Практическая биоинформатика

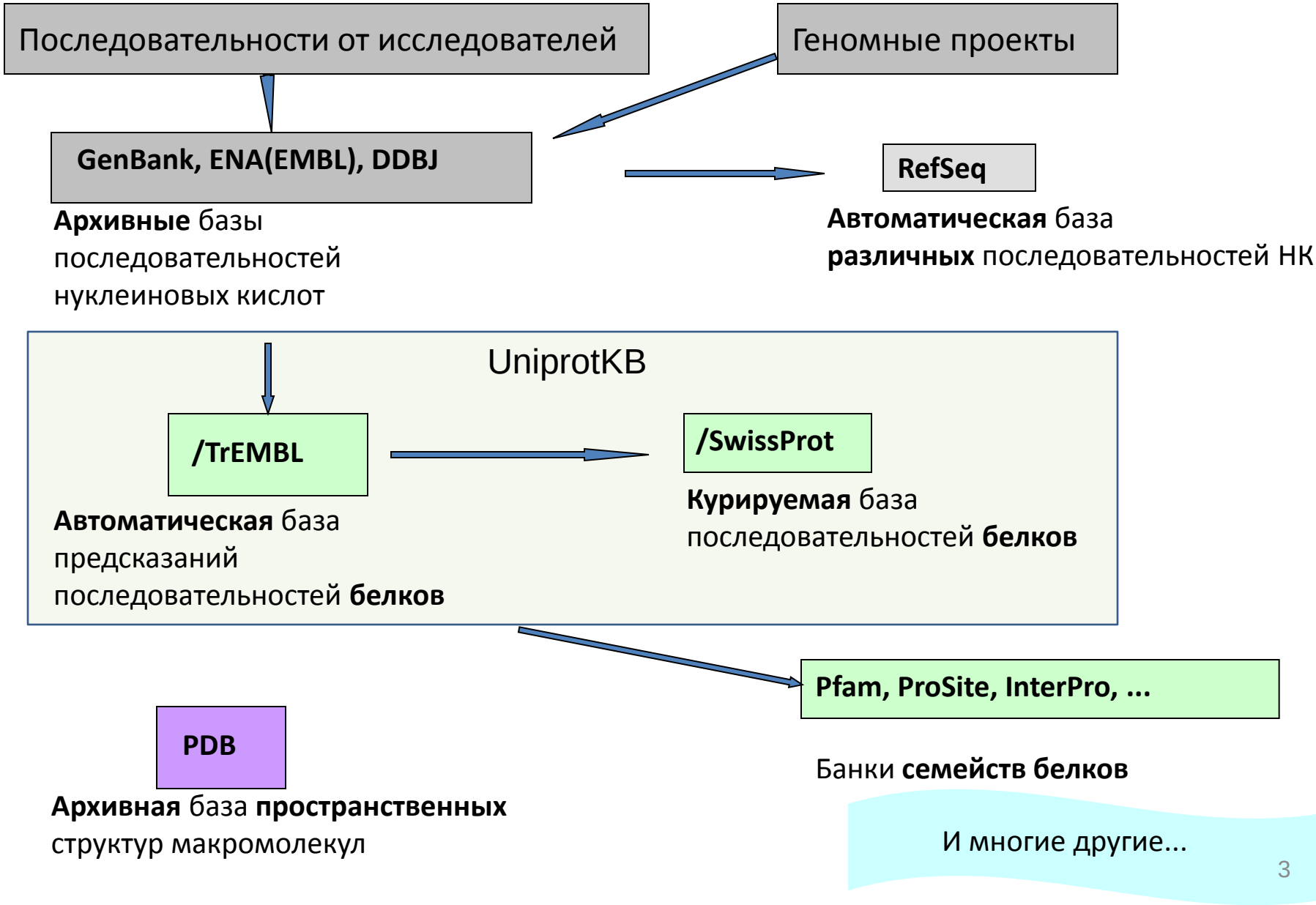
1. Банки данных

Сергей Александрович Спирин
sas@belozersky.msu.ru

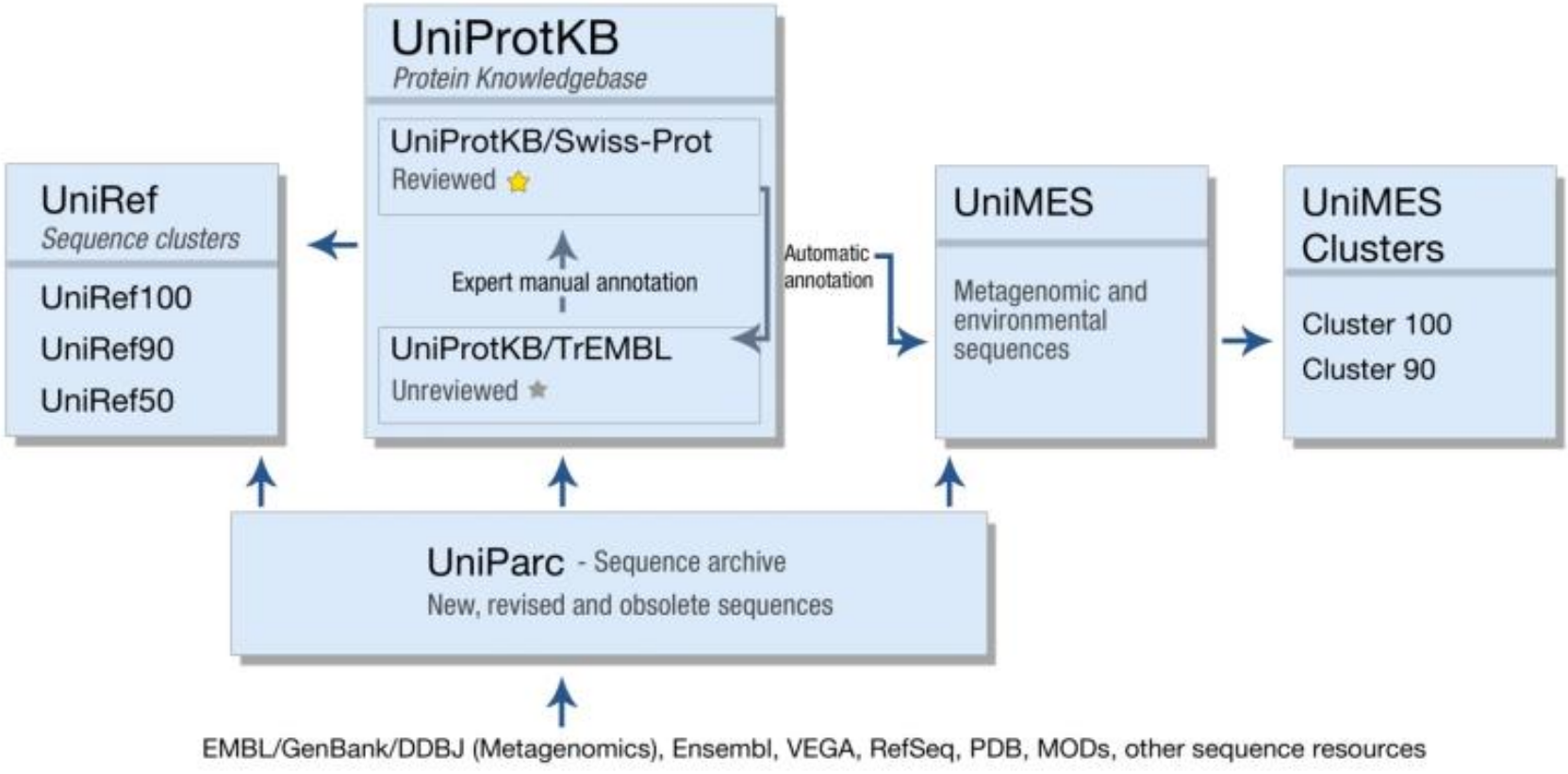
Необходимые программы

- EMBOSS (под Windows mEMBOSS)
- BLAST+
- Jalview
- MEGA
- Jmol

Банки структурной биологической информации



Последовательности белков



Базы данных (общие принципы)

