

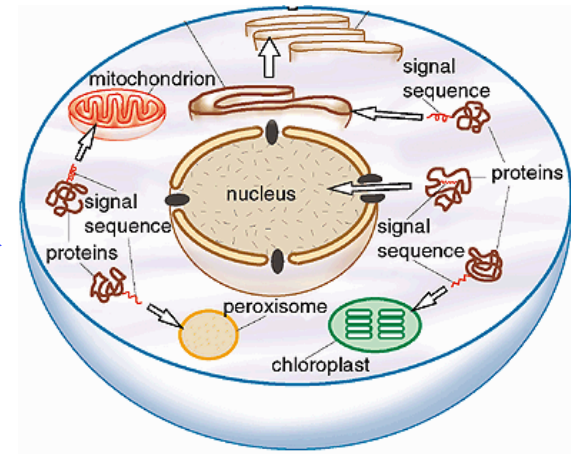
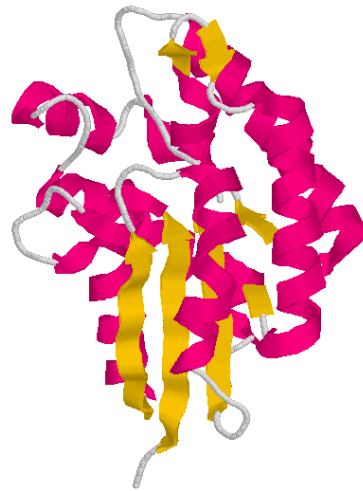
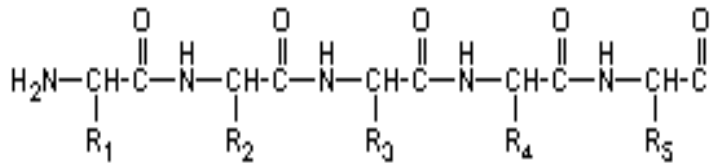
# Template Based Protein Structure Modeling

**Jianlin Cheng, PhD**

Associate Professor  
Computer Science Department  
Informatics Institute  
University of Missouri, Columbia  
2016

# Sequence, Structure and Function

AGCWY.....

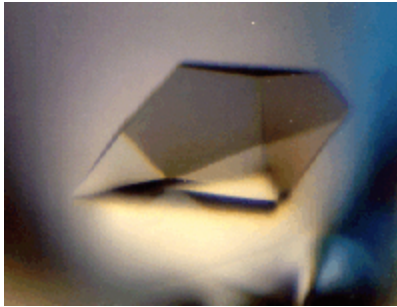


**Cell**

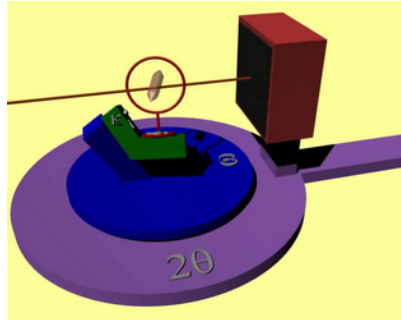
# Protein Structure Determination

- 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

# X-Ray Crystallography



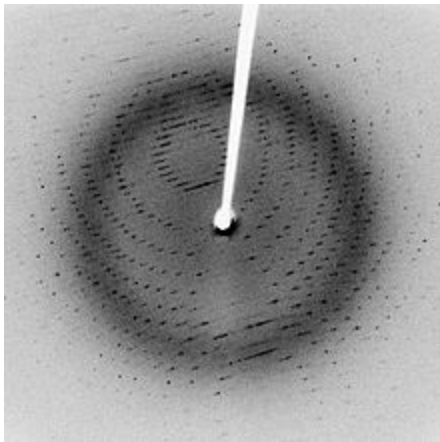
A protein crystal



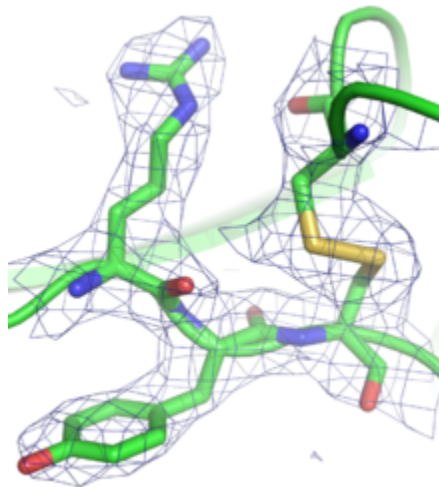
Mount a crystal



Diffractometer



Diffraction



Protein structure



[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

RCSB PDB An Information Portal to 115306 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands

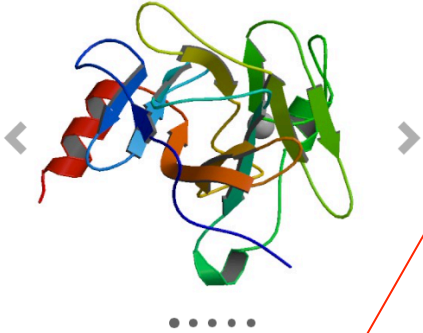
Advanced Search | Browse by Annotations | Search History (1) | Previous Results (12578)

PDB-101 | WORLDWIDE PDB | EMDatabank | NUCLEIC ACID DATABASE | StructuralBiology Knowledgebase

f t y a i s

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

Biological Assembly 1 ?



**4KVP**  
Human p53 Core Domain Mutant V157F  
DOI: 10.2210/pdb4kvp/pdb Entry 4KVP **supersedes** 2QXA  
Classification: **APOPTOSIS**  
Deposited: 2013-05-22 Released: 2013-07-31  
Deposition author(s): [Wallentine, B.D.](#), [Wang, Y.](#), [Luecke, H.](#)  
Organism: [Homo sapiens](#)  
Expression System: Escherichia coli  
Mutation(s): 1  
Structural Biology Knowledgebase: 4KVP (2 models >14 annotations) [SBKB.org](#)

Display Files Download Files

**Experimental Data Snapshot**  
Method: X-RAY DIFFRACTION  
Resolution: 1.5 Å  
R-Value Free: 0.210  
R-Value Work: 0.174

**wwPDB Validation**   
Metric Percentile Ranks Value  
Rfree 0.228  
Clashscore 5  
Ramachandran outliers 0  
Sidechain outliers 1.3%  
RSRZ outliers 1.0%  
Worse Better  
■ Percentile relative to all X-ray structures  
□ Percentile relative to X-ray structures of similar resolution

View in 3D: JSmol or PV (in Browser)  
Standalone Viewers  
Simple Viewer Protein Workshop  
Ligand Explorer Kiosk Viewer  
Protein Symmetry: Asymmetric (View in 3D)  
Protein Stoichiometry: Monomer

