Analysis and Prediction of Protein Structure (I)

Jianlin Cheng, PhD

School of Electrical Engineering and Computer Science
University of Central Florida

2006

Outline

I. **Sequence, Structure, Function Relation**

II. Determination, Storage, Visualization, Analysis, and Comparison

III. Structure Classification

IV. 1D Prediction

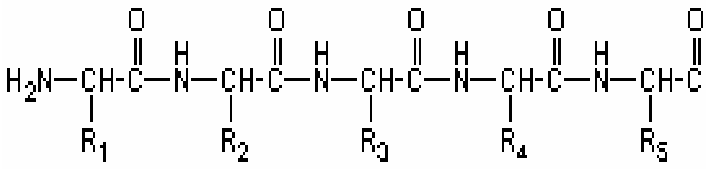
V. 2D Prediction

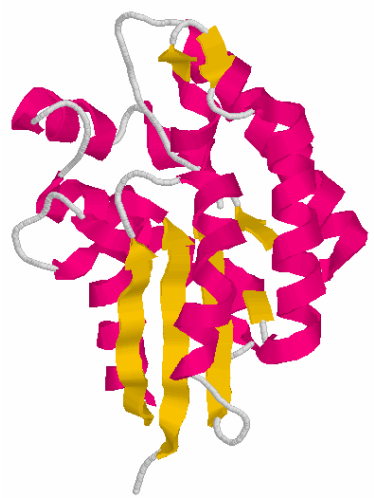
VI. 3D Prediction

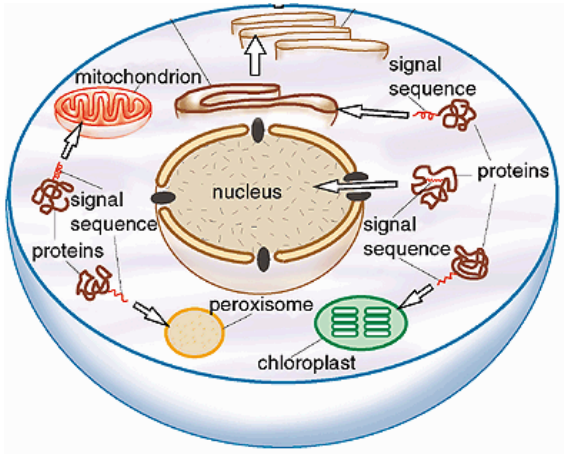
VII. Useful Tools

Sequence, Structure and Function

AGCWY……

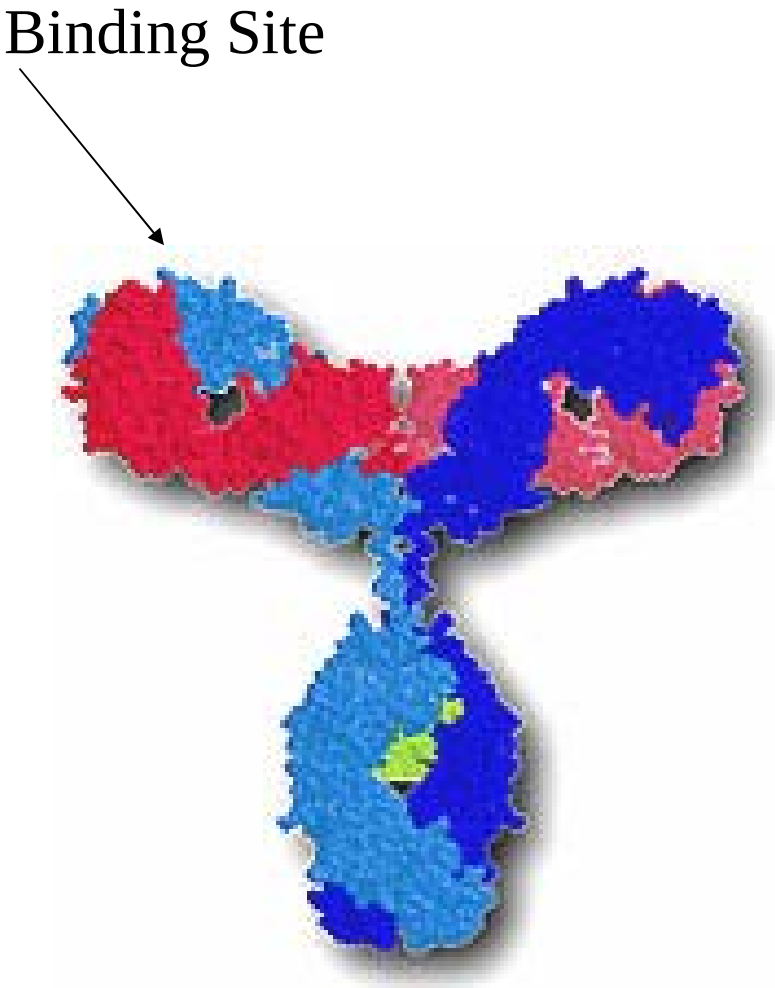






Cell

Protection Function in Immune System



Antibody

High specificity (amino acid sequence segment at binding site)
Strong affinity

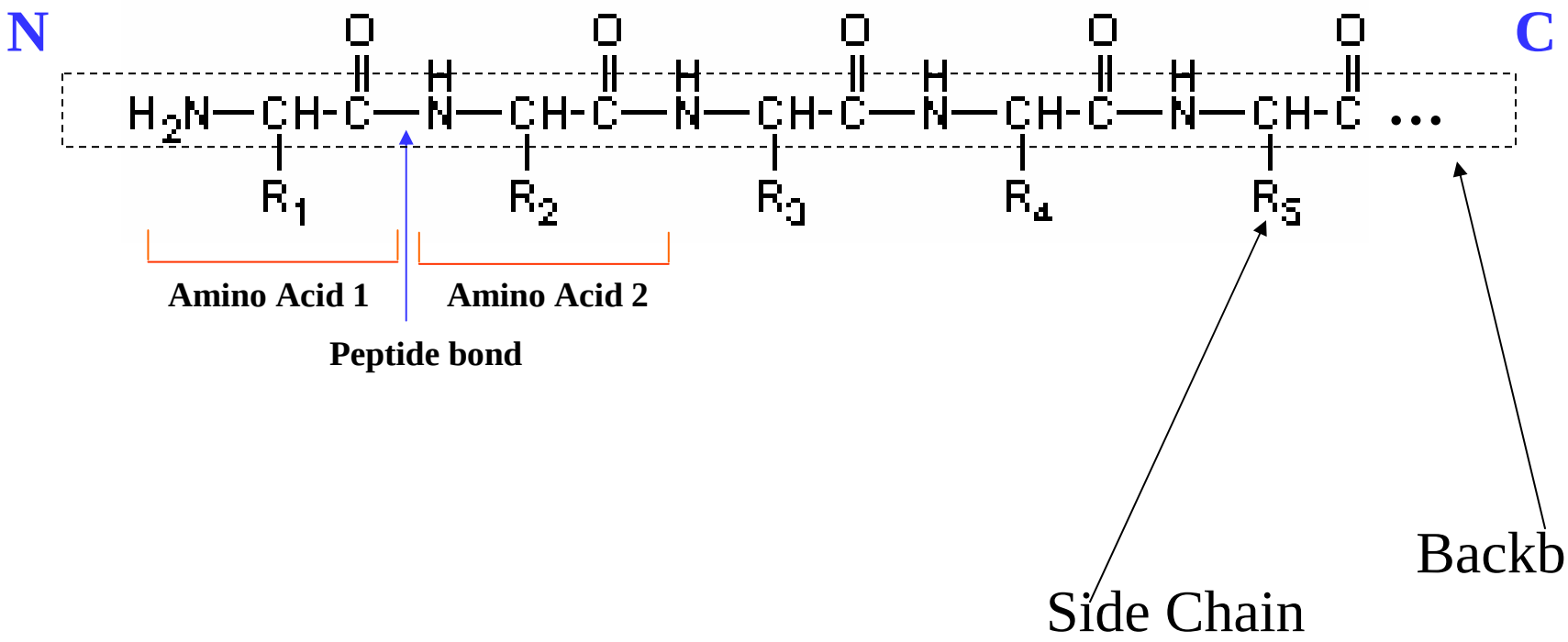
Question: there are so many different viruses not known before hand, how an animal cell figure out an Antibody to bind
to them, but not bind to its own protein? In stock or make on the fly?

Protein Sequence – Primary Structure

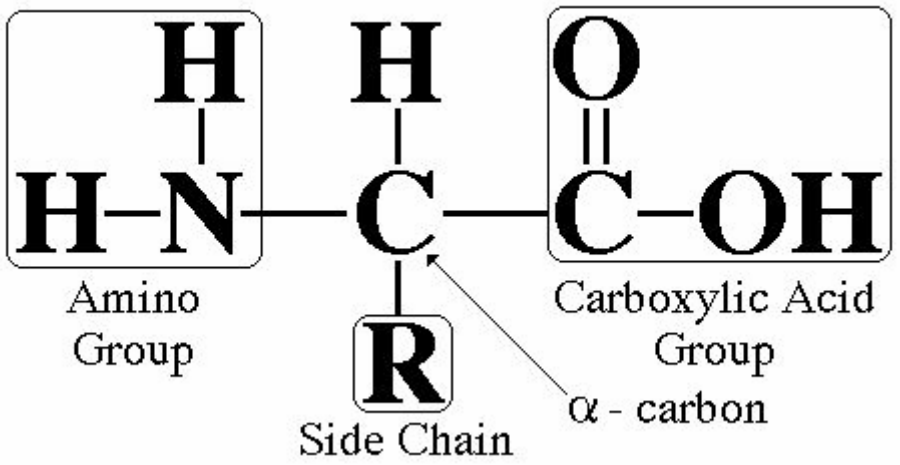
- The first protein was sequenced by Frederick Sanger in 1953.
- Twice Nobel Laureate (1958, 1980) (other: Curie, Pauling, Bardeen).
- Determined the amino acid sequence of insulin and proved proteins have specific primary structure.

Protein Sequence

A directional sequence of amino acids/residues

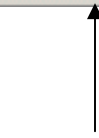


Amino Acid Structure



Amino Acids

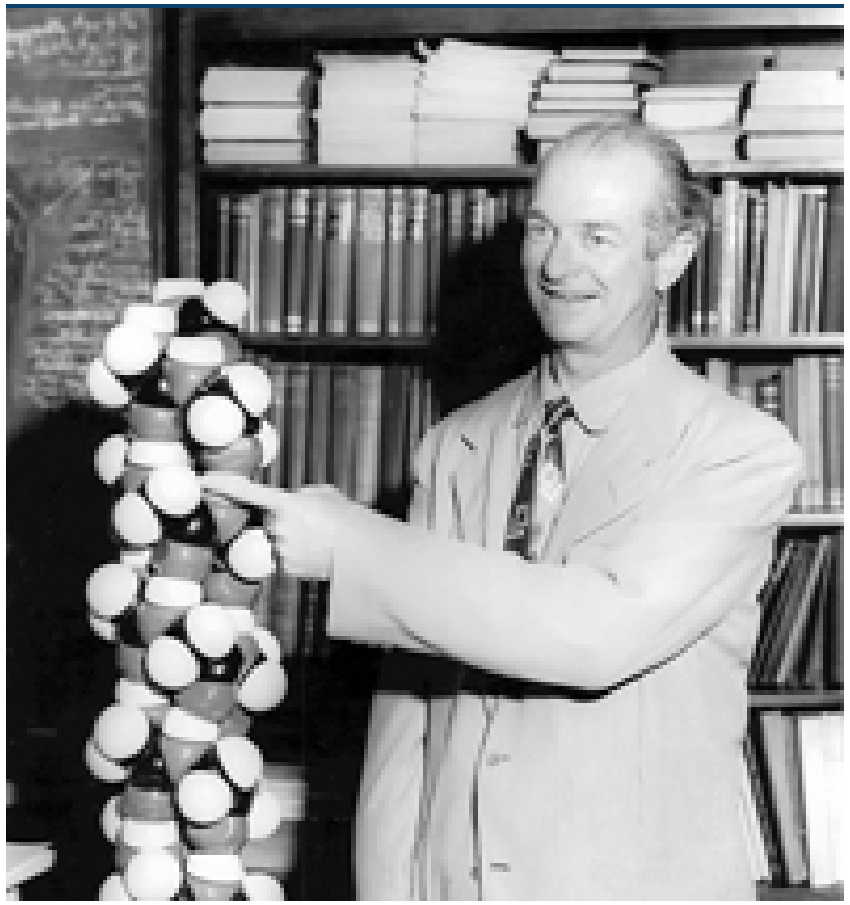
Amino acid	Abbrev.	Side chain	Hydro-phobic	Polar	Charged	Small	Tiny	Aromatic or Aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
Alanine	Ala, A	$-CH_3$	X	-	-	X	X	-	67	GCU, GCC, GCA, GCG	7.8
Cysteine	Cys, C	$-CH_2SH$	X	-	-	X	-	-	86	UGU, UGC	1.9
Aspartate	Asp, D	$-CH_2COOH$	-	X	negative	X	-	-	91	GAU, GAC	5.3
Glutamate	Glu, E	$-CH_2CH_2COOH$	-	X	negative	-	-	-	109	GAA, GAG	6.3
Phenylalanine	Phe, F	$-CH_2C_6H_5$	X	-	-	-	-	Aromatic	135	UUU, UUC	3.9
Glycine	Gly, G	-H	X	-	-	X	X	-	48	GGU, GGC, GGA, GGG	7.2
Histidine	His, H	$-CH_2-C_3H_3N_2$	-	X	positive	-	-	Aromatic	118	CAU, CAC	2.3
Isoleucine	Ile, I	$-CH(CH_3)CH_2CH_3$	X	-	-	-	-	Aliphatic	124	AUU, AUC, AUA	5.3
Lysine	Lys, K	$-(CH_2)_4NH_2$	-	X	positive	-	-	-	135	AAA, AAG	5.9
Leucine	Leu, L	$-CH_2CH(CH_3)_2$	X	-	-	-	-	Aliphatic	124	UUA, UUG, CUU, CUC, CUA, CUG	9.1
Methionine	Met, M	$-CH_2CH_2SCH_3$	X	-	-	-	-	-	124	AUG	2.3
Asparagine	Asn, N	$-CH_2CONH_2$	-	X	-	X	-	-	96	AAU, AAC	4.3
Proline	Pro, P	$-CH_2CH_2CH_2-$	X	-	-	X	-	-	90	CCU, CCC, CCA, CCG	5.2
Glutamine	Gln, Q	$-CH_2CH_2CONH_2$	-	X	-	-	-	-	114	CAA, CAG	4.2
Arginine	Arg, R	$-(CH_2)_3NH-C(NH)NH_2$	-	X	positive	-	-	-	148	CGU, CGC, CGA, CGG, AGA, AGG	5.1
Serine	Ser, S	$-CH_2OH$	-	X	-	X	X	-	73	UCU, UCC, UCA, UCG, AGU,AGC	6.8
Threonine	Thr, T	$-CH(OH)CH_3$	X	X	-	X	-	-	93	ACU, ACC, ACA, ACG	5.9
Valine	Val, V	$-CH(CH_3)_2$	X	-	-	X	-	Aliphatic	105	GUU, GUC, GUA, GUG	6.6
Tryptophan	Trp, W	$-CH_2C_8H_6N$	X	-	-	-	-	Aromatic	163	UGG	1.4
Tyrosine	Tyr, Y	$-CH_2-C_6H_4OH$	X	X	-	-	-	Aromatic	141	UAU, UAC	3.2