- БД состоит из одного или нескольких хранилищ (“таблиц”)
- Единица хранения (строка таблицы) называется *записью* (entry).
- Все записи состоят из *полей* (fields). 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

Банк данных Swiss-Prot

1986

Swiss-Prot – база знаний о белковых последовательностях

- Курируемая база данных
- “Золотой стандарт” аннотации

http://web.expasy.org/docs/swiss-prot_guideline.html

Банк данных Swiss-Prot

С 1987 поддерживается в сотрудничестве между

Swiss **I**nstitute of **B**ioinformatics (**SIB**)
European **B**ioinformatics **I**nstitute (**EBI**)

С 2002 является частью **UniProt knowledgebase**, поддерживаемой UniProt consortium

Амос Байрох

Долговременный руководитель группы Swiss-Prot в Швейцарском Институте Биоинформатики

Физически Swiss-Prot – это один текстовый файл специального формата.

Банк данных TrEMBL

TrEMBL (**Tr**anslated **EMBL**)

Вместе со Swiss-Prot образует UniProtKB.

Формальная трансляция всех кодирующих нуклеотидных последовательностей из банка EMBL.

Автоматическая классификация и аннотация.

Формат записи тот же, что у Swiss-Prot.
Запись можно отличить по слову Unreviewed в первой строке.

Документ (запись, entry) Uniprot

```
ID   Y400_STACT              Reviewed;         188 AA.
AC   P47995; B9DJM1;
DT   01-FEB-1996, integrated into UniProtKB/Swiss-Prot.
DT   05-MAY-2009, sequence version 2.
DT   22-JUL-2015, entry version 67.
DE   RecName: Full=Uncharacterized protein Sca_0400;
DE   AltName: Full=ORF1;
GN   OrderedLocusNames=Sca_0400;
OS   Staphylococcus carnosus (strain TM300).
OC   Bacteria; Firmicutes; Bacilli; Bacillales; Staphylococcus.
OX   NCBI_TaxID=396513;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=TM300;
RX   PubMed=19060169; DOI=10.1128/AEM.01982-08;
RA   Rosenstein R., Nerz C., Biswas L., Resch A., Raddatz G.,
RA   Schuster S.C., Goetz F.;
RT   "Genome analysis of the meat starter culture bacterium Staphylococcus
RT   carnosus TM300.";
RL   Appl. Environ. Microbiol. 75:811-822(2009).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA] OF 28-188.
RX   PubMed=7557338; DOI=10.1111/j.1574-6968.1995.tb07787.x;
RA   Klein M., Meens J., Freudl R.;
RT   "Functional characterization of the Staphylococcus carnosus SecA
RT   protein in Escherichia coli and Bacillus subtilis secA mutant
RT   strains.";
RL   FEMS Microbiol. Lett. 131:271-277(1995).
CC   -!- SIMILARITY: Belongs to the ribosomal protein S30Ae family.
CC       {ECO:0000305}.
CC   -!- SEQUENCE CAUTION:
CC       Sequence=CAA56161.1; Type=Frameshift; Positions=25, 46; Evidence={ECO:0000305};
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   EMBL; X79725; CAA56161.1; ALT_FRAME; Genomic_DNA.
...
PE   3: Inferred from homology;
KW   Complete proteome; Reference proteome.
FT   CHAIN         1    188       Uncharacterized protein Sca_0400.
FT                                /FTId=PRO_0000208591.
FT   CONFLICT     52     52       Missing (in Ref. 2; CAA56161).
FT                                {ECO:0000305}.
...
SQ   SEQUENCE   188 AA;  21886 MW;  25E12963AB332A5F CRC64;
     MIRFEIHGDN LTITDAIRNY IEDKVGKLER YFTNVPNVNA HVKVKTYANS STKIEVTIPL
     NDVTLRAEER NDDLYAGIDL ITNKLERQVR KYKTRVNRKK RKESEHEPFP ATPETPPETA
     VDHDKDDEIE IIRSKQFSLK PMDSEEAVLQ MDLLGHDFFI FNDRETDGTS IVYRRKDGKY
     GLIETVEN
//
```

Описание документа: идентификатор, имя, дата создания и модификации

Аннотация последовательности

Последовательность

Основные поля Uniprot

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. 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

http://www.uniprot.org/uniprot/P00174.txt
http://www.uniprot.org/uniprot/P37869.txt
http://www.uniprot.org/uniprot/P27358.txt

Структура идентификатора записи Swiss-Prot

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 различения 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.

