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.

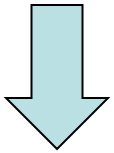
Примеры задач биоинформатики

- 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 ỳже...

Применение компьютерных методов для решения биологических задач



Применение компьютерных методов для решения задач **молекулярной биологии**

... и еще ỳже...

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

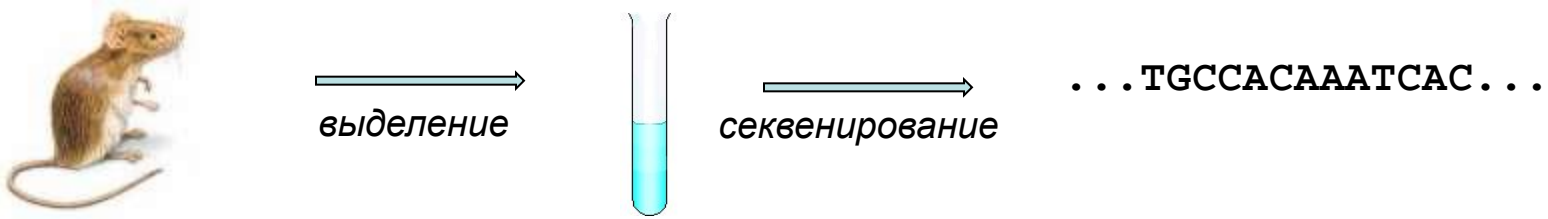
Итак...

Биоинформатика = вычислительная молекулярная биология

Почему так сузился смысл термина?

gatcctccatatacaacggtatctccacctcaggtttagatctcaacaacggaaccattg
ccgacatgagacagttaggtatcgtcgagagttacaagctaaaacgagcagtagtcagct
ctgcatctgaagccgctgaagttctactaagggtggataacatcatccgtgcaagaccaa
gaaccgccaatagacaacatatgtaacatatttaggatatacctcgaaaataataaaccg
ccacactgtcattattataattagaaacagaacgcaaaaattatccactatataattcaa
agacgcgaaaaaaaaagaacaacgcgtcatagaacttttggcaattcgcgtcacaaataa
attttggcaacttatgtttcctcttcgagcagtactcgagccctgtctcaagaatgtaat
aatacccatcgtaggtatggttaaagatagcatctccacaacctcaaagctccttgccga
gagtcgccctcctttgtcgagtaattttcacttttcatatgagaacttattttcttattc
tttactctcacatcctgtagtgattgacactgcaacagccaccatcactagaagaacaga
acaattacttaatagaaaaattatatcttcctcgaaacgatttcctgcttccaacatcta
cgtatatcaagaagcattcacttaccatgacacagcttcagatttcattattgctgacag
ctactatatcactactccatctagtagtggccacgccctatgaggcatatcctatcggaa
aacaataccccccagtggcaagagtcaatgaatcgtttacatttcaaatttccaatgata
cctataaatcgtctgtagacaagacagctcaaataacatacaattgcttcgacttaccga
gctggctttcgtttgactctagttctagaacgttctcaggtgaaccttcttctgacttac
tatctgatgcgaacaccacgttgtatttcaatgtaatactcgagggtacggactctgccg
acagcacgtctttgaacaatacataccaatttgttgttacaaaccgtccatccatctcgc
tatcgtcagatttcaatctattggcgttgttaaaaaactatggttatactaacggcaaaa
acgctctgaaactagatcctaatgaagtcttcaacgtgacttttgaccgttcaatgttca
ctaacgaagaatccattgtgtcgtattacggacgttctcagttgtataatgcgccgttac
ccaattggctgttcttcgattctggcgagttgaagtttactgggacggcaccggtgataa
actcggcgattgctccagaaacaagctacagttttgtcatcatcgctacagacattgaag
gattttctgccgttgaggtagaattcgaattagtcatcggggctcaccagttaactacct
ctattcaaaatagtttgataatcaacgttactgacacaggtaacgtttcatatgacttac
ctctaaactatgtttatctcgatgacgatcctatttcttctgataaattgggttctataa

