Лекции

Лекция 1 Базы данных

План лекции:

1. Введение
2. Краткий обзор основных баз данных по геному человека
3. Работа с BLAST
4. Молекулярно-генетическая реконструкция филогении. Основные сведения
5. Выравнивание молекулярно-генетических последовательностей
6. Методы вычисления расстояний между последовательностями
7. Методы построения филогенетических деревьев
8. Методы оценки филогенетических деревьев

Источники

http://www.jcbi.ru/baza/

http://www.microarray.ru/

Термины

Граф выравнивания, матрица сходства, штраф за делецию, Граф локального выравнивания, локальное выравнивание с аффинными штрафами, Расширенный граф локального выравнивания, штраф за расширение делеции,

Введение

В курсе изучаются методы прикладной математики, прежде всего статистики, и информатики для решения проблем молекулярной биологии, возникающих, в частности при моделировании процессов эволюции и оптимизации селекционного процесса.

Основные задачи курса:

1. Поиск сходства нуклеотидных или аминокислотных последовательностей;
2. Анализ генома (определение белок – кодирующих участков, а также участков, кодирующих тРНК и рРНК; поиск участков ДНК, которые отвечают за регуляцию – сайты связывания регуляторных белков и др.);
3. Предсказание вторичной структуры РНК;
4. Предсказание структуры белков по их аминокислотным последовательностям;
5. Филогенетическое сравнение форм – выяснение их родства.
6. Создание и поддержание баз данных, инструментов для работы с ними, а также методов обработки массовых экспериментов.

На каждом этапе необходимо применение генетико–математических моделей, методов и специальных компьютерных программ. Для предсказания кодирующей части генов используют программы, в основе которых лежит сравнение изучаемой последовательности с последовательностями известных белков, мРНК или ДНК, кодирующей гомологичные гены. Однако такие программы не всегда могут обнаружить гены, специфичные для нового генома, поэтому возникает необходимость дополнительно использовать сложный статистический анализ. Зная предполагаемую структуру гена, можно провести анализ структуры и функции кодируемого им белка.

Решение поставленных задач невозможно без использования баз данных. Но поскольку молекулярно–генетических баз данных большое количество, многие имеют свой формат хранения данных и

средства доступа к содержащейся в ней информации, то существует проблема интеграции. Возникает задача создания стандартов и программных средств, которые позволят пользователю быстро находить информацию на основе компьютерного анализа многих баз данных.

Поскольку решение поставленных задач предполагает использование различных программ и алгоритмов для анализа последовательностей, то возникает задача статистической оценки достоверности, надежности полученных выводов. Для этого можно использовать известные статистические критерии.

В настоящее время в сети Интернет существуют сотни баз данных, которые доступны для поиска данных по молекулярной биологии и другим смежным дисциплинам. Каждая из них имеет свой формат хранения данных, различную степень избыточности, взаимосвязи с родственными или аналогичными базами данных. Каждая база данных имеет также свои средства доступа к информации – различные поисковые программы, программные средства визуализации, пополнения базы.

Крупнейшие хранилища первичных структур ДНК и аминокислотных последовательностей (такие, как EMBL, GenBank, DDBJ, SWISS-PROT, Ensembl и др.) пополняются аннотированными последовательностями непосредственно исследователями, расшифровавшими их, с помощью автоматизированной системы пополнения баз данных по сети Интернет.

Конечно, впоследствии эти данные проверяются персоналом администраций баз данных и существенно пополняются. Вторым основным источником информации во всех базах является специальная научная литература. Многие базы данных, работающие над коллекционированием однородной информации, координируют свои усилия, осуществляя международное разделение труда, это можно проиллюстрировать примером сотрудничества трех всемирных коллекций последовательностей нуклеотидов EMBL (Европа), GenBank (США), DDBJ (Япония).

Наряду с общими базами данных в последнее время появилось много специализированных информационных ресурсов. Многие из них хранят данные, полученные с помощью компьютерных методов, результаты теоретических предсказаний. Большую роль в биоинформатике играют хранилища последовательностей ДНК и кДНК, специализированные базы данных по отдельным регуляторным мотивам нуклеотидных последовательностей, базы данных по экспрессии генов, библиотеки геномов, карт, последовательностей РНК, белков, белковых мотивов, по продукции белков. Есть базы данных по протеомике, структурам белков, мутациям, метаболическим путям и регуляции, по трансгеннным организмам, анатомии, биохимии, а также по научной литературе, по существующему в этих областях исследований программному обеспечению.

Будет дано общее представление о существующих базах данных по геному человека и более детально рассматриваются некоторые БД, список которых будет постоянно пополняться.

