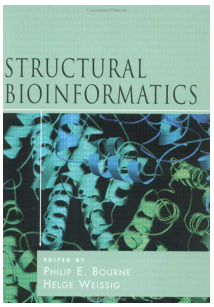
Protein Structure: Data Bases and Classification

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Reference

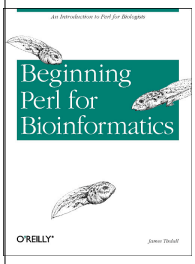


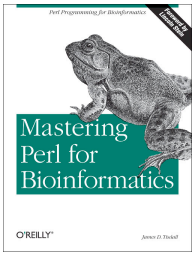
Bourne and Weissig

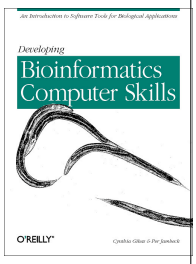
Structural Bioinformatics

Wiley, 2003

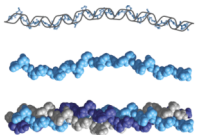
More References

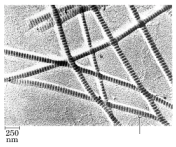




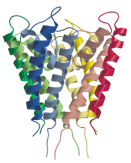


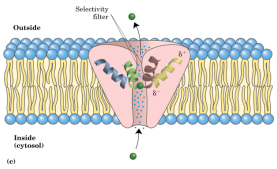
Structural Proteins



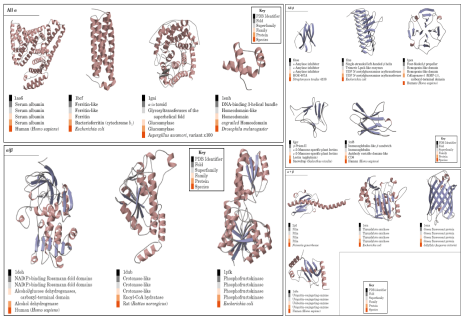


Membrane Proteins





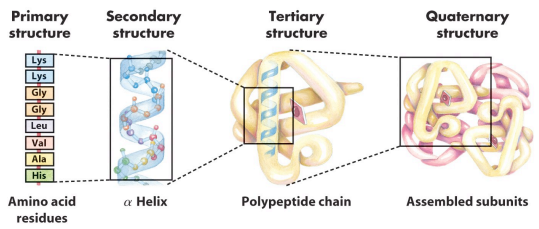
Globular Proteins



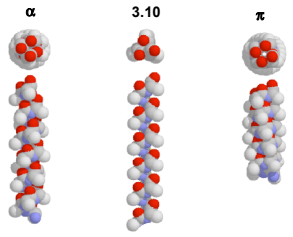
Terminology

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

Hierarchy of Protein Structure

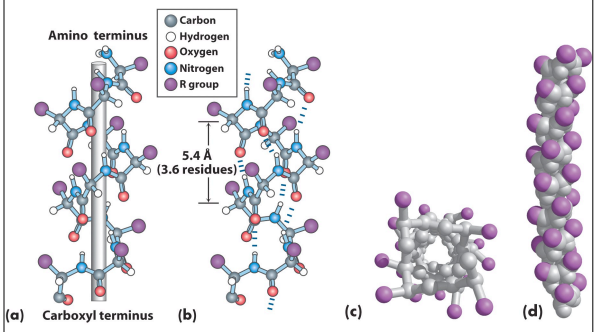


Helices



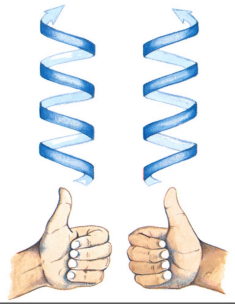
	α	3.10	π
Amino acids/turn:	3.6	3.0	4.4
Frequency	~97%	~3%	rare
H-bonding	*i, i*+4	*i, i*+3	*i, i*+5

