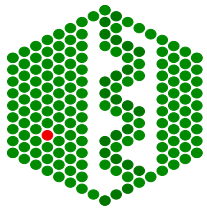
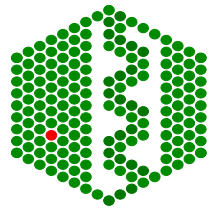
PROTEIN DATABASES



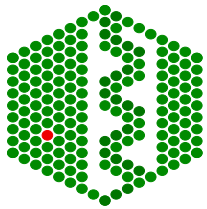
The ideal sequence database for computational analyses and data-mining:

- It must be complete with minimal redundancy
- It must contain as much up-to-date information (annotation) as possible on each sequence
- All the information items must be retrievable by computer programs in a consistent manner
- It must be highly interoperable with other databases



PROTEIN DATABASES

- SWISS-PROT - Manually curated (EBI/SIB)
- TrEMBL - Translation of EMBL (EBI)
- PIR - annotated sequences (NCBI)
- GenPept -GenBank translations
- NRL_3D - Sequences from PDB
- OWL - Non-redundant sequences
- RefSeq - Non-redundant sequence set
- Kabat & IMGT - Immunological proteins



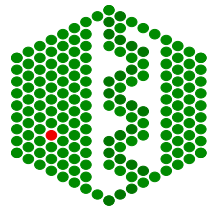
PIR (Protein Information Resource)

mips

- http://pir.georgetown.edu/pirwww/pirhome.shtml
- Sources: GenBank/EMBL/DDBJ translations, literature, direct submissions

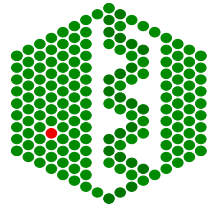
  -PIR-PSD (merging, annotation, classification)

  -PIR-Archive (original sequences)
- Total ~200 000 non-redundant sequences



Annotation in PIR

- Annotation is from literature and available databases
- Uses controlled vocabulary and std nomenclature (Enzyme nomenclature)
- Includes status tags “validated, expt’l, similarity, predicted, absent”
- Classification into superfamilies and homology domain superfamilies
- Classification is used for applying common annotation to similar sequences and integrity checks



Example of a PIR entry (1)

```
ENTRY           CCHU  #type complete
TITLE           cytochrome c [validated] - human
ORGANISM        #formal_name Homo sapiens #common_name man
   #cross-references taxon:9606
DATE            24-Apr-1984 #sequence_revision 30-Sep-1991 #text_change
                28-Jul-2000
ACCESSIONS      A31764; A05676; I55192; A00001
REFERENCE       A31764
   #authors     Evans, M.J.; Scarpulla, R.C.
   #journal     Proc. Natl. Acad. Sci. U.S.A. (1988) 85:9625-9629
   #title       The human somatic cytochrome c gene: two classes of
                processed pseudogenes demarcate a period of rapid
                molecular evolution.
   #cross-references MUID:89071748
   #accession   A31764
      ##molecule_type DNA
      ##residues 1-105 ##label EVA
      ##cross-references GB:M22877; NID:g181241; PIDN:AAA35732.1;
                PID:g181242
REFERENCE       A05676
   #authors     Matsubara, H.; Smith, E.L.
   #journal     J. Biol. Chem. (1963) 238:2732-2753
   #title       Human heart cytochrome c. Chymotryptic peptides, tryptic
                peptides, and the complete amino acid sequence.
   #accession   A05676
      ##molecule_type protein
      ##residues 2-28;29-46;47-100;101-105 ##label MATS
REFERENCE       A00001
   #authors     Matsubara, H.; Smith, E.L.
   #journal     J. Biol. Chem. (1962) 237:3575-3576
   #title       The amino acid sequence of human heart cytochrome c.
   #contents    annotation
   #note        66-Leu is found in 10% of the molecules in pooled protein
REFERENCE       I55192
   #authors     Tanaka, Y.; Ashikari, T.; Shibano, Y.; Amachi, T.;
                Yoshizumi, H.; Matsubara, H.
   #journal     J. Biochem. (1988) 103:954-961
   #title       Construction of a human cytochrome c gene and its
                functional expression in Saccharomyces cerevisiae.
   #cross-references MUID:89008207
   #accession   I55192
      ##status translated from GB/EMBL/DDBJ
      ##molecule_type mRNA
      ##residues 78-105 ##label RES
```

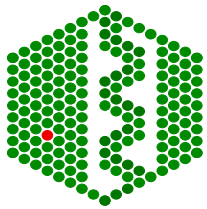
Link to list of entries for this species

Acc no.s of sequences merged with this entry

Links to EMBL/GenBank/DDBJ etc

Link to other entries with same citation

Link creates sequence reported for this reference



Example of a PIR entry (2)

```
GENETICS
   #introns      57/1
CLASSIFICATION  #superfamily cytochrome c; cytochrome c homology
KEYWORDS        acetylated amino end; chromoprotein; electron transfer;
                heme; iron; metalloprotein; mitochondrion; oxidative
                phosphorylation; polymorphism; respiratory chain
FEATURE
   2-105                #product cytochrome c #status experimental #label
                        MAT\
   5-99                 #domain cytochrome c homology #label CYC\
   2                    #modified site acetylated amino end (Gly) (in
                        mature form) #status experimental\
   15,18                #binding site heme (Cys) (covalent) #status
                        experimental\
   19,81                #binding site heme iron (His, Met) (axial
                        ligands) #status predicted
SUMMARY         #length 105 #molecular_weight 11749

SEQUENCE
             5        10        15        20        25        30
    1 M G D V E K G K K I F I M K C S Q C H T V E K G G K H K T G
   31 P N L H G L F G R K T G Q A P G Y S Y T A A N K N K G I I W
   61 G E D T L M E Y L E N P K K Y I P G T K M I F V G I K K K E
   91 E R A D L I A Y L K K A T N E
```

Link of entries classified into this superfamily or with this domain

List of entries with these keywords

List of other PIR entries with this feature

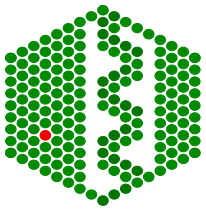
PDB structures most related to CCHU:
1GIW (2-105) 88.5%

Link to PDB entry for this sequence

ALIGNMENTS containing CCHU:
FA2814 cytochrome c - 1.0 1.0
SA2325 cytochrome c superfamily 1.0
M00001 cytochrome c - 1.0 8.0
ALIGNMENTS containing CCHU(5-99):
DA1115 cytochrome c homology

Alignments involving this protein

Link to *i*ProClass (Superfamily classification and Alignment):
*i*ProClass Report for CCHU at PIR.



