Министерство образования и науки Российской федерации
Стерлитамакский филиал
федерального государственного бюджетного образовательного учреждения
высшего образования
«Башкирский государственный университет»
Естественнонаучный факультет
Кафедра биологии

Молекулярные базы данных. Принцип действия и характеристики основных компьютерных программ для сравнения биологических последовательностей

Выполнил: магистрант второго года обучения очной формы по направлению 06.04.01.68 Биология программа Общая биология группа МБИО21 Арефьева Анастасия Александровна

Проверил:
кандидат биологических наук, доцент кафедры биологии
Курамшина Зиля Мухтаровна

Стерлитамак 2016

- **Новейшим методом изучения природных молекул является применение информационных систем.**
- Работая с информационными моделями молекул, исследователь обычно имеет дело с базами, банками данных и инструментами их анализа. Вследствие широкого применения информационных моделей молекул появилось новое направление – биологическая информатика (биоинформатика, компьютерная биология)

- Компьютерным моделированием молекулярно-генетических и смежных процессов занимаются такие науки как биоинформатика, системная биология, геномика, эволюционная генетика, протеомика, транскриптомика, метаболомика и другие, еще более узкоспециализированные дисциплины, в каждой из которых работают тысячи и десятки тысяч исследователей. Такой высокий уровень дифференциации наук связан с колоссальной сложностью и огромным объемом молекулярно-генетических данных. Например, работа с последовательностью ДНК даже простейших эукариот - дрожжей S. cerevisiae – не была бы возможна без использования компьютерных методов, не говоря уже о геноме человека.

Биоинформатика

- Биоинформатика — это область науки, разрабатывающая и применяющая вычислительные алгоритмы для систематизации и анализа генетической информации с целью определения молекулярных основ биологических процессов с последующим использованием этих знаний на практике. Ее основная задача — разработка вычислительных алгоритмов для анализа и систематизации данных о структуре и функциях биологических молекул, прежде всего нуклеиновых кислот и белков. Объем генетической информации, накапливаемой в банках данных, начал увеличиваться с возрастающей скоростью после того, как были разработаны быстрые методы секвенирования (расшифровки нуклеотидных последовательностей ДНК).
- Биоинформатика возникла в 1976-1978 годах, окончательно оформилась в 1980 году со специальным выпуском журнала «Nucleic Acid Research» (NAR).

Биоинформатика включает в себя:

- базы данных, в которых хранится биологическая информация
- набор инструментов для анализа тех данных, которые лежат в таких базах
- правильное применение компьютерных методов для правильного решения биологических задач

Задачи, решаемые биоинформатикой

Источник данных	Объем данных	Задачи
Секвенированные последовательности ДНК	~40 млн. последовательностей, 1012 пар оснований	Функциональная аннотация
Белковые последовательности	~5.5 106 последовательностей (~300 аминокислот каждая)	Сравнительный анализ. Выявление консервативных мотивов
Структуры макромолекул	50000 структур (~3000 атомных координат каждая)	Предсказание, выравнивание, измерение геометрии, докинг
Геномы	Около 1200 геномов прокариот, более 160 геномов эукариот	Сборка полных геномов; Функциональная аннотация; Сравнительный анализ
Экспрессия генов в различных тканях, стадиях развития, состояний организма и т.д.	Сотни тысяч образцов с тысячами вариантов измерений для десятков тысяч генов. ~1013 измерений.	Анализ механизмов регуляции коэкспрессирующихся генов. Связь с последовательностями, структурными и биохимическими данными.
SNP (однонуклеотидные мутации в ДНК)	Только одна база данных dbSNP содержит информацию о 108 мутациях в 23 геномах.	Анализ связи с заболеваниями
Молекулярные взаимодействия, метаболические пути и генные сети	Более 106 молекулярных взаимодействий описано в публикациях. Более ста тысяч метаболических путей и генных сетей представлено в базах данных.	Моделирование молекулярно-генетических процессов и систем
Публикации	Десятки миллионов публикаций	Поиск и извлечение знаний

- Биолог в биоинформатике обычно имеет дело с базами данных и инструментами их анализа. Теперь разберемся, какие базы данных бывают в зависимости от того, что в них помещают.

- Первый тип – архивные базы данных, это большая свалка, куда любой может поместить все, что захочет. К таким базам относятся:
- **GeneBank & EMBL** – здесь хранятся первичные последовательности
- **PDB** – пространственные структуры

- В качестве курьеза можно привести пример: в архивной базе данных указано, что в геноме археи (архебактерии) есть ген, кодирующий белок главного комплекса гистосовместимости, что является полной чепухой, т.к. характерно для позвоночных.

- Второй тип – курируемые базы данных, за достоверность которых отвечает хозяин базы данных. Туда информацию никто не присылает, ее из архивных баз данных отбирают эксперты, проверяя достоверность информации – что записано в этих последовательностях, какие есть экспериментальные основания для того, чтобы считать, что эти последовательности выполняют ту или иную функцию.
- К базам данных такого типа относятся:
- **Swiss- Prot** – наиболее качественная база данных, содержащая аминокислотные последовательности белков
- **KEGG** – информация о метаболизме
- **FlyBase** – информация о *Drosophila*
- **COG** – информация об ортологичных генах (гомологичные гены филогенетически родственных организмов, разошедшихся в процессе видообразования)

- Поддержание базы требует работы кураторов или аннотаторов. Тем не менее, даже в курируемых базах данных могут встречаться курьезные надписи, например такая забавная надпись:
- CAUTION: AN ORF CALLED DSDC WAS ORIGINALLY (REF.3) ASSIGNED TO THE WRONG DNA STRAND AND THOUGHT TO BE A D- SERINE DEAMINASE ACTIVATOR, IT WAS THEN RESEQUENCED BY REF.2 AND STILL THOUGHT TO BE "DSDC", BUT THIS TIME TO FUNCTION AS A D-SERINE PERMEASE. IT IS REF.1 THAT SHOWED THAT DSDC IS ANOTHER GENE AND THAT THIS SEQUENCE SHOULD BE CALLED DSDX. IT SHOULD ALSO BE NOTED THAT THE C-TERMINAL PART OF DSDX (FROM 338 ONWARD) WAS ALSO SEQUENCED (REF.6 AND REF.7) AND WAS THOUGHT TO BE A SEPARATE ORF **(YES, DON'T WORRY, WE ALSO HAD PROBLEMS UNDERSTANDING WHAT HAPPENED!).**
- По крайне мере здесь кураторы базы данных честно признаются, что не знают, как это случилось.

- Третий тип – производные базы данных. Такие базы получаются в результате обработки данных из архивных и курируемых баз данных. Сюда входит:
- **SCOP** – База данных структурной классификации белков (описывается структура белков)
- **PFAM** – База данных по семействам белков
- **GO (Gene Ontology)** – Классификация генов (попытка создания набора терминов, упорядочивания терминологии, чтобы один ген не назывался по разному, и чтобы разным генам не давали одинаковые названия)
- **ProDom** – белковые домены
- **AsMamDB** – альтернативный сплайсинг у млекопитающих

- И интегрированные базы данных, в которых вся информация (курируемая, не курируемая) свалена в кучу, и введя имя гена, можно найти всю связанную с ним информацию – в каких организмах встречается, в каком месте генома локализован, какие функции выполняет и т.д.
- **NCBI Entrez** – доступ к информации о нуклеотидных и аминокислотных последовательностях и структурах
- **Ecocyc** – все о *E. coli* – гены, белки, метаболизм и пр.

- Теперь перейдем к рассмотрению инструментов биоинформатики. Инструменты определяются задачами, которые мы хотим решать.
- Основу биоинформатики составляют сравнения. Если у нас есть, например, аминокислотная последовательность, о которой у нас есть экспериментальные данные, и известны ее функции, и другая, похожая на нее последовательность, мы можем предположить, что эти последовательности выполняют сходные функции. Это задача поиска сходства последовательностей

Компьютерные технологии в науке, С.15-20.

- Как сравнивают последовательности? Запишем одну последовательность под другой:

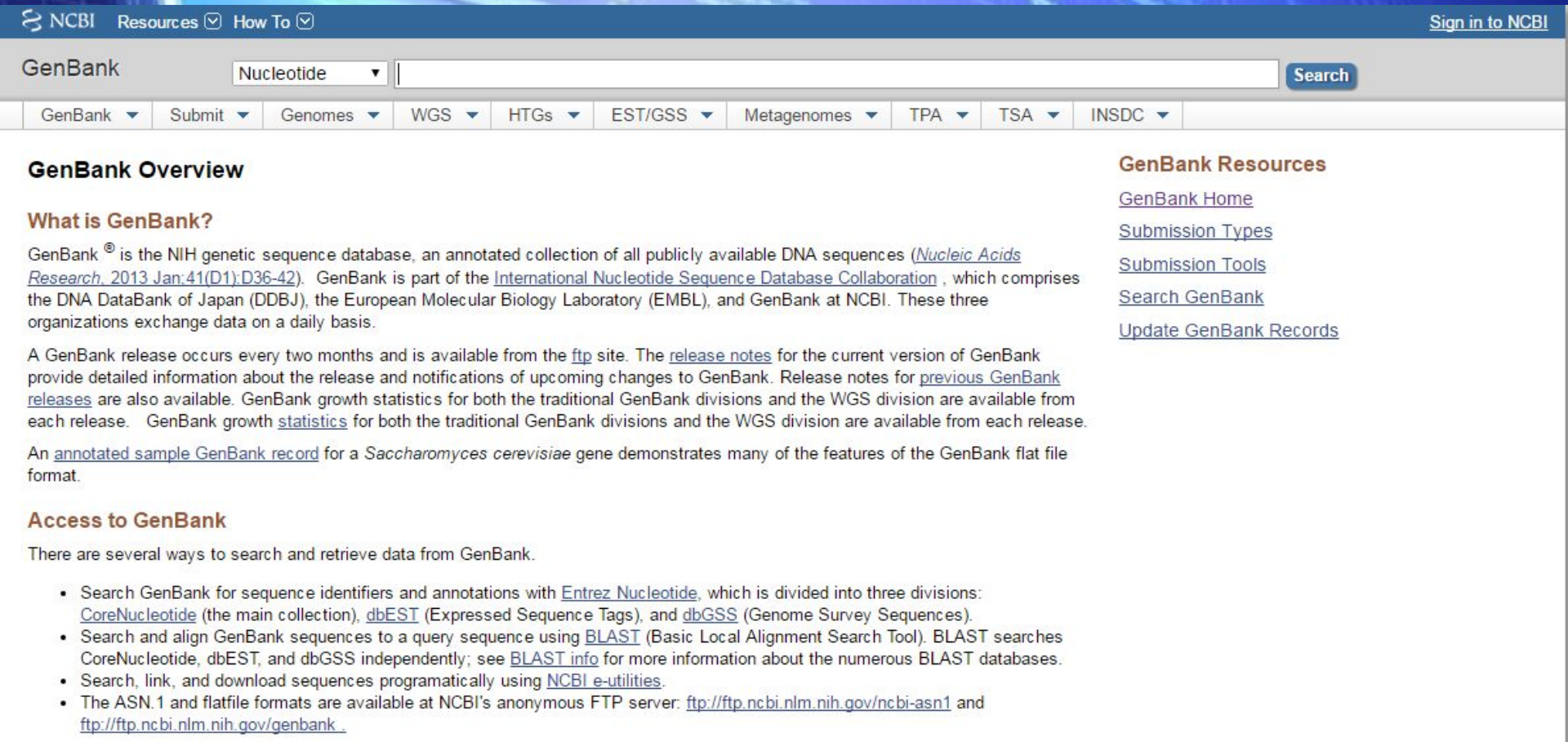
attgt**AC**c**TC**g**T**g**G-AA**----
-----**AC-TC**a**T**a**G**c**AA**ccag

- Нам надо при сравнении найти наилучший вариант, так выровнять эту пару последовательностей, чтобы количество совпадений было максимальным (парное выравнивание). Качество выравнивания оценивают, назначая штрафы за несовпадение букв и за наличие пробелов (когда приходится раздвигать одну последовательность для того, чтобы получить наибольшее число совпадающих позиций)

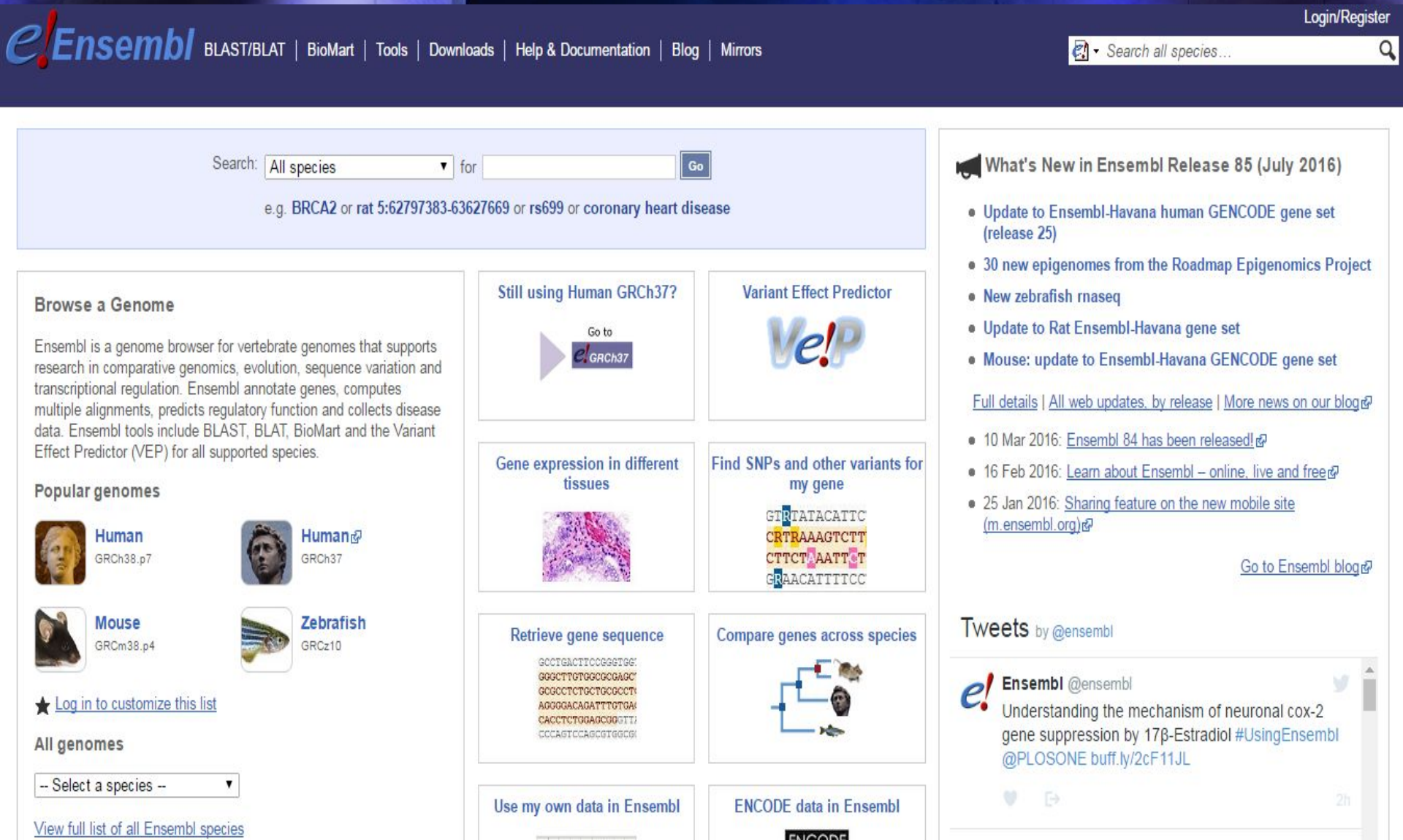
- Таким образом, первым делом после секвенирования последовательности ищут в базах данных похожие последовательности, чтобы после сравнения судить о том, какие функции несет эта последовательность. Если две буквы совпали, значит они находятся под давлением отбора, они функционально важны. Известно, что аминокислоты различаются по своим свойствам, поэтому если произошла аминокислотная замена, это может почти никак не повлиять на работу белка, а может сильно его изменить.
- Например, если лизин (положительно заряженная аминокислота заменится на лейцин (похожий по созвучию, но совершенно несходный по свойствам), то для пространственной структуры и функций белка это может оказаться катастрофой. А вот замена лизина на аргинин (также положительно заряженный) может не сказаться на структуре белка.
- Поэтому при сравнении аминокислотных последовательностей учитывают также матрицу сопоставления аминокислотных остатков (похожих, менее похожих и совсем непохожих).

Молекулярно-генетические данные хранятся в специализированных банках данных (все на английском языке):

