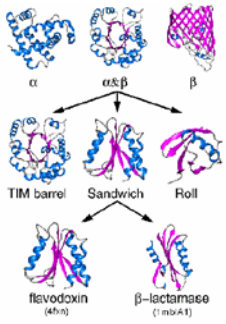
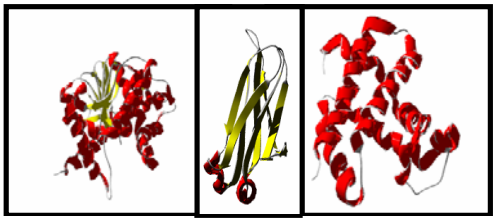
Structural Classification

- Protein structure classification is important because it organizes the protein structure universe that is **independent** of sequence similarity.
- A global picture of the protein universe will help us to understand the functions of proteins better.
- Structures are classified usually by their secondary structure content and how they are arranged.



Classifying Protein Structures



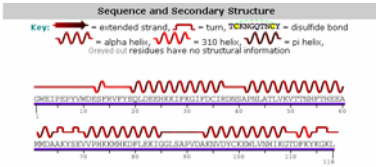
Lactate Dehydrogenase SCOP Class: Mixed α / β

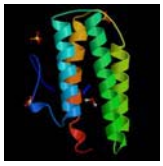
Immunoglobulin SCOP Class: β

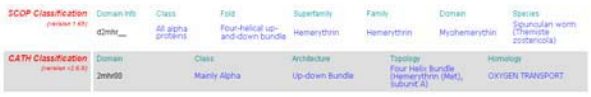
Hemoglobin Beta Chain SCOP Class: α

Classifying Protein Structures

2MHR







Domains

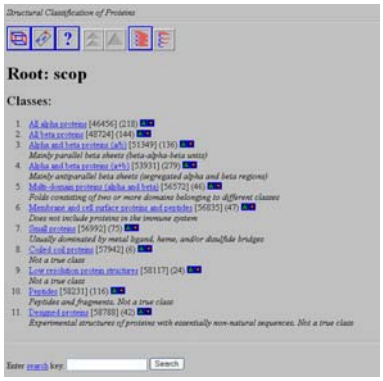
- proteins can be **modular**
- single chain may be divisible into smaller independent units of tertiary structure called **domains**
- domains are the basic unit of structure classification
- different domains in a protein are also often associated with different functions carried out by the protein.
- domains are usually the main units for structural classification

Definition of domain

- "A polypeptide or part of a polypeptide chain that can independently fold into a stable tertiary structure..."
  from Introduction to Protein Structure, by Branden & Tooze
- "Compact units within the folding pattern of a single chain that **look as if** they should have independent stability."
  from Introduction to Protein Architecture, by Lesk

SCOP Database

http://scop.berkeley.edu

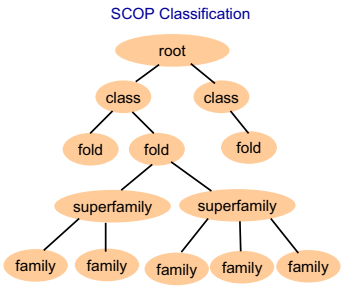


SCOP Hierarchy

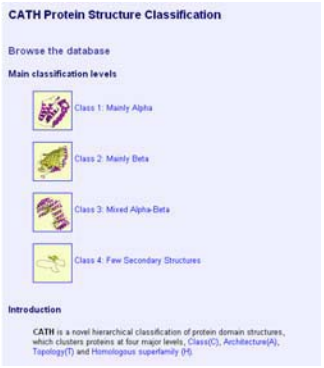
- Class: derived from secondary structure content
- Fold: derived from topological connection, orientation, arrangement and number of secondary structures
- Superfamily: clusters of low sequence similarity but related structures & functions
- Family: clusers of proteins with sequence similarity > 30% with similar structure & function

SCOP Classification

- A hierarchical classification scheme is maintained



CATH Database



http://cathwww.biochem.ucl.ac.uk/latest/index.html

CATH

- Class [C] derived from secondary structure content (automatic)
- Architecture [A] derived from orientation of 2° structures (manual)
- Topology [T] derived from topological connection and # 2° structures
- Homologous Superfamily (H) clusters of similar structures & functions

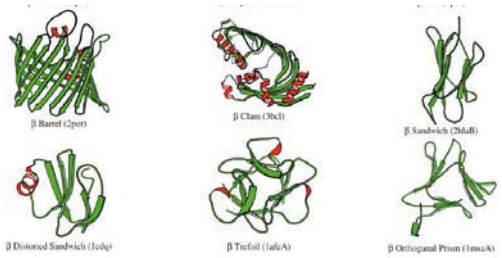
Class

- In the CATH hierarchy, Class simply describes what type of secondary structure is present.
- There are only four classes:
  - mainly α
  - mainly β
  - α & β
  - few secondary structures
- 90% of structures are trivial to assign at this level.

