

Известны последовательности 30 млн белков Как организовать библиотеку белков?!



Источники последовательностей белков

- Получение последовательностей нуклеиновых кислот, ДНК и РНК
 - Геномные проекты
 - Секвенирование интересующих генов и/или участков ДНК в лабораториях
 - Метагеномные проекты
- Идентификация генов
 - Компьютерные предсказания
 - мРНК и их фрагменты – EST и др.
- Экспериментальное подтверждение существования белка
 - мРНК
 - Клонирование и изучение
 - Выделение белка и масс-спектрометрия



2013-09-25

Источники последовательностей белков

Источники последовательностей белков

- Получение последовательностей нуклеиновых кислот, ДНК и РНК
 - Геномные проекты
 - Секвенирование интересующих генов и/или участков ДНК в лабораториях
 - Метагеномные проекты
- Идентификация генов
 - Компьютерные предсказания
 - мРНК и их фрагменты – EST и др.
- Экспериментальное подтверждение существования белка
 - мРНК
 - Клонирование и изучение
 - Выделение белка и масс-спектрометрия

Банки нуклеотидных последовательностей

2013-09-25

Банки нуклеотидных последовательностей

GenBank, EMBL, DDBJ → Refseq

Архивные базы последовательностей нуклеиновых кислот

Автоматическая база различных последовательностей НК

GenBank, EMBL, DDBJ

→ Refseq

Архивные базы последовательностей нуклеиновых кислот

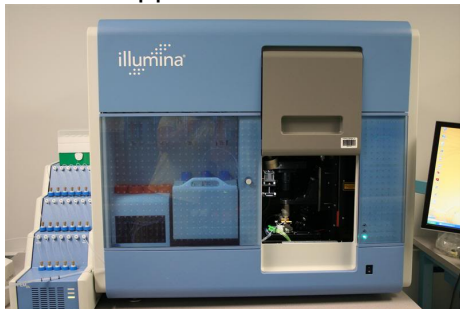
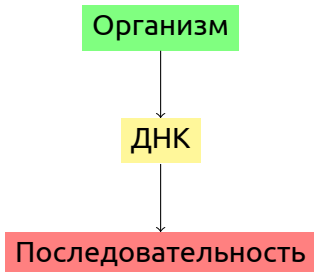
Автоматическая база различных последовательностей НК



Экскурс в историю

Первоначально определяли аминокислотную последовательность белка, отщепляя по одному остатку и определяя какой он.

В конце 1970-х годов был изобретён относительно быстрый и дешёвый метод экспериментального определения последовательности оснований ДНК.

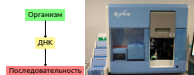


2013-09-25

Экскурс в историю

Экскурс в историю

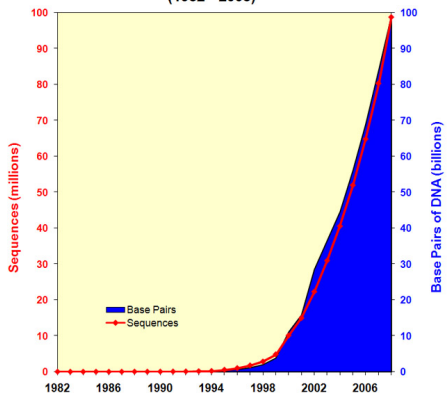
Первоначально определяли аминокислотную последовательность белка, отщепляя по одному остатку и определяя какой он.
В конце 1970-х годов был изобретён относительно быстрый и дешёвый метод экспериментального определения последовательности оснований ДНК.



Экскурс в историю

В 1982 году были открыты публичные электронные банки данных GenBank и EMBL, содержащие последовательности ДНК. В 1984 аналогичный банк данных DDBJ открыт в Японии.

Growth of GenBank
(1982 - 2008)



EMBL

Genbank

DDBJ

2013-09-25

Экскурс в историю

Экскурс в историю

В 1982 году были открыты публичные электронные банки данных GenBank и EMBL, содержащие последовательности ДНК. В 1984 аналогичный банк данных DDBJ открыт в Японии.



EMBL


Genbank

DDBJ

Что включать в описание одной последовательности?



Что хочется (может) знать о белке?