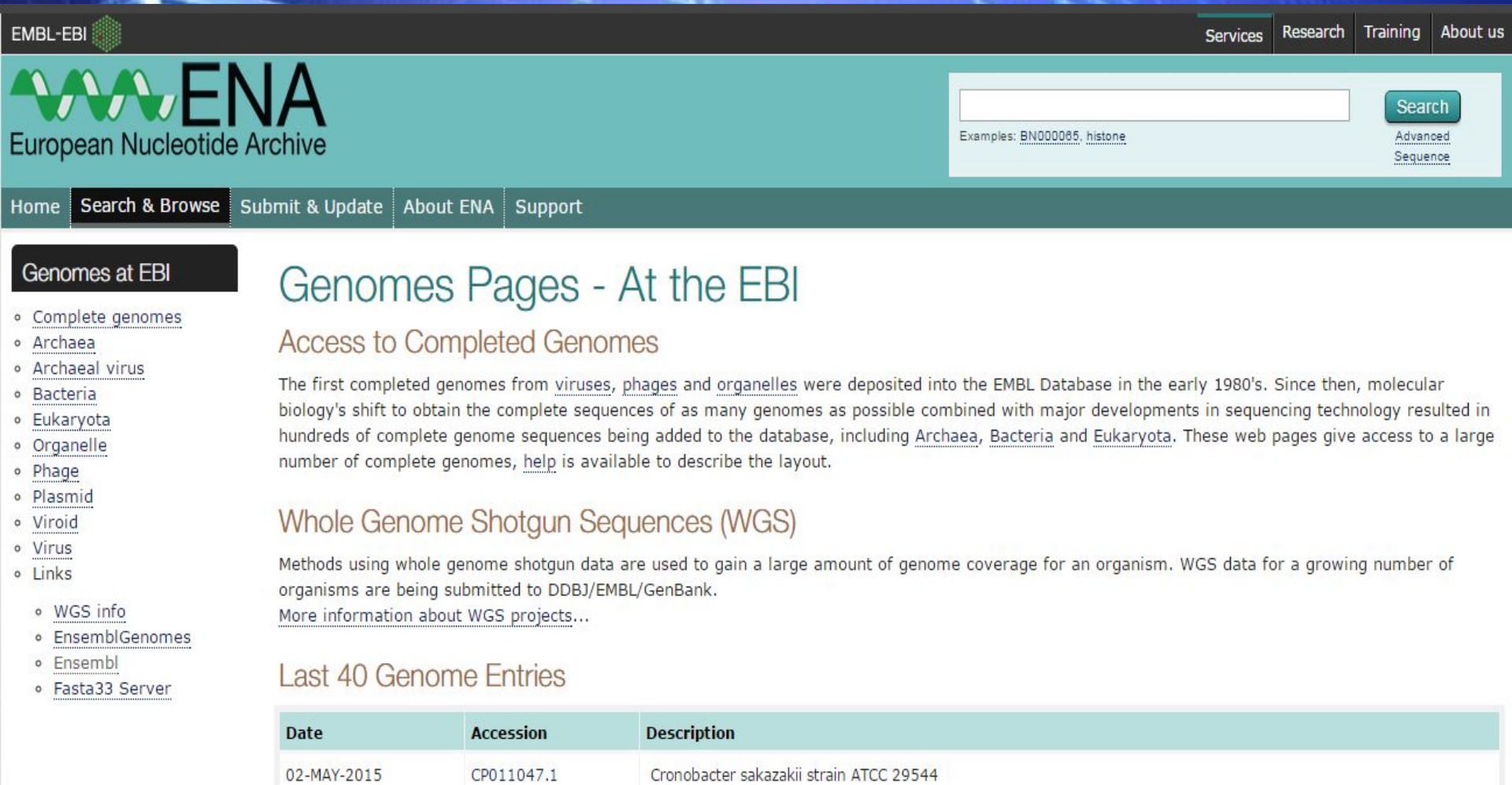
- **крупнейшая база генетических данных – GeneBank**



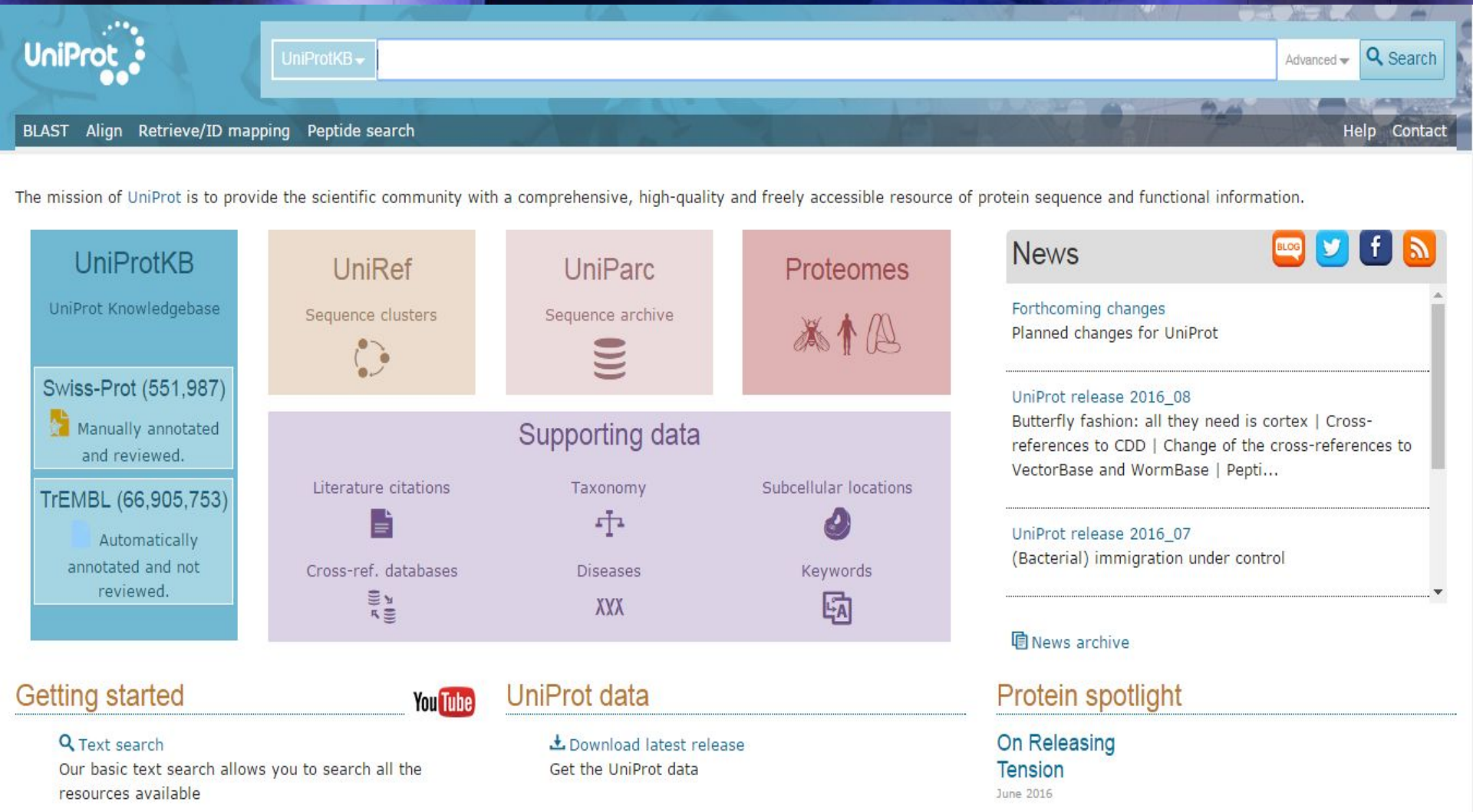
- удобная в навигации база генетических последовательностей – Ensembl



- удобный доступ к полным геномам через сайт Европейского института биоинформатики - http://www.ebi.ac.uk/genomes/



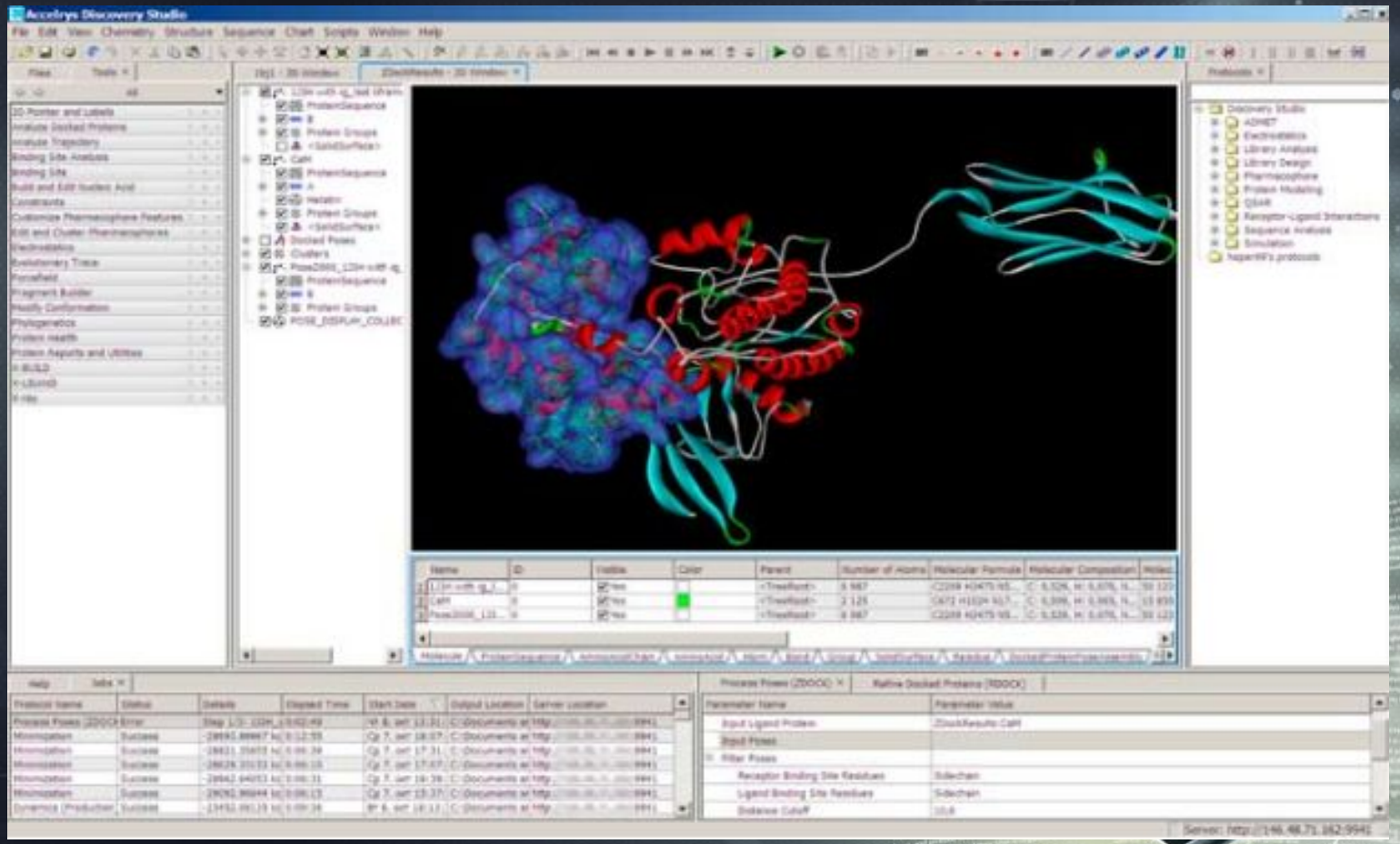
- крупнейший банк белковых данных – UniProt.org



