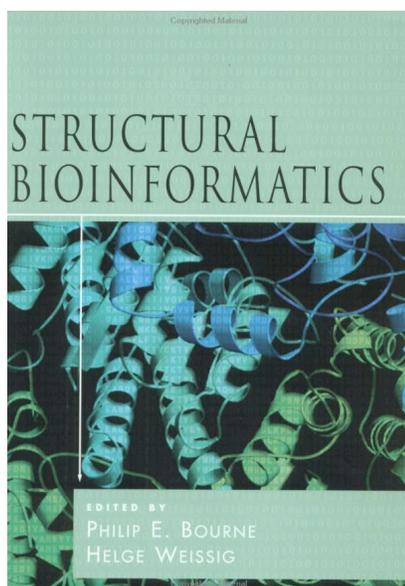


Protein Structure: Data Bases and Classification

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

A Foine Reference

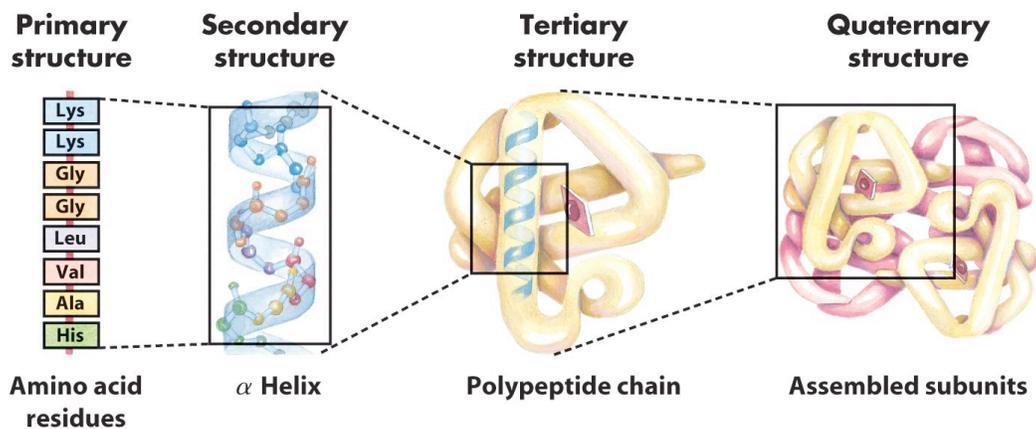


Bourne and Weissig
Structural Bioinformatics
Wiley, 2003

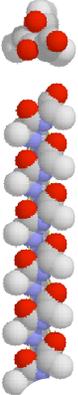
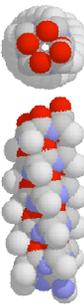
Terminology

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

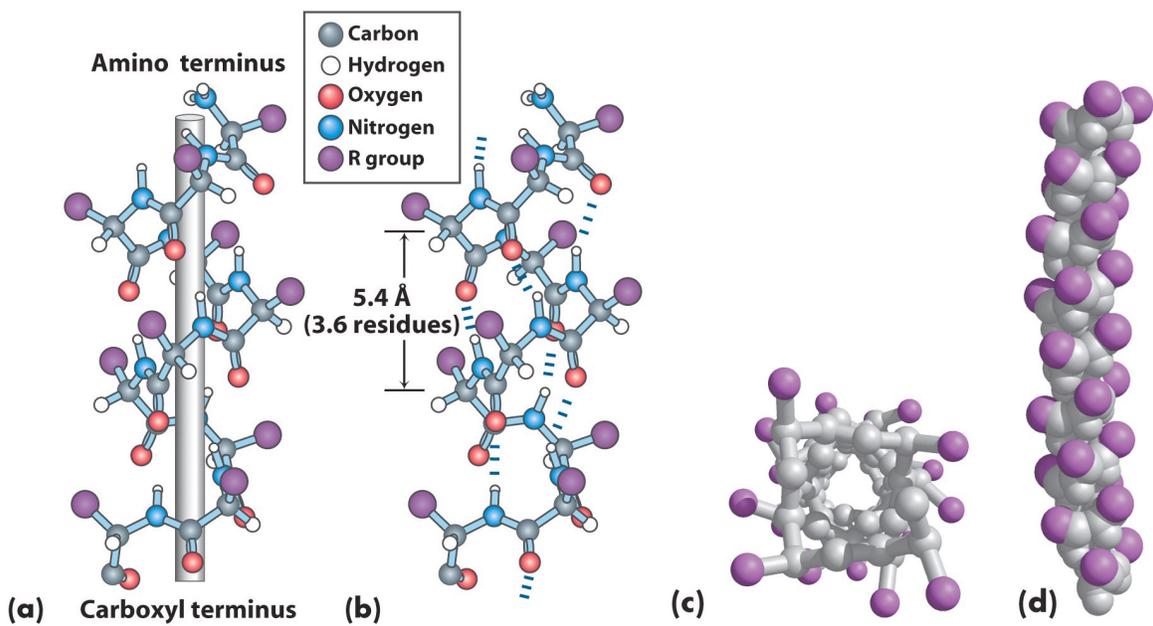
Hierarchy of Protein Structure



Helices

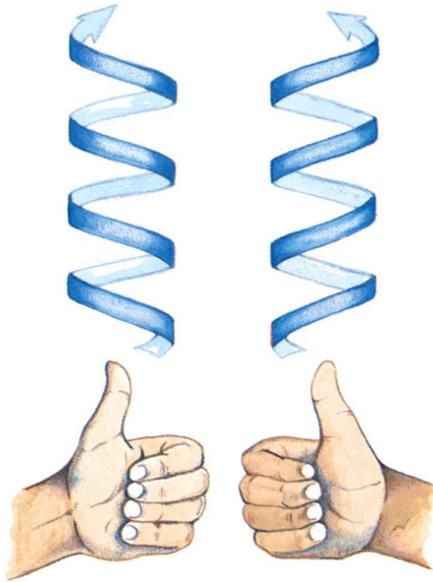
	α	3.10	π
			
Amino acids/turn:	3.6	3.0	4.4
Frequency	~97%	~3%	rare
H-bonding	$i, i+4$	$i, i+3$	$i, i+5$

