

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 Зилья Мухтаровна

- 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, 10 ¹² пар оснований	Функциональная аннотация
Белковые последовательности	~5.5 · 10 ⁶ 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. ~10 ¹³ 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 10 ⁶ 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 крайней 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
- **Есосус** – 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

- Как сравнивают последовательности? Запишем одну последовательность под другой:

attgtACcTCgTgG-AA----
-----AC-TCaTaGcAAccaag

- 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**

NCBI Resources How To Sign in to NCBI

GenBank Nucleotide Search

GenBank Submit Genomes WGS HTGs EST/GSS Metagenomes TPA TSA INSDC

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp](#) site. The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#), which is divided into three divisions: [CoreNucleotide](#) (the main collection), [dbEST](#) (Expressed Sequence Tags), and [dbGSS](#) (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utilities](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

GenBank Resources

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)

• удобная в навигации база генетических последовательностей – Ensembl

Search: All species for

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Popular genomes



Human

GRCm38.p7



Human

GRCm37



Mouse

GRCm38.p4



Zebrafish

GRCz10

★ [Log in to customize this list](#)

All genomes

-- Select a species --

[View full list of all Ensembl species](#)

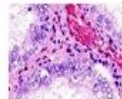
Still using Human GRCh37?



Variant Effect Predictor



Gene expression in different tissues



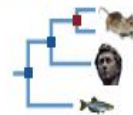
Find SNPs and other variants for my gene

```
GTRFATACATTC  
CRTRAAAGTCTT  
CTTCTAAATTCT  
GRAACATTTTCC
```

Retrieve gene sequence

```
GCCTGACTTCGGGTTGG  
GGGCTTGTGGGGGAGC  
GGGCTCTGCTGGGCTT  
AGGGGACAGATTTGTGA  
CACCTCTGGAGCGGTT  
CCCACTCCAGCCTGGCG
```

Compare genes across species



Use my own data in Ensembl

ENCODE data in Ensembl

What's New in Ensembl Release 85 (July 2016)

- Update to Ensembl-Havana human GENCODE gene set (release 25)
- 30 new epigenomes from the Roadmap Epigenomics Project
- New zebrafish maseq
- Update to Rat Ensembl-Havana gene set
- Mouse: update to Ensembl-Havana GENCODE gene set

[Full details](#) | [All web updates, by release](#) | [More news on our blog](#)

- 10 Mar 2016: [Ensembl 84 has been released!](#)
- 16 Feb 2016: [Learn about Ensembl – online, live and free!](#)
- 25 Jan 2016: [Sharing feature on the new mobile site \(m.ensembl.org\)](#)

[Go to Ensembl blog](#)

Tweets by @ensembl

e! Ensembl @ensembl
Understanding the mechanism of neuronal cox-2 gene suppression by 17β-Estradiol #UsingEnsembl @PLOSONE buff.ly/2cF11JL



2h

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

EMBL-EBI  [Services](#) [Research](#) [Training](#) [About us](#)

 **ENA**
European Nucleotide Archive

[Search](#)
Examples: [BN000085](#), [histone](#) [Advanced Sequence](#)

[Home](#) [Search & Browse](#) [Submit & Update](#) [About ENA](#) [Support](#)

Genomes at EBI

- [Complete genomes](#)
- [Archaea](#)
- [Archaeal virus](#)
- [Bacteria](#)
- [Eukaryota](#)
- [Organelle](#)
- [Phage](#)
- [Plasmid](#)
- [Viroid](#)
- [Virus](#)
- [Links](#)
- [WGS info](#)
- [EnsemblGenomes](#)
- [Ensembl](#)
- [Fasta33 Server](#)

Genomes Pages - At the EBI

Access to Completed Genomes

The first completed genomes from [viruses](#), [phages](#) and [organelles](#) were deposited into the EMBL Database in the early 1980's. Since then, molecular biology's shift to obtain the complete sequences of as many genomes as possible combined with major developments in sequencing technology resulted in hundreds of complete genome sequences being added to the database, including [Archaea](#), [Bacteria](#) and [Eukaryota](#). These web pages give access to a large number of complete genomes, [help](#) is available to describe the layout.

Whole Genome Shotgun Sequences (WGS)

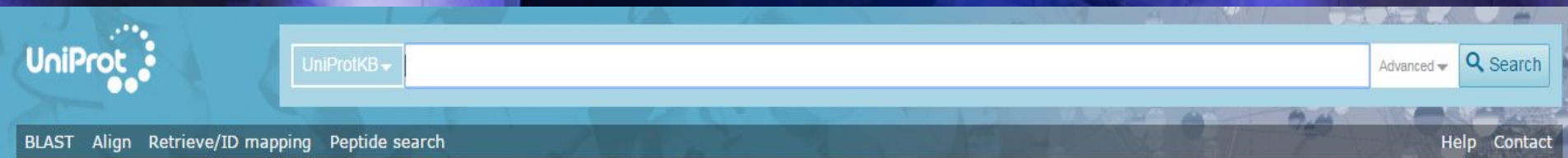
Methods using whole genome shotgun data are used to gain a large amount of genome coverage for an organism. WGS data for a growing number of organisms are being submitted to DDBJ/EMBL/GenBank.

