

Банки последовательностей белков

UniProt

31/03/2020

Банки данных о белках

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

ExPASy (www.expasy.org)

The screenshot shows the ExPASy homepage with a light gray background. At the top left is the SIB logo (a red square with a white stylized mountain peak and the letters 'SIB') next to the text 'ExPASy Bioinformatics Resource Portal'. To the right are links for 'Home', 'About', and 'Contact'. A search bar at the top center contains the placeholder 'Query all databases' and a 'search' button. On the left, a vertical sidebar has red header sections for 'Visual Guidance' and 'Categories'. Under 'Categories', there's a list of life science fields: proteomics, genomics, structure analysis, systems biology, evolutionary biology, population genetics, transcriptomics, biophysics, imaging, IT infrastructure, medicinal chemistry, and glycomics. Below these are two more red header sections: 'Resources A..Z' and 'Links/Documentation'. In the center, a box titled 'Featuring today' highlights 'nfswatch' as an 'NFS traffic monitoring tool' with a [details] link. To the right, a large box titled 'How to use this portal?' features a question mark icon and a bulleted list: 'Features and updates', 'New to ExPASy', and 'Experienced ExPASy users: what is different'. At the bottom center is a graphic for '20 YEARS SIB' with a red background and white text. On the right side, there are two boxes: 'Popular resources' listing UniProtKB, SWISS-MODEL, STRING, and PROSITE; and 'Latest News' listing the UniProt Knowledgebase release (2018_02 - 2018-02-28), release notes, and protein spotlight on side effects.

ExPASy Bioinformatics Resource Portal

Home About Contact

Query all databases help

Visual Guidance

Categories

proteomics
genomics
structure analysis
systems biology
evolutionary biology
population genetics
transcriptomics
biophysics
imaging
IT infrastructure
medicinal chemistry
glycomics

Resources A..Z

Links/Documentation

Featuring today

nfswatch
An NFS traffic monitoring tool
[details]

How to use this portal?

- Features and updates
- New to ExPASy
- Experienced ExPASy users: what is different

Popular resources

- UniProtKB
- SWISS-MODEL
- STRING
- PROSITE

Latest News

UniProt Knowledgebase release
2018_02 - 2018-02-28

Release notes
556,825 UniProtKB/Swiss-Prot entries
(More..)
108,857,716 UniProtKB/TrEMBL entries
(More..)

Protein Spotlight: Side effects -
2018-02-23

Nature tiptoes along a sturdy yet fragile tightrope. DNA is its backbone and provides a basis from which every single living species on this planet emerges and prospers. Time, however, tampers with everything. Silver turns black...
[More](#).

[More news] [SIB news]

ExPASy proteomics

  ExPASy
Bioinformatics Resource Portal

Home About Contact

Query all databases help

Visual Guidance

Categories

proteomics

- protein sequences and identification
- proteomics experiment
- function analysis
- sequence sites, features and motifs
- protein modifications
- protein structure
- protein interactions
- similarity search/alignment

genomics

structure analysis

systems biology

evolutionary biology

population genetics

transcriptomics

biophysics

imaging

IT infrastructure

medicinal chemistry

glycomics

Resources A..Z

Links/Documentation

SIB resources

External resources - (No support from the ExPASy Team)

Databases

-  UniProtKB • functional information on proteins • [more]
-  UniProtKB/Swiss-Prot • protein sequence database • [more]
-  STRING • protein-protein interactions • [more]
-  SWISS-MODEL Repository • protein structure homology models • [more]
-  PROSITE • protein domains and families • [more]
-  ViralZone • portal to viral UniProtKB entries • [more]
-  neXtProt • human proteins • [more]

-  CAZy • Classification of carbohydrate-active enzymes • [more]
-  CSDB • Carbohydrate Structure Database • [more]
-  EMBnet services • bioinformatics tools, databases and courses • [more]
-  ENZYME • enzyme nomenclature • [more]
-  Glyco3D • 3D structures of glyco-related molecules • [more]
-  GlyConnect • Integrated glycodata platform • [more]
-  GlyTouCan • international glycan structure repository • [more]
-  HAMAP • UniProtKB family classification and annotation • [more]
-  iPtgxDBs • integrated proteogenomics search databases • [more]
-  MatrixDB • protein-glycosaminoglycan interactions • [more]
-  MetaNetX • Metabolic Network Repository & Analysis • [more]
-  MIAPEGelDB • MIAPE document edition • [more]
-  MyHits • protein domains database and tools • [more]
-  PaxDb • protein abundance database • [more]
-  Prolune • Popular science articles (in French) • [more]

Tools

-  SWISS-MODEL Workspace • structure homology-modeling • [more]
-  Vital-IT • life science informatics initiative • [more]
-  SwissDock • protein ligand docking server • [more]

