Базы данных последовательностей белков

Откуда берутся последовательности белков
Прошлое

Pehr Edman

- 1950 — N-концевая деградация пептидов:

ди- и трипептиды

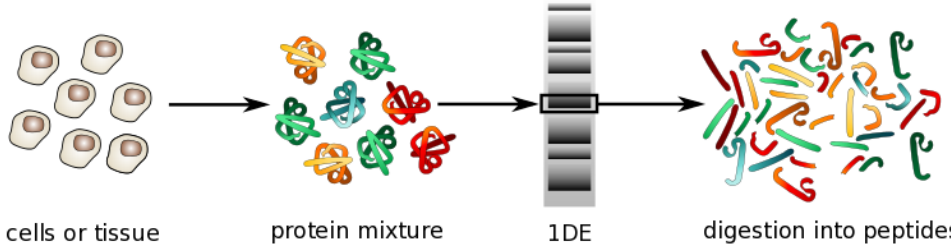
- 1951 и 1952 — Первая последовательность белка:

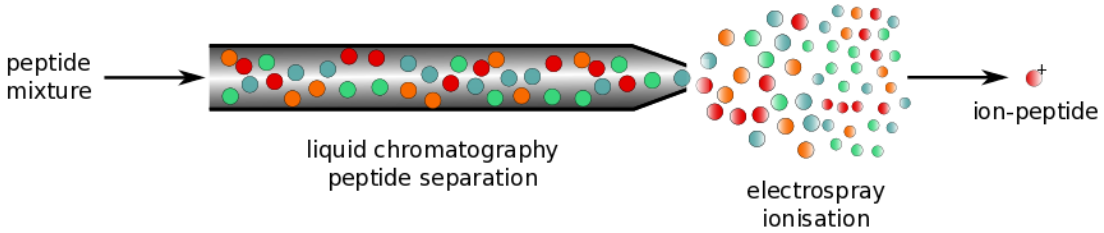
цепи B и A бычьего инсулина, 30 и 21 а.о.

- 1967 — Автоматизация метода Эдмана:

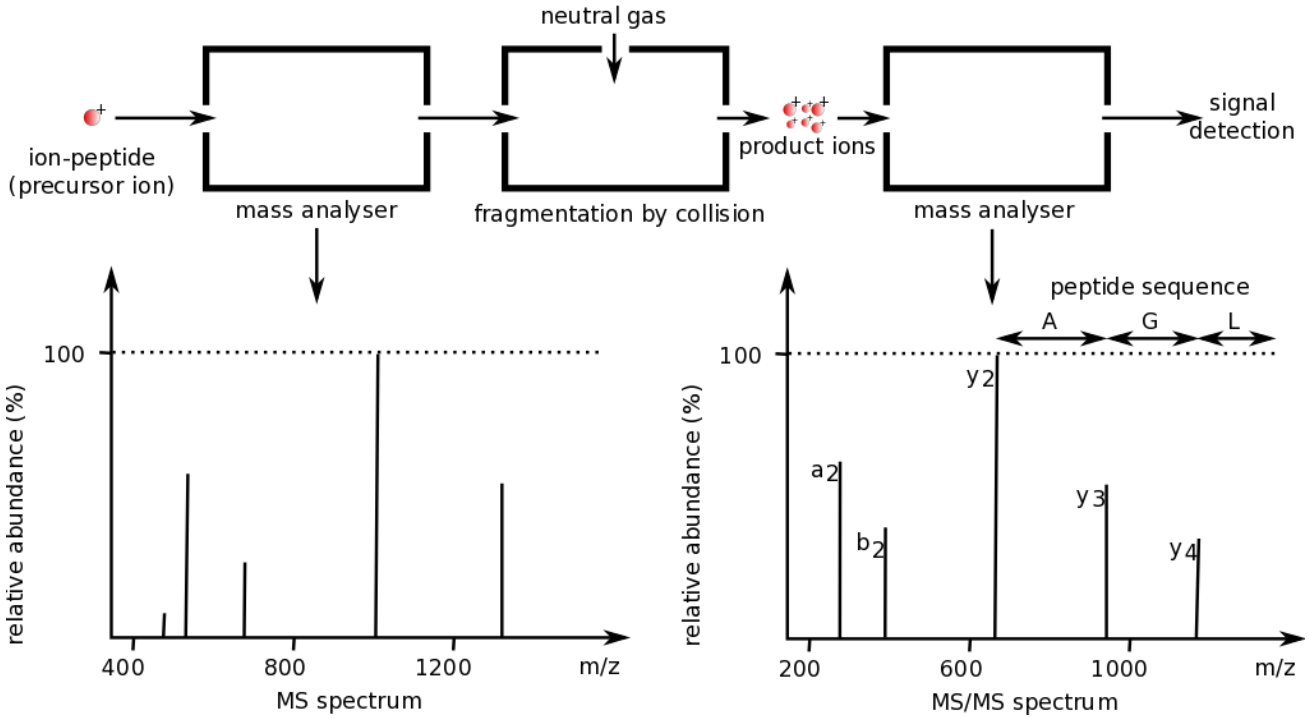
60 а.о. миоглобина кита

Откуда берутся последовательности белков
Настоящее





Белковая масс-спектрометрия



Откуда берутся последовательности белков
Настоящее

Автоматическая трансляция

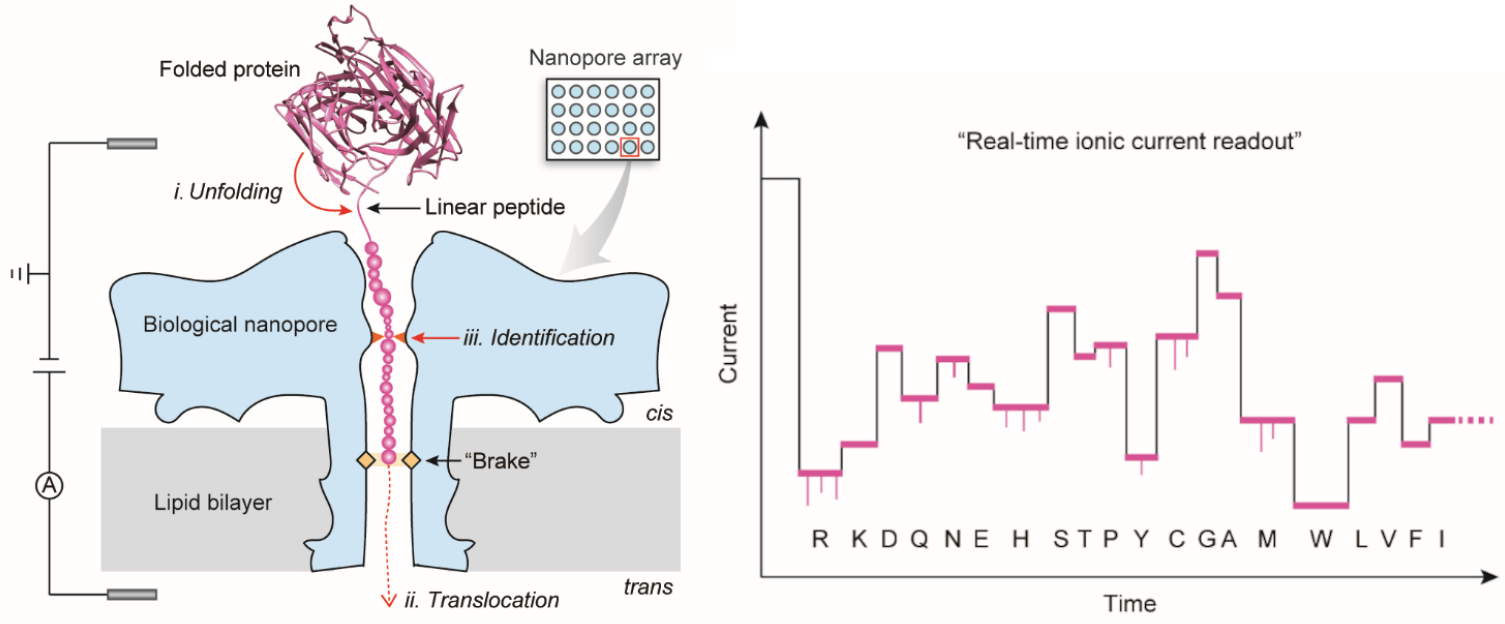
1. Получение нуклеотидной последовательности (геном, экзом, транскриптом, метагеном)
2. Предсказание открытых рамок считывания
3. Аннотация (в основном автоматическая) последовательностей по сходству с известными белками

Так получено подавляющее большинство последовательностей белков

Откуда берутся последовательности белков
Будущее ?