Example of a PIR entry (3)

Link from top of entry page to Composition Table

Composition/Molecular Weight

Entry CCHU

Composition

6	Ala	A	5.7%	2	Gln	Q	1.9%
6	Leu	L	5.7%	2	Ser	S	1.9%
2	Arg	R	1.9%	8	Glu	E	7.6%
18	Lys	K	17.1%	7	Thr	T	6.7%
5	Asn	N	4.8%	13	Gly	G	12.4%
4	Met	M	3.8%	1	Trp	W	1.0%
3	Asp	D	2.9%	3	His	H	2.9%
3	Phe	F	2.9%	5	Tyr	Y	4.8%
2	Cys	C	1.9%	8	Ile	I	7.6%
4	Pro	P	3.8%	3	Val	V	2.9%

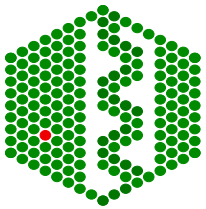
Number of residues = 105
Molecular weight (unmod. chain) = 11,749

CALCULATION NOTES:
1. Let the protein sequence be SEQ(i), for i=1,...,LENGTH.
   Let TOTAL be the sum of the individual weights:
       TOTAL = sum WGT(SEQ(i)), for i=1,...LENGTH.
   Let WATER be the molecular weight of water WGT('-') = 18.015.
   Then, the molecular weight of the protein is
       WEIGHT = TOTAL-(WATER*(LENGTH-1)).

2. For each amino acid x the table gives the molecular weight WGT(x)

A	Ala	89.09	C	Cys	121.15	D	Asp	133.10	E	Glu	147.13
F	Phe	165.19	G	Gly	75.07	H	His	155.16	I	Ile	131.17
K	Lys	146.19	L	Leu	131.17	M	Met	149.21	N	Asn	132.12
P	Pro	115.13	Q	Gln	146.15	R	Arg	174.20	S	Ser	105.09
T	Thr	119.12	V	Val	117.15	W	Trp	204.23	Y	Tyr	181.19
B	Asx	132.61	Z	Glx	146.64	-	H2O	18.015	X	X	128.16

Any other alphabetic character, in particular J, O, or U, is equivalent to X (i.e. WGT(character) = 128.16).



Query squence (length=390)

>Your input sequence

```
MEEPGAQCAPPPPAGSETWVPQANLSSAPSQNCSAKDYIYQDSISLPWKVLLVMLLALIT
LATTLSNAFVIATVYRTRKLHTPANYLIASLAVTDLLVSILVMPISTMYTVTGRWTLGQV
VCDFWLSSDITCCTASILHLCVIALDRYWAITDAVEYSAKRTPKRAAVMIALVWVFSISI
SLPPFFWRQAKAEEEVSECVVNTDHILYTVYSTVGAFYFPTLLLIALYGRIYVEARSRIL
KQTPNRTGKRLTRAQLITDSPGSTSSVTSINSRVPDVPSESGSPVYVNQVKVRVSDALLE
KKKLMAARERKATKTLGIILGAFIVCWLPFFIISLVMPICKDACWFHLAIFDFFTWLGYL
NSLINPIIYTMSNEDFKQAFHKLIRFKCTS
```

Search Results (database=PIR)

Thank you for using the PIR Annotation-Sorted Similarity Search. The results show **Superfamily** and are sorted by **BLAST score**. At PIR, we are committed to provide you with the best annotated protein sequence information. If you want to see the results with other annotation, please make your selection below and click on the submit button.

Select annotation: ◉ Superfamily ○ MIPSfamily ○ Species ○ Tax. group ○ Keywords

Sorted by: ◉ Score ○ Annotation [Submit] [Reset]

Tips

*If all the sequences from only one family, go back and search FAMBASE to see similar sequences from different family groups.

*Click on the color bars to see the full-length alignments. The color reflects the magnitude of the score, %idn = %identity, Ov.lap = size of overlap.

***Overlap ratio (OR) = Overlap / 390**, where 390 is the length of the query seq.

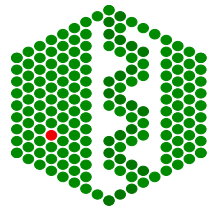
ID	Superfamily	Description	Score	#aa	%idn	Ov.lap	%idn x OR
JN0268	SF002406	serotonin receptor 1B - human	687	390	100	390	100
S68422	SF002406	serotonin receptor 1D beta - rabbit	650	389	92	390	92
S58126	SF002406	serotonin receptor 1-like - rabbit	650	390	93	390	93
S18637	SF002406	serotonin receptor 1B - rat	644	386	93	389	92
A42688	SF002406	serotonin receptor 1B - mouse	638	386	92	388	91
S54153	SF002406	serotonin receptor 1B - Chinese hamster	629	386	90	388	89
I77467	SF005474	serotonin receptor 1D - rat	437	374	66	343	58
A53279	SF005474	serotonin receptor 1D - human	436	377	62	380	60
B30341	SF005474	G protein-coupled receptor RDC4 - dog	430	377	61	381	59
S68423	SF005474	serotonin receptor 1D alpha - rabbit	426	377	61	377	58
A47321	SF005474	serotonin receptor 1F - human	313	366	48	365	44
S26048	SF005474	serotonin receptor 1E-beta - mouse	304	366	47	369	44
A47385	SF005474	serotonin receptor 1E - rat	294	366	46	369	43
A45260	SF005474	serotonin receptor 1E - human	290	365	47	364	43
I49375	SF002406	serotonin receptor 1A - mouse	288	421	42	393	42

