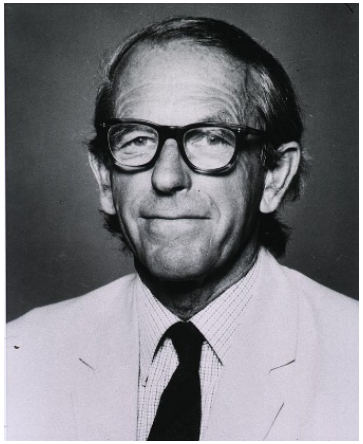
Базы данных последовательностей белков

Массовое определение последовательностей белков

- Выделение всей ДНК / РНК организма или образца из окружающей среды (тотальная ДНК / РНК)
- Секвенирование (в случае РНК сначала обратная транскрипция)
- Сборка генома / транскриптома из ридов (коротких прочтений)
- Предсказание кодирующих последовательностей (генов)
- Предсказание функций белков — продуктов генов

История секвенирования

Первая последовательность белка:
инсулин, цепи А и В.
Frederick Sanger, 1951, 1953. (**PMID: 14886310, 14886311, 13032078, 13032079**)
Нобелевская премия 1958.



F.Sanger 1918-2013

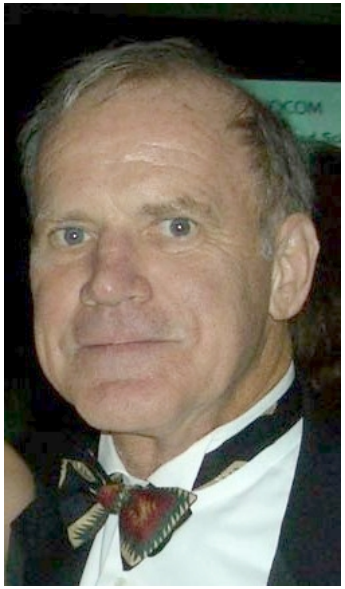
Первая последовательность РНК:
аланиновая т-РНК.
Robert Holley, 1964. (**PMID: 14299636**)
Нобелевская премия 1968.

R.Holley 1922-1993

История секвенирования

Метод секвенирования «по Сенгеру»:
первый полный геном ДНК бактериофага φX174
Frederick Sanger, 1977. (**PMID: 271968**)
Нобелевская премия (вторая!) 1980.

Полимеразная цепная реакция (ПЦР).
Kary Mullis, 1985. (**PMID: 2999980**)
Нобелевская премия 1993.



K.Mullis 1944

История секвенирования

Метод дробовика (Shotgun sequencing):
первый геном бактерии (*Haemophilus influenzae Rd*)
Fleischmann et al., 1995. (**PMID: 7542800**)

Новое поколение секвенаторов (Next-Generation Sequencing, NGS):

- Roche/454 (2004)
- Illumina Solexa (2006)
- SOLiD (2007)
- Pacific Biosciences SMRT (2010)

Почти 30-кратное покрытие генома человека за день!

Хранение информации о геномах

Депонирование информации в базу данных:

- последовательности;
- аннотации (предсказания кодирующих последовательностей и функций белков).

Публикации о геноме или протеоме:

- происходит с задержкой, иногда на годы;
- более достоверная информация, так как проходит стадию рецензирования;
- всё равно может содержать ошибки.

Базы данных (общие принципы)

- 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.
- Записи из разных хранилищ (таблиц) ссылаются друг на друга.

