

## Орг. вопросы

- Сайт <http://intbio.org/bioinf2018>
- **Вводный опросник** – дедлайн 19 февраля 23:59.
- **Форма отчетности:** дифф. зачет с оценкой (!)
- **Контроль успеваемости:** посещение всех лекций + онлайн контрольные работы + вводный опросник = возможность получить отл. и зачет автомат.
- Пропуск лекции без ув. причины, невыполнение работ в срок = оценка не выше 4 баллов.
- **Контроль посещаемости** – способы будут меняться.

## Контроль посещаемости сегодня

Заполнить форму по адресу  
<http://intbio.org/2>  
 Возможность закрывается в 10:50!



либо  
 Записаться в список на перерыве

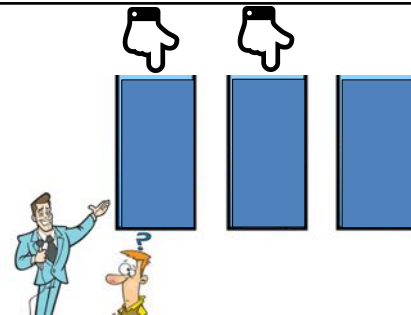
## ВВЕДЕНИЕ В БИОИНФОРМАТИКУ Лекция №2

Природа, передача и хранение информации. Базы данных. Биологические БД. Базы данных NCBI.

Алексей Константинович Шайтан, к.ф.-м.н.

Сайт курса: <http://intbio.org/bioinf2018>

19 февраля 2018



- В результате игрок получает **0.67 бита информации**
- Если бы ведущий открыл дверь в самом начале – только **0.58 бита информации**

[https://en.wikipedia.org/wiki/Monty\\_Hall\\_problem](https://en.wikipedia.org/wiki/Monty_Hall_problem)

## Природа информации

## Информация

ИНФОРМАЦИЯ, 1. Сведения об окружающем мире и протекающих в нем процессах, воспринимаемые человеком или специальным устройством.



Н.Н. Моисеев

... универсального определения информации не только нет, но и быть не может из-за широты этого понятия.

Information is information, not matter or energy.



Norbert Wiener

[http://www.aselibrary.ru/press\\_center/journal/fm/2007/number\\_3/number\\_3\\_6/number\\_3\\_6571/](http://www.aselibrary.ru/press_center/journal/fm/2007/number_3/number_3_6/number_3_6571/)

## Информация



Теперь выберите ёмкость.

32 ГБ<sup>1</sup>

52 990.00 руб.

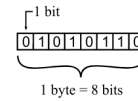
Доставка: на складе

128 ГБ<sup>1</sup>

60 990.00 руб.

Доставка: на складе

## Измерение информации



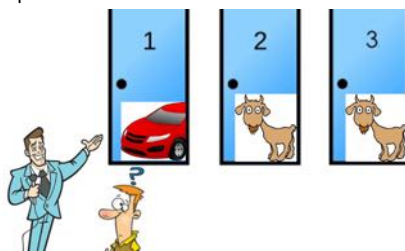
Измерения в байтах				Decimal	Binary
ГОСТ 8.417—2002		Приставки СИ			
Название	Обозначение	Степень	Название	Степень	
байт	Б	$10^0$	-	$10^0$	0000
килобайт	кбайт	$10^3$	кило-	$10^3$	0001
мегабайт	Мбайт	$10^6$	мега-	$10^6$	0010
гигабайт	Гбайт	$10^9$	гига-	$10^9$	0011
терабайт	Тбайт	$10^{12}$	тера-	$10^{12}$	0100
петабайт	Пбайт	$10^{15}$	пета-	$10^{15}$	0101
эксабайт	Эбайт	$10^{18}$	экса-	$10^{18}$	0110
зеттабайт	Збайт	$10^{21}$	зетта-	$10^{21}$	0111
иоттабайт	Ибайт	$10^{24}$	иотта-	$10^{24}$	1000
					1001

Вероятность того, что машина за этой дверью:

1/3

1/3

1/3



- Сколько информации нужно, чтобы закодировать положения машины?

1.5849625007211563... бит

## Теория информации

"the father of [information theory](#)"



**Claude Elwood Shannon**

(April 30, 1916 – February 24, 2001)

[https://youtu.be/z2Whj\\_nL-x8](https://youtu.be/z2Whj_nL-x8)

## Теория информации

Science: A short history of equations

**Without Claude Shannon's information theory there would have been no internet**

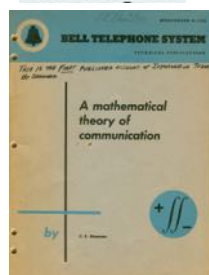
It showed how to make communications faster and take up less space on a hard disk, making the internet possible

$$H = -\sum p(x) \log p(x)$$

Введена мера информации(!)

кг, метр, секунда + БИТ

**The Guardian**



1948

<https://www.khanacademy.org/computing/computer-science/informationtheory/info-theory/v/intro-information-theory>

## Информационная энтропия

[Клод Шеннон](#) предположил, что прирост информации равен утраченной неопределённости, и задал требования к её измерению:

- мера должна быть непрерывной; то есть изменение значения величины вероятности на малую величину должно вызывать малое результирующее изменение функции;
- в случае, когда все варианты (буквы в приведённом примере) равновероятны, увеличение количества вариантов (букв) должно всегда увеличивать значение функции;
- должна быть возможность сделать выбор (в нашем примере букв) в два шага, в которых значение функции конечного результата должно являться суммой функций промежуточных результатов.

$$H = -\sum p(x) \log p(x)$$



## Передача информации

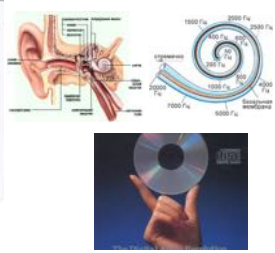
### Связь частоты сигнала и пропускной способности



1908 - 2005



1889 - 1976



#### Теорема Котельникова-(Найквиста-Шенона)

«любую функцию  $F(t)$ , состоящую из частот от 0 до  $f$ , можно непрерывно передавать с любой точностью при помощи чисел, следующих друг за другом через  $1/(2f)$  секунд»

44.1 кГц – частота дискретизации при записи звука

[https://ru.wikipedia.org/wiki/Теорема\\_Котельникова](https://ru.wikipedia.org/wiki/Теорема_Котельникова)

## Передача информации

### Связь частоты сигнала и пропускной способности



1888 - 1970

$$C = B \log_2 \left( 1 + \frac{S}{N} \right),$$

где

$C$  – пропускная способность канала, бит/с;

$B$  – полоса пропускания канала, Гц;

$S$  – полная мощность сигнала над полосой пропускания, Вт или В<sup>2</sup>;

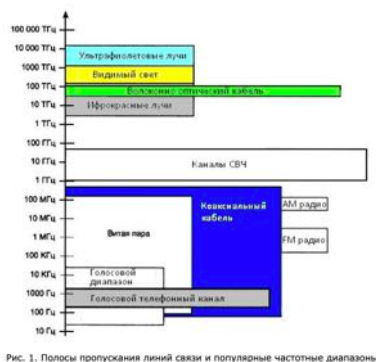
$N$  – полная шумовая мощность над полосой пропускания, Вт или В<sup>2</sup>;

$S/N$  – отношение мощности сигнала к шуму (SNR).

