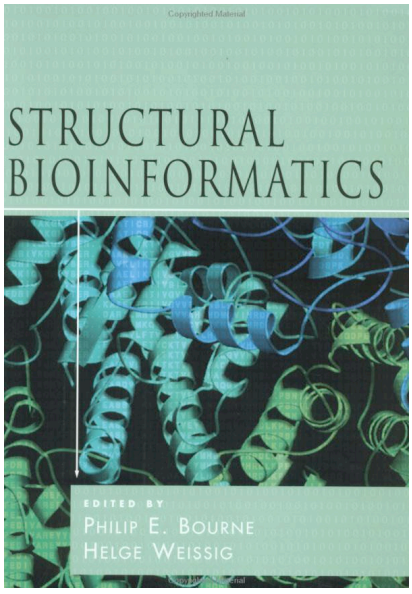
Protein Structure:
Data Bases and Classification

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

A Foine Reference



Bourne and Weissig

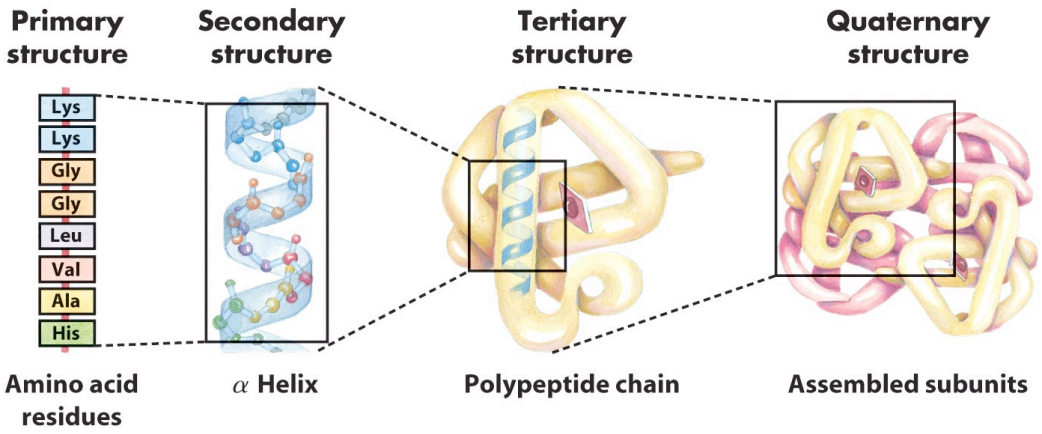
Structural Bioinformatics

Wiley, 2003

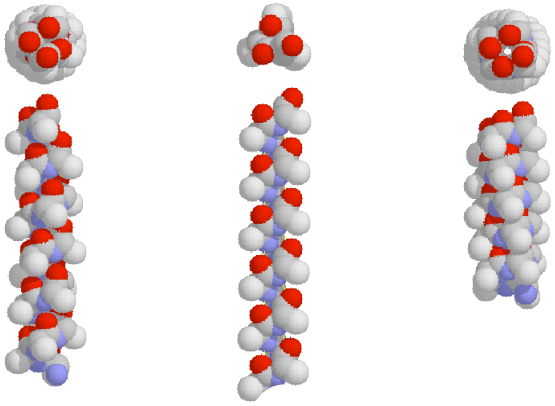
Terminology

- Primary Structure
- Secondary Structure
- Tertiary Structure
- Quatenary Structure
- Supersecondary Structure
- Domain
- Fold

Hierarchy of Protein Structure

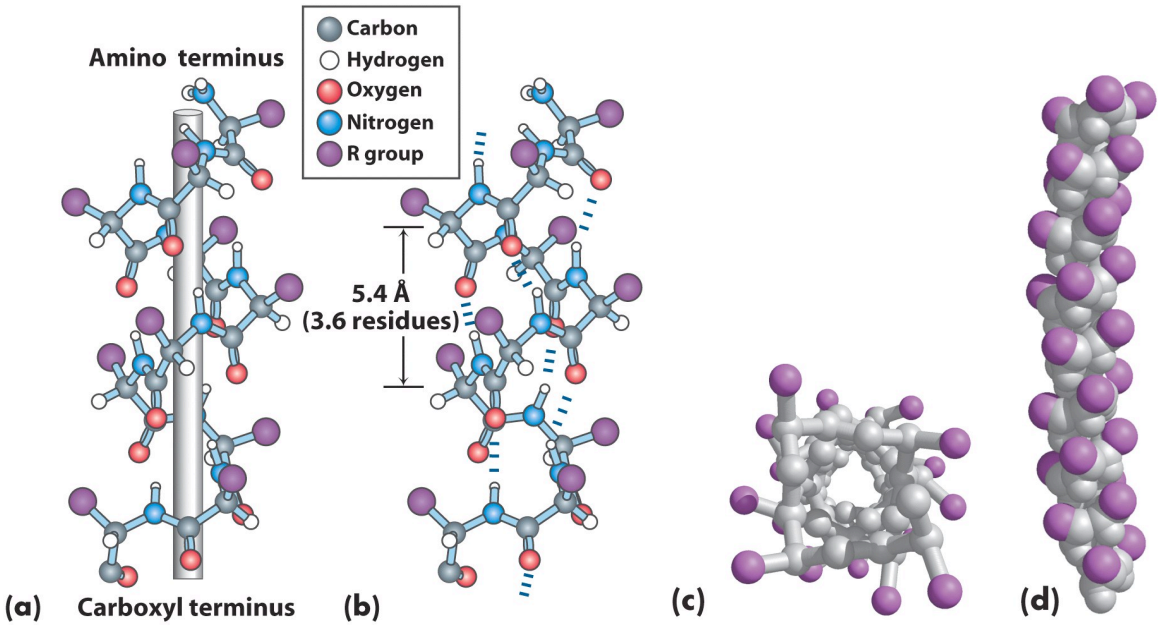


Helices



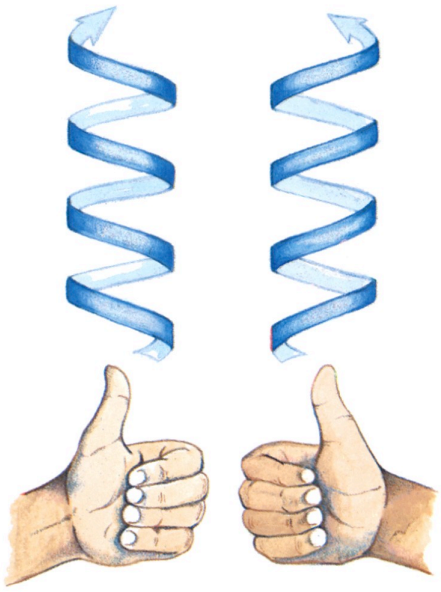
	α	3.10	π
Amino acids/turn:	3.6	3.0	4.4
Frequency	~97%	~3%	rare
H-bonding	*i*, *i*+4	*i*, *i*+3	*i*, *i*+5

α-helices

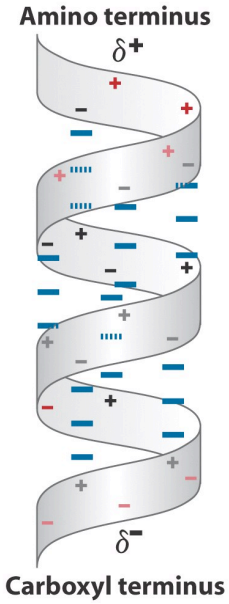


α-helices

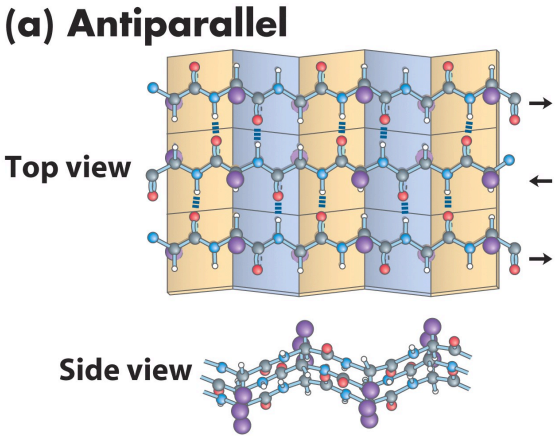
α-helices have handedness:

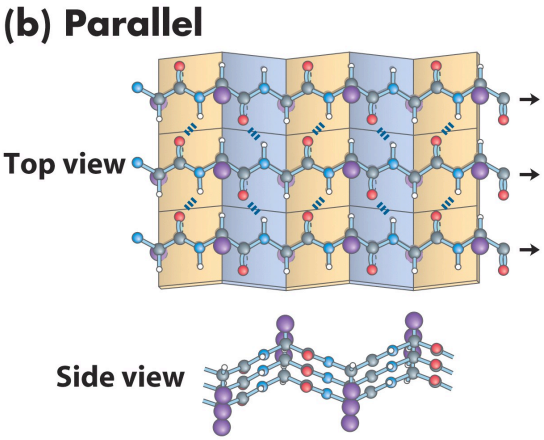


α-helices have a dipole:

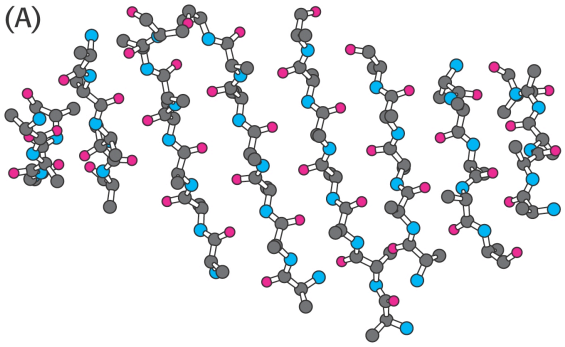


β-sheets

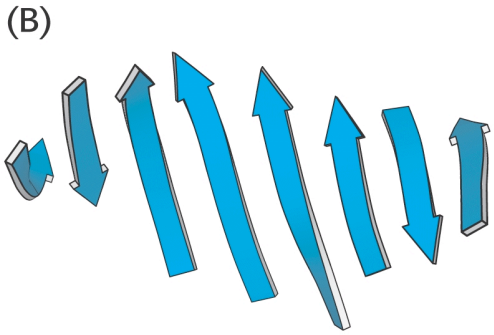


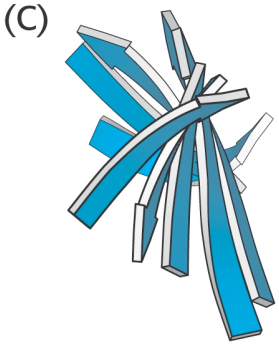


β-sheets

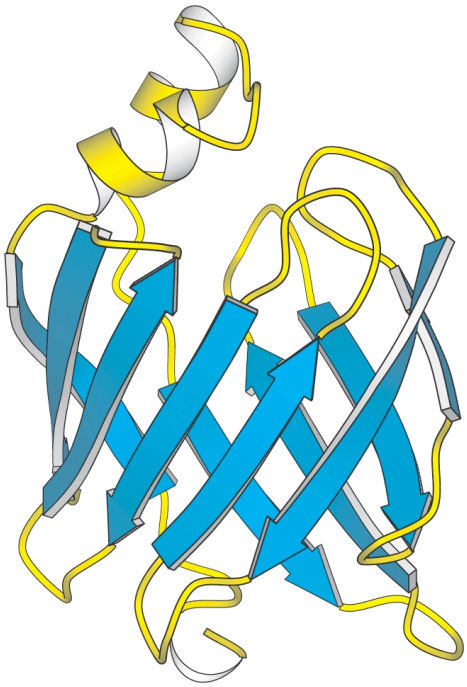


Have a right-handed twist!



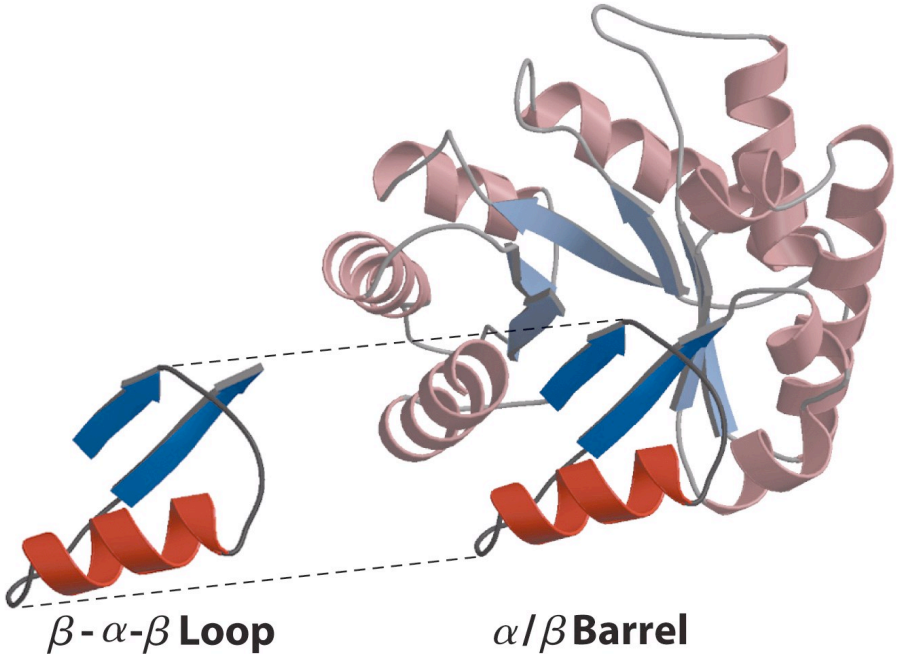


β-sheets

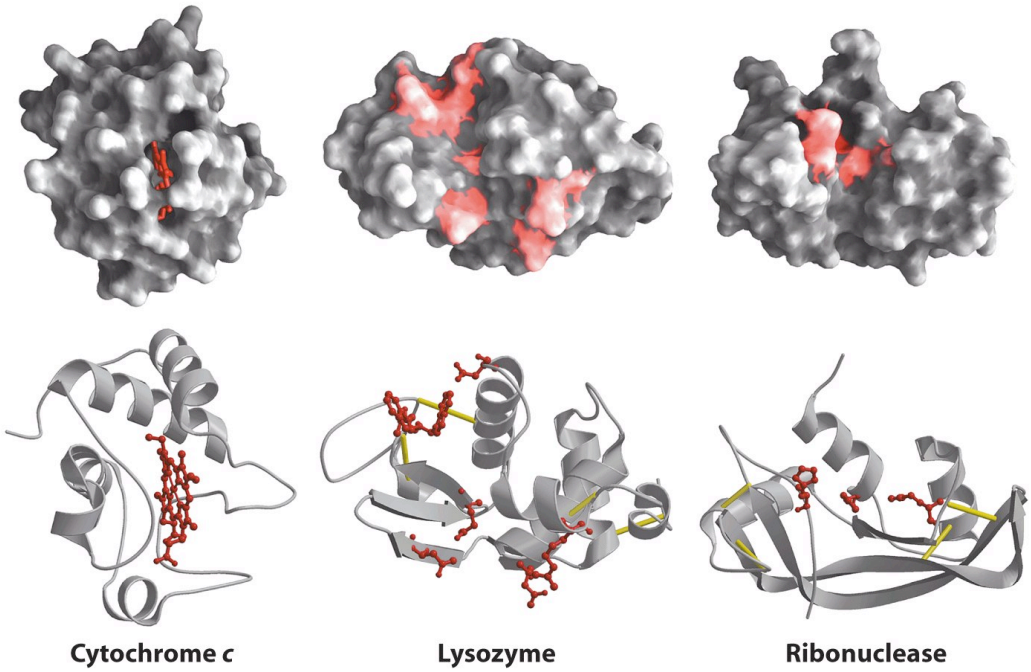


Can form higher level structures!

