Методы изучения филогении прокариот

Учебное пособие

Орлова М.В.

Грабович М.Ю.

Рецензенты: Белоусова Елена Васильевна, к.б.н., биолог (клинико-диагностической лаборатории) бюджетного учреждения здравоохранения Воронежской области "Воронежская областная станция переливания крови"

Лавриненко Игорь Андреевич, к.б.н., доцент кафедры биофизики и биотехнологии Воронежского государственного университета

Место подготовки пособия: Воронежский государственный университет, медико-биологический факультет, кафедра биохимии и физиологии клетки

Точный читательский адрес: рекомендуется для магистрантов 2 курса, специальность 06.04.01 (020400) Биология

Оглавление

Введение

Развитие методов ПЦР и автоматического определения нуклеотидных последовательностей сделало возможным получение генетической информации в непредставимом раньше масштабе.

Одновременное развитие компьютерных технологий привело к тому, что анализ генетических последовательностей, бывший ранее темой работы незначительного числа специалистов, становится повседневной задачей многих научных и практических лабораторий (Лукашов, 2009). В связи с этим является актуальным изучение методов анализа генетической информации.

Также использование генетических данных имеет огромное значение в систематике прокариот. До появления данных о строении ДНК вся бактериальная систематика строилась в основном на морфологических и культуральных признаках. Однако эти признаки далеко не всегда точно отображают эволюционную историю. Использование молекулярных данных существенно расширяет возможности таксономистов и приводит к пересмотру многих систематических групп микроорганизмов.

В данный момент описание новых таксонов прокариот невозможно без молекулярных данных. Поэтому важно освоение основных методов молекулярной филогении в рамках университетского курса.

Глава 1. Данные, используемые в филогенетическом анализе

У всех живых систем, включая доклеточные и клеточные формы жизни, наследственная, или генетическая, информация содержится в геноме, представленном молекулами нуклеиновых кислот. У подавляющего большинства форм жизни генетическая информация передается от поколения к поколению в виде молекул дезоксирибонуклеиновых кислот (ДНК). Ряд вирусов является исключением из этого правила и передает генетическую ин-

формацию в виде рибонуклеиновых кислот (РНК)—РНК-содержащие вирусы.

Нуклеиновые кислоты – это полимерные молекулы, мономерами которых являются 5 различных нуклеотидов: пуриновые - аденин (A) и гуанин (G) - и пиримидиновые – тимин (T), цитозин (C) и урацил (U). Буквенные обозначения являются общепринятыми и используются во всех базах данных. Однако бывает так, что приходится работать с вырожденными последовательностями, когда не известно точно, какой нуклеотид находится в том или ином положении. В этом случае используют коды IUPAC для обозначения нуклеотидов (табл. 1).

Таблица 1.

Коды IUPAC для обозначения нуклеотидов

Код	Обозначает	Комплиментарный нуклеотид
A	A	T(U)
C	C	G
G	G	C
T(U)	T(U)	A
M	A или C	K
R	A или G	Y
W	A или T(U)	W
S	C или G	S
Y	C или T(U)	R
K	C или T(U)	M
V	A или C или G	B
H	A или C или T(U)	D
D	A или G или T (U)	H
B	C или G или T (U)	V
X или N	A или C или G или T (U)	X или N

В процессе репликации нуклеиновых кислот могут происходить ошибки. В результате дочерний геном будет отличаться от родительского генома. Ошибки, происходящие при репликации генома, называют мутациями (mutations). В филогенетическом анализе наибольшее значение имеют точечные мутации (point mutations), которые затрагивают только один или несколько соседних нуклеотидов. Точечные мутации разделяются на следующие виды:

- замена одного нуклеотида на другой—нуклеотидная замена (nucleotide substitution);
- вставка одного или более нуклеотидов (insertion). Частным случаем вставки является удвоение некоего генетического участка—дупликация (duplication);
- удаление одного или нескольких соседних нуклеотидов—делеция (deletion);
- поворот участка нуклеиновой кислоты длиной минимум в два нуклеотида на 180°—инверсия (inversion);
- считывание дочерней молекулы нуклеиновой кислоты не с одной, а с двух и более родительских молекул—рекомбинация (recombination);
- транзиции - замены одного пурина на другой пурин (A→G, G→A) или одного пиримидина на другой пиримидин (C→T, T→C);
- трансверсии - замены между пуринами и пиримидинами (A→T, A→C, G→T, G→C, T→A, T→G, C→A, C→G) (Лукашов, 2009).

Глава 2. Базы данных нуклеотидных и аминокислотных последовательностей

Филогенетические анализы часто основаны на данных, накопленных многими исследователи. Столкнувшись с быстрым увеличением количества доступных последовательностей, невозможно полагаться на печатную литературу. Поэтому ученым пришлось обратиться к оцифрованным базам

данных. Базы данных необходимы в текущих биоинформационных исследованиях, так как они служат местом хранения и поиска информации. Существуют современные базы данных, имеющие мощные инструменты запросов и перекрестные ссылки с другими базами данных. Помимо последовательностей и инструментов поиска базы данных содержат значительное количество сопроводительной информации, так называемой аннотации. К сопроводительной информации относится название организма и типа клеток, из которых была получена последовательность, каким методом она была секвенирована, что за свойства уже известны и т. д. В этой главе мы рассмотрим наиболее важные общедоступные базы данных последовательностей. Список интернет-адресов базы данных, обсуждаемых в этом разделе, приведен в таблице 2.

Для поиска последовательностей баз данных существуют три различные стратегии.

- Чтобы легко получить известную последовательность, можно положиться на уникальные идентификаторы последовательности;
- Собирать полный набор последовательностей, которые разделяют таксономическое происхождение или известное свойство, можно по ключевому слову в аннотации;
- Чтобы найти наиболее полный набор гомологичных последовательностей, можно использовать поиск по подобию с другими последовательностями с использованием таких инструментов как BLAST или FASTA.

Таблица 2.

Интернет-адреса основных баз данных последовательностей и инструментов поиска по базам данных

Название базы данных или инструмента поиска	Интернет-адрес
ACNUC	*http://pbil.univ-lyon1.fr/databases/acnuc/acnuc.html*
BioXL/H	*http://www.biocceleration.com/BioXLH-technical.html*
BLAST	*http://www.ncbi.nlm.nih.gov/blast/*
DDBJ и **DAD**	*http://www.ddbj.nig.ac.jp/*
EMBL	*http://www.ebi.ac.uk/embl/*
EMBL Sequence Version Archive	*http://www.ebi.ac.uk/cgi-bin/sva/sva.pl*
Ensembl	*http://www.ensembl.org/*
Entrez	*http://www.ncbi.nlm.nih.gov/Entrez/*
fastA	*http://fasta.bioch.virginia.edu/fasta/*
GenBank	*http://www.ncbi.nlm.nih.gov/Genbank/*
Gene Ontology	*http://www.geneontology.org/*
HAMAP	*http://www.expasy.org/sprot/hamap/*
HCV database	*http://hcv.lanl.gov/*
HIV database	*http://hiv-web.lanl.gov/*
HOGENOM	*http://pbil.univ-lyon1.fr/databases/hogenom.html*
HOVERGEN	*http://pbil.univ-lyon1.fr/databases/hovergen.html*
IMGT/HLA	*http://www.ebi.ac.uk/imgt/hla/*
IMGT/LIGM	*http://imgt.cines.fr/*
MPsrch, **Scan-**	*http://www.ebi.ac.uk/Tools/ similarity.html*

Название базы данных или инструмента поиска	Интернет-адрес
PS, **WU-BLAST** и **fastA** в **EBI**	
MRS	*http://mrs.cmbi.ru.nl/mrs-3/*
NCBIMap Viewer	*http://www.ncbi.nlm.nih.gov/mapview/*
ORALGEN	*http://www.oralgen.lanl.gov/*
PDB	*http://www.rcsb.org/*
PRF/SEQDB	*http://www.prf.or.jp/*
RefSeq	*http://www.ncbi.nlm.nih.gov/RefSeq/*
Sequin	*http://www.ncbi.nlm.nih.gov/Sequin/*
SRS	*http://www.biowisdom.com/navigation/srs/srs*
сервер **SRS** в **EBI**:	*http://srs.ebi.ac.uk/*
Список публичных серверов SRS	*http://downloads.biowisdomsrs.com/publicsrs.html*
Taxonomy	*http://www.ncbi.nlm.nih.gov/Taxonomy/*
TIGR	*http://www.tigr.org/*
Геномный браузер UCSC	*http://genome.ucsc.edu/*
UniGene	*http://www.ncbi.nlm.nih.gov/UniGene/*
UniProt в **EMBL**	*http://www.ebi.ac.uk/uniprot/*

Название базы данных или инструмента поиска	Интернет-адрес
UniProt в **SIB**	*http://www.expasy.uniprot.org/*
VAST	*http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html*

2.1 Общие базы данных нуклеиновых кислот

В Европе, США и Японии предпринимаются параллельные усилия по поддержанию публичных баз данных со всеми опубликованными последовательностями нуклеиновых кислот:

- База данных EMBL (Европейская лаборатория молекулярной биологии), поддерживаемая в EMBL-EBI (Европейский институт биоинформатики, Хинкстон, Англия, Великобритания).
- GenBank, поддерживаемый NCBI (Национальный центр биотехнологической информации, Бетесда, Мэриленд, США).
- DDBJ (Банк данных ДНК Японии), поддерживаемый NIG / CIB (Национальный институт Генетики, Центр Информационной Биологии, Мишима, Япония).

В начале 1980-х годов кураторы баз данных сканировали печатную литературу по новым последовательностям, но сегодня эти последовательности размещаются авторами через инструменты представления World Wide Web (или по электронной почте после подготовки данных с использованием программного обеспечения Sequin). Существует соглашение между кураторами из трех основных баз данных о перекрестном представлении последовательностей друг другу.

Базы данных содержат как последовательности РНК, так и последовательности ДНК, но, по соглашению, последовательность всегда записывается как ДНК, то есть с T, а не с U. Часто, но не всегда, программное обеспе-

чение для анализа последовательностей обрабатывает U и T, не делая различия (Lemey et al., 2009).

У баз данных есть понятие выпуска – “release”. С регулярными интервалами (2 месяца для GenBank, 3 месяца для двух других) база данных «заморожена» в текущем состоянии до «release». Последовательности, которые отправляются с момента последнего выпуска, доступны как «ежедневные обновления». На протяжении многих лет количество последовательностей в базах данных росло с огромной скоростью (рис. 1).

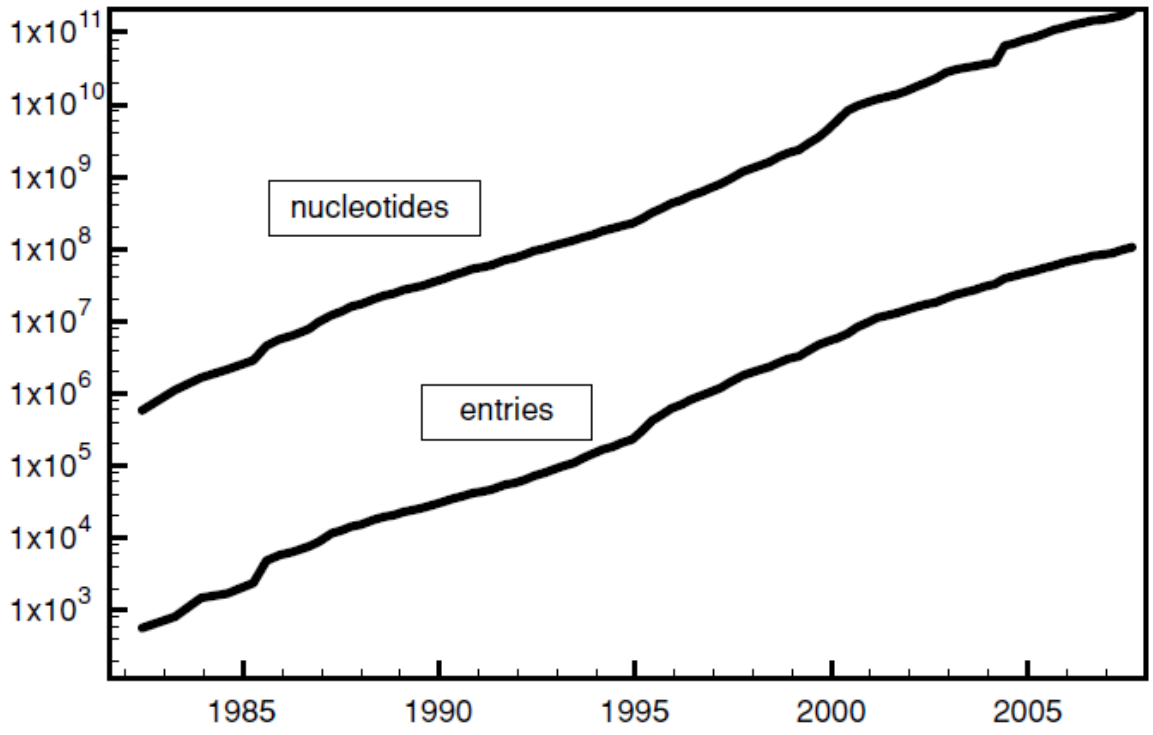

Рис. 1. Рост (по логарифмической шкале) количества нуклеотидов и записей в публичной базе данных EMBL (без избытка) с июня 1982 года по сентябрь 2007 года.

Каждая последовательность имеет ряд уникальных идентификаторов (см. рис. 2), которые позволяют легко осуществить поиск:

```
LOCUS HUMBMYH7 28452 bp DNA linear PRI 30-OCT-2002
DEFINITION Homo sapiens beta-myosin heavy chain (MYH7)
gene, complete cds.
ACCESSION M57965 M30603 M30604 M30605 M57747 M57748
M57749
VERSION M57965.2 GI:24429600
```

Рис. 2. Первые строки записи GenBank. Обратите внимание на разные идентификаторы: имя записи HUMBMYH7, номер первичного доступа M57965 (за которым следуют 6 номеров вторичного доступа), номер версии M57965.2 и номер GI.

2.1.1. Имя записи, имя локуса или идентификатор (ID)

Идентификатор был первоначально разработан как мнемонический (например, ECARGS для гена *ArgS E. coli*), но из-за с быстрого накопления анонимных фрагментов многие последовательности теперь имеют ID, который просто идентичен их AC. ID может измениться от выпуска к выпуску (из-за появления новой информации о том, что представляет собой последовательность) и различаться в трех разных базах данных. EMBL решила отказаться от использования ID с июня 2006 года, но GenBank и DDBJ все еще используют их.

2.1.2. Номер доступа (AC)

В отличие от ID, AC остается постоянным между версиями, и существует соглашение между менеджерами трех баз данных, чтобы давать одинаковый AC одинаковым последовательностям. Помимо своего первичного номера доступа, последовательность может иметь номера вторичного доступа, поскольку при объединении нескольких последовательностей в одну новая последовательность получает новый AC, но наследует все старые AC. Благодаря своему стабильному и универсальному характеру AC является наиболее полезным для поиска последовательности.

2.1.3. Номер версии

Номер версии состоит из AC и числа, которое увеличивается каждый раз, когда последовательность изменяется. Таким образом, номер версии полезен, чтобы найти последовательность, которая использовалась для конкретного исследования. Чтобы было возможно найти «старые последовательности», EBI создал «Архив версий последовательностей», в котором можно выполнять поиск по SV или AC. Однако, номер версии был введен в начале 1999 года и не охватывает период до этого.

2.1.4. Номер GenInfo (только GenBank)

Раньше каждой последовательности, обработанной NCBI, присваивался номер GI, который являлся уникальным для всех баз данных NCBI. В 2016 году было принято решение отказаться от GI, однако при работе со старыми статьями можно встретить ссылки на GI.

2.1.5. Полногеномные последовательности (WGS)

Это последовательности из проектов секвенирования геномов, которые проводились методом дробного секвенирования целого генома. У них нет стабильного номера доступа. Через регулярные интервалы времени все последовательности WGS для одного организма удаляются из базы данных и заменяются новым набором.

2.1.6. Сторонние аннотации (TPA)

Это последовательности с улучшенной документацией и / или последовательностями, построенными путем сборки перекрывающихся записей, представленные лицом, которое не является оригинальным автором последовательностей (Lemey et al., 2009).

2.2. Общие базы данных белковых последовательностей

Белки могут быть секвенированы с использованием классического метода деградации Эдмана, с использованием современных методов, основанных на пептидной масс-спектрометрии, или их последовательности мо-

гут быть выведены из трехмерной структуры, определенной рентгеновской кристаллографией или ЯМР. Однако основная часть записей в белковых базах данных получается путем секвенирования и трансляции кодирующей ДНК / РНК. При этом базы данных могут содержать и содержат ошибки (переводы открытых рамок считывания, которые не соответствуют реальным белкам, ошибки, связанные со сдвигом рамки считывания, ошибочно определенные границы интрон-экзон и т. д.).

В отличие от ситуации для нуклеиновых кислот, которые хранятся в трех идентичных базах данных, белки хранятся в базах данных, которые значительно различаются по содержанию и качеству. Основным источником информации является база данных UniProt, которая должна содержать все белковые последовательности, которые когда-либо были опубликованы, за исключением повторных вариантов одной и той же последовательности, коротких фрагментов и «сомнительных» последовательностей. Она содержит набор трансляций открытых рамок считывания, извлеченных из аннотации последовательностей EMBL, а также данные авторов, которые получили «реальную последовательность белка». Последовательности в UniProt сопровождаются замечательной аннотацией, которая включает перекрестные ссылки на другие базы данных и предварительную идентификацию мотивов на основе автоматизированного поиска по базам данных. UniProt состоит из двух частей: UniProt / SwissProt (поддерживается командой профессора Амоса Байроха в Женевском университете, Швейцария, в сотрудничестве с EMBL-EBI) и UniProt / TrEMBL (поддерживается в EMBL-EBI). TrEMBL – это компьютерная база данных с небольшим вмешательством человека, в то время как SwissProt курируется живыми людьми, которые добавляют информацию, извлеченную из литературы. В настоящее время трансляции из EMBL всегда в первую очередь помещают в TrEMBL, где им присваивают номер доступа. После того как их аннотации были проверены

и обработаны кураторами, они удаляются из TrEMBL и вводятся в SwissProt при сохранении их AC; авторские материалы напрямую отправляются в SwissProt. Последовательности SwissProt имеют идентификатор типа XXXXX YYYYY, где XXXXX обозначает природу белка и YYYYY - уникальный идентификатор организма (например, TAP_DROME для целевого белка Poxn *D. melanogaster*). Последовательности TrEMBL имеют предварительный идентификатор, состоящий из AC и идентификатора организма (Lemey et al., 2009).

