

# Компьютерные технологии в науке

Базы данных.  
Инструменты поиска информации.

"An excellent book for beginners and occasional practitioners"  
Reviews, Journal of the American Medical Association

# Bioinformatics

FOR  
**DUMMIES**

2nd Edition

Updated to cover  
multiple new  
genomes and  
databases

**A Reference  
for the  
Rest of Us!**

FREE eTips at [dummies.com](http://dummies.com)

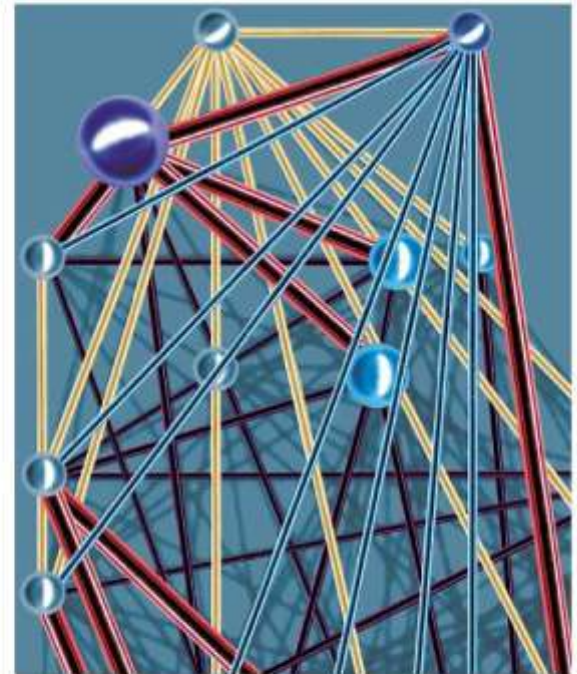
**Jean-Michel Claverie, PhD**  
Research Director, France's Centre National  
de la Recherche Scientifique (CNRS)

**Cedric Notredame, PhD**  
Professor of Bioinformatics, Switzerland's  
Lausanne University and the CNRS



# Bioinformatics

Sequence and Genome Analysis



David W. Mount

COLD SPRING HARBOR LABORATORY PRESS



- Все больше биологических данных, которые не публикуются обычным путем, а помещаются в базы данных с присвоением уникального идентификатора для ссылки при публикации.
- Данные от проектов по секвенированию геномов могут даже и не иметь ссылок в журнальных публикациях. Тем не менее, такие базы данных являются очень важным инструментом для биологических исследований.

# Биологические базы данных

- Биологические базы данных – это архивы согласованных данных, хранящихся в единой форме. Эти базы содержат данные широкого спектра разных областей молекулярной биологии.
- 
- Очень важно, что они, как правило, доступны через интернет и оснащены интуитивно понятно интерфейсом для поиска информации.

# Типы баз данных

- Библиографические (MEDLINE)
- Таксономические
- Нуклеотидные
  - Нуклеотидные последовательности
  - Геномные
  - Microarray Databases
- Белковые
  - Аминокислотные последовательности
  - «Вторичные» базы
- Пространственных структур макромолекул

- Первичные или архивные базы данных содержат аннотированные первичные структуры ДНК и белков, пространственные структуры нуклеиновых кислот и белков, а также **protein expression profiles** – профили экспрессии генов белков клеток.
- Вторичные (**derived**) базы данных содержат результаты анализов первичных источников, включая информацию о специфичных мотивах в последовательностях (**sequence patterns and motifs**), вариантах и мутациях, а также эволюционных связях. К этим же базам данных можно причислить и библиографические базы данных, такие как **Medline**.

# Существуют интегрированные системы для получения всей необходимой информации относительно объекта исследования

- SRS (Sequence Retrieval System) <http://srs.ebi.ac.uk/> является достаточно мощной системой запросов, существующей при Европейском Биоинформационном Институте EBI.
- Обеспечивает информацией из более чем 150 гетерогенных источников.

The screenshot displays the SRS web interface. At the top, there is a navigation bar with 'EMBL-EBI' and 'EBI Search' logos, and a search bar. Below the navigation bar, there are several sections:

- SRS**: A section with a link to 'Start a Permanent Project' and a 'Tips' box containing information about using SRS, linking to SRS, and public SRS servers worldwide.
- Quick Text Search**: A search box with 'Nucleotides' selected and a 'Search' button. Below the search box, it says 'Searches Databases: EMBL Nucleotides'.
- News and Announcements**: A section with a 'Search Tips' link and a list of announcements, including dates and notes about system updates and maintenance.
- List Search**: A section with a 'Search Tips' link and a text area for pasting a list of sequence IDs. Below the text area, there is a 'List file' input field and a 'Search' button.

At the bottom of the page, there is a footer with the text: 'Terms of Use Feedback & Support SRS Release 7.1.3.2 Copyright © 1997-2003 UCLON bioscience AD. All Rights Reserved.'

# Таксономические базы данных



# Таксономическая база – классификация всех организмов

- <http://www.ncbi.nlm.nih.gov/Taxonomy/> - самая популярная база данных. Расположена при NCBI. Иерархическая и основанная на нуклеотидных последовательностях генов.
- Цель – централизовать классификацию всех организмов, представленных в базе хотя бы одной последовательностью гена или белка.
- Может быть использована для определения положения исследуемого организма в иерархии или для получения последовательностей генов данного организма или группы организмов.



Search for As complete name lock Go Clear

# The NCBI Taxonomy Homepage

## Taxonomy Tip of the Day

**Did you know**

that a small number of sequences extracted from extinct organisms have been deposited at GenBank? These include DNA from the Neanderthal man, the woolly mammoth, the saber-toothed cat, and several giant New Zealand birds (moas) among others. A more complete list of extinct organisms that are represented in the public sequence database can be found [here](#).

- Taxonomy browser
- Archaea
- Bacteria
- Eukaryota
- Viroids
- Viruses
- Taxonomy common tree
- Taxonomy information
- Taxonomy resources
- Taxonomic advisors
- Genetic codes
- Taxonomy Statistics
- Taxonomy Name/Id Status Report
- Taxonomy FTP site
- FAQs
- How to reference the NCBI taxonomy database
- How to create links to the NCBI taxonomy
- How to create LinkOut links from the NCBI taxonomy

### These are direct links to some of the organisms commonly used in molecular research projects:

<a href="#">Arabidopsis thaliana</a>	<a href="#">Escherichia coli</a>	<a href="#">Pneumocystis carinii</a>
<a href="#">Bos taurus</a>	<a href="#">Hepatitis C virus</a>	<a href="#">Rattus norvegicus</a>
<a href="#">Caenorhabditis elegans</a>	<a href="#">Homo sapiens</a>	<a href="#">Saccharomyces cerevisiae</a>
<a href="#">Chlamydomonas reinhardtii</a>	<a href="#">Mus musculus</a>	<a href="#">Schizosaccharomyces pombe</a>
<a href="#">Danio rerio (zebrafish)</a>	<a href="#">Mycoplasma pneumoniae</a>	<a href="#">Takifugu rubripes</a>
<a href="#">Dictyostelium discoideum</a>	<a href="#">Oryza sativa</a>	<a href="#">Xenopus laevis</a>
<a href="#">Drosophila melanogaster</a>	<a href="#">Plasmodium falciparum</a>	<a href="#">Zea mays</a>

Comments and questions to [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov)  
 Credits: Joe Bischoff, Mikhail Domrachev, Scott Federhen, Carol Hottom, Detlef Leipe, Vladimir Sousoy, Richard Sternberg, Sean Turner.

# Другие примеры таксономических баз данных:

- NEWT <http://www.ebi.ac.uk/newt/>
- The Tree of Life project  
<http://tolweb.org/tree/phylogeny.html>
- Species 2000 <http://www.sp2000.org/>
- International Organization for Plant Information  
<http://iopi.csu.edu.au/iopi/>
- Integrated Taxonomic Information System  
<http://www.itis.usda.gov/itis/>
- Таксономические базы достаточно противоречивы из-за различных взглядов на классификацию организмов

# Нуклеотидные базы данных

The International Nucleotide Sequence Database Collaboration



<http://www.insdc.org/>

1. EMBL-Bank at the European Bioinformatics Institute (EBI) <http://www.ebi.ac.uk/embl/index.html>
2. The DNA Data Bank of Japan (DDBJ) at the Center for Information Biology (CIB)  
<http://www.ddbj.nig.ac.jp/>
3. GenBank at the National Center for Biotechnology Information (NCBI)  
<http://www.ncbi.nlm.nih.gov/Genbank/>

# EMBL-Bank at the European Bioinformatics Institute (EBI)

<http://www.ebi.ac.uk/embl/index.html>

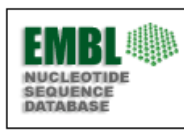
- В Европе большая часть нуклеотидных последовательностей помещается в **EMBL Nucleotide Sequence Database** располагаемый при Европейском Биоинформационном Институте в Кембридже (Великобритания)



EBI > Databases > EMBL-Bank

## EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are [direct submissions](#) from individual researchers, genome sequencing projects and patent applications.



The database is produced in an international [collaboration](#) with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The [current database release](#) (Release 93, Dec 2007), with according [Release notes](#) and [user manual](#) are available from the EBI servers. A sample database entry is shown [here](#).

A publication in [Nucleic Acids Research 2008 Jan \(Database issue\) \[pub ahead of print\]](#), provides further information and details.

The EMBL nucleotide sequence database is part of the [The Protein and Nucleotide Database Group \(PANDA\)](#). This is jointly headed by [Dr. Rolf Apweiler](#) and [Dr. Ewan Birney](#), with Dr. Birney taking responsibility for Nucleotides.

Link	Explanation
<a href="#">Access</a>	<a href="#">Database queries</a> , <a href="#">Completed genomes webservice</a> , <a href="#">FTP archives</a> (EMBL release, alignments etc), <a href="#">EMBL sequence version archive</a> (SVA), <a href="#">Browse by geography</a> .
<a href="#">Submission</a>	Primary sequence submissions, third party annotation, updates and alignment submissions.
<a href="#">Documentation</a>	<a href="#">Release notes user manual</a> , <a href="#">Information for Submitters</a> , <a href="#">FAQ</a> , <a href="#">Release information</a> , <a href="#">Forthcoming Changes</a> , <a href="#">EMBL database statistics</a> , <a href="#">Feature table</a> , <a href="#">XML documentation</a> , <a href="#">Sample entry</a> , <a href="#">Accession Number Prefix Codes</a> , <a href="#">Examples of annotation</a> , <a href="#">EMBL Features &amp; Qualifiers</a> , <a href="#">DE line standards</a> , <a href="#">Database Policies</a>
<a href="#">Publications</a>	Group publications
<a href="#">People</a>	Group members
<a href="#">Contact</a>	How to contact the EMBL Nucleotide Sequence Database
<a href="#">News</a>	List of recent changes on this site

### Contact

For information, comments and/or suggestions, please use the EBI Support Form page <http://www.ebi.ac.uk/support/>

- EMBL-Bank Home
  - Access
  - Documentation
  - News
  - Submission
  - Publications
  - People
  - Contact
- 
- EMBL Fetch**
- Fetch an EMBL record by id
- 
- 
- 
- TPA - Third Party Annotation**
- [Users can now submit re-annotations/re-assemblies of sequences already present in EMBL and owned by other groups.](#)
- 
- Collaborations**
- [INSDC](#) - International Nucleotide Sequence Database Collaboration
  - [NCBI](#) - The Nucleotide Sequence Database is produced in collaboration with GenBank (USA)
  - [DDBJ](#) - The Nucleotide Sequence Database is also produced in collaboration with the DNA Database of Japan (DDBJ)

Каждая запись в банке имеет уникальных идентификатор (*accession number*), который может указываться при публикации статей. Кроме того, запись может иметь номер версии, в случае если обнаружена ошибка. Поскольку за каждую запись отвечает свой собственный автор, то достаточно часто встречаются ошибки в наименовании, могут быть загрязнены, неполны, неправильно аннотированы, содержать ошибки чтения.

# GenBank at the National Center for Biotechnology Information (NCBI)

<http://www.ncbi.nlm.nih.gov/Genbank/>

- Наиболее известная база данных GenBank при Национальном центре Биотехнологической Информации (США)





**What does NCBI do?**

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

[More...](#)

**1 Billion Live Traces**  
The Trace Archive of sequencing traces has reached 1 billion live traces from over 480 organisms. For more information about the Trace Archive database [click here](#).

**NCBI Bookshelf**

Five new medical textbooks are available on the NCBI Bookshelf. Topics include epilepsy, Parkinson's disease, alternative medicine, rehabilitation, and spinal cord medicine. Search these books and many more in the Entrez Books database [here](#).

