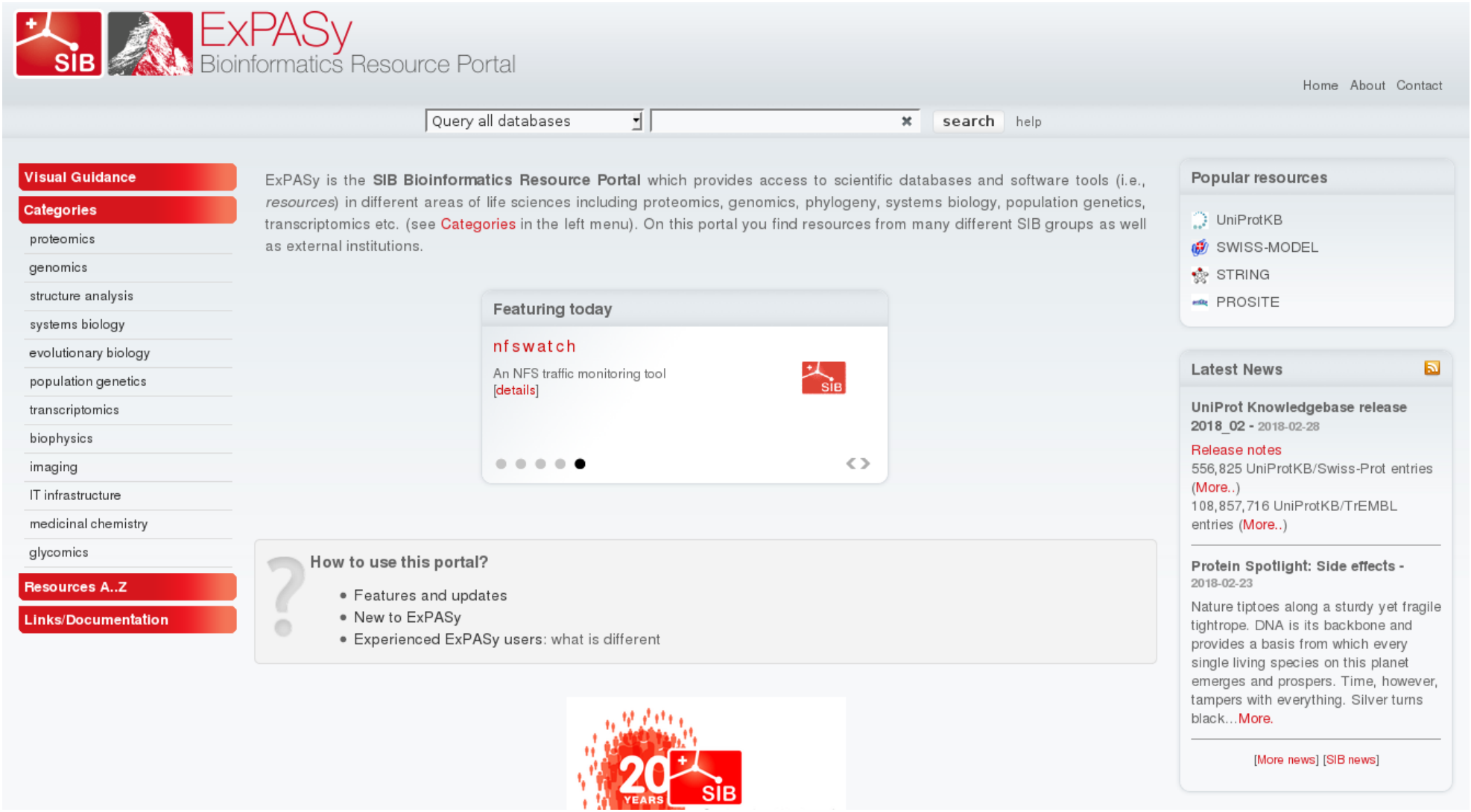
Банки последовательностей белков
UniProt