α-helices

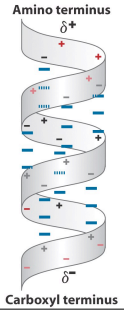


α-helices

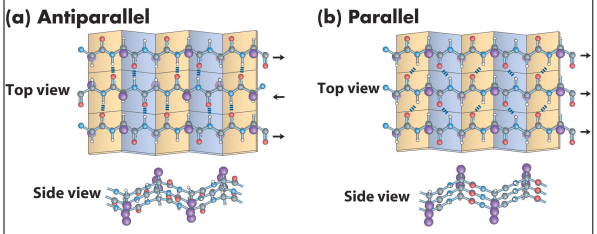
α-helices have handedness:



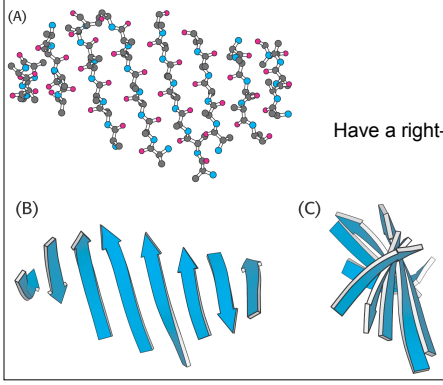
α-helices have a dipole:



β-sheets

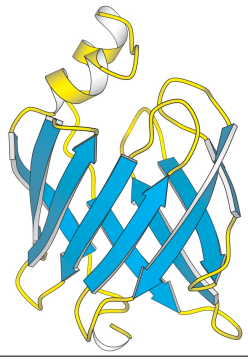


β-sheets



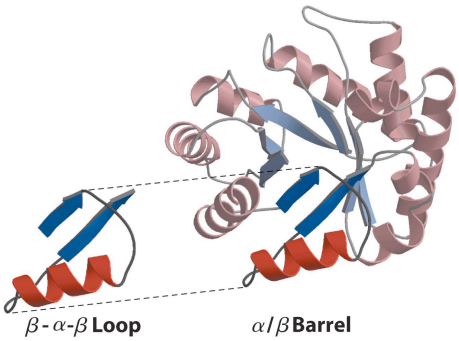
Have a right-handed twist!

β-sheets

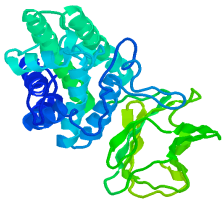


Can form higher level structures!

Super Secondary Structure Motifs



What is a Domain?



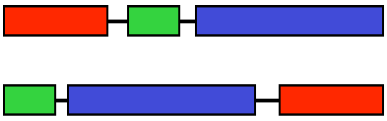
Richardson (1981):

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?



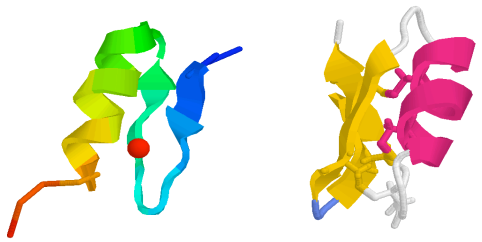
Domains are the currency of protein function!

Domain Size

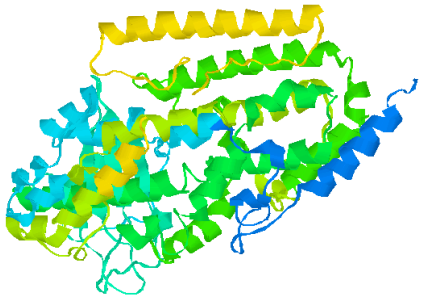
- Domains can be between 25 and 500 residues long.
- Most are less than 200 residues.
- Domains can be smaller than 50 residues, but these need to be stabilized.

Examples are the zinc finger and a scorpion toxin.

