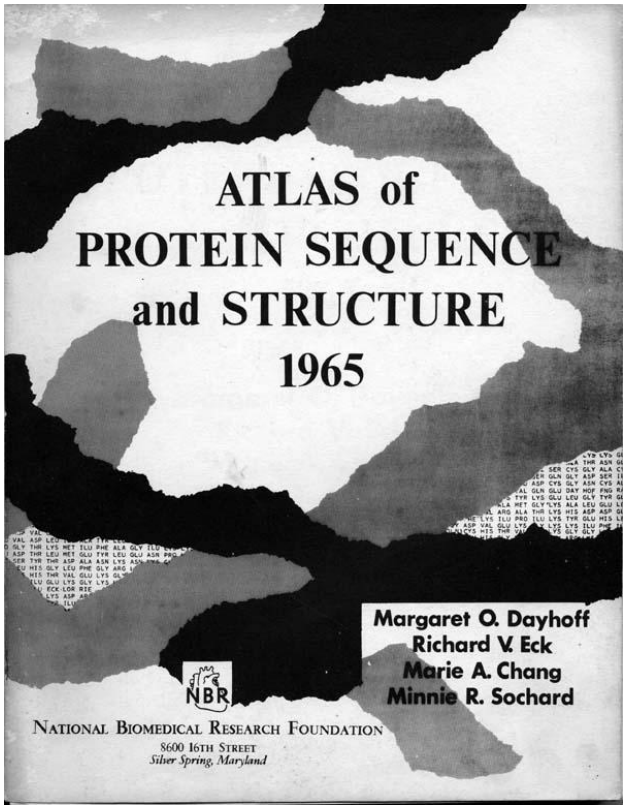
Профессиональные банки последовательностей – UniProt, SwissProt, TrEMBL

О.Занегина
9.02.2009

Первый “банк данных”

1965 -1978



Атлас белковых последовательностей и их структур

Первая версия атласа содержала описание **65 (!)** последовательностей белков

В конце 1970-х годов был изобретён относительно быстрый и дешёвый метод экспериментального определения последовательности оснований в ДНК

Организм

выделение

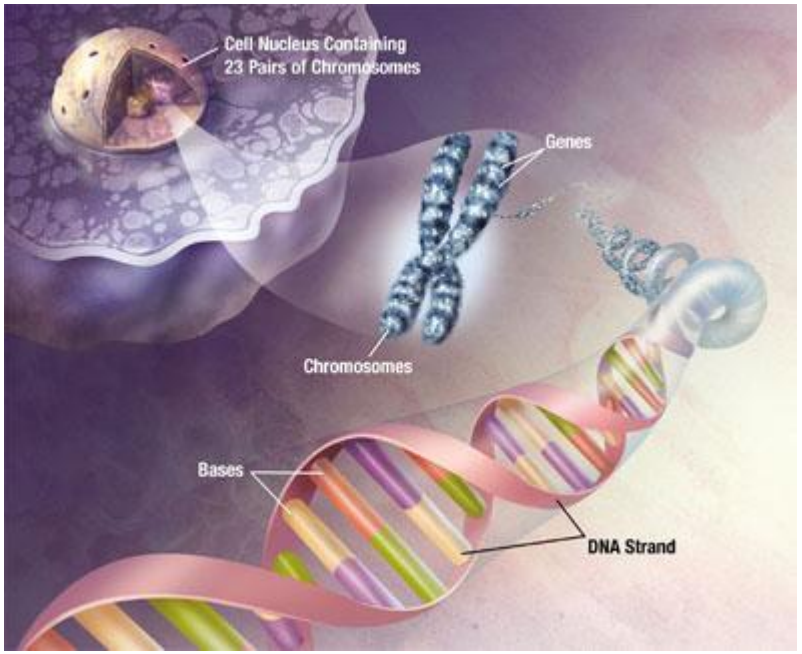
ДНК «в пробирке»

секвенирование

...TGCCACAAATCAC...