Активно развиваются новые подходы к массовому секвенированию белков:

- одномолекулярное нанопоровое секвенирование
- деградация иммобилизованных пептидов по Эдману с флуоресцентной детекцией продуктов
- протеолиз с помощью ClpX
- детекция на основе электронного туннелирования



Hu ZL, et al. Biological Nanopore Approach for Single-Molecule Protein Sequencing. Angew Chem Int Ed Engl. 2021;60(27):14738-14749

Как хранят последовательности белка?

Последовательность белка

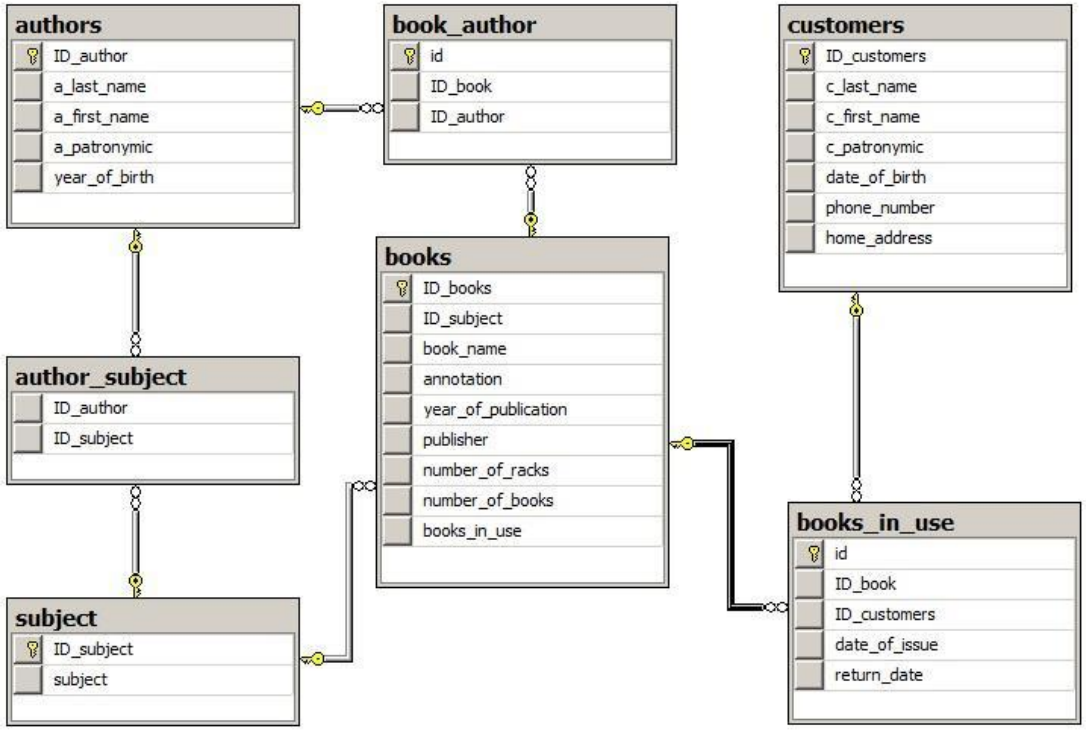
Последовательность аминокислотных остатков

- записанная от N-конца к C-концу,
- с использованием однобуквенных (реже трехбуквенных) обозначений аминокислот по IUPAC
- в виде текста (кодировка ASCII);
- остатки нумеруются, начиная с 1.

MLPGLALLLL AAWTARALEV PTDGNAGLLA EPQIAMFCGR LNMHMNVQNG KWDSDPSGTK
TCIDTKEGIL QYCQEVYPEL QITNVVEANQ PVTIQNWCKR GRKQCKTHPH FVIPYRCLVG
EFVSDALLVP DKCKFLHQER MDVCETHLHW HTVAKETCSE KSTNLHDYGM LLPCGIDKFR
GVEFVCCPLA EESDNVDSAD AEEDDSDVWW GGADTDYADG SEDKVVEVAE EEEVAEVEEE
EADDDEDDED GDEVEEEAEE PYEEATERTT SIATTTTTTT ESVEEVVREV CSEQAETGPC
RAMISRWYFD VTEGKCAPFF YGGCGGNRNN FDTEEYCMAV CGSAMSQSLL KTTQEPLARD
PVKLPTTAAS TPDAVDKYLE TPGDENEHAH FQKAKERLEA KHRERMSQVM REWEEAERQA
KNLPKADKKA VIQHFQEKVE SLEQEAANER QQLVETHMAR VEAMLNDRRR LALENYITAL
QAVPPRPRHV FNMLKKYVRA EQKDRQHTLK HFEHVRMVDP KKAAQIRSQV MTHLRVIYER
MNQSLSLLYN VPAVAEEIQD EVDELLQKEQ NYSDDVLANM ISEPRISYGN DALMPSLTET
KTTVELLPVN GEFSLDDLQP WHSFGADSVP ANTENEVEPV DARPAADRGL TTRPGSGLTN
IKTEEISEVK MDAEFRHDSG YEVHHQKLVF FAEDVGSNKG AIIGLMVGGV VIATVIVITL
VMLKKKQYTS IHHGVVEVDA AVTPEERHLS KMQQNGYENP TYKFFEQMQN

Реляционная база данных

[Реляционную] базу данных можно представить в виде **набора** ссылающихся друг на друга плоских **таблиц**, при условии, что строчки в каждой таблице уникальны, а порядок строк и столбцов не имеет значения.
Единица хранения называется **записью (entry)** и соответствует строке таблицы. Столбцы называют **полями (field)** или атрибутами. В ячейках записаны значения соответствующих полей.



Типы баз данных

На основании того, кто отвечает за достоверность информации, выделяют три типа баз данных.

