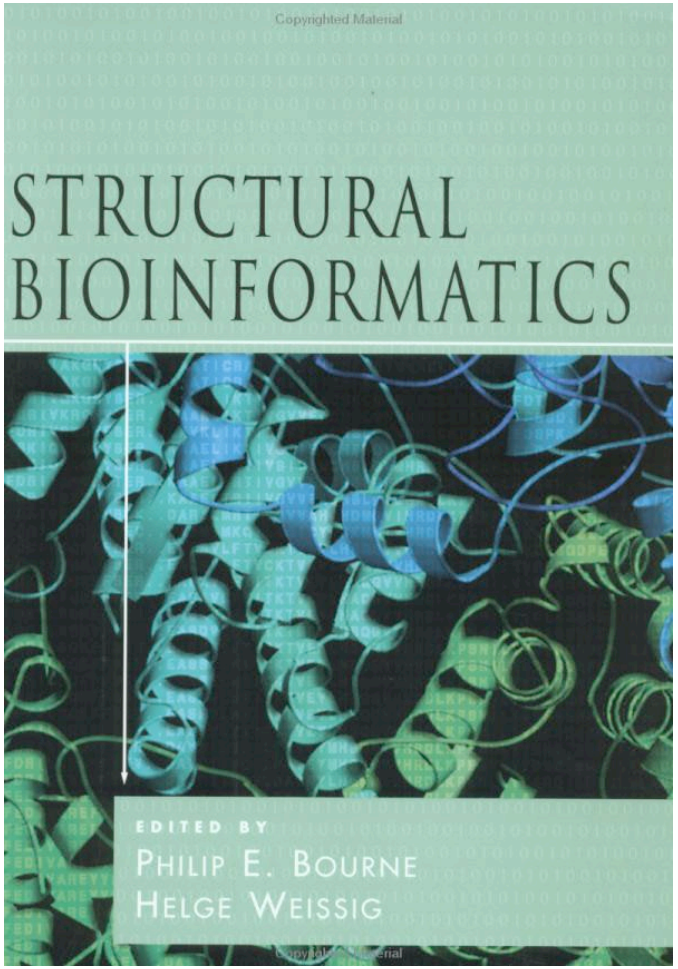
Protein Structure: Data Bases and Classification

