

Что такое биоинформатика?

Банк SwissProt

С.А.Спирин

7, 8, 10 февраля 2006 г., ФББ МГУ

Что такое биоинформатика?

- Исследование информационных процессов в биологических системах (клетках, органах, организме, популяции).
- Изучение и внедрение в компьютерную науку «биологических» методов анализа информации (нейросетей, генетических алгоритмов, нечеткой логики и др.).
- Применение компьютерных методов для решения биологических задач.
- Телепатия, парапсихология, информационные поля и т.п.



Биоинформатика

Исследование информационных процессов в биологических системах (клетках, органах, организме, популяции).

Изучение и внедрение в компьютерную науку «биологических» методов анализа информации (нейросетей, генетических алгоритмов, нечеткой логики и др.).

Применение компьютерных методов для решения биологических задач.

Телепатия, парапсихология, информационные поля и т.п.

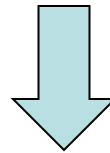
Примеры задач биоинформатики

- Разработка алгоритмов для анализа большого объема биологических данных
 - Алгоритм поиска генов в геноме
- Анализ и интерпретация биологических данных таких, как нуклеотидные и аминокислотные последовательности, структура молекул белков, структура комплексов молекул белков с другими молекулами.
 - Изучение структуры активного центра белка
- Разработка программного обеспечения для управления и быстрого доступа к биологическим данным
 - Создание банка данных аминокислотных последовательностей

Что понимать под биоинформатикой?

Как видим, смысл термина ещё уже...

Применение компьютерных методов для решения биологических задач



Применение компьютерных методов для решения задач
молекулярной биологии

... и ещё уже...

Компьютерный анализ экспериментальных данных о структурах биологических макромолекул (белков и нуклеиновых кислот) с целью получения биологической информации

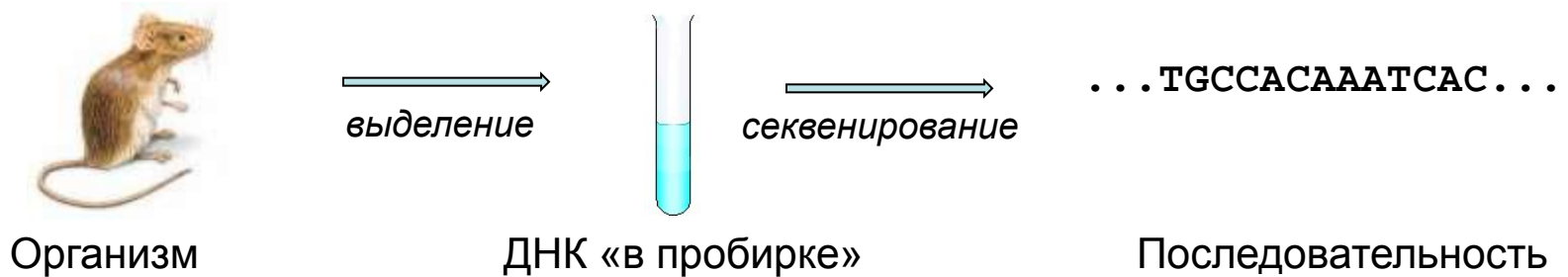
Итак...

Биоинформатика = вычислительная молекулярная биология

Почему так сузился смысл термина?

