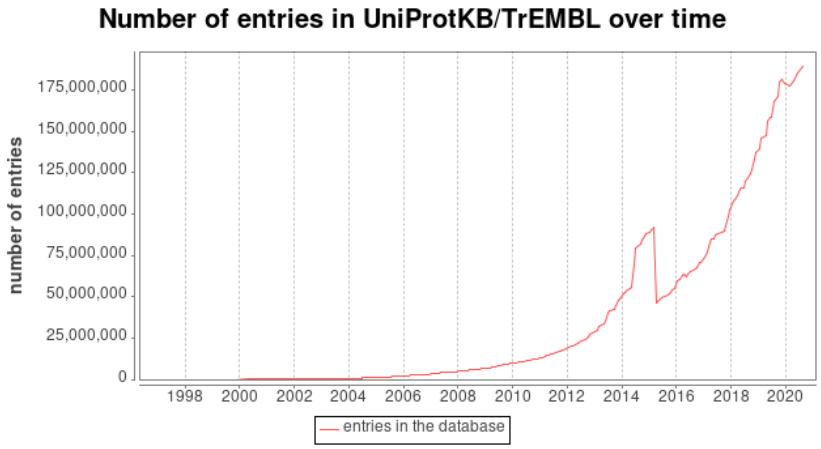
CS612 - Algorithms in Bioinformatics

Databases and Protein Structure Representation

September 29, 2020

Molecular Biology as Information Science

- $> 16,000$ genomes sequenced, mostly bacterial (2019)
- $> 5x10^6$ unique sequences available
- What do we do with them?
  - Compare them to find what is common and different among organisms (Comparative Genomics)
  - Find out how and which genes encode for which proteins
  - Identify changes that lead to disease
  - Associate structural and functional information with new gene sequences



source: `http://www.33rdsquare.com`

- NIH structural genomics project
- Protein Structure Initiative (PSI)
- GOLD (Genomes Online Database) `http://www.genomesonline.org`
- $< 1\%$ of sequences solved
- Experiments lagging behind
- Way too much data for computer scientists to sit around doing nothing

What We Expect From a Biological Databases

- Sequence, functional, structural information, related bibliography
- Well Structured and Indexed
- Well cross-referenced (with other databases)
- Periodically updated and maintained
- Provides tools for analysis and visualization
- Or at least formatted in a compatible way with known tools

http://www.docfoc.com/
biological-databases-pharmamatrix-workshop-2010-philip-winter-ishwar-v-hosamani

Sequence Databases

- International Nucleotide Sequence Database Collaboration (INSDC): `http://www.insdc.org/`
  - NCBI (National Center for Biotechnology Information): `http://ncbi.nih.gov`
  - EMBL-EBI (European Molecular Biology Laboratory, European Bioinformatics Institute): `https://www.ebi.ac.uk/`
  - DDBJ (DNA Data Bank of Japan): `http://www.ddbj.nig.ac.jp/`

Contents of a Database

- Sequences/structures (depends on the database)
- Accession number
- References
- Taxonomic data
- Annotation/curation
- Keywords
- Cross-reference to relevant data in this or other databases.
- Documentation

Organization of a Database

- Hierarchical, where the data is organized at multiple levels.
- Examples: SCOP, CATH, the tree of life.
- Relational: An entry is a set of correspondences between different features of the database (tables).
- It makes it easy to answer queries using operations like union, intersection, difference etc.

NCBI Nucleotide Sequence Databases

- NCBI GenBank (The nucleotide sequence database) – `http://www.ncbi.nlm.nih.gov/genbank/`
- Provides tools for submission (BankIt, Sequin), retrieval (Entrez) and analysis (BLAST, Genome workbench)
- Provides easy access to other NCBI resources

Protein Sequence Databases

- Uniprot – `http://www.uniprot.org/`
- A universal resource, resulting from a merger of several databases.
- Tools: BLAST, align, Retrieve/IDmapping
- Pfam – `http://pfam.xfam.org/`
- A database of protein families based on conserved regions.

Uniprot Entry

www.uniprot.org/uniprot/P0C9R5

UniProt

UniProtKB | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search | Help Contact

UniProtKB - P0C9R5 (36018_ASFWA)

Basket

Display

- Entry
- Publications
- Feature viewer
- Feature table

All None

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

Top

BLAST | Align | Format | Add to basket | History

Feedback | Help video | Other tutorials and videos

Protein: **Protein MGF 360-18R**

Gene: **War-169**

Organism: *African swine fever virus (isolate Warthog/Namibia/Wart80/1980) (ASFV)*

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

Function

Plays a role in virus cell tropism, and may be required for efficient virus replication in macrophages. By similarity

GO - Biological process

- taxis Source: InterPro

Complete GO annotation...

