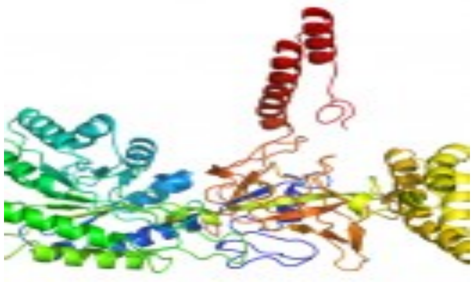
Mizzou

University of Missouri



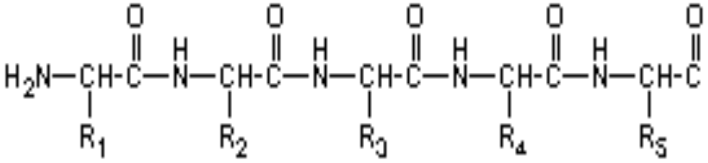
Template Based Protein Structure Modeling

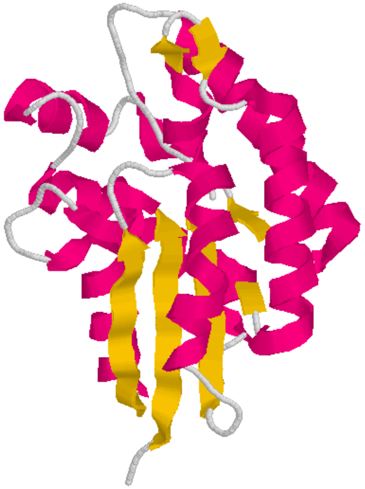
Jianlin Cheng, PhD

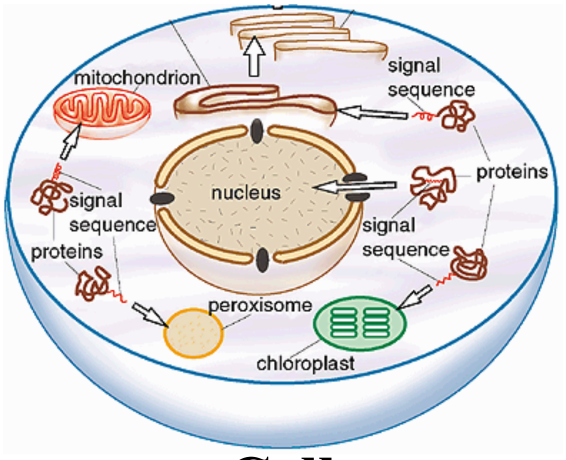
Professor
Department of EECS
Informatics Institute
University of Missouri, Columbia
2019

Sequence, Structure and Function

AGCWY……





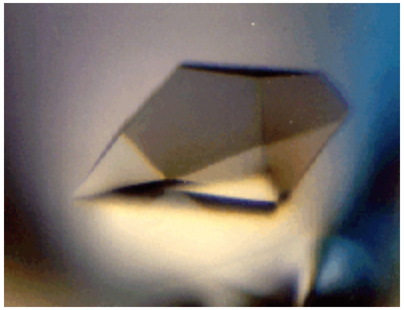


Cell

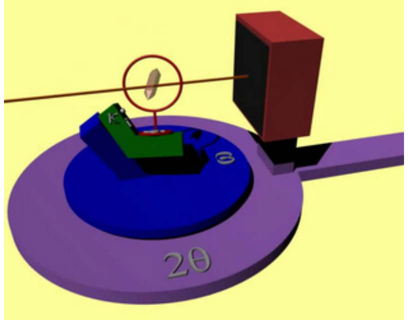
Protein Structure Determination

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy
- Cryo-Electron Microscopy
- 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

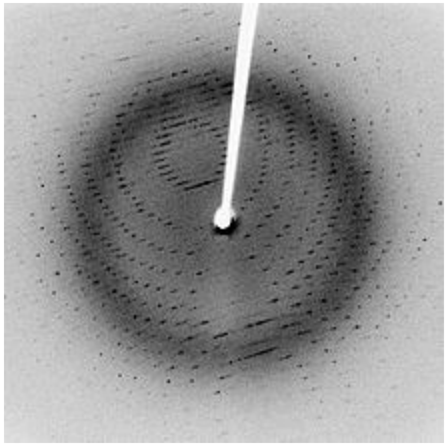


A protein crystal

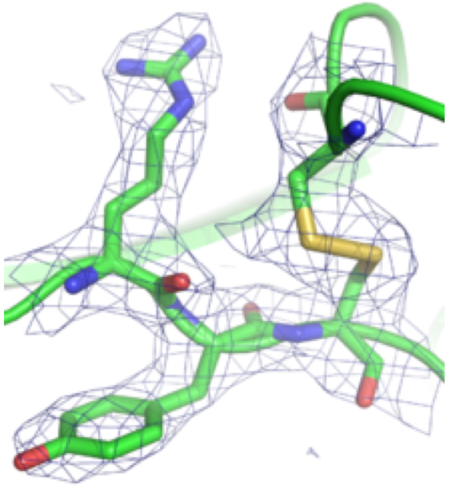


Mount a crystal

Diffractometer



Diffraction



Protein structure

Wikipedia

Kendrew and Perutz won 1962 Nobel Prize

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

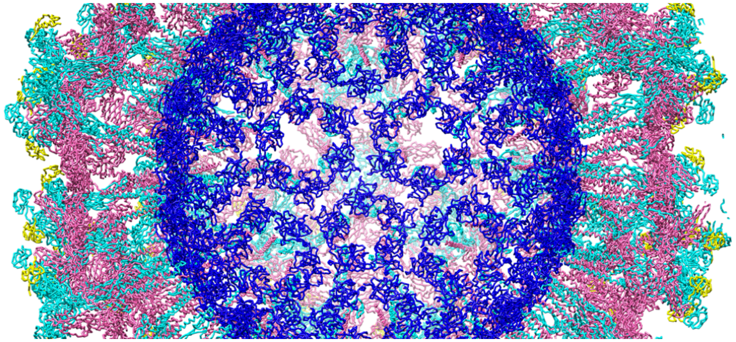
Wikipedia, the free encyclopedia

- **Key idea**: measure the distance between atoms in protein
- Build 3D structures by satisfying the distance between atoms using computational tools such as Crystallography and NMR system (CNS).

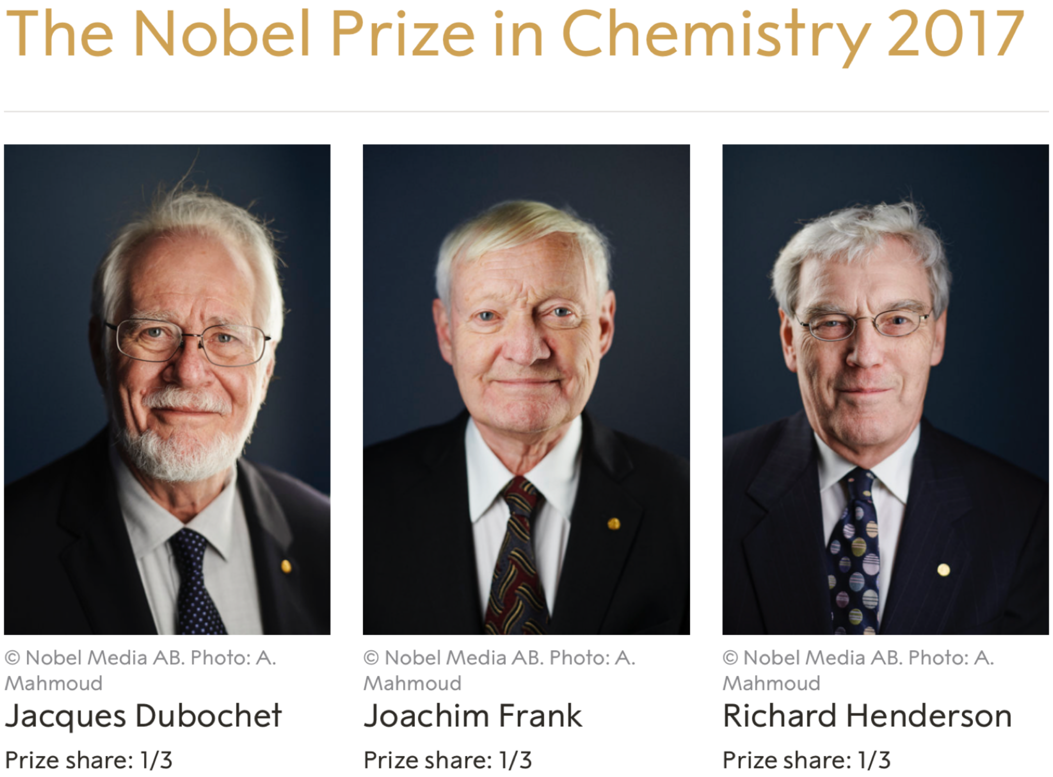
photo Susi Lindig

- **Kurt Wüthrich, Switzerland:** Nobel Prize in Chemistry 2002, "for his development of nuclear magnetic resonance spectroscopy for determining the three-dimensional structure of biological macromolecules in solution"

- **Cryo-EM equipment**



- Key idea: generate 2D images of proteins from different angles, and them assemble them into one 3D structure. A lot of imaging techniques used.



The Nobel Prize in Chemistry 2017

© Nobel Media AB. Photo: A. Mahmoud

Jacques Dubochet

Prize share: 1/3

© Nobel Media AB. Photo: A. Mahmoud

Joachim Frank

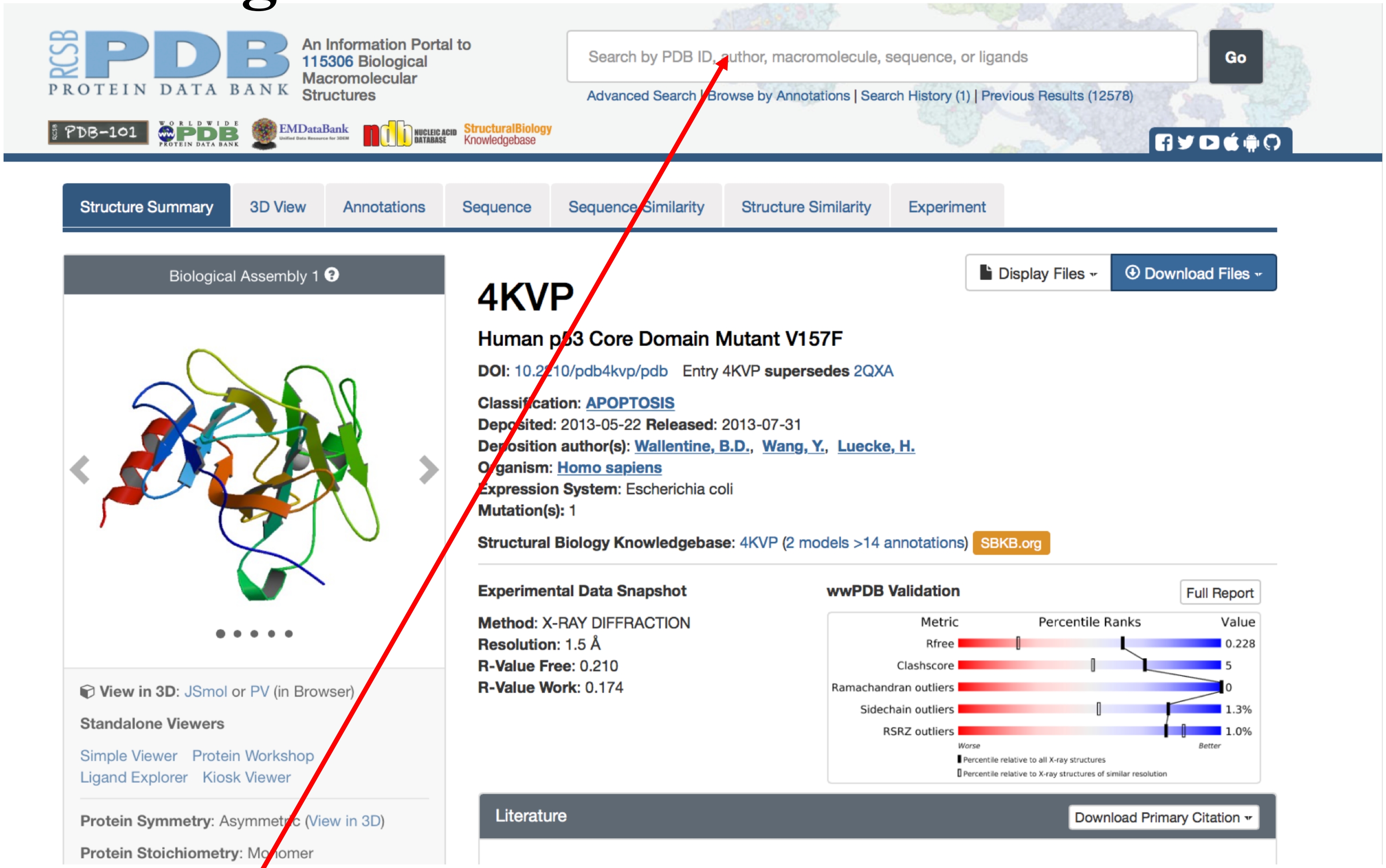
Prize share: 1/3

© Nobel Media AB. Photo: A. Mahmoud

Richard Henderson

Prize share: 1/3

Storage in Protein Data Bank



RCSB PDB PROTEIN DATA BANK

An Information Portal to 115306 Biological Macromolecular Structures

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

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

PDB-101 | wwPDB | EMDataBank | NDB Nucleic Acid Database | StructuralBiology Knowledgebase