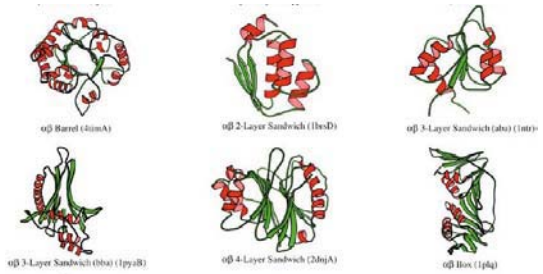
Architecture

- Architecture is hard to define precisely
- In CATH it is defined broadly as describing "general features of protein shape" such as arrangements of secondary structure in 3D space
- It does *not* define connectivities between secondary structural elements--that's what the topology level does. It does not even explicitly define directionality of secondary structure, e.g. parallel or antiparallel beta-sheets.
- in CATH, architectures are presently assigned manually, by visual inspection.

Some mostly beta architectures



Some mixed alpha-beta architectures

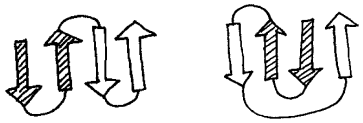


Topology (Fold)

- if two proteins have the same topology, it means they have the same number and arrangement of secondary structures, and the connectivities between these elements are the same.
- this is also sometimes called the fold of a protein.
- in CATH, automated structure alignment is used to group proteins according to topology.

Topology: differences in connectivity

• example: a four-stranded antiparallel beta-sheet can have many different topologies based on the order in which the four beta-strands are connected.



"up-and-down" **"greek key"**

Homologous superfamily/ Sequence family

- The lowest two levels in the CATH hierarchy relate to *common ancestry*
- some, but not all proteins with the same fold show evidence of common ancestry
- the surest way of identifying common ancestry is that two proteins have sequences roughly >30% identical (sequence family level)
- if protein sequences are not that similar, common ancestry may still be inferred on the basis of a combination of structural and functional similarity, and possibly weak sequence similarity (homologous superfamily level)

Comparison of SCOP and CATH Hierarchies

SCOP	CATH
class	class
	architecture
fold	topology
	homologous superfamily
superfamily	⋮
family	sequence family
domain	domain

CATH more directed toward structural classification, SCOP pays more attention to evolutionary relationships

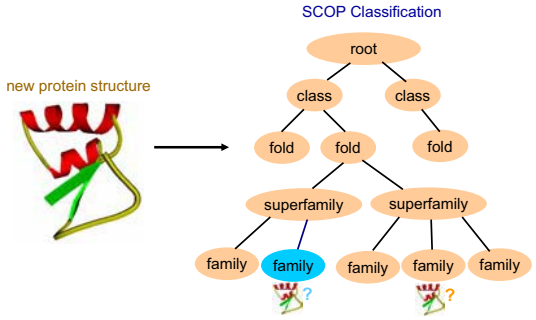
Another SCOP/CATH difference

- in CATH, there is one class to represent mixed alpha-beta
- in SCOP there are two:
  - α/β: beta structure is largely *parallel*, made of βαβ motifs
  - α+β: alpha and beta structure segregated to different parts of structure

SCOP and CATH

- they have in common that they are hierarchical and based on abstractions
- they both include some *manual* aspects and are *curated* by experts in the field of protein structure
- *fully automated* methods for structure classification/comparison?

Problem definition

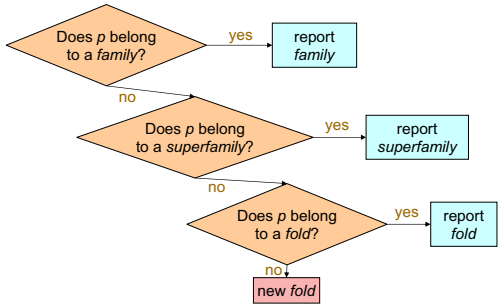


Two problems

- Class membership?
  - Does the query protein belong to a SCOP category? Or does it need a new category to be defined?
  - Binary classification problem:
    - *member*, *non-member*
- Class label assignment?
  - What SCOP category is the query protein assigned to?
  - Multi-class classification problem

Hierarchical classification

- Let *p* be a protein structure, proceed bottom-up from *family* level to *fold* level:



Method

- A consensus classification scheme in which the decisions of a number of individual (component) classifiers are merged together in an intelligent way.
  - Decision Tree Classification is one technique to use as the “intelligent way”

Component classifiers

- Using a sequence/structure comparison tool as a classifier
  - Perform a nearest neighbor query:
    if *similarityScore*(query, NN) < *trained cutoff*
    then not a member of any category
    else member of *class*(NN)
- Comparison tools that we have used:
  Sequence: PSI-Blast, HMMER+SUPERFAMILY database
  Structure: CE, Dali, Vast