12/03/2019

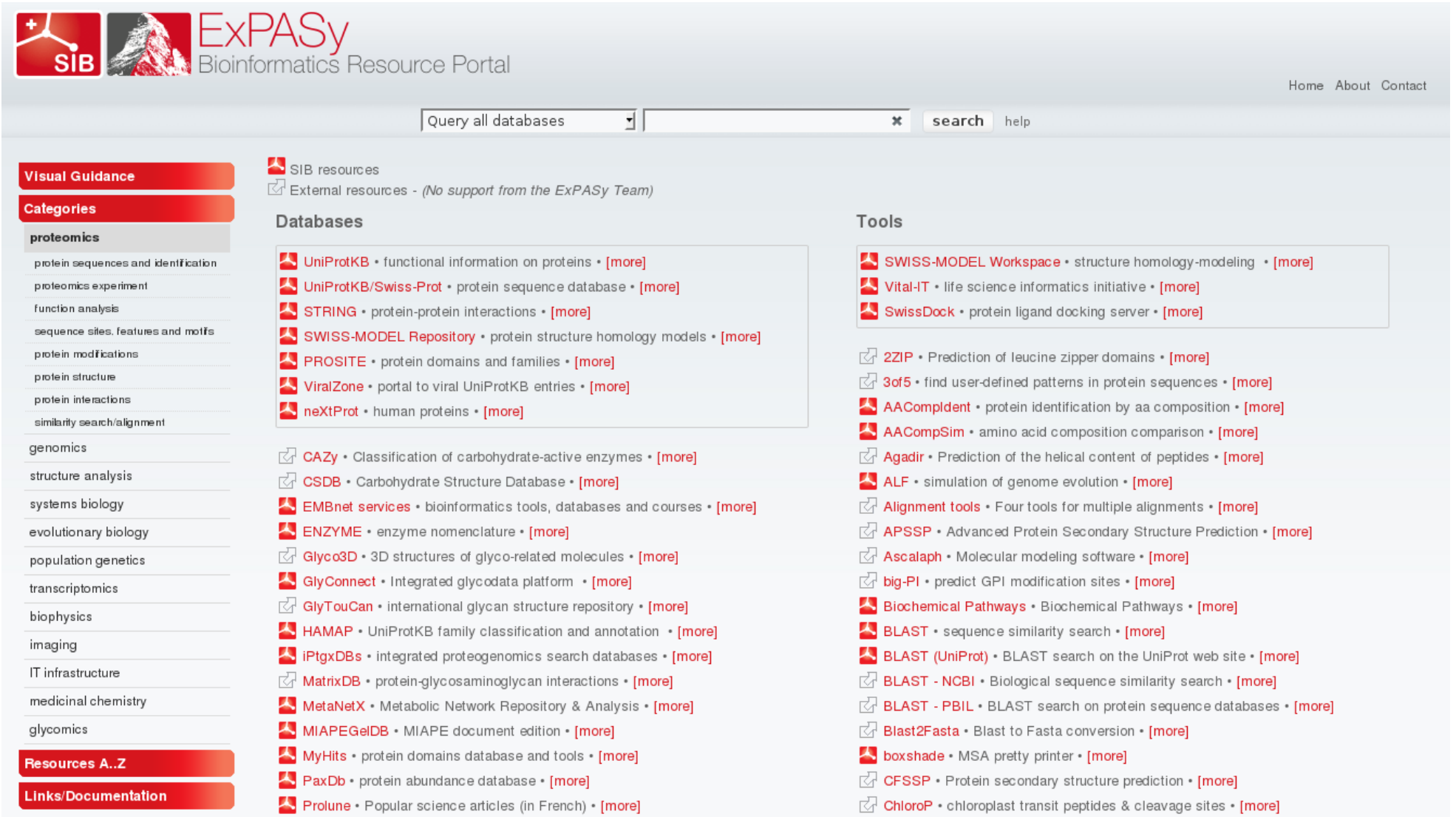
Банки данных о белках

- UniProt – последовательности и аннотации
- RefSeq – последовательности и аннотации
- PDB – пространственные структуры
- PubMed – публикации
- … – еще много чего

ExPASy (www.expasy.org)



ExPASy proteomics



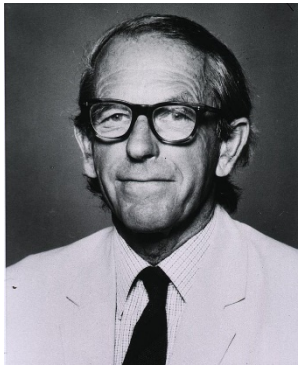
Откуда берется информация?

Экскурс в историю секвенирования

- Первая последовательность белка: инсулин, цепи A и B

  Frederick Sanger, 1951, 1953, нобелевская премия 1958

  До двойной спирали ДНК и кода!



F. Sanger 1918-2013

- Первая последовательность РНК: аланиновая тРНК

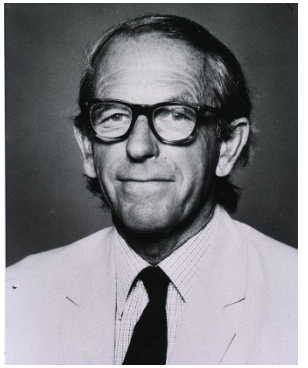
  Robert Holley, 1964, нобелевская премия 1968

R.Holley 1922-1993

- Первый полный геном ДНК бактериофага φX174

  Frederick Sanger, 1977,
  вторая нобелевская премия 1980

  Метод секвенирования “по Сэнгеру”



F. Sanger 1918-2013

- Изобретение ПЦР: полимеразной цепной реакции

  Kary Mullis, 1985,
  нобелевская премия 1993

K. Mullis 1944

- Первый геном бактерии Haemophilus influenzae

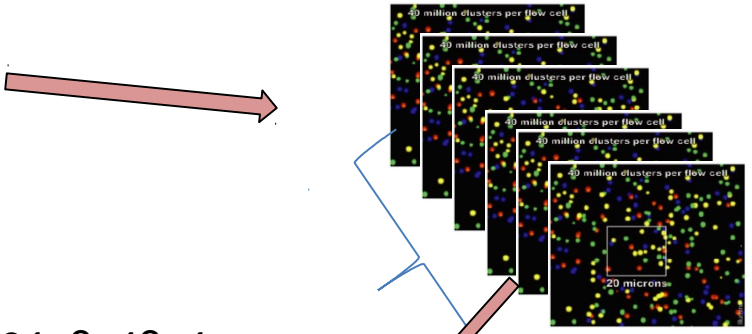
  Метод дробовика (Shotgun sequencing), 1995
  Нужны алгоритмы, программы, компьютер

