Орг. вопросы

- Сайт http://intbio.org/bioinf2019-2020

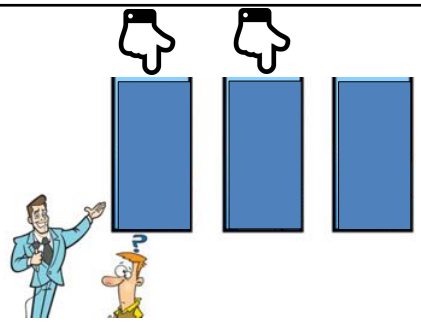
ВВЕДЕНИЕ В БИОИНФОРМАТИКУ
Лекции №2-3

Понятие информации. Способы измерения информации. ДНК как цифровой носитель информации. Системы счисления. Теория информации. Информационная энтропия. Сжатие информации. Теорема Котельникова. Теорема Шеннона-Хартли. Шифрование информации. Хранение информации. Источники больших данные в биомедицине. Проблемы передачи больших данных.

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

Сайт курса: http://intbio.org/bioinf2019-2020

13 сентября 2019



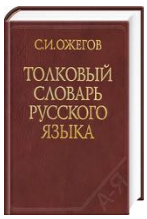
- В результате игрок получает **0.67 бита информации**
- Если бы ведущий открыл дверь в самом начале – только **0.58 бита информации**

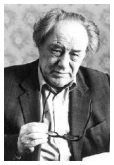
https://en.wikipedia.org/wiki/Monty_Hall_problem

Природа информации

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/irr/2007/number_3/number_3_6/number_3_6571/

Информация

iPhone 7 Plus Серебристый Ёмкость

Теперь выберите ёмкость.

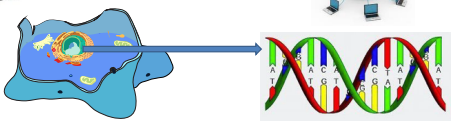
32 ГБ[1] 52 990.00 руб. Доставка: на складе

128 ГБ[1] 60 990.00 руб. Доставка: на складе

Живые системы и цифровые технологии

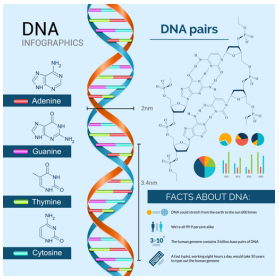
Аналоговые технологии

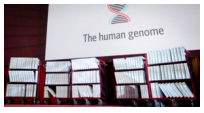
Цифровые технологии



ДНК – цифровой код

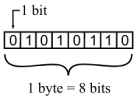
Генетический код





Геном человека
3 млрд нуклеотидов
262 тыс страниц
3 Гбайта

Геном кишечной палочки
4.6 млн нуклеотидов
400 страниц
4.6 Мбайт



Двоичный код

A T G C

Четвертичный код

Код может быть **позиционный** или **непозиционный**

Позиционный код активно используется в **системах счисления**

Соответствие цифр некоторых систем счисления

Основание системы счисления	2	8	10	16
Зелёные ячейки — цифры системы счисления, жёлтые - числа.	0	0	0	0
	1	1	1	1
	10	2	2	2
	11	3	3	3
	100	4	4	4
	101	5	5	5
	110	6	6	6
	111	7	7	7
	1000	10	8	8
	1001	11	9	9
	1010	12	10	A
	1011	13	11	B
	1100	14	12	C
	1101	15	13	D
	1110	16	14	E
	1111	17	15	F

Целое число без знака x в b-ичной системе счисления представляется в виде конечной линейной комбинации степеней числа b[1]:

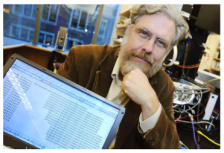
$x = \sum_{k=0}^{n-1} a_k b^k$, где a_k — это целые числа, называемые **цифрами**, удовлетворяющие неравенству $0 \leq a_k \leq b - 1$.

TIME

SCIENCE

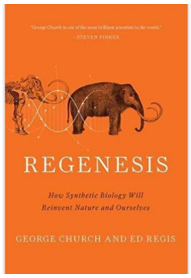
The First Book To Be Encoded in DNA

Two Harvard scientists have produced 70 billion copies of a book in DNA code --and it's smaller than the size of your thumbnail.

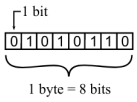


Lisa Poole / AP FILE

In his lab at the Harvard Medical School in Boston, George Church,



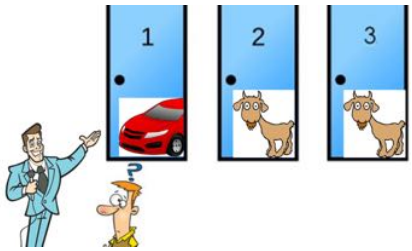
Измерение информации



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

Decimal	Binary
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001

Вероятность того, что машина за этой дверью: 1/3 1/3 1/3

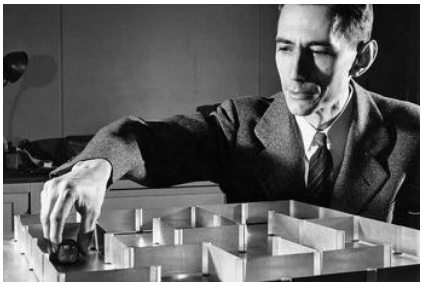


- Сколько информации нужно, чтобы закодировать положении машины?

1.5849625007211563... бит

Теория информации

"the father of information theory"



Claude Elwood Shannon
(April 30, 1916 – February 24, 2001)

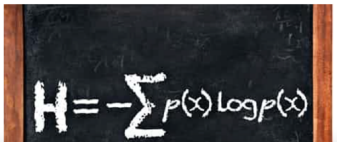
https://youtu.be/z2Whj_nL-x8

Теория информации

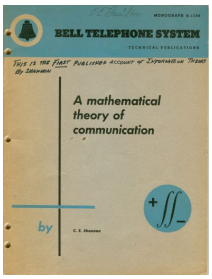
Science A short history of equations

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



The Guardian



Введена мера информации(!)

кг, метр, секунда + БИТ

1948

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

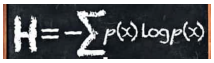
Информационная энтропия

Клод Шеннон предположил, что прирост информации равен утраченной неопределённости, и задал требования к её измерению:

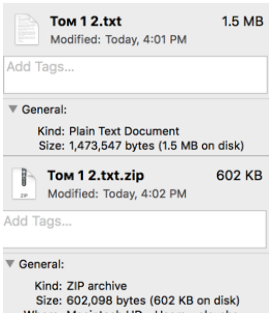
1.мера должна быть непрерывной; то есть изменение значения величины вероятности на малую величину должно вызывать малое результирующее изменение функции;

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.



Сжатие информации



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

Сжатие информации

Кодирование длин серий

`WWWWWWWWWWBWWWWWWWWWWWWBBBWWWWWWWWWWWWWWWWWWWWWWWWBWWWWWWWWWWWWWW`

Посчитаем количество повторяющихся символов:

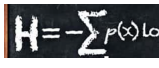
1. 12 символов «W»;
2. 1 символ «B»;
3. 12 символов «W»;
4. 3 символа «B»;
5. 24 символа «W»;
6. 1 символ «B»;
7. 14 символов «W».

