Bioinformatics databases

杨建益

Email: yangjy@nankai.edu.cn

Webpage: http://yanglab.nankai.edu.cn/

Course: http://yanglab.nankai.edu.cn/teaching/bioinformatics/

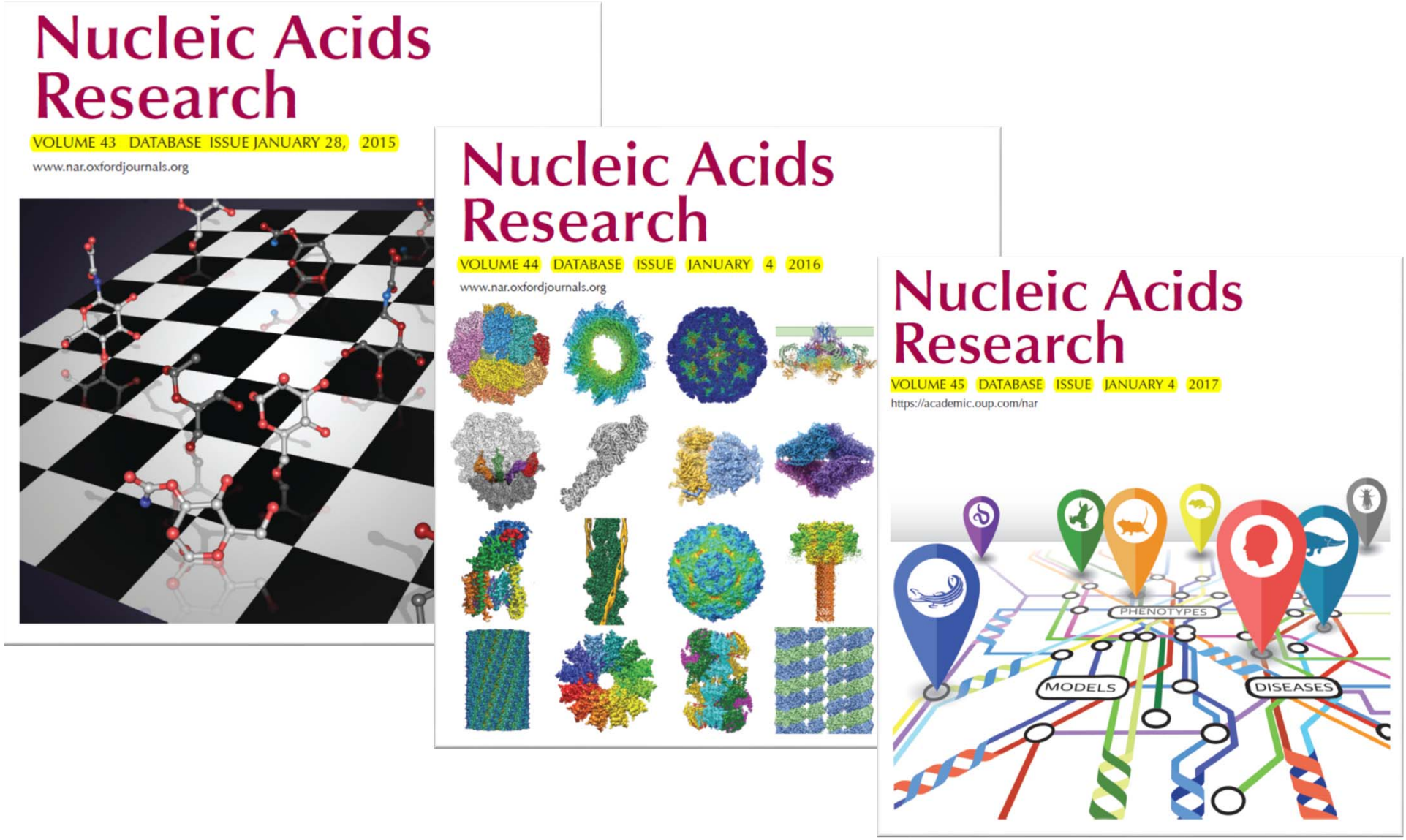
Office: 数学科学学院，419室

Note: the best way to learn for this chapter is to browse the listed websites on you laptop, rather than reading the slides carefully. The data listed on my slides are also outdated.

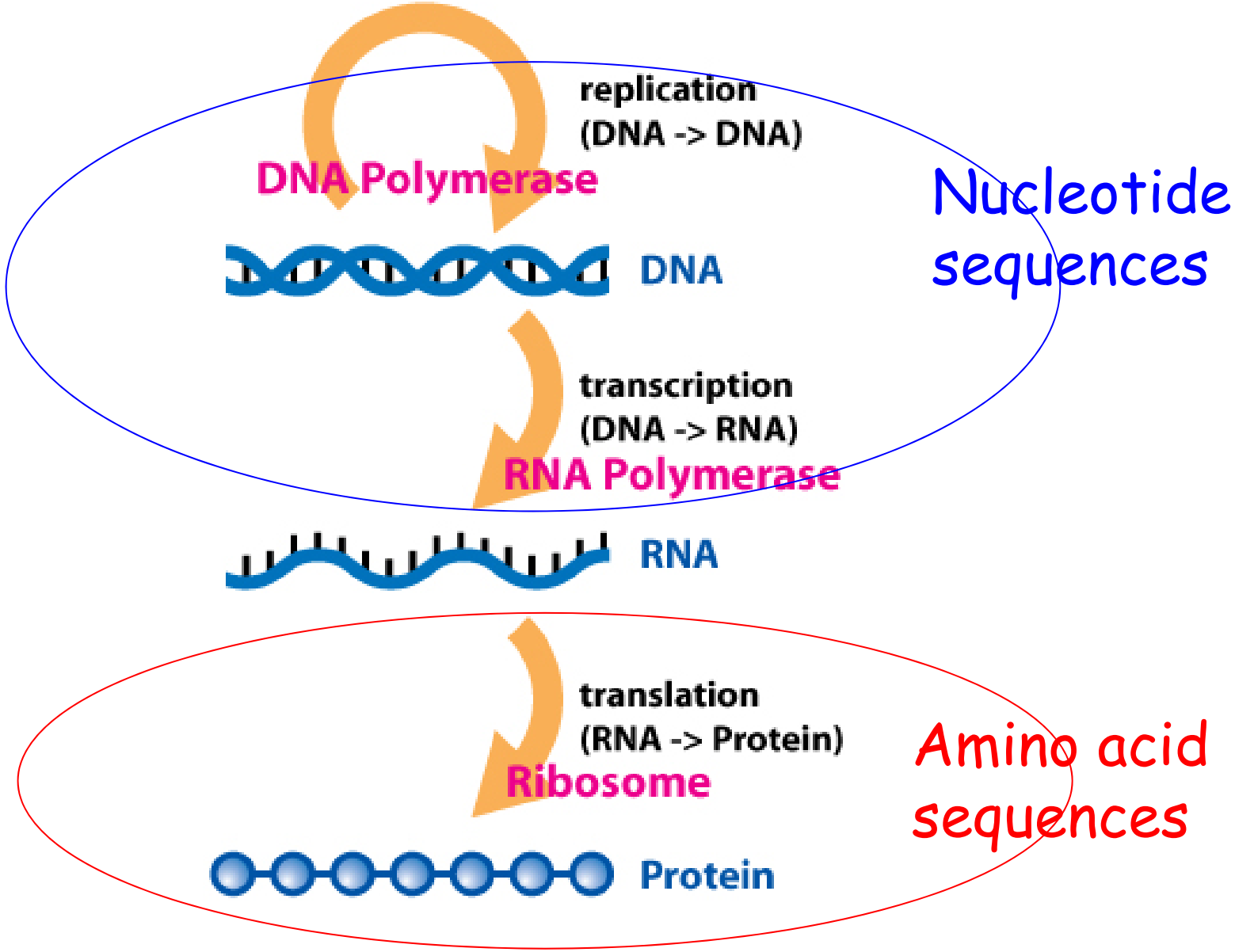
Content

1. Introduction
2. Nucleotide sequence databases
3. Amino acid sequence databases
4. Protein structure databases

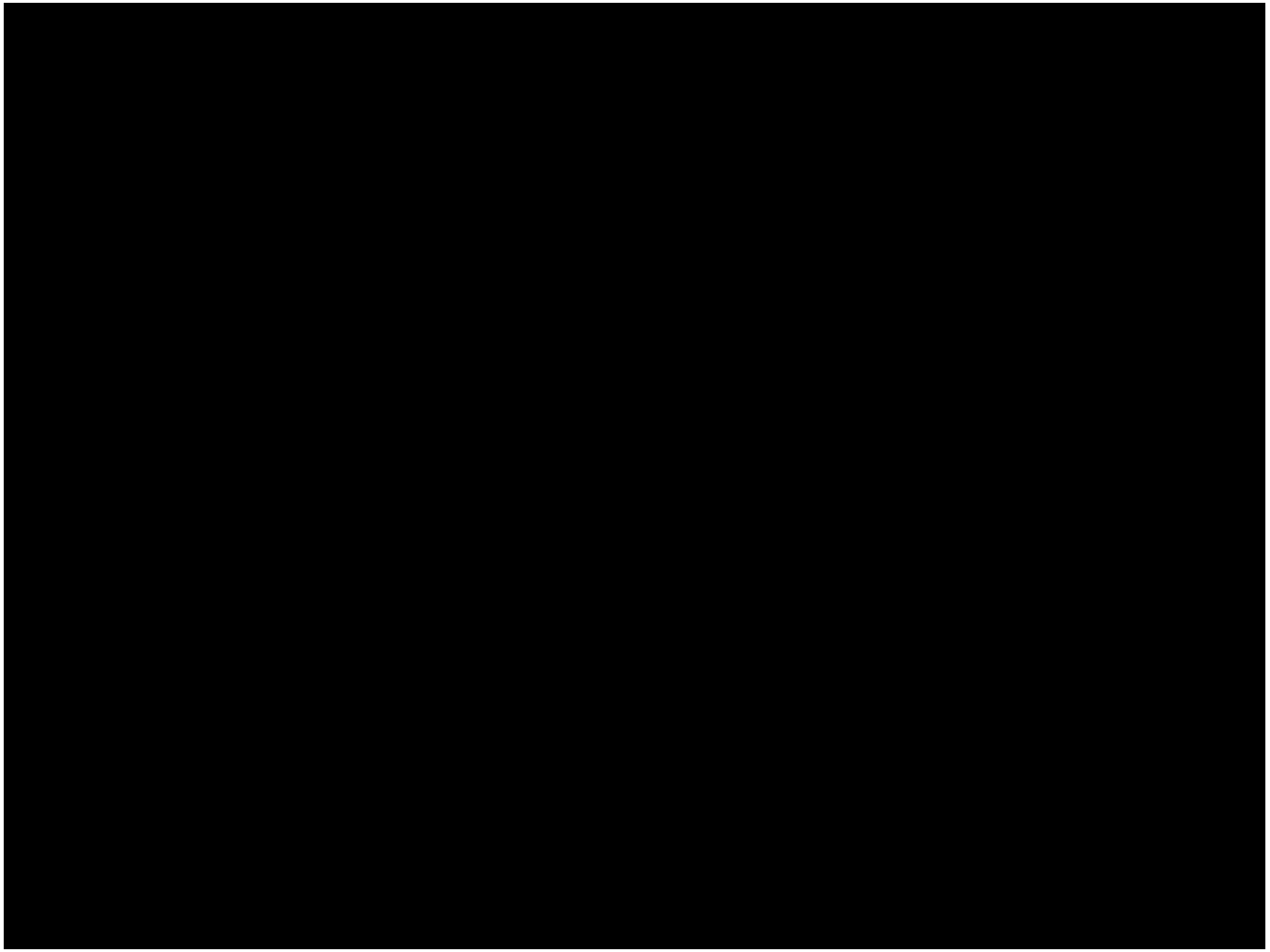
Nucleic Acid Research publishes one database issue every year