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Reference

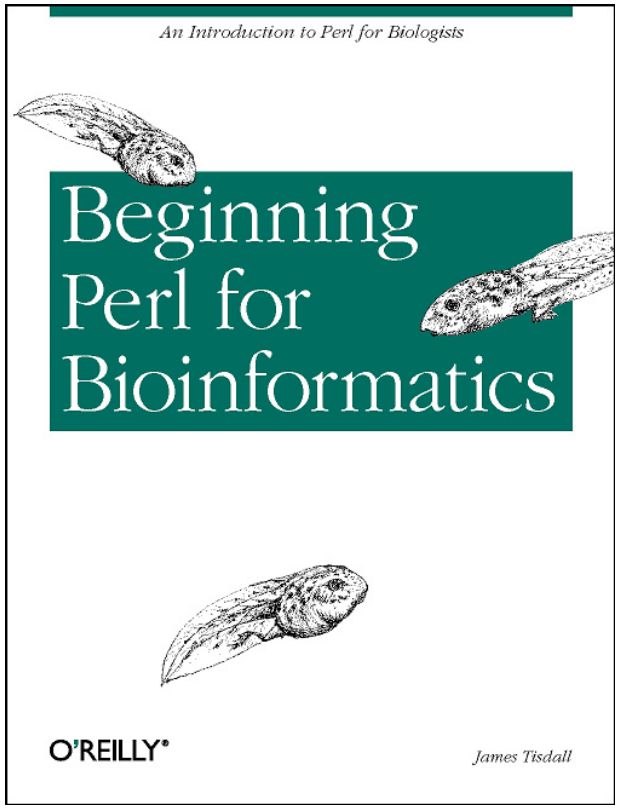


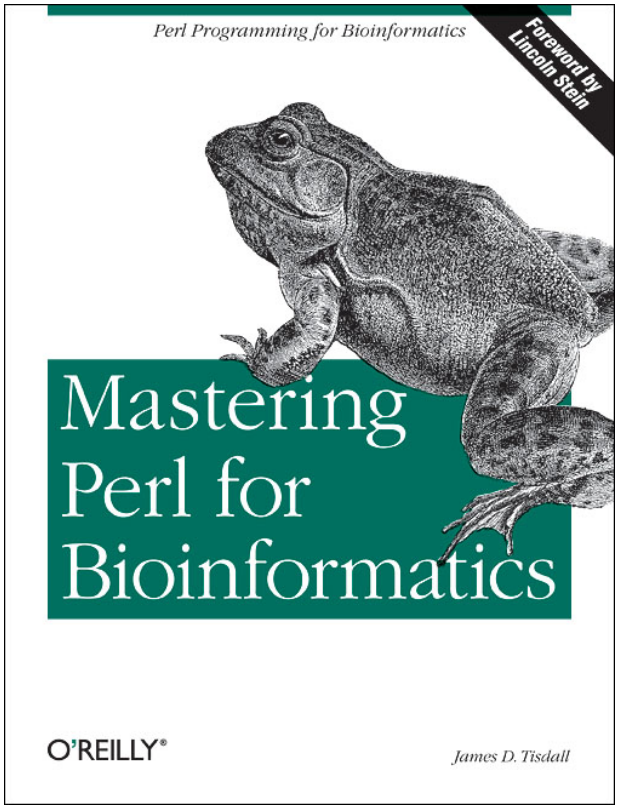
Bourne and Weissig

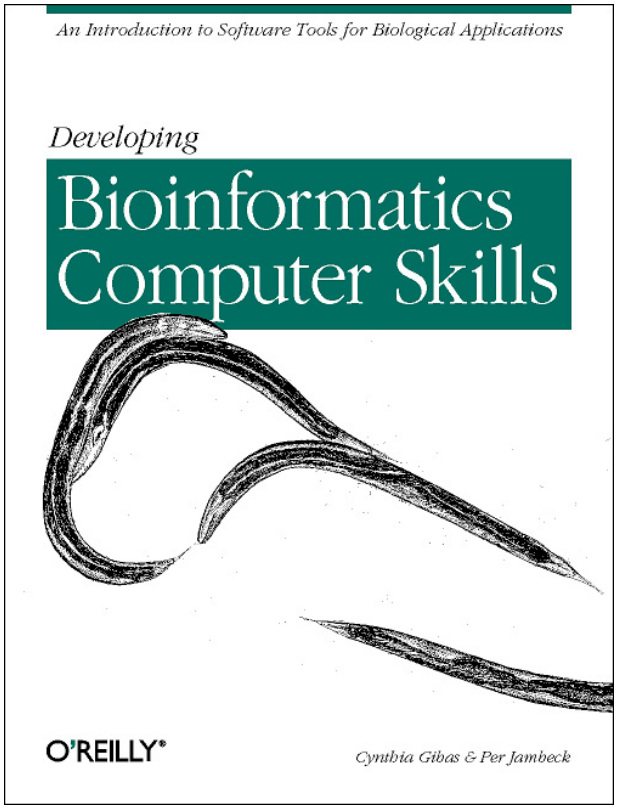
Structural Bioinformatics

Wiley, 2003

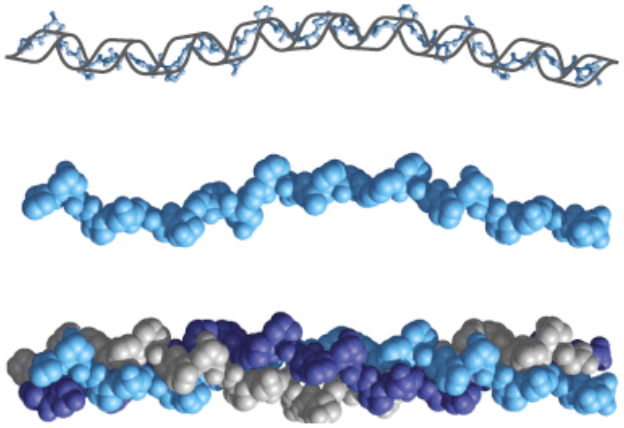
More References

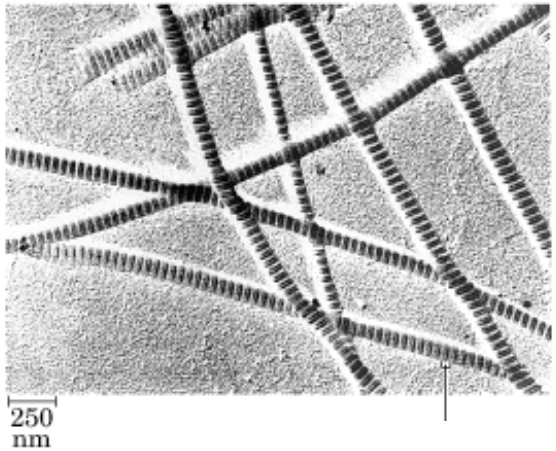




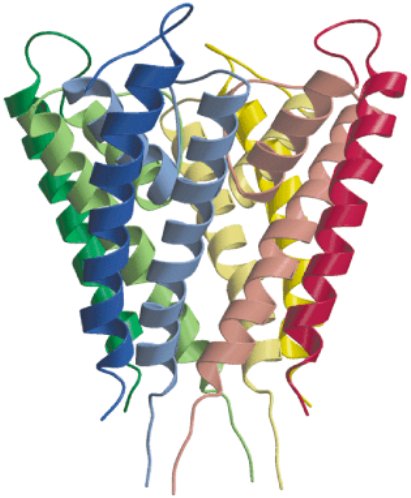


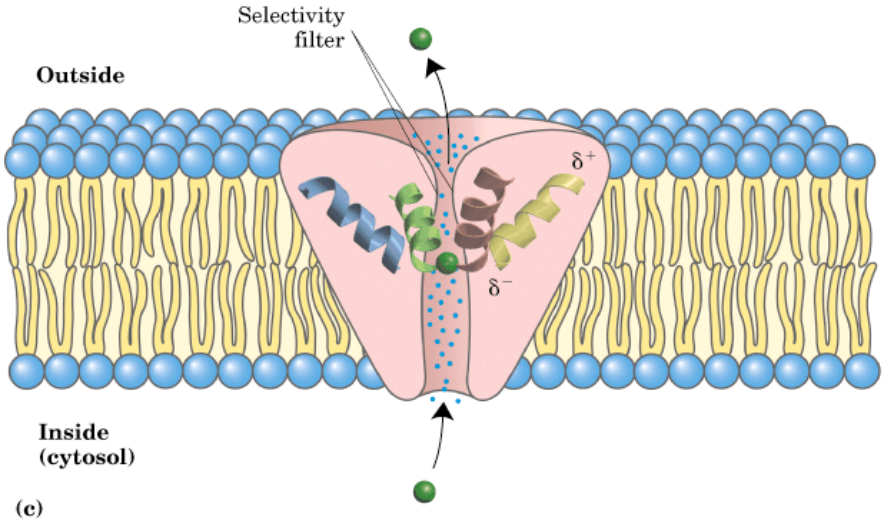
Structural Proteins



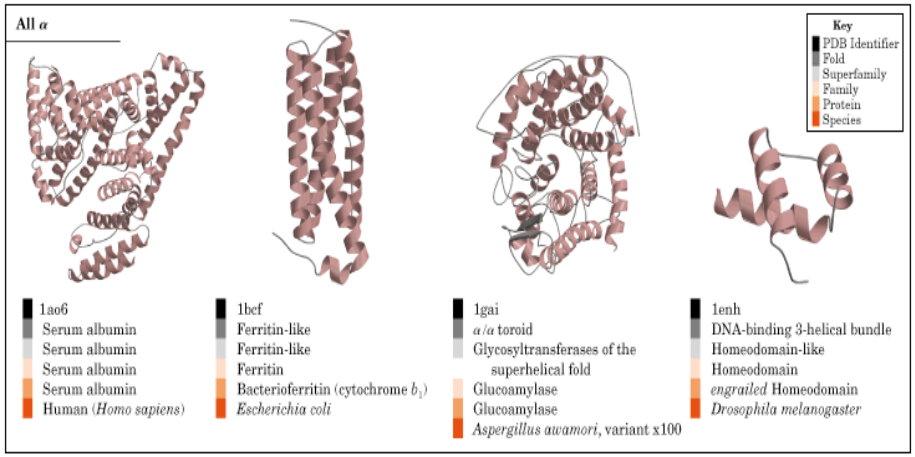


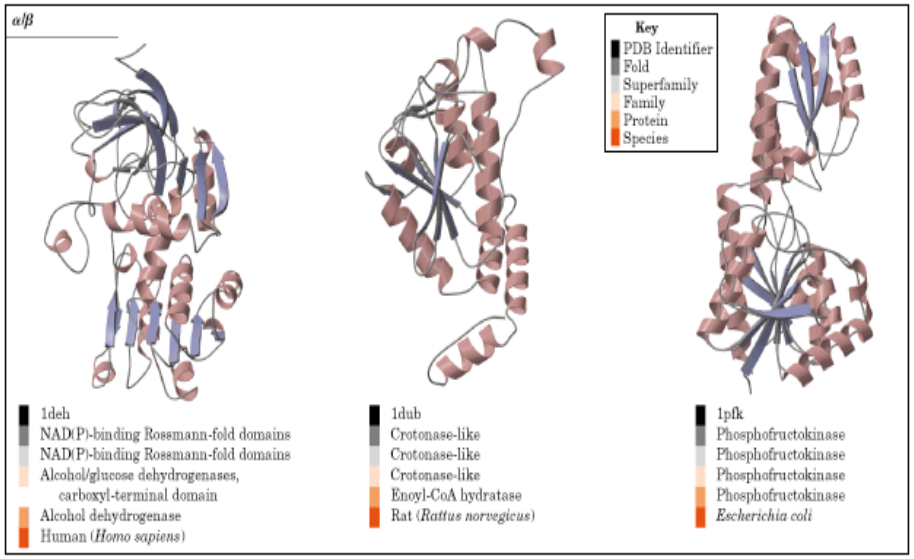
Membrane Proteins

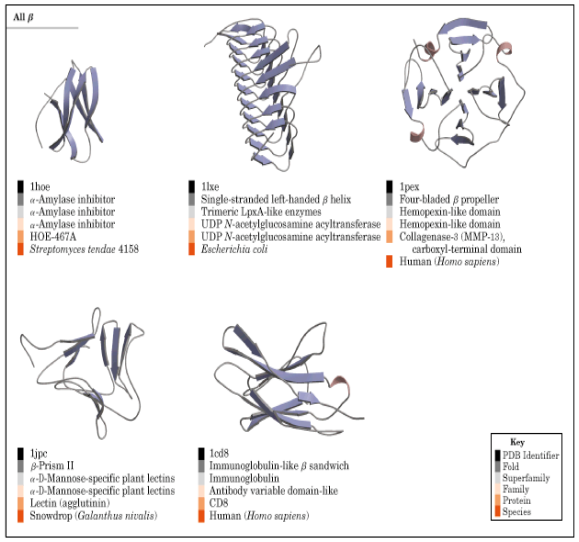


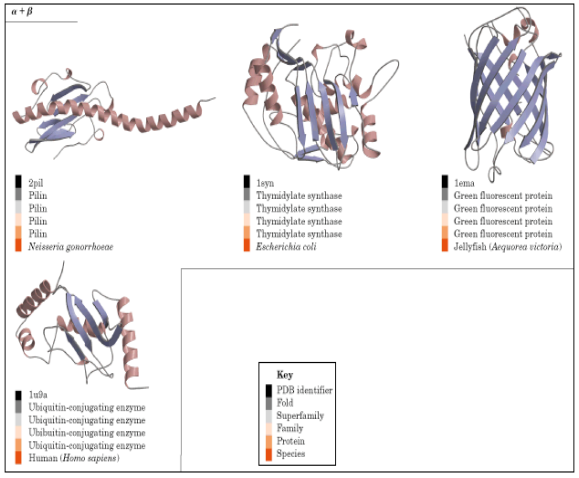


Globular Proteins





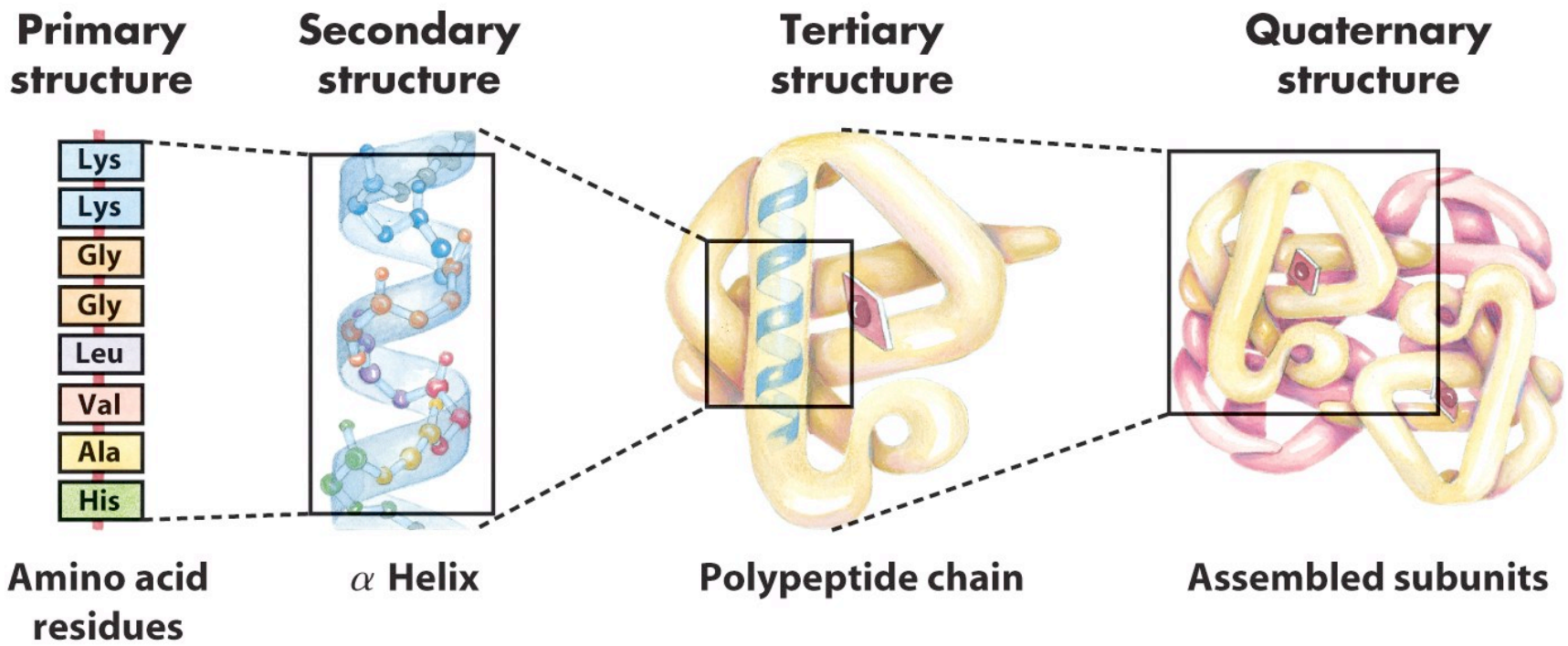




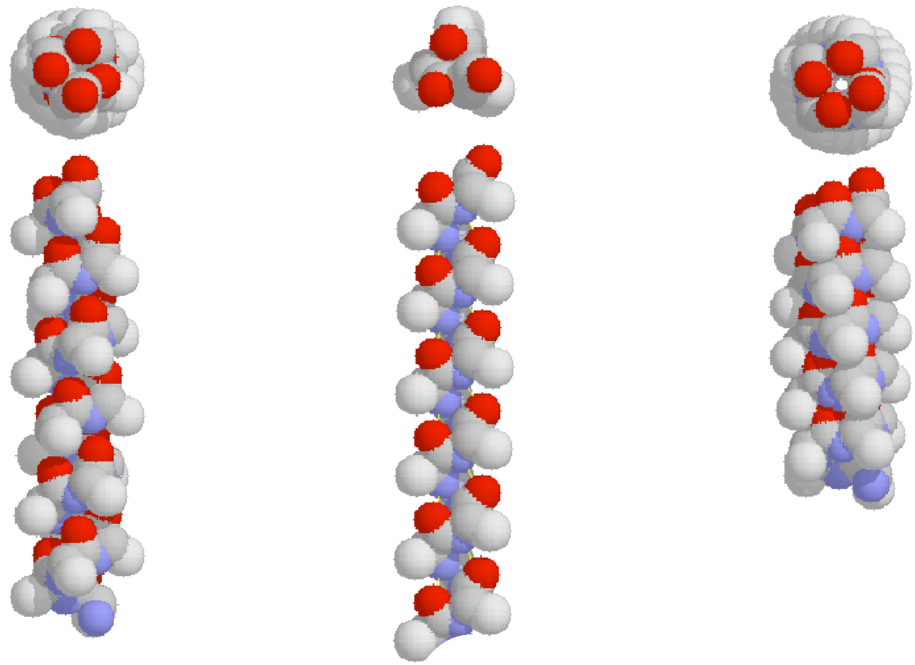
Terminology

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

Hierarchy of Protein Structure

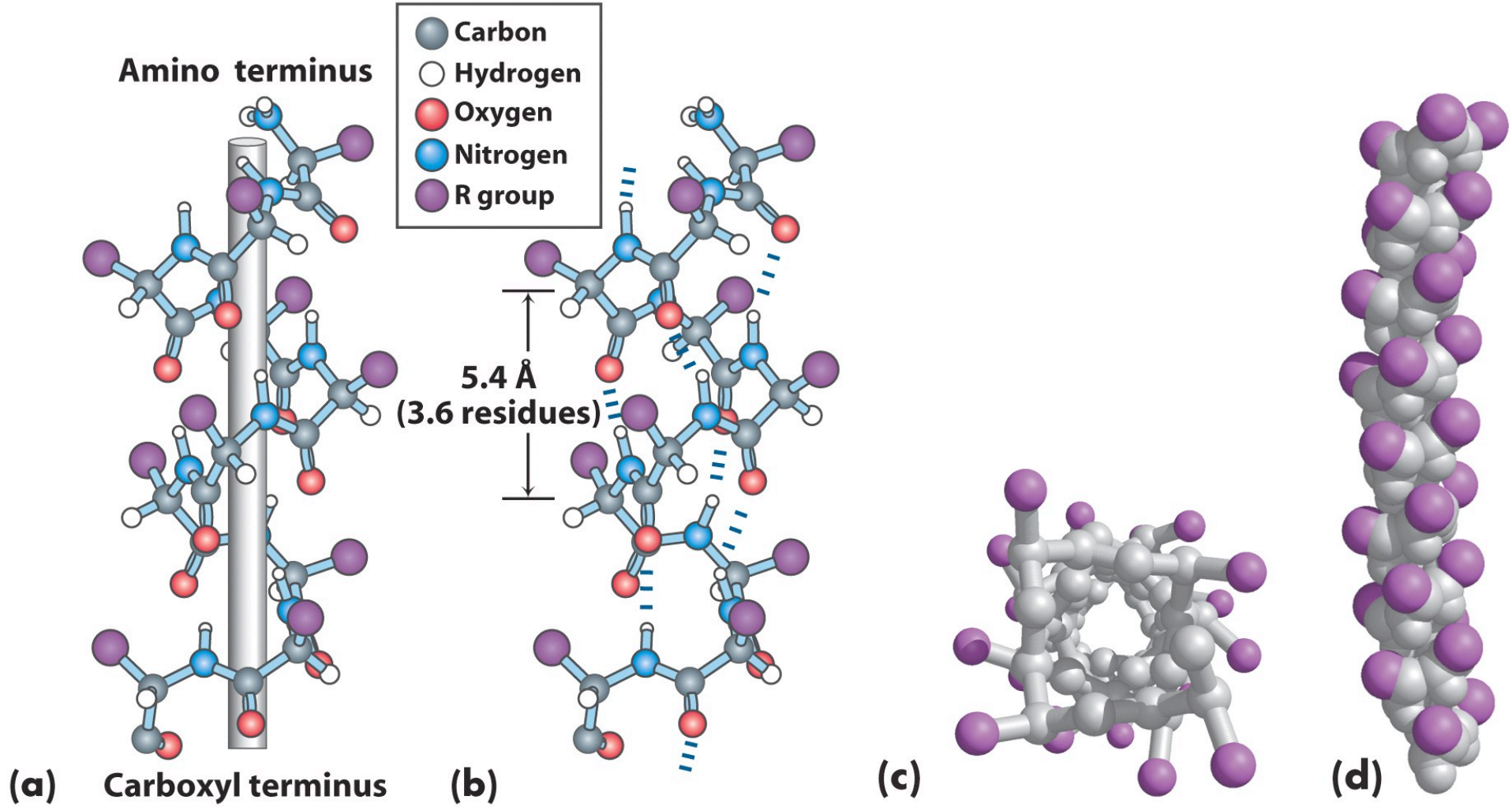


Helices



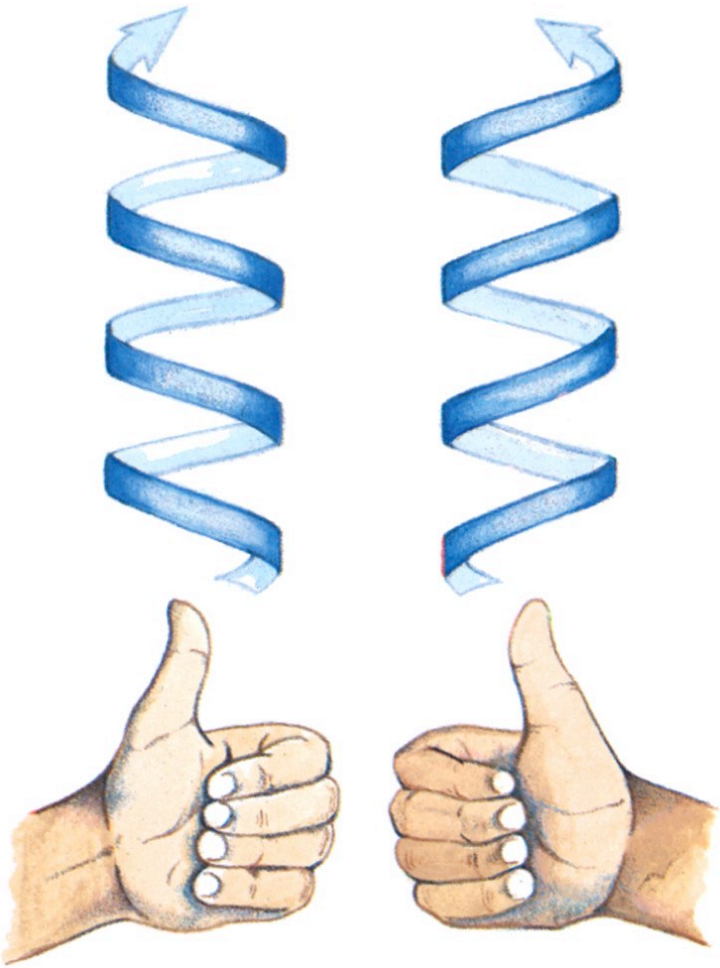
	α	3.10	π
Amino acids/turn:	**3.6**	**3.0**	**4.4**
Frequency	**~97%**	**~3%**	**rare**
H-bonding	***i*, *i*+4**	***i*, *i*+3**	***i*, *i*+5**

α-helices

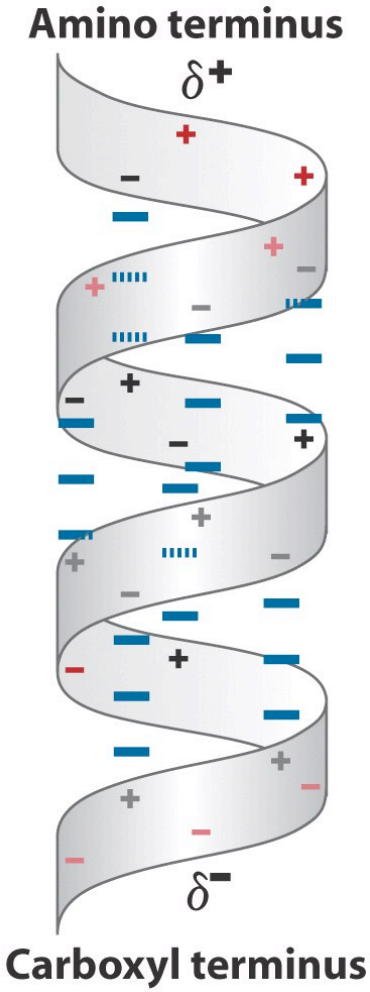

Carbon
Hydrogen
Oxygen
Nitrogen
R group
Amino terminus
Carboxyl terminus
5.4 Å
(3.6 residues)
(a)
(b)
(c)
(d)

