

Базы данных последовательностей белков

UniProt

Откуда берутся последовательности белков?

Прошлое: белковое секвенирование

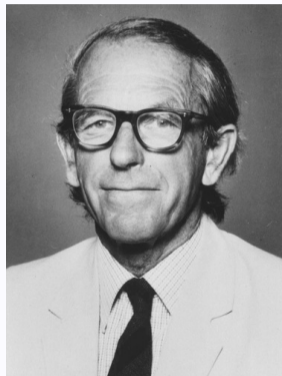
- ▶ 1950 – N-концевая деградация пептидов:
ди- и трипептиды



Pehr Edman

Прошлое: белковое секвенирование

- ▶ 1950 – N-концевая деградация пептидов:
ди- и трипептиды
- ▶ 1951 и 1952 – Первая последовательность белка:
цепи В и А бычьего инсулина, 30 и 21 а.о.



Frederick Sanger

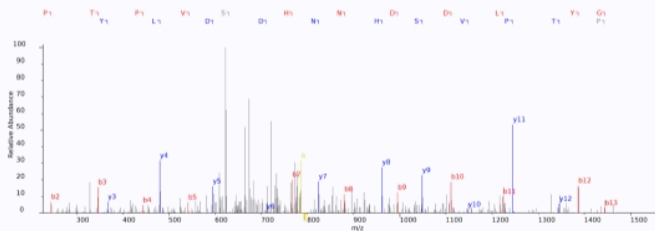
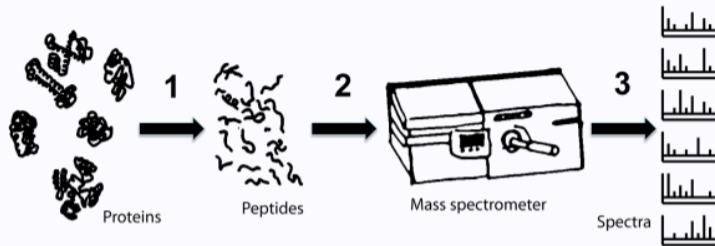
Прошлое: белковое секвенирование

- ▶ 1950 – N-концевая деградация пептидов:
ди- и трипептиды
- ▶ 1951 и 1952 – Первая последовательность белка:
цепи В и А бычьего инсулина, 30 и 21 а.о.
- ▶ 1967 – Автоматизация метода Эдмана:
60 а.о. миоглобина кита



Pehr Edman

Настоящее: белковая масс-спектрометрия



Noble WS and MacCoss MJ. Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput Biol.* 2012;8(1):e1002296

Настоящее: автоматическая трансляция

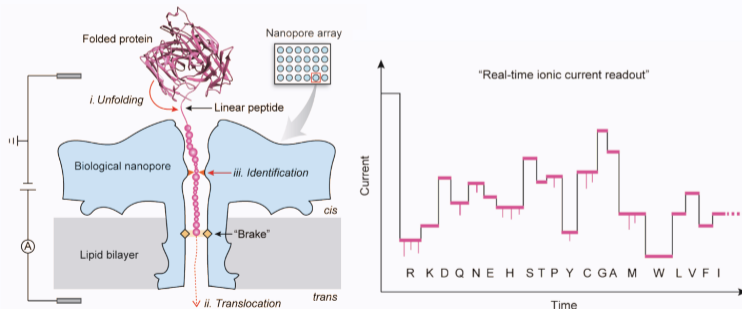
- ▶ Получение нуклеотидной последовательности (геном, экзом, транскриптом, метагеном).
- ▶ Предсказание открытых рамок считывания.
- ▶ Аннотация (в основном автоматическая) последовательностей по сходству с известными белками.

Так получено подавляющее большинство последовательностей белков.

Будущее: белковое секвенирование?

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 белков?

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.

База данных

[Реляционную] **базу данных** можно представить в виде набора ссылающихся друг на друга **плоских таблиц**, при условии, что строки в каждой таблице уникальны, а порядок строк и столбцов не имеет значения.