- крупнейший банк данных о структуре биологических макромолекул http://www.pdb.org/

Информационные системы, касающиеся моделей макромолекул и надмолекулярных структур:

- ***GeneBank & EMBL*** – здесь хранятся первичные последовательности
- ***PDB*** – пространственные структуры белков
- ***Swiss-Prot*** – наиболее качественная база данных, содержащая аминокислотные последовательности белков
- ***KEGG*** – информация о метаболизме (такая, которая представлена на карте метаболических путей)
- ***SCOP*** – база данных структурной классификации белков (описывается структура белков)
- ***PFAM*** – база данных по семействам белков
- ***GO (Gene Ontology)*** – классификация генов (попытка создания набора терминов, упорядочивания терминологии)
- ***ProDom*** – белковые домены
- ***AsMamDB*** – альтернативный сплайсинг у млекопитающих
- ***NCBI Entrez*** – доступ к информации о нуклеотидных и аминокислотных последовательностях и структурах
- ***Ecocyc*** – все о *E. coli* – гены, белки, метаболизм и пр.
- ***Accelrys Discovery Studio***



Окно программы *Discovery Studio*

Видны вторичные и третичные структуры, поверхность белка кальмодулина. Доступные инструменты расположены слева от 3D окна, а протоколы для проведения глубокого изучения взаимодействующих молекул и симуляций – справа. Окна сообщений и вспомогательной информации об исследуемых структурах расположены снизу

- CAZy: Carbohydrate-Active Enzymes Database.

  На сайте представлена современная классификация ферментов синтеза и утилизации углеводов, а также их гомологов. Ферменты (а точнее каждый из их модулей/доменов) разбиты на пять групп: Glycosidases and Transglycosidases, Glycosyltransferases, Polysaccharide Lyases, Carbohydrate Esterases, Carbohydrate-Binding Modules. В пределах каждого из них выделяются семейства, которые нумеруются арабскими цифрами в порядке описания. Каждое семейство объединяет "хорошие" гомологи. Родственные семейства объединены в кланы. Например, гликозидазы и трансгликозидазы образуют 113 семейств (GH1-GH118, кроме GH21, GH40, GH41, GH60 и GH69). Из них 50 семейств объединены в 14 кланов (GH-A–GH-N). Информация о каждом из семейств включает список его представителей из разных организмов с ссылками на базы данных аминокислотных и нуклеотидных последовательностей, указание ферментативных активностей, названий белков, наличия экспериментально определённых трёхмерных структур, данные о молекулярном механизме катализируемой реакции и компонентах активного центра. База данных обновляется примерно раз в месяц.

http://molbiol.edu.ru/review/01_01.html

Молекулярно-биологические ресурсы.

Профессиональные ресурсы.

NAR database

Ежегодно первый номер журнала "Nucleic Acid Research" посвящён обзору молекулярно-биологических баз данных. Обзорную статью этого номера и сортированные (по темам и по алфавиту) списки баз данных в HTML формате можно найти на сайте журнала (кнопка "NAR database issue").

BioMedNet

BioMedNet организован Elsevier Science. Это web-сайт для биологов и медиков. На сегодняшний день он имеет более 600,000 зарегистрированных пользователей (эта цифра увеличивается ежемесячно на более чем 20,000). Регистрация и весь сайт бесплатные. Можно получать новости сайта по E-mail.

* Публикуются обзоры, новости, обзоры конференций;
* Хорошая подборка аннотированных web-ресурсов;
* Список журналов со свободным доступом (часто временно/ради рекламы доступны хорошие журналы), возможность подписаться на "содержание журналов";
* Возможность поиска фирм производителей конкретной медико-биологической продукции;
* Имеются: medline; Technical Tips (коллекция мол. биол. протоколов);
* База вакансий с возможностью поиска.

Medscape Molecular Medicine

Аналогичный описанному выше ресурс (только для медиков) организованный MedicaLogic/Medscape, Inc. На сегодняшний день он имеет более миллиона зарегистрированных пользователей. Можно получать новости сайта по E-mail. Множество полезных ресурсов и ссылок.

Научные конференции

Пока лучший известный нам источник информации о научных конференциях – раздел "Научные конференции" на сайте "Научная инициатива в Интернет".

Патенты.

Два "постоянных" бесплатных адреса и один - теперь уже платный (раньше была бета-тест программа но хорошая).

IBM Patent Server

IBM Intellectual Property Network (IPN). Позволяет искать и просматривать патентные документы. Обеспечивает доступ к:

* United States patents (US). 1971 - до настоящего времени, еженедельное обновление. Полный текст с картинками. От U.S. Patent and Trademark Office.
* European patents - applications (EP-A). 1979 - до настоящего времени, еженедельное обновление. Библиографический текст с картинками. От European Patent Office; Vienna, Austria
* European patents - issued (EP-B). 1980 - до настоящего времени, еженедельное обновление. Библиографический текст с картинками. От European Patent Office; Vienna, Austria.
* Patent Abstracts of Japan (JP). 10/1976- до настоящего времени, еженедельное обновление. Библиографический текст, первая страница. От JAPIO, Japan Patent Information Organization; Toyko, Japan.
* IBM Technical Disclosure Bulletins.

Картинки приходят в "*.pdf" формате, причём их качество обычно весьма плохое.

United States Patent and Trademark Office

Архив содержит все патенты США (только США) с 1976г. Картинки приходят в "*.tif" формате. Для того, чтобы их смотреть надо установить соответствующий plug-in. Список бесплатных источников находится на этом же сайте.

Get^{the}Patent.^{com}

Get^{the}Patent.^{com} организован Cartesian Products, Inc (компания, специализирующаяся в области телекоммуникации и предоставления информации). Архив содержит все патенты США (только США) с 1976г. В настоящее время закончилось бета-тестирование и надо платить. Есть возможность получать файлы по E-mail, а не ждать их загрузки в он-лайн режиме.

Документы приходят с расширением *.cpc. Для их просмотра требуется установить plug-in "CPC Lite" от Cartesian Products. Его можно бесплатно скачать с сайта фирмы-производителя.

Наиболее удобный и быстрый (но теперь уже не бесплатный) способ получения текстов патентов.

Литература.

Журналы.

Более 4000 ссылок на биологические и медицинские журналы содержится на "science.komm" (там же удобные ссылки на полнотекстовые источники, словари, базы данных по абстрактам и т.п.). По web-ссылке вы попадаете на сайт конкретного журнала. На многих журналах можно подписаться на рассылку оглавления по E-mail.

Ссылки.

Citation Matcher. Организован National Center for Biotechnology Information (NCBI) National Institutes of Health (NIH). Позволяет найти статью по библиографическим данным. Возможен поиск сразу же большого количества статей. Можно использовать Citation Matcher через E-mail (E-Mail Citation Matcher); чтобы узнать правила работы достаточно послать письмо с текстом HELP.

Импакт-фактор.

Импакт-фактор журналов можно найти на сайте Journal Citation Reports on the Web (JCR Web). На этом сайте представлены данные по цитированию более 8,400 журналов более чем 3,000 издательств.

Список импакт-факторов биологических журналов за 1999г. имеется на нашем сервере.

Книги.

Получить информацию о книге/аудио/видео/DVD/ (или заказать) можно на сайте "Amazon.com". 4.7 миллионов названий.

Опубликована on line версия книги Benjamin Lewin "Genes". Она основана на тех же материалах, что и печатная версия, но в отличии от неё непрерывно обновляется. Наиболее свежая печатная версия - 7е издание, вышла в январе 2000.

Базы данных по абстрактам.

PubMed. Организован National Center for Biotechnology Information (NCBI) National Institutes of Health (NIH). PubMed содержит полное содержание MEDLINE и PREMEDLINE баз данных и некоторые статьи не входящие в них. Очень удобно доверить регулярный просмотр ссылок на PubMed программе BioMail.

Medline на уже упоминавшемся "BioMedNet".
Medline на "HealthWord Online".
Medline на "Medscape Molecular Medicine".
Другие источники можно найти на "science.komm".

Модельные организмы.

Arabidopsis.

TAIR: совместный проект Carnegie Institution of Washington Department of Plant Biology и National Center for Genome Resources (NCGR).

AIMS: совместный проект Michigan State University и Arabidopsis Biological Resource Center (ABRC) Ohio State University финансируемый NSF. ARABIDOPSIS INFORMATION MANAGEMENT SYSTEM и ARABIDOPSIS BIOLOGICAL RESOURCE CENTER

AFGC - Arabidopsis Functional Genomics Consortium: совместный проект четырёх университетов по изучению функции генов Arabidopsis, объединяющий Expression Microarray Analysis и T-DNA нокаут.

Drosophila

Flybase: информационная база данных по генетике и молекулярной биологии дрозофилы. Содержит литературные данные и данные из Drosophila Genome Projects. FlyBase – совместный проект Berkeley и European Drosophila Genome Projects.