Краткий обзор основных баз данных по геному человека

OMIM – On-line Mendelian Inheritance in Man.
Крупнейшая база данных по человеческим генам и генетическим заболеваниям, создал базу доктор МакКасик (Victor A. McKusick) с коллегами в центре медицинской генетики (Johns Hopkins University, Baltimore, USA), NCBI поддерживает наполнение и обновление базы. Содержит общие обзоры по заболеваниям и конкретным генам, а также ссылки на базы данных ENTREZ.
Адрес: http://www.ncbi.nlm.nih.gov/omim/

GenBank.
GenBank – база данных генетических последовательностей, поддерживается NIH (Национальный Институт Здоровья США), аннотированная база известных последовательностей ДНК, РНК и белков, с литературными ссылками на первоисточники и информацией биологического характера. Обновляется каждые два месяца. Является частью International Nucleotide Sequence Database Collaboration, которая объединяет три крупнейшие коллекции нуклеотидных последовательностей: DDBJ (NIG), EMBL (EBI) и GenBank (NCBI). Три организации осуществляют разделение труда и ежедневно обмениваются новой информацией. Большинство журналов требуют предварительной посылки последовательностей в любую из этих трех баз данных до опубликования статьей о них. В статьях, посвященных очередной порции секвенированных последовательностей, должен упоминаться лишь номер последовательности в базе данных. NCBI постоянно совершенствует и создает новые средства для помещения новых последовательностей в базу, средства эффективного поиска в базе.
Крупнейшая интегрированная поисковая система ENTREZ для нуклеотидных и аминокислотных последовательностей, библиографии (PubMed), полных геномов (Genomes), а также трехмерных структур

белков (MMDB) создана и поддерживается NCBI. При этом поиск ДНК и белков не ограничивается только ресурсами GenBank, но и другими доступными по сети хранилищами информации.
Адрес: http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html

EMBL – the EMBL Nucleotide Sequence Database.
База данных нуклеотидных последовательностей Европейской Молекулярно-Биологической Лаборатории пополняется большей частью непосредственно авторами, определившими первичную структуру фрагмента ДНК или РНК и, кроме последовательности нуклеотидов, содержит разнообразную информацию о каждом фрагменте, включая литературные ссылки, перекрестные ссылки на документы других баз данных, таблицы особенностей и др. Существует с 1982 года. База данных – продукт сотрудничества EMBL (ФРГ), GenBank (США) и DDJP (Япония), каждая из этих трех групп собирает свою порцию информации из всех возможных мировых источников, ежедневно обмениваясь новыми и обновленными документами друг с другом. Удобна своей географической близостью для доступа на территории Европы. В России есть сайт, на котором хранится ежедневно обновляемая копия базы (http://www.genebee.msu.su/, отв. Скулачев В.П.).
Адрес: http://www.ebi.ac.uk/embl/

HGMD – Human Gene Mutation Database.
Содержит информацию обо всех опубликованных повреждениях генов, приводящих к наследственным заболеваниям у человека. Документы базы аннотируют все гены, находящиеся в ядре. Гены митохондриального генома и соматические мутации исключены. Мутации, выявленные на уровне белкового сиквенса, не входят в базу чтобы избежать ошибок из-за отсутствия анализа на уровне ДНК. Молчащие мутации, не приводящие к изменению аминокислотной последовательности тоже исключены. С марта 1999 года включены данные о полиморфизме, связанном с болезнями. Данные берутся из тех же самых журналов, что и данные о мутациях (>250). Сопровождается Институтом медицинской генетики (University of Wales, Cardiff, UK).
Адрес: http://www.hgmd.cf.ac.uk/ac/index.php

KEGG – Kyoto Encyclopedia of Genes and Genomes.
Попытка компьютеризировать все современное знание в молекулярной и клеточной биологии в терминах информационных путей. Это база знаний по систематическому анализу функций генов. Создается институтом химических исследований (Kyoto University, Japan) в рамках японской программы по геному человека. Содержит 6 баз данных – метаболических путей (PATHWAY), генов (GENES) и лигандов (LIGAND), экспериментальных данных по экспрессии генов (EXPRESSION и BRITE), по белкам (SSDB) и обширные возможности для работы со всеми крупными мировыми информационными ресурсами. Базы данных KEGG представляют данные в виде графических диаграмм, включающих большинство метаболических путей и некоторые из наиболее известных регуляторных путей. Кроме того, информация о путях представлена в виде таблиц ортологов, которые содержат как гены-ортологи, так и паралоги из различных организмов. Обновляются базы ежедневно.
Адрес: http://www.genome.ad.jp/kegg/

UniGene.
База данных, которая содержит кластеры похожих последовательностей. Каждый кластер представляет один ген и содержит попутную информацию, например, название ткани, где этот ген экспрессирован. Кроме хорошо известных генов в базу данных включены сотни тысяч новых концов экспрессирующихся последовательностей (EST – expressed sequence tags). Служит для поиска генов в новых последовательностях, а также для определения реагентов при секвенировании генов и их экспрессии. Кластеризация осуществляется автоматически.
Адрес: http://www.ncbi.nlm.nih.gov/unigene