α-helices

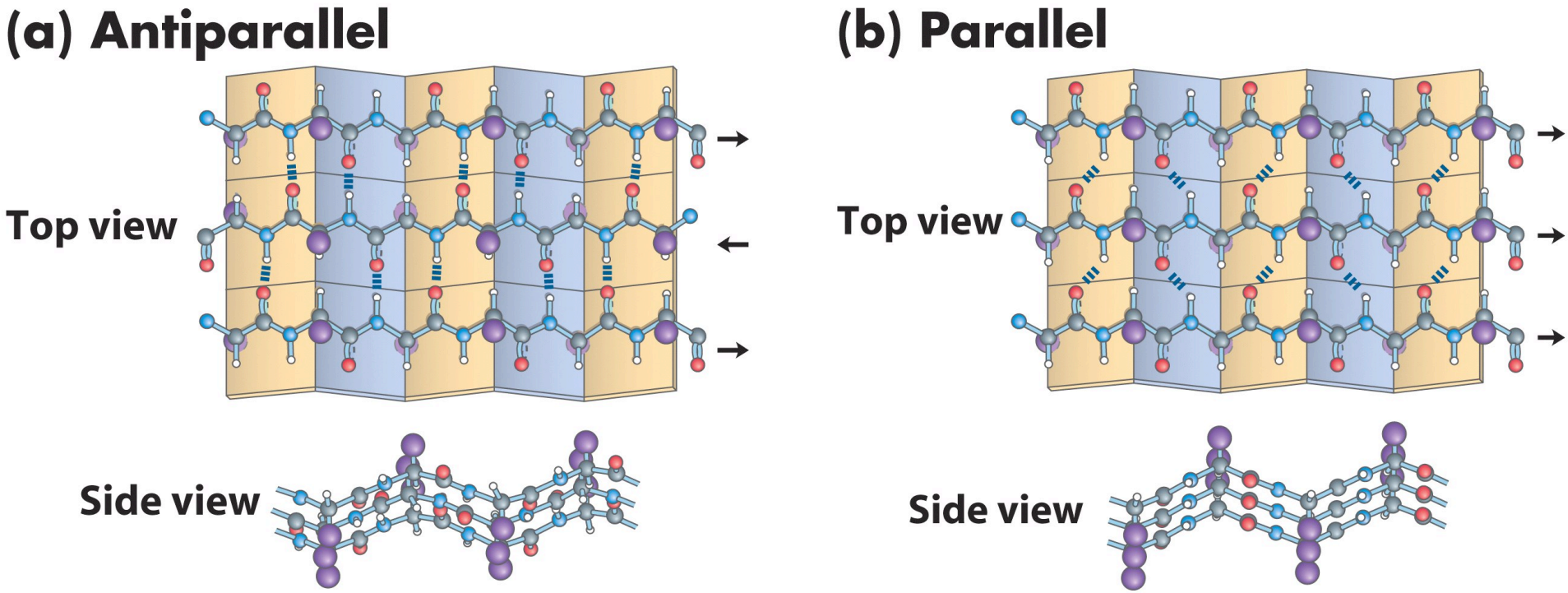
α-helices have handedness:



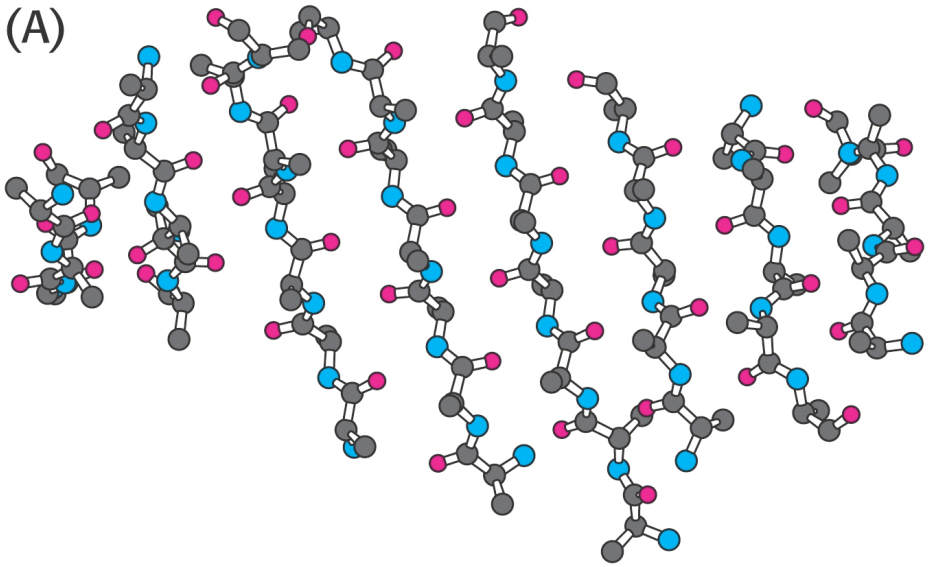
α-helices have a dipole:



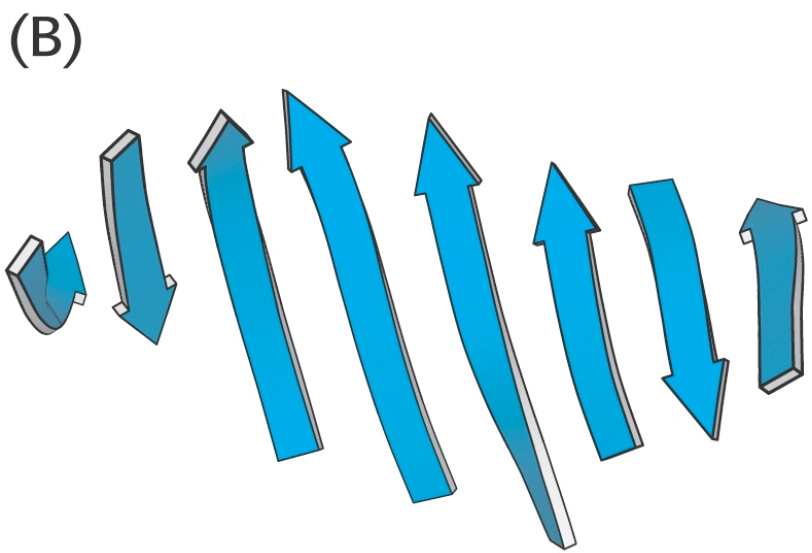
β-sheets

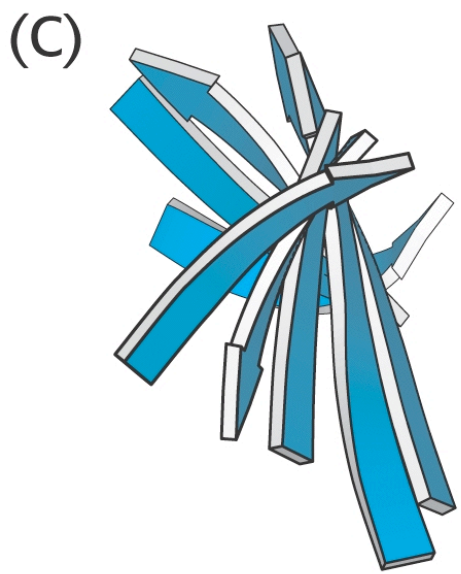

(a) Antiparallel
Top view
Side view
(b) Parallel
Top view
Side view

β-sheets

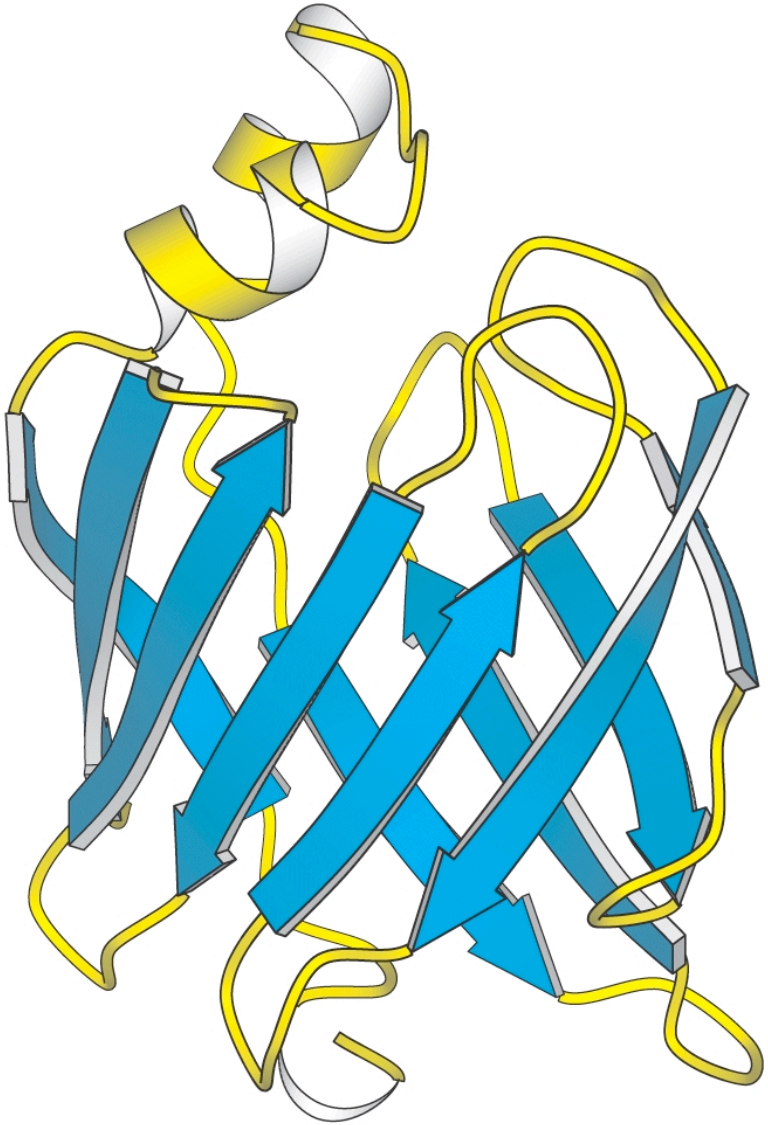


Have a right-handed twist!



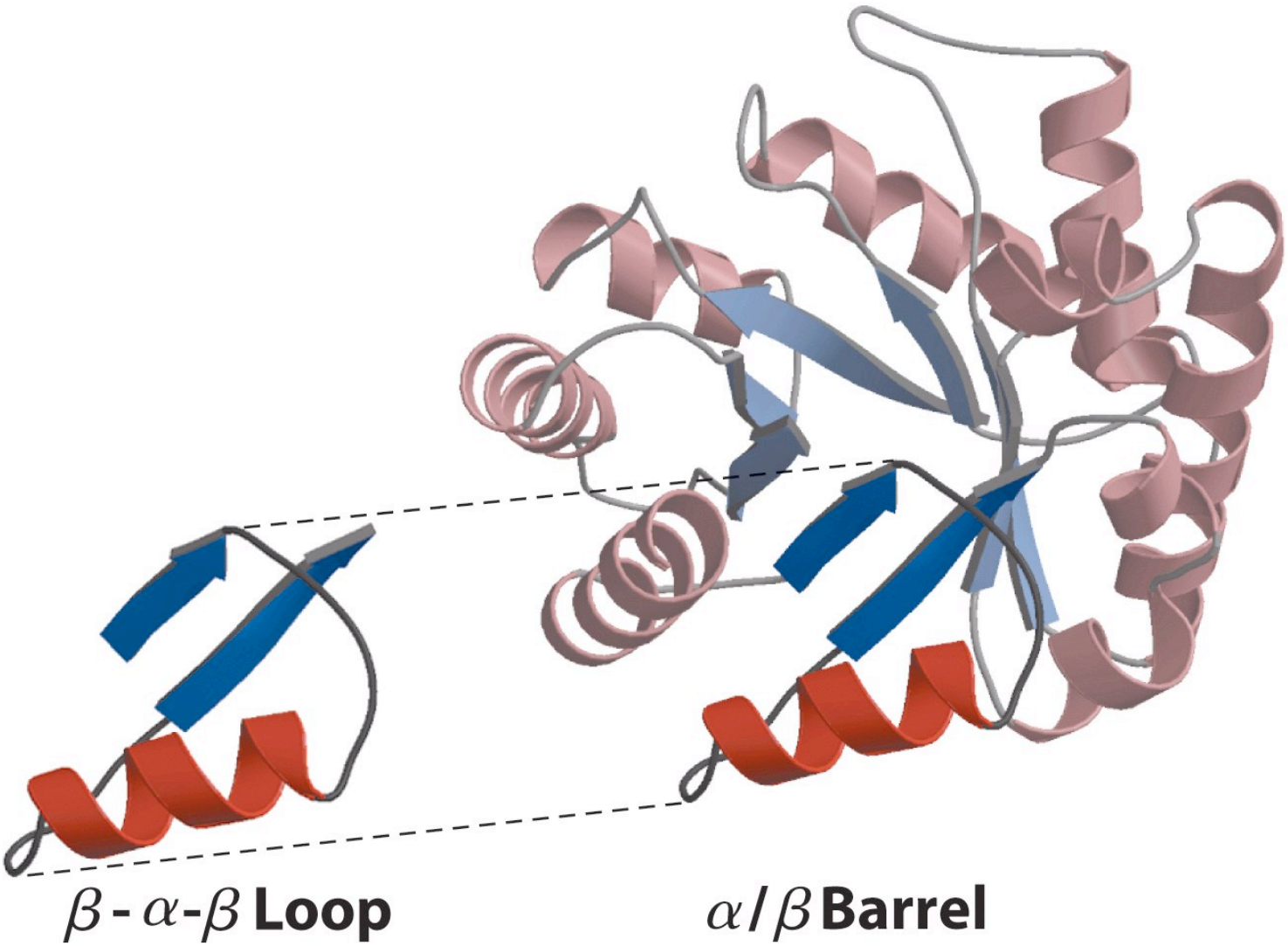


β-sheets



Can form higher level structures!

Super Secondary Structure Motifs



What is a Domain?