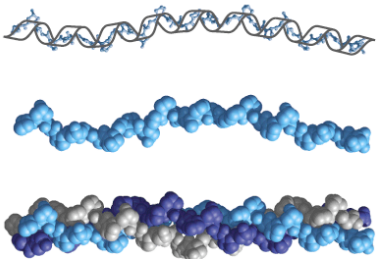
Super Secondary Structure Motifs

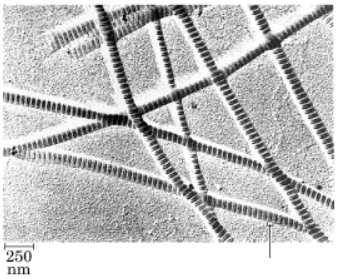


Protein Structure and Function

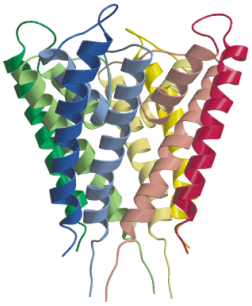


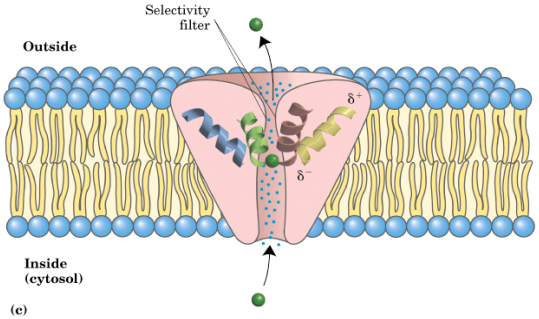
Structural Proteins



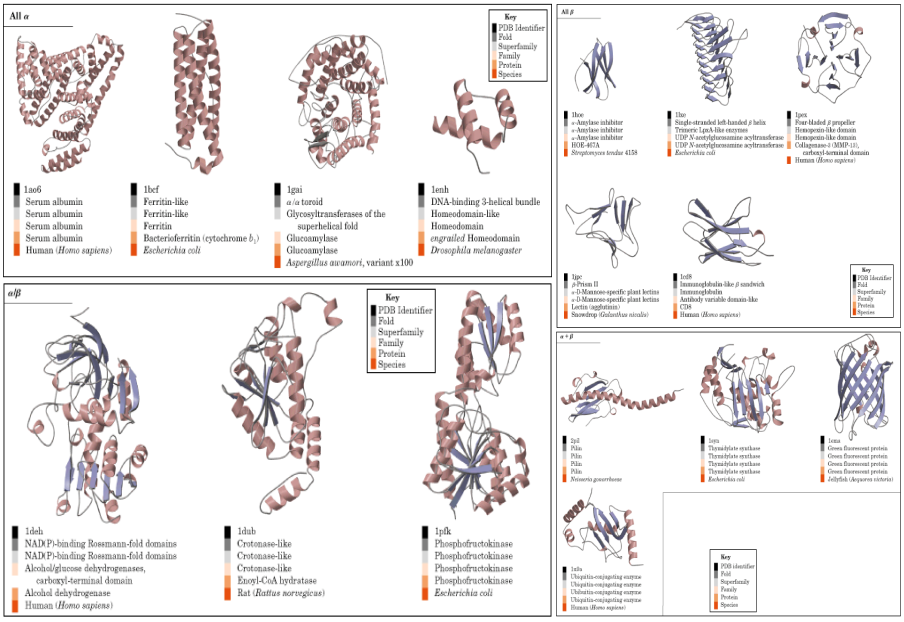


Membrane Proteins

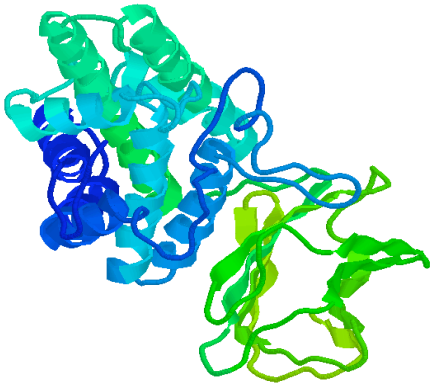




Globular Proteins



What is a Domain?



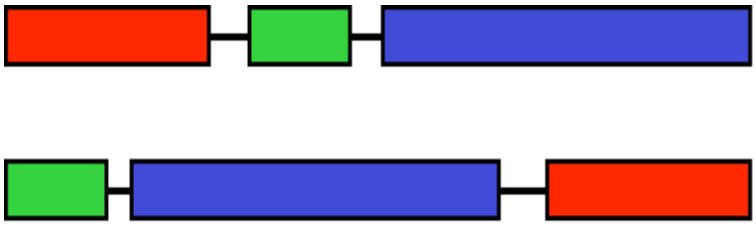
Richardson (1981):

*W*ithin a single subunit [polypeptide chain], contiguous portions of the polypeptide chain frequently fold into compact, local semi-independent units called domains.

More About Domains

- Independent folding units.
- Lots of within contacts, few outside.
- Domains create their own hydrophobic core.
- Regions usually conserved during recombination.
- Different domains of the same protein can have different functions.
- Domains of the same protein may or may not interact.

Why Look for Domains?



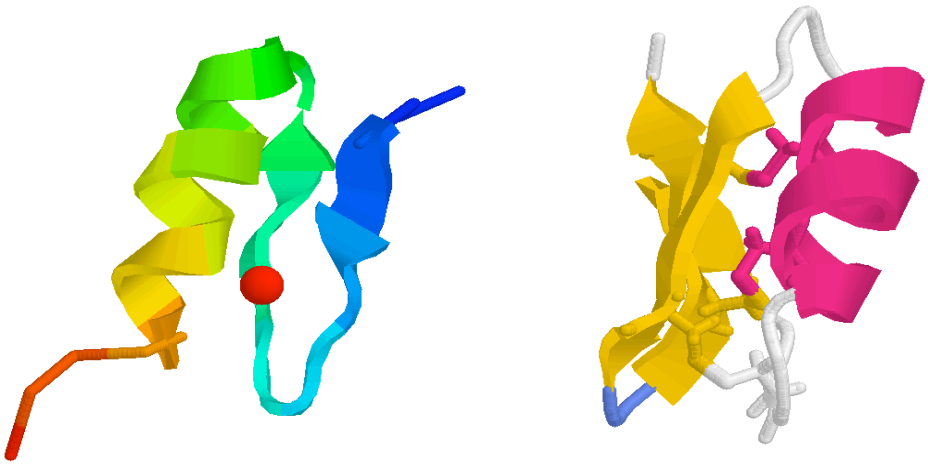
Domains are the currency of protein function!

Domain Size

- Domains can be between 25 and 500 residues long.
- Most are less than 200 residues.
- Domains can be smaller than 50 residues, but these need to be stabilized.