PROSITE – PROtein SITEs and patterns dictionary.
База данных различных паттернов функциональных и регуляторных участков. С помощью этой коллекции можно определить, принадлежит ли, и к какому именно, семейству белков новая последовательность пользователя, или какой важный домен она содержит. Версия 17.21 этой базы, датированная сентябрем 2002 года содержит 11148 единиц хранения, которые описывают 1568 различных паттернов, правил и матриц.
Адрес: http://www.expasy.ch/prosite/

SWISS-PROT|UniProt – the protein sequence data bank.

База данных содержит аннотированные аминокислотные последовательности, транслированные с нуклеотидных последовательностей EMBL; адаптированные последовательности из PIR; а также последовательности, опубликованные в литературе и присланные непосредственно самими авторами. Содержит высококачественные неизбыточные аннотации, перекрестные ссылки на другие родственные базы данных (EMBL, Prosite, PDB). Каждая аннотация содержит описание функции белка, его доменной структуры, особенностей пост-трансляционной модификации, различные варианты. Имеется неаннотированное приложение (TrEMBL). Поодерживается Женевским университетом (Department of Medical Biochemistry of the University of Geneva) и EMBL. (EBI). Для академических пользователей – бесплатна.
Сайт не так давно обновился и немного поменялась структура поисковых запросов. На мой взгляд стало гораздо красивей и удобней.
Адрес: http://www.uniprot.org/

trEMBL – EMBL protein-coding DNA sequence features translated into peptide sequences.
База данных, созданная автоматически, представляет собой приложение к SWISS-PROT. Содержит аминокислотные последовательности, транслированные программно с нуклеотидных кодирующих участков, взятых из базы данных EMBL.
Адрес: http://www.uniprot.org/

ENSEMBL
Ensembl – совместный проект EMBL – EBI и Sanger Centre с целью создания программной системы для автоматической аннотации эукариотических геномов. Осуществляет (бесплатно) следующие возможности: поиск ДНК из человеческого генома, обзор хромосом, поиск белков и белковых семейств. Проект Ensembl стремится обеспечивать соответствие следующим критериям: точный, автоматический анализ данных генома; анализ и аннотациии основаны на текущих, своевременно обновляемых данных; доступность полученных данных для всех через сеть Интернет; предоставление данным другим лабораториям по биоинформатике. Основной акцент в базе данных Ensembl сделан на позвоночных геномах, но другие группы адаптировали систему для использования с растительными и грибковыми геномами.
Адрес: http://www.ensembl.org/

Базы данных можно отнести к следующим типам:
1) Архивные.
К архивным относятся, например, базы данных GeneBank, EMBL, PDB. Любой исследователь может поместить туда свою информацию. За содержание каждой записи в таких базах отвечает сам исследователь. GenBank – база данных генетических последовательностей, основанная в 1982 году. Это аннотированная коллекция всех общедоступных последовательностей ДНК, РНК и белков, снабженных литературными ссылками, и другой биологической информацией. Эта база является частью объединения International Nucleotide Sequence Database Collaboration, которое объединяет три крупнейшие коллекции нуклеотидных последовательностей: DDBJ (DNA Data Bank of Japan), EMBL (European Molecular Biology Laboratory) и GenBank (National Center for Biotechnology Information). Эти три организации ежедневно обмениваются новой информацией. Большинство журналов требуют предварительной посылки новых секвенированных последовательностей в любую из этих трех баз данных до опубликования статьей о них. В статьях, посвященных очередной порции последовательностей, должен упоминаться лишь номер последовательности в базе данных GenBank.
Адрес DDBJ: http://www.ddbj.nig.ac.jp/
Адрес GenBank: http://www.ncbi.nlm.nih.gov/Genbank/
EMBL (European Molecular Biology Laboratory) – эта база данных содержит разнообразную информацию о каждом фрагменте последовательностей, включая литературные ссылки, перекрестные ссылки на документы других баз данных и др.
Адрес EMBL: http://www.ebi.ac.uk/embl/
Еще одна архивная база данных – PDB (Brookhaven Protein DataBank) – содержит данные о коллекции экспериментально определенных трехмерных структур биологических макромолекул (белков и нуклеиновых кислот). С 2002 года в основном депозитарии PDB хранятся структуры, экспериментально определенные с помощью рентгеноструктурного, ядерно–магнитнорезонансного и др. методов. Теоретические структуры выделены в отдельную подбазу PDB.
Адрес: http://www.rcsb.org/pdb/

2) Курируемые базы данных.