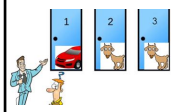
Итого найдено 7 серий. Заменим серии на число повторов и сам повторяющийся символ:

`12W1B12W3B24W1B14W`

https://ru.wikipedia.org/wiki/Кодирование_длин_серий

Информационная энтропия





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

1.5849625007211563... бит

Энтропия

Информационная и физическая энтропии имеют глубинную связь

Ludwig Eduard Boltzmann

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

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

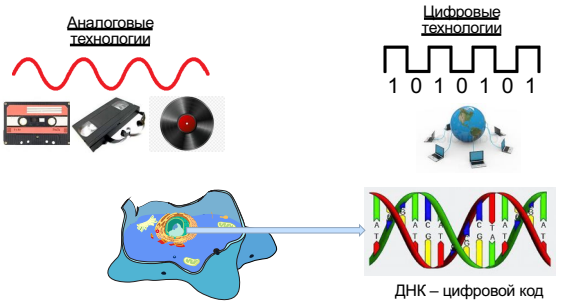
$$S \stackrel{\text{def}}{=} -NkH,$$

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

Живые системы и цифровые технологии - аналогии

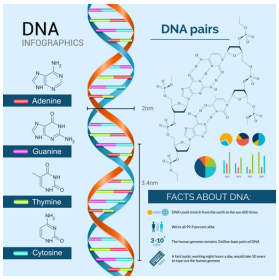
От программирования компьютеров к программированию живых систем

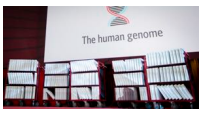
Живые системы и цифровые технологии



ДНК – цифровой код

Генетический код

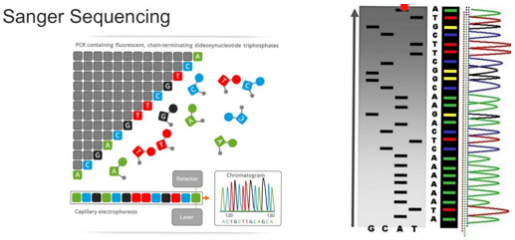




Геном человека
3 млрд нуклеотидов
262 тыс страниц
3 Гбайта

Геном кишечной палочки
4.6 млн нуклеотидов
400 страниц
4.6 Мбайт

Секвенирование ДНК (метод Сэнгера)

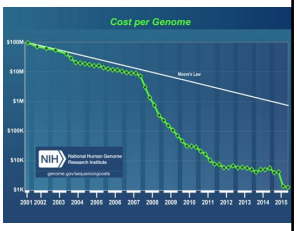


Sanger sequencing uses ddNTPs (dideoxynucleotide triphosphates) which do not have a free 3' OH mixed in with dNTPs
Whenever the DNA polymerase incorporates a ddNTP it won't be able to add any other nucleotides
Then gel electrophoresis is used to separate the DNAvv

https://www.youtube.com/watch?v=593zWZNwbJI

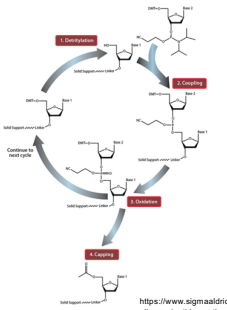
25

Технологии чтения и записи ДНК



26

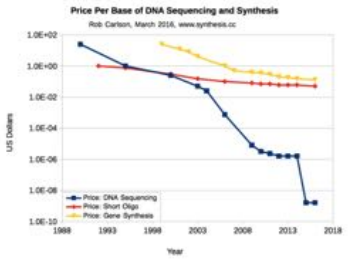
Синтез ДНК

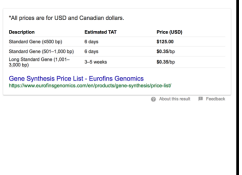


https://www.sigmaaldrich.com/technical-documents/articles/biology/dna-oligonucleotide-synthesis.html

27

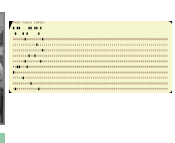
Удешевление синтеза ДНК





28

От программирования компьютеров к программированию жизни



Ядро ОС Linux
~100 Мб

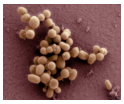
????

Аналогии компьютеров и живых клеток

- Software boots into hardware
- Genetic code boots into cells

Craig Venter

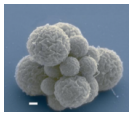
2010 *Mycoplasma mycoides* JCVI-syn1.0



Human	~20,000–25,000
Escherichia coli (K12 strain)	~4500
Syn 1.0	901
*Mycoplasma genitalium**	525
Syn 3.0	473

Первый искусственный геном (~$40 млн)

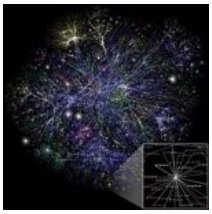
2016 *Syn3.0*



Минимальный геном

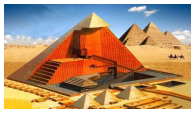
От программирования компьютеров к программированию жизни

~100 млрд. (10^11) нейронов
Каждый нейрон имеет
~7000 синаптических связей



23 млрд. устройств
подключенных к интернету

Сложные инженерные системы



~2500 лет
д.н.э.

Boeing 747
6 млн частей

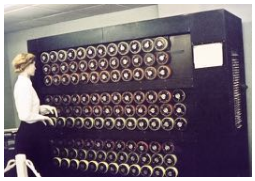
Boeing 777
Первый самолет спроектированный полностью на компьютере

Intel Xeon Phi CPU
8*10^9
транзисторов

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

Аналоговый сигнал

Цифровой сигнал

1 0 1 0 1 0 1

Как конвертировать?

Передача информации

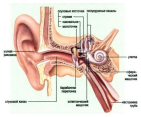
Связь частоты сигнала и пропускной способности

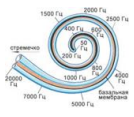
Владимир Александрович Котельников

1908 - 2005

Harry Nyquist

1889 – 1976





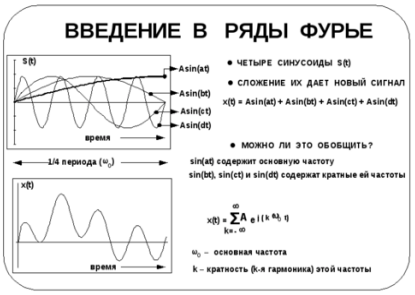
Теорема Котельникова-(Найквиста-Шенона)

«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

ВВЕДЕНИЕ В РЯДЫ ФУРЬЕ



- ЧЕТЫРЕ СИНУСОИДЫ S(t)
- СЛОЖЕНИЕ ИХ ДАЕТ НОВЫЙ СИГНАЛ

x(t) = Asin(at) + Asin(bt) + Asin(ct) + Asin(dt)

- МОЖНО ЛИ ЭТО ОБОБЩИТЬ?

sin(at) содержит основную частоту
sin(bt), sin(ct) и sin(dt) содержат кратные ей частоты

$x(t) = \sum_{k=-\infty}^{\infty} A_k e^{i(k\omega_0 t)}$

ω_0 – основная частота
k – кратность (k-я гармоника) этой частоты

Говоря шире, теорема Котельникова утверждает, что непрерывный сигнал $x(t)$ можно представить в виде интерполяционного ряда:

$$x(t) = \sum_{k=-\infty}^{\infty} x(k\Delta)\,\mathrm{sinc}\left[\frac{\pi}{\Delta}(t - k\Delta)\right],$$

где $\mathrm{sinc}(x) = \sin(x)/x$ – функция sinc. Интервал дискретизации удовлетворяет ограничениям $0 < \Delta \leqslant \frac{1}{2f_c}$. Мгновенные значения данного ряда есть дискретные отсчёты сигнала $x(k\Delta)$.