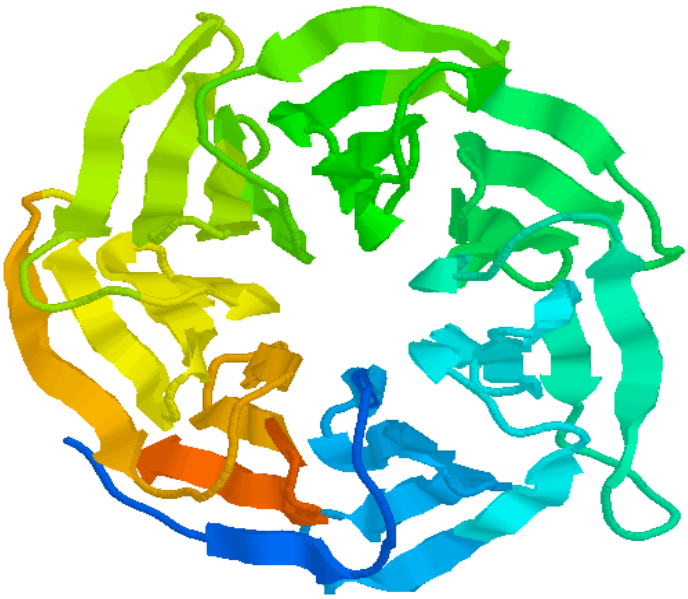
Examples are the zinc finger and a scorpion toxin.

Two Very Small Domains

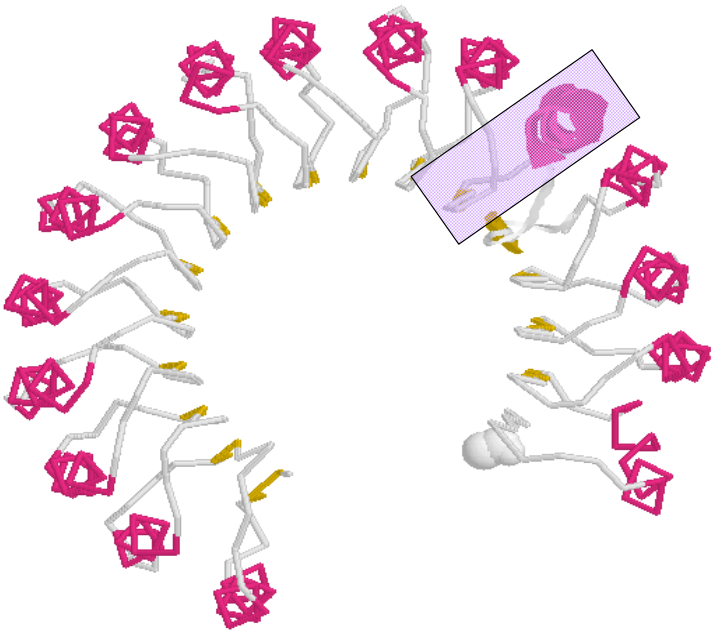


A Humdinger of a Domain

What's the Domain? (Part 1)



What's the Domain? (Part 2)

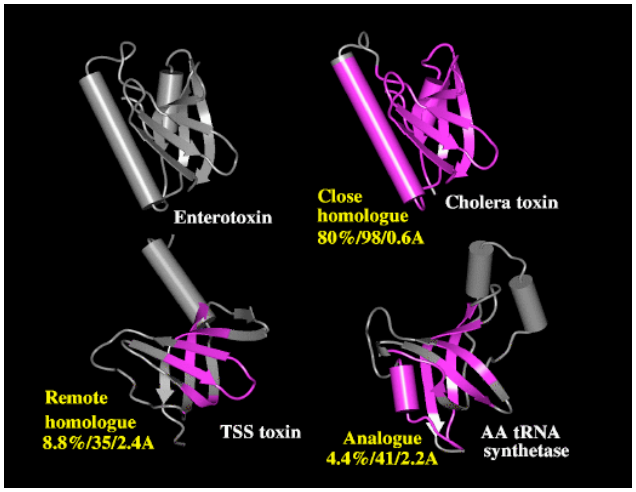


Homology and Analogy

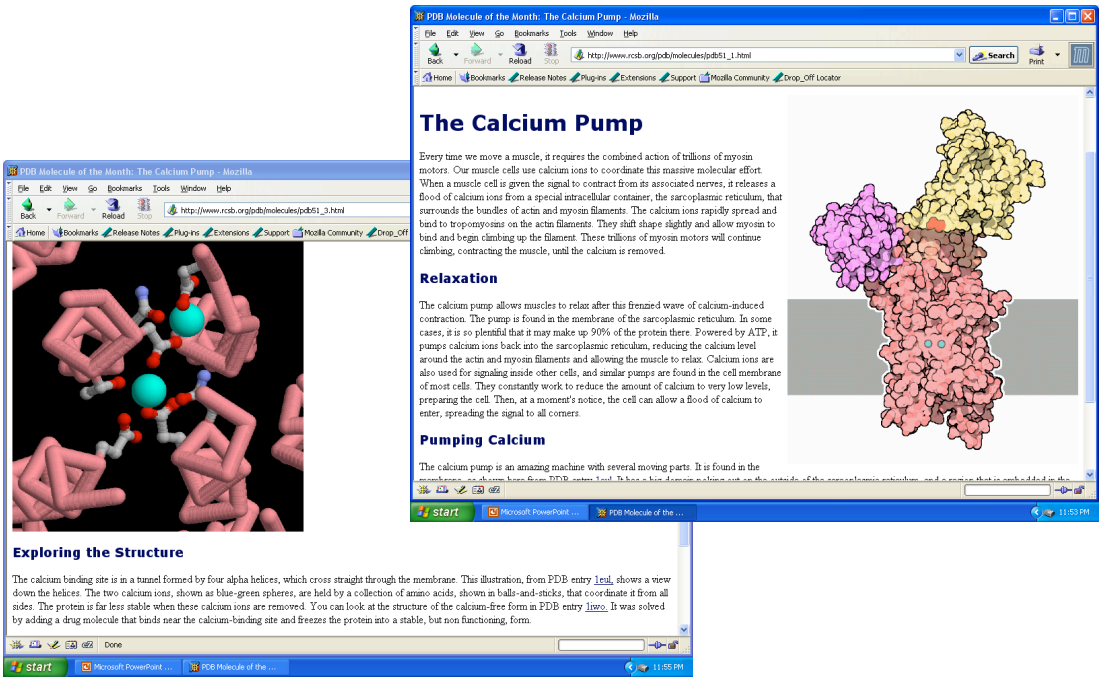
- Homology: Similarity in characteristics resulting from shared ancestry.
- Analogy: The similarity of structure between two species that are not closely related, attributable to convergent evolution.