- Новое поколение секвенаторов последовательностей ДНК (next generation sequencing)

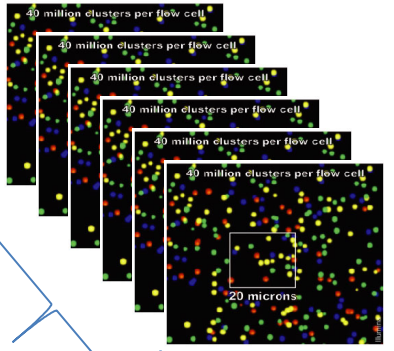
  Illumina Solexa (2006)
  Pacific Biosciences SMRT (2010)
  Roche/454 (2004)

Illumina HiSeq 2000

ДНК в пробирке



101x8x48x4 снимков
прямой проход



Столько же —
на обратный проход

~2x8x200 млн. последовательностей длины 101 в формате fastq,
Итого порядка **300 млрд букв**, 100-кратное покрытие генома человека

Что дальше?

Депонирование последовательности в базу данных

- и её автоматическая аннотация – предсказание кодирующих последовательностей, их продуктов и функции

Публикация о геноме и протеоме (иногда с задержкой на годы)

- более достоверная информация, так как должна проверяться рецензентами
- все же не без ошибок
- её значительно меньше

Последовательности белков

Большинство получены трансляцией предсказанных кодирующих участков в нуклеотидных последовательностях

Правильность предсказаний проверяется:

- лабораторными исследованиями конкретных белков (долго)
- масс-спектрометрией протеома (непросто)
- секвенированием тотальной РНК или конкретной мРНК
- сходством последовательностей с последовательностями известных белков

Классификация банков данных

Архивные – за содержание записи отвечает её автор-экспериментатор (например, GenBank, ENA, PDB)

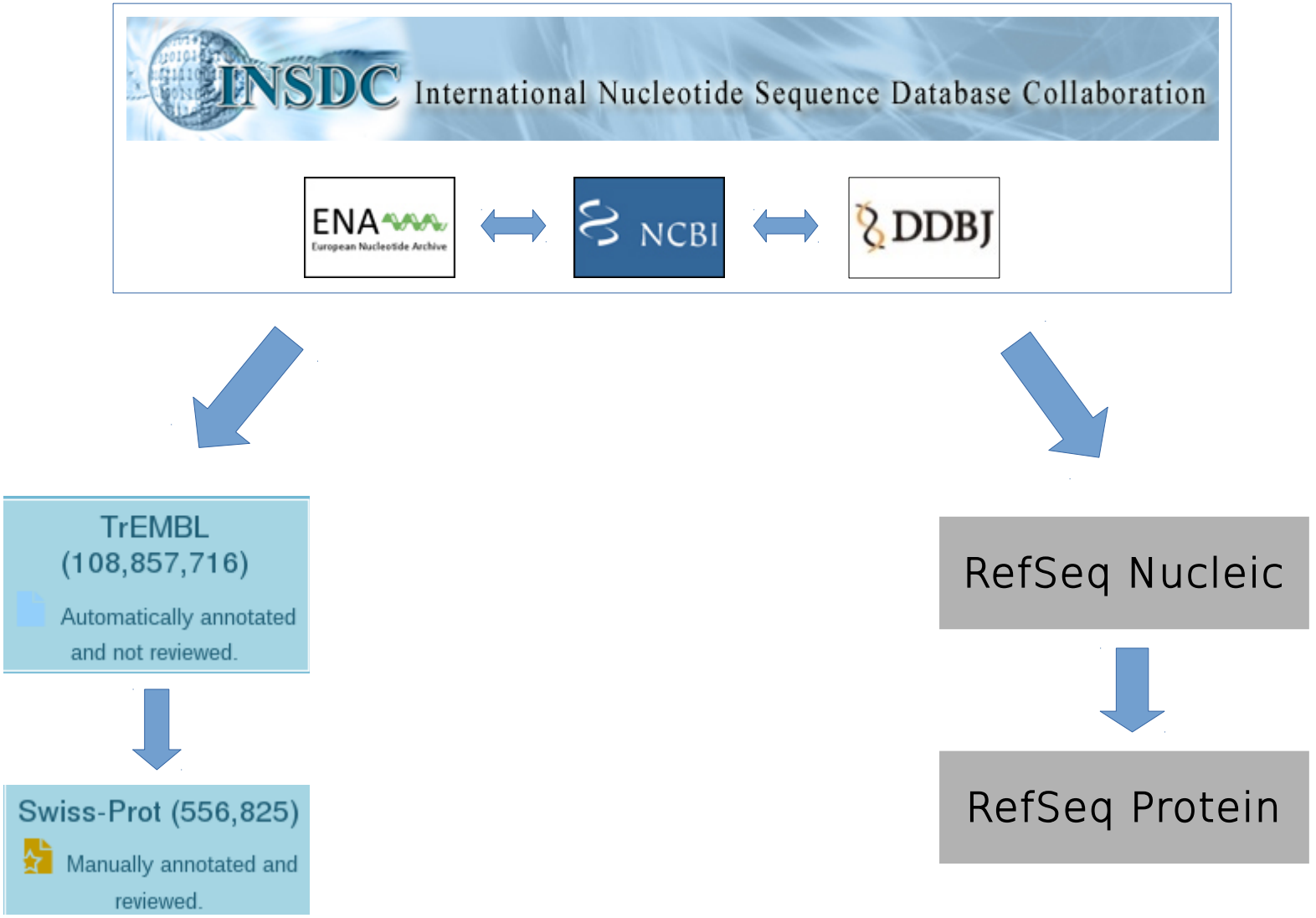
Курируемые – за содержание записи отвечает куратор (например, Swiss-Prot)