**PubMed Central**

An archive of life sciences journals  
● Free fulltext  
● Over 500,000 articles from over 200 journals  
● Linked to PubMed and fully searchable  
Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.

**Hot Spots**

- ▶ Assembly Archive
- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home
- ▶ Entrez Tools
- ▶ Gene expression omnibus (GEO)
- ▶ Human genome resources
- ▶ Malaria genetics & genomics
- ▶ Map Viewer
- ▶ dbMHC
- ▶ Mouse genome resources
- ▶ My NCBI
- ▶ ORF finder
- ▶ Rat genome resources
- ▶ Reference

**GenBank Overview**[PubMed](#) [Entrez](#) [BLAST](#) [OMIM](#) [Books](#) [Taxonomy](#) [Structure](#)

NCBI

SITE MAP

**► What is GenBank?**

GenBank<sup>®</sup> is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research 2004 Jan 1;32\(1\):23-6](#)). There are approximately 37,893,844,733 bases in 32,549,400 sequence records as of February 2004 (see [GenBank growth statistics](#)). GenBank records are annotated using a standard set of biological terms and show these annotations in a [Feature Table](#). As an example, you may view the [record](#) for a *Saccharomyces cerevisiae* gene. The complete [release notes](#) for the current version of GenBank are available. A new release is made every two months. 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.

**► Submissions to GenBank**

Many journals require [submission of sequence information](#) to a database prior to publication so that an accession number may appear in the paper. NCBI has a WWW form, called [BankIt](#), for convenient and quick submission of sequence data. [Sequin](#), NCBI's stand-alone submission software for MAC, PC, and UNIX platforms, is also available by FTP. When using Sequin, the output files for direct submission should be sent to GenBank by electronic mail.

**GenBank Flat File Format**

*Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the [Alphabetical Quicklinks Table](#) or [Resource Guide](#)*

```

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
  source          Location/Qualifiers
                  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="SSIIYNGLISTSGLDLNNGTIADMRQLGIVESYKLRKRAVSSASEA
                  AEVLLRVDNIIIRARPRTANRQHM"
  gene           687..3158
                  /gene="AXL2"
  CDS           687..3158
    
```

## ▸ Submissions to GenBank

Many journals require [submission of sequence information](#) to a database prior to publication so that an accession number may appear in the paper. NCBI has a WWW form, called [BankIt](#), for convenient and quick submission of sequence data. [Sequin](#), NCBI's stand-alone submission software for MAC, PC, and UNIX platforms, is also available by FTP. When using Sequin, the output files for direct submission should be sent to GenBank by electronic mail.

There are specialized, streamlined procedures for batch submissions of sequences, such as [EST](#), [STS](#), and [HTG](#) sequences.

## ▸ Updating or Revising a Sequence

Revisions or updates to GenBank entries can be made at any time and can be accepted as [BankIt](#) or [Sequin](#) files or as the text of an e-mail message. Click on the link for more information about [updating information on GenBank records](#).

## ▸ Access to GenBank

GenBank is available for [searching](#) at NCBI via several methods.

The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess

# NCBI Searching GenBank

- PubMed
- Entrez
- BLAST
- OMIM
- Books
- Taxonomy
- Structure

NCBI  
SITE MAP

## Text and Similarity Searching

### Entrez Browser

GenBank (nucleotides and proteins), PubMed (MEDLINE), 3D structures, genomes, and PopSet databases. GenBank nucleotide records are found in the divisions CoreNucleotide, dbEST or dbGSS. Entrez queries can search these three databases together or separately.

### BLAST Sequence Similarity Searching

Nucleotide or protein query sequences against the specified database using the BLAST suite of algorithms. GenBank nucleotide records are located in separate databases that must be searched independently. These include dbEST and dbGSS, plus multiple databases for the CoreNucleotide division, including nr, htgs, wgs and env\_nt. See the [BLAST info](#) page for more information about the numerous BLAST databases.

### dbEST Searching

dbEST (Database of Expressed Sequence Tags).

### dbSTS Searching

dbSTS (Database of Sequence Tagged Sites).

### dbGSS Searching

dbGSS (Database of Genome Survey Sequences).

## Information about Access to GenBank

### Network Client/Server Applications

Network BLAST



Search across databases  GO CLEAR Help

Welcome to the Entrez cross-database search page

- PubMed: biomedical literature citations and abstracts Books: online books
- PubMed Central: free, full text journal articles OMIM: online Mendelian Inheritance in Man
- Site Search: NCBI web and FTP sites OMIA: online Mendelian Inheritance in Animals

- Nucleotide: sequence database (GenBank) UniGene: gene-oriented clusters of transcript sequences
- Protein: sequence database CDD: conserved protein domain database
- Genome: whole genome sequences 3D Domains: domains from Entrez Structure
- Structure: three-dimensional macromolecular structures UniSTS: markers and mapping data
- Taxonomy: organisms in GenBank PopSet: population study data sets
- SNP: single nucleotide polymorphism GEO Profiles: expression and molecular abundance profiles
- Gene: gene-centered information GEO DataSets: experimental sets of GEO data
- HomoloGene: eukaryotic homology groups Cancer Chromosomes: cytogenetic databases
- PubChem Compound: unique small molecule chemical structures PubChem BioAssay: bioactivity screens of chemical substances
- PubChem Substance: deposited chemical substance records GENSAT: gene expression atlas of mouse central nervous system
- Genome Project: genome project information Probe: sequence-specific reagents

- Journals: detailed information about the journals indexed in PubMed and other Entrez databases MeSH: detailed information about NLM's controlled vocabulary
- NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections

Enter terms and click 'GO' to run the search against ALL the databases, OR  
Click Database Name or icon to go directly to the Search Page for that database, OR  
Click Question Mark for a short explanation of that database.

# 16S rRNA Thermus thermophilus

The Entrez Nucleotides database is a collection of sequences from several sources, including GenBank, RefSeq, and PDB. The number of bases in these databases continues to grow at an exponential rate. As of June 2005, there are over 89 billion bases in GenBank and RefSeq alone.

### Human Genome

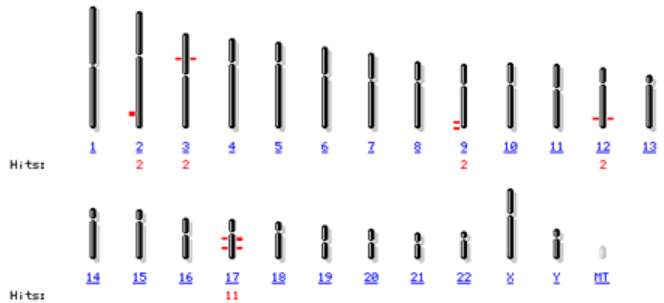
Explore [human genome resources](#) or browse the human genome sequence using the [Map Viewer](#).

### Building the human genome

The Human Genome Reference DNA Sequence was completed in April 2003. The current version is listed as a build number on the [Genome View](#) page and includes an accompanying set of [statistics](#) and [release notes](#).

### Homo sapiens genome view

build 35 version 1 statistics



The chromosomal locations of several genes believed to be associated with the human BRCA1 gene implicated in breast cancer, highlighted using the Map Viewer query "BRCA1" (build [35](#)).


 Search Nucleotide for 16S rRNA Thermus thermophilus [Go](#) [Clear](#) [Save Search](#)
[Limits](#) [Preview/Index](#) [History](#) [Clipboard](#) [Details](#)

Display Summary Show 20 Send to

 All: 63 **bacteria: 59** mRNA: 0 RefSeq: 20

Items 41 - 59 of 59

Previous Page 3 of 3

- |                          |  |         |       |
|--------------------------|--|---------|-------|
| <input type="checkbox"/> | <a href="#">41: AY554280</a>   | Reports | Links |
|                          | Thermus thermophilus 16S ribosomal RNA gene, partial sequence<br>gi 45385854 gb AY554280.1 [45385854]                                      |         |       |
| <input type="checkbox"/> | <a href="#">42: AY497773</a>   | Reports | Links |
|                          | Thermus thermophilus 16S ribosomal RNA gene, partial sequence<br>gi 40744587 gb AY497773.1 [40744587]                                      |         |       |
| <input type="checkbox"/> | <a href="#">43: 1KUQB</a>  | Reports | Links |
|                          | Chain B, Crystal Structure Of T3c Mutant S15 Ribosomal Protein In Complex With 16s Rrna<br>gi 33357250 pdb 1KUQB 33357250]                 |         |       |
| <input type="checkbox"/> | <a href="#">44: X58342</a>   | Reports | Links |
|                          | T.thermophilus 16S ribosomal RNA, part<br>gi 48243 emb X58342.1 TTHE16S[48243]   |         |       |
| <input type="checkbox"/> | <a href="#">45: X58341</a>   | Reports | Links |
|                          | T.flavus 16S ribosomal RNA, part<br>gi 48149 emb X58341.1 TFLA16S[48149]   |         |       |
| <input type="checkbox"/> | <a href="#">46: AJ251938</a>   | Reports | Links |
|                          | Thermus thermophilus 16S rRNA gene, strain CS<br>gi 6901428 emb AJ251938.1 TTH251938[6901428]  |         |       |
| <input type="checkbox"/> | <a href="#">47: 1EG0O</a>  | Reports | Links |
|                          | Chain O, Fitting Of Components With Known Structure Into An 11.5 A Cryo-Em Map Of The E.Coli 70s Ribosome<br>gi 7245465 pdb 1EG0O 7245465] |         |       |
| <input type="checkbox"/> | <a href="#">48: 1EG0M</a>  | Reports | Links |
|                          | Chain M, Fitting Of Components With Known Structure Into An 11.5 A Cryo-Em Map Of The E.Coli 70s Ribosome<br>gi 7245464 pdb 1EG0M 7245464] |         |       |
| <input type="checkbox"/> | <a href="#">49: 1EG0L</a>  | Reports | Links |
|                          | Chain L, Fitting Of Components With Known Structure Into An 11.5 A Cryo-Em Map Of The E.Coli 70s Ribosome<br>gi 7245463 pdb 1EG0L 7245463] |         |       |



1: [AJ251938](#). Reports *Thermus thermophi...* [gi:6901428]

Links

[Features](#) [Sequence](#)

LOCUS TTH251938 1477 bp DNA linear BCT 19-APR-2002

 DEFINITION *Thermus thermophilus* 16S rRNA gene, strain CS.

ACCESSION AJ251938

VERSION AJ251938.1 GI:6901428

KEYWORDS 16S ribosomal RNA; 16S rRNA gene.

 SOURCE *Thermus thermophilus*

 ORGANISM [Thermus thermophilus](#)

 Bacteria; Deinococcus-Thermus; Deinococci; Thermales; Thermaceae; *Thermus*.

REFERENCE 1

AUTHORS Lyon,P.F., Beffa,T., Blanc,M., Auling,G. and Aragno,M.

TITLE Isolation and characterization of highly thermophilic xylanolytic

*Thermus thermophilus* strains from hot composts

 JOURNAL *Can. J. Microbiol.* 46 (11), 1029-1035 (2000)

 PUBMED [11109491](#)

REFERENCE 2 (bases 1 to 1477)

AUTHORS Lyon,P.F.

TITLE Direct Submission

JOURNAL Submitted (16-DEC-1999) Lyon P.F., Microbiology Laboratory,

University of Neuchatel, Case Postale 2, 2007 Neuchatel,

SWITZERLAND

FEATURES Location/Qualifiers

source	1..1477
	/organism="Thermus thermophilus"
	/mol_type="genomic DNA"
	/strain="CS"
	/db_xref="taxon:274"
	/country="Switzerland"
	/note="isolate from hot composts"
<a href="#">gene</a>	8..1477
	/gene="16S rRNA"
<a href="#">rRNA</a>	8..1477
	/gene="16S rRNA"
	/product="16S ribosomal RNA"

ORIGIN

```

1 agagtttgat catggctcag ggtgaacgct ggcggcgtgc ctaagacatg caagtcgtgc
61 gggccgcggg gttttactcc gtggtcagcg gcggacgggt gaataacgcg tgggtgacct
121 accccgaaga gggggacaac ccggggaaac tcgggctaata ccccatatgt gaccgccccc
181 ttgggggtgtg tccaaaggcg tttgcccgct tccggatggg ccgcggtccc atcagctagt
241 tggtaggggta atggcccacc aaggcgcgca cgggtagccc gcctgagagg gtggccggcc
301 acaggggcac tgagacacgg gcccccactcc tacgggagcg agcagttagg aatcttcccg
361 aatggggcgca agcctgacgg agcgcgcgcg cttggaggaa gaagcccttc ggggtgtaaa
421 ctctgaacc cyggacgaaa cccccgcgca ggggactgac ggtaccgggy taatagcgcc
481 ggccaactcc gtgccagcag ccgcggtaat acggagggcy cgagcggtac ccggattcac
541 tgggcgtaaa gggcgtgtag gcggcctggg gcgtcccata tgaagacca cggctcaacc
601 gtgggggagc gtgggatacg ctcaggctag acggtgggag agggtggtgg aattcccgga
661 gtacgggtga aatgcccaga taccgggagg aacgcccgat gcgaaggcag ccacctggtc
721 caccctgtac gctgagggcg gaaagcgtgg ggagcaaac ggattagata cccgggtagt
781 ccacgcccta aacgatgcgc gctaggtctc tgggtctctc gggggccgaa gctaacgcgt
841 taagcgcgcc gcctggggag tacggccgca aggctgaaac tcaaaggaat tgacggggc

```

1: [1QD7A](#) Reports Chain A, Partial ...[gi:6137695]
 Links
[Comment](#) [Features](#) [Sequence](#)

LOCUS 1QD7\_A 271 bp DNA linear BCT 09-JUL-1999  
 DEFINITION Chain A, Partial Model For 30s Ribosomal Subunit.  
 ACCESSION 1QD7\_A  
 VERSION 1QD7\_A GI:6137695  
 KEYWORDS .  
 SOURCE Thermus thermophilus  
 ORGANISM [Thermus thermophilus](#)  
 Bacteria; Deinococcus-Thermus; Deinococci; Thermales; Thermaceae;  
 Thermus.

