

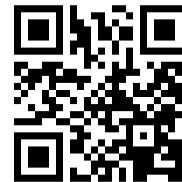
## Орг. вопросы

- 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

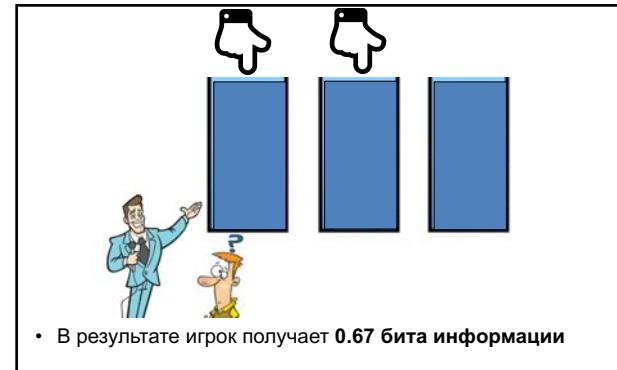
Природа, передача и хранение информации. Базы данных. Биологические БД. Базы данных NCBI.

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

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

19 февраля 2018

## Природа информации



- 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)

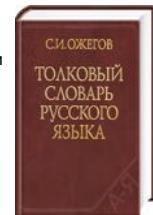
## Информация

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



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

Н.Н. Моисеев



Information is information, not matter or energy.

Norbert Wiener

[http://www.aselibrary.ru/press\\_center/journal/inr/2007/number\\_3/number\\_3\\_6/number\\_3\\_657/](http://www.aselibrary.ru/press_center/journal/inr/2007/number_3/number_3_6/number_3_657/)

## Информация

iPhone 7 Plus Серебристый Ёмкость

Теперь выберите ёмкость.

32 ГБ <sup>1</sup> 52 990.00 руб.	128 ГБ <sup>1</sup> 60 990.00 руб.
--------------------------------------	---------------------------------------

Доставка: на складе Доставка: на складе

## Измерение информации

1 bit  
0 1 0 1 0 1 1 0  
1 byte = 8 bits

Измерения в байтах			
ГОСТ 8.417—2002		Приставки СИ	
Название	Обозначение	Степень	Название
байт	Б	$10^0$	$10^0$
килобайт	кбайт	$10^3$	кило-
мегабайт	Мбайт	$10^6$	mega-
гигабайт	Гбайт	$10^9$	гига-
терабайт	Тбайт	$10^{12}$	тера-
петабайт	Пбайт	$10^{15}$	пета-
экабайт	Эбайт	$10^{18}$	эко-
зеттабайт	Збайт	$10^{21}$	зетта-
иоттабайт	Ибайт	$10^{24}$	иотта-

Вероятность того, что машина за этой дверью:

1/3	1/3	1/3
-----	-----	-----

- Сколько информации нужно, чтобы закодировать положении 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

Введена мера информации(!)  
кг, метр, секунда + БИТ

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:

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

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

## Сжатие информации



	<b>Tom 12.txt</b>	1.5 MB
Modified: Today, 4:01 PM		
<b>General:</b>		
Kind:	Plain Text Document	
Size:	1,473,547 bytes (1.5 MB on disk)	
	<b>Tom 12.txt.zip</b>	602 KB
Modified: Today, 4:02 PM		
<b>General:</b>		
Kind:	ZIP archive	
Size:	602,098 bytes (602 KB on disk)	

Буква в тексте не независимы – одни встречаются чаще других в разных контекстах.

## Сжатие информации

### Кодирование длин серий

Посчитаем количество повторяющихся символов:

1. 12 символов «W»;
2. 1 символ «B»;
3. 12 символов «W»;
4. 3 символа «B»;
5. 24 символа «W»;
6. 1 символ «B»;
7. 14 символов «W».