Names & Taxonomy

Protein names	*Recommended name:* **Protein MGF 360-18R**
Gene names	Ordered Locus Names:War-169
Organism	African swine fever virus (isolate Warthog/Namibia/Wart80/1980) (ASFV)
Taxonomic identifier	561444 [NCBI]
Taxonomic lineage	Viruses › dsDNA viruses, no RNA stage › Asfarviridae › Asfivirus ›
Virus host	Ornithodoros (relapsing fever ticks) [TaxID: 6937] Phacochoerus aethiopicus (Warthog) [TaxID: 85517] Phacochoerus africanus (Warthog) [TaxID: 41426] Potamochoerus larvatus (Bushpig) [TaxID: 273792] Sus scrofa (Pig) [TaxID: 9823]
Proteomes	UP000000858 Component: Genome

Uniprot Search

UniProt

UniProtKB | polymerase organism:"homo sapiens" AND reviewed:yes | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

UniProtKB results

About UniProtKB | Basket

Filter by

Reviewed (3,156) Swiss-Prot

Popular organisms

Human (3,155)

HCYV5 (1)

Search terms

Filter "polymerase" as:

gene ontology (2,859)

keyword (69)

protein family (74)

protein name (172)

View by

Results table

Taxonomy

Keywords

Gene Ontology

BLAST | Align | Download | Add to basket | Columns

1 to 25 of 3,156 | Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
O14802	RPC1_HUMAN	**DNA-directed RNA polymerase III sub...**	**POLR3A**	Homo sapiens (Human)	1,390
P24928	RPB1_HUMAN	**DNA-directed RNA polymerase II subu...**	**POLR2A** POLR2	Homo sapiens (Human)	1,970
P30876	RPB2_HUMAN	**DNA-directed RNA polymerase II subu...**	**POLR2B**	Homo sapiens (Human)	1,174
O15318	RPC7_HUMAN	**DNA-directed RNA polymerase III sub...**	**POLR3G**	Homo sapiens (Human)	223
P19387	RPB3_HUMAN	**DNA-directed RNA polymerase II subu...**	**POLR2C** A-152E5.7	Homo sapiens (Human)	275
O95602	RPA1_HUMAN	**DNA-directed RNA polymerase I subun...**	**POLR1A**	Homo sapiens (Human)	1,720
Q15054	DPOD3_HUMAN	**DNA polymerase delta subunit 3**	**POLD3** KIAA0039	Homo sapiens (Human)	466
P52435	RPB11_HUMAN	**DNA-directed RNA polymerase II subu...**	**POLR2J** POLR2J1	Homo sapiens (Human)	117
Q9BUI4	RPC3_HUMAN	**DNA-directed RNA polymerase III sub...**	**POLR3C**	Homo sapiens (Human)	534
O15514	RPB4_HUMAN	**DNA-directed RNA polymerase II subu...**	**POLR2D**	Homo sapiens (Human)	142
P62487	RPB7_HUMAN	**DNA-directed RNA polymerase II subu...**	**POLR2G** RPB7	Homo sapiens (Human)	172

Protein Structure Databases

- PDB – Protein Data Bank – `http://www.rcsb.org/pdb/`
- SCOP2 – Structural Classification of Proteins v.2 – `http://scop2.mrc-lmb.cam.ac.uk/`
- CATH – Another structural classification database – `http://www.cathdb.info/`
- EMDB – Electron microscopy Database – `https://www.ebi.ac.uk/pdbe/emdb/` (Actually part of the PDB now)