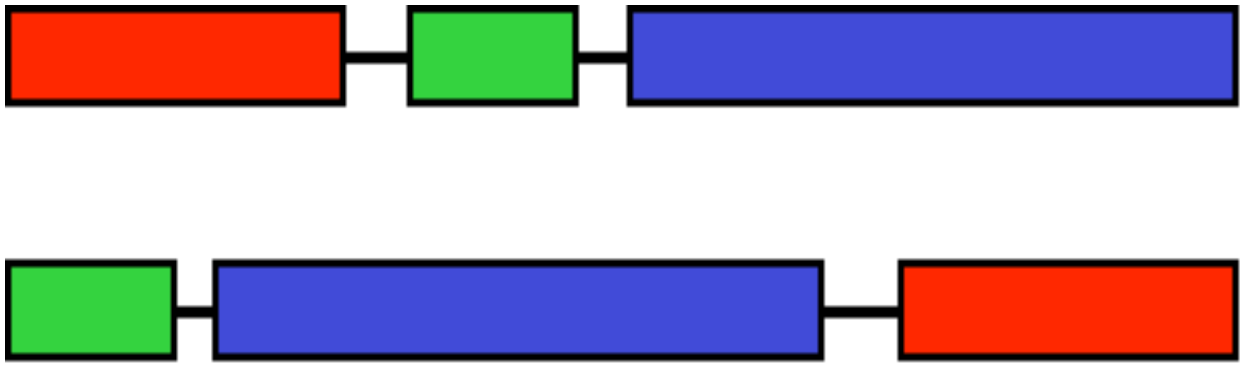
Richardson (1981):

Within 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!

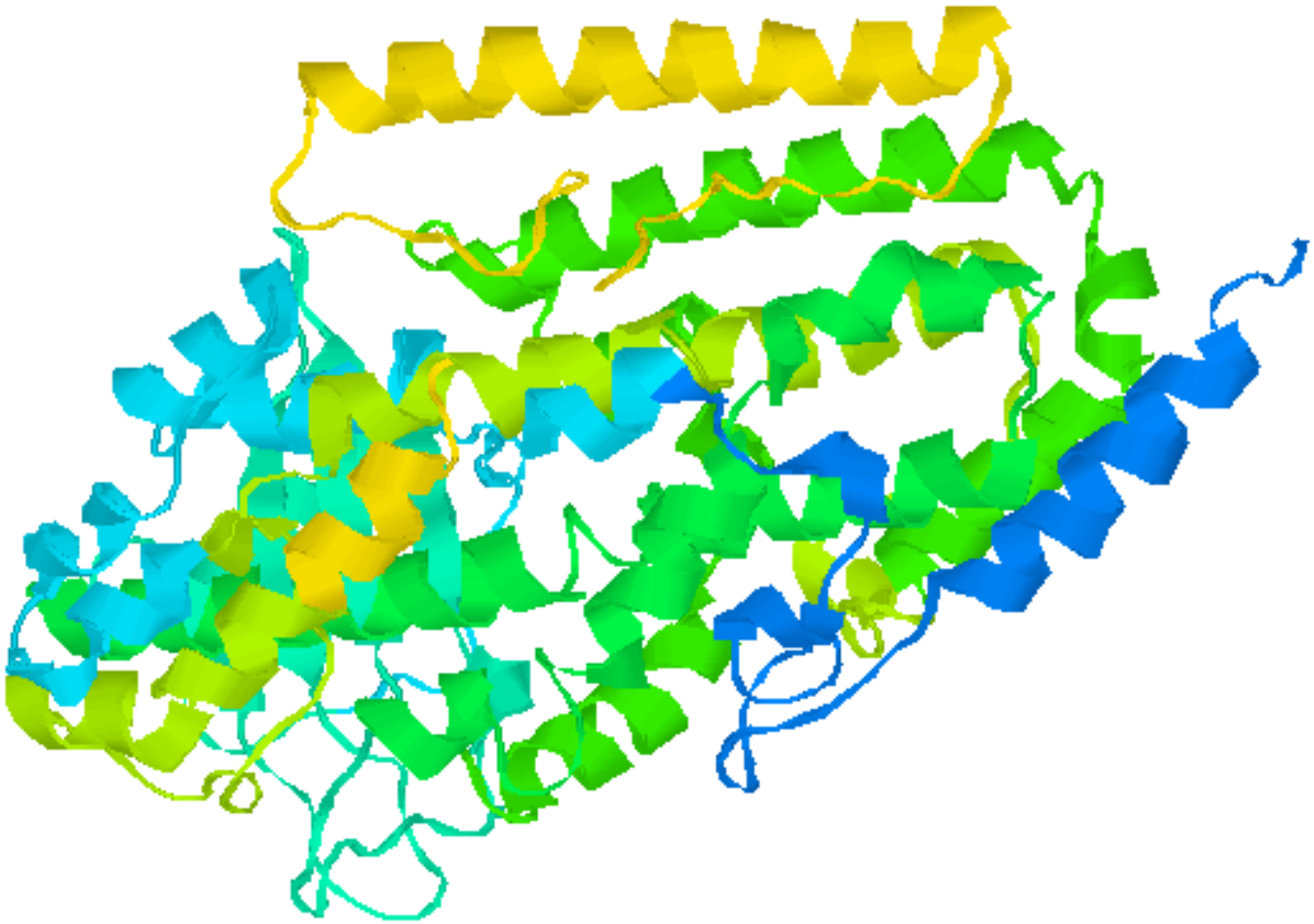
Domain Size

- 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.

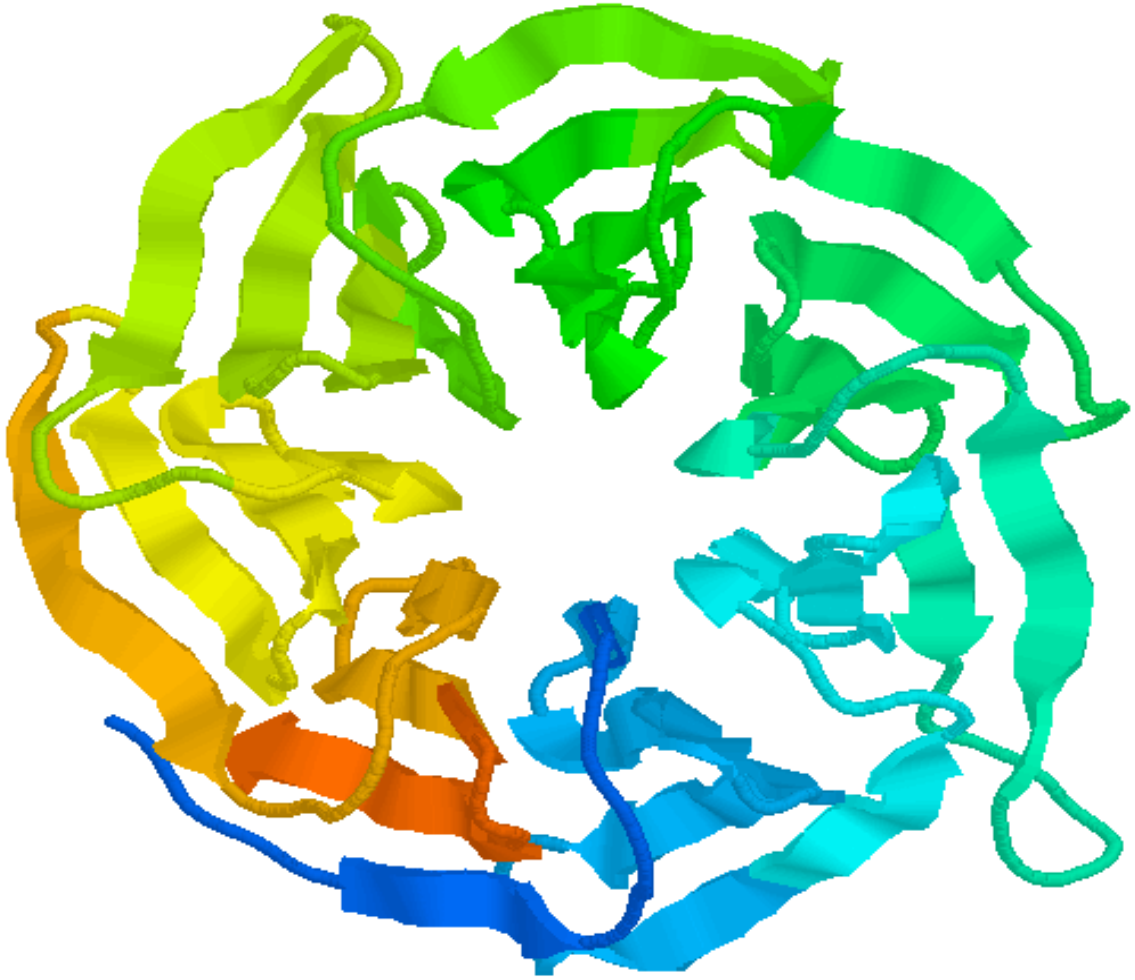
Examples are the zinc finger and a scorpion toxin.

Two Very Small Domains

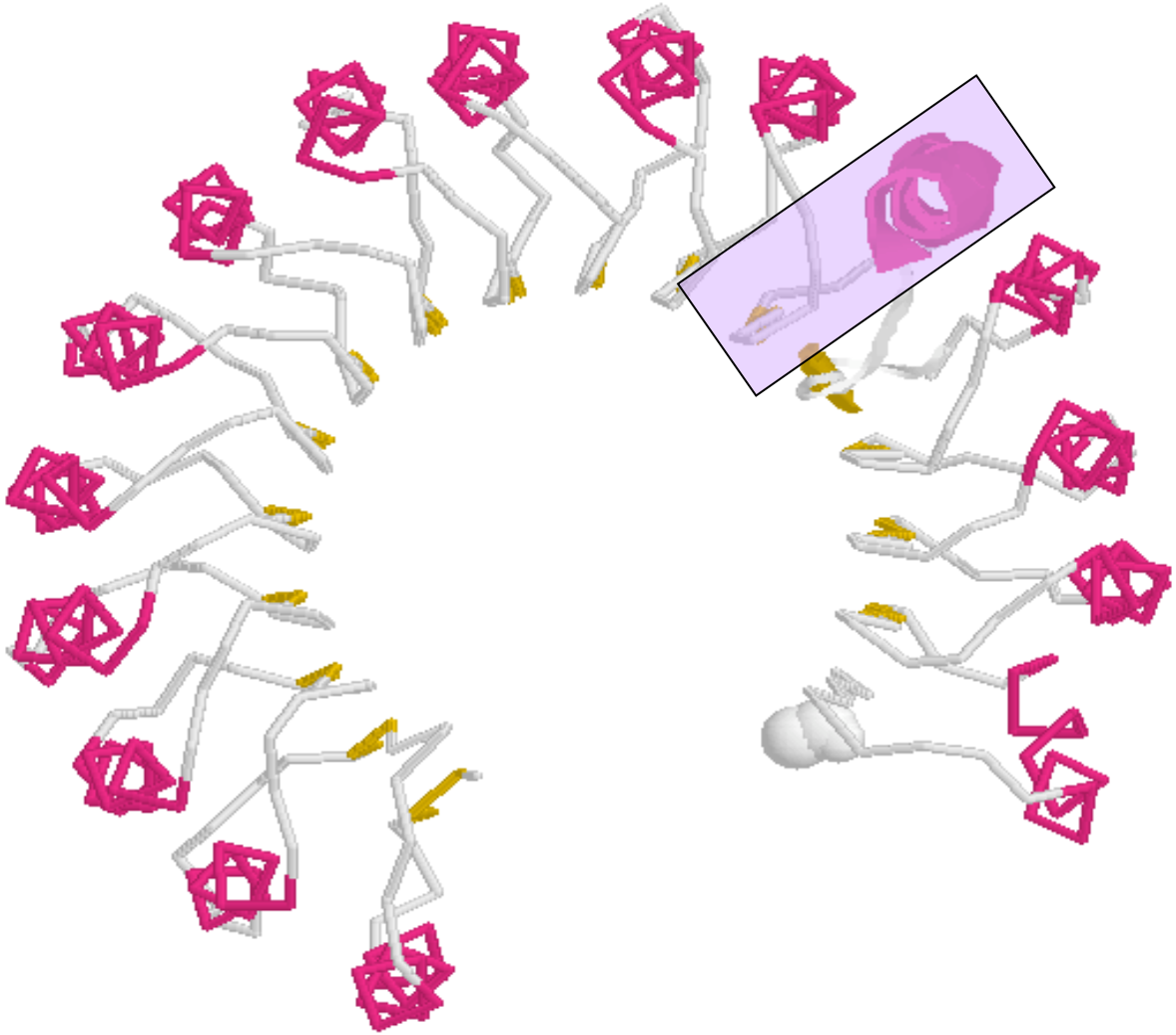
A Humdinger of a Domain



What's the Domain? (Part 1)



What’s the Domain? (Part 2)

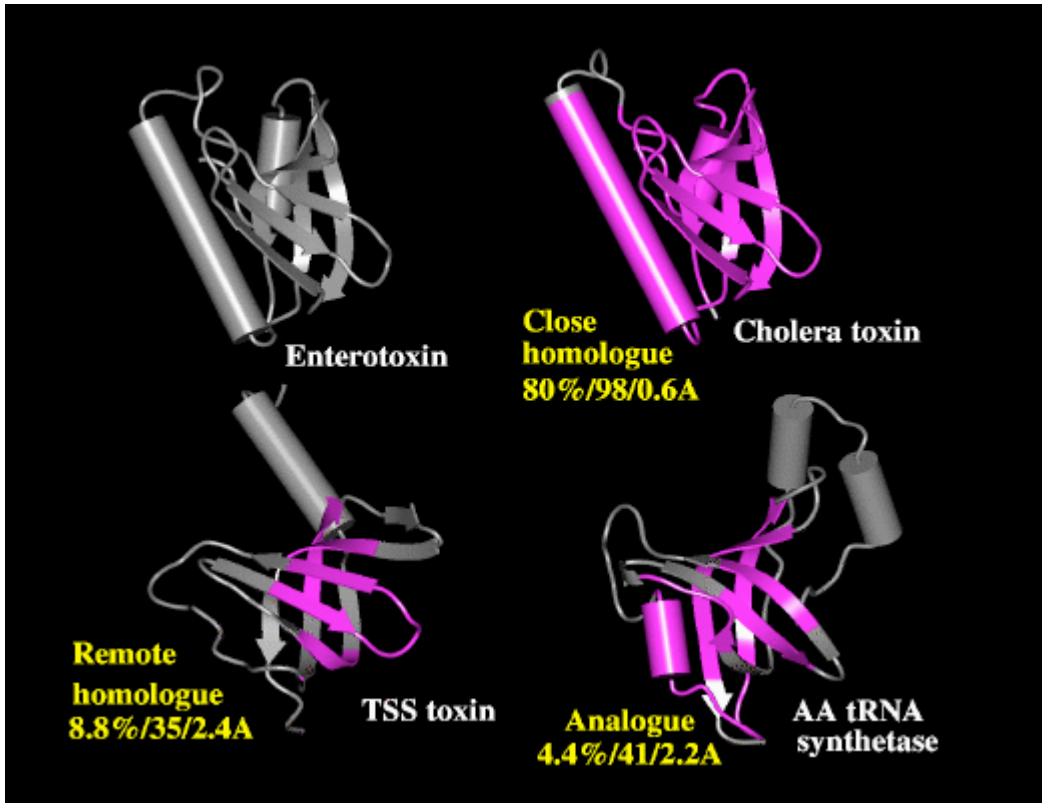


Homology and Analogy

- 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



RCSB Protein Data Bank

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

Ingo's Pond | News | Links | Running | Science | Travel

RCSB PDB PROTEIN DATA BANK

Contact Us | Help | Print Page

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Apr 11, 2006 there are 36012 Structures | PDB Statistics

PDB ID or keyword | Author | SEARCH | Advanced Search

Home | Search

- Home
- Tutorial About This Site
- Getting Started
- Download Files
- Deposit and Validate
- Structural Genomics
- Dictionaries & File Formats
- Software Tools
- Educational Resources
- General Information
- Acknowledgements
- Frequently Asked Questions
- Known Problems
- Report Bugs/Comments

Welcome to the RCSB PDB

The **RCSB** PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the **wwPDB** whose mission is to ensure that the PDB archive remains an international resource with uniform data.