Итого найдено 7 серий. Заменим серии на число повторов и сам повторяющийся символ:

12W12B24W1B14W

[https://ru.wikipedia.org/wiki/Кодирование\\_длин\\_серий](https://ru.wikipedia.org/wiki/Кодирование_длин_серий)

## Информационная энтропия



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

$$H = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{1}{3} \log_2 \frac{1}{3} - \frac{1}{3} \log_2 \frac{1}{3} = -\log_2 \frac{1}{3}$$

1.5849625007211563... бит

## Энтропия

Информационная и физическая энтропии имеют глубинную связь



Ludwig Eduard Boltzmann

<https://ru.wikipedia.org/wiki/H-теорема>

$$H^{\text{def}} = \int P(\ln P) d^3v = \langle \ln P \rangle$$

В термодинамике и кинетической теории, **H**-теорема, полученная **Больцманом** в 1872 году, описывает неубывания энтропии идеального газа в необратимых процессах, исходя из уравнения **Больцмана**.

$$S^{\text{def}} = -NkH,$$

## Передача информации

## Передача информации

- Комплексная область: **теоретические, практические**, физические аспекты
- Вопросы сжатия данных
- Вопросы надежности
- Вопросы шифрования и защиты данных (особенно в медицине и биологии)



[https://en.wikipedia.org/wiki/The\\_Imitation\\_Game](https://en.wikipedia.org/wiki/The_Imitation_Game)

## Передача информации

### Связь частоты сигнала и пропускной способности

**Владимир Александрович Котельников**

**Harry Nyquist**

1908 - 2005      1889 - 1976

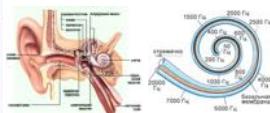
**Теорема Котельникова-(Найквиста-Шеннона)**

«любая 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

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

1888 – 1970

44.1 кГц – частота дискретизации при записи звука





## Передача информации

### Связь частоты сигнала и пропускной способности

**Ralph Hartley**

1888 – 1970

**Теорема Шеннона-Хартли**

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

где

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

$B$  – полоса пропускания канала, Гц;

$S$  – полная мощность сигнала над полосой пропускания, Вт или ВР;

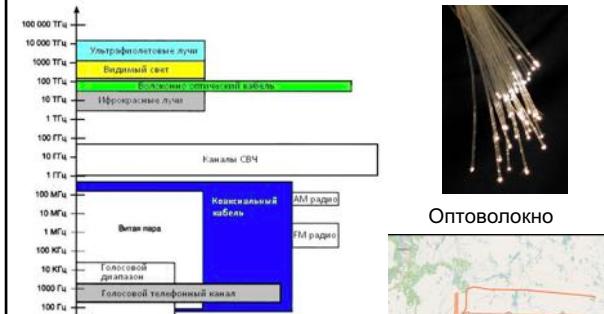
$N$  – полная шумовая мощность над полосой пропускания, Вт или ВР;

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

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

## Передача информации

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



Оптоволокно

Антенны КНЧ

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

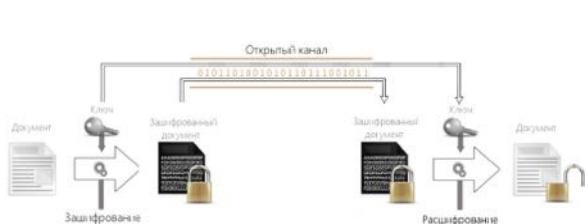
## Каналы связи

Карта подводных кабелей



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

## Шифрование информации



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

## Криптосистемы с открытым ключом

**Alice**

Large Random Number  
Key Generation Program  
Public Private

**Bob**

Hello Alice! Encrypt Alice's public key 6EB69570 0B892CE4  
Hello Alice! Decrypt Alice's private key

Необратимая Хэш функция

```
mbptb:~ alexsha$ md5 -s 'Hello world!!!'
MD5 ("Hello world!!!") = 87ee732d831690f45b8606b1547bd09e
```