α -helices

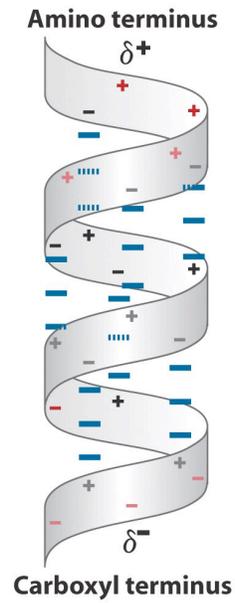


α -helices

α -helices have handedness:

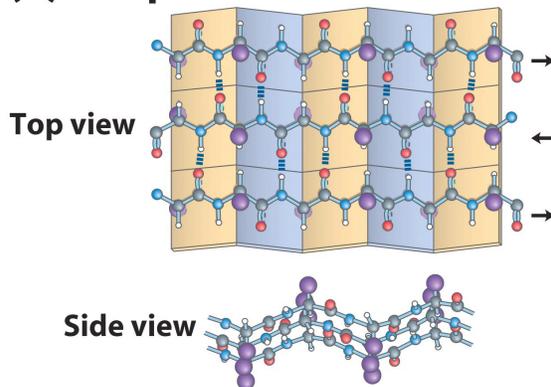


α -helices have a dipole:

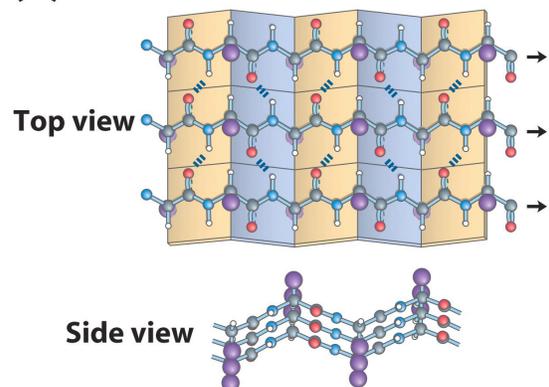


β -sheets

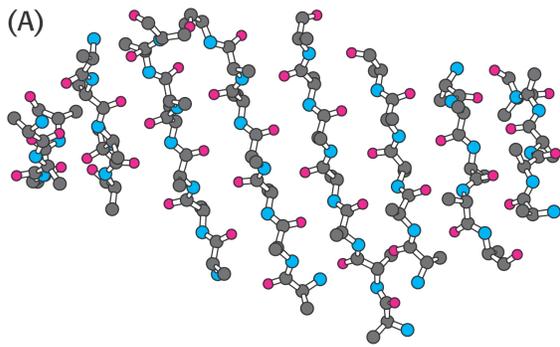
(a) Antiparallel



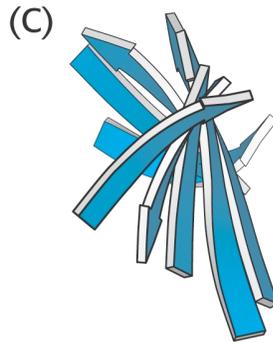
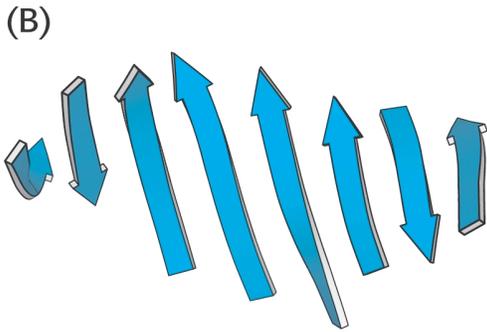
(b) Parallel



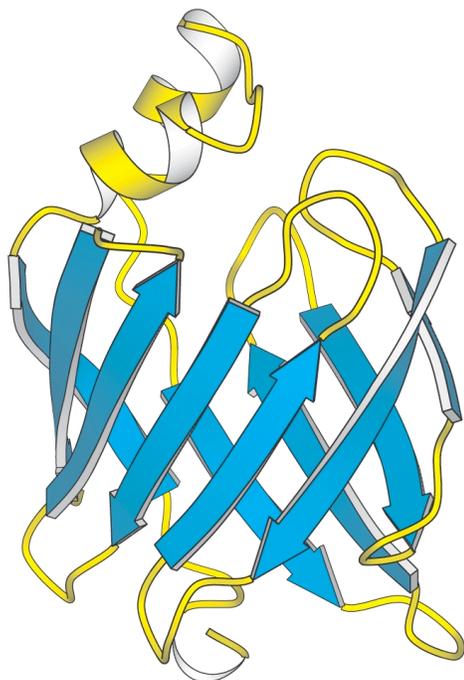
β -sheets



Have a right-handed twist!