Типы баз данных

- Архивные:

за содержание записи отвечает её автор-экспериментатор

NCBI GenBank

ENA European Nucleotide Archive

DDBJ DNA Data Bank of Japan

- Курируемые:

за содержание записей отвечают специальные люди — кураторы

SIB swissprot EBI

- Автоматические:

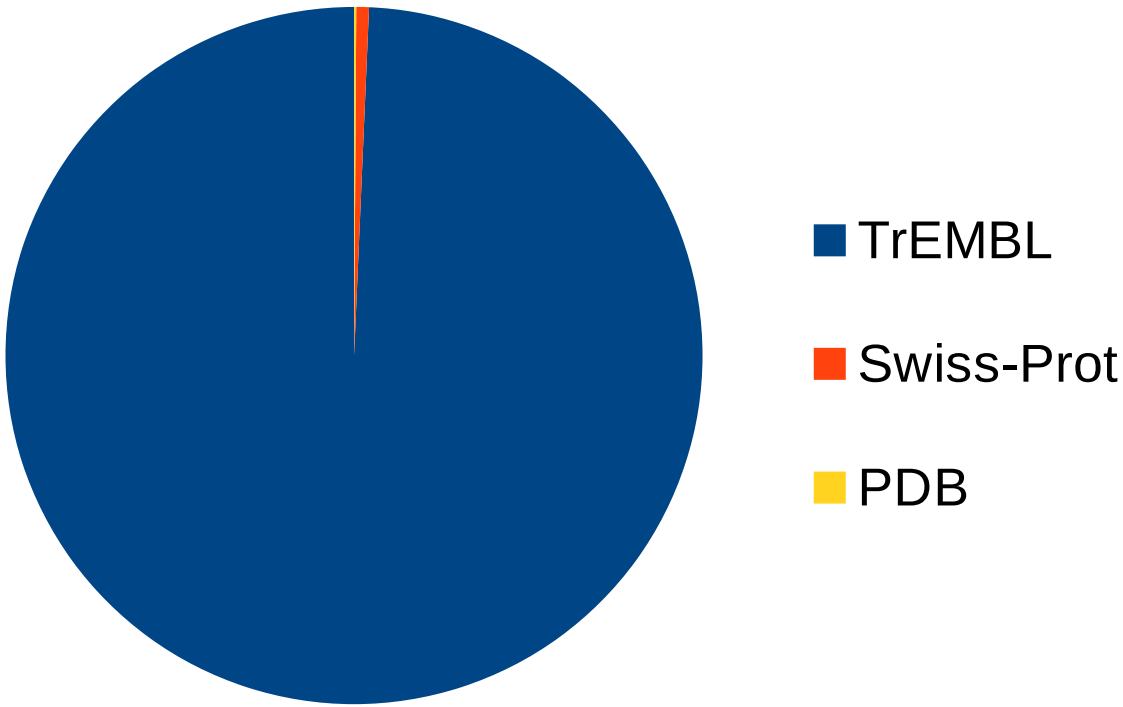
записи генерируются компьютерными программами

TrEMBL TRANSLATED EMBL

Основные базы данных о белках

- Uniprot — последовательности и аннотации
- RefSeq — последовательности и аннотации
- PDB — пространственные структуры
- PubMed — публикации

Соотношение числа белков в разных банках



PDB – пространственные структуры (всего их <100 тыс.)

Swiss-Prot – аннотированные последовательности

TrEMBL – последовательности, полученные компьютерной трансляцией

Последовательностей во много раз больше, чем структур.

Большинство последовательностей не аннотированы.