- Название
- Хим. формулу
 - Последовательность
 - S-S связи
 - Модификации (химические), которым подвергается белок in vivo
- Структуру
 - вторичную
 - пространственную
- Из какого организма получен
 - Полиморфизмы
- Функцию – что умеет делать
 - аминокислотные остатки, ответственные за функцию (функции)
 - другие особенности остатков
- Другую информацию
- Источники информации (правда ли все сказанное о белке?) 

2013-09-25

Что хочется (может) знать о белке?

- Название
- Хим. формулу
 - Последовательность
 - S-S связи
 - Модификации (химические), которым подвергается белок in vivo
- Структуру
 - вторичную
 - пространственную
- Из какого организма получен
 - Полиморфизмы
- Функцию – что умеет делать (функции)
 - аминокислотные остатки, ответственные за функцию
 - другие особенности остатков
- Другую информацию
- Источники информации (правда ли все сказанное о белке?)

Uniprot, база последовательностей белков

Структура записи, в которой хранится одна последовательность



Документ банка данных Swiss-Prot

http://au.expasy.org/uniprot/P47995.txt - Microsoft Inte

File Edit View Favorites Tools Help

Address http://au.expasy.org/uniprot/P47995.txt

```

ID YSEA STACA STANDARD; PRT; 165
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=TH300;
RA Freudl R.;
RL Submitted (JUN-1994) to the EMBL/GenBank
CC -!- SIMILARITY: Belongs to the ribosomal
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; BF8CB91ADE194DDO CRC64;
LERYFTNVFN VNAHVVKVKT Y ANSSKIEVTI PLNDVTLRAE ERNDDIYAGI DKITNKLECC
VRKYKTRVNR KRRKESEHEP FPATPETPPE TAVDHDKDEE IEIIRSKQFS LKPMDEEAV
LQMDLLGTFD FIFNDRETDG TSIVYRRKDG KYGLIETVEK LICDI
//

```

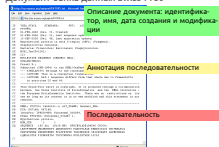
Описание документа: идентификатор, имя, дата создания и модификации

Аннотация последовательности

Последовательность

2013-09-25

Документ банка данных Swiss-Prot



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", особенности частей последовательности



2013-09-25

Основные поля Swiss-Prot

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", особенности частей последовательности

Структура идентификатора записи Swiss-Prot

ENO_BACSU: энолаза из сенной палочки

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:

- 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*));
- вирусы (например, BPP21 для фага P21, MEASY для штамма Yamagata вируса кори (measles) и пр.);
- случаи неопределенного видового названия.



Структура идентификатора записи Swiss-Prot

ENO_BACSU: энолаза из сенной палочки

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:

- 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*));
- вирусы (например, BPP21 для фага P21, MEASY для штамма Yamagata вируса кори (measles) и пр.);
- 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 "VARSPPLIC");
- и т. п.

Имеет строгий формат: Feature Key, FtLocation, FtDescription.

FT DISULFID 334 343 By similarity.

FT CONFLICT 138 138 E -> EE (in Ref. 4; AA sequence).



2013-09-25

Содержимое поля FT

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 "VARSPPLIC");
- и т. п.

Имеет строгий формат: Feature Key, FtLocation, FtDescription.
FT DISULFID 334 343 By similarity.
FT CONFLICT 138 138 E -> EE (in Ref. 4; AA sequence).

Как организовать базу данных?



Белки: БД публикаций или БД последовательностей?

Как создать БД белков? Попробуем:

- БД – большой файл в текстовом формате
- Файл должен читаться и программой, и человеком (“адекватная формализация”)
- Единица хранения – запись, entry один белок – одна запись (что такое – один белок?)
- Записи разбиты на поля
 - Обязательное поле – идентификатор (-ы)
 - Название белка, синонимы
 - Организм
 - Функция
 - Последовательность
 - Пространственная структура?
 - “Особенности” (features) – информация об отдельных остатках
 - Дополнительные сведения
 - Ссылки на источники информации



2013-09-25

Белки: БД публикаций или БД последовательностей?

