Bioinformatics Algorithms

Data sources and formats

David Hoksza
http://siret.ms.mff.cuni.cz/hoksza
https://ksi.mff.cuni.cz
https://bioinformatika.mff.cuni.cz/cusbg

Sequence databases and data formats

Sequence Databases

- DNA
  - GenBank/RefSeq (NCBI), European Nucleotide Archive (EMBL-EBI), DNA Database of Japan (DDBJ)
- Proteins
  - PIR (USA), SwissProt (EMBL-EBI)
  - UniProt (SwissProt + TrEMBL + PIR)
- Derived Databases
  - Pfam, PROSITE, SILVA
- … and MANY more …

GenBank

- Annotated collection of all publicly available DNA sequences and their protein transcripts including mRNA sequences with coding regions, segments of genomic DNA with a single gene or multiple genes, and ribosomal RNA gene clusters
- Maintained by National Center for Biotechnology Information (NCBI)
- Part of the International Nucleotide Sequence Database Collaboration with the European Nucleotide Archive (ENA) operated by European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ)
- 940,513,260,726 bases from 231,982,592 sequences as of August 2021
- More than 100,000 distinct organisms
- Multiple entries for some loci (sequencing can take place under slightly different conditions in various individuals)

RefSeq

- Reference Sequence (RefSeq) database is a **curated collection** of DNA, RNA, and protein sequences built by NCBI
- Provides separate and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts
- Limited to **major organisms** for which sufficient data is available

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for the same loci	Single record for each molecule of major organisms
Records can contradict each other	
No limit to species	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

Searching GenBank with Entrez

- Text-based
  - term1[field1] AND/OR/NOT term2[field2] AND/OR/NOT …
  - find human topoisomerases complexed with dsDNA
    - Topoisomerase[pdbdescr] AND 2[dnachaincount] AND human[organism]
  - Find all fungal structures with bound calcium at 1-2 Å resolution
    - calcium[ligname] AND fungi[organism] AND 1.0:2.0[resolution]
  - 3D Domains: Find all 50-100 kDa strand-only domains published in 2004
    - 0[helixcount] AND 2004[pdat] AND 50000:100000[molwt]

Retrieving GenBank Data

- Entrez
  - federated search engine securing access to multiple health sciences databases maintained by NCBI
    - GenBank, PubMed, PubChem, …
  - all databases can be searched by one query (possible boolean constraints)
  - provides also API interface through defined URL or SOAP – eUtils
  - searching by
    - text
    - accession number (each sequence get accession number when inserted into GenBank)
    - similarity search using BLAST (nucleotide BLAST, protein BLAST, BLASTX, TBLASTN, TBLASTX)
- FTP
  - basically each directory contains a README file about content of that directory

GenBank Flat File Format

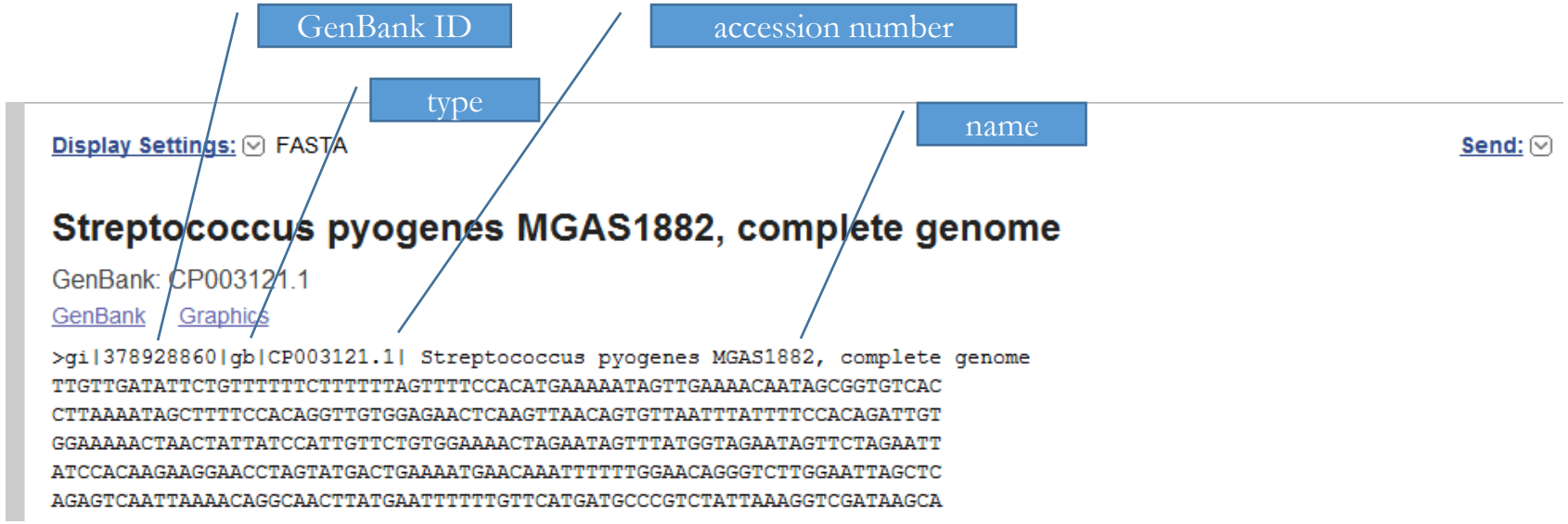
- **Header**
  - **LOCUS** - A short mnemonic name for the entry. The line contains the Accession number, length of molecule, type of molecule (DNA or RNA), a three-letter reference to possible Taxonomy, and the date that the data was made public.
  - **DEFINITION** - description of the sequence
  - **ACCESSION** - accession number is a unique, unchanging code assigned to each entry
  - **VERSION** - primary accession number and a numeric version number associated with the current version of the sequence data in the record. This is followed by an integer key (a "GI") assigned to the sequence by NCBI
  - **KEYWORDS** - gene description
  - **SOURCE** - common name of the organism or the name most frequently used in the literature
  - **ORGANISM** - formal scientific name of the organism (first line) and taxonomic classification levels (second and subsequent lines)
  - **REFERENCE** - articles containing data reported in this entry
  - **AUTHORS** - authors of the citation
  - **TITLE** - full title of citation
  - **JOURNAL** - journal name, volume, year, and page numbers of the citation
  - **MEDLINE** - Medline unique identifier for a citation
  - **PUBMED** - PubMed unique identifier for a citation.
  - **REMARK** - relevance of a citation to an entry
  - **COMMENT** - cross-references to other sequence entries, comparisons to other collections, notes of changes in LOCUS names, and other remarks.
- **Features**
  - **SOURCE** - contains information about organism, mapping, chromosome, tissue alignment, clone identification
  - **CDS** - instructions on how to join sequences together to make an amino acid sequence from the given coordinates. Includes cross references to other databases
  - **GENE Feature** - a segment of DNA identified by a name.
  - **RNA Feature** - used to annotate RNA on genomic sequence (for example: mRNA, tRNA, rRNA)
- **Sequence**