REFERENCE 1 (bases 1 to 271)  
 AUTHORS Ramakrishnan,V. and White,S.W.  
 TITLE The structure of ribosomal protein S5 reveals sites of interaction with 16S rRNA  
 JOURNAL Nature 358 (6389), 768-771 (1992)  
 PUBMED [1508272](#)

REFERENCE 2 (bases 1 to 271)  
 AUTHORS Lindahl,M., Svensson,L.A., Liljas,A., Sedelnikova,I.A., Eliseikina,I.A., Fomenkova,M.P., Nevskaya,N., Nikonov,S.V., Garber,M.B., Muranova,T.A., Rykonova,A.I. and Amons,R.  
 TITLE Crystal structure of the ribosomal protein S6 from Thermus thermophilus  
 JOURNAL EMBO J. 13 (6), 1249-1254 (1994)  
 PUBMED [8137808](#)

REFERENCE 3 (bases 1 to 271)  
 AUTHORS Jaishree,T.N., Ramakrishnan,V. and White,S.W.  
 TITLE Solution structure of prokaryotic ribosomal protein S17 by high-resolution NMR spectroscopy  
 JOURNAL Biochemistry 35 (9), 2845-2853 (1996)  
 PUBMED [8608120](#)

REFERENCE 4 (bases 1 to 271)  
 AUTHORS Wimberly,B.T., White,S.W. and Ramakrishnan,V.  
 TITLE The structure of ribosomal protein S7 at 1.9 Å resolution reveals a beta-hairpin motif that binds double-stranded nucleic acids  
 JOURNAL Structure 5 (9), 1187-1198 (1997)  
 PUBMED [9331418](#)

REFERENCE 5 (bases 1 to 271)  
 AUTHORS Clemons,W.M. Jr., Davies,C., White,S.W. and Ramakrishnan,V.  
 TITLE Conformational variability of the N-terminal helix in the structure of ribosomal protein S15  
 JOURNAL Structure 6 (4), 429-438 (1998)  
 PUBMED [9562554](#)

REFERENCE 6 (bases 1 to 271)  
 AUTHORS Nevskaya,N., Tishchenko,S., Nikulin,A., al-Karadaghi,S., Liljas,A., Ehresmann,B., Ehresmann,C., Garber,M. and Nikonov,S.  
 TITLE Crystal structure of ribosomal protein S8 from Thermus thermophilus reveals a high degree of structural conservation of a specific RNA binding site  
 JOURNAL J. Mol. Biol. 279 (1), 233-244 (1998)  
 PUBMED [9636713](#)

REFERENCE 7 (bases 1 to 271)

Search Nucleotide for Qbeta replicase Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Send to

All: 9 bacteria: 3 mRNA: 2 RefSeq: 0

Show only records from: CoreNucleotide (9), EST (0), GSS (0). [What's this?]

Items 1 - 9 of 9 One page.

- 1: [L07339](#) Reports Links  
Synthetic RNA species SV7, replicated in vitro by Qbeta replicase  
gi|209413|gb|L07339.1|SYNSV7A[209413]
- 2: [L07338](#) Reports Links  
Synthetic RNA species SV5, replicated in vitro by Qbeta replicase  
gi|209412|gb|L07338.1|SYNSV5A[209412]
- 3: [E03191](#) Reports Links  
DNA encoding HF-I(host factorI) of Q beta replicase  
gi|2171408|dbj|E03191.1|pat|JP|1991255097|1[2171408]
- 4: [X54505](#) Reports Links  
E. coli RQ87-3 RNA  
gi|297139|emb|X54505.1|ECRQ873[297139]
- 5: [X54506](#) Reports Links  
E. coli RQ223+1 RNA  
gi|297138|emb|X54506.1|ECRQ2231[297138]
- 6: [L07337](#) Reports Links  
Synthetic RNA species SV11, replicated in vitro by Qbeta replicase  
gi|209395|gb|L07337.1|SYNSV11A[209395]
- 7: [M24871](#) Reports Links  
Bacteriophage Q-beta nanovariant WSIII RNA  
gi|215722|gb|M24871.1|PQBNVWSIC[215722]
- 8: [M24870](#) Reports Links  
Bacteriophage Q-beta nanovariant WSII RNA  
gi|215721|gb|M24870.1|PQBNVWSIB[215721]
- 9: [M24869](#) Reports Links  
Bacteriophage Q-beta nanovariant WSI RNA

1: E03191 Reports DNA encoding HF-I...[gi:2171408]

[Comment](#) [Features](#) [Sequence](#)

LOCUS E03191 309 bp DNA linear PAT 29-SEP-1997  
 DEFINITION DNA encoding HF-I(host factor I) of Q beta replicase.  
 ACCESSION E03191  
 VERSION E03191.1 GI:2171408  
 KEYWORDS JP 1991255097-A/1.  
 SOURCE Escherichia coli  
 ORGANISM [Escherichia coli](#)  
 Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;  
 Enterobacteriaceae; Escherichia.  
 REFERENCE 1 (bases 1 to 309)  
 AUTHORS Kajitani,M., Ishihama,A. and Nakamura,H.  
 TITLE HOST FACTOR-I OF QBETA REPLICASE AND GENE CODING THE SAME  
 JOURNAL Patent: JP 1991255097-A 1 13-NOV-1991;  
 TORAY IND INC  
 COMMENT OS Escherichia coli  
 PN JP 1991255097-A/1  
 PD 13-NOV-1991  
 PF 02-MAR-1990 JP 1990052273  
 PI KAJITANI MASAYUKI, ISHIHAMA AKIRA, NAKAMURA HARUJI PC  
 C07K13/00,C12N15/31//C12P21/00,(C12N15/31,C12R1:19),(C12P21/00, PC  
 C12R1:19);  
 CC strandedness: Double;  
 CC topology: Linear;  
 CC hypothetical: No;  
 CC anti-sense: No;  
 CC \*source: clone=clone 3A1(625);  
 FH Key Location/Qualifiers  
 FT CDS 1..309  
 FT /product='HF-I(host factor I) of Q beta FT  
 replicase'.

FEATURES Location/Qualifiers  
 source 1..309  
 /organism="Escherichia coli"  
 /mol\_type="genomic DNA"  
 /db\_xref="taxon:562"

ORIGIN  
 1 atggctaagg ggcaatcttt acaagatccg ttcctgaacy cactgctgog ggaacgtgtt  
 61 ccagtttcta tttatttggg gaatgggtatt aagctgcaag ggcaaatcga gtcttttgat  
 121 cagttcgtga tctcgttgaa aaacacggtc agccagatgg ttacaagca cgcgatttct  
 181 actgttgytcc cgtctcggcc ggtttctcat cacagtaaca acgcoggtgy cgytaccagc  
 241 agtaactacc atcatggtag cagcgcgcag aatacttccg cgcaacagga cagcgaagaa  
 301 accgaataa

//

- Нуклеотидные базы данных принимают информацию по последовательностям ДНК и предоставляют открытый доступ к ней.
- Доступ к приведенным базам данных бесплатный и возможен через интернет.
- DDBJ, GenBank и EMBL-Bank обмениваются информацией ежедневно, в результате чего содержат практически идентичную информацию.

# Другие нуклеотидные базы

- Genomes Server (<http://www.ebi.ac.uk/genomes/>) – доступ к большому числу полных геномов
- UniGene (<http://www.ncbi.nlm.nih.gov/UniGene/>) – база по кластерам генов, которая направлена на проблему избыточности последовательностей. Соединяет сиквенсы, которые достаточно близки .
- STACK (<http://www.sanbi.ac.za/Dbases.html>) - 'Sequence Tag Alignment and Consensus Knowledgebase' - схожая с предыдущей по задачам база
- EMBL-SVA (<http://www.ebi.ac.uk/embl/sva/>) - 'EMBL Sequence Version Archive' содержит ВСЕ записи с первого выпуска EMBL database. Содержит более чем 100 миллионов записей.

# Специализированные

- RDP (<http://rdp.cme.msu.edu/html/>) - the 'Ribosomal Database Project' – обеспечивает информацией касательно рибосомы, включая анализ данных online, филогенетические деревья по рРНК, выравненные и аннотированные последовательности рРНК
- HIV-SD (<http://hiv-web.lanl.gov/>) - the 'HIV Sequence Database' собирает, курирует и аннотирует ДНК последовательности HIV и SIV; обеспечивает различными инструментами по анализу этих данных
- IMGT (<http://www.ebi.ac.uk/imgt/>) - the 'ImMunoGeneTics database' – специализируется на иммуноглобулинах, Т-клеточных рецепторах и главном комплексе гистосовместимости (МНС) всех видов ПОЗВОНОЧНЫХ

# Специализированные

- TRANSFAC (<http://transfac.gbf.de/TRANSFAC/>) – содержит данные по ДНК транскрипционных факторов и связывающих их участков
- EPD (<http://www.epd.isb-sib.ch/>) - the 'Eukaryotic Promoter Database' - аннотированная коллекция эукариотических POU II промоторов, для которых экспериментально определен стартовый сайт транскрипции
- REBASE (<http://rebase.neb.com/rebase/>) – база по ферментам рестрикции и сайтам рестрикции
- GOBASE (<http://megasun.bch.umontreal.ca/gobase/gobase.html>) специализированная база по геномам органелл



# Геномные базы данных

- Геномные базы значительно различаются по форме и содержанию.
- Для наиболее важных и интересных с точки зрения генетики организмов существуют опубликованные каталоги генов и мутаций в них. В последние годы многие каталоги переведены в электронную форму.
- Кроме того, создано много новых баз отличающихся по подборке генов и способе представления информации.

- Genomes Server <http://www.ebi.ac.uk/genomes/> – доступ к геномам различных организмов
- Proteome Analysis <http://www.ebi.ac.uk/proteome/index.html> – база данных с упором на статистический и сравнительный анализ предсказанных протеомов полностью сиквенированных организмов
- Ensembl <http://www.ebi.ac.uk/ensembl/index.html> – совместный проект EBI и Wellcome Trust Sanger Institute, созданный с целью разработки системы, которая поддерживает автоматическое аннотирование больших эукариотических геномов. Представляет постоянно обновляемую базу с автоматическим аннотированием геномов многоклеточных. Доступны человек, мышь, крыса, рыба фугу, zebrafish, комар, *Drosophila*, *C. elegans* и *C. briggsae*.

# Ensembl <http://www.ebi.ac.uk/ensembl/index.html>

The screenshot shows the Ensembl Genome Browser homepage in a web browser window. The browser's address bar displays <http://www.ebi.ac.uk/ensembl/index.html>. The page layout includes a left-hand navigation menu, a main content area with a header and several sections, and a footer.

**Left-hand navigation menu:**

- Ensembl Home
- Human Genome
- Mouse Genome
- Mosquito Genome
- Trace repository
- Latest Annual Report

**Main Content Area:**

**Ensembl Genome Browser**

Ensembl is a joint project between the [EMBL-EBI](#) and the [Wellcome Trust Sanger Institute](#) that aims at developing a system that maintains automatic annotation of large eukaryotic genomes. Access to all the software and data is free and without constraints of any kind. The project is primarily funded by the [Wellcome Trust](#). It is a comprehensive source of stable annotation with confirmed gene predictions that have been integrated from external data sources: [Ensembl](#) annotates known genes and predicts new ones, with functional annotation from [InterPro](#), [OMIM](#), [SAGE](#) and gene families. [Download Browsing genomes PDF](#)

Ensembl is part of the [The Protein and Nucleotide Database Group @PANDA](#). This is jointly headed by [Dr Rolf Apweiler](#) and [Dr Ewan Birney](#), with [Dr Birney](#) taking responsibility for Nucleotides.