[More information about WGS projects...](#)

Last 40 Genome Entries

Date	Accession	Description
02-MAY-2015	CP011047.1	Cronobacter sakazakii strain ATCC 29544

• крупнейший банк белковых данных – UniProt.org



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (551,987)
Manually annotated and reviewed.

TrEMBL (66,905,753)
Automatically annotated and not reviewed.

UniRef

Sequence clusters

UniParc

Sequence archive

Proteomes

Supporting data

Literature citations 	Taxonomy 	Subcellular locations
Cross-ref. databases 	Diseases XXX	Keywords

News

[BLOG](#) [Twitter](#) [Facebook](#) [RSS](#)

[Forthcoming changes](#)
Planned changes for UniProt

[UniProt release 2016_08](#)
Butterfly fashion: all they need is cortex | Cross-references to CDD | Change of the cross-references to VectorBase and WormBase | Pepti...

[UniProt release 2016_07](#)
(Bacterial) immigration under control

[News archive](#)

Getting started

Text search

Our basic text search allows you to search all the resources available



UniProt data

Download latest release

Get the UniProt data

Protein spotlight

On Releasing Tension

June 2016

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



VALIDATION ▾ DEPOSITION ▾ DATA DICTIONARIES ▾ DOCUMENTATION ▾ TASK FORCES ▾ STATISTICS ▾ ABOUT ▾



Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Learn more about PDB **HISTORY** and **FUTURE**.



Validate Structure

or View validation reports



Deposit Structure

All Deposition Resources



Download Archive

Instructions

wwPDB Members

wwPDB data centers serve as deposition, annotation, and distribution sites of the PDB archive. Each site offers tools for searching, visualizing, and analyzing PDB data.

RCSB PDB

- › Research Collaboratory for Structural Bioinformatics Protein Data Bank



Simple and advanced searching for macromolecules and ligands, tabular reports, specialized visualization tools, sequence-structure comparisons, RCSB PDB Mobile, Molecule of the Month and other educational resources.

wwPDB Resources

Data Dictionaries

- › Macromolecular Dictionary (PDBx/mmCIF)
- › Small Molecule Dictionary (CCD)
- › Peptide-like antibiotic and inhibitor molecules (BIRD)

Annotation

- › Procedures and policies
- › Improvements for consistency and accuracy

Community Input:

News & Announcements

08/05/2016

› wwPDB Validation Server Upgrade

The new wwPDB Validation Server at <https://validate.wwpdb.org> now generates preliminary validation reports for structures solved by NMR and 3D Electron Microscopy, in addition to X-ray crystallography.

[Read more](#)

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.

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.

Профессиональные ресурсы.

NAR database issue
BioMedNet
Medscape Molecular Medicine
Научные конференции

Патенты.

IBM Patent Server
United States Patent and Trademark Office
Get^{the}Patent.com

Литература.

Журналы.
Цитаты.
Импакт-фактор.
Книги.
Базы данных по абстрактам.

Модельные организмы.

Arabidopsis.
Drosophila
Mouse

Сток-центры (культуры клеток и микроорганизмы).

American Type Culture Collection (ATCC)
German Collection of Microorganisms and Cell Cultures (DSMZ)
Адреса других сток-центров.

Базы данных.

GeneCards.

Анализ нуклеотидных и белковых последовательностей.

BCM Search Launcher
ExPASy Molecular Biology Server
National Center for Biotechnology Information

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 *.src. 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.

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="SSINYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
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="MTQLQISLLLTATISLLHLVWATPYEAYPIGKQYPPVARVNESF

TFQISNDTYKSSVDKTAQITYNCFDLPWLSFDSSSRFTSGEPSSDLLSDANTTLYFN
-----//-----
YGSQKTVDTTEKLFDFLEAPEKEKRTSRDVTMSSLDPWNSNISPSVPRKSVTPSPYNVTK
RNRHLQNIQDSQSGKNGITPTTMTSSSDDFVVPKVDGENFCWVHSMEDRRPSSKRL
VDFSNKSNNVNGQVKDIHGRIPLEM"