#### Теорема Шеннона-Хартли

[https://ru.wikipedia.org/wiki/Теорема\\_Шеннона\\_—\\_Хартли](https://ru.wikipedia.org/wiki/Теорема_Шеннона_—_Хартли)

## Передача информации



Оптоволокно

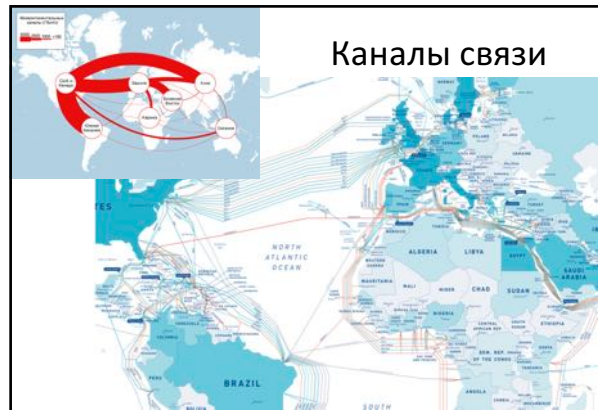


Антенны КНЧ

Рис. 1. Полосы пропускания линий связи и популярные частотные диапазоны

[https://ru.wikipedia.org/wiki/Связь\\_с\\_подводными\\_лодками](https://ru.wikipedia.org/wiki/Связь_с_подводными_лодками)

## Каналы связи



Карта подводных кабелей

<https://habrahabr.ru/company/rootwell/blog/305634/>

## Шифрование информации



[https://en.wikipedia.org/wiki/Public-key\\_cryptography](https://en.wikipedia.org/wiki/Public-key_cryptography)

## Криптосистемы с открытым ключом



#### Необратимая Хэш функция

```
mbptb:~ alexsha$ md5 -s 'Hello world!!!'
MD5 ('Hello world!!!') = 87ee732d831698f45b8606b1547bd09e
```

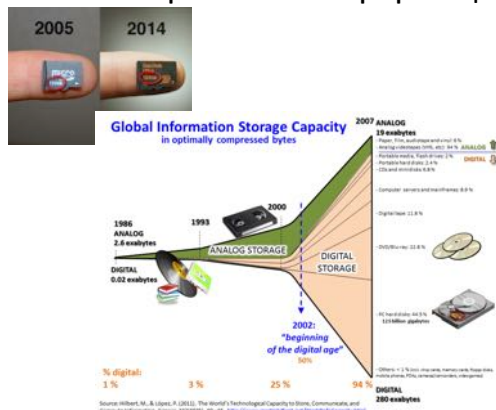
[https://en.wikipedia.org/wiki/Public-key\\_cryptography](https://en.wikipedia.org/wiki/Public-key_cryptography)

## Хранение информации

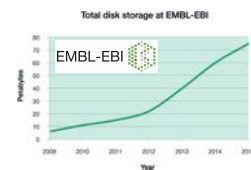
## Хранение информации



## Хранение информации



## Большие данные в биомедицине



2013-2021  
~\$400 млн

Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 billion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (800 Tbps)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 Tbps) data movement

Big Data: Astronomical or Genomic? PLOS 2015

## Источники больших данных в биомедицине

- Омиксные технологии
  - Секвенирование, геномика, транскриптомика, протеомика, метаболомика и т.д.
  - Коннектом мозга
- Медицинская информация
  - Электронные медицинские карты, результаты клинических исследований и т.д.
  - Медицинские изображения, МРТ и т.д.
- Структурная биология и моделирование
  - Данные с лазеров на свободных электронах (XFEL)
  - Моделирование структуры и динамики белков.

## Данные секвенирования, пример Геномы раковых опухолей



Геном человека ~ 3.3 Gb  
x100 секвенирование ~300Gb



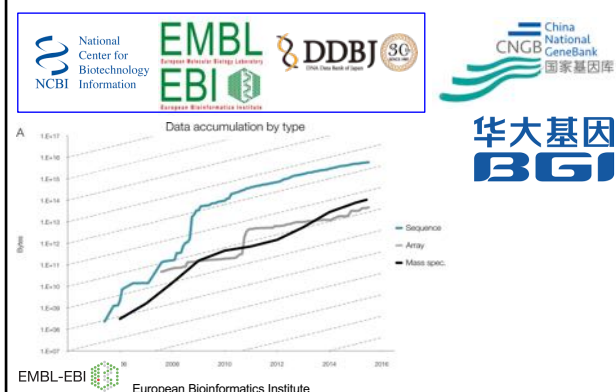
25000 образцов опухолей

Международный проект, данные распределены по миру

File ID	Director	Repository	Project	Study	Storage	Contact	Size
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA1	Aligned	4000	BAW	128,712 GB
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA2	Aligned	4000	BAW	127,212 GB
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA3	Aligned	4000	BAW	128,712 GB
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA4	Aligned	4000	BAW	128,712 GB
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA5	Aligned	4000	BAW	128,712 GB
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA6	Aligned	4000	BAW	128,712 GB
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA7	Aligned	4000	BAW	128,712 GB
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA8	Aligned	4000	BAW	128,712 GB
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA9	Aligned	4000	BAW	128,712 GB
PC0001_10027002	PC0001	London, PC0001, Barcelona, Collaborative, Toronto, USA	BRCA10	Aligned	4000	BAW	128,712 GB



## Централизованные репозитории омических данных



## Genomes en masse

Composite image featuring the Saudi Human Genome Program logo, a Nature journal article snippet about AstraZeneca's project to sequence 2 million genomes, and the Chinese Millionome Database logo.

Text: 5 years ~ 100 000 genomes

Logos for BGI (华大基因) and CNGB (国家基因库) are also visible.

## Структурная биология и моделирование

Composite image showing a satellite view of the European XFEL facility in Hamburg, a diagram of a synchrotron beamline, and a 3D model of a protein structure.

Text: European XFEL, Hamburg

Text: 27000 импульсов в секунду

## Передача больших данных

Выделенные научно-образовательные сети 100Gbit/s

Map of the Asia Pacific Ring (APR) network showing connections between major cities like Tokyo, Seoul, Taipei, and Los Angeles. Logos for aspera (an IBM company) and globus toolkit are shown.

Text: Figure 1: Asia Pacific Ring (APR)

Text: Программные решения

## Базы данных

## Базы данных

- Реляционные базы данных, объектно-ориентированные, RDF
- Системы управления базами данных СУБД
- Языки и стандарты SQL, SPARQL, RDF



## Реляционные базы данных

Клиент				Товар	
Id_кл	Фамилия	Имя	Отчество	Id_тов	Название
15	Иванов	Иван	Иванович	1	Шкаф
16	Петров	Петр	Петрович	2	Стул
17	Николаев	Николай	Николаевич	3	Стол

Заказ				
Id_за	Клиент	Товар	Дата	Количество
1	15	1	15.09.2003	1
2	17	1	17.09.2003	2
3	15	2	20.09.2003	12

SQL      Целостность данных      Транзакции Атомарность      Соответствие требованиям ACID

Изолированность      Надежность

<https://aws.amazon.com/ru/relational-database/>

## Resource Description Framework

### Язык SPARQL

Resource Description Framework (RDF, «среда описания ресурса»<sup>[1]</sup>) — это разработанная консорциумом Всемирной паутины модель для представления данных, в особенности — метаданных<sup>[2]</sup>. RDF представляет утверждения о ресурсах в виде, пригодном для машинной обработки. RDF является частью концепции семантической паутины.