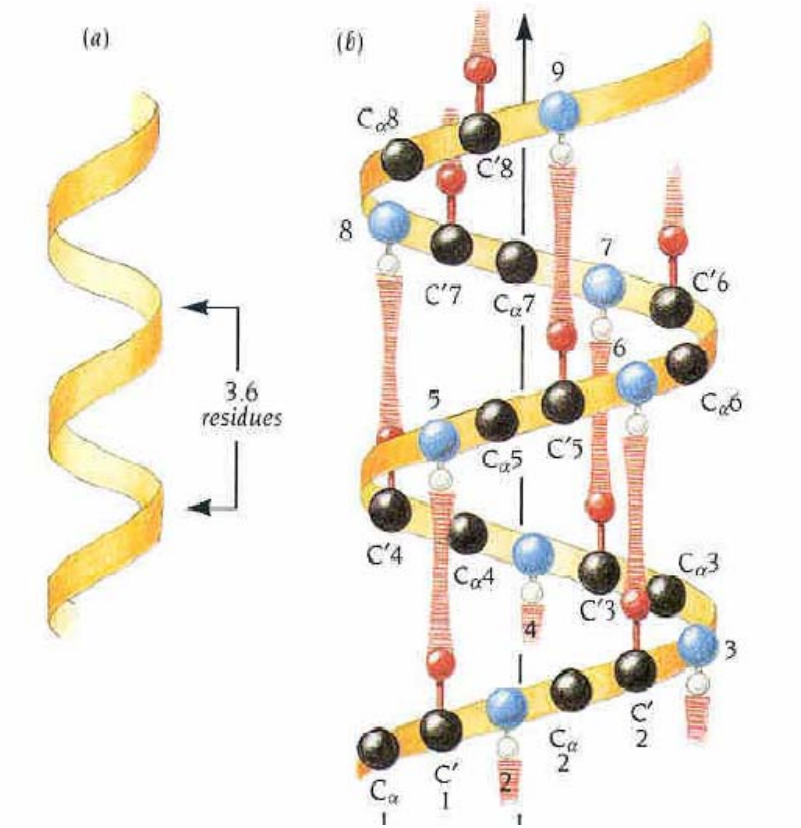
Hydrophilic

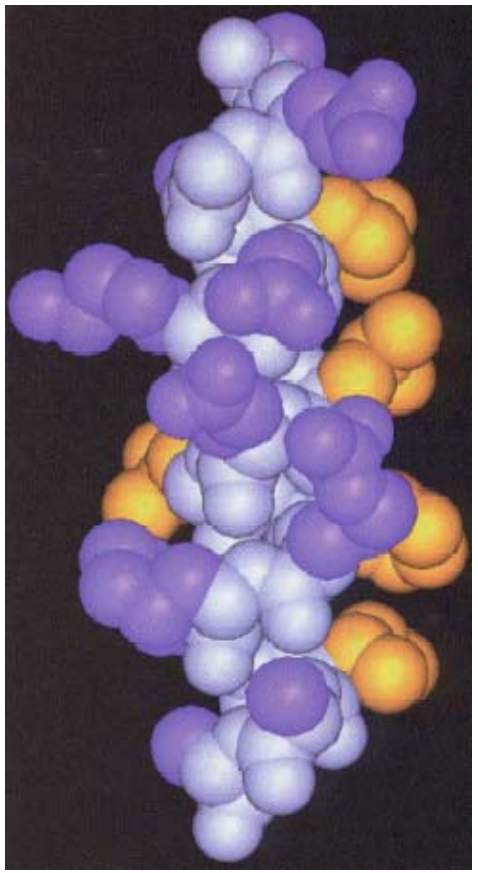
Protein Secondary Structure

- Determined by hydrogen bond patterns
- 3-Class categories: alpha-helix, beta-sheet, loop (or coil)
- First deduced by Linus Pauling et al.



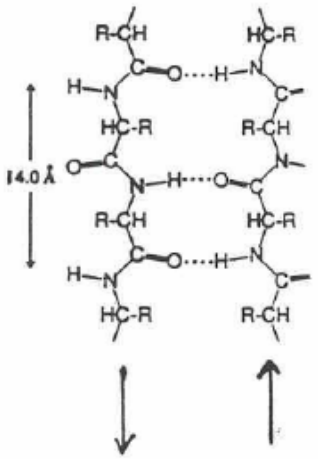
Alpha-Helix

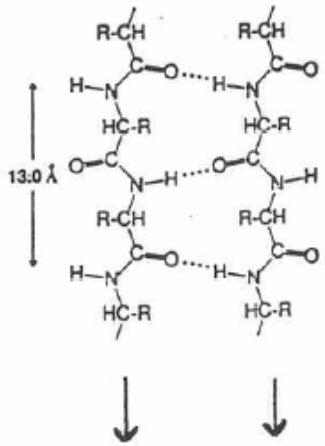


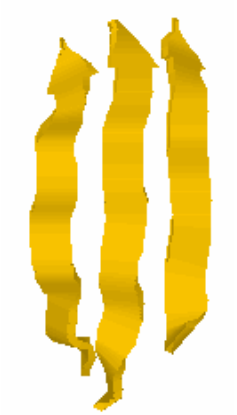


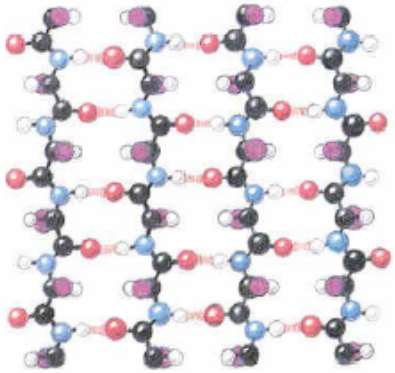
Jurnak, 2003

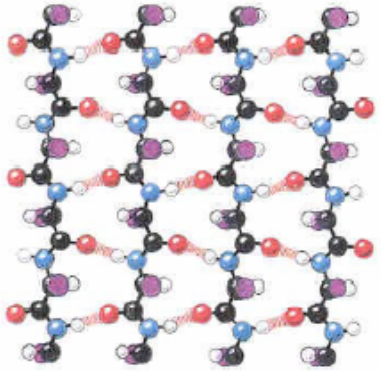
Beta-Sheet

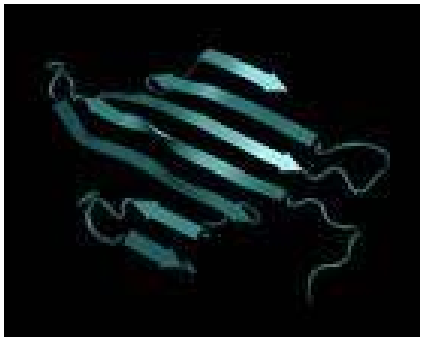








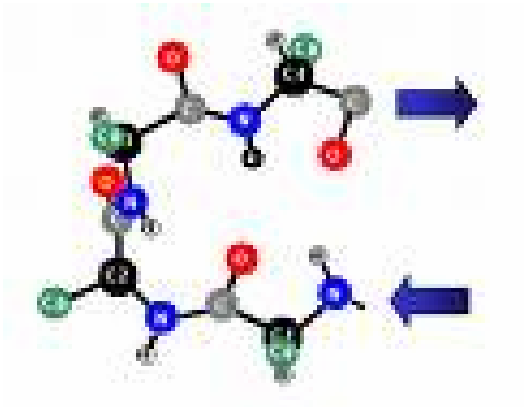




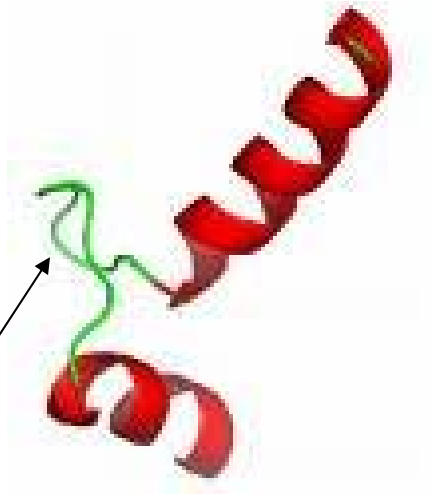
Anti-Parallel

Parallel

Non-Repetitive Secondary Structure



Beta-Turn

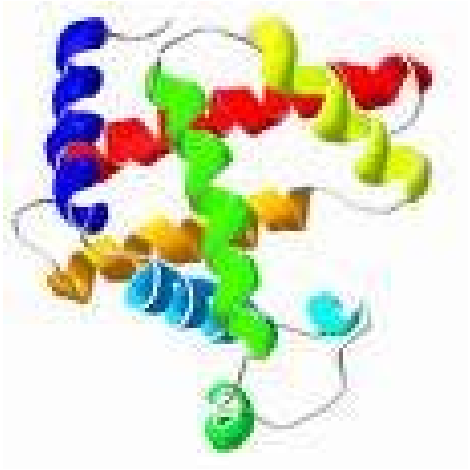


Loop

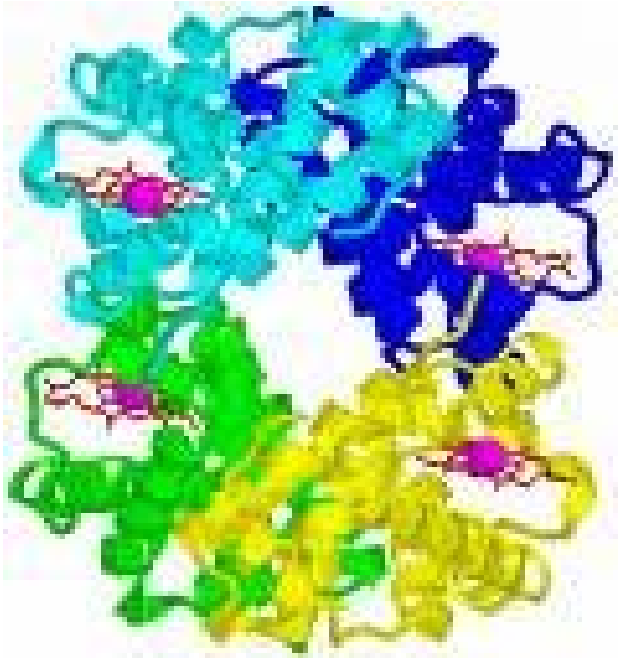
Tertiary Structure

- John Kendrew et al., Myoglobin
- Max Perutz et al., Haemoglobin
- 1962 Nobel Prize in Chemistry