Белки: БД публикаций или БД последовательностей?
 Как создать БД белков? Попробуем:

- БД – большой файл в текстовом формате
- Файл должен читаться и программой, и человеком (“адекватная формализация”)
- Единица хранения – запись, entry один белок – одна запись (что такое – один белок?)
- Записи разбиты на поля
 - Обязательное поле – идентификатор (-ы)
 - Название белка, синонимы
 - Организм
 - Функция
 - Последовательность
 - Пространственная структура?
 - “Особенности” (features) – информация об отдельных остатках
 - Дополнительные сведения
 - Ссылки на источники информации

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



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

Пример: БД “библиотека”

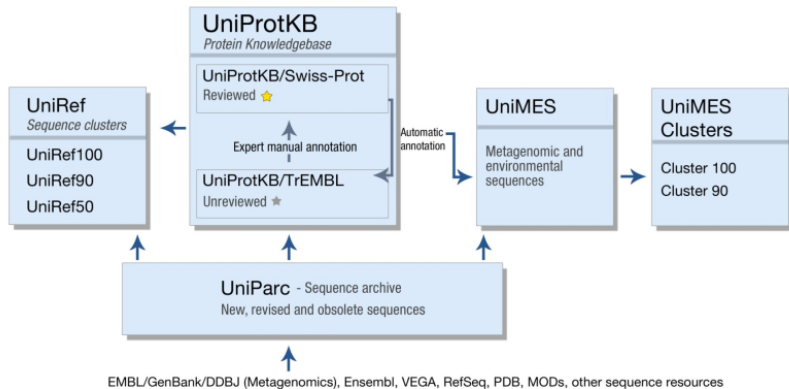
- Запись – книга
- Поля:
 - Название
 - Авторы
 - Год издания
 - Аннотация
 - Текст
- Есть второе хранилище – “картотека”



Пример: БД “библиотека”

- Запись – книга
- Поля:
 - Название
 - Авторы
 - Год издания
 - Аннотация
 - Текст
- Есть второе хранилище – “картотека”

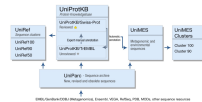
UniProt: из каких частей состоит



2013-09-25

UniProt: из каких частей состоит

UniProt: из каких частей состоит



Банк данных Swiss-Prot



Swiss-Prot – база знаний о белковых последовательностях

- Курируемая база данных
- "Золотой стандарт аннотации проверяет эксперт"
- 534 242 белков



Банк данных TrEMBL



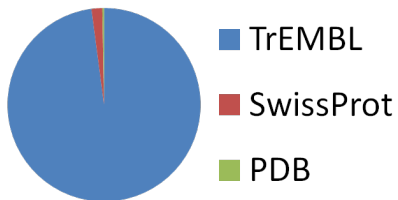
TrEMBL (Translated EMBL)

- Формальная трансляция всех кодирующих нуклеотидных последовательностей из банка EMBL
- Автоматическая классификация и аннотация
- Формат записи тот же, что у Swiss-Prot
- 19 434 245 (19 млн. белков)



Соотношение числа белков

Число записей в банках данных на январь 2013



PDB – пространственные структуры (всего их 88 тыс.)

Swiss-Prot – аннотированные последовательности

TrEMBL – последовательности, полученные компьютерной трансляцией

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



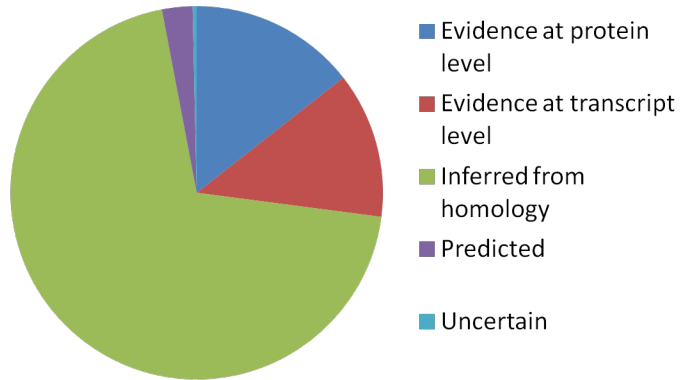
2013-09-25

Соотношение числа белков

Соотношение числа белков

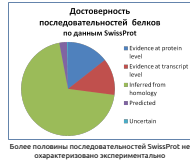


Достоверность последовательностей белков по данным SwissProt



Более половины последовательностей SwissProt не охарактеризовано экспериментально

2013-09-25



"Юридическая" классификация банков данных

2013-09-25

"Юридическая" классификация банков данных

- Архивные, примеры: PDB, GENBANK, EMBL, DDBJ. За содержание каждой записи отвечает её автор-экспериментатор
- Курируемые, за содержание записей отвечают специальные люди — кураторы
- Автоматические, записи генерируются компьютерными программами