- **Архивные**

  записи создают сами экспериментаторы, они же отвечают за достоверность информации (PDB, GenBank, ENA)

- **Курируемые**

  за создание и редактирование записей отвечают кураторы (Swiss-Prot, отчасти RefSeq)

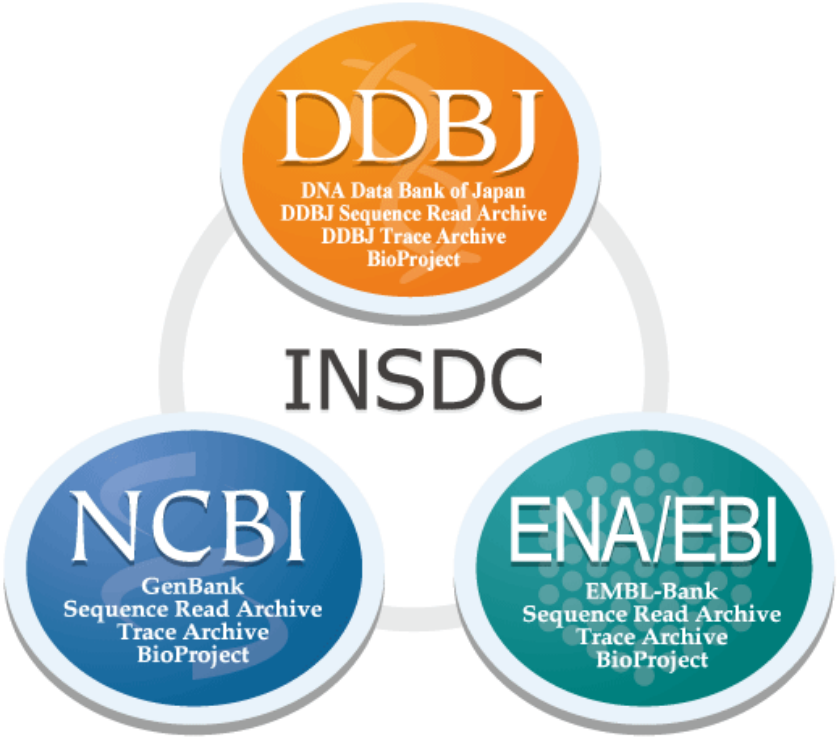
- **Автоматические**

  записи создаются компьютерными программами (TrEMBL, UniParc, основная часть RefSeq)

Потоки данных: INSDC

International Nucleotide Sequence Database Collaboration:
Объединяет 3 крупнейших нуклеотидных архива: GenBank, ENA, DDBJ

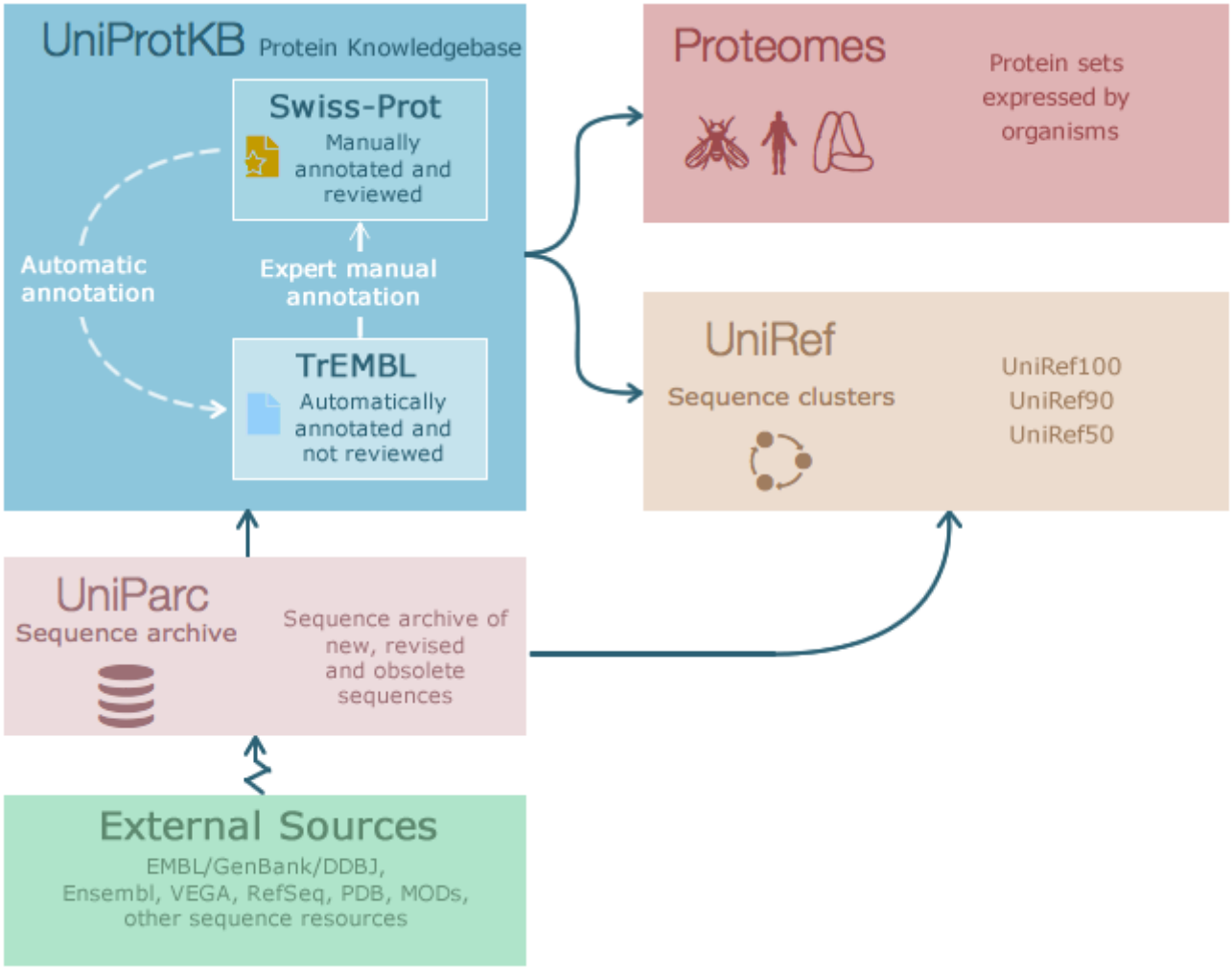
- Ежедневный обмен данными
- Единый формат таблицы локальных особенностей
- Рекомендации по использованию терминов и ключевых слов в аннотациях
- И некоторые прочие унификации (например, таблицы генетического кода и таксономия)



Потоки данных: RefSeq

- Автоматическая (по большей части) база данных на основе GenBank.
- Главная цель – уменьшение избыточности данных и унификация аннотаций.
- Часть записей курируется сотрудниками NCBI и не только (коллаборации со специализированными базами данных).
- Изначально нуклеотидная база, но создаются отдельные записи для закодированных белков (как и в GenBank).

Потоки данных: UniProt



UniParc (Uniprot Archive)

Архив уникальных последовательностей белков.

- Содержит все последовательности белков, которые когда-либо были в UniProtKB, и даже те, которые не были включены в UniProtKB по каким-либо причинам.
- Каждой уникальной последовательности присвоен идентификатор UPI, который никогда не изменяется и не удаляется.
- Запись UniParc содержит только последовательность, её хеш-сумму для проверки, ссылки на базы, в которых в какой-то момент времени хранилась такая же белковая последовательность и чуть-чуть вспомогательной информации.
- Последовательности (почти) неаннотированные.

Например, UPI0000000004

Sequence

Length 304
Mass (Da) 35,446

MD5 Checksum[i] A5A86E0883C185BC3D40A0FF0D49D9C7

```
        10         20         30         40         50         60         70         80         90        100        110        120        130
MPQQLSPINI ETKKAISNAR LKPLDIHYNE SKPTTIQNTG KLVRINFKGG YISGGFLPNE YVLSSLHIYW GKEDDYGSNH LIDVYKYSGE INLVHWNKKK YSSYEEAKKH DDGLIIISIF LQVLDHKNVY

       140        150        160        170        180        190        200        210        220        230        240        250        260
FQKIVNQLDS IRSANTSAPF DSVFYLDNLL PSKLDYFTYL GTTINHSADA VWIIFPTPIN IHSDQLSKFR TLLSSSNHDG KPHYITENYR NPYKLNDDTQ VYYSGEIIRA ATTSPARENY FMRWLSDLRE

       270        280        290        300
TCFSYYQKYI EENKTFAIIA IVFVFILTAI LFFMSRRYSR EKQN
```

Cross references

Customize columns

Database	Identifier	Version	Organism	First seen	Last seen	Active
UniProtKB reviewed	P04195	1	Vaccinia virus (strain Western Reserve) (VACV)	1988-11-01	2025-02-05	Yes
UniProtKB unreviewed	Q6LDV9	1	Vaccinia virus	2006-04-18	2025-02-05	Yes
UniProtKB unreviewed	V5R1H0	1	Vaccinia virus WAU86/88-1	2015-07-22	2025-02-05	Yes
UniProtKB unreviewed	A0A2I2MC48	1	Vaccinia virus (strain Western Reserve) (VACV)	2018-02-28	2025-02-05	Yes
UniProtKB unreviewed	A4GDG8	1	Vaccinia virus (strain Lister) (VACV)	2024-11-27	2025-02-05	Yes
UniProtKB unreviewed	Q76ZR9	1		2004-07-05	2011-10-19	No
RefSeq	YP_232995	1	Vaccinia virus	2005-10-06	2024-09-09	Yes
RefSeq	WP_016017665	1	Vaccinia virus	2021-05-22	2021-07-12	No

UniProtKB (UniProt Knowledgebase)

UniProtKB – две базы аннотированных белковых последовательностей с общим форматом записей.