Central Dogma Of Molecular Biology



Central Dogma Of Molecular Biology



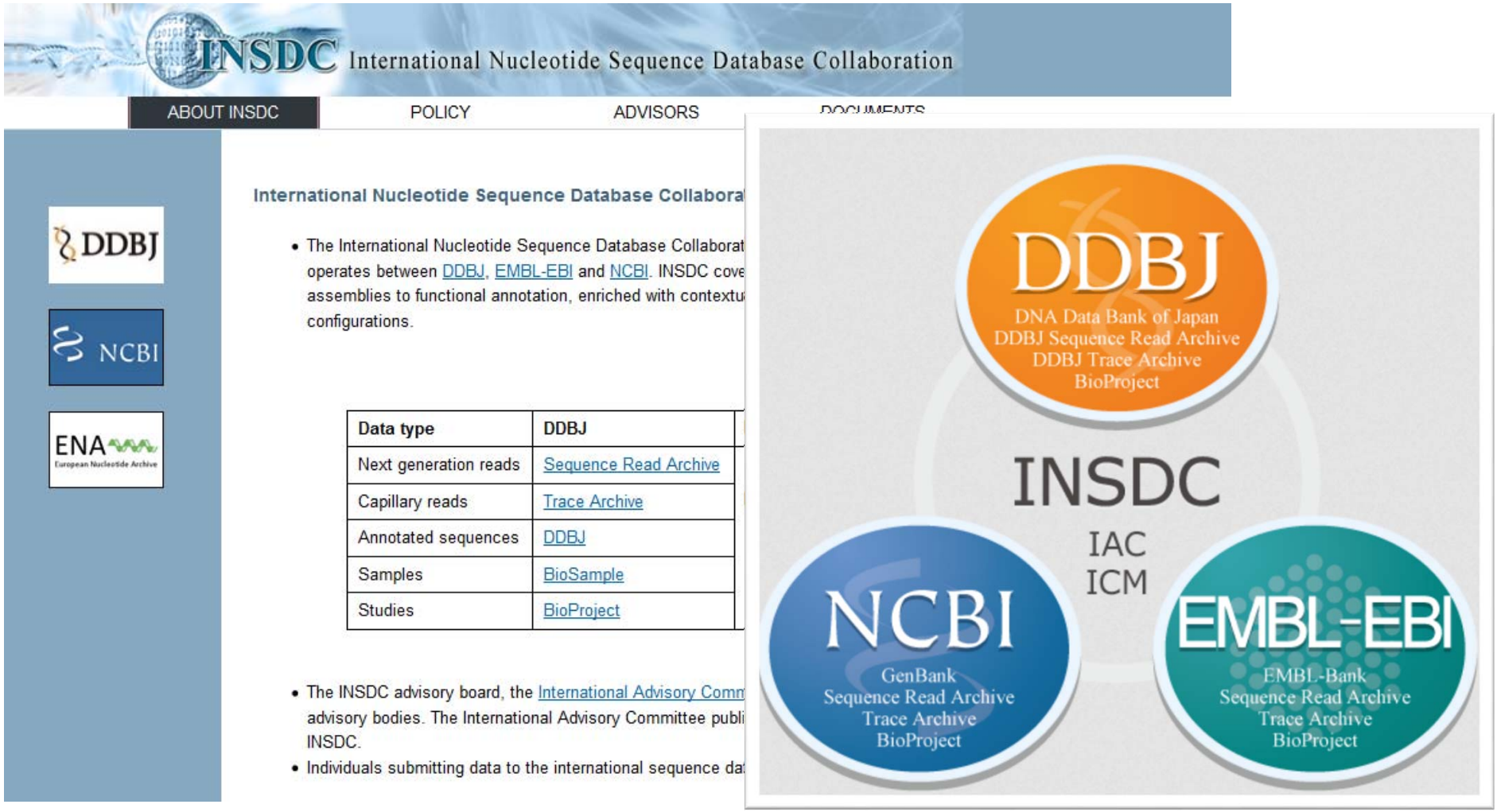
Content

1. Introduction
2. Nucleotide sequence databases
3. Amino acid sequence databases
4. Protein structure databases

Nucleotide sequence databases

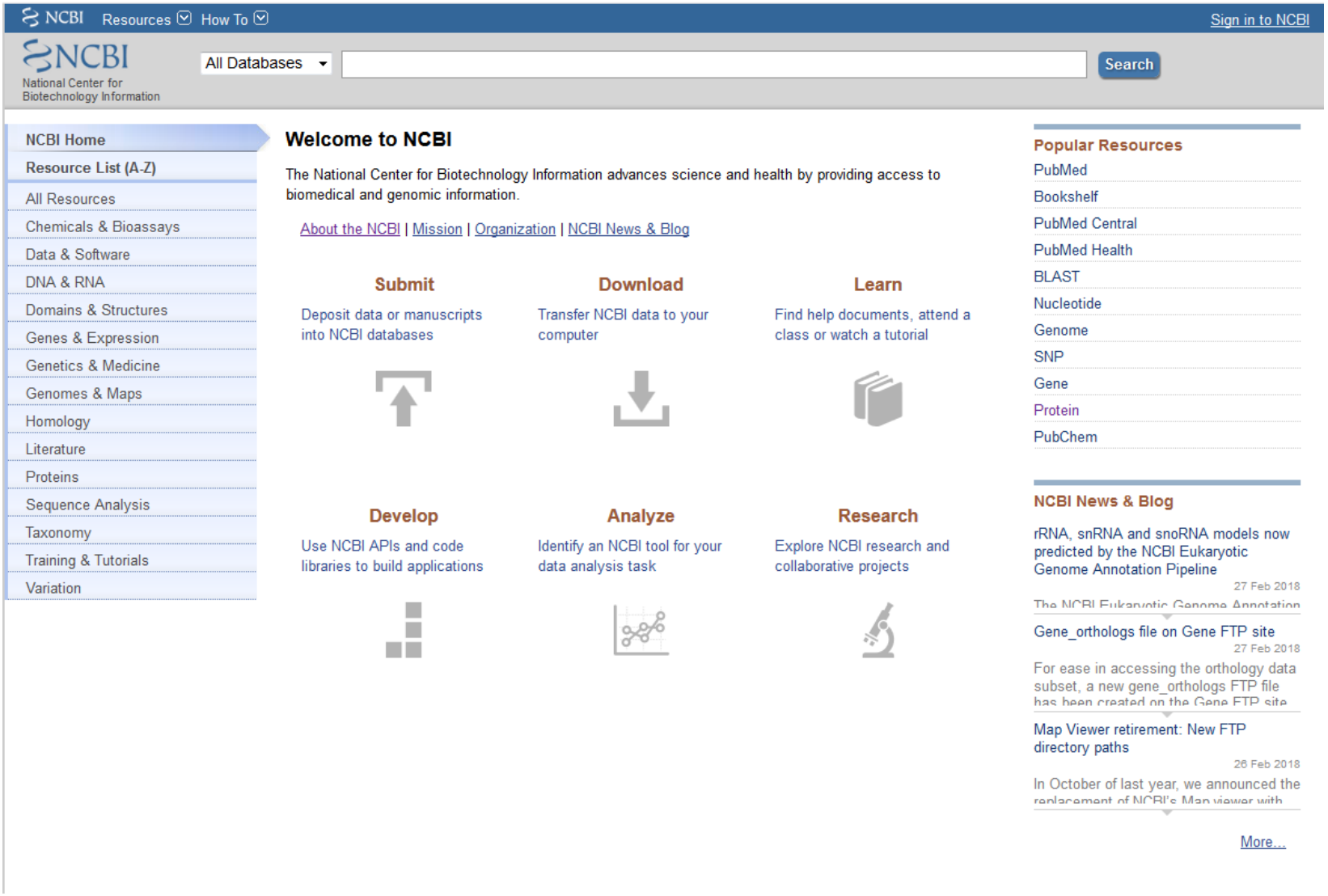
INSDC: International Nucleotide Sequence Database Collaboration

http://www.insdc.org/



Nucleotide sequence databases

NCBI: https://www.ncbi.nlm.nih.gov/

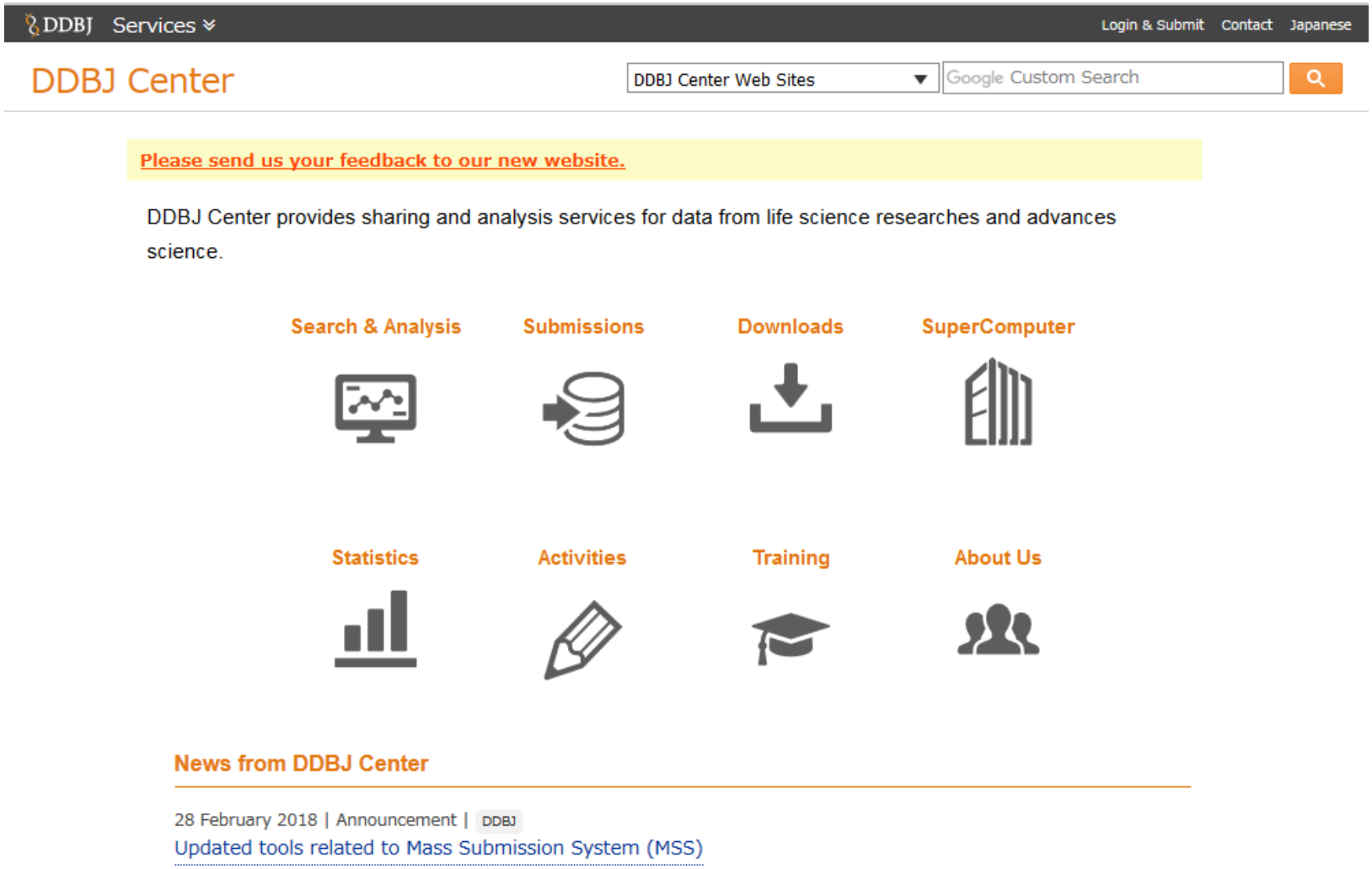


Nucleotide sequence databases

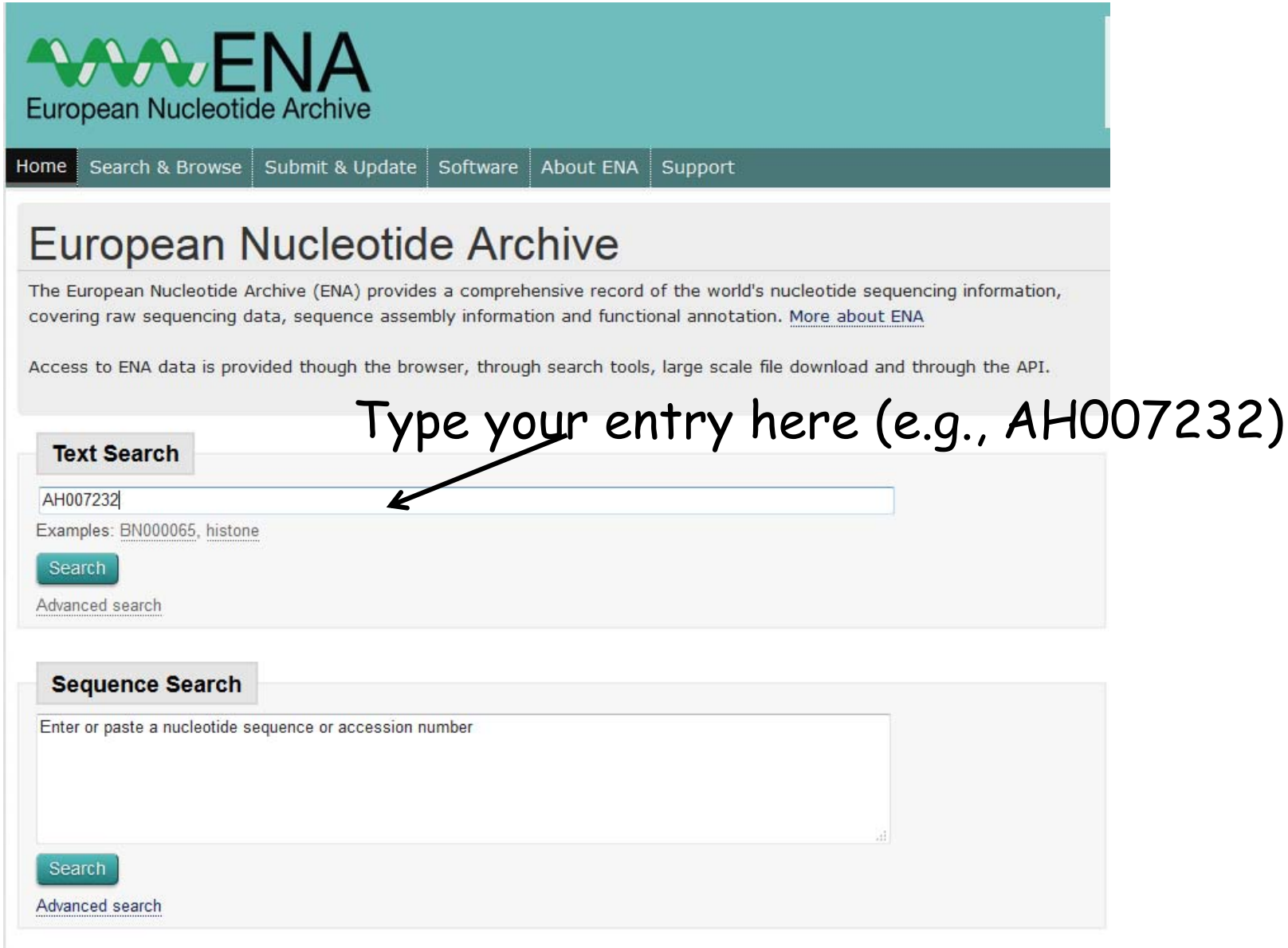
EMBL: https://www.embl.org/

Nucleotide sequence databases

DDBJ: https://www.ddbj.nig.ac.jp/index-e.html

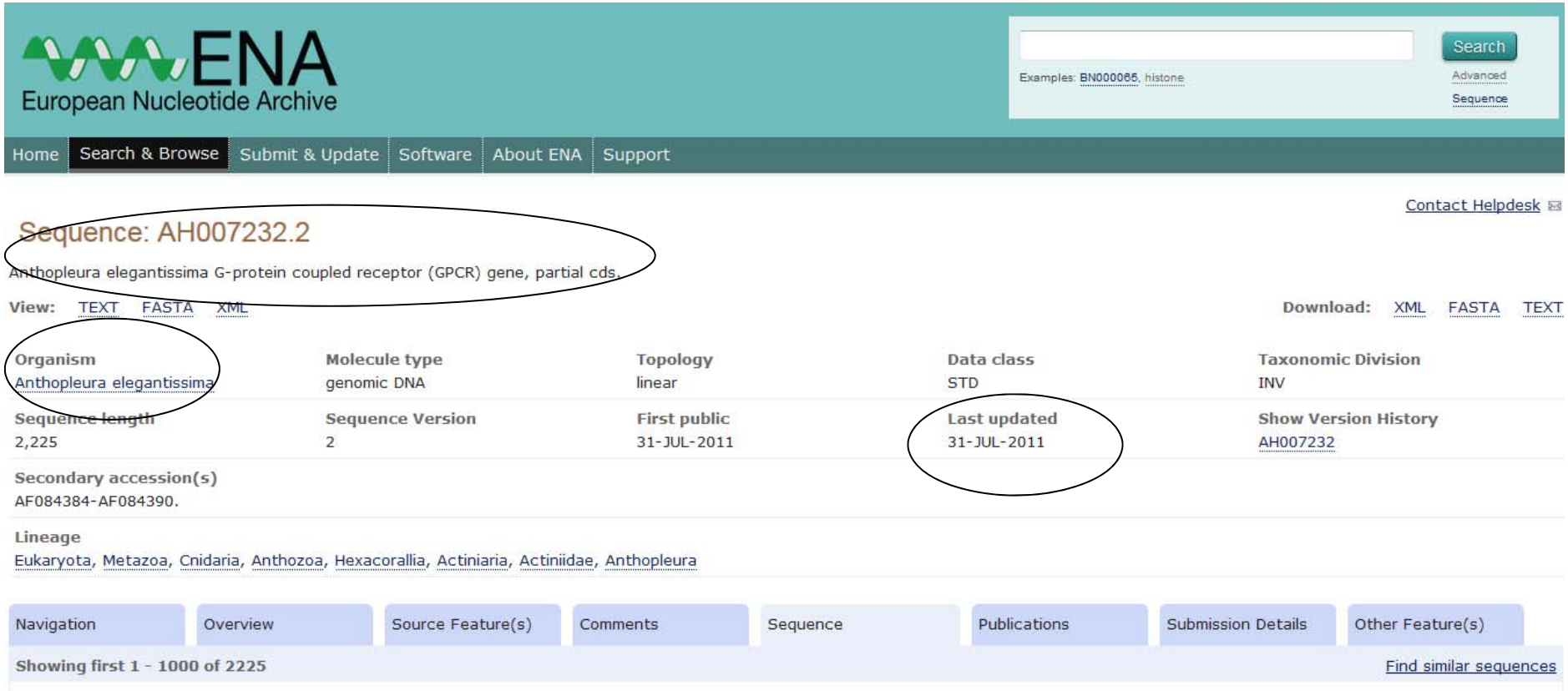


EMBL-EBI, ENA: https://www.ebi.ac.uk/ena



Type your entry here (e.g., AH007232)