EBI предоставляет следующие поисковые инструменты:

UniProt splice variants, полезен в случае, если анализируемый сиквенс является только вариативным, а не репрезентативным в SwissProt.

UniRef100, подмножество UniProt, выбранное таким образом, что никакая последовательность не идентична и не является субфрагментом другой последовательности; полезно для более быстрого поиска.

UniRef90 и UniRef50, еще более урезанное подмножество UniRef100, полученное путем исключения последовательностей короче 11 аминокислот и путем выбора последовательностей таким образом, что нет последовательности идентичной другим более чем на 90% и 50% соответственно

В дополнение к UniProt доступны следующие базы данных:

GenPept: коллекция открытых рамок считывания, извлеченных из GenBank, поддерживается в NCI-ABCC (Национальный институт рака, Передовой Биомедицинский Вычислительный Центр, Фредерик, Мэриленд, США).

DAD: сбор открытых рамок считывания, извлеченных из DDBJ, поддерживаемых в NIG / CIB.

PRF / SEQDB (Protein Resource Foundation), поддерживаемый в ПРФ (Осака, Япония), содержит как трансляции нуклеотидных последовательностей, так и отсеквенированные белки (Lemey et al., 2009).

2.3. Специализированные базы данных последовательностей, справочные базы данных и базы данных генома

В дополнение к общим базам данных существует большое количество (более 100) специализированных баз данных. Они могут быть построены на основе общих баз данных; могут принимать авторские материалы и / или генерироваться путем автоматической аннотации геномов. Они предлагают одно или несколько из следующих преимуществ:

- Данные формируют четко определенный набор последовательностей. Поиск в специализированных базах данных вместо общих может дать тот же результат, но за меньшее время и менее «загрязненный» фоновым шумом.
- База данных часто «очищается», содержит меньше ошибок и меньше избыточных записей.
- Иногда аннотация стандартизована, так что можно легко найти все нужные последовательности, не повторяя поиск с использованием альтернативных ключевых слов.
- Аннотация обычно является более качественной, чем в общих базах данных. Более того, некоторые базы данных содержат последовательности, уже сгруппированные в семейства и / или таблицы с данными, связанными с последовательностями, такие как геномные карты.

Ниже представлен список наиболее интересных специализированных баз данных.

RefSeq: собрание «эталонных» последовательностей, поддерживаемых NCBI. В RefSeq можно найти запись для каждого полностью секвенированного геномного элемента (хромосома, плазмида или другие), геномные последовательности из организмов, для которых процесс секвенирования геномов осуществляется в данный момент, а также геномные последовательности, соответствующие полным генам, продукт их транскрипции (мРНК или некодирующей РНК) и белковый продукт. Некоторые записи

RefSeq создаются полностью автоматически, другие создаются с небольшим вмешательством человека (куратора). Характер записи RefSeq можно отличить по его номеру доступа. Номер доступа начинается с «NC» для кураторского полногеномного элемента и с «NT» для автоматизированной промежуточной сборки.

Некоторые базы данных с автоматически собранными и аннотированными эукариотическими геномами позволяют пользователю перемещаться по геномным картам и извлекать четко определенные части хромосомы:

Ensembl: поддерживается в EMBL-EBI в сотрудничестве с Wellcome Trust, Институт Сэнгера (Хинкстон, Англия, Великобритания)

NCBI Map Viewer: поддерживается в NCBI

UCSC Genome Browser: поддерживается в Университете Калифорнии, Санта-Круз (США)

(**TIGR**): различные геномы, главным образом микробов. На веб-сайте TIGR также содержится коллекция гиперссылок на другие сайты с информацией о геномах.

Базы данных с кластерными последовательностями:

UniGene: собрание последовательностей, сгруппированных по гену, извлеченных из GenBank (в данный момент только для ограниченного числа организмов). Поддерживается NCBI.

Гомологичные последовательности из полной базы геномов **(HOGENOM)**: последовательности из полностью секвенированных геномов (кодирующие последовательности, извлеченные из EMBL и белки, экстрагированные из UniProt), классифицированные по семейству белков. Поддерживается университетом Клода Бернарда (Лион, Франция).

Высококачественная автоматическая и ручная аннотация микробных протеомов **(HAMAP)**: база данных последовательности белка из полностью секвенированных прокариотических геномов (бактерии, археи, пластиды),

экстрагированные из UniProt. Классифицированы по семействам белков и сопровождаются экспертной аннотацией. Поддерживается Женевским университетом (Швейцария).

Базы данных последовательностей, связанных с заболеваниями, передаваемыми половым путем, поддерживаются LANL (Национальная лаборатория Лос-Аламоса, Нью-Мексико, США):

База данных **HIV**: последовательности ВИЧ и SIV.

База данных **HCV**: гепатит С и связанные с ним флавивирусы.

ORALGEN: последовательности вируса простого герпеса и других пероральных патогенов.

Банк данных о белках (**PDB**): 3-D структуры, полученные на основании рентгеновской кристаллографией или ЯМР, белков, нуклеиновых кислот и их комплексов. Поддерживается исследовательской коллаборацией структурной биоинформатики. Некоторые сайты предлагают поиск по последовательностям белков или нуклеиновых кислот, соответствующим записям в PDB (Lemey et al., 2009).

2.4. Комбинированные базы данных, средства зеркального отображения и поиска баз данных.

2.4.1 Entrez

Ни одна из доступных базовых баз данных не является полной. Чтобы компенсировать это были предприняты усилия для создания составных баз данных. Самой популярной является база данных Entrez, поддерживаемая NCBI. Первоначально она содержала только последовательности и рефераты из медицинской литературы, но впоследствии были добавлены новые таблицы (в настоящее время более 30), что делает ее интегрированной базой данных по молекулярной биологии. Отдельные таблицы имеют перекрестные ссылки (рис. 3), так что можно искать, например, последовательность белка и затем, просто перейдя по ссылке, получить нуклеотидную по-

следовательность белок-кодирующего гена, трехмерную структуру белка, ссылки на литературу, в которой описываются последовательности и т. д.

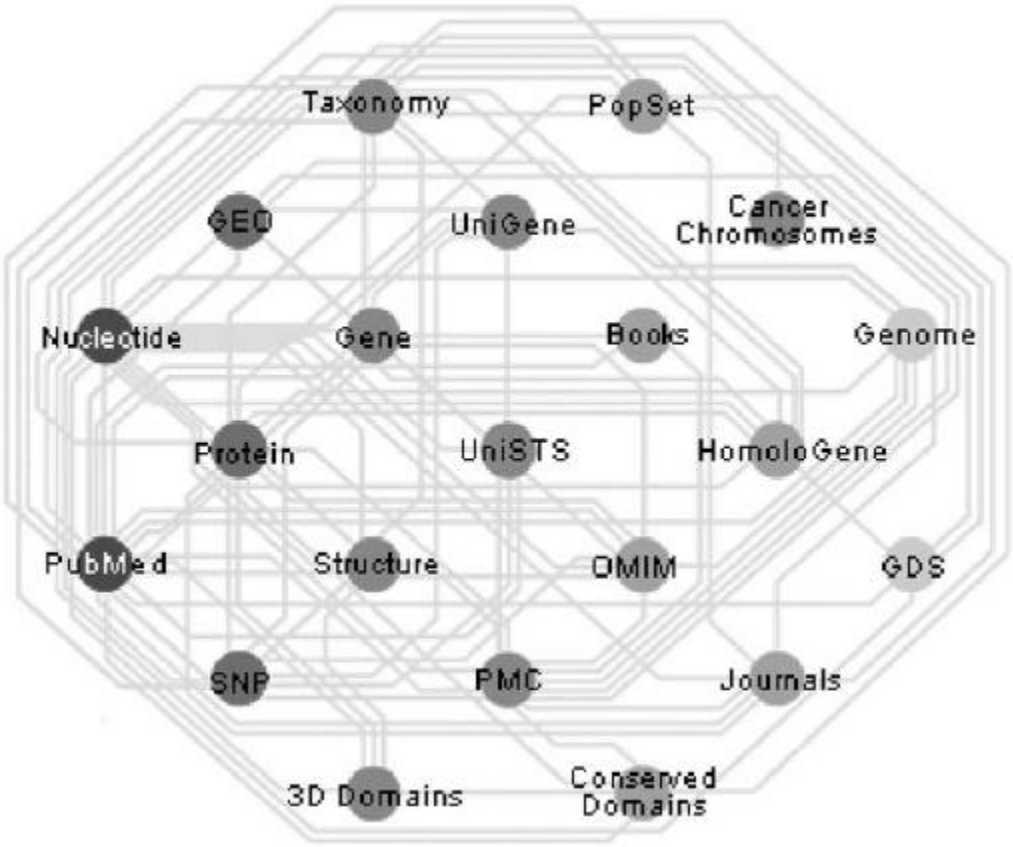

Рис. 3. Отношения между различными базами данных Entrez.

Ниже приведена более подробная информация о базах данных ENTREZ (Lemey et al., 2009).

Nucleotide: полная и не избыточная («nr») база данных последовательности нуклеиновой кислоты. Она содержит последовательности из следующих баз данных: RefSeq, GenBank, EMBL, DDBJ и PDB. Создателями этой базы данных были предприняты значительные усилия для устранения дубликатов. NCBI также предлагает BLAST-поиск исследуемой последовательности в отношении базы данных «nr». Кроме того, NCBI выполняет регулярные интервалы BLAST-поиска каждой последовательности против

полной базы данных, так что Entrez может предоставить ссылку “Related Sequences”, которая позволяет получать очень похожие последовательности.

Protein: белковые последовательности, хранящиеся в RefSeq, трансляции открытых рамок считывания, найденных в GenBank, EMBL и DDBJ, последовательности из PDB, UniProt / SwissProt, PIR и PRF.

UniGene: коллекция последовательностей GenBank, сгруппированных по генам (на данный момент только для ограниченного числа организмов)

Structure: база данных молекулярного моделирования (MMDB): трехмерные структуры белков, нуклеиновых кислот и их комплексов из базы данных PDB, но в улучшенном формате. В NCBI есть инструмент для сравнения трехмерных структур, инструмент для векторного выравнивания (VAST); они предлагают сравнение структуры белка с запросами из MMDB, а также сравнение информации об очень похожих структурах внутри MMDB. NCBI также предоставляет специализированный инструмент для просмотра структур, Cn3D, который может быть установлен как вспомогательное приложение («плагин») в браузере.

Gene: информация о генетических локусах (включая альтернативные названия генов, фенотип, положение на генетической карте и перекрестные ссылки на другие базы данных), соответствующих организмам и последовательностям в RefSeq.

PopSet: база данных множественных выравниваний последовательностей ДНК разных популяций или видов, представленных в GenBank и используемых для филогенетических и популяционных исследований.

PubMed: MEDLINE - это база данных с рефератами из медицинской литературы (начиная с 1966 года), которая хранится в Национальной Медицинской библиотеке (Бетесда, Мэриленд, США). PubMed содержит

MEDLINE плюс еще не аннотированные ссылки и ссылки на статьи, не относящиеся к наукам о жизни, опубликованные в журналах MEDLINE. WWW-интерфейс PubMed предоставляет гиперссылки на онлайн-версии некоторых журналов.

Taxonomy: содержит запись для каждого вида, подвида или более высокого таксона, для которого существует хотя бы одна последовательность в GenBank / EMBL / DDBJ. Запись имеет стандартное имя, которое используется для создания полей «Организм» в базах данных, а также серию альтернативных имен. В настоящее время номенклатура Taxonomy является стандартной для трех основных нуклеотидных баз данных и для UniProt.

Medical SubjectHeadings (**Mesh)**: набор стандартизированных ключевых слов, используемых NLM для индексации MEDLINE.

Задание 1: Поиск с использованием базы данных ENTREZ

ENTREZ - это простой в использовании интегрированный текстовый интерфейс для поиска в NCBI. Он связывает большое количество баз данных, включая нуклеотидные и белковые последовательности, PubMed, 3-D структуры белка, полные геномы, таксономию, и другие. В ENTREZ имеется всестороннее справочное руководство, которое можно просмотреть по адресу http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.

1. На домашней странице ENTREZ (http: //www.ncbi.nlm.nih. Gov / gquery / gquery.fcgi) указать в поле ввода поисковый запрос «Beggiatoa» (рис. 4).

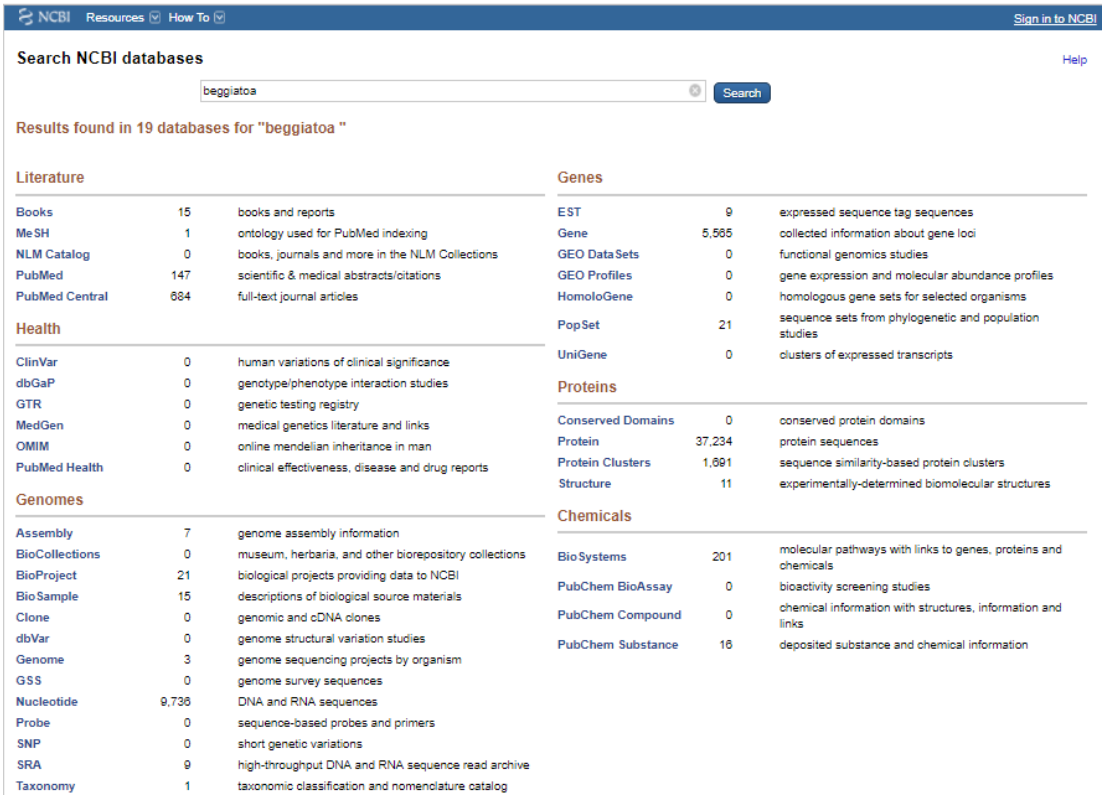

Рис. 4. Результат для поискового термина «Beggiatoa» в базе данных ENTREZ

2. Выполнить параллельный поиск по всем базам данных ENTREZ, введя текст поиска в верхней части страницы, а затем нажать «Go». Выполнить поиск в одной базе данных, введя свой поисковый запрос и щелкнув значок базы данных вместо «Go».

На момент написания методички вы могли найти 684 статьи в научной литературе о Beggiatoa, полный текст которых доступен на PubMed Central. Когда вы нажимаете, например, на пятый элемент в списке, он перепрaвит вас на страницу с этими статьями.

Когда вы будете использовать тот же поисковый термин «Beggiatoa» для поиска и нажмете на кнопку Protein, то увидите, что для представителей рода *Beggiatoa* известно 37234 белковых сиквенса (рис. 5). При нажатии на

ссылку Protein вы переходите на страничку, где приведены все известные белковые последовательности для представителей рода *Beggiatoa*.