- TrEMBL (от **Tr**anslated **EMBL**) – автоматическая база данных, содержащая, в основном, формальные трансляции открытых рамок считывания, предсказанных в нуклеотидных последовательностях.

- Swiss-Prot (раньше была отдельным банком) – курируемая база данных. Кураторы выбирают записи из TrEMBL, проверяют и дополняют их, переносят в Swiss-Prot.

UniRef (UniProt Reference Clusters)

Кластеры записей по сходству последовательностей.
UniProtKB + UniParc без ссылок на UniProtKB

- UniRef100 — идентичные на 100% последовательности и их фрагменты.
- UniRef90 — кластеры самых длинных представителей из кластеров UniRef100, идентичных на 90% и похожих по длине (не короче 80% самой длинной последовательности в кластере). Принадлежность кластеру UniRef90 распространяется на все остальные записи из кластера UniRef100 без проверок.
- UniRef50 — аналогично UniRef90.

Последовательности длины 10 и более короткие включены только в UniRef100, и кластеризуются только при совпадении длины.

Сид (seed)и репрезентативная последовательность

Seed – самая длинная последовательность в кластере, с которой сравниваются остальные последовательности для проверки принадлежности кластеру.

Representative – наиболее хорошо аннотированная последовательность, используется для аннотации кластера (название и длина последовательности).

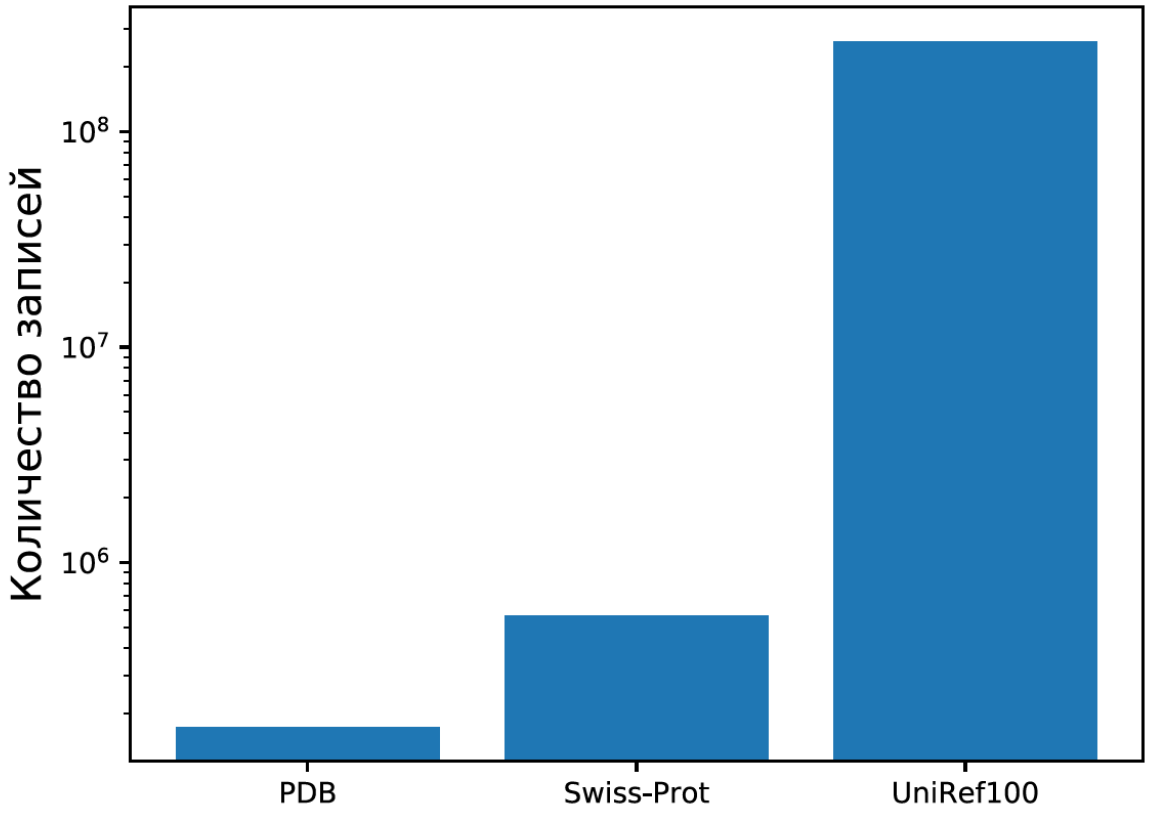
Чего только в кластерах не бывает ☺

кластер UniRef90_P81108

Cluster members	Entry name		Protein names	Organisms	Organism IDs	Related clusters	Length	Role
P81108	**THIO_CLOSG**		**Thioredoxin (Fragment)**	**Clostridium sporogenes**	1509	UniRef100_P81108	**40**	**Representative**
A0A1V9IK41	A0A1V9IK41_CLOSG		Thioredoxin	Clostridium sporogenes	1509	UniRef100_P81108	106	
A0A0B4W3E0	A0A0B4W3E0_CLOBO		Thioredoxin	Clostridium botulinum Prevot_594	1408284	UniRef100_P81108	106	
A0A1J1CWE3	A0A1J1CWE3_CLOBO		Thioredoxin	Clostridium botulinum	1491	UniRef100_A0A1J1CWE3	106	Seed
J7T6P1	J7T6P1_CLOS1		Thioredoxin	Clostridium sporogenes (strain ATCC 15579)	471871	UniRef100_A0A1J1CWE3	106	
A0A0D0ZXA6	A0A0D0ZXA6_CLOBO		Thioredoxin	Clostridium botulinum B2 450	1379739	UniRef100_A0A1J1CWE3	106	
A0A6M0YC23	A0A6M0YC23_CLOSG		Thioredoxin	Clostridium sporogenes	1509	UniRef100_A0A6M0YC23	106	
A0A1S9I145	A0A1S9I145_9CLOT		Thioredoxin	Clostridium tepidum	1962263	UniRef100_A0A1S9I145	106	
A0A6M0T4F3	A0A6M0T4F3_CLOBO		Thioredoxin	Clostridium botulinum	1491	UniRef100_A0A6M0T4F3	106	
A0A6M0XX80	A0A6M0XX80_CLOSG		Thioredoxin	Clostridium sporogenes	1509	UniRef100_A0A6M0T4F3	106	
A0A0M1IUU5	A0A0M1IUU5_9CLOT		Thioredoxin	Clostridium sp. L74	1560217	UniRef100_A0A0M1IUU5	106	
UPI000666DA61			thiol reductase thioredoxin	Clostridium sporogenes	1509	UniRef100_UPI000666DA61	106	
UPI000D0CC3E6			thiol reductase thioredoxin	Clostridium botulinum	1491	UniRef100_UPI000D0CC3E6	106	
UPI001748E097			thioredoxin	Clostridium botulinum	1491	UniRef100_UPI001748E097	106	
UPI0005F06029			thiol reductase thioredoxin	Clostridium sporogenes	1509	UniRef100_UPI0005F06029	106	