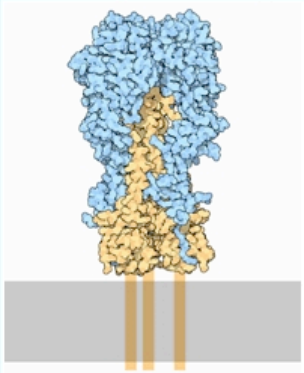
This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found **here**.

A **narrated tutorial** illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Hemagglutinin



Influenza virus is a dangerous enemy. Normally, the immune system fights off infections, eradicating the viruses and causing a few days of miserable flu symptoms. Yearly flu vaccines prime our immune system, making it ready to fight the most common strains of influenza virus. But once every couple of decades, and new strain of influenza appears that is far more pathogenic, allowing it to spread rapidly. This happened at the end of World War I, and the resultant pandemic killed over 20 million people, more than twice the number of people that were killed in the war.

- **More ...**
- **Previous Features**

NEWS

- **Complete News**
- **Newsletter**
- **Discussion Forum**

11-Apr-2006
Validating structures saves deposition time

To lower the number of revisions and problems found during the annotation process, depositors should validate their structure, provide the correct and complete sequence, and run BLAST.

- Full Story ...

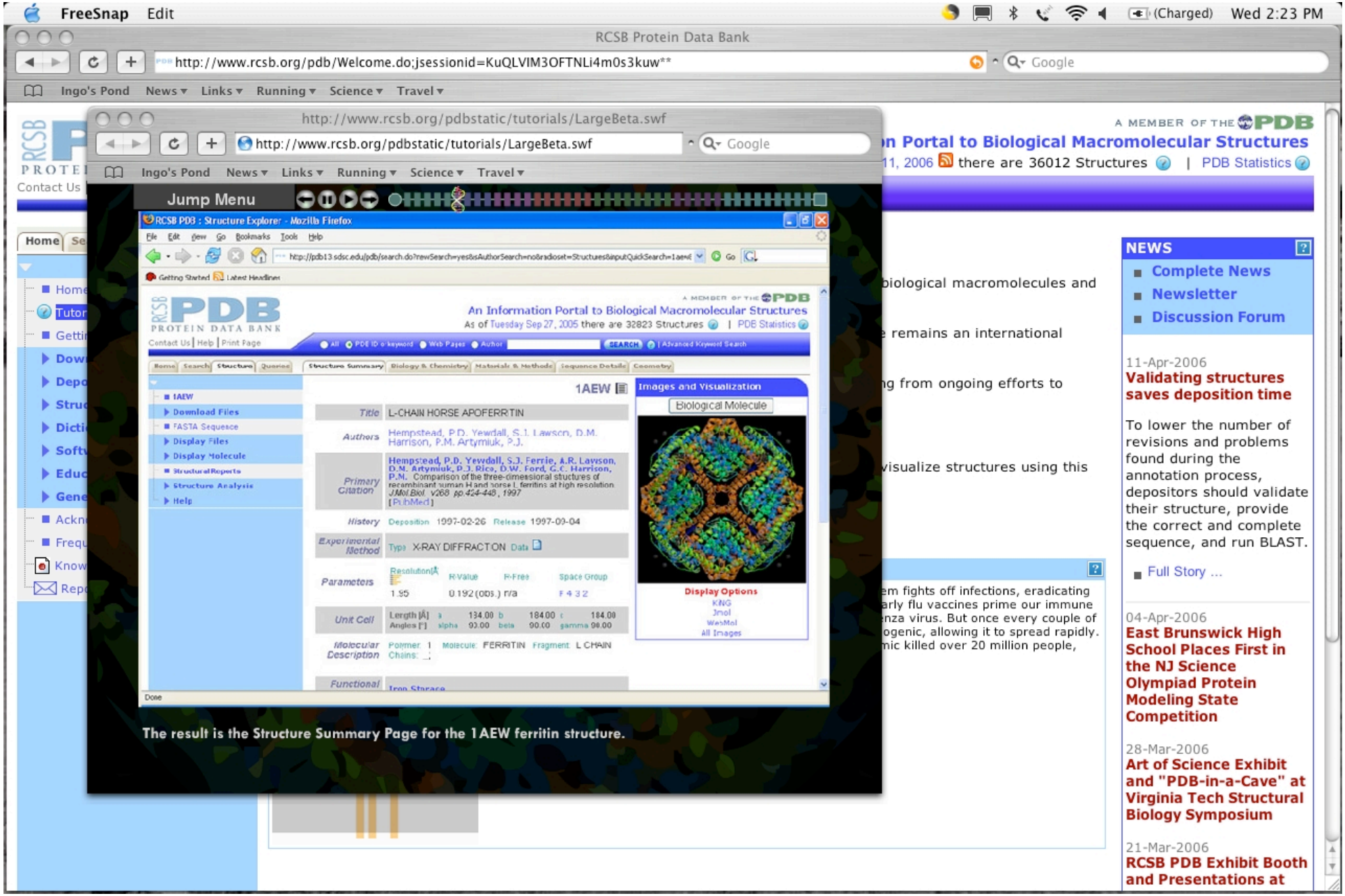
04-Apr-2006
East Brunswick High School Places First in the NJ Science Olympiad Protein Modeling State Competition

28-Mar-2006
Art of Science Exhibit and "PDB-in-a-Cave" at Virginia Tech Structural Biology Symposium

21-Mar-2006
RCSB PDB Exhibit Booth and Presentations at Experimental Biology

The RCSB PDB is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Institute of Neurological Disorders and Stroke (NINDS).

In citing the PDB please refer to: H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242 (2000).


FreeSnap Edit
(Charged) Wed 2:23 PM
RCSB Protein Data Bank
http://www.rcsb.org/pdb/Welcome.do;jsessionid=KuQLVIM3OFTNLi4m0s3kuw**
Ingo's Pond News Links Running Science Travel
http://www.rcsb.org/pdbstatic/tutorials/LargeBeta.swf
Jump Menu
RCSB PDB : Structure Explorer - Mozilla Firefox
An Information Portal to Biological Macromolecular Structures
1AEW
Images and Visualization
Biological Molecule
Display Options
The result is the Structure Summary Page for the 1AEW ferritin structure.
A MEMBER OF THE PDB
Portal to Biological Macromolecular Structures
there are 36012 Structures
PDB Statistics
NEWS
Complete News
Newsletter
Discussion Forum
11-Apr-2006
Validating structures saves deposition time
To lower the number of revisions and problems found during the annotation process, depositors should validate their structure, provide the correct and complete sequence, and run BLAST.
Full Story ...
04-Apr-2006
East Brunswick High School Places First in the NJ Science Olympiad Protein Modeling State Competition
28-Mar-2006
Art of Science Exhibit and "PDB-in-a-Cave" at Virginia Tech Structural Biology Symposium
21-Mar-2006
RCSB PDB Exhibit Booth and Presentations at

RCSB PDB : Structure Explorer

http://www.rcsb.org/pdb/navbarsearch.do?newSearch=yes&isAuthorSearch=no&radioset=All&inputQuickSearch=2mm1&image.x

Ingo's Pond | News | Links | Running | Science | Travel

RCSB PDB PROTEIN DATA BANK

Contact Us | Help | Print Page

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Apr 11, 2006 there are 36012 Structures | PDB Statistics

PDB ID or keyword | Author | SEARCH | Advanced Search

Home | Search | Structure | Queries

Structure Summary | Biology & Chemistry | Materials & Methods | Sequence Details | Geometry

- 2MM1
- Download Files
- FASTA Sequence
- Display Files
- Display Molecule
- Structural Reports
- Structure Analysis
- Help

2MM1

Title	X-RAY CRYSTAL STRUCTURE OF A RECOMBINANT HUMAN MYOGLOBIN MUTANT AT 2.8 ANGSTROMS RESOLUTION
Authors	Hubbard, S.R., Hendrickson, W.A., Lambright, D.G., Boxer, S.G.
Primary Citation	Hubbard, S.R., Hendrickson, W.A., Lambright, D.G., Boxer, S.G. X-ray crystal structure of a recombinant human myoglobin mutant at 2.8 A resolution. *J.Mol.Biol. v213 pp.215-218 , 1990* [Abstract]
History	Deposition 1991-02-19 Release 1993-01-15
Experimental Method	Type X-RAY DIFFRACTION Data
Parameters	Resolution[A] 2.80; R-Value 0.158 (obs.); R-Free n/a; Space Group $P\,3_2\,2\,1$
Unit Cell	Length [A] a 86.20 b 86.20 c 35.60; Angles [°] alpha 90.00 beta 90.00 gamma 120.00
Molecular Description Asymmetric Unit	Polymer: 1 Molecule: MYOGLOBIN Chains: _
Functional Class	Oxygen Transport

Images and Visualization

Biological Molecule / Asymmetric Unit

Display Options

KiNG
Jmol
WebMol
Protein Workshop
QuickPDB
All Images

Source: Polymer: 1 Scientific Name: Homo sapiens

Chemical Component

Identifier	Name	Formula	Drug Similarity	Ligand Structure	Ligand Interaction
HEM	PROTOPORPHYRIN IX CONTAINING FE	$C_{34}H_{32}N_4O_4Fe$	[View]	[View]	[View]

SCOP Classification (version 1.69)

Domain Info	Class	Fold	Superfamily	Family	Domain	Species
d2mm1__	All alpha proteins	Globin-like	Globin-like	Globins	Myoglobin	Human (Homo sapiens)

CATH Classification (version v2.6.0)

Domain	Class	Architecture	Topology	Homology
2mm100	Mainly Alpha	Orthogonal Bundle	Globin-like	Globins

GO Terms

Polymer	Molecular Function	Biological Process	Cellular Component
MYOGLOBIN (2MM1:_)	• binding • oxygen binding • heme binding	• transport • oxygen transport	• none

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

PDB Molecule of the Month: The Calcium Pump - Mozilla

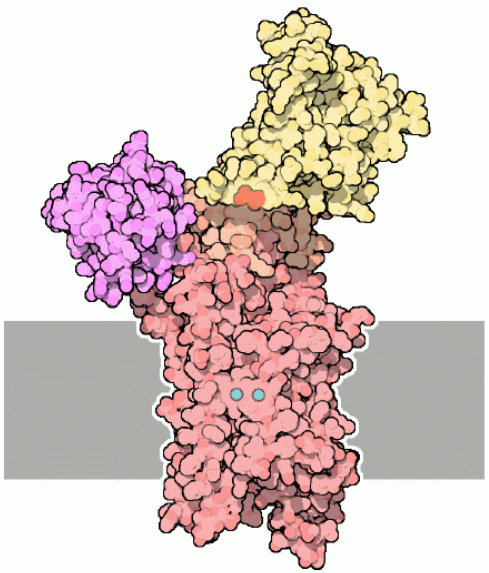
File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.rcsb.org/pdb/molecules/pdb51_1.html Search Print

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

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 tropomyosins 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 climbing, 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

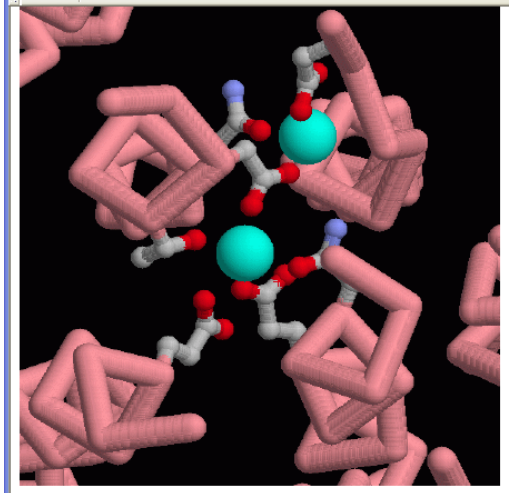
start Microsoft PowerPoint ... PDB Molecule of the ... 11:53 PM

PDB Molecule of the Month: The Calcium Pump - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.rcsb.org/pdb/molecules/pdb51_3.html