GenBank Flat File Format - Example

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     VDFSNKSNVNVGQVKDIHGRIPEML"
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
      841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
      901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
      961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
     1021 acagcacgtc tttgaacaat acataccaat ttgttgttac aaaccgtcca tccatctcgc
     1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa
     1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca
     1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat gcgccgttac
     1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca ccggtgataa
```

FASTA File Format

- Standard text-based format for storing nucleotide/protein sequence information
- Based on format used in FASTA tool for heuristic-based sequence alignment
- Nucleotides/amino acids represented by a single-letter code
- First line contains metadata
  - starts with >
  - standardized within given database



Sequencing-related file formats

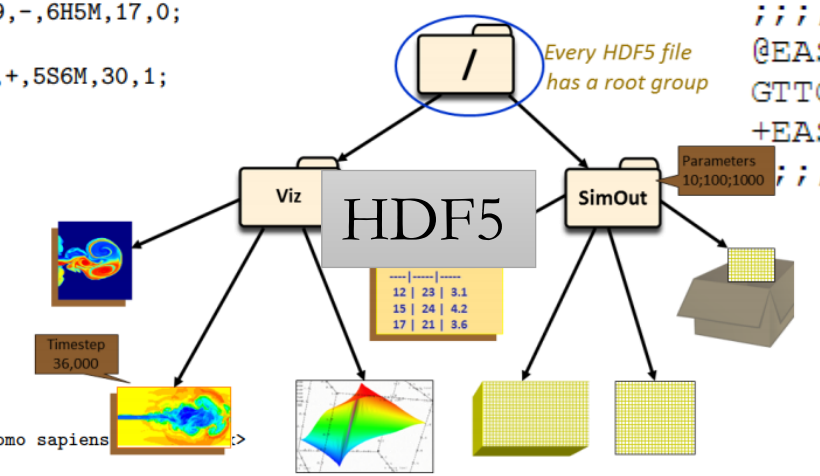
SAM/BAM

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2[illegible]GGATACTG *
r002    0 ref  9 30 3S6[illegible]GGATA    *
r003    0 ref  9 30 5S6[illegible]TAA      * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC      *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC            * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT        * NM:i:1
```

FASTQ

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_[illegible]792
TTGGCAGG[illegible]GATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

HDF5



VCF

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens[illegible]t>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Fr[illegible]">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestra[illegible]">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP memb[illegible]9">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 me[illegible]">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20    14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,[illegible]
20    17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,[illegible]
20    1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,[illegible]
20    1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                  GT:GQ:DP:HQ 0|0:54:7:56,[illegible]
20    1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

BED

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  [illegible]  [illegible]  [illegible]  127472363  127473530  255,0,0
chr7    127473530  127474697  [illegible]  [illegible]  [illegible]  127473530  127474697  255,0,0
chr7    127474697  127475864  [illegible]  [illegible]  [illegible]  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