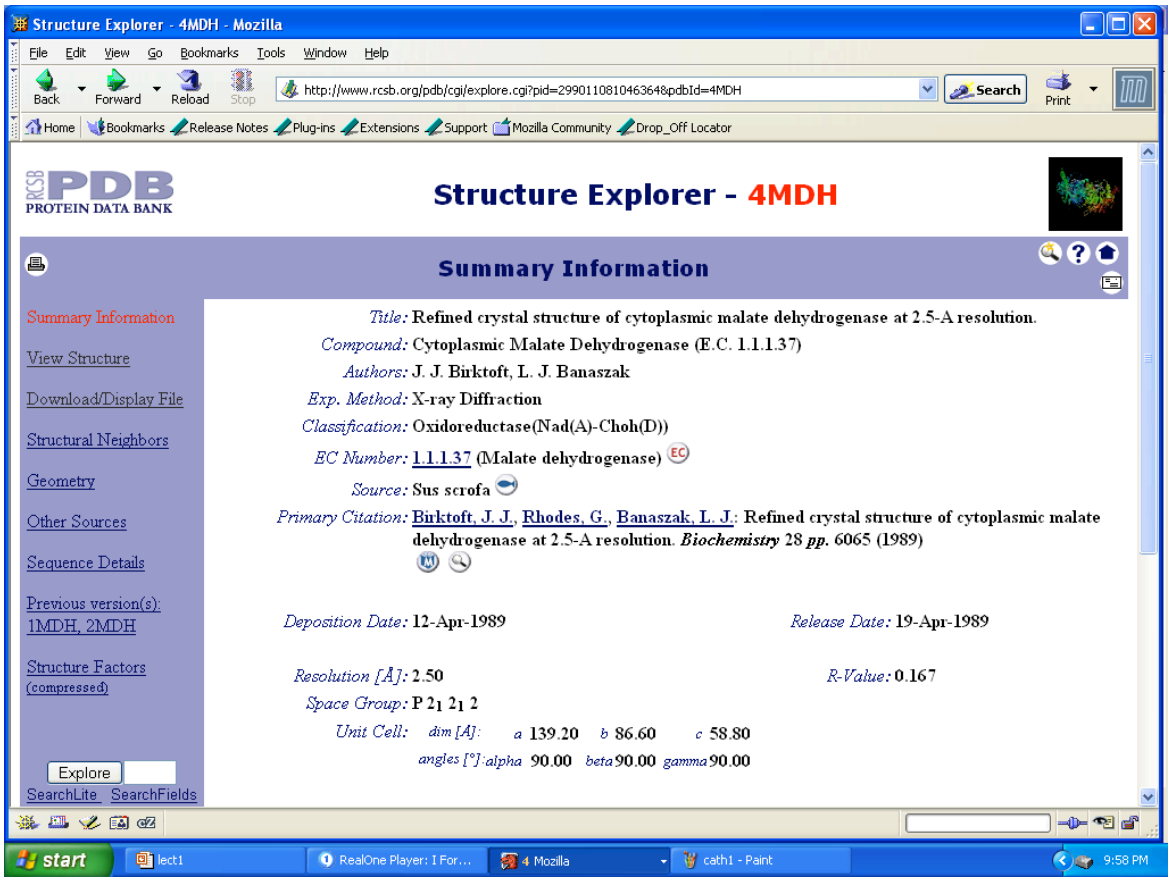
Homologous structures can be devided into orthologues (a result from changes in the same gene between different organisms, such as myoglobin) and paralogues (a result from gene duplication and subsequent changes within an organism and its descendents, such as hemoglobin).

Homology and Analogy



The RCSB Protein Data Bank





PDB File Header

The header contains information about protein and structure, date of the entry, references, crystallographic data, contents and positions of secondary structure elements, etc:

```
HEADER    OXIDOREDUCTASE                          03-OCT-02   1MXT
TITLE     ATOMIC RESOLUTION STRUCTURE OF CHOLESTEROL OXIDASE
TITLE    2 (STREPTOMYCES SP. SA-COO)
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CHOLESTEROL OXIDASE;
COMPND   3 CHAIN: A;
COMPND   4 SYNONYM: CHOD;
COMPND   5 EC: 1.1.3.6;
COMPND   6 ENGINEERED: YES;
COMPND   7 OTHER_DETAILS: FAD COFACTOR NON-COVALENTLY BOUND TO THE
COMPND   8 ENZYME
```

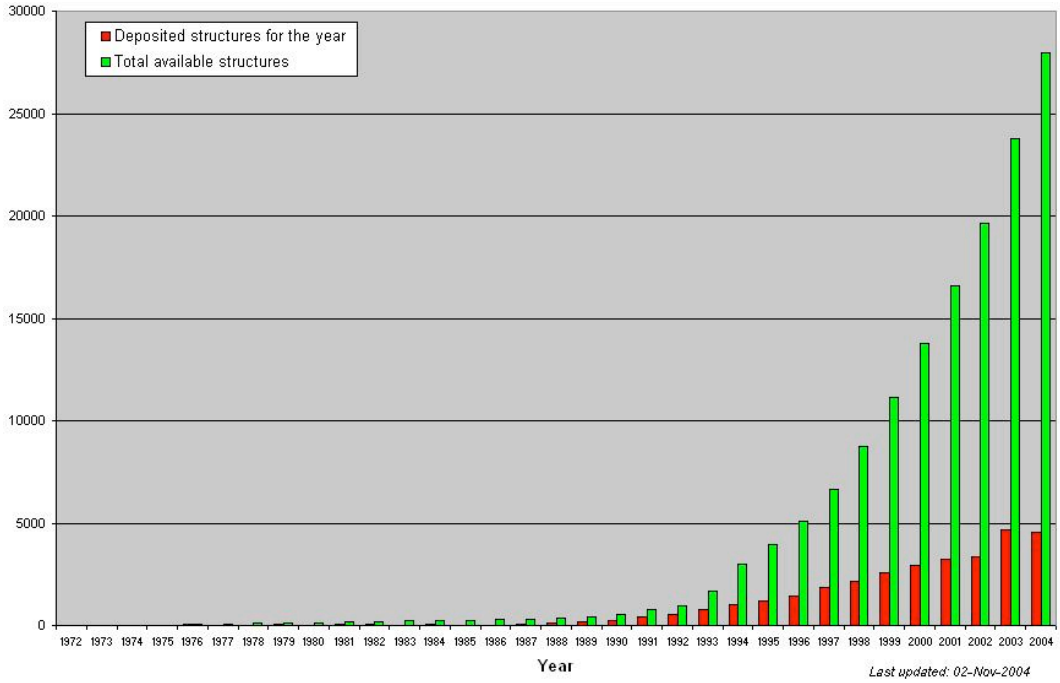
```
AUTHOR    A.VRIELINK,P.I.LARIO
REVDAT   1   25-FEB-03 1MXT    0
JRNL        AUTH   P.I.LARIO,N.SAMPSON,A.VRIELINK
JRNL        TITL   SUB-ATOMIC RESOLUTION CRYSTAL STRUCTURE OF
JRNL        TITL 2 CHOLESTEROL OXIDASE: WHAT ATOMIC RESOLUTION
JRNL        TITL 3 CRYSTALLOGRAPHY REVEALS ABOUT ENZYME MECHANISM AND
JRNL        TITL 4 THE ROLE OF FAD COFACTOR IN REDOX ACTIVITY
JRNL        REF    J.MOL.BIOL.                   V. 326  1635 2003
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836
```