Literature

Search database

## **Search Demo: Human P53 protein – 1KVP**

<http://www.rcsb.org/pdb/explore/explore.do?structureId=4KVP>



# 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
```





# Structure Visualization

- Rasmol (<http://www.umass.edu/microbio/rasmol/getras.htm>)
- MDL Chime (plug-in) (<http://www.mdl.com/products/framework/chime/>)
- **Jmol:** <http://jmol.sourceforge.net/>
- **JSMol: java script version**
- **Pymol:** <http://pymol.sourceforge.net/>
- **Chimera:** <https://www.cgl.ucsf.edu/chimera/>

# JSMol (4KVP, Human P53)

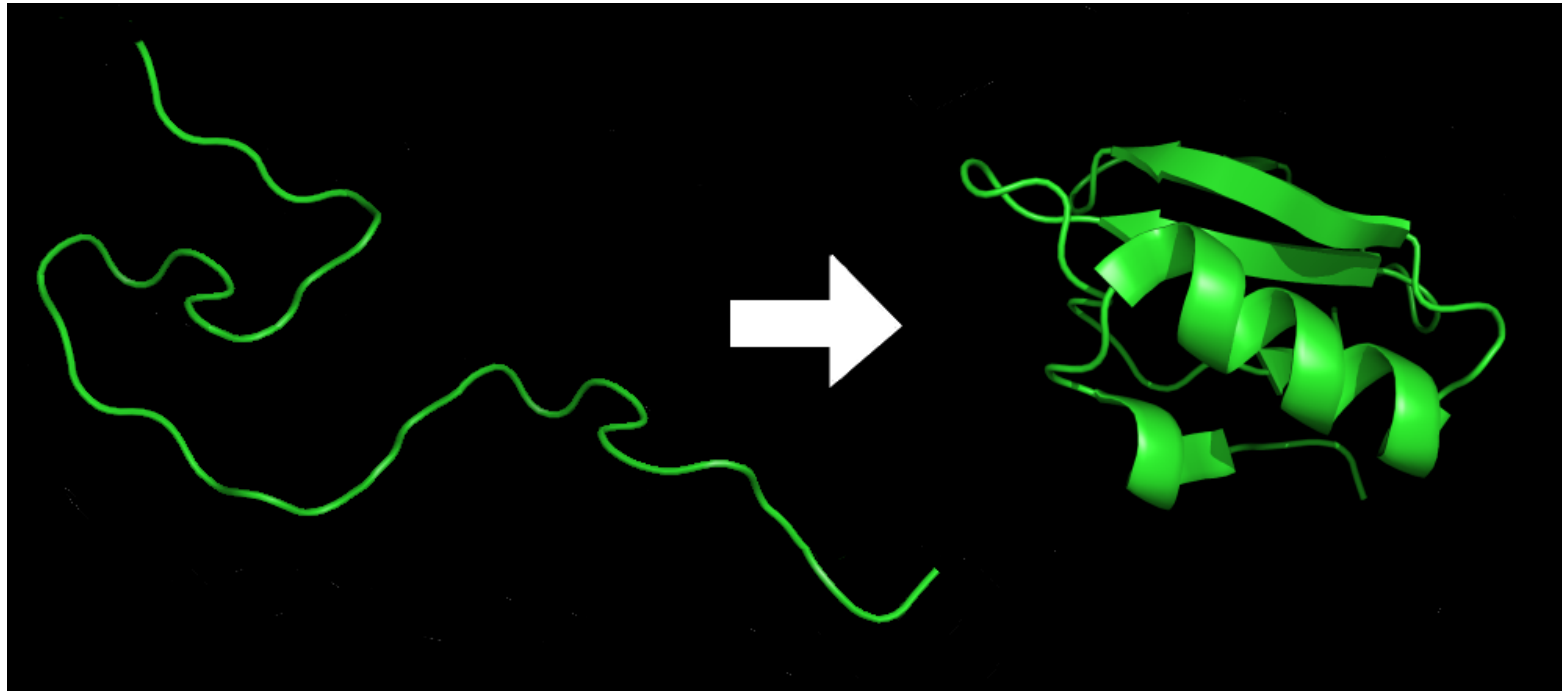
- JSMol:

<http://www.rcsb.org/pdb/explore/jmol.do?structureId=4KVP&bionumber=1>

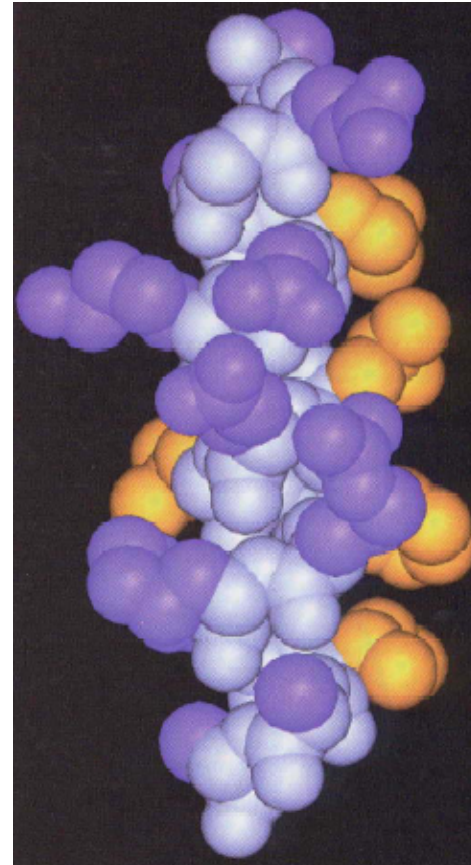
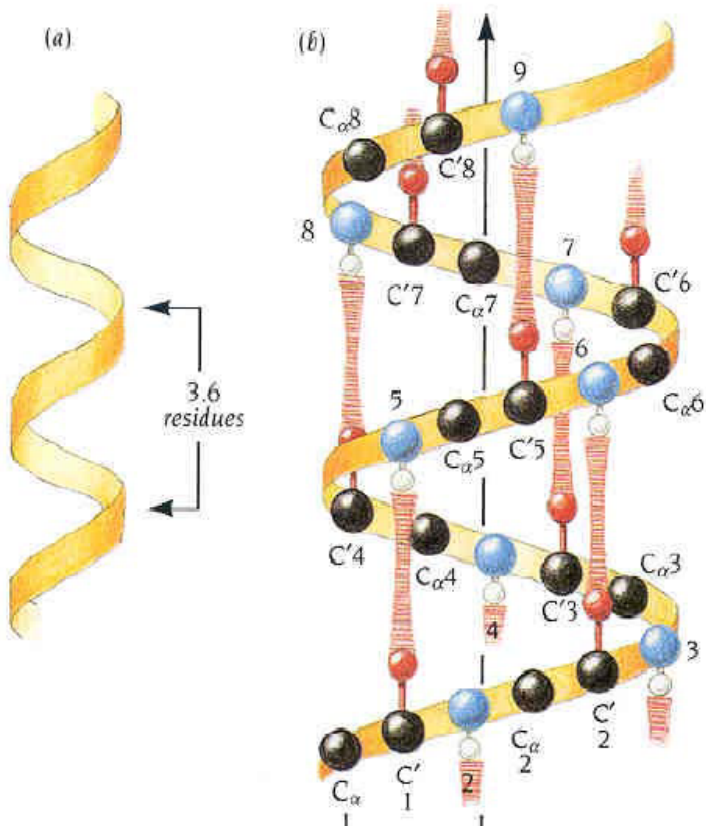
- JMOL: 1VJP
- Identify residues
- Recognize atoms
- Recognize peptide bonds
- Identify backbone
- Identify side chain
- Analyze different visualization style

# Protein Folding

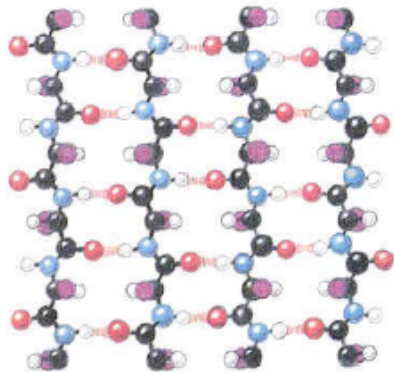
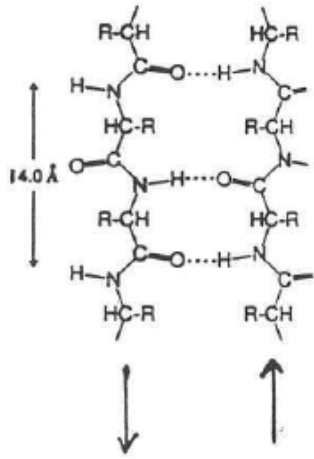
<http://www.youtube.com/watch?v=fvBO3TqJ6FE&feature=fvw>



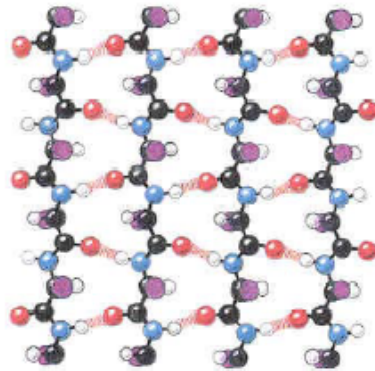
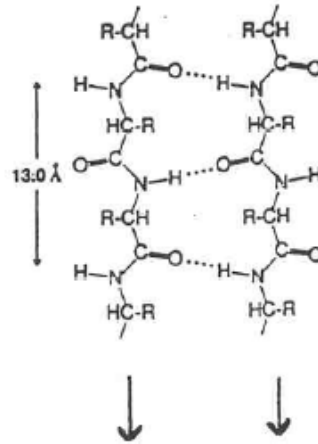
# Alpha-Helix



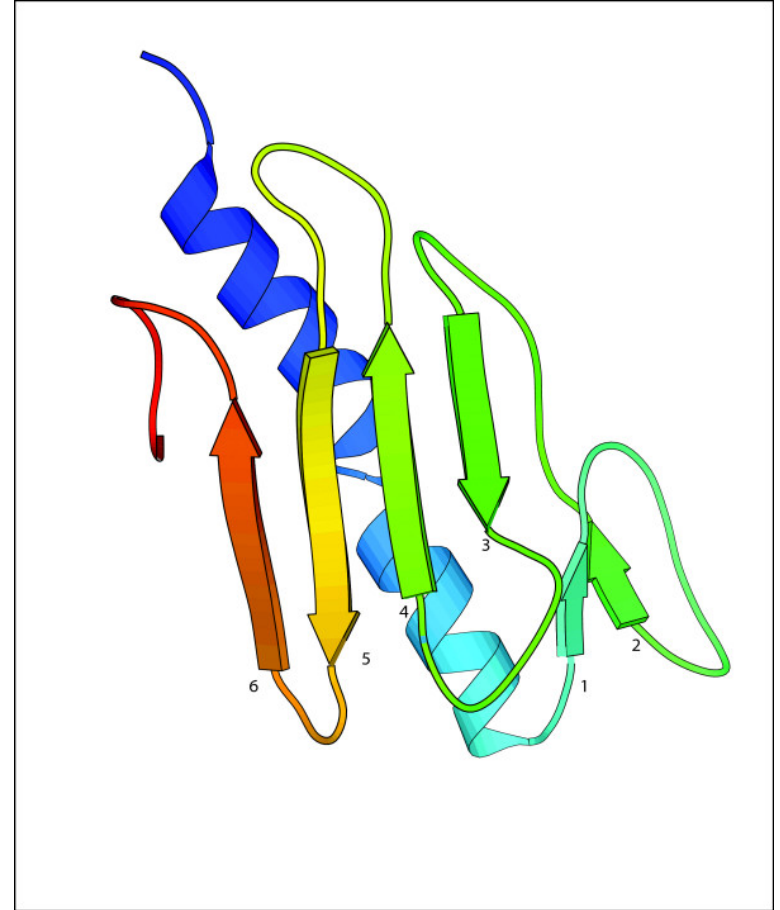
# Beta-Sheet



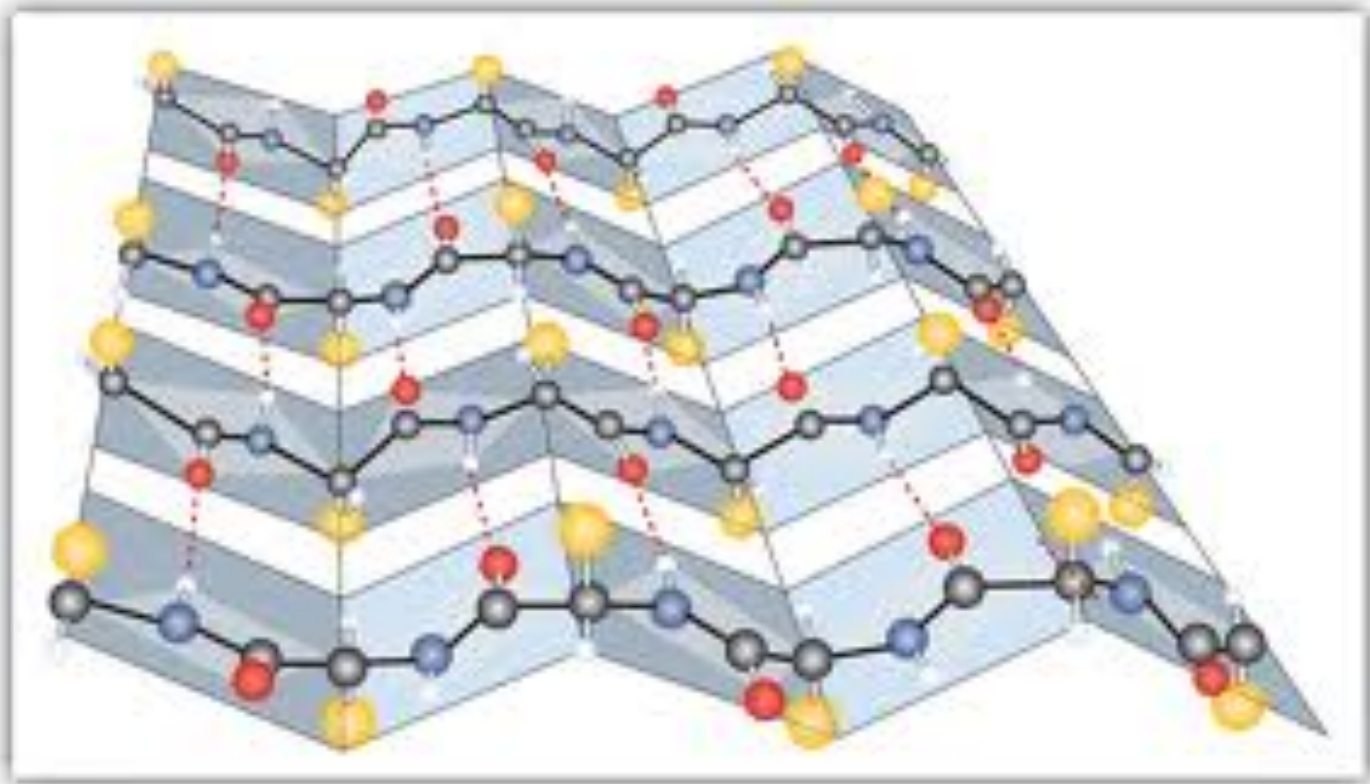
Anti-Parallel



Parallel

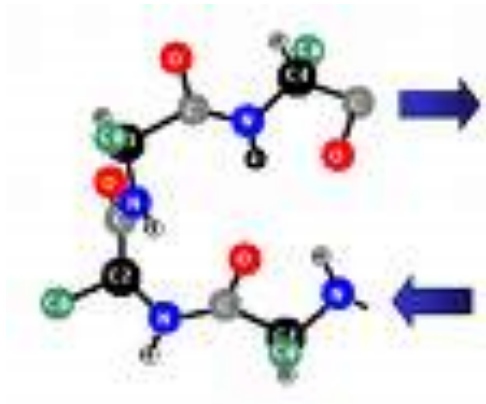


# Beta-Sheet

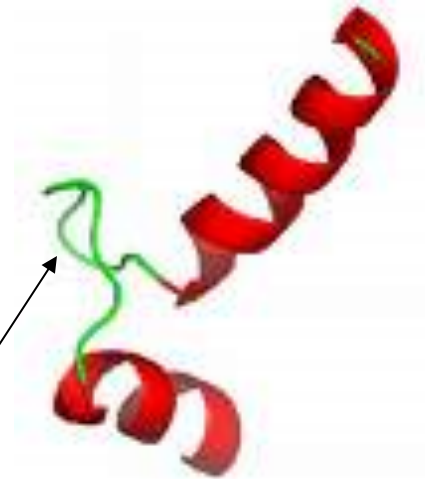




# Non-Repetitive Secondary Structure



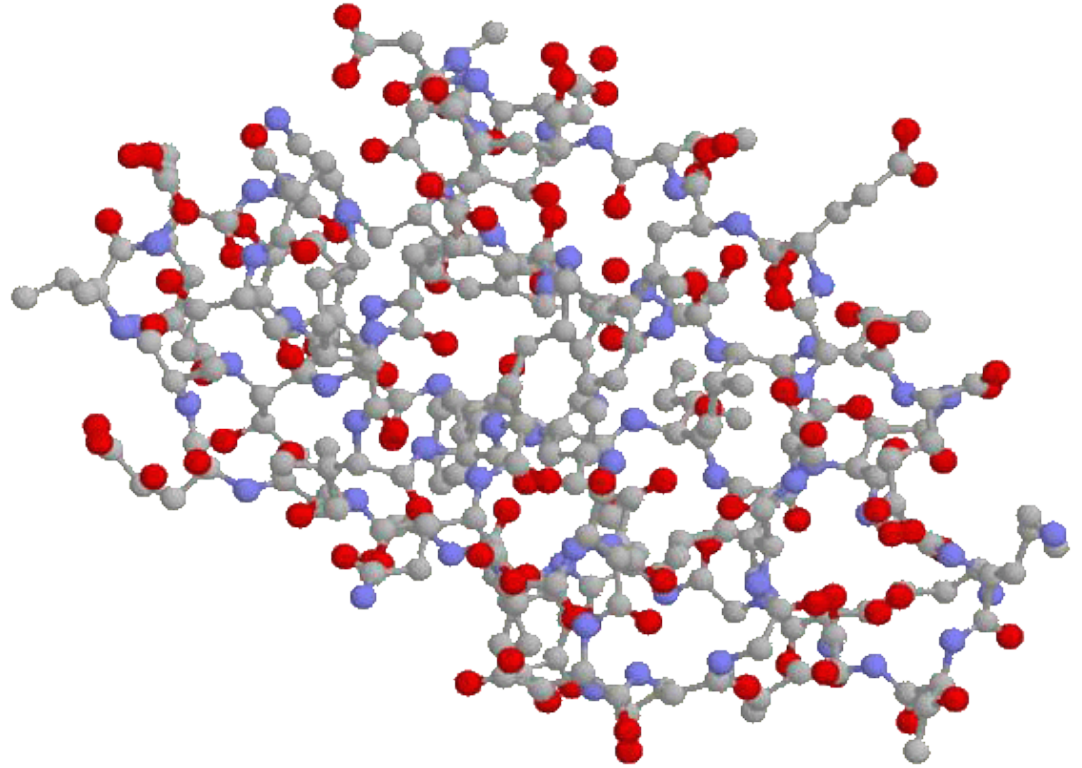
Beta-Turn



Loop



myoglobin



tertiary structure  
(all atom)

# Quaternary Structure: Complex



G-Protein Complex