За содержание записей в таких базах данных отвечают кураторы. Информацию для курируемых баз данных отбирают эксперты из архивных баз.
К курируемым базам относятся, например, SwissProt. Эта база данных белковых последовательностей существует с 1986 года и поддерживается двумя институтами: Swiss Institute of Bioinformatics (SIB) и European Bioinformatics Institute (EBI).
Адрес: http://www.ebi.ac.uk/swissprot/

3) Автоматические базы данных.
В таких базах данных записи генерируются (моделируются) компьютерными программами.
К ним относится, например TrEMBL (Translated EMBL) – автоматическая база предсказаний последовательностей белков. Это формальная трансляция всех кодирующих нуклеотидных последовательностей из банка EMBL.
В 2002 году в результате объединения SwissProt, TrEMBL и PIR был создан банк данных UniProt (Universal Protein Resourse). Это основное хранилище белковых последовательностей и их функций.
UniProt состоит из трех частей:
UniProt Knowlegebase – является центральной базой данных и обеспечивает доступ к обширной курируемой информации по белкам, включая их функцию, классификацию и перекрестные информационные ссылки;
UniProt Archive – UniParc. Отражает хронологию данных определения о всех белковых последовательностях;
UniProt Reference – UniRef. Содержит базы данных, которые объединяют последовательности в кластеры для ускорения поиска.
Адрес UniProt: http://www.ebi.uniprot.org/index.shtml

4) Производные базы данных.
Они получаются в результате компьютерной обработки данных из архивных и курируемых баз данных. Это, например, SCOP, PFAM, GO и др.
SCOP (Structural Classification Of Proteins) – база данных по структурной классификации белков.
Адрес: http://scop.protres.ru/
PFAM (Protein families database of alignments and HMMs) – это большая коллекция семейств белков и доменов, построенных на основании экспертной оценки множественных выравниваний (см. раздел 3). В банке существуют две основные части: PFAMA, содержащая подробно аннотированные белковые семейства, и PFAMB, содержащая различные множественные выравнивания.
Адрес: http://www.sanger.ac.uk/Pfam/
GO (Gene Ontology consortium database). Целью создателей базы было установление контроля за единообразием в описаниях функций, биологических процессов и клеточных компонентов, относящихся к продуктам генов. Унификация описаний в различных базах данных облегчает поиск в них нужного гена. GO – независимая база данных: другие базы данных сотрудничают с ней, помещая ссылки на унифицированные термины GO, либо поддерживают поиск с использованием терминов базы GO, а также стимулируют ее дополнение и уточнение.
Адрес: http://www.geneontology.org/

5) Интегрированные базы данных.
Они объединяют информацию из разных баз. Например, введя имя гена, можно найти всю, связанную с ним информацию.
К таким базам относится ENTREZ (Molecular Biology DataBase and Retrieval System). Эта интегрированная база данных содержит нуклеотидные и аминокислотные последовательности, которые собираются из крупнейших специализированных хранилищ – баз данных. Основой является GenBank, кроме того, информация пополняется из dbEST, dbSTS, SwissProt, PIR, PDB, PRF, GSDB. Данные из перечисленных ресурсов поступают в интегрированную базу данных после 1) присвоения уникального идентификатора последовательности, 2) перевода документов в единый стандарт хранения, 3) проверки данных, 4) проверки всех ссылок по базе данных MedLine, 5) проверки названий организмов по таксономической классификации GenBank Taxonomy.
Адрес ENTREZ: http://www.ncbi.nlm.nih.gov/Database/index.html

Описания многих баз данных по биоинформатике можно найти на русскоязычном сайте, который находится по адресу: http://www.jcbi.ru/index.html
При подаче запросов в большинство существующих программ последовательности должны быть представлены в стандарте IUB/IUPAC. Этот стандарт предусматривает условные обозначения нуклеиновых кислот и аминокислот, представленные в таблицах 2.1, 2.2.

Работа с BLAST

Адрес: http://blast.ncbi.nlm.nih.gov/Blast.cgi

Термин «BLAST» – (Basic Local Alignment Search Tool) означает – поисковый механизм (программу) логического сравнения аминокислотных и нуклеотидных последовательностей. Данный поисковый механизм позволяет находить одинаковые (подобные) области при сравнении последовательностей.

Программа проводит сравнение для нуклеотидной или белковой последовательности, введенной пользователем, со всеми нуклеотидными или протеиновыми последовательностями, имеющимися в базах данных, представленных на сайте NCBI, и затем, подсчитывает в процентах статистику совпадения общих участков для каждой пары сравниваемых последовательностей.

BLAST может быть полезен для оценки функциональных особенностей последовательностей, для установки родственных связей между ними, например, в качестве более поздних модификаций или для идентификации членов генного семейства.

При открытии главной страницы BLAST вверху есть главное меню с 4-мя вкладками:

Home – вкладка для возврата на домашнюю страницу BLAST с любой другой страницы (BLAST home page); Находящаяся под ней выделенная строка – дает переход к новостям, и основным события дня, которые изменяются периодически.