Передача информации

Связь частоты сигнала и пропускной способности

Ralph Hartley

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 В²;

N — полная шумовая мощность над полосой пропускания, Вт или В²;

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

Теорема Шеннона-Хартли

https://ru.wikipedia.org/wiki/Теорема_Шеннона_—_Хартли

Передача информации

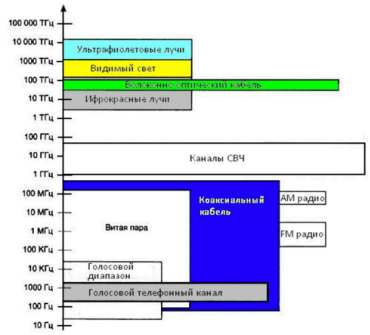


Рис. 1. Полосы пропускания линий связи и популярные частотные диапазоны

Оптоволокно

Антенны КНЧ

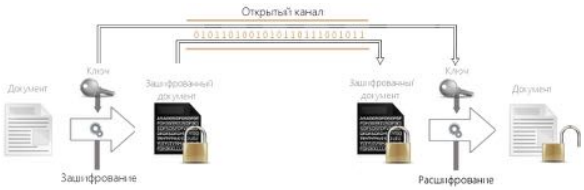
https://ru.wikipedia.org/wiki/Связь_с_подводными_лодками

Каналы связи

Карта подводных кабелей

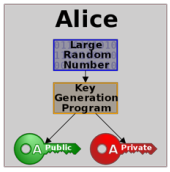
https://habrahabr.ru/company/rootwelt/blog/305634/

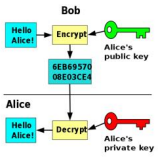
Шифрование информации



https://en.wikipedia.org/wiki/Public-key_cryptography

Криптосистемы с открытым ключом





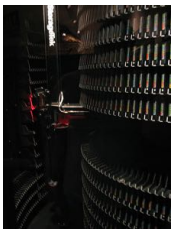
Необратимая Хэш функция

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

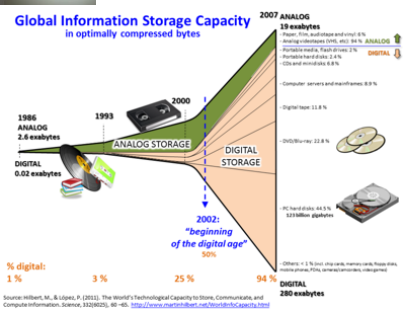
https://en.wikipedia.org/wiki/Public-key_cryptography

Хранение информации

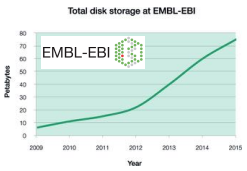
Хранение информации



Хранение информации



Большие данные в биомедицине



NIH Big Data to Knowledge (BD2K)

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 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

Big Data: Astronomical or Genomical?, PLOS 2015

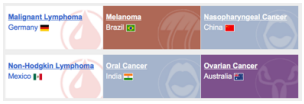
Источники больших данных в биомедицине

- 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

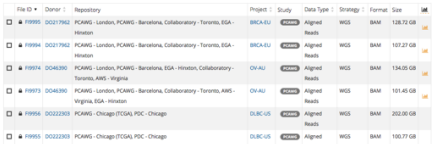
International Cancer Genome Consortium

Геном человека ~ 3.3 Gb
x100 секвенирование ~300Gb



25000 образцов опухолей

Международный проект, данные распределены по миру

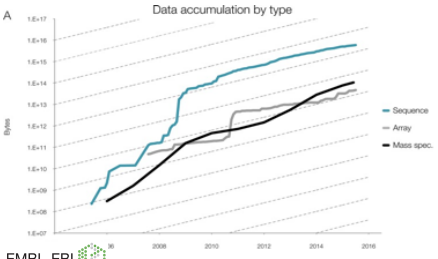


Централизованные репозитории омиксных данных

National Center for Biotechnology Information
NCBI
EMBL
EBI
European Bioinformatics Institute
DDBJ
DNA Data Bank of Japan

China National GeneBank
CNGB
国家基因库

华大基因
BGI



EMBL-EBI
European Bioinformatics Institute

Genomes en masse

برنامج الجينوم البشري السعودي
Saudi Human Genome Program

5 years ~ 100 000 genomes

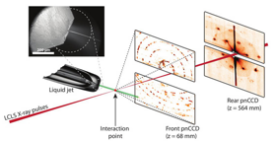
nature
AstraZeneca launches project to sequence 2 million genomes
Drug company aims to pool genomic and medical data in hunt for rare genetic sequences associated with disease.
Heidi Ledford
22 April 2016

CHINESE MILLIONOME DATABASE
SEARCH FOR YOUR INTEREST

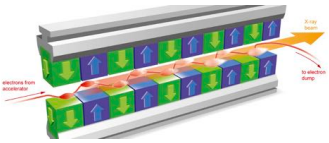
华大基因 BGI
China National GeneBank 国家基因库

Структурная биология и моделирование

European XFEL, Hamburg



27000 импульсов в секунду



Передача больших данных

Выделенные научно-образовательные сети 100Gbit/s

Figure 1: Asia Pacific Ring (APR)

INTERNET2

Программные решения

aspera
an IBM® company

globus toolkit

Базы данных

Базы данных

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

Larry Ellison

Larry Ellison in 2016

Oracle Corporation

Реляционные базы данных

Клиент

Id_кл	Фамилия	Имя	Отчество
15	Иванов	Иван	Иванович
16	Петров	Петр	Петрович
17	Николаев	Николай	Николаевич

Товар

Id_тов	Название
1	Шкаф
2	Стул
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

Атомарность

Изолированность

Надежность

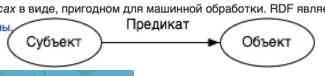
Соответствие требованиям ACID

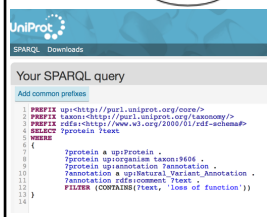
https://aws.amazon.com/ru/relational-database/

Resource Description Framework

Язык SPARQL

Resource Description Framework (RDF, «среда описания ресурса»[1]) — это разработанная консорциумом Всемирной паутины модель для представления данных, в особенности — метаданных[2]. 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/

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.[27][28] 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 государстом
- 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