Perutz Kendrew

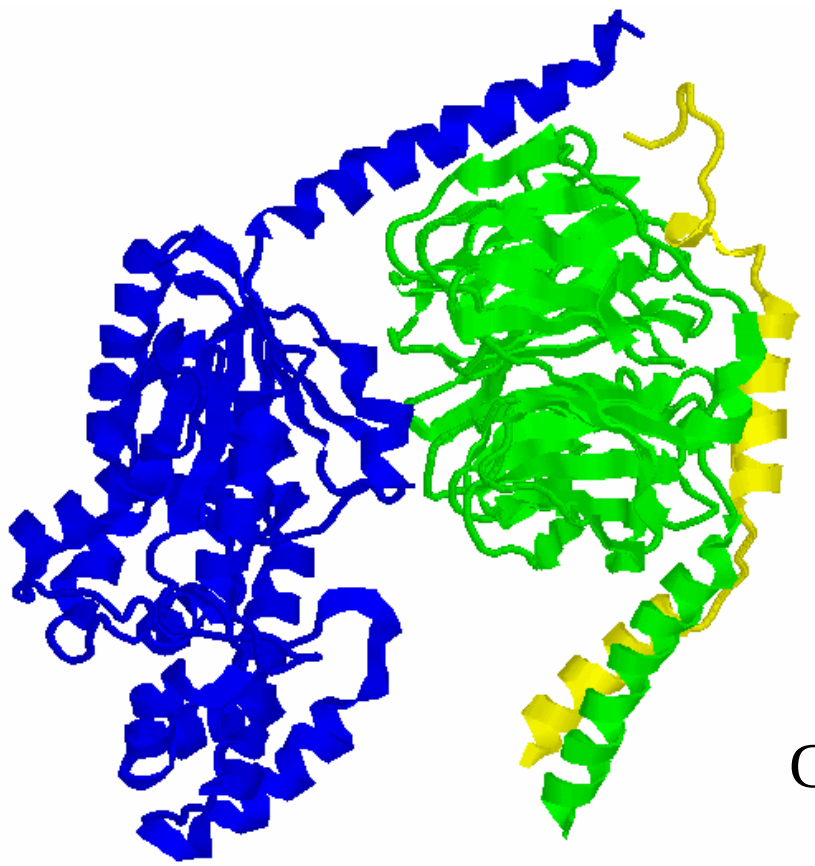


myoglobin



haemoglobin

Quaternary Structure: Complex

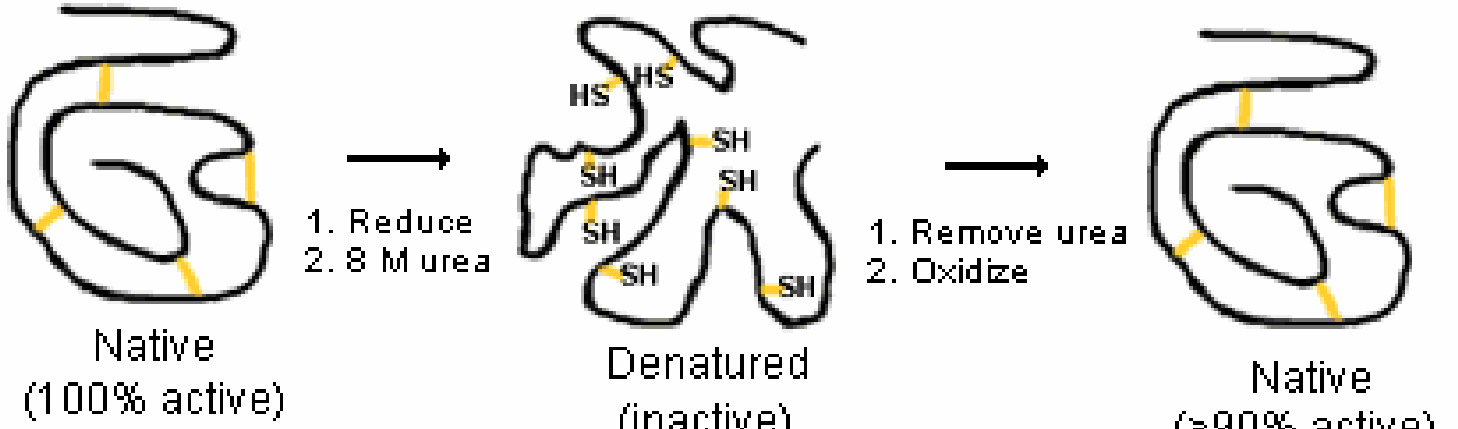


G-Protein Complex

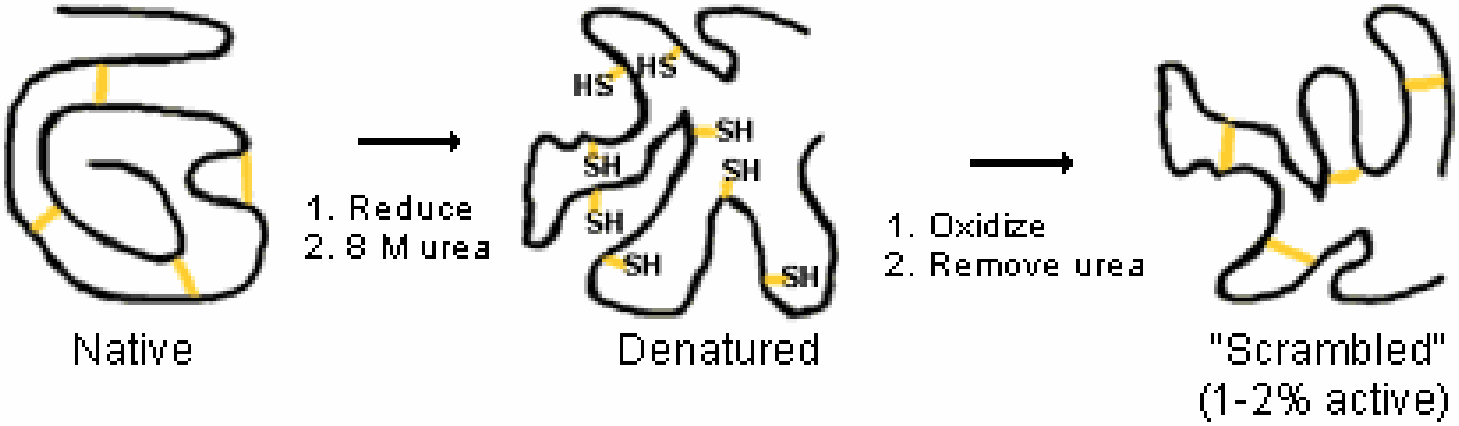
Anfinsen's Folding Experiment

- Structure is uniquely determined by protein sequence
- Protein function is determined by protein structure

The Observation:



The Control:



Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, Analysis, and Comparison

III. Structure Classification

IV. 1D Prediction

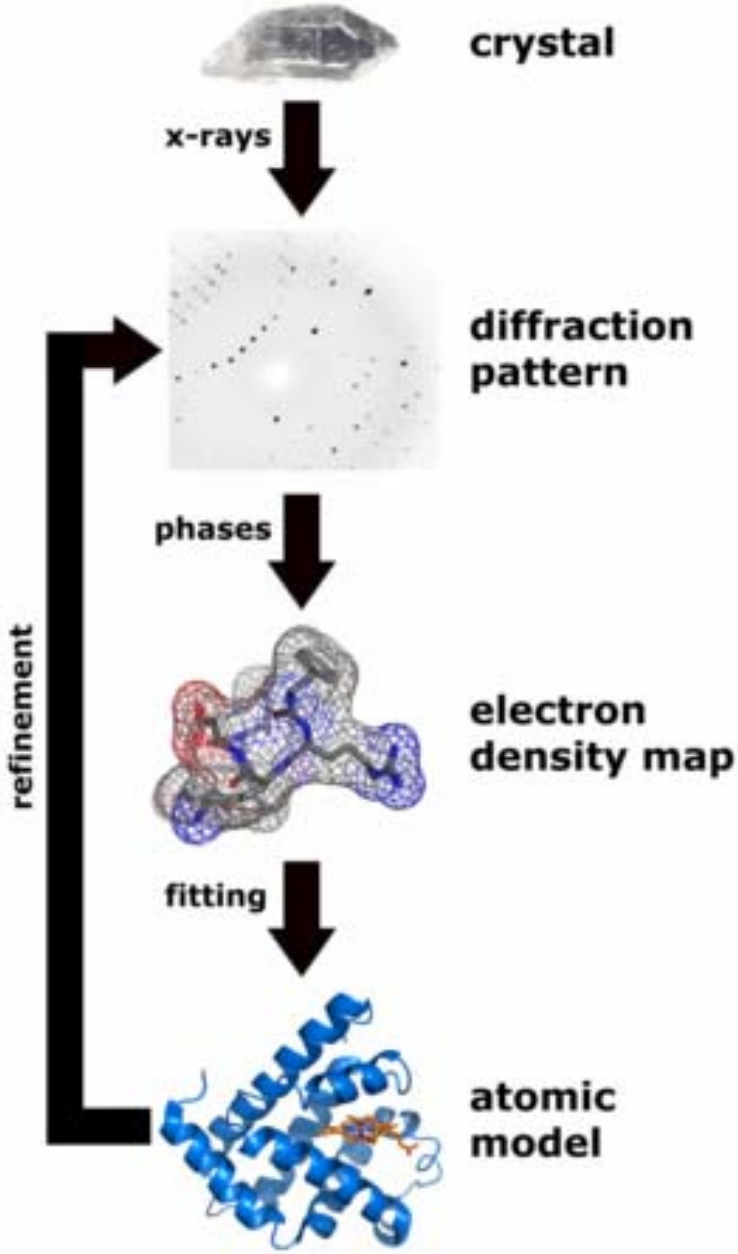
V. 2D Prediction

VI. 3D Prediction

VII. Useful Tools

Protein Structure Determination

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy
- X-ray: any size, accurate (1-3 Angstrom (10^{-10} m)), sometime hard to grow crystal
- NMR: small to medium size, moderate accuracy, structure in solution

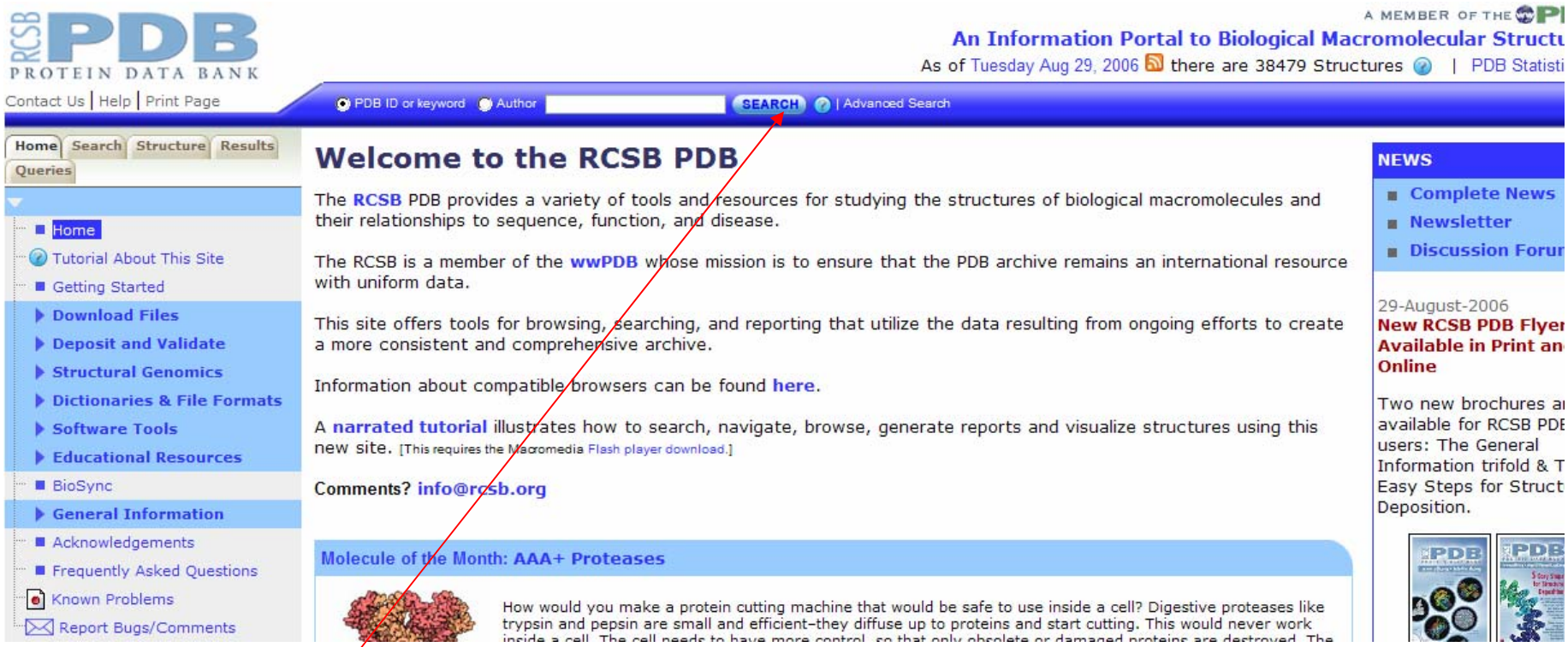


Wikipedia, the free encyclopedia

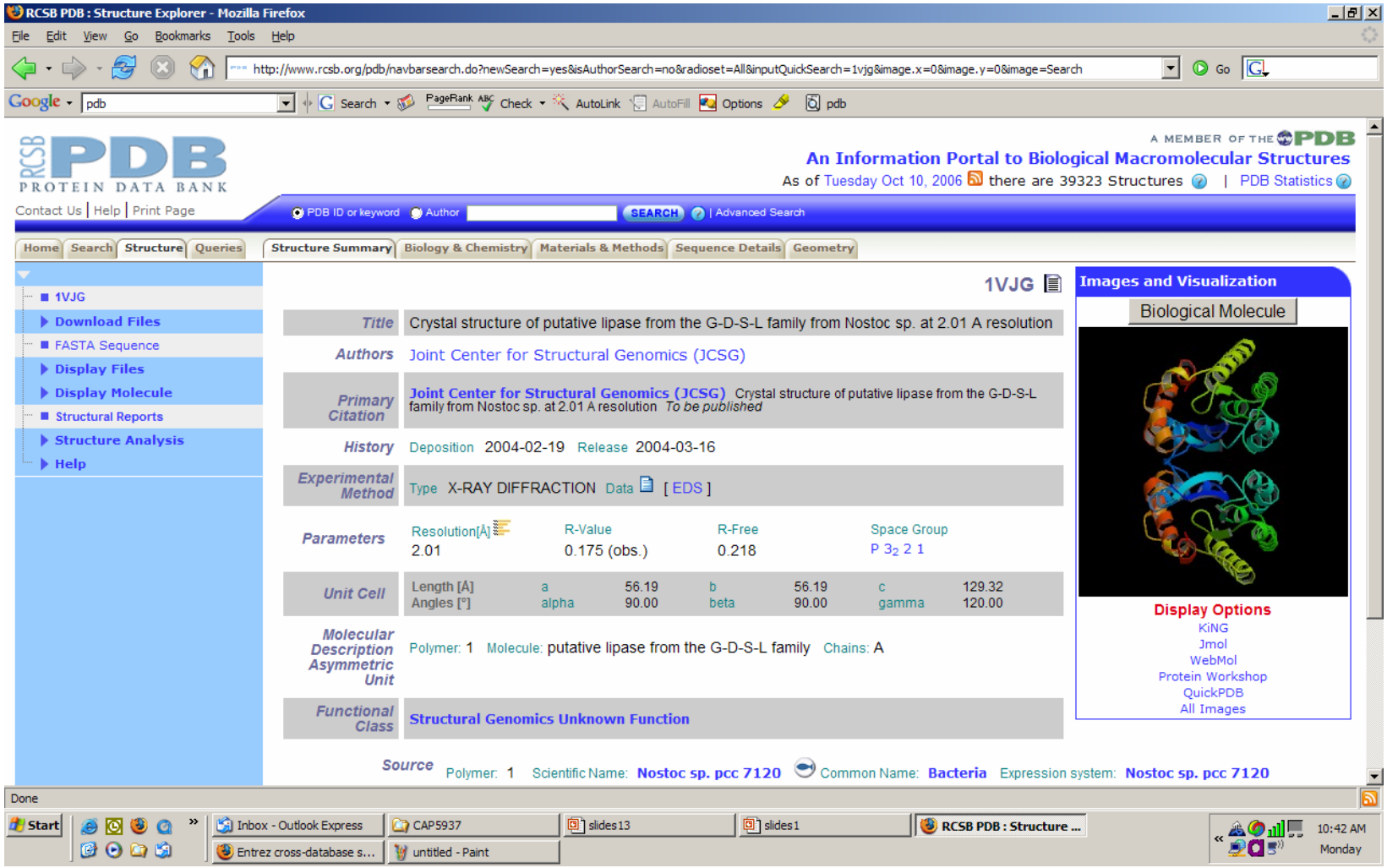
Pacific Northwest National Laboratory's high magnetic field (800 MHz, 18.8 T) NMR spectrometer being loaded with a sample.