- Архивные, примеры: PDB, GENBANK, EMBL, DDBJ. За содержание каждой записи отвечает её автор-экспериментатор
- Курируемые, за содержание записей отвечают специальные люди — кураторы
- Автоматические, записи генерируются компьютерными программами



SRS, Sequence Retrieval System

Основная страница поиска

другие сервера SRS

быстрый поиск

2013-09-25

SRS, Sequence Retrieval System



Начало поиска : Library Page

EMBL-EBI Search

Enter Text Here

Quick Search | Library Page | Other Form

Search Options

- Select the databanks you want to search
- Enter your search terms in the Quick Search box, or choose a query form from below
 - [Standard Query Form](#)
 - [Extended Query Form](#)

You can browse through all the entries in any databanks. First, select the databank you want to browse, then [Browse Entries](#).

Tips

- bookmark this link to return to your project
- Linking to SRS?
- Please read our [Linking to SRS](#) guide for important information regarding linking to our SRS server.

BookMarklets

- Protein Seq
- OMIM/OMIM Seq
- Structures

Available Databases

expand all collapse all Show databanks tooltips:

- Literature, Bibliography and Reference Databases
 - MEDLINE
 - Taxonomy
 - OMIM
 - OMIM Morbid Map
 - Patent Abstracts
 - Karyo's Genomes
- Literature, Bibliography and Reference Databases - subsections
 - MEDLINE (Updates)
 - MEDLINE (Main Release 2009)
 - MEDLINE (Main Release 2008)
 - MED2PUB
- Gene Dictionaries and Ontologies
- Nucleotide sequence databases
 - EMBL
 - Patent DNA
 - EMBL (Contig)
 - EMBL (Contig expanded)
 - EMBL (Annotated Cons)
 - EMBL (Coding Sequences)
 - EMBL ID/Accession Mapping
 - EMBL MGA
 - IMST/IGM-DB
 - IMST/HLS
 - IPD-KIB
 - Genome Reviews
 - GR Genes
 - GR Transcripts
 - RefSeq Genomes
 - UniLists
- Nucleotide sequence databases - subsections
 - EMBL (Updates)
 - EMBL (Release)
 - EMBL (Whole Genome Shotgun)
 - EMBL (Whole Genome Shotgun release)
 - EMBL (Whole Genome Shotgun updates)
 - EMBL (Contig release)
 - EMBL (Contig updates)
 - EMBL (Contig expanded release)
 - EMBL (Contig expanded updates)
 - EMBL (Annotated Cons release)
 - EMBL (Annotated Cons updates)
 - EMBL (Release, Deleted)
 - EMBL (Whole Genome Shotgun Masters)
 - ENA Project
 - RefSeq Genome (Release)
 - RefSeq Genome (Updates)
- Nucleotide related databases
- UniProt Universal Protein Resource
 - UniProtKB
 - UniProtKB/Swiss-Prot
 - UniProtKB/TrEMBL
 - UniRef100
 - UniRef90
 - UniRef50
 - UniParc
- Other protein sequence databases
- Protein function, structure and interaction databases
- Enzymes, reactions and metabolic pathway databases
- Mutation and SNP databases
- Other databases
- User owned databases
- Application result databases
- EMBOSS result databases
- EMBL/CD5 Grouped By

Contains the translations of all coding sequences present in the EMBL sequence database not yet integrated in SWISS-PROT

To obtain comprehensive information on this databank, click the link

Terms of Use Feedback & Support EMBL Release 7.1.13.2 Copyright © 1997-2009 EMBL BioScience AG. All Rights Reserved.

Форма для поиска

Список банков данных, в которых можно искать

Всплывающая подсказка про каждую базу данных

PDB в этом свёрнутом разделе

2013-09-25

Начало поиска : Library Page



Страница поиска : Query Form

2013-09-25

Страница поиска : Query Form



EMBL-EBI

Databases Tools EBI Groups Training Industry About Us Help Site Index

Quick Search Library Page **Query Form** Tools Results Projects Views Databanks [HELP](#)

search UniprotKB/Swiss-Prot

Search Options

Combine search terms with:

Use wildcards

Get results of type:

Result Display Options

View results using:

or

Create a view

Show results per page

Tips

To do more advanced queries, use the [Extended Query Form](#).