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

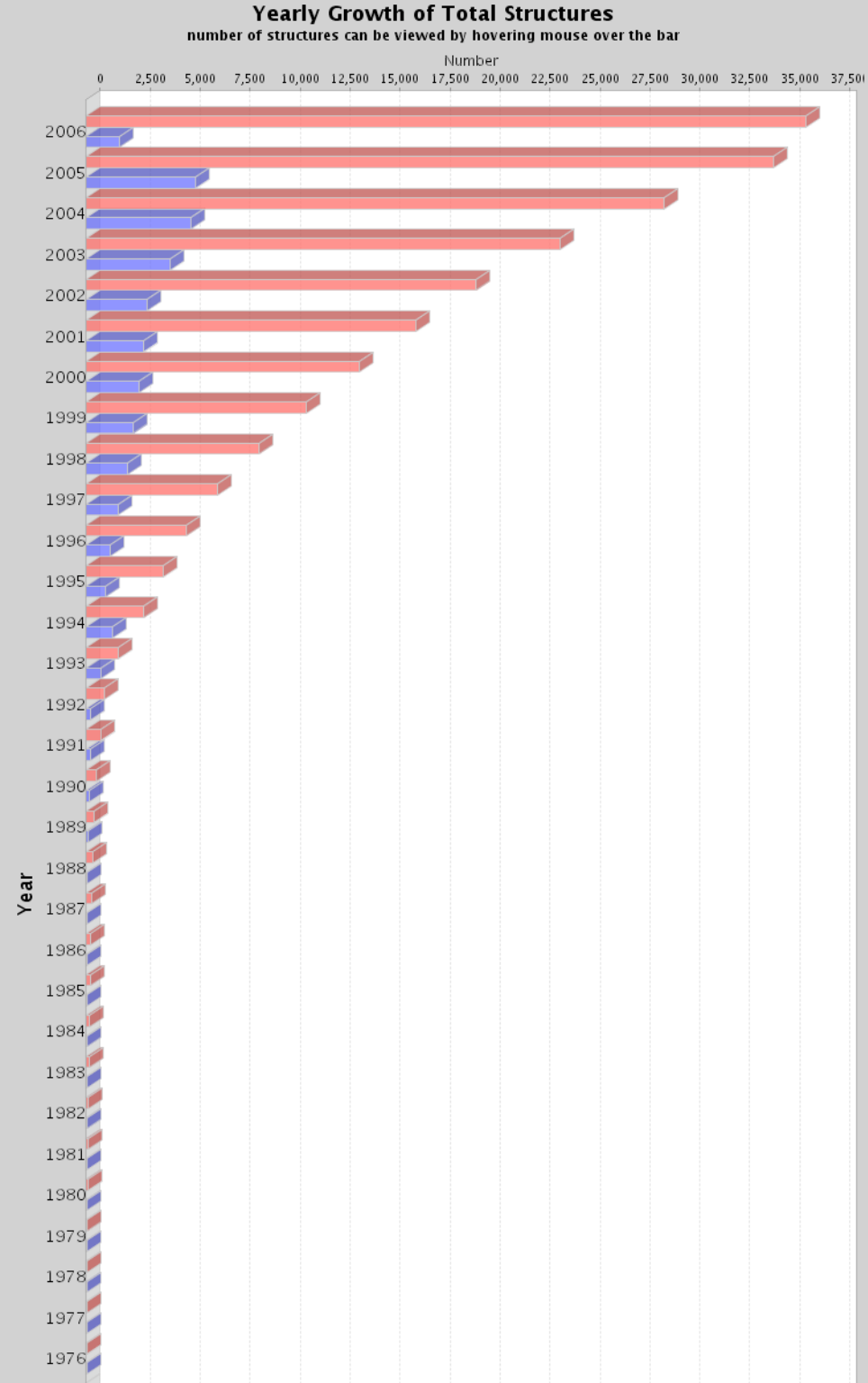


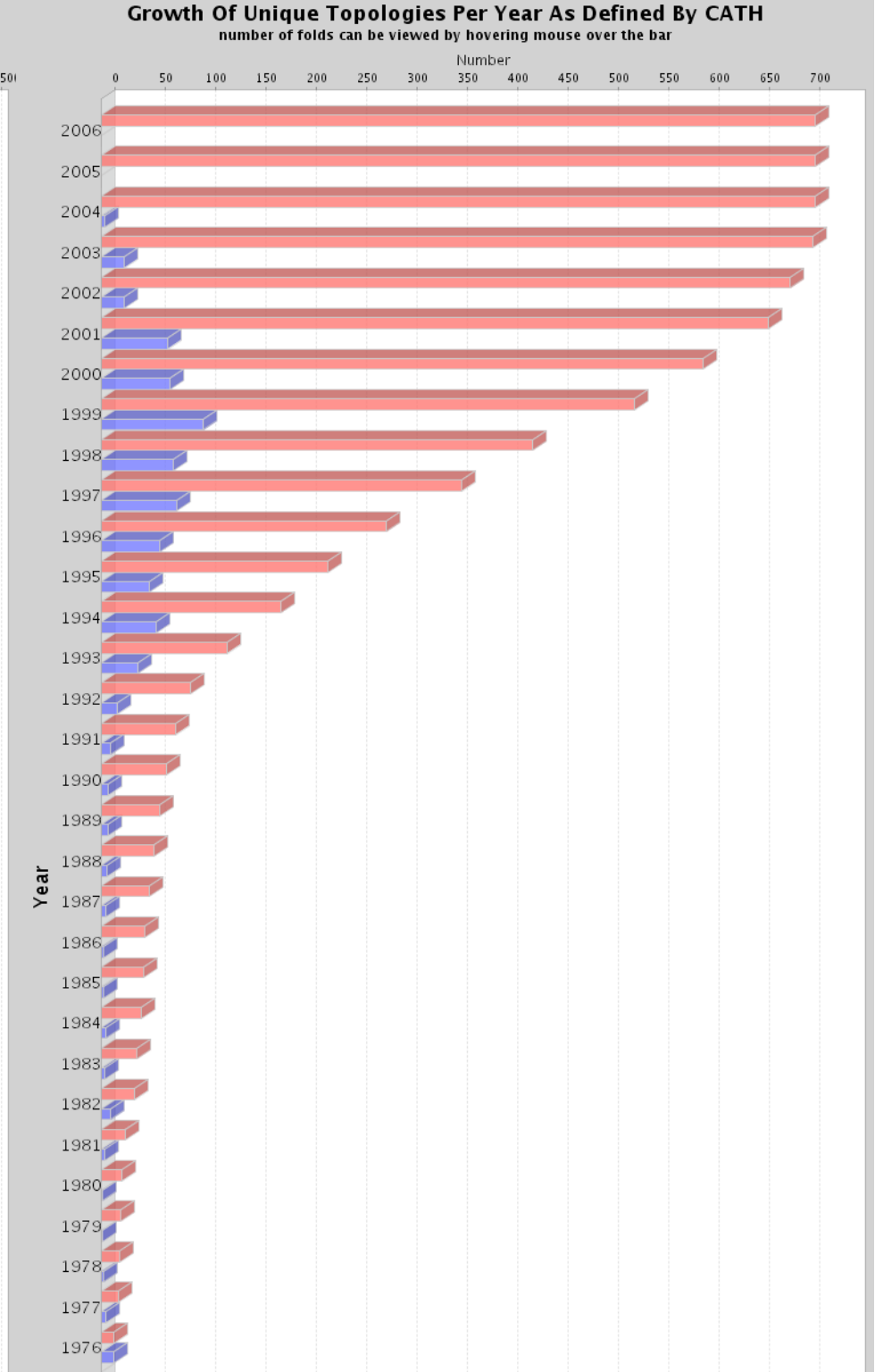
Exploring the Structure

The calcium binding site is in a tunnel formed by four alpha helices, which cross straight through the membrane. This illustration, from PDB entry 1eul, 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 1iwo. 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.

Done

start Microsoft PowerPoint ... PDB Molecule of the ... 11:55 PM


Yearly Growth of Total Structures
number of structures can be viewed by hovering mouse over the bar
Number
0 2,500 5,000 7,500 10,000 12,500 15,000 17,500 20,000 22,500 25,000 27,500 30,000 32,500 35,000 37,500
Year
2006 2005 2004 2003 2002 2001 2000 1999 1998 1997 1996 1995 1994 1993 1992 1991 1990 1989 1988 1987 1986 1985 1984 1983 1982 1981 1980 1979 1978 1977 1976


Growth Of Unique Topologies Per Year As Defined By CATH
number of folds can be viewed by hovering mouse over the bar
Number
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700
Year
2006 2005 2004 2003 2002 2001 2000 1999 1998 1997 1996 1995 1994 1993 1992 1991 1990 1989 1988 1987 1986 1985 1984 1983 1982 1981 1980 1979 1978 1977 1976

Dunbrack Lab Website - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://dunbrack.fccc.edu/Home.php Search Print

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

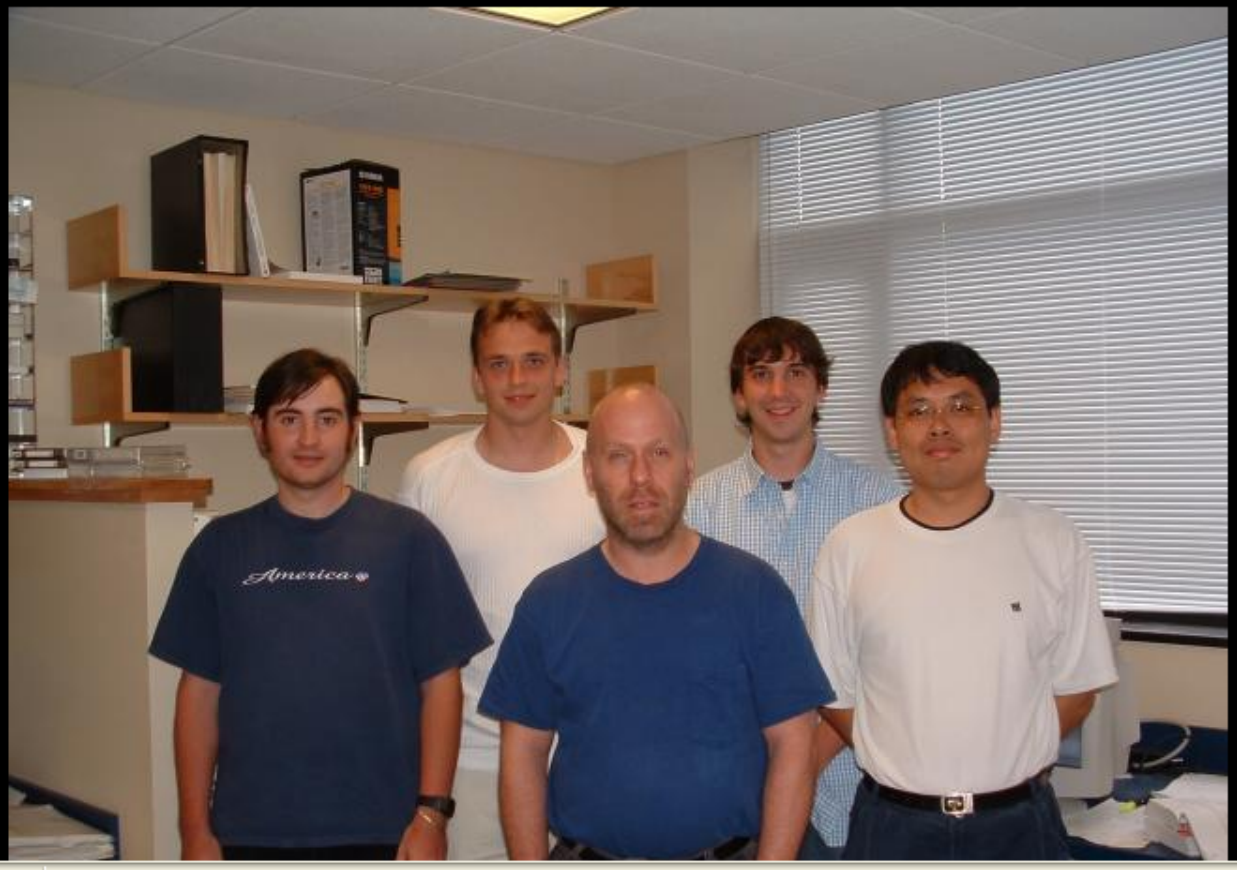
FOX CHASE CANCER CENTER

Dunbrack Lab

- Home
- People
- Projects
- Publications
- Software
- Links

- SCWRL3
- PISCES
- BBDep

>>Home



Done

start Dunbrack Lab Websit... Ingo's Pond - Mozilla 8:07 PM

Cull Protein Sequence List Page - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.fccc.edu/research/labs/dunbrack/cgi-bin/cull_pisces.cgi Search Print

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

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

http://www.fccc.edu/research/labs/dunbrack/pisces

start Dunbrack Lab Websit... Ingo's Pond - Mozilla Cull Protein Sequenc... dunbrack - Paint Microsoft PowerPoint... 8:11 PM

Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.fccc.edu/research/labs/dunbrack/cgi-bin/cull_pisces.cgi Search Print

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

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

http://www.fccc.edu/research/labs/dunbrack/pisces

start Messenger Express - ... Mozilla dunbrack3 - Paint Microsoft PowerPoint ... 8:23 PM

Messenger Express - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop https://webmail.jhsph.edu/en/mail.html?sid=br6p3bo02rt9r6v9e&lang=en&cert=false Search Print

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

Webmail for *Ingo Ruczinski*

Folders Inbox Message Addresses Options Help Log Out

Compose Reply Reply All Forward Delete Prev Next Move message to folder:

From	GL_Wang@fccc.edu Add Sender
Sent	Sunday, April 4, 2004 8:18 pm
To	ingo@jhu.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

Download sequence ID list file http://www.fccc.edu/research/labs/dunbrack/html/cullpdb_pc25_res2.0_R0.20_d040404_chains240.1501
Download fasta format sequence file http://www.fccc.edu/research/labs/dunbrack/html/cullpdb_pc25_res2.0_R0.20_d040404_chains240.fasta.1501
Download similarity log file http://www.fccc.edu/research/labs/dunbrack/html/cullpdb_pc25_res2.0_R0.20_d040404_chains240.log.1501

start Messenger Express -... Cull Protein Sequenc... pisces.html - Mozilla dunbrack2 - Paint Microsoft PowerPoint... 8:20 PM

Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.fccc.edu/research/labs/dunbrack/html/cullpdb_pc25_res2.0_R0.20_d040404_chains240.1501 Search Print

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

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
1O06A	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
1Q08A	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
1HYPO	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
1OKOA	74	XRAY	0.930	0.10	0.13
1G2RA	100	XRAY	1.350	0.16	0.18
1L6KA	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
1NOQA	93	XRAY	1.260	0.17	0.19
1C75A	71	XRAY	0.970	0.12	1.00
1I2TA	61	XRAY	1.040	0.15	0.17
3EBXO	62	XRAY	1.400	0.18	1.00
1MOFO	55	XRAY	1.700	0.17	0.23

Done

start Messenger Express -... Mozilla Mozilla dunbrack3 - Paint Microsoft PowerPoint... 8:25 PM

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.