-  2ZIP • Prediction of leucine zipper domains • [more]
-  3of5 • find user-defined patterns in protein sequences • [more]
-  AAComplidnt • protein identification by aa composition • [more]
-  AACompSim • amino acid composition comparison • [more]
-  Agadir • Prediction of the helical content of peptides • [more]
-  ALF • simulation of genome evolution • [more]
-  Alignment tools • Four tools for multiple alignments • [more]
-  APSSP • Advanced Protein Secondary Structure Prediction • [more]
-  Ascalaph • Molecular modeling software • [more]
-  big-PI • predict GPI modification sites • [more]
-  Biochemical Pathways • Biochemical Pathways • [more]
-  BLAST • sequence similarity search • [more]
-  BLAST (UniProt) • BLAST search on the UniProt web site • [more]
-  BLAST - NCBI • Biological sequence similarity search • [more]
-  BLAST - PBIL • BLAST search on protein sequence databases • [more]
-  Blast2Fasta • Blast to Fasta conversion • [more]
-  boxshade • MSA pretty printer • [more]
-  CFSSP • Protein secondary structure prediction • [more]
-  ChloroP • chloroplast transit peptides & cleavage sites • [more]

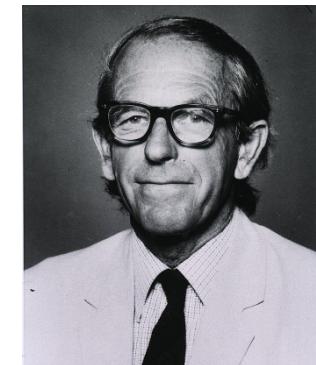
Откуда берется информация?

Экскурс в историю секвенирования

► Первая последовательность белка:
инсулин, цепи А и В

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

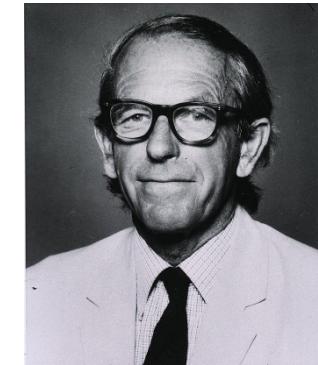


R.Holley 1922-1993

- ▶ Первый полный геном ДНК бактериофага фХ174

Frederick Sanger, 1977,
вторая нобелевская премия 1980

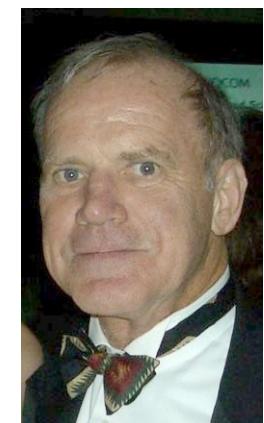
Метод секвенирования “по Сэнгеру”



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)

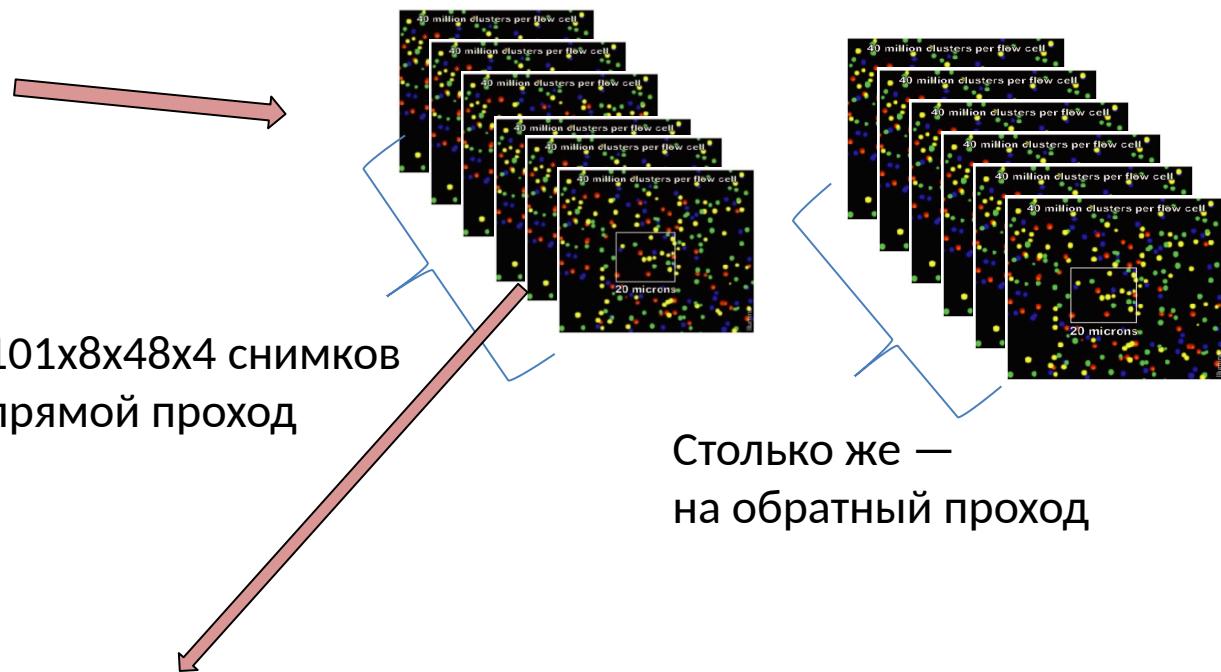
Pacific Biosciences SMRT (2010)

Roche/454 (2004)

Illumina HiSeq 2000



ДНК в пробирке



~**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 функций

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)