Автоматические – записи генерируются компьютерными программами (например, TrEMBL)

Устройство базы данных

- БД состоит из одного или нескольких хранилищ (“таблиц”)
- Единица хранения (строка таблицы) называется *записью* (entry)
- Все записи состоят из *полей* (field). Поля с одним и тем же названием (колонки таблицы) содержат однородную информацию
- Если таблиц больше одной, то записи из разных таблиц ссылаются друг на друга

Базы последовательностей и потоки информации между ними

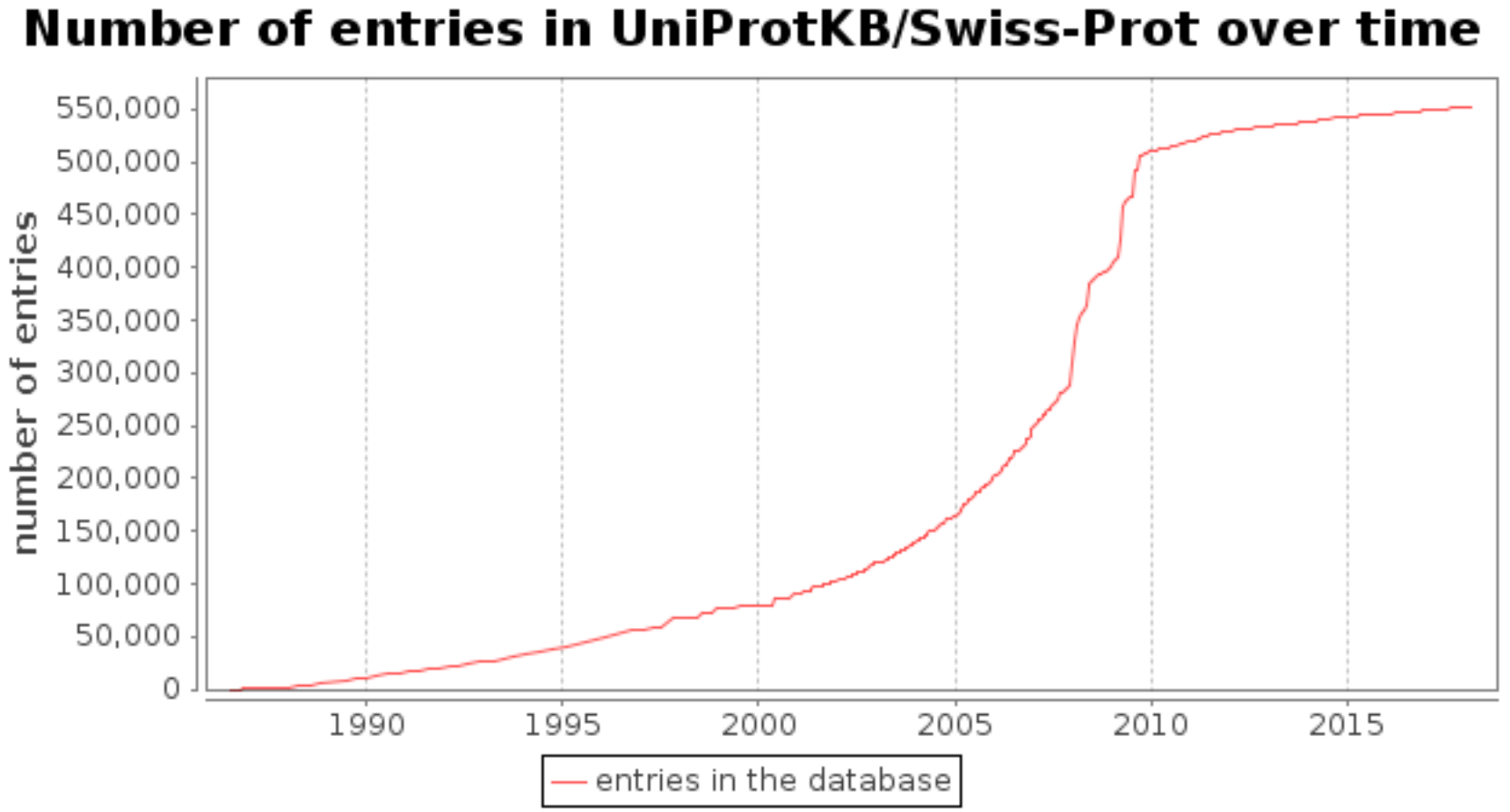


Банк данных Swiss-Prot

- **Swiss-Prot** – база знаний о белковых последовательностях
- Ранее существовал как отдельный банк
- Сейчас – часть Uniprot