IUBio, Bloomington, Indiana USA
Harvard, Cambridge, Mass USA
E.B.I., Cambridge, UK
B.D.G.P., Berkeley, California USA
N.I.G., Japan
NHRI, Taiwan
ANGIS, Sydney, Australia
I.B.M.C., Strasbourg, France
Weizmann, Israel

Mouse

Мы не настолько хорошо знаем ресурсы, связанные с этим организмом, чтобы делать подробный обзор, но не можем не отметить пособие по препарированию мышей "Mickey: The Inside Story" на сайте National Cancer Institute.

Сток-центры (культуры клеток и микроорганизмы).

В сток-центрах можно приобрести штаммы и культуры (заказ на выращивание определённого количества клеток), заказать идентификацию микроорганизмов и т.п. На web-сайтах можно получить информацию об условиях роста, истории линии и ссылки на связанные с ней публикации.

American Type Culture Collection (ATCC)

German Collection of Microorganisms and Cell Cultures (DSMZ)

По крайней мере для бактерий цены раза в два ниже, чем у "ATCC".

Адреса других (в том числе и Российских) сток-центров можно найти на этой странице сайта University of California Museum of Paleontology (UCMP).

Базы данных.

GeneCards

GeneCards: human genes, proteins and diseases. база данных человеческих генов, их продуктов и их участия в генетических заболеваниях. Полезна для тех, кто ищет информацию о генах в контексте функциональных геномных и протеомных проектов.

Анализ нуклеотидных и белковых последовательностей.

BCM Search Launcher

Search Launcher организован Baylor College of Medicine. Сайт предоставляет возможность проводить различные молекулярно-биологические анализы/поиски со стандартного и очень простого интерфейса (реально серьёзные анализы проводятся на других серверах). На сервере можно выполнять простые преобразования последовательностей. Есть возможность организовать анализ сразу пачки последовательностей.

Ресурс просто незаменим для тех, кто не чувствует себя большим мастером по анализу последовательностей.

ExPASy Molecular Biology Server

ExPASy (Expert Protein Analysis System) протеомик сервер Swiss Institute of Bioinformatics (SIB). Сервер предназначен для анализа белковых последовательностей, структур и 2-D PAGE.

National Center for Biotechnology Information

NCBI организован в 1988 как отделение National Library of Medicine(NLM) в National Institutes of Health (NIH). В настоящее время это крупнейшая биологическая база данных (молекулярная биология, биохимия и генетика). NCBI имеет мощные системы обработки и представления этих данных.

По этому адресу имеется очень толковое описание ресурсов сайта.

Форматы файлов, используемых в биоинформатике

FASTA

>roa1_drome Rea guano receptor type III >> 0.1

MVNSNQNQNGNSNGHDDDFPQDSITEPEHMRKLFIGGLDYRTTDENLKAHEKWGNIVDV
VVMKDPRTKRSRGFGFITYSHSSMIDEAQKSRPHKIDGRVEPKRAVPRQDIDSPNAGATV
KKLFVGALKDDHDEQSIRDYFQHFGNIVDNIVIDKETGKKRGFAFVEFDDYDPVDKVVLQ
KQHQLNGKMVDVKKALPKNDQQGGGGGRGGPGGRAGGNRGNMGGGNYGNQNGGGN
WNNGGNNWGNNRGNDNWGNNSFGGGGGGGGGYGGGNNSWGNNNPWDNGNGGGNF
GGGGNNWNGGNDFGGYQQNYGGGPQRGGGNFNNNRMQPYQGGGGFKAGGGNQGNY
GNNQGFNNGGNNRRY

>roa2_drome Rea guano ligand

MVNSNQNQNGNSNGHDDDFPQDSITEPEHMRKLFIGGLDYRTTDENLKAHEKWGNIVDV
VVMKDPTSTSTSTSTSTSTSTSTMIDEAQKSRPHKIDGRVEPKRAVPRQDIDSPNAGATVK
KLFVGALKDDHDEQSIRDYFQHLLLLLLLDLLLLDLLLLDLLLFVEFDDYDPVDKVVLQK
QHQLNGKMVDVKKALPKNDQQGGGGGRGGPGGRAGGNRGNMGGGNYGNQNGGGNW
NNGGNNWGNNRGNDNWGNNSFGGGGGGGGGYGGGNNSWGNNNPWDNGNGGGNFG
GGGNNWNGGNDFGGYQQNYGGGPQRGGGNFNNNRMQPYQGGGGFKAGGGNQGNYG
NNQGFNNGGNNRRY

GenBank

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF

TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
          ----------//----------
YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
 RNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML"
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"

/translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
          FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
 RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
  LISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
        1 gatcctccat     atacaacggt  atctccacct  caggtttaga     tctcaacaac   ggaaccattg
       61 ccgacatgag     acagttaggt  atcgtcgaga  gttacaagct     aaaacgagca   gtagtcagct
      121 ctgcatctga     agccgctgaa  gttctactaa  gggtggataa     catcatccgt   gcaagaccaa
      181 gaaccgccaa     tagacaacat  atgtaacata  tttaggatat     acctcgaaaa   taataaaccg
      241 ccacactgtc     attattataa  ttagaaacag  aacgcaaaaa     ttatccacta   tataattcaa
      301 agacgcgaaa     aaaaaagaac  aacgcgtcat  agaacttttg     gcaattcgcg   tcacaaataa
     ----------//----------
     4621 tcttcgcact     tcttttccca  ttcatctctt  tcttcttcca     aagcaacgat   ccttctaccc
     4681 atttgctcag     agttcaaatc  ggcctctttc  agtttatcca     ttgcttcctt   cagtttggct
     4741 tcactgtctt     ctagctgttg  ttctagatcc  tggtttttct     tggtgtagtt   ctcattatta
     4801 gatctcaagt     tattggagtc  ttcagccaat  tgctttgtat     cagacaattg   actctctaac
     4861 ttctccactt     cactgtcgag  ttgctcgttt  ttagcggaca     aagatttaat   ctcgttttct
     4921 ttttcagtgt     tagattgctc  taattctttg  agctgttctc     tcagctcctc   atatttttct
     4981 tgccatgact     cagattctaa  ttttaagcta  ttcaatttct     ctttgatc
//
```

GenBank. Запись sequence

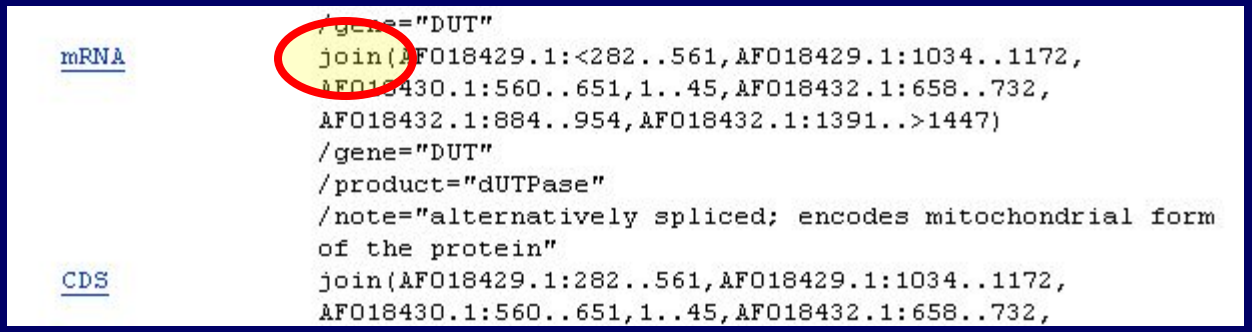
```
                     /protein_id="CAA25860.1"
                     /db_xref="GI:41298"
                     /db_xref="GOA:P0C093"
                     /db_xref="UniProtKB/Swiss-Prot:P0C093"
                     /translation="MAEKQTAKRNRREEILQSLALMLESSDGSQRITTAKLAASVGVS
                     EAALYRHFPSKTRMFDSLIEFIEDSLITRINLILKDEKDTTARLRLIVLLLLGFGERN
                     PGLTRILTGHALMFEQDRLQGRINQLFERIEAQLRQVLREKRMREGEGYTTDETLLAS
                     QILAFCEGMLSRFVRSEFKYRPTDDFDARWPLIAASCSNMTPDDFSSGEFL"
ORIGIN
        1 cagagaaaat caaaaagcag gccacgcagg gtgatgaatt aacaataaaa atggttaaaa
       61 accccgatat cgtcgcaggc gttgccgcac taaaagacca tcgaccctac gtcgttggat
      121 ttgccgccga aacaaataat gtggaagaat acgcccggca aaaacgtatc cgtaaaaacc
      181 ttgatctgat ctgcgcgaac gatgtttccc agccaactca aggatttaac agcgacaaca
      241 acgcattaca ccttttctgg caggacggag ataaagtctt accgcttgag cgcaaagagc
      301 tccttggcca attattactc gacgagatcg tgacccgtta tgatgaaaaa aatcgacgtt
      361 aagattctgg acccgcgcgt tgggaaggaa tttccgctcc cgacttatgc cacctctggc
      421 tctgccggac ttgacctgcg tgcctgtctc aacgacgccg tagaactggc tccgggtgac
      481 actacgctgg ttccgaccgg gctggcgatt catattgccg atccttcact ggcggcaatg
      541 atgctgccgc gctccggatt gggacataag cacggtatcg tgcttggtaa cctggtagga
      601 ttgatcgatt ctgactatca gggccagttg atgatttccg tgtggaaccg tggtcaggac
      661 agcttcacca ttcaacctgg cgaacgcatc gcccagatga tttttgttcc ggtagtacag
      721 gctgaattta atctggtgga agatttcgac gccaccgacc gcggtgaagg cggctttggt
      781 cactctggtc gtcagtaaca catacgcatc cgaataacgt cataacatag ccgcaaacat
      841 ttcgtttgcg gtcatagcgt gggtgccgcc tggcaagtgc ttattttcag gggtattttg
      901 taacatggca gaaaaacaaa ctgcgaaaag gaaccgtcgc gaggaaatac ttcagtctct
      961 ggcgctgatg ctggaatcca gcgatggaag ccaacgtatc acgacggcaa aactggccgc
     1021 ctctgtcggc gtttccgaag cggcactgta tcgccacttc cccagtaaga cccgcatgtt
     1081 cgatagcctg attgagttta tcgaagatag cctgattact cgcatcaacc tgattctgaa
     1141 agatgagaaa gacaccacag cgcgcctgcg tctgattgtg ttgctgcttc tcggttttgg
     1201 tgagcgtaat cctggcctga cccgcatcct cactggtcat gcgctaatgt ttgaacagga
     1261 tcgcctgcaa gggcgcatca accagctgtt cgagcgtatt gaagcgcagc tgcgccaggt
     1321 attgcgtgaa aagagaatgc gtgagggtga aggttacacc accgatgaaa ccctgctggc
     1381 aagccagatc ctggccttct gtgaaggtat gctgtcacgt tttgtccgca gcgaatttaa
     1441 ataccgcccg acggatgatt ttgacgcccg ctggccgcta attgcggcca gttgcagtaa
     1501 tatgacgccg gatgactttt catccggcga gtttctttaa acgccaaact cttcgcgata
     1561 ggccttaacc gccgccagat gttccgccat ttccggcttc tcttccagg
//
```

GenBank. Запись mRNA

GenBank - Поиск в Google | NCBI Sequence Viewer v2.0