Wikipedia, the free encyclopedia

Storage in Protein Data Bank



Search database



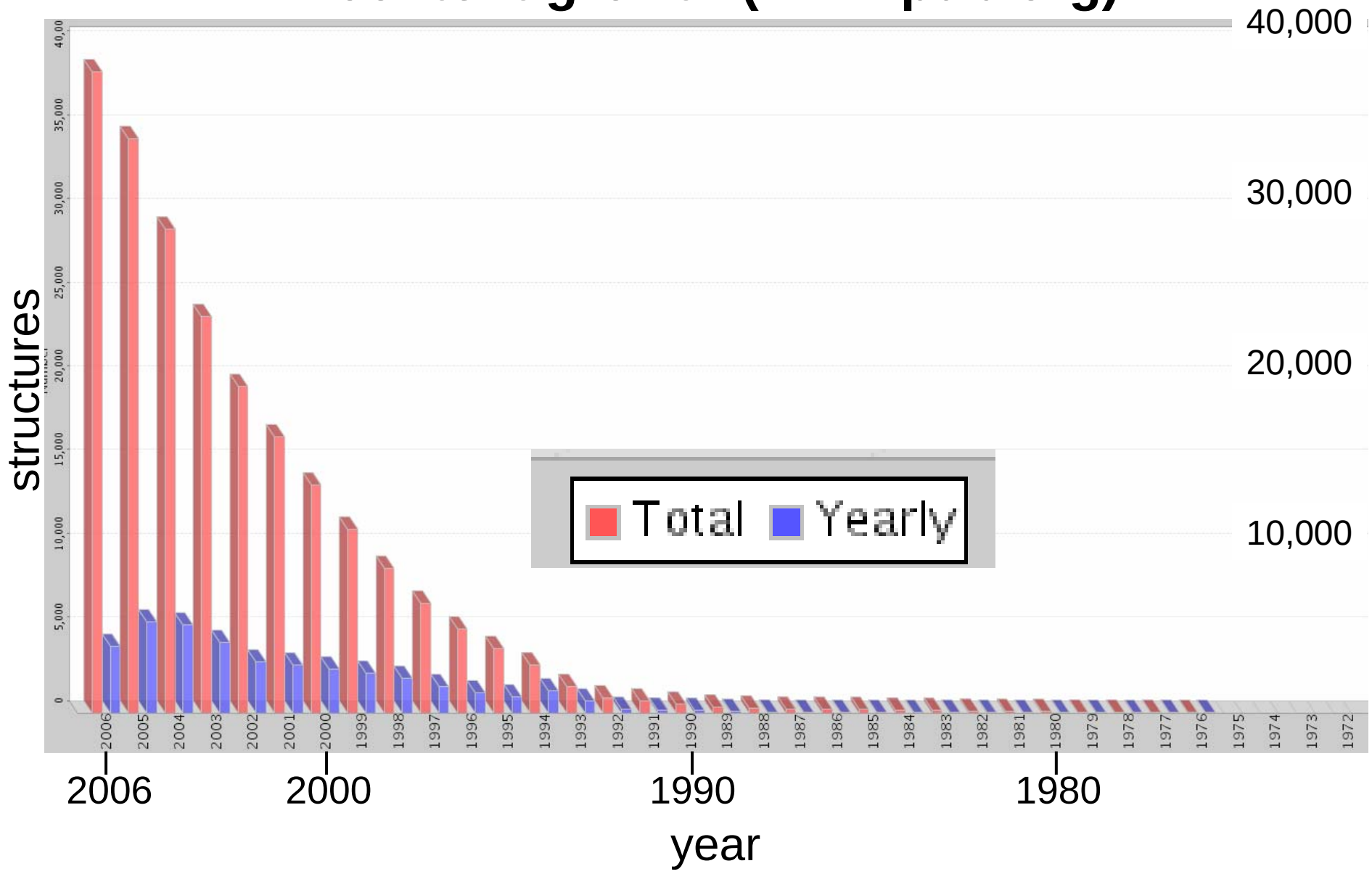
Search protein 1VJG

PDB Format (2C8Q, insulin)

```
HEADER    HORMONE                                 06-DEC-05   2C8Q
TITLE     INSULINE(1SEC) AND UV LASER EXCITED FLUORESCENCE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: INSULIN A CHAIN;
COMPND   3 CHAIN: A;
COMPND   4 MOL_ID: 2;
COMPND   5 MOLECULE: INSULIN B CHAIN;
COMPND   6 CHAIN: B
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGAN: PANCREAS;
SOURCE   5 MOL_ID: 2;
SOURCE   6 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   7 ORGANISM_COMMON: HUMAN;
SOURCE   8 ORGAN: PANCREAS
KEYWDS    LASER, UV, CARBOHYDRATE METABOLISM, HORMONE, DIABETES
KEYWDS   2 MELLITUS, GLUCOSE METABOLISM
EXPDTA    X-RAY DIFFRACTION
AUTHOR    X.VERNEDE,B.LAVAULT,J.OHANA,D.NURIZZO,J.JOLY,L.JACQUAMET,
AUTHOR   2 F.FELISAZ,F.CIPRIANI,D.BOURGEOIS
REVDAT   1   08-MAR-06 2C8Q    0
JRNL        AUTH   X.VERNEDE,B.LAVAULT,J.OHANA,D.NURIZZO,J.JOLY,
JRNL        AUTH 2 L.JACQUAMET,F.FELISAZ,F.CIPRIANI,D.BOURGEOIS
JRNL        TITL   UV LASER-EXCITED FLUORESCENCE AS A TOOL FOR THE
JRNL        TITL 2 VISUALIZATION OF PROTEIN CRYSTALS MOUNTED IN
JRNL        TITL 3 LOOPS.
JRNL        REF    ACTA CRYSTALLOGR.,SECT.D      V.  62   253 2006
JRNL        REFN   ASTM ABCRE6  DK ISSN 0907-4449
REMARK   2
REMARK   2 RESOLUTION. 1.95 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : REFMAC 5.2.0005
REMARK   3   AUTHORS     : MURSHUDOV,VAGIN,DODSON
REMARK   3
REMARK   3   REFINEMENT TARGET : MAXIMUM LIKELIHOOD
```

```
SEQRES   1 A   21  GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU
SEQRES   2 A   21  TYR GLN LEU GLU ASN TYR CYS ASN
SEQRES   1 B   29  PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU
SEQRES   2 B   29  ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR
SEQRES   3 B   29  THR PRO LYS
FORMUL   3  HOH   *31(H2 O1)
HELIX    1   1 GLY A    1  CYS A    7  1                                   7
HELIX    2   2 SER A   12  ASN A   18  1                                   7
HELIX    3   3 GLY B    8  GLY B   20  1                                  13
HELIX    4   4 GLU B   21  GLY B   23  5                                   3
SSBOND   1 CYS A    6    CYS A   11                          1555   1555
SSBOND   2 CYS A    7    CYS B    7                          1555   1555
SSBOND   3 CYS A   20    CYS B   19                          1555   1555
CRYST1   78.608   78.608   78.608  90.00  90.00  90.00 I 21 3        24
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.012721  0.000000  0.000000        0.00000
SCALE2      0.000000  0.012721  0.000000        0.00000
SCALE3      0.000000  0.000000  0.012721        0.00000
ATOM      1  N   GLY A   1      45.324  26.807  11.863  1.00 24.82           N
ATOM      2  CA  GLY A   1      45.123  27.787  12.967  1.00 24.93           C
ATOM      3  C   GLY A   1      43.756  27.627  13.605  1.00 25.16           C
ATOM      4  O   GLY A   1      43.107  26.591  13.438  1.00 25.00           O
ATOM      5  N   ILE A   2      43.313  28.661  14.323  1.00 25.21           N
ATOM      6  CA  ILE A   2      42.050  28.622  15.065  1.00 25.39           C
ATOM      7  C   ILE A   2      40.818  28.303  14.200  1.00 25.69           C
ATOM      8  O   ILE A   2      39.935  27.565  14.635  1.00 25.56           O
ATOM      9  CB  ILE A   2      41.816  29.917  15.917  1.00 25.39           C
```

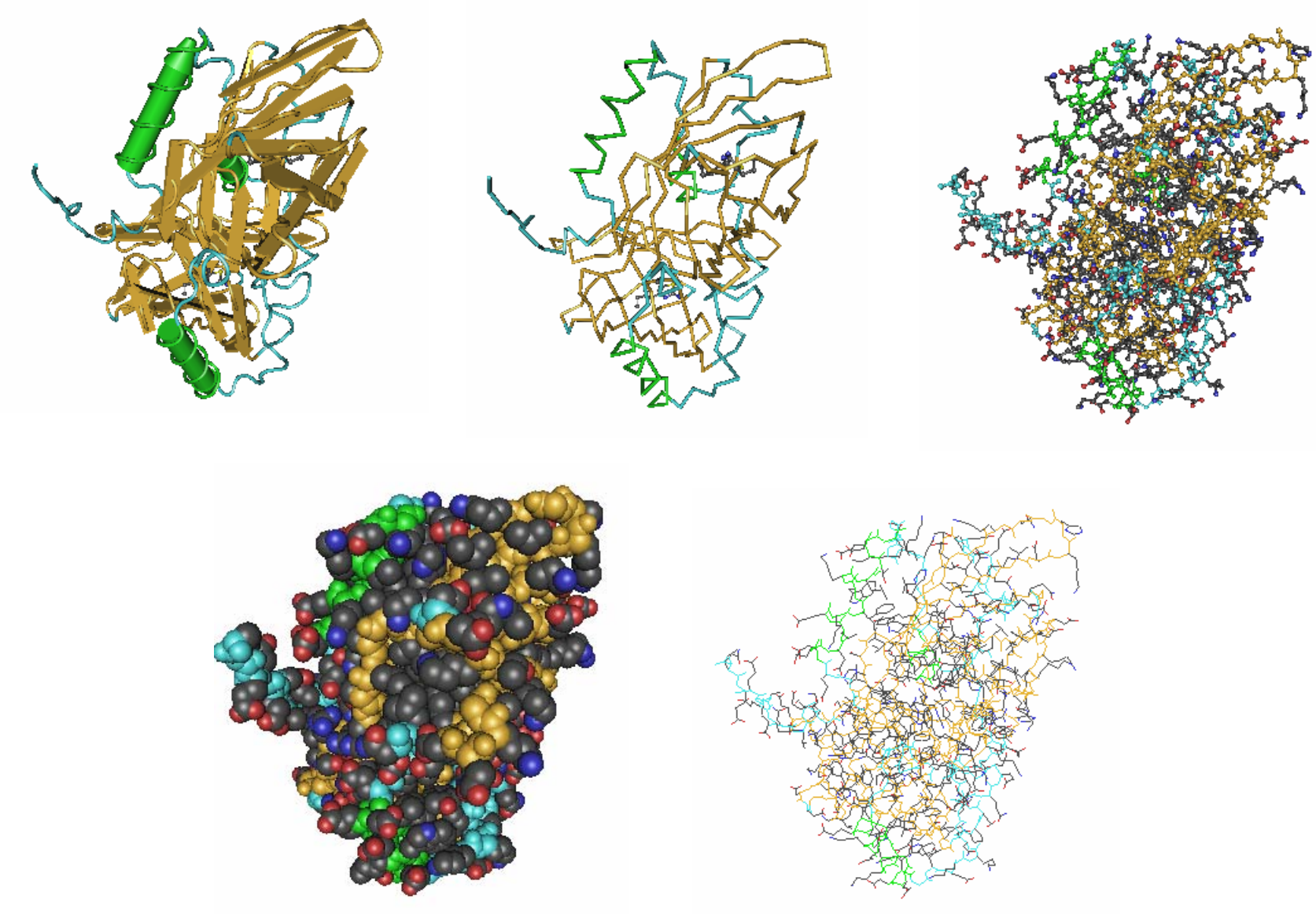
PDB content growth (www.pdb.org)



J. Pevsner, 2005

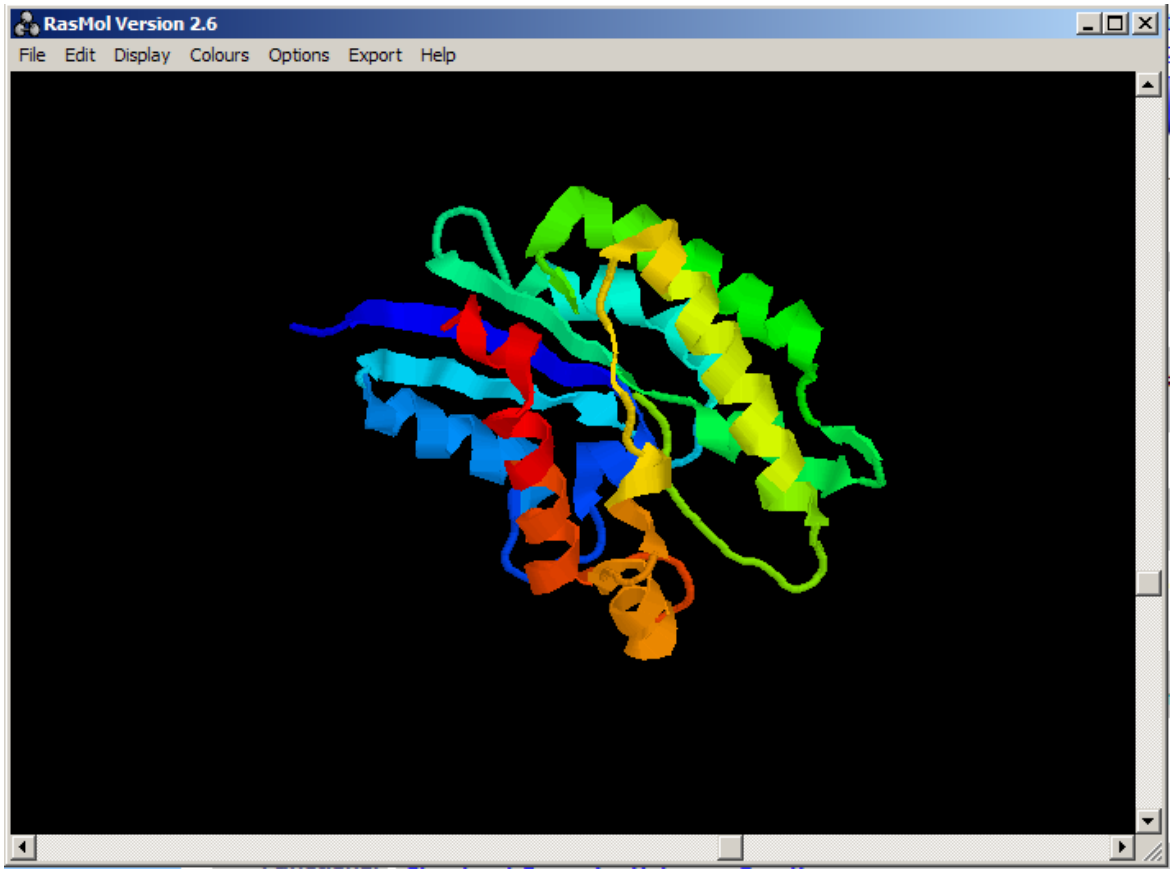
Structure Visualization

- Rasmol (http://www.umass.edu/microbio/rasmol/getras.htm)
- MDL Chime (plug-in) (http://www.mdl.com/products/framework/chime/)
- Protein Explorer (http://molvis.sdsc.edu/protexpl/frntdoor.htm)
- Online tools (PDB sites)



