**BINF 731**

# Protein Structure Analysis

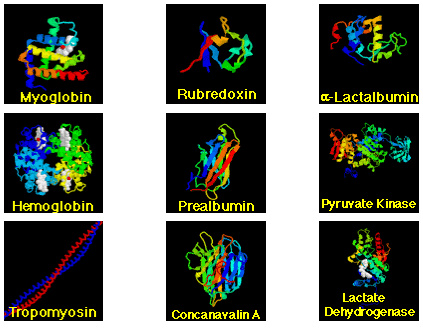
Iosif Vaisman

2013

## Secondary Structure: Computational Problems

Secondary structure characterization
Secondary structure assignment
Secondary structure prediction
Protein structure classification

## Structural classes of proteins




all α  all β  α/β

## Protein Structure Classification

SCOP - Structural Classification of Proteins

FSSP - Fold classification based on Structure-Structure alignment of Proteins

CATH - Class, architecture, topology and homologous superfamily

## SCOP: Structural Classification of Proteins

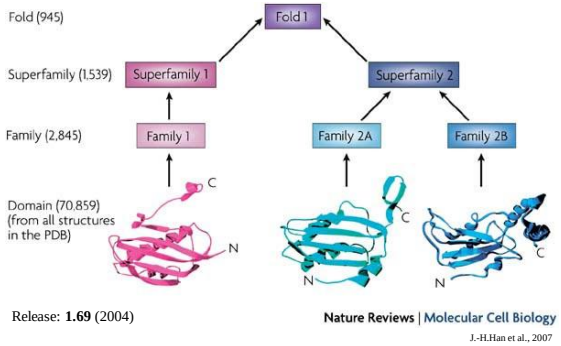
Current release: 1.75
38221 PDB Entries (June 2009). 110800 Domains.

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

The **SCOP** database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy; the principal levels are family, superfamily and fold

**Family**: *Clear evolutionarily relationship*
**Superfamily**: *Probable common evolutionary origin*
**Fold**: *Major structural similarity*

## SCOP: Structural Classification of Proteins



## SCOP: Structural Classification of Proteins

**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%.

## SCOP: Structural Classification of Proteins

**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.

## SCOP: Structural Classification of Proteins

**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.

## SCOP Statistics (2003)

| **Class** | **Folds** | **Super families** | **Families** |
|---|---|---|---|
| All alpha proteins | 179 | 299 | 480 |
| All beta proteins | 126 | 248 | 462 |
| Alpha and beta proteins (a/b) | 121 | 199 | 542 |
| Alpha and beta proteins (a+b) | 234 | 349 | 567 |
| Multi-domain proteins | 38 | 38 | 53 |
| Membrane and cell surface proteins | 36 | 66 | 73 |
| Small proteins | 66 | 95 | 150 |
| **Total** | **800** | **1294** | **2327** |

## SCOP Statistics (2009 – current release)

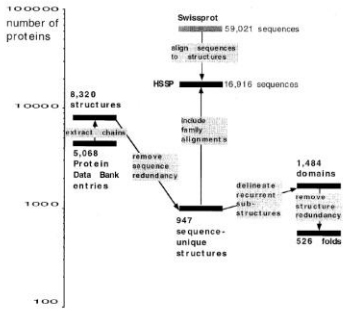
| **Class** | **Number of folds** | **Number of superfamilies** | **Number of families** |
|---|---|---|---|
| All alpha proteins | 284 | 507 | 871 |
| All beta proteins | 174 | 354 | 742 |
| Alpha/beta proteins (a/b) | 147 | 244 | 803 |
| Alpha+beta proteins (a+b) | 376 | 552 | 1055 |
| Multi-domain proteins | 66 | 66 | 89 |
| Membrane and cell surface proteins | 58 | 110 | 123 |
| Small proteins | 90 | 129 | 219 |
| **Total** | **1195** | **1962** | **3902** |

## FSSP (DALI) Database

Current release: April 2009

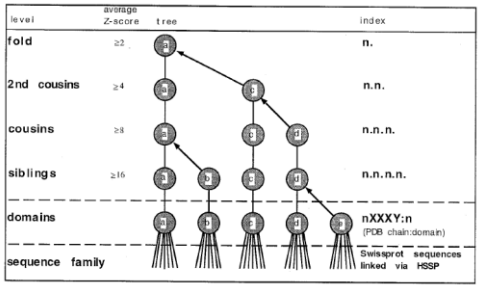
The FSSP database is based on exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB). The classification and alignments are automatically maintained and continuously updated using the Dali search engine.

