Assorted Protein Databases ...

SWISSPROT (UniProtKB/swissprot) – database of protein sequences

Advantage: highly curated (i.e. carefully annotated)

Disadvantage: incomplete due to the careful annotation. However, a complementary "what's missing in swiss-prot" database is also available. This complementary database is known as trEMBL ("translated EMBL", or UniProtKB/TrEmbl)

For SWISSPROT and TrEMBL, see http://www.uniprot.org

PDB – Protein Data Bank (contains experimentally determined protein structures)

see http://www.rcsb.org

CATH – a database that classifies experimentally determined protein domain structures that are found in PDB

(following descriptions taken verbatim from previous or current versions of CATH web page ... or were distilled from these web pages)

"Domains are regions of contiguous polypeptide chain that have been described as compact, local, and semi-independent units"

Class – Architecture – Topology – Homologous Superfamily

Class – derived from secondary structure content of domain

Classes are: "mainly alpha", "mainly beta", "alpha-beta", "few secondary structures"

Architecture: "overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures. ... currently assigned manually ..."

Topology (Fold family): more narrow description of structure, proteins grouped together within same topology by an automatic clustering method

Homologous Superfamily: proteins thought to be evolutionarily related, grouped together if they are pretty similar according to sequence and/or structure.

Sequence families: "Domains clustered in the same sequence families have [high] sequence identities ... indicating highly similar structures and functions."

see http://www.cathdb.info/

SCOP – Structural Classification of Proteins

another hierarchical classification of proteins, classified mainly by manual procedures

Organization of database (quoted from scop web pages)

"1. Family: Clear evolutionarily relationship

Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, in some cases similar functions and structures provide definitive evidence of common descent in the absense of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%.

2. Superfamily: Probable common evolutionary origin

Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies. For example, actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily.

3. Fold: Major structural similarity

Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

Classes of scop:

1. All alpha proteins

2. All beta proteins

3. Alpha and beta proteins (a/b) Mainly parallel beta sheets (beta-alpha-beta units)

4. Alpha and beta proteins (a+b) Mainly antiparallel beta sheets (segregated alpha and beta regions)

5. Multi-domain proteins (alpha and beta). Folds consisting of two or more domains belonging to different classes

6.Membrane and cell surface proteins and peptides. Does not include proteins in the immune system

7.Small proteins. Usually dominated by metal ligand, heme, and/or disulfide bridges

8.Coiled coil proteins.

9.Low resolution protein structures.

10.Peptides. Peptides and fragments

11.Designed proteins. Experimental structures of proteins with essentially non-natural sequences

see http://scop.mrc-lmb.cam.ac.uk/scop/
see also http://scop2.mrc-lmb.cam.ac.uk

HOMSTRAD – Homologous Structure Alignment Database

database of alignments in which protein structures are experimentally determined for all sequences

http://mizuguchilab.org/homstrad/

PANDIT : Protein and Associated Nucleotide Domains with Inferred Trees

database of DNA sequences for PFAM families, along with sequence alignments and inferred evolutionary trees

http://www.ebi.ac.uk/research/goldman/software/pandit

Year	Genbank bases	Swiss-Prot	PDB	Family	SuperFam	Fold
1972	0	0	0	1	1	1
1974	0	0	12	4	4	4
1976	0	0	60	19	18	18
1978	0	0	124	23	22	21
1980	0	0	185	34	32	29
1982	6.8E+05	0	258	55	52	45
1984	3.7E+06	0	337	65	60	50
1986	9.6E+06	4160	396	88	80	68
1988	2.5E+07	8702	503	126	112	93
1990	5.1E+07	18364	754	204	179	141
1992	1.2E+08	28154	1187	349	265	218
1994	2.3E+08	40292	3428	603	438	315
1996	7.3E+08	59021	5654	810	569	397
2000	1.1E+10	85785	10650	1296	820	548
2002	2.9E+10	115106	18948	1827	1073	686
2004	4.5E+10	163235	27969	2327	1294	802
2007	7.7E+10	359942	49760	3464	1777	1086
2010	1.2E+11	514212	62787	3902	1962	1195
2012	1.5E+11	534242	78867		2738	1375
2015		547357	106082		2738	1375
2017	2.3E+11	553474	126447		2737	1375
2021	7.8E+11	564277	174994			

(adapted from Jotun Hein/Ole Lund)