Recent Results – вкладка для открытия результатов поисков, которые Вы совершили в последние 36 часов;

Saved Strategies – вкладка для перехода к сохраненным Вами поисковым запросам на вашей личной страничке «My NCBI» (надо зарегистрироваться);

Help – вкладка для перехода в каталог с документацией по работе с программой BLAST. (Собственно, почти вся информация о работе с бластом переведена оттуда)

В правом верхнем углу основной электронной страницы BLAST расположены опции «Моя личная страница», где можно зарегистрироваться, нажав на «[Sign In]», и в дальнейшем, при очередном открытии страницы BLAST, можно вводить свой логин и пароль для входя в свой личный журнал поисков, нажав на опцию «[Register]».

Пользуясь опцией «моя личная страница» можно сохранять поисковые сессии, получать уведомления с сайта о новом наполнении нужных БД, по своему усмотрению менять фильтры, настройки при проведении поисков, просматривать большее количество ссылок на другие интернет–ресурсы, относящиеся к теме поисков.

Ниже на странице представлены «BLAST Assembled Genomes» – коллекции геномов, совокупностей генов, относящихся к разным видам животных и растительных организмов по которым проводится поиск последовательностей, например, геномы человека, геномы мыши, крысы, геномы шимпанзе, свиней, коров, геномы бактерий, растений, геномы зебра-рыбы, дрозофилы и т и.д. Полный список всех возможных для просмотра геномов можно найти в полной карте геномов, нажав на любую строку из представленных в данной коллекции.

Далее в центре страницы расположен список программ для поиска последовательностей. Их всего пять:

1. nucleotide blast - поиск в БД нуклеотидов, с использованием нуклеотидной формы запроса. Алгоритмы: blastn, megablast, discontiguous megablast
2. protein blast – поиск в белковой БД, с использованием пептидной формы запроса. Алгоритмы: blastp, psi-blast, phi-blast
3. blastx – поиск в базе белков, с использованием формы запроса транслированных нуклеотидов
4. tblastn – поиск в базе транслированных нуклеотидов, с использованием аминокислотного запроса
5. Search translated nucleotide database using a protein query
6. tblastx – поиск в базе транслированных нуклеотидов,с использованием формы запроса транслированных нуклеотидов

Существуют три основных вида преобразования (трансляции), выполняемого для последовательности:

blastx - проводится сравнение нуклеотидной последовательности, которую перемещают (транслируют) во все рамки считывания (при трансляции генетического кода) базы данных протеиновых последовательностей

tblastn – проводится сравнение белковой последовательности, которую динамически транслируют во все рамки считывания базы данных нуклеотидных последовательностей

tblastx - проводится сравнение шести рамочной трансляции (the six-frame translations) нуклеотидной последовательности с шести рамочными трансляциями базы данных нуклеотидных последовательностей. Из-за больших сложностей при проведении этого вида сравнения и значительного поискового «шума» рекомендуется использовать tblastx только, если другие виды сравнения не дают никакого результата.

Пользователям, которые собираются проводить поиски только с tblastx следует инсталлировать командную строку BLAST и запускать приложение со своего компьютера.

Основные базы данных, по которым осуществляется поиск программой BLAST, по своему содержанию сгруппированы на две части: базы данных нуклеотидных последовательностей и БД белковых последовательностей. Эти базы данных, и их подробное описание приведено ниже.

Основные базы данных белковых последовательностей

nr - Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, за исключением того, что имеется в БД «env_nr». Это основная БД по белковым последовательностям, включающая все записи БД GenBank, БД PDB (Протеиновая БД), БД SwissProt (Швейцарское Биохимическое Общество)

refseq – БД протеиновых последовательностей из «NCBI Reference Sequence project»

swissprot – Последняя версия основной публикации БД «SWISS-PROT protein sequence database»

pat - БД Белков из патентного подразделения БД GenBank (Proteins Patent Abstract)

month – БД всех новых или исправленных за последние 30 дней белковых последовательностей БД GenBank CDS translations + PDB + SwissProt + PIR + PRF

pdb – БД последовательностей, извлеченных из записей (3-D структурных) БД «Protein Data Bank»

env_nr – БД основных CDS последовательностей «Non-redundant CDS translations» извлеченных из БД «env_nt»

Smart v4.0 – 663 PSSMs из Smart, которая активно не поддерживается

Pfam v11.0 – 7255 PSSMs из Pfam, но не самое последнее

COG v1.00 – 4873 PSSMs из NCBI COG set

KOG v1.00 – 4825 PSSMs из NCBI KOG set (эукариотидные COG эквиваленты)

CDD v2.05 – 11399 PSSMs из NCBI, взятые с cd set

Поиск по последним 5-ти БД (выделены серым шрифтом) осуществляется только через поисковую страничку rpsblast (поиск по консервативным доменам белков).