AND MANY MORE

Swiss-Prot & TrEMBL & PIR

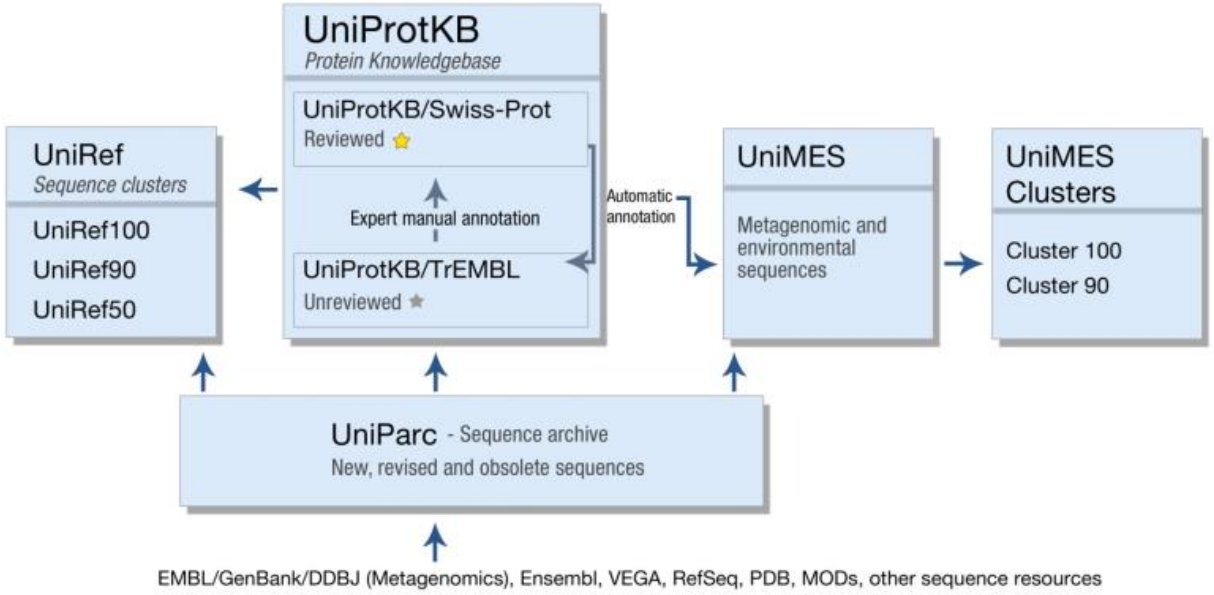
- **Swiss-Prot**
  - **protein sequence database**
  - developed by the Swiss Institute of Bioinformatics (SIB) in 1986 and later on by European Bioinformatics Institute
  - **minimal redundancy**
  - **manually annotated and reviewed**

- **TrEMBL**
  - Translated EMBL Nucleotide Sequence Data Library
  - unreviewed
  - created because sequence data was being generated at a pace that exceeded Swiss-Prot's ability to keep up

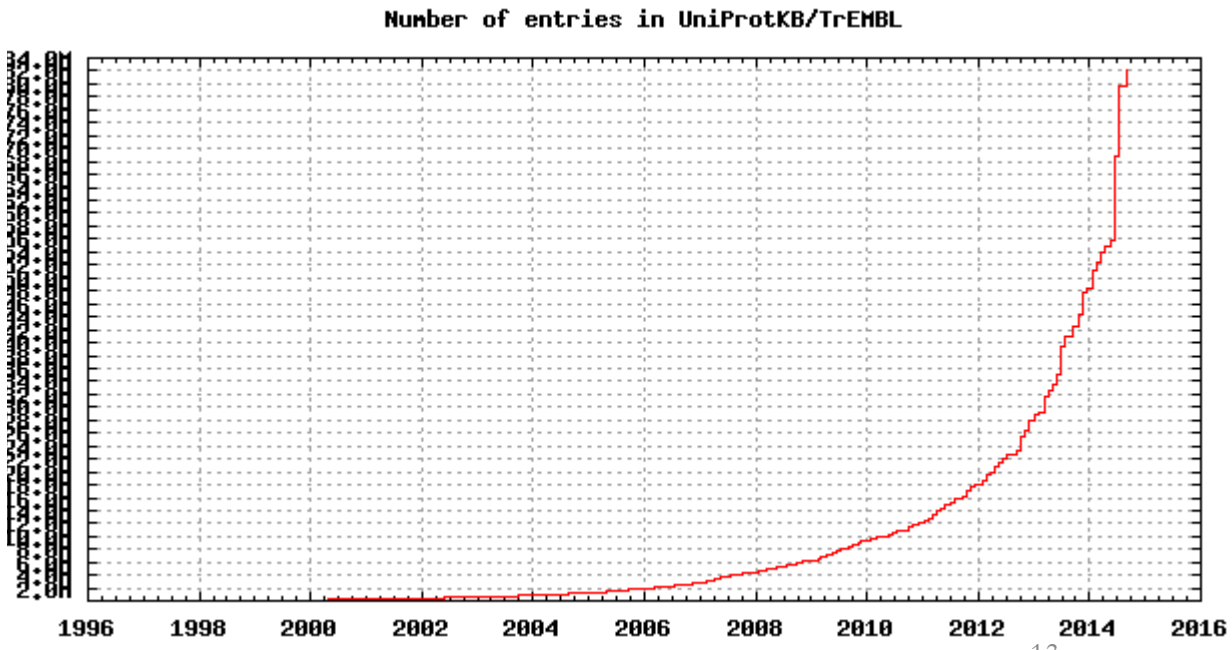
- **PIR (Protein Information Resource)**
  - established in 1984 by the National Biomedical Research Foundation
  - now maintained by Georgetown University Medical Center
  - provides protein **databases and analysis tools** freely accessible to the scientific community
  - includes
    - Protein Sequence Database (PSD) → UniprotKB
      - a database of protein sequences
    - iProClass
      - a database of protein sequences, annotations and curated families
    - PRO (PRotein Ontology), iProLink