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

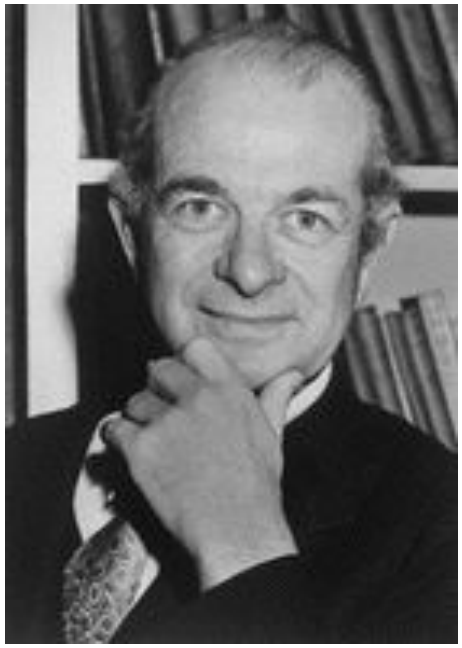
1992: 101 008 486 букв в 78 608 последовательностях

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**

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

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

Пионеры биоинформатики

1965

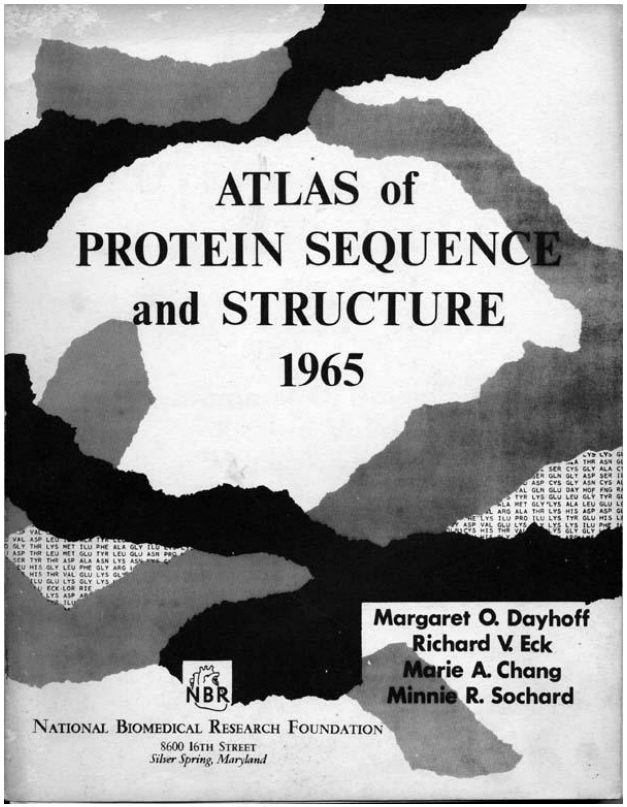
Маргарет Дейхофф

- 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

http://www.expasy.org/sprot/

Банк данных Swiss-Prot

С 1987 поддерживается в сотрудничестве между

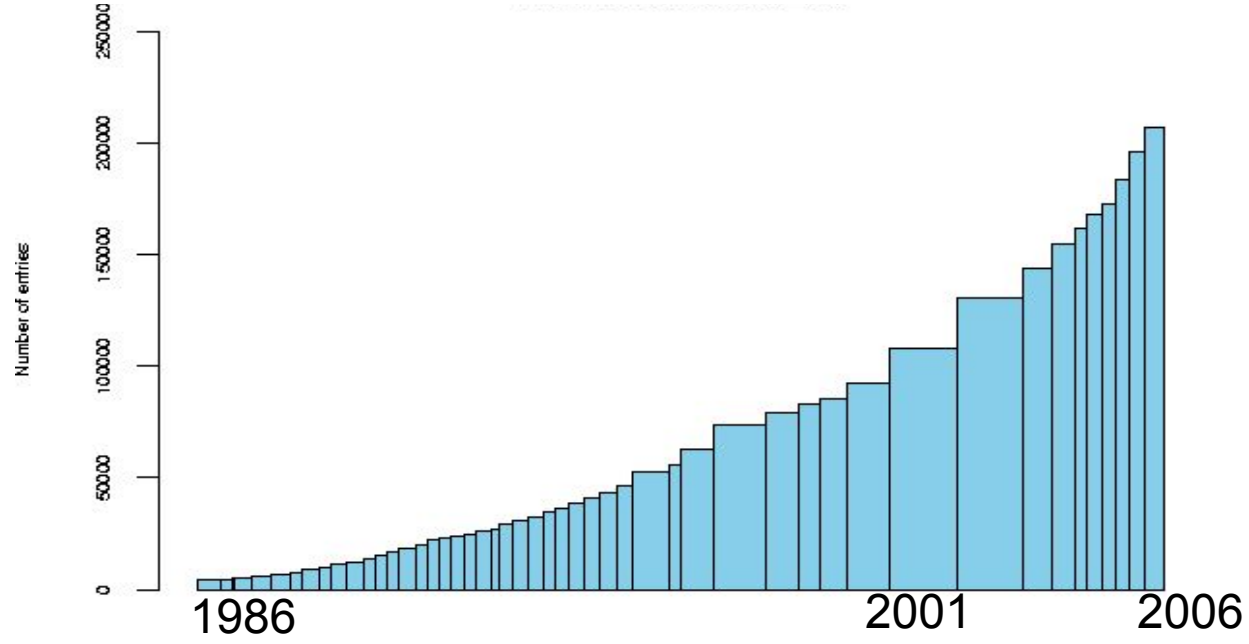
Swiss **I**nstitute of **B**ioinformatics (**SIB**)
European **B**ioinformatics **I**nstitute (**EBI**)