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

## Хранение информации

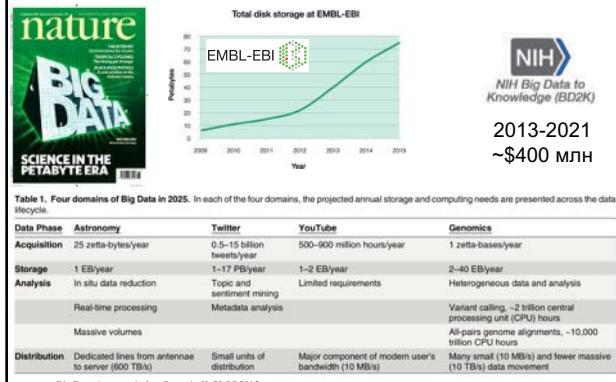
### Хранение информации



## Хранение информации



### Большие данные в биомедицине



## 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.

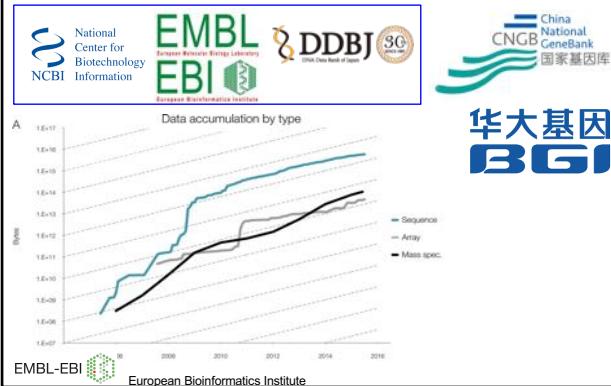
## Данные секвенирования, пример Геномы раковых опухолей



Международный проект, данные распределены по миру

File ID	Owner	Repository	Project	Study	Data Type	Sample	Format	Size	URL
■ PRJEB13021792	PCAWG - London, PCAWG - Berlin, Wellcome, Collaboratory - Toronto, USA - Houston	PRCA.EU	Aligner	WGS	BAM	128.77 GB			
■ PRJEB13021792	PCAWG - London, PCAWG - Berlin, Wellcome, Collaboratory - Toronto, USA - Houston	PRCA.EU	Aligner	WGS	BAM	107.27 GB			
■ PRJEB13021801	PCAWG - London, PCAWG - Berlin, Wellcome, Collaboratory - Toronto, USA - Houston	PRCA.EU	Aligner	WGS	BAM	114.65 GB			
■ PRJEB13021801	PCAWG - London, PCAWG - Berlin, Wellcome, Collaboratory - Toronto, USA - Houston	PRCA.EU	Aligner	WGS	BAM	101.45 GB			
■ PRJEB13021801	PCAWG - Chicago (USA) WGS - Chicago	DBIC.LB	Aligner	WGS	BAM	202.06 GB			
■ PRJEB13021801	PCAWG - Chicago (USA) WGS - Chicago	DBIC.US	Aligner	WGS	BAM	185.77 GB			

## Централизованные репозитории омиксных данных



## Genomes en masse

nature International weekly journal of science

Home News & Comment Research Careers & Jobs Current Issue Archive Audio & Video For Authors

AstroZeneca launches project to sequence 2 million genomes

Drug company aims to pool genomic and medical data in hunt for rare genetic sequences associated with disease.

Heidi Ledford

22 April 2016

CHINESE MILLIONOME DATABASE

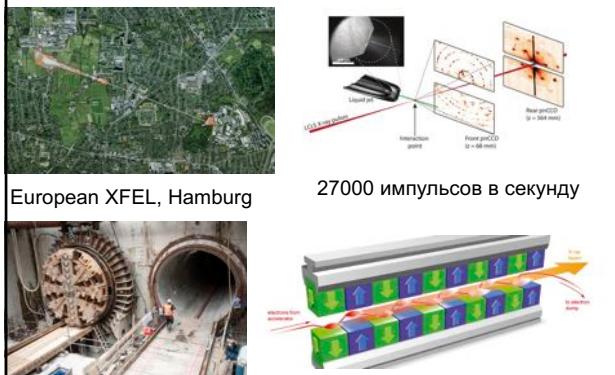
SEARCH FOR YOUR INTEREST

Public Access Coming Soon. Please Hold for Trial

华大基因 BGI

China National GeneBank 国家基因库

## Структурная биология и моделирование



## Передача больших данных

Выделенные научно-образовательные сети 100Gbit/s



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

Клиенты				Товары	
Id_кл	Фамилия	Имя	Отчество	Id_товар	Название
15	Иванов	Иван	Иванович	1	Шкаф
16	Петров	Петр	Петрович	2	Стул
17	Николаев	Николай	Николаевич	3	Стол

Заказ					
Id_зак	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	Соответствие требованиям ACID
		Атомарность	
		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.