Fields you can search

In a single field, you can separate multiple values by: &, | or

1 AllText

1 AllText

1 AllText

1 AllText

Your search terms

Create a view

Select the fields you want displayed in your view and choose the format

Choose 1 or more fields:

- ID
- EntryName
- Data Class
- AccessionNumber
- Primary Accession Number
- Sequence Version
- Creation Date

Display As: Table List

Sequence Format:

Terms of Use Feedback & Support EMBL Release 7.1.13.2 Copyright © 1997-2008 EMBL Bioinformatics AG. All Rights Reserved.

соединяем логическими операторами
ставим ли * в конце запроса
где ищем
что ищем
Список значений полей
Какие столбцы будут в конечной таблице



Логические операторы

Оператор	значение			пример	результат
&	AND	и		phosphorylase & methyltransferase	белки, обладающие обеими функциями одновременно
	OR	или		rat mouse	мышь и крыса
!	BUTNOT	но не		nitrogenase ! reductase	нитрогеназы, не являющиеся редуктазами

(mammalia|aves)!chicken - звери и птицы, но не курица

*-замена любого числа символов: tr*psin = trypsin

В конце каждого запроса по умолчанию стоит * (Use wildcards включено)

2013-09-25

Логические операторы

Оператор	значение	пример	результат
&	AND	и	phosphorylase & methyltransferase Белки, обладающие обеими функциями одновременно
	OR	или	rat mouse мышь и крыса
!	BUTNOT	но не	nitrogenase ! reductase нитрогеназы, не являющиеся редуктазами

(mammalia|aves)!chicken - звери и птицы, но не курица
*-замена любого числа символов: tr*psin = trypsin
В конце каждого запроса по умолчанию стоит * (Use wildcards включено)



Страница поиска : Query Form

EMBL-EBI Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

Quick Search Library Page **Query Form** Tools Results Projects Views Databanks HELP

Reset search UniProtKB/Swiss-Prot

Search Options

Combine search terms with: & (AND)

Use wildcards

Get results of type: Entry

Fields you can search

In a single field, you can separate multiple values by: &, | or ! Search

Description ✓ trypsin

Organism Name ✓ mus

AllText

AllText

Result Display Options

View results using: UniprotView

or

Create a view

Show 30 results per page

Create a view

Select the fields you want displayed in your view and choose the format

Choose 1 or more fields: ID, EntryName, Data Class, AccessionNumber, Primary Accession Number, Sequence Version, Creation Date

Display As: Table List

Sequence Format: swiss

Tips

To do more advanced queries, use the [Extended Query Form](#).

Terms of Use Feedback & Support SRS Release 7.1.3.2 Copyright © 1997-2003 LION bioscience AG. All Rights Reserved.

Ищем: трипсин у мыши



2013-09-25

Страница поиска : Query Form

Страница поиска : Query Form



Ищем: трипсин у мыши

Страница результатов : Query Results

EMBL-EBI All Databases

Databases Tools EBI Groups Training Industry About Us Help Site Index

Quick Search Library Page Query Form Tools Results Projects Views Databanks **HELP**

Query found 3 entries **запрос**

UniProtKB/Swiss-Prot	ID	EntryName	AccNumber	Description
<input type="checkbox"/> UniProtKB/Swiss-Prot:SPIT4_MOUSE	SPIT4_MOUSE	SPIT4_MOUSE	Q9D263	kunitz-type protease inhibitor 4;
<input type="checkbox"/> UniProtKB/Swiss-Prot:TRY2_MOUSE	TRY2_MOUSE	TRY2_MOUSE	P07146	Anionic trypsin-2;
<input type="checkbox"/> UniProtKB/Swiss-Prot:TYSD1_MOUSE	TYSD1_MOUSE	TYSD1_MOUSE	Q9DRA6 Q9VE90	Peroxisomal leader peptide-processing protease;

Apply Options to:

selected results only
 unselected results only

Result Options

Launch analysis tool:
NCBI BLASTP

Show tools relevant to these results:

Link to related information:

Save results:

Display Options

View results using:
SWISSPROT2

Sort results by:
unsorted

ascending
 descending

Show results per page

Printer friendly view

Записи в html (points to UniProtKB/Swiss-Prot column)

Записи в txt (points to AccNumber column)

Обработка результатов (points to Launch, Tools, Link, Save buttons)

Параметры отображения результатов (points to SWISSPROT2, unsorted, ascending, descending, 30, printer friendly view)

Terms of Use | Feedback & Support | SRS Release 7.1.3.2 | Copyright © 1997-2009 EMBL Bioinformatics AG. All Rights Reserved.