- Курируемая база данных
- Аннотации проверяет и дополняет эксперт: использует методы биоинформатики, просматривает публикации.
- 559 228 (полмиллиона) белков

Рост числа записей Swiss-Prot



Банк данных TrEMBL

TrEMBL (**Tr**anslated **EMBL**)

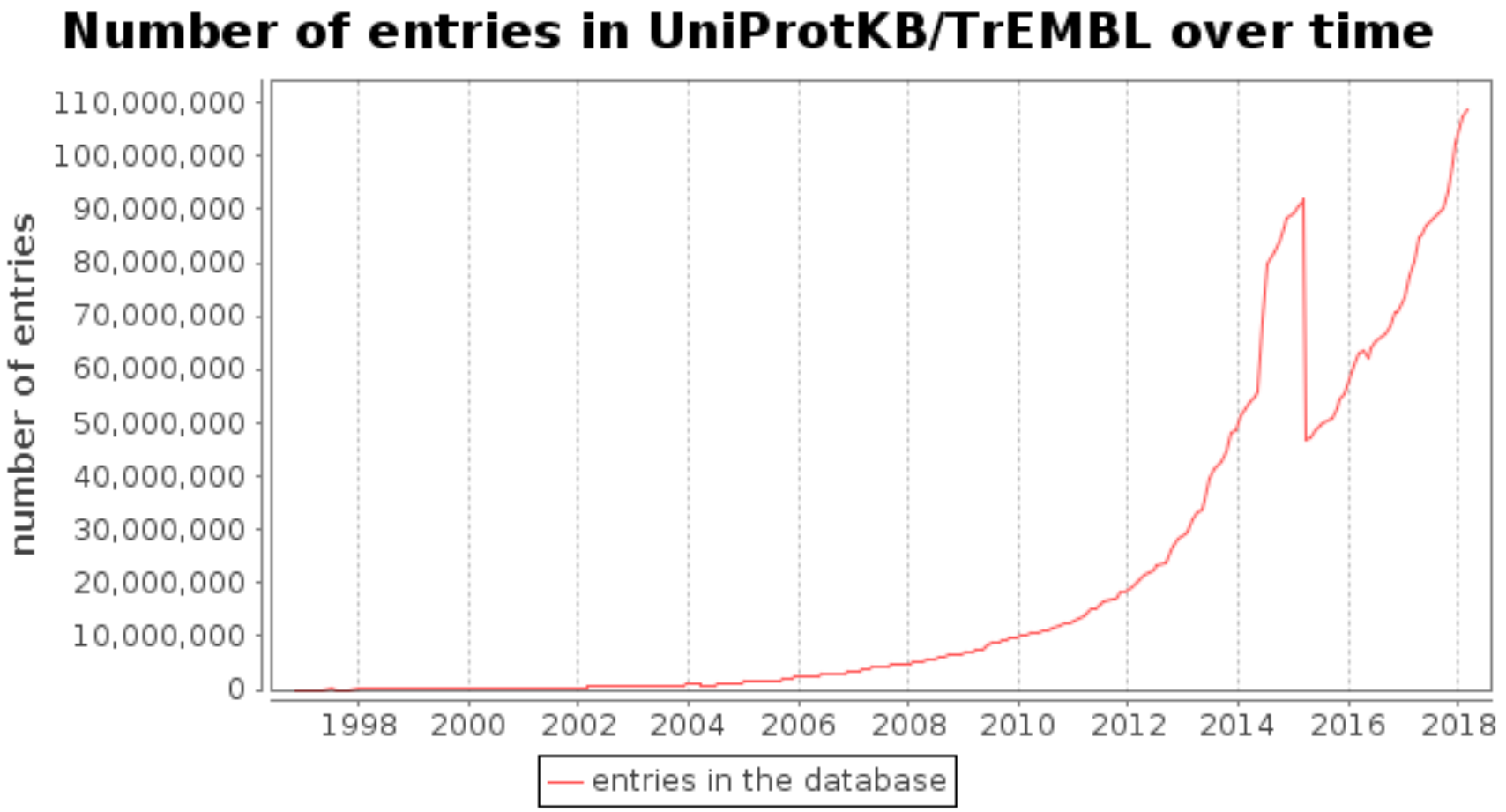
Формальная трансляция всех кодирующих нуклеотидных последовательностей из банка EMBL

Автоматическая классификация и аннотация

Формат записи тот же, что у Swiss-Prot

146 106 279 белков на февраль 2019

Рост числа записей TrEMBL



Известны последовательности десятков миллионов белков

Что такое “один белок”?