Субъект → Предикат → Объект



Select all human UniProt entries with a sequence variant that leads to a 'loss of function'

[https://ru.wikipedia.org/wiki/Resource\\_Description\\_Framework/](https://ru.wikipedia.org/wiki/Resource_Description_Framework/)

## Биологические базы данных

## Biology is a data-intensive science!

- Нужно уметь хранить данные
- Нужно уметь обрабатывать данные
- Нужно уметь обмениваться данными
- Данные должны быть максимально открыты и доступны научному сообществу.
- Data provenance ("происхождение данных")

### Data provenance [ edit ]

Scientific research is generally held to be of good provenance when it is documented in detail sufficient to allow reproducibility.<sup>[27][28]</sup> Scientific workflow systems assist scientists and programmers with tracking their data through all transformations, analyses, and interpretations. Data sets are reliable when the process used to create them are

- Кризис воспроизводимости результатов в науке!?

## Базы данных для биологии

- На данный момент количество не возможно сосчитать — очень много — важно не запутаться и не потеряться при их использовании
- Надежные источники информации о базах данных — научные журналы



Annual Database Issue — информация о ~200 БД каждый год

## Базы данных для биологии

- Бесплатные vs Платные (по подписке)
- Свободно доступные vs Ограниченно доступные
- Большие ресурсы (NCBI, EBI/EMBL, etc.) интегрирующие многие базы данных - поддерживаются государством
- Коллаборации между университетами (напр. PDB)
- Коммерческие компании
- Локальные базы данных, поддерживаемые силами научных групп
- База данных vs Web Server — граница размыта.
- Хорошие БД - информационные ресурсы с возможностями сложного поиска и моделирования.

## Крупные центры биологических БД



- Bethesda, MD USA
- Более 60 БД включая PubMed, GenBank, DBGap, SRA



- European Bioinformatics Institute, Cambridge, UK + Switzerland

## Что храниться?

- БД статей, абстрактов, патентов
- Последовательности ДНК
- Последовательности белков
- 3D структуры молекул
- Геномы
- Данные экспрессии
- Сырые данные с секвенаторов
- Информация о химических соединениях и их активности
- Информация о болезнях, информация о пациентах
- Информация о видах живых организмов
- Информация о метаболических и сигнальных путях
- Информация о взаимодействии молекул
- Много производной информации: базы гомологичных последовательностей, аннотация отдельных классов белков и т.д.

### План

- **Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)**
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии

### Реферативные базы данных

#### Clinical/Biomedical

**PubMed** – US National Library of Medicine database (Medline); refers to >25M articles from 5600 biomedical journals, 1940s to present, with some older items, in medicine, nursing, dentistry, veterinary medicine, allied health & pre-clinical sciences  
- bibliographic database with author-provided abstracts, added indexing terms from **MeSH** (Medical Subject Headings) thesaurus, & links to other resources

[www.pubmed.gov](http://www.pubmed.gov)



FREE

### Реферативные базы данных

#### Clinical/Biomedical

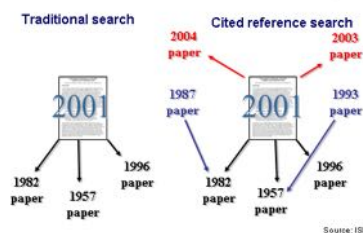
**Embase** – European based, includes all of Medline (database behind PubMed) and more; > 29M records, >8,500 journals, 1940s to present; includes coverage of more basic science journals & pre-clinical topics - especially useful for drug pipeline information, biotechnology, medical devices, conference coverage, toxicology, health policy/management, & alternative/complementary medicine  
EMTREE thesaurus includes almost twice as many terms as PubMed

<https://www.elsevier.com/solutions/embase-biomedical-research>

ELSEVIER PAID

### Реферативные базы данных

#### Cited Reference Searching





### Реферативные базы данных Общенаучные базы данных цитирований

**Web of Science** - covers >12,000 journals from 1900 to present; useful for cited reference, **conference information & affiliations** (institutions)

<https://webofknowledge.com/>



**Scopus** – covers >18,500 journals from 1823 to present, complete citation counts for indexed articles 1996 to present; a general science database, not a specialized database – useful for cited reference, **conference information & affiliations** (institutions)

<https://www.scopus.com/>



### Реферативные базы данных Общенаучные базы данных цитирований



Elibrary.ru/ПИНЦ

### Реферативные базы данных Общенаучные базы данных цитирований

Google Scholar



☒ Articles ☐ Case law

#### Recommended articles

Exploring DNA dynamics within oligonucleosomes with coarse-grained simulations: SIRAH force field extension for protein-DNA complexes  
A Brandner, A Schüller, F Melo, S Pantano - Biochemical and biophysical research ..., 2017

### Базы данных диссертаций

Open DOAR <http://www.openoar.org/index.html> ;  
OpenThesis <http://www.openthesis.org/> ;  
BASE – Bielefeld Academic Search Engine -  
<http://www.base-search.net/>  
> refine search result > document type > theses

ProQuest Dissertations & Theses  
Database <http://www.proquest.com/products-services/pqdt.html> - from 1743  
to present; some fulltext since 1990; fee with some  
free search capability

### Базы данных патентов



[http://www.lens.org/lens/biological\\_search](http://www.lens.org/lens/biological_search) – поиск ДНК последовательностей

Google Patents



☐ Include non patent literature (Google Scholar)

Search and read the full text of patents from around the world.

### План

- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- **Базы данных последовательностей ДНК**
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии

## Базы данных нуклеотидных последовательностей

Нуклеотидные БД – это хранилища, принимающие данные от научного сообщества и представляющие их широкой общественности. Различные БД отличаются по источнику последовательностей, их надежности, широте аннотирования и т.д. В идеале БД должна содержать все известные последовательности.

The International Nucleotide Sequence Database Collaboration – совместный проект EMBL-Bank в Европейском Институте Биоинформатики (EBI), японского банка данных ДНК (DDBJ) в Центре Информационной Биологии (CIB) и GenBank в Национальном Центре Биотехнологической Информации (NCBI).



55

## База данных GenBank

Открытая БД нуклеотидных последовательностей, учреждена в 1982 г.

2017: > 300 000 организмов, ~ 203 млн. последовательностей,  
~ 240 млрд. пар оснований

**Sample GenBank Record**

This page presents an annotated sample GenBank record (accession number **U49845**) in its GenBank Flat File format. You can see the corresponding [live record](#) for U49845, and see [examples of other records](#) that show a range of biological features.