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 “VARSPLIC”);
и т. п.
Имеет строгий формат: Feature Key, FtLocation, FtDescription.

Например:

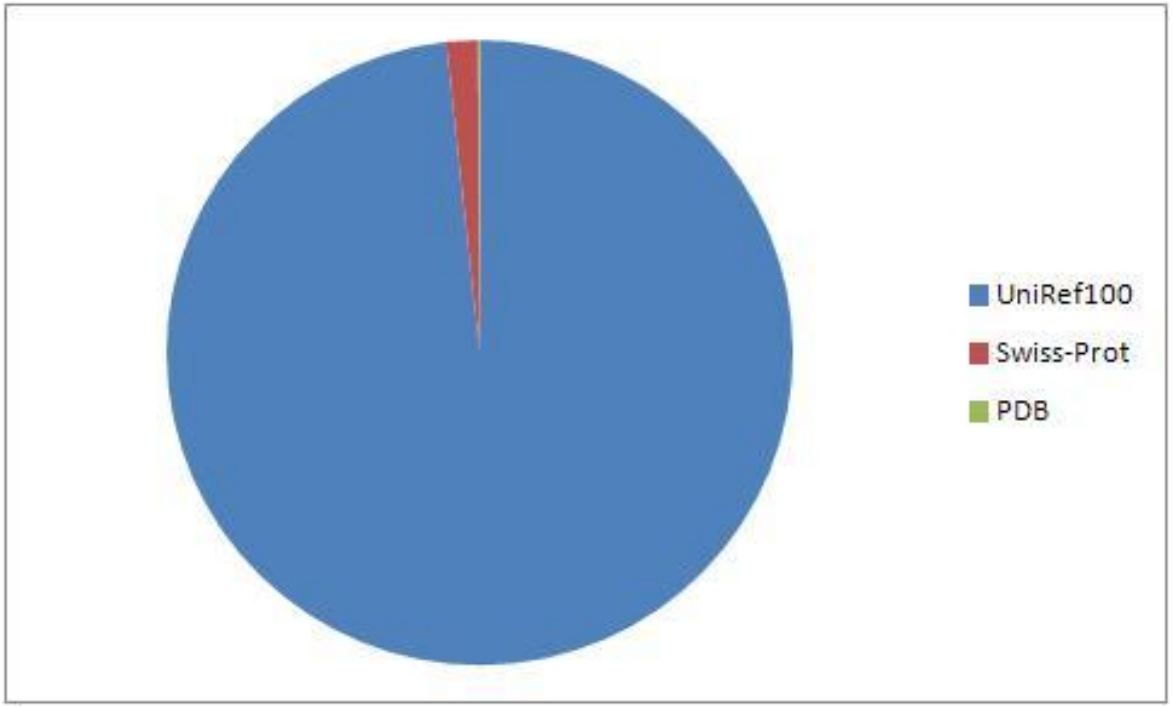
```
FT DISULFID 334 343 By similarity.
FT CONFLICT 138 138 E -> EE (in Ref. 4; AA sequence).
```

UniProt на 22 июля 2015

- SwissProt – 549 008 (~0,5 млн. белков)
- TrEMBL – 50 011 027 (~50 млн. записей)
- UniRef100 – 63 391 201 (~60 млн. различных аминокислотных последовательностей)