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

Biological Assembly 1

Display Files | Download Files

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

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 1.5 Å

R-Value Free: 0.210

R-Value Work: 0.174

wwPDB Validation | Full Report

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

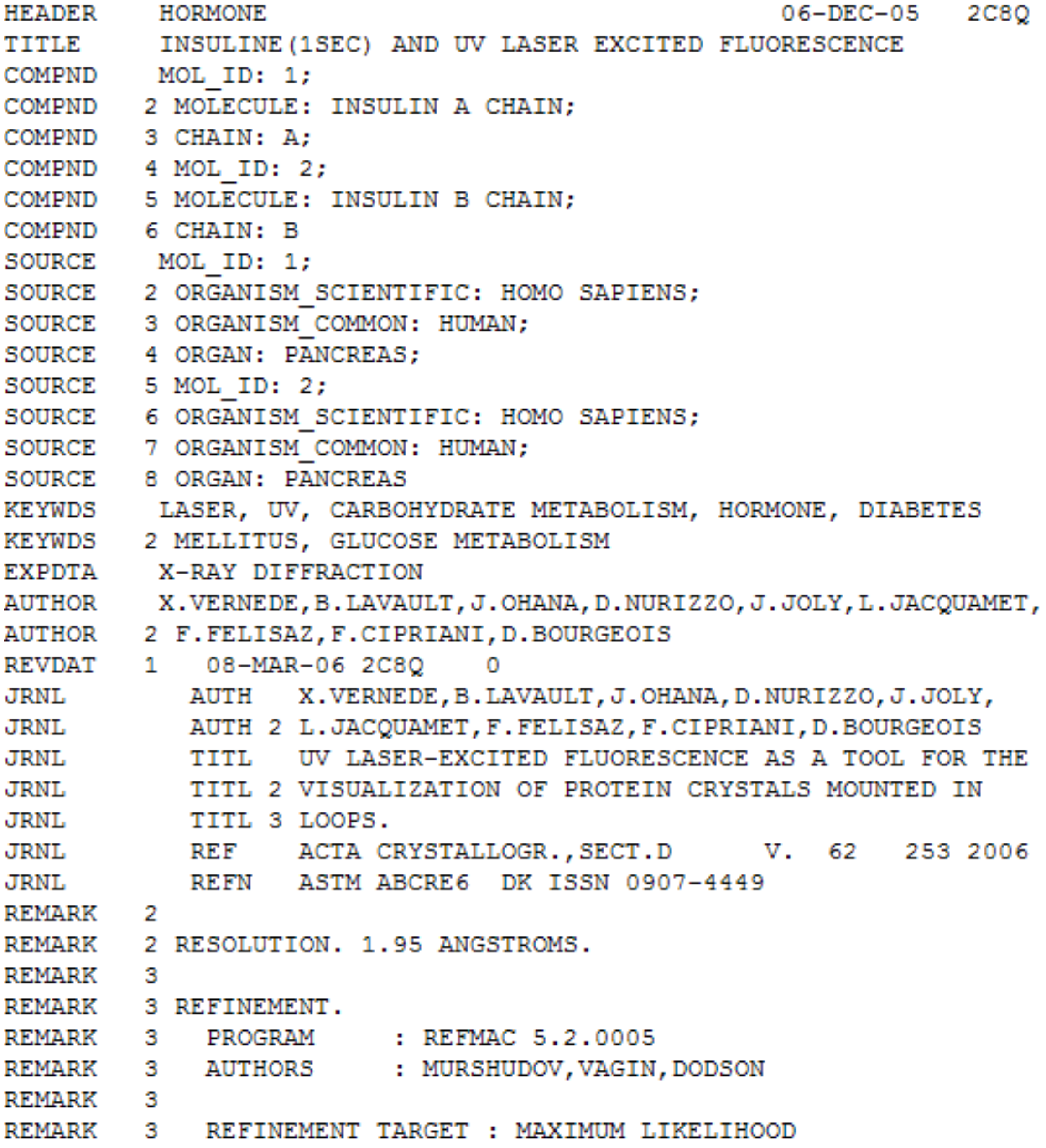
Literature | Download Primary Citation

Search database

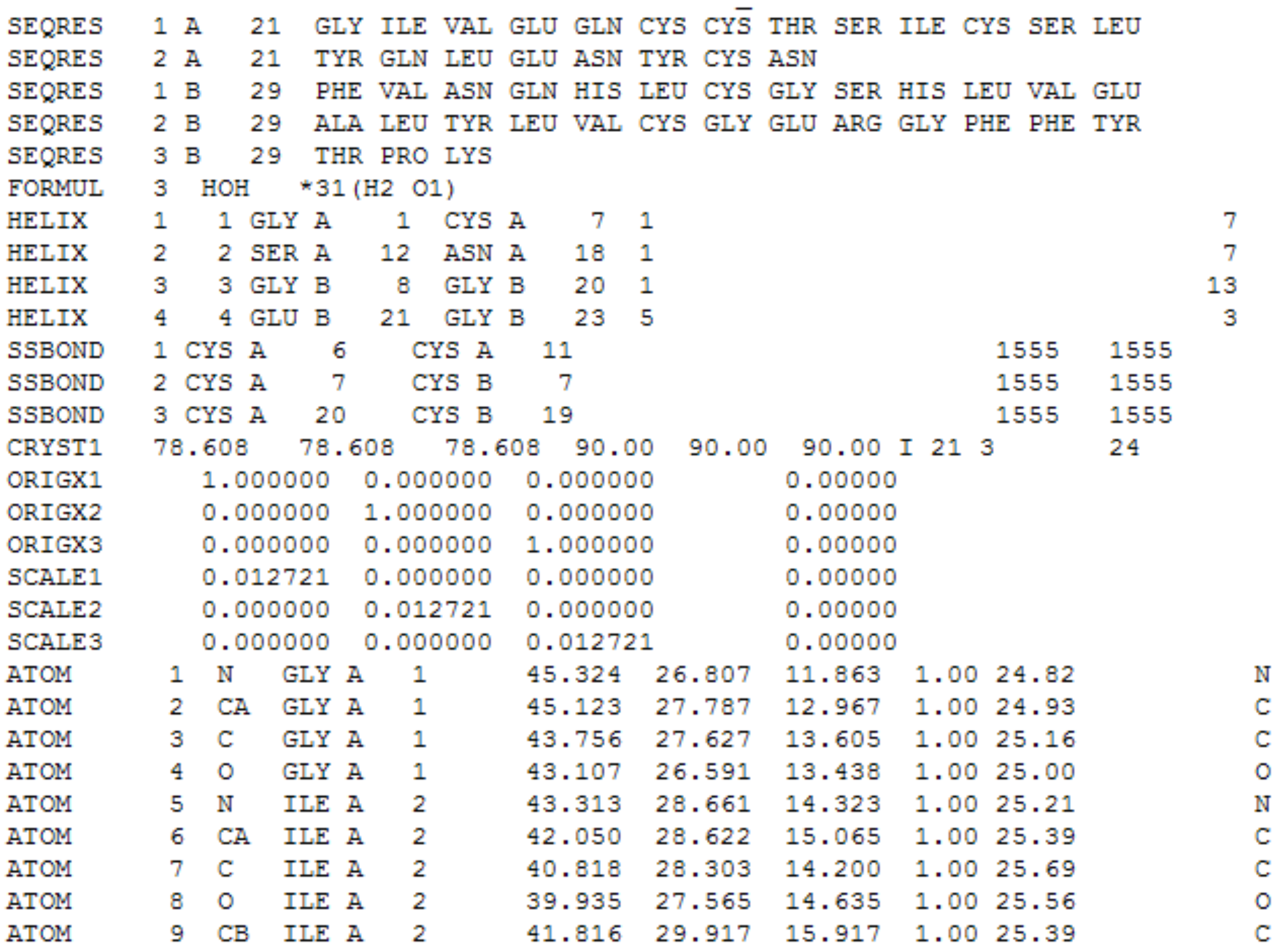
Search Demo: Human P53 protein – 1KVP

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
```



```
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
```

Structure Visualization

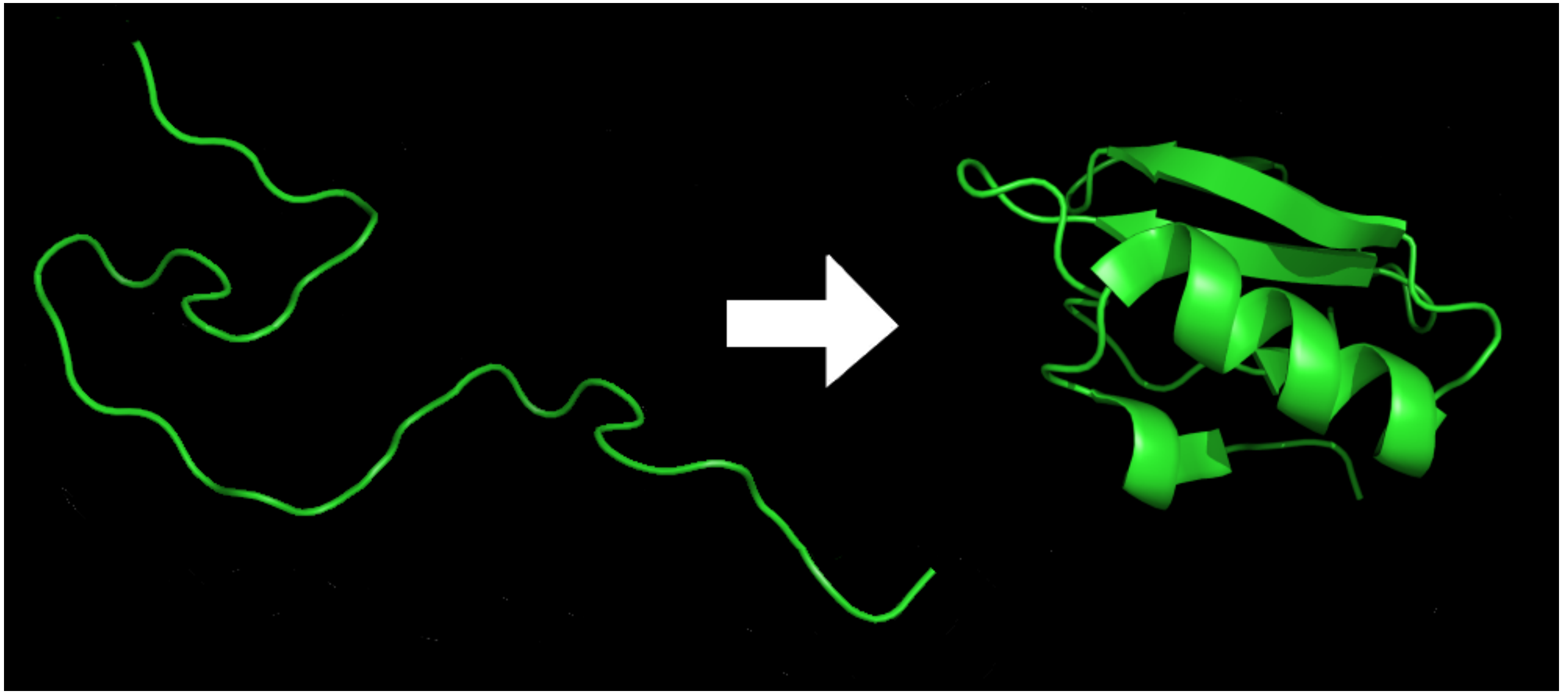