Единица хранения называется **записью (entry)** и соответствует строке таблицы. Столбцы называются **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: RefSeq

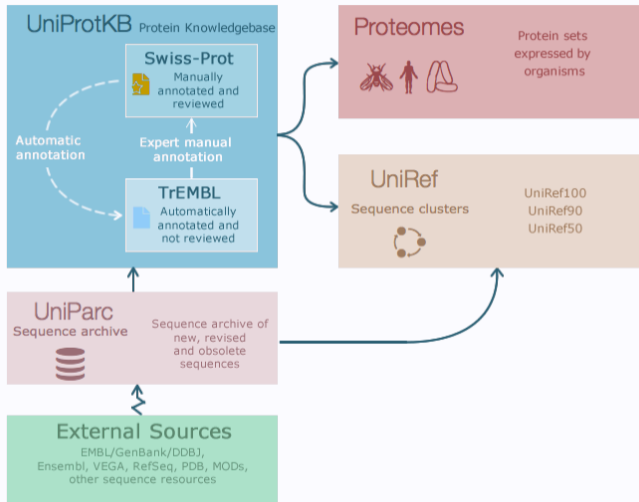


RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

- ▶ 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



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.

Автоматическая база данных!

Например, UPI0000000004

| Database | Identifier | Version | Organism | Proteome | First seen | Last seen | Active |
|----------------------|-------------------------------------|---------|--|-----------------------|------------|------------|--------|
| UniProtKB reviewed | P04195 | 1 | Vaccinia virus (strain Western Reserve) (VACV) | | 1988-11-01 | 2026-01-28 | Yes |
| UniProtKB unreviewed | Q6LDV9 | 1 | Vaccinia virus (VACV) | | 2006-04-18 | 2026-01-28 | Yes |
| UniProtKB unreviewed | V5R1H0 | 1 | Vaccinia virus WAU86/88-1 | | 2015-07-22 | 2026-01-28 | Yes |
| UniProtKB unreviewed | A0A2I2MC48 | 1 | Vaccinia virus (strain Western Reserve) (VACV) | | 2018-02-28 | 2026-01-28 | Yes |
| UniProtKB unreviewed | A4GDG8 | 1 | Vaccinia virus (strain Lister) (VACV) | | 2024-11-27 | 2026-01-28 | Yes |
| UniProtKB unreviewed | Q76ZR9 | 1 | | | 2004-07-05 | 2011-10-19 | No |
| RefSeq | YP_232995 ↗ | 1 | Vaccinia virus (VACV) | | 2005-10-06 | 2025-09-02 | Yes |
| RefSeq | WP_016017665 ↗ | 1 | Vaccinia virus (VACV) | | 2021-05-22 | 2021-07-12 | No |
| EMBL | AAO89392 ↗ | 1 | Vaccinia virus (strain Western Reserve) (VACV) | UP000000344 (Segment) | 2003-06-16 | 2024-07-17 | Yes |
| EMBL | ABD52586 ↗ | 1 | Vaccinia virus (VACV) | UP000180811 (Segment) | 2007-03-31 | 2024-07-17 | Yes |
| PRF | 3315290DX ↗ | | Vaccinia virus (VACV) | | 2007-12-07 | 2009-09-01 | Yes |
| USPTO | ADS58156 ↗ | 1 | | | 2011-03-07 | 2025-09-30 | Yes |
| EMBL | UZL86841 ↗ | 1 | Vaccinia virus (VACV) | | 2023-05-26 | 2025-09-22 | Yes |
| EPO | WO2024US54812_561 ↗ | 1 | | | 2025-05-21 | 2025-11-04 | Yes |
| TREMBLNEW | AAA48264 | | | | 2003-03-29 | 2004-06-11 | No |
| PIRARC | A01146 | | | | 2003-03-31 | 2003-04-04 | No |
| PIRARC | A35014 | | | | 2003-03-31 | 2003-04-04 | No |
| PIR | CRVZW | | | | 2003-04-11 | 2005-01-04 | No |

UniProt Knowledgebase

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

- TrEMBL** (от **T**ranslated **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 с активными ссылками на RefSeq.

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 и репрезентативная последовательность