PFam Entry

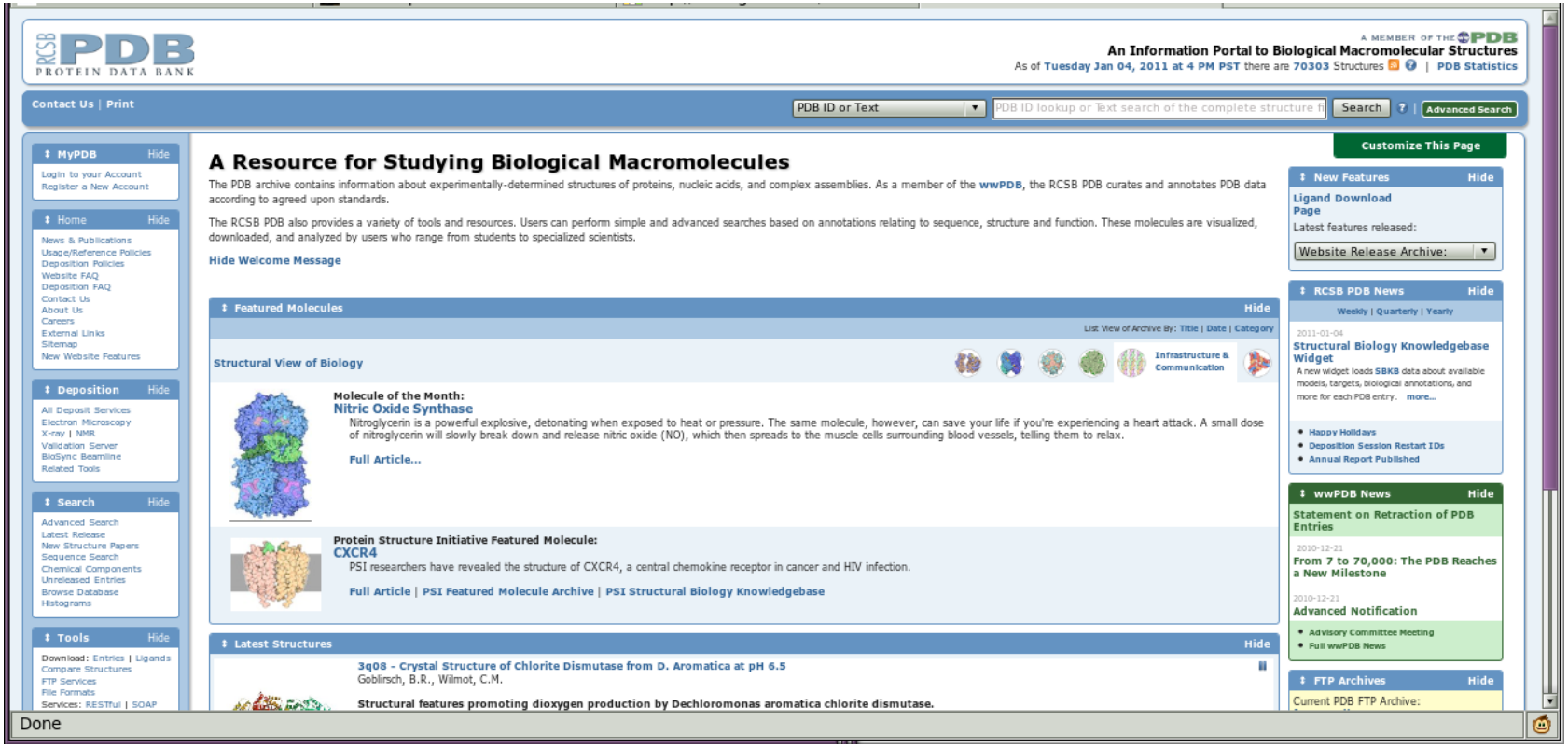
The Protein Databank (PDB)

- Most (all) of the protein structures discovered to date can be found in a large protein repository called the The RCSB Protein DataBank (PDB): `http://www.rcsb.org`.
- PDB is a public domain repository that contains experimentally determined structures of three-dimensional proteins.
- The majority of the proteins in the PDB have been determined by x-ray crystallography.
- The number of proteins determined using NMR methods has been increasing as efficient computational techniques to derive structures from NMR data have been developed.

Retrieving Protein Structures from the PDB

- Starting with 7 structures in 1971, the number has been growing exponentially since then.
- There are over 100,000 structures as of today (early 2016).
- All PDB entries are 4-letter words! 1CRZ, 2BHL . . .
- Sometimes the chain number is added: 1CRZA, 1CRZB . . .
- You can download the coordinates and display the structure
- The BLAST server and other databases contain links to PDB entries if the sequence has a known structure.

