**Тема: Генетические базы данных.**

**Основные вопросы, выносимые на обсуждение семинара.**

1. Основные генетические базы данных сети интернет.
2. Алгоритмы поиска и сравнения нуклеотидных последовательностей в генетических базах данных.
3. Анализ результатов поиска по генетическим базам данных.

**Краткое содержание занятия:**

В настоящее время существует множество Web-сайтов, которые доступны для обзора и 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, PIR и др.) пополняются аннотированными последовательностями непосредственно исследователями, расшифровавшими их, с помощью автоматизированной системы пополнения баз данных по сети Интернет. Впоследствии эти данные проверяются персоналом администраций баз данных и существенно пополняются.

По данным фирмы Intergrated genomics на октябрь 2001 г. опубликованы полные последовательности геномов 74 организмов включая 11 архебактерий, 48 эубактерий, 15 эукариот. На подходе 212 геномов прокариот и 157 эукариот. Общее число нуклеотидов в секвенированных последовательностях превысило 1010.

Базы данных (БД) геномов содержат нуклеотидные последовательности и "транслированные" по ним аминокислотные последовательности белков. В большинстве БД также содержатся дополнительные данные, как экспериментальные (например, значимость гена для выживаемости организма), так и полученные расчетным путем (например, функция белка, кодируемого геном, может быть постулирована на основе сходства его аминокислотной последовательности с первичной структурой уже охарактеризованного белка).

Web-серверы, предоставляющие пользователю генетическую информацию, оснащены комплексом программных средств для поиска в БД и анализа нуклеотидных и аминокислотных последовательностей. В качестве запросов при поиске последовательностей в БД могут использоваться номенклатурные названия генов, организмов, ключевые слова и др.

Ядро любой генетической информационной системы составляет программа поиска в БД гомологов последовательности, заданной пользователем. Часто используются алгоритмы BLAST или Fasta. Кроме того, на Web-серверах БД зачастую представлены дополнительные программные средства, позволяющие рассчитать некоторые физико-химические свойства белка (например, изоэлектрическую точку), предсказывать вторичную структуру, наличие и локализацию трансмембранных участков и т.д. Такие данные часто используются при выполнении широкого круга исследований.

|  |  |
| --- | --- |
| Наименование ГБ | Краткое описание |
| [GenBank](file:///F:\www.ncbi.nlm.nih.gov\genbank\genbank-overview.html) www.ncbi.nlm.nih.gov/genbank | БД по нуклеотидным последовательностям |
| [SwissProt](file:///F:\www.expasy.ch\sprot\sprot-top.html)  www.expasy.ch/sprot | Аннотированная БД по аминокислотным последовательностям белков |
| [PIR](file:///F:\www.nbrf.georgetown.edu\pir\searchdb.html)  www.nbrf.georgetown.edu/pir | Аннотированная БД по аминокислотным последовательностям белков, организованным в соответствии с гомологией и таксономией |
| [PDB](file:///F:\www.rcsb.org\pdb\default.htm)  [www.rcsb.org/pdb](http://www.rcsb.org/pdb) | БД по трехмерной структуре биологических макромолекул |
| [OWL](file:///F:\www.biochem.ucl.ac.uk\bsm\dbbrowser\%20owl\owl.html)  www.biochem.ucl.ac.uk/bsm | Невырожденная комплексная БД по структурам белков из SwissProt, PIR (1-3), GenBank и NRL-3D |
| [NDB](file:///F:\ndbserver.rutgers.edu\default.htm)  www.ndbserver.rutgers.edu | БД по нуклеиновым кислотам, включает структуры ДНК и РНК вместе с их трехмерными изображениями |
| [PROSITE](file:///F:\www.expasy.ch\prosite)  www.expasy.ch/prosite. | БД паттернов функционально значимых участков белков |
| [ProDom](file:///F:\protein.toulouse.inra.fr\prodom.html)  www.protein.toulouse.inra.fr | БД по доменам белков |
| [ProteinMotionsDatabase](file:///F:\hyper.stanford.edu\~mbg\protmotdb)  www.hyper.stanford.edu | БД по динамике белков, включающая многоуровневую классификацию движения петель, доменов и субъединиц |
| [PROMISE](file:///F:\bioinf.leeds.ac.uk\promise)  www.bioinf.leeds.ac.uk | БД по простетическим группам и ионам металла в активных центрах белков |
| [RELIBASE](file:///F:\www.relibase.darmstadt.gmd.de\gmd\default.htm)  www.relibase.darmstadt.gmd.dd | Полная БД по лиганд-рецепторным комплексам |
| [OMIM](file:///F:\www.ncbi.nlm.nih.gov\omim)  www.ncbi.nlm.nih.gov | Каталог генов человека и генетически обусловленных заболеваний |
| [LIGAND](file:///F:\www.genome.ad.jp\dbget\ligand.html)  www.genome.ad.jp | БД по ферментативным реакциям |
| [dbCFC](file:///F:\cytokine.medic.kumamoto.ac.jp\default.htm)  www.cytokine.medic.kumamoto.ac.jp | БД по кодирующим ДНК цитокинов |

Рассмотрим некоторые из БД более подробно :

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.

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.

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 микотическими геномами.