UniProt

- **Uni**versal **Prot**ein Resource
- Integration of Swiss-Prot, TrEMBL, PIR-PSD (and many other) databases



- Project started in 2002 at EBI (European Bioinformatics Institute) and SIB (Swiss Institute of Bioinformatics), and PIR



PROSITE

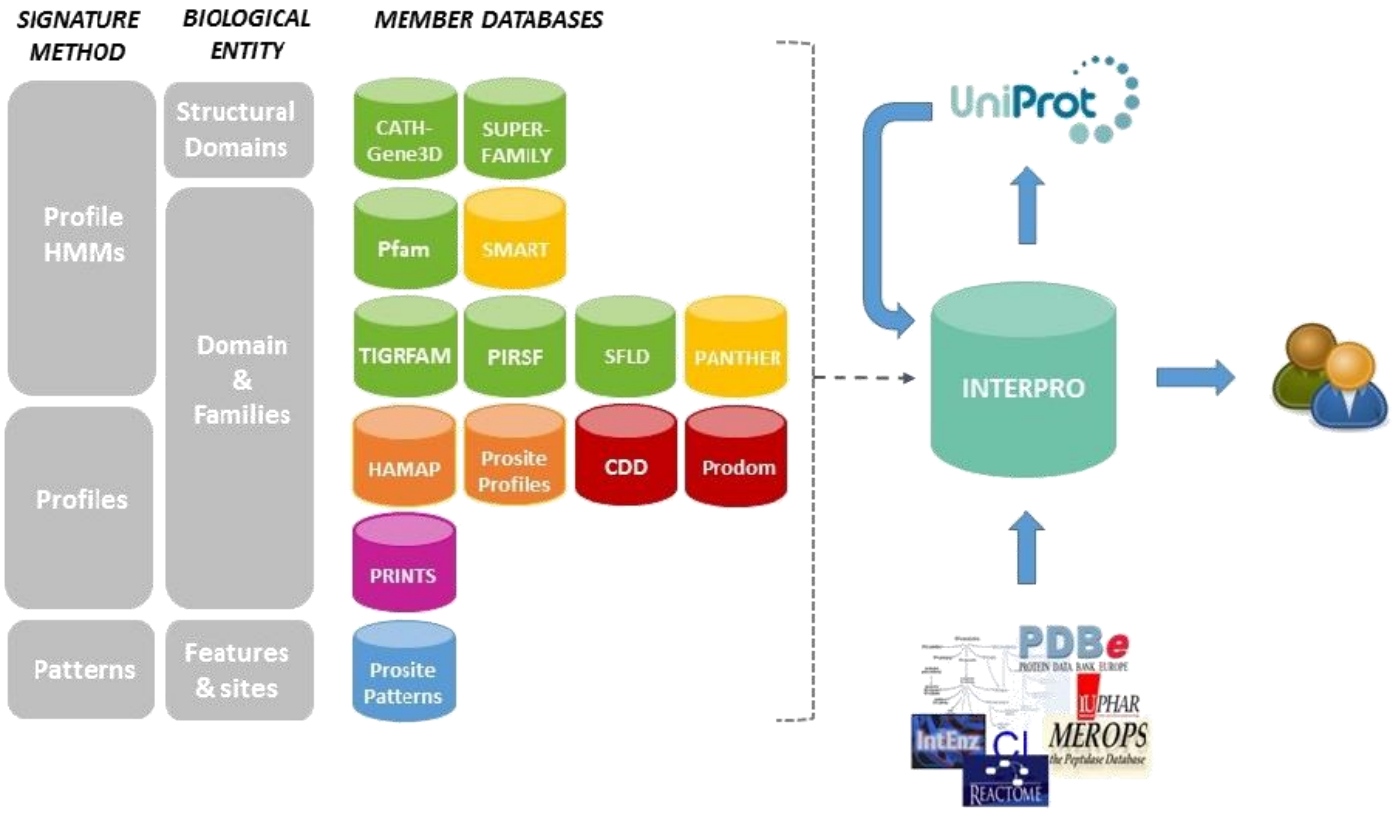
- Database of protein domains, families and functional sites created in 1988
- Available at http://prosite.expasy.org/
- Includes patterns and profiles defining the groups
  - contains tools for motif detection
- Manually curated by SIB
- Can be used to identify new functions or functions of unknown proteins (similarity principle)

Pfam

- Database or protein families based on multiple sequence alignment (MSA)
  - MSAs built using hidden Markov models (HMMS)
  - HMMS part of the database
- Both manually curated (Pfam-A) and automatically classified (Pfam-B)

InterPro

- Functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites
- Integration of member databases into a single searchable database
- Member databases produce signatures which are used to label UniProt entities
- Protein with highly overlapping signatures are grouped into entries