Searching PIR for superfamily annotation

Automated classification of full-length sequences
>99% -families
>70% -superfamilies

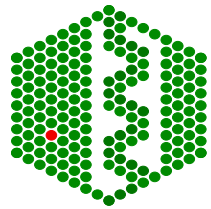
-Use 50% identity for clustering of proteins into families

-Also cluster into homology domain superfamilies



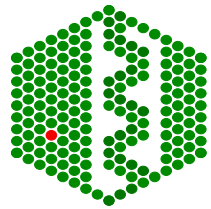
GenPept

```
1-------10--------20--------30--------40--------50--------60--------70------78

LOCUS       ABCAARAA_1
DEFINITION  A.aceti acetic acid resistance protein (aarA) gene, complete cds;
            acetic acid resistance protein (aarA).
DATE        15-SEP-1990
ACCESSION   M34830
ORGANISM    Acetobacter aceti
            Eubacteria; Proteobacteria; alpha subdivision; Acetobacteraceae;
            Acetobacter.
COMMENT     CDS 185..1495
            /db_xref="PID:g141730"
WEIGHT      48238
LENGTH      436
ORIGIN      Translated using phase 1
        1 MSASQKEGKL STATISVDGK SAEMPVLSGT LGPDVIDIRK LPAQLGVFTF DPGYGETAAC
       61 NSKITFIDGD KGVLLHRGYP IAQLDENASY EEVIYLLLNG ELPNKVQYDT FTNTLTNHTL
      121 LHEQIRNFFN GFRRDAHPMA ILCGTVGALS AFYPDANDIA IPANRDLAAM RLIAKIPTIA
      181 AWAYKYTQGE AFIYPRNDLN YAENFLSMMF ARMSEPYKVN PVLARAMNRI LILHADHEQN
      241 ASTSTVRLAG STGANPFACI AAGIAALWGP AHGGANEAVL KMLARIGKKE NIPAFIAQVK
      301 DKNSGVKLMG FGHRVYKNFD PRAKIMQQTC HEVLTELGIK DDPLLDLAVE LEKIALSDDY
      361 FVQRKLYPNV DFYSGIILKA MGIPTSMFTV LFAVARTTGW VSQWKEMIEE PGQRISRPRQ
      421 LYIGAPQRDY VPLAKR
//

1-------10--------20--------30--------40--------50--------60--------70------78
```

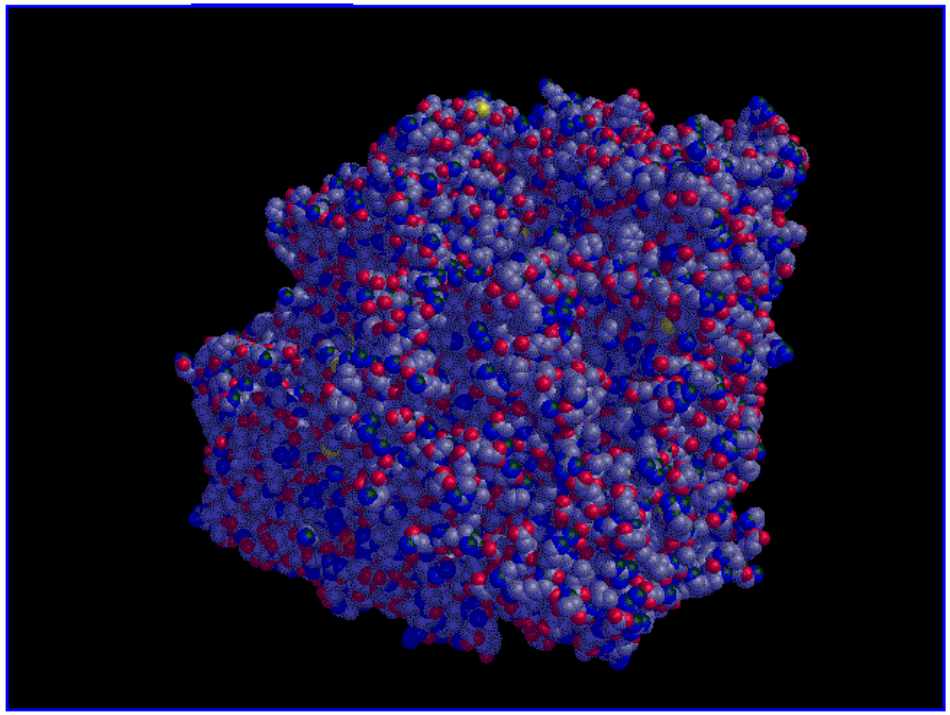



NRL_3D Database

- http://pir.georgetown.edu/pirwww/dbinfo/nrl_3D.html
- Protein database of sequences with 3D structure in PDB



```
ENTRY           1GPAA      #Type Protein
TITLE           glycogen phosphorylase A (r-state), chain A - rabbit
                  #EC-number 2.4.1.1
DATE            20-Dec-1992 #Sequence 20-Dec-1992 #Text 25-Aug-1993
PLACEMENT          0.0    0.0    0.0    0.0    0.0
COMMENT         PDB code: 1GPA
```



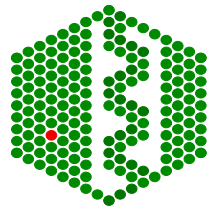
```
SOURCE          Oryctolagus cuniculus #Common-name domestic rabbit
COMMENT         Note: muscle
REFERENCE
   #Authors     Barford D., Hu S.H., Johnson L.N.
   #Citation    coordinates deposited in Brookhaven National
                  Laboratory's Protein Data Bank
REFERENCE
   #Authors     Barford D., Hu S.H., Johnson L.N.
   #Journal     J. Mol. Biol. (1991) 218:233
   #Title       Structural mechanism for glycogen phosphorylase
                  control by phosphorylation and AMP.
REFERENCE
   #Authors     Acharya K.R., Stuart D.I., Varvill K.M., Johnson
                  L.N.
   #Book        Glycogen Phosphorylase B: Description of the Protein
```

NRL_3D Example entry (1)

NRL_3D Example entry (2)