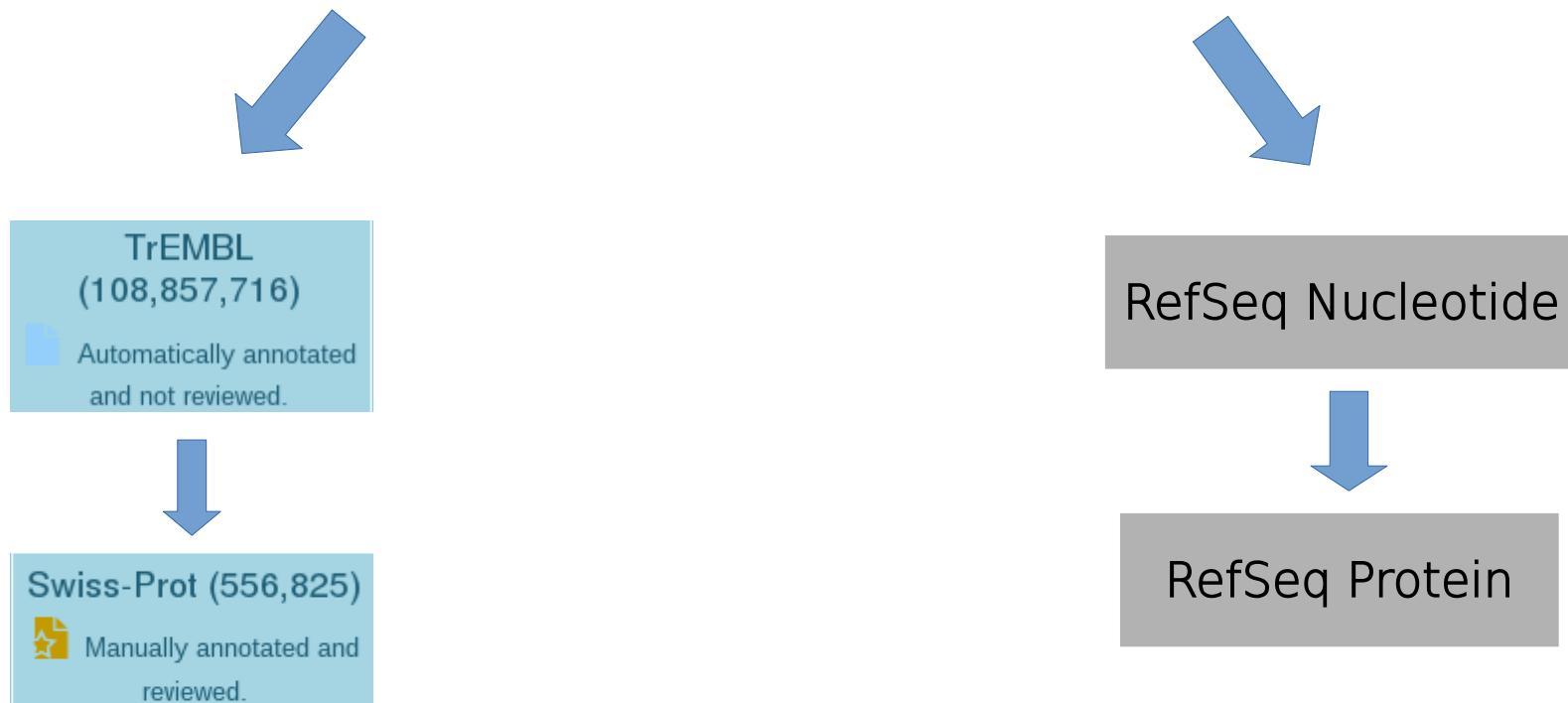
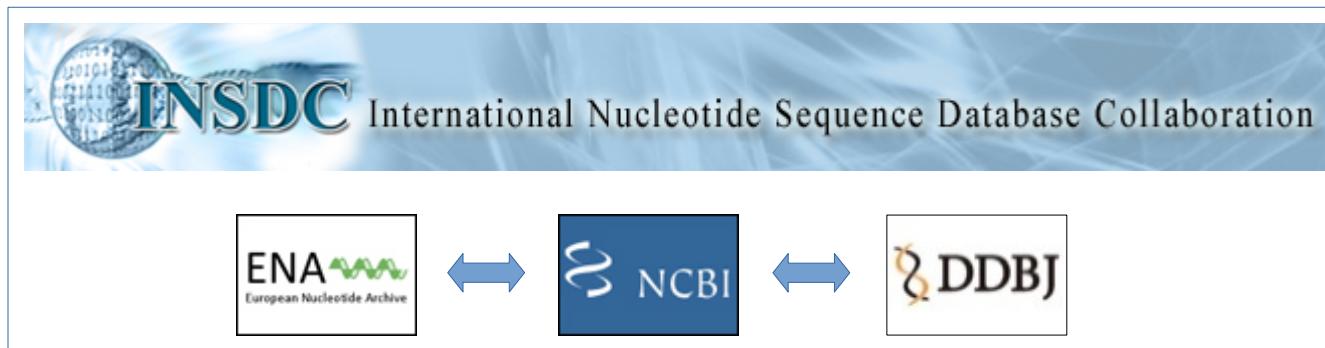
Курируемые – за содержание записи отвечает куратор (например, Swiss-Prot)