**Browse A Genome**

**Mammalian genomes**

- [Bos taurus](#)
- [Canis familiaris](#)
- [Canis porcellus](#)
- [Dasypus novemcinctus](#)
- [Echinops telfairi](#)
- [Erinaceus europaeus](#)
- [Felis catus](#)
- [Homo sapiens](#)
- [Lorodonta africana](#)
- [Macaca mulatta](#)
- [Microcebus murinus](#) **NEW**
- [Monodelphis domestica](#)
- [Mus musculus](#)
- [Myotis lucifugus](#)
- [Orchotona princeps](#) **NEW**
- [Ornithomyces anatinus](#)
- [Oryzias latipes](#)
- [Oryzias latipes](#)
- [Pan troglodytes](#)
- [Rattus norvegicus](#)
- [Spermophilus beletemineatus](#)
- [Tupaia belangeri](#)

**Other species**

- [Aedes aegypti](#)
- [Anopheles gambiae](#)
- [Caenorhabditis elegans](#)
- [Clonostylopsis](#)
- [Clonostylopsis](#)
- [Dario rerio](#)
- [Drosophila melanogaster](#)
- [Fugu rubripes](#)
- [Gallus gallus](#)
- [Gasterosteus aculeatus](#)
- [Oryzias latipes](#)
- [Saccharomyces cerevisiae](#)
- [Takifugu rubripes](#)
- [Tetraodon nigroviridis](#)
- [Xenopus tropicalis](#)

**Ensembl for Schools**

The purpose of these kits is to give you interesting and fun ways to show students how biological research promises to have a huge effect on all of our lives in the very near future, particularly in the field of medicine. The activities here are not only for biology and science teachers, many of them can be done in other classes as well. Contests & Prizes are also available.

- [Malaria Teaching PDF: A teaching kit for high schools](#)
- [Activity packet for teachers and schools](#)
- [Surfing human and mouse genomes on the Internet](#)
- [Genome Browsing Tutorials](#)

**Ensembl version 48 Released**

# Proteome Analysis <http://www.ebi.ac.uk/proteome/index.html>

Browser address bar: <http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do>

EMBL-EBI Search: All Databases [v] Enter Text Here [Go] Reset [?] Advanced Search [Give us feedback]

Navigation: Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Home
- local help ⓘ
- Integr8 News
- Focal Point archive
- Latest Species
- Browse Species
- Inquisitor status
- BioMart
- Proteomes and Genomes FASTA
- About Integr8
- Publications
- Integr8 Web service

**Genome Reviews** [v]  
Curated versions of EMBL entries for complete genome sequences

**IPI** [v]  
A top-level guide to the main databases that describing higher eukaryotic proteomes

EBI > Databases > Integr8

## Integr8 : Access to complete genomes and proteomes

Search for species  [Go] Search for gene/protein  in all species [v] [Go]  
*e.g. "coli", "9606" e.g. "ras1", "P22981", "GO:0007257", "GO:mitosis"*

scope **Bacteria, Archaea, Eukaryota** Change scope

The **Integr8** web portal provides easy access to integrated information about deciphered genomes and their corresponding proteomes. Available data includes DNA sequences (from databases including the EMBL Nucleotide Sequence Database, Genome Reviews, and Ensembl); protein sequences (from databases including the UniProt Knowledgebase and IPI); statistical genome and proteome analysis (performed using InterPro, CluSTR, and GOA); and information about orthology, paralogy, and synteny.



Integr8 data can also be accessed via the [Integr8 FTP](#) site.

**New to Integr8?** The [user guide](#) will show you how to make the most of the data provided by Integr8. Alternatively, you may choose to [start browsing the data](#). We value your feedback! Please [send us your comments](#).

News	Current Status	Focal Point	History	Latest species	GAS top 10
<p>Full access to proteome analysis data from over 450 completely sequenced bacteria has now been made available through Integr8. Data available include gene and protein sets, Genome Reviews files, and InterPro-based proteome analysis. See the <a href="#">Focal Point article</a> from release 74 to find out more.</p> <p>Putative <b>Orthologous Clusters</b> have been made available via the <a href="#">Integr8 FTP site</a>. These clusters, derived from protein similarity data available in the <a href="#">CluSTR</a> database, identify putative orthologues over more than 500 cellular species, such that each gene is a member of a single group, and that each group contains no more than one (non-identical) gene from any species. For more information, see the <a href="#">PORC README</a> file, which explains how the PORC clusters are calculated and how to access the data.</p>					

### Funding



Integr8 has been funded by the European Commission: from March 2006 - February 2009 under FELICS, contract number 021902 (RII3) within the Research Infrastructure Action of the FP6 "Structuring the European Research Area" Programme, and from January 2002 - June 2005 as the TEMBLOR, contract number QLRI-CT-2001-00015 under the RTD programme "Quality of Life and Management of Living Resources"

# Proteome Analysis <http://www.ebi.ac.uk/proteome/index.html>

Browser address bar: <http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do>

EMBL-EBI Search: All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Navigation: Databases Tools EBI Groups Training Industry About Us Help Site Index

- Home
- local help
- Integr8 News
- Focal Point archive
- Latest Species
- Browse Species
- Inquisitor status
- BioMart
- Proteomes and Genomes FASTA
- About Integr8
- Publications
- Integr8 Web service

**Genome Reviews**  
Curated versions of EMBL entries for complete genome sequences

**IPI**  
A top-level guide to the main databases that describing higher eukaryotic proteomes

EBI > Databases > Integr8

## Integr8 : Access to complete genomes and proteomes

Search for species  Go! Search for gene/protein  in all species  Go!  
*e.g. "coli", "9606" e.g. "ras1", "P22981", "GO:0007257", "GO:mitosis"*

scope **Bacteria, Archaea, Eukaryota** Change scope

The **Integr8** web portal provides easy access to integrated information about deciphered genomes and their corresponding proteomes. Available data includes DNA sequences (from databases including the EMBL Nucleotide Sequence Database, Genome Reviews, and Ensembl); protein sequences (from databases including the UniProt Knowledgebase and IPI); statistical genome and proteome analysis (performed using InterPro, CluSTR, and GOA); and information about orthology, paralogy, and synteny.



Integr8 data can also be accessed via the [Integr8 FTP](#) site.

**New to Integr8?** The [user guide](#) will show you how to make the most of the data provided by Integr8. Alternatively, you may choose to [start browsing the data](#). We value your feedback! Please [send us your comments](#).

News	Current Status	Focal Point	History	Latest species	GAS top 10
<p>Full access to proteome analysis data from over 450 completely sequenced bacteria has now been made available through Integr8. Data available include gene and protein sets, Genome Reviews files, and InterPro-based proteome analysis. See the <a href="#">Focal Point article</a> from release 74 to find out more.</p> <p>Putative <b>Orthologous Clusters</b> have been made available via the <a href="#">Integr8 FTP site</a>. These clusters, derived from protein similarity data available in the <a href="#">CluSTR</a> database, identify putative orthologues over more than 500 cellular species, such that each gene is a member of a single group, and that each group contains no more than one (non-identical) gene from any species. For more information, see the <a href="#">PORC README</a> file, which explains how the PORC clusters are calculated and how to access the data.</p>					

### Funding



Integr8 has been funded by the European Commission: from March 2006 - February 2009 under FELICS, contract number 021902 (RII3) within the Research Infrastructure Action of the FP6 "Structuring the European Research Area" Programme, and from January 2002 - June 2005 as the TEMBLOR, contract number QLRI-CT-2001-00015 under the RTD programme "Quality of Life and Management of Living Resources"

- Karyn's Genomes (<http://www.ebi.ac.uk/2can/genomes/index.html>) - предоставляет общую информацию об организмах, чьи геномы полностью сиквенированы. Основная задача сервера – обеспечить лаконичное почему важен полный геном данного организма.
- WormBase (<http://www.wormbase.org/>) – хранилище информации по картированию, сиквенированию и фенотипам *C. elegans* и некоторых других нематод.
- FlyBase (<http://flybase.bio.indiana.edu/>) – база для *Drosophila melanogaster* .
- MGD (<http://www.informatics.jax.org/>) – the 'Mouse Genome Database'
- RGD (<http://rgd.mcw.edu/>) - the 'Rat Genome Database'
- SGD (<http://genome-www.stanford.edu/Saccharomyces/>) - the 'Saccharomyces Genome Database'

- MIPS (<http://www.mips.biochem.mpg.de/proj/yeast/> )  
база по геному дрожжей
- SPGP ([http://www.sanger.ac.uk/Projects/S\\_pombe/](http://www.sanger.ac.uk/Projects/S_pombe/) )  
- 'S. Pombe Genome Project' основанный при Sanger Institute. База по грибку *Schizosaccharomyces pombe*.
- AceDB (<http://www.acedb.org/> )  
- база по генетической информации *Caenorhabditis elegans*. Система управления этой базой очень популярна и легла в основу многих других баз данных. Одноименное название базы и системы управления базой AceDB привело к некоторой путанице среди баз данных по *C.elegans*.
- HIV-SD (<http://hiv-web.lanl.gov/content/hiv-db/mainpage.html> )  
- 'HIV Sequence Database' собирает, курирует и аннотирует ДНК последовательности HIV и SIV



# Базы данных по *E.coli*

- The '*E. coli* Genetic Stock Center' (CGSC)  
<http://cgsc.biology.yale.edu/top.html>  
поддерживает базу по генетической информации, включая генотипы, информацию по штаммам, названиям генов, продуктам генов и специфическим мутациям
- The '*E. coli* Database collection' (ECDC)  
<http://www.uni-giessen.de/~gx1052/ECDC/ecdc.htm>  
курирует основанную на последовательностях генов базу по *E. coli*.
- EcoCyc (<http://ecocyc.org/>)  
'Encyclopedia of *E. coli* Genes and Metabolism' - база по генам *E. coli* и метаболическим путям.

# The 'E. coli Genetic Stock Center' (CGSC)

<http://cgsc.biology.yale.edu/top.html>

**E. coli Genetic Resources at Yale**  
**CGSC, The Coli Genetic Stock Center**

**About the CGSC**  
CGSC Home  
Scope of the Collection  
The Database  
Acknowledgment of Support

**Searching the CGSC Database**  
Which Query should I use?  
Strain Query  
Mutation Query  
Gene/Size Query  
Gene Product Query  
Reference Query  
Query Help

**How to Request Strains or Information**  
Contact Information  
Charges  
Strains Not Found?

**Other CGSC Information**  
FAQ on Procedures  
Current Working Map  
1998 Published Map: MMR  
1998 Gene List: MMR  
Map Diagrams (PDFs)

**Other Links**  
E. coli Links  
Other Stock Collections  
Other Bio Links  
Yale Web

**CGSC DB-WebServer:**  
**Access To and Information About All Query Forms.**

- [Which Query form should I use?](#)
- [Strain Query](#)
- [Mutation Query](#)
- [Gene/Plasmid/Phage/Chrom.Sites Query](#)
- [Gene Product Query](#)
- [Reference Query](#)
- [Query Help](#)

**Due to changes in funding, the CGSC has changed its fees effective March 15th 2008. This will not affect labs with existing subscriptions.**

**[NEW! Addition of the Keio Knockout Collection](#)**

**Note:** The CGSC Collection contains only non-pathogenic laboratory strains, primarily genetic derivatives of *Escherichia coli* K-12, the laboratory strain widely used in genetic and molecular studies. For details, see [About CGSC](#) on the left.

The CGSC Database of *E. coli* genetic information includes **genotypes** and reference information for the **strains** in the CGSC collection, the names, synonyms, properties, and map position for **genes**, **gene product** information, and information on specific **mutations** and **references to primary literature**. The public version of the database includes this information and can be queried directly via this CGSC DB WebServer. For help, use the help links located above and on each query form, or contact us, as indicated below.

**REQUESTING STRAINS** or additional information, as well as questions about the contents, use or alternative access interfaces to the database: [email the CGSC](#) or see other [Contact Information](#) at the left.

