике привело к появлению большого количества инструментов, каждый из которых имеет свои особенности, области применения, а главное – свои форматы для представления данных и результатов вычислений[12]. Все это разнообразие описать в одном пособии невозможно, поэтому мы ограничимся относительно небольшим, но по возможности представительным набором программ и задач.

---

[11] **S**ingle **N**ucleotide **P**olymorphism = однонуклеотидный полиморфизм, произносится как «*снип*».

[12] Громадное количество программ тем не менее не освобождает исследователя от необходимости обращаться к программированию по крайней мере на уровне простых скриптов. Особенно это важно для преобразования форматов данных или в случае необходимости срочно выполнить какие-либо расчеты, например, при чтении статьи коллег.

В отличие от предыдущих разделов, в рамках этого раздела нам потребуются в основном популяционные наборы данных. Для этой цели мы воспользуемся данными, полученными международной группой исследователей под руководством Risto Väinölä из Университета Хельсинки при исследовании байкальской эндемичной амфиподы *Acanthogammarus lappaceus* (PopSet: 353741837,).

8.1. SITES

Загрузка и подготовка набора данных

Поиск и загрузка осуществляются так же, как было описано в § 2.1:

1. Формула запроса в адресной строке портала Генбанка при выбранной базе для поиска PopSet:

```
Acanthogammarus[ORGN] AND COI[Gene]
```

Затем сохраняем 50 последовательностей части митохондриального гена первой субъединицы цитохром b-оксидазы на локальный компьютер в FASTA-формате.

2. Редактируем заголовки так, чтобы в них остались только индивидуальные номера доступа (таксономические названия у всех последовательностей одинаковы, поскольку это – популяционный образец). Для этого используем скрипт:

```
#!/usr/bin/perl

while(<>){
if($_=~/^>/){
@l1=split;
@l2=split('\|', $l1[0]);
print ">$l2[3]\n";
}
else {print "$_";}
}
```

3. Просматриваем файл данных с помощью Seaview (рис. 8.1), чтобы убедиться, что последовательности выровнены и имеют одинаковую длину, а также приблизительно оценить степень полиморфизма. Здесь она очевидно не велика, есть одинаковые последовательности.

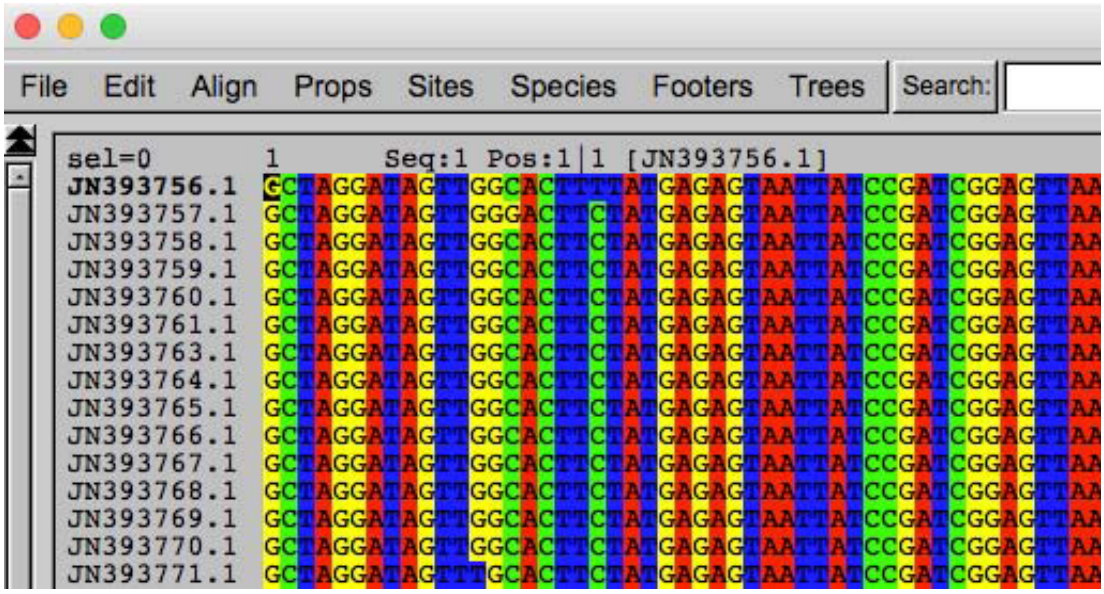


Рис. 8.1. Часть окна Seaview, показывающая выровненные последовательности набора *A. lappaceus*

4. Несмотря на то что мы имеем дело с относительно однородным набором близкородственных последовательностей, следует его кратко охарактеризовать с помощью быстрого филогенетического анализа. ***Необходимо помнить, что само понятие филогении предполагает, что предок не может сосуществовать с потомком. Очевидно, что для наборов популяционных данных это требование не верно!*** Тем не менее филогенетический анализ позволит определиться с дальнейшими действиями. Его можно выполнить в рамках программы Seaview. Метод построения дерева выбираем, как показано на рис. 8.2.

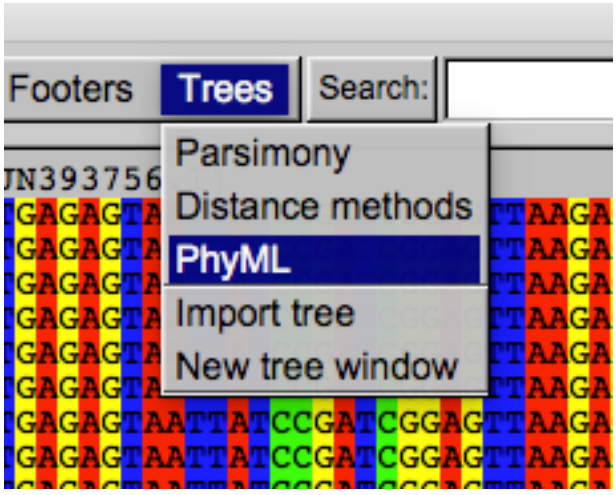


Рис. 8.2. Выбор метода построения дерева

Построение производится с помощью программы phyml, которой данные передает Seaview. Параметры модели при работе с такими маленькими генетическими расстояниями особого значения не имеют, поэтому можно использовать набор «по умолчанию».

Выбор дальнейшей стратегии

Филогенетическое дерево, построенное в окне программы Seaview с помощью phyml, приведено на рис. 8.3. Дерево содержит политомии и ветви нулевой длины.

Именно для анализа таких наборов данных и создана программа Sites.

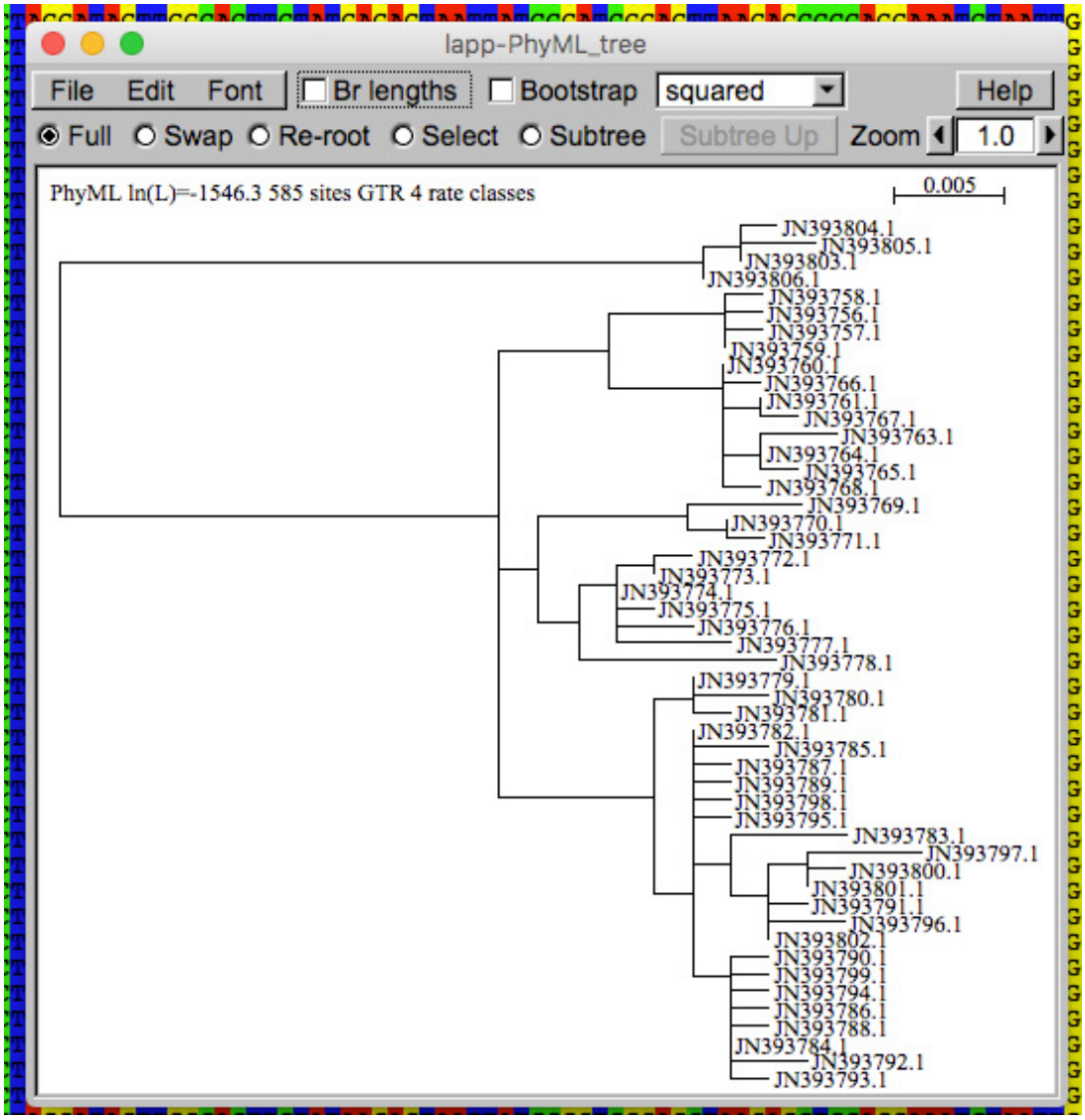


Рис. 8.3. Наиболее правдоподобное филогенетическое дерево нуклеотидных последовательностей COI *A. lappaceus* (PopSet: 353741837) в окне Seaview

Поскольку Sites воспринимает входные файлы в формате PHYLIP, то в заключение используем Seaview для изменения формата данных, как это показано на рис. 8.2.

Необходимо помнить, что Seaview файлы в формате PHYLIP сохраняет как interleaved (см. описание форматов).

Установка программы Sites

Программа Sites создана и доступна с сайта лаборатории Джоди Хея из Университета Рутгерса. Вся документация и ссылки на эту программу расположены по адресу https://bio.cst.temple.edu/hey/program_files/Sites/SITES_Documentation.htm

В документацию включена инструкция по скачиванию программы, а также условия, на которых её можно использовать.

После скачивания и распаковки файла Sites_source.zip следует перейти в директорию Sites_source. Она содержит пять файлов с расширением «.c». Для каждого из них следует выполнить команду компиляции[13]:

```
gcc -c sitemod.c
gcc -c siteread.c
gcc -c siterec.c
gcc -c siteutil.c
gcc -c sites.c
```

При компиляции некоторых из них система может выдавать предупреждения (Warning), на которые обычно не следует обращать внимания.

На заключительном этапе соберём все объектные файлы (с расширением «.o») в один исполняемый:

```
gcc -o sites sitemod.o siteread.o siterec.o siteutil.o sites.o
```

В случае успеха появится новый файл с названием Sites. Его надо скопировать либо в директорию, где находятся ваши данные, либо – в системную директорию из тех, что перечислены в переменной $PATH. Обычно для этого нужны права администратора системы. Вот к нему (к ней) и следует для этого обратиться. В нашем примере исполняемый файл переместили в директорию с данными.

---

[13] Здесь приводится пропись для любого Linux или MacOS X.

Предварительное описание полиморфизма

Одним из форматов, которые воспринимает Sites, является PHYLIP, в который мы перевели файл данных на предыдущем этапе. Им и воспользуемся.

Программу Sites следует запускать из консоля (терминала, командной строки). Если при запуске ей не сообщать никаких параметров, то она будет последовательно предлагать текстовые меню, в который требуется ей эти параметры сообщить. Ниже в таблице приведены параметры программы, которые ей можно сообщить в командной строке.

Таблица

Ключи программы Sites и их значения

Ключ	Значение	Пример использования
I	Название файла данных	-Imydata.phy
R	Название файла результата	-Rmyresult
M	Строка для выходного файла до 78 букв без пробелов	-Мзачем_это_я?
S	Какие полиморфные сайты анализировать?	-Sa
A	Какой тип анализа?	-Aspr
O	Опции для входных данных и для файла с результатами	-Ospr
L	Опции для анализа неравновесия сцепления	-Lypr
C	Определить необычный генетический код	-Cf

В нашем случае используем командную строку

```
./sites -Olapp.phy -Rlapp.res -Sa –A
```

и ответим на несколько дополнительных вопросов, которые задаст программа, в частности о количестве и координатах некодирующих участков (в нашем случае – 0) и т. п. Рассмотрим файл с результатами анализа. Поскольку все последовательности мы исходно считаем принадлежащими к одной-единственной группе, то часть результатов, относящаяся к межгрупповому сравнению, нам пока не интересна.