## Structure processing for Dali/FSSP




Adopted from Holm and Sander, 1998

## Hierarchical clustering of folds in Dali/FSSP




Adopted from Holm and Sander, 1998

## Dali Domain Dictionary

http://www2.ebi.ac.uk/dali/domain

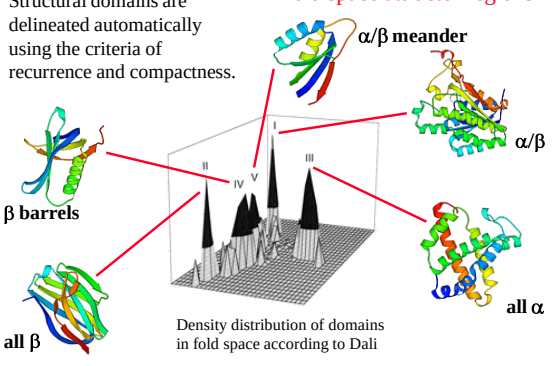
Structural domains are delineated automatically using the criteria of recurrence and compactness. Each domain is assigned a Domain Classification number DC_l_m_n_p , where:

- l - fold space attractor region
- m - globular folding topology
- n - functional family
- p - sequence family

## Dali Domain Dictionary

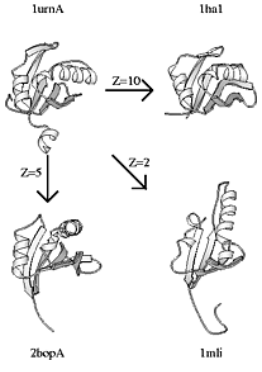
Structural domains are delineated automatically using the criteria of recurrence and compactness.

Fold space attractor regions




Density distribution of domains in fold space according to Dali

## Dali Domain Dictionary




Fold types

Fold types are defined as clusters of structural neighbors in fold space with average pairwise Z-scores (by Dali) above 2.

Structural neighbours of 1urnA (top left). 1mli (bottom right) has the same topology even though there are shifts in the relative orientation of secondary structure elements

## Dali Domain Dictionary

**Functional families**

The third level of the classification infers plausible evolutionary relationships from strong structural similarities which are accompanied by functional or sequence similarities. Functional families are branches of the fold dendrogram where all pairs have a high average neural network prediction for being homologous. The neural network weighs evidence coming from: overlapping sequence neighbours as detected by PSI-Blast, clusters of identically conserved functional residues, E.C. numbers, Swissprot keywords.

## Dali Domain Dictionary

**Sequence families**

The fourth level of the classification is a representative subset of the Protein Data Bank extracted using a 25 % sequence identity threshold. All-against-all structure comparison was carried out within the set of representatives. Homologues are only shown aligned to their representative.

## CATH - Protein Structure Classification

Current release: 3.5 (September 20, 2011)

http://www.cathdb.info

CATH is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels:

- **C**lass
- **A**rchitecture
- **T**opology
- **H**omologous superfamily

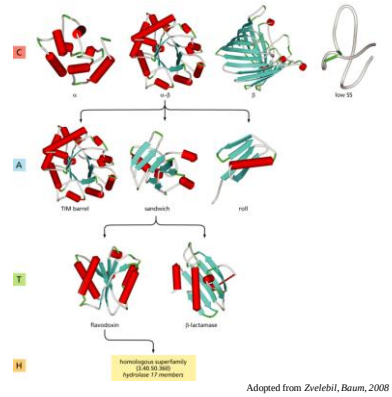
## CATH - Protein Structure Classification

**C**lass

**A**rchitecture

**T**opology

**H**omologous superfamily




Adopted from *Zvelebil, Baum, 2008*

## CATH - Protein Structure Classification

Class, C-level

Class is determined according to the secondary structure composition and packing within the structure. It can be assigned automatically (90% of the known structures) and manually.

Three major classes:

- mainly-alpha
- mainly-beta
- alpha-beta (alpha/beta and alpha+beta)

A fourth class is also identified which contains protein domains which have low secondary structure content.

## CATH - Protein Structure Classification

Architecture, A-level

This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures.

It is currently assigned manually using a simple description of the secondary structure arrangement e.g. barrel or 3-layer sandwich. Reference is made to the literature for well-known architectures (e.g the beta-propellor or alpha four helix bundle).

Procedures are being developed for automating this step.

## CATH - Protein Structure Classification

Topology (Fold family), T-level

Structures are grouped into fold families at this level depending on both the overall shape and connectivity of the secondary structures. This is done using the structure comparison algorithm SSAP.