Gene ID search result



Gene ID search result

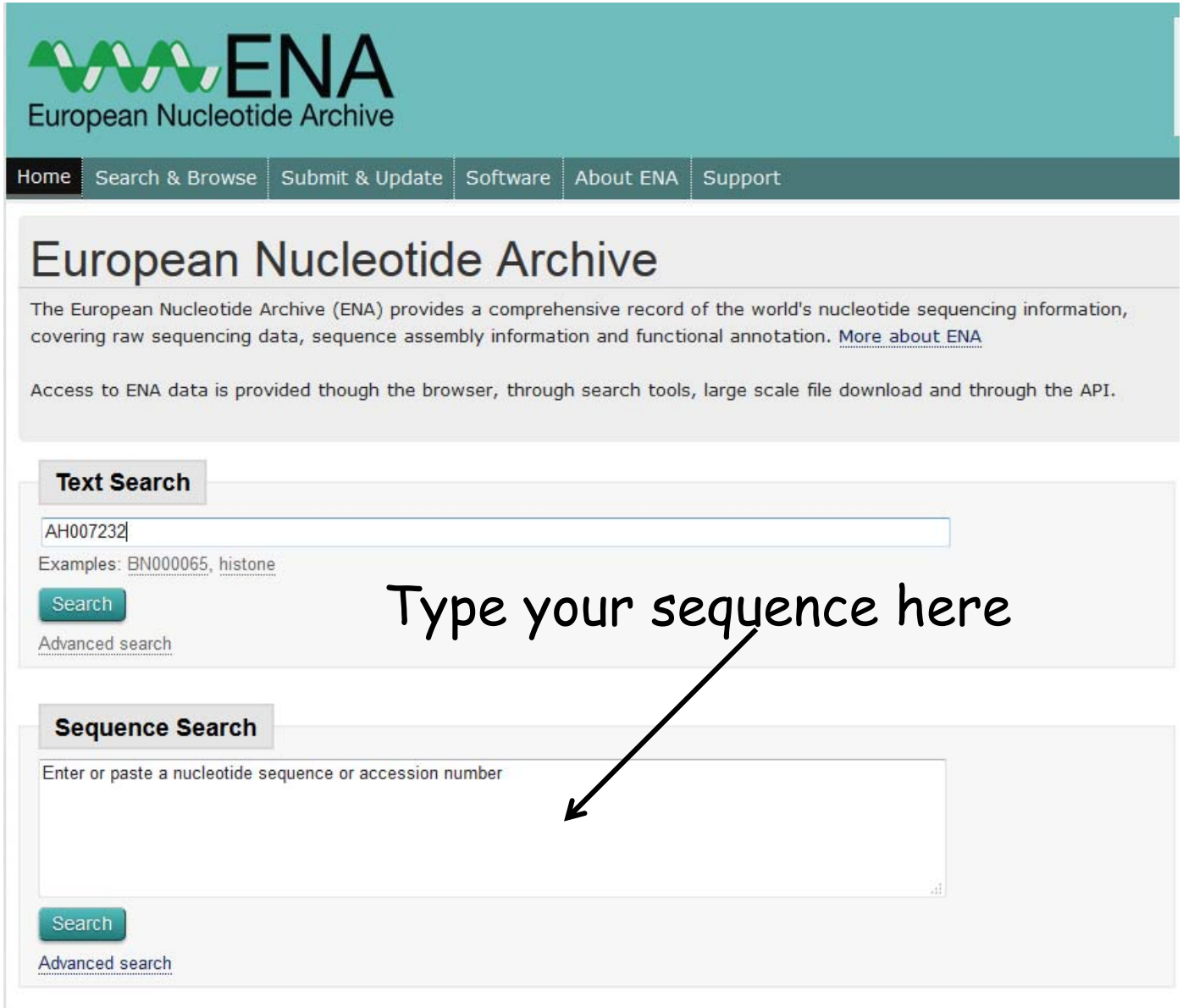
Navigation | Overview | Source Feature(s) | Comments | Sequence | Publications | Submission Details | Other Feature(s)

Showing first 1 - 1000 of 2225 Find similar sequences

```
>ENA|AH007232|AH007232.2 Anthopleura elegantissima G-protein coupled receptor (GPCR) gene, partial cds. :
Location:1..1000
AAACACTACGACATTGTAAGTATTCCTCCACGCTTGGATATGACTGAGTTTAGTACGAGT
AGTAAAACCGAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TCAGGCATGTGTCCTTTTCAGGAGAGATATCCATCCCCTCTGACATACTACAACACTTGG
AGAAATTGAAATATTTGTAAGTATTTGGCGCGGGAAATGTTTGTGTGTGTTAATCTCTAA
CCAAACGTTTCTCTTATATATCGACTTTCCTTCTTTTCAAGAACGCTGAATAACAATAAA
ATCAAGAACATCGCTAAATTTAGAGTCAAAAATGGATATTCCAGTTTGATGTAAGTATGT
TTTAAAAATACATTTTCTGAATTCAAAAGGGGCGCTAGGATACCCATACTCAGAGACTAG
AGACAAGACACTAAGACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCAC
TCTGACTCACCACCTCACTAACTATCTCACCAAATCCTTTTTTCCTATTTTTGCCTTTAG
AACACTAAGTTATACTCACAACATTATAGAGACCATTGAAAATGGTGCCTTCGATGACCT
CCAACAGCTTACTCAACTGTAAGTCCATGCCATTTATATGTACGATATATCTCTGTATTT
ATTACTTATATGCACTTGTGAAAGTTAATTCTTAAATTTCACTCACTCAAAGANNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGATCATGACCCTTATGCATGCTATATA
TATATATTGACTTTCTCTAATGCAATCAATAGACGTGTTT
...
```

DNA sequence

Sequence search



Number of sequences in ENA

ENA release 134

5 Jan 2018 - 16:33

Release 134 of assembled/annotated sequences from ENA is now available on the EBI public ftp server at ftp://ftp.ebi.ac.uk/pub/databases/ena/sequence/release/. It contains 1,157,925,701 sequence entries comprising 2,700,988,919,811 nucleotides. You can see the full release notes at http://bit.ly/LKFtrE.

ENA captures, preserves and presents the world's nucleotide sequence data. New content is included in ENA on a continuous basis and are distributed daily from our browser and RESTful service. The ENA assembled/annotated sequence release provides a quarterly snapshot of content in this important subset of ENA content.

Read more news from this service ›

The search takes 1-2 minutes!

Sequence search output

Results for job ncbiblast-R20180228-083801-0760-51200808-p1m

Summary Table | Tool Output | Visual Output | Result Summary | Submission Details

We want to improve your experience of this tool. Please fill in a short survey on SurveyMonkey to tell us about your experience of this tool. It will take you 2 minutes. Take this survey ›

Selection:

Select All | Invert | Clear

Apply to selection:

Annotations:

Show | Hide

Alignments:

Show | Hide

Entries:

Download in

fasta

format

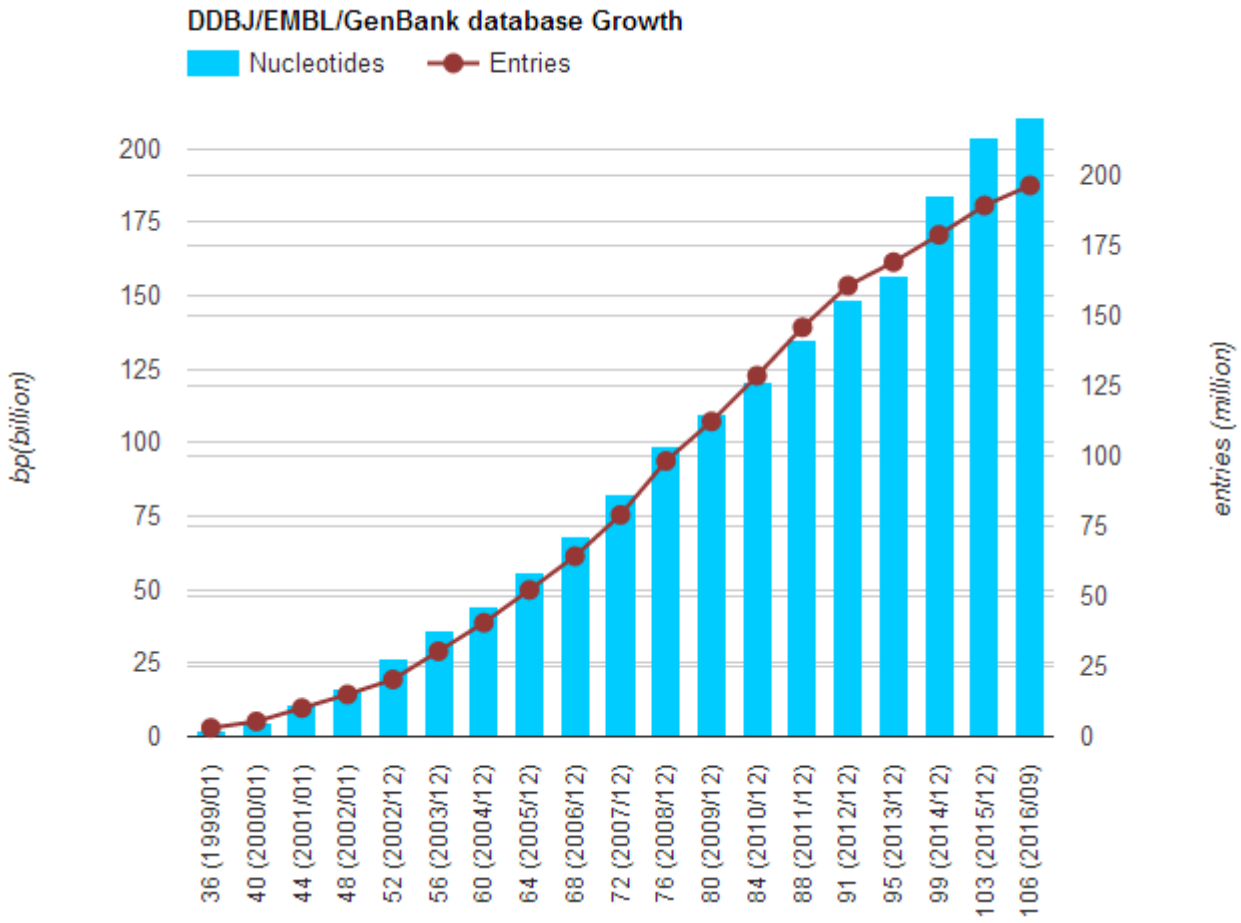
Tools:

Launch

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
1	EM_INV:AH007232	Anthopleura elegantissima G-protein coupled receptor (GPCR) gene, partial cds. *Cross-references and related information in:* Nucleotide sequences, Literature, Protein families, Samples & ontologies, Protein sequences	2225	752.9	100.0	100.0	0.0
2	EM_INV:AY531363	Anthopleura elegantissima isolate eCM_23a G-protein coupled receptor gene, intron 7 and partial cds. *Cross-references and related information in:* Nucleotide sequences, Literature, Samples & ontologies, Protein sequences	242	368.8	98.3	98.3	5.7E-97
3	EM_INV:AY531377	Anthopleura xanthogrammica isolate xHP_11a G-protein coupled receptor gene, intron 7 and partial cds. *Cross-references and related information in:* Nucleotide sequences, Literature, Samples & ontologies, Protein sequences	242	365.7	97.9	97.9	5.1E-96
4	EM_INV:AY531366	Anthopleura sola isolate sNFS_3a G-protein coupled receptor gene, intron 7 and partial cds. *Cross-references and related information in:* Nucleotide sequences, Literature, Samples & ontologies, Protein sequences	242	365.7	97.9	97.9	5.1E-96
5	EM_INV:AY531359	Anthopleura elegantissima isolate eAB_19a G-protein coupled receptor gene, intron 7 and partial cds. *Cross-references and related information in:* Nucleotide sequences, Literature, Samples & ontologies, Protein sequences	242	365.7	97.9	97.9	5.1E-96

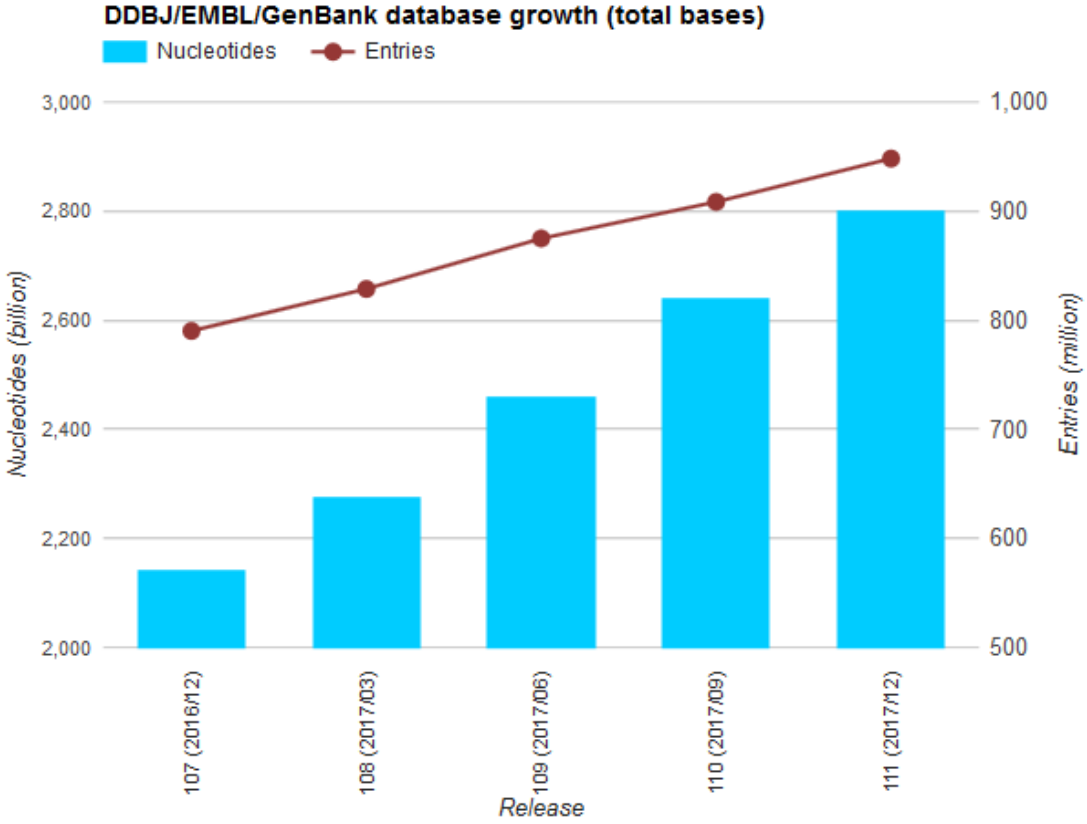
1. How was the search conducted?
2. What does E-Value mean?

Summary of nucleotide sequence databases



http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-old-e.html

Summary of nucleotide sequence databases



Release	Date	Entries	Nucleotides	Comments
107	2016/12	790,211,658	2,144,818,812,438	bulk sequence data inclusion started
108	2017/03	828,693,902	2,277,580,885,713	TLS data type inclusion started
109	2017/06	874,923,909	2,461,362,329,556	
110	2017/09	908,459,458	2,640,554,737,369	
111	2017/12	948,165,315	2,802,943,314,196	

Note: CON and TPA divisions are not included in the release statistics.

https://www.ddbj.nig.ac.jp/stats/dbgrowth-e.html

Content

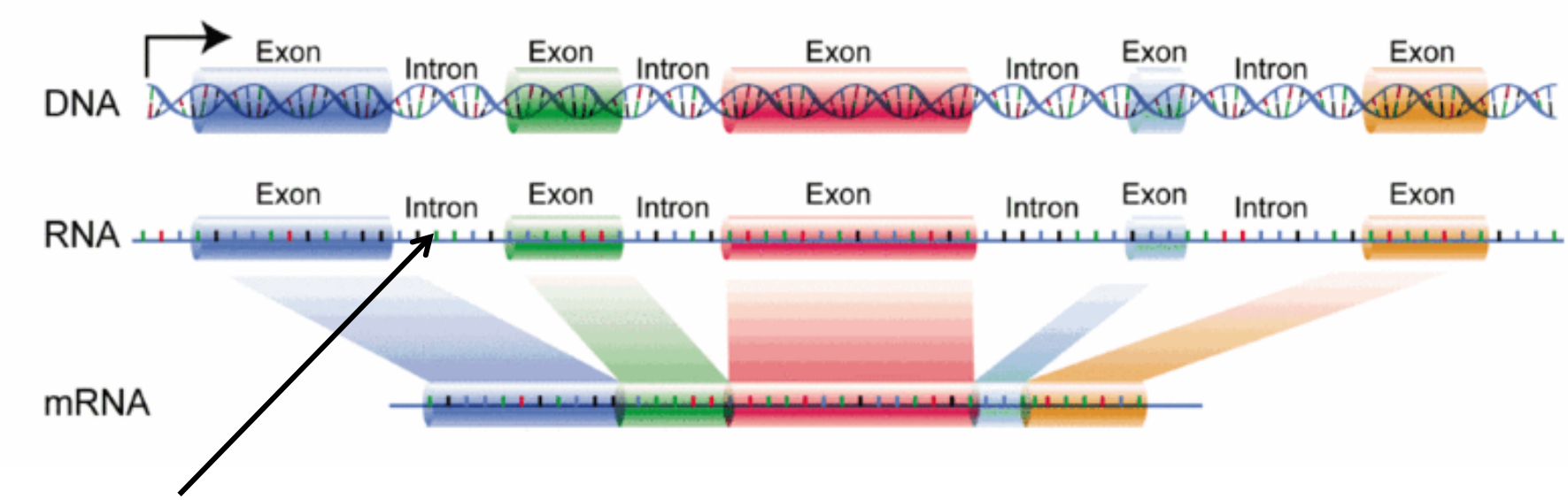
1. Introduction
2. Nucleotide sequence databases
3. Amino acid sequence databases
4. Protein structure databases

From DNA sequences to protein sequences

How exactly do 4-base sequence form 20-amino-acid sequences?