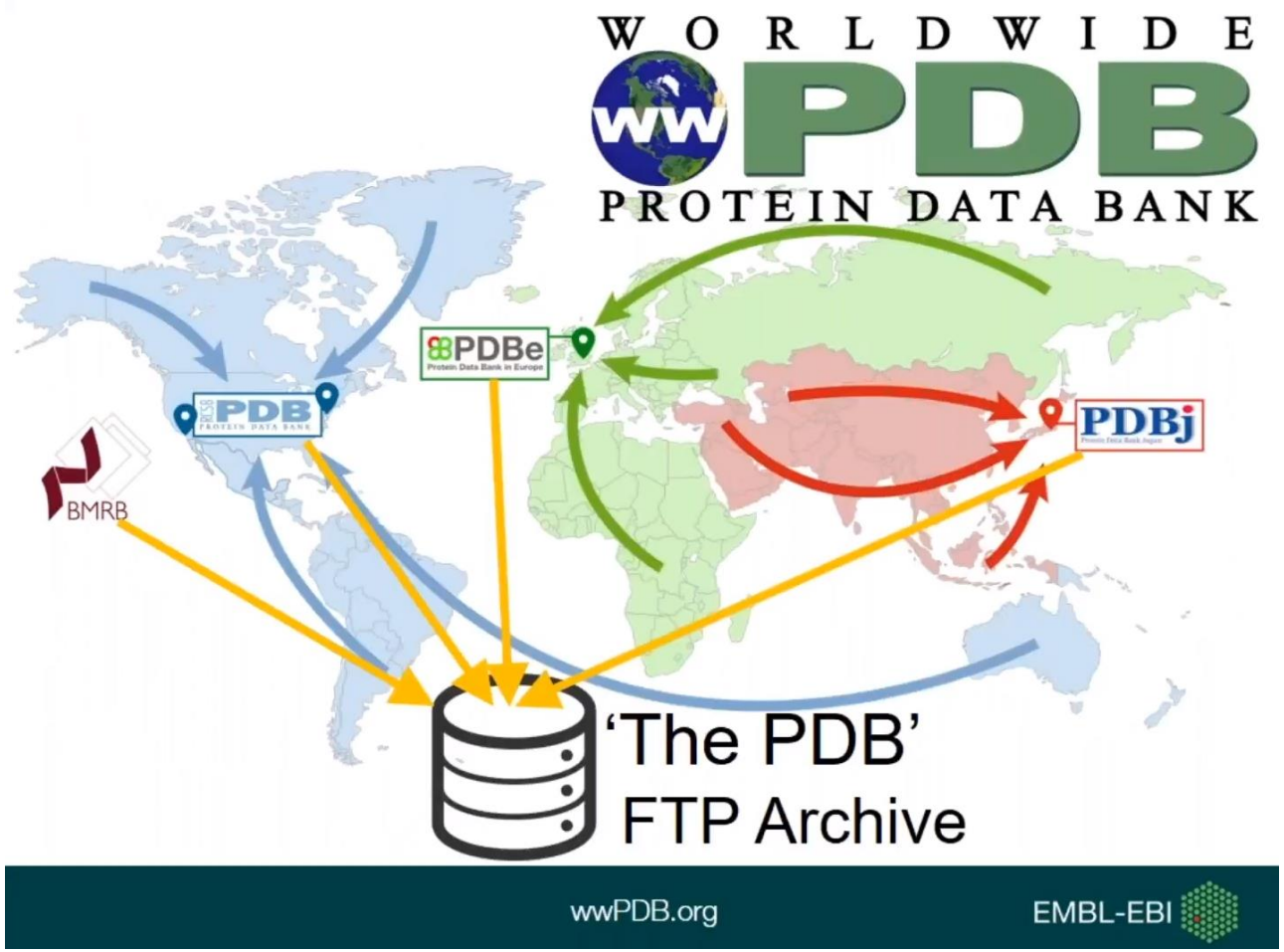
Structure databases and data formats

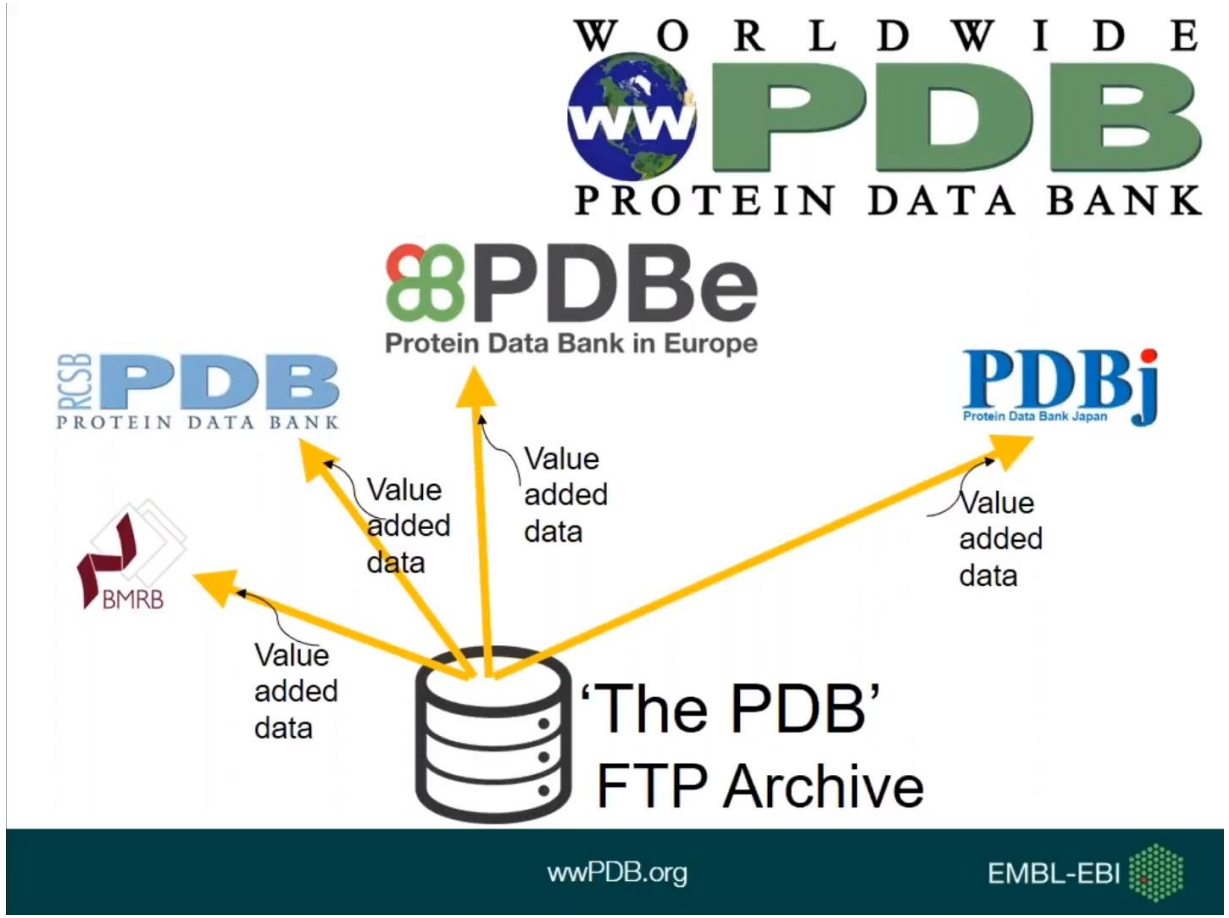
Structure databases

- PDB
  - main depository of protein structural data
- SCOP
  - human-curated hierarchical classification of protein structures built over PDB
- CATH
  - semi-automatic hierarchical classification of protein structures built over PDB
- … and MANY more …

Protein Databank (PDB) (1)

- Established in 1971 as a community-driven effort
- **Primary resource** of (experimental) structure data and related function
- Originally contained protein-only information but nowadays includes **also DNA and RNA** structure information as well as information about complexes





source: https://www.youtube.com/watch?v=PsjAPMd_XN8&index=54&list=WL

Protein Databank (PDB) (2)

- PDB records contain (amongst other information)
  - positions of individual atoms in the 3D space
  - protein sequence
  - secondary structure elements (SSE) information
  - related classification (SCOP, CATH)
  - meta-information such as release date, structure determination data, etc.

- PDB data accessible using
  - web interface
  - FTP
  - API/web services

- Each record is uniquely identified by its PDB ID
  - 4 letter code, e. g., 2AWY

PDB format

- http://www.wwpdb.org/docs.html

- **Text file** containing information about **3D coordinates of atoms** and **supporting information** split into sections