- Rasmol (http://www.umass.edu/microbio/rasmol/getras.htm)
- MDL Chime (plug-in) (http://www.mdl.com/products/framework/chime/)
- **Jmol: http://jmol.sourceforge.net/**
- **JSMol: java script version**
- **Pymol: http://pymol.sourceforge.net/**
- **Chimera: https://www.cgl.ucsf.edu/chimera/**

JSMol (4KVP, Human P53)

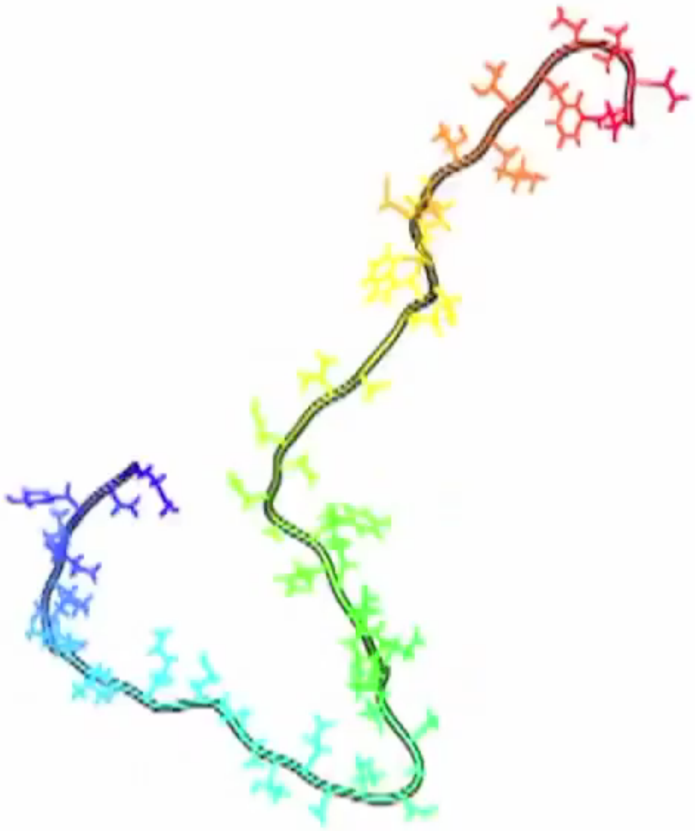
- JSMol: http://www.rcsb.org/pdb/explore/jmol.do?structureId=4KVP&bionumber=1
- JMOL: 1VJP
- Identify residues
- Recognize atoms
- Recognize peptide bonds
- Identify backbone
- Identify side chain
- Analyze different visualization style

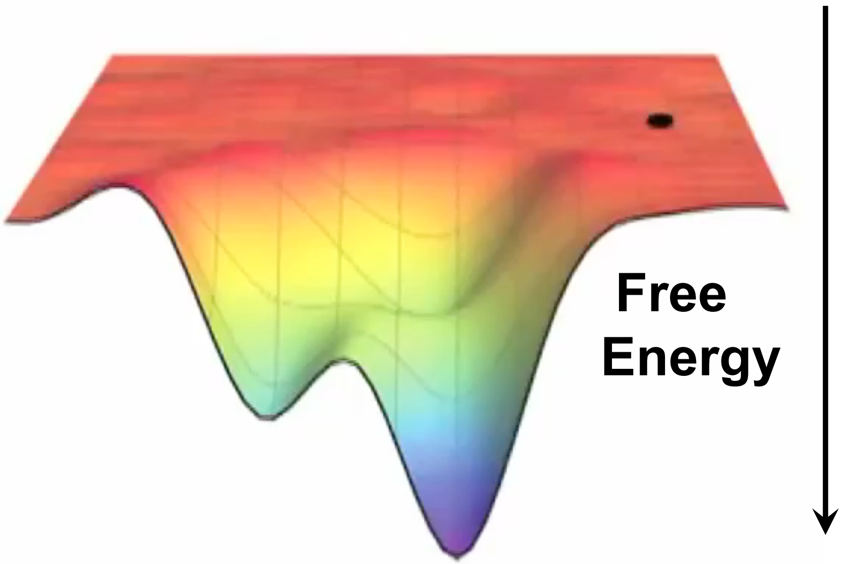
Protein Folding

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