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")

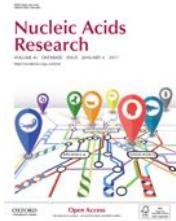
## Data provenance [edit]

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

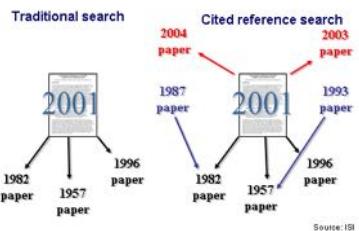
EMTREE thesaurus includes almost twice as many terms as PubMed

<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

**научная электронная библиотека eLIBRARY.RU**

**НОВОСТИ И ОБЪЯВЛЕНИЯ**

- 26.01 Публикации пресс-конференции SCIENCE ONLINE 2018
- 18.01 Участие в научно-практической конференции "Использование РИНЦ и Scopus для оценки научных достижений" 27 февраля
- 12.01 Публикации научных статей в журналах из списка РИНЦ на конференции "Международный научно-практический семинар по проблемам развития научной концепции"
- 05.01 Публикации конференции "Международный научно-практический семинар по проблемам развития научной концепции"
- 0 Другие новости

**ЧИТАЕТЕ СОСТОЯНИЕ**

Число издаваемых журналов	62015
из них российских журналов:	1530
из них выдающихся в науке:	13115
Число издаваемых книг:	13109
Число издаваемых томов:	5206
Число издаваемых сокращенных томов:	14772
из них в открытом доступе:	1537

eLibrary.ru/РИНЦ

**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.opendoar.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**

**LENS.ORG**  
Free, Open and Private Innovation Cartography

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

**Google Patents**

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 (ЕБИ), японского банка данных ДНК (DDBJ) в Центре Информационной Биологии (CIB) и GenBank в Национальном Центре Биотехнологической Информации (NCBI).



55

## База данных GenBank. Структура файла

```

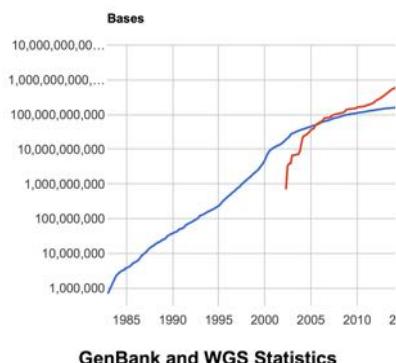
FT          /translation="MGGQPGNGSAFLAPNLNPNSHADPHDVTQQRDEVWVVMGIVMSLIVL
FT          AIVFGNVLVITAIKARFERTQVNTTNTFITSLLACADLVMLGVAUVPFGAAHILMKRMWTGFNE
FT          WCECFWTSIDTCLVTCASITELCIVIAVDRYFAITSPTFKYQSLTCKNKAIRVIIILMVWIVSGL
FT          TSFLPIQMHWYRATHQEAINCYANETCDFFTTQNYAIAISSVFSVFLVIMVQFVYSRV
FT          FQEAKRLQLKIDSEGRFHVNQNSQEVEQDGRTGHGLRRSSKFCLEKEHAKLTGLIIMGTR
FT          FTLCWLPEFVNVHVIQDNLRKEVYVILLNWITCVNSGNPFLYCRSPDPRIARPEL
FT          CLRSSKLAYKGNGYSSNGNTQEWSGYHVEQEKENKLCCEDLPLGTEDFVGHQGTVPSDN
FT          DSQGRNCRSTNDLSL<
FT          variation    46
FT          /gene="ADRB2<
FT          /replace="a<
FT          /note="Arg16 to Gly polymorphism<
XX

```

5

## База данных GenBank. Структура файла

58



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

Genbank – Is an archive! Contains everything.