В начале файла приводится список параметров:

```
RUN PARAMETERS
--- ----------
  SITES INCLUSION STRING   : a
  ANALYSIS STRING       : a
DATA AND OUPUT OPTION STRING: s
ALL INPUT SEQUENCES INCLUDED
NUMBER OF SEQUENCES : 50  LENGTH OF SEQUENCES : 585
CODING AND NONCODING PARTITIONS
------ --- --------- ----------
NUMBER OF NONCODING INTERVALS : 0
APPROXIMATE TOTAL NONCODING LENGTH : 0
APPROXIMATE TOTAL CODING LENGTH  : 585

CODON FRAME POSITION OF FIRST CODING BASE : 1
```

После чего следуют довольно длинные таблицы, которые подробно описывают нуклеотидные замены и их последствия для аминокислотной последовательности, таблицу частот кодонов, попарные различия между последовательностями и т. д. Эта информация нужна, во-первых, для сравнения анализируемого набора данных с другими, а во-вторых, она полезна для анализа

```
POLYMORPHIC SITE FREQUENCIES – not including N’s or indels
----------- ---- -----------
Types of sites: All – all summed; Rep – replacement
Syn – synonymous; NC – noncoding
Sites with 3 or 4 bases in a group are sorted into the ‘rarest base’
vs ‘all others’
Sites that are ambiguous with regard to type are not counted
If an outgroup exists, both folded and rooted results are given, including
ixed differences

FOLDED
GROUP NAME  #seqs Type #Sites Mean  TajD
----------- ----- ---- ------ ----- ------
ALL         50    All   86    4.419 -1.342
ALL         50    NC    0      -     undef
ALL         50    Rep   20    3.050 -1.808
ALL         50    Syn   64    4.453 -1.258

FOLDED DISTRIBUTIONS
------ -------------
The number of lines carrying the rarest base – bases not rooted by outgroup

GROUP NAME  Type  1   2   3   4   5   6   7   8   9  10   11  12
----------- ---- --- --- --- --- --- --- --- --- --- --- --- ---
ALL         All  40   4   3  21   3   0   2   1   0   2   0   2
ALL         NC   0    0   0  0    0   0   0   0   0   0   0   0
```

```
     ALL    Rep  12   1   1  5    0   0   0   0   0   0   0   0
     ALL    Syn  28   3   2 16    3   0   2   1   0   1   0   2
     continued . . .

GROUP NAME  Type  13  14  15  16  17  18  19  20  21  22  23  24
----------- ---- --- --- --- --- --- --- --- --- --- --- --- ---
     ALL    All   1   1   0   0   0   1   0   1   1   0   0   3
     ALL    NC    0   0   0   0   0   0   0   0   0   0   0   0
     ALL    Rep   0   0   0   0   0   0   0   0   0   0   0   1
     ALL    Syn   1   1   0   0   0   1   0   1   1   0   0   1

     SITES WITH MORE THAN TWO BASES SEGREGATING - not including
N's or indels
     ----- ---- ---- ---- --- ----- -----------
     ALL    6 :  117  165  207  291  358  384

     D STATISTICS - test selection within groups. * - without out-
group information
          note - for a site segregating 3 or 4 bases:
     •    the site is counted as only one polymorphic site
     •    but, pairwise differences are calculated using all bases
     that occur

GROUP NAME  #seqs TajimaD Fu&LiD  Fu&LiD* Total-u Ext-u   Ext-u*
----------- ----- ------ -------  ------- ------- ------- ------
    ALL     50   -1.1337 no OutGp -2.1773   86    no OutGp  40

THETA (4Nu) ESTIMATES
GROUP NAME #seqs #bases #SITES ThetaW ThetaW/bp Thetapi  Thetapi/bp
---------- ----- ------ ------ ------- --------- ------- -------
     ALL     50   585.0    86   19.200   0.03282  13.05   0.02230

_____________________________________________________________________
*********************************************************************
     •    end of SITES program output
```

Наибольший интерес представляет последняя строчка – вычисленная, исходя из молекулярного полиморфизма, имеющая фундаментальное значение величина – параметр θ. Напомним, что для диплоидного организма по определению

$$\theta = 4 N_e \mu,$$

где μ – скорость мутирования, для нуклеотидов это число нуклеотидных замен в год на нуклеотид, а N_e – эффективная численность популяции[14]. Эту величину можно вычислить разными

[14] Эффективная численность популяции равна числу особей в идеальной популяции, в которой имеет место такой же уровень дрейфа генов, оцениваемый по выборочной дисперсии частот аллелей на поколение или по скорости уменьшения селективно нейтральной гетерозиготности, что и в реальной популяции.

способами, исходя из разных характеристик генетического разнообразия, а также – исходя из полученных экспериментальным путем экологических оценок физической численности популяций. Для того чтобы вычислить θ разными способами, должны выполняться разные наборы условий. Поэтому несовпадение значений, вычисленных разными способами, может указывать на невыполнение одного из наборов предположений. Именно на измерении достоверности разницы построен, в частности, часто используемый критерий Таджимы (в нашем примере – вторая строка числовых значений снизу), значение которого должно лежать в пределах от –2,5 до +2,5 для того, чтобы можно было утверждать, что у нас либо нет достоверных свидетельств действия положительного отбора на молекулярном уровне, либо – что численность популяции на протяжении весьма длительного времени оставалась примерно постоянной.

Популяционные параметры, оцененные с помощью Sites, могут оказаться очень полезны еще и для того, чтобы их использовать в качестве предварительных оценок (priors) в других программах, которые используют байесовские методы сравнения гипотез, касающихся интересных с общебиологической точки зрения вопросов. Обычно эти вопросы состоят в сравнении вероятности микроэволюционных сценариев, пытающихся объяснить современную картину генетического полиморфизма интересующей исследователя группы организмов. Эти методы требуют очень больших объемов вычислений и тем эффективнее, чем ближе к наиболее вероятным будут исходные значения параметров модели.

Более интересны результаты Sites в случае, если данные относятся к двум или более популяциям. Программа может провести сравнительный анализ полиморфизма, вычислить степень изоляции и оценить миграцию между популяциями. Правда, для этой цели следует (в очередной раз!) преобразовать формат данных, который относительно легко получается при редакции PHYLIP. Главная проблема здесь – это то, что в «родном» формате последовательности должны быть записаны в одну строчку, т. е. – sequential, а Seaview сохраняет их

interleaved, т. е. с переносами[15]. Такого рода проблему можно решить с помощью простого скрипта, а можно воспользоваться тем, что Python (интерпретируемый язык программирования) поддерживает очень удобный диалоговый режим, т. е. можно, находясь в среде Python, последовательно вводить одну команду за другой и наблюдать, как они выполняются. Пример такой сессии приведен ниже:

```
bash-3.2$ python ... from Bio import AlignIO dta = AlignIO.read(open(lapp.fas),fasta) f = open(lapp1.phy, w) f.write(dta.format(phylip-sequential)) f.close()
```

Рассмотрим эти действия подробнее.

1. Запускаем Python.
2. Подгружаем раздел библиотеки Biopython (http://biopython.org), содержащий функции для работы с наборами выровненных последовательностей (alignments). Теперь к этим функциям можно обращаться, ставя вначале название библиотеки (сейчас – AlignIO), а после точки – название функции. В скобках перечисляются значения её аргументов.
3. Используем функцию AlignIO.read() для того, чтобы открыть наш набор данных. Для этого в качестве аргумента ей сообщаем имя файла данных и формат данных.
4. Открываем файл для записи и присваиваем ему название.
5. Записываем в этот файл переформатированные данные.
6. Закрываем файл (необходимо, чтобы избежать возможной потери данных. Закрытие файла означает команду интерпретатору «Допиши всё до конца и поставь знак конца файла!»).
7. После этого можно закрыть сессию Python, напечатав [CTRL]-D.

Конечно, если такую операцию планируют производить неоднократно, то последовательность команд можно преобразовать в скрипт. Для этого следует открыть пустой текстовый файл и скопировать туда команды из диалоговой сессии. Он будет выглядеть так:

---

[15] Кстати, в инструкции авторов об этом – ни слова, ни полслова. Приходится пользователю самому догадываться о причинах ошибки. Подобное разгильдяйство распространено довольно широко, и надо быть к нему готовым.

```
#!/usr/bin/python from Bio import AlignIO dta = AlignIO.read(open(lapp.fas),fasta) f = open(lapp1.phy, w) f.write(dta.format(phylip-sequential)) f.close()
```

Его можно сохранить под каким-нибудь выразительным именем и каждый раз редактировать строки с именами файлов данных.

Итак, теперь мы можем воспользоваться примером из инструкции к Sites (белым на сером фоне обозначены комментарии):

```
SI-CA1 CAGGGTGTCCGACTCGGCCTACTCGAGCA......GCAACAGCC SI-CA2 CAGGGTGTCCGACTCGGCCTACTCGAACAGCTGCAGCAACAGCC . . .
```

и отредактировать файл lapp1.phy, соответственно. При этом, поскольку последовательности близки друг другу, как мы убедились выше, можно их считать некодирующими. На небольших эволюционных расстояниях это вполне корректно делать. Начало входного файла, который обозначим lapp1.SIT, будет выглядеть так:

```
 sequential version 50 585 1 1 1 585 2 group_A 46 group_B 4 JN393756.1GCTAGGAT..
```

Анализ на этой стадии становится сложнее, и поэтому лучше сократить объем информации, сообщаемой в командной строке, за счёт большего использования диалогового режима. Поэтому ограничимся запуском Sites с помощью команды

./sites -Ilapp1.SIT -Rlapp2g

и затем выберем из всех типов анализа только анализ полиморфизма, и включим все позиции последовательностей. В выходном файле тогда эти условия будут суммированы так:

```
RUN PARAMETERS
--- ----------
  SITES INCLUSION STRING   : i
  ANALYSIS STRING       : pm
DATA AND OUPUT OPTION STRING: n
ALL INPUT SEQUENCES INCLUDED
NUMBER OF SEQUENCES : 50  LENGTH OF SEQUENCES : 585
GROUP NAMES AND SEQUENCE NAMES
----- ----- --- -------- -----

big
JN393756.1 JN393757.1 JN393758.1 JN393759.1 JN393760.1 JN393761.1
JN393763.1 JN393764.1 JN393765.1 JN393766.1 JN393767.1 JN393768.1
JN393769.1 JN393770.1 JN393771.1 JN393772.1 JN393773.1 JN393774.1
JN393775.1 JN393776.1 JN393777.1 JN393778.1 JN393779.1 JN393780.1
JN393781.1 JN393782.1 JN393783.1 JN393784.1 JN393785.1 JN393786.1
```

```
JN393787.1 JN393788.1 JN393789.1 JN393790.1 JN393791.1 JN393792.1
JN393793.1 JN393794.1 JN393795.1 JN393796.1 JN393797.1 JN393798.1
JN393799.1 JN393800.1 JN393801.1 JN393802.1
small
JN393803.1 JN393804.1 JN393805.1 JN393806.1
CODING AND NONCODING PARTITIONS
------ --- --------- ----------
NUMBER OF NONCODING INTERVALS : 1
NONCODING SEQUENCE # 1 FROM : 1 TO : 585
APPROXIMATE TOTAL NONCODING LENGTH : 585
APPROXIMATE TOTAL CODING LENGTH  : 0
```

Две таблицы перечисляют, как мы подразделили на группы последовательности. Во вторую (small) попали только четыре из них, те, которые на рис. 8.4 образуют отдельно лежащую группу из четырех OTE (сверху выглядят как внешняя группа и на самом деле ею и являются). Выполняя этот анализ, программа сравнила параметры разнообразия внутри групп и между группами, оценила их эффективные размеры и миграцию между группами. Разберем подробнее, как это выглядит в выходном файле, который, как и было сказано в командной строке, называется lapp2g.SIT (расширение программа добавляет автоматически, чтобы это отключить, надо править код, а авторы в этой особенности не сознались – только чтобы жизнь была интересней).

```
GROUP DIFFERENCES – not including N's or indels (не включая N или индели)
----- -----------
•     над и на диагонали – средние попарные различия
•     под диагональю: суммарные средние попарные расстояния
1     2
                ----------
      1: big       |  9.50 |  33.36
                ---------------------
      2: small      27.61 |  2.00
```

т. е. числа на диагонали (верхний левый и нижний правый углы) означают средний внутригрупповой полиморфизм, а верхний правый и нижний левый углы – различные способы измерения межгрупповых различий, когда в любой сравниваемой паре один элемент принадлежит к одной группе, а другой – ко второй. Здесь мы видим, что межгрупповые различия существенно больше внутригрупповых, и, следовательно, разбиение данных на именно эти две группы имеет биологический смысл.

Следующая важная деталь содержится в двух таблицах. Первая из них:

```
FIXED DIFFERENCES - not including N's or indels
----- -----------
1     2
1: big      -  20
2: small    -   -
```

означает, что после разделения групп возникло 20 **фиксированных замен**. Другими словами, в двадцати положениях последовательности **все** члены группы big имеют один нуклеотид, а члены группы small – другой. Очевидно, что из-за маленького размера второй группы число фиксированных замен завышено, и скорее всего с увеличением выборки оно будет уменьшаться. Но сам факт наличия таких диагностических замен, во-первых, свидетельствует о длительной репродуктивной изоляции, а во-вторых – о возможности построить простой молекулярный тест на принадлежность вновь пойманной особи к той или иной группе. Последняя задача облегчается и приводимым ниже списком таких позиций.

Особенно интересно количество фиксированных замен сравнить с приведенным ниже фрагментом результирующей распечатки, который означает, что осталась только одна-единственная позиция, которая содержит нуклеотид полиморфный в обеих группах. То есть между этими группами практически отсутствует ***унаследованный общий полиморфизм***, наличие которого вполне распространено для пар сестринских видов. Сравнение числа фиксированных замен и практически исчезающе малого общего полиморфизма уже наводит на предположение о том, что мы имеем дело с двумя близкими, но тем не менее разными видами.

```
SHARED POLYMORPHISMS - not including N's or indels
------ -------------
1      2
 1: big       -   1
 2: small     -   -
```

И наконец, последний принципиально важный результат, который мы можем почерпнуть из анализа полиморфизма с помощью Sites, приведен ниже:

```
Fst AND POPULATION MIGRATION RATES (Nm, assuming diploidy)
--- --- -------------------------
•     above the diagonal: Nm estimates assuming diploidy
•     below the diagonal: Fst estimates
1     2
          ----------
 1: big       |     |  0.052
          --------------------
 2: small     0.828 |
```

Если оценка миграции между двумя группами (правый верхний угол таблицы) имеет мало смысла, поскольку вычисляется в предположении о диплоидности организмов по анализируемому маркеру, а мы имеем дело с митохондриальным геном, т. е. по этому маркеру организмы *приблизительно* гаплоидны, то $F_{ST} = 0{,}828$ очень важен. *F*-статистика, или F_{ST}, характеризует, какую долю от общего полиморфизма системы из двух групп организмов составляют межгрупповые различия. Эта величина изменяется от 0 (две группы составляют одну панмиктическую популяцию) до того, как между ними полностью исчезает сходство. Такие высокие величины, как получилась у нас, характерны для сестринских, но довольно давно разделившихся видов.