National Center for Biotechnology Information
NCBI

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

EMBL-EBI

- 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

NATIONAL LIBRARY OF MEDICINE

FREE

Реферативные базы данных

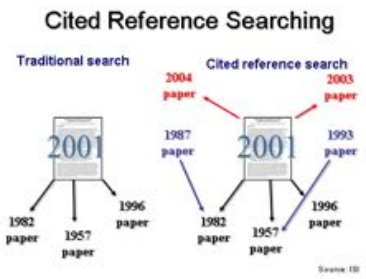
Clinical/Biomedical

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



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/

THOMSON REUTERS

PAID

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/

ELSEVIER

PAID

Реферативные базы данных

Общенаучные базы данных цитирований

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

http://www.lens.org/lens/biological_search – поиск ДНК последовательностей

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

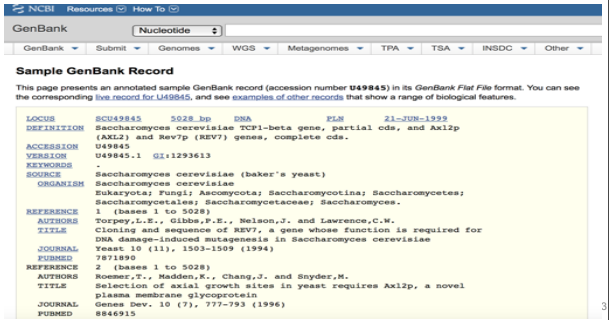
Нуклеотидные БД – это хранилища, принимающие данные от научного сообщества и представляющие их широкой общественности. Различные БД отличаются по источнику последовательностей, их надежности, широте аннотирования и т.д. В идеале БД должна содержать все известные последовательности.

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).

72

База данных GenBank

Открытая БД нуклеотидных последовательностей, учреждена в 1982 г.
2017: > 300 000 организмов, ~ 203 млн. последовательностей,
~ 240 млрд. пар оснований

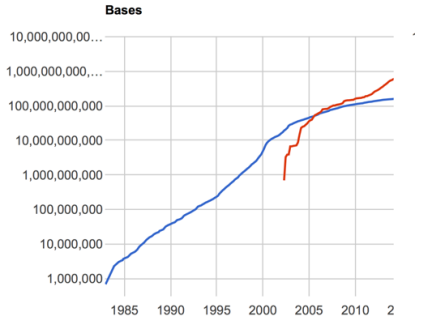


База данных GenBank. Структура файла

```
...
FT          /translation="MGQPGNGSAFLLAPNGSHAPDHDVTQQRDEVWVVGMGIVMSLIVL
FT          AIVFGNVLVITAIAKFERLQTVTNYFITSLACADLVMGLAVVPFGAAHILMKMWTFGNF
FT          WCEFWTSIDVLCVTASIETLCVIAVDRYFAITSPFKYQSLLTKNKARVIILMVWIVSGL
FT          TSFLPIQMHWYRATHQEAINCYANETCCDFFTNQAYAIASSIVSFYVPLVIMVFVYSRV
FT          FQEAKRQLQKIDKSEGRFHVQNLSQVEQDGRTGHGLRRSSKFCLKEHKALKTLGIIMGT
FT          FTLCWLPFFIVNIVHVIQDNLIRKEVYILLNWIGYVNSGFNPLIYCRSPDFRIAFQELL
FT          CLRRSSLKAYGNGYSSNGNTGEQSGYHVEQEKENKLLCEDLPGTEDFVGHQGTVPSDNI
FT          DSQGRNCSTNDSLL«
FT  variation       46
FT          /gene="ADRB2«
FT          /replace="a«
FT          /note="Arg16 to Gly polymorphism«
XX
...
```


База данных GenBank. Структура файла

```
...
SQ Sequence 1242 BP; 275 A; 331 C; 326 G; 310 T; 0 other;
atggggcaac ccgggaacgg cagcgccttc ttgctggcac ccaatggaag ccatgcgccg   60
gaccacgacg tcacgcagca aagggacgag gtgtgggtgg tgggcatggg catcgtcatg  120
tctctcatcg tcctggccat cgtgtttggc aatgtgctgg tcatcacagc cattgccaag  180
ttcgagcgtc tgcagacggt caccaactac ttcatcactt cactggcctg tgctgatctg  240
gtcatgggcc tggcagtggt gccctttggg gccgcccata ttcttatgaa aatgtggact  300
tttggcaact tctggtgcga gttttggact tccattgatg tgctgtgcgt cacggccagc  360
attgagaccc tgtgcgtgat cgcagtggat cgctactttg ccattacttc acctttcaag  420
taccagagcc tgctgaccaa gaataaggcc cgggtgatca ttctgatggt gtggattgtg  480
tcaggcctta cctccttctt gcccattcag atgcactggt accgggccac ccaccaggaa  540
gccatcaact gctatgccaa tgagacctgc tgtgacttct tcacgaacca agcctatgcc  600
attgcctctt ccatcgtgtc cttctacgtt cccctggtga tcatggtctt cgtctactcc  660
agggtctttc aggaggccaa aaggcagctc cagaagattg acaaatctga gggccgcttc  720
catgtccaga accttagcca ggtggagcag gatgggcgga cggggcatgg actccgcaga  780
tcttccaagt tctgcttgaa ggagcacaaa gccctcaaga cgttaggcat catcatgggc  840
actttcaccc tctgctggct gcccttcttc atcgttaaca ttgtgcatgt gatccaggat  900
aacctcatcc gtaaggaagt ttacatcctc ctaaattgga taggctatgt caattctggt  960
ttcaatcccc ttatctactg ccggagccca gatttcagga ttgccttcca ggagcttctg 1020
tgcctgcgca ggtcttcttt gaaggcctat gggaatggct actccagcaa cggcaacaca 1080
ggggagcaga gtggatatca cgtggaacag gagaaagaaa ataaactgct gtgtgaagac 1140
ctcccaggca cggaagactt tgtgggccat caaggtactg tgcctagcga taacattgat 1200
tcacaaggga ggaattgtag tacaaatgac tcactgctgt aa              1242
//
```

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

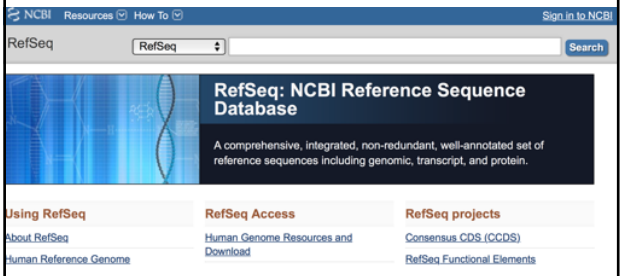
GenBank: Z71230.1

FASTA Graphics

```
FEATURES             Location/Qualifiers
     source          1..124
                     /organism="Nicotiana tabacum"
                     /organelle="plastid:chloroplast"
                     /mol_type="genomic DNA"
                     /isolate="Cuban cahibo cigar, gift from President Fidel
                     Castro"
                     /db_xref="taxon:4097"
```