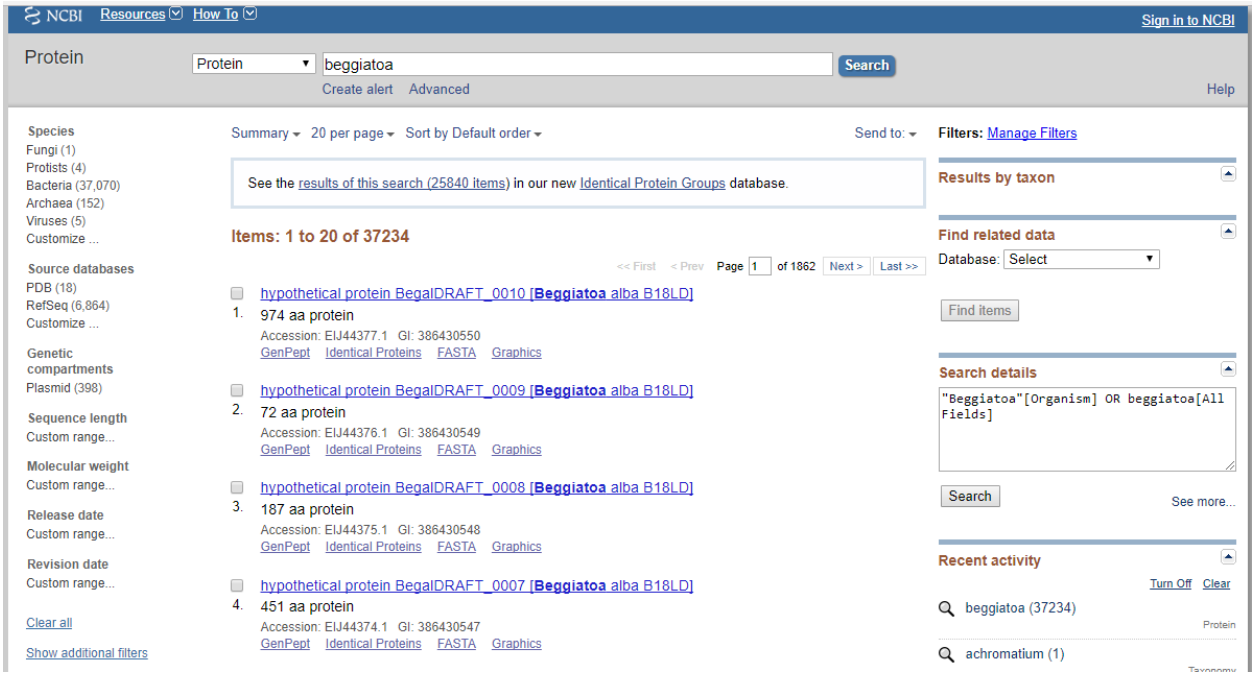

Рис. 5. Результаты по запросу “Beggiatoa” в базе данных Protein

При нажатии на гиперссылку на синий номер доступа в списке вы попадете на файл последовательности в формате GenBank. Формат Genbank хорошо подходит для размещения единичной последовательности и ее аннотации. Однако для филогенетического анализа необходимо объединение нескольких последовательностей в одном файле. К сожалению, разные программы использовали различные форматы входных файлов, что привело к тому, что в настоящий момент в молекулярной биологии используется около 18 форматов (см. Глава 3). Следовательно, при проведении комплексного анализа последовательностей надо потратить много усилий на подготовку и / или преобразованию файлов.

Глава 3. Форматы файлов

FASTA является самым простым текстовым форматом для представления последовательности белка или нуклеиновой кислоты. Формат начи-

нается с однострочного описания последовательности, которому предшествует «>». Остальную часть строки можно использовать для описания, но рекомендуется использовать для этого не больше 80 символов. На следующей строке представлена фактическая последовательность аминокислот или нуклеиновых кислот в стандартных кодах IUB / IUPAC. В случае если это выравненные последовательности, в них также могут содержаться пробелы «-». Формат FASTA иногда упоминается как формат Pearson / Fasta. Многие программы, такие как ClustalW, Paup, HyPhy, Rdp и Dambe, могут читать или импортировать формат FASTA. Для указания на этот тип формата, обычно используются расширения «.fas» или «.fasta» (Lemey et al., 2009).

Пример файла в формате FASTA приведен ниже.

```
> Taxon1
AATTCCCCAGCTTTCCACCAAGCTC
> Taxon2
AATTCCACAGCTTTCCACCAAGCTC
> Taxon3
AACTCCAGCACATTCCACCAAGCTC
> Taxon4
AACTCCACAACATTCCACCAAGCTC
```

NEXUS является модульным форматом для систематических данных и может содержать как последовательности данных, так и филогенетические деревья в блочных единицах (Maddison et al., 1997). Пример файла NEXUS, содержащего блок данных, показан ниже.

Данные и деревья являются публичными блоками, содержащими информацию, которая может использоваться несколькими программами. Частные блоки, содержащие информацию, подходящую для конкретных программ, используются Paup *, MrBayes, FigTree, SplitsTree и т. д.

Комментарии в файле NEXUS добавляются с помощью «[]»; в приведенном ниже примере используется две строки комментариев для указания позиций последовательности в выравнивании. Файлы в формате NEXUS обычно имеют расширение «.nex», «.nexus» или «.nxs».

```
#NEXUS
Begin DATA;
Dimensions ntax = 4 nchar = 25;
Format datatype = NUCLEOTIDE gap = -;
Matrix
[    1     11  21   ]
[    |     |    |   ]
Taxon1 AATTCCCCAGCTTTCCACCAAGCTC
Taxon2 AATTCCACAGCTTTCCACCAAGCTC
Taxon3 AACTCCAGCACATTCCACCAAGCTC
Taxon4 AACTCCACAACATTCCACCAAGCTC
;
End;
```

PHYLIP

Первоначально PHYLIP являлся стандартным форматом входного файла для программ в пакете Phylip. Впоследствии он был реализован как формат входных файлов для многих других программ, таких как Tree-Puzzle, PhyML и Iqpnni. Первая строка входного файла содержит количество таксонов и количество символов (в данном случае сайты выравнивания), разделенные пробелами. Во многих случаях имя таксона должно быть ограничено десятью символами. Если оно короче десяти символов, недостающие символы заполняются пробелами. Следует избегать специальных символов (“(“and”)”), квадратные скобки (“[“and”]”), двоеточие (“: “), точка с запятой (“; “) и запятая (","); это справедливо для большинства форматов

последовательности. Имя таксона должно быть в той же строке, что и первый символ данных для этого таксона. Файлы в формате PHYLIP имеет расширение «.phy».

```
4 25
Taxon1 AATTCCCCAGCTTTCCACCAAGCTC
Taxon2 AATTCCACAGCTTTCCACCAAGCTC
Taxon3 AACTCCAGCACATTCCACCAAGCTC
Taxon4 AACTCCACAACATTCCACCAAGCTC
```

Формат PHYLIP может быть «последовательным» или «чередующимся». В последовательном формате данные могут в любой момент перейти на новую строку:

```
4 65
Taxon1
AATTCCCCAGCTTTCCACCAAGCTCTGCAAGATCCCAGAGTCAAGGGCCTGTAT
TTTCCTGCTGG
Taxon2
AATTCCACAGCTTTCCACCAAGCTCTGCAAGATCCCAGAGTCAGGGGCCTGTAT
TTTCCTGCTGG
Taxon3      AACTCCAGCACATTCCACCAAGCTCTGCTAGATCC---
AGTGAGGGGCCTATACGTTCCTGCTGG
Taxon4
AACTCCACAACATTCCACCAAGCTCTGCTAGATCCCAGAGTGAGGGGCCTTTAT
TATCCTGCTGG
```

Формат с чередованием PHYLIP имеет первую часть каждой из последовательностей (50 сайтов в пример ниже), затем некоторые строки, дающие следующую часть каждой последовательности и т. д.

```
4 65
Taxon1 AATTCCCCAG CTTTCCACCA AGCTCTGCAA GATCCCAGAG
TCAAGGGCCT
```

```
    Taxon2 AATTCCACAG CTTTCCACCA AGCTCTGCAA GATCCCAGAG
TCAGGGGCCT
    Taxon3 AACTCCAGCA CATTCCACCA AGCTCTGCTA GATCC---AG
TGAGGGGCCT
    Taxon4 AACTCCACAA CATTCCACCA AGCTCTGCTA GATCCCAGAG
TGAGGGGCCT
    GTATTTTCCT GCTGG
    GTATTTTCCT GCTGG
    ATACGTTCCT GCTGG
    TTATTATCCT GCTGG
```

Чтобы облегчить чтение выравнивания, в примере выше имеется пробел через каждые десять символов (любые такие пробелы разрешены). В некоторых случаях чередующийся формат может или должен иметь спецификацию «I» в первой строке (например, «4 65 I»). Обратите внимание, что выравнивания в формате NEXUS также могут быть чередующимися или последовательными. Программы в пакете Paml, используют последовательный формат PHYLIP, но допускают большее количество символов в имени таксона (до 30). Paml рассматривает два последовательных пробела как конец имени вида, так что имя вида не должно иметь ровно 30 (или 10) символов (Lemey et al., 2009).

CLUSTAL

CLUSTAL широко не поддерживается в качестве входного формата в филогенетических программах, но он включен в данный обзор, поскольку является стандартным форматом вывода популярного программного обеспечения выравнивания. Формат распознается по слову CLUSTAL в начале файла. CLUSTAL - это чередующийся формат с блоками, которые повторяют имена таксонов, которые не должны содержать пробелов или превы-

шать 30 символов, а за ними следует пустое пространство. Вывод выравнивания последовательностей из программного обеспечения Clustal обычно задается по умолчанию в формате ".aln.". Clustal также указывает консервативные остатки в выравнивании использованием значка "*" внизу каждого блока (Lemey et al., 2009).

```
Taxon1 AATTCCCCAGCTTTCCACCAAGCTC
Taxon2 AATTCCACAGCTTTCCACCAAGCTC
Taxon3 AACTCCAGCACATTCCACCAAGCTC
Taxon4 AACTCCACAACATTCCACCAAGCTC
       ** *** * ************
```

MEGA

В этом формате ключевое слово «#Mega» указывает, что файл данных подготовлен для анализа с использованием программы Mega. Ключевое слово должно присутствовать на первой строке данных файла. Во второй строке должно быть записано слово «Title», за которым может следовать небольшое описание данных не более одной строки. Комментарии могут быть написаны на одной или нескольких строках сразу после строки заголовка и перед данными. Каждая метка таксона должна быть написана на новой строке с надписью «#» (без пробелов или вкладок). Последовательности могут быть отформатированы как в последовательном, так и в чередующемся формате.

```
#Mega
Title: example data set
! commenting
#Taxon1
AATTCCCCAGCTTTCCACCAAGCTC
#Taxon2
AATTCCACAGCTTTCCACCAAGCTC
#Taxon3
```

```
AACTCCAGCACATTCCACCAAGCTC
#Taxon4
AACTCCACAACATTCCACCAAGCTC
```

Некоторые продвинутые программы для популяционной генетики, такие как Beast и Lamarc, используют формат файла XML. Однако эти программы обычно предоставляют инструменты преобразования файлов, которые принимают формат FASTA или NEXUS.

Хотя форматы файлов и могут быть преобразованы вручную в любом текстовом редакторе, существуют программы для автоматического преобразования форматов последовательностей, такие как Readseq, написанный Доном Гилбертом. Онлайн-версия этого инструмента доступна по адресу http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi. К счастью, многие программы для выравнивания и филогенетического анализа принимают на вход и выдают на выходе файлы разных форматов. Новые программы обычно используют PHYLIP, NEXUS или FASTA в качестве стандартного формата ввода (Lemey et al., 2009).

Глава 4. Этапы филогенетического анализа

Филогенетический анализ молекулярных данных является одним из подходов к теоретическому изучению структуры и функции генетических макромолекул (РНК, ДНК, белков) и их эволюционного преобразования. Основная цель филогенетического анализа - изучение эволюционного порядка дивергенции последовательностей генов и белков или их частей, а также восстановление списков эволюционных событий (замен нуклеотидов, делеций и вставок) в предковых линиях этих макромолекул.

Филогенетический анализ состоит из следующих этапов:

1. Выравнивание генетических последовательностей;
2. Расчет генетических дистанций;
3. Выбор модели нуклеотидных или аминокислотных замен;

4. Построение филогенетического дерева.

Глава 5. Выравнивание генетических последовательностей

В эволюции генетических последовательностей происходят замены, вставки и делеции. Первым этапом филогенетического анализа является идентификация вставок и делеций, имевших место в эволюционной истории анализируемой группы последовательностей. Выравнивание последовательностей направлено на выявление гомологичных (имеющих общее эволюционное происхождение) позиций анализируемых последовательностей, установление наиболее вероятного, т.е. требующего наименьшего числа эволюционных событий, сценария эволюции анализируемой группы.

Чем больше пробелов вводится при выравнивании, тем меньше различий (замен) имеют последовательности после выравнивания.

Минимизация числа замен и минимизация числа пробелов находятся в противоречии друг с другом, чтобы решить эту задачу вводится система положительных и отрицательных штрафов за каждый пробел или замену.

Штрафы бывают положительными за совпадение нуклеотида и отрицательными за:

- несовпадение нуклеотида или аминокислоты;
- начало пробела;
- продолжение пробела.

5.1 Выравнивание BLAST

Поскольку не всегда можно позволить себе высокопроизводительную машину или длительное время поиска на стандартном компьютере, биоинформатики разработали так называемые «эвристические» алгоритмы, которые позволяют осуществлять поиск в базе данных за значительно меньшее время, однако не дают абсолютной гарантии найти все последовательности с наивысшей оптимальной оценкой выравнивания. Самой популярной про-

граммой для выравнивания является Blast. Более ранняя версия 1 Blast, разработанная в NCBI, обеспечивала выравнивание без пробелов (Altschul et al., 1990). После того, как Уоррен Гиш из команды «Blast» перебрался в Вашингтонский университет (Сент-Луис, Миссури, США), параллельно с новыми версиями в NCBI (interchangeably Blast, Blast 2, «Gapped Blast» или NCBI Blast (Altschul et al., 1997)) стал разрабатываться WU-Blast в Вашингтонском университете. По сравнению с NCBI Blast, WU-Blast имеет больше дополнительных параметров, которые позволяют увеличить скорость в ущерб чувствительности.

В текущем пакете Blast содержится программа «все-в-одном» BLASTALL, которая предлагает пять типов поиска в базе данных:

BLASTN сравнивает последовательность нуклеиновой кислоты со всеми последовательностями нуклеиновых кислот в базах данных и их дополнениях.

BLASTP сравнивает последовательность белка со всеми последовательностями в белковых базах данных.

BLASTX сравнивает последовательность нуклеиновой кислоты, транслируемой с шести рамок считывания, со всеми последовательностями белковых баз данных.

TBLASTN сравнивает последовательность белка со всеми последовательностями нуклеиновых кислоты из база данных, которые «на лету» транслируются по шести рамкам считывания.

TBLASTX сравнивает последовательность нуклеиновой кислоты, транслируемой по шести рамкам считывания, со всеми последовательностями базы данных нуклеиновой кислоты, транслируемых на лету по шести рамкам считывания. TBlastX не делает выравнивание с пробелами.

Существует также программа Blastpgp (только для поиска белка), который вычисляет E () используя более точную формулу с учетом состава

запроса и последовательностей в базе данных, а также предлагает следующие типы поиска:

Position Specific Iterated Blast (PSI-Blast) сначала просто ищет белок против белка, затем выбирает наилучшее соотношение последовательность запроса / выравнивание с последовательностями из базы данных, объединяет их во множественное выравнивание, преобразует множественное выравнивание последовательностей в профиль частоты аминокислот и осуществляет поиск профиля относительно базы данных. Процедуру можно повторить до тех пор, пока профиль не сходится, то есть до тех пор, пока последовательности, найденные профилем, не будут такими же, как последовательности, используемые для его создания.

Pattern-Hit Initiated Blast (PHI-Blast) выполняет поиск одновременно шаблона (определенного в формате PROSITE) и последовательности белка против белковой базы данных, находит последовательности, которые соответствуют шаблону и показывают значительное местное сходство в исследуемом регионе.

На первом этапе работы можно использовать PHI-Blast, а затем продолжить с PSI-Blast (Lemey et al., 2009).

Задание 2:

1. Зайти на сайт https://www.ncbi.nlm.nih.gov

2. Выбрать Resources -> DNA&RNA -> BLAST (рис. 6)

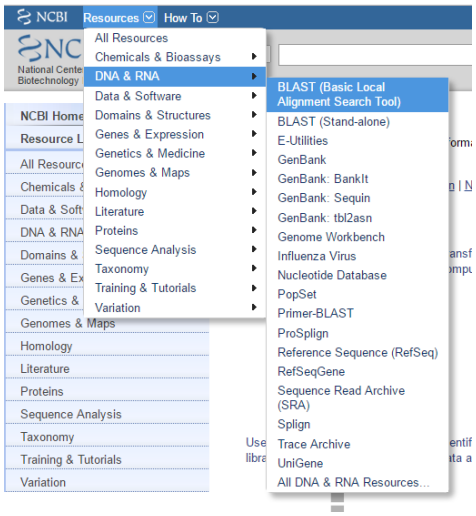

Рис. 6. Последовательность открытия вкладок для выполнения задания 2 п.2

3. В открывшемся окне выбрать Protein BLAST

4. В новом окне поставить галочку напротив align two or more sequences (рис. 7)

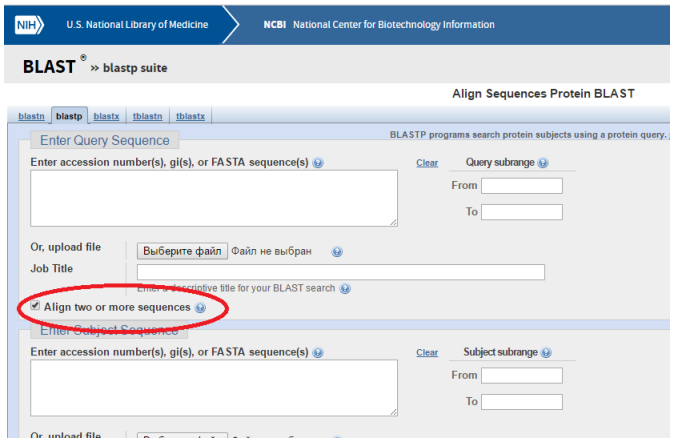

Рис. 7. Внешний вид страницы, которая появляется при выполнении задания 2 п.4.

5. В оба окошка вставить последовательности, которые необходимо выровнять

6. Нажать “BLAST”

5.2. Множественное выравнивание

Одним из самых широко распространенных подходов к выравниванию нескольких последовательностей является алгоритм прогрессивного множественного выравнивания (progressive multiple alignment), основанный на проведении выравнивания в три этапа. На первом этапе все последовательности попарно выравниваются между собой с использованием того или иного алгоритма выравнивания, и среди них выявляют группы схожих между собой последовательностей. На втором этапе производится выравнивание последовательностей в каждой такой группе, после чего – выравнивание групп между собой.

При выравнивании кодирующих нуклеотидных последовательностей в общем случае удобнее и эффективнее транслировать их в кодирующие аминокислотные последовательности и провести выравнивание на аминокислотном уровне.