В общем случае проблема разделения набора последовательностей на группы, или другими словами – поиск популяционной структуры – очень сложная задача, не имеющая какого-то общего решения. Можно для иллюстрации этого утверждения использовать наш пример – полиморфизм *Acanthogammarus lappaceus*. Если внимательно рассматривать рис. 8.3, то большую группу можно разбить еще на три группы и провести подобный вышеописанному анализ распределения полиморфизма между ними. Существуют специальные методы, предназначенные для того, чтобы найти достоверное дробление набора последовательностей на подмножества и для статистического обоснования такого разбиения. Необходимо отметить, что такого рода задачи помимо чисто научного интереса могут представлять и некоторую практическую ценность, если они касаются экономически важных организмов.

Контрольные вопросы

1. Перечислите основные параметры, которые характеризуют генетическое разнообразие популяций.

2. Какие основные параметры применяются для описания взаимодействия между популяциями?

9. ФИЛОГЕНЕТИЧЕСКИЕ СЕТИ И ПРОСТИРАЮЩИЕСЯ ДЕРЕВЬЯ

Привычная форма представления эволюционных процессов – филогенетические деревья. Они хорошо сочетаются с представлением об эволюции как о процессе накопления различий в результате мутаций и череды видообразовательных событий. Построение филогенетических деревьев требует соблюдения ряда условий. Основным из них можно считать унаследованное из кладистики правило, согласно которому предок и потомок не могут сосуществовать. То есть во внутренних узлах дерева могут располагаться только вымершие формы.

Очевидно, что для случаев внутривидового полиморфизма или даже унаследованного полиморфизма у недавно дивергировавших сестринских видов это требование соблюдаться не может. Более того, и для более масштабных процессов порой невозможно пренебрегать такими явлениями, как потеря генов и их образование в результате дупликаций, межвидовая гибридизация, горизонтальный перенос генетического материала и рекомбинация. Все эти явления довольно сложно или невозможно представить в виде филогении. Именно здесь нужны филогенетические сети и простирающиеся деревья.

9.1. SplitsTree4

Программу и инструкцию к ней можно загрузить с сайта **http://www.splitstree.org**. Она бесплатна, и авторы её активно совершенствуют. Поэтому имеет смысл периодически проверять, нет ли обновлений. SplitsTree в её современном виде написана на языке Java и поэтому работает на всех платформах – Windows, Linux и MacOS X. С этим же связаны и ее недостатки – относительно низкое быстродействие и проблемы с интерфейсом.

Теория филогенетических сетей

Теория филогенетических сетей, сравнительный анализ различных форм их использования в эволюционных исследованиях и соотношение с более привычными способами изображения эволюционных событий были предложены Бандельтом (1995), программно реализованы и систематизированы в работе Хусона и Брайанта (2006)[16]. Они определили филогенетическое дерево как ***любой граф, в котором узлы представляют таксоны, а ребра соответствуют эволюционным взаимоотношениям между ними***. Таким образом, обычное филогенетическое дерево представляет собой частный случай филогенетической сети. Для того чтобы сеть была деревом, требуется соблюдение довольно многих дополнительных ограничений. Об одном из них мы уже упоминали выше, а именно – о том, что сосуществование предка и потомка на дереве запрещено. В результате современные таксоны (OTE) по определению не могут располагаться во внутренних узлах. В филогенетической сети в общем случае это требование не является обязательным.

Другое бросающееся в глаза отличие состоит в том, что в филогенетическом дереве от одного узла к другому ведет единственный путь (естественно, от любого узла к любому можно «пройти», не выходя за пределы дерева/сети). В сети между двумя узлами может быть более чем один путь. Более того, не всегда существует единственный кратчайший путь. В результате в филогенетической сети могут присутствовать циклы, что совершенно немыслимо у филогенетических деревьев.

Интересный частный случай филогенетических сетей – ретикулярные сети. **Ретикулярная сеть – это филогенетическое дерево, содержащее *дополнительные ребра*, соединяющие узлы**. Это приводит к тому, что у одного потомка может быть несколько предков. Такие ситуации отображают явления ретикулярной эволюции – гибридизацию, рекомбинацию или горизонтальный перенос генов.

---

[16] Bandelt H. J. Combination of data in phylogenetic analysis // Plant Syst. Evol. Suppl. 1995. Vol. 9. P. 355–361; Huson D. H., Bryant D. Application of Phylogenetic Networks in Evolutionary Studies // Mol. Biol. Evol. 2006.Vol. 23, N 2. P. 254–267.

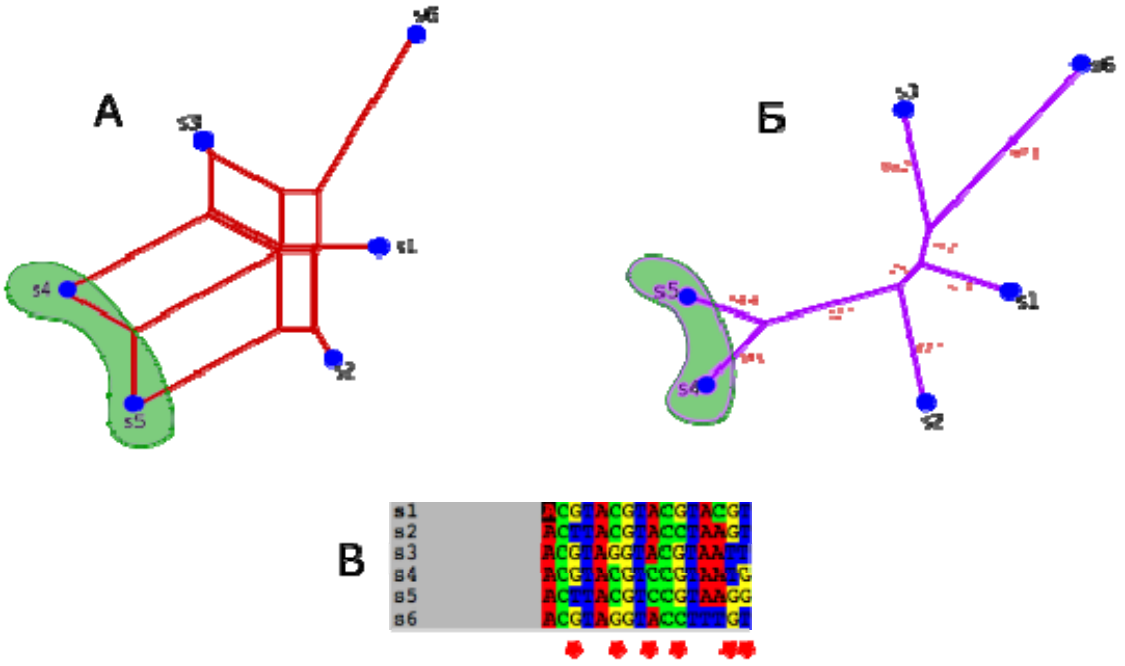


Рис. 9.1. Различные формы филогений. *А* – сеть разбиений, построенная для набора последовательностей, представленных на панели В; *Б* – NJ-дерево, построенное для тех же последовательностей с использованием P-дистанций (доля замен независимо от нуклеотида)

Сети разбиений (split networks) служат для отражения противоречивых сигналов в данных, а также ситуаций, когда информации недостаточно для доказательства единственной филогенетической гипотезы. В таких сетях узлы связывают между собой многочисленные параллельные ребра, каждое из которых соответствует определенному разбиению данных. Практически в любом наборе данных присутствуют противоречащие друг другу разбиения. Для того чтобы отобразить их все, приходится вводить в такую сеть дополнительные внутренние узлы, которые не соответствуют общим предкам.

Важное отличие сетей от филогенетических деревьев состоит в том, что у них запрещено вращение вокруг внутренних ветвей. Сравним положения s4 и s5 на рис. 9.1, *А* и *Б*. В случае дерева вращение вокруг ветви, ведущей к ближайшей внешней группе s2, ничего не изменяет с точки зрения эволюционной истории и топологии дерева. В терминах ньюикского формата

текстового представления топологии дерева это можно записать так:

((*s*4, *s*5),*s*2); = ((*s*5, *s*4),*s*2); = (*s*2,(*s*4, *s*5); = (*s*2,(*s*5, *s*4));

для сети разбиений это будут принципиально разные топологии:

((*s*4, *s*5),*s*2); ≠ ((*s*5, *s*4),*s*2); ≠ (*s*2,(*s*4, *s*5); ≠ (*s*2,(*s*5, *s*4));

Интерпретация сетей разбиений

Во многих случаях возникает потребность суммировать большое количество разных деревьев. Это могут быть бутстрепные реплики, могут быть результаты анализа разных генов и т. д. Традиционный метод решения этой проблемы состоит в построении разного рода консенсусных деревьев. Это может быть строгий консенсус, когда результирующее дерево содержит только те элементы топологии, которые присутствуют во всех суммируемых деревьях. Все остальные ветви схлопываются в политомии[17]. Иногда требование 100%-ной представленности смягчают, допуская, например 50%-ную. Либо просто оставляя те элементы, которые набирают при сравнении большинство[18]. Последний способ суммирования информации используется при построении привычных многим «бутстрепных» деревьев. Важно понимать, что в результате его применения *филограммы превращаются в кладограммы*, т. е. длины ветвей теряют смысл, а дерево иллюстрирует только порядок ветвления или очередность эволюционных событий.

Принципиально иной подход к решению задачи суммирования многих деревьев основан на сетях разбиений и состоит в том, чтобы:

1) перекодировать каждое дерево в набор разбиений;
2) найти в списках разбиений общую часть;
3) представить эту общую часть в виде сети разбиений.

---

[17] То есть узлы, у которых один предок, но более двух потомков. В ньюикской нотации это записывается как (*S*1,*S*2,...,*Sn*), где *n* > 2.

[18] Такие консенсусные деревья называются построенными по «правилу большинства» (majority rule).

В результате этой процедуры получается сеть – строгий консенсус. От неё легко перейти к частичному консенсусу уровня *X*, когда в «общую часть» включают все разбиения, которые встречаются с частотой, больше либо равной *X*. Необходимо отметить, что такие сети сохраняют гораздо больше информации, чем консенсусные деревья.

Существуют модификации обычных методов исследования статистической поддержки сетей разбиений (как и других разновидностей филогенетических сетей). Один из вариантов – такой же непараметрический бутстреп, как и в случае обычных филогенетических деревьев. Напомним, что этот метод состоит из следующих этапов:

1) **создание бутстрепной реплики**. Для набора последовательностей длиной *N* элементов генерируется последовательность случайных чисел, принадлежащих равномерному распределению от 0 до *N*, т. е. производится выборка *с возвращением*;

2) в соответствии с этой последовательностью выбираются колонки (вертикальные блоки шириной в один элемент последовательности) и подставляются справа к новому синтетическому набору данных;

3) в результате получается набор данных, в котором позиции исходного набора не только перемешаны в случайном порядке (это не важно!), но некоторые из них представлены несколько раз, а некоторые – не представлены совсем;

4) эту процедуру повторяют заданное число раз, получая таким образом большое число бутстрепных реплик.

В результате объединения расщеплений получается бутстрепная сеть, которую можно использовать аналогично консенсусным деревьям. Обычно такие действия приводят к некоторому упрощению сети и очень хорошо помогают проследить и объяснить неопределенности, обычно присутствующие в базальной части дерева. Оказалось, что построение сетей разбиений очень устойчиво к весьма неприятному артефакту традиционных методов построения филогенетических деревьев – слипанию длинных ветвей (long branch attraction), которое существенно затрудняет эволюционные исследования событий, далеко отстоящих друг от друга на временной шкале.

Интерес представляют и специальные проблемы, для которых хорошо подходят сети разбиений – детекция рекомбинации или конверсии генов в эволюционном масштабе времени, а также – для случаев сетчатого видообразования в результате межвидовой гибридизации. В программе SplitsTree предусмотрены специальные возможности для такого рода исследований, которые, однако, выходят за пределы данного пособия.

Контрольный вопрос

Что такое филогенетические сети и простирающиеся деревья, чем они отличаются от филогенетических деревьев?

10. ВРЕМЯ ЭВОЛЮЦИОННЫХ СОБЫТИЙ

Оценка времени расхождения генетических линий – один из важнейших этапов филогенетического анализа. Очень часто возникает потребность «привязать» эволюционные процессы к определенным этапам развития Земли и, соответственно, определить скорость их развития. В случае, если для анализа используют сравнение последовательностей нуклеиновых кислот, оценка скорости эволюционных процессов выражается в оценке **средней** скорости фиксации замен нуклеотидов. Или, после построения филогенетического дерева, оценка времени события сводится к оценке суммы длин ветвей, ведущих от узла-события к современным ОТЕ.

Инструментом определения времени события уже полвека является гипотеза «молекулярных часов», которую в простейшей форме можно записать как

$$t = kd,$$

где t – время; d – ожидаемое генетическое расстояние между видами; k – калибровочный коэффициент. Ожидаемое расстояние между видами вычисляют с использованием модели молекулярной эволюции $d = f(D)$, связывающей наблюдаемые замены с D и d. Калибровочный коэффициент можно оценить с помощью внешней информации – геологических или палеонтологических данных.

Довольно редко удается подобрать такую модель молекулярной эволюции, чтобы расстояния от современных видов до корня дерева были хотя бы примерно одинаковы, и к тому же эта модель имела бы достаточную статистическую поддержку. Поэтому в общем случае приходится использовать модель, согласно которой скорости фиксации замен различаются в различных линиях, т. е. вместо единственного k использовать большее количество: $k_1, \ldots, k_n$, где n в пределе может быть равно числу ветвей (терминальных и внутренних) дерева.

«Расслабленные» молекулярные часы

Скорость накопления нейтральных замен может зависеть от многих факторов – размера популяции, скорости возникновения мутаций, длительности поколения и т. п. Соответственно, существует большое количество моделей, описывающих вариабельность скорости молекулярной эволюции в пределах одного дерева.