Computational Protein Folding by MULTICOM (Demo)





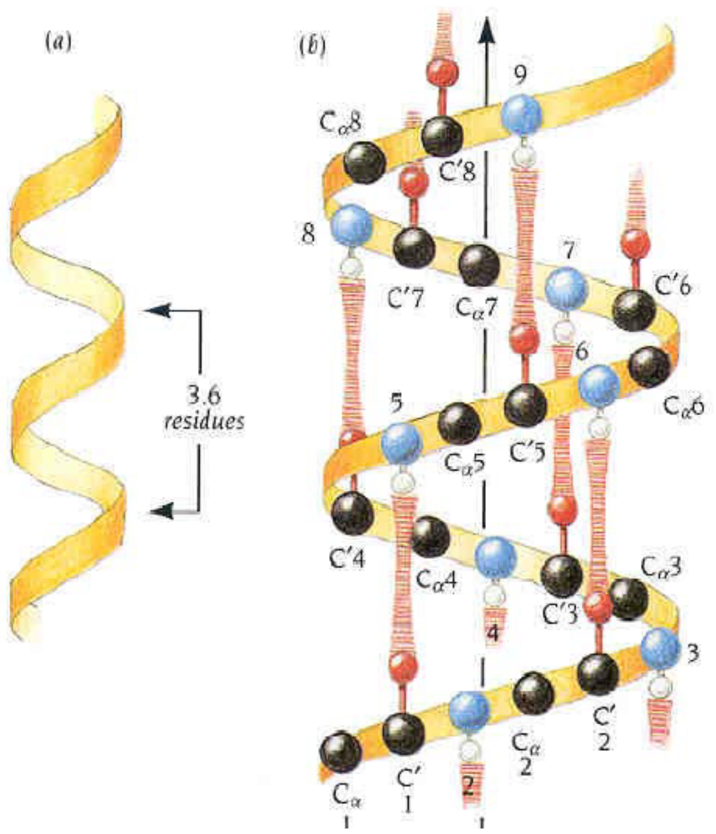
Funnel-shaped energy landscape

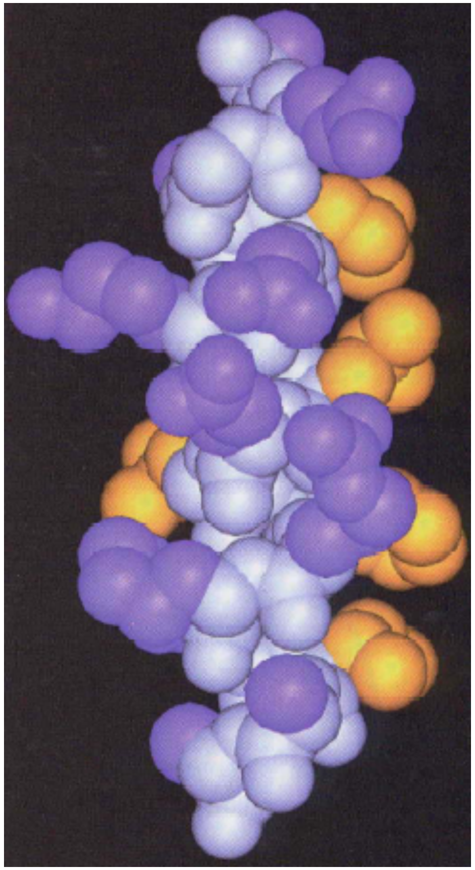
Bhattacharya & Cheng, 2015

AlphaFold Movie

- https://deepmind.com/blog/alphafold/#gif-242

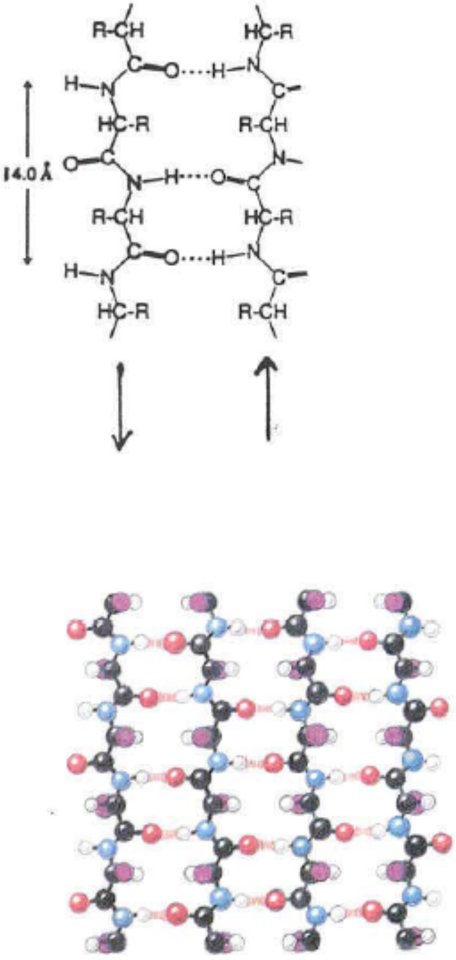
Alpha-Helix



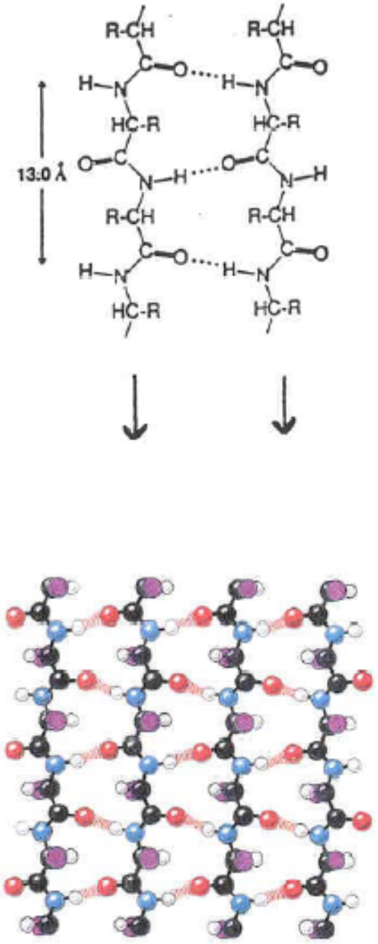


Jurnak, 2003

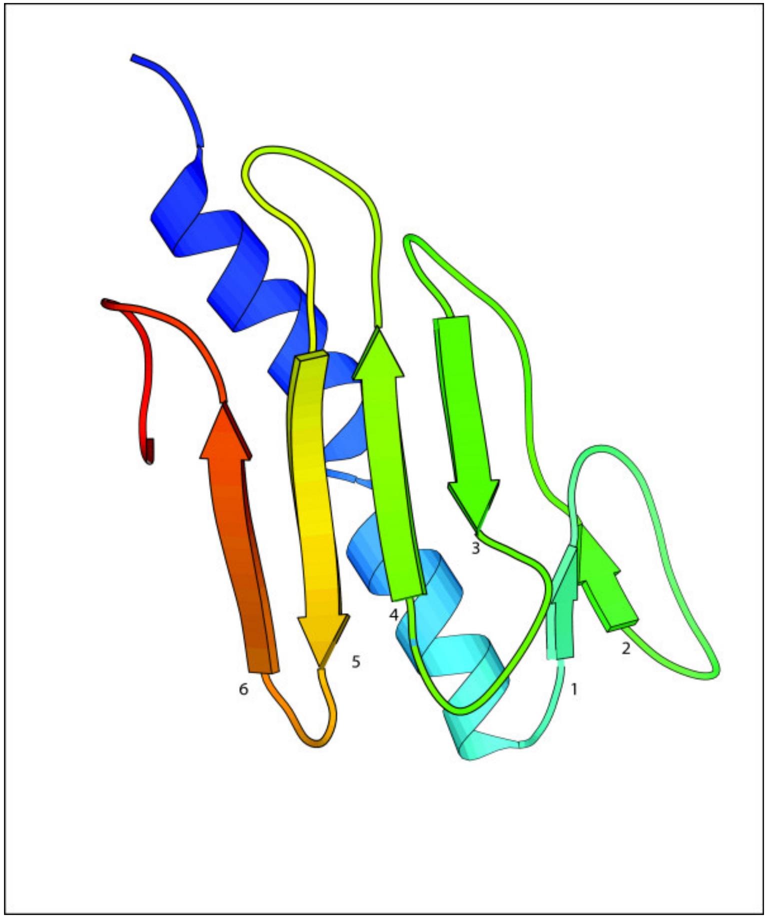
Beta-Sheet



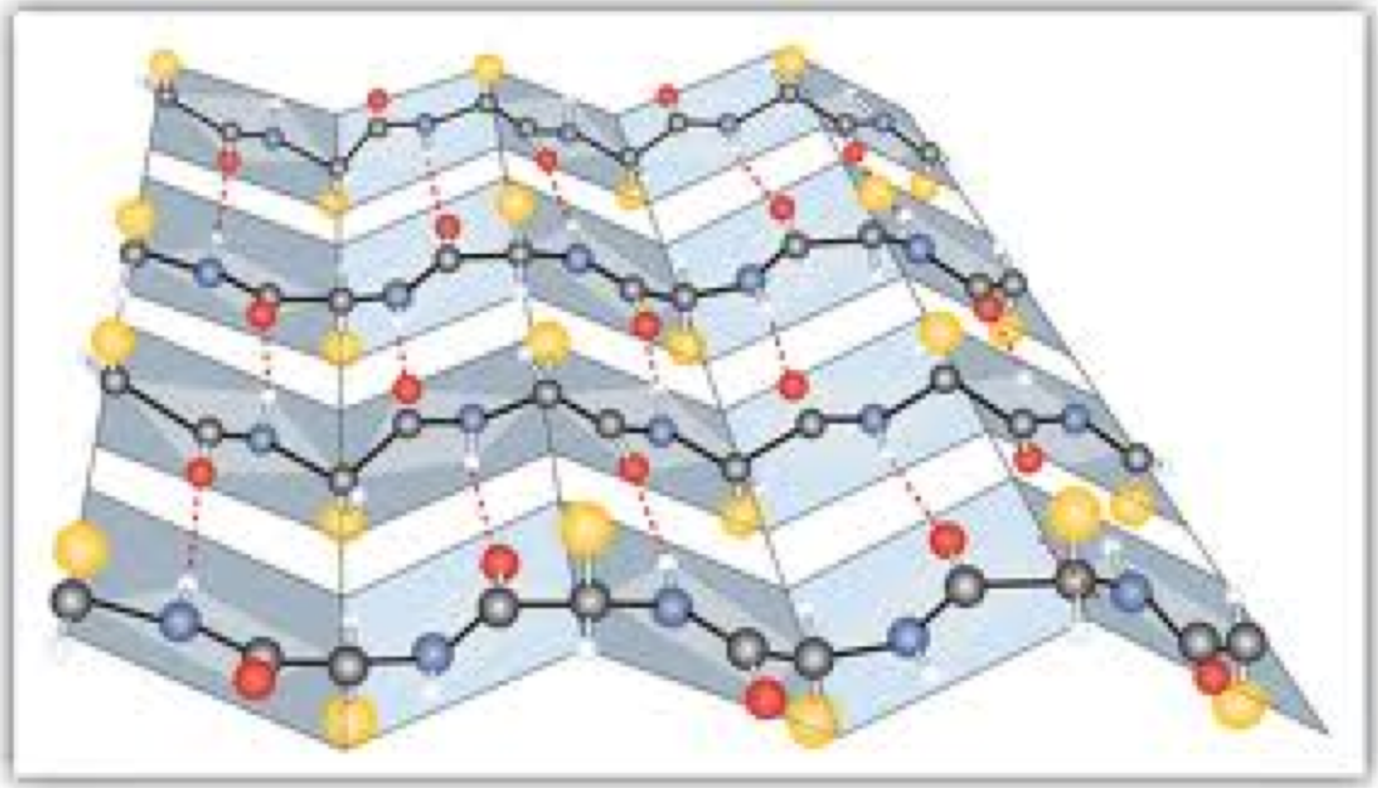
Anti-Parallel



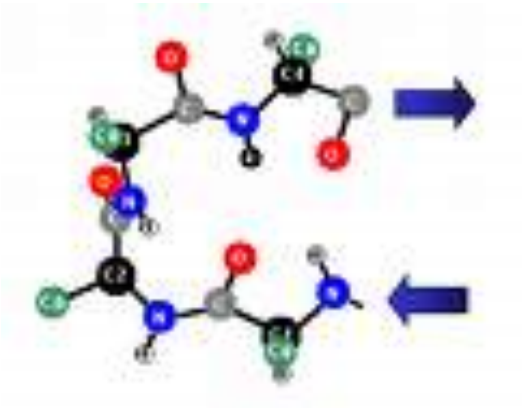
Parallel



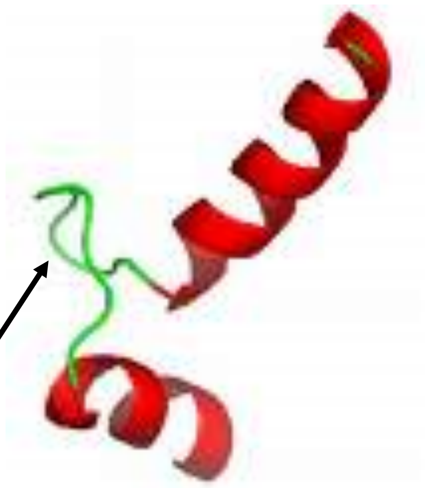
Beta-Sheet



Non-Repetitive Secondary Structure



Beta-Turn



Loop

Announcement – Next Class

Data-driven modeling of protein structure, 3D genome and gene regulatory network