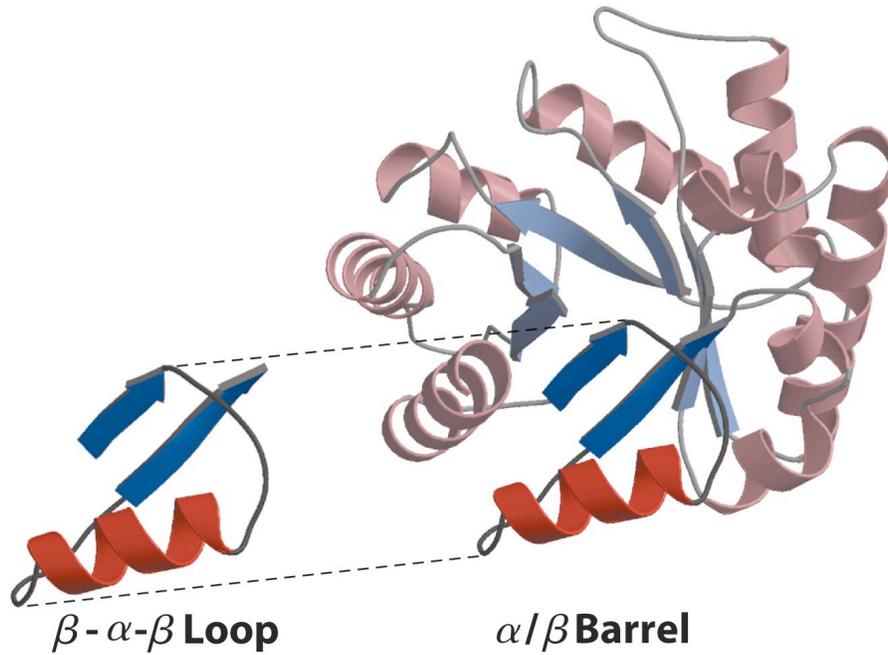


β -sheets

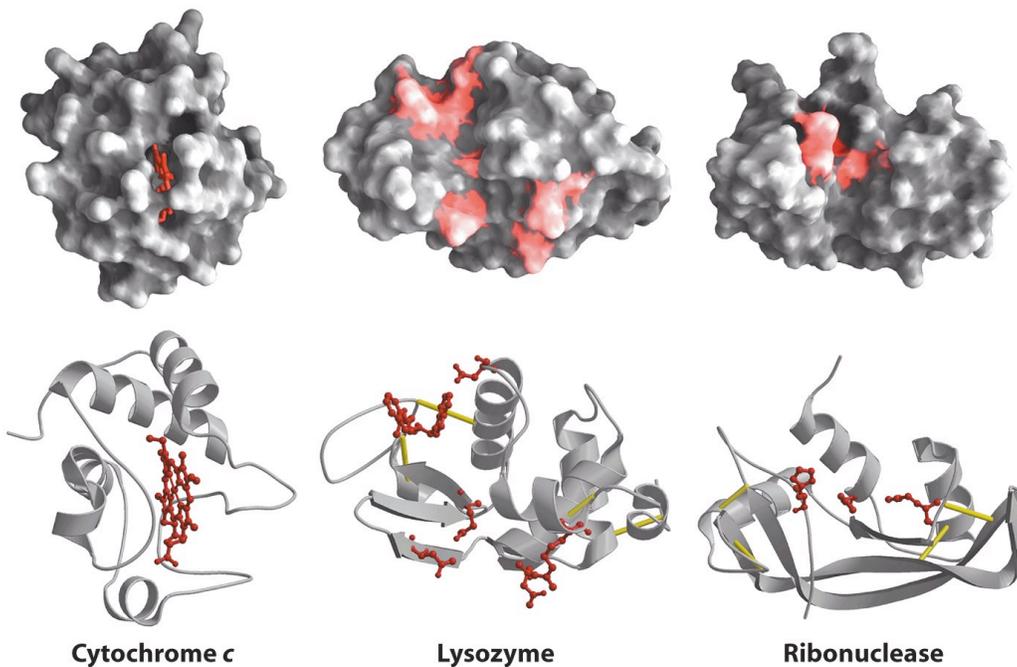


Can form higher level structures!

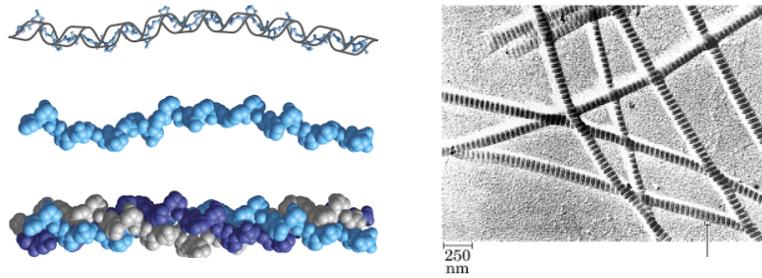
Super Secondary Structure Motifs



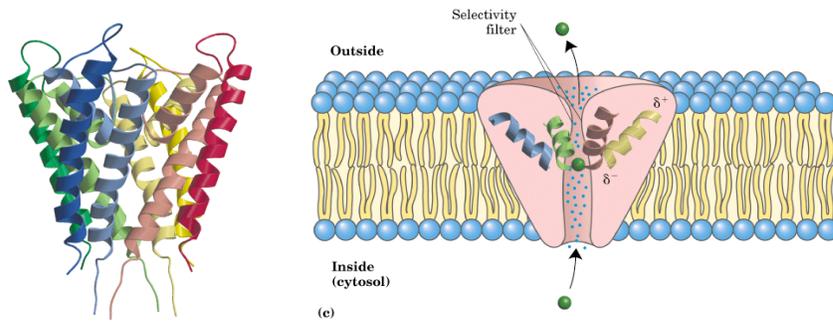
Protein Structure and Function



Structural Proteins



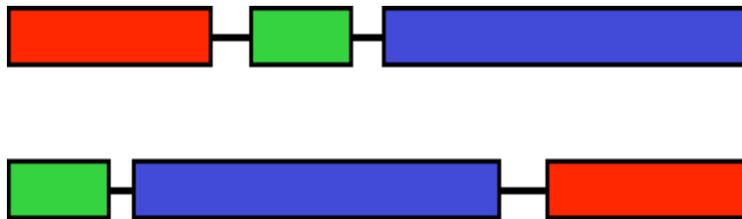
Membrane Proteins



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?



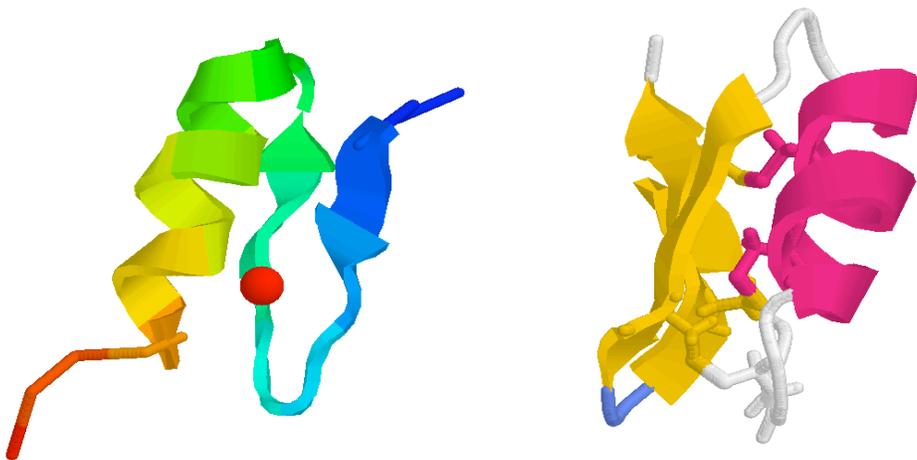
Domains are the currency of protein function!