# 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, 1983**)

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

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

**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)**

your e-mail

PDB File

or you can instead enter a PDB id.

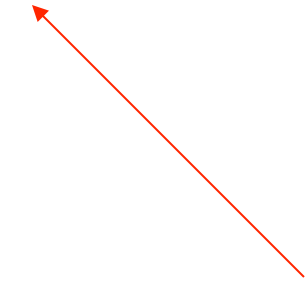
**<http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html>**

#	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			

Amino  
Acids

Secondary  
Structure

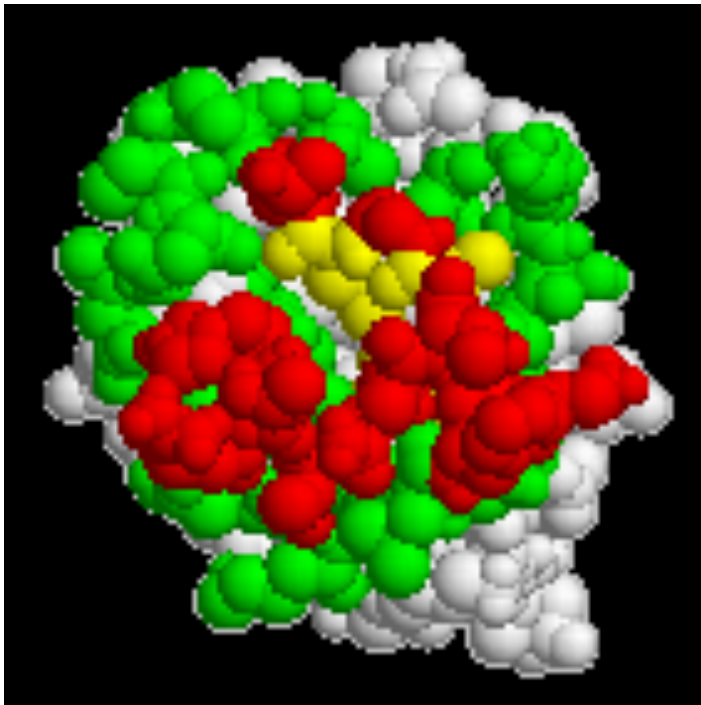
Solvent  
Accessibility





# Solvent Accessibility

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

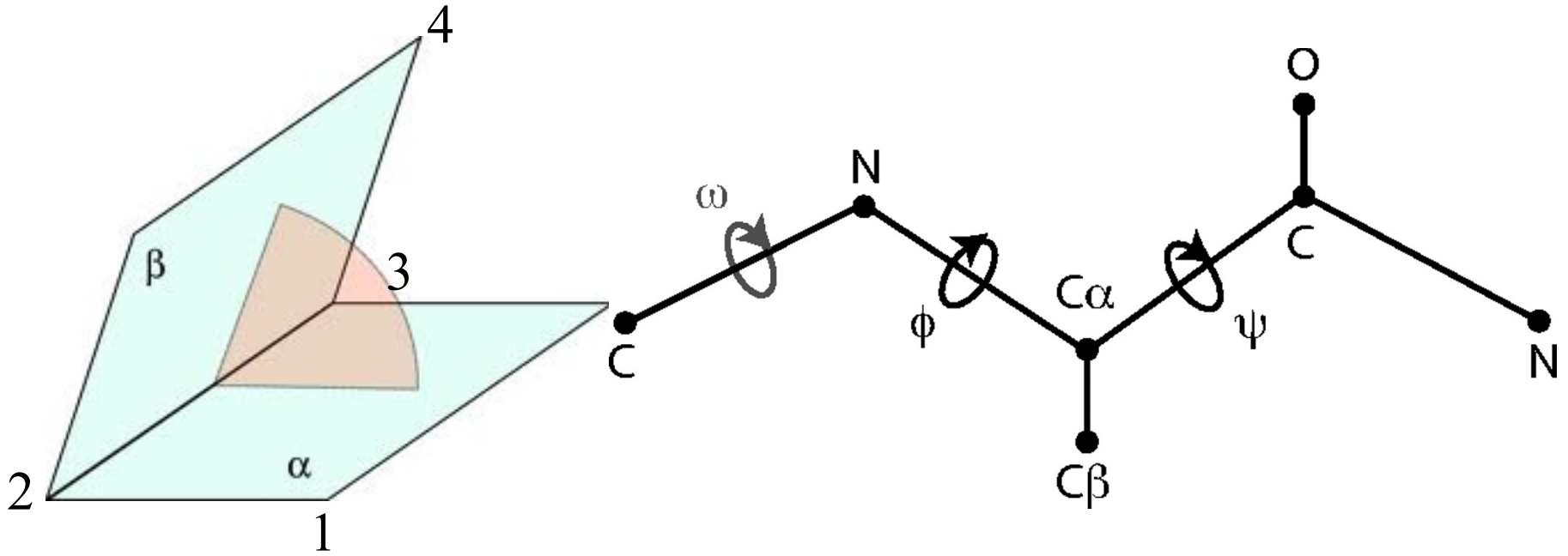


Maximum solvent accessible area for each amino acid is its whole surface area.

Hydrophobic residues like to be Buried inside (interior).

Hydrophilic residues like to be exposed on the surface.

# Dihedral Angle



# Project Groups

- 20 students
- Form 4 groups

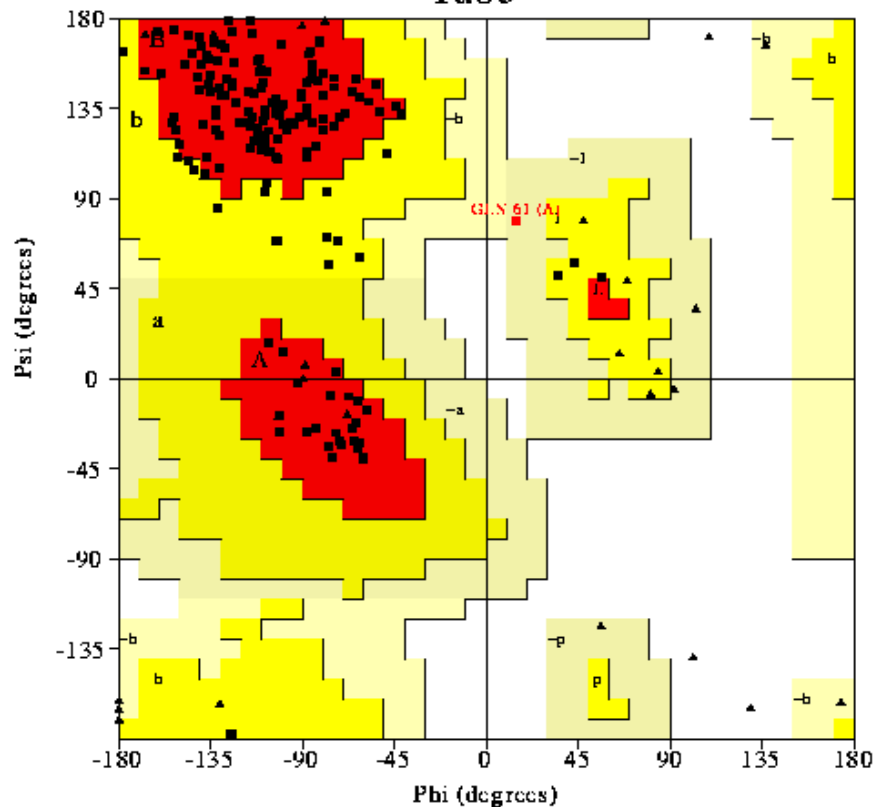
# Dihedral / Torsion Angle

$\phi$  (phi, involving the backbone atoms C'-N-C $\alpha$ -C'),  $\psi$  (psi, involving the backbone atoms N-C $\alpha$ -C'-N)

- [http://en.wikipedia.org/wiki/Dihedral\\_angle](http://en.wikipedia.org/wiki/Dihedral_angle)

# Ramachandran Plot

## 1abc

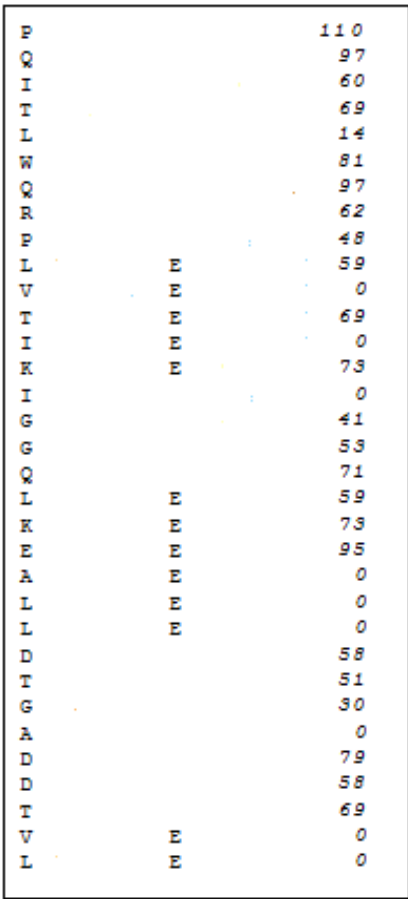


### Plot statistics

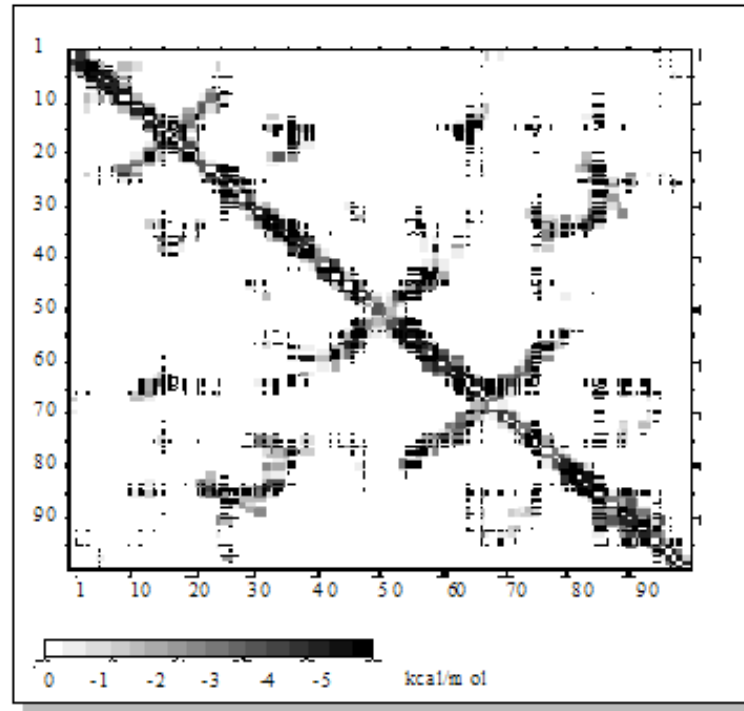
Residues in most favoured regions [A,B,L]	141	69.9%
Residues in additional allowed regions [a,b,l,p]	15	9.4%
Residues in generously allowed regions [-a,-b,-l,-p]	1	0.6%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	159	100.0%
Number of end residues (excl. Gly and Pro)	5	
Number of glycine residues (shown as triangles)	26	