PDB File Body

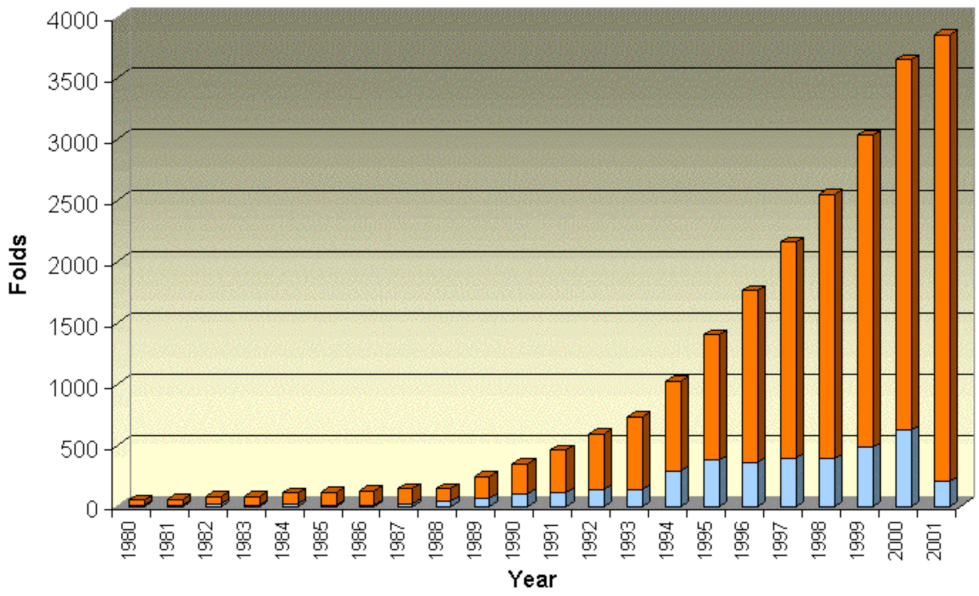
The body of the PDB file contains information about the atoms in the structure:

```
ATOM     76  N   PRO A  12      31.129  -4.659  43.245  1.00  9.00           N
ATOM     77  CA  PRO A  12      32.426  -4.662  42.542  1.00  9.00           C
ATOM     78  C   PRO A  12      32.423  -4.009  41.182  1.00  8.02           C
ATOM     79  O   PRO A  12      33.267  -3.177  40.892  1.00  8.31           O
ATOM     80  CB  PRO A  12      32.791  -6.126  42.592  1.00 10.02           C
ATOM     81  CG  PRO A  12      32.190  -6.663  43.857  1.00 10.12           C
ATOM     82  CD  PRO A  12      30.850  -5.927  43.925  1.00  9.87           C
ATOM     90  N   ALA A  13      31.485  -4.468  40.316  1.00  8.06           N
ATOM     91  CA  ALA A  13      31.357  -3.854  39.004  1.00  7.28           C
ATOM     92  C   ALA A  13      29.947  -3.309  38.814  1.00  7.21           C
ATOM     93  O   ALA A  13      28.969  -3.932  39.200  1.00  7.56           O
ATOM     94  CB  ALA A  13      31.636  -4.879  37.897  1.00  8.54           C
```

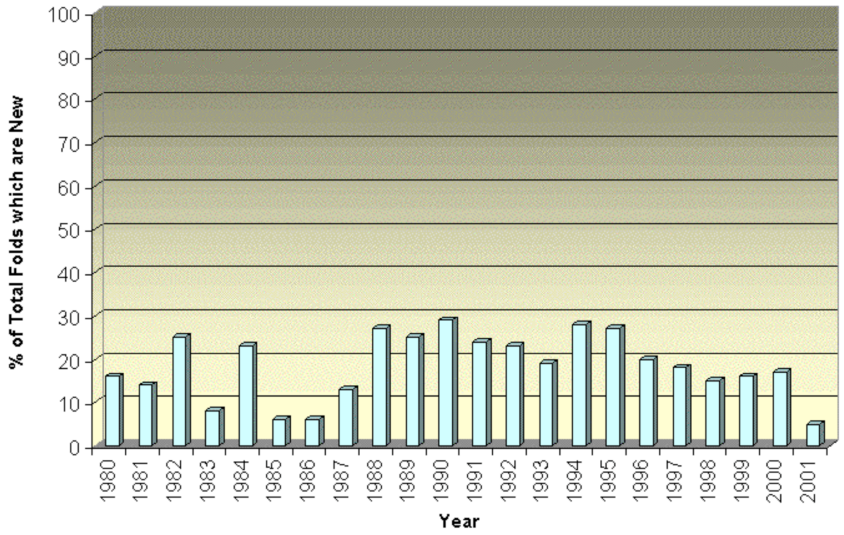
Growth of Structural Data



Unique Folds in the PDB



New Folds Become Rare



SCOP

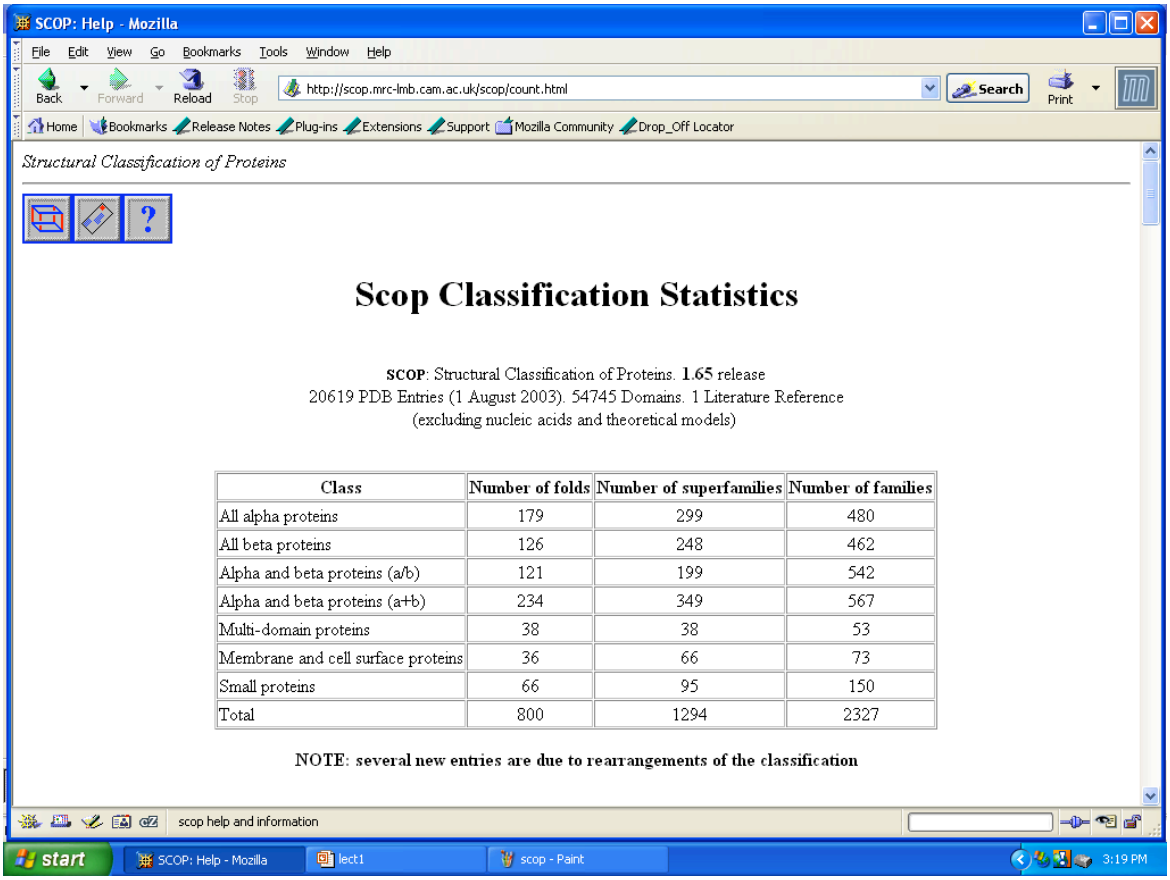
Structural Classification of Proteins

- Proteins are classified (manually!) taking both structural and evolutionary relationship into account.
- There are 7 classes of proteins, the main ones being all alpha, all beta, alpha/beta, and alpha+beta.
- The principle levels in the hierarchy are fold, superfamily, and family.

Hubbard, Murzin, Brenner and Chothia (1997)

SCOP Levels

- **Family**: Clear evolutionarily relationship. In general >30% pairwise residue identities between the proteins.
- **Superfamily**: Probable common evolutionary origin. Proteins have low sequence identities, but structural and functional features suggest that a common evolutionary origin is probable.
- **Fold**: Major structural similarity. Proteins have the same major secondary structures in same arrangement and with the same topological connections.