The PDB


RCSB PDB
PROTEIN DATA BANK
A MEMBER OF THE PDB
An Information Portal to Biological Macromolecular Structures
As of Tuesday Jan 04, 2011 at 4 PM PST there are 70303 Structures | PDB Statistics
Contact Us | Print
PDB ID or Text
Search
Advanced Search
Customize This Page
MyPDB Hide
Login to your Account
Register a New Account
Home Hide
News & Publications
Usage/Reference Policies
Deposition Policies
Website FAQ
Deposition FAQ
Contact Us
About Us
Careers
External Links
Sitemap
New Website Features
Deposition Hide
All Deposit Services
Electron Microscopy
X-ray | NMR
Validation Server
BioSync Beamline
Related Tools
Search Hide
Advanced Search
Latest Release
New Structure Papers
Sequence Search
Chemical Components
Unreleased Entries
Browse Database
Histograms
Tools Hide
Download: Entries | Ligands
Compare Structures
FTP Services
File Formats
Services: RESTful | SOAP
A Resource for Studying Biological Macromolecules
The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.
The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.
Hide Welcome Message
Featured Molecules Hide
List View of Archive By: Title | Date | Category
Structural View of Biology
Infrastructure & Communication
Molecule of the Month:
Nitric Oxide Synthase
Nitroglycerin is a powerful explosive, detonating when exposed to heat or pressure. The same molecule, however, can save your life if you're experiencing a heart attack. A small dose of nitroglycerin will slowly break down and release nitric oxide (NO), which then spreads to the muscle cells surrounding blood vessels, telling them to relax.
Full Article...
Protein Structure Initiative Featured Molecule:
CXCR4
PSI researchers have revealed the structure of CXCR4, a central chemokine receptor in cancer and HIV infection.
Full Article | PSI Featured Molecule Archive | PSI Structural Biology Knowledgebase
Latest Structures Hide
3q08 - Crystal Structure of Chlorite Dismutase from D. Aromatica at pH 6.5
Goblirsch, B.R., Wilmot, C.M.
Structural features promoting dioxygen production by Dechloromonas aromatica chlorite dismutase.
New Features Hide
Ligand Download Page
Latest features released:
Website Release Archive:
RCSB PDB News Hide
Weekly | Quarterly | Yearly
2011-01-04
Structural Biology Knowledgebase Widget
A new widget loads SBKB data about available models, targets, biological annotations, and more for each PDB entry. more...
Happy Holidays
Deposition Session Restart IDs
Annual Report Published
wwPDB News Hide
Statement on Retraction of PDB Entries
2010-12-21
From 7 to 70,000: The PDB Reaches a New Milestone
2010-12-21
Advanced Notification
Advisory Committee Meeting
Full wwPDB News
FTP Archives Hide
Current PDB FTP Archive:
Done

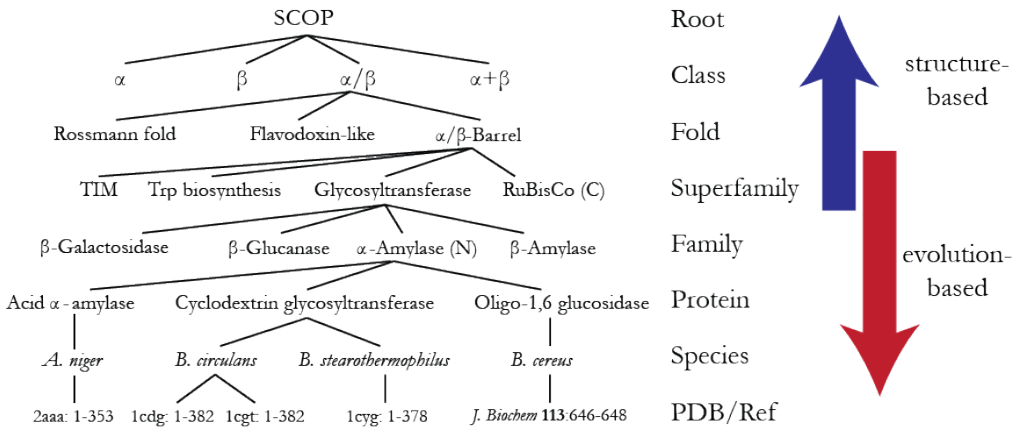
The wwPDB

- In recent years, the major database for macromolecular structures is the worldwide PDB (wwPDB) at `http://www.wwpdb.org/`.
- It is a joint effort of the RCSB, the Protein Data Bank Europe (at the European Bioinformatics Institute, EBI), the Protein Databank Japan (based at Osaka University), and the Biological Magnetic Resonance Data Bank (BMRB).