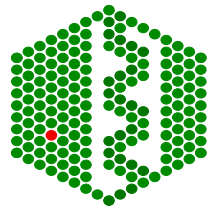
```
COMMENT         Resolution: 2.9 angstroms
COMMENT         R-value: 0.176
COMMENT         Determination: X-ray diffraction
KEYWORDS        Glycogen phosphorylase
FEATURE
    14-29                   #Region helix (right hand alpha)\
    38-69                   #Region helix (right hand alpha)\
    85-93                   #Region helix (right hand alpha)\
    95-106                  #Region helix (right hand alpha)\
    109-115                 #Region helix (right hand alpha)\
    125-141                 #Region helix (right hand alpha)\
    252-265                 #Region helix (right hand alpha)\
    280-305                 #Region helix (right hand alpha)\
    319-324                 #Region helix (right hand alpha)\
    335-346                 #Region helix (right hand alpha)\
    351-363                 #Region helix (right hand alpha)\
    379-387                 #Region helix (right hand alpha)\
    387-409                 #Region helix (right hand alpha)\
    411-420                 #Region helix (right hand alpha)\
    431-439                 #Region helix (right hand alpha)\
    447-457                 #Region helix (right hand alpha)\
    459-466                 #Region helix (right hand alpha)\
    479-486                 #Region helix (right hand pi)\
    487-499                 #Region helix (right hand alpha)\
    500-505                 #Region helix (right hand 3-10)\
    505-516                 #Region helix (right hand 3-10)\
    518-544                 #Region helix (right hand alpha)\
    566-584                 #Region helix (right hand alpha)\
    604-622                 #Region helix (right hand alpha)\
    640-648                 #Region helix (right hand alpha)\
    667-675                 #Region helix (right hand alpha)\
    686-695                 #Region helix (right hand alpha)\
    705-716                 #Region helix (right hand alpha)\
    719-726                 #Region helix (right hand alpha)\
    726-738                 #Region helix (right hand alpha)\
    749-759                 #Region helix (right hand alpha)\
    767-783                 #Region helix (right hand alpha)\
    784-797                 #Region helix (right hand alpha)\
    803-816                 #Region helix (right hand alpha)\
    182-184,213-223,
    228-238,144-151,72-77,
    323-330,362-367,
    442-445,469-475         #Region beta sheet\
    158-162,165-169         #Region beta sheet\
    153-154,267-270         #Region beta sheet\
    189-200,203-214         #Region beta sheet\
    80-83,120-122           #Region beta sheet\
    376-380,428-432,
    421-423                 #Region beta sheet\
    630-636,591-599
    584-587                 #Region turn (type I)\
    601-604                 #Region turn (type II)\
    603-605                 #Region turn (reverse gamma)\
    621-624                 #Region turn (type I)\
    624-627                 #Region turn (type III)\
    648-651                 #Region turn (type I)\
    659-662                 #Region turn (type II)\
    684-687                 #Region turn (type II')\
    695-698                 #Region turn (type III)\
    696-699                 #Region turn (type I)\
    739-742                 #Region turn (type I)\
    745-748                 #Region turn (type I)\
    747-750                 #Region turn (type I)\
    760-763                 #Region turn (type I)\
    764-767                 #Region turn (type III)\
    765-768                 #Region turn (type I)\
    796-799                 #Region turn (type I)\
    799-802                 #Region turn (type III)\
    800-803                 #Region turn (type I)
SUMMARY       #Molecular-weight 95534  #Length 828  #Checksum  7271
SEQUENCE
                5        10        15        20        25        30
      1 R K Q I S V R G L A G V E N V T E L K K N F N R H L H F T L
     31 V K D R N V A T P R D Y Y F A L A H T V R D H L V G R W I R
     61 T Q Q H Y Y E K D P K R I Y Y L S L E F Y M G R T L Q N T M
     91 V N L A L E N A C D E A T Y Q L G L D M E E L E E I E E D A
    121 G L G N G G L G R L A A C F L D S M A T L G L A A Y G Y G I
    151 R Y E F G I F N Q K I C G G W Q M E E A D D W L R Y G N P W
    181 E K A R P E F T L P V H F Y G R V E H T S Q G A K W V D T Q
    211 V V L A M P Y D T P V P G Y R N N V V N T M R L W S A K A P
    241 N D F N L K D F N V G G Y I Q A V L D R N L A E N I S R V L
    271 Y P N D N F F E G K E L R L K Q E Y F V V A A T L Q D I I R
    301 R F K S S K F G C R D P V R T N F D A F P D K V A I Q L N D
    331 T H P S L A I P E L M R V L V D L E R L D W D K A W E V T V
    361 K T C A Y T N H T V I P E A L E R W P V H L L E T L L P R H
    391 L Q I I Y E I N Q R F L N R V A A A F P G D V D R L R R M S
    421 L V E E G A V K R I N M A H L C I A G S H A V N G V A R I H
    451 S E I L K K T I F K D F Y E L E P H K F Q N K T N G I T P R
    481 R W L V L C N P G L A E I I A E R I G E E Y I S D L D Q L R
    511 K L L S Y V D D E A F I R D V A K V K Q E N K L K F A A Y L
    541 E R E Y K V H I N P N S L F D V Q V K R I H E Y K R Q L L N
    571 C L H V I T L Y N R I K K E P N K F V V P R T V M I G G K A
    601 A P G Y H M A K M I I K L I T A I G D V V N H D P V V G D R
```

EMBL
European Bioinformatics Institute



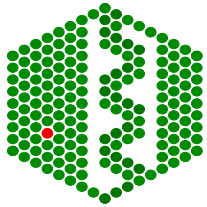
OWL

- http://www.bioinf.man.ac.uk/dbbrowser/OWL/
- Non-redundant protein database derived from SWISS-PROT, PIR, GenBank (translations) and NRL_3D
- 279,796 entries, small because of strict redundancy criteria
- All identical and trivially-different sequences (i.e. those having a single amino acid change) are removed
- SWISS-PROT is highest priority, NRL_3D lowest



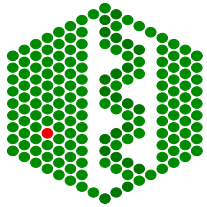
RefSeq

- http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
- Reference sequence standards for genomes, transcripts and proteins for human, mouse and rat
- Manually curated, non-redundant, status (genome annotation, predicted, provisional, reviewed)
- Includes data from NCBI Human Genome Annotation Project