  - title
  - primary structure
  - heterogen
  - secondary structure
  - connectivity annotation
  - miscellaneous features
  - crystallographic and coordinate transformation
  - **coordinates**
  - connectivity
  - bookkeeping

- **Individual records** in the sections are **string data types** with **fixed-length parts** (e.g., date in the HEADER record appears on position 51-59)

- Valid **not only for proteins** but also for other molecules (DNA, RNA, ligands)

PDB format – title section

- **Description** of the **experiment** and the **biological macromolecules** present in the entry
- Records
  - HEADER, OBSLTE, TITLE, SPLIT, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, AUTHOR, REVDAT, SPRSDE, JRNL, REMARK
- **HEADER**
  - class
  - deposition date
  - identifier

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
HEADER    PHOTOSYNTHESIS                          28-MAR-07   2UXK
HEADER    TRANSFERASE/TRANSFERASE INHIBITOR       17-SEP-04   1XH6
HEADER    MEMBRANE PROTEIN, TRANSPORT PROTEIN     20-JUL-06   2HRT
```

- **TITLE**

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
TITLE     RHIZOPUSPEPSIN COMPLEXED WITH REDUCED PEPTIDE INHIBITOR
TITLE     STRUCTURE OF THE TRANSFORMED MONOCLINIC  LYSOZYME BY
TITLE    2 CONTROLLED DEHYDRATION
TITLE     NMR STUDY  OF OXIDIZED THIOREDOXIN MUTANT (C62A,C69A,C73A)
TITLE    2 MINIMIZED  AVERAGE STRUCTURE
```