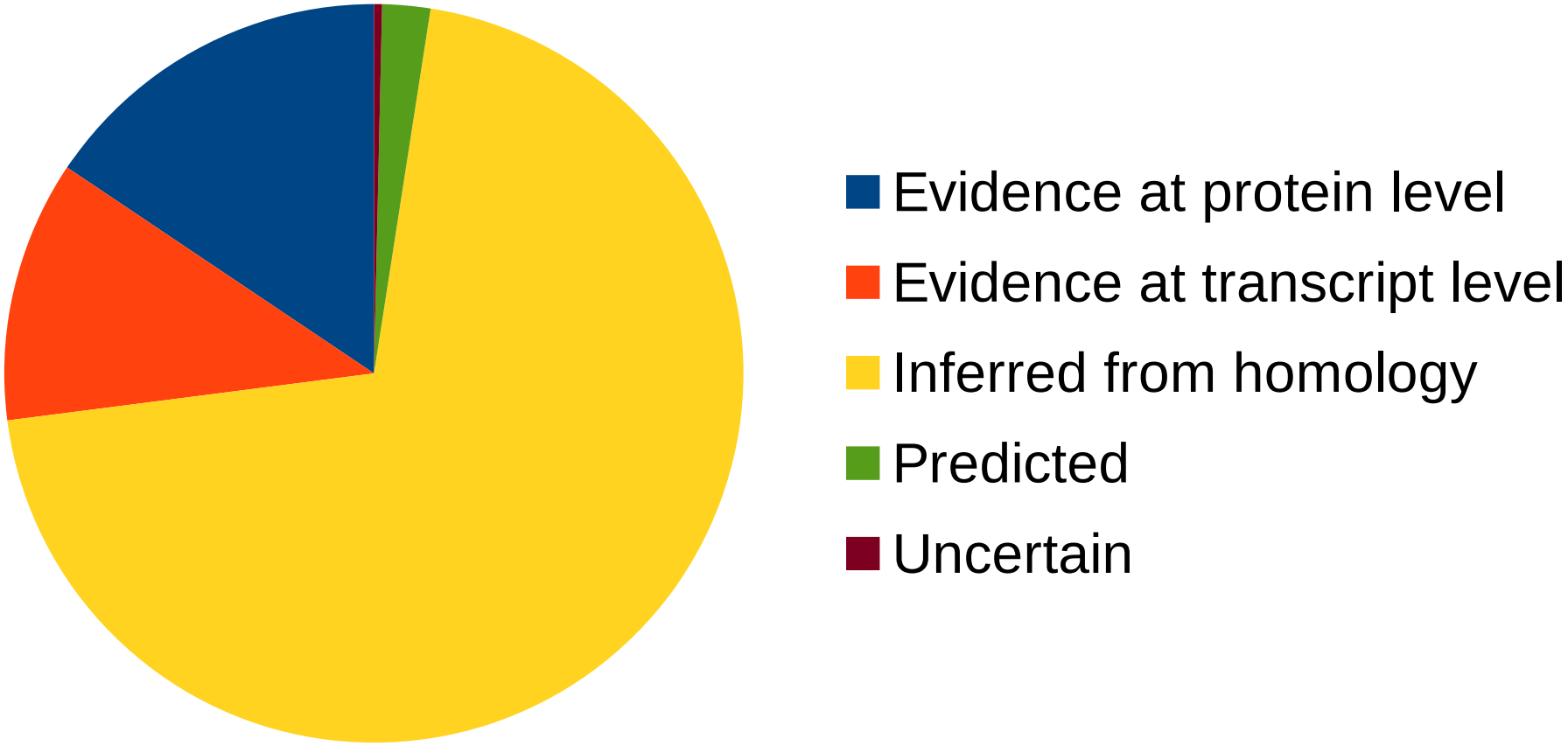
Последовательности белков

Большинство получены трансляцией предсказанных кодирующих участков в нуклеотидных последовательностях.

Правильность предсказаний проверяется:

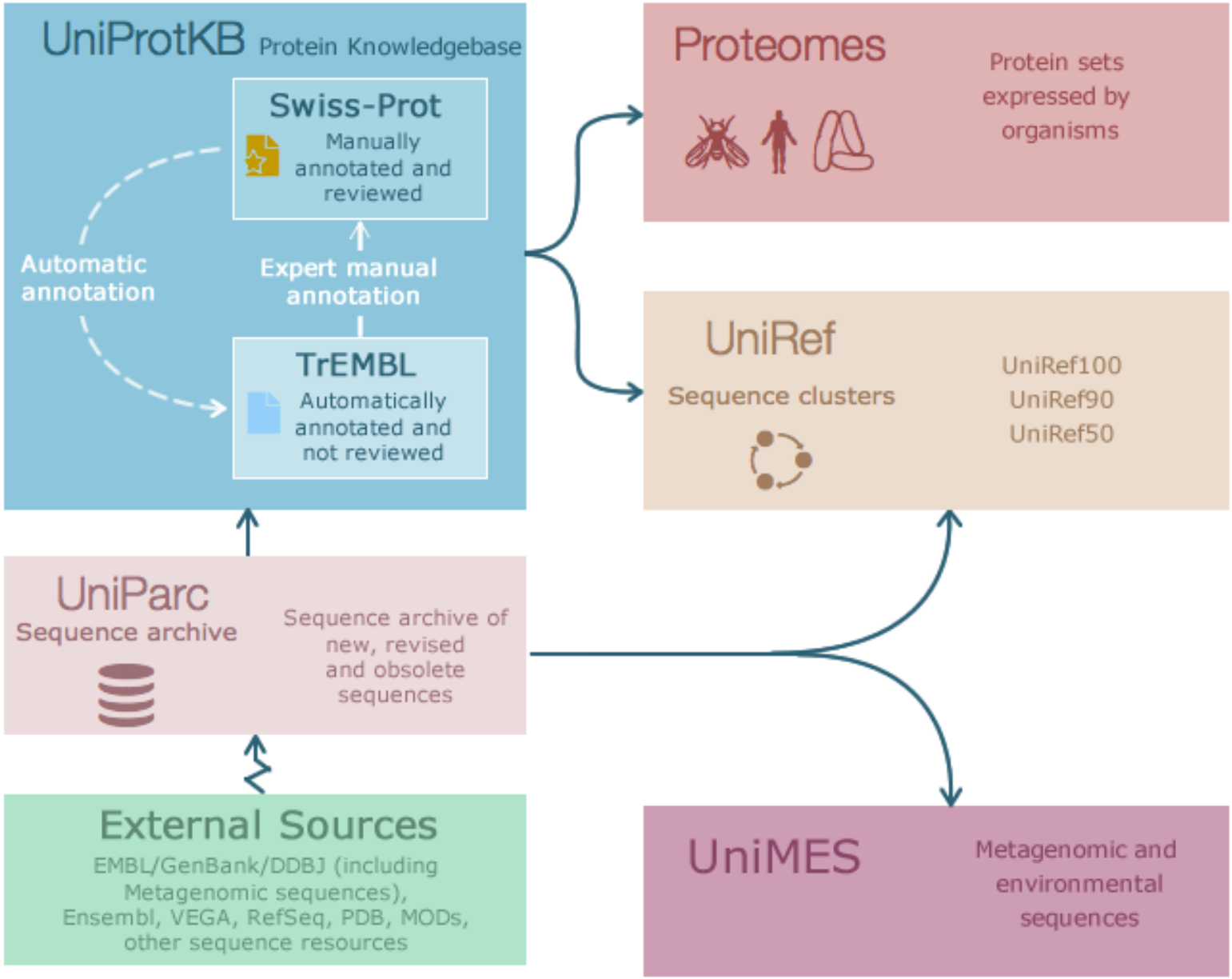
- лабораторными исследованиями конкретных белков (долго);
- масс-спектрометрией протеома (сложно);
- секвенированием тотальной РНК или конкретной мРНК;
- сходством последовательностей с последовательностями известных белков.

Достоверность последовательностей белков



Более половины последовательностей Swiss-Prot не охарактеризовано экспериментально.