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

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

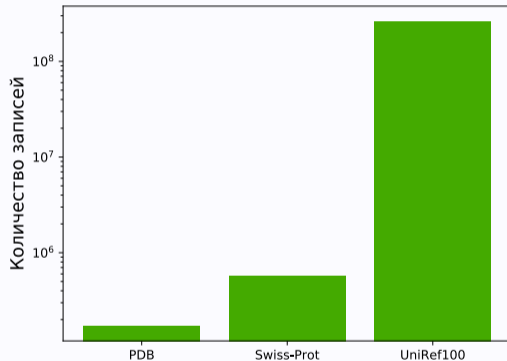
Случаются приколы 😞

Например, кластер UniRef90_P81108

| <input type="checkbox"/> | Cluster members | Entry name | Protein names | Organisms | Organism IDs | Related clusters | Length | Role |
|--------------------------|-----------------|-------------------|---|--|--------------|-------------------------|-----------|-----------------------|
| <input type="checkbox"/> | P81108 | THIO_CLOSG |  Thioredoxin (Fragment) | Clostridium sporogenes | 1509 | UniRef100_P81108 | 40 | Representative |
| <input type="checkbox"/> | A0A1V9IK41 | A0A1V9IK41_CLOSG |  Thioredoxin | Clostridium sporogenes | 1509 | UniRef100_P81108 | 106 | |
| <input type="checkbox"/> | A0A0B4W3E0 | A0A0B4W3E0_CLOBO |  Thioredoxin | Clostridium botulinum Prevot_594 | 1408284 | UniRef100_P81108 | 106 | |
| <input type="checkbox"/> | A0A1J1CWE3 | A0A1J1CWE3_CLOBO |  Thioredoxin | Clostridium botulinum | 1491 | UniRef100_A0A1J1CWE3 | 106 | Seed |
| <input type="checkbox"/> | J7T6P1 | J7T6P1_CLOS1 |  Thioredoxin | Clostridium sporogenes (strain ATCC 15579) | 471871 | UniRef100_A0A1J1CWE3 | 106 | |
| <input type="checkbox"/> | A0A0D0ZX46 | A0A0D0ZX46_CLOBO |  Thioredoxin | Clostridium botulinum B2 450 | 1379739 | UniRef100_A0A1J1CWE3 | 106 | |
| <input type="checkbox"/> | A0A6M0YC23 | A0A6M0YC23_CLOSG |  Thioredoxin | Clostridium sporogenes | 1509 | UniRef100_A0A6M0YC23 | 106 | |
| <input type="checkbox"/> | A0A1S9I145 | A0A1S9I145_9CLOT |  Thioredoxin | Clostridium tepidum | 1962263 | UniRef100_A0A1S9I145 | 106 | |
| <input type="checkbox"/> | A0A6M0T4F3 | A0A6M0T4F3_CLOBO |  Thioredoxin | Clostridium botulinum | 1491 | UniRef100_A0A6M0T4F3 | 106 | |
| <input type="checkbox"/> | A0A6M0XX80 | A0A6M0XX80_CLOSG |  Thioredoxin | Clostridium sporogenes | 1509 | UniRef100_A0A6M0T4F3 | 106 | |
| <input type="checkbox"/> | A0A0M1IUU5 | A0A0M1IUU5_9CLOT |  Thioredoxin | Clostridium sp. L74 | 1560217 | UniRef100_A0A0M1IUU5 | 106 | |
| <input type="checkbox"/> | UPI000666DA61 | |  thiol reductase thioredoxin | Clostridium sporogenes | 1509 | UniRef100_UPI000666DA61 | 106 | |
| <input type="checkbox"/> | UPI000D0CC3E6 | |  thiol reductase thioredoxin | Clostridium botulinum | 1491 | UniRef100_UPI000D0CC3E6 | 106 | |
| <input type="checkbox"/> | UPI001748E097 | |  thioredoxin | Clostridium botulinum | 1491 | UniRef100_UPI001748E097 | 106 | |
| <input type="checkbox"/> | UPI0005F06029 | |  thiol reductase thioredoxin | Clostridium sporogenes | 1509 | UniRef100_UPI0005F06029 | 106 | |

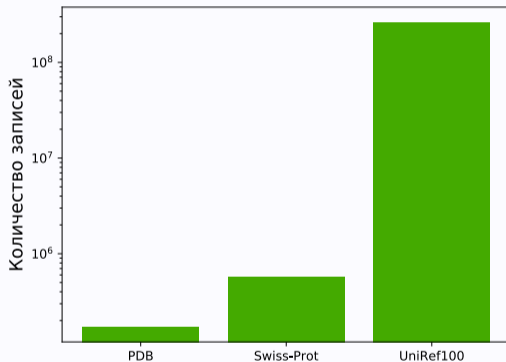
Масштаб проблемы

Число записей про белки в разных базах данных.



