МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ»

М.В. Орлова,
М.Ю. Грабович

МЕТОДЫ ИЗУЧЕНИЯ ФИЛОГЕНИИ ПРОКАРИОТ

Учебное пособие

Воронеж
Издательский дом ВГУ
2017

ηqdbqүеʏ bɆ dеbƎbʏзе d єɑɑбɔʎ ɟqddʏqʎè

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 5

1. ДАННЫЕ, ИСПОЛЬЗУЕМЫЕ В ФИЛОГЕНЕТИЧЕСКОМ АНАЛИЗЕ 6

2. БАЗЫ ДАННЫХ НУКЛЕОТИДНЫХ И АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ 8

2.1. Общие базы данных нуклеиновых кислот 10

2.1.1. Имя записи, имя локуса или идентификатор (ID) 12

2.1.2. Номер доступа (AC) 12

2.1.3. Номер версии 12

2.1.4. Номер GenInfo (только GenBank) 13

2.1.5. Полногеномные последовательности (WGS) 13

2.1.6. Сторонние аннотации (TPA) 13

2.2. Общие базы данных белковых последовательностей 13

2.3. Специализированные базы данных последовательностей, справочные базы данных и базы данных генома 15

2.4. Комбинированные базы данных, средства зеркального отображения и поиска баз данных. 17

2.4.1. Entrez 17

3. ФОРМАТЫ ФАЙЛОВ 22

4. ЭТАПЫ ФИЛОГЕНЕТИЧЕСКОГО АНАЛИЗА 27

5. ВЫРАВНИВАНИЕ ГЕНЕТИЧЕСКИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ 28

5.1. Выравнивание BLAST 28

5.2. Множественное выравнивание 31

5.3. Выравнивание Clustal 32

5.4. Выравнивание T-Coffee 34

5.5. Выравнивание MUSCLE 35

6. РАСЧЕТ ГЕНЕТИЧЕСКИХ ДИСТАНЦИЙ 37

7. МОДЕЛИ НАКОПЛЕНИЯ ЗАМЕН. 38

8. ФИЛОГЕНЕТИЧЕСКИЕ ДЕРЕВЬЯ 41

8.1. Структура филогенетического дерева 41

8.2. Количество возможных деревьев 42

ηqdbqүеʏ bɆ dеbƎbʏзе d єɑɑбɔʎ ɟqddʏqʎè

ŋ̥qdqyeY bƎ dəqƎbУзe d əɑɔqɔɣ ɟdqqYqyə

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 информацию 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

ŋ̥qdqyeY bƎ dəqƎbУзe d əɑɔqɔɣ ɟdqqYqyə

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).

2. БАЗЫ ДАННЫХ НУКЛЕОТИДНЫХ И АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Филогенетические анализы часто основаны на данных, накопленных многими исследователями. 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.

1. Чтобы легко получить известную последовательность, можно положиться на уникальные идентификаторы последовательности.

2. Собирать полный набор последовательностей, которые разделяют таксономическое происхождение или известное свойство, можно по ключевому слову в аннотации.

3. 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-PS, WU-BLAST и fastA в EBI	*http://www.ebi.ac.uk/Tools/ similarity.html*
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/*

ђqdbyeY bƎ debƎbY3e d eaobcy ЗqdbYqyè

Окончание табл. 2

Название базы данных или инструмента поиска	Интернет-адрес
Геномный браузер UCSC	*http://genome.ucsc.edu/*
UniGene	*http://www.ncbi.nlm.nih.gov/UniGene/*
UniProt в EMBL	*http://www.ebi.ac.uk/uniprot/*
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 обеспечение для анализа последовательностей обрабатывает U и T, не делая различия (Lemey et al., 2009).

ђqdbyeY bƎ debƎbY3e d eaobcy ЗqdbYqyè