Структура UniProt



UniProtKB/TrEMBL

TrEMBL — **Tr**anslated **EMBL**

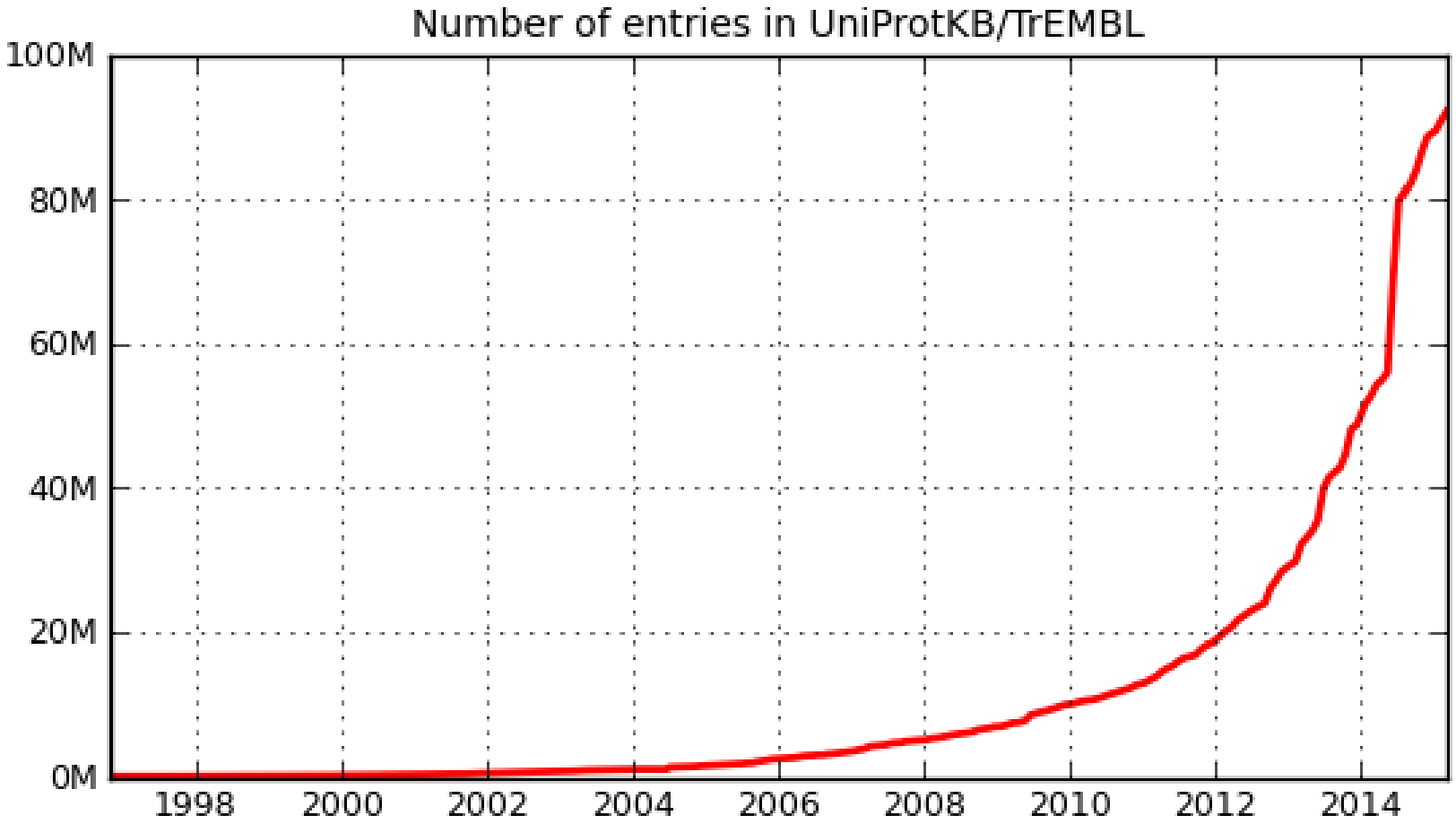
Формальная трансляция всех кодирующих нуклеотидных последовательностей из ENA (банк EMBL).

Автоматическая классификация и аннотация.