Масштаб проблемы

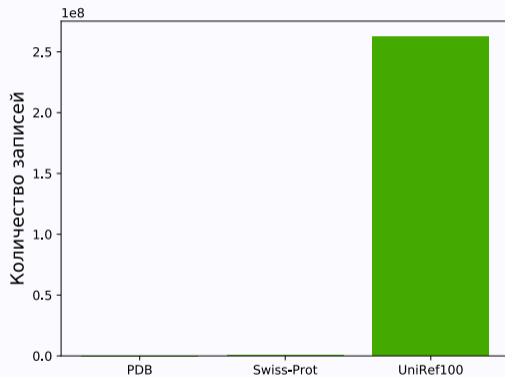
Число записей про белки в разных базах данных.



Логарифмическая шкала!

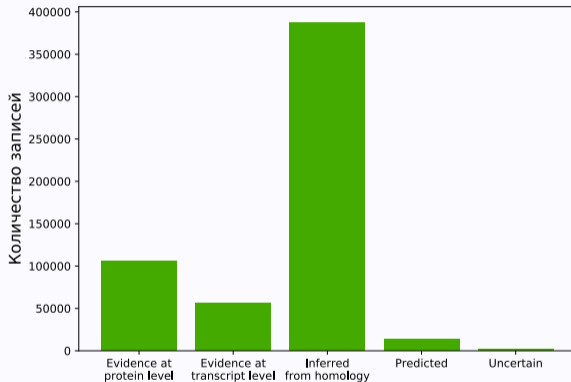
Масштаб проблемы

Число записей про белки в разных базах данных.



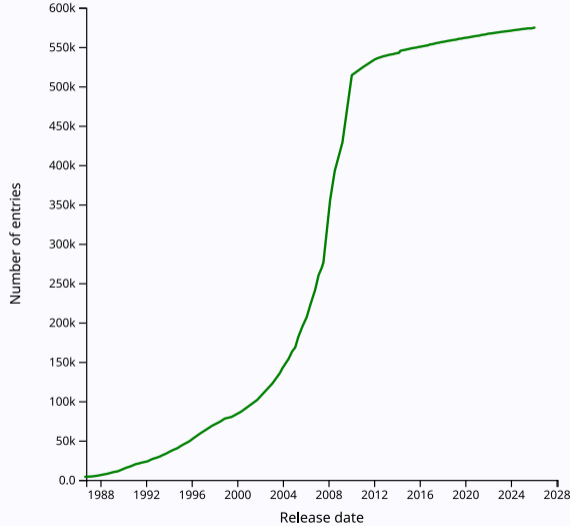
Масштаб проблемы

Достоверность белков в Swiss-Prot.

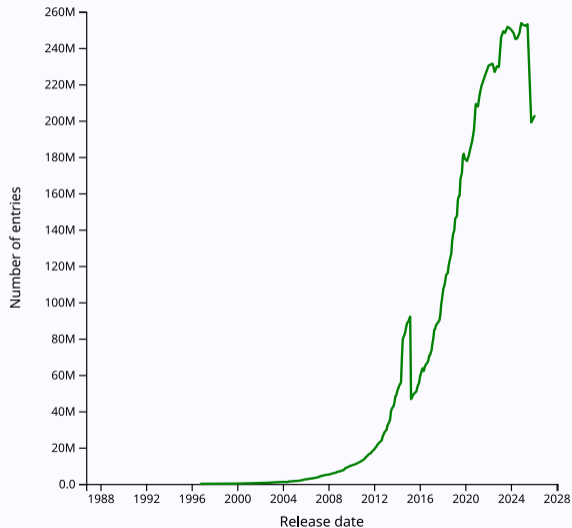


Среди аннотированных вручную белков большая часть не изучена экспериментально.

Рост числа записей в Swiss-Prot



Рост числа записей в TrEMBL



Что такое «один белок»?

Какие есть проблемы?

Консервативный ген ⇒ одинаковые по последовательности белки из разных видов.

Недавняя дупликация ⇒ разные участки генома кодируют одинаковые белки.

Полиморфизм ⇒ последовательность отличается у разных особей одного вида.

Соматические различия ⇒ разные варианты гена в клетках одного организма.

Альтернативный сплайсинг ⇒ ген один, а продукты разные.

Транссплайсинг ⇒ в сплайсинг вовлечены транскрипты разных генов.

Запись UniProtKB

Одна запись – все продукты одного гена из организмов одного вида или штамма.

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

Если для гена известно множество изоформ, сильно отличающихся по последовательности и функциям, могут создать несколько записей Swiss-Prot.