The PDB File Format

```
                                Chain name
    Amino Acid                /   Sequence Number
                \           /   /
     Element     \         /   /      -----Coordinates-----
                  \   \   /   /       X       Y       Z      (etc.)
ATOM      1  N   ASP L   1        4.060   7.307   5.186  ...
ATOM      2  CA  ASP L   1        4.042   7.776   6.553  ...
ATOM      3  C   ASP L   1        2.668   8.426   6.644  ...
ATOM      4  O   ASP L   1        1.987   8.438   5.606  ...
ATOM      5  CB  ASP L   1        5.090   8.827   6.797  ...
ATOM      6  CG  ASP L   1        6.338   8.761   5.929  ...
ATOM      7  OD1 ASP L   1        6.576   9.758   5.241  ...
ATOM      8  OD2 ASP L   1        7.065   7.759   5.948  ...
                \\
                 Element position within amino acid
```

Classification of Protein Structures - The SCOP Database

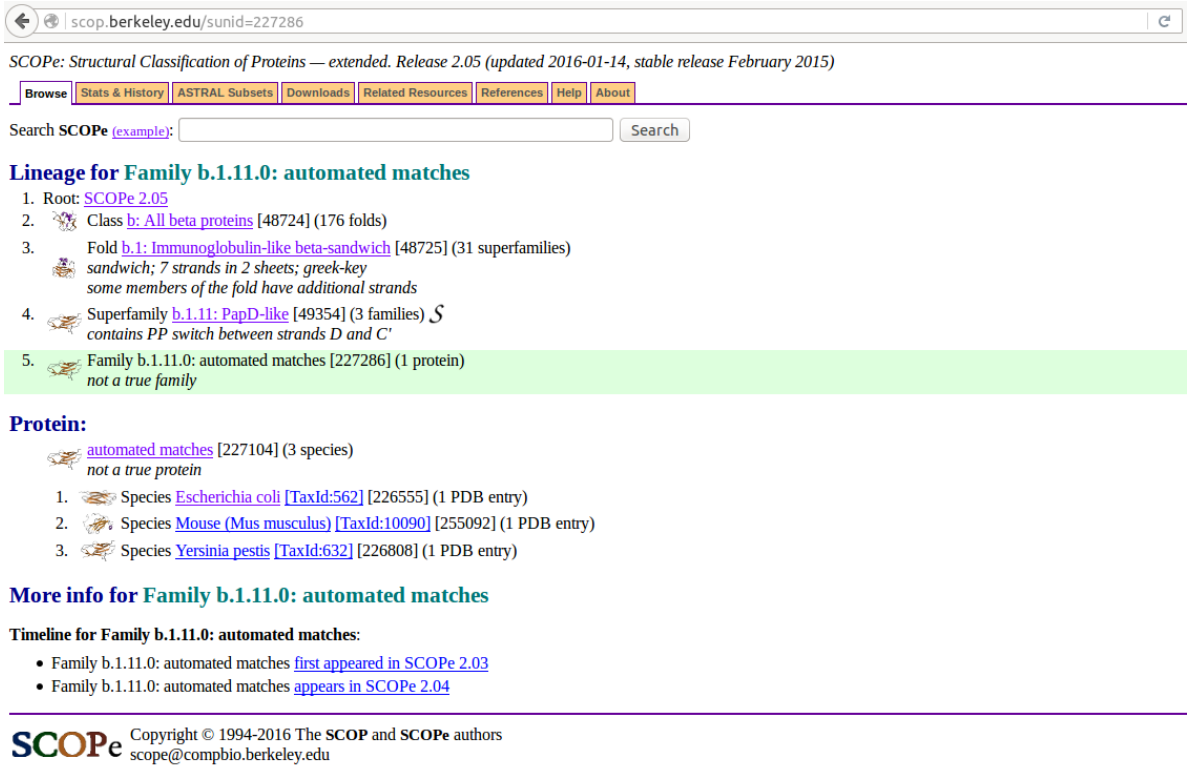


Chothia, Murzin (Cambridge)

Hand-curated hierarchical taxonomy of proteins based on their structural and evolutionary relationships.

- Classes
- Fold Level
- Super Family
- Family
- Domain

The SCOPe Database



scop.berkeley.edu/sunid=227286

SCOPe: Structural Classification of Proteins — extended. Release 2.05 (updated 2016-01-14, stable release February 2015)

Browse | Stats & History | ASTRAL Subsets | Downloads | Related Resources | References | Help | About

Search **SCOPe** (example): Search

Lineage for Family b.1.11.0: automated matches

1. Root: SCOPe 2.05
2. Class b: All beta proteins [48724] (176 folds)
3. Fold b.1: Immunoglobulin-like beta-sandwich [48725] (31 superfamilies)
   sandwich; 7 strands in 2 sheets; greek-key
   some members of the fold have additional strands
4. Superfamily b.1.11: PapD-like [49354] (3 families) $\mathcal{S}$
   contains PP switch between strands D and C'
5. Family b.1.11.0: automated matches [227286] (1 protein)
   not a true family