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

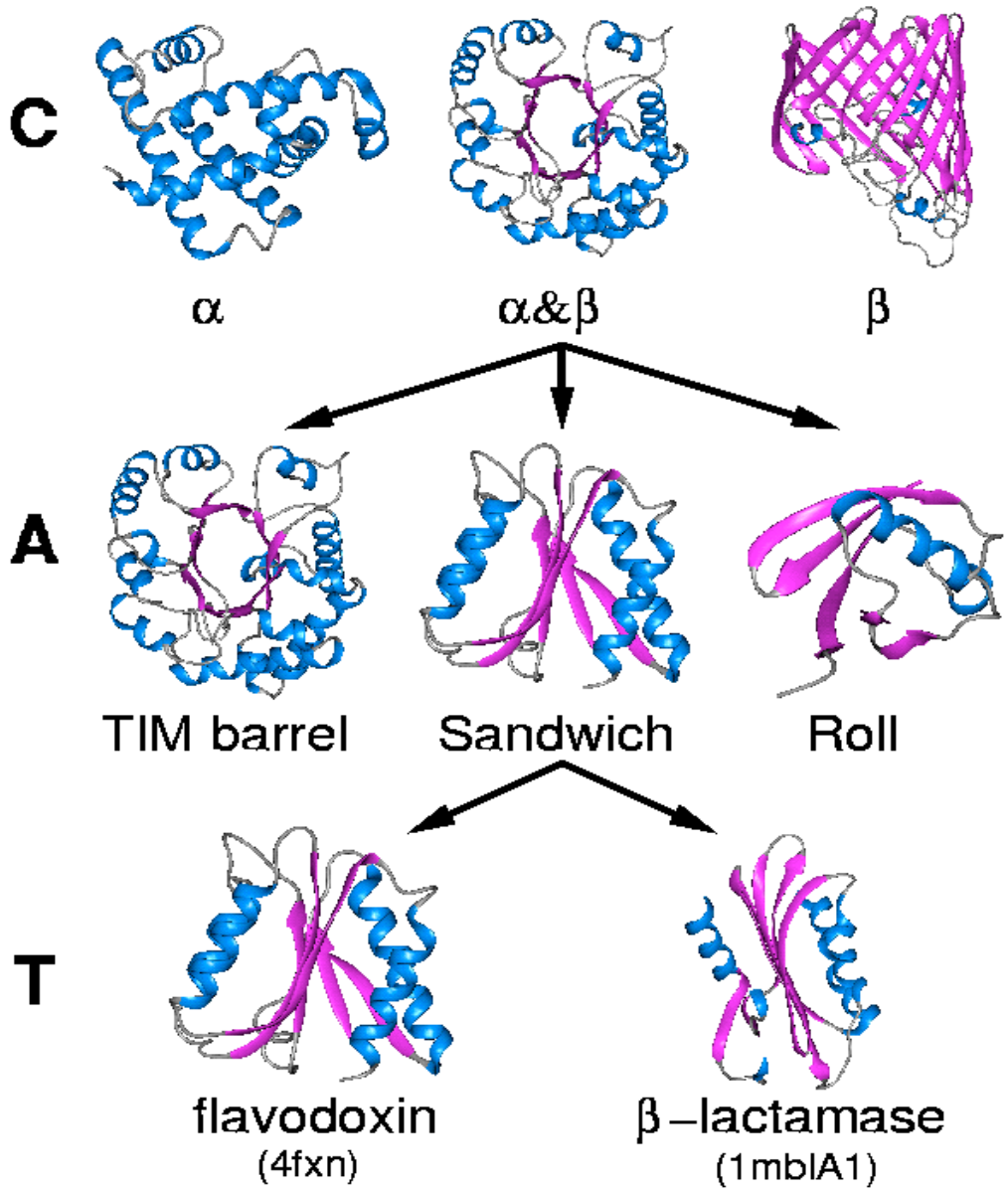
CATH

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

CATH Releases

http://www.cathdb.info/latest/releases.html

Ingo's Pond News Links Running Science Travel

CATH Protein Structure Classification

CATH DHS Gene3D Impala FTP Internal

Search

Go!

- PDB Code
- CATH Code
- General Text

Goto

SSAP Server
GRATH Server
DHS
Gene3D

Navigation

Home
Top of hierarchy

Home > Top Releases

CATH Releases

This page provides information on the official CATH releases.

CATH v2.6.0

Version	2.6.0
Date	11-04-2005

C	A	T	H	S	N	I	D
Mainly Alpha	5	251	465	1402	2189	3705	14105
Mainly Beta	19	160	311	1443	2961	4329	18771
Alpha Beta	14	414	706	3014	4781	7660	33080
Few Secondary Structures	1	82	90	144	232	285	1098
Preliminary single domain assigments	10	808	809	906	967	1090	3012
Multi-domain domains	1	12	12	16	25	36	109
CATH-35 Sequence families	1	4707	4707	4719	4768	4862	6168
	1	22	22	27	33	38	198

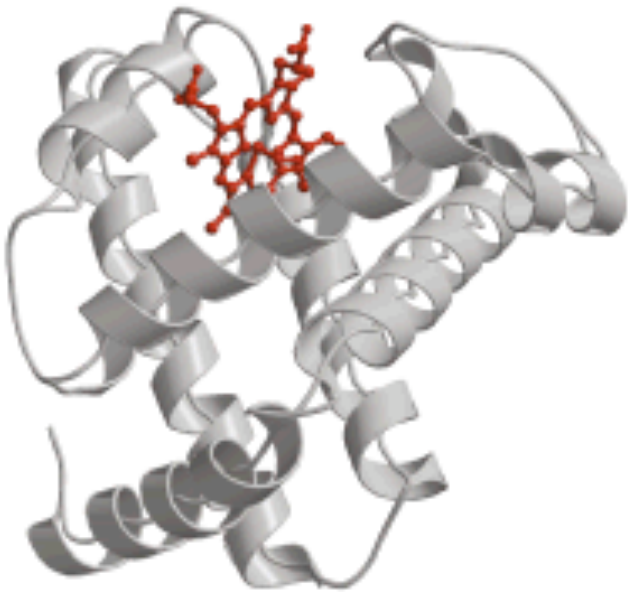
DALI

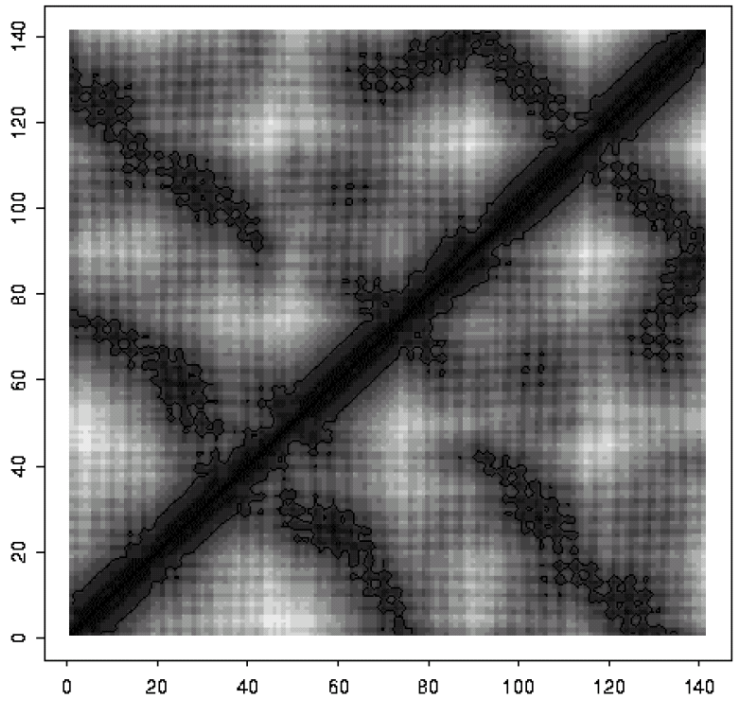
Distance Matrix Alignment

- 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)

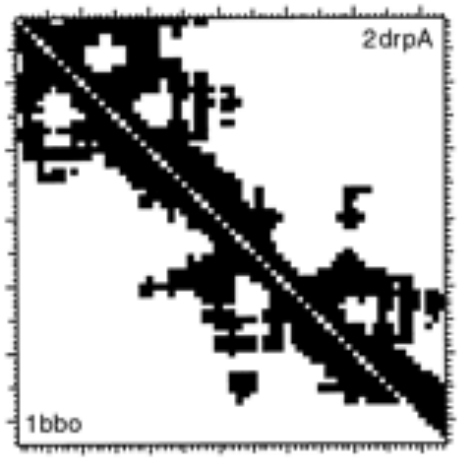
DALI



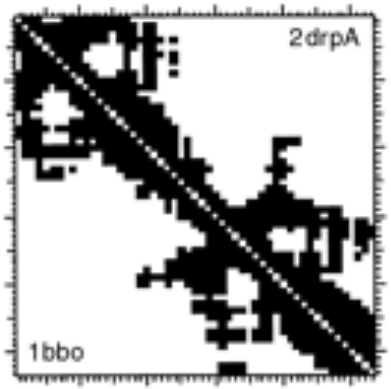


DALI

Unaligned:



Aligned:



Unaligned:

```
1bbo     1 KYICEECGIRXKKPSMLKKHIRTHTDVRPYHCTYCNFSFKTKGNLTKHMKSKAHSKK   57

2drpA  103 FTKEGEHTYRCKVCSRVYTHISNFCRHYVTSHKRNVKVYPCPFCFKEFTRKDNMTAHVKIIHK 165
```

Aligned:

```
1bbo     1 .......KYICEECGIRXKKPSMLKKHIRTHt..DVRPYHCTYCNFSFKTKGNLTKHMKSKAHskk   57
                   | |  |        |    |  |     |  | |  |   |  | | | | |
2drpA  103 ftkegehTYRCKVCSRVYTHISNFCRHYVTShkrNVKVYPCPFCFKEFTRKDNMTAHVKIIHK... 165
```

FSSP and DDD

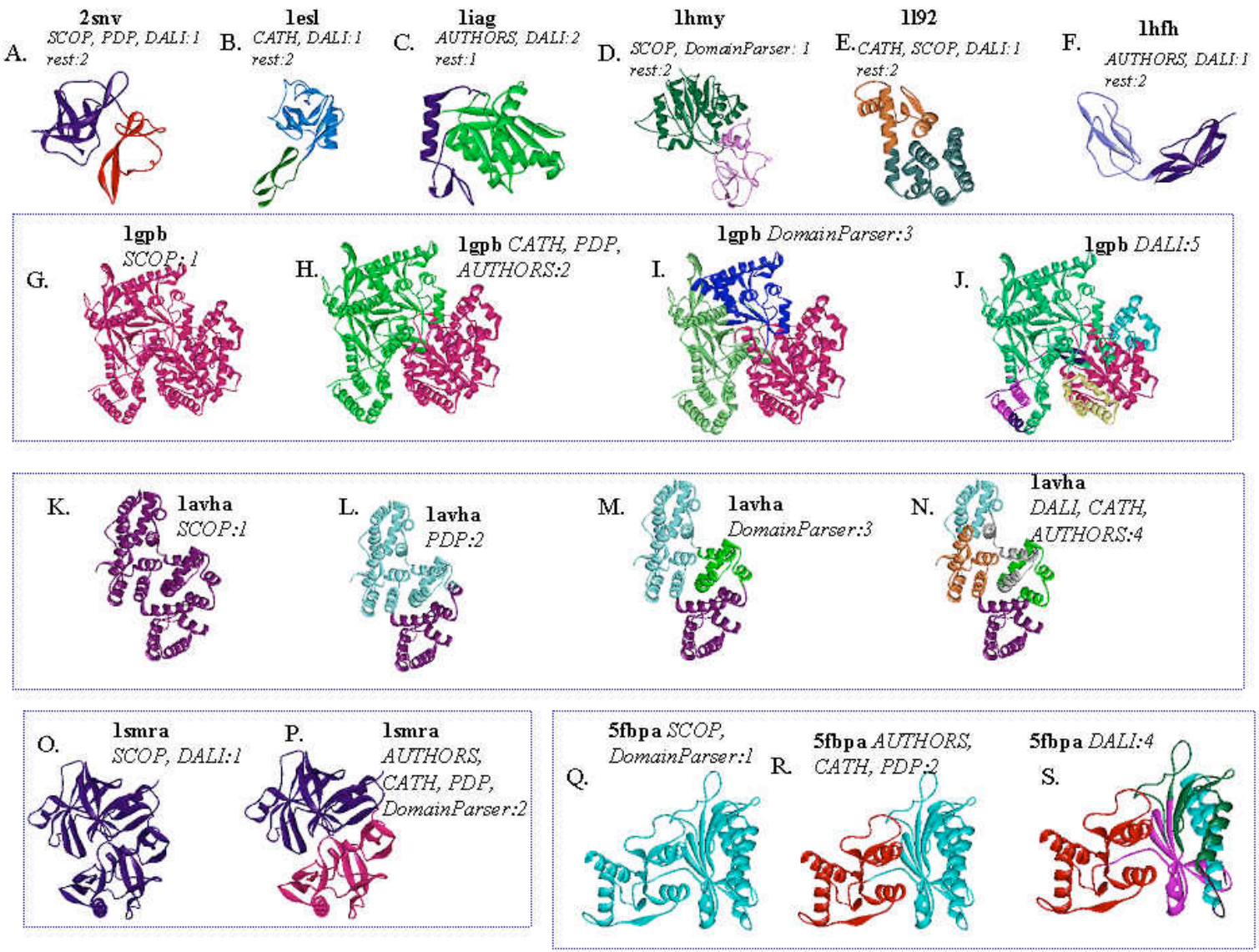
- The families of structurally similar proteins (FSSP) is a database of structural alignments of proteins in the protein data bank (PDB). It presents the results of applying DALI to (almost) all chains of proteins in the PDB.
- The DALI domain dictionary (DDD) is a corresponding classification of recurrent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

Other Algorithms for Domain Decomposition

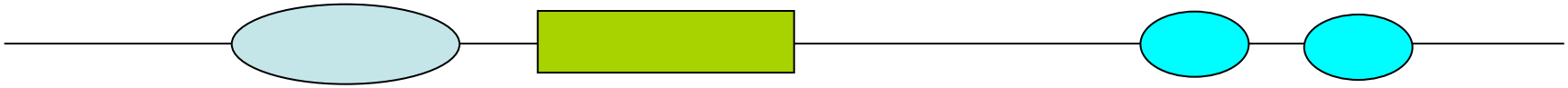
- The Protein Domain Parser (PDP) uses compactness as a chief principle.
  `http://123d.ncifcrf.gov/pdp.html`

- 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…


A. 2snv SCOP, PDP, DALI:1 rest:2
B. 1esl CATH, DALI:1 rest:2
C. 1iag AUTHORS, DALI:2 rest:1
D. 1hmy SCOP, DomainParser: 1 rest:2
E. 1l92 CATH, SCOP, DALI:1 rest:2
F. 1hfh AUTHORS, DALI:1 rest:2
G. 1gpb SCOP: 1
H. 1gpb CATH, PDP, AUTHORS:2
I. 1gpb DomainParser:3
J. 1gpb DALI:5
K. 1avha SCOP:1
L. 1avha PDP:2
M. 1avha DomainParser:3
N. 1avha DALI, CATH, AUTHORS:4
O. 1smra SCOP, DALI:1
P. 1smra AUTHORS, CATH, PDP, DomainParser:2
Q. 5fbpa SCOP, DomainParser:1
R. 5fbpa AUTHORS, CATH, PDP:2
S. 5fbpa DALI:4

Parsing Sequence into Domains



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

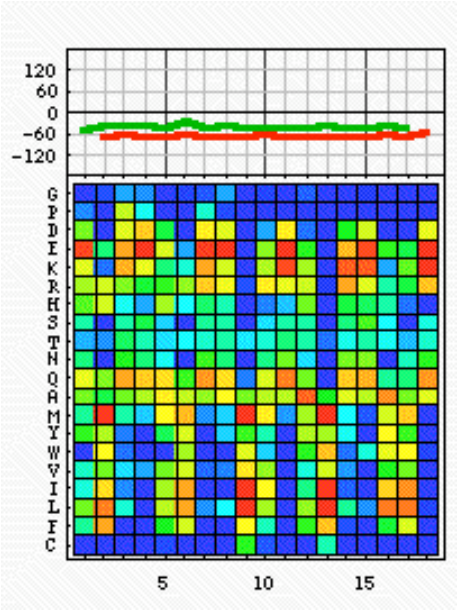
Why is That Important?