Как уже отмечалось выше, простейшая модель была предложена еще в 1962 г. Полингом и Цукеркандлем. В современных терминах её можно назвать «строгой моделью молекулярных часов». К сожалению, к эмпирическим данным она применима довольно редко. В настоящее время разработано довольно много моделей, описывающих изменчивость скоростей эволюции между линиями. Обычно эти модели используют в качестве предварительных условий (priors) для байесовских методов. Применение численных методов типа MCMC (Monte-Carlo Markov Chains) для аппроксимации постериорных распределений вероятностей значений параметров сделало байесовские методы самым, пожалуй, мощным инструментом филогенетического анализа.

Модели изменчивости скоростей молекулярной эволюции

- *Гипотеза глобальных строгих молекулярных часов*. Скорость фиксации замен постоянна во времени.
- *Локальные молекулярные часы*. Близкородственные клады эволюционируют с одинаковой скоростью. Чем дальше клады, тем больше могут у них различаться скорости накопления замен.
- *Составной пуассоновский процесс*. По мере эволюции каждой из линий в некоторых точках могут происходить изменения скорости. Новая скорость при этом является произведением старой и случайной величины, распределенной в соответствии с гамма-распределением.
- *Автокоррелированные скорости*. Скорости эволюции плавно изменяются по ветвям дерева. В одном варианте скорость изменяется в каждом узле дерева, новая скорость является случайной величиной, распределенной согласно логнормальному распределению, математическим ожиданием которого ста-

новится «родительская» скорость накопления замен. В другом варианте используется процесс Кокса – Ингерсолла – Росса, когда скорость в дочерней ветви определяется нецентральным распределением χ^2. Процесс включает переменную, которая определяет инстенсивность, с которой он стремится к стационарному распределению.

- *Некоррелированные скорости*. Скорости, поставленные в соответствие каждому узлу дерева, являются случайными величинами, принадлежащими какому-либо параметрическому распределению, например экспоненциальному или логнормальному.

Большое разнообразие моделей изменчивости скоростей молекулярной эволюции, которые можно использовать в рамках модели расслабленных часов[19], может серьёзно затруднить их использование или по крайней мере – осложнить аргументацию предпочтения одной модели другой. Ещё больше осложняет картину то, что исследования применимости моделей расслабленных часов к разным наборам данных дают противоречивые результаты. Поэтому в рамках исследования обычно требуется сравнить несколько моделей и выбрать наиболее подходящую. Облегчить выбор можно также, если учесть масштаб эволюционных событий. Для описания эволюции относительно близких видов, представленных большим количеством последовательностей, скорее всего подойдут модели с автокорреляцией скоростей, тогда как анализ небольшого числа далеких друг от друга таксонов следует начать с применения некоррелированных скоростей.

Палеонтологические и геологические калибровки молекулярных часов

Филогенетический анализ, не принимающий во внимание внешней информации относительно палеонтологических находок или геологических событий, может давать только относительные оценки времени узлов дерева. Часто информации о по-

[19] Наверное, эту модель можно называть «моделью неточных часов». Действительно, оценки времени с её помощью имеют гораздо более широкие доверительные интервалы, чем если бы была использована строгая модель. Но, во-первых, уж лучше широкий интервал, который с высокой вероятностью включает верное значение, и во-вторых, «неточные часы» – это плохой маркетинг.

рядке эволюционных событий во времени бывает вполне достаточно для достижения целей исследования (например, при анализе тенденций в изменениях количественных признаков). Тем не менее гораздо чаще от филогенетического анализа ожидают именно временных привязок. Источниками данных для калибровки молекулярных часов могут быть данные палеонтологии, другие, уже калиброванные часы (с помощью их перекалибровки совместно с новыми данными), геологические события вроде образования Панамского перешейка, разделившего обитавших по обе его стороны особей или датированные образцы вирусных нуклеотидных последовательностей.

10.1. Анализ времени эволюционных процессов с применением BEAST

Скорость молекулярной эволюции и датировки

Данное упражнение основано на анализе набора предварительно выровненных последовательностей вируса папилломы кошачьих (Feline Papilloma Virus, FPV). Наша цель – рассчитать скорость эволюции для каждой генетической линии, основываясь на времени дивергенции видов-хозяев.

На первом этапе необходимо конвертировать файл с данными в формате NEXUS в файл BEAST формата XML. Для этого нам понадобится программа BEAUti[20] (Bayesian Evolutionary Analysis Utility). На втором этапе мы запустим BEAST с файлом данных, моделью и установками. Последний этап состоит в том, что суммировать полученные результаты.

BEAUTi: подготовка исходных данных.

1. Чтобы загрузить файл данных, содержащий последовательности ДНК в формате NEXUS, надо выбрать Import Alignment... из меню File, как указано на рис. 10.1:

[20] Пользователи последних версий MacOS X могут столкнуться с проблемами из-за совместимости версий Java, на котором написаны все используемые ниже программы. Если компьютер станет «выносить мозг», ищите решения на https://support.apple.com/kb/DL1572?locale=ru_RU.

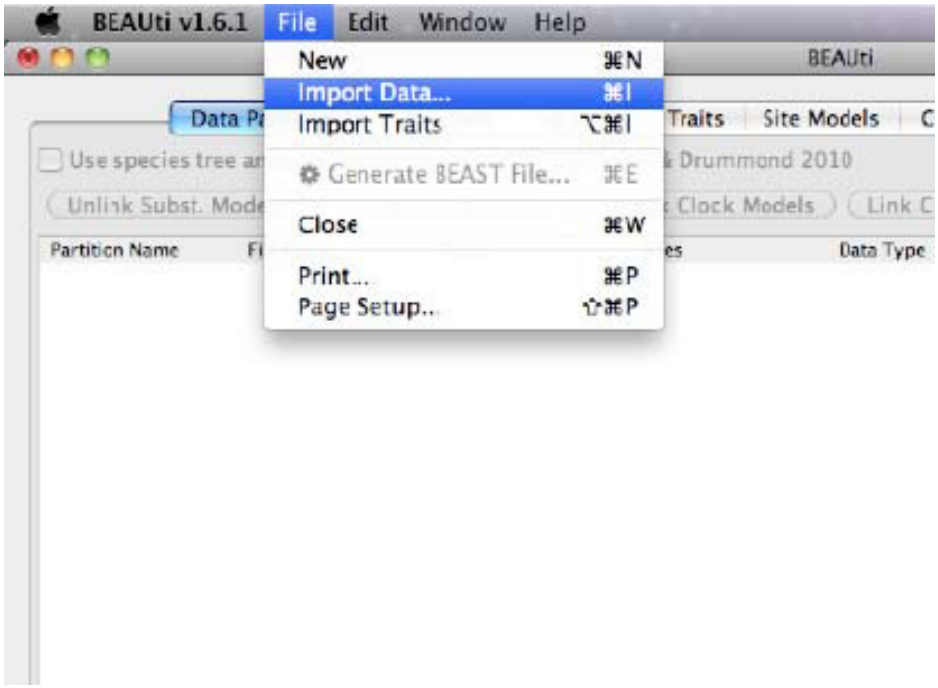


Рис. 10.1. Импорт набора нуклеотидных последовательностей с помощью BEAUTi.

Выбираем файл FPV.nex из папки BEAST/examples/VirusPractical/FPV. Этот файл содержит набор выровненных последовательностей фрагмента генома вируса папилломы пяти видов кошачьих и родственных вирусов енота и собаки. Выглядит файл в формате следующим образом:

```
#NEXUS
Begin data;
Dimensions ntax=7 nchar=1434;
Format datatype=nucleotide gap=-;
Matrix
CanineOralPV  ATGGCAAGGAAAAGACGCGCAGCCCCTCAAGATATATACCCTGCTTGTAAA
FelisPV1     ATGCTTAGGCAAAAACGTGCAGCCCCAAAAGATATTTACCCACAATGCAAG
LynxPV1     ATGCTACGGCGAAAACGTGCAGCCCCCCATGATATCTACCCCCAATGCAAA
PumaPV1     ATGCTTAGGCGAAAACGTGCAGCCCCCAAAGATATTTACCCCCAATGCAAA
RacoonPV1    ATGACTCGCAAACGCCGCGCCGCTCCTCGTGATATATACCCCTCTTGCAAA
AsianLionPV1  ATGCTAAGGCGAAAACGTGCAGCCCCCTCAGATATCTACCCCCAATGCAAA
SnowLeopardPV1 ATGCTAAGGCGAAAACGTGCAGCCCCTTCTGATATTTACCCACAATGCAAA
;
End;
```

2. Определим калибровочные узлы. Для этого выберите вкладку Taxon Sets в основном меню. Вы увидите панель, которая позволит вам создавать наборы таксонов. Создав набор таксонов, вы сможете позже добавлять информацию по калибровке для последнего общего предка (most recent common ancestor,

MRCA). Нажмите маленький «плюс» на левой панели внизу, это создаст новый набор таксонов. Двойным кликом переименуйте файл untitled1 в Felis/Lynx/Puma. Выделите таксоны FelisPV1, LynxPV1 и PumaPV1 и нажмите на зеленую стрелку.

Повторите эту же процедуру, создав набор Lion/Leopard, который будет содержать таксоны SnowLeopardPV1 и AsianLionPV1. Создайте группу таксонов Cats, которая содержит все последовательности вируса папилломы кошачьих (т. е. за исключением RacoonPV1 и CanineOralPV). Экран будет выглядеть, как показано на рис. 10.2.

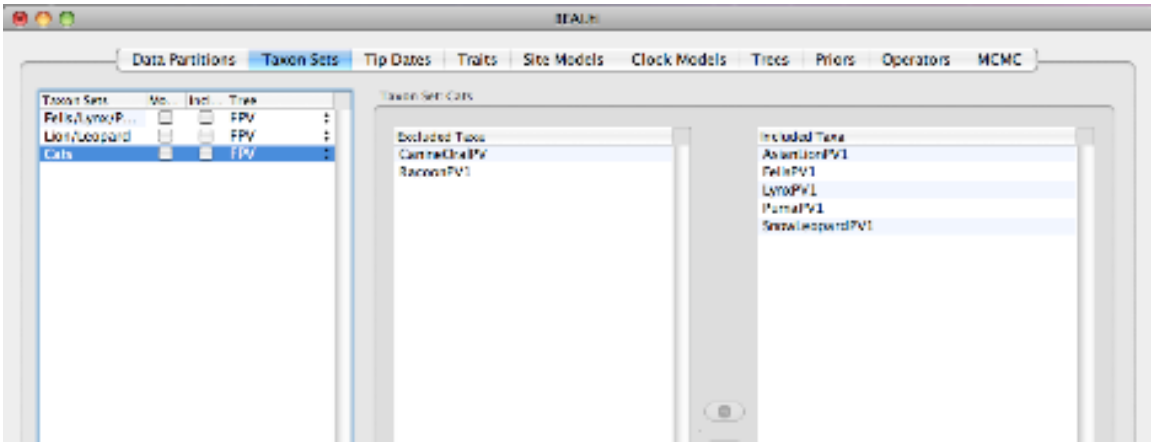


Рис. 10.2. Разбиение OTE на группы

Если бы мы хотели указать, что последовательности вирусов папилломы енота и собаки являются аутгруппой, мы должны были бы выделить колонку **Monophyletic?**. Это подтвердило бы то, что группа в процессе MCMC анализа Cats остается монофилитичной. Однако для нашего анализа мы не будем этого делать, так как мы хотим подтвердить, что дерево папилломы вирусов такое же, как и дерево хозяев. Таким образом мы **закладываем в расчеты гипотезу, что дивергенция вирусов происходила одновременно с дивергенцией хозяев**.

3. На следующем шаге предстоит выбрать модель молекулярной эволюции. Для этого выбрать вкладку Site Models, которая содержит установки эволюционной модели, которую будет использовать BEAST. Какие именно параметры будут отображаться, зависит от данных (нуклеотиды или аминокислоты) (рис. 10.3). Значения, которые автоматически будут установлены после загрузки набора данных FPV, будут весьма произвольны, поэтому нам надо внести некоторые изменения.

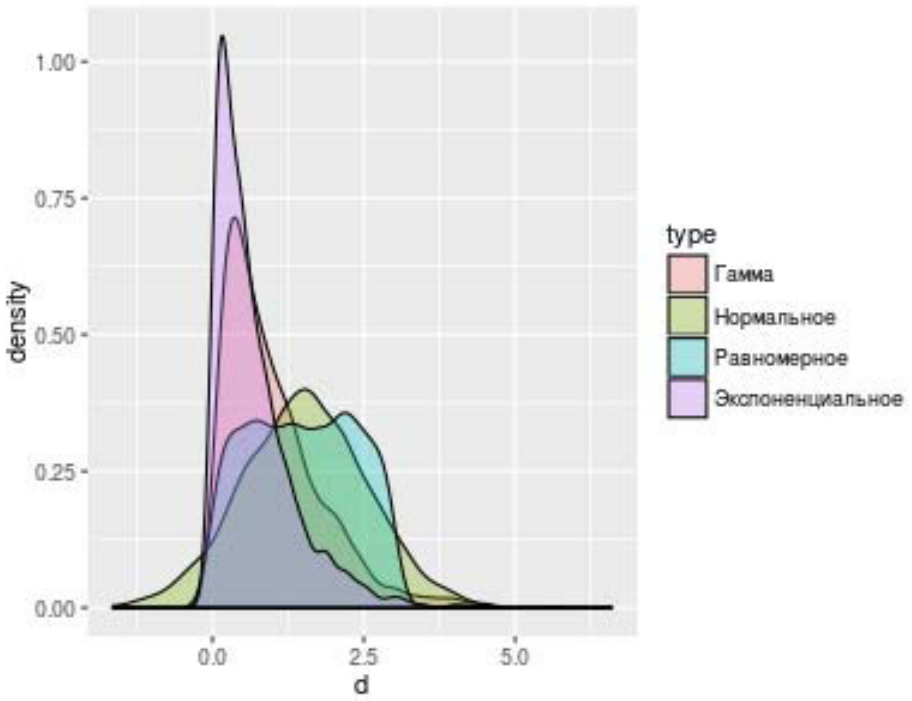


Рис. 10.3. Различные распределения с математическим ожиданием $\bar{x} = 1.5$

Большинство моделей вам известно. Для этого анализа мы сделаем следующие изменения:

- выберите Empirical в меню Base frequencies. Это означает, что мы не намерены пренебрегать различиями частот нуклеотидов, их исходные значения будут определены из анализа последовательностей[21] (рис. 10.4);
- отметьте Gamma в меню Site Heterogeneity Model. В результате модель учтет гетерогенность скоростей эволюции в разных положениях последовательностей. Для большинства популяционных выборок этого делать не стоит;
- **установка молекулярных часов.** Следующее, что нам надо сделать, это кликнуть на Clock Models и изменить модель молекулярных часов на Relaxed Clock: Uncorrelated Log-normal для того, чтобы оценить гетерогенность скорости молекулярных часов в эволюционных линиях. Опции вашей модели должны выглядеть, как указано на рис. 10.5.