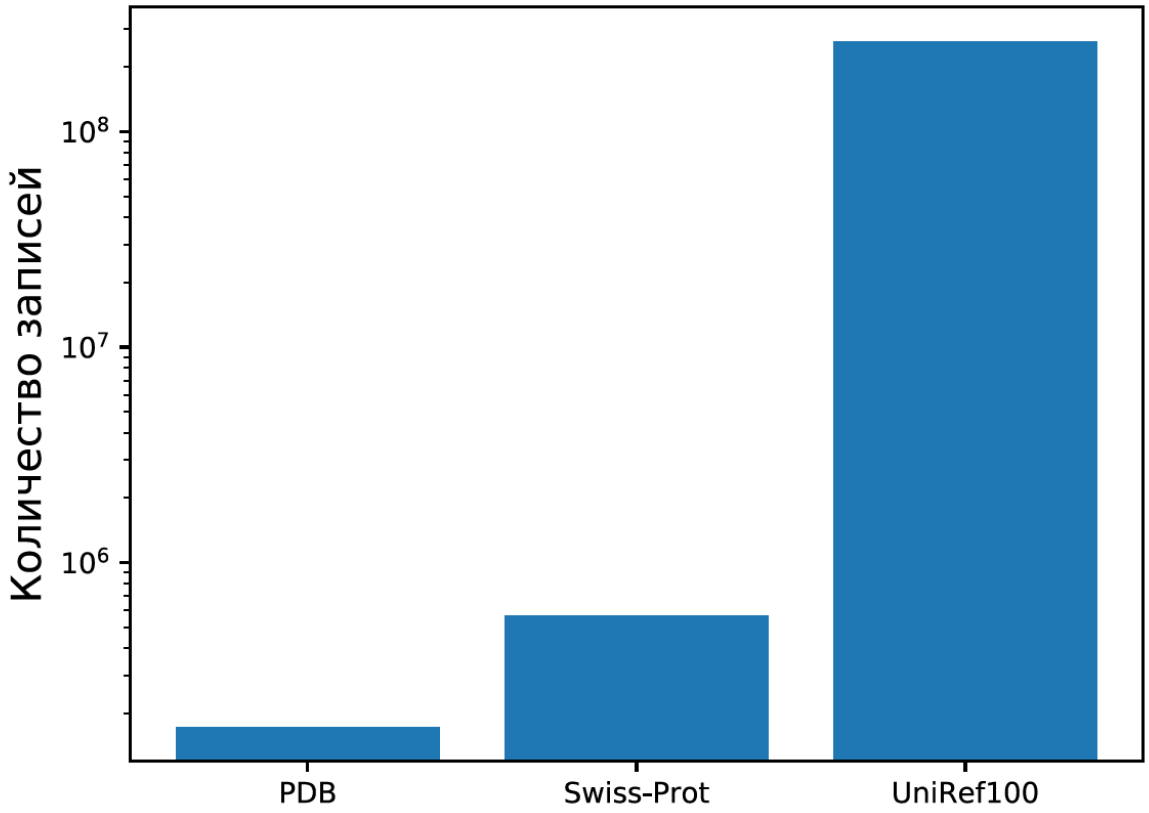
Вспомним , что говорили про источники последовательностей

Число записей про белки в разных базах данных



Вспомним , что говорили про источники последовательностей

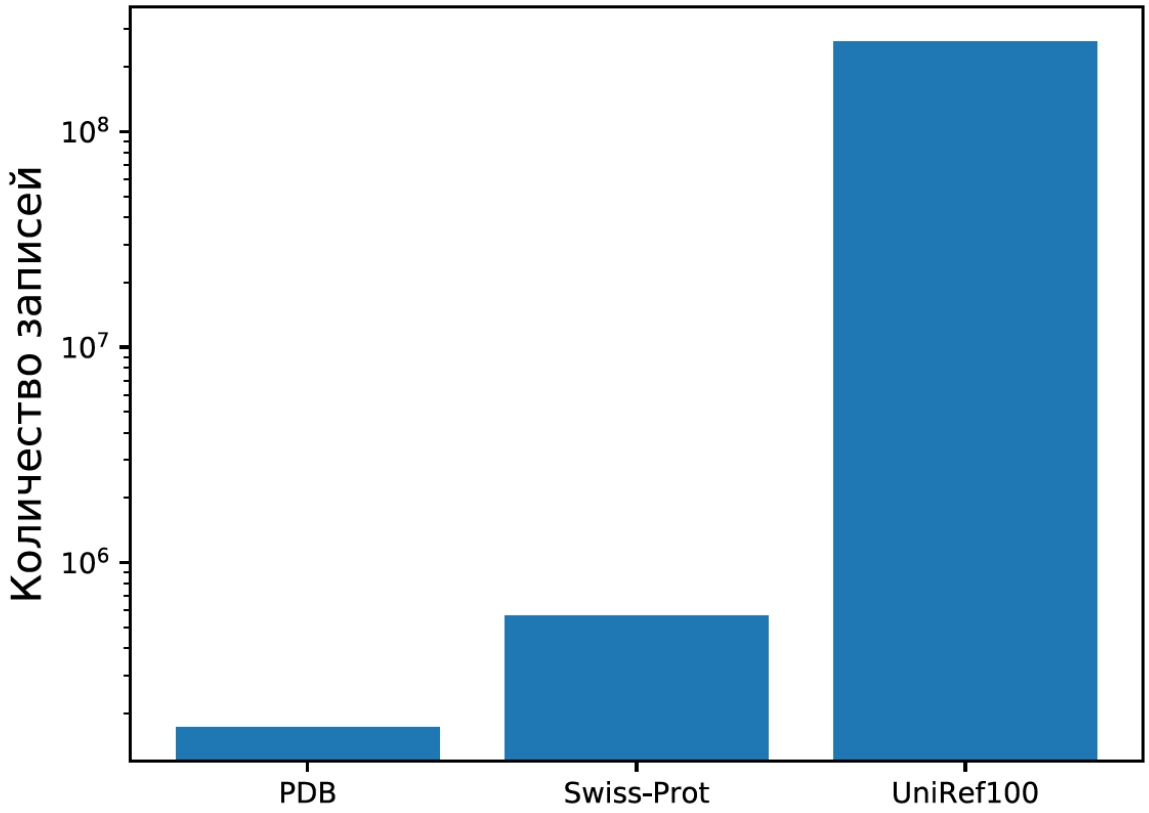
Число записей про белки в разных базах данных



Что самое страшное в этой диаграмме?