Jianin Cheng

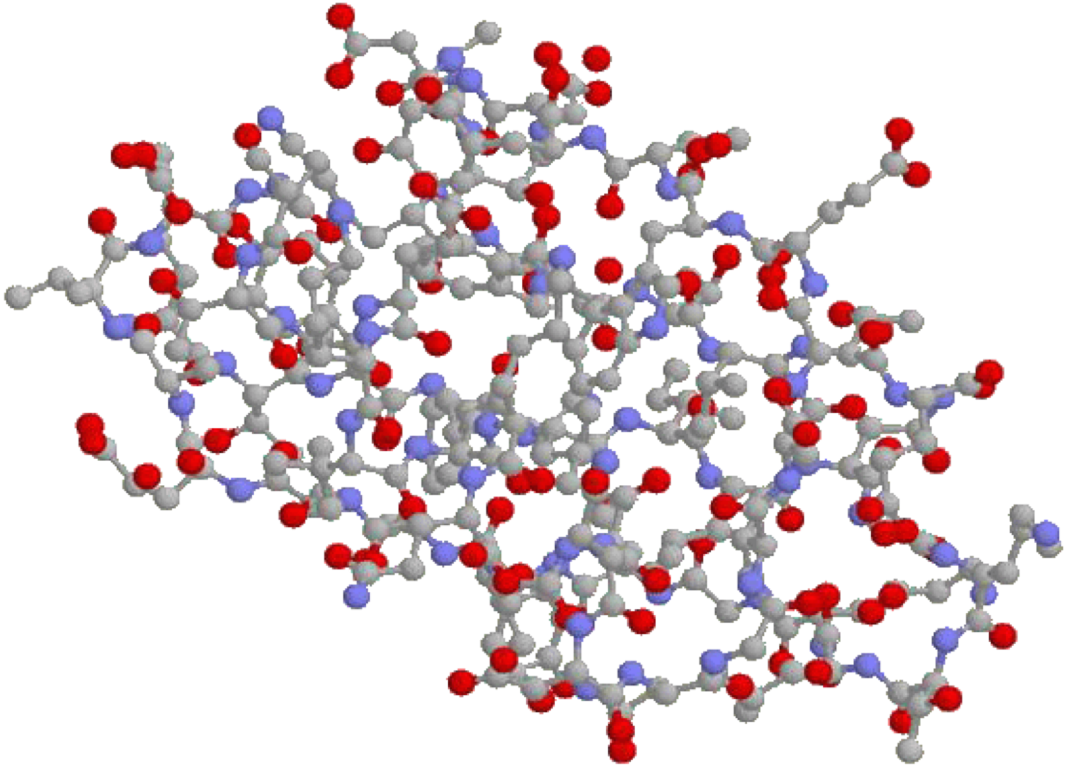
Hosted by:
Dr. Zezong Gu

Monday, Feburary, 11, 2019
4:00 p.m.

Pathology Conference Room
MA223 Medical Sciences Building Annex

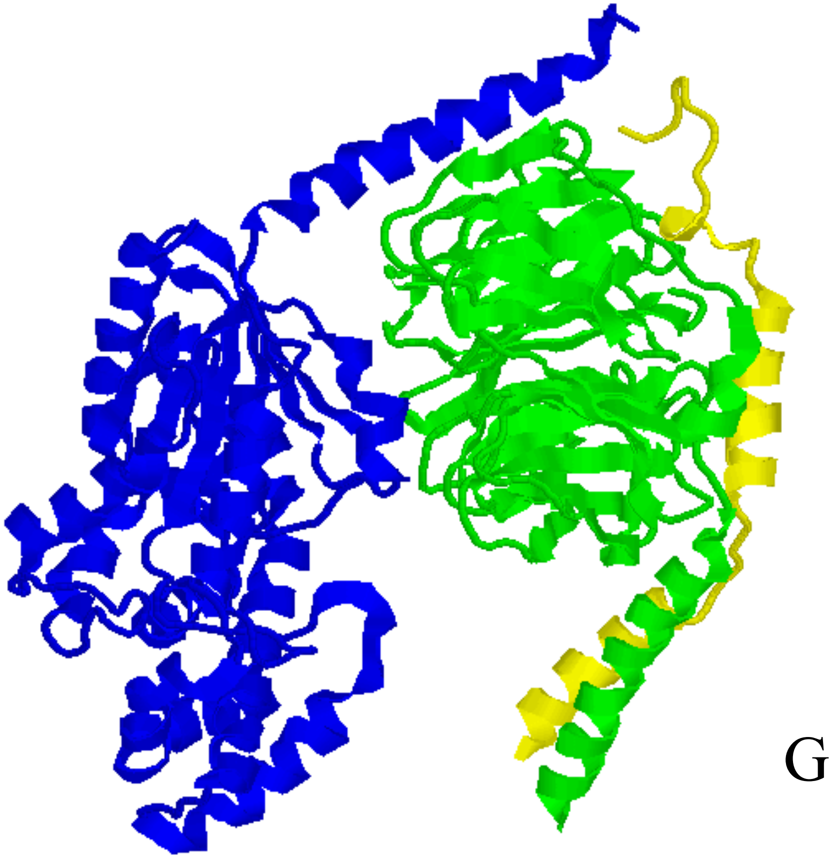
Refreshments provided at 3:50 pm

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

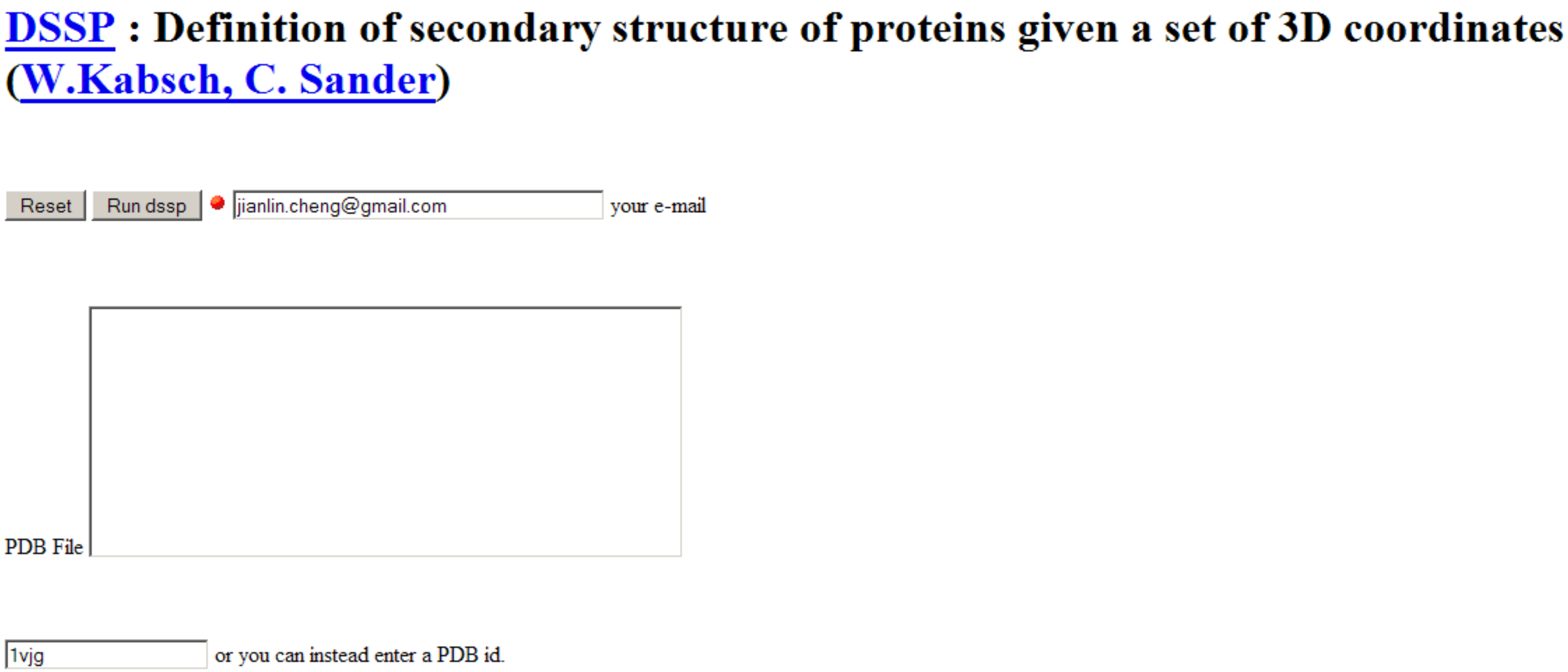
E = extended strand, participates in beta ladder

T = hydrogen bonded turn

S = bend

Blank = loop

DSSP Web Service



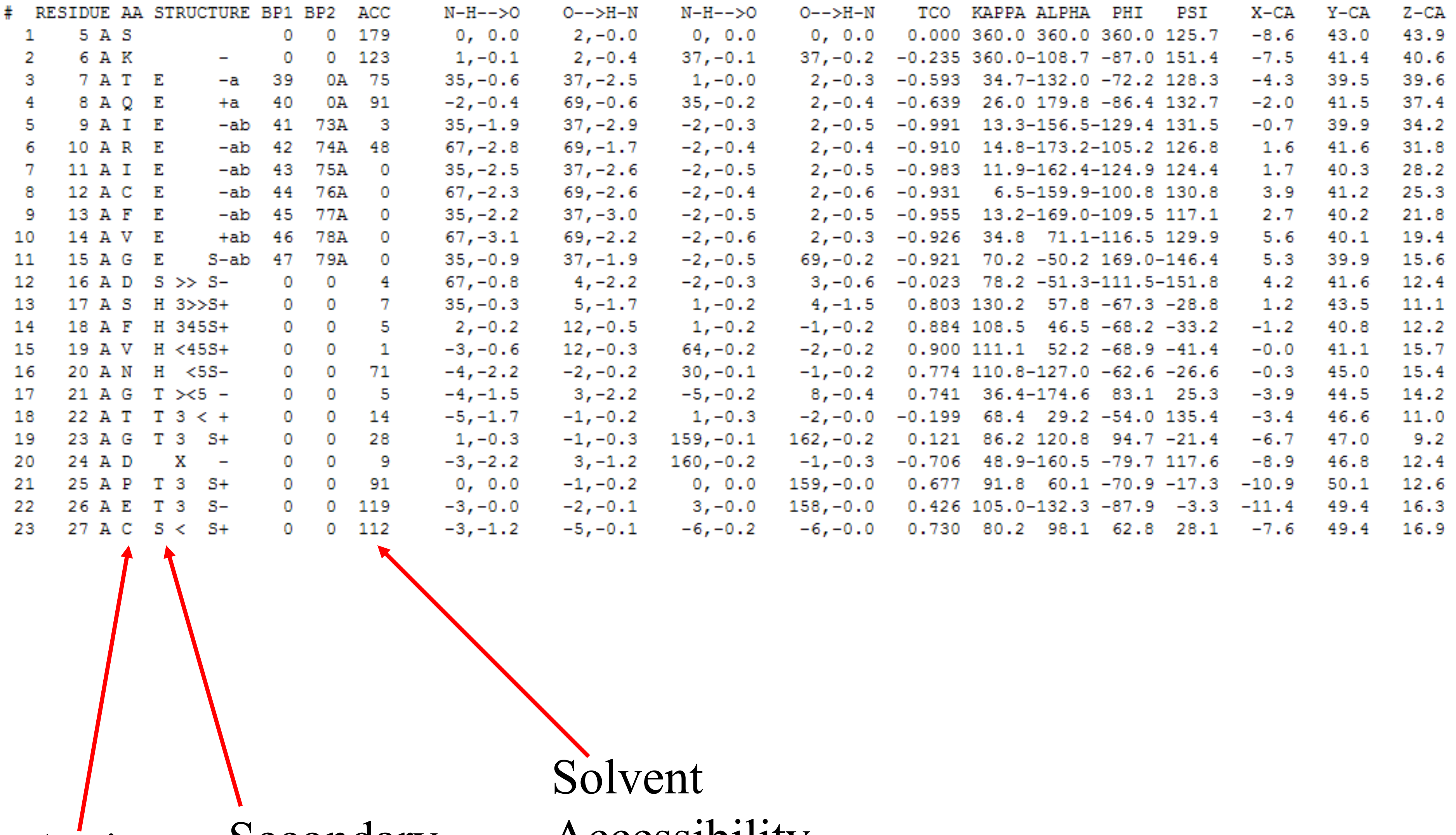
DSSP : Definition of secondary structure of proteins given a set of 3D coordinates (W.Kabsch, C. Sander)

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