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





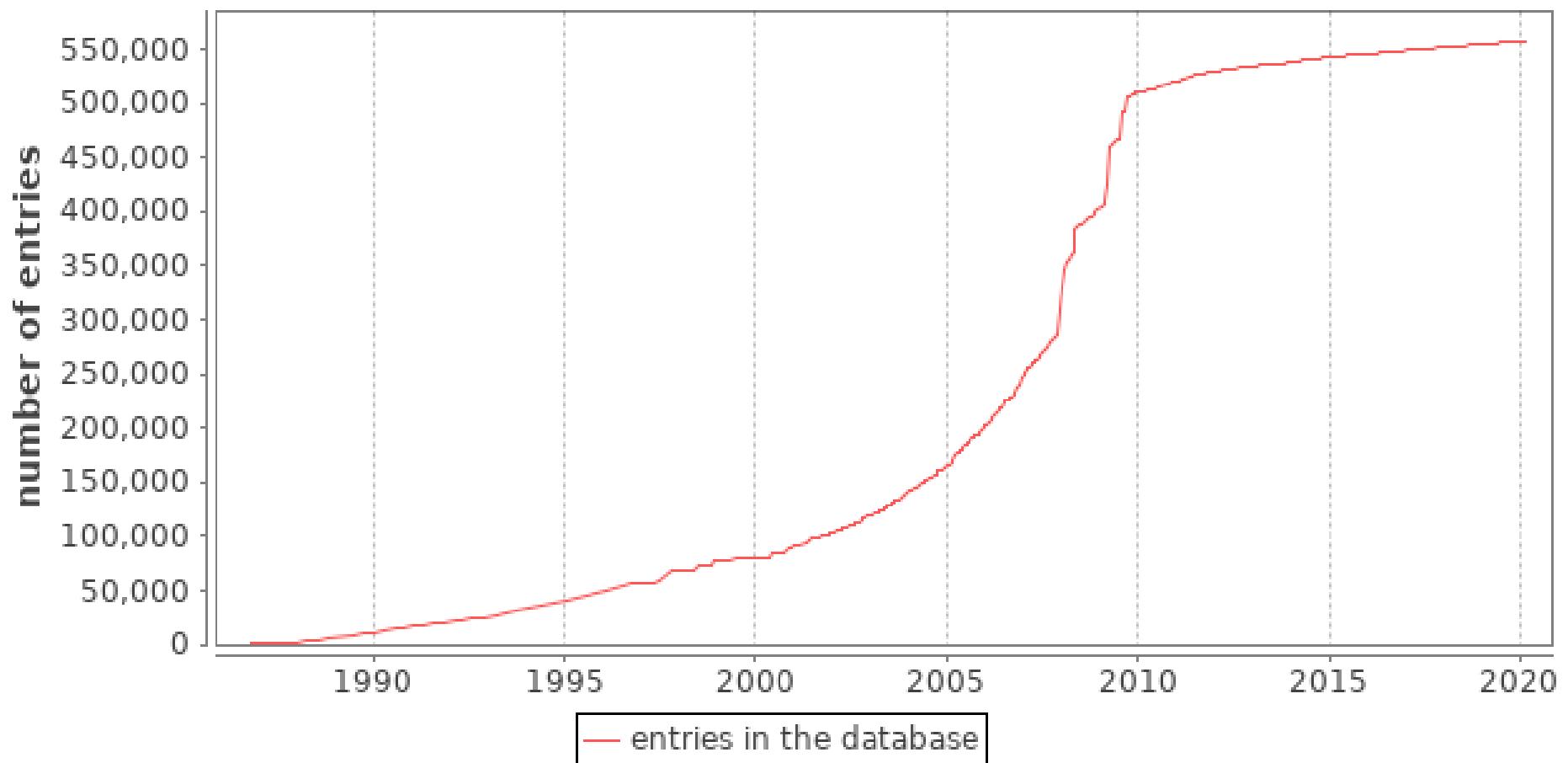
Swiss-Prot

Swiss-Prot – база знаний о последовательностях белков

- Ранее существовал как отдельный банк
- Сейчас – часть Uniprot
- Курируемая база данных
- Аннотации проверяет и дополняет эксперт: использует методы биоинформатики, просматривает публикации.
- 561 911 (половина миллиона) белков

Рост числа записей Swiss-Prot

Number of entries in UniProtKB/Swiss-Prot over time





TrEMBL

TrEMBL (Translated EMBL)

Формальная трансляция всех кодирующих
нуклеотидных последовательностей из банка EMBL