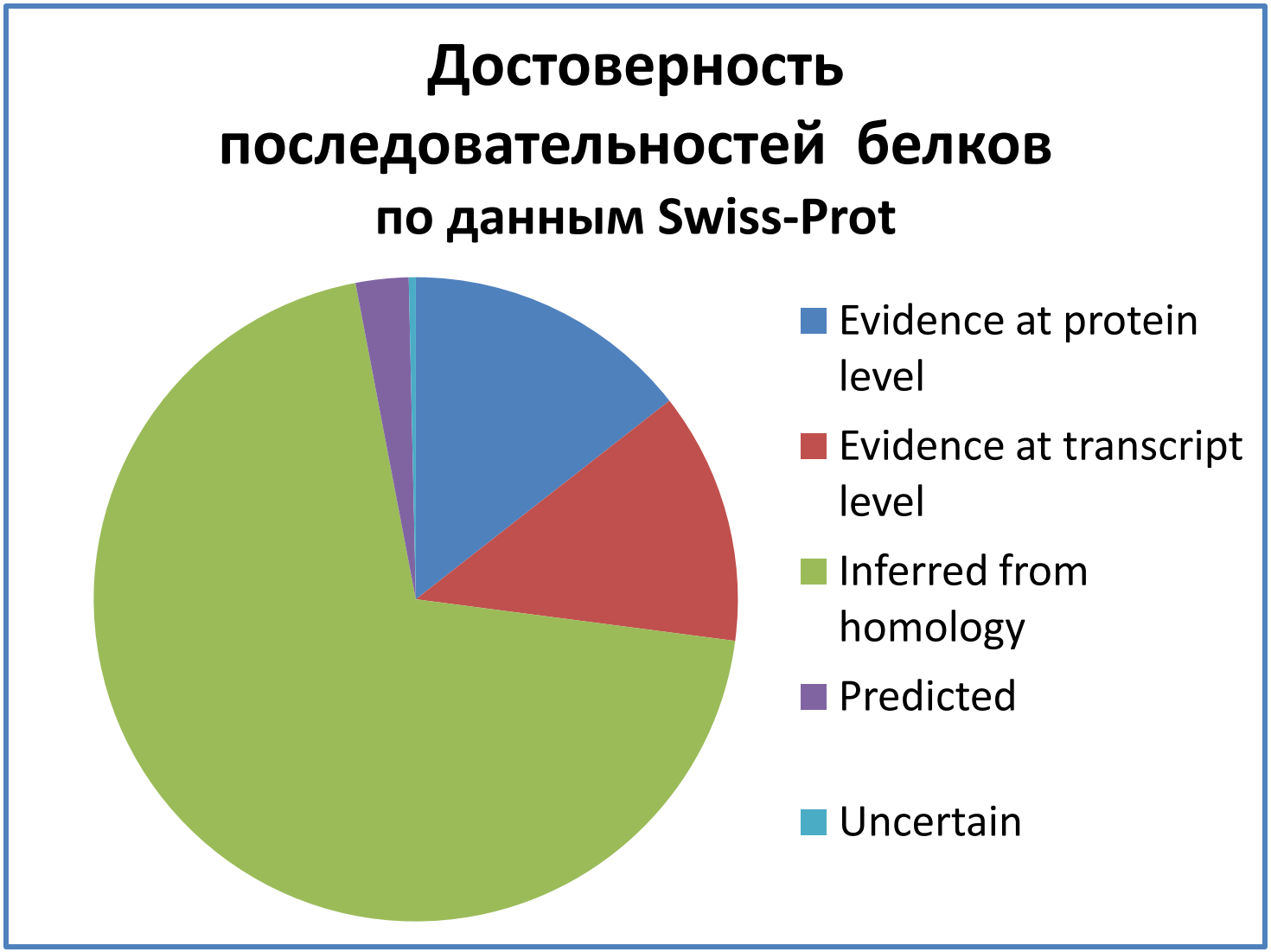
Для сравнения: банк PDB (пространственные структуры) содержит 111 749 записей, представляющих около 35 000 различных белков.

Число белков в разных БД



Последовательностей во много раз больше, чем структур!

Большинство последовательностей не аннотированы!



Более половины последовательностей Swiss-Prot не охарактеризовано экспериментально

Банки GenBank, EMBL, DDBJ

Содержат результаты работ по секвенированию нуклеиновых кислот.

Архивные банки: за содержание записей несут ответственность только их авторы.

С конца 1980-х годов журналы не публикуют работы о секвенировании последовательностей ДНК и РНК, если сами эти последовательности не депонированы в один из этих банков.

Ежедневный обмен данными.