```
  #  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
```

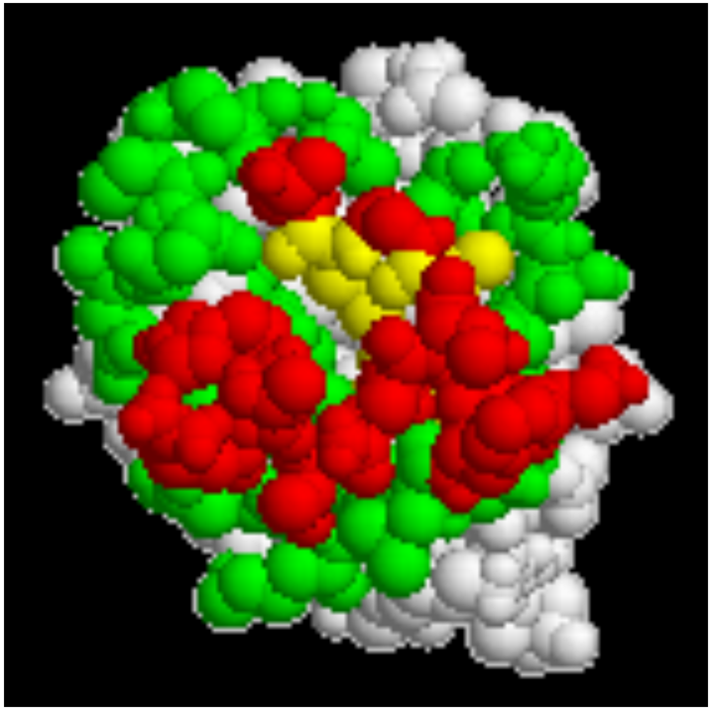
Amino Acids

Secondary Structure

Solvent Accessibility

Solvent Accessibility

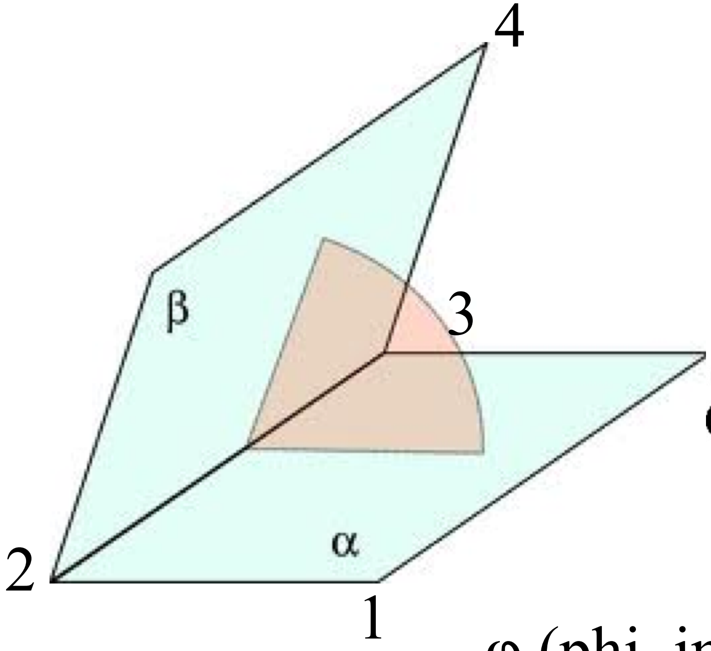
Size of the area of an amino acid that is exposed to solvent (water).

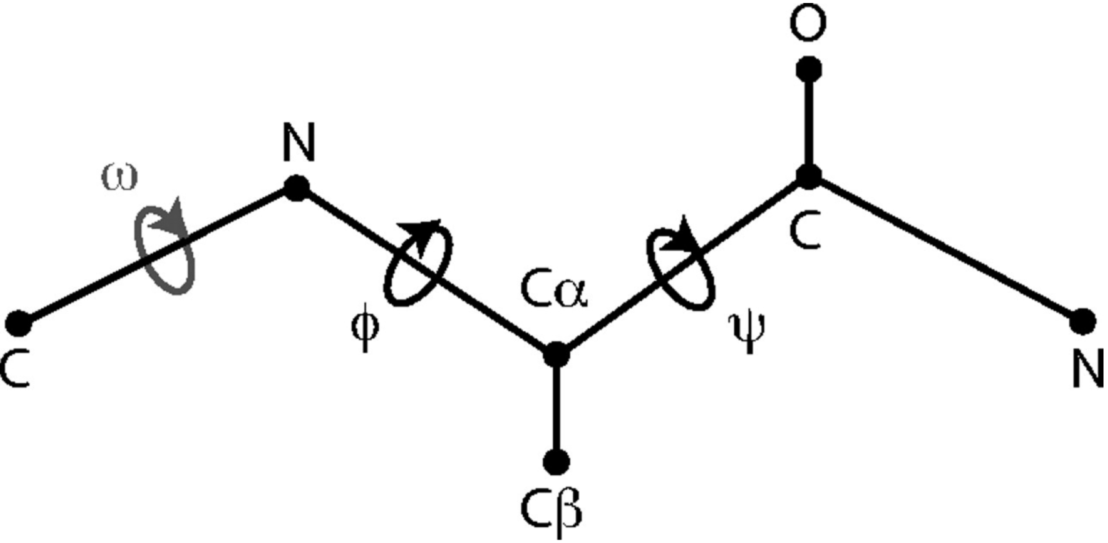


Maximum solvent accessible area for each amino acid is its whole surface area.

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

Dihedral / Torsional Angle





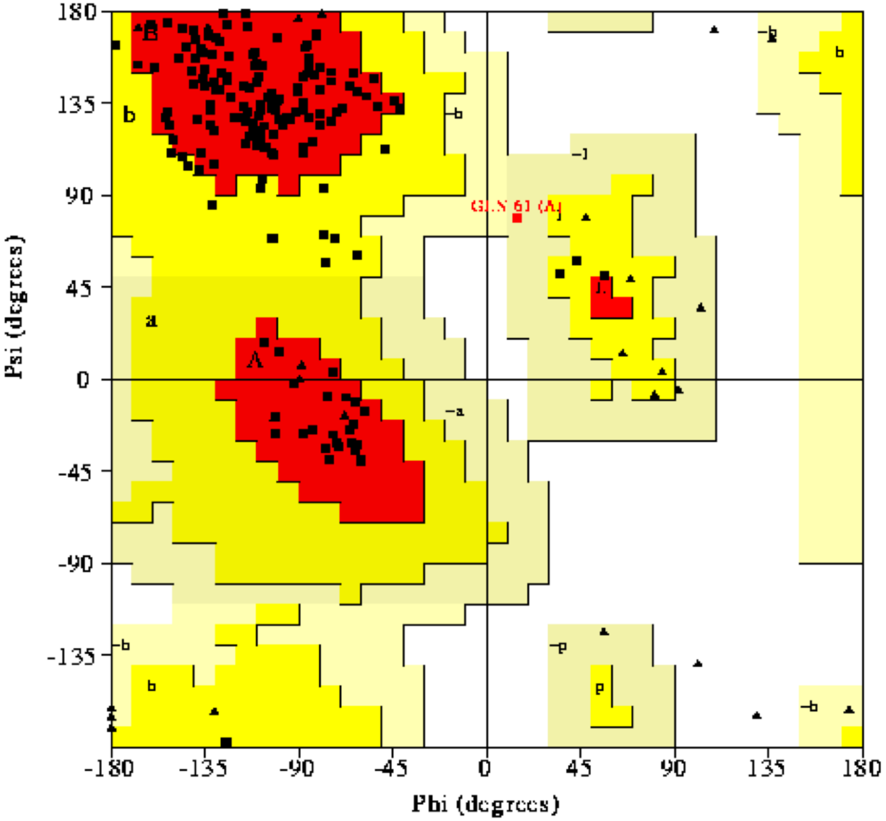
φ (phi, involving the backbone atoms C'-N-Cα-C'), ψ (psi, involving the backbone atoms N-Cα-C'-N)

- http://en.wikipedia.org/wiki/Dihedral_angle

PROCHECK

Ramachandran Plot

1abc



Plot statistics

Residues in most favoured regions [A,B,L]	143	89.9%
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	

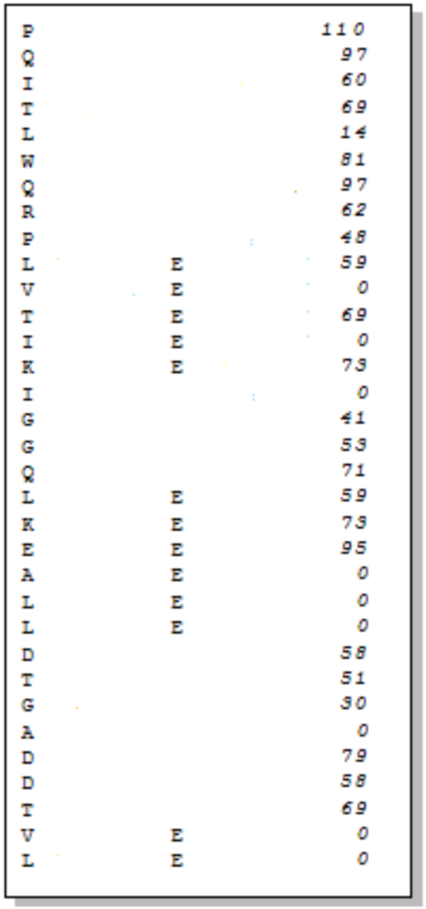
Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.

1abc_01.ps

Project Groups

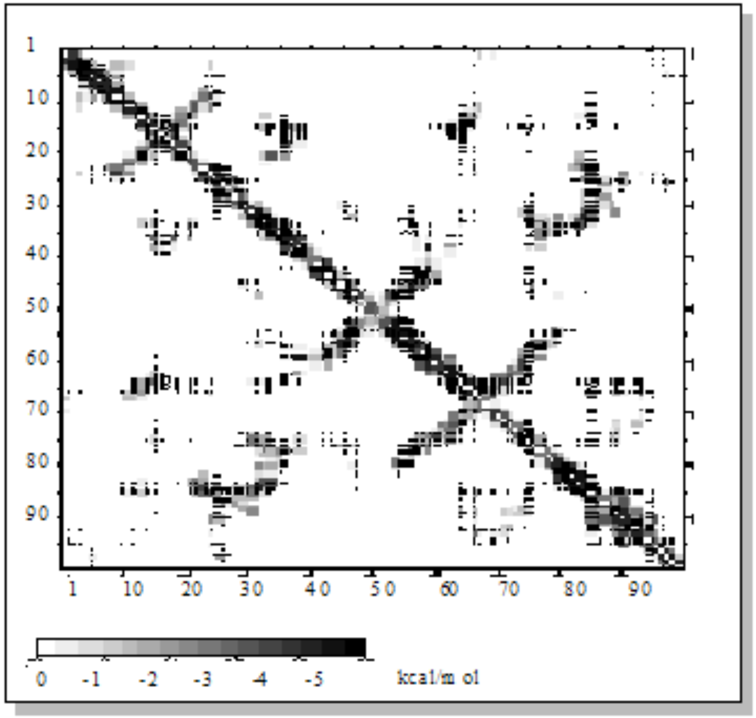
- 19 students?
- Form 4 groups (4-5 students per group)

Protein Structure 1D, 2D, 3D

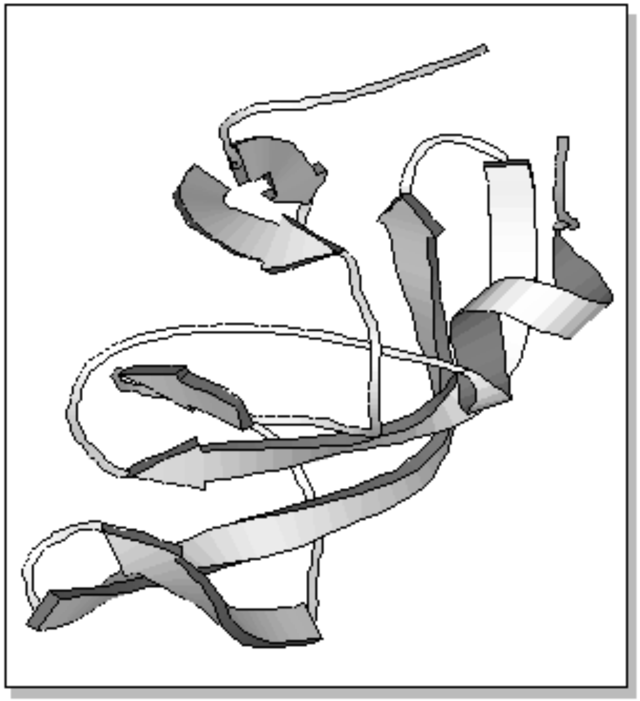


P		110
Q		97
I		60
T		69
L		14
W		81
Q		97
R		62
P		48
L	E	59
V	E	0
T	E	69
I	E	0
K	E	73
I		0
G		41
G		53
Q		71
L	E	59
K	E	73
E	E	95
A	E	0
L	E	0
L	E	0
D		58
T		51
G		30
A		0
D		79
D		58
T		69
V	E	0
L	E	0

1D



2D



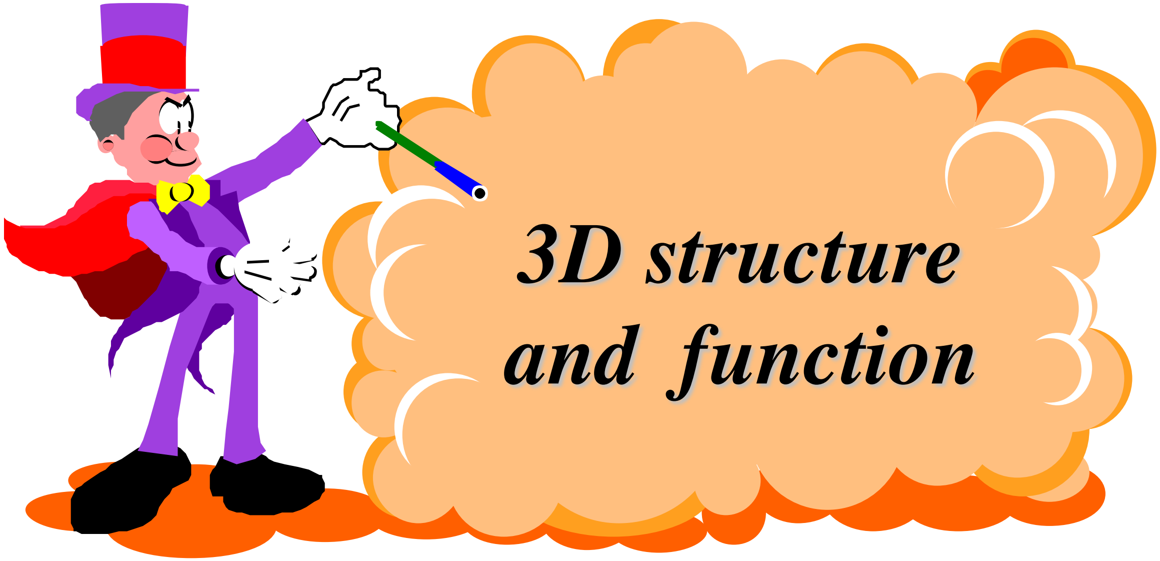
3D

B. Rost, 2005

Goal of Structure Prediction

- Epstein & Anfinsen, 1961:
  sequence uniquely determines structure

- INPUT: sequence
- OUTPUT:



3D structure and function

This is a Nobel Prize Winning Problem!!!

B. Rost, 2005

CASP – Olympics of Protein Structure Prediction

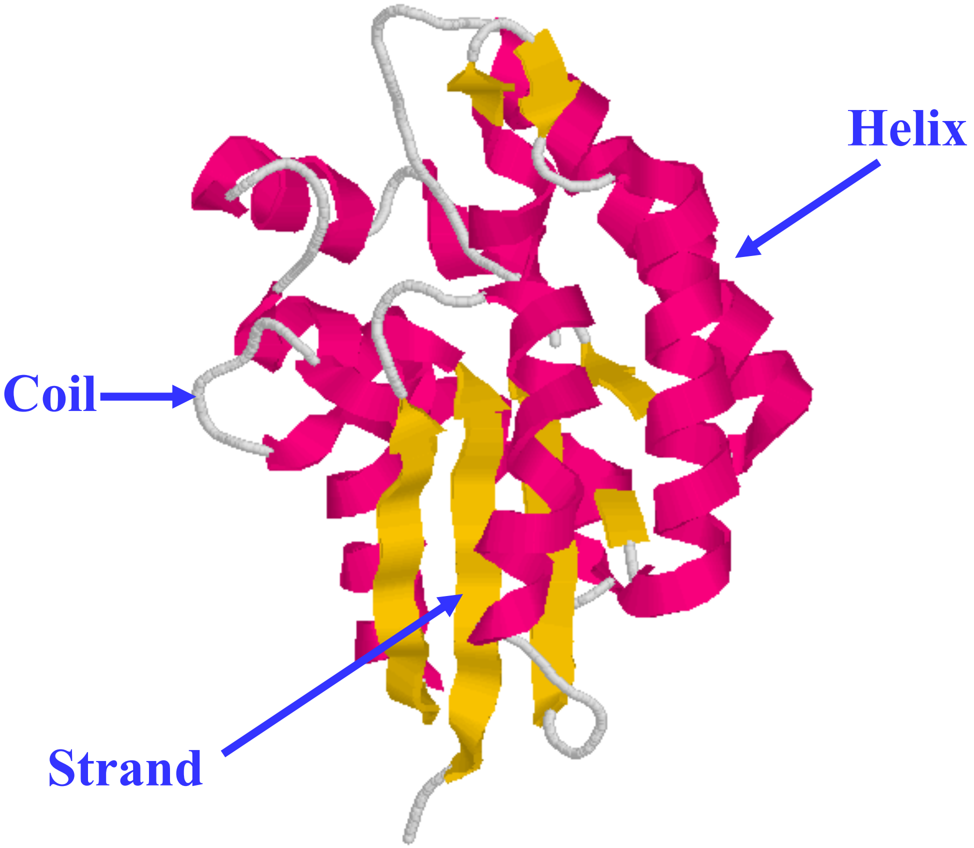
- Critical Assessment of Techniques of Protein Structure Prediction