## Nicotiana tabacum chloroplast JLA region, sequence 2

FASTA Graphic

FEATURES Location/Qualifiers  
source 1..124  
/organism="Nicotiana tabacum"  
/organelle="plastidchloroplast"  
/mol\_type="genomic DNA"  
/isolate="Cuban cañabo cigar, gift from President Fidel Castro"  
/db\_xref="taxon:4027"

RefSeq – is a reference sequence database!

RefSeq – is a reference sequence database!

Ifи 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="MGQPQNGSAFLAPNGSHAPDHVTQQRDEVWVVMGIVMSLIVL
FT      AIVFGNVLVITAIKFERLQTNTVYFITSACADLVMLGAIVPFGAAHILMKMWTGFNF
FT      WCEFWTSIDVLCVTVTASIELTLCVIAVDRYFAITSPFKYQSLLTKNARVIIIMWWIVSGL
FT      TSPLPIQMHWYRATHQEAINCYANETCCDFFTNQAYAASSIVSFYVPLVIMWFVYSRV
FT      FQEAKRLQLQKIDKSEGRFHVNLSQVEQDGRTGHGLRRSKFCLEHEKALKTLGIIMG
FT      FTLCNLPPFIVNVHVIQDNLIRKEVYIILNNWIGYVNNSGFNPILYCRSPDFRIAFQELL
FT      CLRSSSLKAYGNGYSSNNTGEQSGYHVEQEKENKLLEDLPGTEDFVGHQGTVPDSNI
FT      DSQGRNCSTNDSSL<
FT variation 46
FT      /gene="ADRB2<
FT      /replace="a<
FT      /note="Arg16 to Gly polymorphism<
XX
...
Protein sequences are annotate within GB records63
```

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> ProClass, and iProLINK.

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



## 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

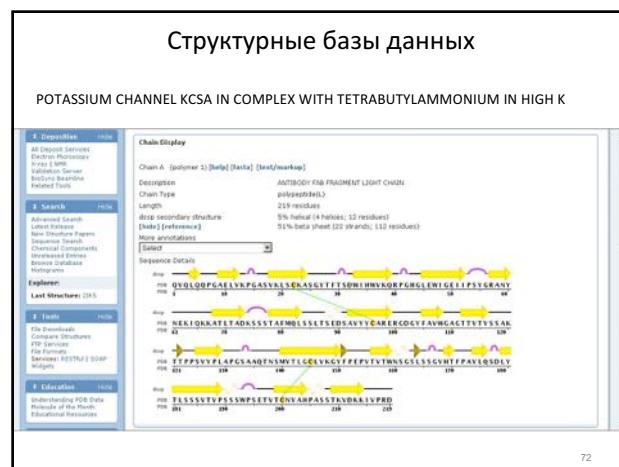
## Структурные базы данных

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



## Структурные базы данных

POTASSIUM CHANNEL KCSA IN COMPLEX WITH TETRABUTYLMONIUM IN HIGH K



## База данных PDB. Структура файла

```

HEADER IMMUNE SYSTEM/METAL TRANSPORT          15-AUG-08 2JK5
TITLE POTASSIUM CHANNEL KC8A IN COMPLEX WITH TETRAUBUTYLAMMONIUM
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,
...
EXPTPA X-RAY DIFFRACTION
AUTHOR M. J. LENNARZ, P. J. FOCIA, T. WAGNER, A. GROSS
REVDAT 1 17-NOV-09 2JK5 0
JRNL AUTH M. J. LENNARZ, P. J. FOCIA, T. WAGNER, A. GROSS
JRNL TITL STRUCTURE OF A POTASSIUM CHANNEL IN COMPLEX WITH TETRAUBUTYLAMMONIUM
JRNL TITL 2 SYMMETRICAL QUATERNARY AMMONIUM COMPOUNDS REVEALS
JRNL TITL 3 A NOVEL HYDROPHOBIC BINDING SITE
JRNL REF TO BE PUBLISHED
JRNL REFN
REMARK 2
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 сигналных 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

PubChem Compound search results for epinephrine and epinephrine hydrochloride. The interface shows chemical structures, IUPAC names, and various biological activity annotations like BioAssays, BioSystems, and BioActivity Analysis.