Some Maybe Surprising Results

5NLL 1AMO 1CHN 1FNB

Flavodoxin Cytochrome reductase Protein CHEY Ferredoxin reductase

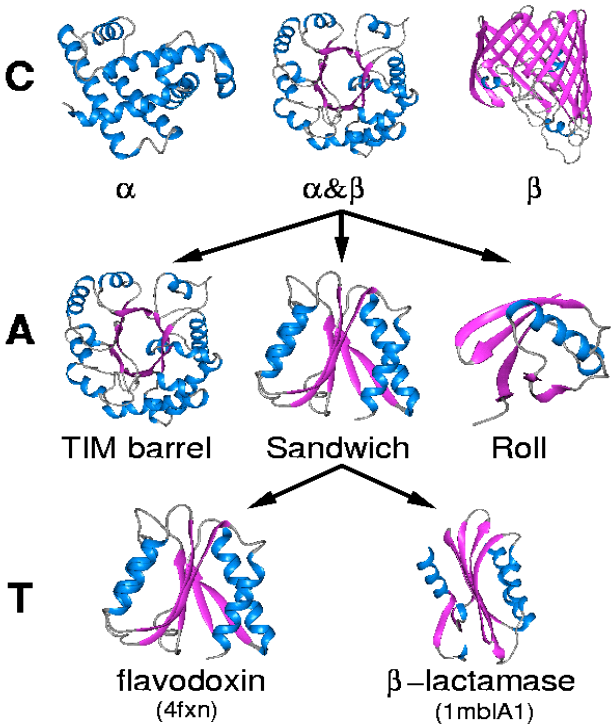
CATH

Protein Structure Classification

- The CATH database is a hierarchical domain classification of protein structures in the Brookhaven protein databank. Only NMR structures and crystal structures solved to resolution better than 3.0 angstroms are considered.
- There are four major levels in this hierarchy: Class, Architecture, Topology (fold family) and Homologous superfamily.
- Multidomain proteins are subdivided into their domains using a consensus procedure. All the classification is performed on individual protein domains.

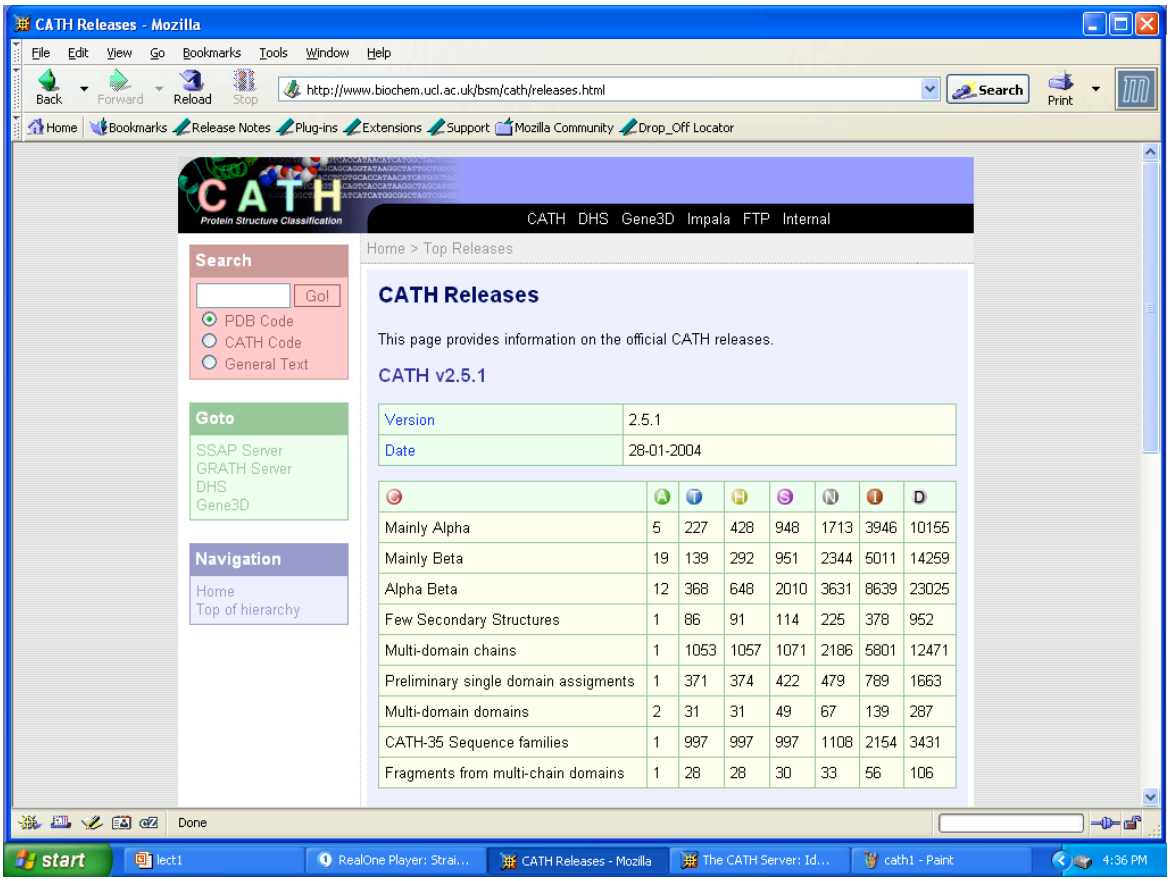
Orengo, Michie, Jones, Jones, Swindells, Thornton (1997)

The CATH Hierarchy



SCOP versus CATH

Correspondencce between SCOP and CATH hierarchies	
SCOP	CATH
Class	Class
	Architecture
Fold	Topology
	Homologous superfamily
Superfamily	
Family	Sequence family
Domain	Domain



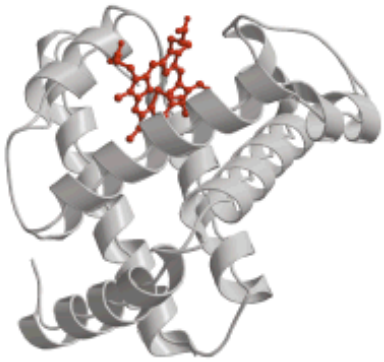
DALI

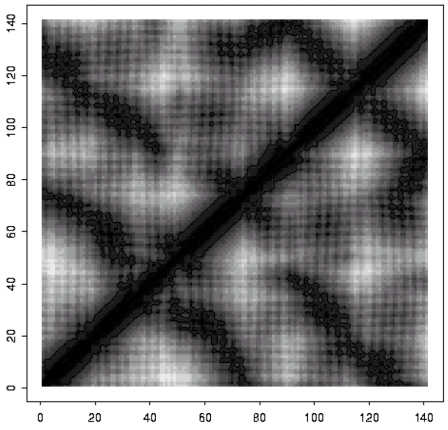
Distance Matrix Alignment

- DALI generates alignments of structural fragments, and is able to find alignments involving chain reversals and different topologies.
- The algorithm uses distance matrices to represent each structure to be compared.
- Application of DALI to the entire PDB produces two classifications of structures: FSSP and DDD (3D).

Holm and Sander (1993)

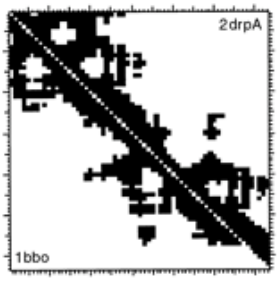
DALI



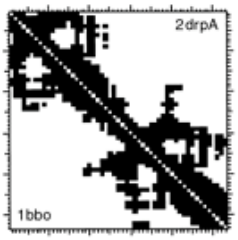


DALI

Unaligned:



Aligned:



Unaligned:

```
1bbo     1 KYICEECGIEXKKKPSMLKKHIRTHTDVRPYHCTYCNFSFKTKGNLTKHMKSKAHSKK  57

2drpA 103 FTKEGEHTYECKVCSRVYTHISNFCEHYVTSHKRNVKVYPCPFCFKEFTRKDNMTAHVKIIHK 165
```

Aligned:

```
1bbo     1 .......KYICEECGIEXKKKPSMLKKHIRTHt..DVRPYHCTYCNFSFKTKGNLTKHMKSKAHskk  57
                 | | |       |   | |      | | | |   |  | | | | |
2drpA 103 ftkegehTYECKVCSRVYTHISNFCEHYVTShkrNVKVYPCPFCFKEFTRKDNMTAHVKIIHK... 165
```

FSSP and DDD

- The families of structurally similar proteins (FSSP) is a database of structural alignments of proteins in the protein data bank (PDB). It presents the results of applying DALI to (almost) all chains of proteins in the PDB.
- The DALI domain dictionary (DDD) is a corresponding classification of recurrent domains automatically extracted from known proteins.