- 1994,1996,1998,2000,2002,2004,2006, 2008, 2010, 2012, 2014, 2016, 2018
- Blind Test, Independent Evaluation

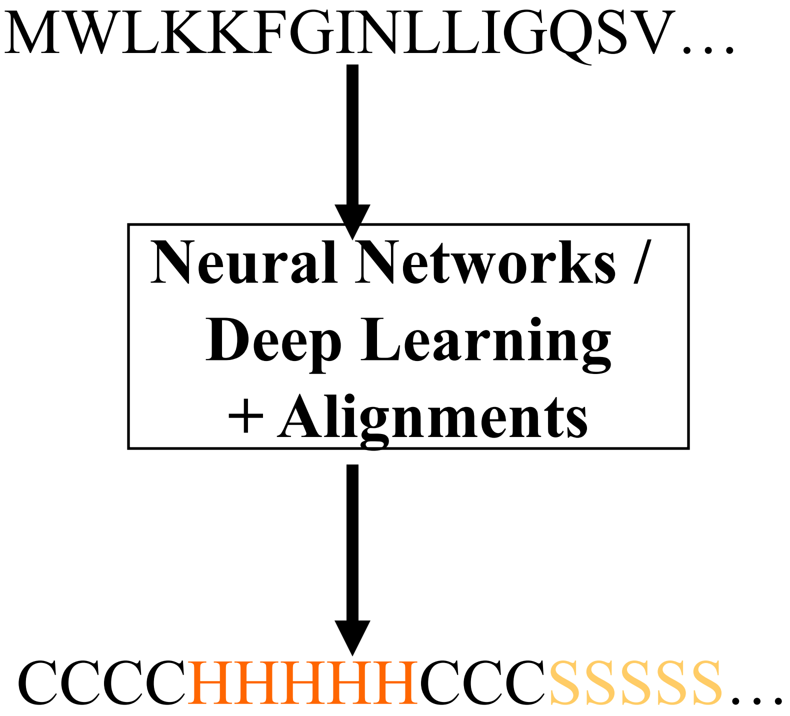
- CASP13 (http://predictioncenter.org/casp13/index.cgi)

CASP13 Demo

- http://predictioncenter.org/casp13/index.cgi

1D: Secondary Structure Prediction





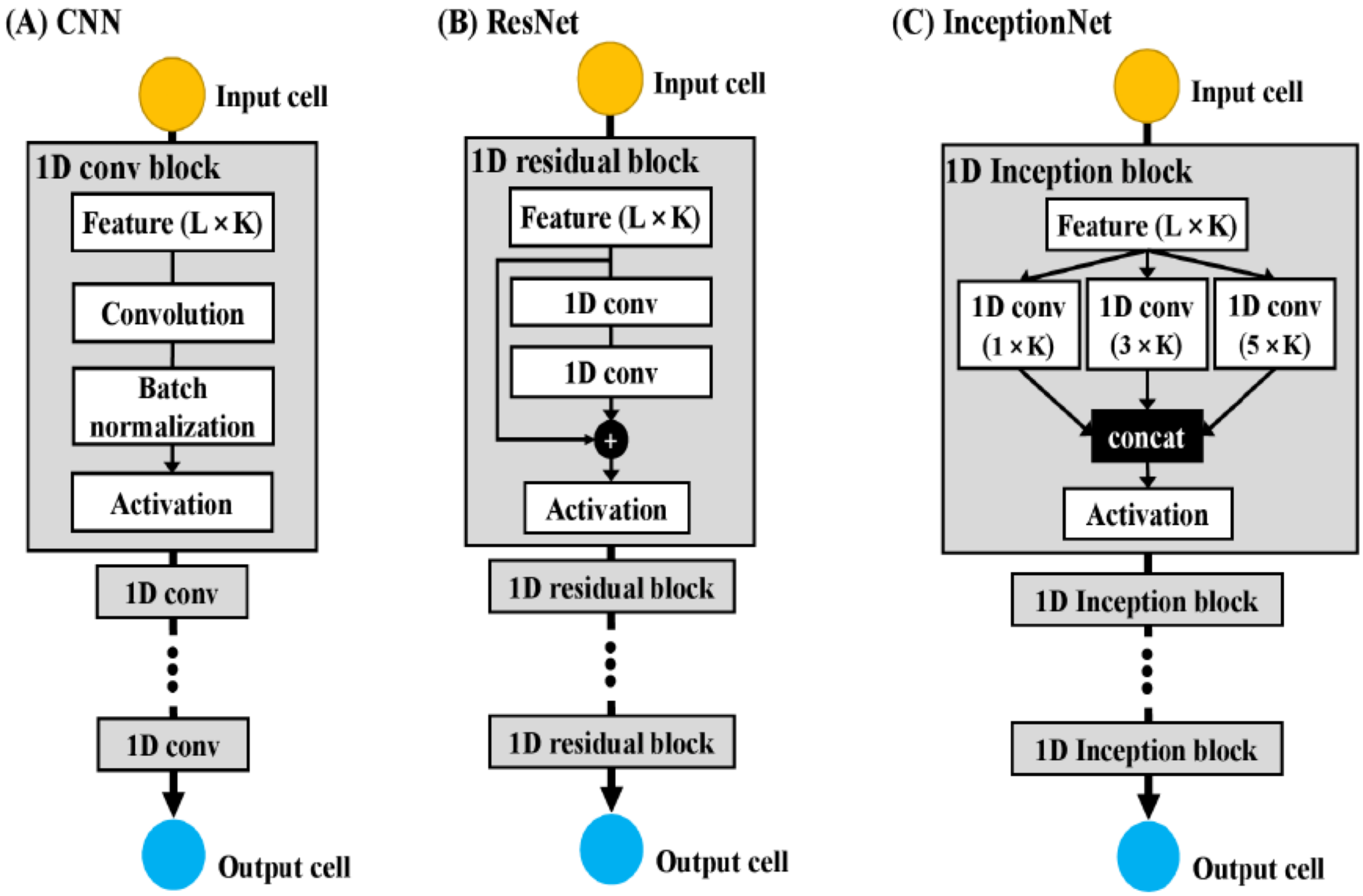
MWLKKFGINLLIGQSV…

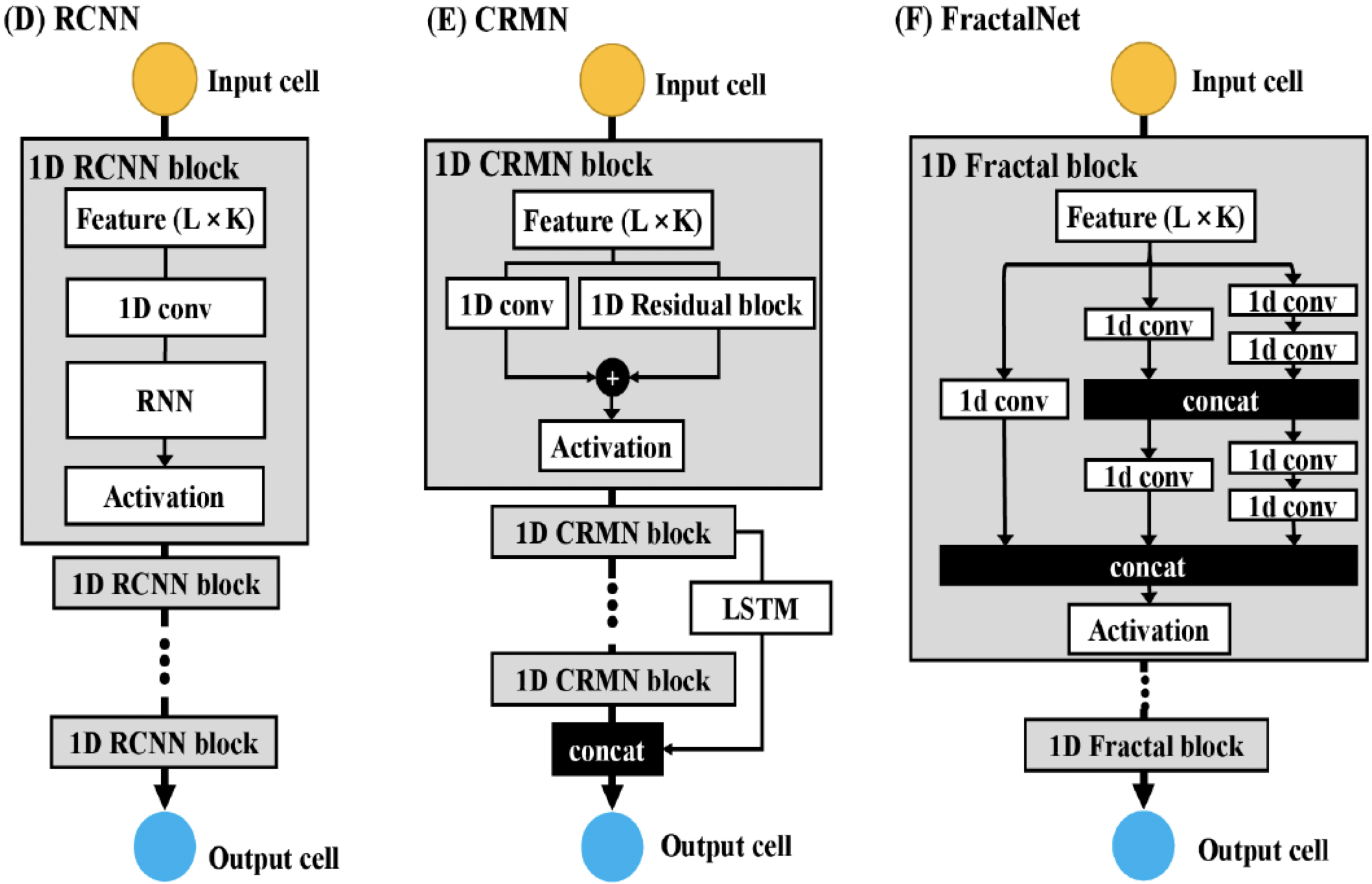
Neural Networks / Deep Learning + Alignments

CCCCHHHHHCCCSSSSS…

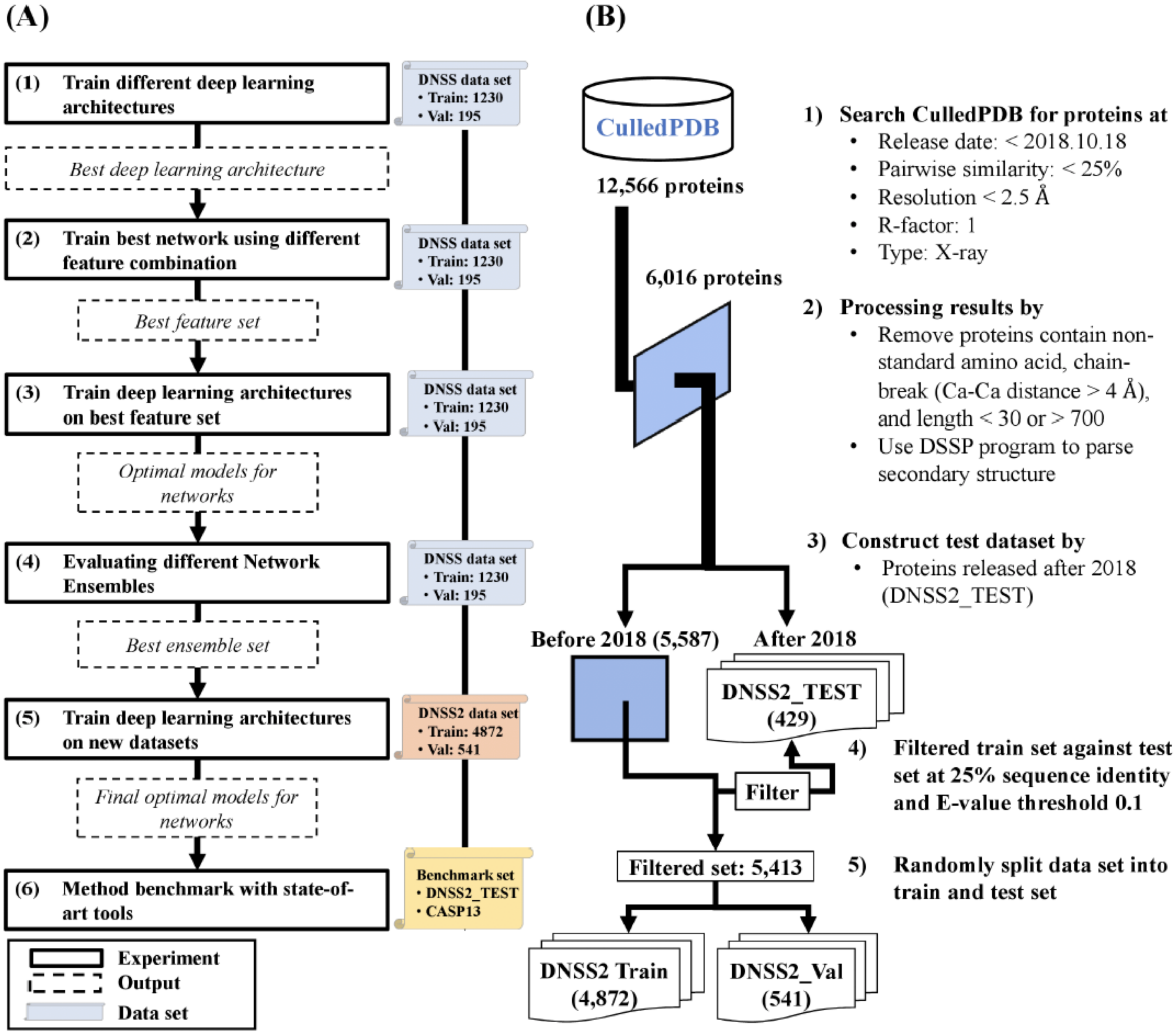
Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005

Deep Learning




(D) RCNN
Input cell
1D RCNN block
Feature (L × K)
1D conv
RNN
Activation
1D RCNN block
1D RCNN block
Output cell
(E) CRMN
Input cell
1D CRMN block
Feature (L × K)
1D conv
1D Residual block
+
Activation
1D CRMN block
LSTM
1D CRMN block
concat
Output cell
(F) FractalNet
Input cell
1D Fractal block
Feature (L × K)
1d conv
1d conv
1d conv
1d conv
concat
1d conv
1d conv
1d conv
concat
Activation
1D Fractal block
Output cell

Machine Learning Workflow



(A)

(1) Train different deep learning architectures

DNSS data set
- Train: 1230
- Val: 195

Best deep learning architecture

(2) Train best network using different feature combination

DNSS data set
- Train: 1230
- Val: 195

Best feature set

(3) Train deep learning architectures on best feature set

DNSS data set
- Train: 1230
- Val: 195

Optimal models for networks

(4) Evaluating different Network Ensembles

DNSS data set
- Train: 1230
- Val: 195

Best ensemble set

(5) Train deep learning architectures on new datasets

DNSS2 data set
- Train: 4872
- Val: 541

Final optimal models for networks

(6) Method benchmark with state-of-art tools

Benchmark set
- DNSS2_TEST
- CASP13

▭	Experiment
⬚	Output
▬	Data set

(B)

CulledPDB

12,566 proteins