Funded by a grant from the National Science Foundation

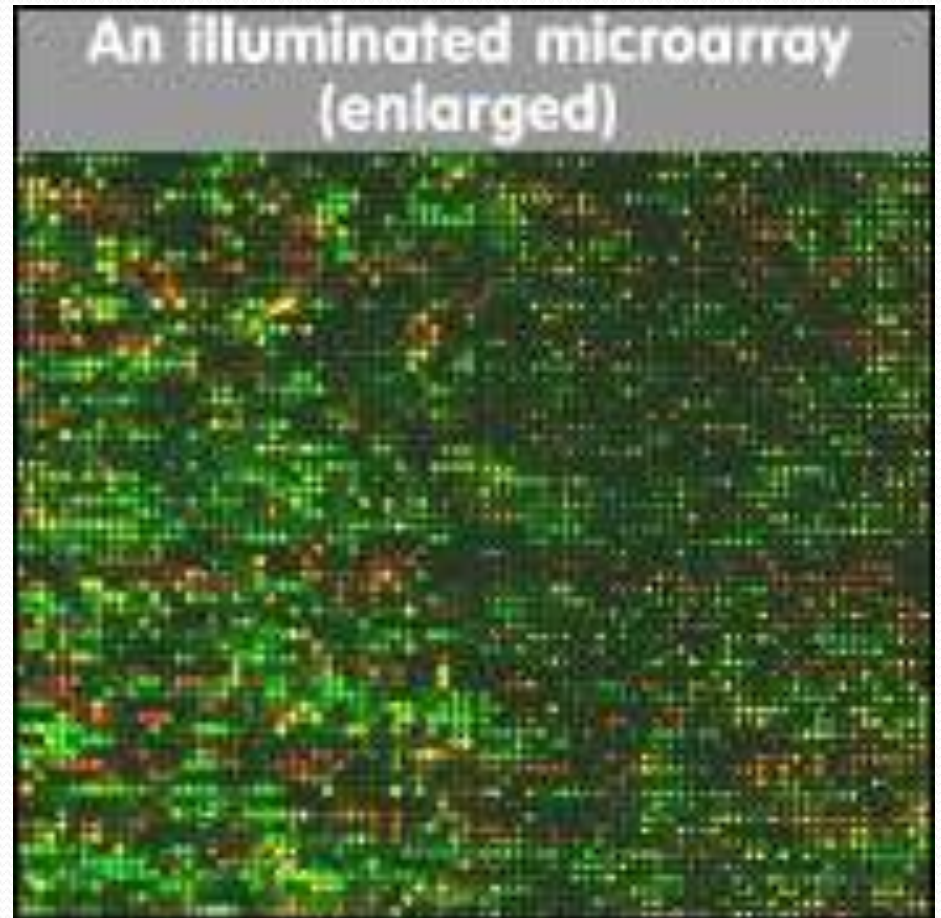
# Базы геномов растений

- MaizeDB (<http://www.agron.missouri.edu/>)
- The 'Plant Genome Information Resource' (PGDIC)  
<http://www.nal.usda.gov/pgdic/>  
доступ к базам геномов многих растений, включая хлопок, люцерна, пшеница, ячмень, рожь, рис, просо, сорго, пасленовые и деревья
- MENDEL (<http://www.mendel.ac.uk/>)  
обширная база данных по генам растений

# Microarray Databases

# Что такое microarray?

- Microarray – это другое название «биочипа»
- «Биочип» - маленькая пластинка из стекла или кремния с нанесенным матрицей биомолекул, используемых в качестве биосенсора
- **array** – матрица или таблица



# Что такое microarray?

- В основе технологии классических биочипов лежит процесс связывания комплементарных одноцепочечных последовательностей нуклеиновых кислот.
- Технология биочипов позволяет использовать информацию, полученную при выполнении геномных проектов, для получения ответов *какие гены экспрессируются в данных типах клеток в заданное время при определенных условиях.*
  - Биочипы позволяют сравнивать экспрессию генов в нормальных клетках и больных (например раковых).
- Данная технология имеет несколько названий:
  - DNA microarrays, DNA arrays, DNA chips, gene chips и другие.
  - Иногда эти технологии разделяют, но по сути это синонимы.

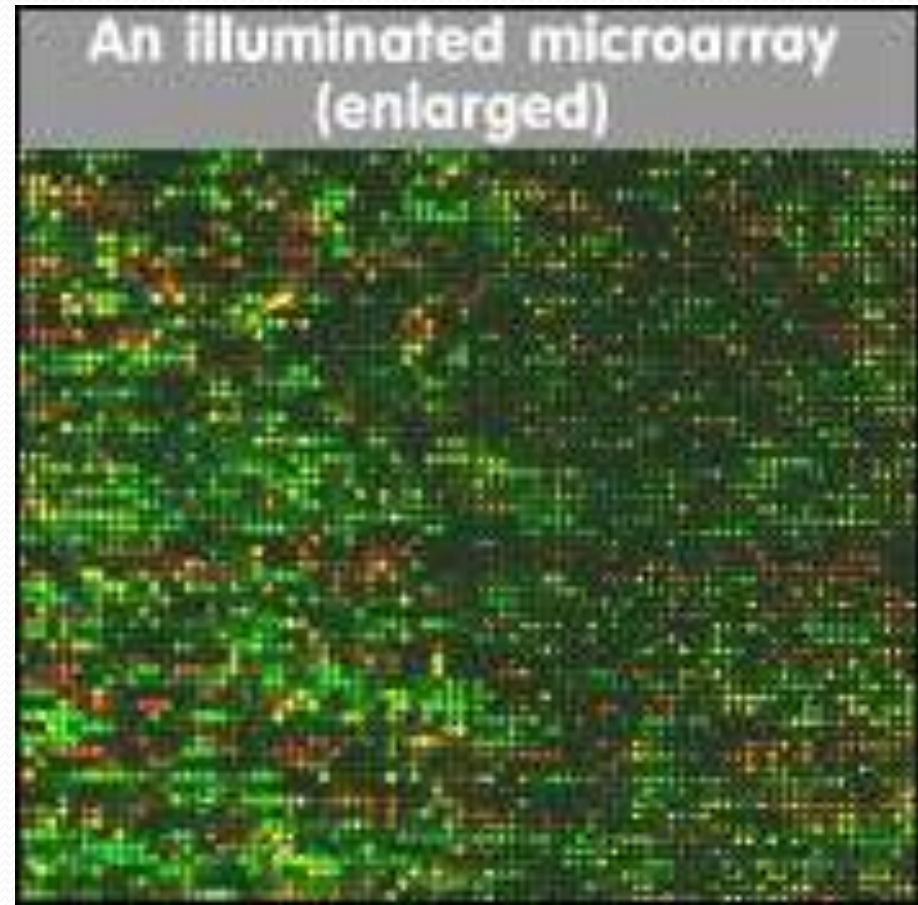
Биочип – обычно стеклянная пластина на которую наносятся молекулы ДНК .  
Количество пятен – от *тысяч до сотен тысяч*.

Каждое пятно содержит идентичные молекулы ДНК длиной от *20 до сотен* нуклеотидов.

Для исследования экспрессии генов каждая молекула ДНК должна в идеале идентифицировать один ген или экзон генома, что не всегда выполнимо из-за семейств сходных генов.

В 1997 г. были доступны биочипы содержащие примерно 6000 генов из генома дрожжей.

Пятна на биочип наносятся роботом, либо синтезируются фотолитографией (как и компьютерные чипы), либо «чернильной» печатью.

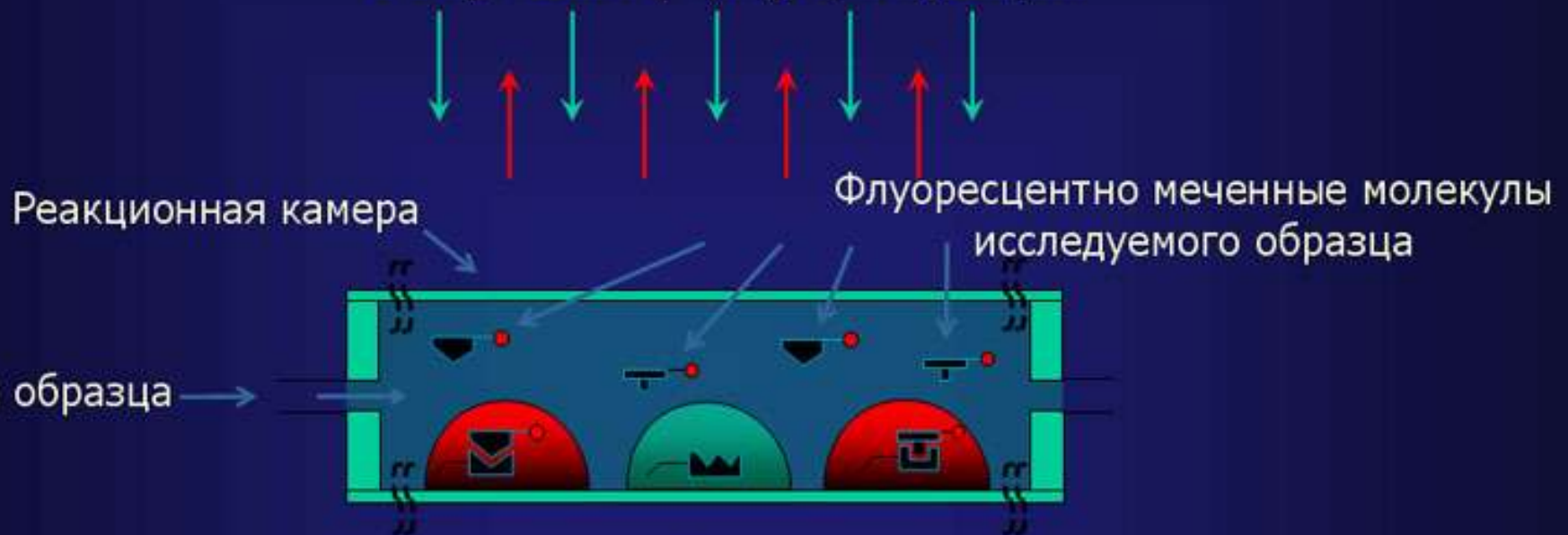


~2.5 см  
пятно ~0.1 мм

# Принцип работы биочипа

Институт молекулярной биологии  
РАН им. В.А. Энгельгардта

## Облучение / флуоресценция



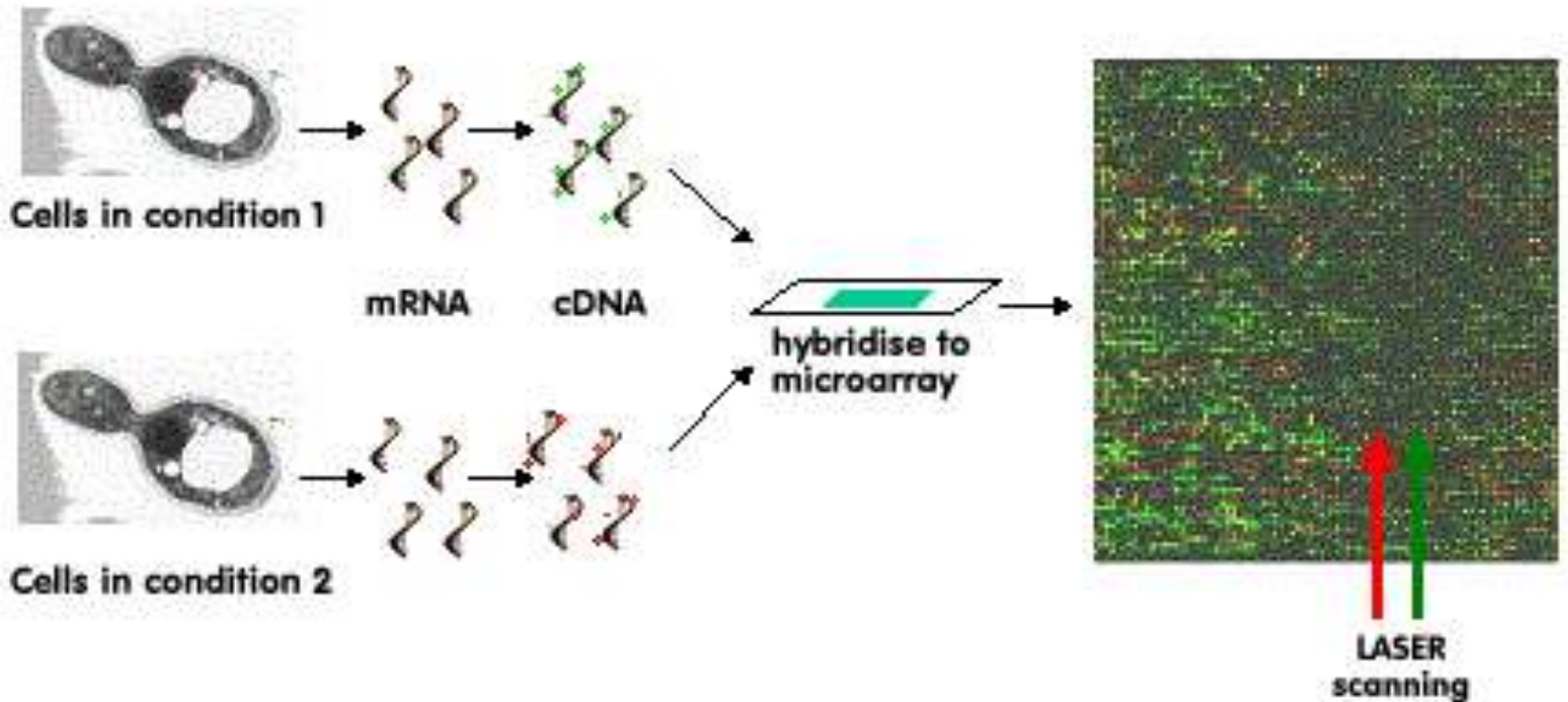
Элементы биочипа с иммобилизованными молекулами