```
LOCUS       HSU90223                 960 bp    mRNA    linear   PRI 03-JAN-1998
DEFINITION  Human deoxyuridine triphosphate nucleotidohydrolase precursor mRNA,
            nuclear gene encoding mitochondrial protein, complete cds.
ACCESSION   U90223
VERSION     U90223.1  GI:2735291
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 960)
  AUTHORS   Ladner,R.D. and Caradonna,S.J.
  TITLE     The Human dUTPase Gene Encodes Both Nuclear and Mitochondrial
            Isoforms: Differential Expression of the Isoforms and
            Characterization of a cDNA Encoding the Mitochondrial Species
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 960)
  AUTHORS   Ladner,R.D. and Caradonna,S.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (19-FEB-1997) Dept. of Molecular Biology, Univ. of Med.
            and Dent. of NJ-School of Osteopathic Medicine, 2 Medical Center
            Drive, Stratford, NJ 08084, USA
FEATURES             Location/Qualifiers
     source          1..960
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
     CDS             63..821
                     /note="mitochondrial dUTPase isoform; DUT-M"
                     /codon_start=1
                     /product="deoxyuridine triphosphate nucleotidohydrolase
                     precursor"
                     /protein_id="AAB94642.1"
                     /db_xref="GI:2735292"
                     /translation="MTPLCPRPALCYHFLTSLLRSAMQNARGTAEGRSRGTLRARPAP
                     RPPAAQHGIPRPLSSAGRLSQGCRGASTVGAAGWKGELPKAGGSPAPGPETPAISPSK
                     RARPAEVGGMQLRFARLSEHATAPTRGSARAAGYDLYSAYDYTIPPMEKAVVKTDIQI
                     ALPSGCYGRVAPRSGLAAKHFIDVGAGVIDEDYRGNVGVVLFNFGKEKFEVKKGDRIA
                     QLICERIFYPEIEEVQALDDTERGSGGFGSTGKN"
     sig peptide     63..269
                     /note="mitochondrial targeting presequence"
     mat peptide     270..818
                     /product="deoxyuridine triphosphate nucleotidohydrolase"
ORIGIN
        1 ggtggaagcc tggcgcacgt ccggaggtgc cgaggaccca accagcccaa actctggggg
```

Сплайсинг и восстановление последовательности mRNA



mRNA seq=(AF018429.1:282-561)+(AF018429.1:1034-1172)+(AF018430.1:560-651)+(AF018430.1:1-45)+……

GenBank. Запись genomic DNA

1: AF018431. Reports Homo sapiens dUTP...[gi:2443577]

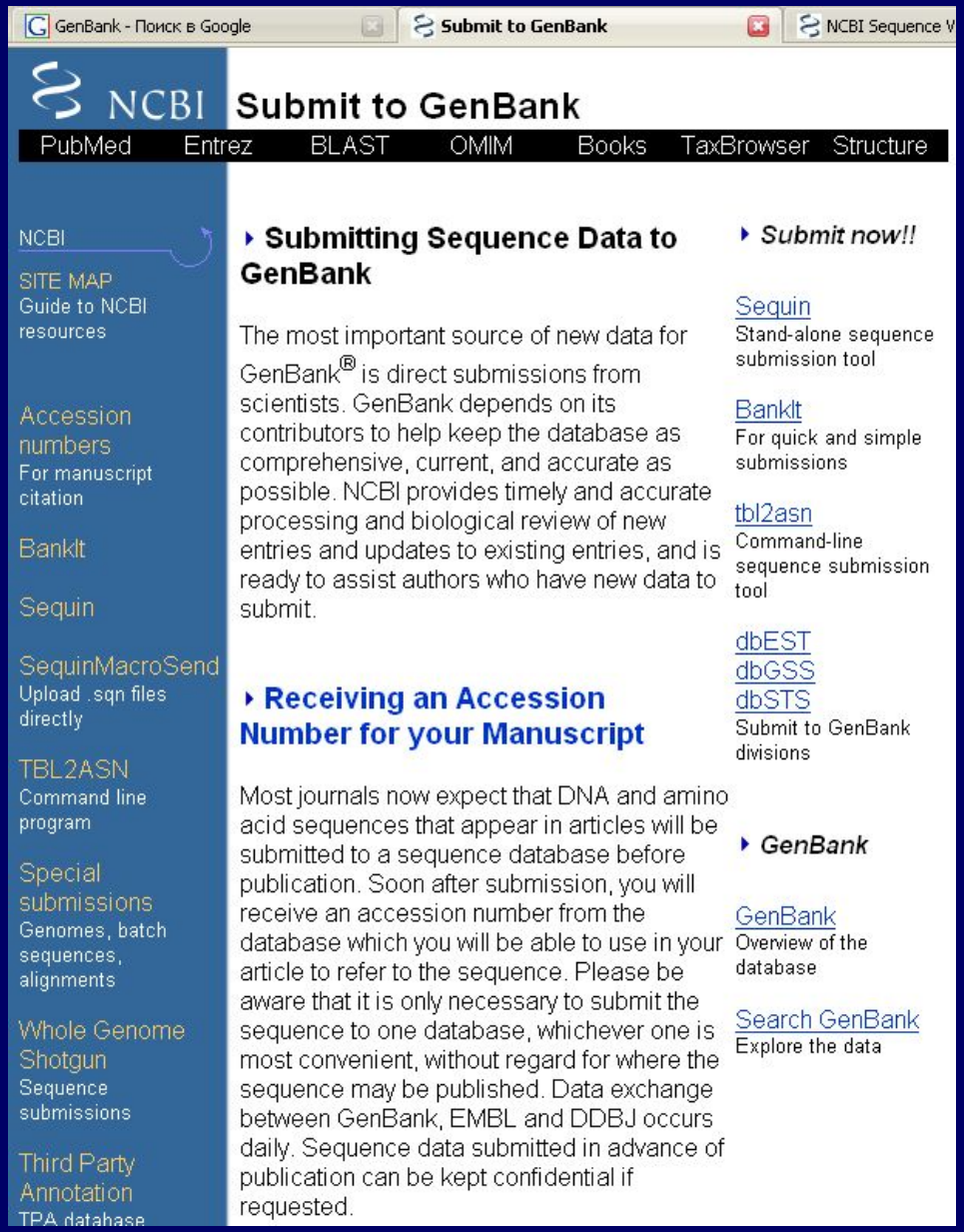
Features Sequence

```
LOCUS       HSDUT3                   577 bp    DNA     linear   PRI 28-SEP-1997
DEFINITION  Homo sapiens dUTPase (DUT) gene, exon 4.
ACCESSION   AF018431
VERSION     AF018431.1  GI:2443577
KEYWORDS    .
SEGMENT     3 of 4
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 577)
  AUTHORS   Pearlman,R.E.
  TITLE     Human genomic nuclear and mitochondria dUTPase gene
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 577)
  AUTHORS   Pearlman,R.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-AUG-1997) Biology, York University, 4700 Keele St.,
            North York, ONT M3J 1P3, Canada
FEATURES             Location/Qualifiers
     source          1..577
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /map="15q15-q21.1"
     gene            order(AF018429.1:<1..1735,AF018430.1:1..1177,1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..>1447)
                     /gene="DUT"
     mRNA            join(AF018429.1:<282..561,AF018429.1:1034..1172,
                     AF018430.1:560..651,1..45,AF018432.1:658..732,
                     AF018432.1:884..954,AF018432.1:1391..>1447)
                     /gene="DUT"
```