5.3. Выравнивание Clustal

Форматы файлов и их доступность

Как обсуждалось в главе 3, наиболее распространенными форматами файлов являются Genbank, EMBL, SWISS-PROT, FASTA, PHYLIP, NEXUS и Clustal. Форматы базы данных Genbank, EMBL и SWISS-PROT обычно используются для единичной последовательности, и большая часть каждой записи посвящена информации о последовательности. Как правило, эти форматы не используются для множественного выравнивания. Тем не менее, Clustal может читать эти форматы и выдавать выравнивания (включая пробелы, «-») в разных форматах, таких как PHYLIP и Clustal (используется исключительно для множественных выравниваний). На самом деле программы Clustal могут использоваться в качестве преобразователей выравнивания, они предоставляют возможность считывать файлы последовательностей в следующих форматах: NBRF / PIR, EMBL / SWISS-PROT, FASTA, GDE, Clustal, GCG / MSF и NBRF / PIR; и записывать файлы выравнивания

во всех следующих форматах: NBRF / PIR, GDE, Clustal, GCG / MSF и PHYLIP.

ClustalW и ClustalX свободно доступны и могут быть загружены с файлового сервера EMBL / EBI (ftp://ftp.ebi.ac.uk/pub/software/) или с сервера ICGEB в Страсбурге, Франция (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/ и ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/). Эти сайты также доступны через веб-сайт программ филогении, поддерживаемый Джо Фелсенштейном (http://evolution.genetics.washington.edu/phylip/software.html). Программы написаны под операционные системы MSDOS или Windows, Macintosh, VAX VMS и Unix / Linux. В каждом случае ClustalX (X означает X окон) обеспечивает пользователям графический интерфейс с ярким отображением выравниваний. ClustalW имеет более старый текстовый интерфейс и менее привлекателен для случайного использования. Тем не менее, в нем имеются обширные возможности командной строки, что делает его чрезвычайно полезным для высокопроизводительного использования. Алгоритм Clustal также реализуется в ряде коммерческих пакетов.

Clustal также напрямую доступен с нескольких серверов через Интернет, что особенно привлекательно для случайных пользователей. Однако интернет-сайт не лишен недостатков. Во-первых, пользователи обычно не имеют доступа ко всему комплексу функций, которые обеспечивает Clustal. Во-вторых, может быть сложно отправить и получить большое количество последовательности или выравниваний. В-третьих, для выполнения больших выравниваний может потребоваться много времени; может быть даже предел числа последовательностей, которые можно выровнять. Тем не менее, превосходные серверы Clustal вполне можно использовать через сайт EBI (http://www.ebi.ac.uk/clustalw/) и поисковую установку BCM на сайте http://searchlauncher.bcm.tmc.edu/.

Задание 3:

Сделать выравнивание аминокислотных последовательностей в Clustal

1. Зайти на сайт https://www.ncbi.nlm.nih.gov
2. Ввести в поисковую строку gyrB Beggiatoa и выбрать поиск по базе данных “Protein”.
3. Сохранить последовательности в формате fasta: send to -> file -> format -> fasta
4. Перейти на сайт http://www.ebi.ac.uk/Tools/msa/clustalo/
5. Загрузить файл со скачанными последовательностями
6. Установить следующие настройки:

Enter or paste a set of -> Protein

Output format -> Clustal w/o numbers

7. Нажать кнопку “submit”
8. Чтобы результаты были более выразительные, можно выделить цветом одинаковые буквы, нажав кнопку “Show colors” (рис 8).

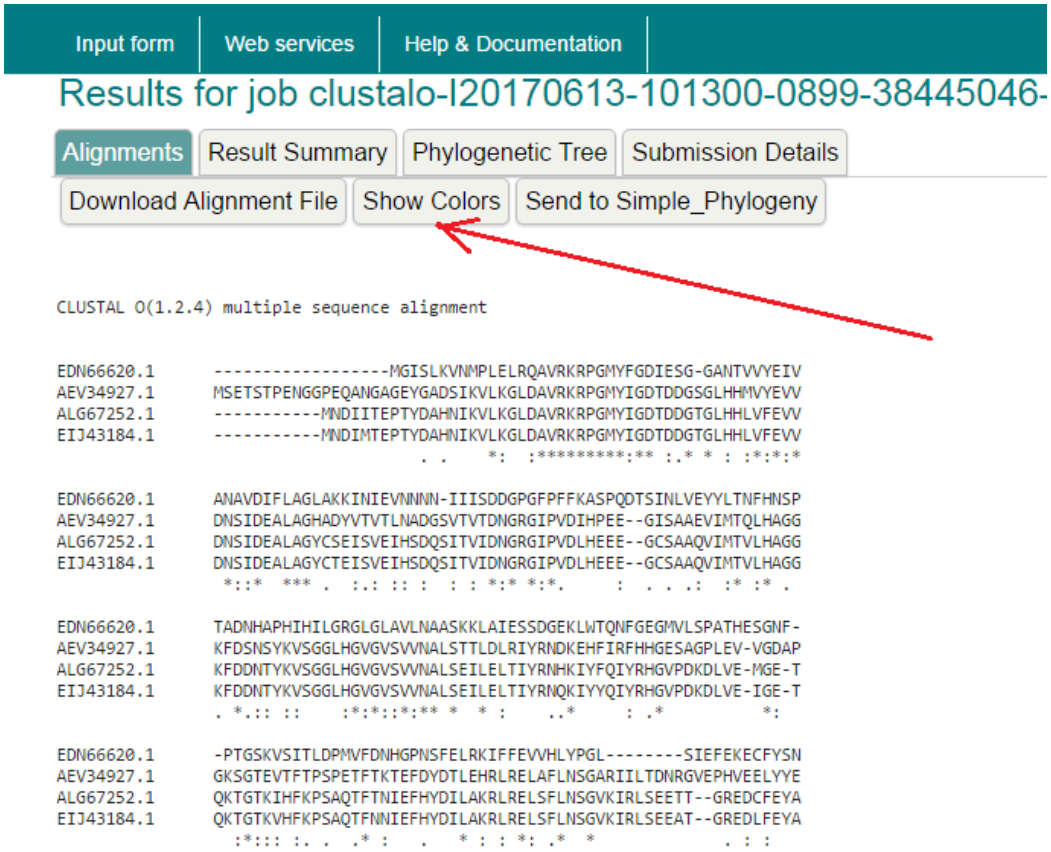

Рис. 8. Результат выравнивания аминокислотных последовательностей в программе Clustal Omega. * - совпадение аминокислоты во всех выравниваемых последовательностях, : - совпадение аминокислоты во всех последовательностях кроме одной, . - совпадение аминокислоты во всех последовательностях кроме двух.

5.4. Выравнивание T-Coffee

Хотя для различных систем (Windows, Unix / linux и MacosX) доступно автономное программное обеспечение T-Coffee, мы разберем как выполнять выравнивание с использованием веб-сервера T-Coffee (доступно на http://www.tcoffee.org/). Стандартная форма представления требует от нас только загрузить файл в формате FASTA или вставить последовательности в представленное окно (рис. 9). Если ввести адрес электронной почты, то на указанный почтовый ящик придет уведомление со ссылкой на страницу результатов, но это не обязательно. Расчеты T-Coffee занимают больше времени, чем прогрессивное выравнивание Clustal.

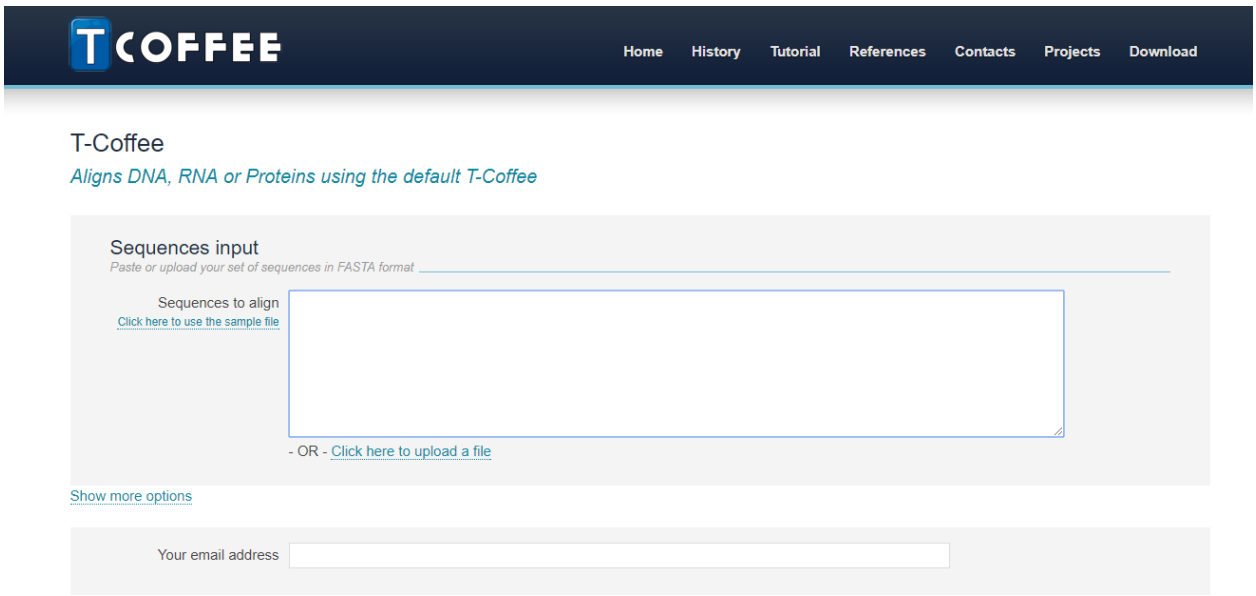

Рис. 9. Страница сайта T-Coffe, на которую надо загружать последовательности

Задание, отправленное на веб-сервер, должно занимать менее 2 минут. Когда процедура выравнивания будет завершена, появится новая страница со ссылками на выходные файлы. Файл можно сохранить в формате FASTA, PDF или HTML.

5.5. Выравнивание MUSCLE

Автономное программное обеспечение для MUSCLE, доступное для Windows, Unix / linux и MacosX, можно скачать по адресу http: //www.drive5.com/ (рис. 10.). Muscle - это программа для командной строки и, следовательно, требует использование терминала (Unix / Linux / MacosX) или DOS-окна в операционной системе Windows. Для того чтобы запустить программу нужно скопировать исходный файл с последовательностями в формате FASTA в папку Muscle, открыть терминал / DOS-окно и перейти в папку Muscle (используя команду "CD"). Чтобы запустить программу Muscle нужно ввести команды `-in названние_файла.fasta -out название_файла muscle.fasta.`

В DOS-окне необходимо указать исполняемый файл с расширением: `muscle.exe -in названние_файла.fasta -out названние_файла muscle.fasta.` Программа завершает процедуру выравнивания за несколько секунд и выдает файл вывода в формате fasta (`"названние_файла muscle.fasta"`).

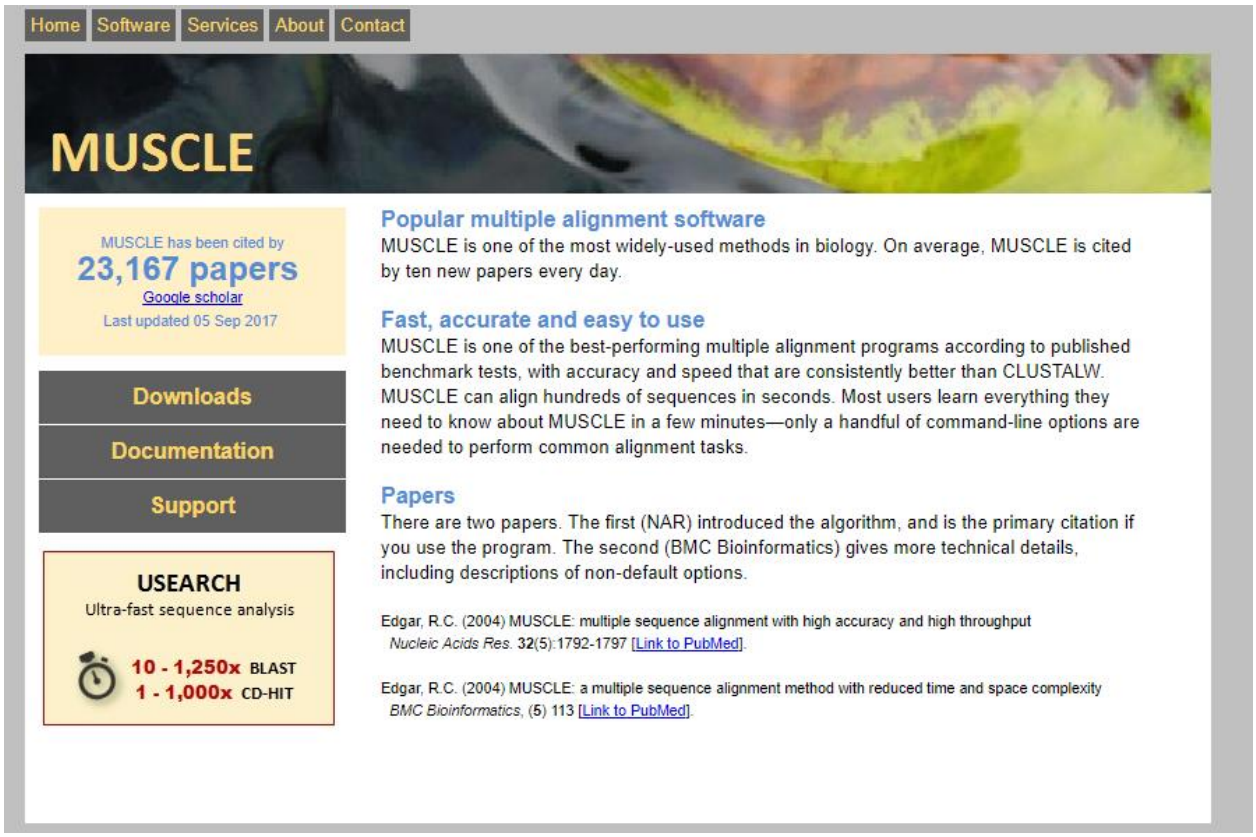

Рис. 10. Страница сайта, с которого можно скачать программное обеспечение для выравнивания Muscle

Глава 6. Расчет генетических дистанций

В процессе независимой эволюции двух последовательностей ДНК, произошедших от общего предка, в них будут накапливаться различия в результате мутационного процесса. Предположим, что мы сравниваем две последовательности длинной в *N* нуклеотидов (после выравнивания). Если мы обнаружим в этих последовательностях *S* несовпадающих нуклеотидов, то генетическую дистанцию *p* можно определить по формуле:

P=S/N

Дистанцию *p* называют долей замен или долей несовпадающих нуклеотидов в последовательностях. Ожидается, что дистанция *p* не может превышать величину 0.75. Это связано с тем, что последовательности состоят из нуклеотидов четырех типов. Если все положения заменились хотя бы один раз, то в среднем каждое четвертое положение будет совпадать, а

каждые три положения будут отличаться, т.е. *p* достигнет максимального значения 0.75. В особых случаях, если нуклеотидный состав сравниваемых последовательностей неодинаков, отличия могут превысить порог 0.75. Другая проблема состоит в том, что каждое положение в последовательности может заменяться не один раз. Это приводит к тому, что генетическая дистанция *p* не изменяется пропорционально времени. Для проведения филогенетического анализа наоборот требуются такие меры генетического расстояния, которые бы увеличивались прямо пропорционально времени прошедшему с момента разделения предковой линии. Решить проблему расчета линеализированной, относительно времени прошедшего с момента разделения предковой линии, меры генетического расстояния можно с помощью одной из моделей накопления замен (Темралеева и др., 2014).

Глава 7. Модели накопления замен

1) raw или p-дистанция: доля несовпадающих нуклеотидов при попарном сравнении выровненных последовательностей. Данный вид дистанций может применяться для проверки насыщения последовательностей заменами.

2) JC69 – модель Джукса-Кантора: модель производит линеаризацию p-дистанции с учетом того, что все замены имеют одинаковую вероятность. Частоты встречаемости нуклеотидов всех четырех типов не различаются.

3) K80 – двухпараметрическая модель Кимуры: имеет те же основные предположения, что и модель Джукса-Кантора за исключением
того, что два вида замен транзиций (A ↔ G, C ↔ T), и трансверсий (A ↔ C, A ↔ T, C ↔ G, G ↔ T) происходят с разной вероятностью. Частоты встречаемости нуклеотидов всех четырех типов не различаются.

4) F81 – модель Фельзенштейна: обобщенный вариант модели JC69, учитывающей поправку на разные частоты встречаемости нуклеотидов.

5) K81 – обобщенная модель Кимуры: предполагает различные вероятности для двух видов трансверсий: A ↔ C и G ↔ T с одной стороны, и A ↔ T и C ↔ G с другой стороны. Частоты встречаемости нуклеотидов всех четырех типов не различаются.

6) F84 – модифицированная модель Кимуры K80, с предположением о разных частотах встречаемости нуклеотидов.

7) T92 – модель Тамуры: обобщенная модель K80 с учетом смещения нуклеотидного состава последовательностей в область увеличения концентрации GC-пар.

8) TN93 – модель Тамуры-Нея: модель предполагает различные вероятности для обоих видов транзиций (A ↔ G по сравнению с C ↔ T), и трансверсий. Частоты встречаемости нуклеотидов различаются.

9) HKY – модель Хасигавы-Кишино-Яно: объединение предположений модели K80 и F81, близка к модели F84 и доступна только в программах, производящих кластеризацию на основе дискретных методов, например, таких как метод максимального правдоподобия.

10) GTR – обобщенная модель с полным временным преобразованием: вероятности всех возможных вариантов замен (за исключением обратных) отличаются друг от друга и частоты встречаемости нуклеотидов разные. Модель доступна только в программах, производящих кластеризацию на основе дискретных методов таких, например как метод максимального правдоподобия.