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

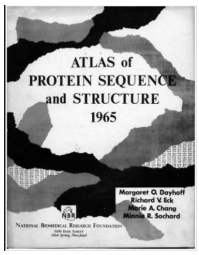
...

```
...
FT                   /translation="MGQPGNGSAFLLAPNGSHAPDHDVTQQRDEVWVVGMGIVMSLIVL
FT                   AIVFGNVLVITAIAKFERLQTVTNYFITSLACADLVMGLAVVPFGAAHILMKMWTFGNF
FT                   WCEFWTSIDVLCVTASIETLCVIAVDRYFAITSPFKYQSLLTKNKARVIILMVWIVSGL
FT                   TSFLPIQMHWYRATHQEAINCYANETCCDFFTNQAYAIASSIVSFYVPLVIMVFVYSRV
FT                   FQEAKRQLQKIDKSEGRFHVQNLSQVEQDGRTGHGLRRSSKFCLKEHKALKTLGIIMGT
FT                   FTLCWLPFFIVNIVHVIQDNLIRKEVYILLNWIGYVNSGFNPLIYCRSPDFRIAFQELL
FT                   CLRRSSLKAYGNGYSSNGNTGEQSGYHVEQEKENKLLCEDLPGTEDFVGHQGTVPSDNI
FT                   DSQGRNCSTNDSLL«
FT   variation       46
FT                   /gene="ADRB2«
FT                   /replace="a«
FT                   /note="Arg16 to Gly polymorphism«
XX
...
```

Protein sequences are annotate within GB records

80

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

Protein Centric Sequence Databases

PIR A UniProt CONSORTIUM MEMBER
Protein Information Resource

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,[8] iProClass, and iProLINK.

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

Белковые базы данных

UniProt

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

83

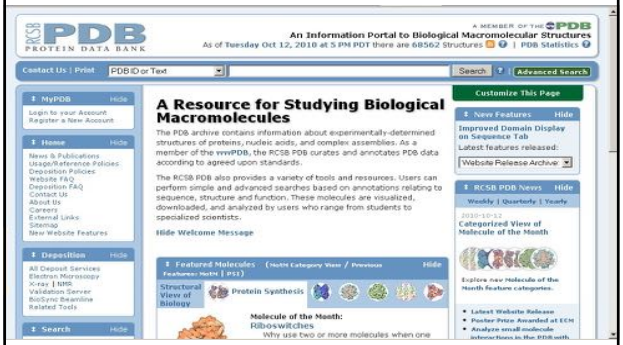
Белковые базы данных

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

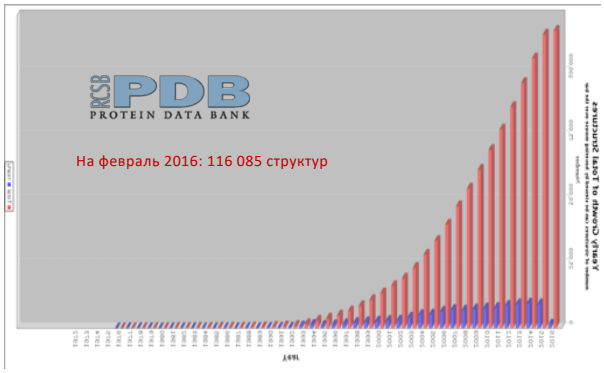
Структурные базы данных



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

86

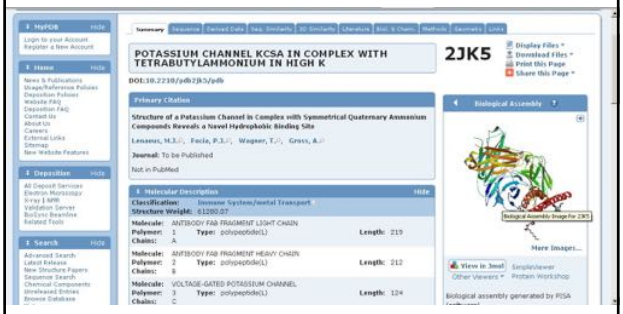
Структурные базы данных



87

Структурные базы данных

POTASSIUM CHANNEL KCSA IN COMPLEX WITH TETRABUTYLAMMONIUM IN HIGH K



88

Структурные базы данных

POTASSIUM CHANNEL KCSA IN COMPLEX WITH TETRABUTYLAMMONIUM IN HIGH K

89

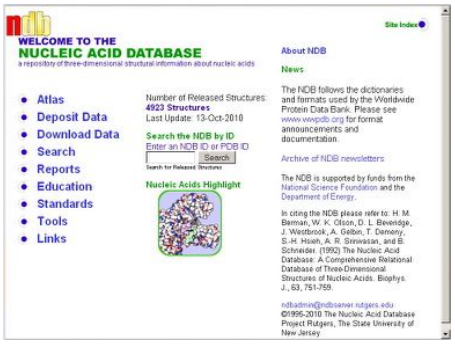
База данных PDB. Структура файла

```
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        REFN
REMARK    2
REMARK    2 RESOLUTION.    2.4  ANGSTROMS.
REMARK    3
REMARK    3 REFINEMENT.
REMARK    3   PROGRAM     : REFMAC 5.5.0051
...
```