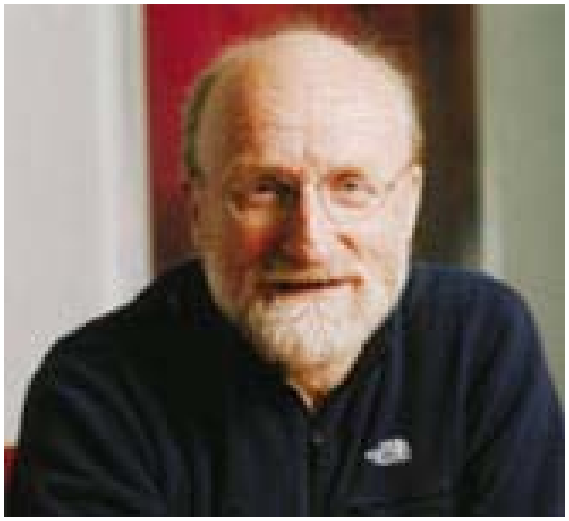
J. Pevsner, 2005

Demo Using Rasmol (1VJG)



Structure Analysis

- Assign secondary structure for amino acids from 3D structure
- Generate solvent accessible area for amino acids from 3D structure
- Most widely used tool: DSSP (Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. **Kabsch and Sander**)



DSSP server: http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html

DSSP download: http://swift.cmbi.ru.nl/gv/dssp/

Chris Sander
One of the founders of Bioinformatics and Computational Biology

DSSP Code:

H = alpha helix

G = 3-helix (3/10 helix)

I = 5 helix (pi helix)

B = residue in isolated beta-bridge

E = extended strand, participates in beta ladder

T = hydrogen bonded turn

S = bend

Blank = loop

DSSP Web Service

DSSP : Definition of secondary structure of proteins given a set of 3D coordinates (W.Kabsch, C. Sander)

Reset | Run dssp | jianlin.cheng@gmail.com your e-mail

PDB File

1vjg or you can instead enter a PDB id.

http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html

DSSP : Definition of secondary structure of proteins given a set of 3D coordinates (W.Kabsch, C. Sander)

Results:

dssp.out (31.16 Ko)

predator | Run the selected program on dssp.out

standard error file

From now, this files will remain accessible for 10 days at: http://bioweb.pasteur.fr/seqanal/tmp/dssp/A58289116100642/
You can save them individually by the **Save file** function if needed.

Job summary | default format

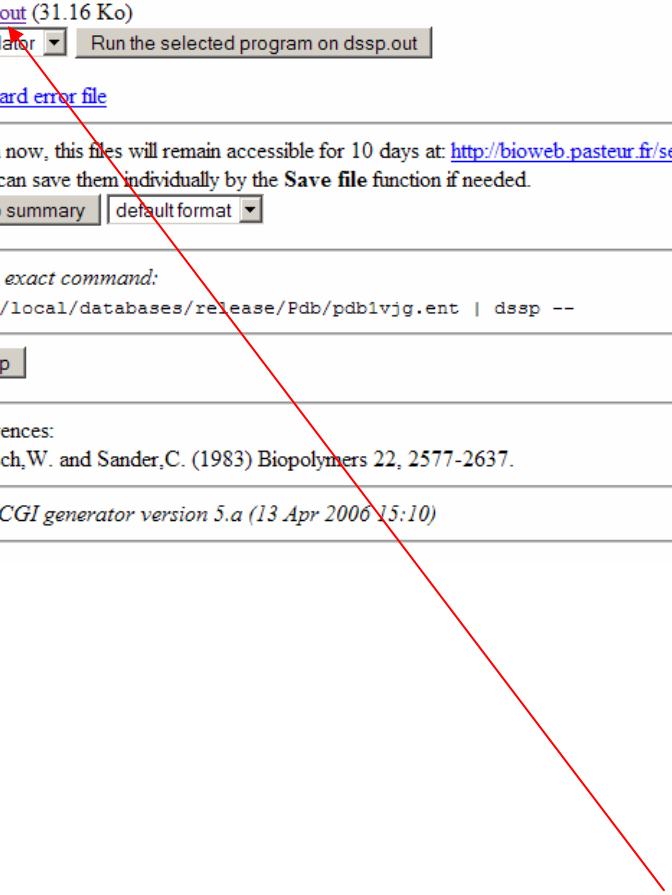
Unix exact command:

```
cat /local/databases/release/Pdb/pdb1vjg.ent | dssp --
```

Help

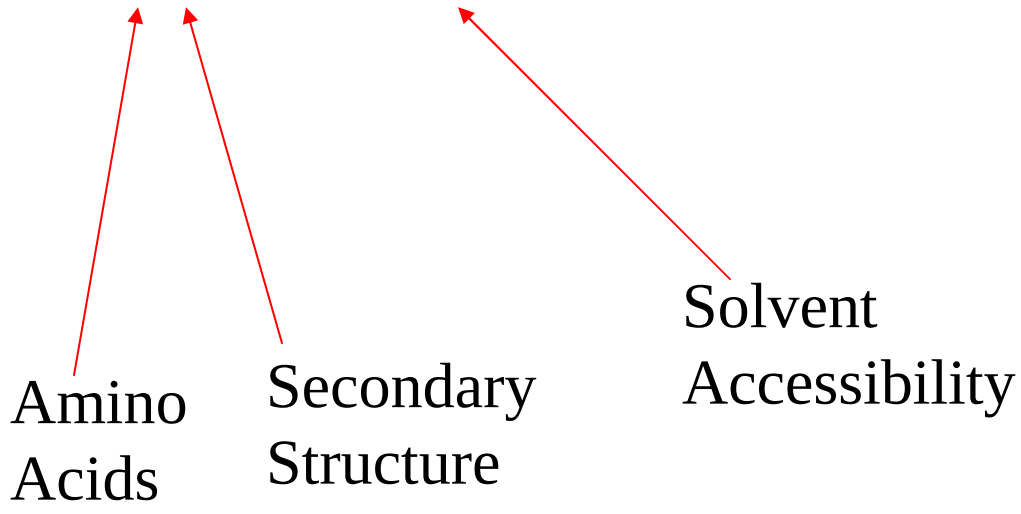
References:
Kabsch,W. and Sander,C. (1983) Biopolymers 22, 2577-2637.

Pise CGI generator version 5.a (13 Apr 2006 15:10)



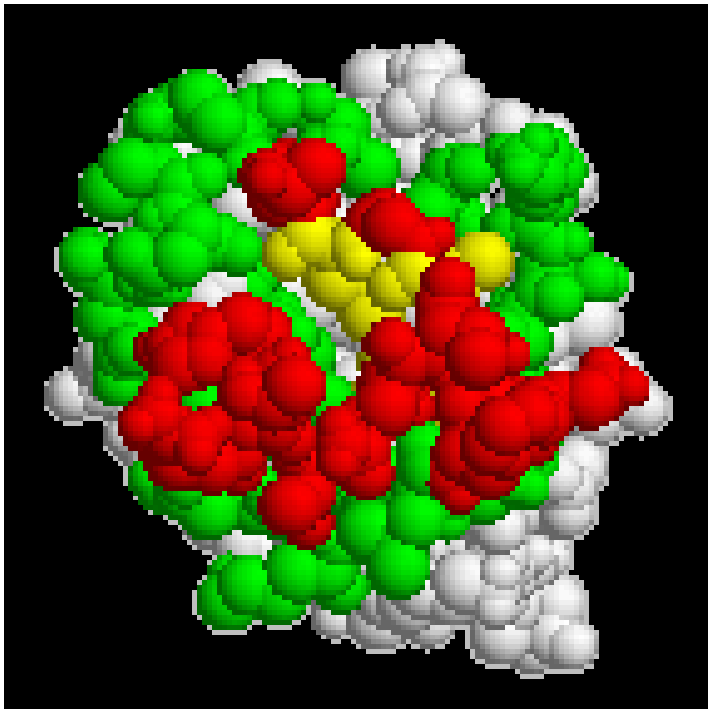
Click here

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N	TCO	KAPPA	ALPHA	PHI	PSI	X-CA	Y-CA	Z-CA
1	5	A	S		0	0	179	0, 0.0	2,-0.0	0, 0.0	0, 0.0	0.000	360.0	360.0	360.0	125.7	-8.6	43.0	43.9
2	6	A	K	-	0	0	123	1,-0.1	2,-0.4	37,-0.1	37,-0.2	-0.235	360.0	-108.7	-87.0	151.4	-7.5	41.4	40.6
3	7	A	T	E -a	39	0A	75	35,-0.6	37,-2.5	1,-0.0	2,-0.3	-0.593	34.7	-132.0	-72.2	128.3	-4.3	39.5	39.6
4	8	A	Q	E +a	40	0A	91	-2,-0.4	69,-0.6	35,-0.2	2,-0.4	-0.639	26.0	179.8	-86.4	132.7	-2.0	41.5	37.4
5	9	A	I	E -ab	41	73A	3	35,-1.9	37,-2.9	-2,-0.3	2,-0.5	-0.991	13.3	-156.5	-129.4	131.5	-0.7	39.9	34.2
6	10	A	R	E -ab	42	74A	48	67,-2.8	69,-1.7	-2,-0.4	2,-0.4	-0.910	14.8	-173.2	-105.2	126.8	1.6	41.6	31.8
7	11	A	I	E -ab	43	75A	0	35,-2.5	37,-2.6	-2,-0.5	2,-0.5	-0.983	11.9	-162.4	-124.9	124.4	1.7	40.3	28.2
8	12	A	C	E -ab	44	76A	0	67,-2.3	69,-2.6	-2,-0.4	2,-0.6	-0.931	6.5	-159.9	-100.8	130.8	3.9	41.2	25.3
9	13	A	F	E -ab	45	77A	0	35,-2.2	37,-3.0	-2,-0.5	2,-0.5	-0.955	13.2	-169.0	-109.5	117.1	2.7	40.2	21.8
10	14	A	V	E +ab	46	78A	0	67,-3.1	69,-2.2	-2,-0.6	2,-0.3	-0.926	34.8	71.1	-116.5	129.9	5.6	40.1	19.4
11	15	A	G	E S-ab	47	79A	0	35,-0.9	37,-1.9	-2,-0.5	69,-0.2	-0.921	70.2	-50.2	169.0	-146.4	5.3	39.9	15.6
12	16	A	D	S >> S-	0	0	4	67,-0.8	4,-2.2	-2,-0.3	3,-0.6	-0.023	78.2	-51.3	-111.5	-151.8	4.2	41.6	12.4
13	17	A	S	H 3>>S+	0	0	7	35,-0.3	5,-1.7	1,-0.2	4,-1.5	0.803	130.2	57.8	-67.3	-28.8	1.2	43.5	11.1
14	18	A	F	H 345S+	0	0	5	2,-0.2	12,-0.5	1,-0.2	-1,-0.2	0.884	108.5	46.5	-68.2	-33.2	-1.2	40.8	12.2
15	19	A	V	H <45S+	0	0	1	-3,-0.6	12,-0.3	64,-0.2	-2,-0.2	0.900	111.1	52.2	-68.9	-41.4	-0.0	41.1	15.7
16	20	A	N	H <5S-	0	0	71	-4,-2.2	-2,-0.2	30,-0.1	-1,-0.2	0.774	110.8	-127.0	-62.6	-26.6	-0.3	45.0	15.4
17	21	A	G	T ><5 -	0	0	5	-4,-1.5	3,-2.2	-5,-0.2	8,-0.4	0.741	36.4	-174.6	83.1	25.3	-3.9	44.5	14.2
18	22	A	T	T 3 < +	0	0	14	-5,-1.7	-1,-0.2	1,-0.3	-2,-0.0	-0.199	68.4	29.2	-54.0	135.4	-3.4	46.6	11.0
19	23	A	G	T 3 S+	0	0	28	1,-0.3	-1,-0.3	159,-0.1	162,-0.2	0.121	86.2	120.8	94.7	-21.4	-6.7	47.0	9.2
20	24	A	D	X -	0	0	9	-3,-2.2	3,-1.2	160,-0.2	-1,-0.3	-0.706	48.9	-160.5	-79.7	117.6	-8.9	46.8	12.4
21	25	A	P	T 3 S+	0	0	91	0, 0.0	-1,-0.2	0, 0.0	159,-0.0	0.677	91.8	60.1	-70.9	-17.3	-10.9	50.1	12.6
22	26	A	E	T 3 S-	0	0	119	-3,-0.0	-2,-0.1	3,-0.0	158,-0.0	0.426	105.0	-132.3	-87.9	-3.3	-11.4	49.4	16.3
23	27	A	C	S < S+	0	0	112	-3,-1.2	-5,-0.1	-6,-0.2	-6,-0.0	0.730	80.2	98.1	62.8	28.1	-7.6	49.4	16.9



Solvent Accessibility

Size of the area of an amino acid that is exposed to solvent (water).



Maximum solvent accessible area for each amino acid is its whole surface area.

Hydrophobic residues like to be Buried inside (interior).
Hydrophilic residues like to be exposed on the surface.