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 (**Tr**anslated **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 (**Uni**versal **Prot**ein Resource)

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

GenBank EMBL DDBJ

компьютерный поиск гена, трансляция и компьютерная аннотация

Базы данных научной литературы

TrEMBL TRANSLATED EMBL ~2 500 000 последовательностей

Экспертиза

swissprot ~200 000 последовательностей

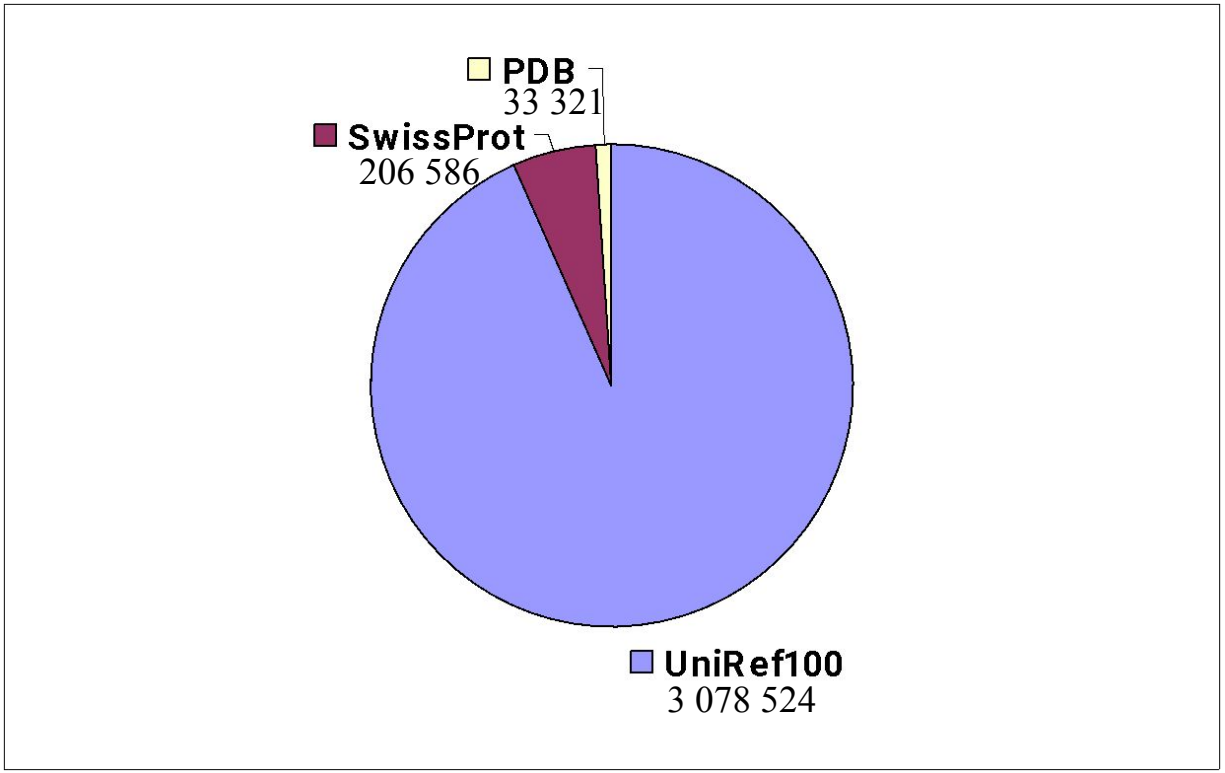
UniParc (UniProt Archive)

PIR-PSD pir.georgetown.edu

UniProt the universal protein resource

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   ---------------------------------------------------------------------
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
- AC
- DE
- OS
- OC

И сама последовательность, конечно.