Формат записи такой же, как у Swiss-Prot.

92124243 (92 млн) записей в начале 2015 года.

UniProtKB/TrEMBL
рост числа записей



UniProtKB/Swiss-Prot

Swiss-Prot – банк знаний о белковых последовательностях.

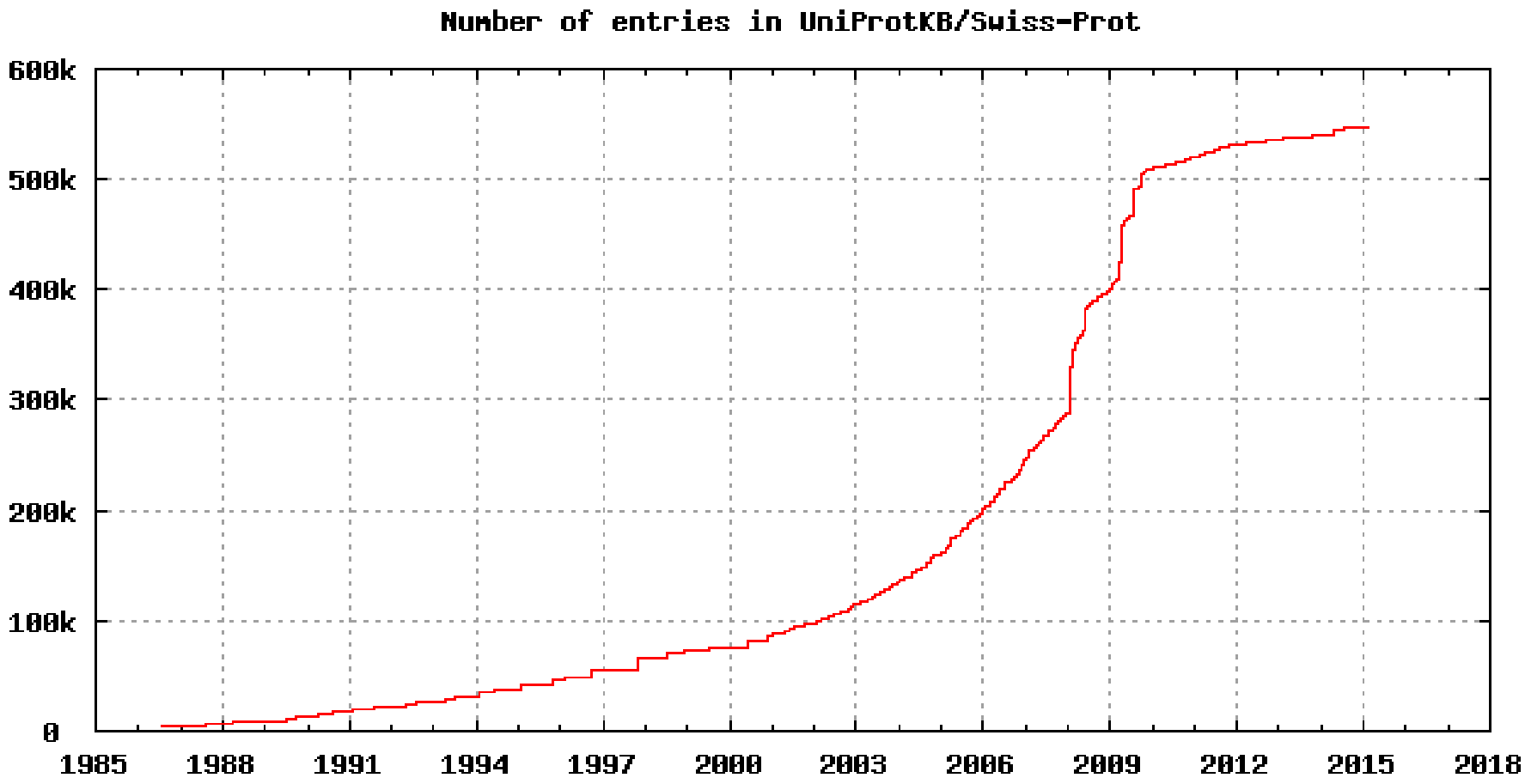
Ранее существовал как отдельный банк, сейчас – часть Uniprot.