FIRST POSITION	SECOND POSITION: T	C	A	G	THIRD POSITION
T	TTT (F)	TCT (S)	TAT (Y)	TGT (C)	T
	TTC (F)	TCC (S)	TAC (Y)	TGC (C)	C
	TTA (L)	TCA (S)	**TAA STOP**	**TGA STOP**	A
	TTG (L)	TCG (S)	**TAG STOP**	TGG (W)	G
C	CTT (L)	CCT (P)	CAT (H)	CGT (R)	T
	CTC (L)	CCC (P)	CAC (H)	CGC (R)	C
	CTA (L)	CCA (P)	CAA (Q)	CGA (R)	A
	CTG (L)	CCG (P)	CAG (Q)	CGG (R)	G
A	ATT (I)	ACT (T)	AAT (N)	AGT (S)	T
	ATC (I)	ACC (T)	AAC (N)	AGC (S)	C
	ATA (I)	ACA (T)	AAA (K)	AGA (R)	A
	ATG (M)	ACG (T)	AAG (K)	AGG (R)	G
G	GTT (V)	GCT (A)	GAT (D)	GGT (G)	T
	GTC (V)	GCC (A)	GAC (D)	GGC (G)	C
	GTA (V)	GCA (A)	GAA (E)	GGA (G)	A
	GTG (V)	GCG (A)	GAG (E)	GGG (G)	G

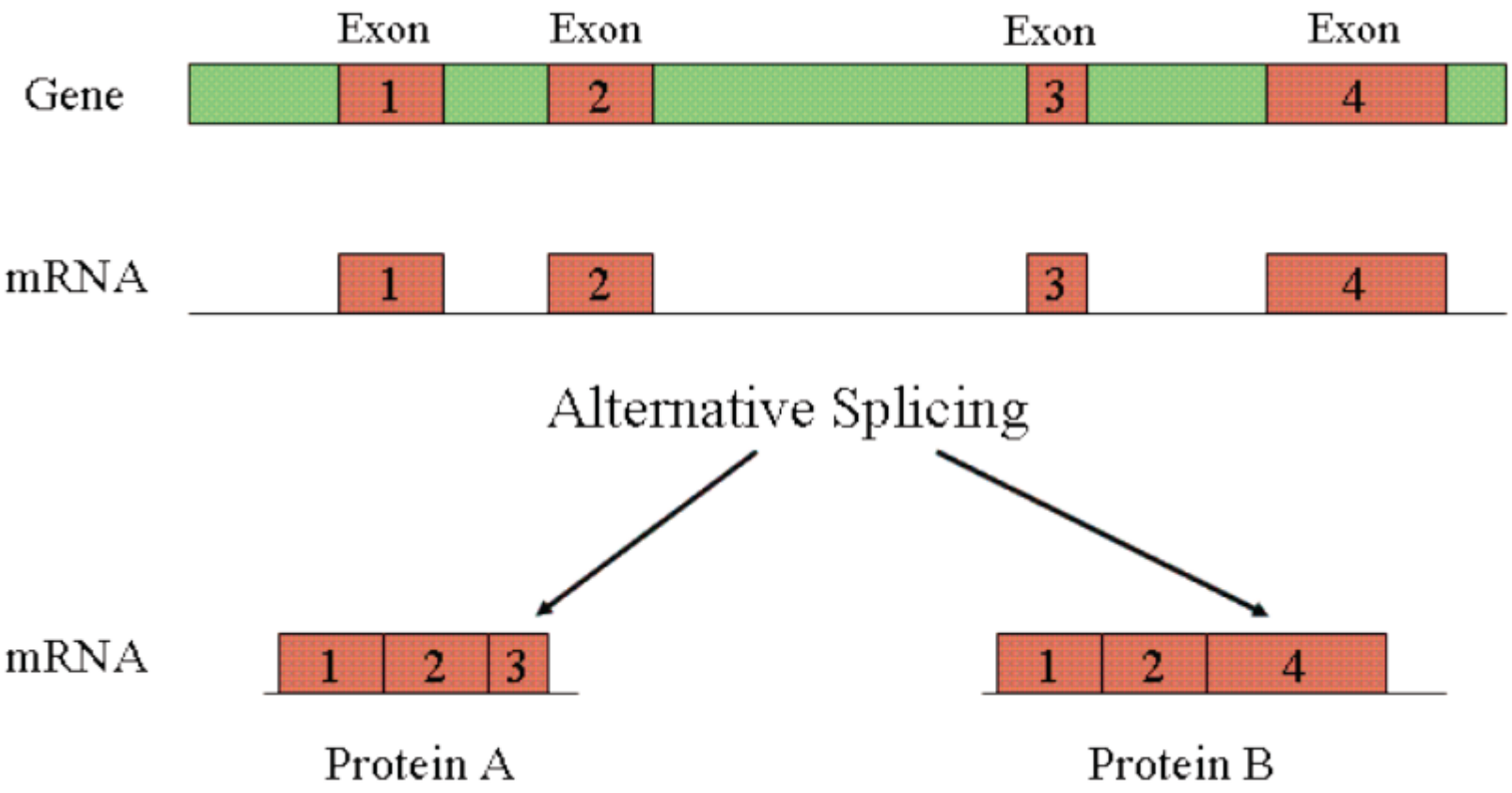
Protein cannot be simply translated from DNA sequence



Introns are not translated into mRNA

- One has to know exon location to translate DNA to protein sequence

Alternative RNA Splicing



- Even if one knows exon locations, proteins can not be translated uniquely due to splicing

Protein sequence databases

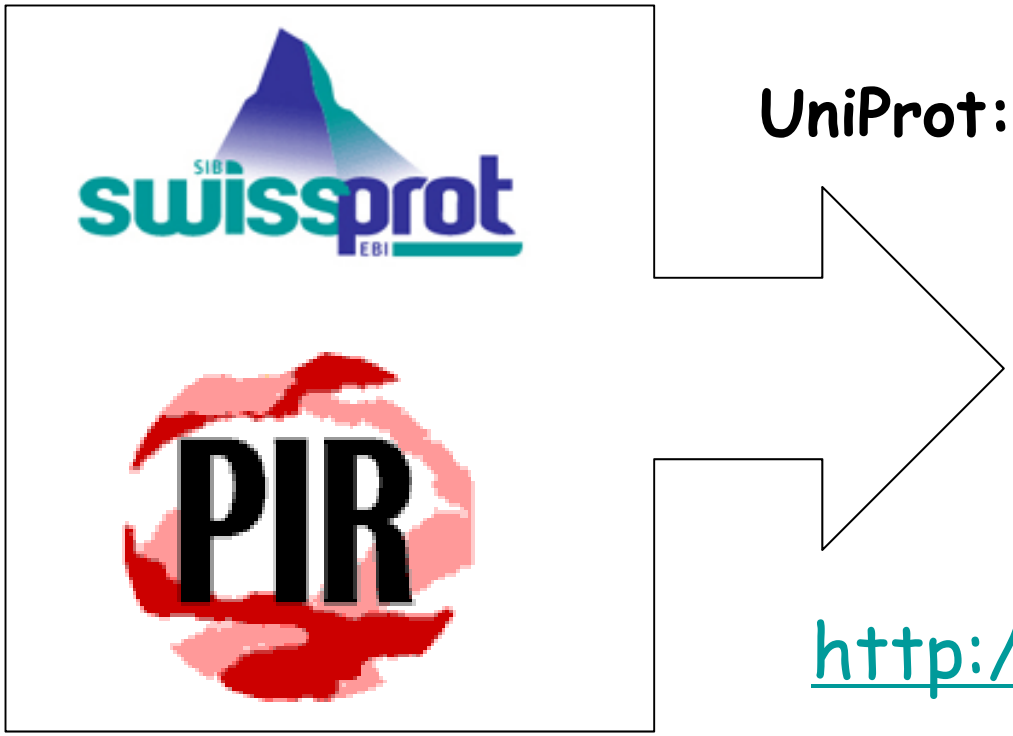
(http://www.expasy.ch/sprot): A curated protein sequence database organized by Swiss Institute of Bioinformatics and EBI.

- **Swiss-Prot** is a protein sequence database to provide a high level of manual annotations (such as function, domain structure etc), started by Amos Bairoch at **1986**.

- **TrEMBL** is a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries.

(http://pir.georgetown.edu): A curated protein sequence database organized by National Biomedical Research Foundation (NBRF) in the US. Started by Marget Dayhoff in **1965**.

Two protein sequence databases are merged in 2002

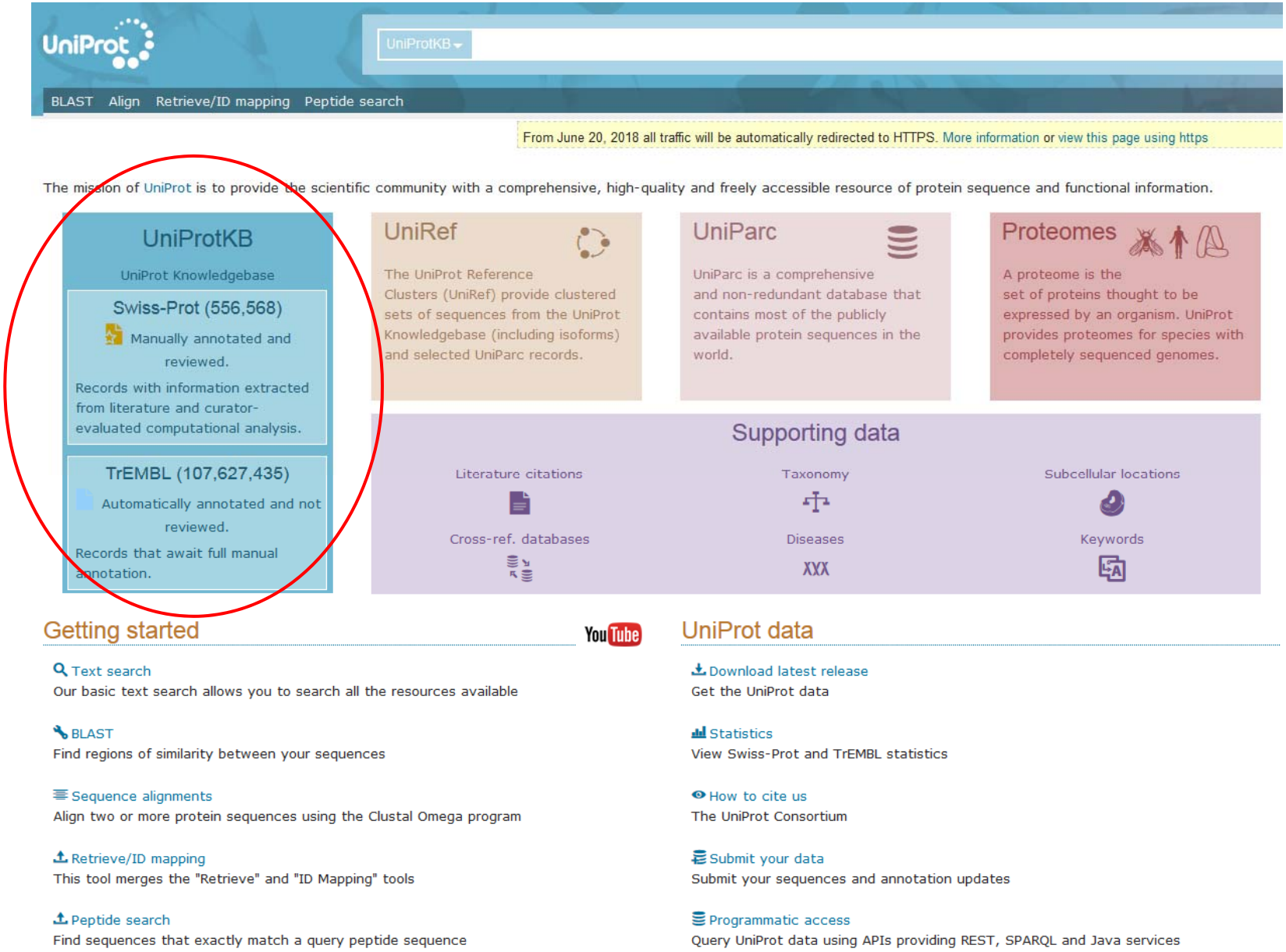


UniProt: universal protein resource

http://www.uniprot.org/

It was generally believed that PIR does not reach the level of completeness in the entry annotation as does SWISS-PROT. Although SWISS-PROT and PIR overlap extensively, there are still many sequences which can be found in only one of them. A new database of UniProt emerges in 2002 which combines SwissProt, TrEMBL, and PIR.

UniProt Homepage: http://www.uniprot.org/



Where do the UniProtKB protein sequences come from?

More than 95% of the protein sequences provided by UniProtKB come from the translations of coding sequences (CDS) submitted to the EMBL-Bank/GenBank/DDBJ nucleotide sequence resources (International Nucleotide Sequence Database Collaboration (INSDC)). These CDS are either generated by gene prediction programs or are experimentally proven. A protein identifier ("protein_id") is assigned to the translated CDS and can be found in the original EMBL-Bank/GenBank/DDBJ record and in the relevant UniProtKB entry.

The translated CDS sequences are automatically transferred to the TrEMBL section of UniProtKB. The TrEMBL records can be selected for further manual annotation and then integrated into the UniProtKB/Swiss-Prot section. The "protein_id" are listed in the cross-reference part of the 'Sequence' section, of the UniProtKB entries (see for example P13744 'Translation').

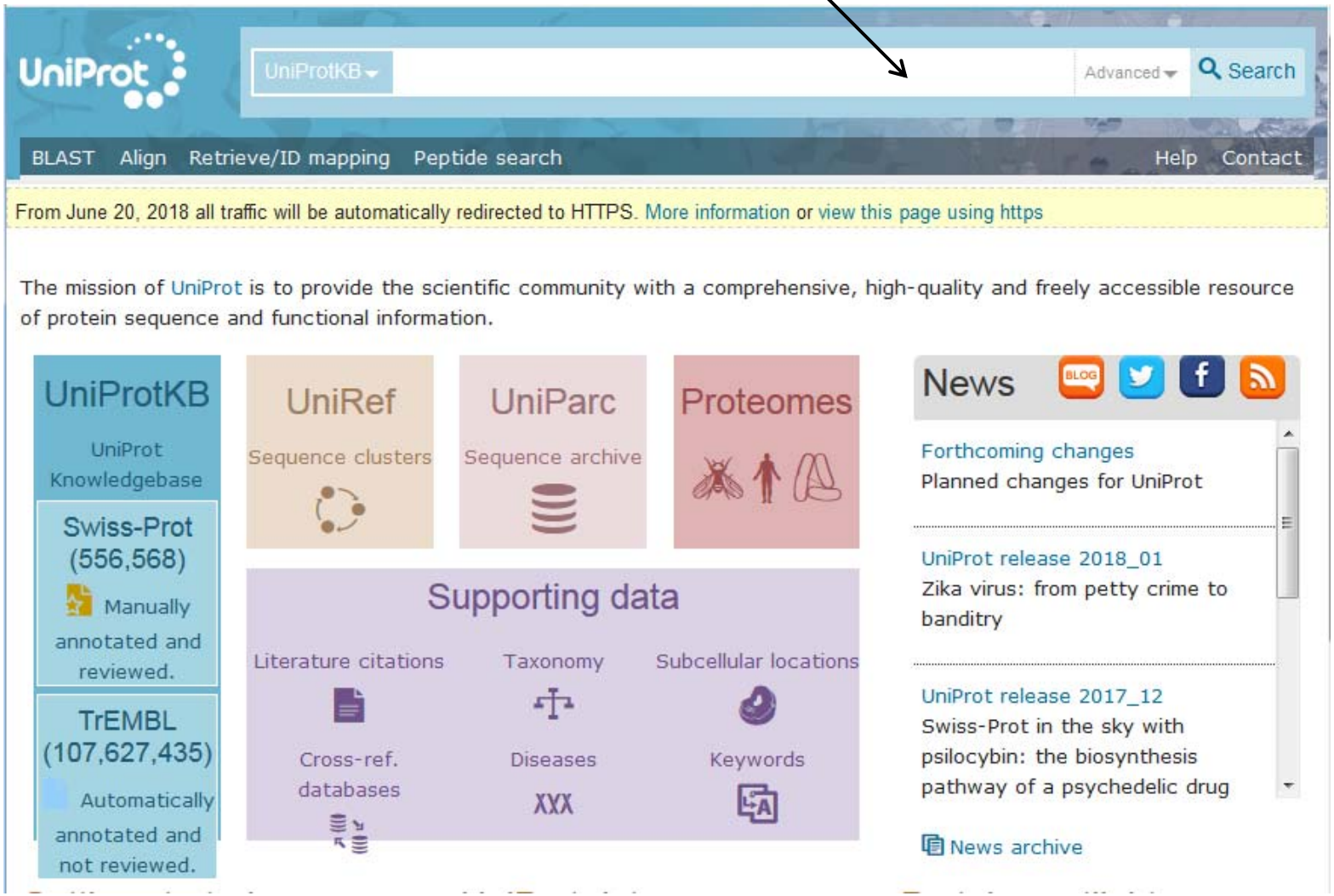
In addition to translated CDS, UniProtKB protein sequences may come from:

- the PDB database.
- sequences experimentally obtained by direct protein sequencing, by Edman degradation or MS/MS experiments and submitted to UniProtKB/Swiss-Prot. Only about 5% of the UniProtKB/Swiss-Prot entries contain sequence data obtained by direct protein sequencing (list of entries with the keyword `'Direct protein sequencing'`).
- sequences scanned from the literature (i.e. PRF or other journal scan project).
- sequences derived from gene prediction, not submitted to EMBL-Bank/GenBank/DDBJ (Ensembl, Ensembl Genomes, WormBase ParaSite or VectorBase) (1), RefSeq, CCDS, etc).
- sequences derived from in-house gene prediction, in very specific cases.