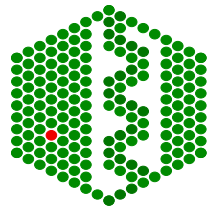
SWISS-PROT

- A curated protein sequence data bank established in July 1986 by Amos Bairoch in Geneva and now maintained collaboratively with EMBL
- Contains 94 000 manually annotated protein sequence entries (but >60% of all seq with some basic biochemical characterisation)
- Distinguishes between expt'l and comput'l derived annotation



SWISS-PROT STATISTICS

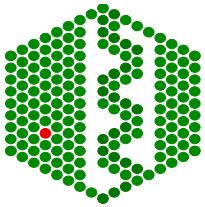
- 94 000 SWISS-PROT entries
- 32 000 000 amino acids
- abstracted from > 70 000 references
- linked by > 420 000 direct pointers to 35 related or specialized data collections



Example of a SWISS-PROT entry

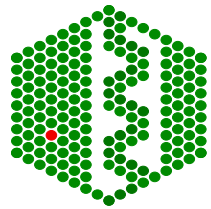
```
ID   GUNB_CLOTM     STANDARD;      PRT;   563 AA.
AC   P04956;
DT   13-AUG-1987 (Rel. 05, Created)
DT   13-AUG-1987 (Rel. 05, Last sequence update)
DT   01-FEB-1995 (Rel. 31, Last annotation update)
DE   ENDOGLUCANASE B PRECURSOR (EC 3.2.1.4) (EGB) (ENDO-1,4-BETA-GLUCANASE)
DE   (CELLULASE B).
GN   CELB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=86148508; PubMed=3453102;
RA   Grepinet O., Beguin P.;
RT   "Sequence of the cellulase gene of Clostridium thermocellum coding for
RT   endoglucanase B.";
RL   Nucleic Acids Res. 14:1791-1799(1986).
CC   -!- FUNCTION: THIS ENZYME CATALYZES THE ENDOHYDROLYSIS OF 1,4-BETA-
CC       GLUCOSIDIC LINKAGES IN CELLULOSE, LICHENIN AND CEREAL BETA-D-
CC       GLUCANS.
CC   -!- CATALYTIC ACTIVITY: ENDOHYDROLYSIS OF 1,4-BETA-D-GLUCOSIDIC
CC       LINKAGES IN CELLULOSE.
CC   -!- DOMAIN: A 24 RESIDUES DOMAIN IS REPEATED TWICE IN THIS ENZYME AS
CC       WELL AS IN OTHER C.THERMOCELLUM CELLULOSOME ENZYMES. THIS DOMAIN
CC       MAY FUNCTION AS THE BINDING LIGAND FOR THE SL COMPONENT.
CC   -!- SIMILARITY: BELONGS TO CELLULASE FAMILY A (FAMILY 5 OF GLYCOSYL
CC       HYDROLASES).
DR   EMBL; X03592; CAA27266.1; -.
DR   PIR; A23512; CZCLBM.
DR   HSSP; P54583; 1ECE.
DR   InterPro; IPR002105; Dockerin_1.
DR   InterPro; IPR002048; EF-hand.
DR   InterPro; IPR001547; Glyco_hydro_F5.
DR   Pfam; PF00150; cellulase; 1.
DR   Pfam; PF00404; Dockerin_1; 2.
DR   PROSITE; PS00018; EF_HAND; UNKNOWN_1.
DR   PROSITE; PS00448; CLOS_CELLULOSOME_RPT; 2.
DR   PROSITE; PS00659; GLYCOSYL_HYDROL_F5; 1.
KW   Cellulose degradation; Hydrolase; Glycosidase; Repeat; Signal.
FT   SIGNAL        1     27       OR 31.
FT   CHAIN        28    563       ENDOGLUCANASE B.
FT   ACT SITE     204    204      PROTON DONOR (BY SIMILARITY).
FT   ACT SITE     363    363      NUCLEOPHILE (BY SIMILARITY).
FT   DOMAIN       502    557      2 X 24 AA APPROXIMATE REPEATS.
FT   REPEAT       502    526      1.
FT   REPEAT       534    557      2.
SQ   SEQUENCE   563 AA;  63929 MW;  866FE55704A1DE4B CRC64;
     MKKFLVLLIA LIMIATLLVV PGVQTSAEGS YADLAEPDDD WLHVEGTNIV DKYGNKVWIT
     GANWFGFNCR ERMLLDSYHS DIIADIELVA DKGINVVRMP IATDLLYAWS QGIYPPSTDT
     SYNNPALAGL NSYELFNFML ENFKRVGIKV ILDVHSPETD NQGHNYPLWY NTTITEEIFK
     KAWVWVAERY KNDDTIIGFD LKNEPHTNTG TMKIKAQSAI WDDSNHPNNW KRVAEETALA
     ILEVHPNVLI FVEGVEMYPK DGIWDDETFD TSPWTGNNDY YGNWWGGNLR GVKDYPINLG
     KYQSQLVYSP HDYGPIVYEQ DWFKGDFITA NDEQAKRILY EQCWRDNWAY IMEEGISPLL
     LGEWGGMTEG GHPLLDLNLK YLRCMRDFIL ENKYKLHHTF WCINIDSADT GGLFTRDEGT
     PFPGGRDLKW NDNKYDNYLY PVLWKTEDGK FIGLDHKIPL GRNGISISQL SNYTPSVTPS
     PSATPSPTTI TAPPTDTVTY GDVNGDGRVN SSDVALLKRY LLGLVENINK EAADVNVSGT
     VNSTDLAIMK RYVLRSISEL PYK
//
```

EMBL
European Bioinformatics Institute



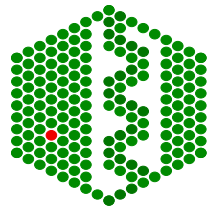
The annotation is mainly found in:

- Comment (CC) lines
- Feature table (FT)
- Keyword (KW) lines
- Description (DE) lines