Последовательность

Молекулярная биология



Молекулярная биология
Molecular biology

+

Компьютер
Computer

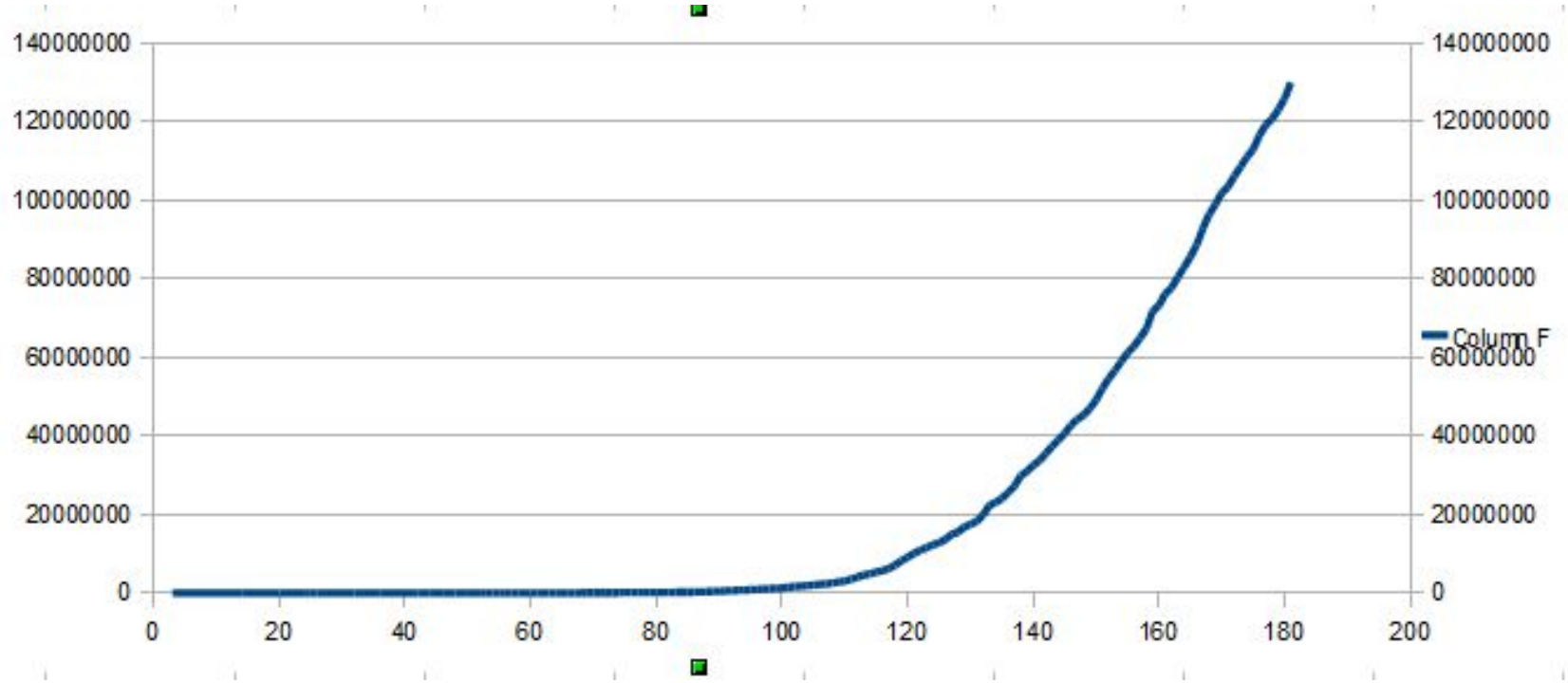
=

Биоинформатика

Computational Molecular biology

Для хранения все возрастающей информации о последовательностях ДНК в 1982 году был основан GenBank

NCBI GenBank — хранилище последовательностей нуклеиновых кислот
в виде компьютерных файлов