Domain Size

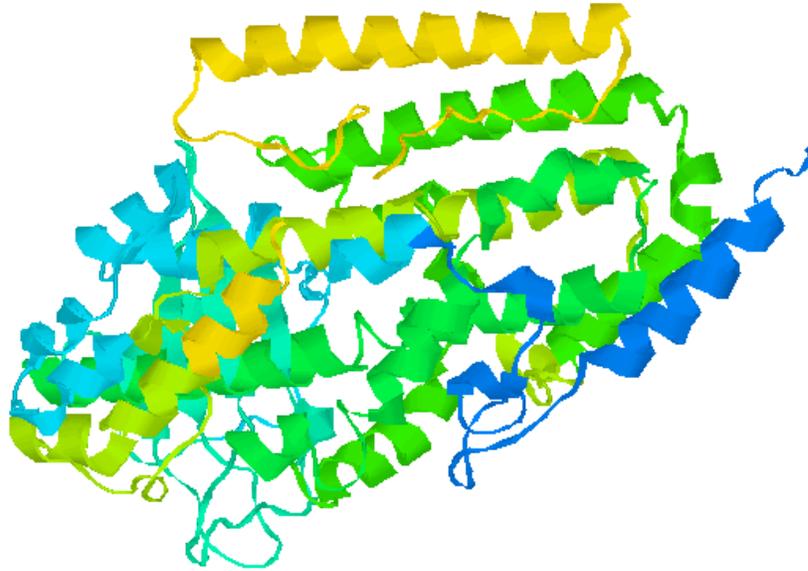
- Domains can be between 25 and 500 residues long.
- Most are less than 200 residues.
- Domains can be smaller than 50 residues, but these need to be stabilized.

Examples are the zinc finger and a scorpion toxin.

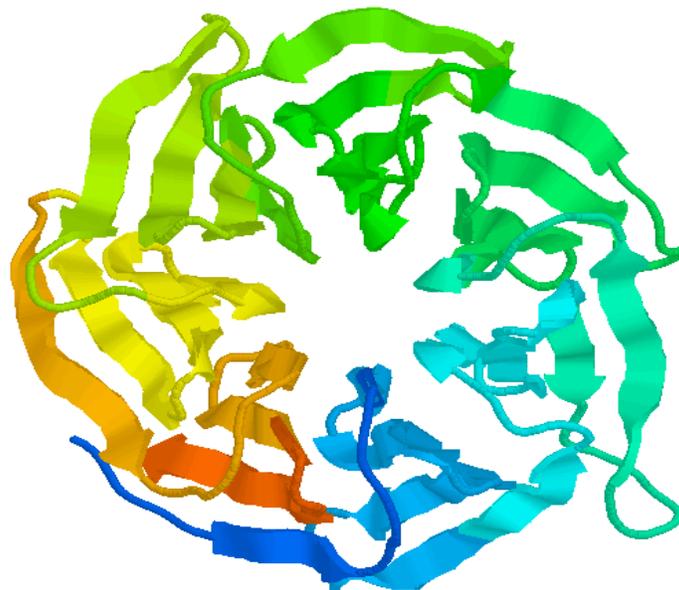
Two Very Small Domains



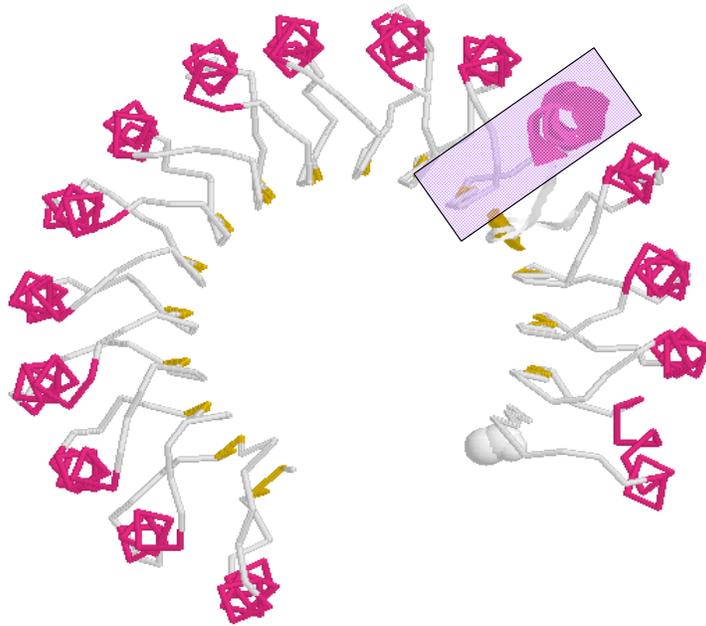
A Humdinger of a Domain



What's the Domain? (Part 1)



What's the Domain? (Part 2)

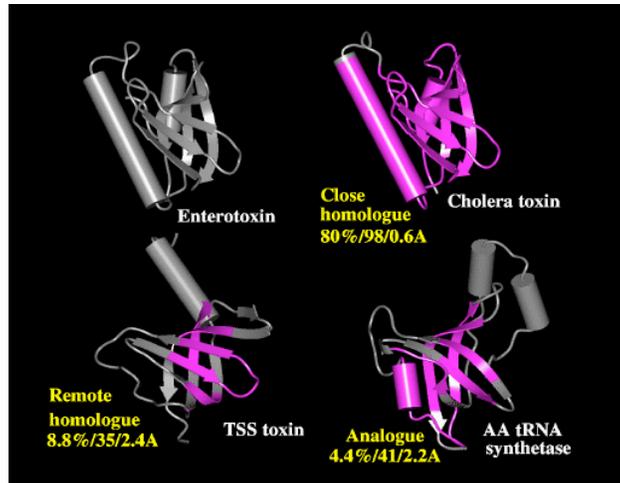


Homology and Analogy

- 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 divided 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 descendants, such as hemoglobin).