Number of proline residues	15	
Total number of residues	205	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

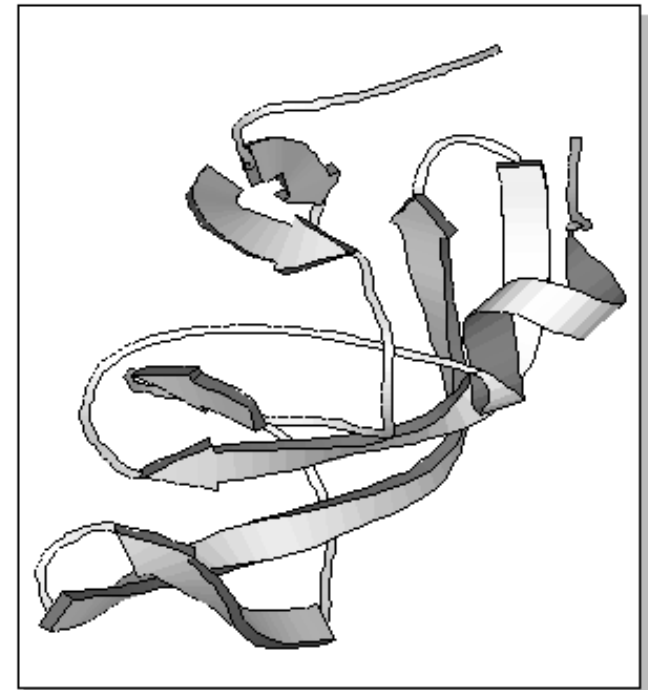
# Protein Structure 1D, 2D, 3D



1D



2D



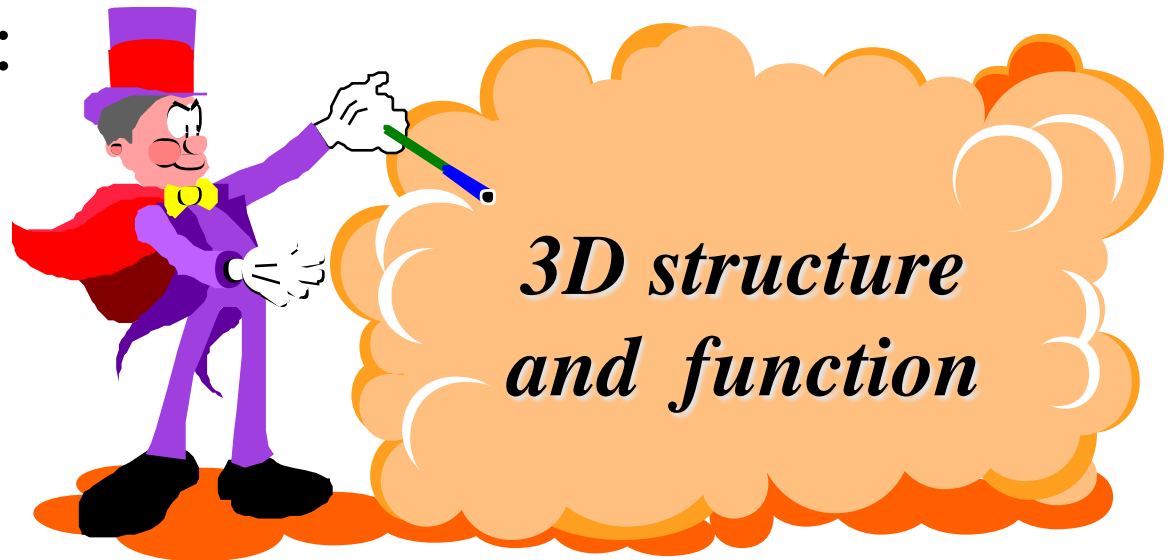
3D

# Goal of Structure Prediction

- Epstein & Anfinsen, 1961:  
sequence uniquely determines structure

- INPUT: sequence

- OUTPUT:





# CASP – Olympics of Protein Structure Prediction

- Critical Assessment of Techniques of Protein Structure Prediction
- 1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010, **2012, 2014**
- Blind Test, Independent Evaluation

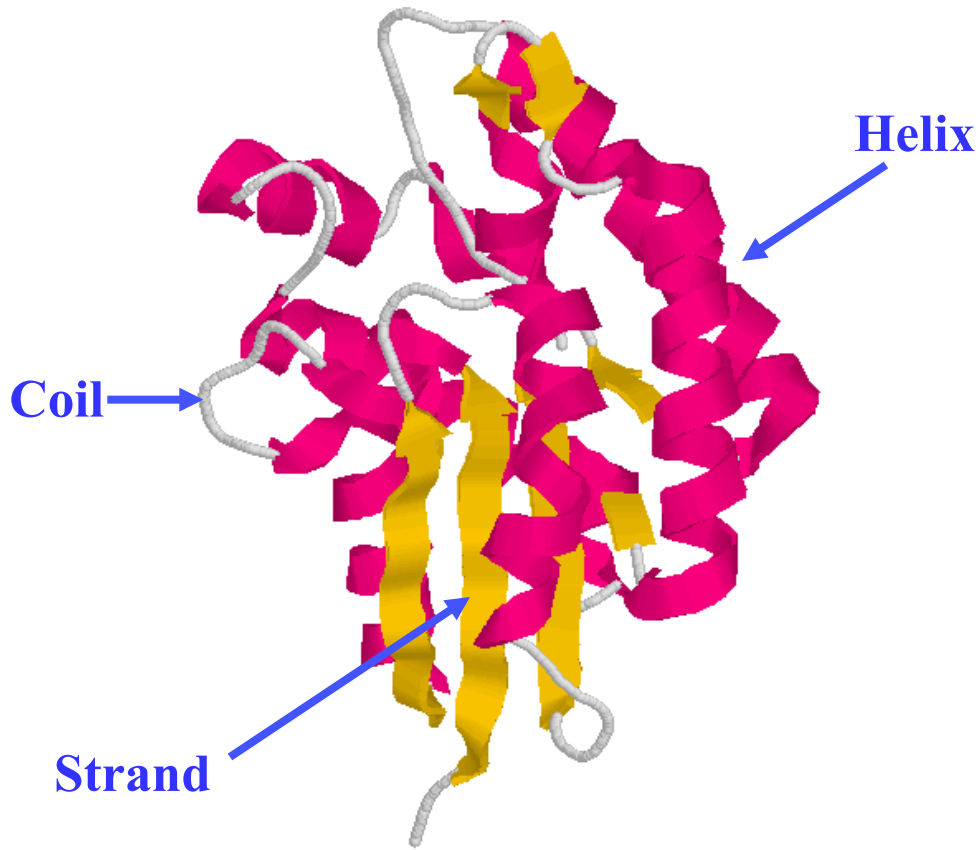


- **CASP11** (<http://www.predictioncenter.org/casp11/index.cgi>)

# CASP11 Demo

- <http://www.predictioncenter.org/casp11/index.cgi>

# 1D: Secondary Structure Prediction



MWLKKFGINLLIGQSV...

Neural Networks  
+ Alignments

CCCCHHHHCCCCSSSS...

# Widely Used Tools (~78-80%)

**Sspro 5:** <http://download.igb.uci.edu>

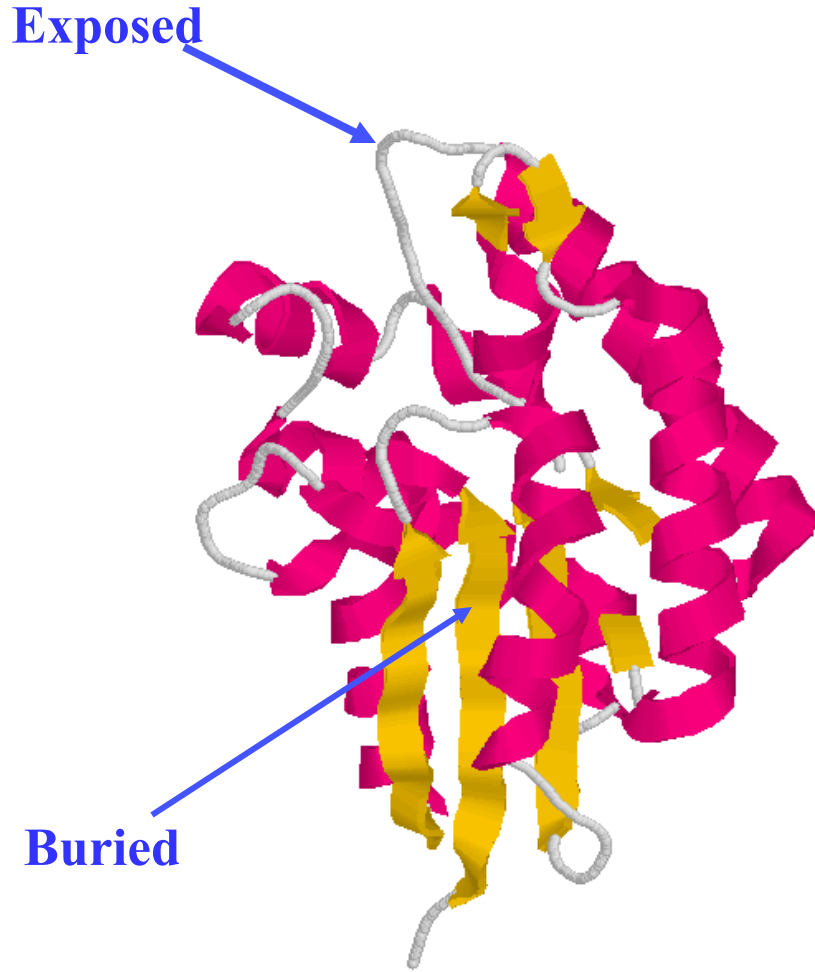
**Distill:** <http://distill.ucd.ie/porter/>

**PSI-PRED:** <http://bioinf.cs.ucl.ac.uk/psipred/psiform.html>  
software is also available

**SAM:** [http://compbio.soe.ucsc.edu/SAM\\_T08/T08-query.html](http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html)

**PHD:** <http://www.predictprotein.org/>

# 1D: Solvent Accessibility Prediction



MWLKKFGINLLIGQSV...

Neural Networks  
+ Alignments

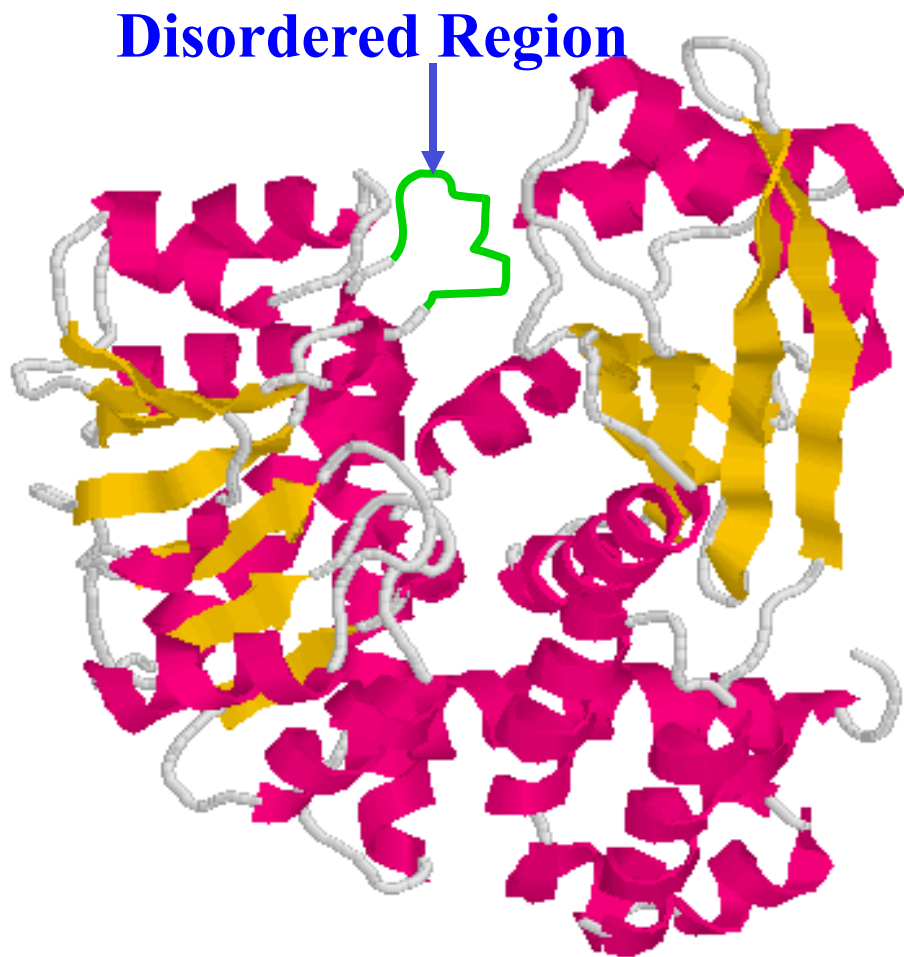
eeeeeeebbbbbbbbeeebbb...

Accuracy: 79% at 25% threshold

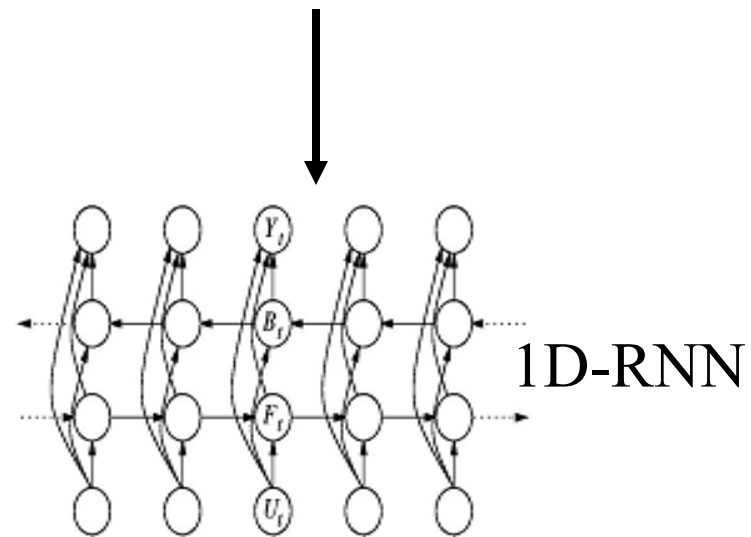
# Widely Used Tools (78%)

- ACCpro 5: <http://download.igb.uci.edu>
- SCRATCH: <http://scratch.proteomics.ics.uci.edu/>