Функция CDD Search работает только для белковых последовательностей BLAST. Проводится сравнение введенной белковой последовательности с последовательностями Главной резервной БД – Conserved Domain Database (CDD). Найденные пары последовательностей обеспечивают дополнительную возможность понимания сути запроса. CDD содержит коллекцию сгруппированных белковых профилей, которую сформировали из коллекций двух внешних БД Smart and Pfam, плюс внутренние наработки NCBI: БД COG и БД cd. Более подробно об этих БД можно прочесть на страничке «CDD homepage».

Базы данных, используемых программой BLAST для поиска нуклеотидных последовательностей

nr – Весь GenBank + EMBL + DDBJ + PDB БД последовательностей (за исключением БД EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences) – БД основная по нуклеотидным последовательностям, включающая все записи БД GenBank, БД EMBL + БД DDBJ, БД PDB (Протеиновая БД)

refseq_mrna – БД нуклеотидных последовательностей mRNA из «NCBI Reference Sequence project

refseq_genomic – БД геномных последоательностей «Genomic sequences» из «NCBI Reference Sequence project»

est – БД последовательностей «GenBank + EMBL + DDBJ» из подраздела EST Division

est_human – Коллекция последовательностей подраздела EST Division по человеку

est_mouse – Коллекция последовательностей подраздела EST Division по мыши

est_others – Коллекция всех других последовательностей подраздела EST Division, исключая мышь и человека

gss «Genome Survey Sequence» – БД «Обзор геномных последовательностей», включающие в себя single-pass genomic data, exon-trapped seq., and Alu PCR seq.

htgs «Unfinished High Throughput Genomic Sequences» – незаконченные высокореактивные геномные последовательности с фазами 0, 1 и 2. (Законченные аналоги с фазой 3 HTG представлены в БД «nr»)

pat – БД нуклеотидов из патентного подразделения БД GenBank (Nucleotide Patent Abstract).

pdb – Последовательности, взятые из 3-dimensional structurezаписей БД Банка Протеиновых данных (Brookhaven Protein Data Bank). Записи не являются кодированными последовательностями соответств. белков, найденных в той же записи БД Protein Data Bank(PDB)

month – БД всех новых или исправленных за последние 30 дней нуклеотидных последовательностей БД GenBank + БД EMBL +DDBJ+PDB

alu_repeats – БД отобранных Alu-повторений (repeats) из REPBASE, пригодные для наложения этих же Alu-повторений из запрашиваемой последовательности. Подробно см. статью “Alu alert”. Авторы - Claverie and Makalowski, Nature 371: 752 (1994)

dbsts – БД последовательностей с меченными участками (Sequence Tag Site) полученная из STS division (раздела) БД GenBank + EMBL + DDBJ

chromosome – БД полных геномов (Complete genomes), и полных хромосом из NCBI Reference Sequence project. БД частично перекрывает записи БД refseq_genomic.

wgs – БД Whole Genome Shotgun sequences

env_nt – БД последовательностей природных (environmental samples), таких как образцы неокультуренных бактерий, выделенные из почвы или морской воды. Самый известный источник таких образцов – это БД Sagarsso Sea project.

Всех БД, доступных для использования с BLAST, достаточно много. Некоторые из них, например, SwissProt и PDB работают отдельно от NCBI (внешние БД). Другие, такие как ecoli, dbEST and month, – являются подмножеством БД, предоставляемых NCBI (внутренние БД). Все остальные «виртуальные БД» можно просмотреть и открыть, используя опцию “Limit by Entrez Query”.

Баз данных, используемых программой BLAST для поиска геномов различных видов животного и растительного мира

genome (all assemblies) - полная коллекция БД текущих известных геномов. Формат номера следующий

RefSeq Accession Numbers – NT_?????? , или NW_?????? (6 цифр). В БД входят коллекции, основанные на клонах и коллекции полного геномной фрагментации (whole genome shotgun) или композитная коллекция. Это наиболее полная геномная БД.

genome (reference only) – ссылочная БД БД содержащая все, что и БД genome (all assembliesно только в краткой реферативной форме. БД обновляется, сразу же по мере публикации информации.

HTGS – БД содержит коллекцию геномных последовательностей из GenBank, с ключевым словом «HTG keyword». БД позволяет искать одновременно htgs_phase3 последовательности (обычно ищут в БД NR) и htgs_phase0, 1 и 2 последовательности (обычно ищут в БД HTGS).

RefSeq RNA Коллекция ссылочных данных mRNAs разработана NCBI RefSeq project. БД обновляется ежедневно.

RefSeq protein Коллекция ссылочных данных белков разработана NCBI RefSeq project. БД обновляется ежедневно.