Homology and Analogy



The RCSB Protein Data Bank - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.rcsb.org/pdb/

Back Forward Reload Stop

Home Bookmarks Release Notes Plug-ins Extensions Support Mozilla Community Drop_Off Locator

RCSB PDB
PROTEIN DATA BANK

RCSB Home wwPDB Home Contact Us

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

DEPOSIT data
DOWNLOAD files
browse LINKS
BETA TEST new features
BETA XML files

Current Holdings
24785 Structures
Last Update: 23-Mar-2004
PDB Statistics

Search the Archive

Enter a PDB ID or keyword [Query Tutorial](#)

PDB ID Authors Full Text Search
 match exact word remove similar sequences

[SearchLite](#) keyword search form with examples
[SearchFields](#) customizable search form
[Status Search](#) find entries awaiting release

News [Complete News Newsletter](#) [pdb-I Archive Subscribe](#)

23-Mar-2004
PDB Focus: XML Data Files All released PDB entries are available in XML format from the PDB beta FTP site at

PDB Mirrors
Please bookmark a mirror site
[San Diego Supercomputer Center, UCSD*](#)
[Rutgers University*](#)
[Center for Advanced Research in Biotechnology, NIST*](#)
[Cambridge Crystallographic Data Centre, UK](#)
[National University of Singapore](#)
[Osaka University, Japan](#)
[Max Delbrück Center for Molecular Medicine, Germany](#)

[OCA / PDB Lite](#) [MORE...](#)

*RCSB partner

In citing the PDB please refer to:
H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Blat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The

Molecule of the Month:
The Calcium Pump

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced

start Microsoft PowerPoint ... The RCSB Protein Dat... 11:29 PM

The RCSB Protein Data Bank

The Calcium Pump

Every time we move a muscle, it requires the combined action of trillions of myosin motors. Our muscle cells use calcium ions to coordinate this massive molecular effort. When a muscle cell is given the signal to contract from its associated nerves, it releases a flood of calcium ions from a special intracellular container, the sarcoplasmic reticulum, that surrounds the bundles of actin and myosin filaments. The calcium ions rapidly spread and bind to troponin on the actin filaments. They shift shape slightly and allow myosin to bind and begin climbing up the filament. These trillions of myosin motors will continue clanking, contracting the muscle, until the calcium is removed.

Relaxation

The calcium pump allows muscles to relax after this frenzied wave of calcium-induced contraction. The pump is found in the membrane of the sarcoplasmic reticulum. In some cases, it is so plentiful that it may make up 90% of the protein there. Powered by ATP, it pumps calcium ions back into the sarcoplasmic reticulum, reducing the calcium level around the actin and myosin filaments and allowing the muscle to relax. Calcium ions are also used for signaling inside other cells, and similar pumps are found in the cell membrane of most cells. They constantly work to reduce the amount of calcium to very low levels, preparing the cell. Then, at a moment's notice, the cell can allow a flood of calcium to enter, spreading the signal to all corners.

Pumping Calcium

The calcium pump is an amazing machine with several moving parts. It is found in the membrane of the sarcoplasmic reticulum. It has a binding site on the outside of the sarcoplasmic reticulum and a pump that pushes calcium ions back into the sarcoplasmic reticulum.

Exploring the Structure

The calcium binding site is in a tunnel formed by four alpha helices, which cross straight through the membrane. This illustration, from PDB entry [1mdh](#), shows a view down the helices. The two calcium ions, shown as blue-green spheres, are held by a collection of amino acids, shown in balls-and-sticks, that coordinate it from all sides. The protein is far less stable when these calcium ions are removed. You can look at the structure of the calcium-free form in PDB entry [1jwo](#). It was solved by adding a drug molecule that binds near the calcium-binding site and freezes the protein into a stable, but non-functioning, form.

Structure Explorer - 4MDH

Summary Information

Title: Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5-Å resolution.
Compound: Cytoplasmic Malate Dehydrogenase (E.C. 1.1.1.37)
Authors: J. J. Birktoft, L. J. Banaszak
Exp. Method: X-ray Diffraction
Classification: Oxidoreductase(Nad(A)-Choh(D))
EC Number: [1.1.1.37](#) (Malate dehydrogenase)

Source: [Sus scrofa](#)

Primary Citation: [Birktoft, J. J., Rhodes, G., Banaszak, L. J.:](#) Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5-Å resolution. *Biochemistry* 28 pp. 6065 (1989)

Deposition Date: 12-Apr-1989 **Release Date:** 19-Apr-1989

Resolution [Å]: 2.50 **R-Value:** 0.167

Space Group: P 2₁ 2₁ 2

Unit Cell: dim [Å]: a 139.20 b 86.60 c 58.80
angles [°]: alpha 90.00 beta 90.00 gamma 90.00

Summary Information
[View Structure](#)
[Download/Display File](#)
[Structural Neighbors](#)
[Geometry](#)
[Other Sources](#)
[Sequence Details](#)
[Previous version\(s\): 1MDH, 2MDH](#)
[Structure Factors \(compressed\)](#)

Explore
[SearchLite](#) [SearchFields](#)