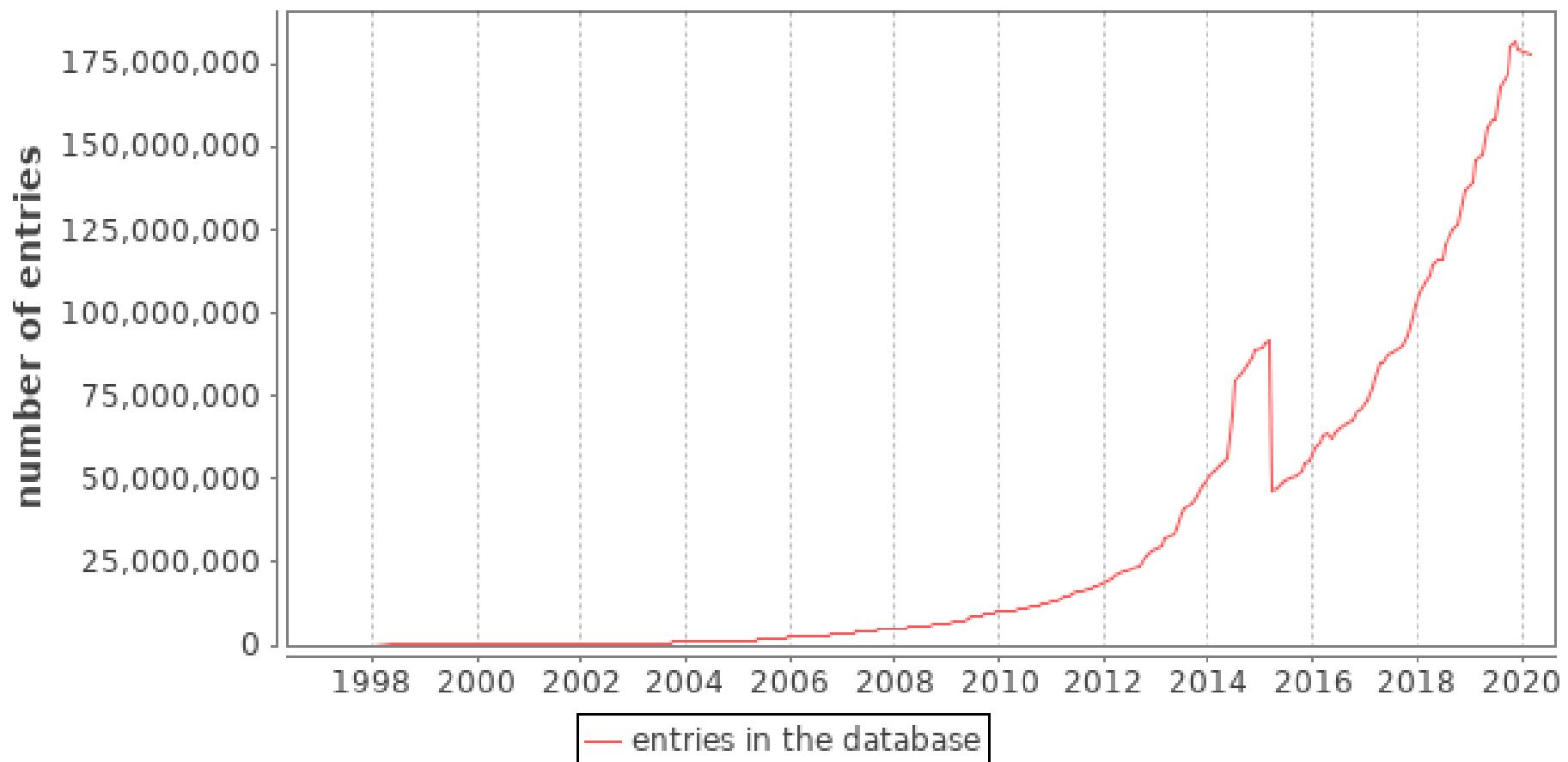
Автоматическая классификация и аннотация

Формат записи тот же, что у Swiss-Prot

177 754 527 белков на февраль 2020

Рост числа записей TrEMBL

Number of entries in UniProtKB/TrEMBL over time



Известны последовательности
десятков миллионов белков

Что такое “один белок”?

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

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_XXXXXX`. В аннотации перечислены геномы, в которых такая последовательность закодирована

Проблемы:

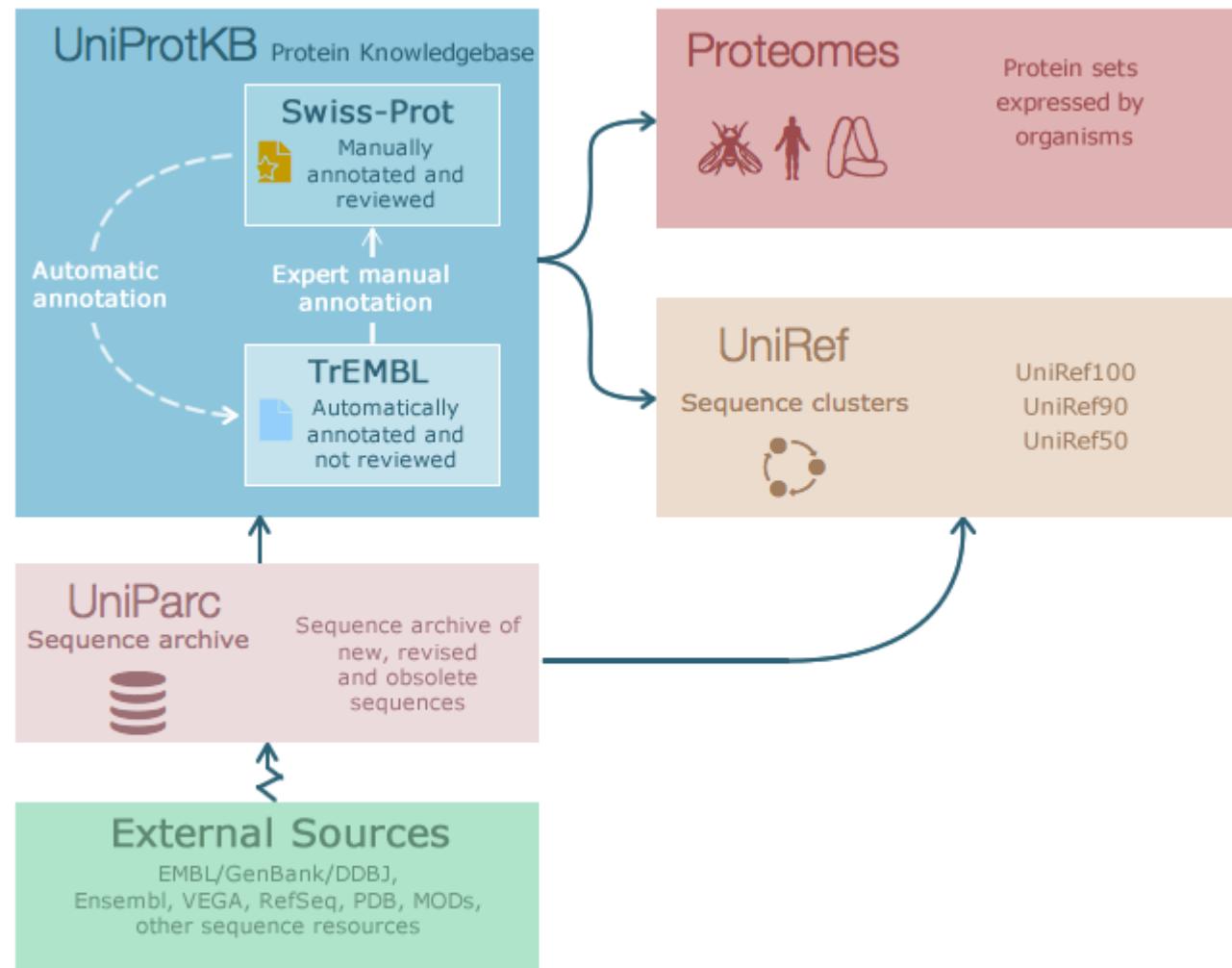
- Два гена из одного генома кодируют один и тот же белок (недавняя дупликация)
- Два гена из разных видов кодируют белки с одинаковой последовательностью
- Полиморфизм: последовательность белка из организма Пети отличается от таковой из организма Коли (или в штаммах бактерий)
- Альтернативный сплайсинг: 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

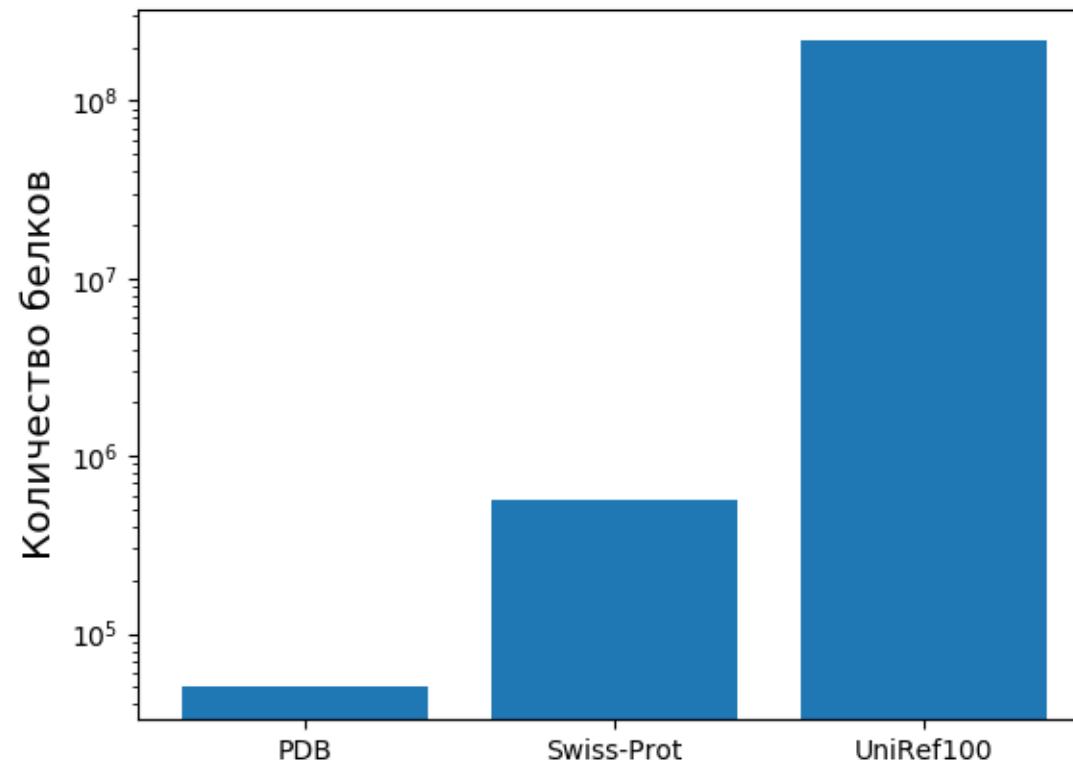
Одна и та же последовательность может попасть в банк несколько раз.