LOCUS	SCU49845	5028 bp	DNA	PLN	21-JUN-1999
DEFINITION	Saccharomyces cerevisiae TCP1-beta gene, partial cde, and Axl2p (Axl2) and Rev7p (REV7) genes, complete cde.				
ACCESSION	U49845				
VERSION	U49845.1 GI:1293613				
KEYWORDS	-				
SOURCE	Saccharomyces cerevisiae (baker's yeast)				
ORGANISM	Saccharomyces cerevisiae				
	Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomycetes.				
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)				
FUNDED	7871890				
REFERENCE	2. (bases 1 to 5028)				
AUTHORS	Romer, F., Madden, R., Chang, J. and Snyder, M.				
TITLE	Selection of axial growth sites in yeast requires Axl2p, a novel plasma membrane glycoprotein				
JOURNAL	Gene Dev. 10 (7), 777-793 (1996)				
FUNDED	8846915				

## База данных GenBank. Структура файла

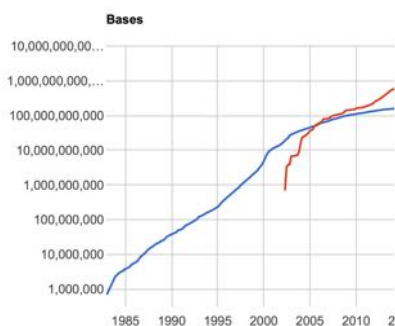
```
...
FT      /translation="MQQPGNGSAFLAPNGSHAPDHVTDQRDEVVVVGMGIVMSLIVL
FT      AIVFGNVLVITAIKFERLTQTVNMFITSLACADLVMLAVVPPGAHIIMRWMTFGNF
FT      WCEFWTSIDVLCVTASITLCVIAVDYFAITSPFKYQSLTKNKARVILMVIVSGL
FT      TSFLPIQMHWYRATHQEAICYANETCCDFFTNQAYAIASSIVSFYVPLVMVFVYSRV
FT      FQEAQRQLQKIDKSEGRFHVQNLQVEQDGRTHGLRSSKFKLKEHKALKTLGIIMGT
FT      FTLCWLPPFFIVNIHVIQDNLIRKEVYILLNWIGYVNSGFNPLIYCRSPDFRIAFQELL
FT      CLRBSLSLKAYGNGYSSNGNTGRQSGVHVEQEKENKLLCEDLPGETDFVGHQGTVPDSNI
FT      DSQGRNCSTNDSLLK
FT      variation 46
FT      /gene="ADRB2K
FT      /replace="ak
FT      /note="Arg16 to Gly polymorphism"
XX
...
```

57

## База данных GenBank. Структура файла

```
...
SQ Sequence 1242 BP; 275 A; 331 C; 326 G; 310 T; 0 other;
atggggcgaac cggggaacg cagcgccctt ttgtggcga caatggaa caatgcgcg 60
gaccagcagc tcacgcagca aagggcagcg gtgtgggtgg tgggcattgg catgtccatg 120
tctctcatcg tctggccat cgtgttggg aatgtctgg tcatcagc catggccaa 180
tcagacgctg tgcacaggt caccactac tcatcactt caatggcgt tgcgtatctg 240
gtcatgggoc tggcagtggt gcccttggg gccgccaca ttcttatga aatgtggact 300
tttggcaact tctgttgcga gttttggat tcatgtgat tgcgtgctg caaggccagc 360
attgagcccc tgtgctgat cgcagtggt cgtactctt caattactt ccatttcaag 420
taccagagcc tgcagacca gaataagcc cgggtgatc ttctgatgt gtggattgtg 480
tcaggcccta cctctctctt gccattcag atgcactgt accggccca ccaacaggaa 540
gccatcaact gctatgcaa tgagacctg tgtgactct tcacagcaa agcctatgcc 600
attgctcttt ccatcgtgt cttctacgt cccctgttg tcatgtctt cgtctactcc 660
agggtctttt agggagccaa aaggcagct cagaagattg acaaatctga gggccgcttc 720
catgtccaga accctagcaa ggtggagcg gatggcgga cggggcatgg actccgcaga 780
tcttccaaat tctgcttga ggagacaaa gccctcaag cgttagcat catcatgggc 840
acttccacc tctgctggt gcccttctt atcgttaaa ttgtgatgt gatccaggat 900
aacctcatcc gtaaggaagt ttacatctc taaatttga taggtatgt caattctggt 960
ttcaatcccc ttatctactg ccggagccca gatttcaga ttgccttca ggagctctgt 1020
tgcttcgca ggtctctt gaagcccat gggaatggt actccagca cggcaacaca 1080
ggggagcaga gtgatatca cgtggaacg gagaagaaa ataaactgt gttgtaagc 1140
ctccaggca cggaaagatt tgtgggcat caagtactg tgcttagcga taacattgat 1200
tcacaaggga ggaattgtg tacaatgac tcaactcgt aa 1242
//
```

58



GenBank and WGS Statistics

<https://www.ncbi.nlm.nih.gov/genbank/statistics/>

Genbank – is an archive! Contains everything.

### Nicotiana tabacum chloroplast JLA region, sequence 2

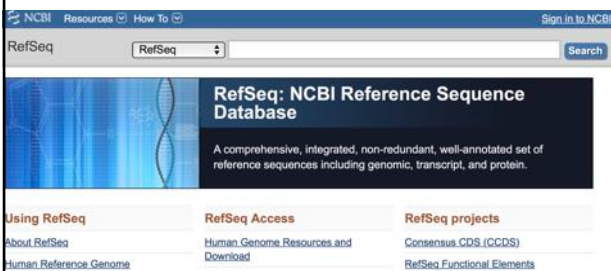
GenBank: Z71230.1

FASTA Graphics

FEATURES	Location/Qualifiers
source	1..124
	/organism="Nicotiana tabacum"
	/organism="plastid:chloroplast"
	/mol_type="genomic DNA"
	/isolate="Cuban habano cigar, gift from President Fidel Castro"
	/db_xref="taxon:597"

RefSeq – is a reference sequence database!

### RefSeq – is a reference sequence database!



Если нужен список последовательностей всех генов человека – это вопрос к RefSeq, а не GenBank!

### План

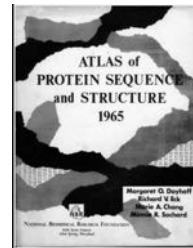
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- **Базы данных последовательностей белков**
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии

### GenBank/RefSeq is nucleotide centric, but

```
...
...
FT      /translation="MGQPGNGSAFLAPNGSHAPDHDVTQQRDEVWVVGIMVSLIVL
FT      AIVFGNVLVITAIKFERLQVTNYFITSLACADLVGLAVVFGAAHILMRMTFGNF
FT      WCEFWTSIDVLCVTASITLCVIAVDYFAITSPFKYQSLTGNKARVILMVIVSGL
FT      TSFLPIQMHWRATHQEAICYANETCCDFFTNQAYAIASSIVSYVPLVIMVFYSRV
FT      FQBAKRQLQIKDSEGRFHVQNLISQVEQDGRGTGGLRSSKFKLKEHKALKTLGIMGT
FT      FTLCLWLPFFIVNIVHVQDMLIRKEVYILLNWIGYVNSGFNPLIYCRSPDFRIAFQELL
FT      CLRRSSLKAYNGYSSNGTGEQSGYHVEQEKENKLLCEDLPGETDFVGHQGTVPSPNI
FT      DSQGRNCSTNDSLL«
FT      variation      46
FT      /gene="ADRB2«
FT      /replace="a«
FT      /note="Arg16 to Gly polymorphisme
XX
...
```

Protein sequences are annotate within GB records <sup>63</sup>

### Protein Centric Sequence Databases



Margaret Oakley Dayhoff  
1925-1983

Margaret Dayhoff, a founder of the field of bioinformatics

Invented one-letter amino acid code, substitution matrices, etc.