[21] Этот и следующий шаги кажутся очевидными, но они добавляют в модель пять параметров вместо одного, что может неоправданно и сильно увеличить объемы вычислений.

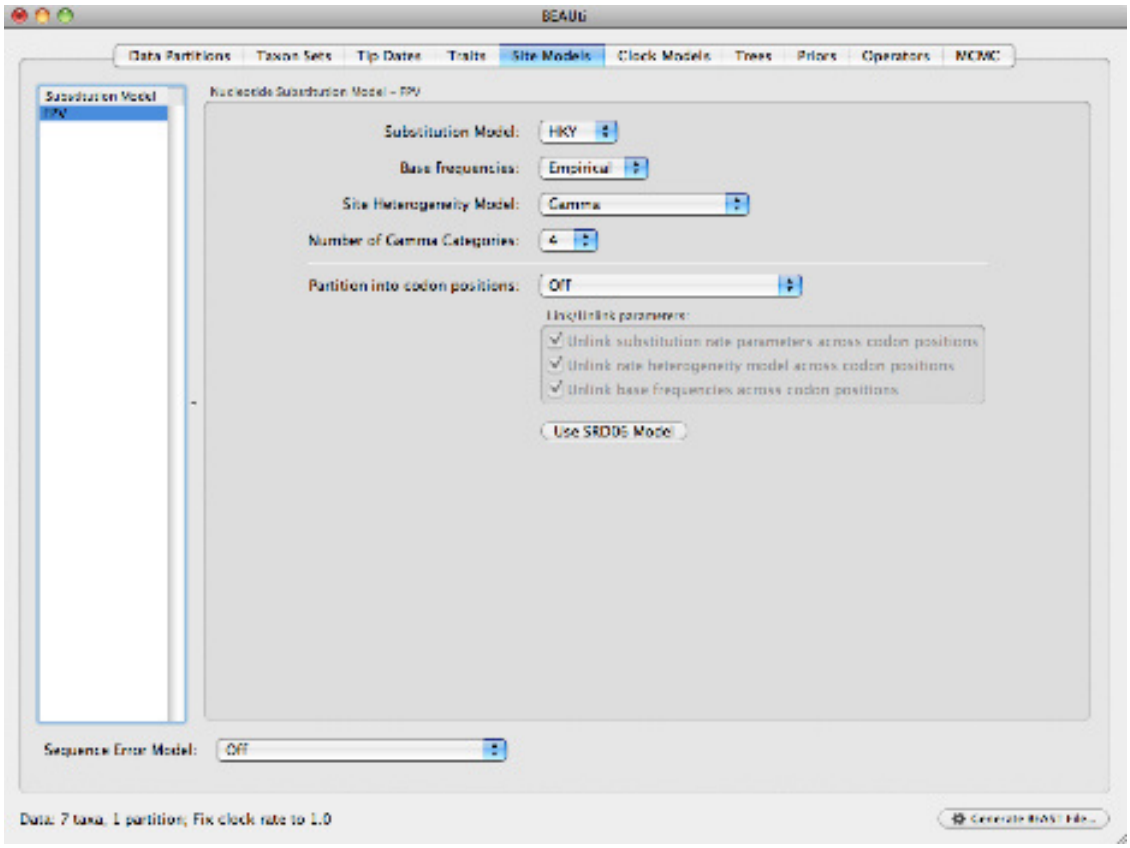


Рис. 10.4. Определение параметров модели молекулярной эволюции

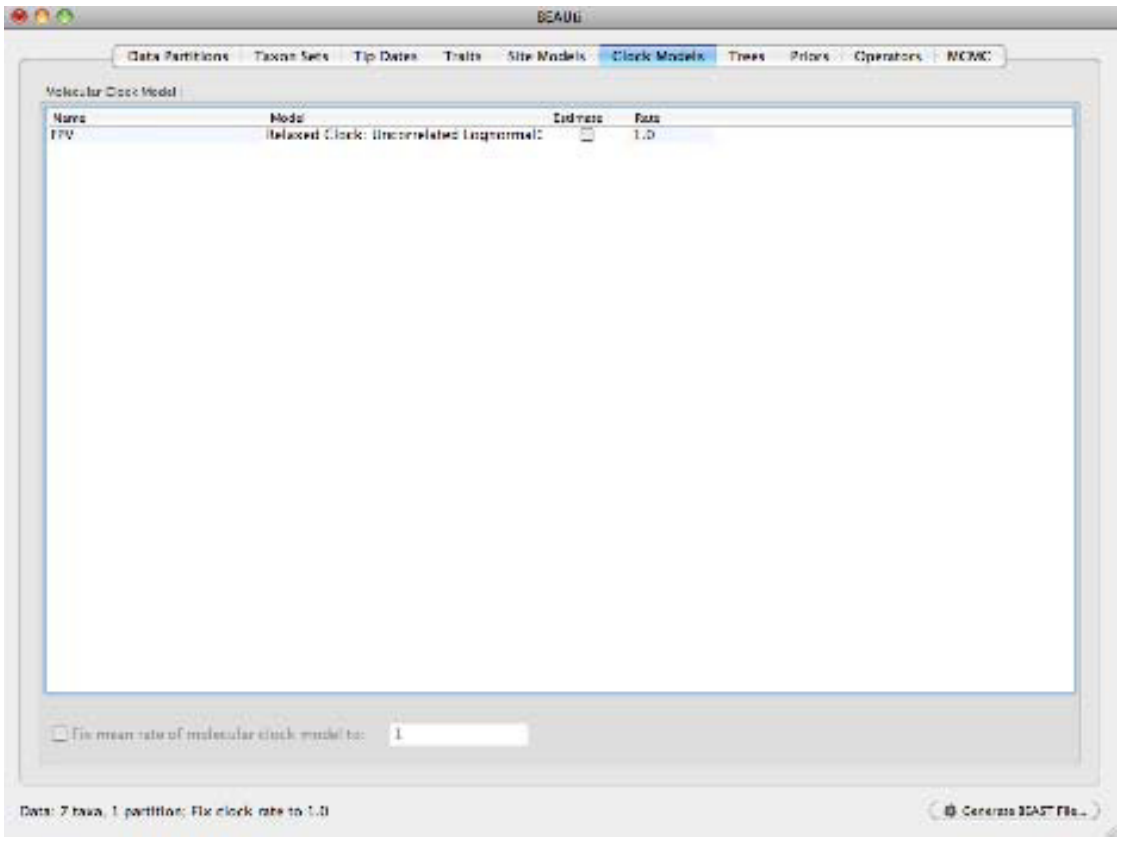


Рис. 10.5. Определение модели молекулярных часов

Необходимо установить флажок во вкладке Estimate, потому что мы хотим калибровать молекулярные часы (и, таким образом, времена дивергенции). Это будет установлено автоматически, в данном случае по статистике tmrca, которая появилась в панели Priors.

- Вкладка Trees позволяет уточнить каждый параметр модели. Первое, что должно быть установлено, это модель Yule. Это простая модель видообразования, которая, как правило, более целесообразна при рассмотрении последовательностей из различных видов. От других моделей её отличает некоторое преимущество, отдаваемое «сбалансированным» деревьям и более равномерному распределению узлов по длине дерева. Выберите ее в меню Tree prior. Остальные модели относятся, скорее, к вопросам, касающимся сравнения демографических сценариев при сравнении различных популяций.
- **Предварительные условия (priors)**. Как и при любом вычислении, использующем MCMC, перед началом работы требуется определить стартовые значения параметров, а также формы и пределы распределений их значений, которые будет перебирать программа BEAST. Что касается стартовых значений, то их можно оценивать с помощью простых инструментов вроде обсуждавшихся выше. Характер распределения следует выбрать один из представленных на рис. 10.6. В данном случае выбираем нормальное распределение (normal), согласно которому, по нашему мнению, распределены оценки времён событий, используемых для калибровки часов. В рассматриваемом примере таких событий два[22].

Кликните мышью кнопку tmrca у группы Felis/Lynx/Puma. Появится диалоговое окно, позволяющее установить предварительные значения (priors) для времени существования общего предка на основании имеющихся палеонтологических данных.

---

[22] Неправильная форма распределения приведет к замедлению расчетов в случае, если оптимальное значение параметра будет близко к краю распределения. То же относится и к слишком широкому диапазону предполагаемой вариации. С другой стороны, задавая более узкие пределы и «острые» распределения, исследователь рискует промахнуться мимо оптимума. Об этом можно судить по форме *постериорных* распределений, о чём речь пойдёт ниже.

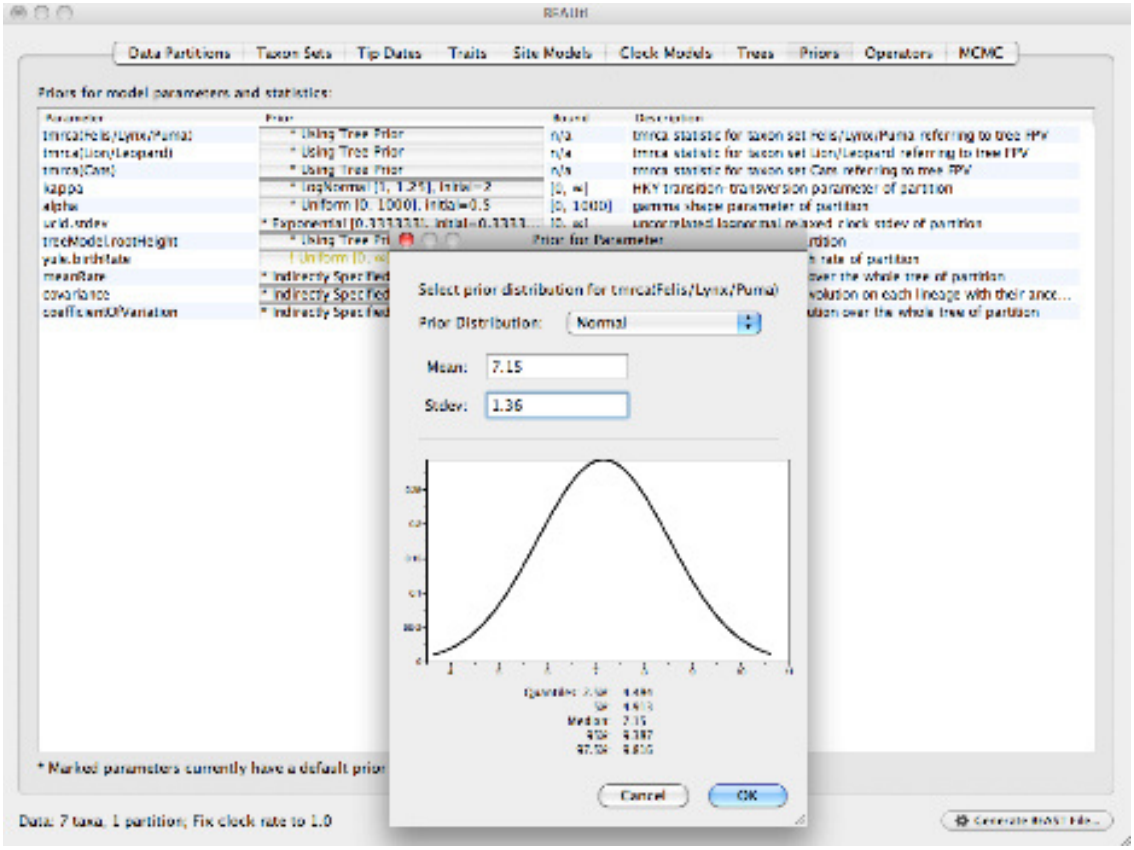


Рис. 10.6. Выбор формы распределения скоростей, из которого будут случайно выбираться значения

На рис. 10.6 показано окно, в котором изображено распределение t_{MRCA} для кошачьих. Его центр (математическое ожидание) приходится на 7,15 млн лет, а стандартное отклонение равно 1,36 млн лет. Эти параметры подобраны так, чтобы 95 % распределения укладывалось в интервал от 4,5 до 9,8 млн лет.

- Аналогично определяем предварительные условия для второго отмеченного ранее узла – дивергенции льва и леопарда. Время этого события палеонтологи оценивают в 3,72±1,05 млн лет.
- Последний параметр, который предстоит определить и *который не имеет присваиваемого автоматически значения,* – это степень изменчивости скорости молекулярной эволюции в пределах дерева. Этот параметр называется **uld.mean**, и в рамках рассматриваемого предела предлагается установить для него довольно широкие пределы варьирования и равномерное распределение[23]. На этом этапе именно так и следует поступать,

[23] Обратите внимание, что в BEAUTi во многих случаях по умолчанию для переменных предлагается по умолчанию равномерное распределение с предела-

поскольку обычно нет никаких предварительных данных, которые помогли бы оценить ожидаемый диапазон скоростей молекулярной эволюции.

В итоге вкладка priors должна выглядеть как на рис. 10.7.