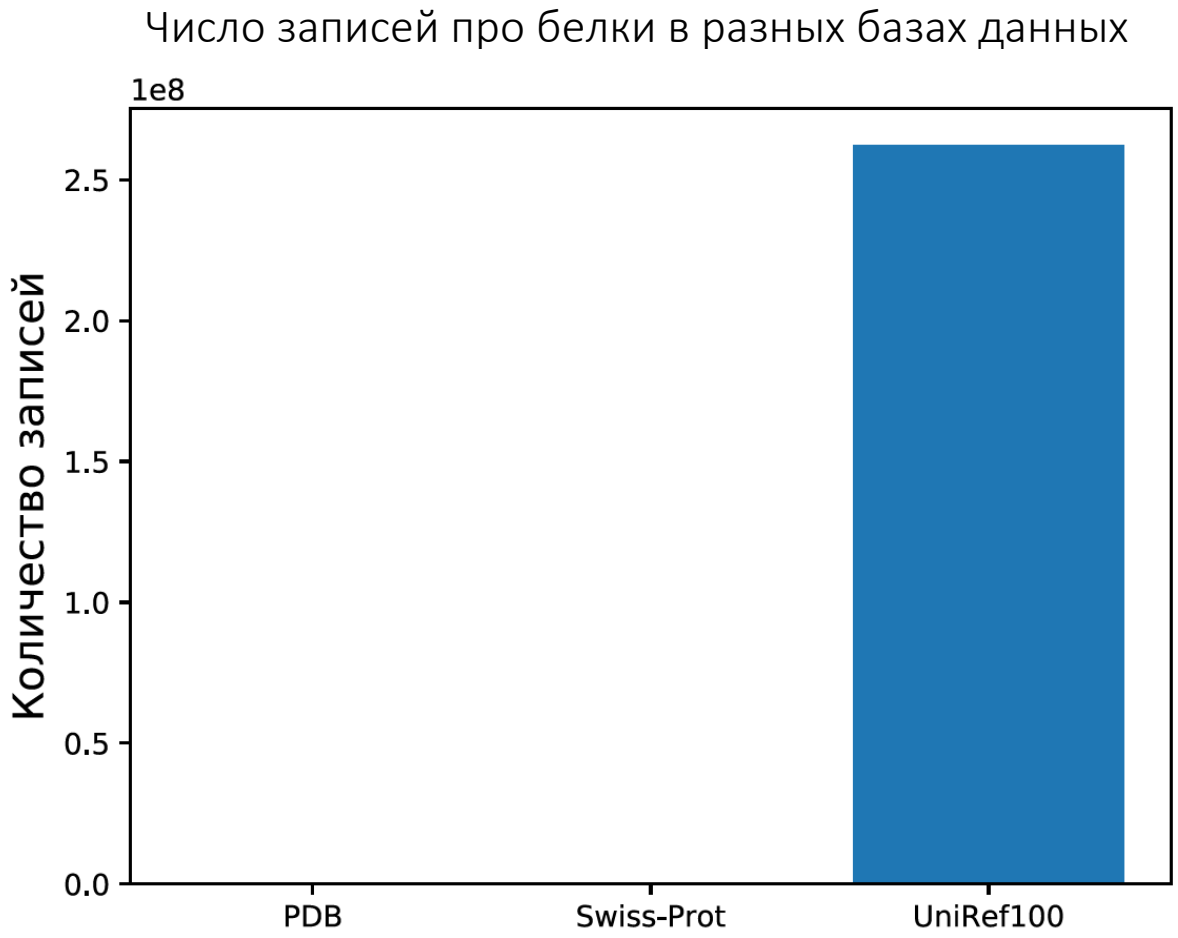
Вспомним , что говорили про источники последовательностей

Число записей про белки в разных базах данных



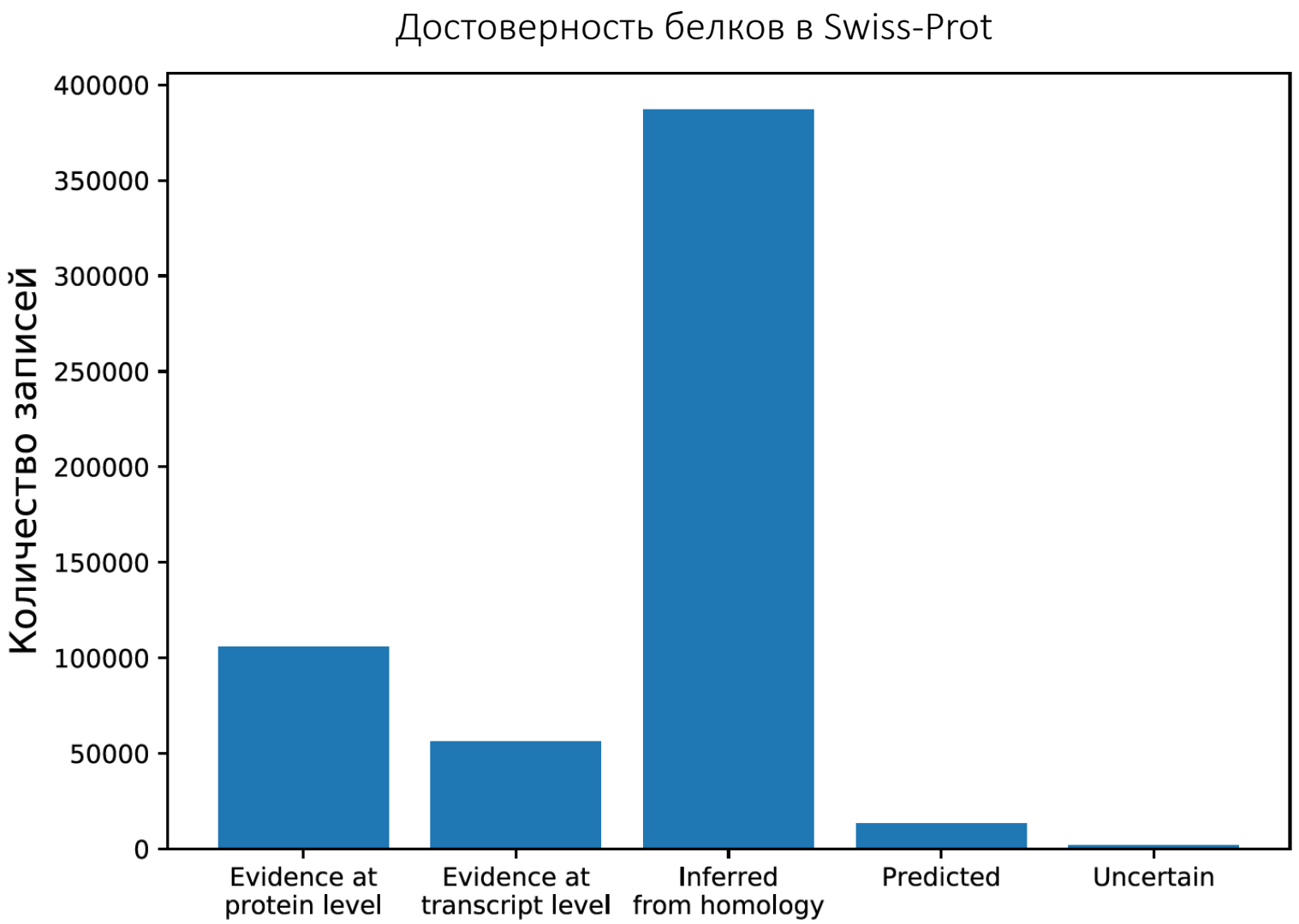
Это логарифмическая шкала!

На самом деле всё вот так!