More than 95% of the protein sequences provided by UniProtKB come from the translations of CDS submitted to the EMBL-Bank/GenBank/DDBJ nucleotide sequence resources. These CDS are either generated by gene prediction programs or are experimentally proven.

Search database by UniProt ID

Type your ID here (e.g., Q754G5)



Search result

UniProt

UniProtKB | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

From June 20, 2018 all traffic will be automatically redirected to HTTPS. More information or view this page using https

UniProtKB - Q754G5 (ARP4_ASHGO)

Basket

Display

Entry
Publications
Feature viewer
Feature table

All None

- Function
- Names & Taxonomy
- Subcell. location
- Pathol./Biotech
- PTM / Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar proteins
- Cross-references

BLAST | Align | Format | Add to basket | History | Feedback | Help video | Other tutorials and videos

Protein: **Actin-related protein 4**

Gene: **ARP4**

Organism: *Ashbya gossypii (strain ATCC 10895 / CBS 109.51 / FGSC 9923 / NRRL Y-1056) (Yeast) (Eremothecium gossypii)*

Status: Reviewed - Annotation score: ●●●○○ - Protein inferred from homology

Function description

Function

Chromatin interaction component of the NuA4 histone acetyltransferase complex which is involved in transcriptional activation of selected genes principally by acetylation of nucleosomal histone H4 and H2A. The NuA4 complex is also involved in DNA repair. Is required for NuA4 complex integrity. Component of the SWR1 complex which mediates the ATP-dependent exchange of histone H2A for the H2A variant HZT1 leading to transcriptional regulation of selected genes by chromatin remodeling. Component of the INO80 complex which remodels chromatin by shifting nucleosomes and is involved in DNA repair (By similarity). By similarity

GO - Molecular function

- ATP-dependent 3'-5' DNA helicase activity — Source: EnsemblFungi
- histone acetyltransferase activity — Source: EnsemblFungi
- nucleosomal histone binding — Source: EnsemblFungi

View the complete GO annotation on QuickGO ...

GO - Biological process

- chromatin remodeling — Source: EnsemblFungi
- DNA repair — Source: UniProtKB-KW

Search result

Amino acid sequence

Publications

Feature viewer

Feature table

All None

- [x] Function
- [x] Names & Taxonomy
- [x] Subcellular location
- [ ] Pathology & Biotech
- [x] PTM / Processing
- [ ] Expression
- [x] Interaction
- [x] Structure
- [x] Family & Domains
- [x] Sequence
- [x] Similar proteins
- [ ] Cross-references
- [x] Entry information
- [x] Miscellaneous

▲Top

Sequence

Sequence status: Complete.

Q754G5-1 [UniParc] FASTA Add to basket

« Hide

```
        10         20         30         40         50
MSNSALQVYG GDEITAVVID PGSFTTNIGY SGTDCPQAIL PSCYGKYTEG
        60         70         80         90        100
EKDELFSEQS IGLPRKDYEI HNIVQNGEVV DWEKAEKQWD WAIRSELRFE
       110        120        130        140        150
TNSGMPALLT EPIWNSEENR KKSLEVLLES MDFSACYLVP TATAVSFAMG
       160        170        180        190        200
RPTCLVVDIG HDVTSVCPVV DGMTLSKSSM RSYIAGSLLN ELIRSQLAPR
       210        220        230        240        250
KVIPLFQVAQ RRPVFMERKF DYEIHPSLQK FVNERQFFQE FKETMLQVAP
       260        270        280        290        300
TSISKFKSEI ETTSKRSIEA PWGEELVYDS LQRLEFAEQL FTPDLSQFPE
       310        320        330        340        350
DWPISKDGVV ETWHNDYVPL KRNKPGTNVK DKEGTLDATP VPDENSVTSA
       360        370        380        390        400
DQPNDNGKRN LEETTPDQKN EVSGLADLIY SSIMSTDVDL RTTLSHNVVI
       410        420        430        440        450
TGGTSSLPGL MDRISAELNR SLPALKFRML TSGQLRERQY QGWLGGSILA
       460        470
SLGTFHQLWV GKQEYAEVGA DRLLKDRFR
```

Search database by sequence

Type your protein sequence here

UniProt

UniProtKB | Advanced | Search

BLAST | Align | Retrieve/ID mapping | Peptide search | Help | Contact

From June 20, 2018 all traffic will be automatically redirected to HTTPS. More information or view this page using https

BLAST

How to use this tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

1. Enter either a protein or nucleotide sequence or a UniProt identifier (e.g.P00750 or A4_HUMAN or UPI0000000001) into the form field.
2. Optionally, change the program parameters with the dropdown menus under the form.
3. Click the *Run BLAST* button.

Help | BLAST help video | Other tutorials and videos | Downloads

```
>sp|Q754G5|ARP4_ASHGO Actin-related protein 4 OS=Ashbya gossypii (strain ATCC 10895 / CBS 109.51 / FGSC 9923 / NRRL Y-1056)
GN=ARP4 PE=3 SV=1
MSNSALQVYGGDEITAVVIDPGSFTTNIGYSGTDCPQAILPSVGKYTEGEKDELFSEQS
IGLPRKDYEIHNIVQNGEVVDWEKAEKQWDWAIRSELRFETNSGMPALLTEPIWNSEENR
KKSLEVLLESMDFSACYLVPTATAVSFAMGRPTCLVVDIGHDVTSVCPVVDGMTLSKSSM
RSYIAGSLLNELIRSQLAPRKVIPLFQVAQRRPVFMERKFDYEIHPSLQKFVNERQFFQE
FKETMLQVAPTSISKFKSEIETTSKRSIEAPWGEELVYDSLQRLEFAEQLFTPDLSQFPE
DWPISKDGVVETWHNDYVPLKRNKPGTNVKDKEGTLDATPVPDENSVTSADQPNDNGKRN
```

Target database	E-Threshold	Matrix	Filtering	Gapped	Hits
UniProtKB	10	Auto	None	yes	250

☐ Run BLAST in a separate window.

Clear | Run BLAST

Search result

Alignments

BLAST | Align | Download | Add to basket | Columns

1 to 25 of 250 ▶ Show 25

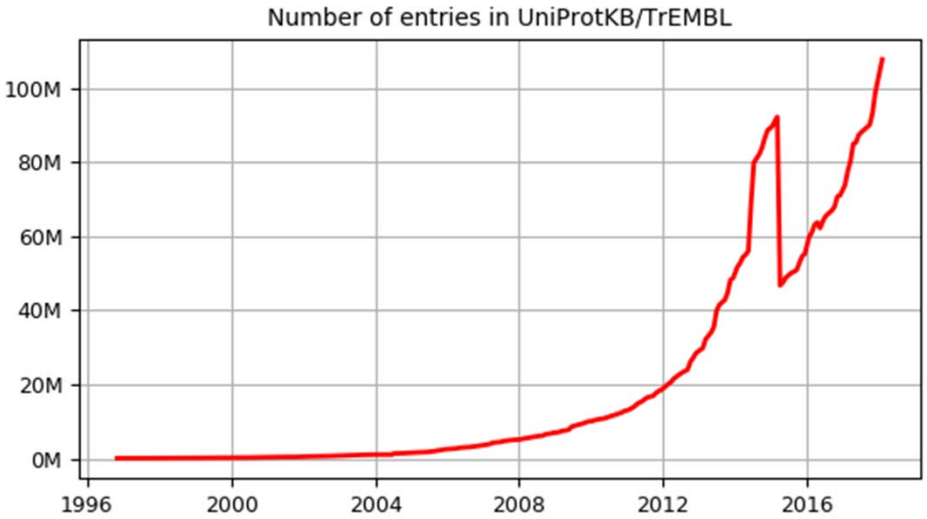
Entry	Alignment overview	Info	Status
Query: sp|Q754G5|ARP4_ASHGO B20180228F725F458AC8690F874DD868E4ED79B88EEE5968			
Q754G5	ARP4_ASHGO - Actin-related protein 4 - Ashbya gossypii ... - View alignment	E-value: 0.0 Score: 2,493 Ident.: 100.0%	
R9XI71	R9XI71_ASHAC - AaceriAFR105Cp Ashbya aceri (Yeast) - View alignment	E-value: 0.0 Score: 2,415 Ident.: 96.2%	
G8JUC6	G8JUC6_ERECY - Uncharacterized protein - Eremothecium cym... - View alignment	E-value: 0.0 Score: 2,081 Ident.: 80.8%	
A0A109UXL6	A0A109UXL6_9SACH - HBR427Wp - Eremothecium sin... - View alignment	E-value: 0.0 Score: 1,962 Ident.: 75.1%	
A0A1G4MKL3	A0A1G4MKL3_LACFM - LAFE_0H13146g1_1 - Lachancea fermen... - View alignment	E-value: 0.0 Score: 1,815 Ident.: 69.7%	
C5E324	C5E324_LACTC - KLTH0H09702p - Lachancea thermo... - View alignment	E-value: 0.0 Score: 1,680 Ident.: 64.5%	
A0A0P1KNH3	A0A0P1KNH3_9SACH - LAQU0S02e06766g1_1 - Lachancea quebec... - View alignment	E-value: 0.0 Score: 1,671 Ident.: 64.1%	

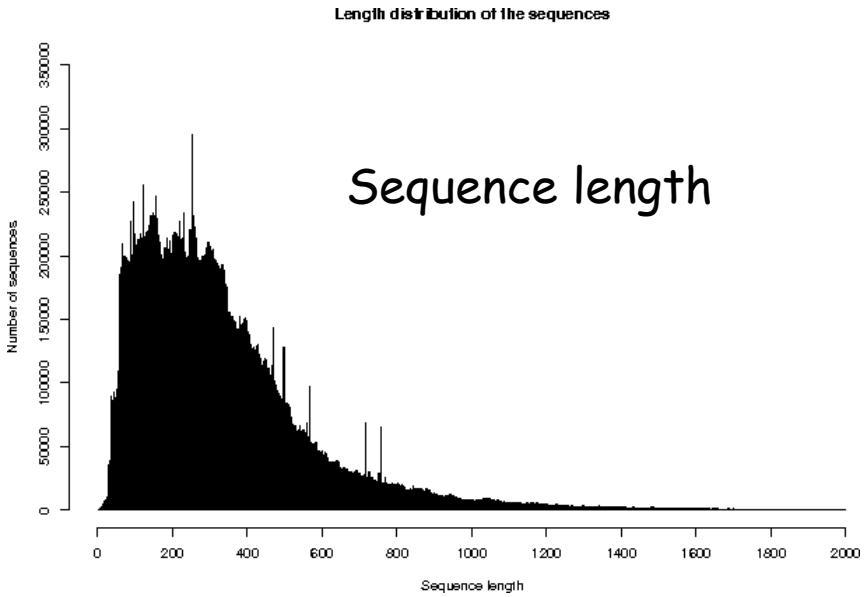
?

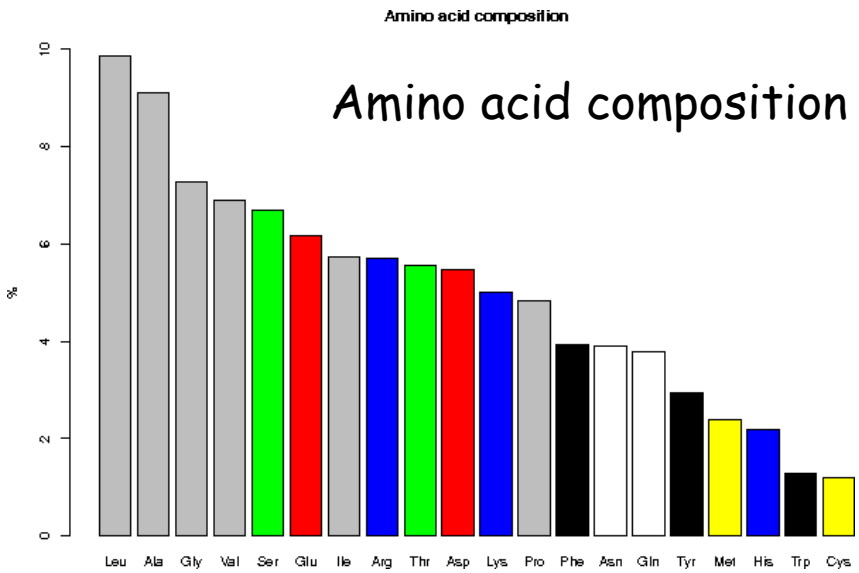
Statistics in UniProtKB/TrEMBL

https://www.ebi.ac.uk/uniprot/TrEMBLstats

107,627,435 sequences,
36,161,263,380 amino acids.