- PHD: <http://www.predictprotein.org/>
- Distill: <http://distill.ucd.ie/porter/>

# 1D: Disordered Region Prediction Using Neural Networks



MWLKKFGINLLIGQSV...



OOOOODDDDOOOOO...

93% TP at 5% FP



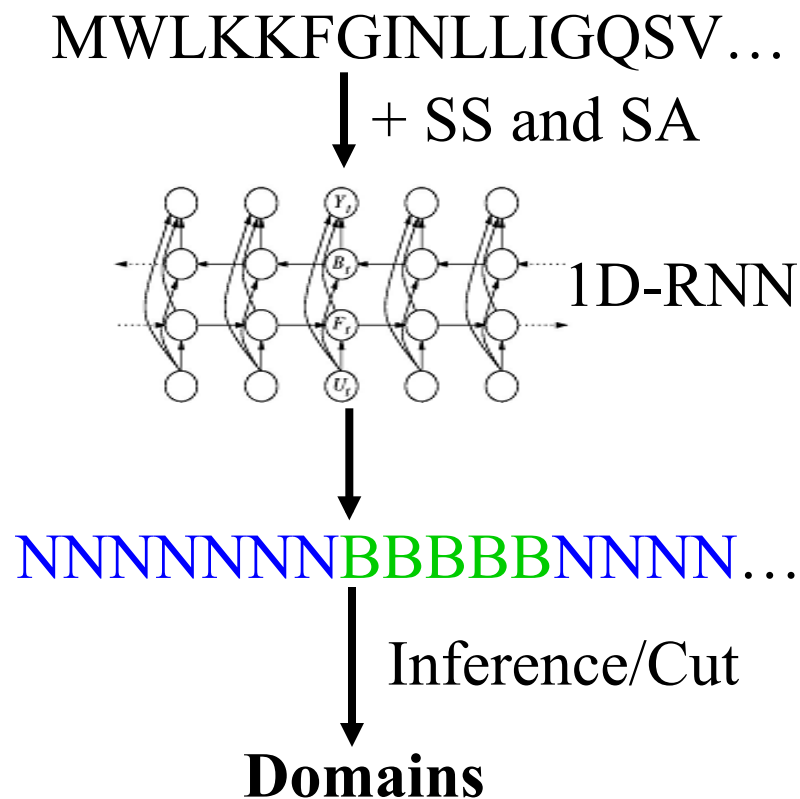
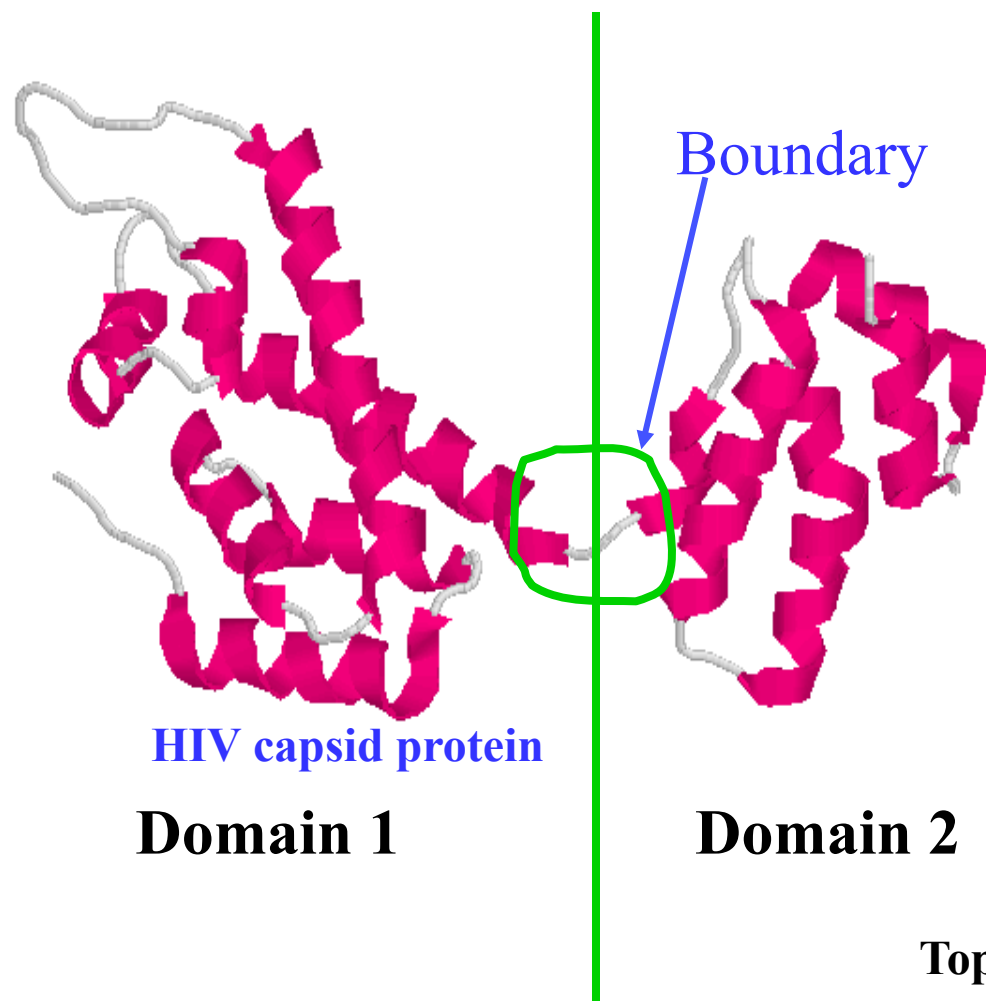
# Tools

**PreDisorder:** [http://sysbio.rnet.missouri.edu/multicom\\_toolbox/](http://sysbio.rnet.missouri.edu/multicom_toolbox/)

**A collection of disorder predictors:**

<http://www.disprot.org/predictors.php>

# 1D: Protein Domain Prediction Using Neural Networks



Top *ab-initio* domain predictor in CAFASP4

# DoBo

## Protein domain boundary prediction by integrating evolutionary signals and machine learning


Have a question? Maybe it's answered in the [FAQ](#)

**Job Details**


**Job title (optional)**

**Sequence**

Plain sequence. Spaces, newlines and any FASTA header will be ignored.  
Minimum sequence length is 90 residues.

**Confidence level**  

Set a minimum threshold for the confidence of domain boundary predictions.

**Single/multi-domain classification**  

Run an additional check to classify query as a single or multi-domain protein.

Web: [http://sysbio.rnet.missouri.edu/multicom\\_toolbox/index.html](http://sysbio.rnet.missouri.edu/multicom_toolbox/index.html)

### Reference:

J. Eickholt, X. Deng, and J. Cheng. **DoBo: Protein Domain Boundary Prediction by Integrating Evolutionary Signals and Machine Learning.** *BMC Bioinformatics*. 12:43, 2011.

### 1. Input query

LNKGQRHIKIREIIMS...

### 2. Identify homologous sequences w/ PSI-BLAST



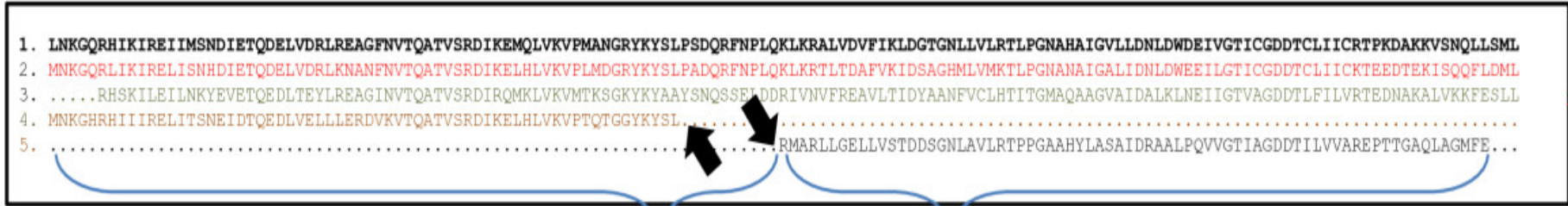
### 3. Extract pairwise alignments