- **EXPDATA**
  - information about the experiment

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
EXPDTA    X-RAY DIFFRACTION
EXPDTA    NEUTRON DIFFRACTION; X-RAY DIFFRACTION
EXPDTA    SOLUTION NMR
EXPDTA    ELECTRON MICROSCOPY
```

- **JRNL**
  - primary literature citation that describes the experiment which resulted in the deposited coordinate set

PDB format – primary structure

- **Sequence information**
- Records
  - DBREF, DBREF1/DBREF2, SEQADV, SEQRES, MO

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
DBREF  2JHQ A    1   226  UNP    Q9KPK8   UNG_VIBCH        1    226

DBREF  3AKY A    1   219  UNP    P07170   KAD1_YEAST       3    221

DBREF  1HAN A    2   298  UNP    P47228   BPHC_BURCE       1    297

DBREF  3D3I A    0   760  UNP    P42592   YGJK_ECOLI      23    783
DBREF  3D3I B    0   760  UNP    P42592   YGJK_ECOLI      23    783

DBREF  3C2J A    1     8  PDB    3C2J     3C2J             1      8
DBREF  3C2J B  101   108  PDB    3C2J     3C2J           101    108

DBREF  1FFK 0    2  2923  GB     3377779  AF034620      2597   5518
```

- **DBREF**
  - link to corresponding database sequence
- **SEQADV**
  - differences between PDB record and corresponding seq DB record
- **SEQRES**
  - listing of the consecutive chemical components covalently linked in a linear fashion to form a polymer
    - line number for given chain
    - chain ID
    - # residues in chain
    - residues

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
SEQRES   1 A   21  GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU
SEQRES   2 A   21  TYR GLN LEU GLU ASN TYR CYS ASN
SEQRES   1 B   30  PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU
SEQRES   2 B   30  ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR
SEQRES   3 B   30  THR PRO LYS ALA
SEQRES   1 C   21  GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU
SEQRES   2 C   21  TYR GLN LEU GLU ASN TYR CYS ASN
SEQRES   1 D   30  PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU
SEQRES   2 D   30  ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR

SEQRES   1 A    8   DA  DA  DC  DC  DG  DG  DT  DT
SEQRES   1 B    8   DA  DA  DC  DC  DG  DG  DT  DT

SEQRES   1 X   39    U   C   C   C   C   C   G   U   G   C   C   C   A
SEQRES   2 X   39    U   A   G   C   G   G   C   G   U   G   G   A   A
SEQRES   3 X   39    C   C   A   C   C   C   G   U   U   C   C   C   A
```

PDB format – heterogen section

- Description of **non-standard residues** in the entry
- Groups are considered HET if they **are not part of a biological polymer** described in SEQRES but are rather **bound** to it

```
                 1         2         3
12345678901234567890123456789012
HET    TRS  B 975       8

HET    UDP  A1457      25
HET    B3P  A1458      19

HET    NAG  Y   3      15
HET    FUC  Y   4      10
HET    NON  Y   5      12
HET    UNK  A 161       1
```

- Records
  - HET, FORMUL, HETNAM, HETSYN
- **HET**
  - het ID
  - chain
  - sequence number
  - insertion code
  - number of atoms
- **HETNAM**
  - continuation
  - het ID

```
         1         2         3         4         5         6
12345678901234567890123456789012345678901234567890123456789012345678901234
HETNAM     NAG N-ACETYL-D-GLUCOSAMINE
HETNAM     SAD BETA-METHYLENE SELENAZOLE-4-CARBOXAMIDE ADENINE
HETNAM  2  SAD DINUCLEOTIDE

HETNAM     UDP URIDINE-5'-DIPHOSPHATE

HETNAM     UNX UNKNOWN ATOM OR ION
HETNAM     UNL UNKNOWN LIGAND

HETNAM     B3P 2-[3-(2-HYDROXY-1,1-DIHYDROXYMETHYL-ETHYLAMINO)-
HETNAM   2 B3P  PROPYLAMINO]-2-HYDROXYMETHYL-PROPANE-1,3-DIOL
```

PDB format – coordinate section

- Collection of **atomic coordinates**

- Records
  - MODEL, ATOM, ANISOU, TER, HETATM, ENDMDL

- **MODEL/ENDMDL**
  - each structure can be captured multiple times → multiple models

- **TER**
  - end of model

- **ATOM/HETATM**
  - atom serial number, atom name, residue name, alternate location, residue name, chain identifier, residue sequence
    number, insertion code, x, y, z coordinates, …

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM     32  N  AARG A  -3      11.281  86.699  94.383  0.50 35.88           N
ATOM     33  CA AARG A  -3      12.353  85.696  94.456  0.50 36.67           C
ATOM     34  C  AARG A  -3      13.559  86.257  95.222  0.50 37.37           C
ATOM     35  O  AARG A  -3      13.753  87.471  95.270  0.50 37.74           O
ATOM     36  CB AARG A  -3      12.774  85.306  93.039  0.50 37.25           C
ATOM     37  CG AARG A  -3      11.754  84.432  92.321  0.50 38.44           C
ATOM     38  CD AARG A  -3      11.698  84.678  90.815  0.50 38.51           C
ATOM     39  NE AARG A  -3      12.984  84.447  90.163  0.50 39.94           N
ATOM     40  CZ AARG A  -3      13.202  84.534  88.850  0.50 40.03           C
ATOM     41  NH1AARG A  -3      12.218  84.840  88.007  0.50 40.76           N
ATOM     42  NH2AARG A  -3      14.421  84.308  88.373  0.50 40.45           N
ATOM     43  N  BARG A  -3      11.296  86.721  94.521  0.50 35.60           N
ATOM     44  CA BARG A  -3      12.333  85.862  95.041  0.50 36.42           C
ATOM     45  C  BARG A  -3      12.759  86.530  96.365  0.50 36.39           C
ATOM     46  O  BARG A  -3      12.924  87.757  96.420  0.50 37.26           O
ATOM     47  CB BARG A  -3      13.428  85.746  93.980  0.50 36.60           C
ATOM     48  CG BARG A  -3      12.866  85.172  92.651  0.50 37.31           C
ATOM     49  CD BARG A  -3      13.374  85.886  91.406  0.50 37.66           C
ATOM     50  NE BARG A  -3      12.644  85.487  90.195  0.50 38.24           N
ATOM     51  CZ BARG A  -3      13.114  85.582  88.947  0.50 39.55           C
ATOM     52  NH1BARG A  -3      14.338  86.056  88.706  0.50 40.23           N
```