Для более обоснованного выбора модели нуклеотидных замен можно воспользоваться программой **jModelTest** (Darriba et al., 2012)

Принципы, на основании которых выбирается одна из моделей замен:

- если различные модели дают приблизительно одинаковые результаты, следует использовать более простую модель, имеющую меньшую дисперсию;

- при дистанциях до 0,05 можно использовать модель Джукса-Кантора;
- при возрастании дистанции до 0,3, можно использовать модель Джукса-Кантора при небольшом соотношении числа транзиций и трансверсий (скажем, менее 2). Если соотношение транзиций и трансверсий больше 2, и число сравниваемых нуклеотидов велико, то следует использовать модель Кимуры;
- если дистанции выше, а содержание четырех нуклеотидов значительно отличается от 25%, при небольшом соотношении числа транзиций и трансверсий следует использовать модель Таджимы-Неи, при большом – модель Тамуры или Тамуры-Неи;
- при дистанциях выше 0,3 и вариациях в частоте замен в различных позициях, следует использовать Г-дистанции;
- хорошим правилом является использовать несколько различных моделей, и в случае, если их результаты значительно отличаются друг от друга, установить причину этого (Лукашов, 2009).

При возрастании дистанции выше 1, т.е. когда в каждой позиции в среднем произошла более чем одна замена, ни одна из моделей не будет давать надежных результатов. В таких случаях говорят о влиянии эффекта насыщения (saturation) нуклеотидных замен. Собственно, влияние этого эффекта начинает сказываться уже при меньших нуклеотидных дистанциях между последовательностями. Это будет связано и с трудностями (ненадежностью) выравнивания таких последовательностей, и с большим стандартным отклонением при расчете дистанции. В таких случаях для установления эволюционных отношений между формами жизни следует использовать менее изменчивые участки генома, или использовать анализ либо аминокислотных последовательностей, либо несинонимичных позиций.

Вторым фактором, определяющим генетическую дистанцию между последовательностями, является неодинаковость скоростей накопления за-

мен в различных участках сравниваемых последовательностей ДНК. Причиной этому может быть, например, кодирование участком ДНК активного центра белка, который обладает высокой степенью консервативности относительно аминокислотного состава и функции. Для корректировки неодинаковой скорости накопления замен в различных участках последовательности служит гамма распределение с параметром *G*. Подобная корректировка получила название *G*-коррекция, она предусмотрена в моделях *JC69, K80, F81, K81, F84, T92, TN93, HKY* и *GTR*. Если внимательно посмотреть на приведенное описание 10-ти моделей, то по мере продвижения от начала списка к его концу заметно постепенное усложнение алгоритма расчета генетической дистанции. При выборе более сложной модели происходит увеличение расчетного компьютерного времени для проведения филогенетического анализа, особенно при больших выборках (Темралеева и др., 2014).

Одним из наиболее универсальных средств для вычисления матриц попарных генетических дистанций между последовательностями является язык программирования R в комплекте с пакетом APE (Paradis et al., 2004).

Глава 8. Филогенетические деревья

8.1. Структура филогенетического дерева

Общая структура филогенетического дерева изображена на рис. 11. В состав дерева входят следующие структурные единицы:

Оперативные таксономические единицы (OTU) = листья – объекты филогенетического исследования (гены, участки генов, нуклеотидные или аминокислотные последовательности).

Узел – последний общий предок двух последовательностей, момент дихотомии.

Корень – предок всех анализируемых последовательностей. Деревья с корнями отражают направление эволюции, без корней – только родственные отношения между анализируемыми последовательностями. Для опре-

деления положения корня можно использовать внешнюю группу – одну или несколько OTU, которые отпочковались от общего дерева заведомо раньше (но не на много раньше) анализируемых OTU. Например если мы анализируем нуклеотидные последовательности генов 16S рРНК альфапротеобактерий, то в качестве внешней группы можно взять последовательность бета- или гаммапротеобактерии.

Ветвь – связь между узлами или между узлом и листом.

Порядок всех ветвей дерева называется топологией.

Кластер, клада – группа OTU, имеющих общего предка.

Монофилетическая группа – группа OTU, объединяющая все OTU, происходящие от общего предка.

Полифилетическая группа – группа OTU, имеющая разных последних общих предков.

Парафилетическая группа – группа, состоящая из OTU, имеющих общего предка, но не включающая в себя все OTU, происходящие от этого предка.

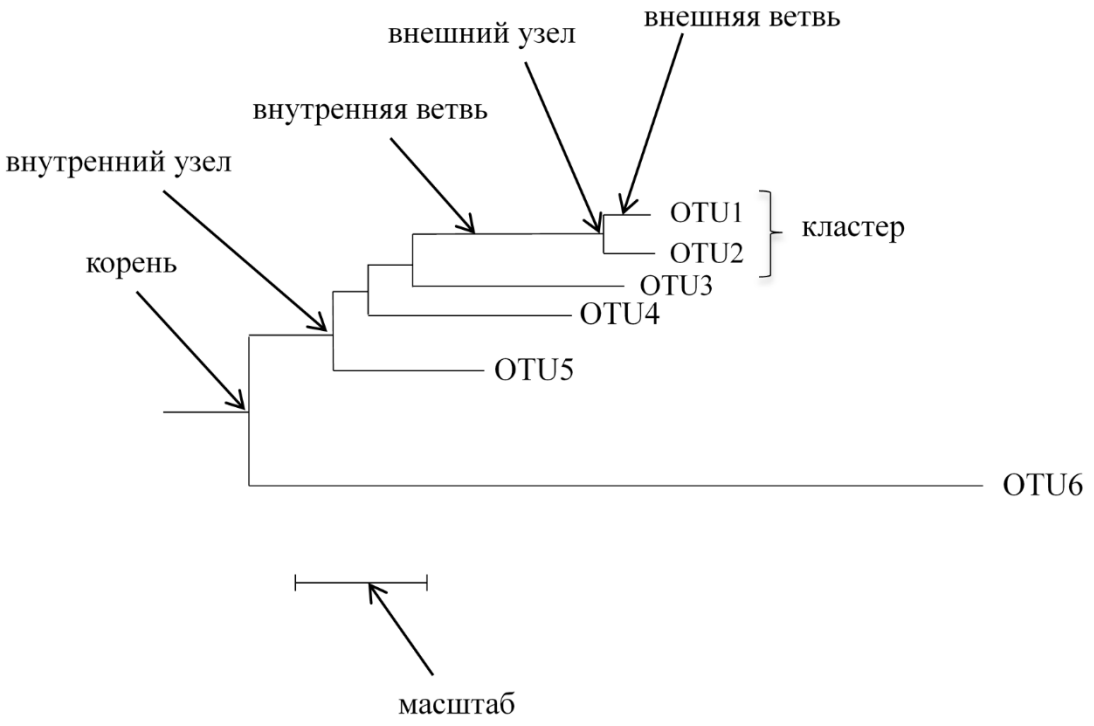

Рис. 11. Структура филогенетического дерева.

8.2. Количество возможных деревьев

Рассмотрим различные варианты филогенетических деревьев для трех OTU – последовательностей A, B и C (рис. 12). Отметим, что топология первых двух деревьев на рисунке одинакова, так как оба этих дерева показывают филогенетические отношения между тремя последовательностями, порядок их бифуркаций, одинаковым образом. Такие деревья называются конгруэнтными. Для трех последовательностей возможно три неконгруэнтных дерева, в которых последовательности A и B (рис. 12, первые 2 дерева), A и C (рис. 12, третье дерево), B и C (рис. 12, четвертое дерево) более родственны друг другу.

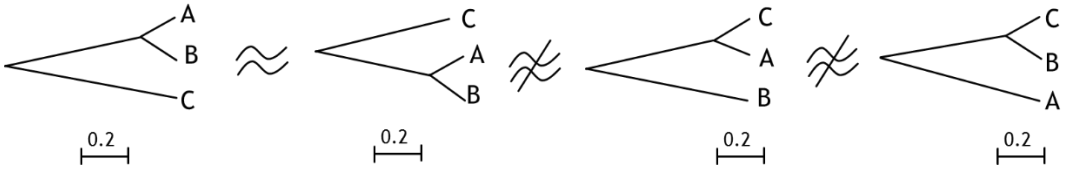

Рис. 12. Филогенетические деревья для трех OTU.

Таким образом, для трех OUT существуют три различных (неконгруэнтных) дерева. При увеличении числа анализируемых OTU число возможных деревьев возрастает согласно формулам:

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!} \quad N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

N_R – число неконгруэнтных укорененных деревьев для n OTU,

N_U – число неконгруэнтных неукорененных деревьев для n OTU

Уже для 10 OTU число укорененных и неукорененных деревьев составляет 34 миллиона и 2 миллиона соответственно, и только одно из этих деревьев является истиным (Лукашов, 2009).

Задание 4: Рассчитать количество укорененных и неукорененных конгруэнтных деревьев для 6 последовательностей

8.3. Топология деревьев

Топология деревьев различается в зависимости от эволюционной истории входящих в состав дерева OTU.

Звездообразная филогения – короткие внутренние ветви при относительно длинных внешних ветвях (близость к политомии). Отражает эволюционную радиацию – значительное и быстрое возрастание генетического разнообразия в группе.

Кактусообразная филогения – длинные внутренние ветви. Отражает более медленное увеличение генетического разнообразия в группе.

Политомия – отделение трех и более последовательностей от одного узла сразу. Ее возможность дискуссионна.

8.4. Формат для сохранения деревьев

Филогенетические деревья почти всегда сохраняются в одном из двух форматов: NEWICK или NEXUS. Стандарт NEWICK для машиночитаемого формата дерева использует соответствия между деревьями и вложенными круглыми скобками; пример для дерева с четырьмя таксонами показано на рис. 13. В этих обозначениях дерево в основном представляет собой строку, в которой приведены пары в круглых скобках, каждая пара в круглых скобках представляет собой внутренний узел. Длины ветвей для терминальных ветвей и внутренних узлов написаны после двоеточия. Формат NEXUS включает форматирование NEWICK вместе с другими командами и обычно имеет отдельный блок определения таксонов (Lemey et al., 2009). Эквивалент NEXUS для дерева на рисунке 13 с длинами ветвей:

```
#NEXUS
Begin trees;
Translate
```

```
    1 A,
    2 B,
    3 C,
    4 D,
    ;
    tree PAUP_1 = [&U]
((1:0.1,2:0.2):0.2,(3:0.3,4:0.4));
    End;
```

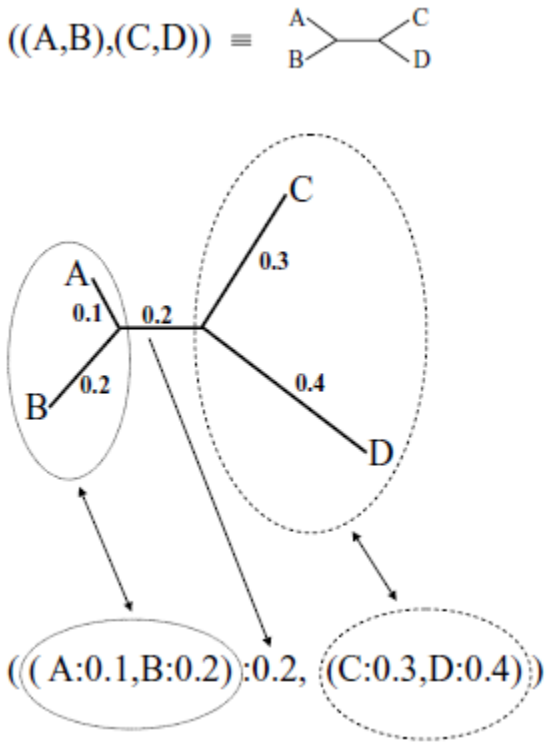

Рис. 13. Филогенетическое дерево в формате NEWICK. Гипотетическое неукорененное дерево для четырех таксонов (A, B, C и D) с числами вдоль ветвей, которые обозначают предполагаемые генетических расстояний и их описанием в формате NEWICK.

8.5. Методы построения филогенетических деревьев

Все методы построения филогенетических деревьев делятся на 2 большие группы – дистанционные и методы анализа дискретных признаков.

8.5.1. Дистанционные методы построения филогенетических деревьев

Построение филогенетических деревьев дистанционными методами включает в себя 2 этапа:

- Первый этап – установление попарных эволюционных дистанций между анализируемыми последовательностями (очень важен выбор эволюционной модели и установление оптимального метода расчета эволюционных дистанций!). На выходе – матрица дистанций.
- Второй этап – перевод матрицы дистанций в графический формат (собственно дерево)

К дистанционным методам относятся:

- Метод минимума эволюции
- Метод присоединения соседей

Метод минимума эволюции

- Наиболее вероятным сценарием эволюции группы является тот, который требует наименьшего числа эволюционных событий.
- Сумма длин всех ветвей (S) – общее число эволюционных событий
- Метод подразумевает построение всех возможных деревьев и выбор из них дерева с наименьшей суммой длин всех ветвей
- Число теоретически возможных топологий очень высоко, поэтому необходим промежуточный отсев топологий, не ведущих к минимизации S. Основан на связях между соседями.

Метод присоединения соседей (Neighbour-Joining)

Самый популярный среди дистанционных методов — это метод ближайших соседей (neighbour joining). Среди анализируемых видов находят два с минимальными различиями в последовательности (т.е., максимально похожие). Исходя из составленной матрицы, данные об этих видах «объединяются», и далее они участвуют в анализе в объединенном состоянии (рис. 14). Виды один за другим проходят эту процедуру до тех пор, пока не будет найдено одно, полностью разрешенное дерево. Этот алгоритм хорош

тем, что он относительно прост и подходит для обработки больших наборов данных (Baum & Smith ,2012).

Разные авторы, однако, перечисляют некоторые минусы метода ближайших соседей. Например, есть мнение, что этот метод хуже работает с таксонами, которые филогенетически далеки друг от друга (Bruno et al., 2000; Yang & Rannala, 2012). Также недостатком можно считать и то, что метод всегда выдает дерево с одним-единственным возможным вариантом ветвления (Baum & Smith ,2012). Это происходит потому, что алгоритм подразумевает построение одной филогении без сравнения с другими, тогда как в кладистических методах оцениваются деревья с различным порядком ветвления. Несмотря на то, что в серьезных филогенетических анализах методы матрицы расстояний сейчас почти не используются, они применяются, например, для быстрого построения филогений близкородственных бактерий и вирусов.

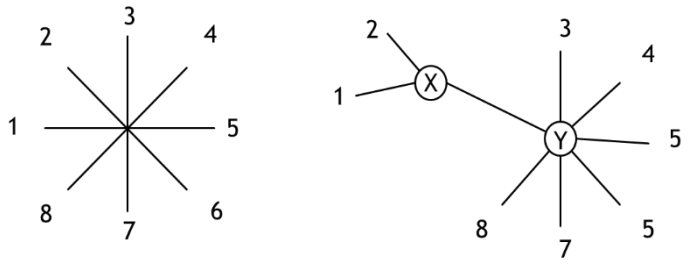

Рис. 14. Схема построения филогенетического дерева методом NJ

8.5.2. Методы анализа дискретных признаков

- Рассматривают отличия между последовательностями в конкретных позициях.
- Цель – реконструкция сценария наиболее вероятно объясняющего порядок конкретных нуклеотидных замен.

К методам анализа дискретных признаков относятся:

- метод максимальной экономии (Maximum Parsimony);

- метод максимального правдоподобия (Maximum Likelihood).

Maximum Parsimony (Метод наибольшей экономии)

- Направлен на поиск филогенетического дерева, требующего наименьшего числа эволюционных изменений. Подразумевает, что для данной группы последовательностей может существовать несколько деревьев, равновероятно объясняющих наблюдаемые различия.
- Анализируются нуклеотиды или аминокислоты, находящиеся в информативных позициях (parsimonious sites).
- Информативные позиции – позиции, состояние которых позволяет отдать предпочтение тому или иному филогенетическому дереву. Это те позиции, в которых у анализируемых последовательностей находятся два или больше различных нуклеотидов и как минимум два из них присутствуют в двух или более последовательностях.

Maximum Likelihood (Метод максимального правдоподобия)

- Базируется на использовании моделей эволюции.
- Использует эти модели для построения филогенетического дерева, исходя из оценки вероятности (правдоподобия) нахождения каждого конкретного нуклеотида в каждой конкретной позиции.
- Вероятность – функция, направленная в будущее. Мы знаем начальные условия и можем предположить результат.
- Правдоподобие – функция, направленная в прошлое. Мы знаем результат и предполагаем исходные условия.
- Одна из самых совершенных программ для анализа – PAML.

8.6. Статистическая оценка дерева

- Важный критерий – проверка надежности каждой ветви дерева.

- Самый надежный способ статистической оценки достоверности дерева – бутстрэп-анализ.
- Для проведения бутстрэп-анализа генерируются случайные выборки позиций (100-1000), по которым также строятся деревья.

8.7. Сравнение филогенетических методов

- При сравнении дистанционных методов и методов анализа дискретных признаков большинство ученых предпочитают последние.
- Методы анализа дискретных признаков учитывают распределение нуклеотидов в каждой позиции последовательности отдельно, а дистанционные методы усредняют изменения по всем позициям.
- Метод MP анализирует только филогенетически информативные изменения, не учитывая остальные.
- Метод NJ занимает меньше компьютерного времени и позволяет использовать и сравнивать различные эволюционные модели.
- При анализе данных ДНК-ДНК гибридизации, рестрикционного анализа и т.д. можно использовать только дистанционные методы.
- Проблемы при несоблюдении модели молекулярных часов серьезнее в методе MP, чем в NJ (Лукашов, 2009).