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

Структура UniProt

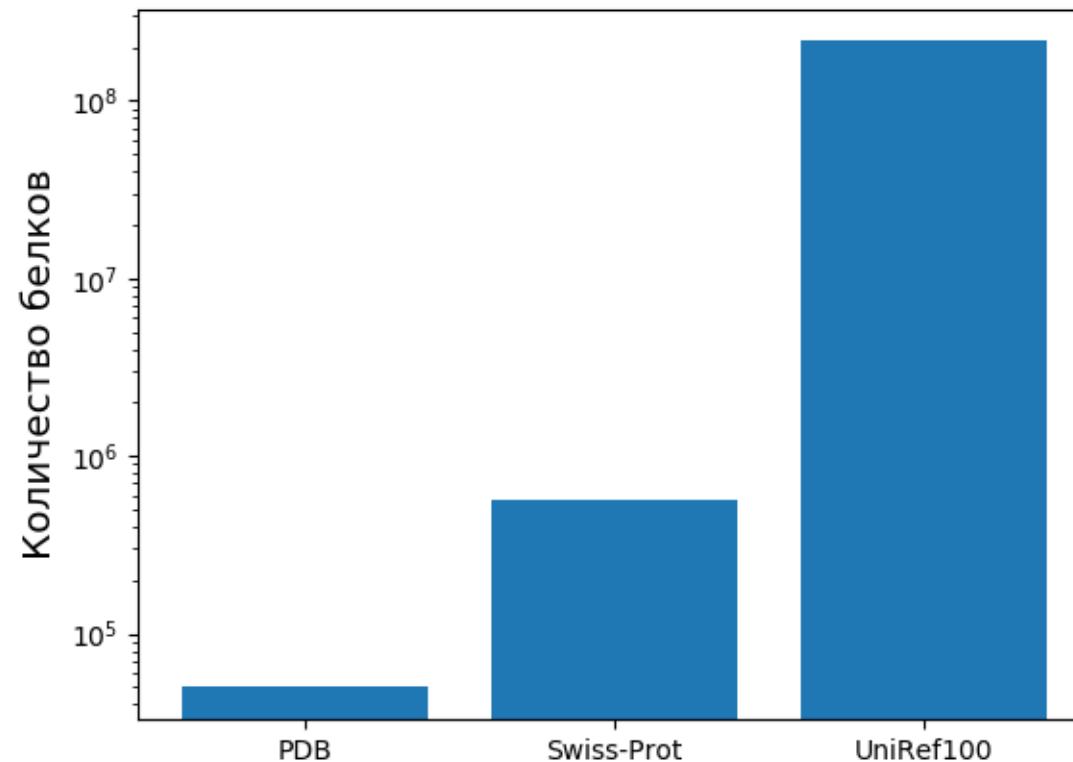


Число белков в разных банках

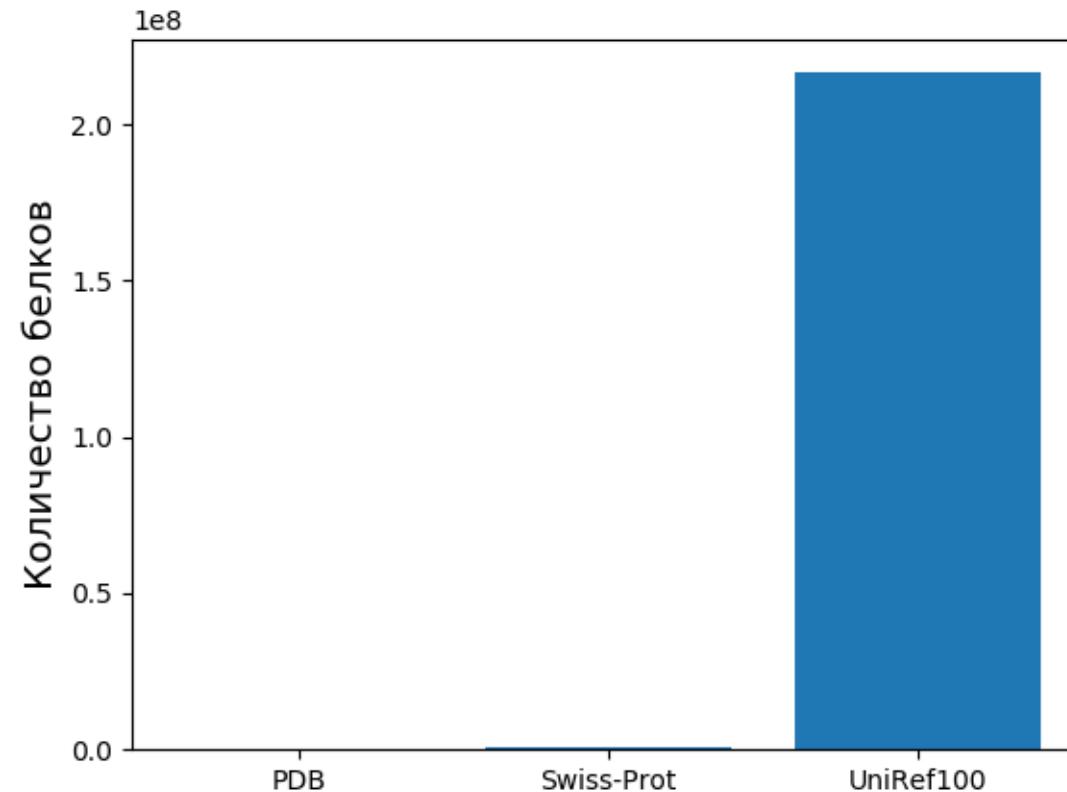


Число белков в разных банках

Логарифмическая шкала!