90

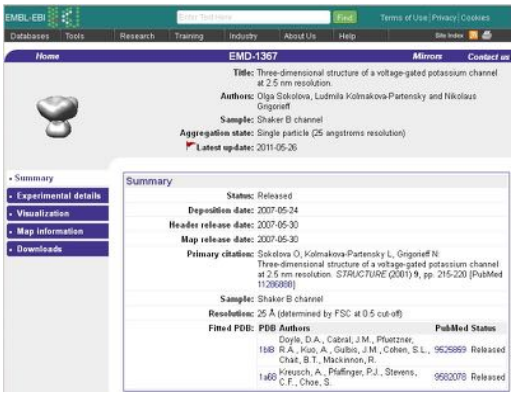
Структурные базы данных

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



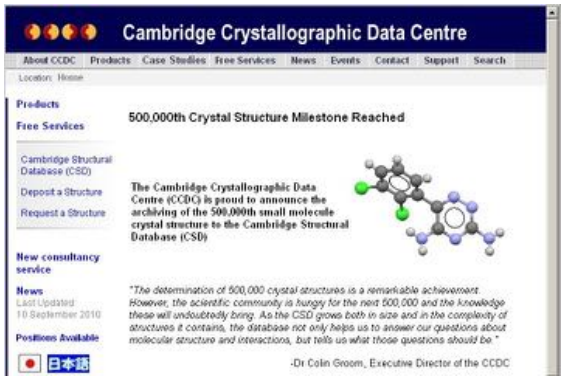
91

Структурные базы данных



92

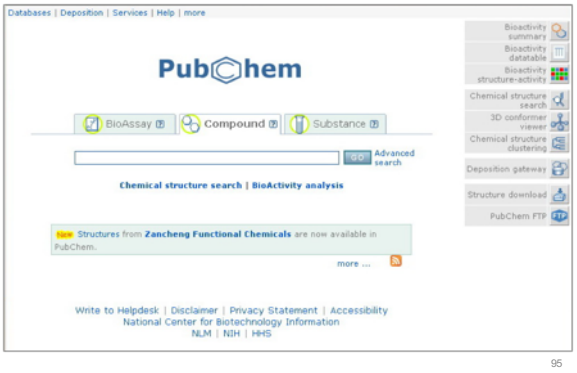
Структурные базы данных



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

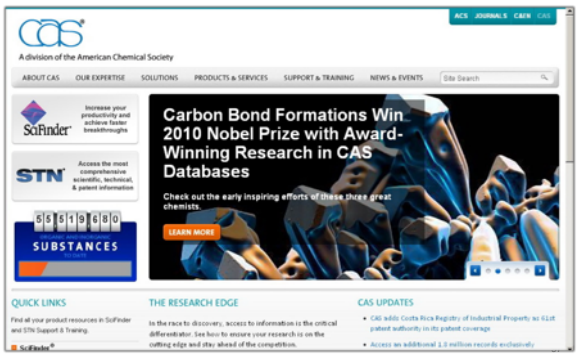
Базы данных химических соединений



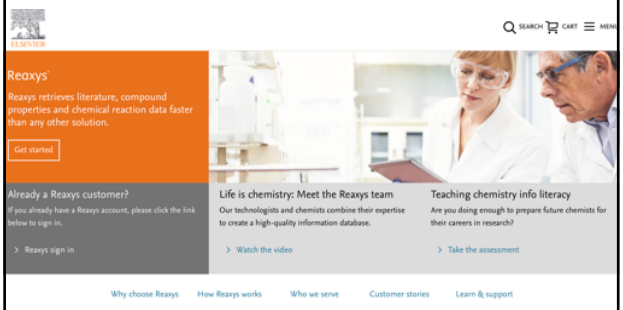
Базы данных химических соединений

Базы данных химических соединений

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



Базы данных химических соединений



98

Базы данных углеводов

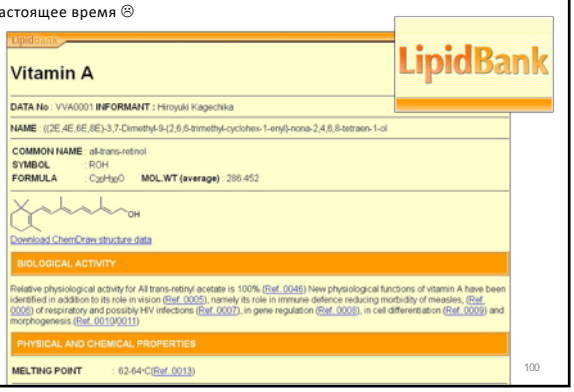
For 2017:

7005 publications for 18924 compounds from 8859 organisms

99

Структурные базы данных

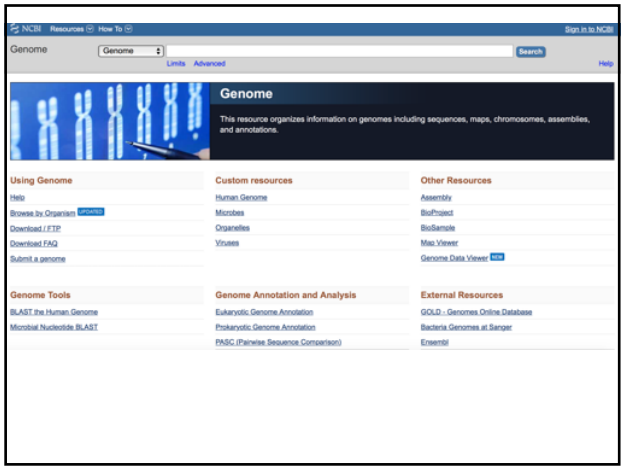
7009 структур липидов и сходных соединений – не поддерживается в настоящее время ☹

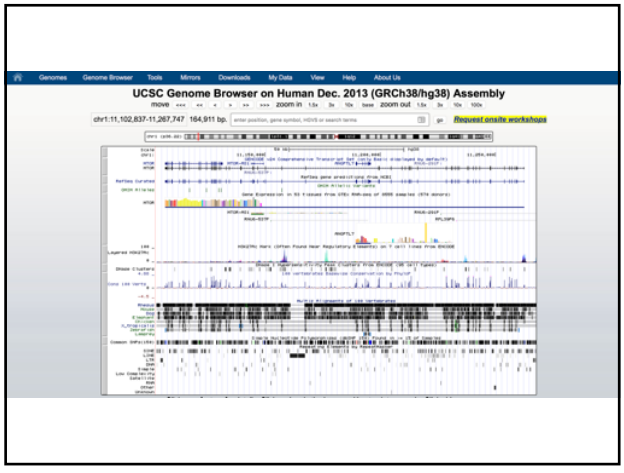


100

План

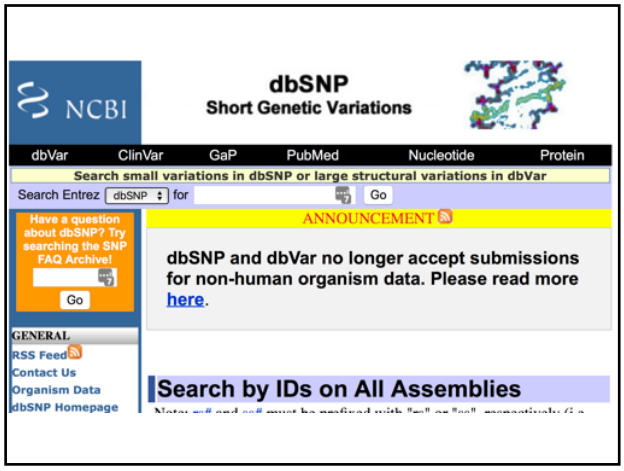
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- **Базы данных геномов и аннотаций**
- Базы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