Формат записи UniProtKB

Структура записи UniProtKB

```
ID  NU4LM_BALMU          Reviewed;          98 AA.
AC  P41301;
DT  01-FEB-1995, integrated into UniProtKB/Swiss-Prot.
DT  26-FEB-2020, entry version 75.
DE  RecName: Full=NADH-ubiquinone oxidoreductase chain 4L;
DE      EC=7.1.1.2;
GN  Name=MT-ND4L; Synonyms=MTND4L, NADH4L, ND4L;
OS  Balaenoptera musculus (Blue whale).
OG  Mitochondrion.
OC  Eukaryota; Metazoa; Chordata; ...; Balaenopteridae; Balaenoptera.
OX  NCBI_TaxID=9771;
RN  [1]
RP  NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX  PubMed=8308901; DOI=10.1007/bf00178861;
RA  Arnason U., Gullberg A.;
RT  "Comparison between the complete ... that can hybridize in nature.";
RL  J. Mol. Evol. 37:312-322(1993).
CC  -!- SUBCELLULAR LOCATION: Mitochondrion membrane {ECO:0000250}; Multi-pass
CC      membrane protein {ECO:0000250}.
CC  -!- SIMILARITY: Belongs to the complex I subunit 4L family. {ECO:0000305}.
DR  EMBL; X72204; CAA51003.1; -; Genomic_DNA.
DR  RefSeq; NP_007064.1; NC_001601.1.
DR  GO; GO:0016021; C:integral component of membrane; IEA:UniProtKB-KW.
DR  Pfam; PF00420; Oxidored_q2; 1.
PE  3: Inferred from homology;
KW  Translocase; Transmembrane; Transmembrane helix; Transport; Ubiquinone.
FT  CHAIN           1..98
FT                  /note="NADH-ubiquinone oxidoreductase chain 4L"
FT                  /id="PRO_0000118394"
FT  TRANSMEM       1..21
FT                  /evidence="ECO:0000255"
SQ  SEQUENCE      98 AA; 10747 MW;  9F770651FE65ED1B CRC64;
    MTLIHMNVLMAFSMSLVGLL MYRSHLSAL LCLEGMMLSL FVLAALTILN SHFTLANMMP
    IILLVFAAYV AAIGLALLVM VSNTYGTDYV QSLNLLQC
```

метаданные

аннотация
последовательности

последовательность

Основные поля записи UniProtKB

- ID** – название записи (уникальный идентификатор)
- AC** – код доступа (стабильный идентификатор)
- DE** – description, название (функция) белка
- OS** – видовое название организма–источника белка
- OC** – таксономическое положение организма (по NCBI Таксоному)
- DR** – ссылки на записи в других базах данных
- 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.

АС – код доступа (accession number), *стабильный* идентификатор:

- ▶ не изменяется и не удаляется;
- ▶ у записи может быть несколько АС;
- ▶ может повторяться у нескольких записей;
- ▶ случайная комбинация букв и цифр.

Примеры: A2BC19, P12345, A0A023GPI8.

Когда ссылаетесь на запись, указывайте основной (primary, первый в списке) код доступа!

Он стабильный и уникален в рамках текущего релиза UniProtKB.

Таблица локальных особенностей (Feature table, FT)

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

```
FT CHAIN 1..188
FT /note="Isochorismatase family protein YecD"
FT /id="PRO_0000201831"
FT HELIX 6..8
FT /evidence="ECO:0000244|PDB:1J2R"
FT STRAND 9..14
FT /evidence="ECO:0000244|PDB:1J2R"
FT REGION 5..34
FT /note="Interaction with RNase E"
FT ACT_SITE 209
FT /note="Proton donor"
FT /evidence="ECO:0000255|HAMAP-Rule:MF_00318,
FT ECO:0000269|PubMed:15003462"
FT METAL 246
FT /note="Magnesium"
FT /evidence="ECO:0000269|PubMed:11676541,
FT ECO:0000269|PubMed:16516921"
FT BINDING 159
FT /note="Substrate"
FT /evidence="ECO:0000255|HAMAP-Rule:MF_00318"
FT MOD_RES 257
FT /note="N6-acetyllysine"
FT /evidence="ECO:0000269|PubMed:18723842"
FT CONFLICT 180..188
FT /note="SVEEILNAL -> TWKRSSTRYDLHRSTAMVAS (in Ref. 1)"
FT /evidence="ECO:0000305"
FT MUTAGEN 168
FT /note="E->Q: 5% activity; not secreted."
FT /evidence="ECO:0000269|PubMed:15003462"
```

вторичная структура

сайты связывания

модифицированные остатки

разночтения в
последовательности

и т.д.