Compound	Count
Epinephrine, Adrenalin, adrenaline	39
Epinephrine hydrochloride	60
Protein 3D Structures	6
All Crystal Structure Of	1

## Базы данных химических соединений

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

The CAS homepage features a banner about the 2010 Nobel Prize in Chemistry, sections for SciFinder and STN, and links to various databases and resources.

## Базы данных химических соединений

The Reaxys search interface includes sections for 'Why choose Reaxys', 'How Reaxys works', 'Who we serve', 'Customer stories', and 'Learn & support'.

## Базы данных углеводов



For 2017:

7005 publications for 18924 compounds from 8859 organisms

The search interface includes sections for 'Database search' (Structures, Composition, Organisms, Publications, NMR signals) and 'Useful tools' (Predict NMR, Elucidate, Fragments, Cluster taxa, GT activities, Examples).

## Структурные базы данных

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

The LipidBank interface for Vitamin A shows its chemical structure, physical properties, and biological activity. It notes that the database no longer supports 7009 lipid structures.

## 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

NCBI Resources | How To | Sign in to NCBI

Genome | Help

## Genome

This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

**Using Genome**

- File
- Browse by Organism
- Download FTP
- Download FAQ
- Submit a genome

**Custom resources**

- Human Genome
- Micobes
- Orphanets
- Viruses

**Other Resources**

- Assembly
- BioMart
- BioSeme
- Mac View
- Genome Data Viewer

**Genome Tools**

- BLAST: the Human Genome
- Microbial Nucleotide BLAST

**Genome Annotation and Analysis**

- Eukaryotic Genome Annotation
- Prokaryotic Genome Annotation
- PASC (Panwise Sequence Conservation)

**External Resources**

- GOLD: Genomes Online Databases
- Bacteria Genomes at Sanger
- Ensembl

Ensembl BLAST/BLAT | BioMart | Tools | More | Search all species... | Login/Register

## Ensembl

Search: All species for Go

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

**What's New in Ensembl Release 91 (December 2017)**

- New Primate Species
- Update of cat assembly and genebuild to *Felis\_catus\_8.0*
- Mouse: update to Ensembl-Havana GENCODE gene set
- New and updated probe mapping data for primates
- Microarray Probe Mapping Update

[Full details](#) | [All web updates](#) by

**Browse a Genome**

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate *homo\_sapiens* *mmusculus* *mus musculus*

**Find a Data Display**

Not sure how to find the data visualisation you need? With our

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | View | Help | About Us | Request onsite workshops

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr1:11,502,837-11,267,747 164,911 bp. Enter position, gene symbol, refseq or search terms

### 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

NCBI

## dbSNP Short Genetic Variations

dbVar ClinVar GaP PubMed Nucleotide Protein

Search small variations in dbSNP or large structural variations in dbVar

Search Entrez dbSNP for Go

ANNOUNCEMENT

dbSNP and dbVar no longer accept submissions for non-human organism data. Please read more [here](#).

GENERAL RSS Feed Contact Us Organism Data dbSNP Homepage

Search by IDs on All Assemblies

NCBI Resources | How To | Sign in to NCBI

dbVar | Advanced | Search | Help

## dbVar

dbVar is NCBI's database of human genomic structural variation — insertions, deletions, duplications, inversions, mobile elements, and translocations