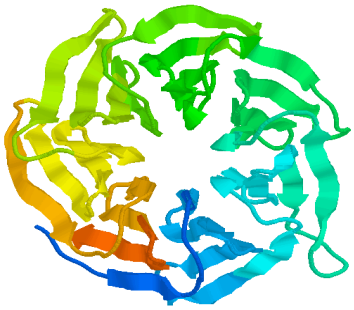
Two Very Small Domains



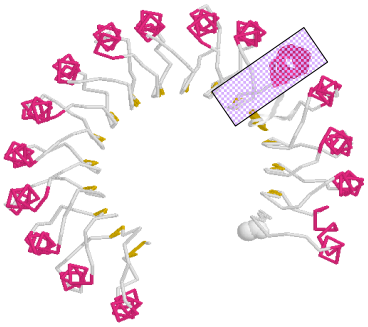
A Humdinger of a Domain



What's the Domain? (Part 1)



What's the Domain? (Part 2)

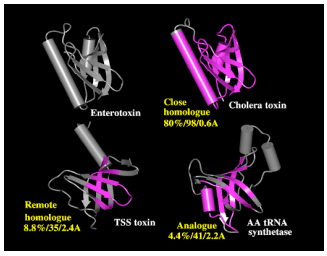


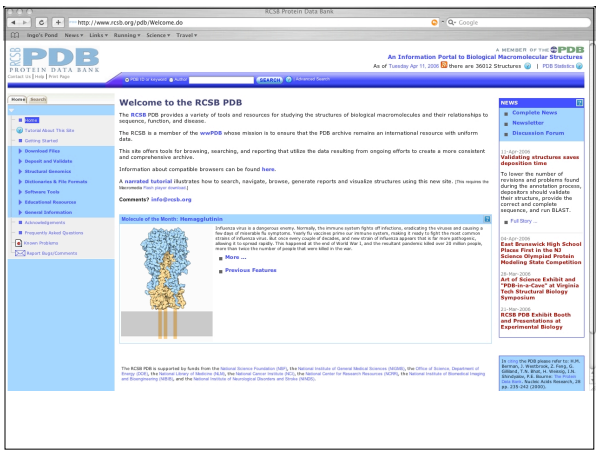
Homology and Analogy

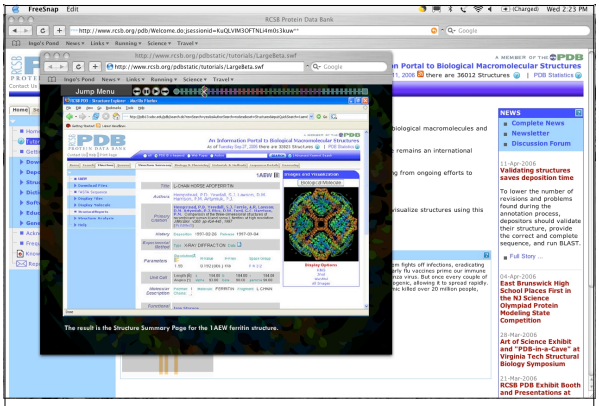
- Homology: Similarity in characteristics resulting from shared ancestry.
- Analogy: The similarity of structure between two species that are not closely related, attributable to convergent evolution.

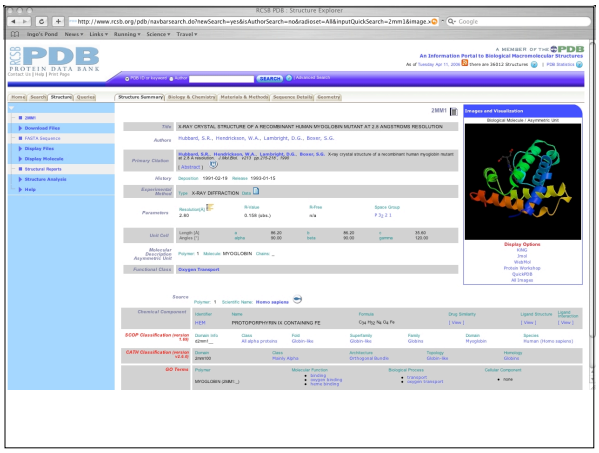
Homologous structures can be devided into orthologues (a result from changes in the same gene between different organisms, such as myoglobin) and paralogues (a result from gene duplication and subsequent changes within an organism and its descendents, such as hemoglobin).