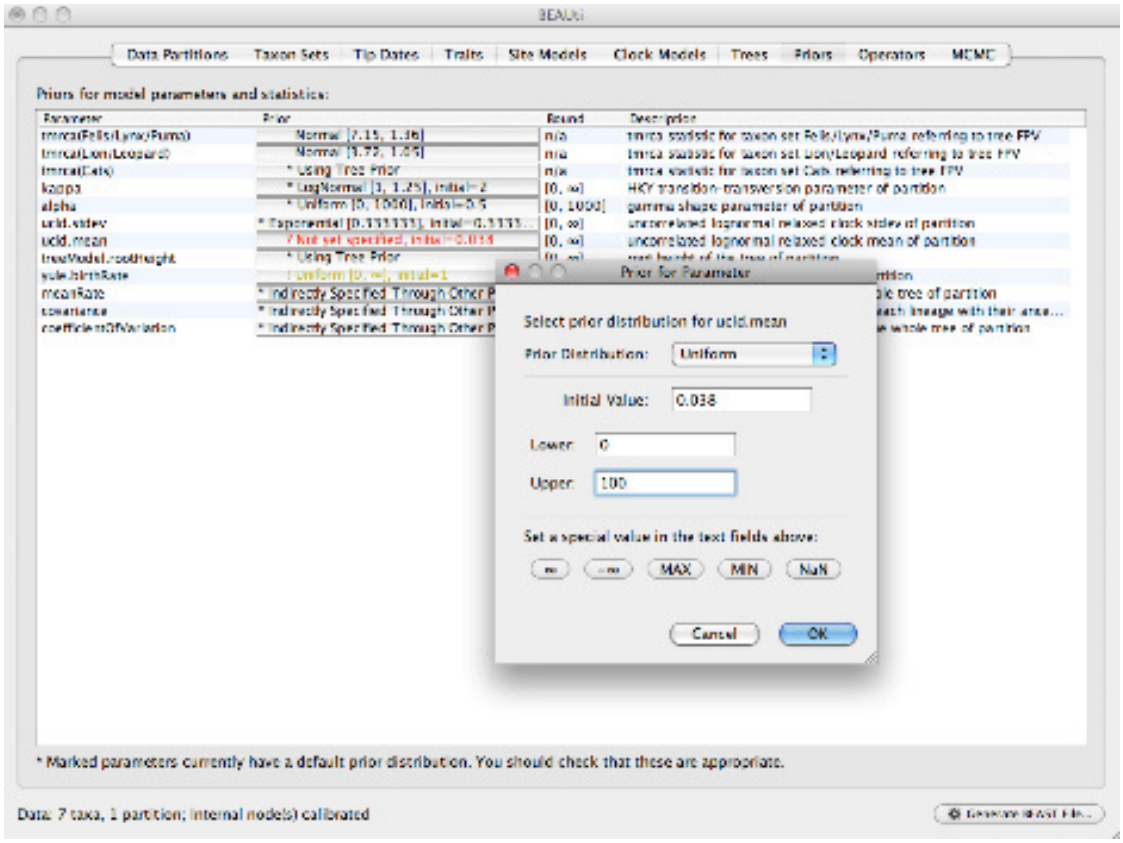


Рис. 10.7. Обзор предварительных установок параметров анализа

4. Определение параметров MCMC

Игнорируйте вкладку Operators, там содержатся технические установки, влияющие на производительность программы MCMC.

Следующая вкладка MCMC обеспечивает более общие настройки для контроля длины MCMC и имен файлов (рис. 10.8). Первое, у нас есть Length of chain. Это количество шагов MCMC, которое будет сделано в цепочке до окончания процесса. Насколько длинной эта цепочка может быть, зависит от размера набора данных, сложности модели и качества ожидаемого

ми $(-\infty, +\infty)$ или, например, $(0, +\infty)$. Это очевидно нереальные предположения, их не сразу удается заметить, но это одна из основных причин, по которой «ничего не работает не понятно почему. Я вроде все правильно делаю!».

ответа. Значение по умолчанию 10 000 000 является совершенно произвольным и должно быть установлено в соответствии с размером из набора данных. Для этого набора данных давайте сначала установим длину цепи до 800 000, так как это будет работать достаточно быстро на большинстве современных компьютеров (несколько минут).

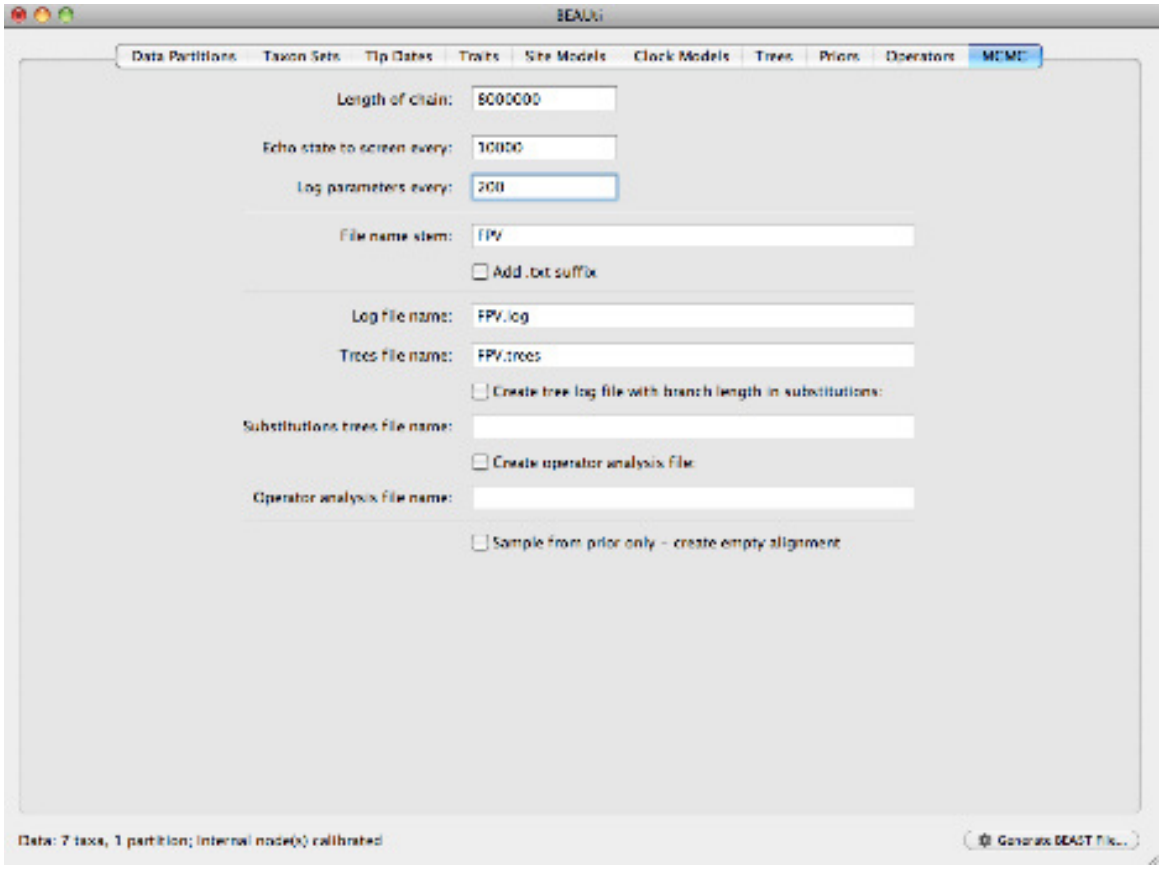


Рис. 10.8. Определение длительности анализа (число шагов)

Следующие параметры определяют, как часто значения параметров марковских цепей будут отображаться на экране, и записываются в log-файл. Этот экран предназначен просто для отслеживания прогресса программы и может быть установлен на любое значение. Для log-файла значение должно быть установлено относительно общей длины цепи. Слишком частые выборки приводят к созданию очень больших файлов с небольшим дополнительным преимуществом с точки зрения точности анализа. Слишком редкие выборки приведут к тому, что выходной файл будет содержать слишком мало информации для построения достаточно гладких кривых постериорных распреде-

лений (см. ниже). Разумный размер выборки составляет не менее 10 000 образцов. Соответственно, частоту выборки следует установить как одну десятитысячную от длины цепи.

Для этого упражнения мы установим 10 000 для журнала экрана и 200 для log-файла. Последние две опции дают имена файлов и log-файлов для выбранных параметров и деревьев. Они будут установлены как значение по умолчанию на основе имени импортируемого NEXUS файла.

5. Создание файла BEAST XML

Теперь мы готовы создать файл BEAST XML. Чтобы сделать это, выберите либо опцию Generate BEAST File... из меню File либо щелкните мышью на такую же кнопку внизу окна. Сохраните файл с подходящим именем, например FPV.xml. Теперь мы готовы к запуску BEAST.

Запуск BEAST

Теперь запускаем BEAST с входным файлом XML, который мы создали.

BEAST запустится, когда закончит выводить отчет на экран. Файлы с результатами сохранятся на диск в то же самое место, где находится исходный файл. Отчет о работе будет выглядеть следующим образом:

```
BEAST v1.6.1, 2002-2010
Bayesian Evolutionary Analysis Sampling Trees
Designed and developed by
Alexei J. Drummond, Andrew Rambaut and Marc A. Suchard
Department of Computer Science
University of Auckland
alexei@cs.auckland.ac.nz
Institute of Evolutionary Biology
University of Edinburgh
a.rambaut@ed.ac.uk
David Geffen School of Medicine
University of California, Los Angeles
msuchard@ucla.edu
Downloads, Help & Resources:
http://beast.bio.ed.ac.uk
Source code distributed under the GNU Lesser General Public Li-
cense:
http://code.google.com/p/beast-mcmc
BEAST developers:
Alex Alekseyenko, Erik Bloomquist, Joseph Heled, Sebastian Hoehna,
Philippe Lemey, Wai Lok Sibon Li, Gerton Lunter, Sidney Markowitz,
Vladimir Minin, Michael Defoin Platel, Oliver Pybus, Chieh-Hsi Wu,
Walter Xie
```

```
Thanks to:
Roald Forsberg, Beth Shapiro and Korbinian Strimmer

Random number seed: 1312153611069
Parsing XML file: FPV.xml
File encoding: MacRoman
Read alignment: alignment
Sequences = 7
Sites = 1434
Datatype = nucleotide
Site patterns 'patterns' created from positions 1-1434 of alignment
'alignment'
pattern count = 535
Using Yule prior on tree
Creating the tree model, 'treeModel'
initial tree topology = ((((FelisPV1,PumaPV1),AsianLionPV1),
(LynxPV1,SnowLeopardPV1)),(CanineOralPV,RacoonPV1))
tree height = 132.97768408757648
Using discretized relaxed clock model.
over sampling = 1
parametric model = logNormalDistributionModel
rate categories = 12
Creating state frequencies model: Using empirical frequencies from
data = {0.25455, 0.24404, 0.24965, 0.25175}
Creating HKY substitution model. Initial kappa = 2.0
Creating site model.
4 category discrete gamma with initial shape = 0.5
TreeLikelihood(treeModel) using native nucleotide likelihood core
Ignoring ambiguities in tree likelihood.
With 535 unique site patterns.
Branch rate model used: discretizedBranchRates
Creating swap operator for parameter branchRates.categories
(weight=10.0)
Creating the MCMC chain:
chainLength=8000000
autoOptimize=true
autoOptimize delayed for 80000 steps
\# BEAST v1.6.1, Build r3651
\# Generated Mon Aug 01 11:07:43 NZST 2011 [seed=1312153611069]

state   Posterior    Prior         Likelihood   rootHeight   ucld.mean
0       -13903.4023  -4438.3864 -9465.0159  132.978  3.8E-2 -
10000   -8124.7719   -27.2840   -8097.4879  21.9550   2.47017E-2 -
20000    -8126.6850  -27.4416   -8099.2434   19.9261 4.85395E-2 0.04
hours/
million
states   30000  -8124.4149  -26.8353  -8097.5796  20.2383  3.45595E-2
0.03 hours/million states
7980000  -8129.3648  -26.3189  -8103.0459  22.3010  3.53831E-2  0.03
hours/million states
7990000  -8128.8040  -28.2878  -8100.5162  24.6816  3.25836E-2  0.03
hours/million states
8000000  -8126.1647  -27.8020  -8098.3627  20.2114  2.48247E-2  0.03
hours/million states
```

```
Operator analysis
Operator                Tuning   Count    Time Time/Op Pr(accept)Performance
                                                                 suggestion
scale(kappa)             0.646    7104    1323   0.19   0.2466   good
scale(alpha)             0.675    7183    1258   0.18   0.2557   good
scale(ucld.mean)         0.706    213313  38354  0.18   0.2318   good
scale(ucld.stdev)        0.239    213711  38168  0.18   0.3243   good
subtreeSlide
treeModel)               2.525   1068995  114848 0.11   0.2676   good
Narrow Exchange(treeModel)       1066542  104496 0.1    0.0739   good
Wide Exchange(treeModel)         213823   14185  0.07   0.0112   good
wilsonBalding(treeModel)         213626   17113  0.08   0.0037   low
scale(treeModel.
rootHeight)               0.7    213764   9532   0.04   0.2643   good
uniform(nodeHeights
treeModel))                      2136378  270540 0.13   0.1846   good
scale(yule.birthRate)   0.216 213420   649    0.0    0.275    good

up:ucld.mean down:nodeHeights(treeModel)
                0.402   213885 38623  0.18   0.2368  slightly high
Try setting scaleFactor to about 0.3973
swapOperator(branchRates.categories)
                712956  100235  0.14  0.6203  high  No suggestions
randomWalkInteger(branchRates.categories)
                711666  83686  0.12  0.8843   high  Try increasing
windowSize uniformInteger(branchRates.categories)
                713634   83246 0.12  0.7204   high
                16.47148333333333 minutes
```

Анализ результатов

Прежде чем приступить к анализу результатов, желательно определить, стоит ли вообще это делать. Вполне может оказаться, что результат получился крайне низкого качества и в лучшем случае потребуется опять использовать BEAST с другими параметрами – как с измененными предварительными условиями, так и с измененными характеристиками MCMC. В худшем же случае может оказаться, что с помощью имеющихся в нашем распоряжении данных просто невозможно получить достоверный ответ и следует вернуться в «мокрую» лабораторию и получить больше нуклеотидных последовательностей или вообще поискать другой молекулярный маркер.

Запускаем программу Tracer, чтобы проанализировать файл после работы BEAST. Когда откроется основное окно программы, выбираем Import Trace File... из меню File и открываем файл, сделанный BEAST, который называется FPV.log.txt. Вы увидите такое окно, изображенное на рис. 10.9.