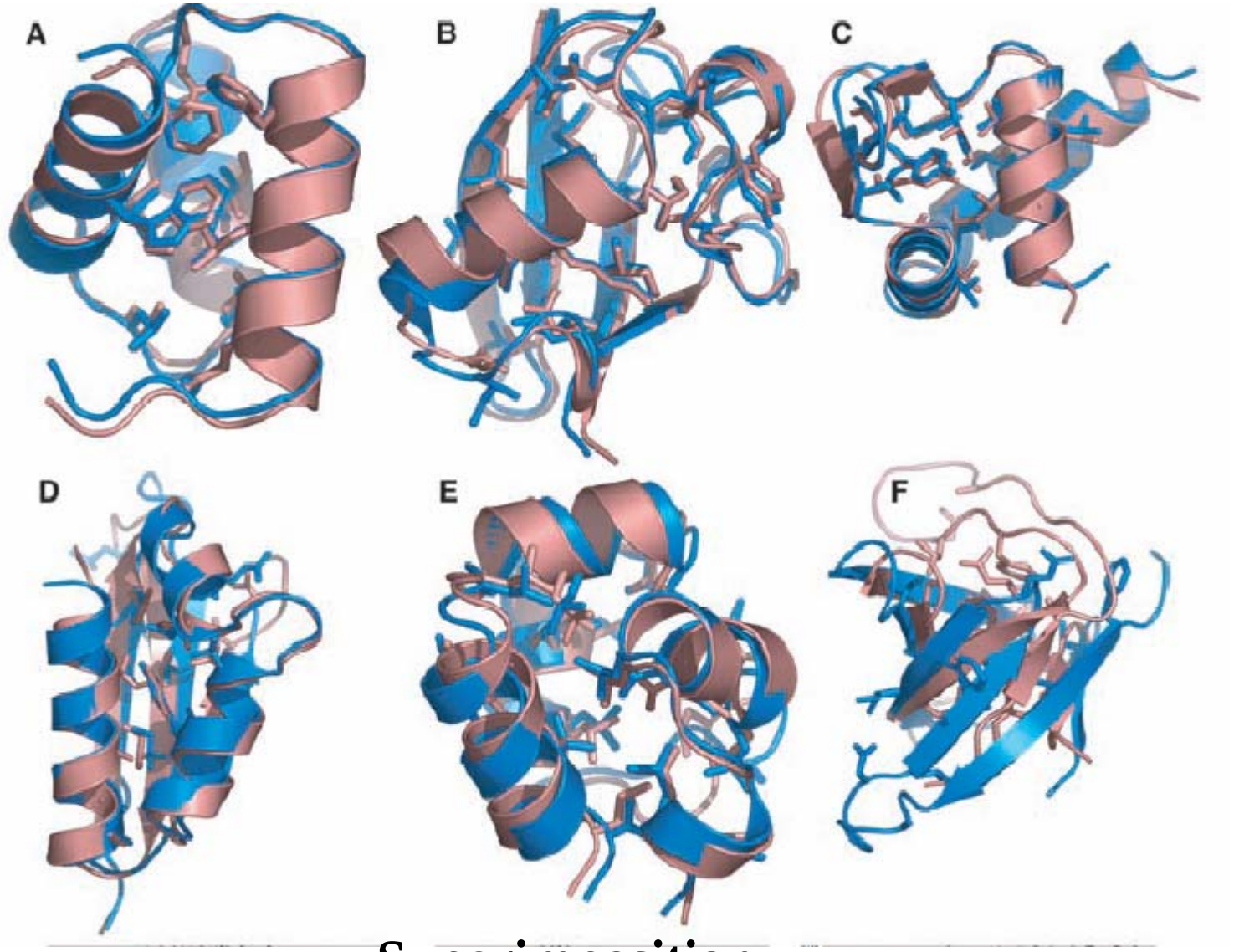
Structure Comparison (Alignment)

- Are the structures of two protein similar?
- Are the two structure models of the same protein similar?
- Different measures (RMSD, GDT-TS (Zemla et al., 1999), MaxSub (Siew et al., 2000), TM score (Zhang and Skolnick,2005))

Basic Idea

- Try to superimpose two structures to minimize the distances between corresponding atoms (residues)
- Hard geometric optimization problem
- Alignment problem with very complex scoring function (treat each position as one character)

Root Mean Square Deviation $$RMSD = \sqrt{\frac{\sum_{i=1}^{i=n} ((a_{ix} - b_{ix})^2 + (a_{iy} - b_{iy})^2 + (a_{iz} - b_{iz})^2)}{n}}$$



Superimposition

David Baker, 2005

Useful Structure Alignment Tools

- CE (http://cl.sdsc.edu/)
- DALI (http://www.ebi.ac.uk/dali/)
- VAST (NCBI)

CE CALCULATE TWO CHAINS

Calculate structural alignment for two polypeptide chains either from the PDB or uploaded by the user.

Specify two polypeptide chains and optionally the similarity level and use of sequence information and then press the "Calculate Alignment" button. Selecting the appropriate ? will provide help on that spe

Calculate Alignment | Reset Form

Select Similarity Level: Medium ?

☐ Use Sequence Information *(optional)* ?

Chain 1:	○ PDB: 4HHB:A ? OR ● User File: C:\casp7\301\foldpro1.pdb Browse... Chain ID: ? ☐ Use Fragment From: To: *(optional)* ? Sequence numbering
Chain 2:	○ PDB: 4HHB:B ? OR ● User File: C:\casp7\301\ROBETTA_TS1.pdl Browse... Chain ID: ? ☐ Use Fragment From: To: *(optional)* ? Sequence numbering

USR1:_(size=395) vs USR2:_(size=395) Structure Alignment

Rmsd = 2.4Å Z-Score = 6.6
Sequence identity = 42.8%
Aligned/gap positions = 332/105

Sequence alignment based on structure alignment.

Sequence alignment based on structure alignment. Position numbers according to sequence (starting from 1) and according to PDB are given as SSSS/PPPP, SSSS - sequence, PPPP - PDB.

USR1:_ -

USR2:_ -

```
USR1:_     4/5     PPQIRIPATYLRGGTSKGVFFRLEDLPE-------SCRVPGEARDRLFMRVIGSPDPYAA
USR2:_     6/7     QIRIPATYLRGGTSKGVFFRL------EDLPESCRVPGEARDRLFMRVIGSPDPYA---A

USR1:_    57/58    HIDGMGGATSSTSKCVILSKSSQPGHDVDYLYGQVSIDKPFVDWSGNCGNLSTGAGAFAL
USR2:_    57/58    HIDGMGGATSSTSKCVILSKSSQPGHDVDYLYGQVSIDKPFVDWSGNCGNLSTGAGAFAL

USR1:_   117/118   HAGLVDPARIPEDGICEVRIWQANIGKTIIAHVPVSGGQVQETGDFELDGVTFPAAEIVL
USR2:_   117/118   HAGLVDPARIPEDGICEVRIWQANIGKTIIAHVPVSGGQVQETGDFELDGVTFPAAEIVL
```

Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, and Comparison

III. Structure Classification

IV. 1D Prediction

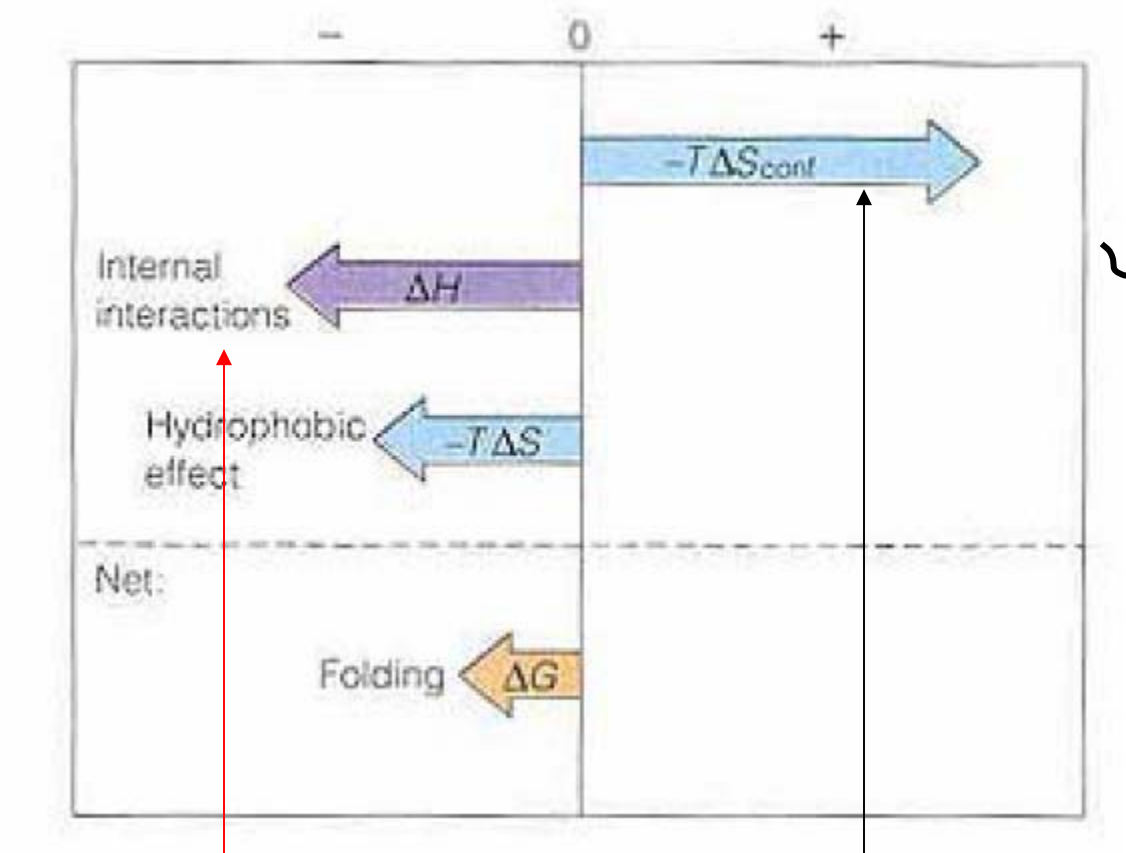
V. 2D Prediction

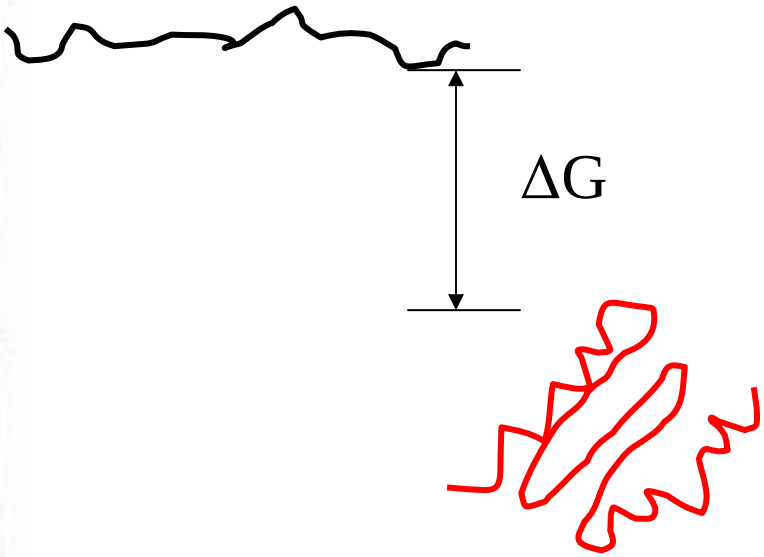
VI. 3D Prediction

VII. Useful Tools

Protein Folding

Favor **Disfavor**





H-bond, Charge-Charge Interaction
Charge-Polar interaction
Polar-Polar interaction

Entropy

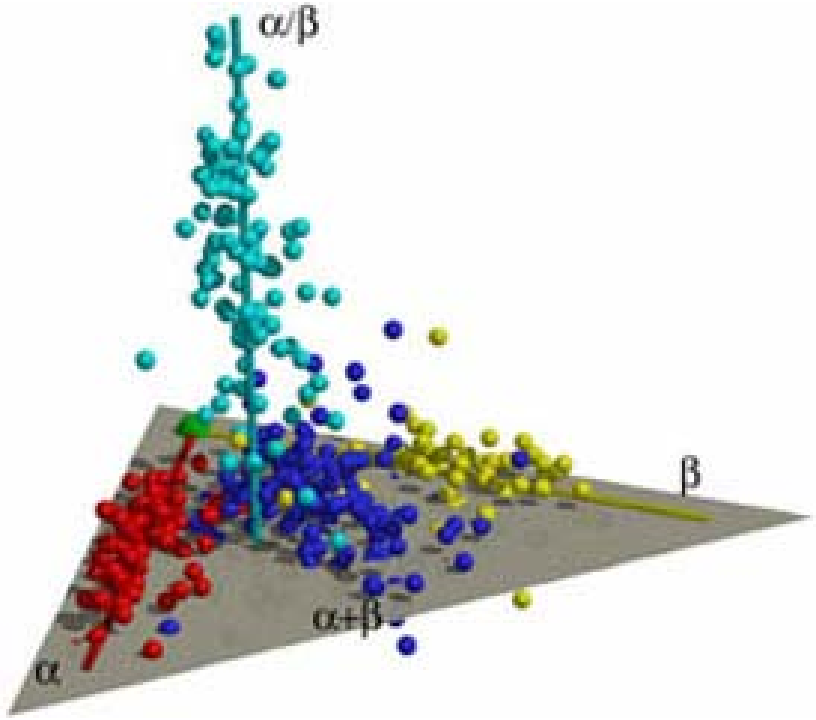
Magnitude of ΔG

- ΔG is small, only about 1-3 H-Bonds.
- A small ΔG is critical for maintenance of the conformation flexibility of proteins in biochemical processes.
- Question: can any amino acid sequences fold into a stable protein structure?