Курируемая база данных: аннотации проверяет и дополняет эксперт.

- Теоретически, просматривает публикации.
- Практически, использует методы биоинформатики.

547599 (0.5 млн) записей в начале 2015 года.

UniProtKB/Swiss-Prot рост числа записей



Что такое один белок?

Одна запись **Uniprot** ~ продукт одного гена из одного вида или подвида.

В **RefSeq** в 2013 году появились последовательности белков, не привязанные строго к одному виду (идентификаторы начинаются с WP_...). В аннотации перечислены все геномы, кодирующие такую последовательность.

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.

Структура идентификатора записи Swiss-Prot

ENO_BACSU: енолаза из сенной палочки.

мнемоника функции белка

мнемоника организма

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

Содержание записи Swiss-Prot

Идентификаторы.

Даты.

Название и синонимы.

Организм и таксономия.

Публикации.

CC:

- функция;
- локализация в клетке;
- биологический процесс;
- и т.д.

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.

Содержание записи Swiss-Prot

```
ID   Y400_STACT              Reviewed;         188 AA.
AC   P47995; B9DJM1;
DT   01-FEB-1996, integrated into UniProtKB/Swiss-Prot.
DT   07-JAN-2015, entry version 64.
DE   RecName: Full=Uncharacterized protein Sca_0400;
GN   OrderedLocusNames=Sca_0400;
OS   Staphylococcus carnosus (strain TM300).
OC   Bacteria; Firmicutes; Bacilli; Bacillales; Staphylococcus.
OX   NCBI_TaxID=396513;
RN   [1]
RX   PubMed=19060169; DOI=10.1128/AEM.01982-08;
RA   Schuster S.C., Goetz F.;
RT   "Genome analysis of the bacterium Staphylococcus carnosus TM300.";
RL   Appl. Environ. Microbiol. 75:811-822(2009).
CC   -!- SIMILARITY: Belongs to the ribosomal protein S30Ae family.
CC       {ECO:0000305}.
CC   -!- SEQUENCE CAUTION:
CC       Sequence=CAA56161.1; Type=Frameshift; Positions=25, 46;
CC   ---------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ---------------------------------------------------------------------
DR   EMBL; AM295250; CAL27314.1; -; Genomic_DNA.
DR   PIR; S47148; S47148.
DR   RefSeq; WP_015899659.1; NC_012121.1.
DR   KEGG; sca:Sca_0400; -.
DR   eggNOG; COG1544; -.
DR   TIGRFAMs; TIGR00741; yfiA; 1.
PE   3: Inferred from homology;
KW   Complete proteome; Reference proteome.
FT   CHAIN         1    188       Uncharacterized protein Sca_0400.
FT                               /FTId=PRO_0000208591.
SQ   SEQUENCE   188 AA;  21886 MW;  25E12963AB332A5F CRC64;
     MIRFEIHGDN LTITDAIRNY IEDKVGKLER YFTNVPNVNA HVKVKTYANS STKIEVTIPL
     NDVTLRAEER NDDLYAGIDL ITNKLERQVR KYKTRVNRKK RKESEHEPFP ATPETPPETA
     GLIETVEN
//
```

Описание документа: идентификатор, имя, дата создания и модификации.

Аннотация.

Последовательность.

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 другие).

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", особенности частей последовательности.