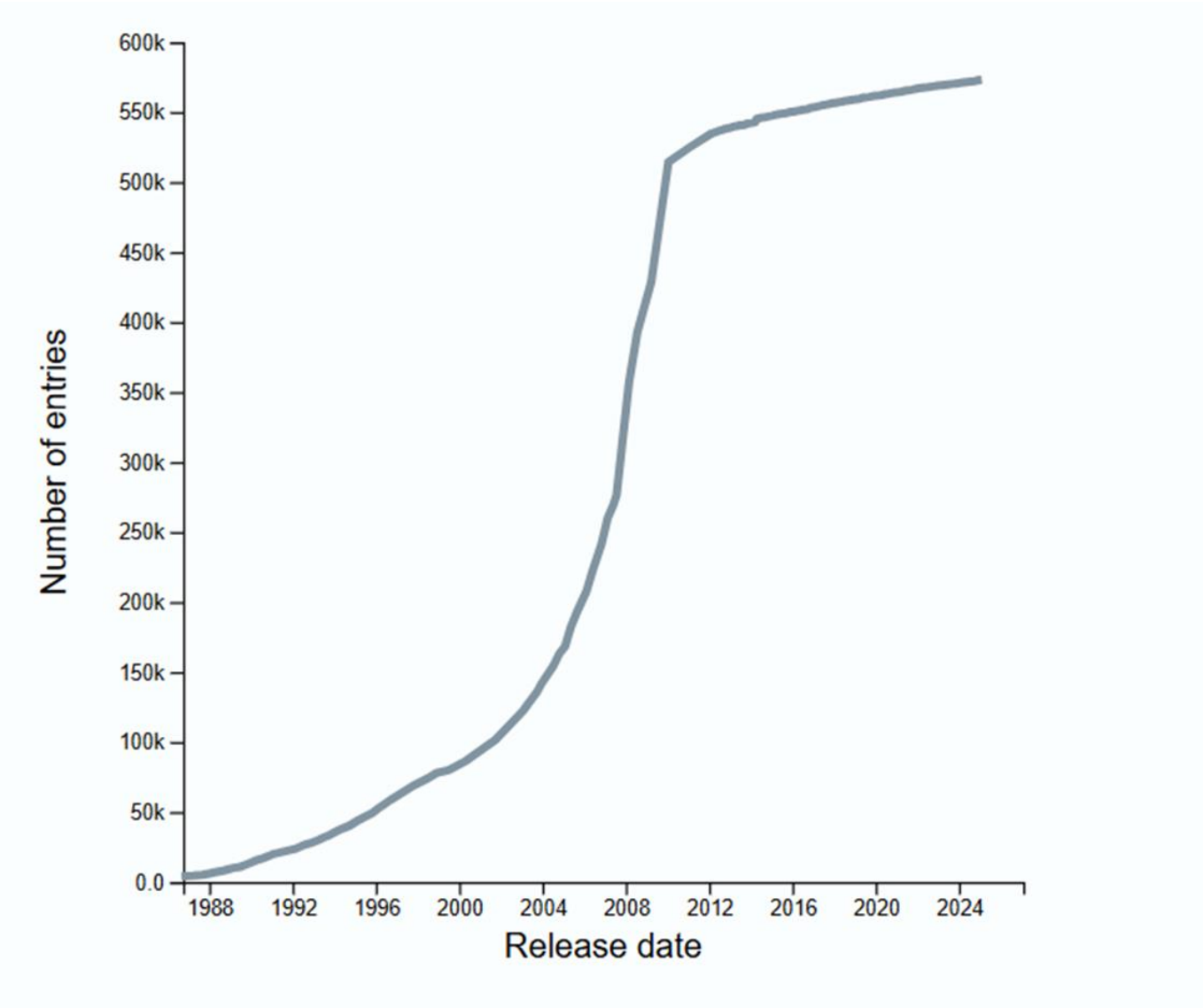
Известных структур во много раз меньше, чем последовательностей.
Большинство последовательностей предсказано и аннотировано лишь автоматически.

На самом деле всё вот так!

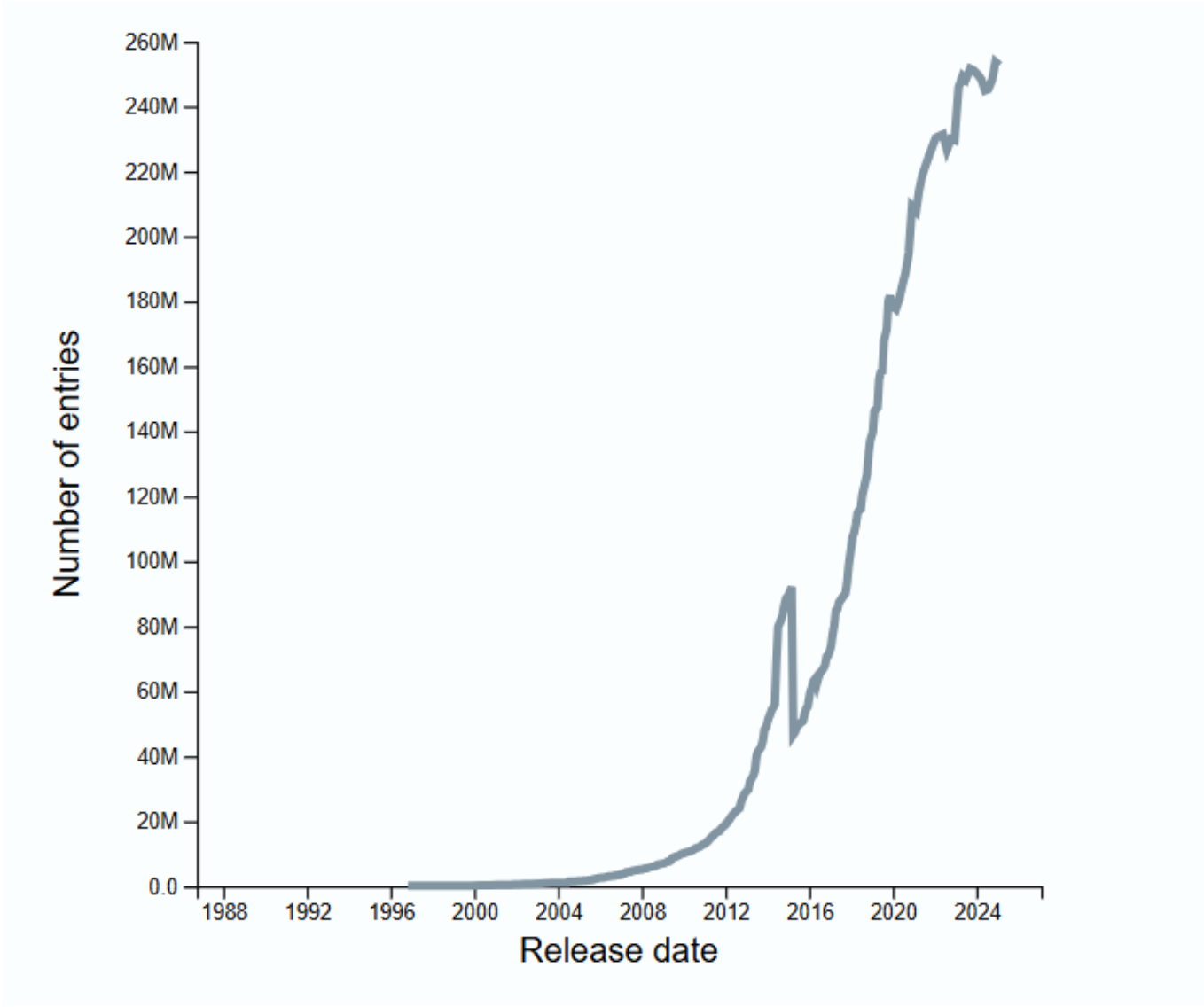


Среди аннотированных вручную белков большая часть не изучена экспериментально даже на уровне транскрипта.

Рост числа записей в Swiss-Prot



Рост числа записей в TrEMBL



Future releases

Release 2026_02

Currently expected in the late first or second quarter of 2026

Changes that we are making in this release:

- Removal of all remaining non-reference proteome protein entries (except reviewed (Swiss-Prot) or biologically highly relevant protein entries) from UniProtKB.

Changes in numbers:

- Removal of approximately 60 million protein entries from UniProtKB, which will result in an estimated total of 139 million entries remaining.

How this might affect you:

- Protein entries that do not belong to a Reference Proteome or that are not considered to be of high biological relevance can still be found in UniParc, where users can find some information such as the protein sequence, gene and protein names, InterPro signatures, etc.