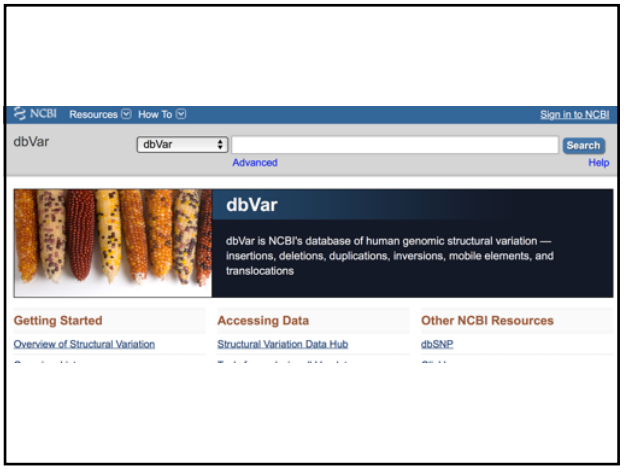


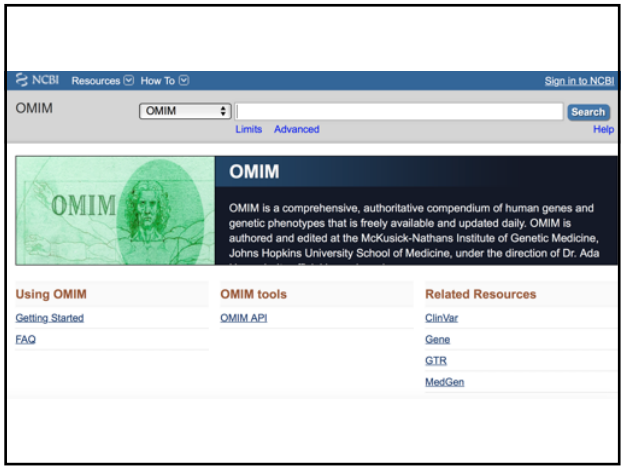


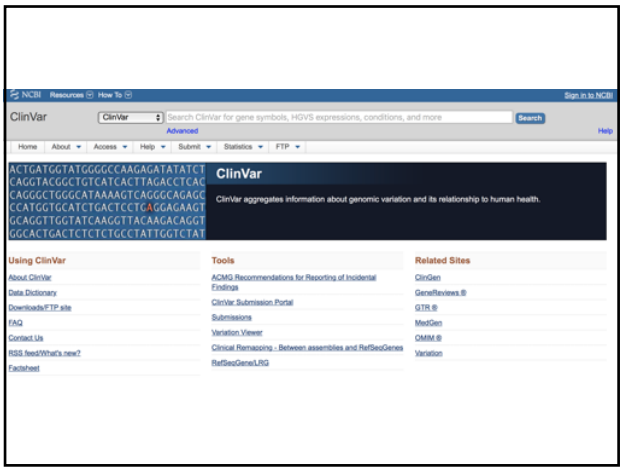
План

- 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



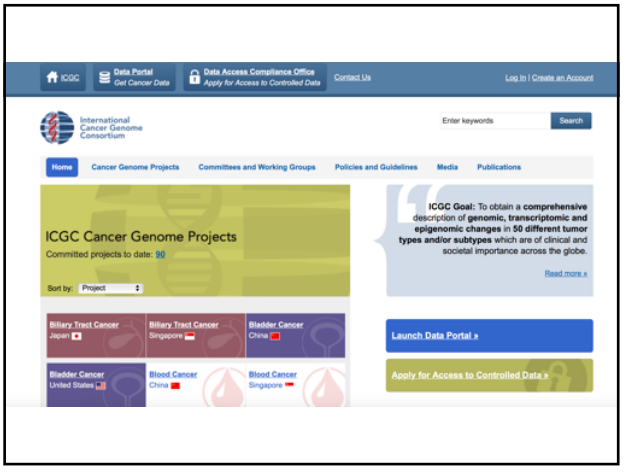






План

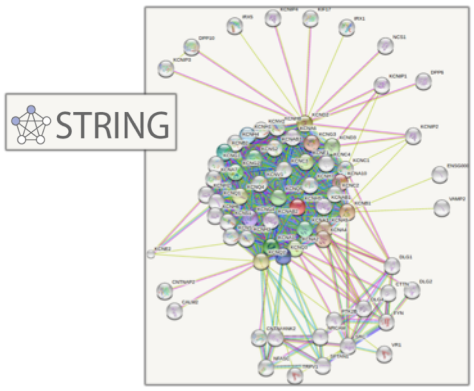
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы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

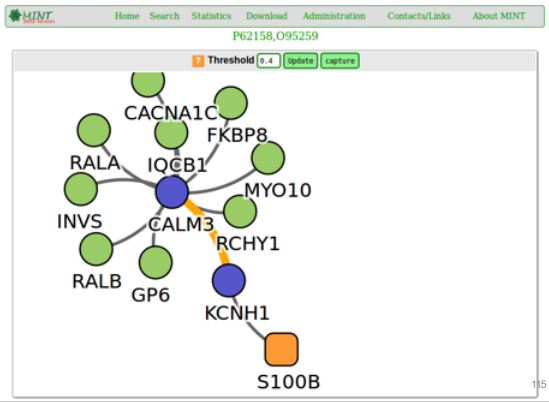
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы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