Флуоресцентная картина ячеек биочипа





Micorarray applications allows to compare gene expression levels in two different samples



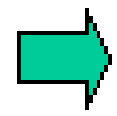
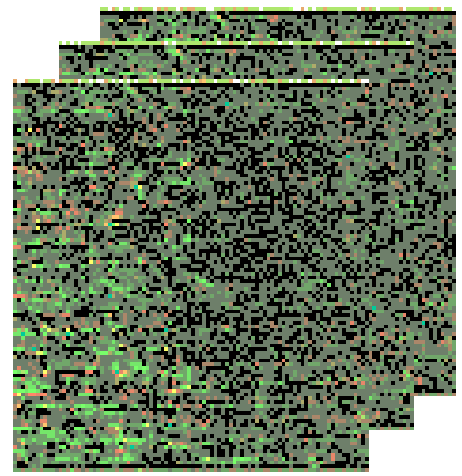
Измерение уровня экспрессии генов в двух разных образцах

1. Экспрессия генов
2. Синтез кДНК с флуоресцентными метками
3. Нанесение образцов на биочип и гибридизация
4. Считывание результатов лазерным сканером

# Image analysis of the raw microarray data

**RAW DATA**

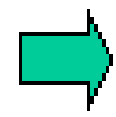
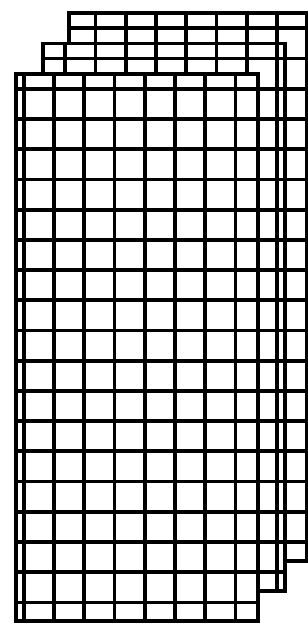
Array scans



Spots

**QUANTITATION  
MATRICES**

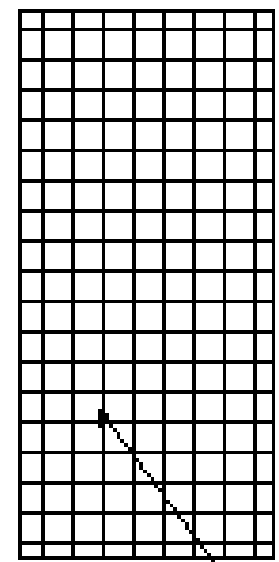
Quantifications



Genes

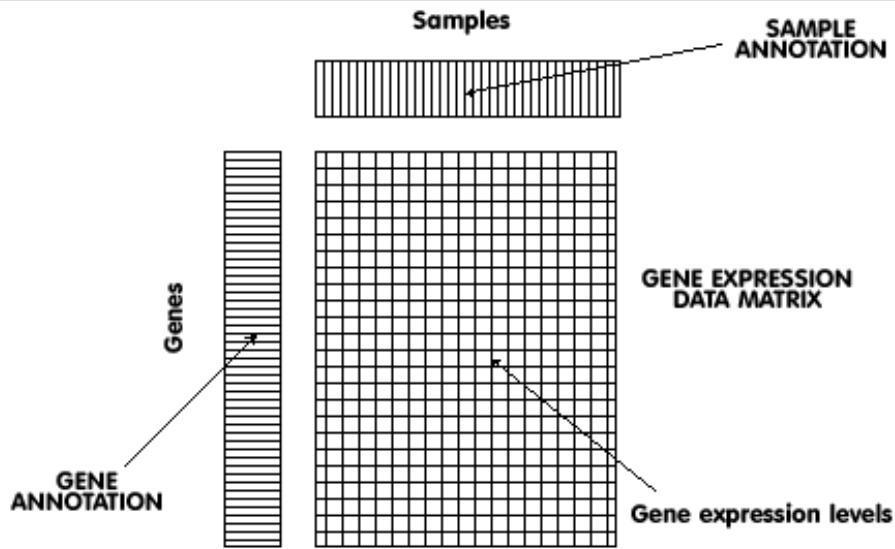
**GENE EXPRESSION  
DATA MATRIX**

Conditions



Gene Expression levels

A gene expression database can be regarded as consisting of three parts – the gene expression data matrix, gene annotation and sample annotation



Биочип можно рассматривать как матрицу из трех компонентов: матрицу экспрессии генов, матрицу отнесения генов, матрицу идентификации образцов.

Данный метод анализа экспрессии генов уже сейчас позволил накопить огромный массив информации.

GEO <http://www.ncbi.nlm.nih.gov/geo/>

ArrayExpress <http://www.ebi.ac.uk/microarray-as/ae/>

Во многих смыслах такую базу сложнее организовывать по сравнению с геномными базами.

Кроме того существует проблема стандартизации.

# GEO

## Total holdings

	Public	Unreleased	Total
Platforms	4390	361	4751
Samples	199436	45485	244921
Series	7804	1618	9422

## Browse public holdings

- ◆ All contacts
- ◆ All platforms
  - ◊ in situ oligonucleotide (1253)
  - ◊ spotted oligonucleotide (1096)
  - ◊ spotted DNA/cDNA (1849)
  - ◊ antibody (5)
  - ◊ tissue (0)
  - ◊ MS (10)
  - ◊ SARST (1)
  - ◊ MPSS (12)
  - ◊ RT-PCR (7)
  - ◊ oligonucleotide beads (48)
  - ◊ mixed spotted oligonucleotide/cDNA (6)
  - ◊ spotted protein (3)
  - ◊ SAGE (54)
- ◆ All samples
  - ◊ RNA (165767)
  - ◊ genomic (30003)
  - ◊ protein (651)
  - ◊ SAGE (993)
  - ◊ mixed (901)
- ◆ All series

2008

## Total holdings

	Public	Unreleased	Total
Platforms	6439	665	7104
Samples	350114	75843	425957
Series	13687	2815	16502

## Browse public holdings

- All contacts
- All platforms
  - ◊ in situ oligonucleotide (2016)
  - ◊ spotted oligonucleotide (1616)
  - ◊ spotted DNA/cDNA (2242)
  - ◊ antibody (8)
  - ◊ tissue (0)
  - ◊ MS (15)
  - ◊ SARST (2)
  - ◊ MPSS (23)
  - ◊ RT-PCR (20)
  - ◊ oligonucleotide beads (108)
  - ◊ mixed spotted oligonucleotide/cDNA (11)
  - ◊ spotted protein (14)
  - ◊ SAGE (67)
- All samples
  - ◊ RNA (287861)
  - ◊ genomic (54736)
  - ◊ protein (1464)
  - ◊ SAGE (1644)
  - ◊ mixed (1599)
  - ◊ SRA (223)
- All series

2009

# ArrayExpress



ArrayExpress is a public archive for **functional genomics data** compliant with [MIAME](#)- and [MINSEQE](#) requirements in accordance with compliant data in accordance with [MGED](#) recommendations. The Gene Expression Atlas uses curated, re-annotated subset of data from the Archive to provide information about **gene expression** under various biological conditions.

## Experiments Archive

9045 experiments, 256636 assays



Experiment, citation, sample and factor annotations

[Browse experiments](#)  
[Advanced query interface](#)

[Submitter/reviewer login](#)

 [ArrayExpress Query Help](#)

## Gene Expression Atlas

1134 experiments, 31275 assays, 5877 conditions

Genes

up/down in

Conditions

Any species

[Gene Expression Atlas Home](#)

## News



- **25 Aug 2009 - MGED 12 - with student bursaries**  
[Student bursaries](#) are available for US and EC students attending the MGED 12 conference in Phoenix, Arizona, USA.
- **16 Jun 2009 - Gene Expression Atlas - Release 1.1.0**  
New features include an ontology driven interface using [EFO](#), newly added datasets, expression profile similarity searching and top 10 variable genes per experiment...[try them now](#).

## Links

- [ArrayExpress User Survey](#)
- [Help](#) | [Training](#) | [FAQ](#) | [Citing](#)
- [Submit Data](#) (array based and re-sequencing)
- [Programmatic Access](#) | [FTP Access](#)
- [Software Downloads](#) and [Statistics](#)
- [EFO](#) | [Bioconductor Package](#) | [Quality Metrics](#)
- [ArrayExpress Scientific Advisory Board](#)
- [Microarray Informatics Group](#)

# ArrayExpress

EMBL-EBI [Home](#) [About Us](#) [Contact Us](#) [Help](#) [Feedback](#) [Site Index](#)

Search:

Match whole words   
  Loaded in Gene Expression Atlas

Filter on [reset]   
 Display options [reset]

  
 25 experiments per page

  
 Detailed view

Submitter/reviewer login    [ArrayExpress Browser Help](#)

ID	Title	Assays	Species	Date	Processed	Raw	Atlas
E-GEOD-13524	Transcription profiling of rat nucleus accumbens of alcohol-preferring animals following chronic ethanol consumption	29	Rattus norvegicus	2009-09-24			--
E-GEOD-6614	Transcription profiling of mouse brains following nicotine-induced seizures	28	Mus musculus	2009-09-23			--
E-BUGS-94	Comparative genomic hybridization of <i>S. aureus</i> to examine Genetic Diversity in CC398 Methicillin-Resistant Staph	4	Mycobacterium tuberculosis	2009-09-22			--
E-GEOD-4773	Transcription profiling of human SK-N-BE cell line model of Parkinson's disease	21	Homo sapiens	2009-09-22			--
E-GEOD-6285	Transcription profiling of brains of mice fed four different diets for a 2-week duration	24	Mus musculus	2009-09-22			--
E-GEOD-8150	Transcription profiling of mouse brain to identify age-related transcriptional changes and the effect of dietary suppl	20	Mus musculus	2009-09-22			--
E-GEOD-10748	Transcription profiling of rat brain treated with D-serine	24	Rattus norvegicus	2009-09-22			--
E-GEOD-13793	Transcription profiling of rat cerebellum and hippocampus following exposure to neurotoxicant Arsenic 1254	24	Rattus norvegicus	2009-09-22			--
E-TABM-756	Transcription profiling of mouse liver after partial hepatectomy in time course	366	Mus musculus	2009-09-22			--
E-TABM-555	Transcription profiling of rat to investigate technical and biological variabilities on the Agilent platform	96	Rattus norvegicus	2009-09-21			--
E-TABM-783	Transcription profiling of human non anaplastic T-cell lymphoma	35		2009-09-21			--
E-BUGS-79	Transcription profiling of Staphylococcus aureus to subinhibitory concentrations of a sequence-selective, DNA mino	30	Staphylococcus aureus	2009-09-18			--
E-MEXP-2358	Transcript profiling of Arabidopsis thaliana transgenic seedlings constitutively overexpressing UGT74E2 (35S::UGT)	6	Arabidopsis thaliana	2009-09-18			--
E-MTAB-144	Methylation profiling of human pediatric acute lymphoblastic leukemia	20	Homo sapiens	2009-09-18			--
E-GEOD-11974	Re-sequencing of rice seed	1	Oryza sativa	2009-09-16			--
E-GEOD-12297	Re-sequencing of human post mortem cerebellum reveals altered synaptic vesicular transport in	20	Homo sapiens	2009-09-16			--
E-GEOD-12346	ChIP-Seq of mouse Stat5a and Stat5b with an IgG control reveals priming for Th2 differentiation requires IL-2-mei	9	Mus musculus	2009-09-16			--
E-GEOD-13750	Re-sequencing of yeast ribosomally bound sequences	8	Saccharomyces cerevisiae	2009-09-16			--
E-GEOD-14600	ChIP-Seq of Arabidopsis to identify target genes of the MADS transcription factor SEPALLATA3	5	Arabidopsis thaliana	2009-09-16			--
E-GEOD-14605	Re-sequencing of mouse single cell - wild-type oocytes, two single Dicer knockout oocyte, and one single Ago2 kno	6	Mus musculus	2009-09-16			--
E-GEOD-15922	Methylation profiling of Arabidopsis endosperm	4	Arabidopsis thaliana	2009-09-16			--
E-GEOD-16916	Re-sequencing of maize tip and base of a developing	2	Zea mays	2009-09-16			--
E-GEOD-17454	ChIP-Seq of C. elegans to identify transcription factor CEH-14 binding sites - ModEncode	3	Caenorhabditis elegans	2009-09-16			--
E-MEXP-2343	Transcription profiling of human colorectal carcinoma cells retrotransfected by Cox2 to analyse gene alteratio	4	Homo sapiens	2009-09-16			--
E-MEXP-2362	Transcription profiling of a yeast strain containing 2 extra copies of Abf2p compared to its wild type counterpart	2	Saccharomyces cerevisiae	2009-09-16			--

9045 experiments, 256636 assays. Displaying experiments 1 to 25. Pages: 1 2 3 4 5 6 7 8 9 10 ... 360

Terms of Use    EBI Funding    Contact EBI    © European Bioinformatics Institute 2009. EBI is an Organisation of the European Molecular Biology Laboratory.