gatcctccatatacaacggtatctccacctcaggtttagatctcaacaacggaaccattg
ccgacatgagacagttaggtatcgtcgagagttacaagctaaaacgagcagtagtcagct
ctgcatctgaagccgctgaagttctactaaggggtggataacatcatccgtgcaagaccaa
gaaccgccaatagacaacatatgtaacatatttaggatatacctcgaaaataataaccg
ccacactgtcattattataattagaaacagaacgcaaaaattatccactatataattcaa
agacgcgaaaaaaaaagaacaacgcgtcatagaacttttggcaattcgcgtcacaataa
at tt tggcaacttatgtttcctcttcgagcagtagctcgagccctgtctcaagaatgtaat
aatacccatcgtaggtatggttaagatagcatctccacaacctcaaagctccttgccga
gagtcgccctcctttgtcgagtaat t t t cact t t t c a t a t g a g a a c t t a t t t t c t t a t t c
t t t a c t c t c a c a t c c t g t a g t g a t t g a c a c t g c a a c a g c c a c c a t c a c t a g a a g a a c a g a
a c a a t t a c t t a a t a g a a a a t t a t a t c t t c c t c g a a a c g a t t t c c t g c t t c c a a c a t c t a
c g t a t a t c a a g a a g c a t t c a c t t a c c a t g a c a c a g c t t c a g a t t t c a t t a t t g c t g a c a g
c t a c t a t a t c a c t a c t c c a t c t a g t a g t g g c c a c g c c c t a t g a g g c a t a t c c t a t c g g a a
a a c a a t a c c c c c c a g t g g c a a g a g t c a a t g a a t c g t t t a c a t t t c a a a t t t c c a a t g a t a
c c t a t a a a t c g t c t g t a g a c a a g a c a g c t c a a a t a a c a t a c a a t t g c t t c g a c t t a c c g a
g c t g g c t t t c g t t t g a c t c t a g t t c t a g a a c g t t c t c a g g t g a a c c t t c t t c t g a c t t a c
t a t c t g a t g c g a a c a c c a c g t t g t a t t t c a a t g t a a t a c t c g a g g g t a c g g a c t c t g c c g
a c a g c a c g t c t t t g a a c a a t a c a t a c c a a t t t g t t g t t a c a a a c c g t c c a t c c a t c t c g c
t a t c g t c a g a t t t c a a t c t a t t g g c g t t g t t a a a a a c t a t g g t t a t a c t a a c g g c a a a a
a c g c t c t g a a a c t a g a t c c t a a t g a a g t c t t c a a c g t g a c t t t t g a c c g t t c a a t g t t c a
c t a a c g a a g a a t c c a t t g t g t c g t a t t a c g g a c g t t c t c a g t t g t a t a a t g c g c c g t t a c
c c a a t t g g c t g t t c t t c g a t t c t g g c g a g t t g a a g t t t a c t g g g a c g g c a c c g g t g a t a a
a c t c g g c g a t t g c t c a g a a a c a a g c t a c a g t t t t g t c a t c a t c g c t a c a g a c a t t g a a g
g a t t t t c t g c c g t t g a g g t a g a a t t c g a a t t a g t c a t c g g g g c t c a c c a g t t a a c t a c c t
c t a t t c a a a a t a g t t t g a t a a t c a a c g t t a c t g a c a c a g g t a a c g t t t c a t a t g a c t t a c
c t c t a a a c t a t g t t t a t c t c g a t g a c g a t c c t a t t t c t t c t g a t a a a t t g g g t t c t a t a a

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



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

GenBank — хранилище последовательностей нуклеиновых кислот в виде компьютерных файлов

Объем GenBank'a:

1982: 680 338 букв в 606 последовательностях

1992: 101 008 486 букв в 78 608 последовательностях

2002: 28 507 990 166 букв в 22 318 883 последовательностях

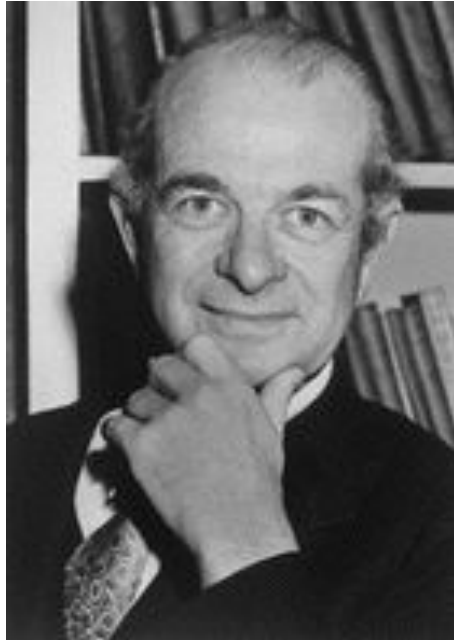
2004: 44 575 745 176 букв в 40 604 319 последовательностях

2005: 56 037 734 462 букв в 52 016 762 последовательностях
(из ~165 000 организмов)

Размер файлов — 196 Gb

Пионеры биоинформатики

1962



Лайнус Полинг

- Анализ аминокислотных последовательностей глобинов нескольких позвоночных
- Гипотеза **молекулярных часов**

Zuckerlandl, E., and L. Pauling. **1962**. Molecular disease, evolution, and genic heterogeneity. Horizons in Biochemistry, Academic Press, New York, 189-225.