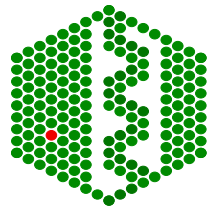
The topics of the CC lines are:

- ALTERNATIVE PRODUCTS
- CATALYTIC
- CAUTION
- COFACTOR
- DEVELOPMENTAL STAGE
- DISEASE
- DOMAIN
- ENZYME REGULATION
- FUNCTION
- INDUCTION
- MASS SPECTROMETRY
- PATHWAY
- PHARMACEUTICALS
- POLYMORPHISM
- PTM
- SIMILARITY
- SUBCELLULAR LOCATION
- SUBUNIT
- TISSUE SPECIFICITY



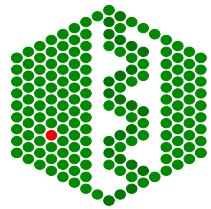
The FT keys are handling:

- Change indicators
- Amino-acid modifications
- Regions
- Secondary structure
- Other features



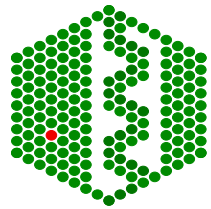
Change indicators are:

- CONFLICT - Different papers report differing sequences
- VARIANT - Authors report that sequence variants exist
- VARSPLIC - Description of sequence variants produced by alternative splicing
- MUTAGEN - Site which has been experimentally altered



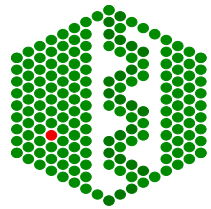
Amino-acid modifications are:

- MOD_RES - Post-translational modification of a residue
- LIPID - Covalent binding of a lipidic moiety
- DISULFID - Disulfide bond
- THIOLEST - Thiolester bond
- THIOETH - Thioether bond
- CARBOHYD - Glycosylation site
- METAL - Binding site for a metal ion
- BINDING - Binding site for any chemical group (co-enzyme, prosthetic group, etc.)



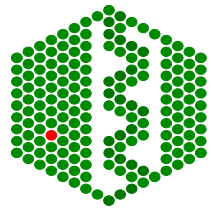
Regions:

- SIGNAL
- TRANSIT
- PROPEP
- CHAIN
- PEPTIDE
- DOMAIN
- CA_BIND
- DNA_BIND
- NP_BIND
- TRANSMEM
- ZN_FING
- SIMILAR
- REPEAT



Other features are:

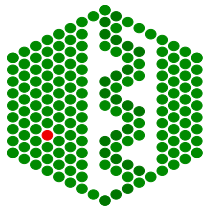
- ACT_SITE - Amino acid(s) involved in the activity of an enzyme
- SITE - Any other interesting site on the sequence
- INIT_MET - The sequence is known to start with an initiator methionine
- NON_TER - The residue at an extremity of the sequence is not the terminal residue
- NON_CONS - Non consecutive residues
- UNSURE - Uncertainties in the sequence



The KW lines:

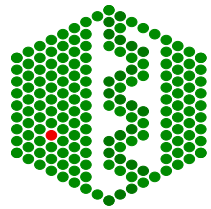
- around 800 different keywords
- keyword dictionary available
- Controlled use of the keywords has cross-references

- **DBXREFS** – crossreferences to about 30 databases including pattern dbs, specialised genome dbs, other sequence dbs

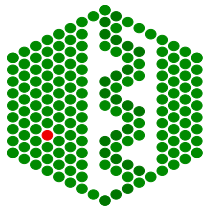


Annotation sources:

- publications that report new sequence data
- review articles to periodically update the annotation of families or groups of proteins
- external experts