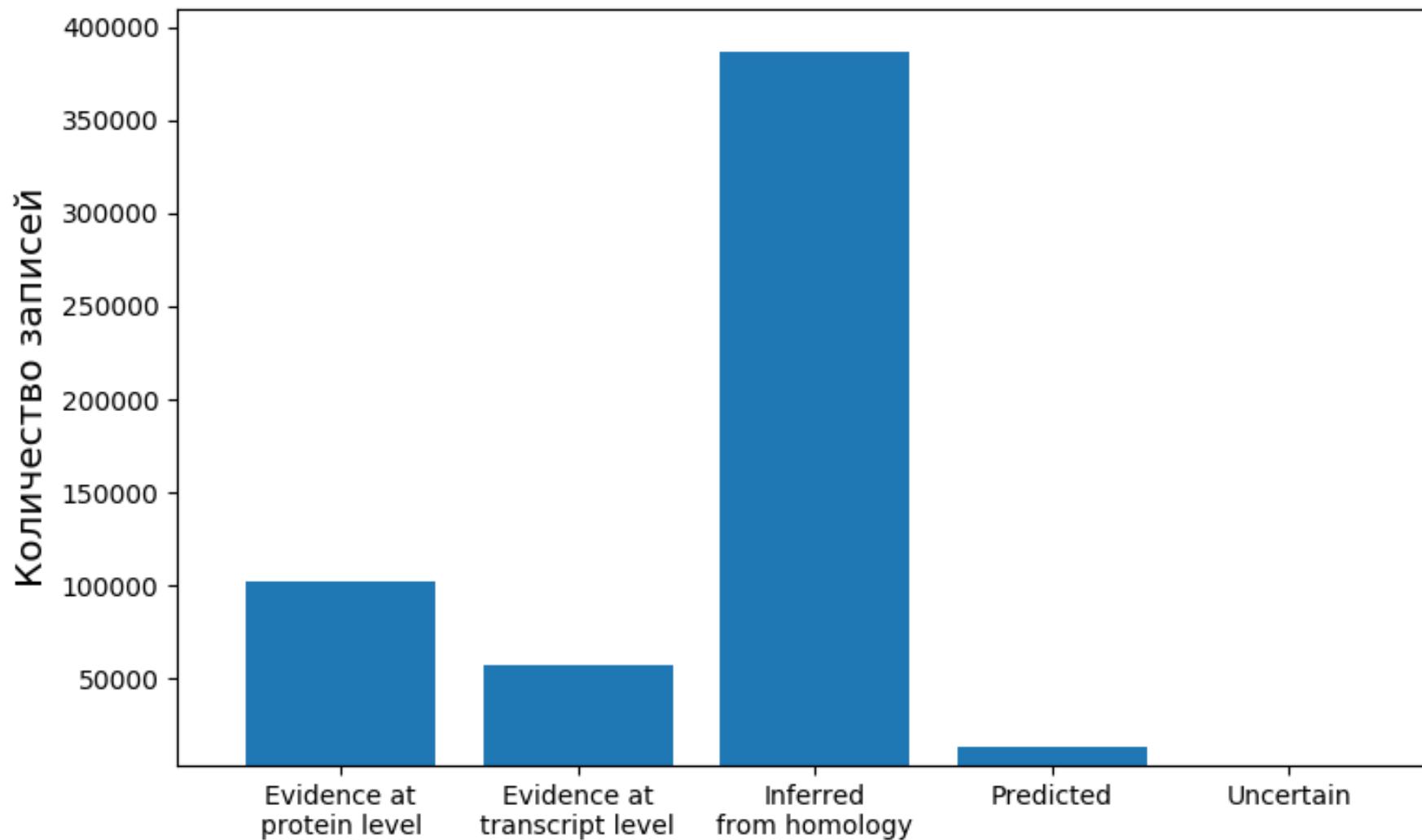


Число белков в разных банках



Последовательностей во много раз больше, чем структур.
Большинство последовательностей не аннотированы.

Достоверность белков Swiss-Prot



Какая информация может быть
указана в аннотации записи Uniprot?

Структура записи 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_S30SS54.  
DR  Pfam; PF02482; Ribosomal_S30AE; 1.  
KW  Hypothetical protein.  
FT  NON_TER    1    1  
SQ  SEQUENCE    165 AA;  19138 MW;  BF8CB91ADE194DD0 CRC64;  
LERYFTNVPN VNAHVKVKTG ANSSKIEVTI PLNDVTLRAE ERNDDIYAGI DKITNKLECQ  
VRKYKTRVNR KKRKESEHEP FPATPETPPE TAVDHDKDDE IEIIRSKQFS LKPMDSSEEAV  
LQMDLLGTDF FIFNDRETDG TSIVYRRKDQ KYGLIETVEK LICDI
```

Описание документа: идентификатор, имя, дата создания и модификаций

Аннотация последовательности

Последовательность

В аннотации записи есть:

- Идентификаторы
- Даты
- Название и синонимы
- Организм и таксономия
- Публикации
- СС:
 - 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>

Swiss-Prot ID

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 различия 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.

Feature table (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;
- разнотечения в последовательности (“CONFLICT”);
- варианты (напр., альтернативный сплайсинг “VARSPPLIC”);

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) .
```