Zuckerlandl, E., and L. Pauling. **1965**. Evolutionary divergence and convergence in proteins. Evolving Genes and Proteins, Academic Press, New York, 97-166.

Пионеры биоинформатики

1965



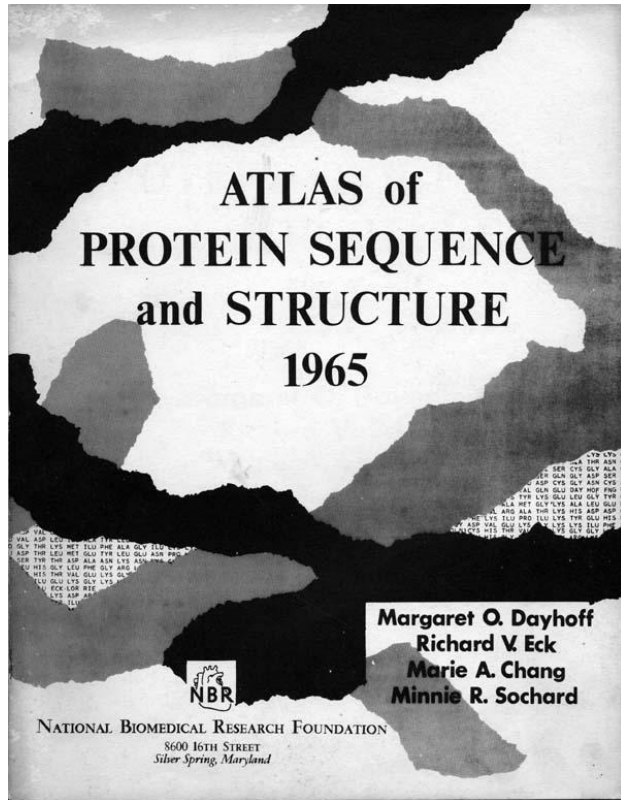
Маргарет Дейхофф

- Однобуквенный код аминокислот
A,C,D,E,F,G,H...
- Матрицы аминокислотных замен
PAM (Point Accepted Mutation)

Атлас последовательностей белков и их структур (1965)

Первый “банк данных”

1965 -1978



Атлас белковых
последовательностей и
их структур

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

Банки данных

- **Архивные**
(примеры: PDB, GenBank)
за содержание каждой записи отвечает её автор-экспериментатор
- **Курируемые**
за содержание записей отвечают специальные люди — кураторы
- **Автоматические**
записи генерируются компьютерными программами

Банк данных Swiss-Prot

1986



Swiss-Prot – база знаний о
белковых последовательностях

- Курируемая база данных
- “**Золотой стандарт**” аннотации

Банк данных Swiss-Prot



С 1987 поддерживается в сотрудничестве между

Swiss Institute of Bioinformatics (**SIB**)
European Bioinformatics Institute (**EBI**)