Этот вопрос стал нетривиальным и актуальным в последние годы из-за революции в технологии сенквенирования

Одна запись Uniprot

- Примерно: продукт одного гена из одного вида или подвида

- В RefSeq в 2013 ввели новое понятие: последовательность белка, не привязанная строго к одному виду. Идентификаторы вида WP_XXXXXXXX. В аннотации перечислены геномы, в которых такая последовательность закодирована

Проблемы:

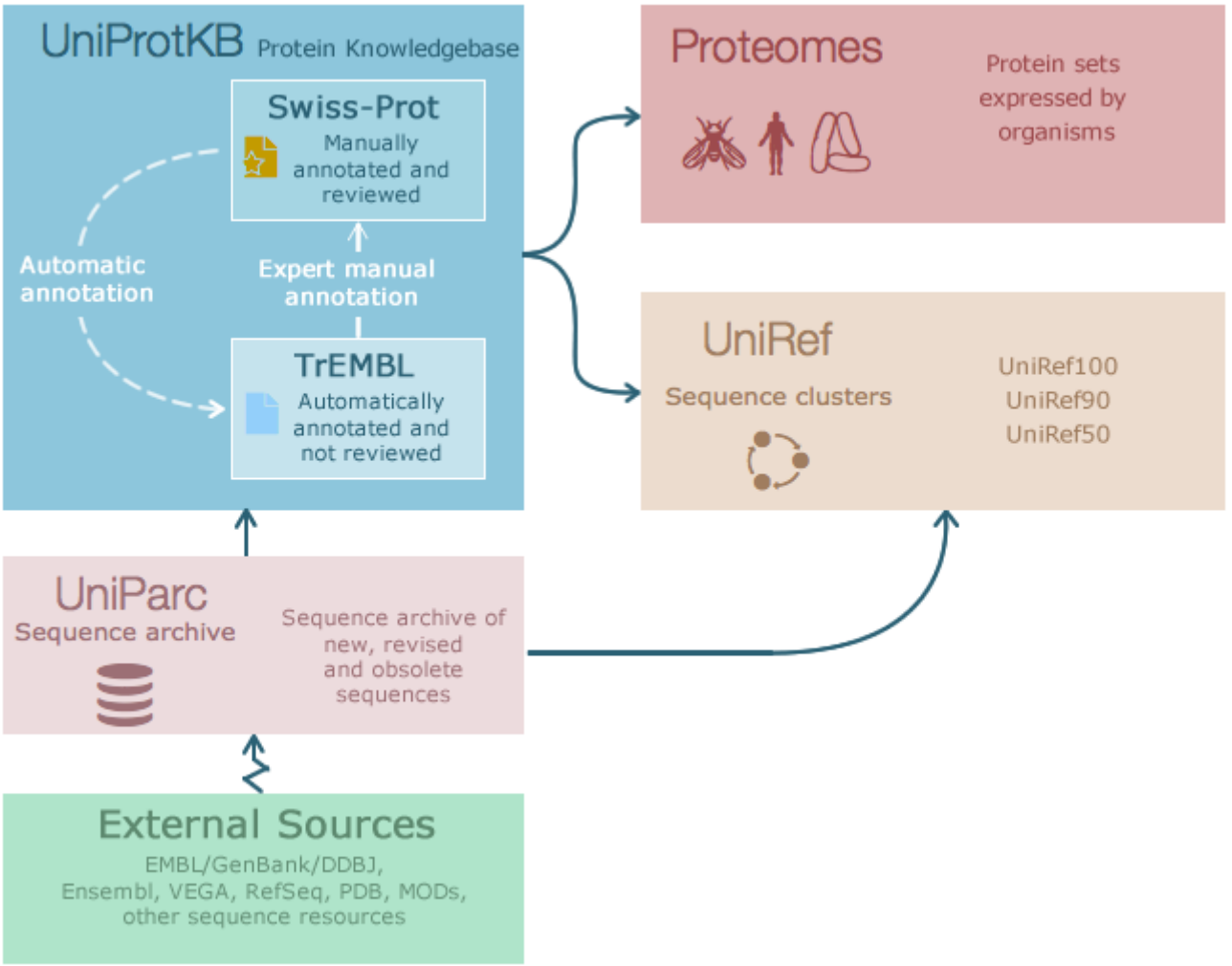
- Два гена из одного генома кодируют один и тот же белок (недавняя дупликация)
- Два гена из разных видов кодируют белки с одинаковой последовательностью
- Полиморфизм: последовательность белка из организма Пети отличается от таковой из организма Коли (или в штаммах бактерий)
- Альтернативный сплайсинг: один ген кодирует несколько изоформ белка, разных по последовательности
- Трансплайсинг: сплайсинг происходит между разными генами! Получающийся белок не закодирован в одном гене
- Соматические различия: разные клетки одного организма кодируют белки с отличающимися последовательностями; иммуноглобулины в лимфоцитах, нормальные и раковые клетки, мутации соматических клеток (?)
- …