Банки данных

- Архивные

  (примеры: RCSB PDB PROTEIN DATA BANK, NCBI GenBank

  за содержание каждой записи отвечает её автор-экспериментатор

- Курируемые

  за содержание записей отвечают специальные люди — кураторы

- Автоматические

  записи генерируются компьютерными программами

Банки структурной биологической информации

NCBI GenBank

DDBJ DNA Data Bank of Japan

EMBL NUCLEOTIDE SEQUENCE DATABASE

Архивные базы последовательностей НК

RefSeq genomes·transcripts·proteins

Автоматическая база **различных** последовательностей НК

TrEMBL TRANSLATED EMBL

Автоматическая база предсказаний последовательностей **белков**

swissprot

Курируемая база последовательностей **белков**

Pfam prosite **InterPro, …**

Банки **семейств белков**

RCSB PDB PROTEIN DATA BANK

Архивная база **пространственных** структур макромолекул

И многие другие...

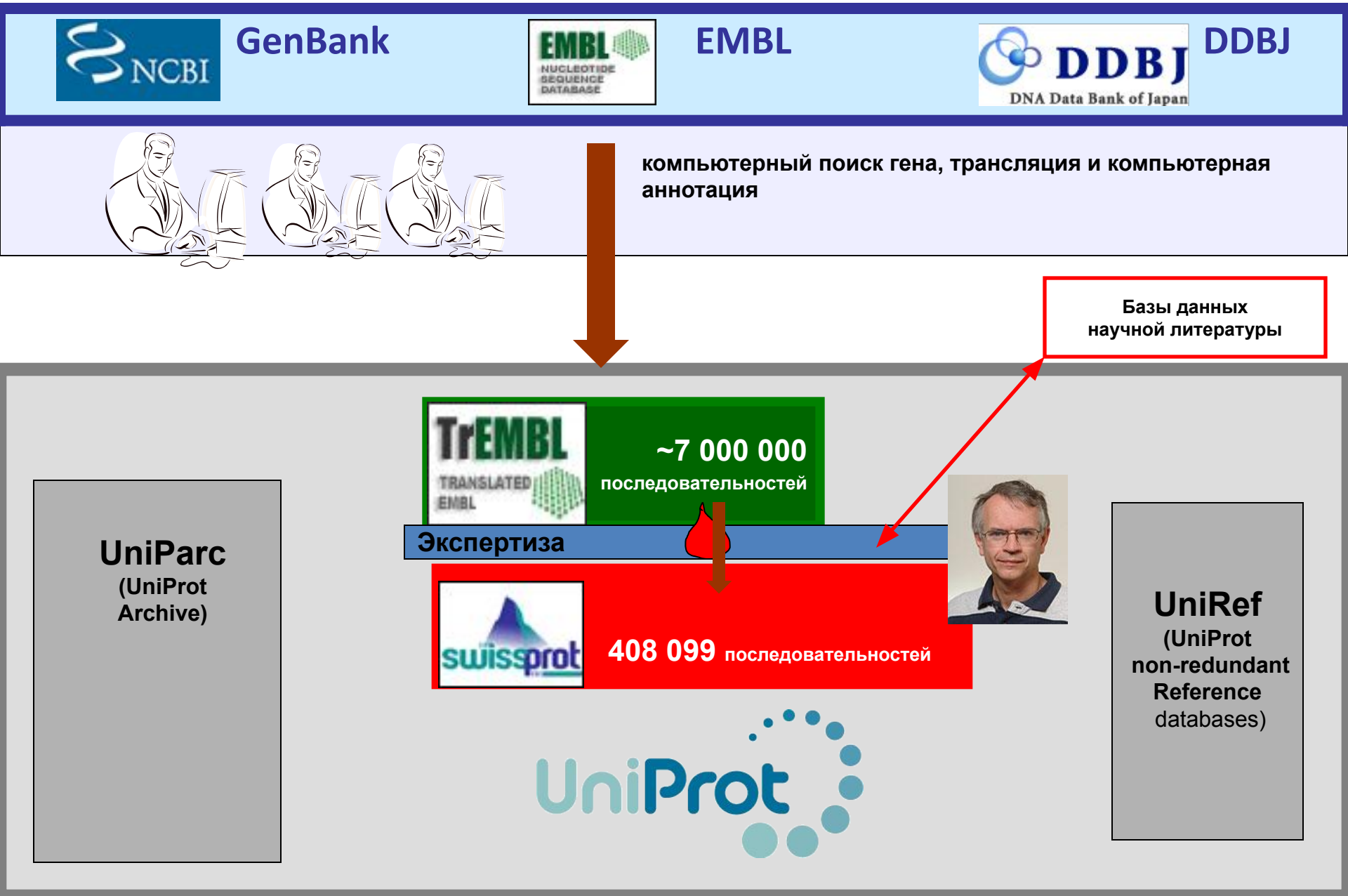
Банк данных

Universal **Prot**ein Resource

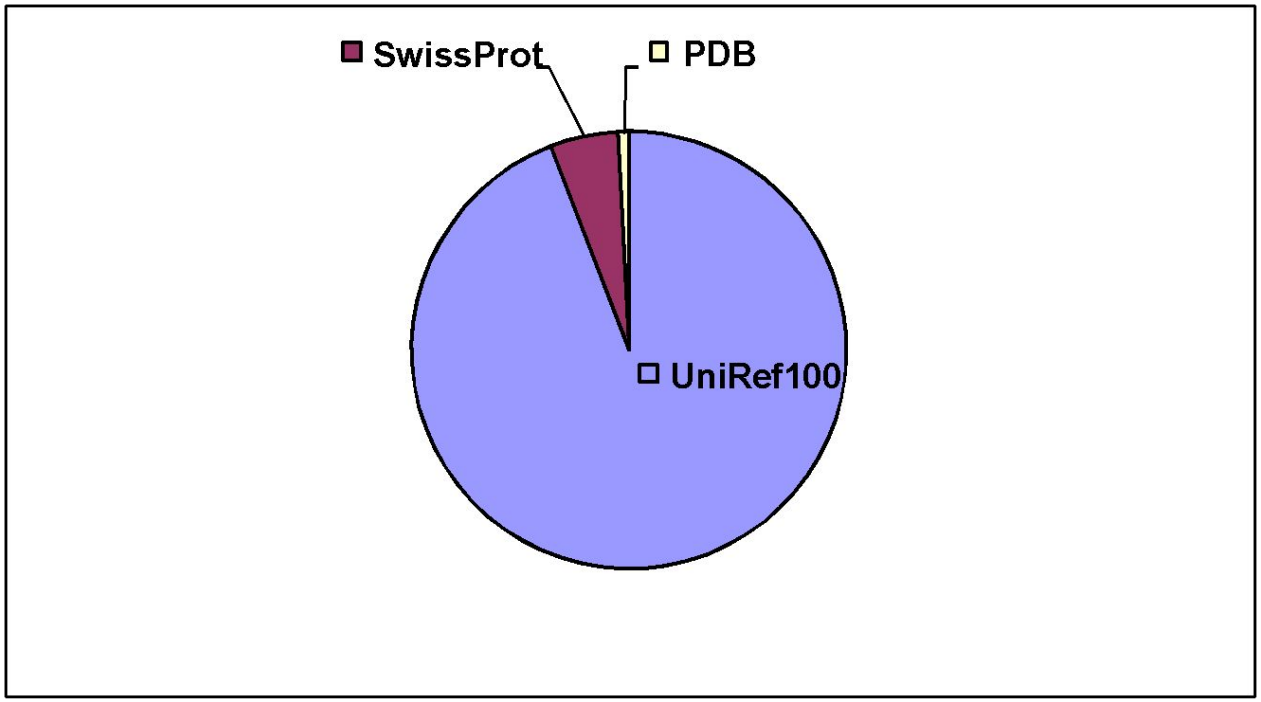
- UniProt Knowlegebase

- UniProt Archive – **UniParc**
  - Все доступные белковые последовательности из разных БД
  - Свой неизменный ID
  - История записей про каждую последовательность
- UniProt Reference – **UniRef**
  - Избавление от избыточности