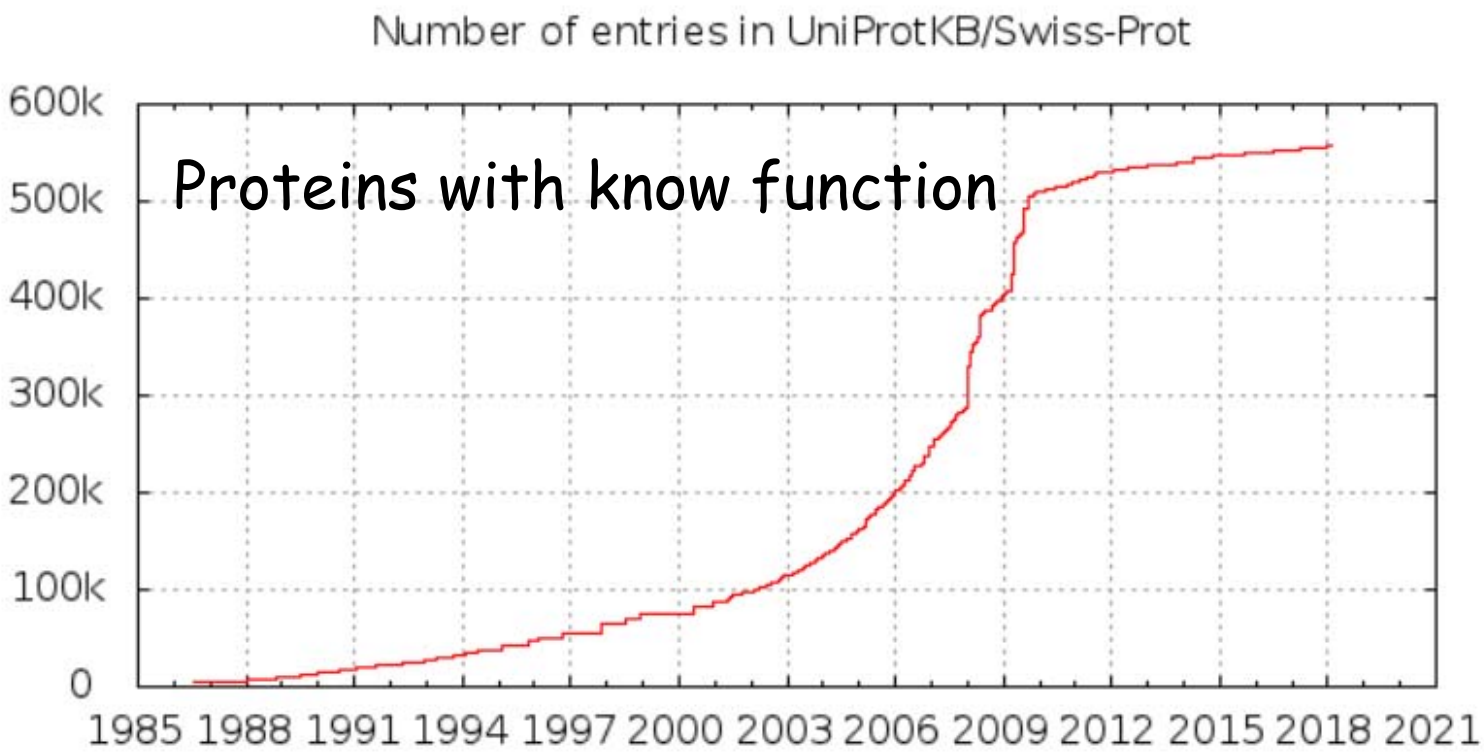






Statistics in UniProtKB/Swiss-Prot

https://web.expasy.org/docs/relnotes/relstat.html



Release 2018_01 of 31-Jan-18 of UniProtKB/Swiss-Prot contains 556568 sequence entries, comprising 199530821 amino acids abstracted from 257937 references

Several numbers to remember

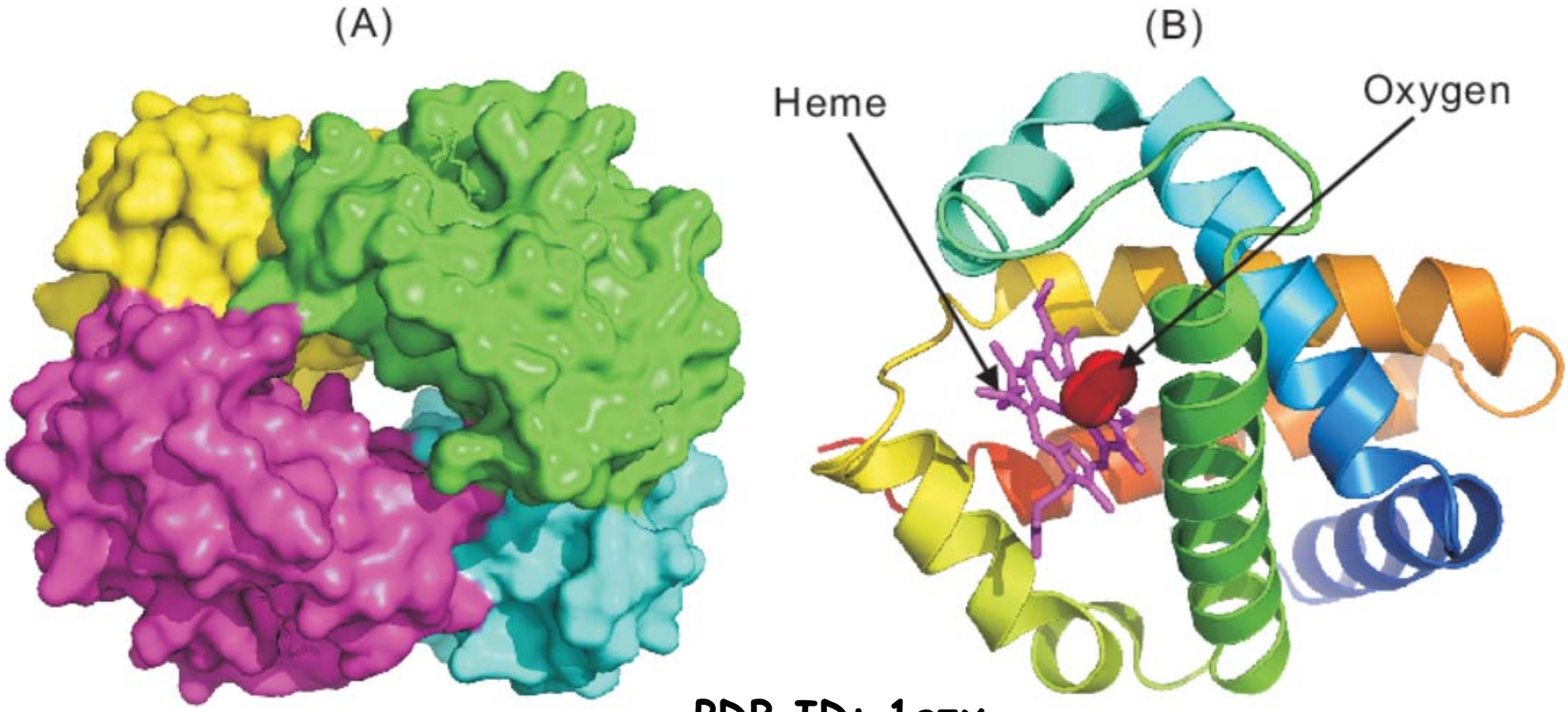
- **950M** Nucleotide sequences in GeneBank (known DNA sequences)
- **100M** protein sequences in TrEMBL (translated proteins)
- **550k** protein sequences in Swiss-Prot (proteins with know function)
- ?? Proteins with known structure

Content

1. Introduction
2. Nucleotide sequence databases
3. Amino acid sequence databases
4. Protein structure databases

The first atomic-level protein structure

Haemoglobin (血红蛋白)

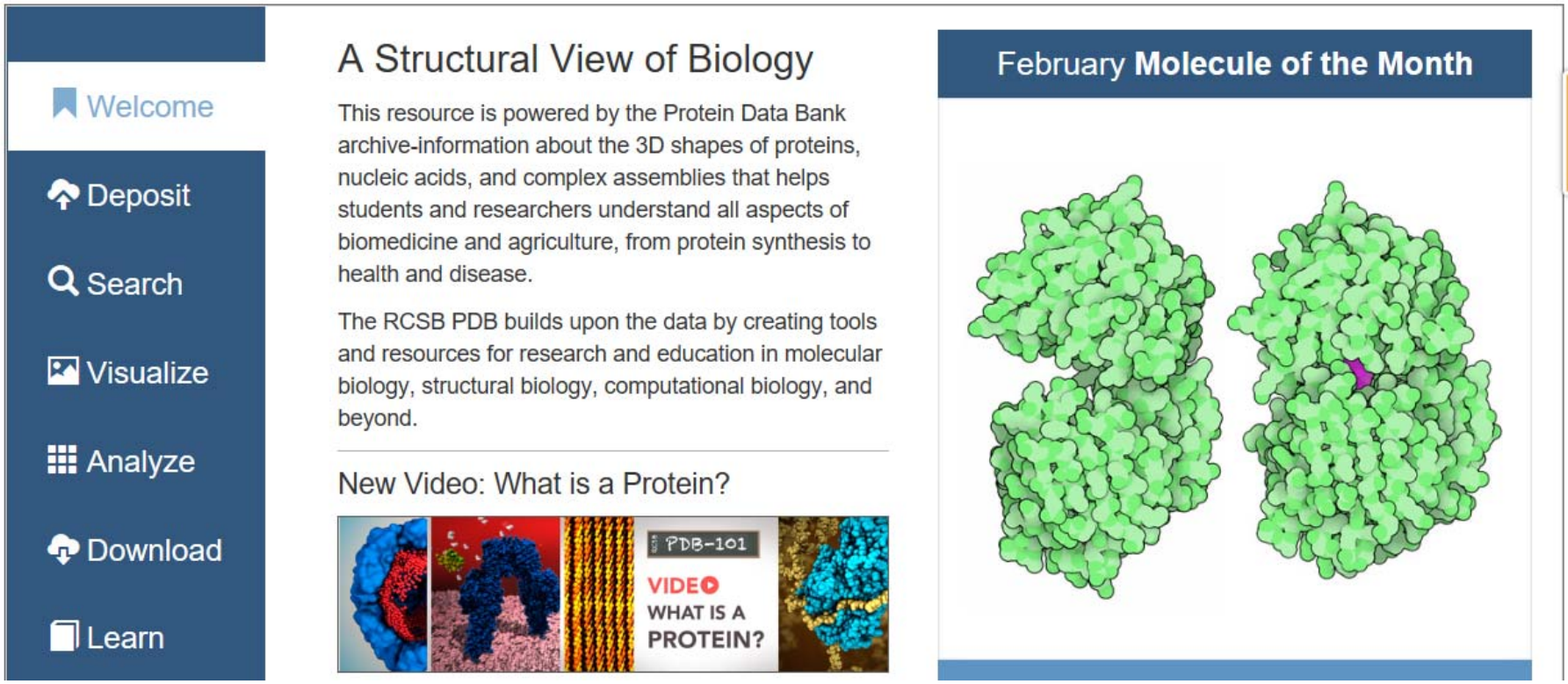


PDB ID: 1gzx

the Nobel Prize in Chemistry in 1962 to Perutz and Kendrew

Protein Data Bank

Type your keywords or PDB ID here, eg, 1gzx

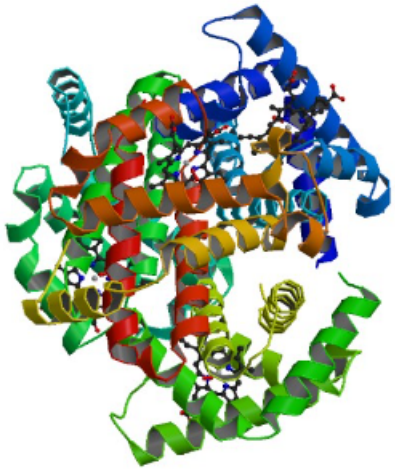


Search result

View .pdb file

Display Files | Download Files

Biological Assembly 1



3D View: Structure | Electron Density | Ligand Interaction

Standalone Viewers
Protein Workshop | Ligand Explorer

Global Symmetry: Cyclic - C2 (3D View)
Global Stoichiometry: Hetero 4-mer - A2B2

Pseudo Symmetry: Dihedral - D2 (3D View)
Pseudo Stoichiometry: Homo 4-mer -A4

1GZX

oxy T state haemoglobin: oxygen bound at all four haems

DOI: 10.2210/pdb1GZX/pdb

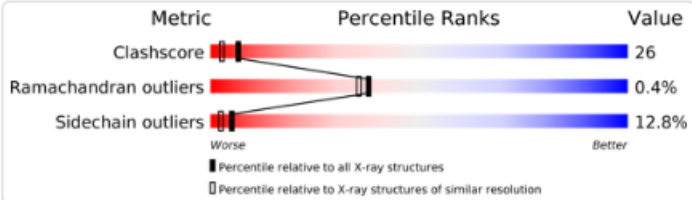
Classification: OXYGEN TRANSPORT
Organism(s): Homo sapiens

Deposited: 2002-06-07 **Released:** 2002-07-08
Deposition Author(s): Paoli, M., Liddington, R., Tame, J., Wilkinson, A., Dodson, G.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.1 Å
R-Value Free: 0.221
R-Value Work: 0.199

wwPDB Validation 3D Report | Full Report



This is version 1.2 of the entry. See complete history.

Literature Download Primary Citation

Crystal Structure of T State Haemoglobin with Oxygen Bound at All Four Haems.

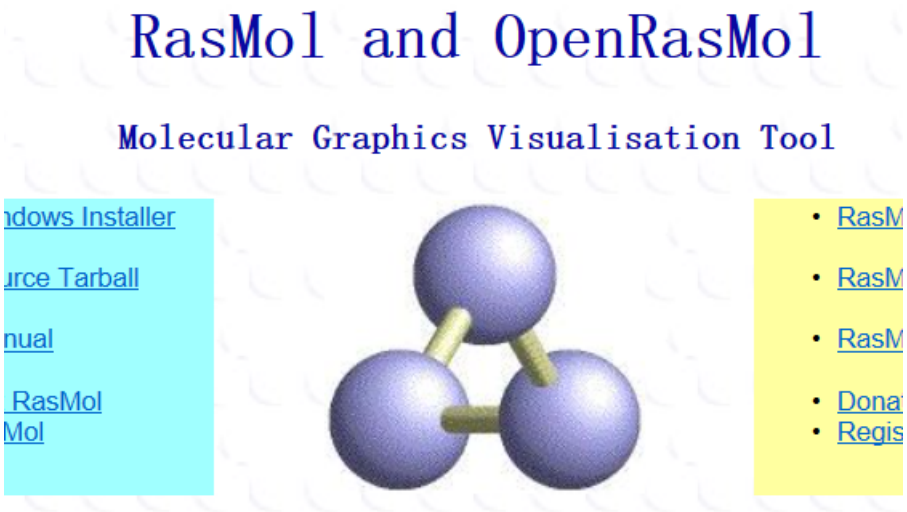
Paoli, M., Liddington, R., Tame, J., Wilkinson, A., Dodson, G.
(1996) J.Mol.Biol. **256**: 775