Protein:

automated matches [227104] (3 species)
not a true protein

1. Species Escherichia coli [TaxId:562] [226555] (1 PDB entry)
2. Species Mouse (Mus musculus) [TaxId:10090] [255092] (1 PDB entry)
3. Species Yersinia pestis [TaxId:632] [226808] (1 PDB entry)

More info for Family b.1.11.0: automated matches

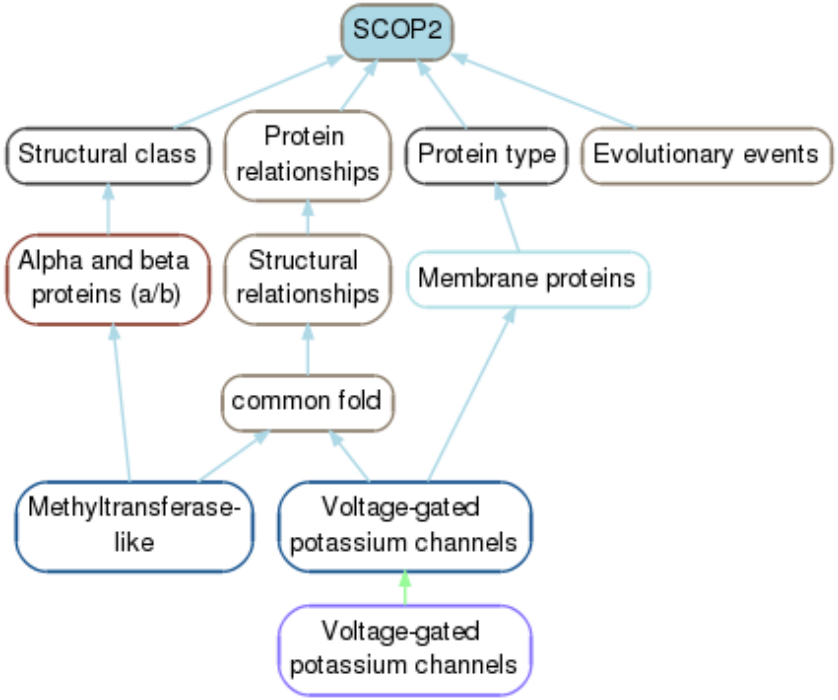
Timeline for Family b.1.11.0: automated matches:

- Family b.1.11.0: automated matches first appeared in SCOPe 2.03
- Family b.1.11.0: automated matches appears in SCOPe 2.04

SCOPe Copyright © 1994-2016 The **SCOP** and **SCOPe** authors
scope@compbio.berkeley.edu

- The successor of SCOP (which is no longer maintained/updated).
- Rather similar, combination of hand-curated and automated methods.

The SCOP2 Database Prototype



- Similar to SCOP(e), but different.
- Adding evolutionary events and protein types among others.
- Several new hierarchical categories.
- The evolutionary relationships induce a graph-like structure rather than rigid hierarchy.

The CATH Database

- Another database which classifies protein structures downloaded from the Protein Data Bank.
- It is a semi-automatic, hierarchical classification of protein domains initially published in 1997.
- CATH is an acronym of the four main levels in the classification.

#	Level	Description
1	**C**lass	Overall secondary-structure content of the domain. (Equivalent to SCOP class)
2	**A**rchitecture	High structural similarity but no evidence of homology. (Equivalent to SCOP fold)
3	**T**opology	A large-scale grouping of topologies which share particular structural features
4	**H**omologous superfamily	Indicative of a demonstrable evolutionary relationship. (Equivalent to SCOP superfamily)

The CATH Database

- Much of the work is done by automatic methods, however there are important manual elements to the classification.
- First – separate the proteins into domains. It is difficult to produce an unequivocal definition of a domain and this is one area in which CATH and SCOP differ.
- The domains are automatically sorted into classes and clustered on the basis of sequence similarities.
- These groups form the **H** levels of the classification. The topology level is formed by structural comparisons of the homologous groups.
- Finally, the **A**rchitecture level is assigned manually.

The CATH Database

Class Level classification is done on the basis of 4 criteria:

1. Secondary structure content;
2. Secondary structure contacts;
3. Secondary structure alternation score; and
4. Percentage of parallel strands.

CATH defines four classes: mostly-α, mostly-β, α and β, few secondary structures.