Что мы называем «одним белком»?

Какие есть проблемы?

Два гена из одного генома кодируют один белок (недавняя дупликация).

Два гена из разных видов (штаммов, родов, ...) кодируют один белок.

Полиморфизм: последовательность белка отличается у разных особей одного вида.

Соматические различия: разные белки в разных клетках организма (иммунные клетки, раковые клетки, соматические мутации).

Альтернативный сплайсинг: у одного гена может быть несколько продуктов, разных по последовательности.

Транссплайсинг: сплайсинг между разными генами, белок не закодирован в одном гене.

Одна запись UniProtKB

Одна запись – все **продукты одного гена** из организмов **одного вида**. Известные изоформы, полиморфизмы и т.д. указывают в аннотации записей.

Изоформы указаны в полях CC (подраздел "Alternative products") и FT (конкретные участки различий), полиморфизмы указывают в поле FT.

Правило не строгое, из него есть исключения. Например, если для гена известно множество изоформ, сильно отличающихся по последовательности и функциям, то для них создадут несколько записей.

Формат записи UniProtKB

```
ID   CA1AB_CONAN             Reviewed;         17 AA.
AC   P0C1V7;
DT   19-SEP-2006, integrated into UniProtKB/Swiss-Prot.
DT   19-SEP-2006, sequence version 1.
DT   14-DEC-2022, entry version 40.

DE   RecName: Full=Alpha-conotoxin AnIB {ECO:0000303|PubMed:14971903};
DE   Contains:
DE     RecName: Full=Alpha-conotoxin AnIA {ECO:0000303|PubMed:14971903};
OS   Conus anemone (Anemone cone).
OC   Eukaryota; Metazoa; Spiralia; Lophotrochozoa; Mollusca; Gastropoda;
OC   Caenogastropoda; Neogastropoda; Conoidea; Conidae; Conus; Floraconus.
OX   NCBI_TaxID=101285;
RN   [1]
RP   PROTEIN SEQUENCE, SULFATION AT TYR-16, AMIDATION AT CYS-17, MASS
RP   SPECTROMETRY, SYNTHESIS, SUBCELLULAR LOCATION, AND MUTAGENESIS OF GLY-1.
RC   TISSUE=Venom;
RX   PubMed=14971903; DOI=10.1021/jm031010o;
RA   Loughnan M.L., Nicke A., Jones A., Adams D.J., Alewood P.F., Lewis R.J.;
RT   "Chemical and functional identification and characterization of novel
RT   sulfated alpha-conotoxins from the cone snail Conus anemone.";
RL   J. Med. Chem. 47:1234-1241(2004).
RN   [2] {ECO:0007744|PDB:7N20, ECO:0007744|PDB:7N21, ECO:0007744|PDB:7N22, ECO:0007744|PDB:7N23}
RP   STRUCTURE BY NMR (SULFATED ALPHA-CONOTOXIN ANIB), SYNTHESIS (SULFATED;
RP   NON-SULFATED; AMIDATED AND NON-AMIDATED ALPHA-CONOTOXIN ANIB CONOTOXIN),
RP   AND PTM.
RX   PubMed=34671739; DOI=10.1039/d1md00182e;
RA   Ho T.N.T., Lee H.S., Swaminathan S., Goodwin L., Rai N., Ushay B.,
RA   Lewis R.J., Rosengren K.J., Conibear A.C.;
RT   "Posttranslational modifications of alpha-conotoxins: sulfotyrosine and C-
RT   terminal amidation stabilise structures and increase acetylcholine receptor
RT   binding.";
RL   RSC Med. Chem. 12:1574-1584(2021).
                                        ...
FT   MUTAGEN         1
FT                   /note="Missing: 1.6-fold increase in inhibitory potency on
FT                   alpha-3-beta-2/CHRNA3-CHRNB2."
FT                   /evidence="ECO:0000269|PubMed:14971903"
FT   HELIX           3..5
FT                   /evidence="ECO:0007829|PDB:7N20"
FT   HELIX           7..12
FT                   /evidence="ECO:0007829|PDB:7N20"
FT   TURN            14..16
FT                   /evidence="ECO:0007829|PDB:7N20"


SQ   SEQUENCE   17 AA;  1714 MW;  76F84AFDFAC99005 CRC64;
     GGCCSHPACA ANNQDYC
```

Метаданные

Аннотация последовательности

Последовательность

Основные поля записи

ID — название записи (идентификатор)

AC — код доступа (еще один идентификатор)

DE — description, описание (функция) белка

OS — видовое название организма–источника белка

OC — таксономическое положение организма (по NCBI Taxonomy)

DR — ссылки на записи в других базах данных о данном белке

PE — protein existence, 5 уровней достоверности существования белка

KW — ключевые слова

FT — feature table, таблица локальных особенностей

CC — comments, другая полезная информация, плохо поддающаяся формализации

SQ — последовательность

Идентификаторы записи UniProtKB

ID – имя записи (**entry name**), уникальный идентификатор

- единственный у записи и уникальный
- может изменяться со временем
- человекочитаемый, включает мнемонику функции и мнемонику организма
- примеры: INS_HUMAN, INS1_MOUSE, A0A1S2PNH5_9ACTN

AC – код доступа (**accession number**), стабильный идентификатор

- не изменяется и не удаляется
- у записи может быть несколько AC
- может повторяться у нескольких записей (основной AC всегда уникальный)
- случайная комбинация букв и цифр
- примеры: A2BC19, P12345, A0A023GPI8

Когда ссылаетесь на запись, указывайте **основной** (primary) **код доступа**!

Таблица локальных особенностей (Feature table, FT)

Имеет строгий формат, список и описание всех возможных ключей доступно на сайте UniProt.

```
FT   PEPTIDE         1..17
FT                   /note="Alpha-conotoxin AnIB"
FT                   /evidence="ECO:0000269|PubMed:14971903"
FT                   /id="PRO_0000249780"
FT   PEPTIDE         3..17
FT                   /note="Alpha-conotoxin AnIA"
FT                   /evidence="ECO:0000269|PubMed:14971903"
FT                   /id="PRO_0000249781"
FT   REGION          5..7
FT                   /note="Ser-Xaa-Pro motif, crucial for potent interaction
FT                   with nAChR"
FT                   /evidence="ECO:0000250|UniProtKB:P56636"
FT   MOD_RES         16
FT                   /note="Sulfotyrosine"
FT                   /evidence="ECO:0000269|PubMed:14971903"
FT   MOD_RES         17
FT                   /note="Cysteine amide"
FT                   /evidence="ECO:0000269|PubMed:14971903"
FT   DISULFID        3..9
FT                   /evidence="ECO:0000305|PubMed:34671739,
FT                   ECO:0007744|PDB:7N20"
FT   DISULFID        4..17
FT                   /evidence="ECO:0000305|PubMed:34671739,
FT                   ECO:0007744|PDB:7N20"
FT   MUTAGEN         1
FT                   /note="G->GG: 1.4-fold increase in inhibitory potency on
FT                   alpha-3-beta-2/CHRNA3-CHRNB2."
FT                   /evidence="ECO:0000269|PubMed:14971903"
FT   MUTAGEN         1
FT                   /note="Missing: 1.6-fold increase in inhibitory potency on
FT                   alpha-3-beta-2/CHRNA3-CHRNB2."
FT                   /evidence="ECO:0000269|PubMed:14971903"
FT   HELIX           3..5
FT                   /evidence="ECO:0007829|PDB:7N20"
FT   HELIX           7..12
FT                   /evidence="ECO:0007829|PDB:7N20"
FT   TURN            14..16
FT                   /evidence="ECO:0007829|PDB:7N20"
```

Конец лекционной части