[https://en.wikipedia.org/wiki/Margaret\\_Oakley\\_Dayhoff](https://en.wikipedia.org/wiki/Margaret_Oakley_Dayhoff)

### Protein Centric Sequence Databases

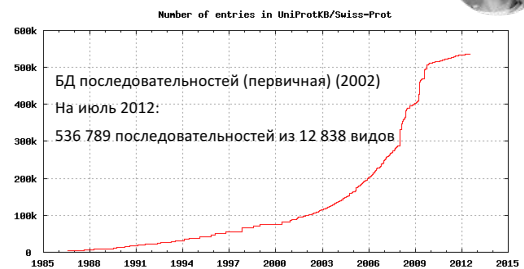


<http://pir.georgetown.edu>

In 2002, PIR along with its international partners, EBI (European Bioinformatics Institute) and SIB (Swiss Institute of Bioinformatics), were awarded a grant from NIH to create UniProt, a single worldwide database of protein sequence and function, by unifying the PIR-PSD, Swiss-Prot, and TrEMBL databases. As of 2010, PIR offers a wide variety of resources mainly oriented to assist the propagation and standardization of protein annotation: PIRSF,<sup>[8]</sup> iProClass, and iProLINK.

The Protein Ontology (PRO) is another popular database released by the Protein Information Resource.<sup>[9][10]</sup>

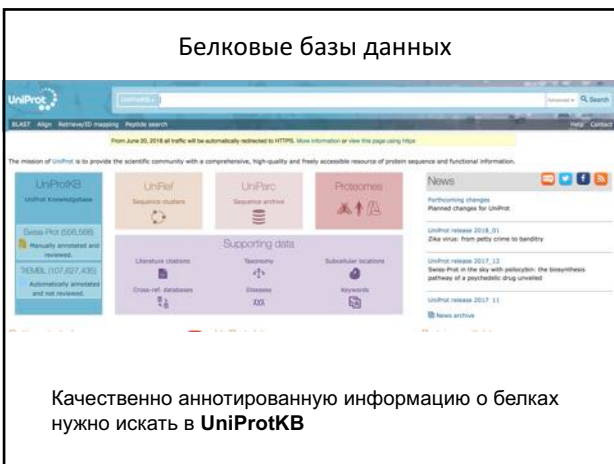
### Белковые базы данных



UniProt – наиболее всеобъемлющий каталог информации о белках, объединяющий в себе данные из UniProtKB/Swiss-Prot, UniProtKB/TrEMBL и PIR.

66

## Белковые базы данных



Качественно аннотированную информацию о белках нужно искать в **UniProtKB**

## План

- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- **Базы данных 3D структур**
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии

## Структурные базы данных



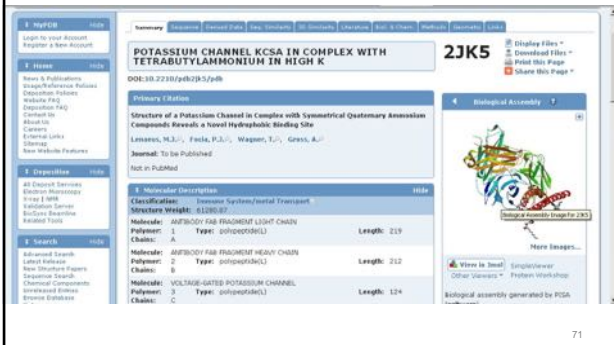
PDB – содержит информацию об экспериментально определенных структурах белков, нуклеиновых кислот и различных комплексов.

## Структурные базы данных



## Структурные базы данных

### POTASSIUM CHANNEL KCSA IN COMPLEX WITH TETRABUTYLAMMONIUM IN HIGH K



## Структурные базы данных

### POTASSIUM CHANNEL KCSA IN COMPLEX WITH TETRABUTYLAMMONIUM IN HIGH K





## База данных PDB. Структура файла

```

HEADER      IMMUNE SYSTEM/METAL TRANSPORT          15-AUG-08   2JK5
TITLE       POTASSIUM CHANNEL KCSA IN COMPLEX WITH TETRABUTYLAMMONIUM
TITLE       2 IN HIGH K
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: ANTIBODY FAB FRAGMENT LIGHT CHAIN;
COMPND      3 CHAIN: A;
COMPND      4 ENGINEERED: YES;
COMPND      5 MOL_ID: 2;
...
KEYWDS      IMMUNE SYSTEM METAL TRANSPORT COMPLEX, QUATERNARY AMMONIUM,
...
EXPDTA      X-RAY DIFFRACTION
AUTHOR      M.J.LENAEUS,P.J.FOCIA,T.WAGNER,A.GROSS
REVDAT      1 17-NOV-09 2JK5 0
JRNL        AUTH  M.J.LENAEUS,P.J.FOCIA,T.WAGNER,A.GROSS
JRNL        TITL  STRUCTURE OF A POTASSIUM CHANNEL IN COMPLEX WITH
JRNL        TITL  2 SYMMETRICAL QUATERNARY AMMONIUM COMPOUNDS REVEALS
JRNL        TITL  3 A NOVEL HYDROPHOBIC BINDING SITE
JRNL        REF   TO BE PUBLISHED
JRNL        REFM  REFM
REMARK      2 RESOLUTION.      2.4  ANGSTROMS.
REMARK      3
REMARK      3 REFINEMENT.
REMARK      3 PROGRAM      : REFMAC 5.5.0051
...

```

73

## Структурные базы данных

NDB – основана в 1992 г. для сбора и распространения информации о структуре нуклеиновых кислот. Формат хранения данных идентичен PDB.



74

## Структурные базы данных

75

## Структурные базы данных

76

## План

- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- **Базы данных хим. соединений**
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии

## Базы данных химических соединений

78



## Базы данных химических соединений

The screenshot shows the PubChem website interface. At the top, there's a search bar with 'epinephrine' entered. Below the search bar, there are tabs for 'Advanced Search', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The main content area displays search results for 'epinephrine', including its chemical structure, molecular weight (183.204420 g/mol), and various links to related structures, bioassays, and literature. A sidebar on the right lists 'Selected Compounds' with counts for different categories like 'BioAssays', 'Protein 3D Structures', and 'Crystal Structures'.

## Базы данных химических соединений

Chemical Abstract Service – в регистре содержится 130 млн соединений (2018)

The screenshot shows the CAS website, an initiative of the American Chemical Society. It features a navigation bar with links like 'ABOUT CAS', 'OUR EXPERTISE', 'SOLUTIONS', 'PRODUCTS & SERVICES', 'SUPPORT & TRAINING', and 'NEWS & EVENTS'. The main banner highlights 'Carbon Bond Formations Win 2010 Nobel Prize with Award-Winning Research in CAS Databases'. Below the banner, there are sections for 'QUICK LINKS', 'THE RESEARCH EDGE', and 'CAS UPDATES'.

## Базы данных химических соединений

The screenshot shows the Reaxys website, which is described as a platform that retrieves literature, compound properties, and chemical reaction data. It features a 'Get started' button and a section for 'Already a Reaxys customer?' with a link to sign in. The website also promotes its role in 'Life is chemistry' and 'Teaching chemistry info literacy'.