Борьба с избыточностью

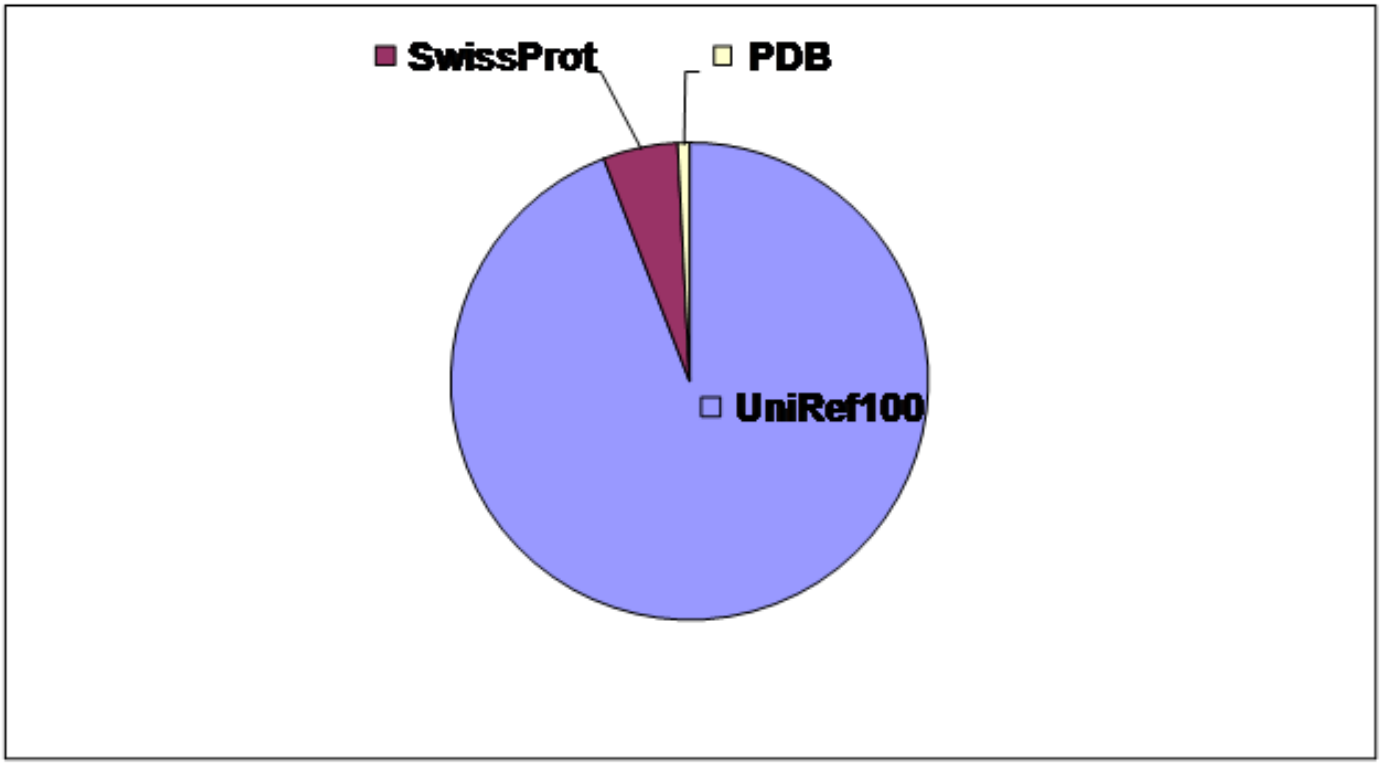
Одна и та же последовательность может попасть в банк несколько раз.

- В UniProt: UniRef – кластеры близких последовательностей (UniRef100, UniRef90, UniRef50).
- В NCBI: RefSeq + Protein Clusters

Структура UniProt



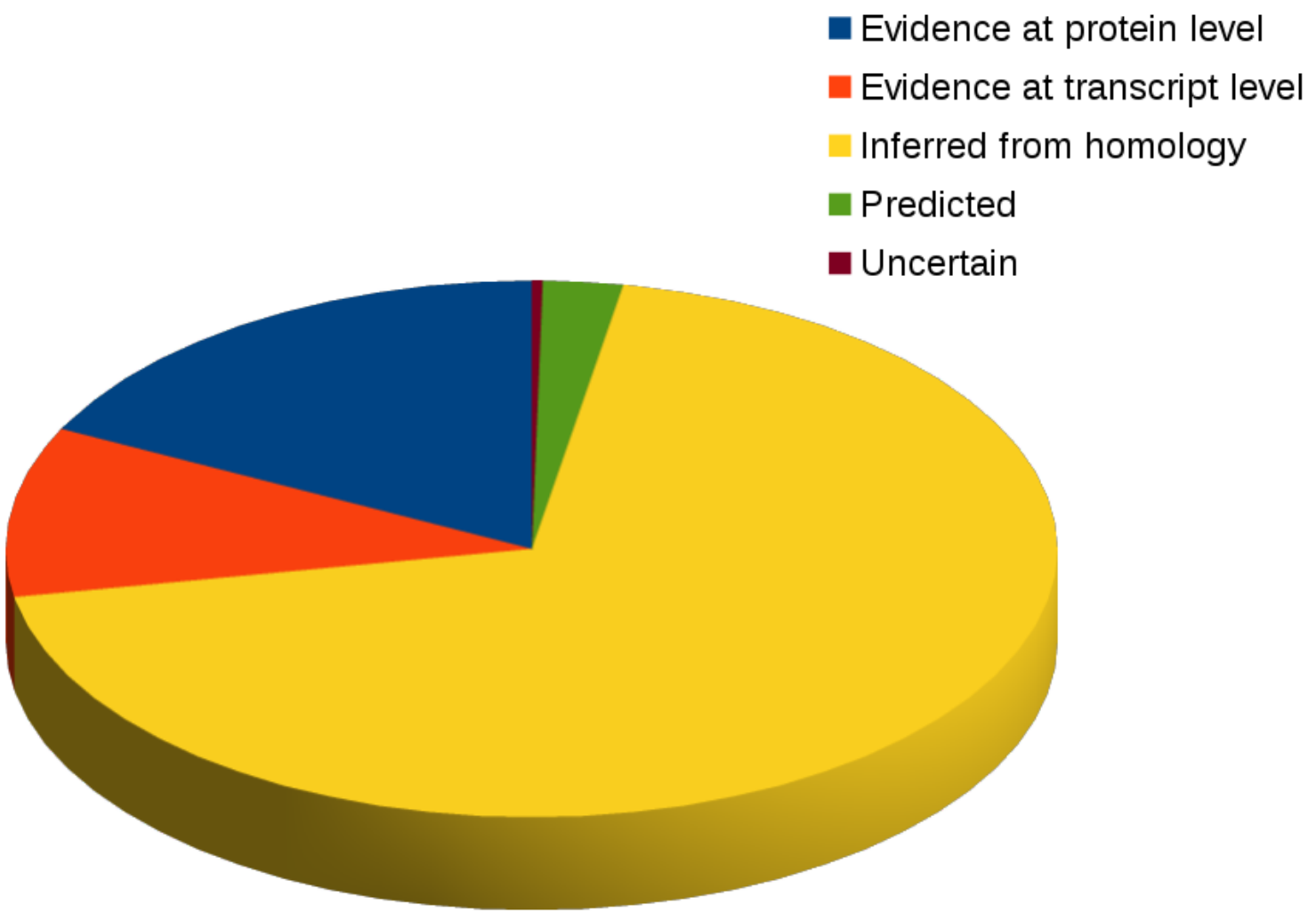
Соотношение числа белков,
представленных в разных банках