PDB file format

http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html#ATOM

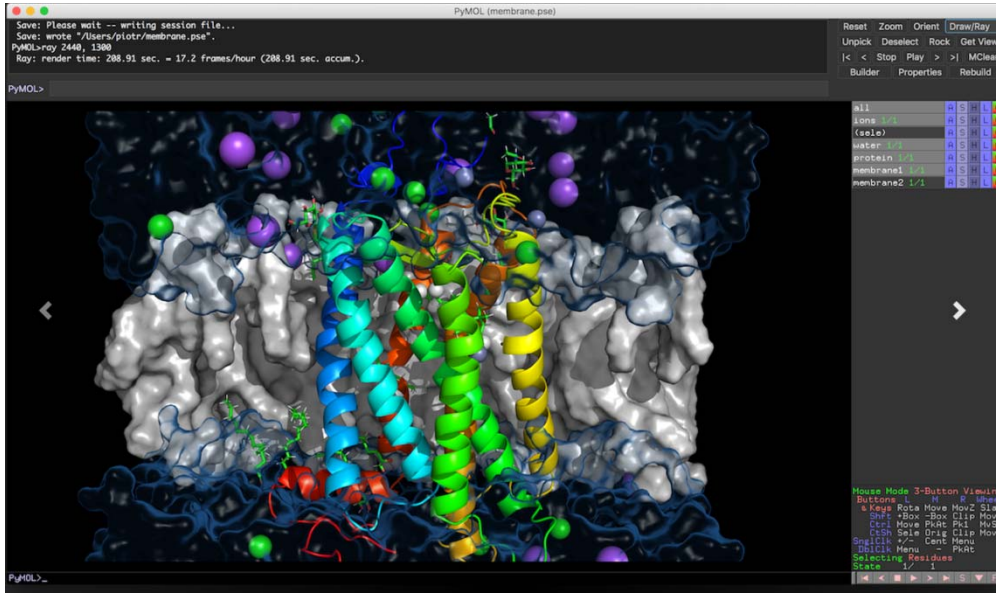
```
ATOM      51  N   THR A   8       20.765   4.947  -6.642  1.00 18.59           N
ATOM      52  CA  THR A   8       20.630   5.778  -7.845  1.00 20.94           C
ATOM      53  C   THR A   8       19.438   6.731  -7.855  1.00 20.07           C
ATOM      54  O   THR A   8       19.504   7.900  -8.272  1.00 19.45           O
ATOM      55  CB  THR A   8       20.618   4.897  -9.154  1.00 22.44           C
ATOM      56  OG1 THR A   8       21.849   4.150  -9.312  1.00 28.00           O
ATOM      57  CG2 THR A   8       20.529   5.834 -10.338  1.00 25.81           C
ATOM      58  N   ASN A   9       18.307   6.163  -7.389  1.00 22.23           N
ATOM      59  CA  ASN A   9       16.976   6.854  -7.314  1.00 20.96           C
ATOM      60  C   ASN A   9       16.992   8.013  -6.334  1.00 18.95           C
ATOM      61  O   ASN A   9       16.385   9.031  -6.629  1.00 18.11           O
ATOM      62  CB  ASN A   9       15.843   5.891  -6.952  1.00 19.87           C
ATOM      63  CG  ASN A   9       15.529   4.873  -7.997  1.00 21.16           C
ATOM      64  OD1 ASN A   9       15.700   5.218  -9.167  1.00 24.31           O
ATOM      65  ND2 ASN A   9       15.059   3.674  -7.718  1.00 20.09           N
ATOM      66  N   VAL A  10       17.659   7.753  -5.230  1.00 20.53           N
ATOM      67  CA  VAL A  10       17.776   8.750  -4.158  1.00 20.63           C
ATOM      68  C   VAL A  10       18.701   9.851  -4.547  1.00 22.18           C
ATOM      69  O   VAL A  10       18.380  10.964  -4.201  1.00 22.12           O
ATOM      70  CB  VAL A  10       18.222   8.039  -2.861  1.00 21.95           C
ATOM      71  CG1 VAL A  10       18.610   9.066  -1.812  1.00 20.26           C
ATOM      72  CG2 VAL A  10       17.194   6.969  -2.505  1.00 21.76           C
ATOM      73  N   LYS A  11       19.819   9.560  -5.212  1.00 26.40           N
ATOM      74  CA  LYS A  11       20.741  10.649  -5.639  1.00 26.50           C
ATOM      75  C   LYS A  11       20.061  11.571  -6.660  1.00 27.10           C
ATOM      76  O   LYS A  11       20.201  12.835  -6.654  1.00 27.46           O
ATOM      77  CB  LYS A  11       21.948  10.059  -6.323  1.00 26.87           C
ATOM      78  CG  LYS A  11       22.830   9.106  -5.534  1.00 30.08           C
ATOM      79  CD  LYS A  11       23.507   9.803  -4.361  1.00 32.89           C
ATOM      80  CE  LYS A  11       24.973   9.361  -4.310  1.00 33.95           C
ATOM      81  NZ  LYS A  11       25.883  10.328  -3.614  1.00 32.73           N
ATOM      82  N   ALA A  12       19.338  10.858  -7.536  1.00 26.72           N
ATOM      83  CA  ALA A  12       18.594  11.540  -8.623  1.00 28.18           C
ATOM      84  C   ALA A  12       17.457  12.435  -8.129  1.00 27.39           C
ATOM      85  O   ALA A  12       17.417  13.612  -8.664  1.00 28.67           O
ATOM      86  CB  ALA A  12       18.195  10.522  -9.693  1.00 30.03           C
ATOM      87  N   ALA A  13       16.649  11.979  -7.177  1.00 26.03           N
ATOM      88  CA  ALA A  13       15.529  12.803  -6.630  1.00 23.44           C
```

Software for visualizing PDB file



RasMol

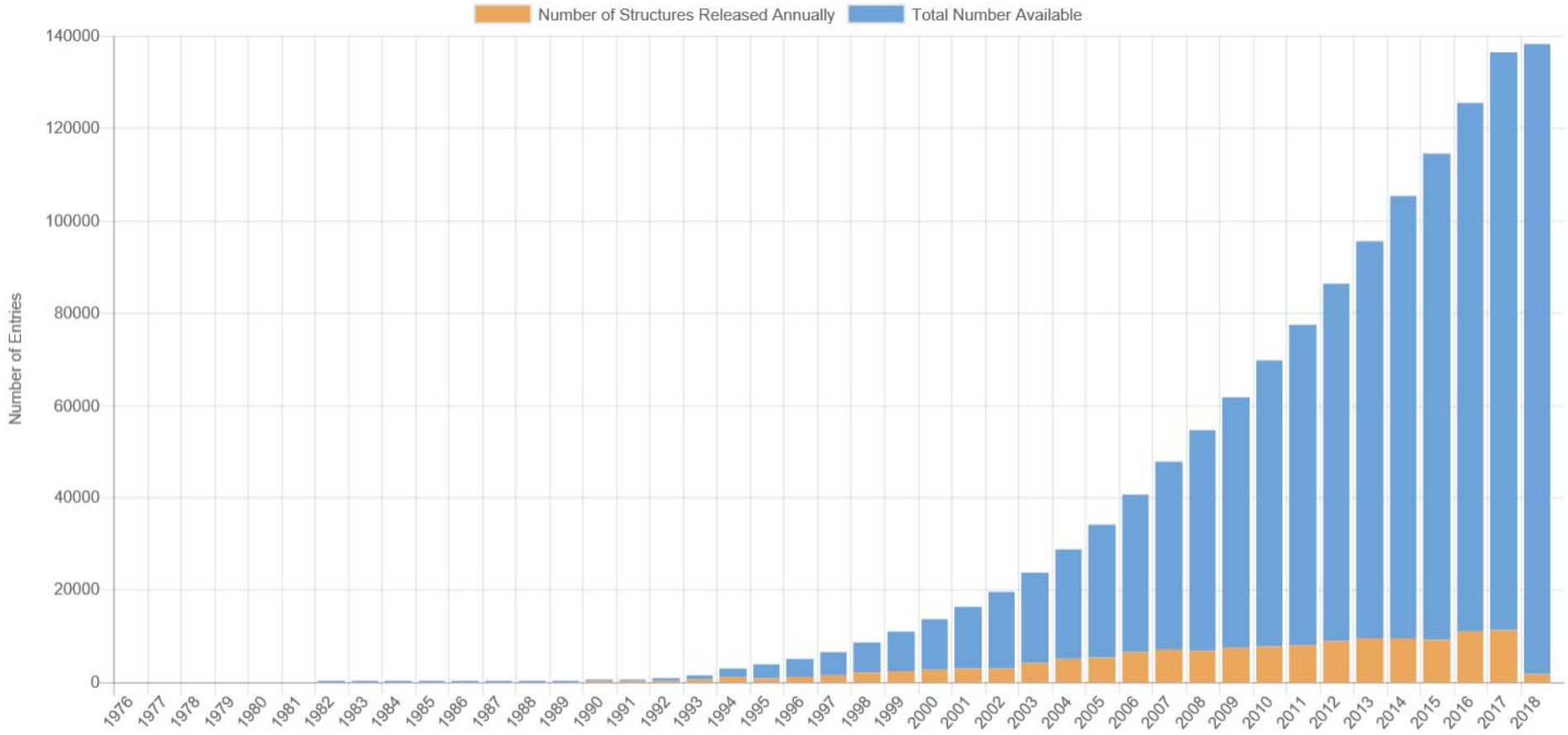
http://www.openrasmol.org/



PyMOL

https://pymol.org/2/

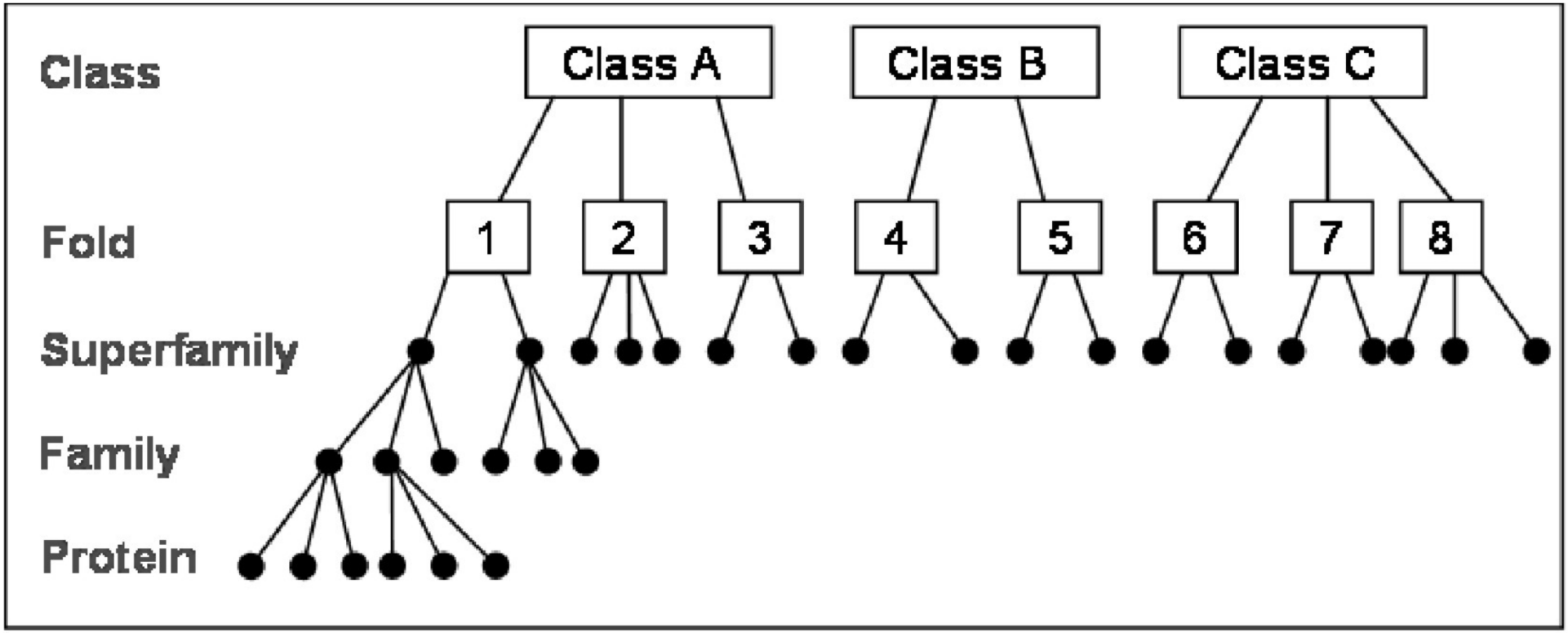
Statistics in PDB



https://www.rcsb.org/stats/growth/overall

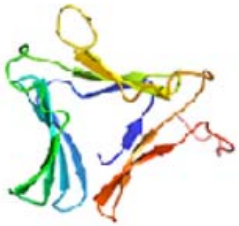
SCOP database

http://scop.mrc-lmb.cam.ac.uk/scop/



SCOP fold

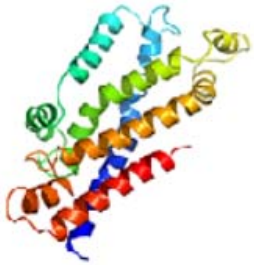
Ammonium transporter



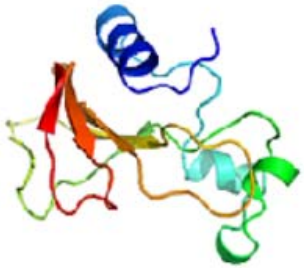
beta-Prism II

FAS1 domain

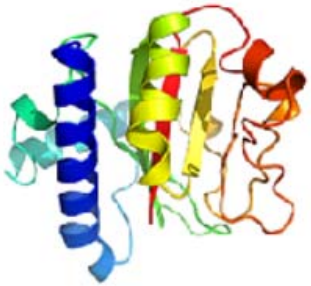
Histone-fold



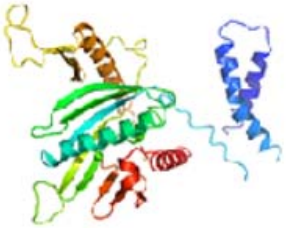
MetI-like



Microbial ribonucleases



Pili subunits

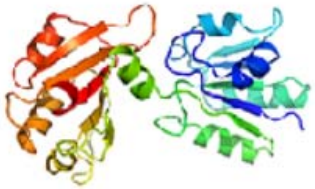


Release factor

Ribokinase-like

Serpins

Streptavidin-like



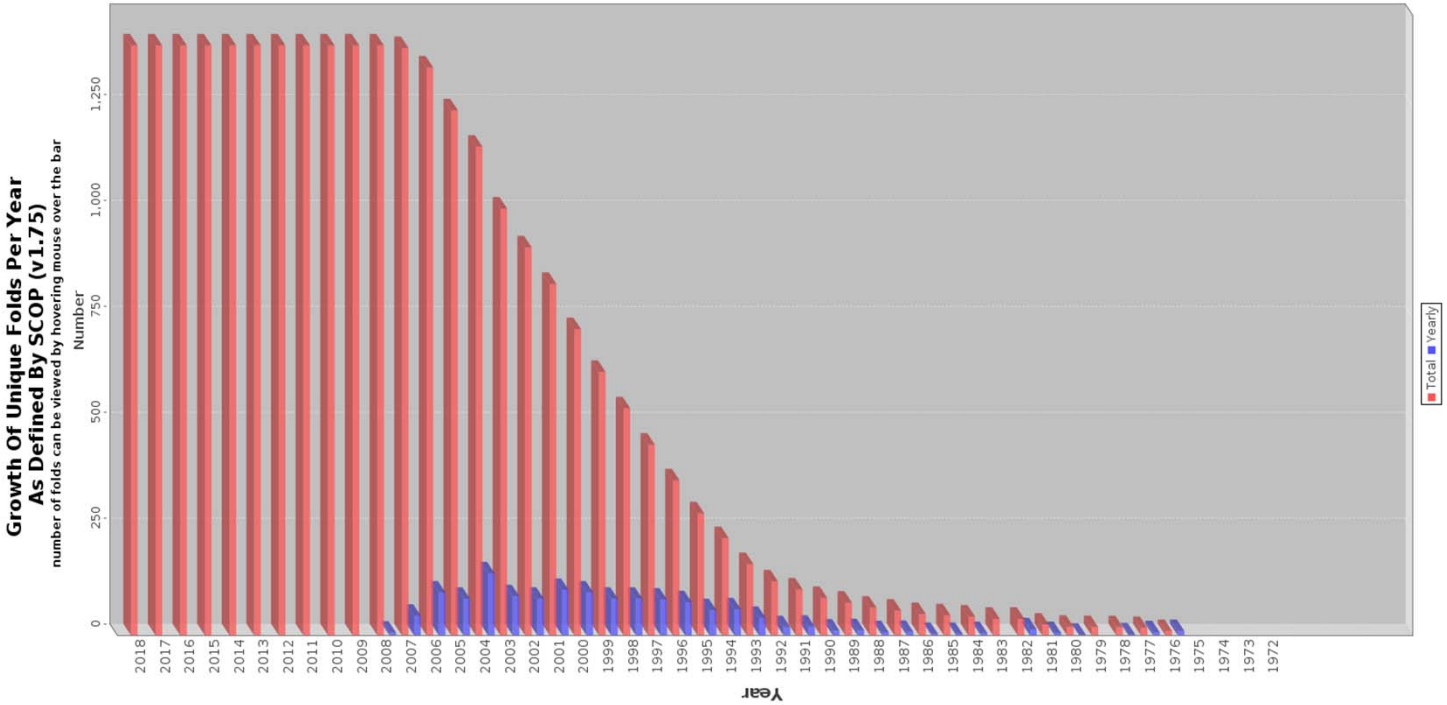
Tetrapyrrole methylase

Statistics in SCOP

SCOPe 2.06-stable statistics:

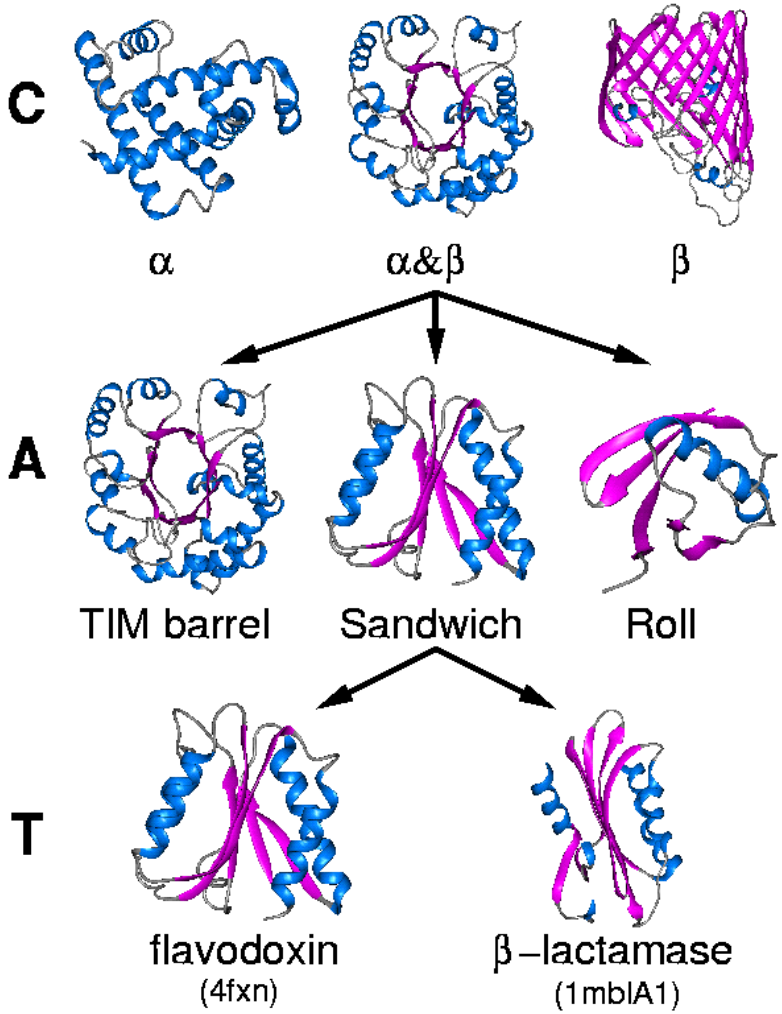
77439 PDB entries (released/updated prior to 2016-01-13). 244326 Domains. 1 Literature reference.