Protein Structure Classification

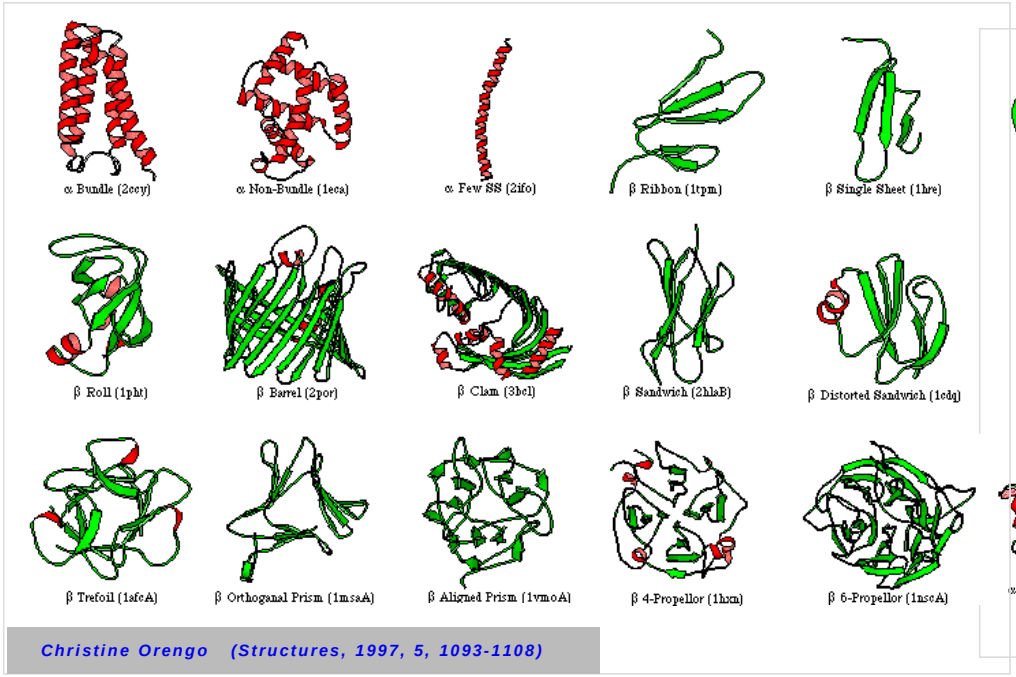
- About 3 million known protein sequences
- About 30,000 known protein structures in PDB
- Many protein structures are similar due to evolutionary relationship
- Many protein structures are similar due to convergent evolution
- Number of unique structure topologies is estimated to be limited (1000 – 1500?)
- Number of protein sequences is huge (20^{300})

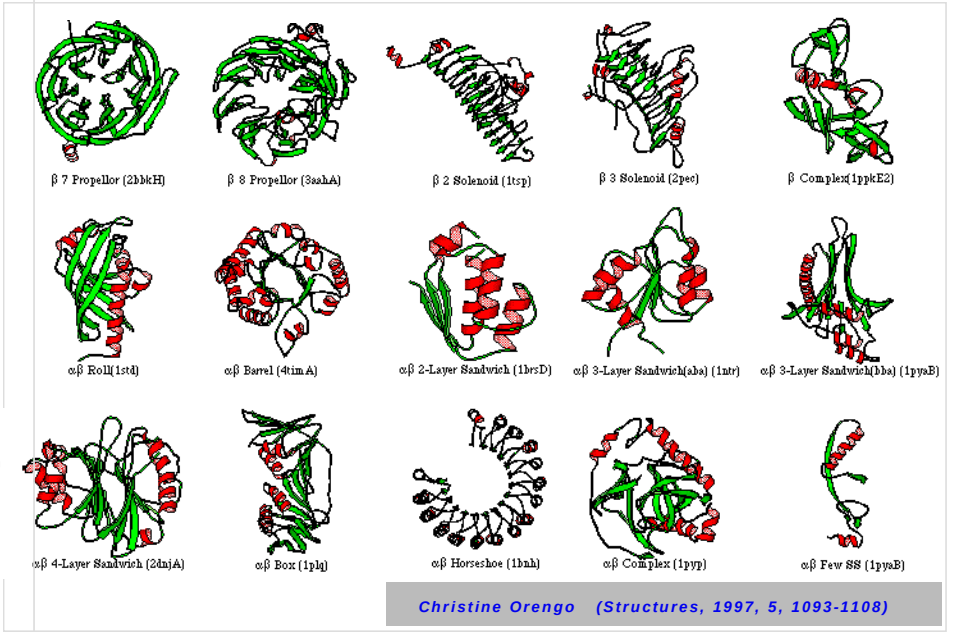
Protein Structure Universe



Proteins. One thousand families for the molecular biologist. C. Chothia. Nature, 1992.

Colors in the universe of protein structures

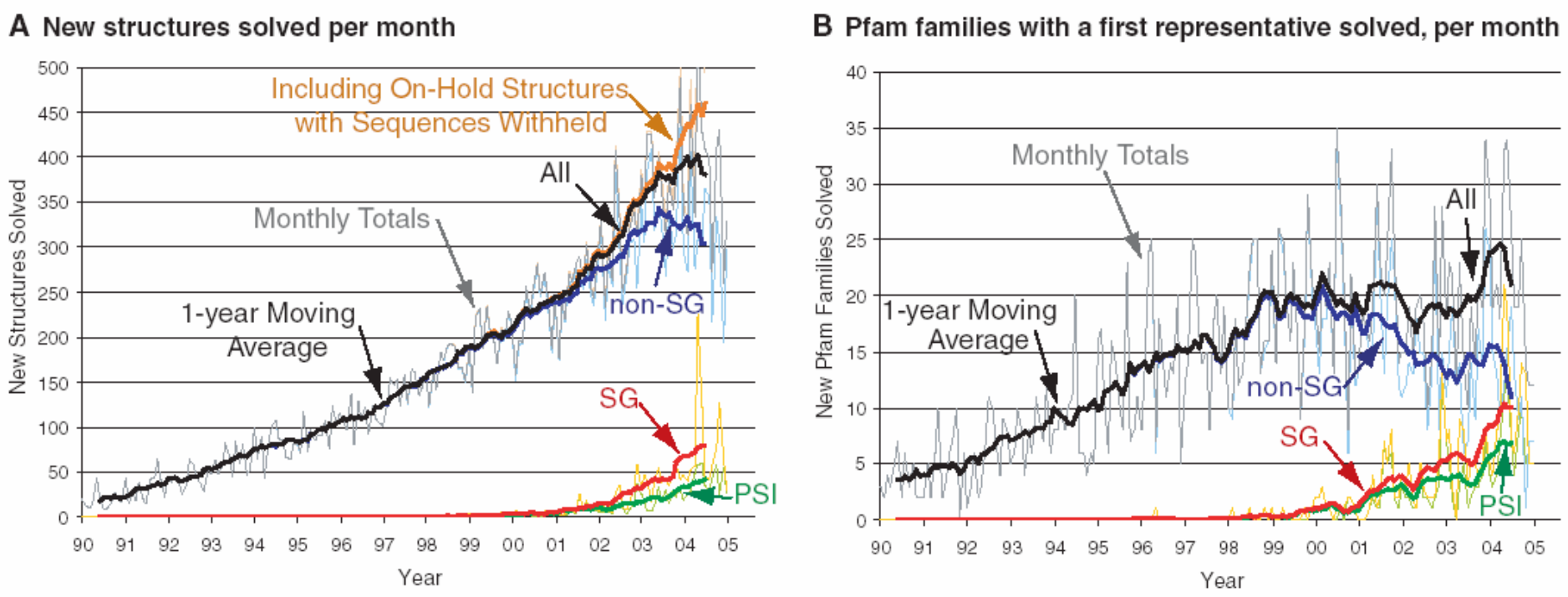




Christine Orengo **1997** *Structures* **5** 1093-1108

B. Rost, 2005

Mapping Protein Universe: Structural Genomics



Chandonia and Brenner, 2005

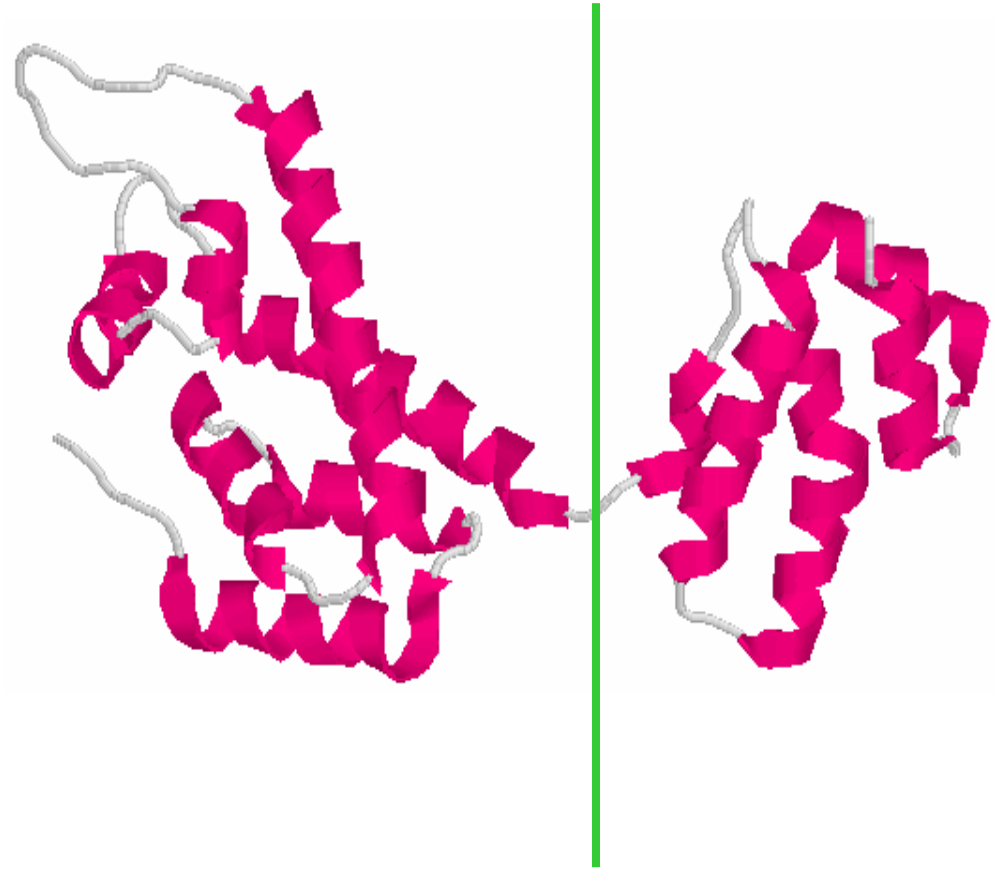
Why is the evolution of protein structure so slow?

- Protein sequence evolves faster than protein structure
- Protein structure is more conserved due to function constraints
- Nature reuses existing folds for new functions (pretty much like programming paradigm in computer science)

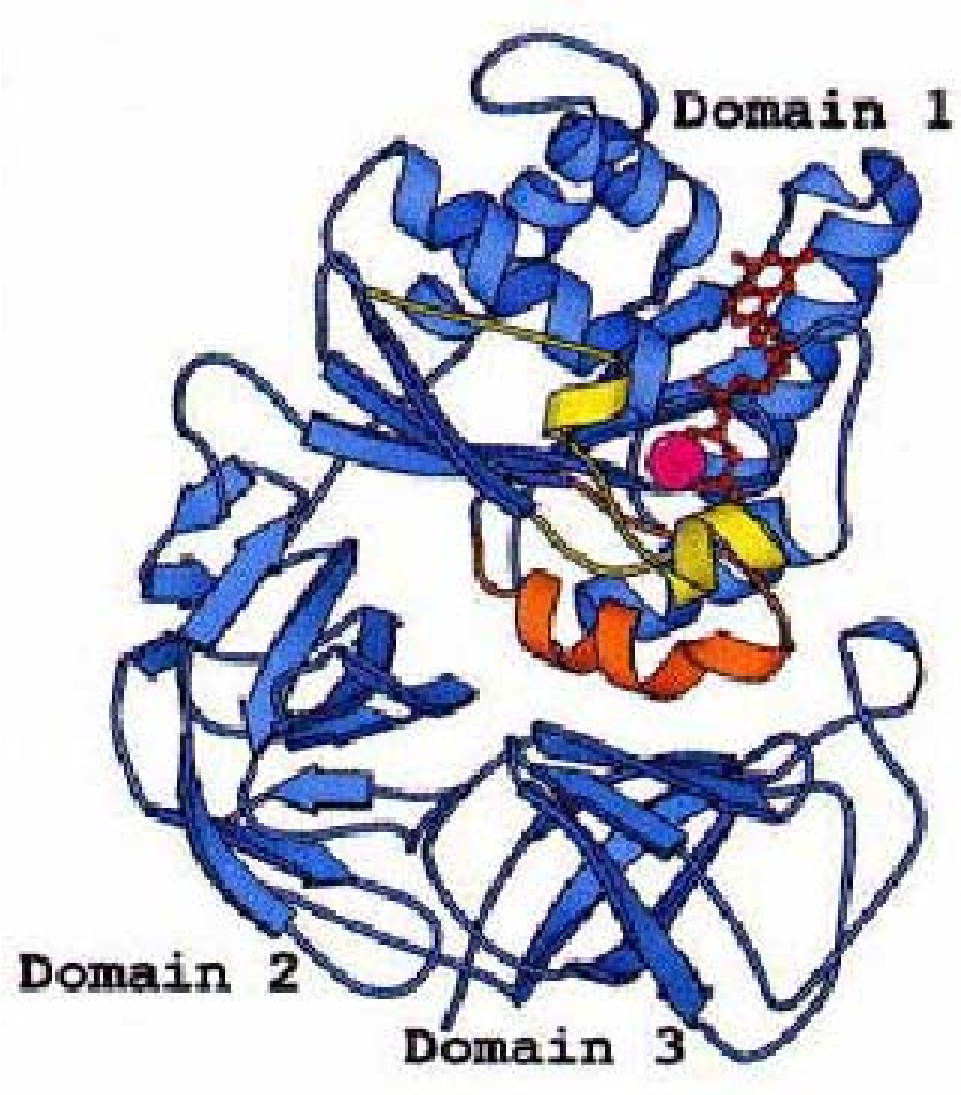
Domain and Fold

- Protein domain is the structural (also functional) unit.
- Protein domain is usually defined as a chunk of protein (usually continuous sub-sequences, but not always) that can fold independently into its tertiary structure without other parts of the protein
- Fold is the topology of a protein or domain: connectivity of secondary structure elements.