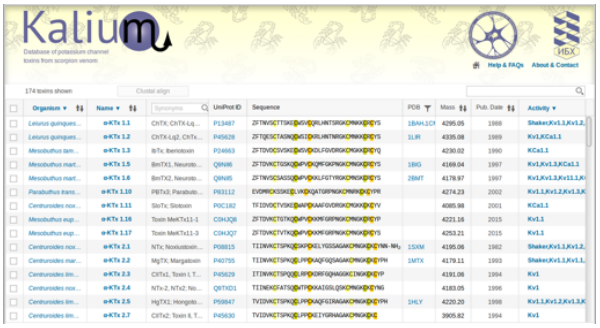
114

Базы данных взаимодействий



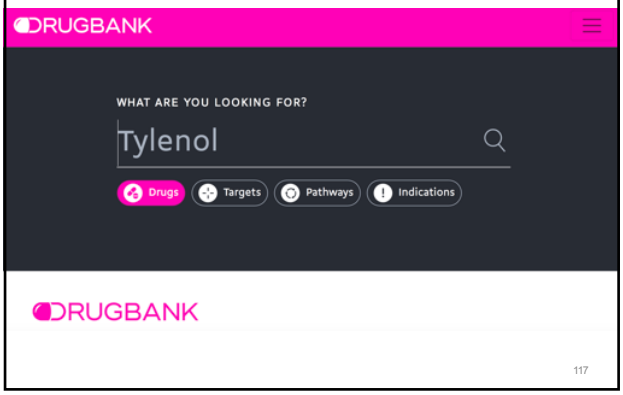
115

Базы данных взаимодействий



116

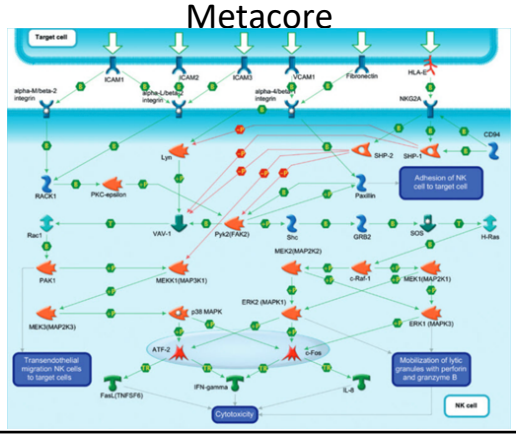
Базы данных химических соединений



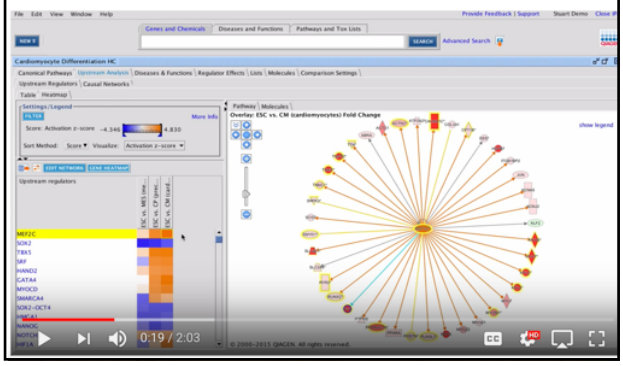
План

- 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

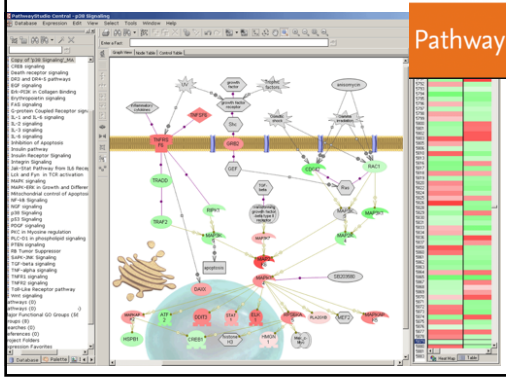


Ingenuity Pathway Analysis



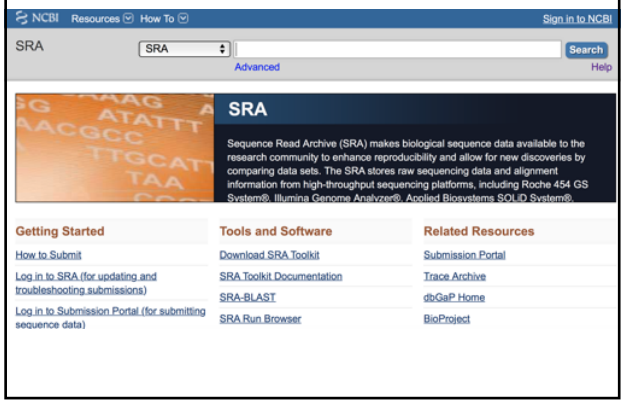
ELSEVIER

Pathway Studio®



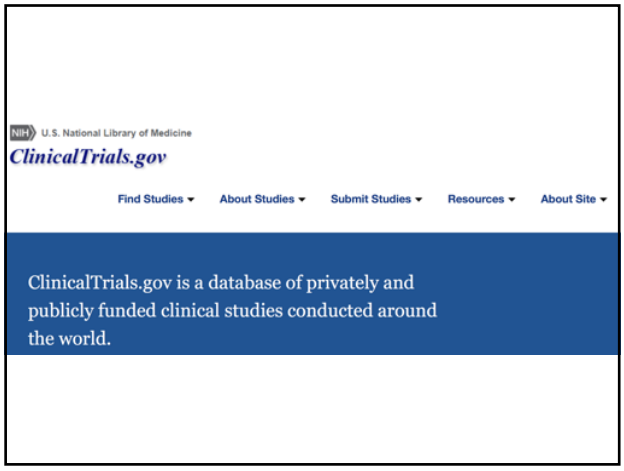
План

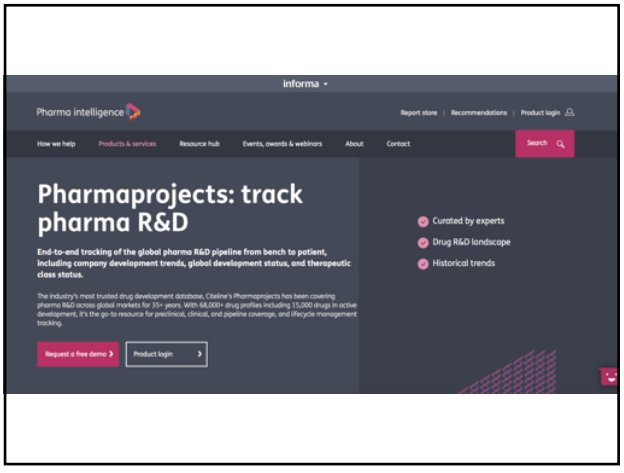
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы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

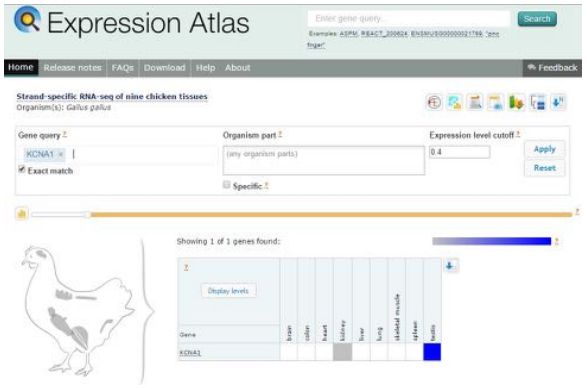




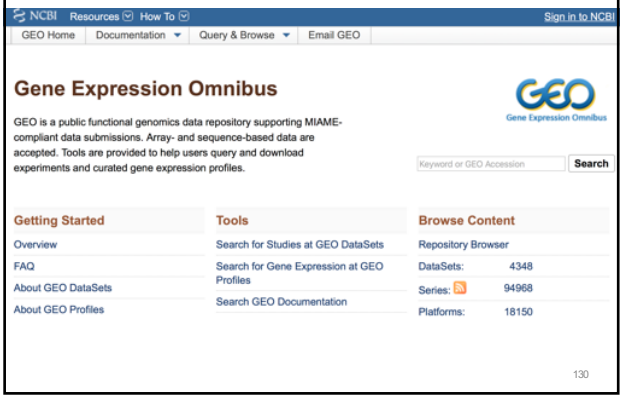
План

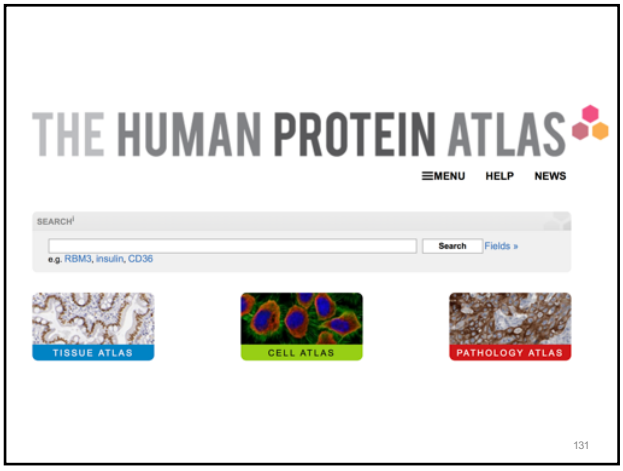
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы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

Базы данных экспрессии генов



Базы данных экспрессии генов





План

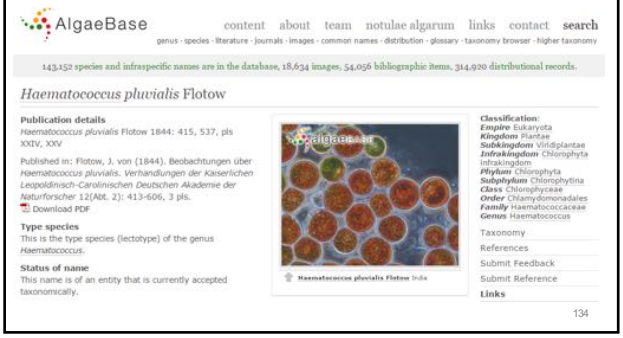
- Библиографические/реферативные базы данных литературных источников (статьи, тезисы, патенты, материалы конференций и т.д.)
- Базы данных последовательностей ДНК
- Базы данных последовательностей белков
- Базы данных 3D структур
- Базы данных хим. соединений
- Базы данных геномов и аннотаций
- Базы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



Видовые базы данных
https://plant.depo.msu.ru