8.8. Монофилетическая и полифилетическая группы

Алгоритмы, обсуждавшиеся выше, обычно генерируют строго бифуркационные деревья (т. е. деревья, где любой внутренний узел всегда связан только с тремя другими узлами). Это стандартный способ представления эволюционных отношений между организмами, но он предполагает, что в ходе эволюции любая предковая последовательность (внутренние узлы дерева) может порождать только две отдельные линии (листья). Однако в природе существуют такие явления, как взрывное эволюционное излучение ВИЧ или HCV, которое может быть наилучшим образом представлено с помощью многоуровневого дерева, такого как показано на рисунке 15 а,

или деревом, которое допускает некоторую степень мультифуркации (рис. 15 b). Мультифуркации на филогенетическом дереве также известны как политомии и могут быть классифицированы как твердые политомии и мягкие политомии. Твердые политомики представляют собой взрывное излучение, в котором один общий предок почти мгновенно дал несколько разных линий в одно и то же время. Трудные политомии трудно доказать, и даже сомнительно, на самом ли деле они происходят (подробное обсуждение см. в статьях Li, 1997, и Page & Homes, 1998). С другой стороны, мягкие политомы представляют собой неразрешенные топологии деревьев. Они отражают неопределенность в отношении точного шаблона ветвления, который наилучшим образом описывает данные. Наконец, бывают ситуации - например, в случае рекомбинации, в которых данные, по-видимому, в некоторой степени поддерживают две или более разных топологий дерева. В таких случаях исследуемые последовательности могут быть лучше представлены сетью, такой как та, что изображена на рисунке 15 c.

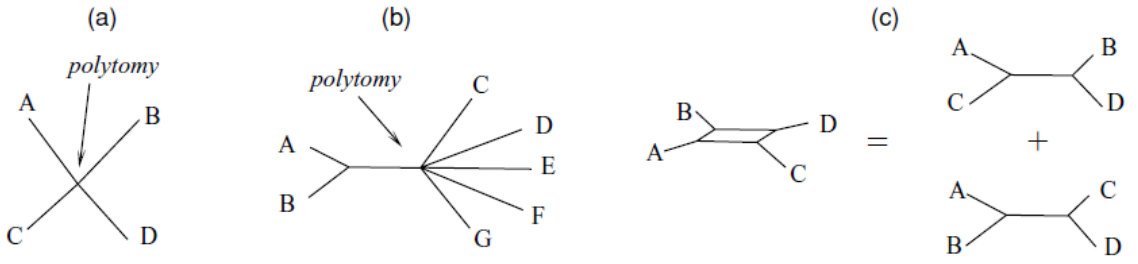

Рис. 15. Недиверсирующие деревья и сети; стрелки указывают на политомию. (A) звездообразное дерево; (B) дерево с внутренней политомией; (C) сеть.

8.9. Программы для построения филогенетических деревьев

Для вычисления генетических расстояний из последовательностей ДНК доступно большое количество программных пакетов. Полный список поддерживается Джо Фелсенштейном на странице http://evolution.genetics.washington.edu/PHYLIP/software.html.

Пакет Phylip был одним из первых пакетов бесплатного программного обеспчения для филогенеза (Felsenstein, 1993). Это пакет состоит из нескольких программ для расчета генетических расстояний и вывода филогенетических деревьев по разным алгоритмам. Полное описание пакета, в том числе инструкции по установке на разных машинах, можно найти по адресу http://evolution.gs.washington.edu/phylip.html. Основные программные модули Phylip кратко излагаются в таблице.

Таблица 3

Основные программные модули Phylip

Команда Phylip	Входные данные	Тип анализа
DNAdist.exe	Выровненные последовательности ДНК	вычисляет генетические расстояния, используя разные модели нуклеотидных замен
ProtDist.exe	Выровненные последовательности белков	вычисляет генетические расстояния, используя различные матрицы аминокислотных замен
Neighbor.exe	Генетические дистанции	строит NJ или UPGMA деревья
Fitch.exe	Генетические дистанции	строит Fitch–Margoliash деревья
Kitch.exe	Генетические дистанции	строит Fitch–Margoliash деревья на основании молекулярных часов
DNAML.exe	Выровненные последовательности ДНК	строит maximum likelihood деревья

ProtPars.exe	Выровненные последовательности белков	строит maximum parsimony деревья
SeqBoot.exe	Выровненные последовательности ДНК или белков	Генерирует результаты бутстрэп-анализа или выровненные последовательности ДНК или белков
Consense.exe	филогенетические деревья	Генерирует консенсусное дерево

Tree-Puzzle была первоначально разработана для восстановления филогенетических деревьев из молекулярной последовательности с использованием максимального правдоподобия с быстрым алгоритмом поиска деревьев (Strimmer & von Haeseler, 1995). Программа также вычисляет попарные расстояния максимального правдоподобия на основании ряда моделей нуклеотидных замен. Версии Tree-Puzzle для UNIX, MacOSX и Windows можно бесплатно загружать со страницы Tree-Puzzle по адресу: http://www.TREE-PUZZLE.de/ (рис. 16).

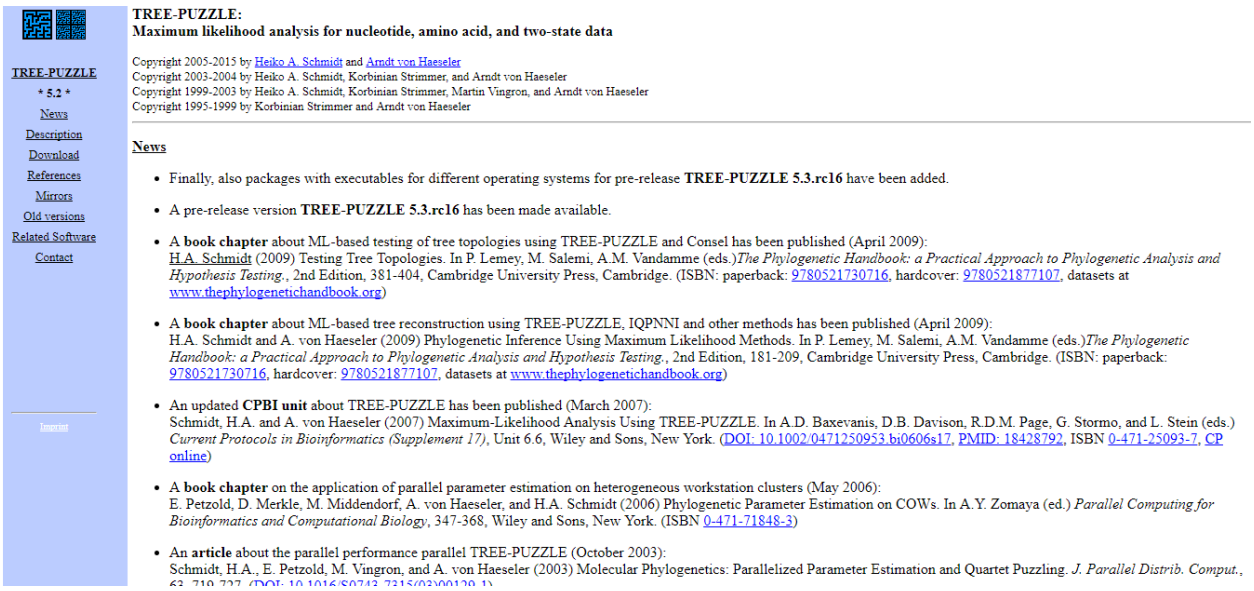

Рис. 16. Стартовая страница веб-сайта программы Tree-Puzzle

Любые программные модули в Phylip и Tree-Puzzle работает одним и тем же способом: им нужен файл, содержащий входные данные, например, выровненные последовательности ДНК в Phylip, который должен быть помещен в тот же каталог, где находится программа. Далее программа создает один или несколько выходных файлов в текстовом формате (обычно эти файлы называются outfile и outtree), содержащих результат анализа.

MEGA (Молекулярный эволюционный генетический анализ) - сложная программа, первоначально разработанная для проведения анализа дистанций и экономии как нуклеотидных, так и аминокислотных последовательностей (Kumar et al., 2012). Одно из преимуществ Mega - возможность вычисления стандартных ошибок оценок расстояния либо используя аналитические формулы, полученные для конкретной эволюционной модели, либо используя бутстрэп-анализ. Последняя версия программы также включает в себя отличный редактор данных, который позволяет выровнять несколько последовательностей с использованием встроенной реализации алгоритма Clustal и вручную отредактировать выровненные и невыровненные последовательности. Программное обеспечение является бесплатным и может быть загружено по ссылке http://www.megasoftware.net/overview.html. На веб-сайте также содержится подробный обзор возможностей программы, инструкции по установке и обширный онлайн-доступ к документации. Mega работает только под Windows, но его можно запустить на Mac с помощью эмулятора ПК (что, к сожалению, делает программу очень медленной) или в разделе Windows, установленном на новых компьютерах Mac с процессорами Intel.

8.10. Использование программы MEGA7 для построения филогенетических деревьев

Задание 5: построение филогенетического дерева по аминокислотным последовательностям *gyrB* для *Beggiatoa leptomitoformis* D-402

1. При помощи BLAST найти всех ближайших соседей исследуемой последовательности и сохранить их в формате FASTA (рис. 17)

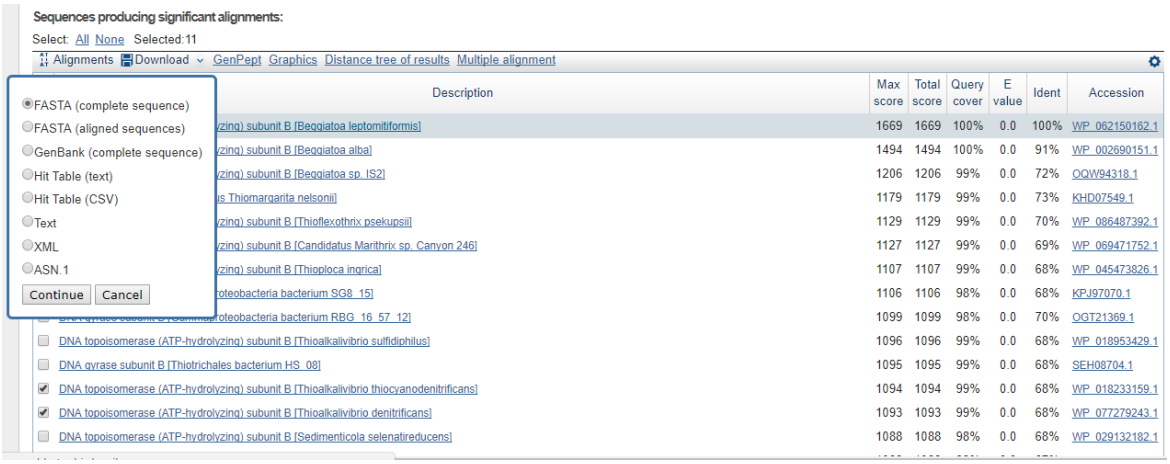

Рис. 17. Сохранение списка анализируемых последовательностей.

2. Запустить программу MEGA7 и создать новое выравнивание: Align – Edit/Build Alignment (рис. 18). В появившемся диалоговом окне выбрать «Create a new alignment» (рис. 19)

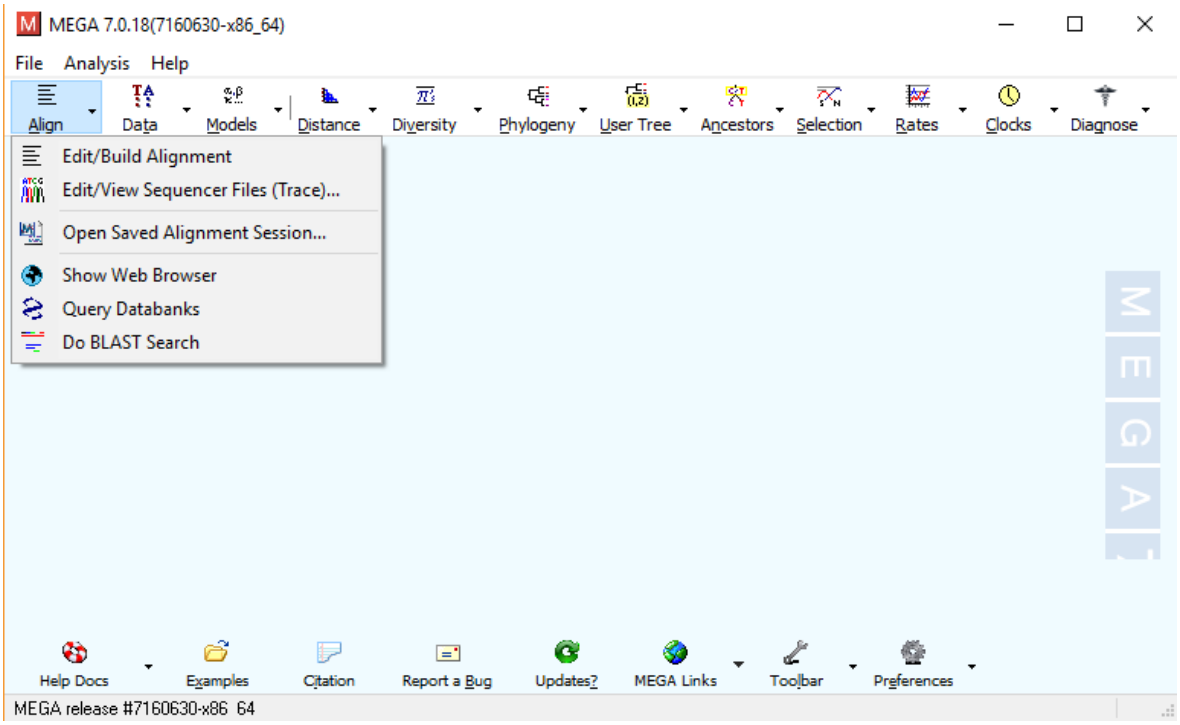

Рис. 18. Первый этап создания выравнивания в программе MEGA7

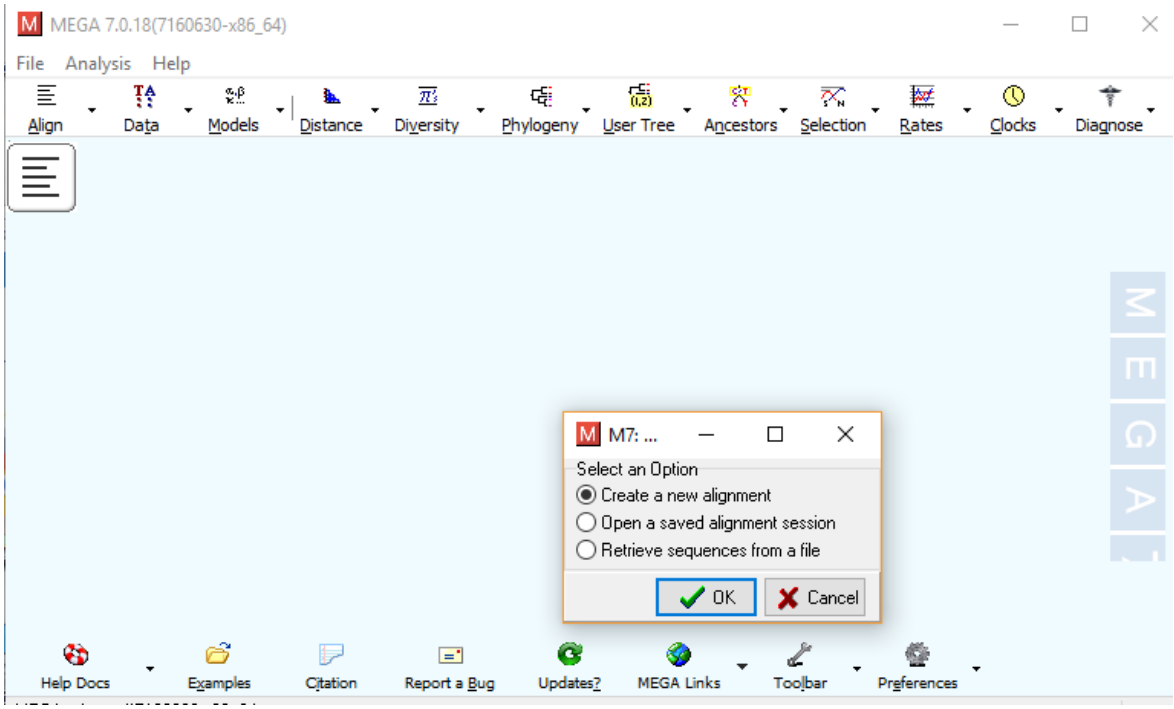

Рис. 19. Второй этап создания выравнивания в программе MEGA7

Появится диалоговое окно, в котором предлагается выбрать тип данных для выравнивания. Выбираем «Protein» (рис. 20).

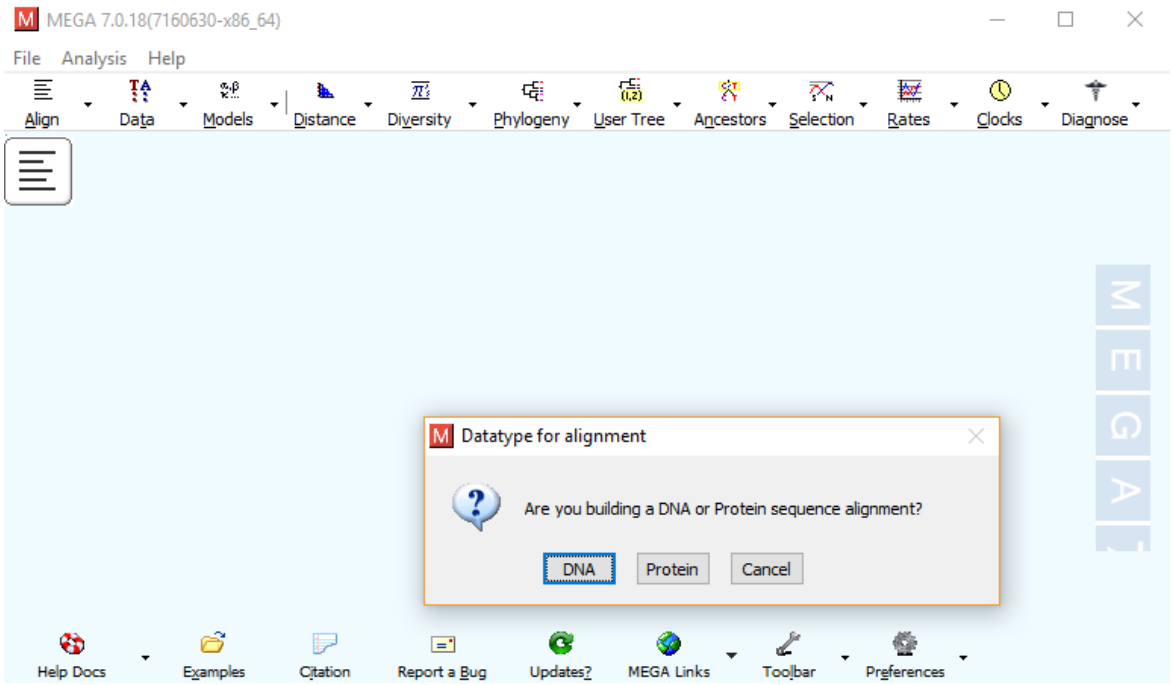

Рис. 20. Выбор типа данных для выравнивания.