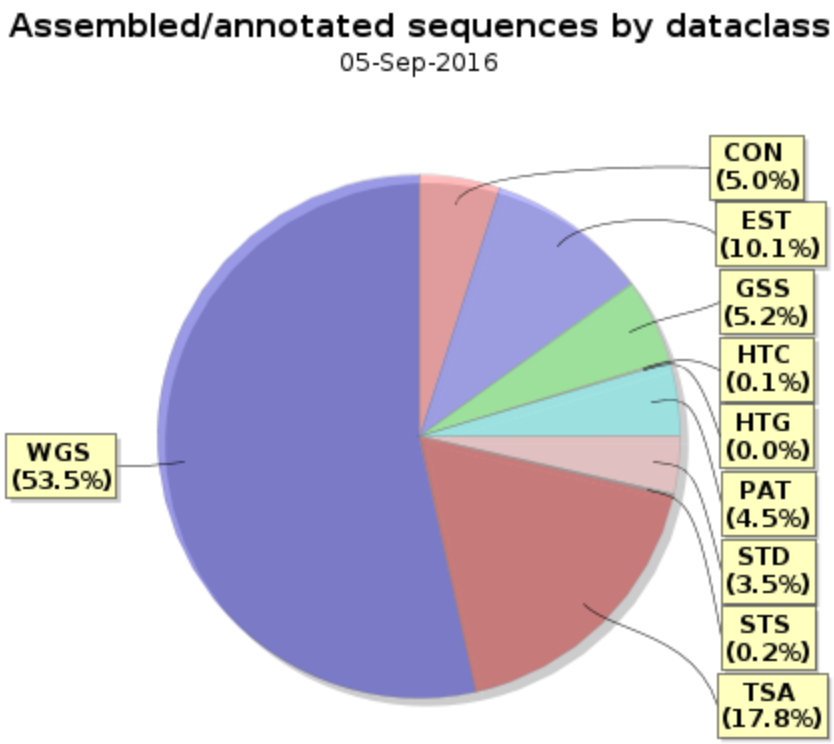
Версия EMBL от 5 сентября 2016 г. содержит 759,3 млн. последовательностей и 1854,7 млрд. нуклеотидов (http://www.ebi.ac.uk/ena/about/statistics).

Помимо EMBL, Европейский нуклеотидный архив (ENA, http://www.ebi.ac.uk/ena/) включает ещё SRA (Sequence Read Archive): ~ 24,5 трлн. последовательностей, 3,35 квадрильона нуклеотидов.

Разделы EMBL

HUM: Human
MUS: Mus musculus
ROD: Other Rodents
MAM: Other Mammals
VRT: Other Vertebrates
INV: Invertebrates
FUN: Fungi
PLN: Plants
PRO: Prokaryotes
VRL: Viruses
PHG: Bacteriophage
ENV: Environmental Samples
SYN: Synthetic
TGN: Transgenic
UNC: Unclassified

Классы данных EMBL



http://www.ebi.ac.uk/ena/about/statistics

http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html

RefSeq

- Поддерживается NCBI: http://www.ncbi.nlm.nih.gov/refseq/
- Не содержит повторений (в отличие от GenBank!)
- Состоит из трёх частей: RefSeq genomic, RefSeq RNA (только мРНК!), RefSeq protein
- Призван навести порядок в сумбуре секвенируемых последовательностей. Но, конечно, в связи с этим отстаёт…

Геномные браузеры

- USCS Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway)
  - продвинутый графический интерфейс для аннотированных геномов избранных эукариот
- NCBI (http://www.ncbi.nlm.nih.gov/genome/browse)
  - все полные геномы
- EnsEMBL (http://www.ensembl.org/)
  - продвинутый графический интерфейс для хорошо аннотированных геномов избранных эукариот (два десятка животных и дрожжи)
- EnsEMBL genomes (http://www.ensemblgenomes.org/)
  - расширение возможностей EnsEMBL на все геномы (в процессе…)

Форматы хранения последовательностей

- Swiss-Prot – для белков
- EMBL, GenBank – для нуклеотидных последовательностей
- Fasta – универсальный формат для хранения одной или многих последовательностей.

Понимается подавляющим большинством программ работы с последовательностями.