Build RNA Коллекция ссылочных данных mRNAs разработана NCBI как часть геномного реферативного канала. БД обновляется по мере публикации информации.

Build protein Коллекция ссылочных данных белков разработана NCBI как часть геномного реферативного канала. БД обновляется по мере публикации информации.

Ab Initio RNA Коллекция ab initio RNA прогнозирования (predictions) разработана NCBI как часть геномного реферативного канала. БД обновляется по мере публикации информации.

Ab Initio protein Коллекция ab initio белкового прогнозирования (predictions) разработана NCBI как часть геномного реферативного канала. БД обновляется по мере публикации информации.

ESTs БД считывания однократных последовательностей из библиотек cDNA. БД обновляется ежедневно.

BAC ends БД концевых последовательностей клонов BAC. БД обновляется ежедневно.

Traces-WGS БД всех исходных (простейших) организмов в WGS следах. БД обновляется по мере необходимости.

Traces-ESTs БД всех исходных (простейших) организмов в EST следах. БД обновляется по мере необходимости.

Traces-other БД всех исходных (простейших) организмов в non-WGS и non-EST следах. БД обновляется по мере необходимости

WGS contigs Если организм был собран с использованием полной геномной стратегии shotgun (WGS), то в этой БД он будет доступен (if the WGS assembly is in GenBank). БД обновляется по мере необходимости.

Gene Trap Clones (Mouse Only) Коллекция последовательностей сгенерированных путем представления «Gene Trap» вставок. БД обновляется еженедельно.

Reference Dog Assembly (boxer) БД геномов «supercontigs» из Whole Genome Shotgun (WGS) собранных из 7.6X наполнения полной геномной библиотеки. Эта коллекция создана в Broad Institute , с использованием Arachne ассемблера (механизм сборки).

Celera Dog Assembly (Poodle) Коллекция геномов «contigs» из Whole Genome Shotgun (WGS) созданных из 1.5X наполнения полной геномной библиотеки. Описание этой коллекции можно посмотреть в статье автора Kirkness et al (2003).

Celera Dog Extra (Poodle) Коллекция геномов Whole Genome Shotgun (WGS) которые не были воссозданы из «contigs» (the Celera Dog Assembly). Описание этой коллекции можно посмотреть в статье автора Kirkness et al (2003).

Ref Chimp Assembly Коллекция геномов «contig» из Whole Genome Shotgun (WGS) созданных с использованием программы Arachne. Эти геномы «contigs» были созданы из 4.5X наполнения из группы WGS записей. Публикации по этой теме доступны с 2004 года.

Alt Chimp Assembly БД геномов «contigs» из Whole Genome Shotgun (WGS) с использованием программы PCAP. Эти геномы «contigs» были созданы из 4.5X наполнения из группы WGS записей. Публикации по этой теме доступны с 2004 года.

Celera CSA БД Celera на Январь 2001 г. Собрание геномов человека методом «compartmental shotgun assembly» (CSA). БД получена из 27 млн. записей Celera's 5.3X «whole genome shotgun» и 16 млн. записей расщепленных данных из БД GenBank и других геномных проектов. Более подробно смотри статью в Pubmed – (Nature 2001. 409:860-921).

Celera cWGA БД Celera на Ноябрь 2000 г содержит полный набор геномов (WGA) человека. БД получена из 27 млн. записей Celera's 5.3X «whole genome shotgun» и 16 млн. записей расщепленных данных из БД GenBank и других геномных проектов. Более подробно смотри статью в Pubmed – (Nature 2001. 409:860-921).

Celera WGA БД Celera на Декабрь 2001 г. содержит полный набор геномов (WGSA) человека БД получена из 27 млн. записей Celera's 5.3X «whole genome shotgun» только для «shotgun» данных и 104,000 BAC концевых последовательностей спаренных с последовательностями из GenBank и других геномных проектов. Более подробно смотри статьи – (Nature 1996. 381:364-366; Genomics 2000. 63:321-332).

hsc_tcag БД «The Hospital for Sick Children Center for Applied Genomics assembly of Human Chromosome 7». Это комбинация последовательностей WGS, взятых из Celera and HTGS последовательностей, взятых из «Human Genome Sequencing Consortium». Более подробно смотри статью Scherer et al (2003).

Броузеры генома

Геном состоит из набора хромосом, а хромосома — это две цепочки, свёрнутые в спираль. Каждая из цепочек содержит последовательность нуклеотидов с четырьмя типами азотистых оснований — аденин (A), гуанин (G), цитозин (C) и тимин (T). По одной цепочке легко определить вторую, если помнить, что аденин соединяется в пару с тимином, а гуанин с цитозином. Некоторые участки ДНК называются генами, с них считывается РНК, по которой потом кодируются белки. Белки состоят из аминокислот 20 видов (плюс пара экзотических), каждая из которых кодируется по трём нуклеотидам.