Domain Example (HIV-1 Capsid)



Another Domain Example



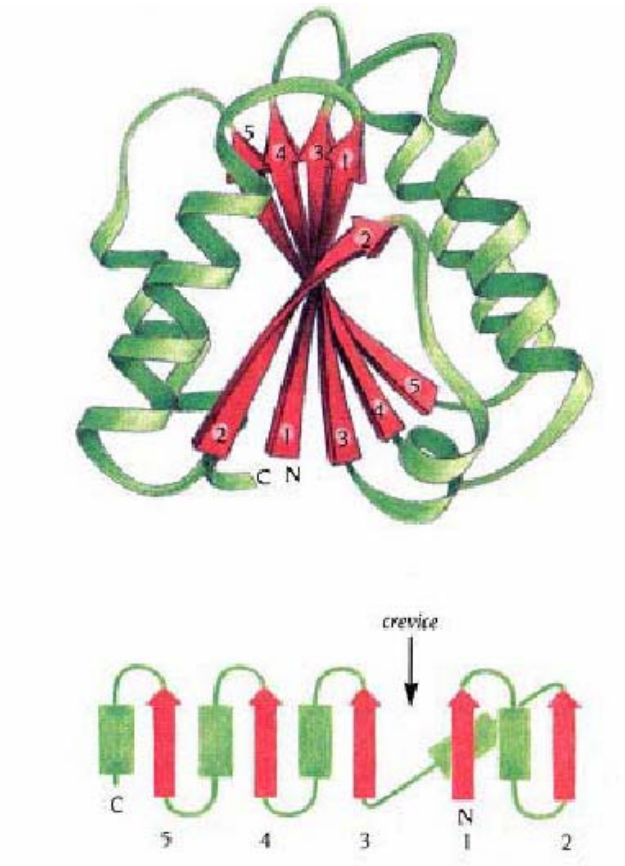
http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1eft;pr=213-312

Domain Parsing Tools

- PDP (protein domain parser) (http://123d.ncifcrf.gov/pdp.html)
- Domain Parser (http://csbl.bmb.uga.edu/downloads/)

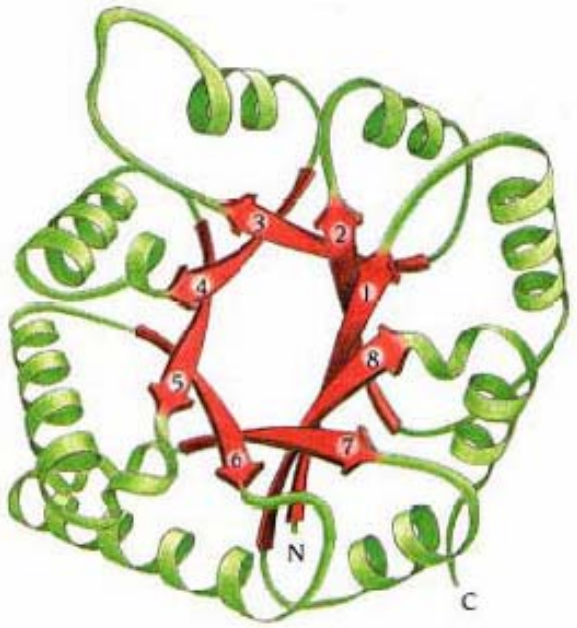
Typical Folds

- Fold: connectivity or arrangement of secondary structure elements.
- NAD-binding Rossman fold
- 3 layers, a/b/a, parallel beta-sheet of 3 strands. Order: 321456

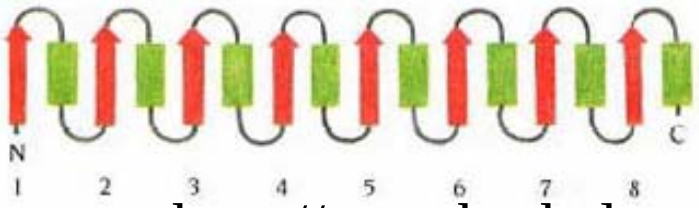


http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1b16;pc=a

Another Fold Example: TIM beta-alpha barrel

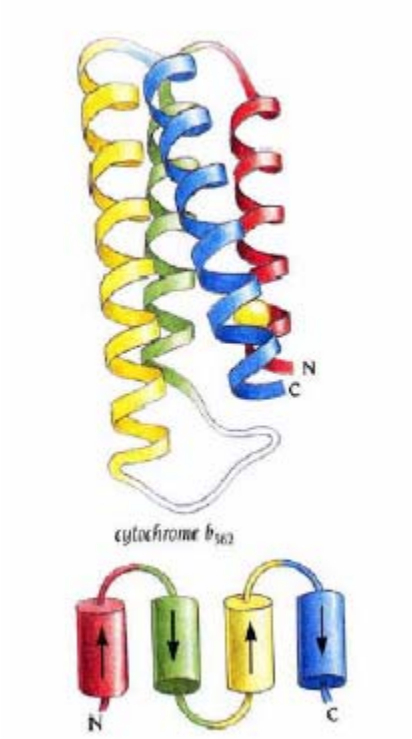


Contains parallel beta-sheet Barrel, closed. 8 strands. Order 1,2,3,4,5,6,7,8.



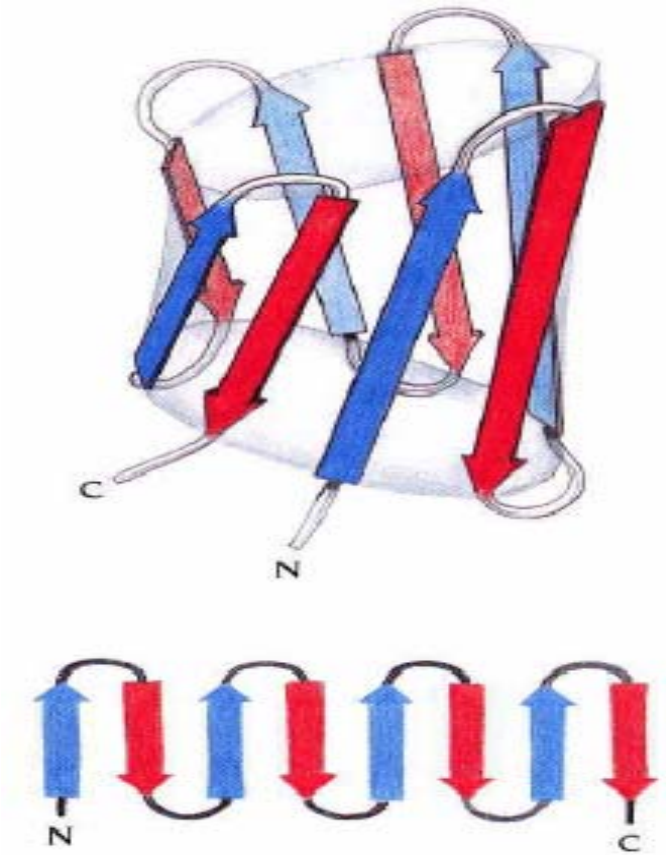
http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1hti;pc=a

Fold: Helix Bundle (human growth factor)



http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1hgu

Fold: beta barrel

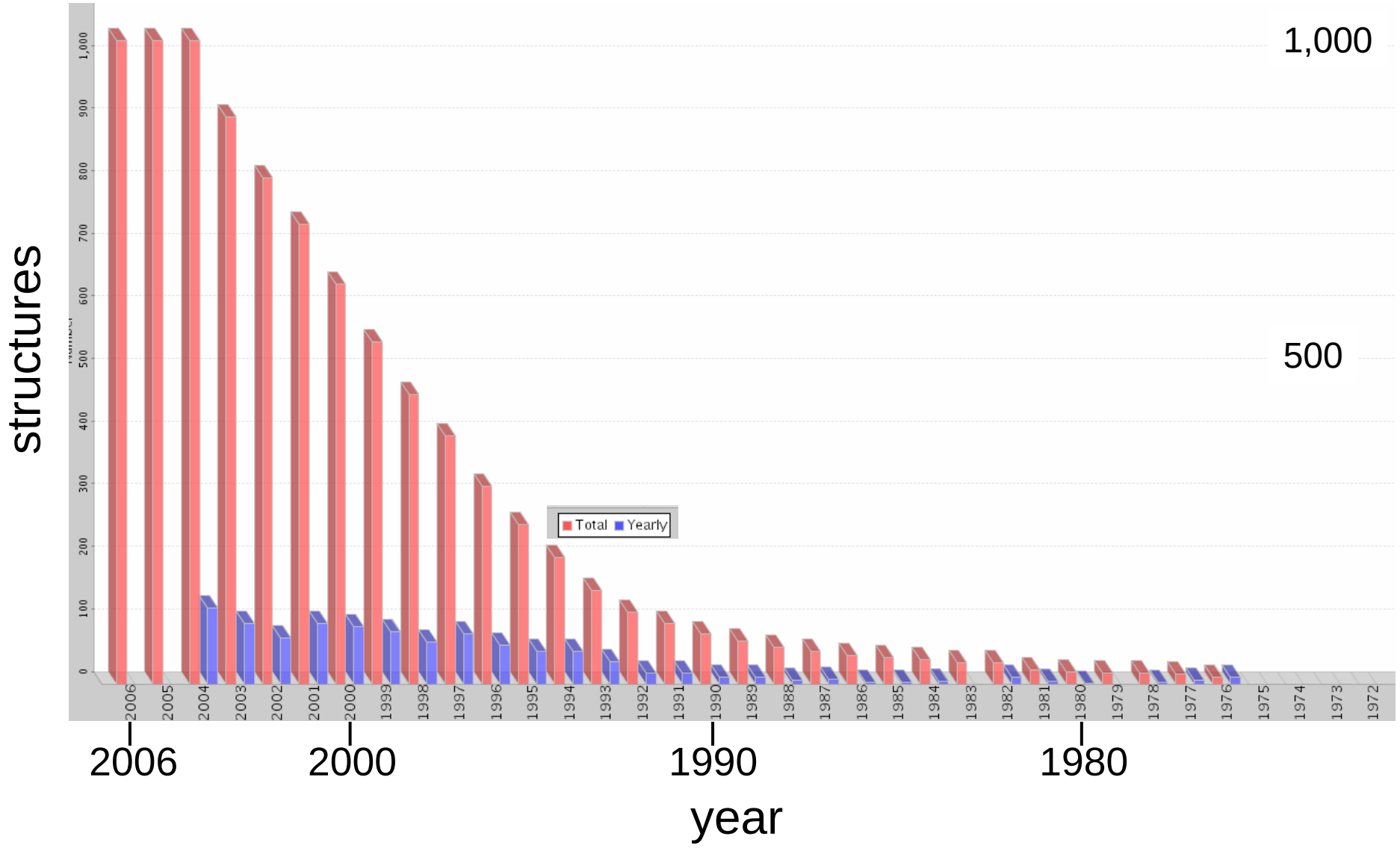


http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1rbp

Fold Example: Lamda repressor DNA Binding

http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=5cro;pc=0

Number of unique folds (defined by SCOP) in PDB

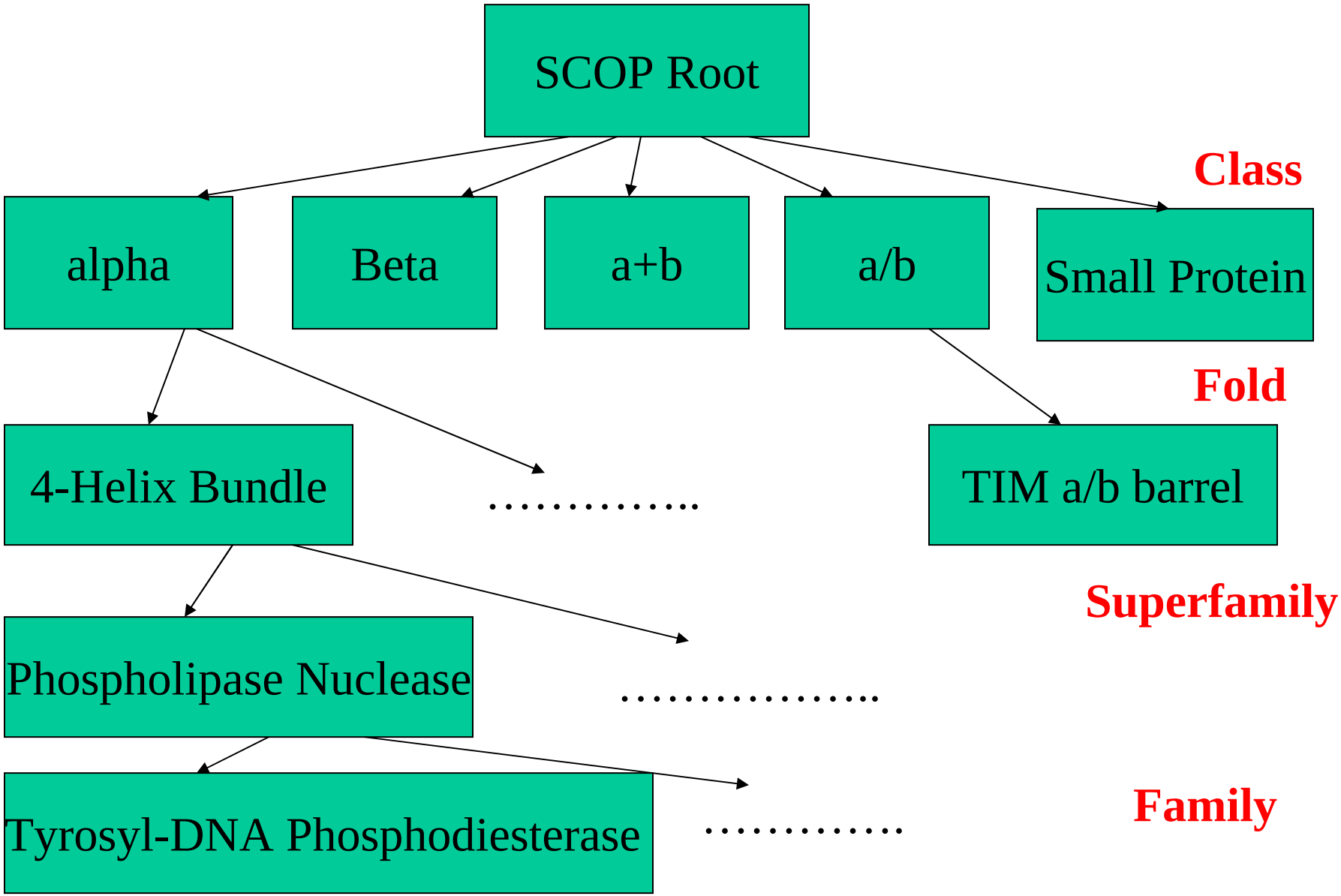


J. Pevsner, 2005

Structure Classification Database

- SCOP (http://scop.berkeley.edu/)
- CATH (http://cathwww.biochem.ucl.ac.uk/latest/index.html)
- Dali/FSSP (http://ekhidna.biocenter.helsinki.fi/dali/start)

SCOP Classification Levels



Structural Classification of Proteins

Root: scop

Classes:

1. All alpha proteins (138)
2. All beta proteins (93)
3. Alpha and beta proteins (a/b) (97)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) (184)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) (28)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides (11)
   Does not include proteins in the immune system
7. Small proteins (54)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins (5)
9. Low resolution protein structures (12)
10. Peptides (77)
    Peptides and fragments
11. Designed proteins (24)
    Experimental structures of proteins with essentially non-natural sequences

Enter search key: Search

Generated from scop database 1.55 with scopm 1.095 on Mon Jul 9 18:35:50 2001

J. Pevsner 2005

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