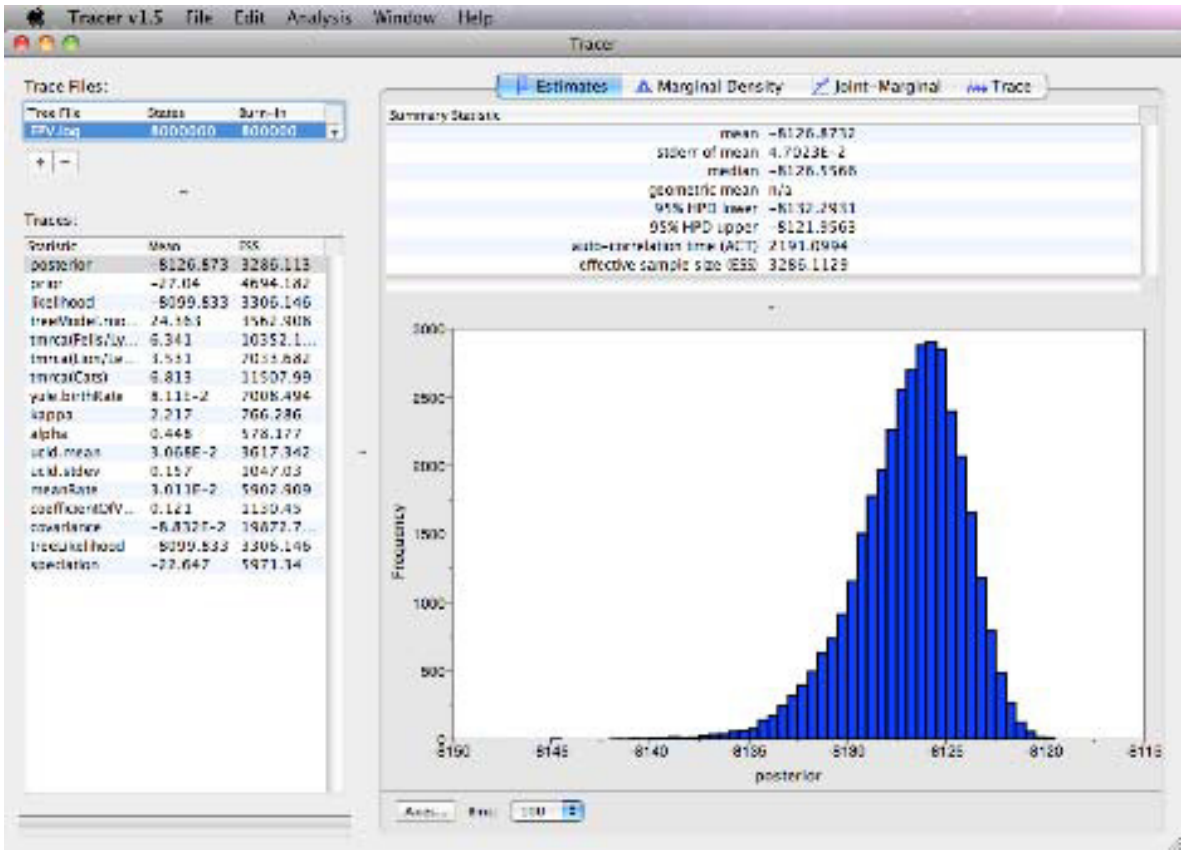


Рис. 10.9. Вид вкладки параметров MCMC

Помните, что MCMC – стохастический алгоритм, так что реальные результаты вычислений будут отличаться от одного запуска программы к другому. Для того чтобы получать абсолютно воспроизводимые результаты, требуется инициировать датчик случайных величин одним и тем же числом. Интерфейс программы позволяет это сделать.

Оценить качество анализа (и данных!) помогают два связанных друг с другом параметра – автрокорреляция (autocorrelation) и эффективный размер выборки (ESS, effective sample size). Для того чтобы понять, что это такое, вспомним, что такое MCMC.

С левой стороны есть список различных величин, которые вычислил BEAST. Каждой из этих величин соответствует предварительное условие, которое было определено во входном файле. В процессе работы программы происходит оптимизация значения этих параметров и генерируются так называемые постериорные распределения, которые можно видеть в правом окне. В тех случаях, когда предварительные распределения были заданы правильно, постериорное распределение находится в

середине окна и превращается в ноль и справа и слева. В случае если предварительный интервал был определен неправильно, постериорное распределение может совсем не иметь максимума или упираться в один из краев окна. Обычно после выполнения программы на экране появляется диагностика, содержащая рекомендации изменений, которые следует сделать в предварительных условиях.

Контрольные вопросы

1. Перечислите типы «расслабленных» молекулярных часов.
2. Опишите преимущества использования палеонтологической (исторической) информации при филогенетическом анализе.

11. ИССЛЕДОВАНИЕ ПРОСТРАНСТВЕННОЙ СТРУКТУРЫ ВИДОВ

В рамках исследований биоразнообразия важнейшее значение имеет выявление генетической структурированности организмов определенного вида в пространстве. Другими словами, важно знать, каким образом исследуемый вид разбит на популяции или другие подвидовые единицы. Границы между популяциями в большинстве случаев определяются географическими барьерами, как правило, затрудняющими, но не полностью исключающими миграцию и, соответственно – обмен генами. На протяжении большей части XX в. методы генетики популяций позволяли анализировать только пары популяций с потоком генов, близким к симметричному. Появление коалесцентной теории и «быстрых» генетических маркеров в последние десятилетия позволило перейти к анализу более сложных систем популяций с асимметричными потоками генов. Такая возможность оказалась исключительно полезной в практическом смысле, например, при анализе естественной подразделенности промысловых рыб, нарушение которой может приводить к отрицательным последствиям вроде общего снижения жизнеспособности.

Поиск оптимального количества популяций или генетических кластеров, соответствующего характеру наблюдаемой изменчивости, обычно производится с использованием байесовского метода целым рядом программ: Geneland, TESS, BAPPS, Structure, Structurama. Несмотря на очевидное методологическое единство, перечисленные программы имеют и некоторые отличия, касающиеся допущений моделей и вычислительных стратегий. Наибольшей популярностью у исследователей пользуются Geneland, позволяющая совместное использование генетических и географических сведений, и Structure, использующая только генетические данные.

Geneland выпускается как надстройка или дополнение к бесплатной статистической среде программирования R и имеет удобный пользовательский интерфейс. Главной ее целью является детекция популяционной структуры на основании вариации частот аллелей, которая может быть выражена в отклонении от равновесия Харди – Вайнберга для диплоидных маркеров. Программа может использовать несколько моделей, а также может объединять генетическую, географическую и фенотипическую информацию для расчета количества популяций и очерчивания их пространственного распределения. Возможно применение различных типов ДНК-маркеров: полиморфизм длин рестрикционных фрагментов (ПДРФ, микросателлитов, наборов однонуклеотидных полиморфизмов или даже сведений о полиморфизме нуклеотидных последовательностей митохондриальной ДНК или НК вирусов.

Кроме того, в программе доступны две модели: коррелирующих и некоррелирующих аллельных частот. В основе первой лежит некоторое упрощение: предполагается, что аллели, редкие для одной из популяций, также редко будут встречаться и в других. Показано, что данная модель частот позволяет выявлять более тонкие генетические отличия, но в то же время более чувствительна к отклонениям от ожидаемых распределений, например, в результате «изоляции дистанциями», и, таким образом, подходит лучше для видов с низкой внутривидовой вариабельностью. Рекомендуется сначала использовать модель без корреляции, а затем сравнивать результаты. В основе первой лежит некоторое упрощение: предполагается, что аллели, редкие для одной из популяций, также редко будут встречаться и в других. Показано, что данная модель частот позволяет выявлять более тонкие генетические отличия, но в то же время более чувствительна к отклонениям от ожидаемых распределений, например, в результате «изоляции дистанциями», и, таким образом, подходит лучше для видов с низкой внутривидовой вариабельностью. Рекомендуется сначала использовать модель без корреляции, а затем сравнивать результаты.

Запуск программы производится в консоли языка R с помощью команды:

```
Geneland.GUI()
```

Появится интерактивное окно, в котором можно добавлять требующиеся для анализа данные (рис. 11.1). В самом простом случае необходимы 3 файла в формате .txt: файл с именами анализируемых организмов (все имена в одном столбце), файл с координатами для каждого образца (2 столбца с десятичными широтой и долготой), а также файл с генетической информацией.

При использовании гаплоидных маркеров, таких как последовательности митохондриального генома, все вариабельные сайты записываются как целые числа в строку для каждого организма:

```
3 2 4 1 3 4 2 2
3 2 4 4 3 4 2 2 и тд.
```

Для диплоидных маркеров (например, набор однонуклеотидных полиморфных сайтов) последовательность генетических признаков записывается в следующем виде:

```
2 2 4 4 3 4 2 2
2 2 4 4 4 4 2 2 и тд.
```

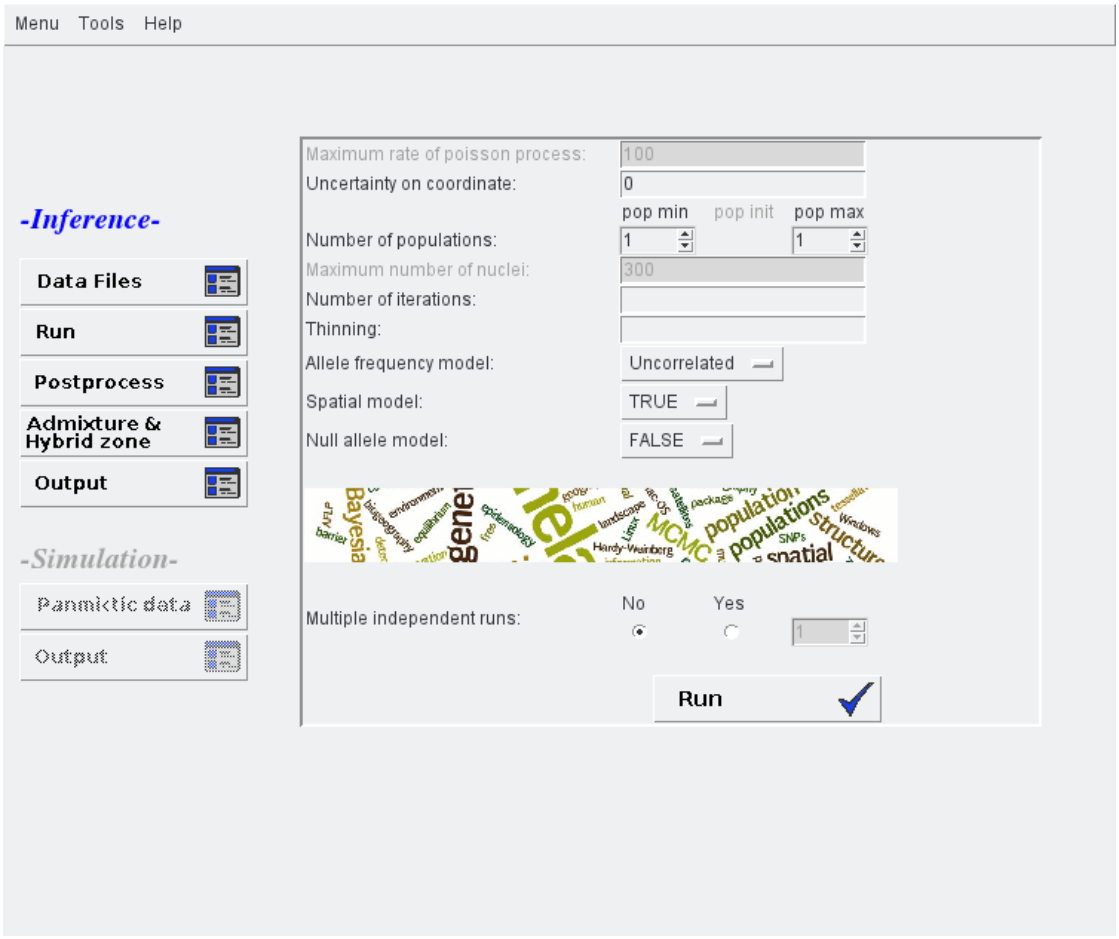


Рис. 11.1. Стартовое окно Geneland

Переходим на следующее окно для задания параметров байесовского анализа (рис. 11.2):

1) неточность определения координат (обычно задается для активно передвигающихся животных);

2) количество тестируемых популяций;

3) количество итераций MCMC (длина цепи, например, 50 000);

4) прореживание (thinning) (например, сохраняется только каждая 50-я итерация);

5) количество повторностей проведения анализа (независимых прогонов);

6) флажок использования географических данных.

В Geneland возможно использование как модели с географическими корреляциями, так и без них. Сравнение результатов, полученных обоими методами, также несет важную информацию о популяционной и пространственной структуре вида. Важнейшим допущением в географической модели является то, что особи, собранные в одной точке, обязательно должны быть отнесены к одному кластеру (исключение составляют только случаи, когда мы задаем неточность определения координат и, таким образом, можем предполагать присутствие мигрантов).

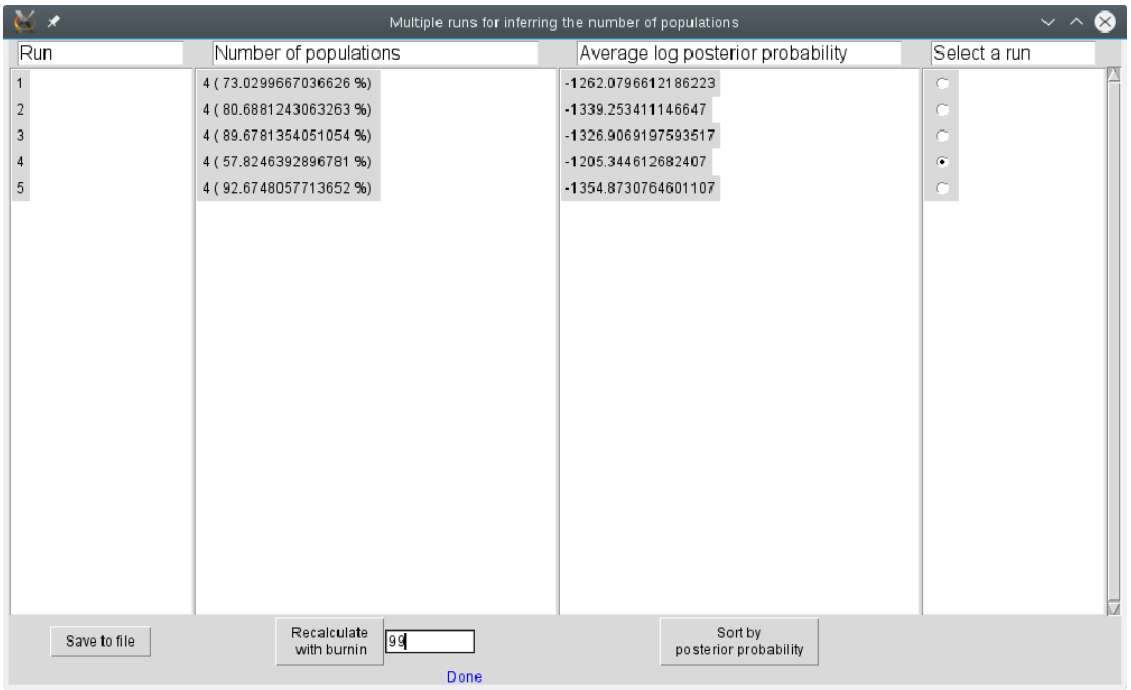


Рис. 11.2. Окно вывода статистики по цепям. Выбор лучшего прогона

На начальном этапе рекомендуется задавать, например, 10 «ранов» с небольшим количеством итераций (50 000–100 000). При запуске анализа появится окно, где будет выводиться основная статистика по цепям, здесь же возможно пересчитать апостериорные вероятности с учетом отбрасываемых первых итераций (около 25 %) с помощью флажка параметра burn in. Если набор данных неоднозначен или организмы имеют сложную иерархическую популяционную структуру, то цепи могут отличаться по количеству предсказанных кластеров, в таком случае выбирают обычно цепь с наибольшим средним значением апостериорных вероятностей. Статистика по разным ранам позволяет определить, какое же количество генетических кластеров лучше соответствует наблюдаемому полиморфизму.