```

Query 1 LNKGQRHIKIREIIMSNDIETQDELVDRLREAGFNVTQATVSRDIKEMQLVKVPMANGRY 60
Sbjct 1 MNKGQRHIKIREIIANKEIETQDELVDILRNEGFNVTQATVSRDIKELHLVKVPLHDGRY 60
...
Query 6 RHIKIREIIMSNDIETQDELVDRLREAGFNVTQATVSRDIKEMQLVKVPMANGRYKYSLP 65
Sbjct 5 RHSKILEILNKYEVEVETQEDLTEYLREAGINVTQATVSRDIRQMKLVKVMTKSGKYKYAAY 64
...
Query 1 LNKGQRHIKIREIIMSNDIETQDELVDRLREAGFNVTQATVSRDIKEMQLVKVPMANGRY 60
Sbjct 1 MNKGQRHIKIREIIANKEIETQDELVDILRNEGFNVTQATVSRDIKELHLVKVPLHDGRY 60

```

### 4. Form multiple sequence alignment



### 5. Identify domain boundary signals

Gap 45 residues or longer



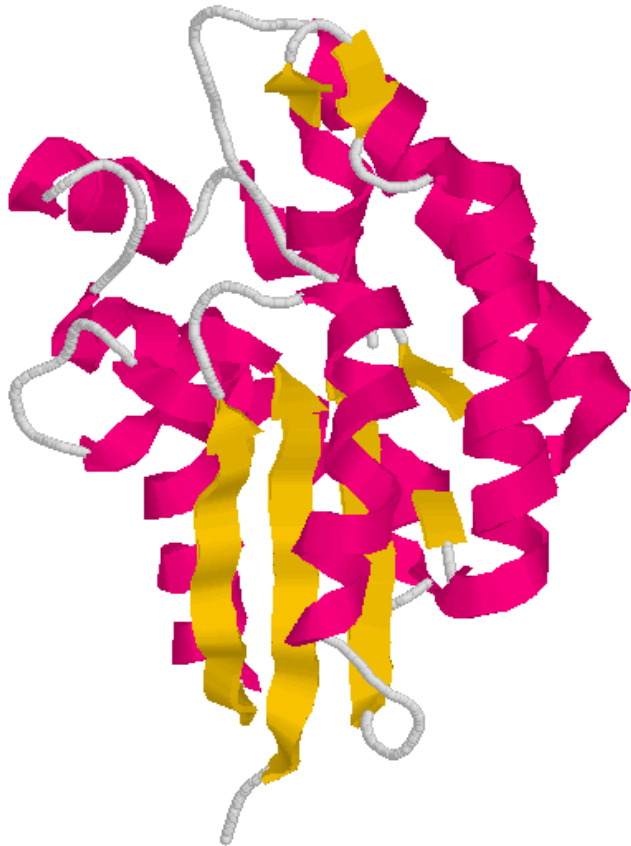
Remaining sequence longer than 45 residues



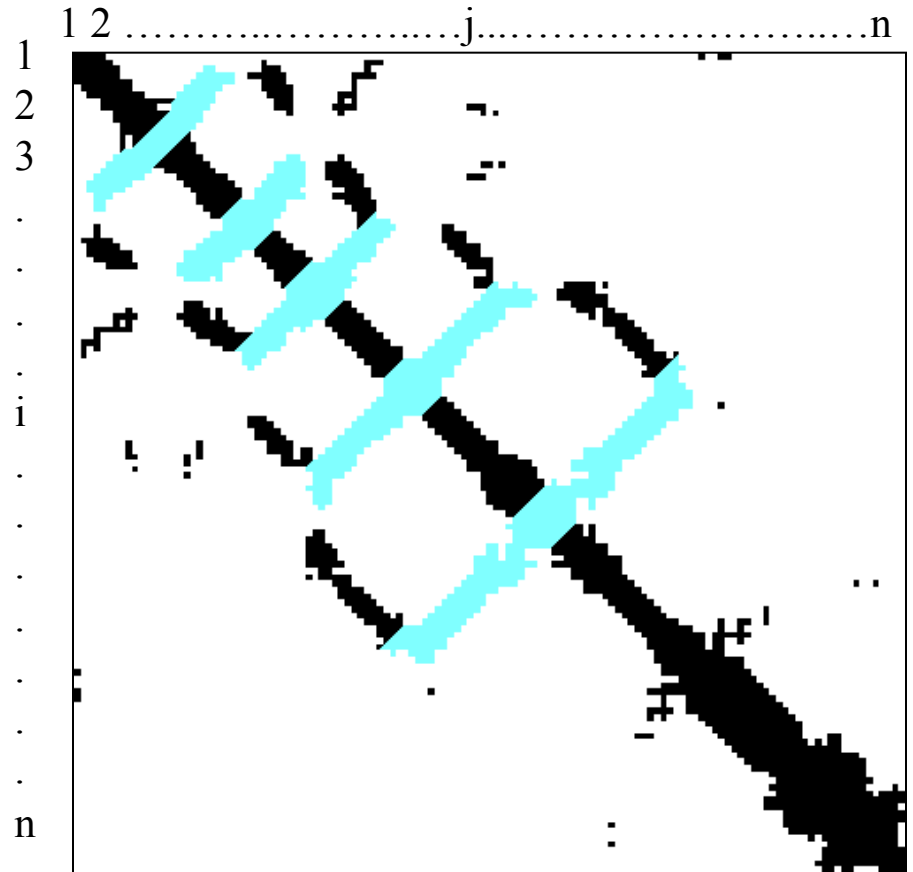
Domain boundary signal (indicated by large arrows)

# 2D: Contact Map Prediction

3D Structure



2D Contact Map



Distance Threshold = 8Å

Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005

# Contact Prediction

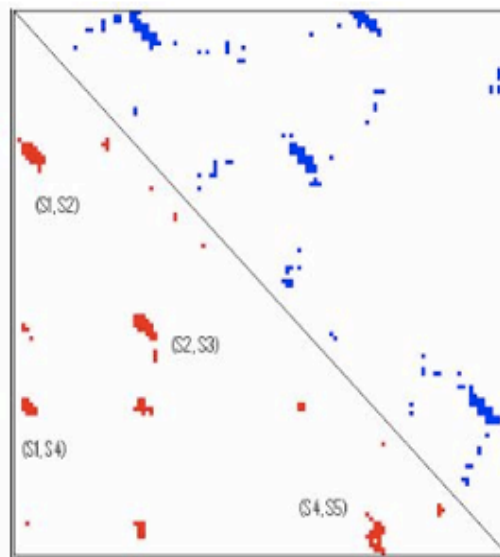
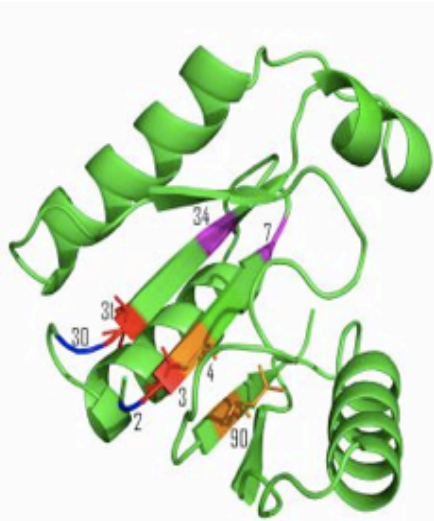
- SVMcon/NNcon/DNcon:

<http://casp.rnet.missouri.edu/svmcon.html>

- PISCOV:

<http://bioinfadmin.cs.ucl.ac.uk/downloads/>

[PISCOV/](#)



## NNcon: Protein Contact Map Prediction Using Artificial Neural Networks ([Help](#))

Email address(where the prediction will be sent):

Target Name(required):

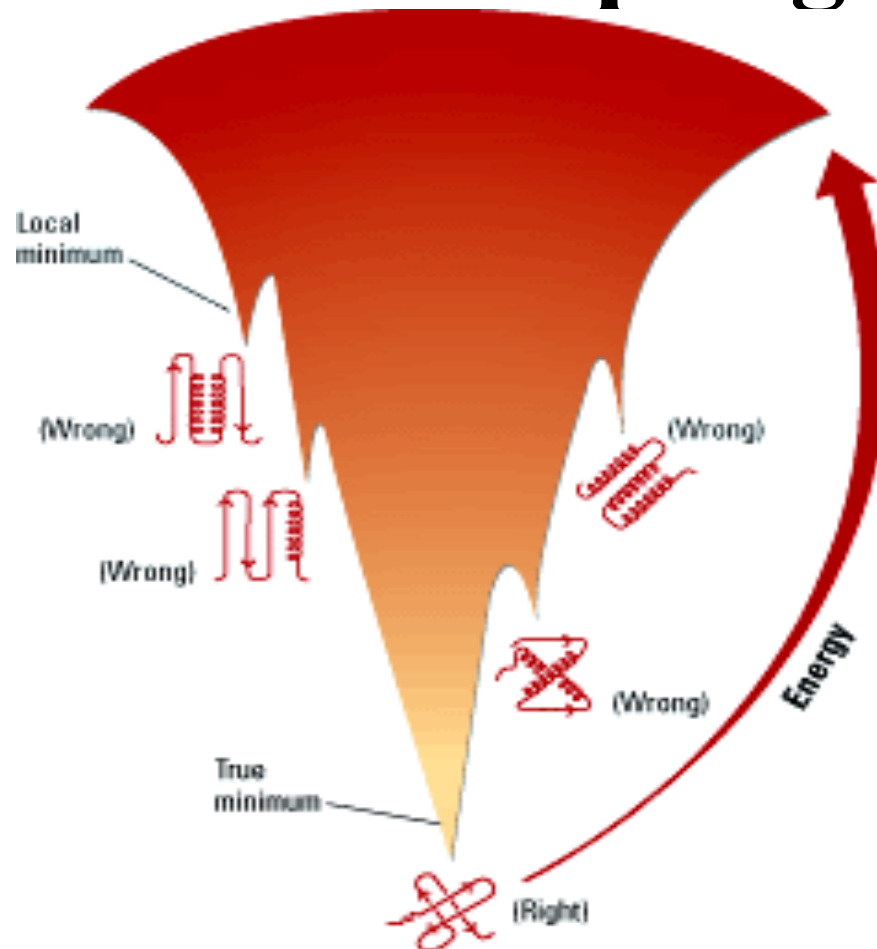
Protein sequence(one plain sequence, no headers, and length < 1000 amino acids; an example sequence is [here](#)):

Predict

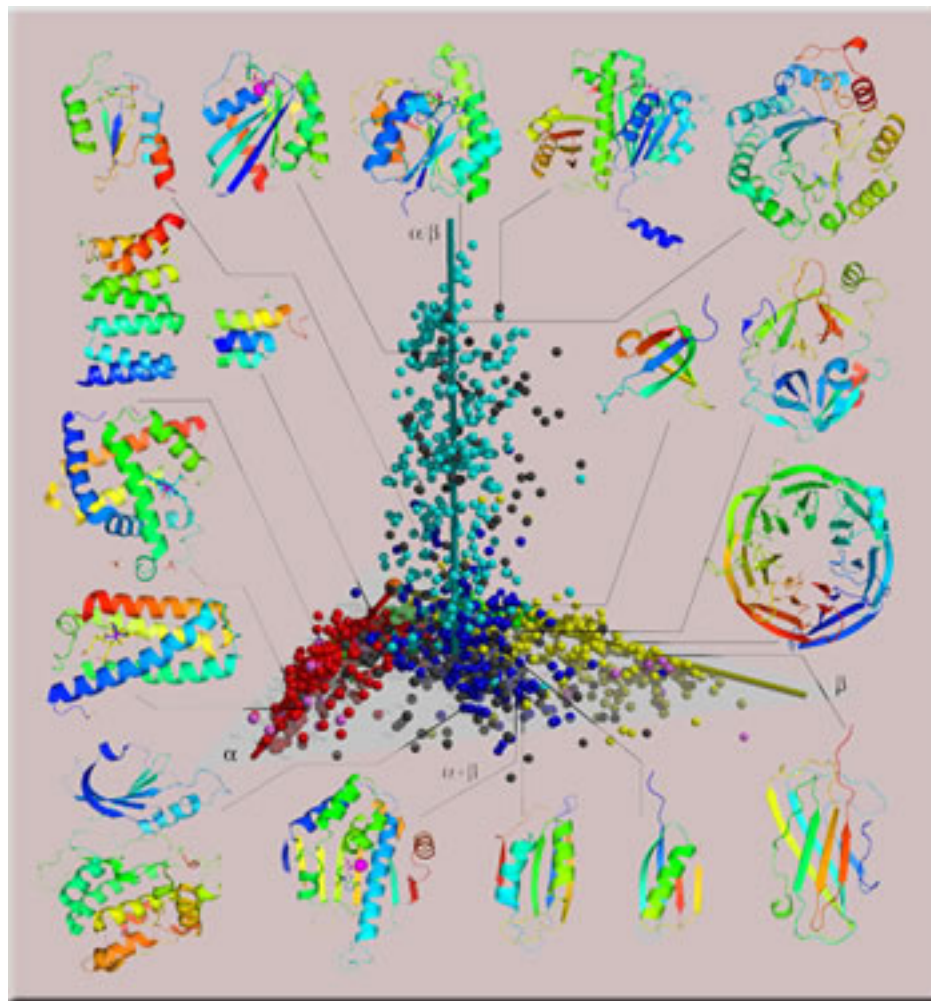
**Protein tertiary structure prediction is a space sampling problem.**



# Protein Energy Landscape & Free Sampling



# Protein Structure Space & Target Sampling

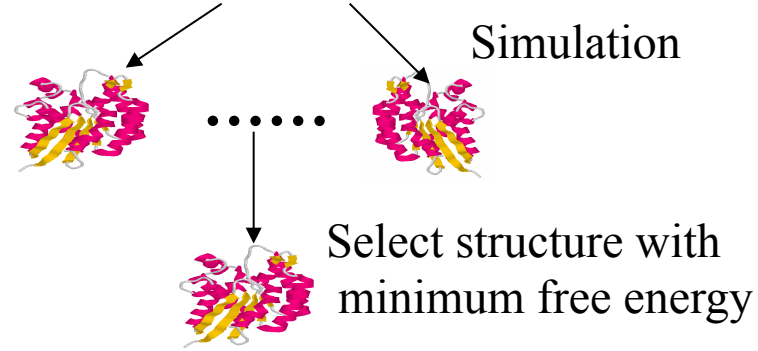


# Two Approaches for 3D Structure Prediction

## • Ab Initio Structure Prediction

Physical force field – protein folding  
Contact map - reconstruction

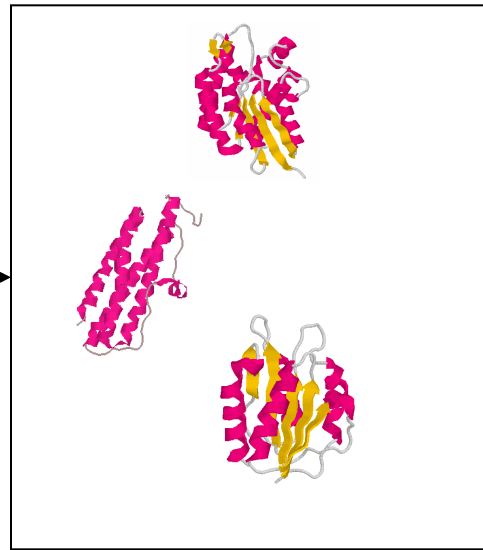
MWLKKFGINLLIGQSV...



## • Template-Based Structure Prediction

Query protein

MWLKKFGINKH...



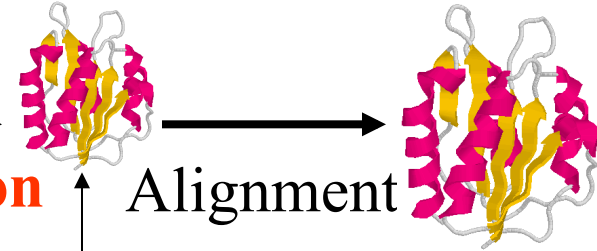
Protein Data Bank

**Fold**

**Recognition**

Alignment

Template



# Template-Based Structure Prediction

1. Template identification
2. Query-template alignment
3. Model generation
4. Model evaluation
5. Model refinement

Notes: if template is easy to identify, it is often called **comparative Modeling or homology** modeling. If template is hard to identify, it is often called **fold recognition**.

# TARGET

ASILPKRLFGNCEQTSDEGLK  
IERTPLVPHISAQNVCLKIDD  
VPERLIPERASFQWMNDK

# TEMPLATE



ASILPKRLFGNCEQTSDEGLK IERTPLVPHISAQNVCLKIDD VPERLIPE  
MSVIPKRLYG NCEQTSEE AIRIEDSPIV --- TADLVCLKIDEIPERLVGE



Copy  
Loop Modeling  
Optimization

How to find templates?  
How to get alignments?

# Modeller

- Need an alignment file between query and template sequence in the PIR format
- Need the structure (atom coordinates) file of template protein
- You need to write a simple script (Python for version 8.2) to tell how to generate the model and where to find the alignment file and template structure file.
- Run Modeller on the script. Modeller will automatically copy coordinates and make necessary adjustments to generate a model.



# Structure File Example (1SDMA.atm)

ATOM	1	N	LYS	1	-3.978	26.298	113.043	1.00	31.75	N
ATOM	2	CA	LYS	1	-4.532	25.067	113.678	1.00	31.58	C
ATOM	3	C	LYS	1	-5.805	25.389	114.448	1.00	30.38	C
ATOM	4	O	LYS	1	-6.887	24.945	114.072	1.00	32.68	O
ATOM	5	CB	LYS	1	-3.507	24.446	114.631	1.00	34.97	C
ATOM	6	CG	LYS	1	-3.743	22.970	114.942	1.00	36.49	C
ATOM	7	CD	LYS	1	-3.886	22.172	113.644	1.00	39.52	C
ATOM	8	CE	LYS	1	-3.318	20.766	113.761	1.00	41.58	C
ATOM	9	NZ	LYS	1	-1.817	20.761	113.756	1.00	43.48	N
ATOM	10	N	ILE	2	-5.687	26.161	115.522	1.00	26.16	N
ATOM	11	CA	ILE	2	-6.867	26.500	116.302	1.00	22.75	C
ATOM	12	C	ILE	2	-7.887	27.226	115.439	1.00	21.35	C
ATOM	13	O	ILE	2	-7.565	28.200	114.770	1.00	20.95	O
ATOM	14	CB	ILE	2	-6.513	27.377	117.523	1.00	21.68	C
ATOM	15	CG1	ILE	2	-5.701	26.563	118.526	1.00	21.13	C
ATOM	16	CG2	ILE	2	-7.782	27.875	118.200	1.00	18.96	C
ATOM	17	CD1	ILE	2	-5.368	27.325	119.787	1.00	21.39	C
ATOM	18	N	ARG	3	-9.120	26.737	115.461	1.00	22.04	N
ATOM	19	CA	ARG	3	-10.214	27.327	114.693	1.00	23.95	C
ATOM	20	C	ARG	3	-10.783	28.563	115.400	1.00	22.82	C
ATOM	21	O	ARG	3	-10.771	28.645	116.629	1.00	22.62	O
ATOM	22	CB	ARG	3	-11.327	26.290	114.510	1.00	26.34	C
ATOM	23	CG	ARG	3	-11.351	25.586	113.161	1.00	30.68	C
ATOM	24	CD	ARG	3	-10.004	25.034	112.771	1.00	35.43	C
ATOM	25	NE	ARG	3	-10.104	24.072	111.672	1.00	43.37	N
ATOM	26	CZ	ARG	3	-10.575	24.350	110.458	1.00	46.04	C
ATOM	27	NH1	ARG	3	-10.997	25.572	110.168	1.00	48.68	N
ATOM	28	NH2	ARG	3	-10.627	23.400	109.532	1.00	48.37	N
ATOM	29	N	VAL	4	-11.278	29.524	114.630	1.00	20.49	N
ATOM	30	CA	VAL	4	-11.853	30.724	115.225	1.00	17.59	C
ATOM	31	C	VAL	4	-13.082	31.211	114.471	1.00	18.31	C
ATOM	32	O	VAL	4	-13.030	31.446	113.264	1.00	16.37	O
ATOM	33	CB	VAL	4	-10.834	31.872	115.272	1.00	19.94	C
ATOM	34	CG1	VAL	4	-11.512	33.168	115.759	1.00	15.64	C
ATOM	35	CG2	VAL	4	-9.668	31.489	116.168	1.00	15.45	C



# Modeller Python Script (bioinfo.py)

```
# Homology modelling by the automodel class
```

```
from modeller.automodel import * # Load the automodel class
```

```
log.verbose() # request verbose output
```

```
env = environ() # create a new MODELLER environment to build this model in
```