Performance of component classifiers

- Database: SCOP 1.59
- Query: SCOP 1.61 – SCOP 1.59

Class membership

	HMM	BLAST	CE	Dali	Vast	At least one
family	**94.5%**	**92.6%**	**89%**	**89%**	**89%**	**98.2%**
superfamily	**78.6%**	**66.1%**	**72.2%**	**77.6%**	**78.4%**	**96%**
fold	**73%**	**60.7%**	**78.5%**	**82%**	**85%**	**100%**

Performance of component classifiers

- Database: SCOP 1.59
- Query: SCOP 1.61 – SCOP 1.59

Class label assignment

	HMM	BLAST	CE	Dali	Vast	At least one
family	**94.8%**	**92.3%**	**91%**	**88%**	**92%**	**97.9%**
superfamily	**69%**	**12%**	**81%**	**80.4%**	**81.7%**	**93.9%**
fold	**40.5%**	**0%**	**40.5%**	**46%**	**54%**	**64.9%**

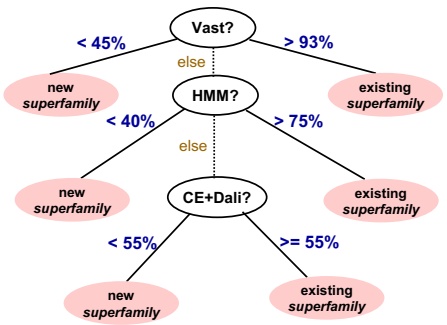
Normalization of similarity scores

- Universal confidence levels instead of tool-specific scores
- Perform nearest neighbor queries
  - Database: SCOP 1.59
  - Query: SCOP 1.61 – SCOP 1.59
- Partition score space of tools into confidence levels
  - e.g. CE *z-score* of 5.4 → we are 80% confident that the query protein is a member of an existing fold.

Consensus Decision

- Each component classifier reports a confidence level for the query protein:
  - $c = [C_1, C_2, C_3, C_4, C_5]$
- What is the best way to combine these probabilistic decisions?
  - A solution: decision trees.

decision tree: *superfamily* level

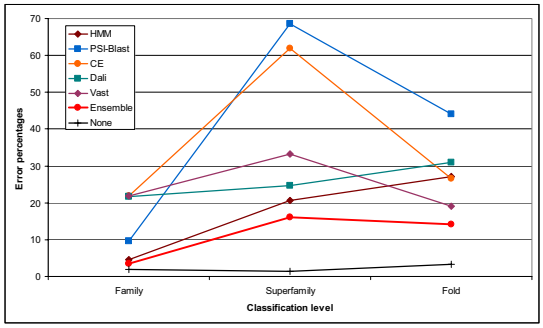


Experimental evaluation

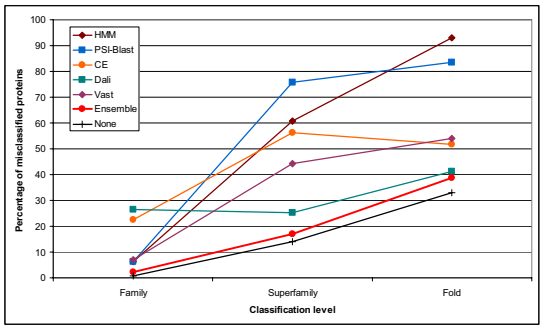
- The dataset:

	Training	Evaluation
Database	**v1.59 (*20449*)**	**v1.61 (*22724*)**
Query	**v1.61 – v1.59 (*2241*)**	**v1.63 – v1.59 (*2825*)**
new *family*	***248***	***618***
new *superfamily*	***84***	***424***
new *fold*	***47***	***339***

Test results: class membership



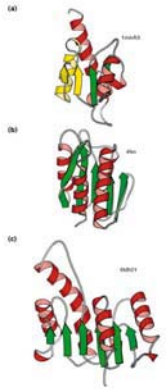
Test results: class label assignment



Difficult cases

Russian doll effect:

- **A small structure is contained exactly in a larger structure, but they are classified in different groups in SCOP hierarchy.**



Conclusions

- Automated classification of a protein structure is feasible.
- More accurate classifiers can be built by combining less accurate individual classifiers.
- The presented classification technique can also be used as an aid to manual classification by providing clues.

Other Servers/Databases

- Dali - http://www.ebi.ac.uk/dali/
- VAST - http://ww.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
- FSSP - http://www.ebi.ac.uk/dali/fssp/fssp.html
- PDBsum - http://ww.biochem.ucl.ac.uk/bsm/pdbsum/

Discovery of New Folds

- structural taxonomy reveals that although structures are being solved more rapidly than ever, fewer and fewer of them have new folds! Will we get them all soon?