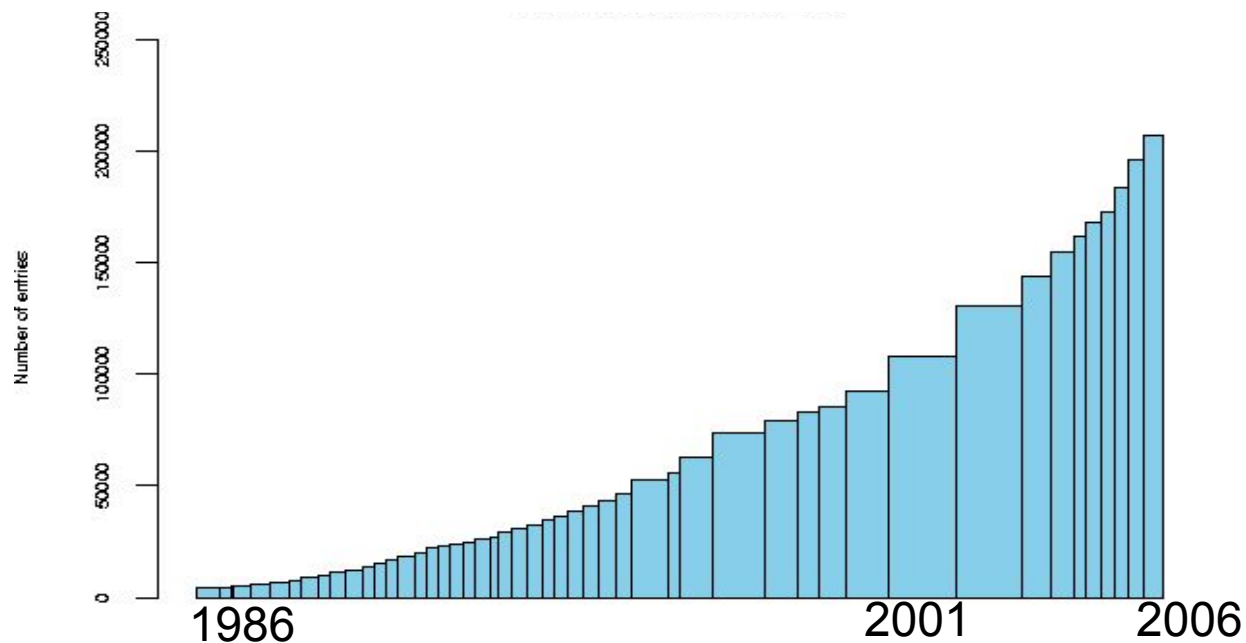


Амос Байрох

Руководитель группы Swiss-Prot в
Швейцарском Институте Биоинформатики

Банк данных Swiss-Prot

Статистика роста количества документов



Текущий релиз **48.9** (24 января 2006) содержит 206586 документов

Банк данных TrEMBL



TrEMBL (Translated **EMBL)**

Формальная трансляция всех кодирующих
нуклеотидных последовательностей из банка EMBL

Автоматическая классификация и аннотация

Текущий релиз **31.9** (24 января 2006) содержит 2 586 884 документа

Тенденция объединения

2002



PIR Protein Information Resource



Банк данных UniProt



UniProt (Universal Protein Resource)

- UniProt Knowledgebase – **SwissProt+TrEMBL**
- UniProt Archive – **UniParc**
- UniProt Reference – **UniRef**

ttttacctcttttagtgatattgtgatagagcaaaaatcccgcacattgtgctgggattgttttaaaccttgttgatttaattttcaatcgcttctttataaagaagtagtggtgccc
acaacactcacattgcatatcaatacggcctttatgttcggctaataatttcgcaatttcttcatcagagatgagcagtagatgcagaactagaacgctcagcagagcagccaca
gaaaaattgtacatcttgtgctggataaagattaacggtttctctgtagataaacgataggagtaacttctgacgggagaccaataattcttcatctttactgttgctgagc
gtagttaaatgctcaaaatcttctggtgtaccagaaccatcaggcataattgtaataacatacctgctgccaactggctgaccttcatattctccagtacgaataattaattgagttg