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
```

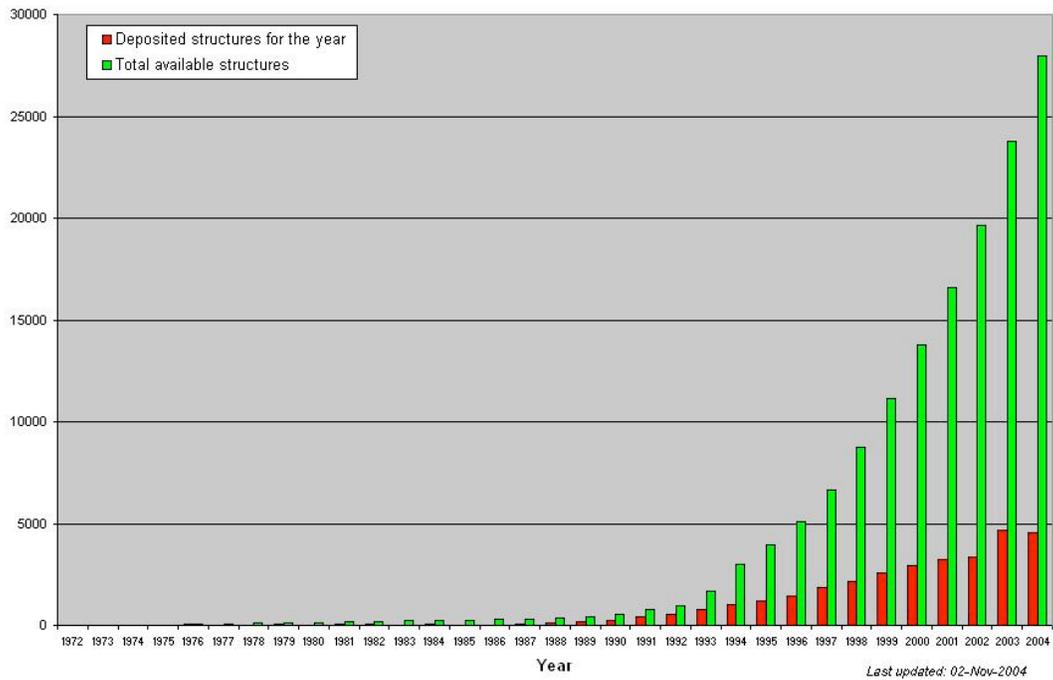
```
AUTHOR    A.VRIELINK,P.I.LARIO
REVDATE   1 25-FEB-03 1MXT 0
JRNL      AUTH P.I.LARIO,N.SAMPSON,A.VRIELINK
JRNL      TITL SUB-ATOMIC RESOLUTION CRYSTAL STRUCTURE OF
JRNL      TITL 2 CHOLESTEROL OXIDASE: WHAT ATOMIC RESOLUTION
JRNL      TITL 3 CRYSTALLOGRAPHY REVEALS ABOUT ENZYME MECHANISM AND
JRNL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

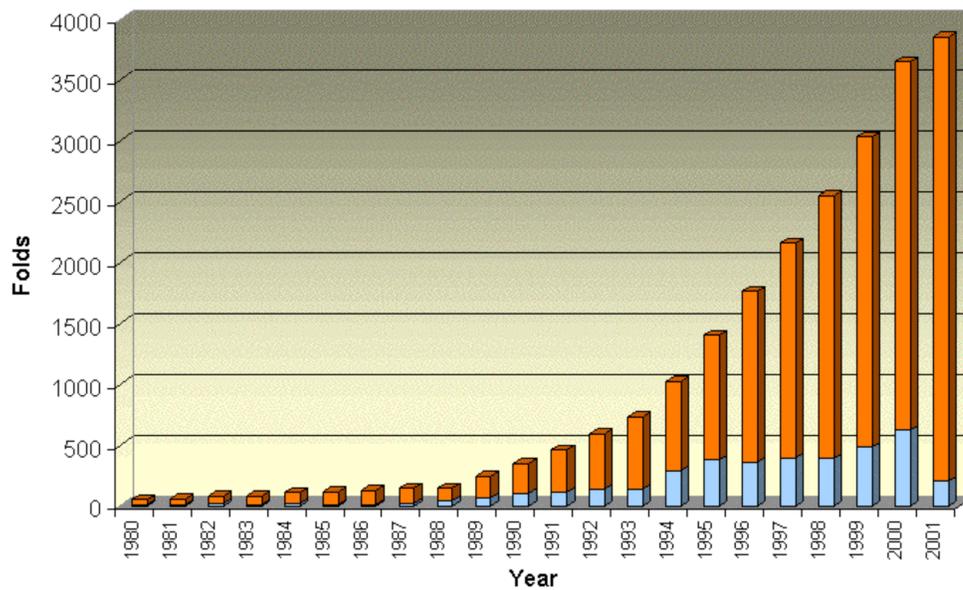
The body of the PDB file contains information about the atoms in the structure:

```
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
```