# Анализ данных и профили экспрессии

- Количество информации даже с одного биочипа очень велико, обработка которой требует очень мощное программное обеспечение
- Кластеризация и предсказание классов – это типичные методы анализа таких данных.
- Expression Profiler <http://ep.ebi.ac.uk/> является одной из популярных программ.

## Books

### Books Menu

- Books Home
- Author Information
- Exam Copy Policy
- Customer Services
- Information About ISBNs
- Rights & Permissions
- Recommend to Library

### This Book

- **Book Information**
- Reviews
- Table of Contents
- About the author

### See Also

- Related Books
- Related Journals
- More by this author
- Blackwell Molecular and Cell Biology

### Related Websites

- Life & Physical Sciences

### Book Information

## Microarray Gene Expression Data Analysis

### A Beginner's Guide

**By:** Helen Causton (Imperial College), J Quackenbush and Alvis Brazma (The European Bioinformatics Institute)

### Reviews

"Quite a few recently published books discuss analysis of microarray gene expression data for beginners. *Microarray Gene Expression Data Analysis ... is arguably the best of its kind in this regard.*" *Terry Speed, The Walter & Eliza Hall Institute of Medical Research, Nature Cell Biology, December 2003*

[▶▶ More reviews](#)

### Description

This guide covers aspects of designing microarray experiments and analysing the data generated, including information on some of the tools that are available from non-commercial sources. Concepts and principles underpinning gene expression analysis are emphasised and wherever possible, the mathematics has been simplified. The guide is intended for use by graduates and researchers in bioinformatics and the life sciences and is also suitable for statisticians who are interested in the approaches currently used to study gene expression.

- Microarrays are an automated way of carrying out thousands of experiments at once, and allows scientists to obtain huge amounts of information very quickly.
- Short, concise text on this difficult topic area.
- Clear illustrations throughout.
- Written by well-known teachers in the subject.
- Provides insight into how to analyse the data produced from microarrays.

### Table of Contents

[Top ▲](#)

Part I: Introduction.  
Part II: Aspects Of Experimental Design.  
Part III: Data Analysis.  
Part IV: Glossary.

[▶▶ Detailed contents](#)

### About the Author

[Top ▲](#)

**Helen Causton** is an experimental biologist who carried out some of the early studies on genome-wide transcriptional regulation in yeast using microarrays. She is Head of the Clinical Sciences Microarray Centre at Imperial College, University of London.

**John Quackenbush** is a principal investigator at The Institute for Genomic Research (TIGR). His research interests include development of software for microarray data analysis, gene indices and comparative genomics.

**Alvis Brazma** is a computer scientist who has been involved in microarray data analysis since 1998. He heads microarray informatics at the European Bioinformatics Institute and is in charge of establishing a public repository for microarray data.



[E-mail Alerting](#)

**Paperback**

Status: Available

ISBN: 9781405106825

ISBN10: 1405106824

[▶▶ Buy Now](#) [▶▶ Exam Copy](#)

**Publication Dates**

USA: Apr 2003  
Rest of World: Mar 2003  
Australia: May 2003

**Format**

244 x 172 mm , 6.75 x 9.75 in

**Details**

176 pages, 46 illustrations.

Blackwell Publishing is now part of John Wiley & Sons

All customers will be directed to this book's page on Wiley.com for the latest price and purchase options.

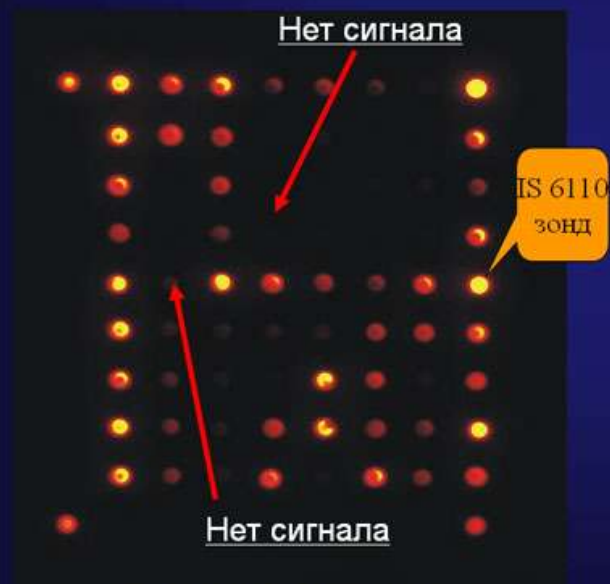




Тест-система «ТБ-Биочип» позволяет обнаруживать не менее 95% рифампицин- и свыше 80% изониазид-устойчивых штаммов возбудителя туберкулёза менее чем за сутки

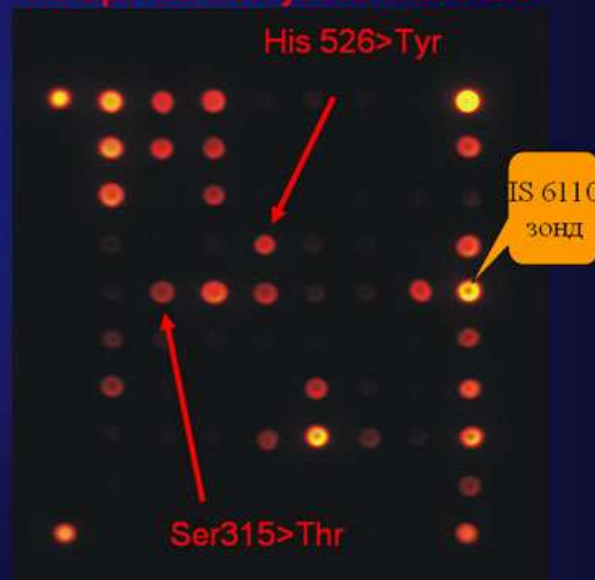
Тест-системы 'ТБ-Биочип' для обнаружения возбудителя туберкулеза и определения его лекарственной устойчивости

Чувствительный штамм



Стандартная терапия

Штамм с множественной лекарственной устойчивостью



Рекомендуется лечение резервными препаратами

С помощью тест-системы «ЛейкоГен-Биочип» можно подобрать наиболее оптимальную терапию при раке крови (лейкемии)

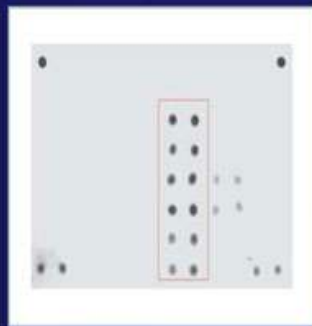
## Тест-система «ЛейкоГен-БИОЧИП»

регистрационное удостоверение № ФС 012a2005/2668-06

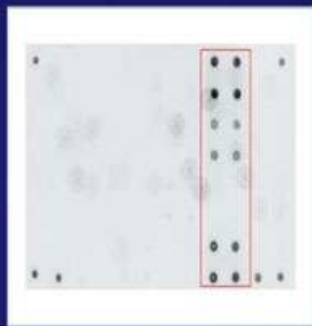
Идентификация хромосомных перестроек позволяет установить точный диагноз, сделать прогноз заболевания и подобрать адекватную терапию



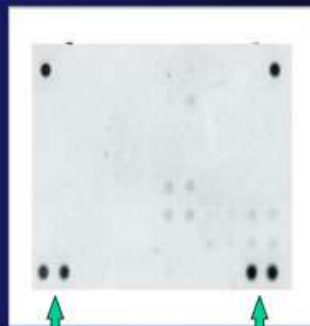
Хронический миелолейкоз.  
Неблагоприятный прогноз.  
Специфичная терапия ST1571.  
Трансплантация костного мозга.



Острый промиелоцитарный лейкоз.  
Благоприятный прогноз.  
Специфичная терапия препаратами ретиноевой кислоты.



Острый лимфобластный лейкоз.  
Крайне неблагоприятный прогноз.  
Высокодозная химиотерапия



Контрольный ген ABL

# Белковые базы данных

- Существуют универсальные базы данных и специализированные, которые покрывают узкие семейства или группы белков, или белки из определенного организма.
- **Примеры:**
- База по аминокислотным последовательностям (первичной структуре) белков UniProtKB/Swiss-Prot
- Специализированная база по первичной структуре белков GOA
- Специализированная база ENZYME
- Вторичная база InterPro
- Пространственные структуры белков (и других макромолекул) PDB

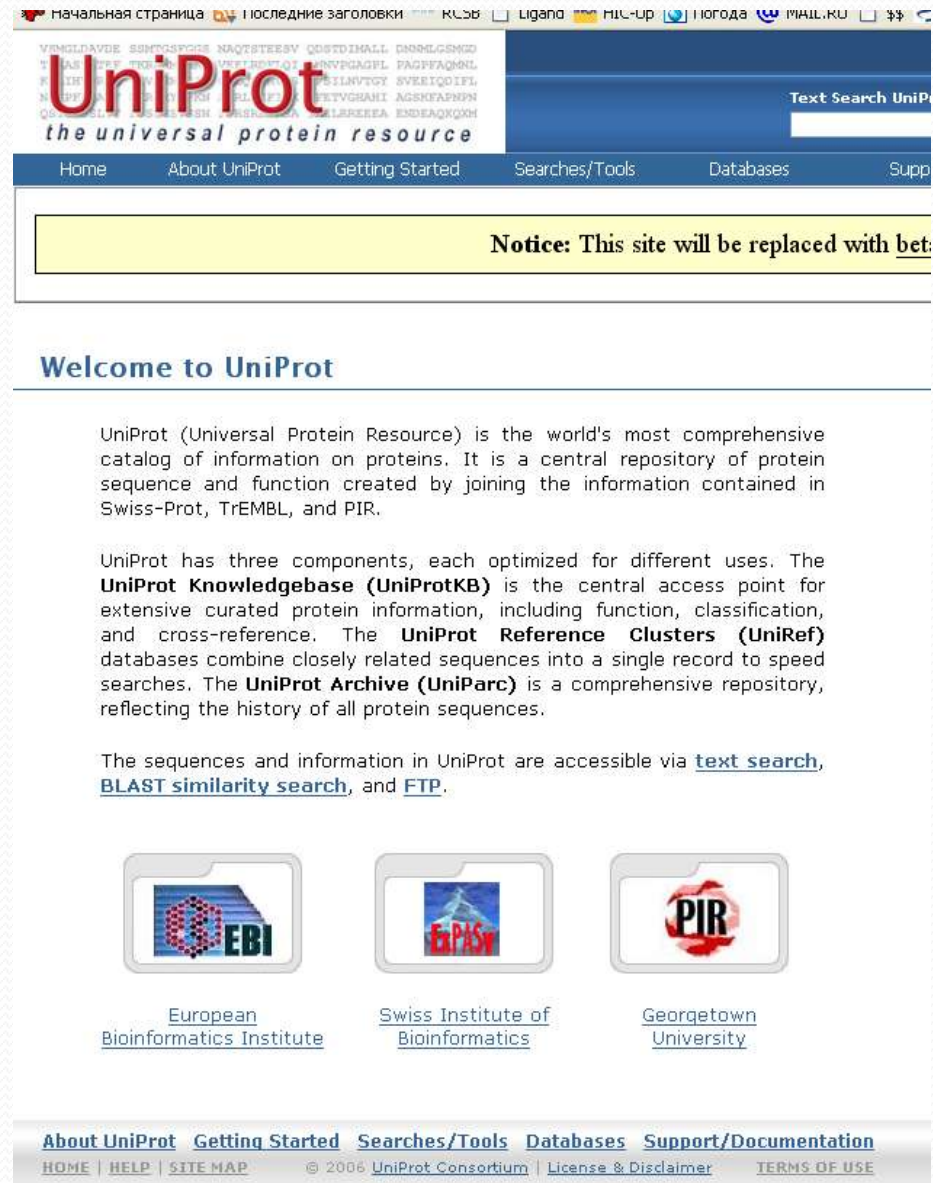
# База первичных структур белков - UniProtKB



- UniProt (Universal Protein Resource; <http://www.ebi.ac.uk/uniprot/index.html>) является одним из наиболее всеобъемлющих каталогов информации о белках и их функциях, а также центральным хранилищем первичных структур.
- Данные о белках из более 6000 видов организмов. Половина информации – о белках из 20 наиболее изучаемых организмов.
- UniProt объединяет базы UniProtKB/Swiss-Prot, UniProtKB/TrEMBL и PIR.