GenBank



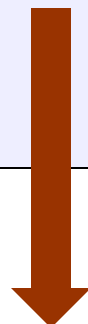
EMBL



DDBJ



компьютерный поиск гена, трансляция и компьютерная аннотация



Базы данных научной литературы



~2 500 000 последовательностей

Экспертиза



~200 000 последовательностей

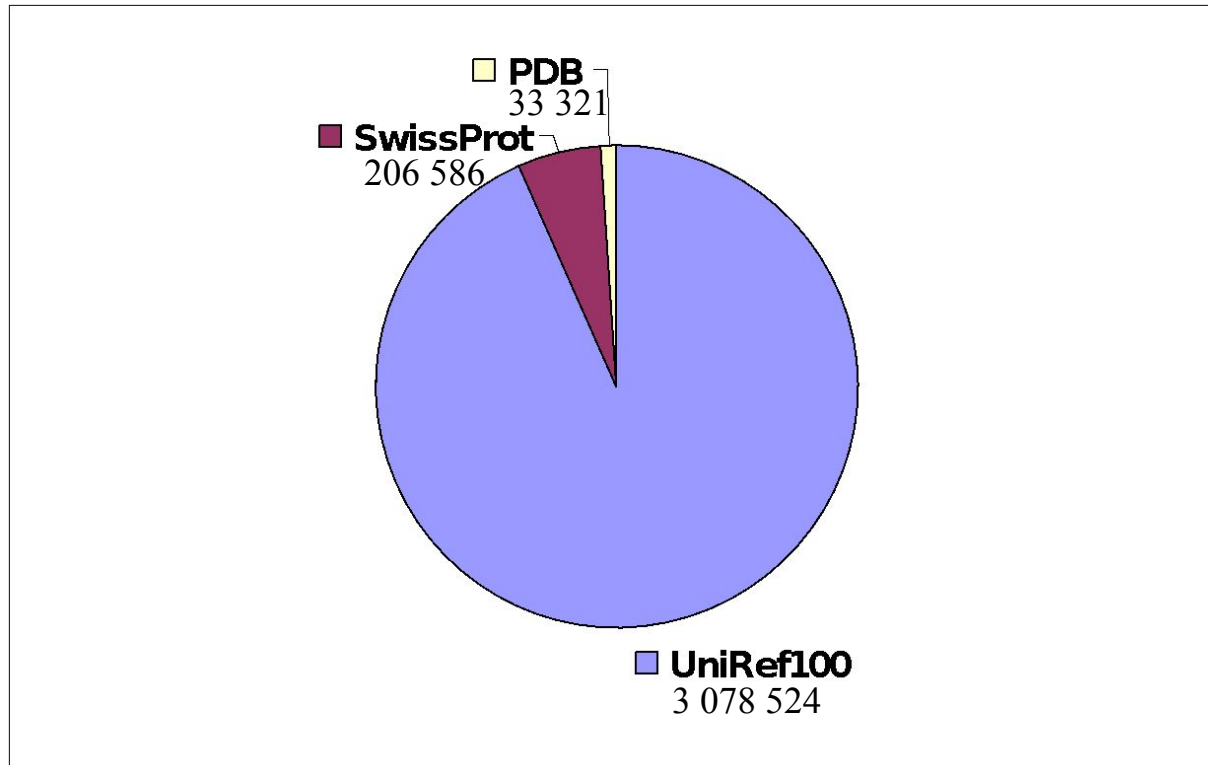


UniParc
(UniProt Archive)

UniRef
(UniProt non-redundant Reference databases)



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 -----
CC DR EMBL; X79725; CAA56161.1; ALT_FRAME; Genomic_DNA.
CC DR PIR; S47148; S47148.
CC DR InterPro; IPR003489; Ribosomal_S30S54.
CC DR Pfam; PFO2482; Ribosomal_S30AE; 1.
CC KW Hypothetical protein.
CC FT NON_TER 1 1
CC SQ SEQUENCE 165 AA; 19138 MW; BF8CB91ADE194DDO CRC64;
CC LERYFTNVPN VNAHVKVKTY ANSSKIEVTI PLNDVTLRAE ERNDDIYAGI DKITNKLECG
CC VRKYKTRVNR KKRKESSEHEP FPATPETPPE TAVDHDKDDE IEIIRSKQFS LKPMDSEEAV
CC LQMDLLGTDF FIFNDRETDG TSIVYRRKDG KYGLIETVEK LICDI
```

Описание документа: идентификатор,
ИМЯ, дата создания и модификации

Аннотация
последовательности

Последовательность

Основные поля записи SwissProt

- ID
- AC
- DE
- OS
- OC

И сама последовательность, конечно.