Homology and Analogy









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
```

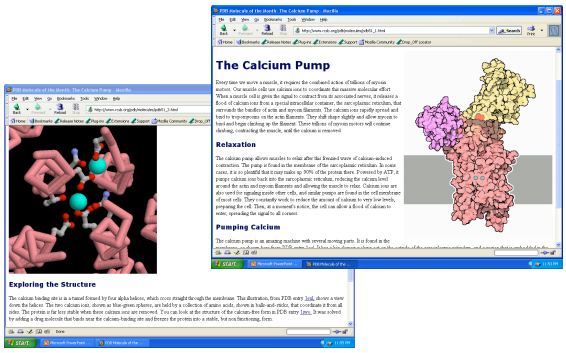
```
AUTHOR    A.VRIELINK,P.I.LARIO
REVDAT   1   25-FEB-03 1MXT    0
JRNL        AUTH   P.I.LARIO,N.SAMPSON,A.VRIELINK
JRNL        TITL   SUB-ATOMIC RESOLUTION CRYSTAL STRUCTURE OF
JRNL       TITL 2 CHOLESTEROL OXIDASE: WHAT ATOMIC RESOLUTION
JRNL       TITL 3 CRYSTALLOGRAPHY REVEALS ABOUT ENZYME MECHANISM 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
```

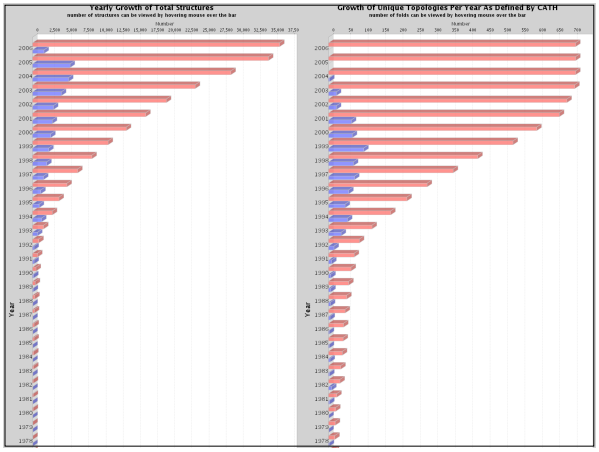
PDB File Body

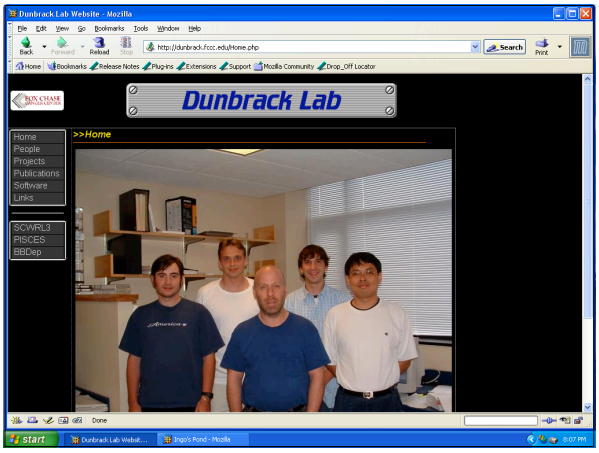
The body of the PDB file contains information about the atoms in the structure:

```
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
```

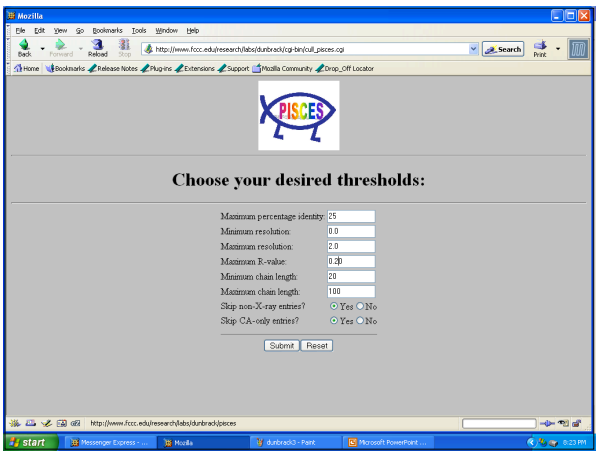
Molecule of the Month

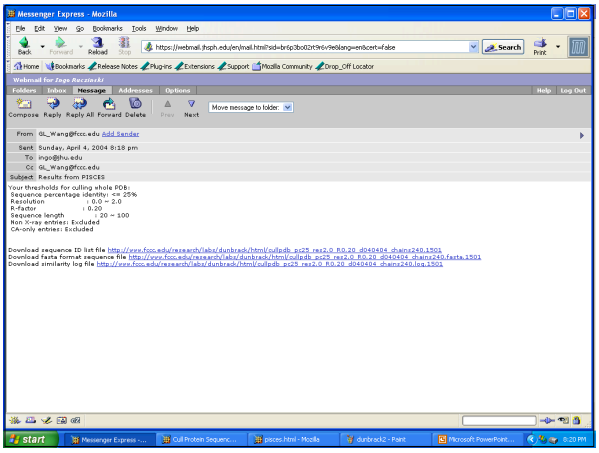


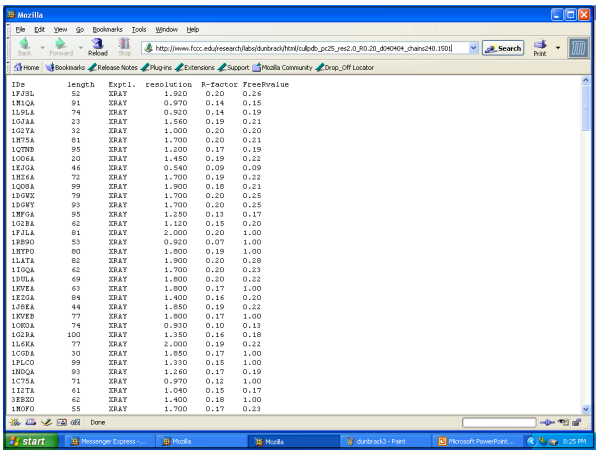





Cull Protein Sequence List Page - Mozilla
PISCES
What do you want to do?
Cull sequences from the whole PDB by resolution, sequence identity, R-factor, etc.
Cull from your own list of PDB chains
Cull from your own list of GenBank, SwissProt, etc. identifiers. For instance, you can paste the hits listed at the top of BLAST output, we can go get the whole sequences from GenBank.
Cull from your own file of sequences in FASTA format or from BLAST output (i.e., we use the fragments of sequences from the Sbjct line in the BLAST output which you will upload)
Submit
Reset


Mozilla
PISCES
Choose your desired thresholds:
Maximum percentage identity: 25
Minimum resolution: 0.0
Maximum resolution: 2.0
Maximum R-value: 0.20
Minimum chain length: 20
Maximum chain length: 100
Skip non-X-ray entries? Yes No
Skip CA-only entries? Yes No
Submit
Reset


Messenger Express - Mozilla
From GL_Wang@fccc.edu
Sent Sunday, April 4, 2004 8:18 pm
To ingo@hu.edu
Cc GL_Wang@fccc.edu
Subject Results from PISCES
Your thresholds for culling whole PDB:
Sequence percentage identity: <= 25%
Resolution : 0.0 ~ 2.0
R-factor : 0.20
Sequence length : 20 ~ 100
Non X-ray entries: Excluded
CA-only entries: Excluded



IDs	length	Exptl.	resolution	R-factor	FreeRvalue
1FJSL	52	XRAY	1.920	0.20	0.26
1M1QA	91	XRAY	0.970	0.14	0.15
1L9LA	74	XRAY	0.920	0.14	0.19
1GJAA	23	XRAY	1.560	0.19	0.21
1G2YA	32	XRAY	1.000	0.20	0.20
1H75A	81	XRAY	1.700	0.20	0.21
1QTNB	95	XRAY	1.200	0.17	0.19
1OO6A	20	XRAY	1.450	0.19	0.22
1EJGA	46	XRAY	0.540	0.09	0.09
1HZ6A	72	XRAY	1.700	0.19	0.22
1QO8A	99	XRAY	1.900	0.18	0.21
1DGWX	79	XRAY	1.700	0.20	0.25
1DGWY	93	XRAY	1.700	0.20	0.25
1MFGA	95	XRAY	1.250	0.13	0.17
1G2BA	62	XRAY	1.120	0.15	0.20
1FJLA	81	XRAY	2.000	0.20	1.00
1RB9O	53	XRAY	0.920	0.07	1.00
1NYPO	80	XRAY	1.800	0.19	1.00
1LATA	82	XRAY	1.900	0.20	0.28
1IGQA	62	XRAY	1.700	0.20	0.23
1DULA	69	XRAY	1.800	0.20	0.22
1KVEA	63	XRAY	1.800	0.17	1.00
1EZGA	84	XRAY	1.400	0.16	0.20
1J8EA	44	XRAY	1.850	0.19	0.22
1KVEB	77	XRAY	1.800	0.17	1.00
1OBGA	74	XRAY	0.930	0.10	0.13
1G2RA	100	XRAY	1.350	0.16	0.18