ttttacctctttttagtgatattgtgatatagagcaaaaatcccgacattgtgtcgggattgtttttaaactcttgttgattttaatttttcaatcgcttctttattaaagaagtagtgtgtgccacaacactcacattg
catatcaatacggcctttatgttcggctaatatttcgtcaatttcttcatcagagatgagcagtagatgcagaactagaacgctcagcagagcagccacagaaaaattgtacatcttgtgctggataaagattaa
cggtttcttcgtgatataaacgataggagtaactcttctgcagggagaccaaataattcttcatcttttactgttgctgcgagcgtagttaaatgctcaaaatcttctggtgtaccagaaccatcaggcataatttg
taataacatacctgctgccactggcttgccttcatattctccagtacgaataattaattgagtttgaagactcatattttcagtgaagtttcgatcgcccttaggaggggccgcgctttctctttcaa



Соотношение числа белков,
представленных в разных банках



Последовательностей во много раз больше, чем структур!

Большинство последовательностей не аннотированы!

Документ банка данных Swiss-Prot

```
ID   YSEA_STACA     STANDARD;      PRT;   165 AA.
AC   P47995;
DT   01-FEB-1996 (Rel. 33, Created)
DT   01-FEB-1996 (Rel. 33, Last sequence update)
DT   13-SEP-2005 (Rel. 48, Last annotation update)
DE   Hypothetical protein in secA 5'region (ORF1) (Fragment).
OS   Staphylococcus carnosus.
OC   Bacteria; Firmicutes; Bacillales; Staphylococcus.
OX   NCBI_TaxID=1281;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RC   STRAIN=TM300;
RA   Freudl R.;
RL   Submitted (JUN-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- SIMILARITY: Belongs to the ribosomal protein S30Ae family.
CC   -!- CAUTION: This is a conceptual translation.
CC   -!- CAUTION: Ref.1 sequence differs from that shown due to frameshifts
CC       in positions 25 and 46.
CC   ---------------------------------------------------------------------
CC   This Swiss-Prot entry is copyright. It is produced through a collaboration
CC   between  the Swiss Institute of Bioinformatics  and the  EMBL outstation -
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use as long as its content is in no way modified and this statement is not
CC   removed.
CC   ---------------------------------------------------------------------
DR   EMBL; X79725; CAA56161.1; ALT_FRAME; Genomic_DNA.
DR   PIR; S47148; S47148.
DR   InterPro; IPR003489; Ribosomal_S30S54.
DR   Pfam; PF02482; Ribosomal_S30AE; 1.
KW   Hypothetical protein.
FT   NON_TER       1       1
SQ   SEQUENCE   165 AA;  19138 MW;  BF8CB91ADE194DD0 CRC64;
     LERYFTNVPN VNAHVKVKTY ANSSKIEVTI PLNDVTLRAE ERNDDIYAGI DKITNKLECQ
     VRKYKTRVNR KKRKESEHEP FPATPETPPE TAVDHDKDDE IEIIRSKQFS LKPMDSEEAV
     LQMDLLGTDF FIFNDRETDG TSIVYRRKDG KYGLIETVEK LICDI
//
```

Описание документа: идентификатор, имя, дата создания и модификации

Аннотация последовательности

Последовательность

Основные поля записи SwissProt

- ID - Идентификатор последовательности, часто кодирует биологически осмысленную информацию, меняется от выпуска к выпуску БД
- AC - "Код доступа" — уникальный идентификатор последовательности, не меняющийся от выпуска к выпуску БД
- DE - Название (краткое описание) белка, часто указывающее на его функцию
- OS - Организм, в котором найден белок
- OC - Полная таксономия организма

И сама последовательность, конечно.