Some fold families are very highly populated and are currently subdivided using a higher cutoff on the SSAP score.

# CATH - Protein Structure Classification

Homologous Superfamily, H-level

This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are identified first by sequence comparisons and subsequently by structure comparison using SSAP.

Structures are clustered into the same homologous superfamily if they satisfy one of the following criteria:

- Sequence identity >= 35%, 60% of larger structure equivalent to smaller
- SSAP score >= 80.0 and sequence identity >= 20% 60% of larger structure equivalent to smaller
- SSAP score >= 80.0, 60% of larger structure equivalent to smaller, and domains which have related functions

# CATH - Protein Structure Classification

Sequence families, S-level

Structures within each H-level are further clustered on sequence identity. Domains clustered in the same sequence families have sequence identities >35% (with at least 60% of the larger domain equivalent to the smaller), indicating highly similar structures and functions.

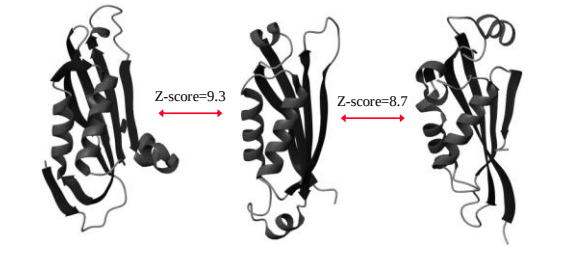
# CATH Statistics

**Version** 3.5 **Date** 9-20-2011

| | |
|---|---|
| **Number of Domains** | 173,536 |
| **Number of Superfamilies** | 2,626 |
| **Number of PDBs** | 51,334 |

| **C** | **A** | **T** | **H** | **S** | **O** | **L** | **I** | **D** |
|---|---|---|---|---|---|---|---|---|
| Mainly Alpha | 5 | 305 | 652 | 1850 | 2329 | 3001 | 5587 | 19729 |
| Mainly Beta | 20 | 191 | 415 | 1860 | 2531 | 3846 | 6503 | 25537 |
| Alpha Beta | 14 | 496 | 922 | 3922 | 5303 | 6659 | 12998 | 47193 |
| Few Sec Struct | 1 | 92 | 102 | 162 | 200 | 275 | 403 | 1426 |
| **Total** | **40** | **1084** | **2091** | **7794** | **10363** | **13781** | **25491** | **93885** |

# Classification limitations



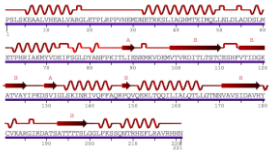

1a8rA, αβ three-layer (ββα) sandwich (CATH 3.50.11)

1dhn, αβ three-layer (ββα) sandwich (CATH 3.50.11)

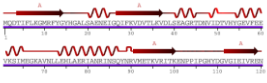
1b66A, αβ two-layer sandwich (CATH 3.30.479)

Adopted from Getz et al., 2002

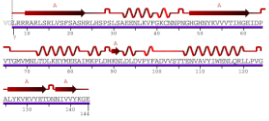
# Classification limitations



1a8rA

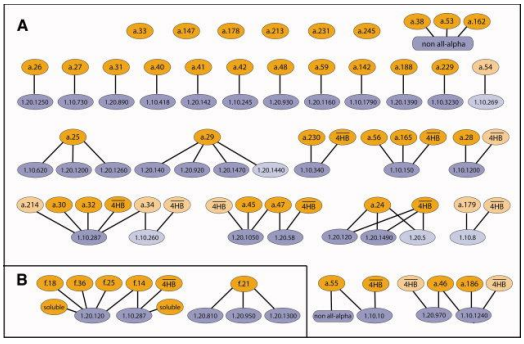


1dhn



1b66A

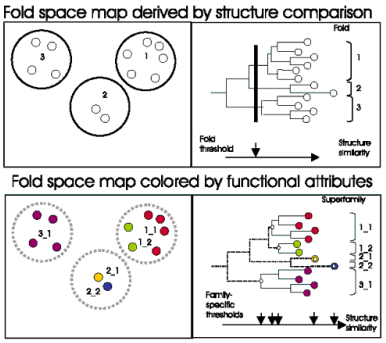
# Classification limitations




Relationships between four helix bundle folds in SCOP (orange) and CATH (blue).

S. Neumann et al., 2010

# Homologous and Analogous Proteins




Adopted from Dietmann & Holm, 2001

# Homologous and Analogous Proteins

- Homologous: same fold, same or similar function, common ancestry.
- Analogous: same fold, different function, ancestral origin unknown.