1L6EA	77	XRAY	2.000	0.19	0.22
1CGDA	30	XRAY	1.850	0.17	1.00
1PLCO	99	XRAY	1.330	0.15	1.00
1NDQA	93	XRAY	1.260	0.17	0.19
1C75A	71	XRAY	0.970	0.12	1.00
1IZTA	61	XRAY	1.040	0.15	0.17
3EEIO	62	XRAY	1.400	0.18	1.00
1MOFO	55	XRAY	1.700	0.17	0.23

SCOP
Structural Classification of Proteins

- Proteins are classified (manually!) taking both the 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 of SCOP are fold, superfamily, and family.

Murzin AG, Brenner SE, Hubbard T, and Chothia C (1995)

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.

SCOP Help

http://scop.mrc-lmb.cam.ac.uk/scop/count.html#scop-1.69

Ingo's Pond News Links Running Science Travel

Structural Classification of Proteins

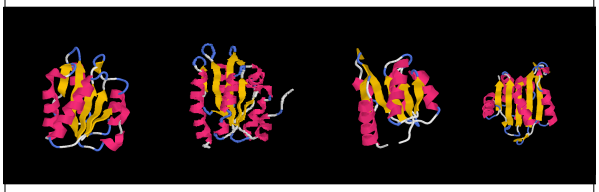
Scop Classification Statistics

SCOP: Structural Classification of Proteins. **1.69** release
25973 PDB Entries (1 Oct 2004). 70859 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	218	376	608
All beta proteins	144	290	560
Alpha and beta proteins (a/b)	136	222	629
Alpha and beta proteins (a+b)	279	409	717
Multi-domain proteins	46	46	61
Membrane and cell surface proteins	47	88	99
Small proteins	75	108	171
Total	945	1539	2845

Some Maybe Surprising Results

5NLL 1AMO 1CHN 1FNB



Flavodoxin Cytochrome reductase Protein CHEY Ferredoxin reductase

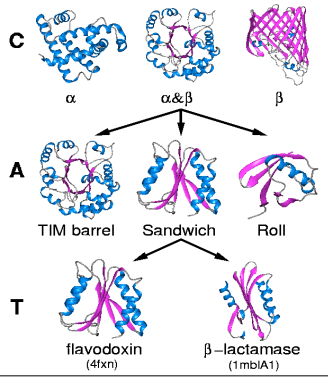
CATH

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 CA, Michie AD, Jones S, Jones DT, Swindells MB, and Thornton JM (1997)

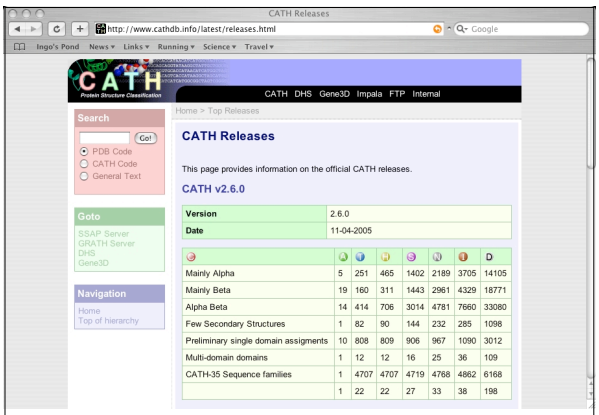
The CATH Hierarchy



SCOP versus CATH

Correspondencce between SCOP and CATH hierarchies

SCOP	CATH
Class	Class
	Architecture
Fold	Topology
	Homologous superfamily
Superfamily	
Family	Sequence family
Domain	Domain

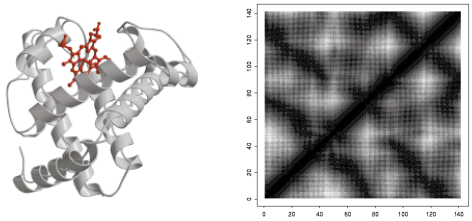


DALI
Distance Matrix Alignment

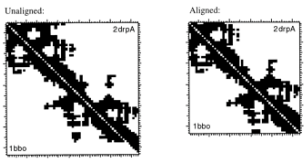
- DALI generates alignments of structural fragments, and is able to find alignments involving chain reversals and different topologies.
- The algorithm uses distance matrices to represent each structure to be compared.
- Application of DALI to the entire PDB produces two classifications of structures: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

Holm L, and Sander C (1993)

DALI



DALI



Unaligned:

```
1bbo     1 KYICEECGIEKKKPSMLKHMIRTHTDVRPYHCTYCNFSFKTKGNLTKHMKSKAHSKK  57
2drpA  103 PTKEGEHTYECKVCSRVYTHISNFCRHYVTSHKRNVKVYECPFCFKETRKDNMTAHVKLIHK 165
```

Aligned:

```
1bbo     1 .......KYICEECGIEKKKPSMLKHMIRTHt...DVRPYHCTYCNFSFKTKGNLTKHMKSKAHsKK  57
                   |||  |  |||||||||||||||   |||||||||||||||||||||||||||||||
2drpA  103 ftkegehTYECKVCSRVYTHISNFCRHYVTShkrNVKVYECPFCFKETRKDNMTAHVKLIHK... 165
```

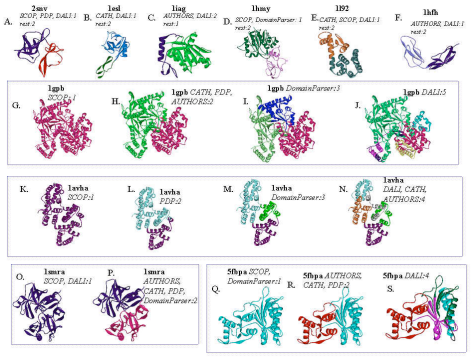
FSSP and DDD

- 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.

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.

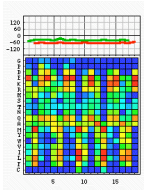
Why is That Important?

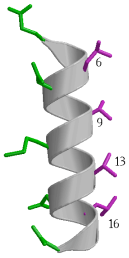
- Functional insights.
- Improved database searching.
- Fold recognition.
- Structure determination.

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

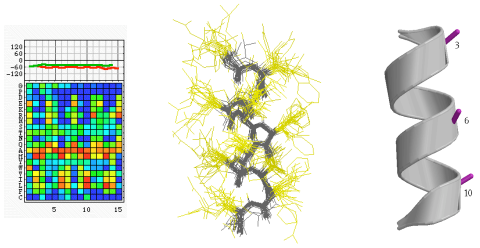
PFAM:
`http://www.sanger.ac.uk/Software/Pfam/`

I-Sites

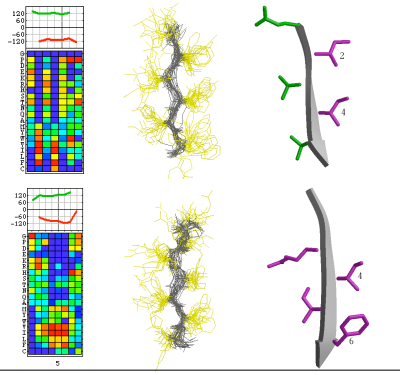




I-Sites



I-Sites



I-Sites