2013-09-25

Страница результатов : Query Results





Описание полей (fields)

2013-09-25

Описание полей (fields)

EMBL-EBI search interface showing search results for a query: `(([swissprot-Description]tr*psin) & [swissprot-Organism]tr*mouse)`. The query is labeled "запрос" (query).

The search results table is annotated with red brackets and labels:

- UniProtKB/Swiss-Prot**: A red bracket groups the first column of the table, labeled "Записи в html" (Records in HTML).
- ID**: A red bracket groups the second column, labeled "Записи в txt" (Records in TXT).
- Apply Options to:** A red bracket groups the "Apply Options to:" section, labeled "Обработка результатов" (Result processing).
- Display Options**: A red bracket groups the "Display Options" section, labeled "Параметры отображения результатов" (Result display parameters).

UniProtKB/Swiss-Prot	ID	EntryName	AccNumber	Description
UniProtKB/Swiss-Prot:SPIT4_MOUSE	SPIT4_MOUSE	SPIT4_MOUSE	Q9D263	kunitz-type protease inhibitor 4;
UniProtKB/Swiss-Prot:TRY2_MOUSE	TRY2_MOUSE	TRY2_MOUSE	P07146	Anionic trypsin-2;
UniProtKB/Swiss-Prot:TYSD1_MOUSE	TYSD1_MOUSE	TYSD1_MOUSE	Q9DRA6 Q9VE90	Peroxisomal leader peptide-processing protease;

Additional interface elements shown include "Apply Options to:" (selected/unselected results only), "Result Options" (Launch analysis tool: NCBI BLASTP, Show tools relevant to these results, Link to related information, Save results), and "Display Options" (View results using: SWISSPROT2, Sort results by: unsorted, Show 30 results per page, Printer friendly view).



Проект GO (Gene Ontology)

- Цель: Создание унифицированной терминологии для аннотации генов
- БД GO Включает три независимых словаря
 - Молекулярные функции (molecular Function) (Как? С чем?)
Например, carbohydrate binding или ATPase activity
 - Биологические процессы (biological Process) (Зачем?)
Например, митоз или биосинтез пуринов
 - Клеточные компоненты (cellular Component) (Где?) Например, ядро или холофермент РНК-полимераза II
- В консорциум GO входит EBI (БД Uniprot, Interpro) и много других организаций GOA – проект описания записей Uniprot терминами GO 94% записей Swissprot и 65% TrEMBL имеют хотя бы один термин GO



2013-09-25

Проект GO (Gene Ontology)

Проект GO (Gene Ontology)

- Цель: Создание унифицированной терминологии для аннотации генов
- БД GO Включает три независимых словаря
 - Молекулярные функции (molecular Function) (Как? С чем?)
Например, carbohydrate binding или ATPase activity
 - Биологические процессы (biological Process) (Зачем?)
Например, митоз или биосинтез пуринов
 - Клеточные компоненты (cellular Component) (Где?) Например, ядро или холофермент РНК-полимераза II
- В консорциум GO входит EBI (БД Uniprot, Interpro) и много других организаций GOA – проект описания записей Uniprot терминами GO 94% записей Swissprot и 65% TrEMBL имеют хотя бы один термин GO

Запись GO называется "термин GO"

- Термины имеют определение и перечень синонимов.
- Термины в пределах одной онтологии (словаря) связаны отношениями "is_a" "is_part_of"
- Термины имеют стандартные идентификаторы: GO:0000093 (пример)



└─ Запись GO называется "термин GO"

- Термины имеют определение и перечень синонимов.
- Термины в пределах одной онтологии (словаря) связаны отношениями "is_a" "is_part_of"
- Термины имеют стандартные идентификаторы: GO:0000093 (пример)

tricarboxylic acid cycle

Accession: GO:0006099

Ontology: biological_process

Synonyms:

exact: citric acid cycle

exact: Krebs cycle

exact: TCA cycle

Definition: A nearly universal metabolic pathway in which the acetyl group of acetyl coenzyme A is effectively oxidized to two CO₂ and four pairs of electrons are transferred to coenzymes. The acetyl group combines with oxaloacetate to form citrate, which undergoes successive transformations to isocitrate, 2-oxoglutarate, succinyl-CoA, succinate, fumarate, malate, and oxaloacetate again, thus completing the cycle. In eukaryotes the tricarboxylic acid is confined to the mitochondria. See also glyoxylate cycle.