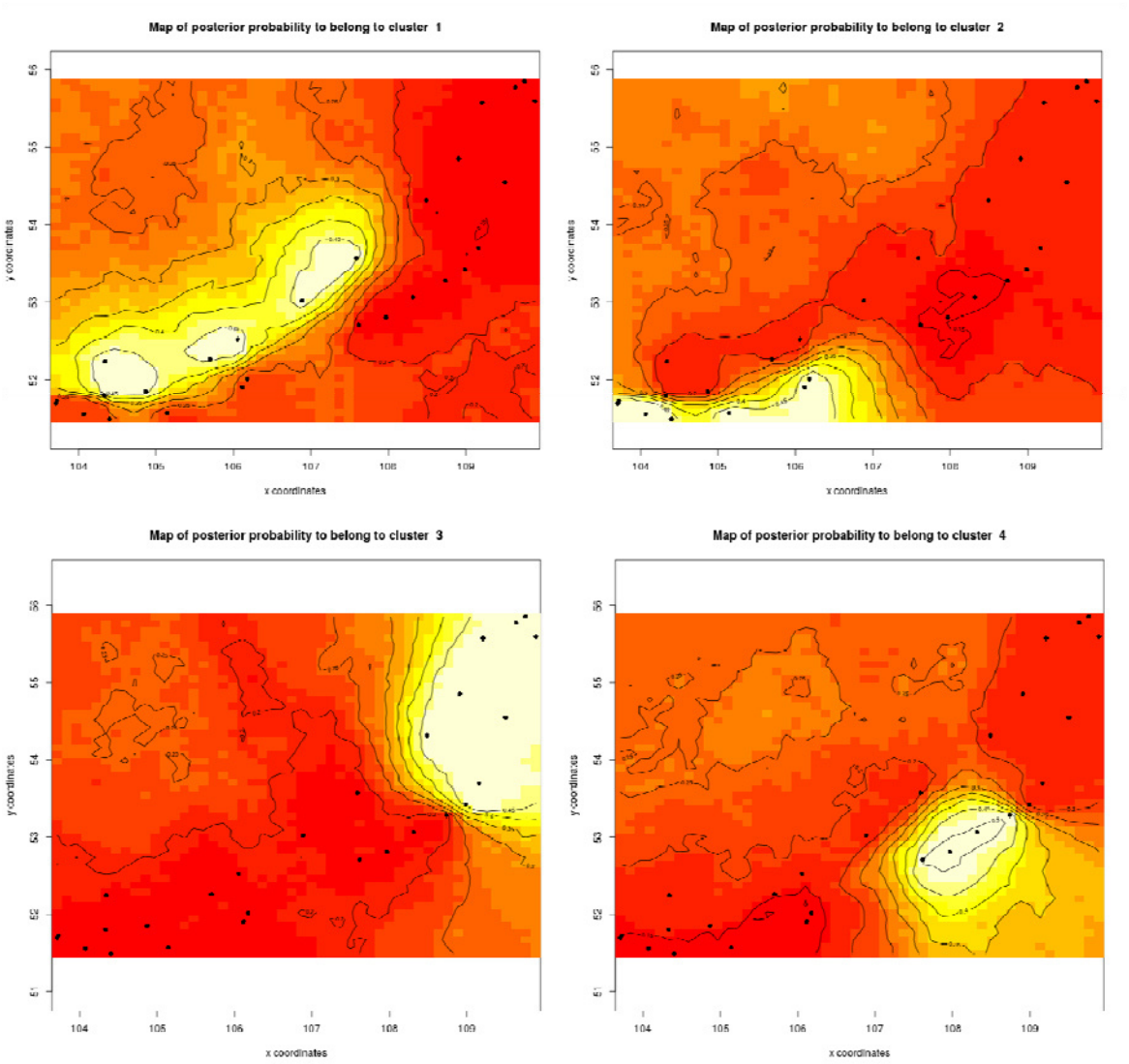
После того как расчеты сделаны и записаны в отдельные файлы формата .txt в изначально заданную директорию, можно задать параметры пост-процессинга: разрешение пространственной модели в пикселах (например, 50×50) и количество неучитываемых первых итераций.

На данном этапе возможно получить и сравнить как табличную, так и графическую информацию по каждой из цепей, доступную на вкладке Output.

Затем можно задать определенное ранее количество популяций для более длительного анализа, в котором более точно будут определены апостериорные вероятности принадлежности каждого организма и каждой точки сбора к каждому из кластеров.

После этого можно визуализировать всю соответствующую данному рану статистику, карты с апостериорными вероятностями для каждого пиксела, просмотреть таблицы значений Fst и вероятностей принадлежности к каждому из кластеров для каждого организма. Визуализация пространственных структур осуществляется с помощью статистической модели разбиения Вороного. Предполагается, что каждый кластер географически может быть изображен объединением нескольких полигонов на плоскости. Точки, которые станут центрами полигонов, генерируются как случайные пуассоновские величины. Полигоны окрашиваются в соответствии с апостериорной вероятностью их принадлежности к каждому из кластеров.

На рис. 11.3 изображен пример полученных карт популяций. Изолинии содержат подписи со значениями вероятностей.



11.3. Карта принадлежности к популяциям

Контрольный вопрос

Методы использования географической информации при анализе биоразнообразия и разбиения ареалов видов на ареалы популяций.

ЗАКЛЮЧЕНИЕ

В заключение следует призвать читателя/пользователя этого пособия к осторожности. Современная популяционная и экологическая молекулярная генетика использует колоссальный набор методов биоинформатики. Списки программ, посвященных различным разделам этой дисциплины, содержат сотни наименований. Более того, как и в любой быстро развивающейся отрасли науки, на стыке эволюционной теории, экологии и биоинформатики происходит бурное развитие – появление и исчезновение новых систем взглядов и концепций, зачастую противоречащих друг другу и далеко не всегда – очевидным образом. Следует помнить, что авторы программ – живые люди, которые формализуют и превращают в код сложные и спорные идеи, и пользователь постоянно должен отдавать себе отчет в том, что каждая программа, каждый алгоритм построены на целом наборе предположений, которые опираются на биологический смысл, каким понимают его авторы. Использование воплощенных в программы чужих идей и взглядов, таким образом, таит в себе многие опасности, но с другой стороны – создает массу возможностей и делает работу в этой области биологии интересной совершенно по-особому.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

Гладков Л. Генетические алгоритмы / Л. Гладков, В. Курейчик. – 2-е изд., испр. и доп. – М. : Физматлит, 2006. – 320 с.

Ли Ч. Введение в популяционную генетику / Ч. Ли. – М. : Мир, 1978. – 560 с.

Мандель Б. Р. Основы современной генетики / Б. Р. Мандель. – М. ; Бердин : Директ-Медиа, 2016. – 334 с.

Павлинов И. Введение в современную филогенетику / И. Павлинов. – М. : Litres, 2018. – 51 с.

Павлинов И. Я. Кладистический анализ (методологические проблемы) / И. Павлинов. – М. : МГУ, 1990. – 160 с.

Павлинов И. Я. Систематика современных млекопитающих / И. Павлинов. – М. : Изд. Моск. ун-та, 2003. – 297 с.

Характеристика генетического разнообразия медоносных пчел (Apis mellifera L.) Томской популяции по комплексу ДНК-маркеров / Н. В. Островерхова [и др.] // Чтения памяти Алексея Ивановича Куренцова. – Владивосток, 2015. – Т. 26. – С. 227–240.

Хедрик Ф. Генетика популяций : пер. с англ. / Ф. Хедрик. – М. : Техносфера, 2003. – 592 с.

Aerts D. The GTR-model: a universal framework for quantum-like measurements / D. Aerts, M. S. D. Bianchi // Probing the Meaning of Quantum Mechanics: Superpositions, Dynamics, Semantics and Identity. – 2016. – P. 91–140.

Clement M. TCS: a computer program to estimate gene genealogies / M. Clement, D. Posada, K. A. Crandall // Molecular ecology. – 2000. – Vol. 9, N 10 .– P. 1657–1659.

Coordinators N. R. Database resources of the national center for biotechnology information // Nucleic acids research. – 2016. – Vol. 44, N Database issue. – C. D7.

Drosophila Polymorphism Database (DPDB) A Portal for Nucleotide Polymorphism in Drosophila / S. Casillas [et al.] // Fly. – 2007. – Vol. 1, N 4. – P. 205–211.

Robust Estimation of Evolutionary Distances with Information Theory / M. D. Cao [et al.] // Molecular biology and evolution. – 2016. – Vol. 33, N 5. – C. 1349–1357.

Galtier N. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny / N. Galtier, M. Gouy, C. Gautier // Bioinformatics. – 1996. – Vol. 12, N 6. – P. 543-548.

Spectronet: a package for computing spectra and median networks / K. T. Huber [et al.] // Applied bioinformatics. – 2002. – Vol. 1, N 3. – P. 159–161.

Huson D. H. SplitsTree: analyzing and visualizing evolutionary data // Bioinformatics (Oxford, England). – 1998. – Vol. 14, N 1. – P. 68–73.

Identifying optimal models of evolution / L. S. Jermiin [et al.] // Bioinformatics. – New York : Humana Press, 2017. – P. 379–420.

Kloepper T. H. Drawing explicit phylogenetic networks and their integration into SplitsTree / T. H. Kloepper, D. H. Huson // BMC evolutionary biology. – 2008. – Vol. 8, N 1. – P. 22

Koski L. B. The closest BLAST hit is often not the nearest neighbor / L. B. Koski, G. B. Golding // J. of Molecular Evolution. – 2001. – P. 540–542.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Базыкин А. Д.* Нелинейная динамика взаимодействующих популяций / А. Д. Базыкин. – М. ; Ижевск : Ин-т компьютер. исслед., 2003; 368 с.

2. *Гладков Л.* Генетические алгоритмы / Л. Гладков, В. Курейчик. – 2-е изд., испр. и доп. – М. : Физматлит, 2006. – 320 с.

3. *Голубовский М. Д.* Век генетики: эволюция идей и понятий / М. Д. Голубовский. – СПб. : Борей Арт, 2000. – 262 с.

4. *Кайданов Л. З.* Генетика популяций / Л. З. Кайданов. – М. : Высш. шк., 1996. – 320 с.

5. *Картавцев Ю. Ф.* Молекулярная эволюция и популяционная генетика / Ю. Ф. Картавцев. – 2-е изд. – Владивосток : Изд-во Дальневост. гос. ун-та, 2008. – 25 печ. л.

6. *Кержнер И. М.* Прошлое, настоящее и будущее таксономии / И. М. Кержнер, Б. А. Коротяев // Рус. орнитол. журн. – 2015. – Т. 24, № 1190. – С. 10–19.

7. *Ли Ч.* Введение в популяционную генетику / Ч. Ли. – М. : Мир, 1978. – 560 с.

8. *Лукашов В. В.* Молекулярная эволюция и филогенетический анализ / В. В. Лукашов. – М.: Бином : Лаборатория знаний, 2009. – URL: http://molbiol.ru/forums/index.php?showtopic=361358

9. *Лухтанов В. А.* Принципы реконструкции филогенезов: признаки, модели эволюции и методы филогенетического анализа // Тр. Зоол. Ин-та РАН. Прил. – 2013. – № 2. – С. 39–52.

10. *Марков А.* Рождение сложности. Эволюционная биология сегодня: неожиданные открытия и новые вопросы [Электронный ресурс] / А. Марков. – URL: https://www.litmir.me/br/?b=182746&p=1 Litres.

11. *Ней М.* Молекулярная эволюция и филогенетика / М. Ней, С. Кумар. – Киев : КВІЦ, 2004. – 418 с.

12. *Павлинов И. Я.* Методики кладистики / И. Я. Павлинов. – М. : Изд-во МГУ, 1989. – 118 с.

13. *Хедрик Ф.* Генетика популяций : пер. с англ. / Ф. Хедрик. – М. : Техносфера, 2003. – 592 с.

14. *Шмальгаузен И. И.* Кибернетические вопросы биологии / И. И. Шмальгаузен. – М.: Книга по Требованию, 2012. – 223 с.

Учебное издание

Щербаков Дмитрий Юрьевич
Адельшин Ренат Викторович
Коваленкова Мария Владимировна

АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОВРЕМЕННОЙ ГЕНЕТИКИ

БИОИНФОРМАЦИОННЫЕ МЕТОДЫ АНАЛИЗА БИОРАЗНООБРАЗИЯ

ISBN 978-5-9624-1600-7

Редактор *А. Н. Шестакова*
Дизайн обложки: *П. О. Ершов*

Темплан 2018. Поз. 59
Подписано в печать 08.06.2018. Формат 60x90 1/16
Уч.-изд. л. 5,1. Усл. печ. л. 7,5. Тираж 100 экз. Заказ 82

ИЗДАТЕЛЬСТВО ИГУ
664074, г. Иркутск, ул. Лермонтова, 124