Содержание поля FT

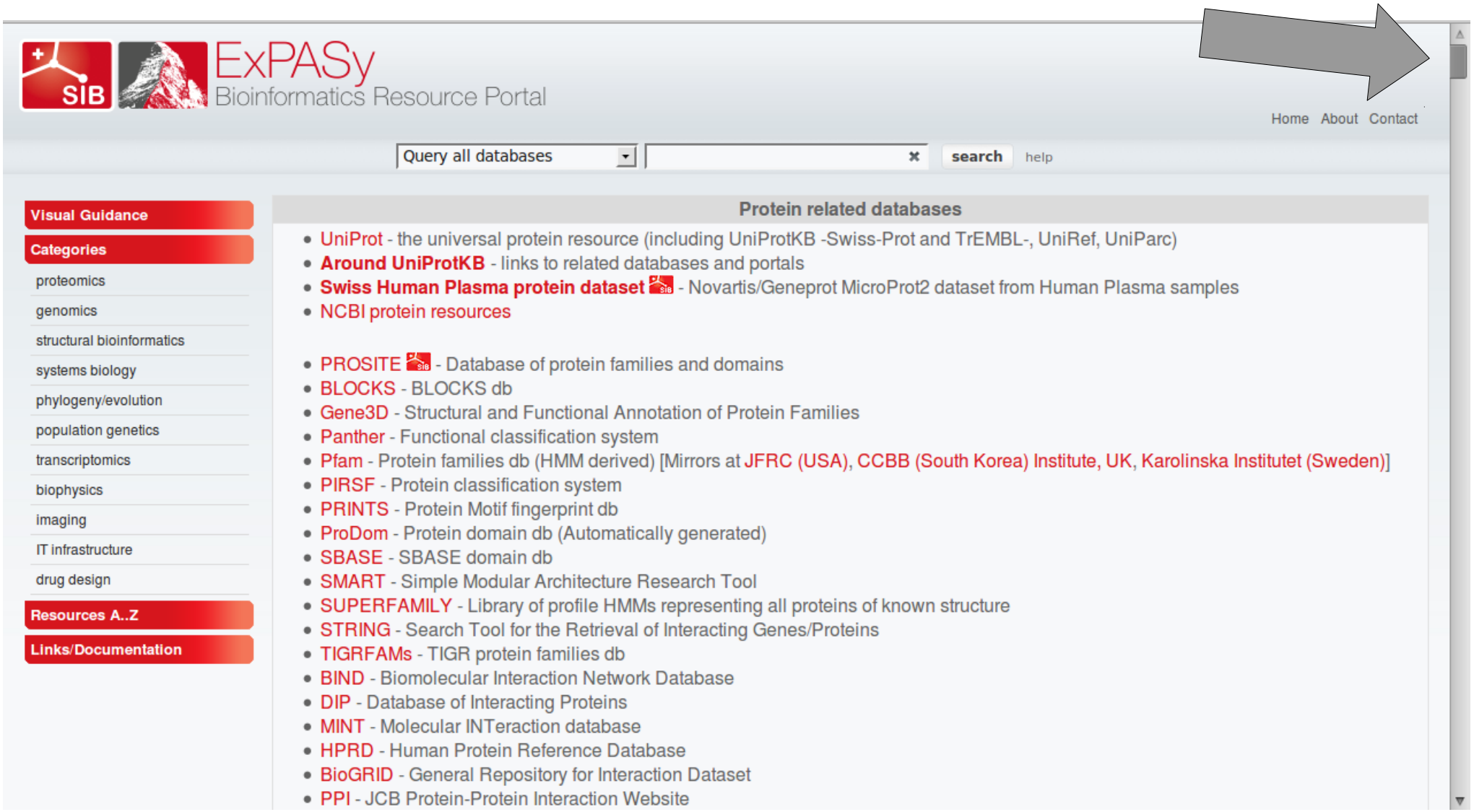
Feature Table — характеристики участков последовательности
В частности:

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

Имеет строгий формат: Key, Location, Description.

Пример:

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

Портал биоинформатический ресурсов на сайте SIB

http://www.expasy.org/old_links