References: Holm and Sander

- *Protein Structure Comparison by Alignment of Distance Matrices*, Journal of Molecular Biology 233, pp 123-138, 1993.
- *The FSSP Database of Structurally Aligned Protein Fold Families*, Nucleic Acids Research 22 (17), pp 3600-3609, 1994.
- *Mapping the Universe*, Science 273 (5275), pp 595-602, 1996.
- *Touring Protein Fold Space with Dali/FSSP*, Nucleic Acids Research 26 (1), pp 316-319, 1998.

Other Algorithms for Domain Decomposition

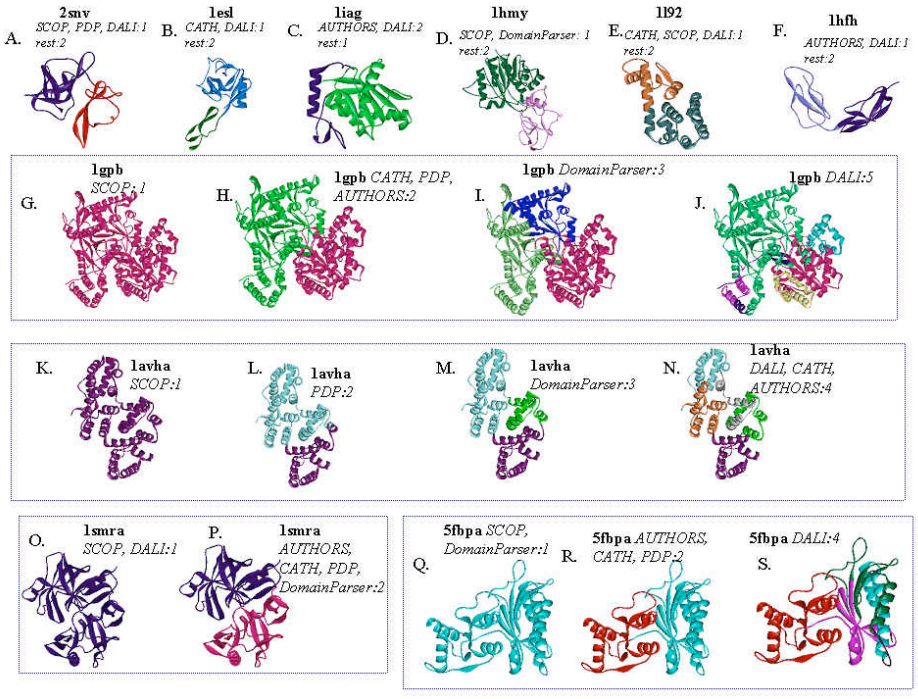
- The Protein Domain Parser (PDP) uses compactness as a chief principle.

  `http://123d.ncifcrf.gov/pdp.html`

- DomainParser is graph theory based. The underlying principle used is that residue-residue contacts are denser within a domain than between domains.

  `http://compbio.ornl.gov/structure/domainparser/`

Oh Dear…



Parsing Sequence into Domains

- Look for internal duplication.
- Look for low complexity segments.
- Look for transmembrane segments.

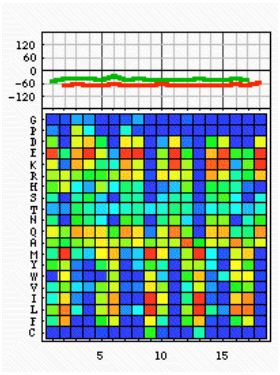
Why is That Important?

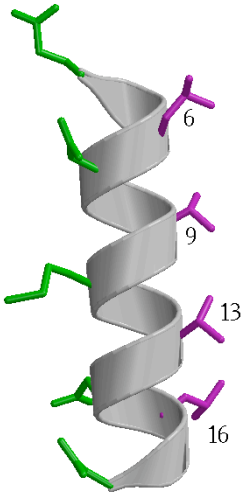
- Functional insights.
- Improved database searching.
- Fold recognition.
- Structure determination.

PRODOM:
`http://protein.toulouse.inra.fr/prodom/current/html/home.php`

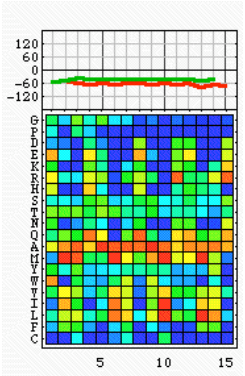
PFAM:
`http://www.sanger.ac.uk/Software/Pfam/`

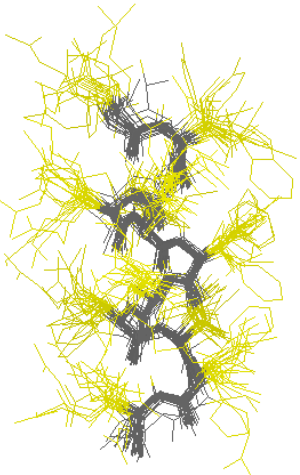
I-Sites

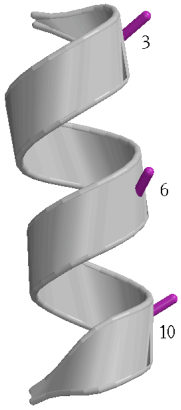




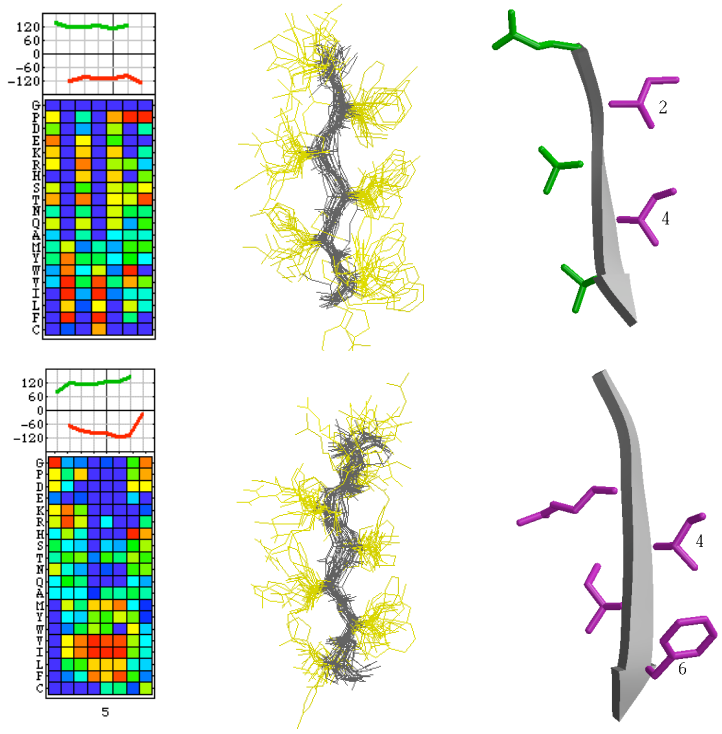
I-Sites







I-Sites



I-Sites