# База первичных структур белков - UniProtKB

- Состоит из трех компонент, оптимизированных под разные задачи:
  - The UniProt Knowledgebase (UniProt) – центр доступа, включая функции, классификацию и перекрестные ссылки
  - The UniProt Reference Clusters (UniRef) – комбинирует близкие структуры в одной записи для ускорения обработки
  - The UniProt Archive (UniParc) – всеобъемлющая база, отражающая историю всех первичных структур белков



мачальная страница | последние заголовки | UniProt | Ligand | NCBI-Up | погода | MAIL.RU | \$

## UniProt

the universal protein resource

Text Search UniProt

Home About UniProt Getting Started Searches/Tools Databases Support


**Notice: This site will be replaced with bet:**

### Welcome to UniProt


UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt has three components, each optimized for different uses. The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information, including function, classification, and cross-reference. The **UniProt Reference Clusters (UniRef)** databases combine closely related sequences into a single record to speed searches. The **UniProt Archive (UniParc)** is a comprehensive repository, reflecting the history of all protein sequences.


The sequences and information in UniProt are accessible via [text search](#), [BLAST similarity search](#), and [FTP](#).



[European Bioinformatics Institute](#)



[Swiss Institute of Bioinformatics](#)



[Georgetown University](#)

[About UniProt](#) [Getting Started](#) [Searches/Tools](#) [Databases](#) [Support/Documentation](#)

HOME | HELP | SITE MAP | © 2006 UniProt Consortium | [License & Disclaimer](#) | [TERMS OF USE](#)

# База первичных структур белков - UniProtKB/Swiss-Prot

Search  for

 **Swiss-Prot**  
Protein knowledgebase  
**TrEMBL**  
Computer-annotated supplement to Swiss-Prot



<http://au.expasy.org/sprot/>

The UniProt Knowledgebase consists of

- **UniProtKB/Swiss-Prot**, a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases ([More details](#) / [References](#) / [Linking to Swiss-Prot](#) / [User manual](#) / [Recent changes](#) / [Disclaimer](#))
- **UniProtKB/TrEMBL**, a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups [at SIB](#) and [at EBI](#).

UniProt Knowledgebase Release 12.8 consists of:  
**UniProtKB/Swiss-Prot Release 54.8 of 05-Feb-2008: 349480 entries** ([More statistics](#))

**UniProtKB/TrEMBL Release 37.8 of 05-Feb-2008: 5329119 entries** ([More statistics](#))

> **Swiss-Prot headlines**  
Over 20,000 fungal proteins manually annotated in UniProtKB/Swiss-Prot (Read more...)

### Access to the UniProt Knowledgebase

- **SRS** - Access to UniProtKB/Swiss-Prot, UniProtKB/TrEMBL and other databases using the Sequence Retrieval System
- Full text search in the UniProt Knowledgebase
- Advanced search in the UniProt Knowledgebase by description, gene name and organism (can be used to create html links to UniProt Knowledgebase queries)
- Taxonomy browser (NEW!)
- **BLAST** similarity search

- by description or identification (any word in the DE, OS, OG, GN and ID lines)
- by citation (RL line; UniProtKB/Swiss-Prot only)

- Retrieve a list of UniProtKB entries
- Randomly retrieve a UniProtKB entry
- UniProtKB Sequence/Annotation Version Database **\*\*\***
- Swiss-Prot ID tracker

### Documents and services

-  **Swiss-Prot documents** - user manual, release notes, indices and lots of other **important** documents and lists
- **Swiss-Shop** - a service that allows you to automatically obtain (by email) new UniProtKB/Swiss-Prot sequence entries relevant to your field(s) of interest
- Updates and submissions:
  - Report form for updates or corrections of an existing Swiss-Prot entry or of a family of entries
  - Sequence data submission to Swiss-Prot
- **FTP** - How to obtain a local copy of Swiss-Prot and TrEMBL

# База первичных структур белков - UniProtKB/TrEMBL

Дополнение базы Swiss-Prot, называемое TrEMBL ([Translation of EMBL Nucleotide Sequence Database](#)) состоит из аминокислотных последовательностей, полученных трансляцией всех кодирующих последовательностей ДНК из базы EMBL Nucleotide Sequence Database за исключением тех, которые уже включены в UniProtKB/Swiss-Prot.



# База первичных структур белков - UniProtKB/TrEMBL

*TrEMBL* состоит из двух основных секций:

- SP-TrEMBL содержит записи, которые в конечном счете будут включены в *Swiss-Prot*.
- REM-TrEMBL (REMAining TrEMBL) содержит записи, которые не будут включены в *Swiss-Prot*.

**REM-TrEMBL** содержит последовательности, которые получены синтетически, неполные (*truncated*), запатентованы, являются псевдогенами, иммуноглобулинами или Т-клеточными рецепторами. Они не интересны для аннотирования и не включаются в базу *Swiss-Prot*.

# База первичных структур белков - UniProtKB/PIR



- PIR '*Protein Information Resource*' <http://pir.georgetown.edu/>  
Основан National Biomedical Research Foundation в 1984; публиковался как '*Atlas of Protein Sequence and Structure*' (Dayhoff et al., 1965; Dayhoff, 1979). С 1988 г. является международной базой.
- Разделена на 4 секции: PIR<sub>1</sub>, PIR<sub>2</sub>, PIR<sub>3</sub> и PIR<sub>4</sub>.
  - PIR<sub>1</sub> полностью классифицирована по суперсемействам и аннотирована
  - PIR<sub>2</sub> является переходным разделом к PIR<sub>1</sub> от PIR<sub>3</sub>
  - PIR<sub>3</sub> служит временным разделом для новых поступлений
  - PIR<sub>4</sub> включает неклассифицированные последовательности



INTEGRATED PROTEIN INFORMATICS RESOURCE  
FOR GENOMIC AND PROTEOMIC RESEARCH



The Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information.

UniProtKB | UniRef | UniParc

12.8

Current release:

PIRSF

Protein Family Classification System



- Classification reflecting evolutionary relationships of full-length proteins
- Functional site and protein name rules
- \*Sample family report\*

iProClass

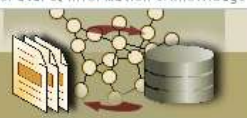
Integrated Protein Knowledgebase



- Value-added reports for UniProtKB and unique UniParc proteins
- Functional analysis and protein ID mapping
- \*Sample protein report\*

iProLINK

Literature, Information & Knowledge



- Source for text mining and ontology development
- RLIMS-P text mining tool, BioThesaurus, and PProtein Ontology
- Bibliography mapping

OTHER RESOURCE

- Proteomics: NIAID Biodefense Proteomics Admin. Center
- PIR Grid-Enablement: Data node on NCI's caBIG

PEPTIDE SEARCH

DATABASE: UniProtKB

Use single letter amino acid code

TEXT SEARCH

DATABASE: iProClass

<http://pir.georgetown.edu/>

# Специализированные базы

- Существует огромное количество специализированных белковых баз данных,
- Коллекция ссылок на базы:
- <http://www.expasy.ch/alinks.html>
- <http://www.ebi.ac.uk/integr8/> - портал ссылок на информацию по расшифрованным геномам и соответствующие протеомы
- Такие базы ценны тем, что, как правило, дополняют базы по первичным структурам дополнительной аналитической информацией

- GOA – GO «Gene Ontology Annotation»  
<http://www.ebi.ac.uk/GOA/index.html>
- MEROPS <http://merops.sanger.ac.uk/> - каталог и структурная классификация пептидаз в семейства на основе «пептидазных единиц»
- GPCRDb <http://www.gpcr.org/> - база по G-protein coupled receptors (GPCRs)  
большом семействе белков, являющихся важными компонентами различных сигнальных систем животных
- YPD «the yeast protein database»  
<http://www.incyte.com/sequence/proteome/databases/YPD.shtml>  
– база для белков из *S. cerevisiae*. Содержит более 50,000 строк аннотаций относительно 6,000 белков на основе обзора примерно 8,500 публикаций.

- ENZYME <http://www.expasy.ch/enzyme/> - аннотированное расширение UniProtKB/Swiss-Prot по ферментам
- BRENDA <http://www.brenda.uni-koeln.de/> - база по свойствам ферментов
- LIGAND <http://www.genome.ad.jp/ligand/> - Ligand Chemical Database for Enzyme Reactions
- EMP <http://emp.mcs.anl.gov/> - 'Enzymes and Metabolic Pathways'
- <http://www.expasy.ch/ch2d/> - база данных по двумерным гелям
- <http://prowl.rockefeller.edu/> - база данных по масс-спектропии белков и протеолитических фрагментов.



# ENZYME

## Enzyme nomenclature database

<http://www.expasy.ch/enzyme>

**ENZYME** is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided [[More details](#) / [References](#) / [Linking to ENZYME](#) / [Disclaimer](#)].

**Release of 05-Feb-2008 (4063 active entries)**

### Access to ENZYME

- by EC number:  .  .  .
- by enzyme class
- by description (official name) or alternative name(s):
- by chemical compound
- by cofactor
- by search in comments lines
- [SRS](#) - Sequence Retrieval System

### Documents

- [ENZYME user manual](#)
- [How to obtain ENZYME](#)

### Services

- [Report forms for a new ENZYME entry or for an error/update in an existing entry](#)
- [Downloading ENZYME by FTP](#)

### Related tools and databases

- [Biochemical Pathways](#) - Interactive access to Roche Applied Science "Biochemical Pathways"
- [BRENDA](#) - Comprehensive Enzyme Information system
- [EMP](#) - Enzymes and Metabolic Pathways database
- [KEGG](#) - Kyoto Encyclopedia of Genes and Genomes
- [MetaCyc](#) - Metabolic Encyclopedia of enzymes and metabolic pathways
- [IUBMB Enzyme Nomenclature](#)
- [BioCarta](#) - Pathways of Life

### Acknowledgements



## KEGG LIGAND Database

Molecular building blocks of life in the chemical space

[KEGG2](#) [ATLAS](#) [PATHWAY](#) [BRITE](#) [GENES](#) [SSDB](#) [LIGAND](#) [DBGET](#)

### Chemical Substances and Reactions

**KEGG LIGAND** contains our knowledge on the universe of chemical substances and reactions that are relevant to life. It is a composite database currently consisting of COMPOUND, DRUG, GLYCAN, REACTION, RPAIR, and ENZYME databases. ENZYME is derived from the Enzyme Nomenclature, but the others are internally developed and maintained.

Database	Identifier	Content	Specialized entry point	
LIGAND	<a href="#">COMPOUND</a>	C number	Chemical compound structures	<a href="#">KEGG COMPOUND</a>
	<a href="#">DRUG</a>	D number	Drug structures	<a href="#">KEGG DRUG</a>
	<a href="#">GLYCAN</a>	G number	Glycan structures	<a href="#">KEGG GLYCAN</a>
	<a href="#">REACTION</a>	R number	Biochemical reactions	<a href="#">KEGG REACTION</a>
	<a href="#">RPAIR</a>	A number	Reactant pair alignments	
	<a href="#">ENZYME</a>	EC number	Enzyme nomenclature	

**Search**  for

bfind mode  bget mode

### LIGAND Relational Database

The primary database of KEGG LIGAND is a relational database with the [KegDraw](#) interface, which is used to generate the secondary (flat file) database for DBGET. A read-only copy of the relational database is also made publicly accessible.

#### Search COMPOUND

#### Search DRUG

#### Search GLYCAN

#### Search REACTION

### Computational Tools



# Компьютерные технологии в науке

Работа с информацией.