Последовательностей во много раз больше, чем структур!

Большинство последовательностей не аннотированы!

Достоверность белков Swiss-Prot



Какая информация может быть указана в аннотации записи Uniprot?

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

В аннотации записи есть:

- Идентификаторы
- Даты
- Название и синонимы
- Организм и таксономия
- Публикации
- CC:
  - Функция
  - Локализация в клетке
  - Биологический процесс
  - И др.
- Ссылки на записи этого белка из других БД
- Обоснования существования белка (Protein Evidence, PE) и его свойств
- Ключевые слова
- Особенности, привязанные к а.к.о. или участкам последовательности

Основные поля Swiss-Prot

ID – идентификатор в текущем релизе. Всегда один, но может меняться от релиза к релизу.

AC – так называемый «номер доступа» (Accession number). Раз появившись, не исчезнет (поэтому именно на AC надо указывать при использовании данных Swiss-Prot в публикациях). Может быть не один (по разным причинам).

DE – «description», описание белка. В последних релизах имеет внутреннюю структуру, т.е. делится на подполя (краткое рекомендуемое название, полное рекомендуемое название, синонимы и др.)

OS – видовое название организма – источника данного белка

OC – таксономия организма (в соответствии с текущим стандартом NCBI)

DR – ссылки на другие базы данных

FT – “feature table”, локальные особенности последовательности

http://www.uniprot.org/uniprot/P00174.txt
http://www.uniprot.org/uniprot/P37869.txt
http://www.uniprot.org/uniprot/P27358.txt

Структура идентификатора записи Swiss-Prot

ENO_BACSU: энолаза из сенной палочки

Мнемоника организма

Мнемоника функции белка

Как правило, мнемоника организма состоит из 3 букв родового названия и 2 букв видового (*Bacillus subtilis* → BACSU).

Для штаммов бактерий из видового названия берётся одна буква, а последний символ используется для различения штаммов.
Исключения:
а) 16 наиболее представленных организмов
(BOVIN for Bovine, CHICK for Chicken, ECOLI for *Escherichia coli*, HORSE for Horse, HUMAN for Human, MAIZE for Maize (*Zea mays*) , MOUSE for Mouse, PEA for Garden pea (*Pisum sativum*), PIG for Pig, RABIT for Rabbit, RAT for Rat, SHEEP for Sheep, SOYBN for Soybean (*Glycine max*), TOBAC for Common tobacco (*Nicotiana tabacum*), WHEAT for Wheat (*Triticum aestivum*), YEAST for Baker's yeast (*Saccharomyces cerevisiae*));
б) вирусы (например, BPP21 для фага P21, MEASY для штамма Yamagata вируса кори (measles) и пр.);

в) случаи неопределенного видового названия.

Содержимое поля FT

Feature Table — характеристики участков последовательности
В частности:

- трансмембранные участки;
- сигнальные последовательности
- сайты связывания разнообразных лигандов, ионов, нуклеиновых кислот;
- сайты посттрансляционной модификации;
- вторичная структура;
- домены;
- разночтения в последовательности (“CONFLICT”);
- варианты (напр., альтернативный сплайсинг “VARSPLIC”);

и т. п.
Имеет строгий формат: Feature Key, FtLocation, FtDescription.

Например:

```
FT DISULFID 334 343 By similarity.
FT CONFLICT 138 138 E -> EE (in Ref. 4; AA sequence).
```