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="MNRWVEKWLRVYLKCYINLILFYRNVPYPPQSFYDITYQSFNLQP
FVPINRHPALIDYIEELILDVLSKLTHTVYRFSICIHKNKNDLCIEKYVLDVDFSELQHVD
KDDQIITETEVFDEFRRSSLNLSLIMHLEKLPKVNDDTITFEAVINAIIELELGHKLDNRN
RVDSLEEKAEIERDSNWWKCQEDENLPDNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
LISGDDKILNGVYSQYEEGESIFGSLF"

ORIGIN
1 gatcctccat atacaacggt atccacact caggtttaga tctcaacaac ggaaccattg
61 ccgacatgag acagttagggt atcgtcgaga gttacaagct aaaacgagca gtatgcagct
121 ctgcatctga agccgctgaa gttctactaa ggggtggataa catcatccgt gcaagaccaa
181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccc
241 ccacactgtc attattataa ttgaaaacag aacgcaaaaa ttatccacta tataattcaa
301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa

-----//-----
4621 tcttcgcact tctttccca ttcattctt tcttctcca aagcaacgat ctttctacc
4681 atttgctcag agtccaatc ggcctcttc agttatcca ttgcttctt cagtttggct
4741 tcactgtctt cttagctgtt ttctagatcc tggttttct tgggttagtt ctcattatta
4801 gatcacaagt tattggagtc ttcagccaat tgctttgat cagacaattg actctctaac
4861 ttctccact cactgtcgag ttgctcgtt ttacgagaca aagattaat ctcgctttt
4921 ttttcagtgt tagatgctc taattcttg agctgtctc tcagctcctc atattttct
4981 tgccatgact cagattctaa ttttaagcta tcaatttct ctttgatc

//

GenBank. Запись sequence

```
/protein_id="CAA25860.1"  
/db_xref="GI:41298"  
/db_xref="GOA:P0C093"  
/db_xref="UniProtKB/Swiss-Prot:P0C093"  
/translation="MAEKQTAKRNRREEILQSLALMLESSDGSQRITTAKLAASVGVSEAAALYRHFPSKTRMFDSLIEFIEDSLITRINLILKDEKDTARLRLIVLLLLLGFGERNPGLTRILTGHALMFEQDRLQGRINQLFERIEAQLRQVLREKRMREGEGYTTDETLLASQILAFCEGMLSRFVRSEFKYRPTDDFDARWPLIAASCNMTPDDFSSGEFL"
```

ORIGIN

```
1 cagagaaaat caaaaagcag gccacgcagg gtgatgaatt aacaataaaa atggttaaaa  
61 accccgatat cgtcgcaggc gttgccgcac taaaagacca tcgaccctac gtcggttgat  
121 ttgccgccga aacaaataat gtggaagaat acgccggca aaaacgtatc cgtaaaaacc  
181 ttgatctgat ctgcgcgaac gatgtttccc agccaactca aggatttaac agcgacaaca  
241 acgcattaca ccttttctgg caggacggag ataaagtctt accgcttgag cgcaaagagc  
301 tccttggccca attattactc gacgagatcg tgacccgta tgatgaaaaa aatcgacggt  
361 aagattctgg acccgcgcgt tgggaaggaa tttccgctcc cgacttatgc cacctctggc  
421 tctgccggac ttgacctgcy tgcctgtctc aacgacgccg tagaactggc tccgggtgac  
481 actacgctgg ttccgaccgg gctggcgatt catattgccg atccttcaact ggccgcaatg  
541 atgctgccgc gctccggatt gggacataag cacggtatcg tgcttggtaa cctggtagga  
601 ttgatcgatt ctgactatca gggccagttg atgatttccg tgtggaaccg tggtcaggac  
661 agcttcacca ttcaacctgg cgaacgcac gcccagatga tttttgttcc ggtagtagac  
721 gctgaattta atctggtgga agatttccgac gccaccgacc gcggtgaagg cggctttggt  
781 cactctggtc gtcagtaaca catacgcac cgaataacgt cataacatag ccgcaaacat  
841 ttcgtttgcg gtcatagcgt ggggtgccgc tggcaagtgc ttattttcag gggatatttg  
901 taacatggca gaaaaacaaa ctgcgaaaag gaaccgtcgc gaggaaat ac ttcagtctct  
961 ggcgctgatg ctggaatcca gcgatggaag ccaacgtatc acgacggcaa aactggccgc  
1021 ctctgtcggc gtttccgaag cggcactgta tcgccacttc cccagtaaga cccgcatggt  
1081 cgatagcctg attgagtta tcgaagatag cctgattact cgcatcaacc tgattctgaa  
1141 agatgagaaa gacaccacag ccgcctgcy tctgattgtg ttgctgcttc tcggttttg  
1201 tgagcgtaat cctggcctga cccgcacact cactggtcat gcgctaagt ttgaacagga  
1261 tcgcctgcaa gggcgcacat accagctggt cgagcgtatt gaagcgcagc tgcgccaggt  
1321 attgcgtgaa aagagaatgc gtgaggggtga aggttacacc accgatgaaa ccctgctggc  
1381 aagccagatc ctggccttct gtgaaggat gctgtcacgt tttgtccgca gcgaatttaa  
1441 ataccgcccg acggatgatt ttgacgcccg ctggccgcta attgcccgca gttgcagtaa  
1501 tatgacgccg gatgactttt catccggcga gtttctttaa acgccaaact cttcgcgata  
1561 ggccttaacc gccgccagat gttccgccat ttccggettc 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
source Location/Qualifiers
1..960
/organism="Homo sapiens"
/mol_type="mRNA"
/db_xref="GeneID: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="MTPLCPRPALCYHFLTSLLRSAMQNARGTAEGRSRGTLRARPA
RPPAAQHGI PRPLSSAGRLSQGCRGASTVGAAGWKGELPKAGGS PAPGPET PAISPSK
RARP AEVGGMQLRFARLSEHATAPTRGSARAAGYDLYSAYDYTI PPMEKAVVKTDIQI
ALPSGCYGRVAPRSGLAAKHFIDVGAGVIDEDYRGNVGVVLFNFGKEKFEVKKGDRIA
QLICERIFYPEIEEVQALDDTERGSGGFGSTGKN"
[sig peptide](#)
63..269
/note="mitochondrial targeting presequence"
[mat peptide](#)
270..818
/product="deoxyuridine triphosphate nucleotidohydrolase"
ORIGIN
1 qgtqqaagcc tggcgcacgt ccggaaggtgc cgaagaccca accagccaa actctqqgg

Сплайсинг и восстановление последовательности mRNA