Появляется вот такое окно (рис. 21):

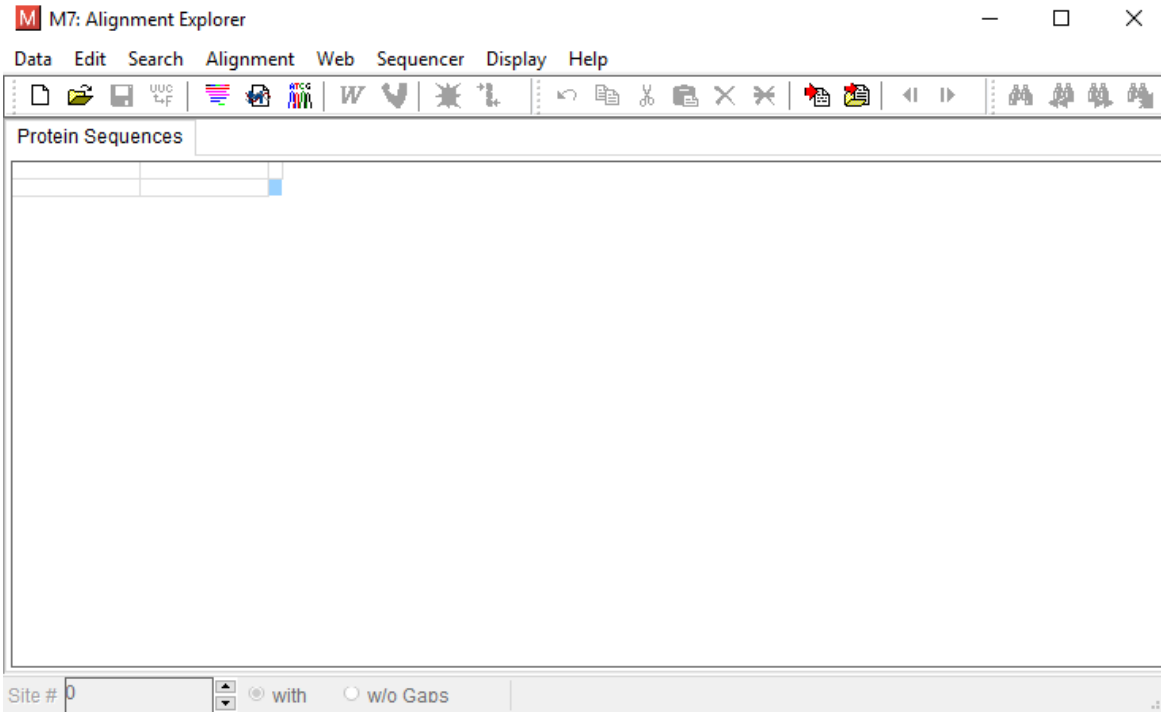

Рис. 21. Окно для вставки нового выравнивания в программе MEGA7.

Вставляем в него скачанные ранее последовательности и выравниваем. Для этого выбираем Alignment – Align by ClustalW (рис. 22).

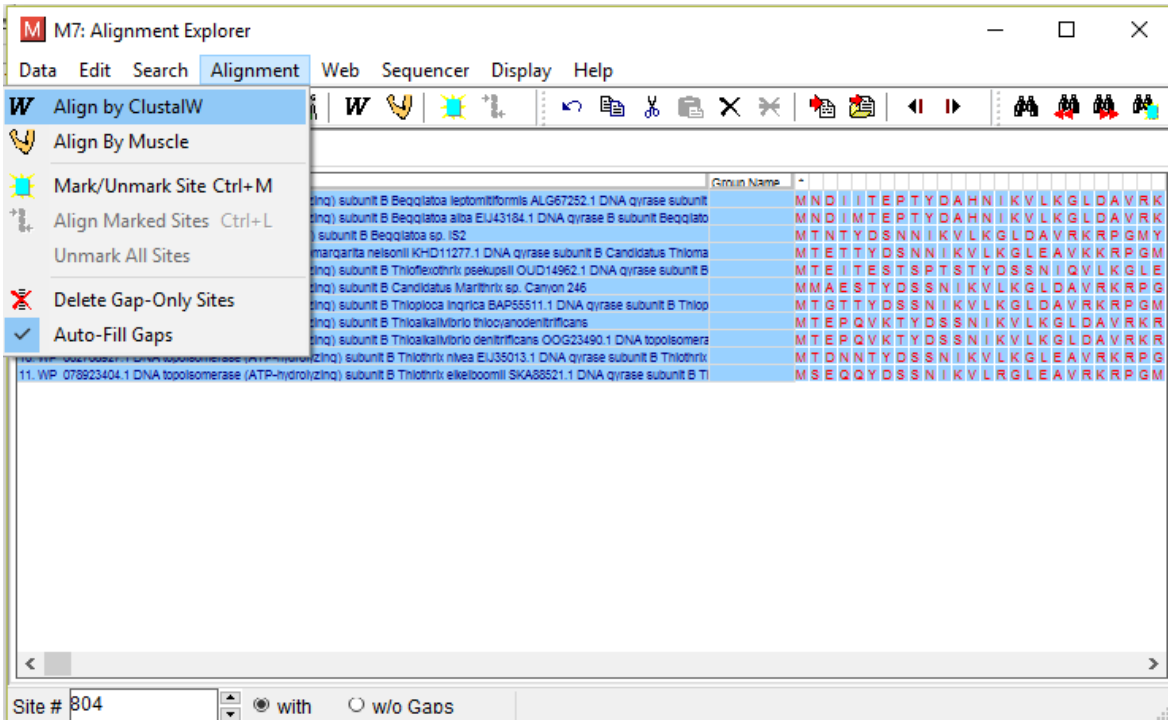

Рис. 22. Выравнивание последовательностей в программе MEGA7

После окончания выравнивания сохраняем его в формате MEGA в той папке, с которой мы будем работать. Data – Export Alignment – MEGA Format (рис. 23).

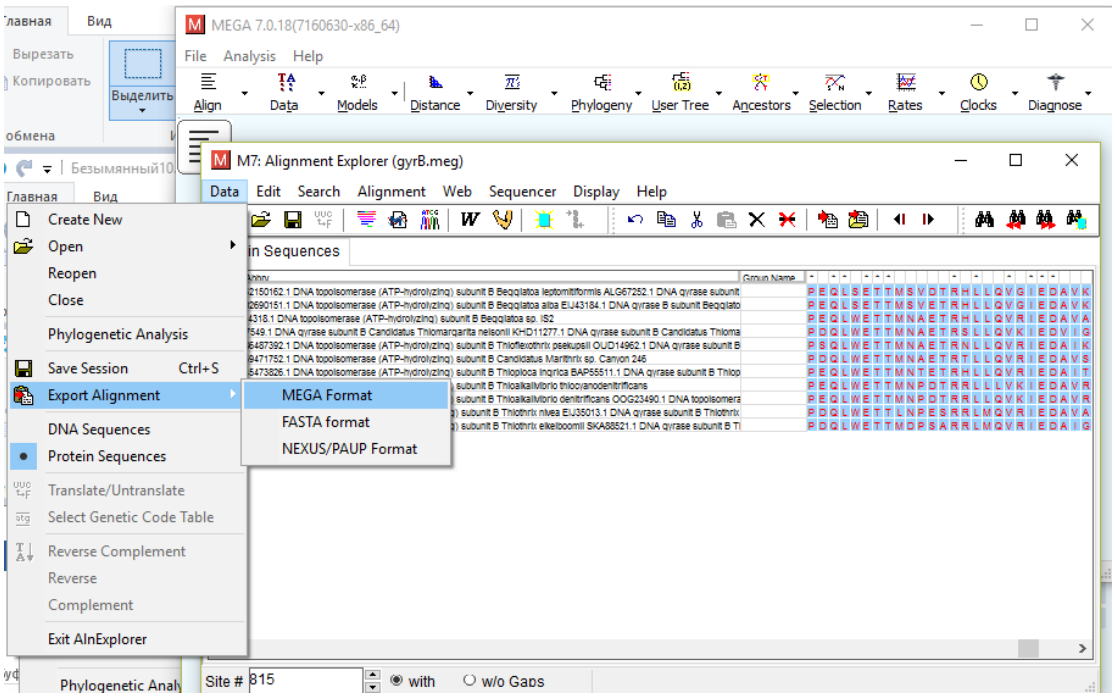

Рис. 23. Сохранение выравнивания в программе MEGA7

С сохраненным выравниванием можно работать несколько раз. Если уже после построения дерева мы решили, что какие-то гены можно исключить из анализа, не обязательно проводить заново всю процедуру. Достаточно открыть выравнивание и выделить только те последовательности, с которыми мы будем работать. Если же нам нужно включить в дерево последовательность, которой не было в выравнивании, необходимо найти ее, вставить в FASTA –файл с остальными последовательностями и снова провести выравнивания в программе MEGA. Поэтому лучше изначально сохранить избыточное количество сиквенсов, а затем исключать лишние из анализа.

Для построения филогенетического дерева выбираем Phylogeny – Construct/Test Neighbor-Joining Tree (рис. 24). Когда мы работаем с какими-

либо последовательностями в первый раз, лучше всего начинать именно с метода Neighbor-Joining, модель Джукса-Кантора, поскольку он наиболее быстрый и у него наименьшие требования к производительности компьютера, однако он дает лишь общие представления о филогении данной группы. Для подтверждения правильности топологии данной группы необходимо совпадение топологии деревьев, построенных двумя разными методами.

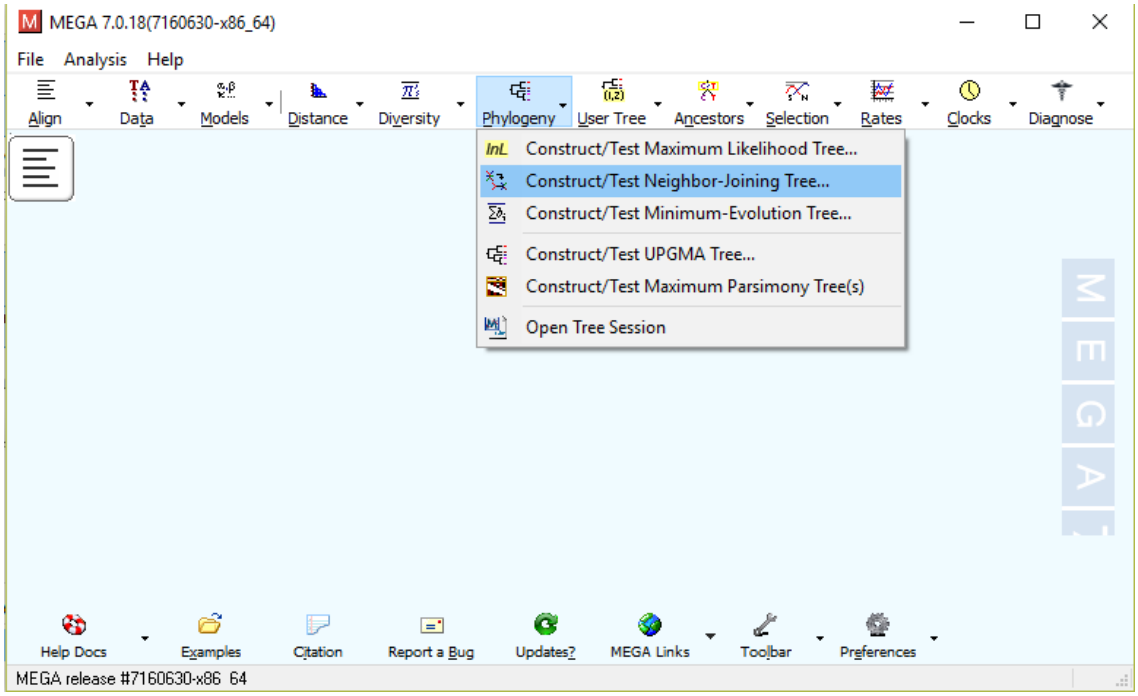

Рис. 24. Запуск филогенетического анализа в программе MEGA7

После того как мы выбрали эволюционную модель и запустили построение филогенетического дерева, появляется диалоговое окно, в котором мы можем указывать различные параметры конструируемого дерева. Строки, выделенные белым, редактировать нельзя. Если же строки выделены желтым, можно выбирать параметры из предложенного списка (рис. 25).

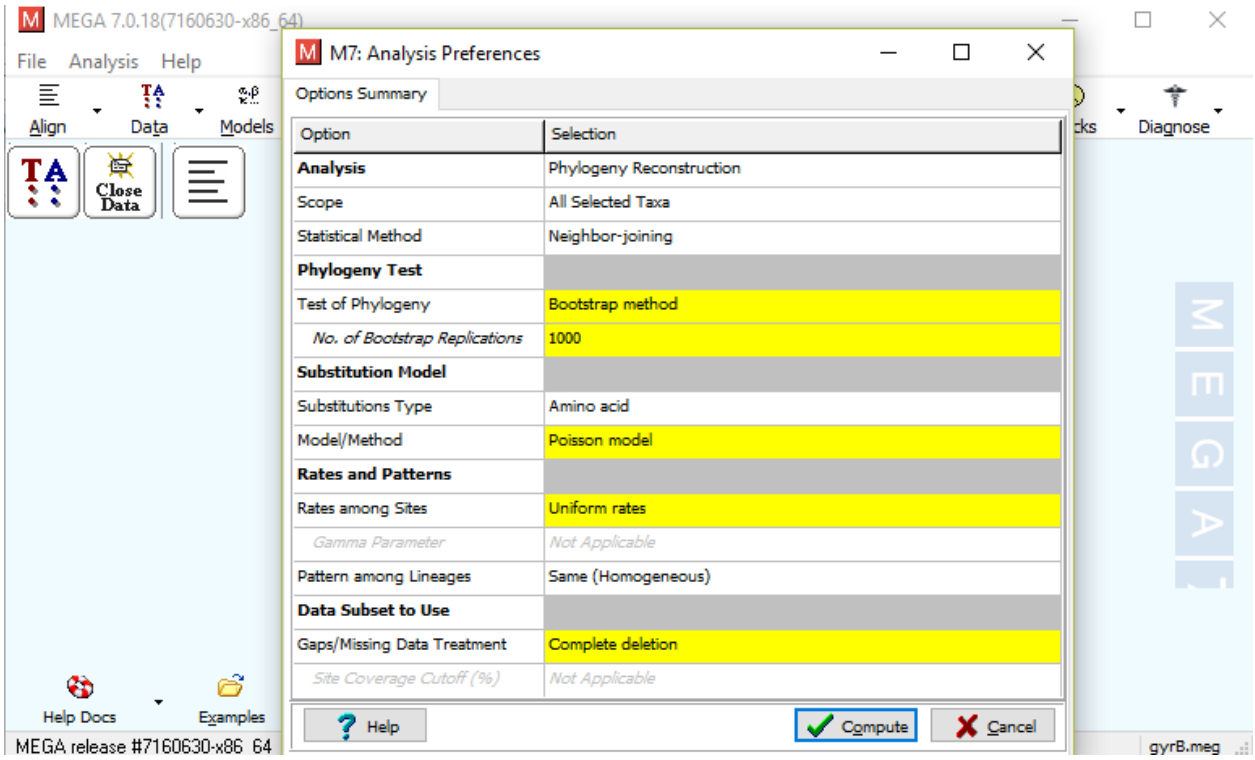

Рис. 25. Выбор параметров для построения дерева в программе MEGA7

В результате мы получаем эволюционное дерево (рис. 26)

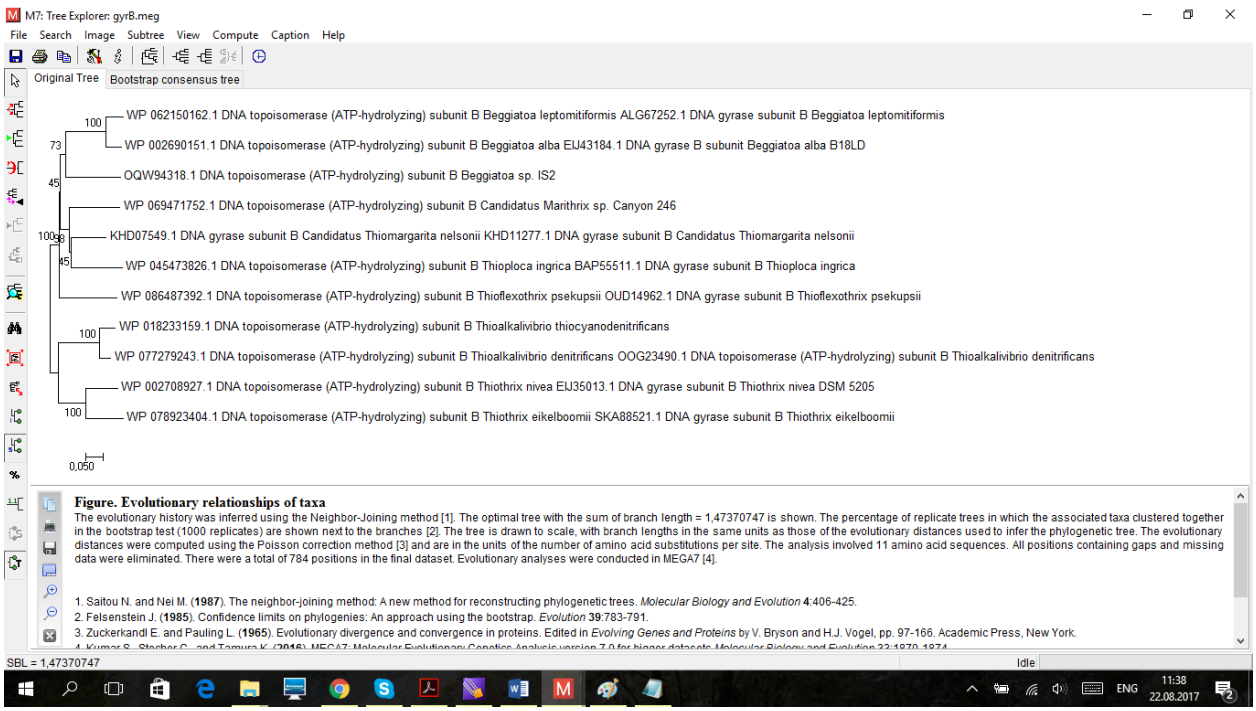

Рис. 26. Результат филогенетического анализа в программе MEGA7

Его можно скопировать и вставить в Microsoft Word. Оно вставится в виде рисунка. Щелкнуть по нему правой кнопкой мыши, выбрать изменение рисунка. Теперь можно форматировать наше дерево! Можно изменить размер и тип шрифта, удалить какую-то излишнюю часть информации о последовательностях, переставить цифры значения бутстрэпа, чтобы они не сливались с ветвями дерева и т.д. (рис. 27).