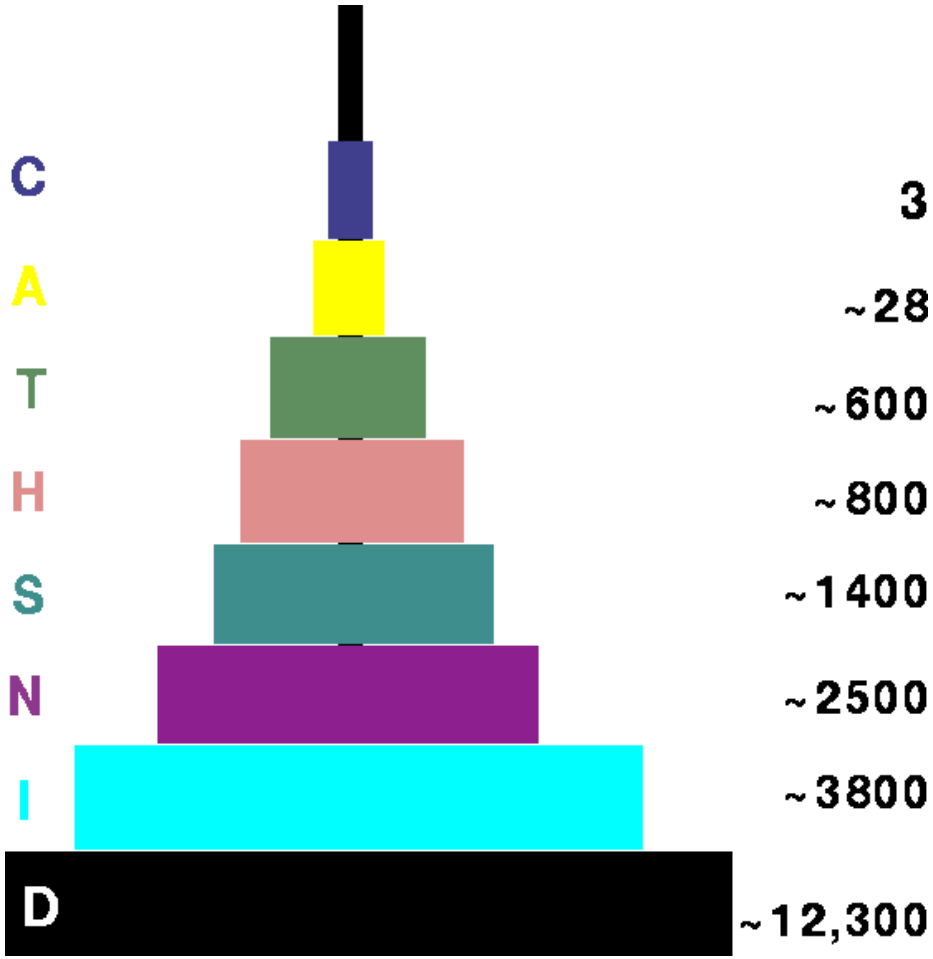
Class	Number of folds	Number of superfamilies	Number of families
a: All alpha proteins	289	513	1049
b: All beta proteins	177	365	952
c: Alpha and beta proteins (a/b)	148	246	984
d: Alpha and beta proteins (a+b)	385	562	1319
e: Multi-domain proteins (alpha and beta)	69	69	111
f: Membrane and cell surface proteins and peptides	59	118	169
g: Small proteins	94	135	267
Totals	**1221** (13 new)	**2008** (24 new)	**4851** (47 new)



CATH database

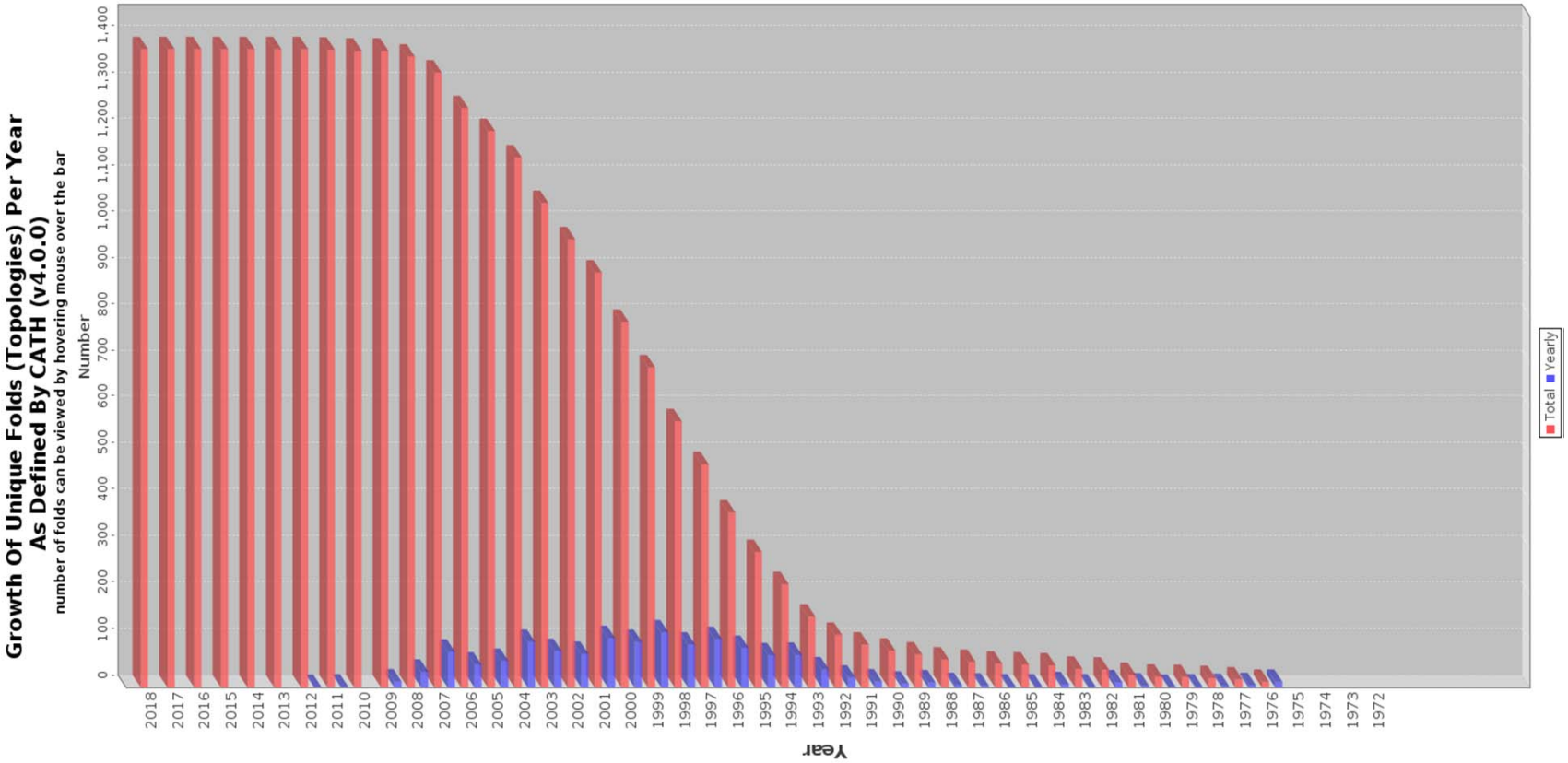
(C)lass, (A)rchitecture, (T)opology or fold, (H)omologous family





Statistics in CATH

Class	Architecture	Topology	Homologous Superfamily	S35 Family	S60 Family	S95 Family	S100 Family	Domains
Class 1	5	405	2174	7771	9948	12797	24176	90302
Class 2	21	244	1395	7011	9627	15169	26904	110260
Class 3	14	634	2428	16196	23313	29670	60020	229776
Class 4	1	108	122	311	399	542	876	4519
TOTAL	**41**	**1391**	**6119**	**31289**	**43287**	**58178**	**111976**	**434857**



Total number of folds is limited

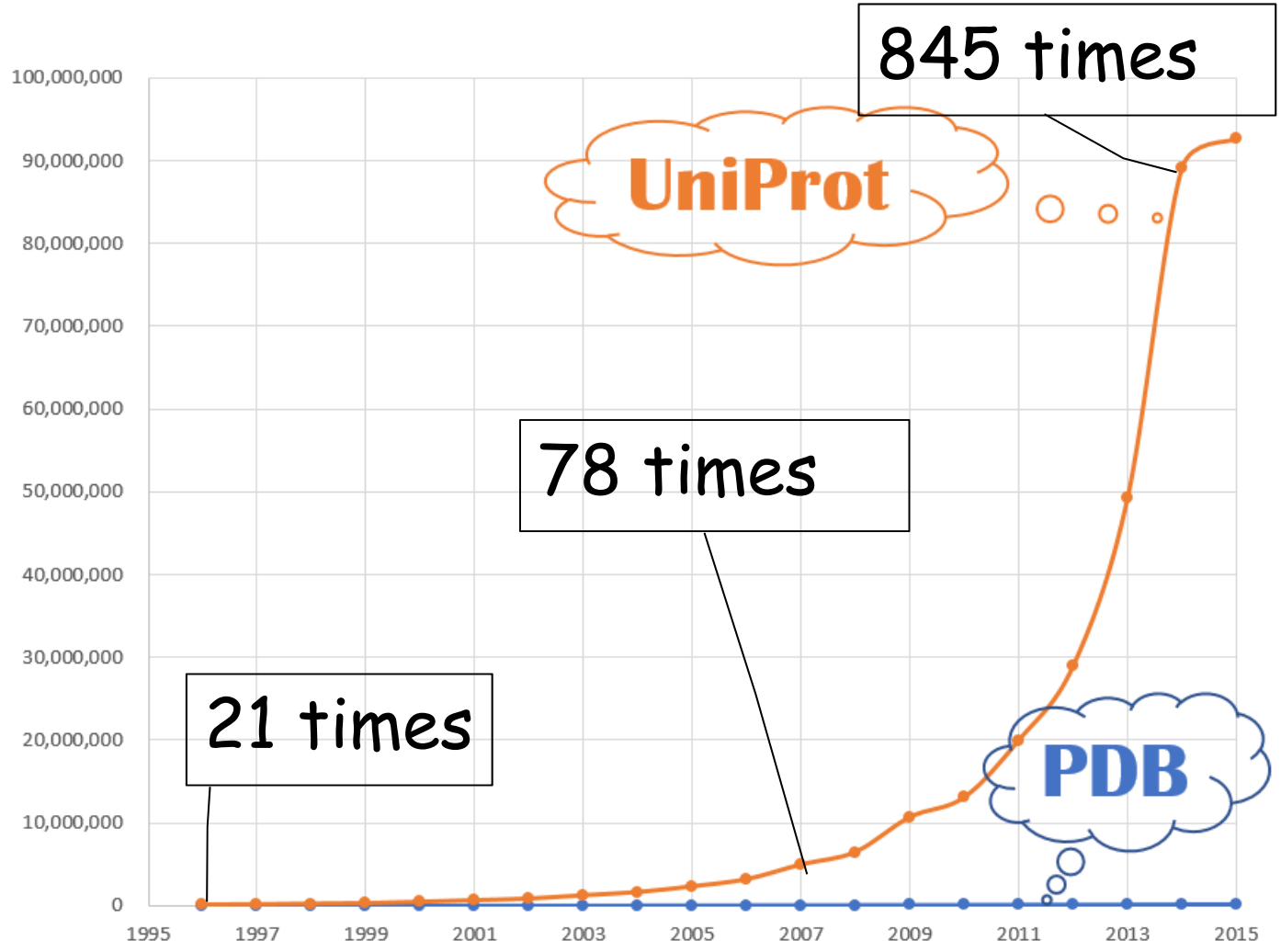
Although the number of protein entries increases exponentially, the increasing speed of the number of new folds keeps decreasing. This indicates that the number of protein folds in nature should be limited (~2,000, by estimation). The fact that many different sequences adopt a similar fold is the base of template-based protein structure prediction algorithms.

Summary of proteomics

- **950M** Nucleotide sequences in GeneBank (known DNA sequences)
- **100M** protein sequences in TrEMBL (translated proteins)
- **550k** protein sequences in Swiss-Prot (proteins with know function)
- **140k** protein structure in PDB (proteins with known structure)
- **~2k** different folds in nature

- **0.55%** of known protein sequences have known function
- **0.14%** of known protein sequences have known structure

Sequence-structure gap



How to bridge the gap?

Many other popular and useful databases…

Pfam: protein sequence family built based on hidden Markov models

Gene Ontology: a database for protein function description

STRING: a database for protein-protein interaction

BioGrid: a curated biological database of protein-protein interactions

InterPro: a database of protein families, domains and functional sites

Drugbank: a database for druggable small molecules

Read through the NAR's annual database issue to find more…

Summary

- Introduction (Central dogma?)
- Nucleotide sequence databases

Names of three databases? How many nucleotide sequences?

- Amino acid sequence databases

Names of the databases and relationship?

- Protein structure databases

Names of the databases and relationship?

Project 1

基于本次课所介绍的数据库，搜索新冠状病毒的信息。需要报告的内容有：
1. 核酸序列
2. 蛋白质序列
3. 蛋白质结构

注：以上信息可能会来自不同的病人，因此若发现有多个样本（病人），请注意区分开，若可能的话，可以考察一下不同样本之间的异同。

1. 本课程分小组学习，请自行组合2-3人一组。
2. 本次作业两周后提交给助教，之后我会公布助教联系方式。

Papers needed to read for the next class

All can be download at:
http://yanglab.nankai.edu.cn/teaching/bioinformatics/papers/

BLOSUM: (Most often-used score matrix for protein sequence alignment):
Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992 Nov 15;89(22):10915-9

PAM matrix (construct a score matrix for guide protein sequence alignment):
DAYHOFF, M., R. SCHWARTZ, AND B. ORCUTT. 1978. A model of evolutionary change in proteins. Pages 345--352 in Atlas of protein sequence and structure, Volume 5 (M. Dayhoff, ed.). National Biomedical Research Foundation, Washington, D.C.

Needleman-Wunsch: (A dynamic programming algorithm for sequence alignments)
S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol., vol.48, pp.443-453 (1970).

Smith-Waterman: (An extension of Needleman-Wunsch)
T F Smith & M S Waterman, Identification of common molecular subsequences. J Mol Biol (1981) 147, 195-197.

BLAST (The most often-used heuristic alignment)
Altschul et al Basic Local Alignment Search Tool. J Mol Biol (1990) 215, 403.