Формат Fasta

```
>sp|P00174|CYB5_CHICK Cytochrome b5 OS=Gallus gallus GN=CYB5A PE=1 SV=4
MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
GGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKPAETLITTVQSNSSSWSNWVI
PAIAAIIVALMYRSYMSE
```

Формат Fasta

Знак “>” – показатель строки с названием

```
>sp|P00174|CYB5_CHICK Cytochrome b5 OS=Gallus gallus GN=CYB5A PE=1 SV=4
MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
GGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKPAETLITTVQSNSSSWSNWVI
PAIAAIIVALMYRSYMSE
```

Формат Fasta

Знак “>” – показатель строки с названием

Имя последовательности
(до первого пробела).
В данном случае состоит из трёх «полей».

```
>sp|P00174|CYB5_CHICK Cytochrome b5 OS=Gallus gallus GN=CYB5A PE=1 SV=4
MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
GGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKPAETLITTVQSNSSSWSNWVI
PAIAAIIVALMYRSYMSE
```

Формат Fasta

Знак “>” – показатель строки с названием

Имя последовательности
(до первого пробела).
В данном случае состоит из трёх «полей».

Описание последовательности
(от первого пробела до конца строки).
Может отсутствовать.

```
>sp|P00174|CYB5_CHICK Cytochrome b5 OS=Gallus gallus GN=CYB5A PE=1 SV=4
MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
GGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKPAETLITTVQSNSSSWSNWVI
PAIAAIIVALMYRSYMSE
```

Формат Fasta

Знак “>” – показатель строки с названием

Имя последовательности
(до первого пробела).
В данном случае состоит из трёх «полей».

Описание последовательности
(от первого пробела до конца строки).
Может отсутствовать.

```
>sp|P00174|CYB5_CHICK Cytochrome b5 OS=Gallus gallus GN=CYB5A PE=1 SV=4
MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
GGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKPAETLITTVQSNSSSWSNWVI
PAIAAIIVALMYRSYMSE
```

Последовательность в однобуквенном коде, в одну или несколько строк.

Формат Fasta

(много последовательностей)

```
>sp|P00167|CYB5_HUMAN Cytochrome b5 OS=Homo sapiens GN=CYB5A PE=1 SV=2
MAEQSDEAVKYYTLEEIQKHNHSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQAGGDAT
ENFEDVGHSTDAREMSKTFIIGELHPDDRPKLNKPPETLITTIDSSSSWWTNWVIPAISA
VAVALMYRLYMAED
>sp|O43169|CYB5B_HUMAN Cytochrome b5 type B OS=Homo sapiens GN=CYB5B PE=1
SV=2
MATAEASGSDGKGQEVETSVTYYRLEEVAKRNSLKELWLVIHGRVYDVTRFLNEHPGGEE
VLLEQAGVDASESFEDVGHSSDAREMLKQYYIGDIHPSDLKPESGSKDPSKNDTCKSCWA
YWILPIIGAVLLGFLYRYYTSESKSS
>sp|P04166|CYB5B_RAT Cytochrome b5 type B OS=Rattus norvegicus GN=Cyb5b PE=1
SV=2
MATPEASGSGRNGQGSDPAVTYYRLEEVAKRNTAEETWMVIHGRVYDITRFLSEHPGGEE
VLLEQAGADATESFEDVGHSPDAREMLKQYYIGDVHPNDLKPKDGDKDPSKNNSCQSSWA
YWIVPIVGAILIGFLYRHFWADSKSS
>sp|P00173|CYB5_RAT Cytochrome b5 OS=Rattus norvegicus GN=Cyb5a PE=1 SV=2
MAEQSDKDVKYYTLEEIQKHKDSKSTWVILHHKVYDLTKFLEEHPGGEEVLREQAGGDAT
ENFEDVGHSTDARELSKTYIIGELHPDDRSKIAKPSETLITTVESNSSWWTNWVIPAISA
LVVALMYRLYMAED
>sp|P00174|CYB5_CHICK Cytochrome b5 OS=Gallus gallus GN=CYB5A PE=1 SV=4
MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
GGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKPAETLITTVQSNSSSWSNWVI
PAIAAIIVALMYRSYMSE
```