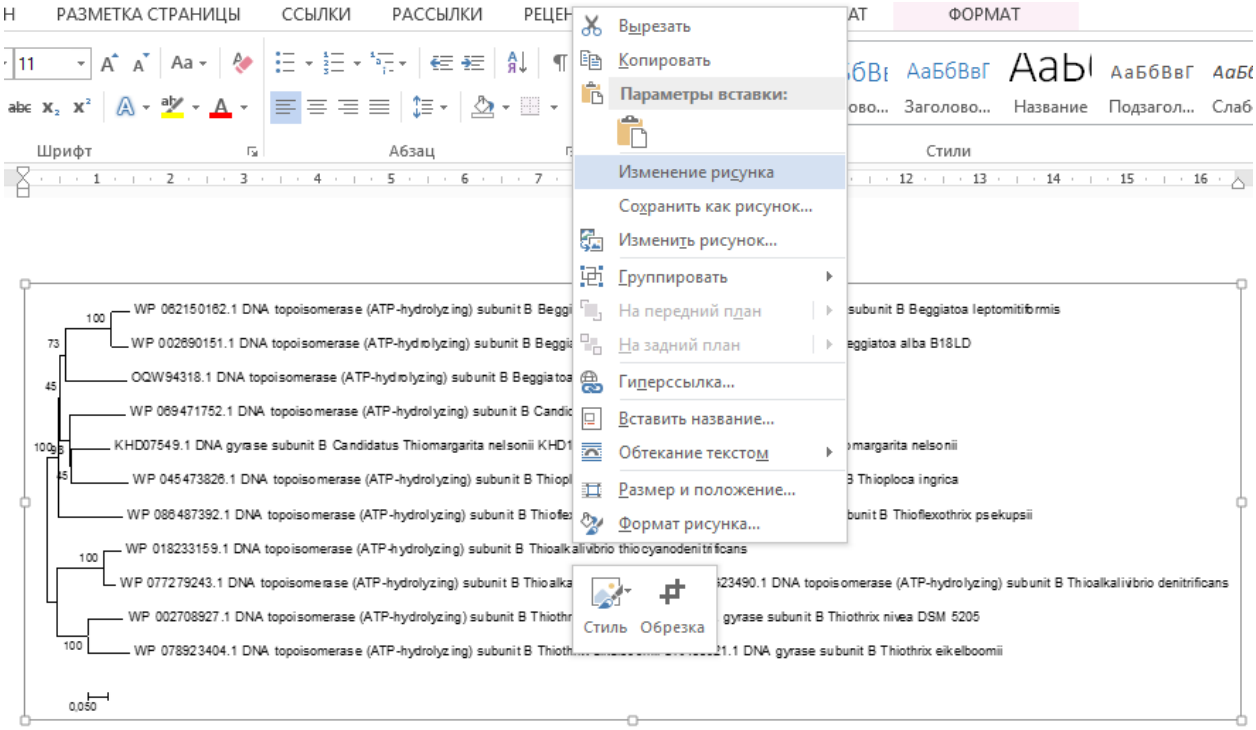

Рис. 27. Изменение филогенетического дерева в программе Microsoft Word

Можно не вставлять полученное дерево в Word, а сохранить его в формате PDF, TIFF или PNG: Image – Save as PNG/TIFF/PDF file. Но это не так удобно, потому что его нельзя будет отформатировать.

Перед выходом из программы MEGA обязательно сохраняем полученное дерево: File – Save Current Session. Получаем файл в формате .meg, который можно использовать для дальнейшего редактирования в программе.

8.10.1. Редактирование филогенетического дерева в программе MEGA7

Программа MEGA предоставляет широкие возможности по редактированию готового филогенетического дерева. Большое количество инструментов для манипуляции с ветвями дерева расположено во вкладке Subtree.

1. Можно укоренить дерево при помощи команды Root. Для этого выбираем соответствующую функцию и щелкаем мышкой по ветви дерева, которая по нашему мнению должна вести к внешней группе.
2. При помощи функций Flip и Swap можно вращать ветви дерева вокруг оси.
3. Команда Compress/Expend позволяет объединить все OTU внутренней ветви. Если нажать на внутреннюю ветку, MEGA предложит вам указать имя для группы, которая будет сформирована. Затем он сжимает все линии, определяемые этой ветвью, в сплошной удлиненный треугольник, толщина которого пропорциональна числу сконденсированных таксонов. Потом изменять имя группы можно при помощи команды Draw options.
4. При помощи команды Display in a window можно отобразить какую-либо внутреннюю ветвь в отдельном окне.

   Многие из этих функций также доступны на панели инструментов в левой части отображаемого дерева.

В меню View отображаются несколько вариантов просмотра:

1. Topology only: отображает дерево в виде отношений между таксонами, игнорируя длины ветвей.
2. Root on midpoint: укореняет дерево в средней точке самого длинного пути между двумя таксонами.
3. Arrange taxa: это позволяет упорядочить таксоны в дереве на основе порядка таксонов в файле входных данных или создать дерево, которое выглядит «сбалансированным».

4. Tree / Branch Style: позволяет выбрать отображение дерева в одном из трех стилей: традиционный, радиационный или круговой. Для традиционных существует три дополнительных варианта: прямоугольный, прямой или изогнутый (рис. 27).

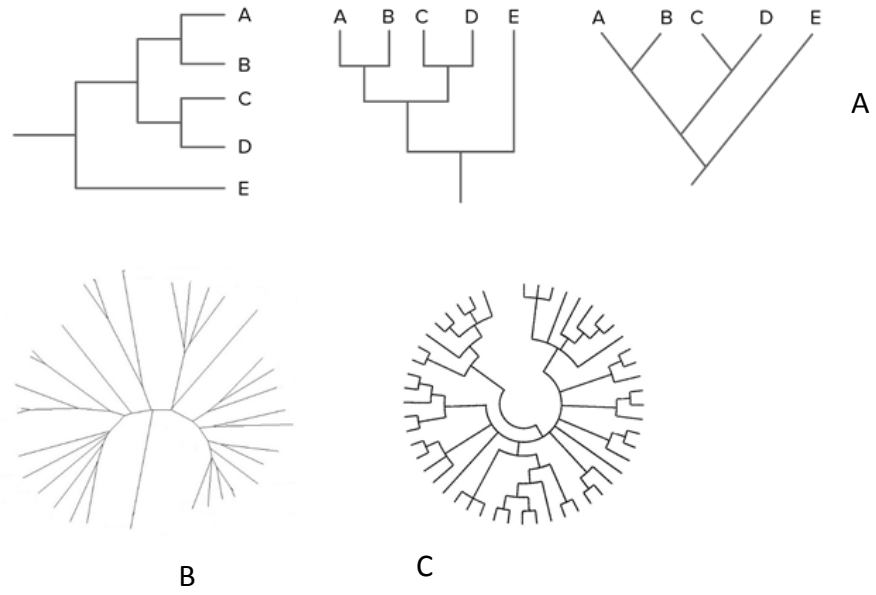

Рис. 27. Стили отображения дереьев. A – прямоугольный, B – прямой, C – изогнутый.

5. Show / Hide: позволяет отображать или скрывать следующую информацию: маркер таксона, статистика (например, значения бутстрэпа), длины ветвей, идентификаторы узлов, время расхождения (для временных графиков), охват данных (для временных графиков) или масштаб бар.
6. Fonts: эта команда позволяет выбирать такие функции, как тип и размер шрифта для указанной на дереве информации, включая подпись таксона, статистику и шкалу.
7. Goto Note/Branch: позволяет перейти к ветке с определенным номером. При этом выбранная ветка подсвечивается синим цветом.
8. Options: эта команда открывает диалоговое окно Option, которое обеспечивает контроль над различными аспектами рисования дерева, включая отдельные ветви, имена таксонов и шкалу шкалы.

Меню Compute: В этом меню можно рассчитать различные варианты дерева, включая сконденсированное дерево, временное дерево и консенсусное дерево.

1. Сконденсированное дерево. Когда внутренние ветви филогенетического дерева не имеют статистически значимой длины, выбор этой команды уплотняет дерево в топологию, в которой каждая ветвь с меньшей статистической значимостью удаляется.
2. Временное дерево. Временные деревья можно построить в MEGA, где время расходимости оценивается для всех точек ветвления в дереве, используя подход, основанный на методе RelTime (RelTime описан в Kumar et al., 2012), который не требует предположений о вариациях скорости линии.
3. Консенсусное дерево. Метод MP дает много одинаково экономных деревьев. Выбор этой команды создает составное дерево, которое является консенсусом среди всех таких деревьев (подробнее см. Nei & Kumar, 2000, с. 130).

Глава 9. Выбор для анализа ДНК или белка.

9.1. Интроны и некодирующая ДНК

При работе с ДНК-последовательностью необходимо выяснить, какие ее части используются для кодирования белка. Это может быть особенно проблематично для эукариот, геном которых содержит гораздо больше ДНК, чем требуется для кодирования белков; последовательность случайного фрагмента ДНК, скорее всего, не кодирует никакого белка. Эукариотические гены, в основном, состоят из экзонов с вкраплениями интронов. Из-за различий в эволюционном давлении на экзоны и интроны скорость появления базовых замен в этих двух элементах эукариотических генов может быть совершенно разной. Поэтому изучение эволюции белка с использованием сиквенса его ДНК должно включать только кодирующие по-

следовательности. Для этого требуется отредактировать все интроны в последовательности ДНК. Многие из этих проблем можно избежать, если использовать последовательность соответствующей мРНК вместо ДНК, поскольку отдельные экзоны объединяются в один непрерывный участок РНК, который содержит намного меньше посторонних материалов. Для большинства белок-кодирующих генов не только нуклеотидные последовательности, но и переведенная последовательность белка доступны в базе данных нуклеиновых кислот или в базе данных вторичного белка

9.2. Выбор ДНК или белка?

Если это возможно, рекомендуется проанализировать набор данных и для ДНК, и для белка; однако следует иметь в виду, что для очень отдаленных таксонов нуклеотидные последовательности, вероятно, потеряли филогенетическое значение. Напротив, для группы близкородственных таксонов, рекомендуется проводить анализ на основе ДНК, поскольку может быть меньше проблем, таких как различия в смещении кодонов или насыщении третьей позиции кодонов. Но все же стоит проводить параллельный анализ нуклеотидных и белковых последовательностей. Более того, там, где есть двусмысленности в выравнивании последовательностей генов, рекомендуется переводить последовательности сначала в белковые, затем выравнивать белковые последовательности и определять положения пропусков в последовательностях ДНК согласно более надежному выравниванию белков (Lemey et al., 2009).

Заключение

У молекулярных методов есть много преимуществ перед морфологическим анализом. Во-первых, ДНК содержит в себе множество данных, которые можно использовать в расчетах — ведь в генах могут содержаться сотни нуклеотидов. Чаще всего для оценки родства используют больше од-

ного гена, тогда как для анализа на основе морфологических данных используют несколько десятков признаков. Во-вторых, анализ ДНК считается более объективным. Дело в том, что морфологические признаки разные люди могут трактовать по-разному, тогда как нуклеотиды всегда одинаковы. В-третьих, ДНК можно использовать как для анализа групп высоких рангов, так и для выяснения отношений между видами, и даже между отдельными индивидами. Морфологический же анализ более достоверен при работе с таксонами высоких рангов, чем на уровне видов, просто потому, что чем выше ранг, тем лучше отличаются группы, и тем легче отличить аналогичный признак от гомологичного.

В частности, анализ консервативных последовательностей рибосомальных РНК микроорганизмов позволил установить, что все живое на Земле делится не на два царства, как считали несколько десятилетий назад, — эукариот и бактерий, — а на три: эукариот, бактерий и архей. Морфологическое сходство бактерий и архей с лихвой окупается огромной разницей их молекулярного устройства. Честь этого открытия принадлежит Карлу Вёзе.

Несмотря на то, что преимущество молекулярного анализа кажется вполне обоснованным, есть все же и несколько причин, по которым морфологию нельзя отправить «в отставку».)

Первая причина заключается в том, что не каждый организм подходит для выделения ДНК. Он должен быть собран и сохранен специальным образом, иначе эта молекула просто разрушается. Множество редких и интересных видов было описано много десятков лет назад, когда еще даже про ДНК ничего не знали, и в наши дни не очень понятно, где их искать и как собирать.

Вторая причина заключается в том, что далеко не всегда результаты молекулярных филогенетических методов вызывают доверие. Иногда быва-

ет так, что они не совпадают с устоявшимися «классическими» взглядами. Это, конечно, не означает, что именно молекулярные данные неверны, просто такие несовпадения являются «звоночком», что где-то закралась ошибка. Несовпадения могут быть не только из-за ошибок в самом анализе, но и из-за того, что были неправильно выбраны гены. Гены, мутирующие с высокой скоростью, подходят для выяснения родства между видами, но не походят для анализа групп более высоких рангов. Но гомологичные гены в разных группах организмов могут меняться с разной скоростью, поэтому гены, подходящие для анализа одной группы, могут не подходить для другой группы того же ранга. В общем, подбор нужных участков ДНК может оказаться не очень легкой работой, особенно если учесть, что далеко не все гены у всех видов хорошо изучены.

Третья причина — это высокая стоимость секвенирования генов. Для построения филогении одного небольшого рода можно легко потратить пару тысяч долларов. А если учесть, что гены не всегда подбирают правильно с первого раза, или некоторые экземпляры оказываются непригодными для секвенирования, то анализ надо проводить повторно, и цена может быть больше, чем предполагалось изначально. Анализ же на основе морфологических признаков обходится гораздо дешевле.

Библиографический список

3. Лукашов В.В. Молекулярная эволюция и филогенетический анализ [Текст]: учеб. Пособие / Лукашов В.В. — М.: БИНОМ, 2009. – 256 с.
4. Современные методы выделения, культивирования и идентификации зеленых водорослей (Chlorophyta) [Текст] / А.Д. Темралеева [и др.]. - Кострома: Костромской печатный дом. – 215 с.
5. Altschul S.F. Basic local alignment search tool [Text] / Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. // J Mol Biol. - №3. – p. 403-410
6. Altschul S.F. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [Text] / Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. // Nucleic Acids Res. – 1997. - №25. – p. 3398-3402
7. Baum D.A. Tree thinking: an introduction to phylogenetic biology [Text] / Baum D.A., Smith S.D. - Greenwood Village, CO: Roberts and Company Publishers, Inc. - 2012. – 496 p.
8. Bruno W.J. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction [Text] / Bruno W.J., Socci N.D., Halpern A.L. // Mol. Biol. Evol. – 2000. - №17. – p. 189-197.
9. Darriba D. jModelTest 2: more models, new heuristics and parallel computing [Text] / Darriba D., Taboada G.L., Doallo R., Posada D. // Nat Methods. – 2012. - №9. – p. 1-4.
10. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. / Department of Genetics, University ofWashington, Seattle. – 1993.
11. Kumar S. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis [Text] / Kumar S., Stecher G., Peterson D., Tamura K. // Bioinformatics. – 2012. - №20.- p. 2685-2686

12. Lemey P. The Phylogenetic Handbook [Text] / Lemey P., Salemi M., Vandamme A.-M. - New York: Cambridge University Press. – 2009. – 751 p.
13. Li,W.-H. Molecular Evolution. / Li,W.-H. - Sunderland, MA, USA: Sinauer Associates, Inc. – 1997. – 487 p.
14. Maddison D.R. NEXUS: an extensible file format for systematic information. / Maddison D.R., Swofford D.L., Maddison W.P. // Syst Biol. – 1997. - №4. P.590-621.
15. Nei M. Molecular Evolution and Phylogenetics [Text] / Nei M., Kumar S. –Oxford: Oxford University Press. – 2000. – 333p.
16. Page R. Molecular Evolution: A Phylogenetic Approach [Text] / Page R., Homes E. - Oxford, UK: Blackwell Science Ltd. – 1998. – 352 p.
17. Paradis E. APE: Analyses of Phylogenetics and Evolution in R language [Text] / Paradis E., Claude J., Strimmer K. // Bioinformatics. - №2. – p. 289–290
18. Strimmer K. Performance of the maximum likelihood, neighbor joining, and maximum parsimony methods when sequence sites are not independent [Text] / Strimmer K., von Haeseler A. // Systematic Biology. 1995 - №4 – p. 533-547.
19. Yang Z. Molecular phylogenetics: principles and practice [Text] / Yang Z.,Rannala B. // Nat. Rev. Genet. – 2012. - № 13. – p. 303-314