```
# directories for input atom files
```

```
env.io.atom_files_directory = './../atom_files'
```

```
a = automodel(env,
```

```
   alnfile = 'bioinfo.pir', # alignment filename
```

```
    knowns = '1SDMA', # codes of the templates
```

```
    sequence = 'bioinfo') # code of the target
```

```
a.starting_model= 1 # index of the first model
```

```
a.ending_model = 1 # index of the last model
```

```
    # (determines how many models to calculate)
```

```
a.make() # do the actual homology modelling
```

Where to find structure file

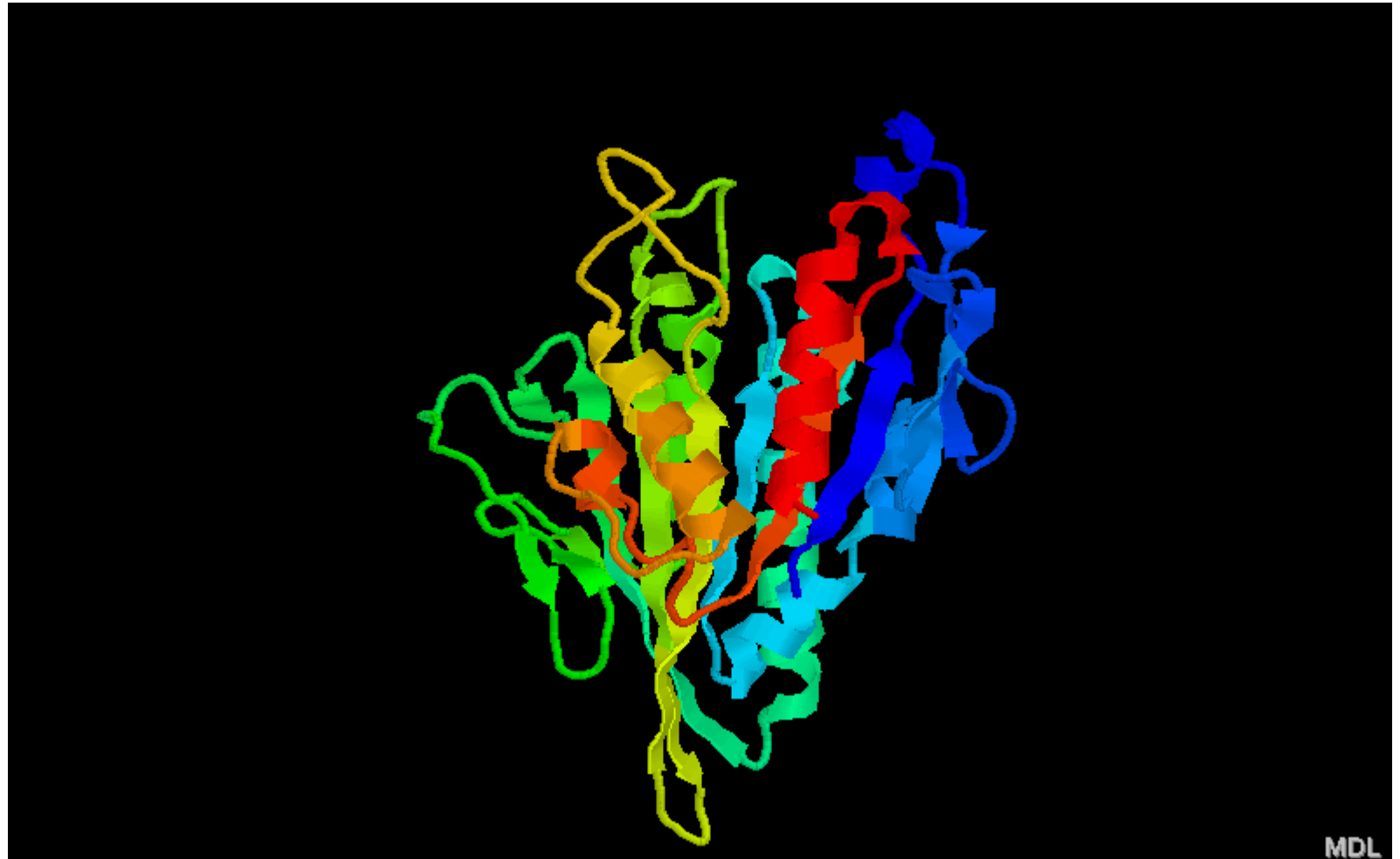
PIR alignment file name

Template structure file id

Query sequence id

# Output Example

Command: mod8v2 bioinfo.py



# Template Based Modeling Methods

- Comparative Protein Modeling by Satisfaction of Spatial Restraints by Andrej Sali and Tom L. Blundell
- 3D Model is obtained by satisfying spatial restraints derived from alignment with a known structure, which are expressed as probability density functions (pdfs) of the restraints.
- Pdfs serve as an objective function for optimization

# Probability Density Functions of Features

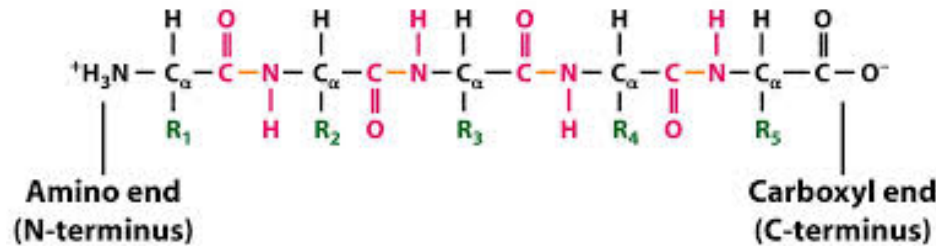


Figure 3-3b  
Molecular Cell Biology, Sixth Edition  
© 2008 W. H. Freeman and Company

- Ca – Ca distances
- Main-chain N-O distance
- Main-chain dihedral angles
- Side-chain dihedral angles
- A protein pdf is a combination of individual pdfs of features of the whole protein

# Optimization Procedure

- Objective: the pdf of a protein derived from restraints extracted from templates and alignments
- Initial input: initial (x, y, z) of each residue satisfying bond length / angle restraints
- Optimization: adjust x, y, z to maximize the pdf (i.e. probability), i.e. reduce the violations of feature restraint as much as possible

# Topic 1 – Template Based Modeling

- CASP11 TBM targets
- Known templates at CASP11 web sites
- Develop a homology-based algorithm / tool to build models from templates
- Assess the quality of models
- Implement from scratch
- **Form your group (by Wednesday)**

# Feature Restraints from Templates

- A database of 17 family alignments including 80 proteins was constructed to obtain feature statistics.
- Feature constraint is represented as conditional distribution. E.g.  $P(\text{ca-ca distance in target} \mid \text{ca-ca distance in template, ...})$ ,  $P(\text{psi angle of a residue in target} \mid \text{psi angle of an equivalent residue in template, ...})$

# Side Chain & Main Chain

- $P(X1 \mid \text{residue type, } \phi, \psi)$
- Main-chain and side-chain modeling can be separated or carried out simultaneously



# Function Fitting from Known Data

- $P(x|a,b,c) \sim W_{x,a,b,c} \sim f(x,a,b,c,q)$
- $W$  is a multi-dimensional table calculated from relative frequencies in the data
- $f$  is a function that fits  $W$  by minimizing root mean square difference
- Fitting algorithm: *Levenberg-Marquardt* algorithm for non-constrained least-squares fitting of a non-linear multidimensional model implemented in the program LSQ.

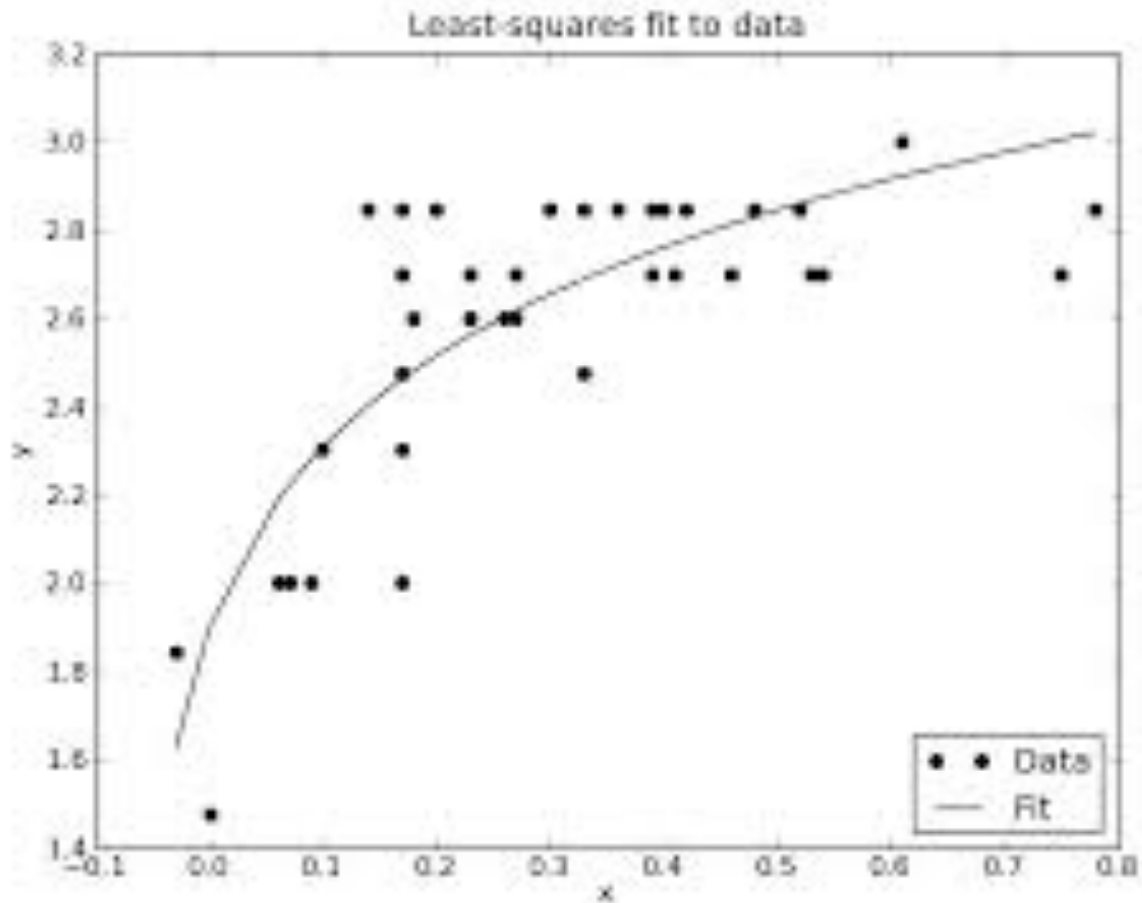
# Features in Multi-Dimensional Table (MDT) Program

- Residue type
- Main-chain dihedral angle of a residue
- Secondary structure of a residue

# An Example of Generating a pdf for one feature (phi angle)

Residue A in target	Residue B in template	Angle in Template	Angle in Target
A	C	50	58, 60, 49, ...
A	C	70	67, 82, 87
A	K	10	95, 11, 10.8...
...	...	...	...

# *Levenberg-Marquardt* algorithm



1	$r$	Amino acid residue type
2	$\Phi$	Main-chain dihedral angle $\Phi$
3	$\Psi$	Main-chain dihedral angle $\Psi$
4	$t$	Secondary structure class of a residue
5	$M$	Main-chain conformation class of a residue
6	$\alpha$	Fractional content of residues in the main-chain conformation class A
7	$\chi_i$	Side-chain dihedral angle $\chi_i$ , $i = 1, 2, 3, 4$
8	$c_i$	Side-chain dihedral angle $\chi_i$ class, $i = 1, 2, 3, 4$
9	$a$	Residue solvent accessibility
10	$\bar{a}$	Average accessibility of two residues in one protein
11	$s$	Residue neighbourhood difference between two proteins
12	$\bar{s}$	Average residue neighbourhood difference between two proteins
13	$i$	Fractional sequence identity between two proteins
14	$d$	C $^\alpha$ -C $^\alpha$ distance
15	$\Delta d$	Difference between two C $^\alpha$ -C $^\alpha$ distances in two proteins
16	$h$	Main-chain N-O distance
17	$\Delta h$	Difference between two main-chain N-O distances in two proteins
18	$b$	Average residue $B_{\text{iso}}$
19	$R$	Resolution of X-ray analysis
20	$g$	Distance of a residue from a gap in alignment
21	$\bar{g}$	Average distance of a residue from a gap

# Main Chain Conformation Class

*Parameters of the main-chain conformation classes*

	Mean ( $^{\circ}$ )		Standard deviation ( $^{\circ}$ )	
	$\bar{\Phi}_i$	$\bar{\Psi}_i$	$\sigma_i(\Phi)$	$\sigma_i(\Psi)$
A	-65	-41	15	15
B	-130	135	15	20
P	-65	140	15	15
G	60	40	10	10
L	90	-10	15	10
E	130	180	25	25

# Usefulness of Features

- The most useful pdf is the one that predicts the unknown feature most accurately
- Measured by the entropy of a pdf

# **Stereochemical Restraints (Generic)**

- Obtained from sequence of a protein
- Bond distance, bond angle, planarity of peptide groups, side-chain rings, chiralities of C $\alpha$  atoms and side-chains, van der Waals contact distance (radii values)
- Mean value and standard deviations for bond lengths, bond angles, and dihedral angles are obtained from GROMOS86



# Bond Length and Angles (harmonic model)

The classical harmonic model for the bond length between two atoms gives the vibrational potential energy of the bond as:

$$E(b) = \frac{1}{2}c(b - b_o)^2. \quad (19)$$

$$p^b(b) = \frac{1}{\sigma_b \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{b - \bar{b}}{\sigma_b} \right)^2 \right] = N(\bar{b}, \sigma_b).$$

# Van der Waals Repulsion (only non-harmonic feature)

(ii) *van der Waals repulsion*

van der Waals repulsion is the only stereochemical feature which is not described by the harmonic model. Instead, the following pdf is used for two atoms:

$$p^v(d) = c \cdot \begin{cases} N(d_o, \sigma_w); & d \leq d_o \\ \frac{1}{\sigma_w \sqrt{2\pi}}; & d_o < d < d_{\max}, \end{cases} \quad (22)$$

where  $d$  is the distance between the two atoms,  $d_o$  is the sum of their van der Waals radii and  $\sigma_w$  is the standard deviation of the Gaussian part of the whole pdf (usually 0.05 Å).  $d_{\max}$  is the maximal possible linear dimension of a protein and constant  $c$  is chosen so that  $p^v(d)$  integrates to 1. This pdf does not differentiate between contact distances larger than  $d_o$ , but it does select against distances smaller than  $d_o$ . This is achieved by imposing a repulsive harmonic potential on atoms that are less than  $d_o$  apart.