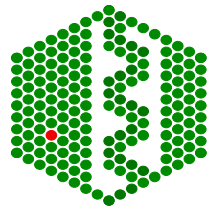


1.9.1998: SWISS-PROT ceased to be in the public domain

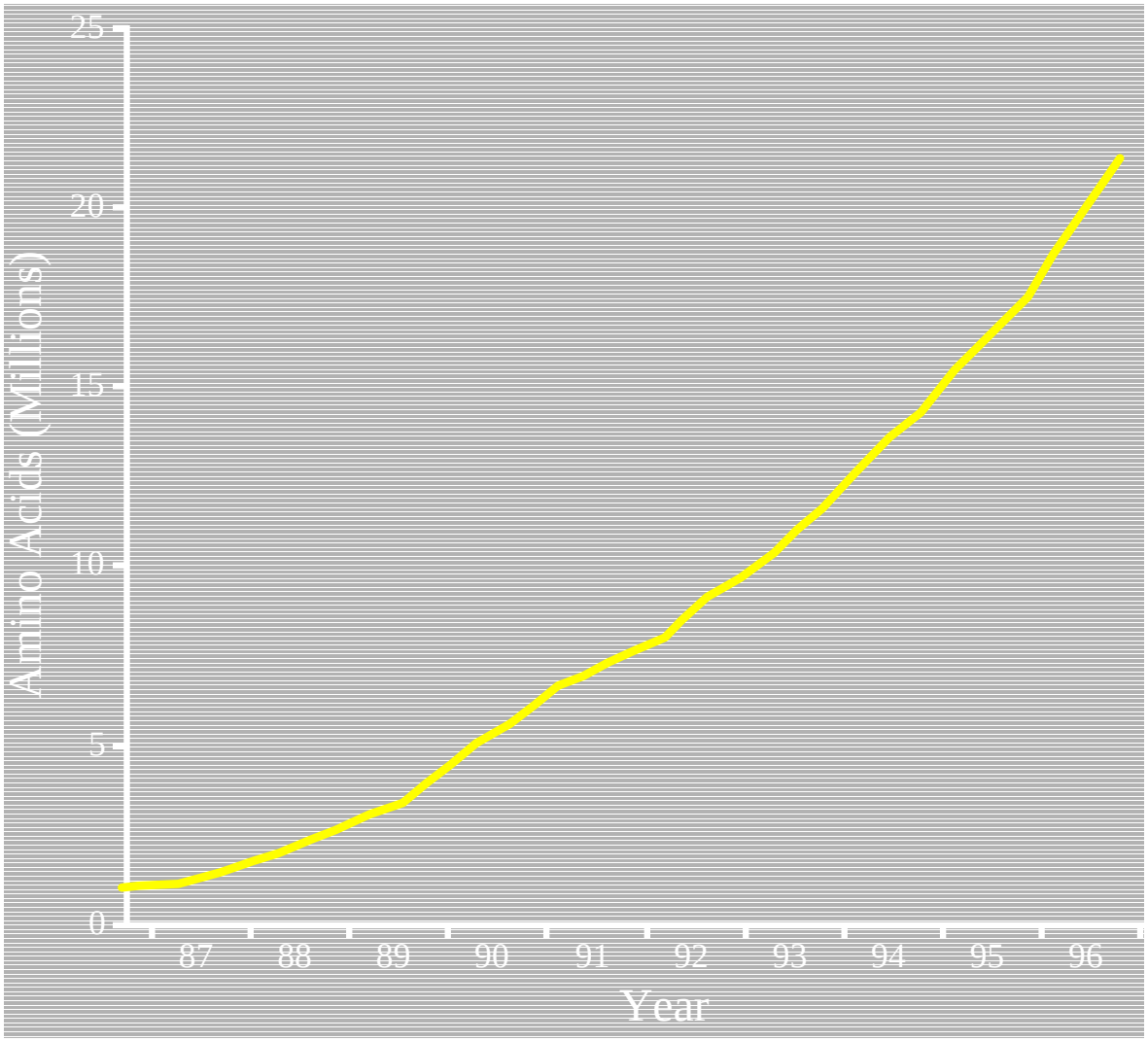


What has changed

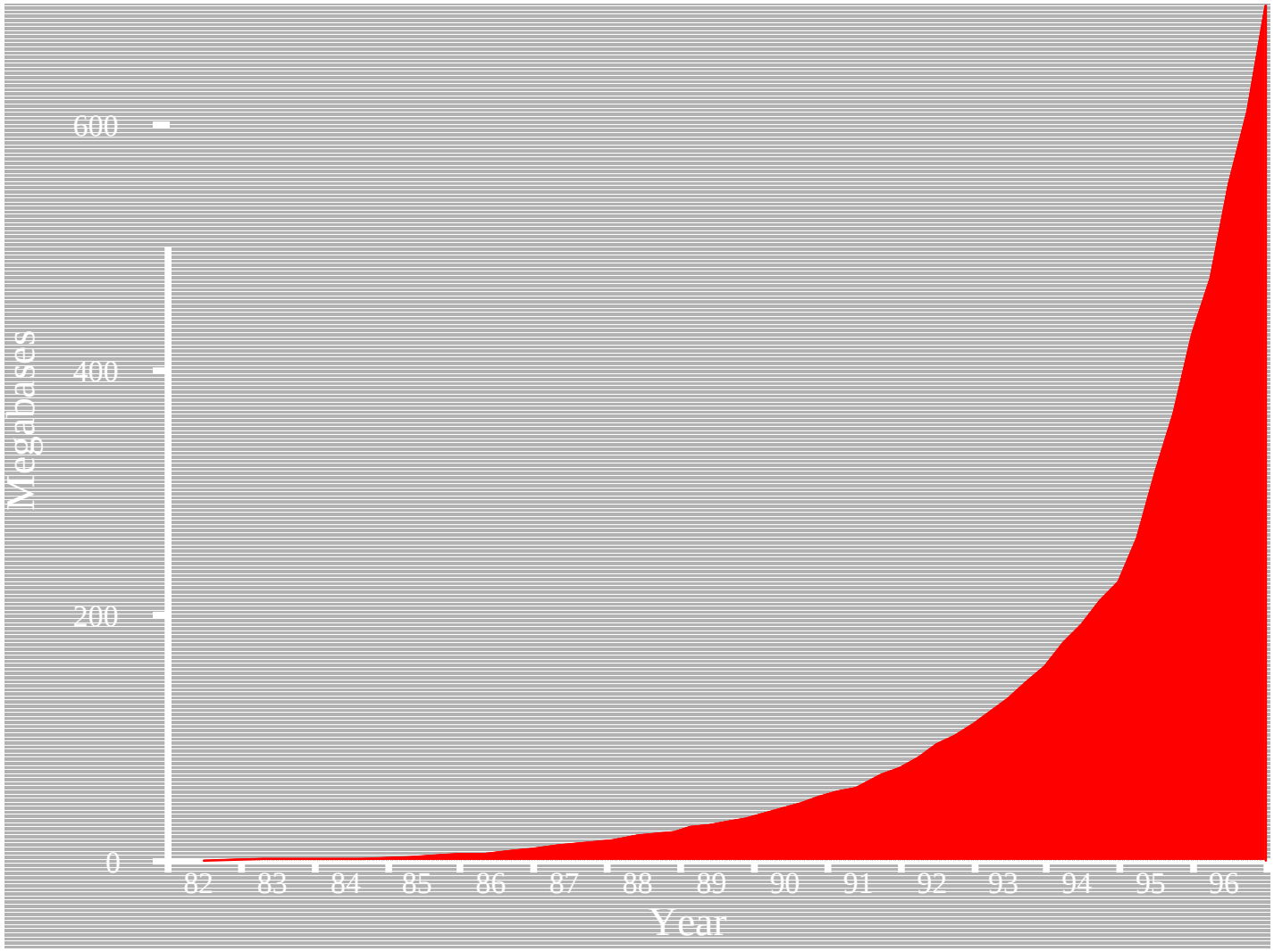
- No changes for academic users
- Almost no restrictions on the redistribution of SWISS-PROT by academic servers or software companies
- Commercial users are required to pay yearly subscription fees. These fees will be used to complement the existing grants in order to provide stable long-term funding