```
mRNA      /gene="DUT"  
            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,
```

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
source Location/Qualifiers
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="MTPLCPRPALCYHFLTSLLRSAMQONARGTAEGRSRGTLRARPAP
                     RPPAAQHGIPIRPLSSAGRLSQGCRGASTVGAAGWKGELPKAGGSPAPGPETPAISPSK
                     RARPAEYVGGMQLRFARLSEHATAPTRGSARAAGYDLYSAYDYTIIPPMKAVVKTDIQI
                     ALPSGCYGRVAPRSGLAAKHFIDVGAGVIDEDYRGNVGVVLFNFGKEKFEVKKGDRIA
                     QLICERIFYPEIEEVQALDDTERGSGGGFGSTGKN"
     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="MPCSEETPAISPSKRARPAEYVGGMQLRFARLSEHATAPTRGSAR
                     AAGYDLYSAYDYTIIPPMKAVVKTDIQIALPSGCYGRVAPRSGLAAKHFIDVGAGVID
                     EDYRGNVGVVLFNFGKEKFEVKKGDRIAQLICERIFYPEIEEVQALDDTERGSGGGFGS
                     TGKN"
     exon            1..45
                     /gene="DUT"
                     /number=4

```

GenBank. Аннотация

Как добавить данные в GB?

GenBank - Поиск в Google Submit to GenBank NCBI Sequence V

NCBI **Submit to GenBank**

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

NCBI

Submitting Sequence Data to GenBank [Submit now!!](#)

The most important source of new data for GenBank® is direct submissions from scientists. GenBank depends on its contributors to help keep the database as comprehensive, current, and accurate as possible. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries, and is ready to assist authors who have new data to submit.

[Sequin](#)
Stand-alone sequence submission tool

[BankIt](#)
For quick and simple submissions

[tbl2asn](#)
Command-line sequence submission tool

[dbEST](#)
[dbGSS](#)
[dbSTS](#)
Submit to GenBank divisions

Receiving an Accession Number for your Manuscript

Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before publication. Soon after submission, you will receive an accession number from the database which you will be able to use in your article to refer to the sequence. Please be aware that it is only necessary to submit the sequence to one database, whichever one is most convenient, without regard for where the sequence may be published. Data exchange between GenBank, EMBL and DDBJ occurs daily. Sequence data submitted in advance of publication can be kept confidential if requested.

GenBank

[GenBank](#)
Overview of the database

[Search GenBank](#)
Explore the data

[GenBank](#)
Guide to NCBI resources

Accession numbers
For manuscript citation

BankIt

Sequin

SequinMacroSend
Upload .sqn files directly

TBL2ASN
Command line program

Special submissions
Genomes, batch sequences, alignments

Whole Genome Shotgun
Sequence submissions

Third Party Annotation
TPA database

Зачем?

- информация в community;
- Журналы требуют это ДО публикации

Долго ли это?

2 рабочих дня

Данные могут 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 1RRXA 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 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)

PDB-XML

PDBML: the representation of archival macromolecular structure data in XML. John Westbrook, 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!**