# Ca-Ca Distance Features

$$p^d(d|\bar{g}, i, \bar{a}', d') = \frac{1}{\sigma(\bar{g}, i, \bar{a}', d')\sqrt{2\pi}} \times \exp\left[-\frac{1}{2}\left(\frac{d-d'}{\sigma(\bar{g}, i, \bar{a}', d')}\right)^2\right]$$

Standard deviation depends on solvent accessibility, gaps of alignment, and sequence identity.

# Combine pdfs of a Feature (Ca-Ca distance) from Multiple Templates

- Weighted sum of the same type of pdfs from multiple known structures

The last step in the derivation of the feature pdf is to include the van der Waals restraint. Since all stereochemical restraints have to be satisfied in all structures, these restraints are multiplied into the feature pdf and we obtain the final feature pdf:

$$p^D(d) = [\omega_1 p_1^d(d) + \omega_2 p_2^d(d)] p^v(d).$$

# Derivation of a molecular pdf from feature pdfs

- Combine all feature pdfs into a molecular pdf 
$$P = \prod_i p^F(f_i). \quad (34)$$
- 3D structure of a protein is uniquely determined if a sufficient large number of its features,  $f_i$ , are specified
- The goal is to find the 3D structure that is consistent with the most probable values of individual features  $f_i$ , i.e. to maximize the molecular pdf.

# Optimization

- Optimize the logarithm of molecular pdf – the objective function

F. 
$$F = -\ln(P), \quad (35)$$

- All the features of the molecular pdf is expressed in terms of atomic Cartesian coordinates (x, y, z)
- F is more suitable for optimization because multiplication is converted into addition and the problem of floating point overflow is smaller for F.

# Successive Optimization

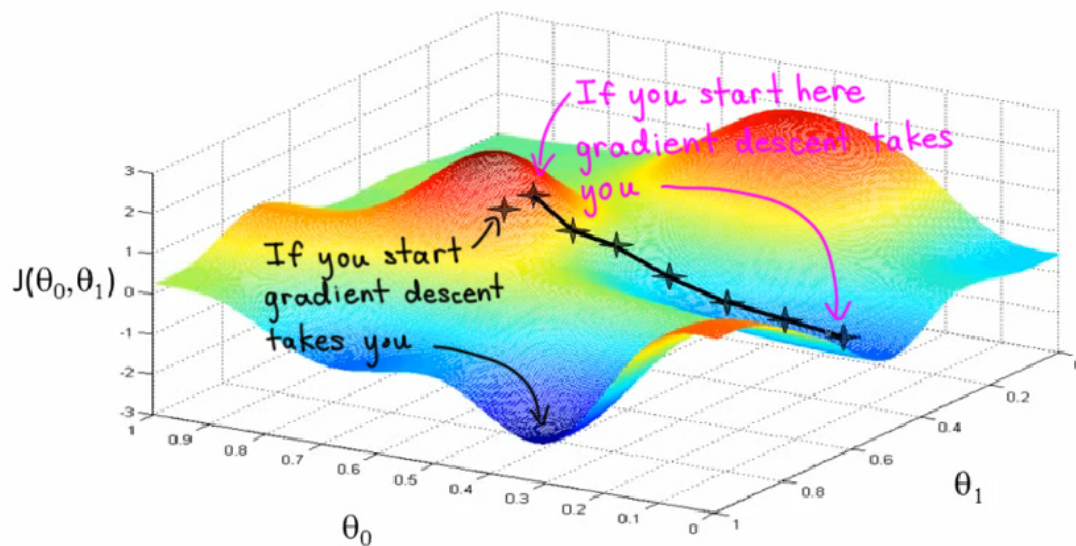
- The optimum of the molecular pdf is found by successive optimization of increasingly more complex target function till the whole molecular pdf.
- From local restraints to long-range restraints to all the restraints
- Restraints is ordered by the sequence distance between atoms / residues (1, 2, ... N-1), N is the sequence length.
- Successively adding restraints with  $\leq$  sequence distance  $i$  at each step  $i$ .

# Initial Conformation of Step $i$

- At step 1, initial conformation can be an extended chain, or a conformation derived from the extended chain by rotation of dihedral angles
- At step  $i$ , the initial conformation is the final conformation of step  $i - 1$ .
- An ensemble of conformations will be produced by using different initial conformations.



# Optimization: Gradient Descent



# Gradient Descent

$$x^{t+1} = x^t + d^t$$

$$d^t = -\eta \frac{\partial f}{\partial x^t}$$

# Gradient Descent

- **Random Initialization:**  $(x_1^0, y_1^0, z_1^0), (x_2^0, y_2^0, z_2^0), \dots, (x_N^0, y_N^0, z_N^0)$
- **Update:**

$$\mathbf{X}_1^{t+1} = \mathbf{X}_1^t - \eta^* \Delta \mathbf{X} \quad \mathbf{Y}_1^{t+1} = \mathbf{Y}_1^t - \eta^* \Delta \mathbf{Y} \quad \mathbf{Z}_1^{t+1} = \mathbf{Z}_1^t - \eta^* \Delta \mathbf{Z}$$

$$\mathbf{X}_2^{t+1} = \mathbf{X}_2^t - \eta^* \Delta \mathbf{X} \quad \mathbf{Y}_2^{t+1} = \mathbf{Y}_2^t - \eta^* \Delta \mathbf{Y} \quad \mathbf{Z}_2^{t+1} = \mathbf{Z}_2^t - \eta^* \Delta \mathbf{Z}$$

•  
•  
•

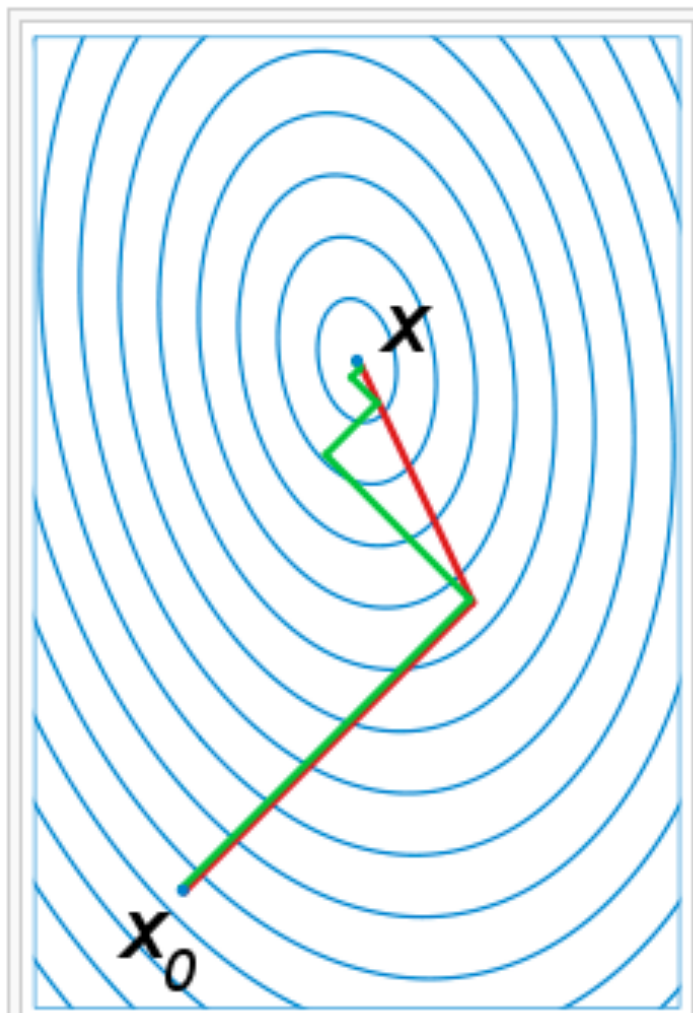
$$\mathbf{X}_N^{t+1} = \mathbf{X}_N^t - \eta^* \Delta \mathbf{X} \quad \mathbf{Y}_N^{t+1} = \mathbf{Y}_N^t - \eta^* \Delta \mathbf{Y} \quad \mathbf{Z}_N^{t+1} = \mathbf{Z}_N^t - \eta^* \Delta \mathbf{Z}$$

# Conjugate Gradient Descent

$$x^{t+1} = x^t + \eta d^t$$

$$d^t = -\frac{\partial f^t}{\partial x^t} + d^{t-1}$$

$$d^t = -\frac{\partial f^t}{\partial x^t}$$



A comparison of the convergence of [gradient descent](#) with optimal step size (in green) and conjugate vector (in red) for minimizing a quadratic function associated with a given linear system. Conjugate gradient, assuming exact arithmetic, converges in at most  $n$  steps where  $n$  is the size of the matrix of the system (here  $n=2$ ).

*Spatial restraints used to model trypsin*

Type	Basis pdfs <sup>a</sup>	Feature pdfs <sup>b</sup>	Violations <sup>c</sup>	r.m.s. <sup>d</sup>	r.m.s. <sup>e</sup>
Bond lengths	1659	1659	0 (0.1 Å)	0.005 Å	0.005 Å
Bond angles	2250	2250	5 (10°)	2.00°	2.00°
Dihedral angles <sup>f</sup>	919	919	1 (20°)	3.40°	3.40°
van der Waals contacts <sup>g</sup>	531	531	0 (0.2 Å)	0.02 Å	0.02 Å
C <sup>α</sup> -C <sup>β</sup> distances	23,538	11,914	26 (1.5 Å)	0.22 Å	0.47 Å
Main-chain N-O distances	7480	3832	19 (1.5 Å)	0.31 Å	0.51 Å
Main-chain Φ dihedral angles	1110	222	2 (20°)	10.8°	21.2°
Main-chain Ψ dihedral angles	1332	222	9 (20°)	10.6°	20.3°
Side-chain χ <sub>1</sub> dihedral angles	528	176	5 (25°)	8.4°	16.8°
Side-chain χ <sub>2</sub> dihedral angles	264	103	3 (25°)	10.2°	13.0°
Side-chain χ <sub>3</sub> dihedral angles	92	32	2 (25°)	11.9°	48.1°
Side-chain χ <sub>4</sub> dihedral angles	48	16	0 (25°)	4.5°	21.9°
Disulphide bridge bonds	6	6	0 (0.1°)	0.007 Å	0.007 Å
Disulphide bridge angles	12	12	0 (10°)	3.7°	3.7°
Disulphide bridge dihedral angles	6	12	0 (20°)	10.0°	12.9°
<i>cis</i> -Peptides <sup>h</sup>	0	0			

# Group Formation

- **Group 1:** Avimanyou Vasta, Jie Hou, Muhammad Yasin, Chao Fang, Tosin Oluwadare
- **Group 2:** Vasanthi Mandhadi, Sowmya Ankilli, Shirish Gajul, Shashank Chary, Sai Amsanpally
- **Group 3:** Chao He, Xutong Zhang, Yiran Tian, Yiming Liang, Zhen Lyu
- **Group 4:** Zhaoyu Li, Peng Sun, Junlin Wang, Xiaohui Yang, Tianbo Wang

# Project 1

- Design and develop a template-based protein structure modeling tool
- Assess its performance on a few TBM targets used in CASP11 benchmark



# Project Directory

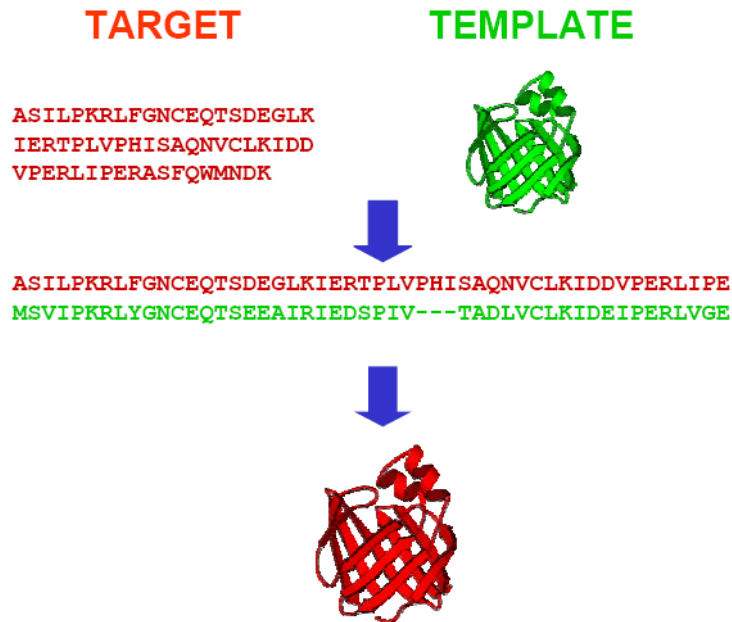
- Project1
- ---- src: source code
- ---- bin: binary
- ---- lib: library
- ---- data: data
- ---- training: training
- ---- test: test cases
- ---- doc: document / references / presentation / report
- ---- other: third-party programs

# Discussion of Your Project Plan

- Data preparation & data sharing (cloud computing)
- Algorithm development (initialization, restraints extraction & representation, sampling, optimization): creative, alternative, plural
- Implementation: interface, design, platform, languages, code base / from scratch, task assignment, timeline, progress track
- Evaluation plan (metrics, tools, data, objective, comprehensive, expectation)
- Challenges, Technical Hurdles, Feasibility, Strength, weakness, Risks
- Visualization
- Software Package (installation, test cases)

# Useful Tools

- Loop modeling: <http://www.math.unm.edu/~vageli/codes/codes.html>



- Tools convert between (x,y,z) coordinates and (phi, psi) angles :Rosetta function
- ModLoop a web server for loop modeling based on Modeller

# **Key Milestones of the Project**

- Class discussion (Feb. 10, Wednesday)
- Presentation of your plan on Feb. 15 (Monday)
- Presentation of your results on Feb. 24 (Wednesday)