SWISS-PROT Growth

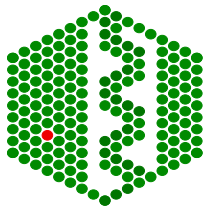


DNA sequence database growth



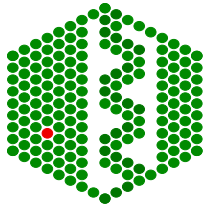
EMBL
European Bioinformatics Institute

The Bottleneck: Manual annotation

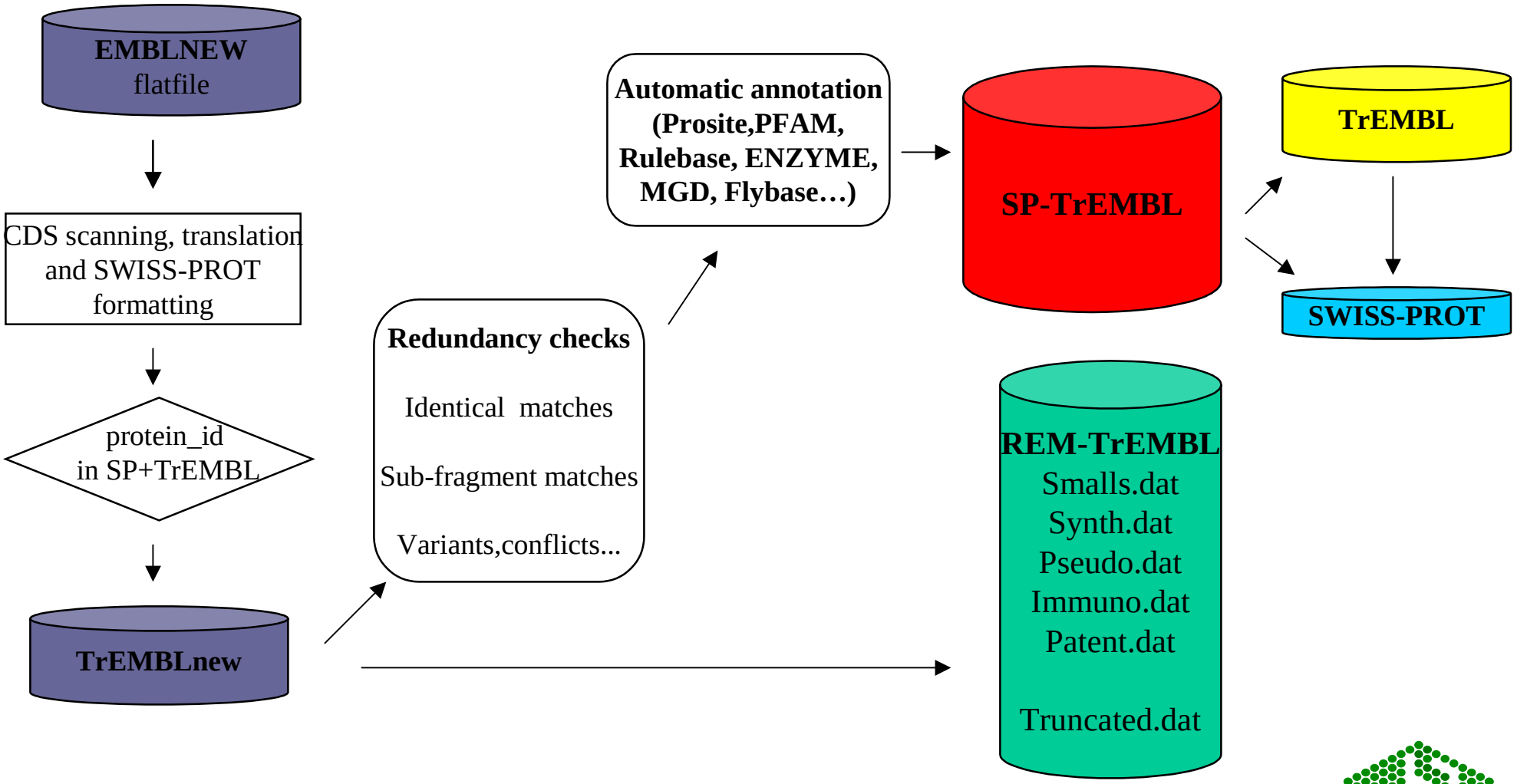


TrEMBL

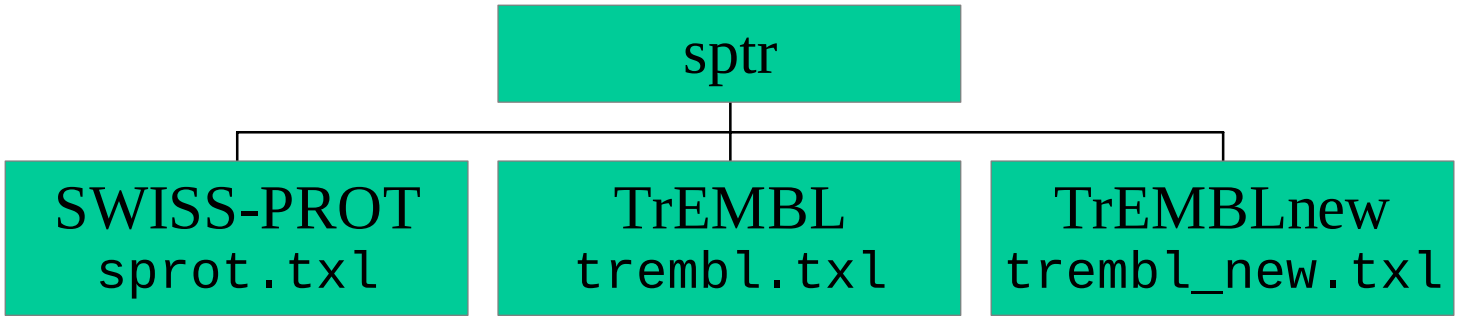
- We cannot cope with the speed with which new data is coming out
- We do not want to dilute the quality of SWISS-PROT
- Solution: TrEMBL (TRanslation of EMBL): contains all translations of CDS in the Nucleotide Sequence Database not in SWISS-PROT
- TrEMBL is automatically generated and annotated using software tools



TrEMBL production



SWISS-PROT + TrEMBL



- 94 000 SWISS-PROT entries
- 425 000 TrEMBL entries
- weekly production of a non-redundant and comprehensive protein sequence database consisting of SWISS-PROT, TrEMBL, and TrEMBLnew:

ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/