```
FEATURES             Location/Qualifiers
     source          1..577
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /map="15q15-q21.1"
     gene            order(AF018429.1:<1..1735,AF018430.1:1..1177,1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..>1447)
                     /gene="DUT"
     mRNA            join(AF018429.1:<282..561,AF018429.1:1034..1172,
                     AF018430.1:560..651,1..45,AF018432.1:658..732,
                     AF018432.1:884..954,AF018432.1:1391..>1447)
                     /gene="DUT"
                     /product="dUTPase"
                     /note="alternatively spliced; encodes mitochondrial form
                     of the protein"
     CDS             join(AF018429.1:282..561,AF018429.1:1034..1172,
                     AF018430.1:560..651,1..45,AF018432.1:658..732,
                     AF018432.1:884..954,AF018432.1:1391..1447)
                     /gene="DUT"
                     /note="DUT-M; alternatively spliced; mitochondrial form of
                     the protein; similar to H. sapiens dUTPase encoded by
                     GenBank Accession Number U90224"
                     /codon_start=1
                     /product="dUTPase"
                     /protein_id="AAB71393.1"
                     /db_xref="GI:2443580"
                     /translation="MTPLCPRPALCYHFLTSLLRSAMQNARGTAEGRSRGTLRARPAP
                     RPPAAQHGIPRPLSSAGRLSQGCRGASTVGAAGWKGELPKAGGSPAPGPETPAISPSK
                     RARPAEVGGMQLRFARLSEHATAPTRGSARAAGYDLYSAYDYTIPPMEKAVVKTDIQI
                     ALPSGCYGRVAPRSGLAAKHFIDVGAGVIDEDYRGNVGVVLFNFGKEKFEVKKGDRIA
                     QLICERIFYPEIEEVQALDDTERGSGGFGSTGKN"
     mRNA            join(AF018429.1:<1018..1172,AF018430.1:560..651,1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..>1447)
                     /gene="DUT"
                     /product="dUTPase"
                     /note="alternatively spliced; encodes nuclear form of the
                     protein"
     CDS             join(AF018429.1:1018..1172,AF018430.1:560..651,1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..1447)
                     /gene="DUT"
                     /note="DUT-N; alternatively spliced; nuclear form of the
                     protein; similar to H. sapiens dUTPase encoded by GenBank
                     Accession Number U90224"
                     /codon_start=1
                     /product="dUTPase"
                     /protein_id="AAB71394.1"
                     /db_xref="GI:2443581"
                     /translation="MPCSEETPAISPSKRARPAEVGGMQLRFARLSEHATAPTRGSAR
                     AAGYDLYSAYDYTIPPMEKAVVKTDIQIALPSGCYGRVAPRSGLAAKHFIDVGAGVID
                     EDYRGNVGVVLFNFGKEKFEVKKGDRIAQLICERIFYPEIEEVQALDDTERGSGGFGS
                     TGKN"
     exon            1..45
                     /gene="DUT"
                     /number=4
```

GenBank. Аннотация

Как добавить данные в GB?



Зачем?
- информация в community;
- Журналы требуют это ДО публикации

Долго ли это?
2 рабочих дня

Данные могу быть закрыты до выхода статьи (по запросу)

Что нужно?
Последовательность, ее описание (аннотация), описание источника

http://www.ncbi.nlm.nih.gov/Genbank/submit.html

Форматы описания белков

PDB

PDB-XML

MMDB-Cn3D

PDB – Protein Data Bank

```
HEADER    LUMINESCENT PROTEIN                   09-DEC-03   1RRX
TITLE     CRYSTALLOGRAPHIC EVIDENCE FOR ISOMERIC CHROMOPHORES IN 3-
TITLE    2 FLUOROTYROSYL-GREEN FLUORESCENT PROTEIN
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: SIGF1-GFP FUSION PROTEIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES;
COMPND   5 OTHER_DETAILS: CONTAINS 3-FLUORO-TYROSINE
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: AEQUOREA VICTORIA;
SOURCE   3 ORGANISM_COMMON: FUNGI;
SOURCE   4 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   5 EXPRESSION_SYSTEM_COMMON: BACTERIA;
SOURCE   6 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID
KEYWDS    BETA-BARREL, EGFP, NON-CANONICAL AMINO ACID, CHROMOPHORE
KEYWDS   2 ISOMERISATION
EXPDTA    X-RAY DIFFRACTION
AUTHOR    J.H.BAE,P.PARAMITA PAL,L.MORODER,R.HUBER,N.BUDISA
REVDAT   1   08-JUN-04 1RRX    0
JRNL        AUTH   J.H.BAE,P.PARAMITA PAL,L.MORODER,R.HUBER,N.BUDISA
JRNL        TITL   CRYSTALLOGRAPHIC EVIDENCE FOR ISOMERIC
JRNL        TITL 2 CHROMOPHORES IN 3-FLUOROTYROSYL-GREEN FLUORESCENT
JRNL        TITL 3 PROTEIN.
JRNL        REF    CHEMBIOCHEM                V.   5   720 2004
JRNL        REF  2 EUROP.J.CHEM.BIOL.
JRNL        REFN              GE ISSN 1439-4227
REMARK   1
REMARK   2
REMARK   2 RESOLUTION. 2.10 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
----------------------------------------//----------------------------------------------------------
REMARK 500  M RES CSSEQI ATM1   ATM2   ATM3
REMARK 500    LEU A  44   CA  -  CB  -  CG  ANGL. DEV. = 13.7 DEGREES
REMARK 500    LEU A  64   N   -  CA  -  C   ANGL. DEV. =-16.6 DEGREES
REMARK 500    LEU A  64   CA  -  C   -  O   ANGL. DEV. =-16.0 DEGREES
REMARK 500    LEU A  64   CA  -  C   -  N   ANGL. DEV. = 31.6 DEGREES
REMARK 500    LEU A  64   O   -  C   -  N   ANGL. DEV. =-15.9 DEGREES
REMARK 500    THR A  97   N   -  CA  -  C   ANGL. DEV. =-14.0 DEGREES
REMARK 500    GLU A 115   N   -  CA  -  C   ANGL. DEV. =-13.1 DEGREES
REMARK 900
REMARK 900 RELATED ENTRIES
REMARK 900 RELATED ID: 1EMG   RELATED DB: PDB
REMARK 900 THE WILD TYPE OF STUDIED NON-CANONICAL AMINO ACID-
REMARK 900 CONTAINING GFP
DBREF  1RRX A    2   227  UNP    P42212   GFP_AEQVI      290    517
SEQADV 1RRX YOF A   39  UNP  P42212    TYR   327 MODIFIED RESIDUE
SEQADV 1RRX MFC A   66  UNP  P42212    THR   353 MODIFIED RESIDUE
SEQADV 1RRX MFC A   66  UNP  P42212    TYR   354 MODIFIED RESIDUE
SEQADV 1RRX MFC A   66  UNP  P42212    GLY   355 MODIFIED RESIDUE
SEQADV 1RRX YOF A   74  UNP  P42212    TYR   362 MODIFIED RESIDUE
SEQADV 1RRX YOF A   92  UNP  P42212    TYR   380 MODIFIED RESIDUE
SEQADV 1RRX YOF A  106  UNP  P42212    TYR   394 MODIFIED RESIDUE
SEQADV 1RRX YOF A  143  UNP  P42212    TYR   431 MODIFIED RESIDUE
SEQADV 1RRX YOF A  143  UNP  P42212    TYR   433 MODIFIED RESIDUE
SEQADV 1RRX YOF A  151  UNP  P42212    TYR   439 MODIFIED RESIDUE
SEQADV 1RRX YOF A  182  UNP  P42212    TYR   470 MODIFIED RESIDUE
SEQADV 1RRX YOF A  200  UNP  P42212    TYR   488 MODIFIED RESIDUE
SEQRES   1 A  226  SER LYS GLY GLU GLU LEU PHE THR GLY VAL VAL PRO ILE
SEQRES   2 A  226  LEU VAL GLU LEU ASP GLY ASP VAL ASN GLY HIS LYS PHE
SEQRES   3 A  226  SER VAL SER GLY GLU GLY GLU GLY ASP ALA THR YOF GLY
SEQRES   4 A  226  LYS LEU THR LEU LYS PHE ILE CYS THR THR GLY LYS LEU
SEQRES   5 A  226  PRO VAL PRO TRP PRO THR LEU VAL THR THR LEU MFC
VAL
SEQRES   6 A  226  GLN CYS PHE SER ARG YOF PRO ASP HIS MET LYS GLN HIS
SEQRES   7 A  226  ASP PHE PHE LYS SER ALA MET PRO GLU GLY YOF VAL GLN
SEQRES   8 A  226  GLU ARG THR ILE PHE PHE LYS ASP ASP GLY ASN YOF LYS
SEQRES   9 A  226  THR ARG ALA GLU VAL LYS PHE GLU GLY ASP THR LEU VAL
SEQRES  10 A  226  ASN ARG ILE GLU LEU LYS GLY ILE ASP PHE LYS GLU ASP
SEQRES  11 A  226  GLY ASN ILE LEU GLY HIS LYS LEU GLU YOF ASN YOF ASN
SEQRES  12 A  226  SER HIS ASN VAL YOF ILE MET ALA ASP LYS GLN LYS ASN
SEQRES  13 A  226  GLY ILE LYS VAL ASN PHE LYS ILE ARG HIS ASN ILE GLU
SEQRES  14 A  226  ASP GLY SER VAL GLN LEU ALA ASP HIS YOF GLN GLN ASN
SEQRES  15 A  226  THR PRO ILE GLY ASP GLY PRO VAL LEU LEU PRO ASP ASN
SEQRES  16 A  226  HIS YOF LEU SER THR GLN SER ALA LEU SER LYS ASP PRO
SEQRES  17 A  226  ASN GLU LYS ARG ASP HIS MET VAL LEU LEU GLU PHE VAL
SEQRES  18 A  226  THR ALA ALA GLY ILE
MODRES 1RRX YOF A   39  TYR  3-FLUOROTYROSINE
MODRES 1RRX YOF A   74  TYR  3-FLUOROTYROSINE
MODRES 1RRX YOF A   92  TYR  3-FLUOROTYROSINE
MODRES 1RRX YOF A  106  TYR  3-FLUOROTYROSINE
MODRES 1RRX YOF A  143  TYR  3-FLUOROTYROSINE
MODRES 1RRX YOF A  145  TYR  3-FLUOROTYROSINE
MODRES 1RRX YOF A  151  TYR  3-FLUOROTYROSINE
MODRES 1RRX YOF A  182  TYR  3-FLUOROTYROSINE
MODRES 1RRX YOF A  200  TYR  3-FLUOROTYROSINE
MODRES 1RRX MFC A   66  GLY  CYCLIZED
MODRES 1RRX MFC A   66  TYR  CYCLIZED
HETNAM     YOF 3-FLUOROTYROSINE
HETNAM     MFC 5-[1-(3-FLUORO-4-HYDROXY-PHENYL)-METH-(Z)-YLIDENE]-3,
HETNAM   2 MFC  5-DIHYDRO-IMIDAZOL-4-ONE
FORMUL   1  YOF    9(C9 H10 F N O3)
FORMUL   1  MFC    C15 H16 F N3 O5
FORMUL   2  HOH   *61(H2 O)
```