Браузер генома — это такая одномерная карта, которая отображает какую-нибудь нуклеотидную последовательность (скажем, хромосому или отдельный ген) с сопутствующей информацией. Информация обычно структурируется в блоки, называемые треками (tracks). К примеру, может быть трек с генами или с отдельными нуклеотидами. Отдельные сущности на треках часто называют фичами (features).

Бывают браузеры геномов, рассчитанные на маленькие бактериальные геномы, но универсальному браузеру необходимо показывать и длинные хромосомы позвоночных целиком, и отдельные нуклеотиды. Самая длинная хромосома человека (первая) содержит около 250 миллионов пар оснований, то есть масштаб должен меняться примерно в миллион раз. Конечно, в разном масштабе информация отображается по-разному. В самом детальном масштабе можно увидеть отдельные нуклеотиды, как на прямой, так и на обратной спирали ДНК.

http://ugene.unipro.ru/ Ugene

Разработчики: Центр информационных технологий «УниПро» , Новосибирск,
http://unipro.ru/ru/about/overview.html

Стандартный генетический код

Основания ДНК (РНК)

```
A    Adenine    Аденин
T    Thymine    Тимин      (U    Uracil    Урацил)
G    Guanine    Гуанин
C    Cytosine   Цитозин
```

Таблица генетического кода

	T(U)	C	A	G
T(U)	TTT Phe	TCT Ser	TAT Tyr	TGT Cys
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys
	TTA Leu	TCA Ser	TAA Stop	TGA Stop
	TTG Leu	TCG Ser	TAG Stop	TGG Trp
C	CTT Leu	CCT Pro	CAT His	CGT Arg
	CTC Leu	CCC Pro	CAC His	CGC Arg
	CTA Leu	CCA Pro	CAA Gln	CGA Arg
	CTG Leu	CCG Pro	CAG Gln	CGG Arg
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser
	ATC Ile	ACC Thr	AAC Asn	AGC Ser
	ATA Ile	ACA Thr	AAA Lys	AGA Arg
	ATG Met	ACG Thr	AAG Lys	AGG Arg
G	GTT Val	GCT Ala	GAT Asp	GGT Gly
	GTC Val	GCC Ala	GAC Asp	GGC Gly
	GTA Val	GCA Ala	GAA Glu	GGA Gly
	GTG Val	GCG Ala	GAG Glu	GGG Gly

Аминокислоты

```
A   Ala   Alanine          Аланин
```

R	Arg	Arginine	Аргинин
N	Asn	Asparagine	Аспарагин
D	Asp	Aspartic Acid	Аспарагиновая кислота
C	Cys	Cysteine	Цистеин
Q	Gln	Glutamine	Глютамин
E	Glu	Glutamic Acid	Глютаминовая кислота
G	Gly	Glycine	Глицин
H	His	Histidine	Гистидин
I	Ile	Isoleucine	Изолейцин
L	Leu	Leucine	Лейцин
K	Lys	Lysine	Лизин
M	Met	Methionine	Метионин
F	Phe	Phenylalanine	Фенилаланин
P	Pro	Proline	Пролин
S	Ser	Serine	Серин
T	Thr	Threonine	Треонин
W	Trp	Thryptophan	Триптофан
Y	Tyr	Tyrosine	Тирозин
V	Val	Valine	Валин

"Stop" в таблице кода означает стоп-кодон — сигнал окончания трансляции.

Таблица 1. Обозначения, принятые для нуклеиновых кислот по стандарту IUB/IUPAC.

Обозна–чение	**Название**	**Обозна–чение**	**Название**	**Обозна–чение**	**Название**
A	adenine	**R**	G, A (purine)	**B**	G, T, C
C	cytosine	**Y**	T, C (pyrimidine)	**D**	G, A, T
G	guanine	**W**	A, T	**H**	A, C, T
T	thymine	**K**	G, T (keto)	**N**	A, G, C, T
U	uracil	**S**	G, C		
M	A, C (amino)	**V**	A, C, G		

Таблица 2. Обозначения, принятые в однобуквенном коде аминокислот по стандарту IUB/IUPAC.

Обозна-чение	**Название (англ.)**	**Название (рус.)**	**Обозна-чение**	**Название (англ.)**	**Название (рус.)**
G	glycine	глицин	**U**	selenocysteine	селеноцистеин
H	histidine	гистидин	**V**	valine	валин
I	isoleucine	изолейцин	**W**	tryptophan	триптофан
K	lysine	лизин	**Y**	tyrosine	тирозин
L	leucine	лейцин	**Z**	glutamate or glutamine	глютамат или глютамин
M	methionine	метионин			
*****	translation stop	стоп–кодон	**X**	any	любая