## Базы данных углеводов



For 2017:

7005 publications for 18924 compounds from 8859 organisms

The screenshot shows the search interface of the Carbohydrate Structure Database. It includes a 'Database search' section with icons for 'Structures', 'Composition', 'Organisms', 'Publications', and 'NMR signals'. Below this, there's a 'Useful tools' section with icons for 'Predict NMR', 'Elucidate', 'Fragments', 'Cluster taxa', 'GT activities', and 'Examples'.

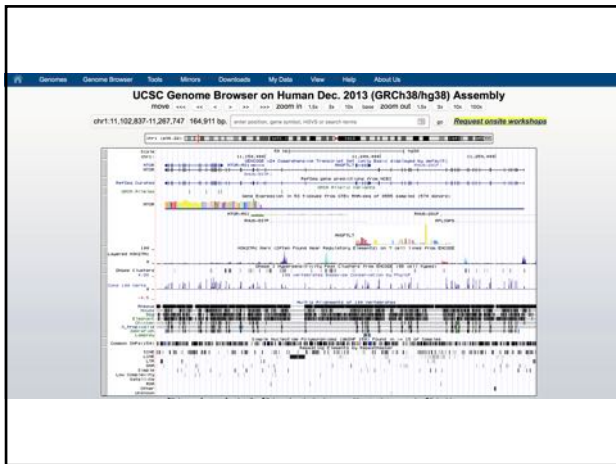
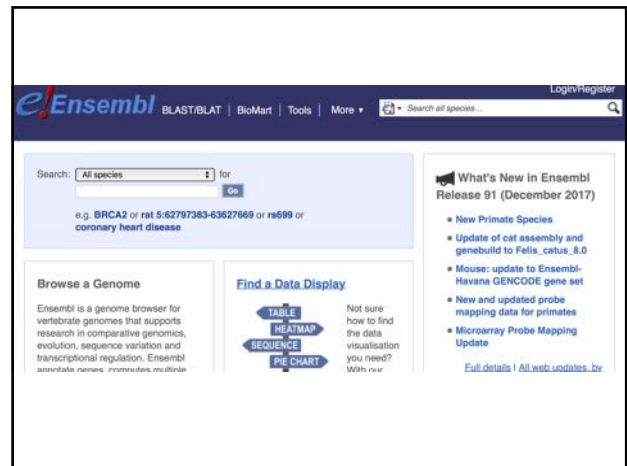
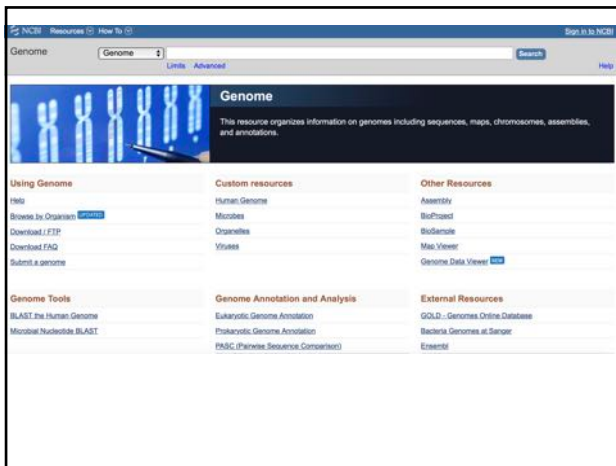
## Структурные базы данных

7009 структур липидов и сходных соединений – не поддерживается в настоящее время ☹

The screenshot shows the LipidBank website, which provides detailed information about lipids. The example shown is for 'Vitamin A'. It includes the chemical structure, molecular weight (286.452), and a section for 'BIOLOGICAL ACTIVITY' describing its role in vision and immune defense. The website also mentions that it does not currently support 7009 lipid structures and related compounds.

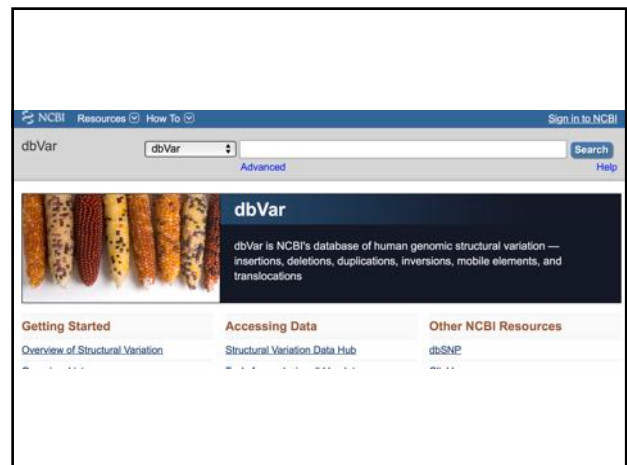
## План

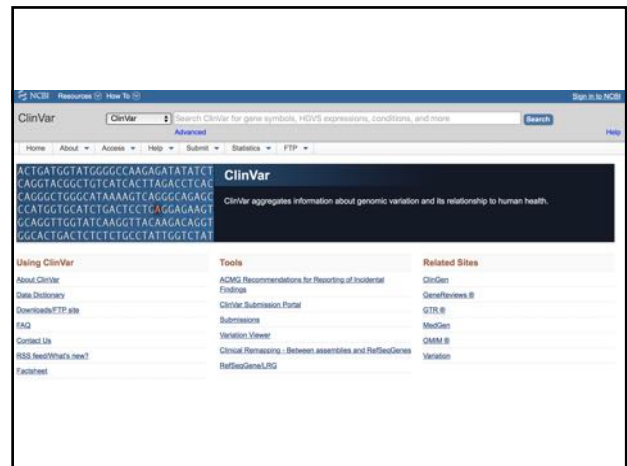
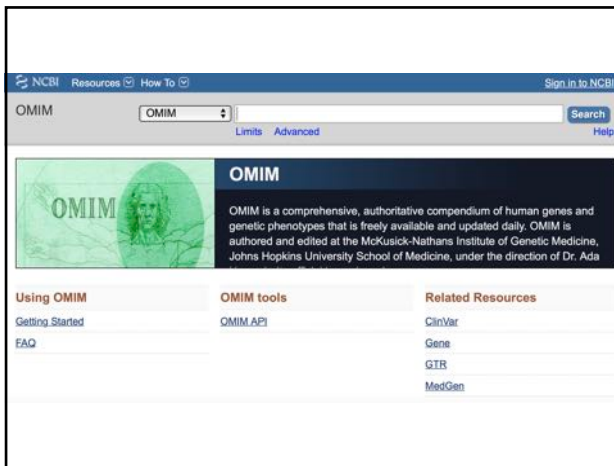
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- **Базы данных геномов и аннотаций**
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии



#### План

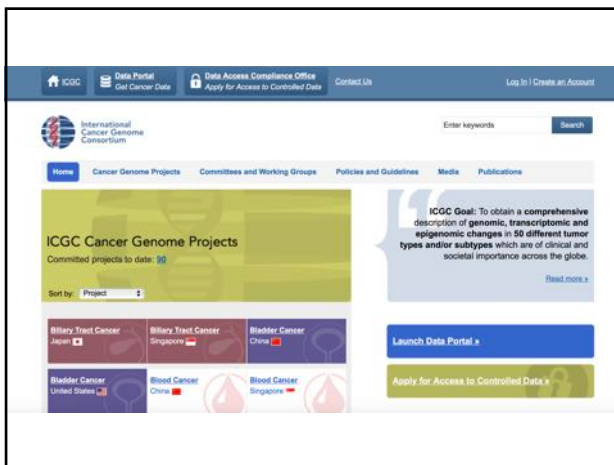
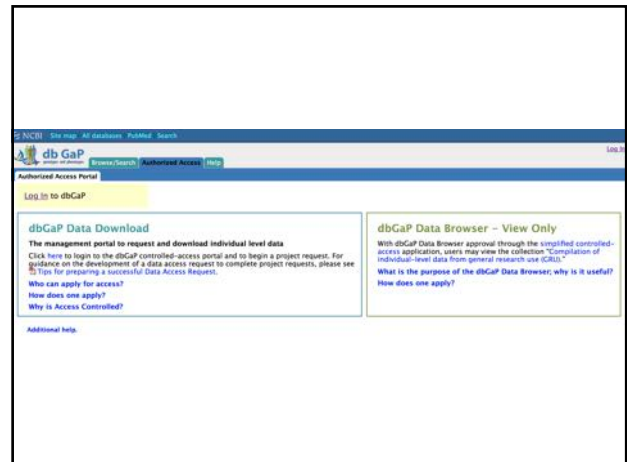
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- **Базы данных вариаций генома**
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии





### План

- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- **Базы данных геном-фенотип**
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии



### План

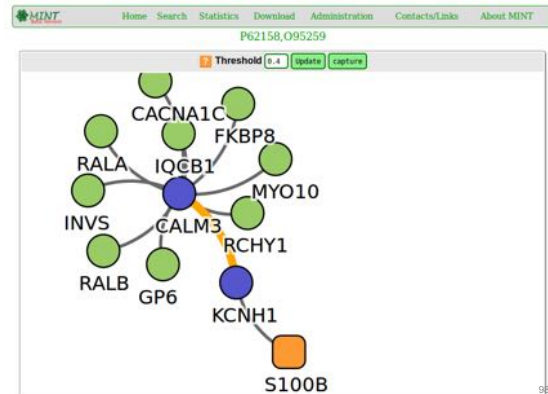
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- **Базы данных взаимодействий**
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии

## Базы данных взаимодействий



97

## Базы данных взаимодействий



98

## Базы данных взаимодействий

Kalium

174 items shown

Organism	Name	UniProt ID	Sequence	POS	Mass	Pub Date	Activity
Leishmania panamensis	LeishTn 1.1	C9TXA	CTTAAAG...	12844117	4295.05	1999	ShakerKv4.2/Kv4.2
Leishmania panamensis	LeishTn 1.2	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2/Kv4.2
Leishmania panamensis	LeishTn 1.3	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.4	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.5	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.6	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.7	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.8	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.9	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.10	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.11	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.12	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.13	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.14	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.15	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.16	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.17	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.18	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.19	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.20	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.21	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.22	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.23	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.24	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.25	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.26	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2
Leishmania panamensis	LeishTn 1.27	C9TXA	CTTAAAG...	12844117	4295.05	1999	Kv4.2

99

## Базы данных химических соединений

DRUGBANK

WHAT ARE YOU LOOKING FOR?

Tylenol

Drugs Targets Pathways Indications

DRUGBANK

100

## План

- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- **Базы данных сигнальных/метаболических путей**
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии



## KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

Menu PATHWAY BRITE MODULE KO GENES LIGAND NETWORK DISEASE DRUG DBGET

Select prefix map Organism Enter keywords Go Help

[ New pathway maps | Update history ]

**Pathway Maps**

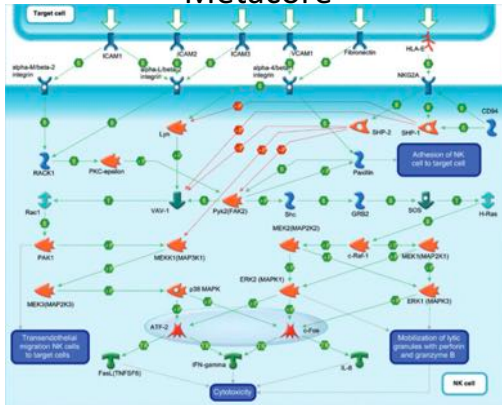
KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for:

1. Metabolism
  - Global/overview
  - Carbohydrate
  - Energy
  - Lipid
  - Nucleotide
  - Amino acid
  - Other amino
  - Glycan
  - Cofactor/vitamin
  - Terpenoid/PK
  - Other secondary metabolite
  - Xenobiotics
  - Chemical structure
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Organismal Systems
6. Human Diseases
7. Drug Development

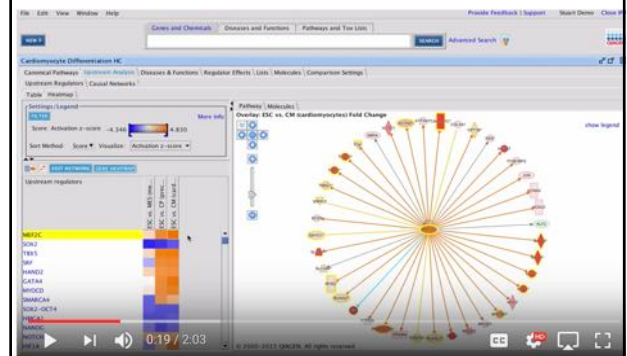
KEGG PATHWAY is a reference database for **Pathway Mapping**.



## Metacore

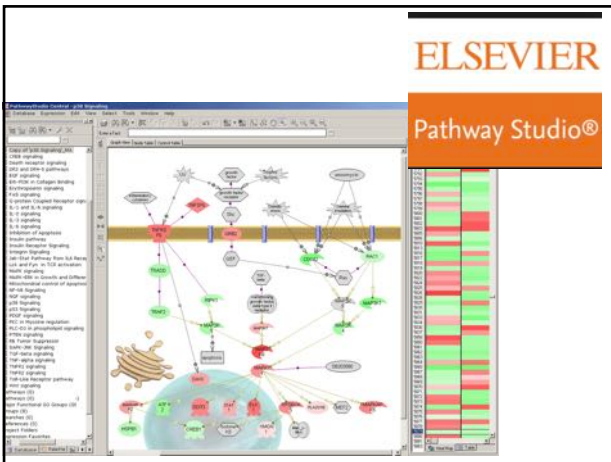


## Ingenuity Pathway Analysis



ELSEVIER

Pathway Studio®



### План

- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- **Базы данных секвенирования**
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии

### План

- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- **Базы данных клинических исследований и лекарств**
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии

NCBI Resources | How To | Sign In to NCBI

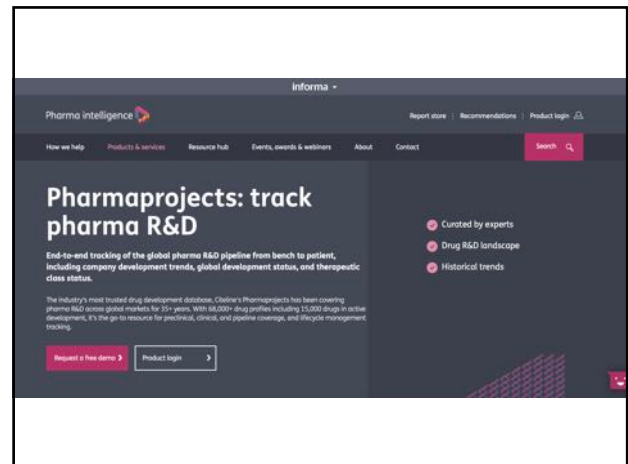
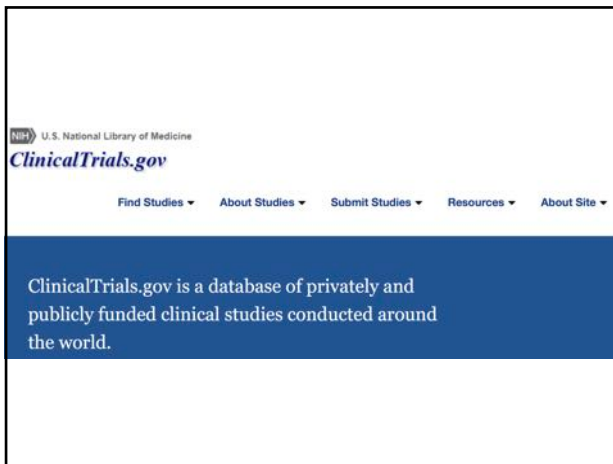
SRA | SRA | Advanced | Search | Help

**SRA**

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, and Applied Biosystems SOLiD System®.

Getting Started	Tools and Software	Related Resources
<a href="#">How to Submit</a>	<a href="#">Download SRA Toolkit</a>	<a href="#">Submission Portal</a>
<a href="#">Log in to SRA (for updating and troubleshooting submissions)</a>	<a href="#">SRA Toolkit Documentation</a>	<a href="#">Trace Archive</a>
<a href="#">Log in to Submission Portal (for submitting sequence data)</a>	<a href="#">SRA-BLAST</a>	<a href="#">dbGap Home</a>
	<a href="#">SRA Run Browser</a>	<a href="#">BioProject</a>

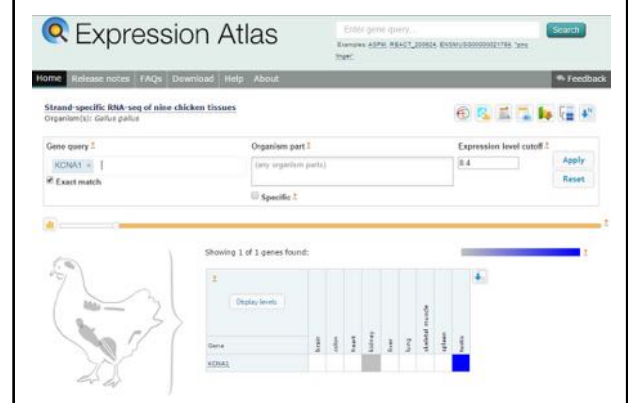




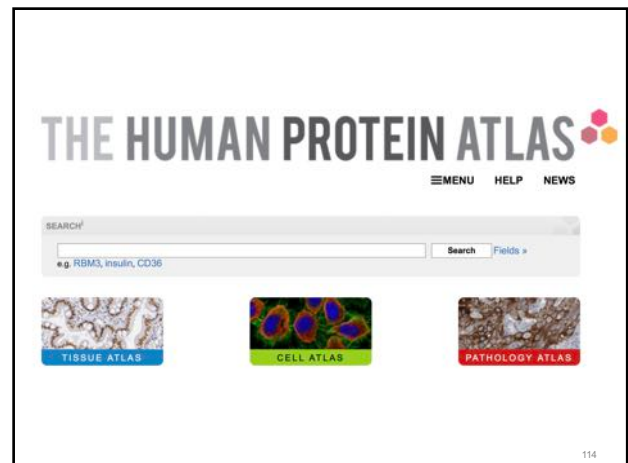
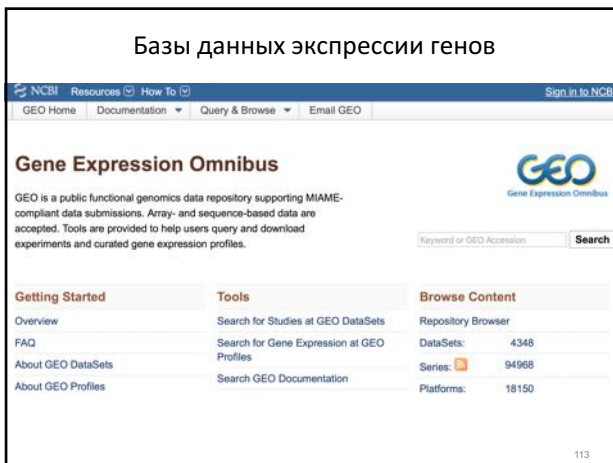
### План

- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- **Базы данных по экспрессии генов/гистологии**
- Базы данные по таксономии

### Базы данных экспрессии генов



### Базы данных экспрессии генов



### План

- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы данных вариаций генома
- Базы данных геном-фенотип
- Базы данных взаимодействий
- Базы данных сигнальных путей
- Базы данных секвенирования
- Базы данных заболеваний и медицинской информации
- Базы данных по экспрессии генов/гистологии
- **Базы данных по таксономии**

### Таксономические базы данных

**Taxonomy Browser** – знаменитая таксономическая БД, имеющая иерархическую структуру, основанную на анализе последовательностей и призванная упорядочить классификацию организмов, для которых известна хотя бы одна последовательность ДНК или белка.



### Видовые базы данных

Содержат таксономическую, библиографическую, географическую, визуальную и прочую информацию

**AlgaeBase** content about team notulae algarum links contact search

genus - species - literature - journals - images - common names - distribution - glossary - taxonomy browser - higher taxonomy

143,552 species and infraspecific names are in the database, 18,634 images, 54,056 bibliographic items, 314,920 distributional records.

**Haematococcus pluvialis** Flotow

**Publication details**  
Haematococcus pluvialis Flotow 1844: 415, 537, pls. XXIV, XXV  
Published in: Flotow, J. von (1844). Beobachtungen über Haematococcus pluvialis. Verhandlungen der Kaiserlichen Leopoldinisch-Carolinischen Deutschen Akademie der Naturforscher 12 (Abt. 2): 413-606, 3 pls.  
Download PDF

**Type species**  
This is the type species (lectotype) of the genus Haematococcus.

**Status of name**  
This name is of an entity that is currently accepted taxonomically.

**Classification**  
Eukarya  
Kingdom Plantae  
Subkingdom Viridiplantae  
Infrakingdom Chlorophyta  
Infrakingdom Rhodophyta  
Subphylum Chlorophytina  
Class Chlorophyceae  
Order Chloromonadales  
Family Haematococcaceae  
Genus Haematococcus

**Taxonomy**  
References  
Submit Feedback  
Submit Reference  
Links

117

### Видовые базы данных

<https://plant.depo.msu.ru>

**Депозитарий живых систем «НОВЕ КОСКИ»**

О системе Контакты Ссылки

Национальный банк-депозитарий живых систем

Проект Московского университета «Нове коски» посвящен созданию многофункционального центра хранения биологического материала.

Планируется работа с материалами всех возможных типов - от отдельных биологических материалов до целых живых организмов.

Создание депозитария позволит собирать, безвредно хранить, изучать и создавать новые способы пользования биологическим материалом.

**Фото дня** **Организм недели**

Подайте образец

118