```
HELIX    1   1 GLU A    5  THR A    9  5                                    5
HELIX    2   2 ALA A   37  YOF A   39  5                                    3
HELIX    3   3 PRO A   56  VAL A   61  5                                    6
HELIX    4   4 VAL A   68  SER A   72  5                                    5
HELIX    5   5 PRO A   75  HIS A   81  5                                    7
HELIX    6   6 ASP A   82  ALA A   87  1                                    6
SHEET          1          A12   VAL   A       12        VAL    A       22        0
SHEET    2     A12 HIS A   25   ASP A   36 -1  O   GLY A   31    N   VAL A   16
SHEET    3     A12 LYS A   41   CYS A   48 -1  O   THR A   43    N   GLU A   34
SHEET    4     A12 HIS A 217   ALA A 227 -1  O   LEU A 220     N   LEU A   44
SHEET    5     A12 HIS A 199   SER A 208 -1  N   SER A 202     O   THR A 225
SHEET    6     A12 ASN A 149   ASP A 155 -1  N   ILE A 152     O   HIS A 199
SHEET    7     A12 GLY A 160   ASN A 170 -1  O   GLY A 160     N   ASP A 155
SHEET    8     A12 VAL A 176   PRO A 187 -1  O   GLN A 177     N   HIS A 169
SHEET    9     A12 YOF A   92  PHE A 100 -1  N   GLU A   95    O   GLN A 184
SHEET   10     A12 ASN A 105   GLU A 115 -1  O   YOF A 106     N   ILE A   98
SHEET   11     A12 THR A 118   ILE A 128 -1  O   LYS A 126     N   LYS A 107
SHEET   12     A12 VAL A   12  VAL A   22   1  N   ASP A   21    O   GLY A 127
CISPEP     1  MET   A       88        PRO   A       89                    0                 0.50
CRYST1    51.003     62.430     70.931    90.00    90.00    90.00 P 21  21  21         4
ORIGX1                1.000000       0.000000       0.000000                        0.00000
ORIGX2                0.000000       1.000000       0.000000                        0.00000
ORIGX3                0.000000       0.000000       1.000000                        0.00000
SCALE1                0.019607       0.000000       0.000000                        0.00000
SCALE2                0.000000       0.016018       0.000000                        0.00000
SCALE3                0.000000       0.000000       0.014098                        0.00000
ATOM      1  N    SER A   2       28.277   8.150  50.951  1.00 57.00           N
ATOM      2  CA   SER A   2       27.454   9.223  51.584  1.00 55.40           C
ATOM      3  C    SER A   2       25.972   8.992  51.295  1.00 55.44           C
ATOM      4  O    SER A   2       25.576   7.932  50.799  1.00 54.37           O
ATOM      5  CB   SER A   2       27.883  10.601  51.046  1.00 70.82           C
ATOM      6  OG   SER A   2       27.150  11.676  51.622  1.00 71.45           O
ATOM      7  N    LYS A   3       25.157   9.993  51.619  1.00141.28           N
ATOM      8  CA   LYS A   3       23.716   9.932  51.398  1.00140.16           C
-----------------------------------//-----------------------------------------------
ATOM     47  CA   PHE A   8       26.551  11.090  41.294  1.00 19.27           C
ATOM     48  C    PHE A   8       27.751  10.357  40.676  1.00 21.43           C
ATOM     49  O    PHE A   8       28.562  10.924  39.938  1.00 21.44           O
ATOM     50  CB   PHE A   8       27.022  12.362  41.991  1.00 21.68           C
ATOM     51  CG   PHE A   8       25.909  13.297  42.288  1.00 17.60           C
ATOM     52  CD1  PHE A   8       25.488  14.212  41.321  1.00 14.95           C
ATOM    495  CA   VAL A  68       23.860  22.610  40.452  1.00 14.12           C
ATOM    496  C    VAL A  68       25.259  22.196  40.854  1.00 13.41           C
ATOM   1164  CA   SER A 147       37.123  31.083  35.325  1.00 21.88           C
ATOM   1819  CD1  ILE A 229       38.888  21.450  53.055  1.00 29.11           C
ATOM   1820  OXT  ILE A 229       43.220  19.637  50.148  1.00 25.25           O
TER                  1821                                      ILE       A       229
HETATM 1822  O   HOH     1      30.450  20.682  37.367  1.00 15.75           O
HETATM 1823  O   HOH     2      26.443  24.175  38.999  1.00 18.82           O
-----------------------------------//-----------------------------------------------
HETATM 1831  O   HOH    10      29.132  18.648  45.101  1.00 13.77           O
HETATM 1832  O   HOH    11      24.076  46.248  42.794  1.00 22.62           O
HETATM 1833  O   HOH    12      31.870  32.426  52.146  1.00 36.77           O
HETATM 1880  O   HOH    59      37.243  14.571  53.463  1.00 31.12           O
HETATM 1881  O   HOH    60      40.360  20.483  56.144  1.00 32.74           O
HETATM 1882  O   HOH    61      13.483  49.374  33.179  1.00 30.77           O
CONECT  267  268
CONECT  268  267  269  271
CONECT  819  820
CONECT 1594 1592 1596 1598
CONECT 1595 1593 1596
CONECT 1596 1594 1595 1597
CONECT 1597 1596
CONECT 1598 1594
MASTER      259    0   10    6   12    0    0    6 1881    1  140   18
END
```

PDB-XML

PDBML: the representation of archival macromolecular structure data in XML. John Wesbrook, Nobutoshi Ito, Haruki Nakamura, Kim Henrick and Helen M. Berman, Bioinformatics, 21(7), 988-992, 2005.

```
<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock datablockName="1CFC"
  xmlns:PDBx="http://pdbml.pdb.org/schema/pdbx-v32.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:schemaLocation="http://pdbml.pdb.org/schema/pdbx-v32.xsd
pdbx-v32.xsd">
  <PDBx:atom_siteCategory>
    <PDBx:atom_site id="1">
      <PDBx:B_iso_or_equiv>1.43</PDBx:B_iso_or_equiv>
      <PDBx:B_iso_or_equiv_esd xsi:nil="true" />
      <PDBx:Cartn_x>14.550</PDBx:Cartn_x>
      <PDBx:Cartn_x_esd xsi:nil="true" />
      <PDBx:Cartn_y>12.461</PDBx:Cartn_y>
      <PDBx:Cartn_y_esd xsi:nil="true" />
      <PDBx:Cartn_z>-10.584</PDBx:Cartn_z>
      <PDBx:Cartn_z_esd xsi:nil="true" />
      <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
      <PDBx:auth_atom_id>N</PDBx:auth_atom_id>
      <PDBx:auth_comp_id>ALA</PDBx:auth_comp_id>
      <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
      <PDBx:group_PDB>ATOM</PDBx:group_PDB>
      <PDBx:label_alt_id></PDBx:label_alt_id>
      <PDBx:label_asym_id>A</PDBx:label_asym_id>
      <PDBx:label_atom_id>N</PDBx:label_atom_id>
      <PDBx:label_comp_id>ALA</PDBx:label_comp_id>
      <PDBx:label_entity_id>1</PDBx:label_entity_id>
      <PDBx:label_seq_id>1</PDBx:label_seq_id>
      <PDBx:occupancy>1.00</PDBx:occupancy>
      <PDBx:occupancy_esd xsi:nil="true" />
      <PDBx:pdbx_PDB_ins_code xsi:nil="true" />

<PDBx:pdbx_PDB_model_num>1</PDBx:pdbx_PDB_
model_num>
      <PDBx:pdbx_formal_charge xsi:nil="true" />
      <PDBx:type_symbol>N</PDBx:type_symbol>
    </PDBx:atom_site>
    <PDBx:atom_site id="2">
```


Спасибо за внимание!