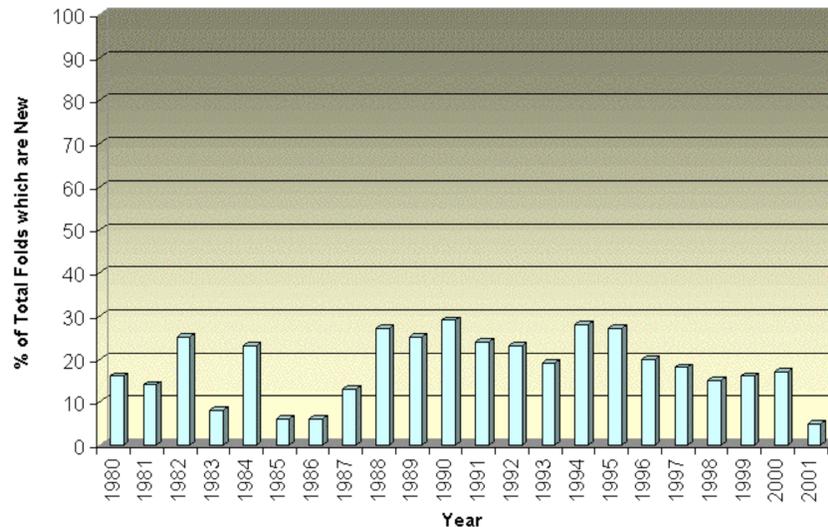
Growth of Structural Data



Unique Folds in the PDB



New Folds Become Rare



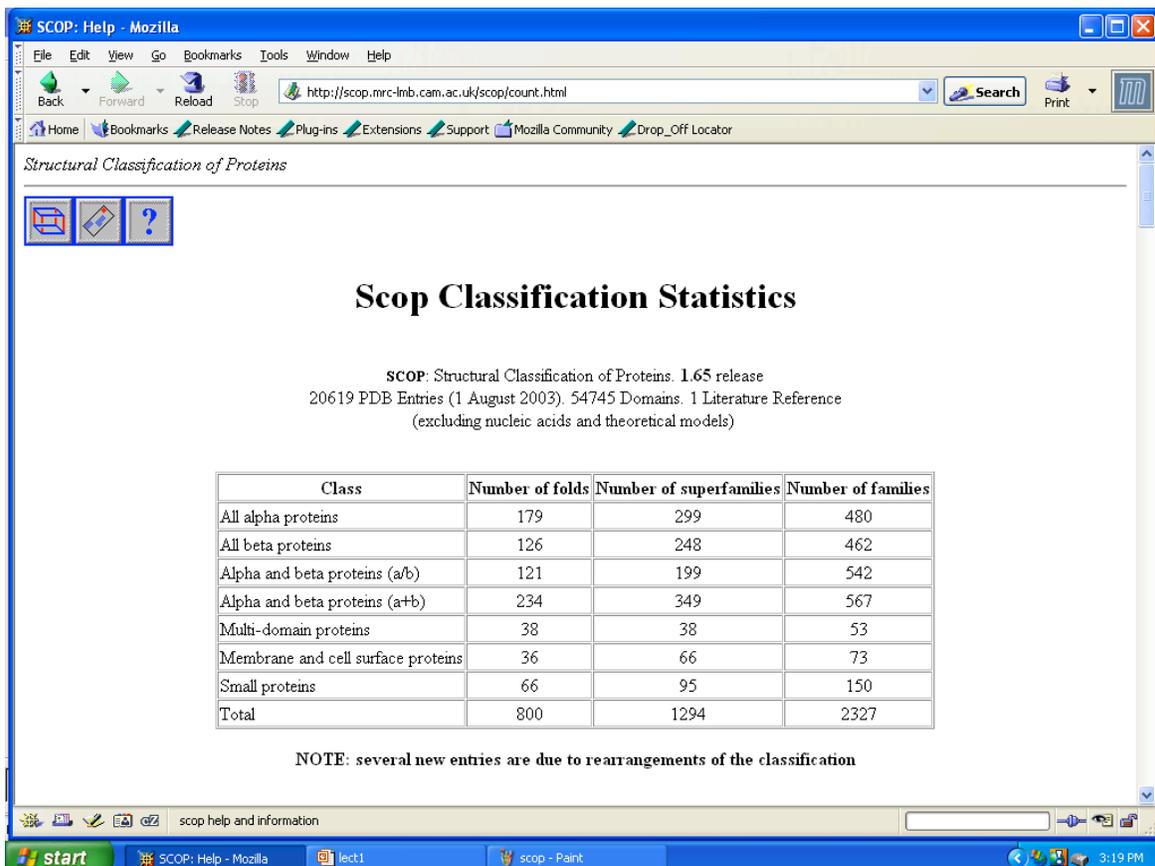
SCOP

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.

SCOP Levels

- **Family:** Clear evolutionary 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.



Structural Classification of Proteins

Scop Classification Statistics

SCOP: Structural Classification of Proteins. 1.65 release
20619 PDB Entries (1 August 2003). 54745 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of 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

NOTE: several new entries are due to rearrangements of the classification

Some Maybe Surprising Results

5NLL

1AMO

1CHN

1FNB



Flavodoxin

Cytochrome reductase

Protein CHEY

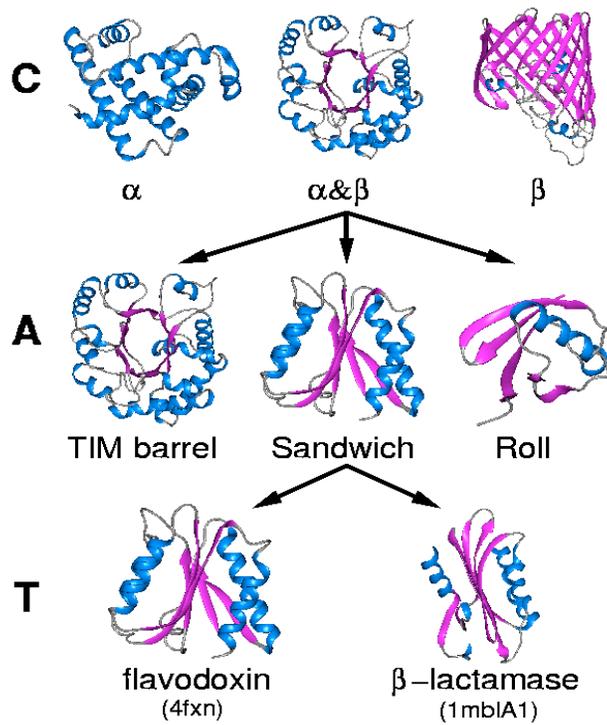
Ferredoxin reductase

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.

The CATH Hierarchy



SCOP versus CATH

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

The screenshot shows the CATH Releases website in a Mozilla browser window. The page title is "CATH Releases" and the URL is "http://www.biochem.ucl.ac.uk/bsm/cath/releases.html". The page content includes a search box, navigation links, and a table of release statistics for CATH v2.5.1.

CATH v2.5.1

Version	2.5.1							
Date	28-01-2004							
Mainly Alpha	5	227	428	948	1713	3946	10155	
Mainly Beta	19	139	292	951	2344	5011	14259	
Alpha Beta	12	368	648	2010	3631	8639	23025	
Few Secondary Structures	1	86	91	114	225	378	952	
Multi-domain chains	1	1053	1057	1071	2186	5801	12471	
Preliminary single domain assignments	1	371	374	422	479	789	1663	
Multi-domain domains	2	31	31	49	67	139	287	
CATH-35 Sequence families	1	997	997	997	1108	2154	3431	
Fragments from multi-chain domains	1	28	28	30	33	56	106	

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).