**Getting Started**

Overview of Structural Variation

**Accessing Data**

Structural Variation Data Hub

**Other NCBI Resources**

dbSNP

### 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

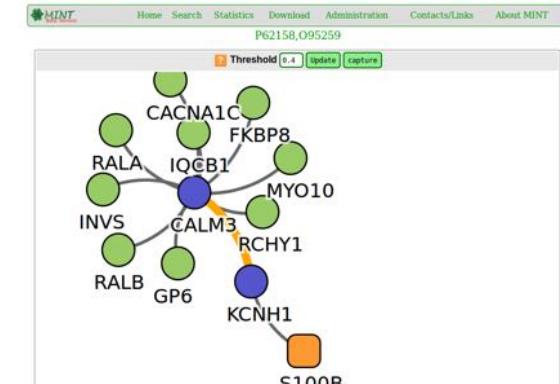
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы 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

99

### Базы данных химических соединений

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

### 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 клинических исследований и лекарств**
- Базы данных по экспрессии генов/гистологии
- Базы данных по таксономии

The screenshot shows the main navigation bar of the ClinicalTrials.gov website. It includes the NLM logo, the text "U.S. National Library of Medicine", the "ClinicalTrials.gov" logo, and a navigation menu with links for "Find Studies", "About Studies", "Submit Studies", "Resources", and "About Site".

The screenshot shows the homepage of the Pharmaprojects website. At the top, there's a navigation bar with links for 'Pharma intelligence', 'Informa', 'Report store', 'Recommendations', 'Product login', and a search bar. Below the navigation is a main banner featuring a large orange button labeled 'Request a free demo' and a 'Product login' button. The central part of the page has a large heading 'Pharmaprojects: track pharma R&D' with a subtext 'End-to-end tracking of the global pharma R&D pipeline from bench to patient, including company development trends, global development status, and therapeutic class status.' To the right, there are three circular icons with text: 'Curated by experts', 'Drug R&D landscape', and 'Historical trends'. At the bottom, there's a dark footer section with a grid of small icons.

**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

The screenshot shows the Expression Atlas homepage with a search bar at the top containing the query "Enter gene query... Examples: ASPN, REACT\_200004, ENSG0000021784, 'pig'". Below the search bar is a "Search" button. The main content area features a large blue "Expression Atlas" logo with a magnifying glass icon. A navigation bar below the logo includes links for "Home", "Release notes", "FAQs", "Download", "Help", and "About", along with a "Feedback" link.

The main search results are displayed under the heading "Strand-specific RNA-seq of nine chicken tissues". The results are organized by tissue: Brain, Colon, Heart, Liver, Lung, Oviduct, Testis, Uterus, and Yolk sac. Each tissue entry includes a small thumbnail image of a chicken, the gene name "KCNQ1", and a "Display levels" button. A horizontal progress bar at the bottom indicates the search results are "Showing 1 of 1 genes found".

**Базы данных экспрессии генов**

NCBI Resources ▾ How To ▾ Sign in to NCBI

GEO Home Documentation ▾ Query & Browse ▾ Email GEO

## Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Keyword or GEO Accession  Search

Getting Started	Tools	Browse Content
Overview	Search for Studies at GEO DataSets	Repository Browser
FAQ	Search for Gene Expression at GEO Profiles	DataSets: 4348
About GEO DataSets	Search GEO Documentation	Series: 94968
About GEO Profiles		Platforms: 18150

The image shows the homepage of The Human Protein Atlas. At the top center is the title "THE HUMAN PROTEIN ATLAS" in large, bold, grey letters. To the right of the title is a colorful logo consisting of several overlapping circles in shades of pink, orange, and yellow. Below the title is a navigation bar with three items: "MENU", "HELP", and "NEWS". Underneath the navigation bar is a search bar with the placeholder text "SEARCH". Below the search bar is a suggestion "e.g. RBM3, insulin, CD36". To the right of the search bar are two buttons: "Search" and "Fields >". Below the search area are three large, rounded rectangular boxes representing different atlases. The first box on the left is labeled "TISSUE ATLAS" at the bottom and contains a grayscale image of tissue sections. The middle box is labeled "CELL ATLAS" at the bottom and contains a color-coded image of cells. The third box on the right is labeled "PATHOLOGY ATLAS" at the bottom and contains a grayscale image of tissue sections.

### 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

### Видовые базы данных

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