PDB format – example (1AOI)

1AOI.pdb

mmCIF

- macromolecular Crystallographic Information File
  - Extension of CIF format
  - Data match mmCIF dictionary

- PDB format is not capable of capturing some more complex structures
  - mmCIF includes features which are either not available in the PDB format (description of the biological active molecule) or are not structured (experimental details from REMARK records)

```
HEADER    PLANT SEED PROTEIN                      11-OCT-91   1CBN

          _struct.entry_id                    '1CBN'
          _struct.title                       'PLANT SEED PROTEIN'

          _struct_keywords.entry_id           '1CBN'
          _struct_keywords.text               'plant seed protein'

          _database_2.database_id              PDB
          _database_2.database_code            1CBN

          _database_PDB_rev.num                       1
          _database_PDB_rev.date_original 1991-10-11

loop_
    _atom_site.group_PDB
    _atom_site.type_symbol
    _atom_site.label_atom_id
    _atom_site.label_comp_id
    _atom_site.label_asym_id
    _atom_site.label_seq_id
    _atom_site.label_alt_id
    _atom_site.cartn_x
    _atom_site.cartn_y
    _atom_site.cartn_z
    _atom_site.occupancy
    _atom_site.B_iso_or_equiv
    _atom_site.footnote_id
    _atom_site.entity_id
    _atom_site.entity_seq_num
    _atom_site.id
  ATOM N  N   VAL  A  11 . 25.360  30.691  11.795  1.00  17.93 . 1  11 1
  ATOM C  CA  VAL  A  11 . 25.970  31.965  12.332  1.00  17.75 . 1  11 2
  ATOM C  C   VAL  A  11 . 25.569  32.010  13.881  1.00  17.83 . 1  11 3
#                [data omitted]
```

SCOP (**S**tructural **C**lassification **o**f **P**rotein Structures)

- Curated hierarchical classification (gold standard) built over PDB established in 1995
- Classifies proteins by domains (not whole structures)
  - independent subunits of protein structure which can each show function by its own (loose definition)

- Next to function discovery, it can be used for testing quality of similarity methods
  - one can take structure from PDB (SCOP)
  - identify most similar protein in SCOP (according to given pairwise similarity measure)
  - check whether, e.g., the most similar structure share classification with the query
  - when this is done for all structures, one can see in how many per cents the predicted classification was correct → quality of the measure

SCOP – hierarchy

1. **Family**
   - proteins in the same family can have **high sequence similarity** (> 30%) **or** lower sequence similarity (> 15%) with **very similar function or structure**
2. **Superfamily**
   - proteins sharing **common evolutionary origin** (based on structural and functional features) but **differing in sequence**
3. **Fold**
   - structures sharing **major secondary structures** in similar topological distribution
4. **Class**
   - structures with **similar folds**
     - **all α** - proteins containing mainly (but not exclusively) α helices
     - **all β** - proteins containing mainly (but not exclusively) β sheets
     - **α/β** - proteins containing β sheet surrounded by α helices
     - **$\alpha + \beta$** - proteins containing α helices separated by β sheets
     - **small proteins, low resolution protein structures**, …

CATH (**C**lass, **H**ierarchy, **T**opology, **H**omologous superfamily)

- Semi automatic, hierarchical classification of protein **domain** structures

- Classification procedure uses a combination of automated and manual techniques which include computational algorithms, empirical and statistical evidence, literature review and expert analysis

- Similar classification to SCOP

CATH - hierarchy

1. **Homologous superfamily**
   - groups together protein domains which are thought to share a **common ancestor** and can therefore be described as homologous
2. **Topology**
   - structures grouped into fold groups at this level depending on both the **overall shape and connectivity of the secondary structures**.
3. **Architecture**
   - structures classified according to their **overall shape** as determined by the **orientations of the secondary structures in 3D space** but ignores the connectivity between them
4. **Class**
   - structures classified according to their **secondary structure composition**
     - mostly α
     - mostly β
     - mixed α/β
     - few secondary structures

Programmatic access to data sources

- UniProt API
  - retrieve individual records by ids or queries
  - mapping between different formats and databases
- Proteins API
  - Mapping of data from large scale studies to UniProt
- PDBe API
  - Access to PDB records
  - Mapping between UniProt and PDB (SIFTS)
- NCBI APIs