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

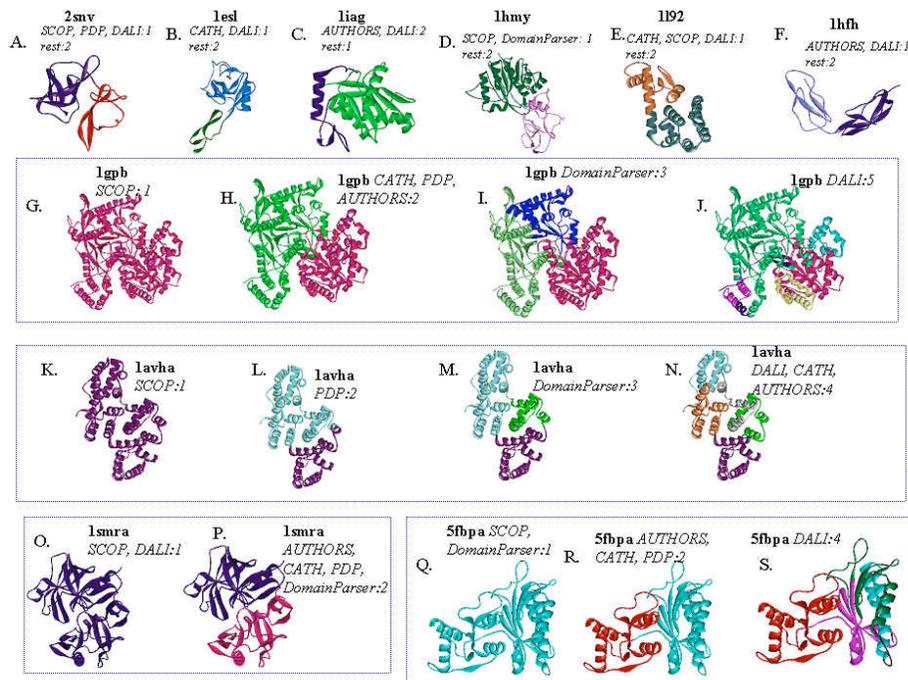
- The Protein Domain Parser (PDP) uses compactness as a chief principle.

<http://123d.ncifcrf.gov/pdp.html>

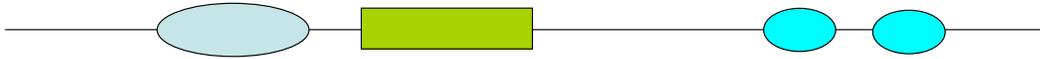
- DomainParser is graph theory based. The underlying principle used is that residue-residue contacts are denser within a domain than between domains.

<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.

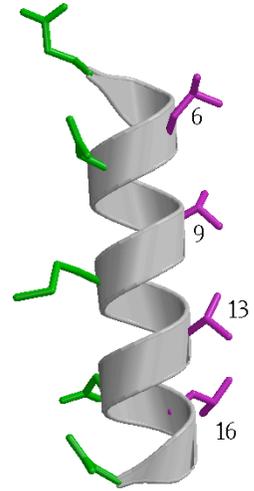
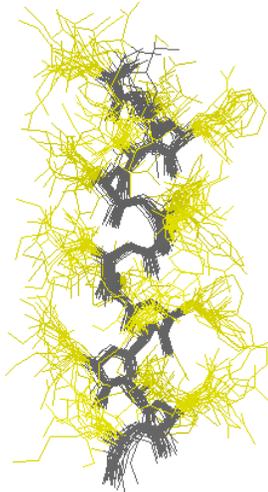
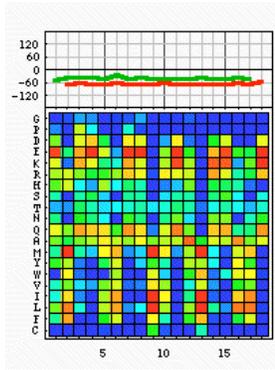
PRODOM:

<http://protein.toulouse.inra.fr/prodom/current/html/home.php>

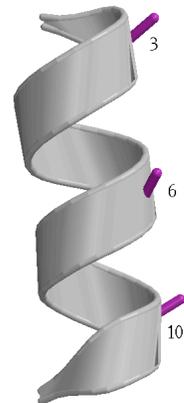
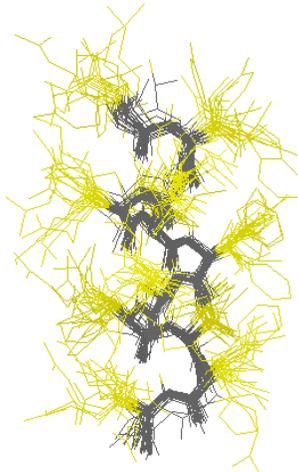
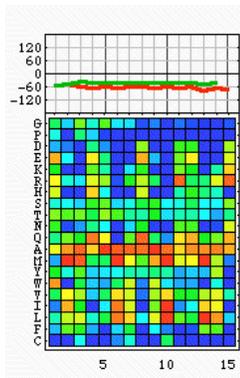
PFAM:

<http://www.sanger.ac.uk/Software/Pfam/>

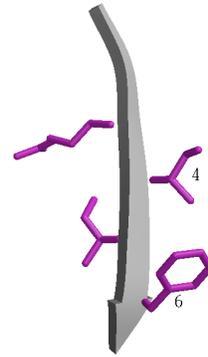
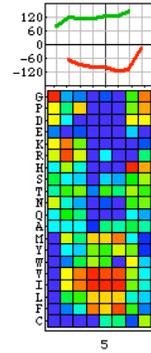
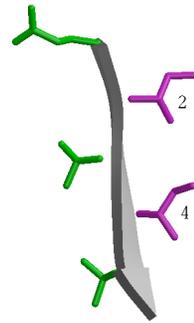
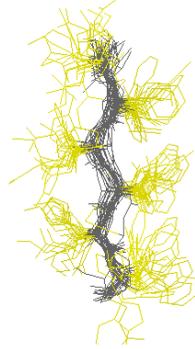
I-Sites



I-Sites



I-Sites



I-Sites