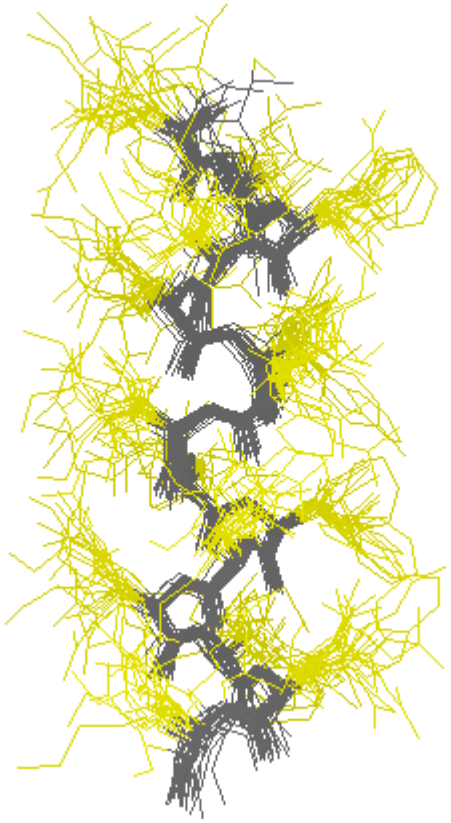
- Functional insights.
- Improved database searching.
- Fold recognition.
- Structure determination.

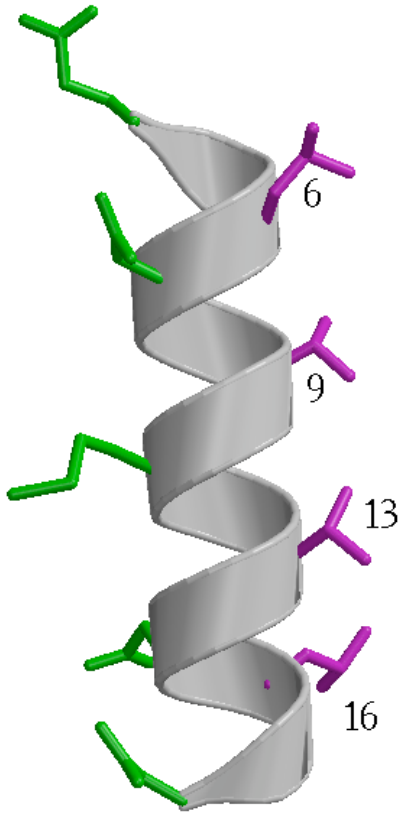
PRODOM: http://protein.toulouse.inra.fr/prodom/current/html/home.php

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

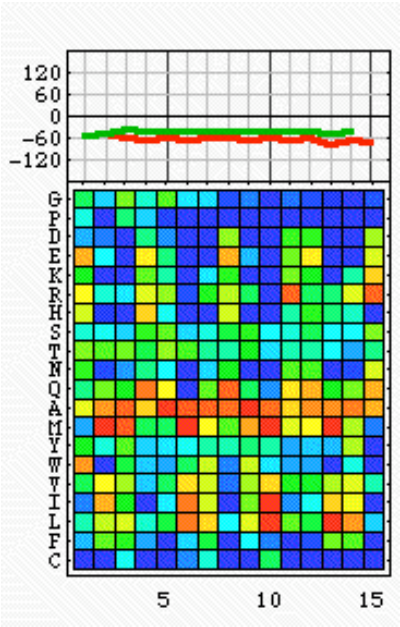
I-Sites

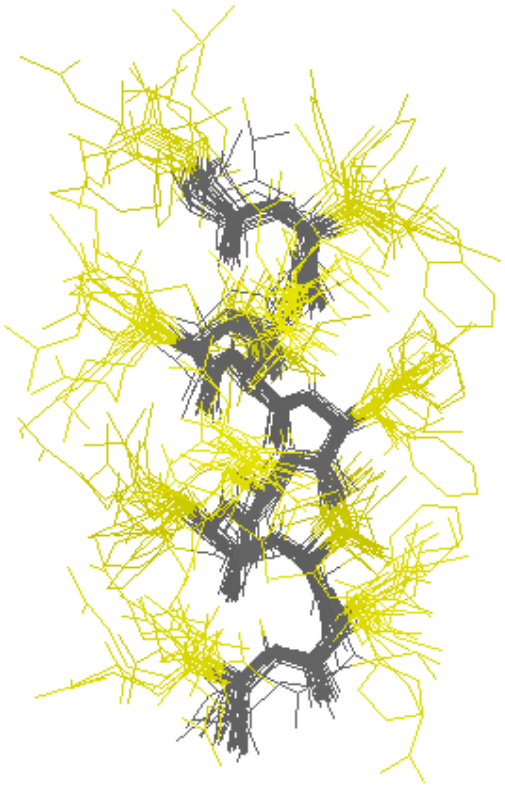


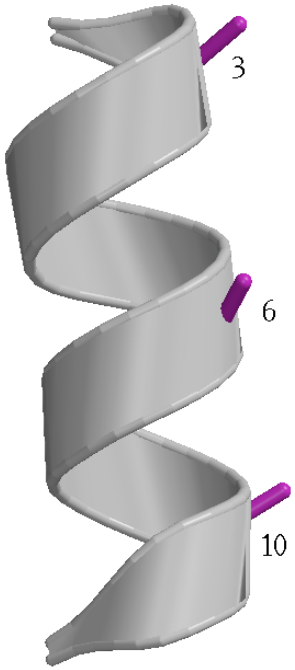




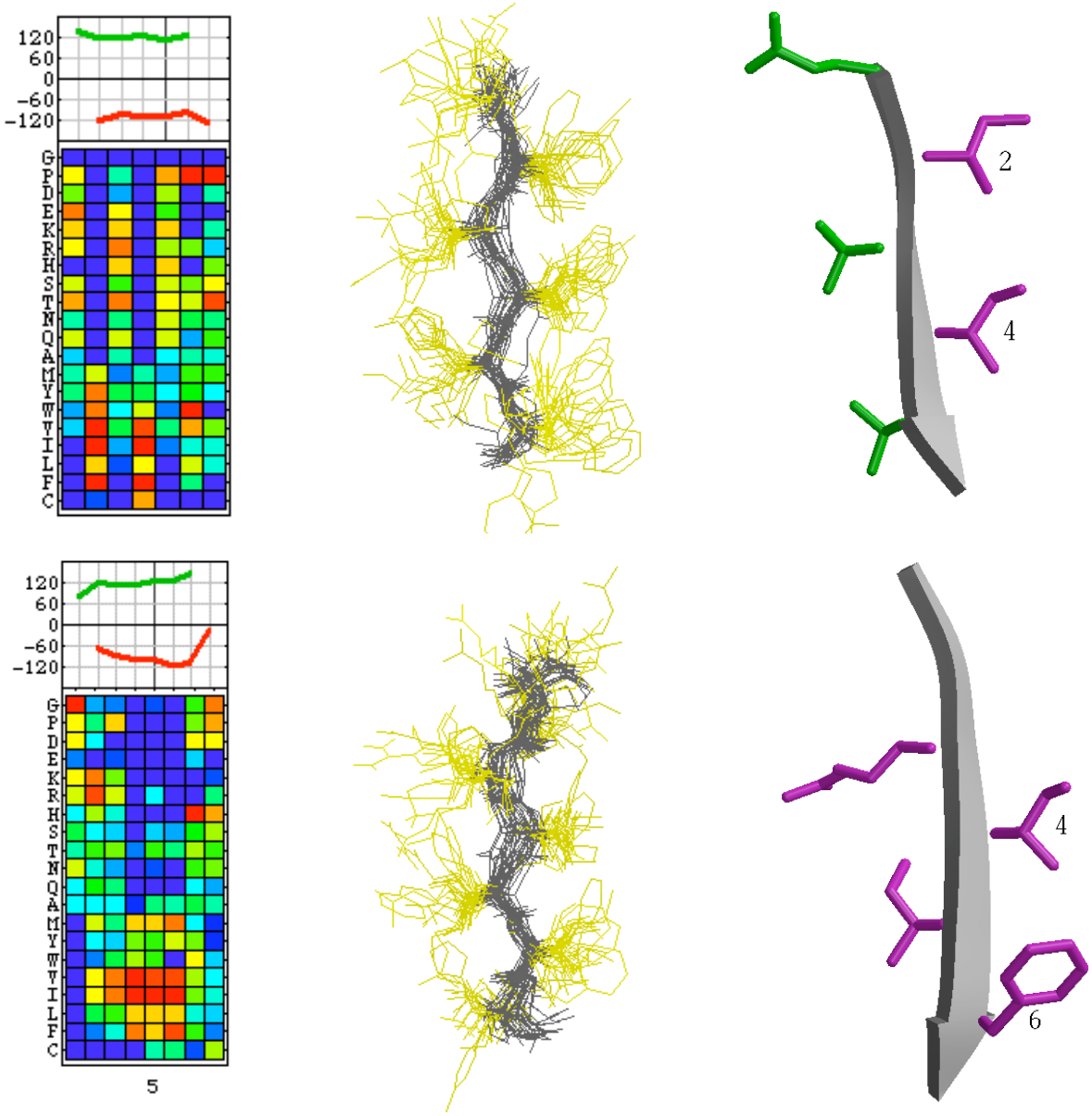
I-Sites



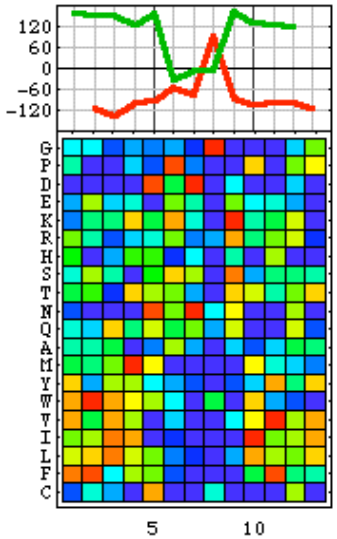


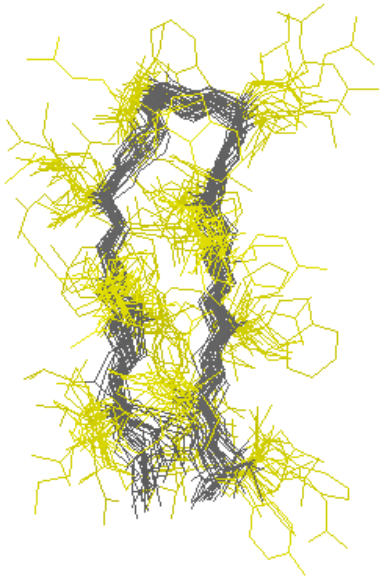


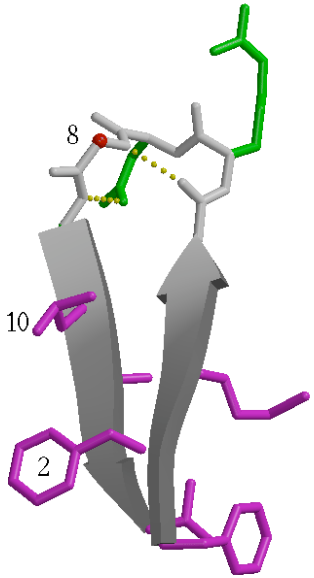
I-Sites

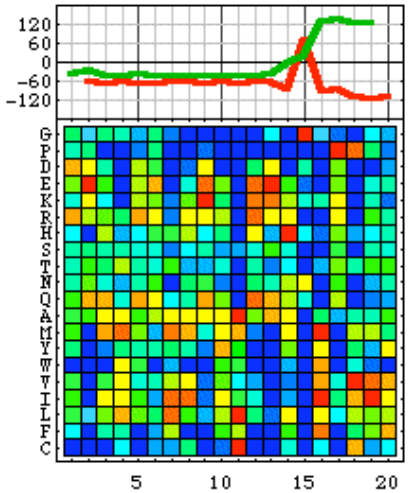


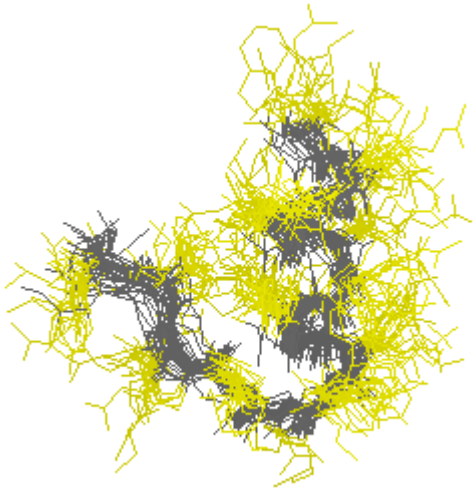
I-Sites

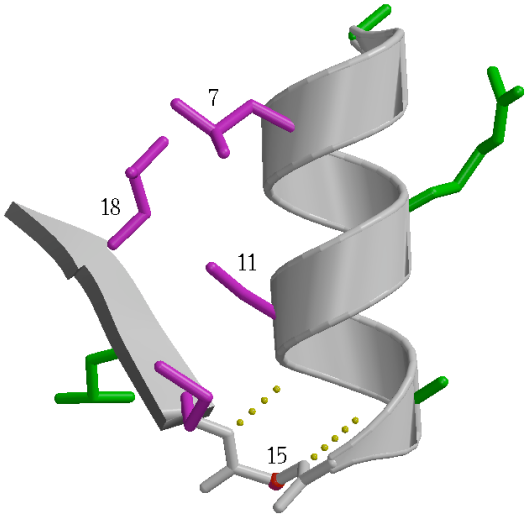

120
60
0
-60
-120
G
P
D
E
K
R
H
S
T
N
Q
A
M
Y
W
V
I
L
F
C
5
10




8
10
2


120
60
0
-60
-120
G
P
D
E
K
R
H
S
T
N
Q
A
M
Y
W
V
I
L
F
C
5
10
15
20




7
18
11
15