2013-09-25

tricarboxylic acid cycle

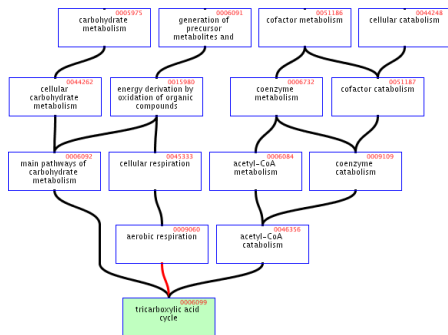
tricarboxylic acid cycle

Accession: GO:0006099
 Ontology: biological_process
 Synonyms:
 exact: citric acid cycle
 exact: Krebs cycle
 exact: TCA cycle

Definition: A nearly universal metabolic pathway in which the acetyl group of acetyl coenzyme A is effectively oxidized to two CO₂ and four pairs of electrons are transferred to coenzymes. The acetyl group combines with oxaloacetate to form citrate, which undergoes successive transformations to isocitrate, 2-oxoglutarate, succinyl-CoA, succinate, fumarate, malate, and oxaloacetate again, thus completing the cycle. In eukaryotes the tricarboxylic acid is confined to the mitochondria. See also glyoxylate cycle.

Ориентированный ациклический граф

- отношение "_is_a": "A is B" означает, что A — частный случай B;
- отношение "is_part_of": "A is part of B" означает, что A — часть B, но B не обязательно содержит A.



2013-09-25

Ориентированный ациклический граф

- отношение "_is_a": "A is B" означает, что A — частный случай B;
- отношение "is_part_of": "A is part of B" означает, что A — часть B, но B не обязательно содержит A.



Аннотация GO записей Uniprot

2013-09-25

Аннотация GO записей Uniprot

- Два принципа:
 - каждая аннотация должна ссылаться на источник:
 - литературная ссылка
 - другая база данных
 - компьютерное предсказание
 - аннотация должна указывать на достоверность применимости термина GO к данному белку в источнике (kind of evidence)

- Два принципа:
 - каждая аннотация должна ссылаться на источник:
 - литературная ссылка
 - другая база данных
 - компьютерное предсказание
 - аннотация должна указывать на достоверность применимости термина GO к данному белку в источнике (kind of evidence)



Evidence Codes

- IDA Inferred from Direct Assay
- TAS - Traceable Author Statement
- IMP - Inferred from Mutant Phenotype
- IGI - Inferred from Genetic Interaction
- IPI - Inferred from Physical Interaction
- RCA - Inferred from Reviewed Computational Analysis
- ISS - Inferred from Sequence Similarity
- IEP - Inferred from Expression Pattern
- NAS - Non-traceable Author Statement
- IEA - Inferred from Electronic Annotation
- IC - Inferred by Curator
- ND - No biological Data available



2013-09-25

Evidence Codes

Evidence Codes

- IDA Inferred from Direct Assay
- TAS - Traceable Author Statement
- IMP - Inferred from Mutant Phenotype
- IGI - Inferred from Genetic Interaction
- IPI - Inferred from Physical Interaction
- RCA - Inferred from Reviewed Computational Analysis
- ISS - Inferred from Sequence Similarity
- IEP - Inferred from Expression Pattern
- NAS - Non-traceable Author Statement
- IEA - Inferred from Electronic Annotation
- IC - Inferred by Curator
- ND - No biological Data available

Предостережение:

GO не является номенклатурой генов или их продуктов (белков).
Словари описывают биологические феномены (например,
программируемую клеточную смерть), а не конкретные
биологические объекты



2013-09-25

└ Предостережение:

Предостережение:

GO не является номенклатурой генов или их продуктов (белков).
Словари описывают биологические феномены (например,
программируемую клеточную смерть), а не конкретные
биологические объекты

Вопросы

2013-09-25

↳ Вопросы