6,016 proteins

Before 2018 (5,587)

After 2018

DNSS2_TEST (429)

Filter

Filtered set: 5,413

DNSS2 Train (4,872)

DNSS2_Val (541)

1) Search CulledPDB for proteins at
- Release date: < 2018.10.18
- Pairwise similarity: < 25%
- Resolution < 2.5 Å
- R-factor: 1
- Type: X-ray

2) Processing results by
- Remove proteins contain non-standard amino acid, chain-break (Ca-Ca distance > 4 Å), and length < 30 or > 700
- Use DSSP program to parse secondary structure

3) Construct test dataset by
- Proteins released after 2018 (DNSS2_TEST)

4) Filtered train set against test set at 25% sequence identity and E-value threshold 0.1

5) Randomly split data set into train and test set

Method	Q3(%)	Sov(%)
DNSS2_CNN	80.29	72.1
DNSS2_RCNN	81.83	73.97
DNSS2_ResNet	81.53	73.71
DNSS2_CRMN	81.91	73.37
DNSS2_FractalNet	82.02	73.8
DNSS2_InceptionNet	82.74	75.3
DNSS2	83.84	75.5

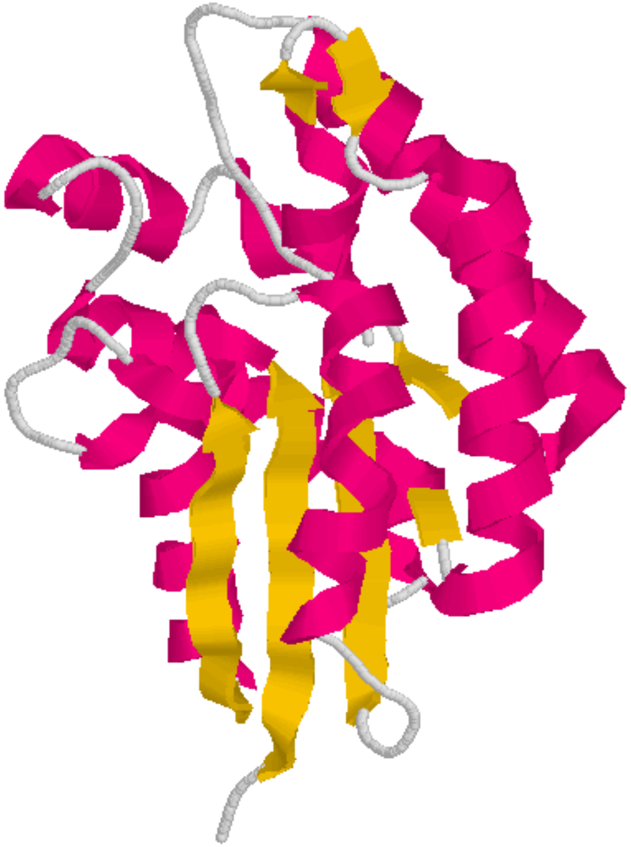
Table 3. Performance of the six different deep learning architectures (CNN, RCNN, ResNet, CRMN, FractalNet, and InceptionNet) and their ensemble (DNSS2) on DNSS1 validation dataset and the updated protein sequence database.

	All		TBM		FM	
Method	Q3 (%)	SOV (%)	Q3 (%)	SOV (%)	Q3 (%)	SOV (%)
SSPro5.2	76.73	69.94	78.16	71.32	76.12	70.88
PSSpred	78.8	67.85	81.32	72.11	76.99	64.55
MUFOLD	79.58	71.74	79.71	74.13	79.8	70.79
DeepCNF	80.24	69.5	82.34	73.68	78.36	65.55
PSIPRED	80.7	72	83.67	76.72	78.41	68.14
SPIDER3	81.73	74.39	84.84	78.31	78.89	71.1
Porter5	82.07	74.61	84.79	78.98	79.42	70.3
DNSS1	77.06	70.40	79.48	73.58	75.46	68.79
DNSS2	82.2	73.03	85.37	76.98	79.82	70.56

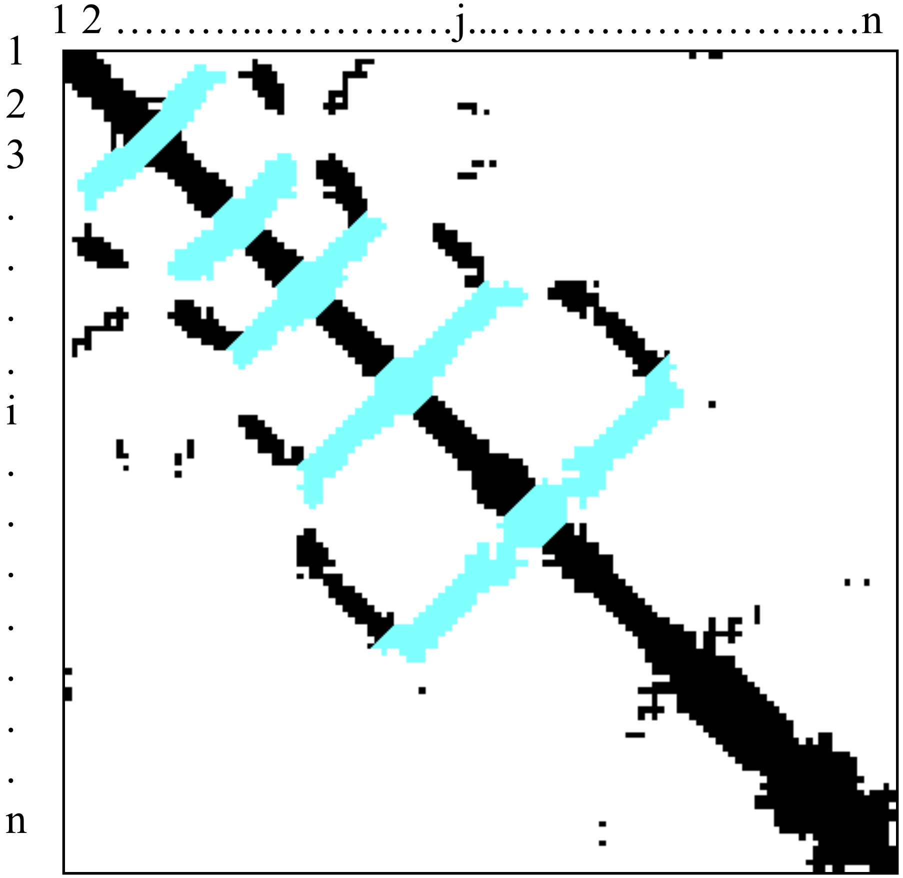
Table 5. Comparison of methods on the CASP13 dataset in terms of all CASP13 targets, template-based targets, and template-free targets.

2D: Contact Map Prediction

3D Structure

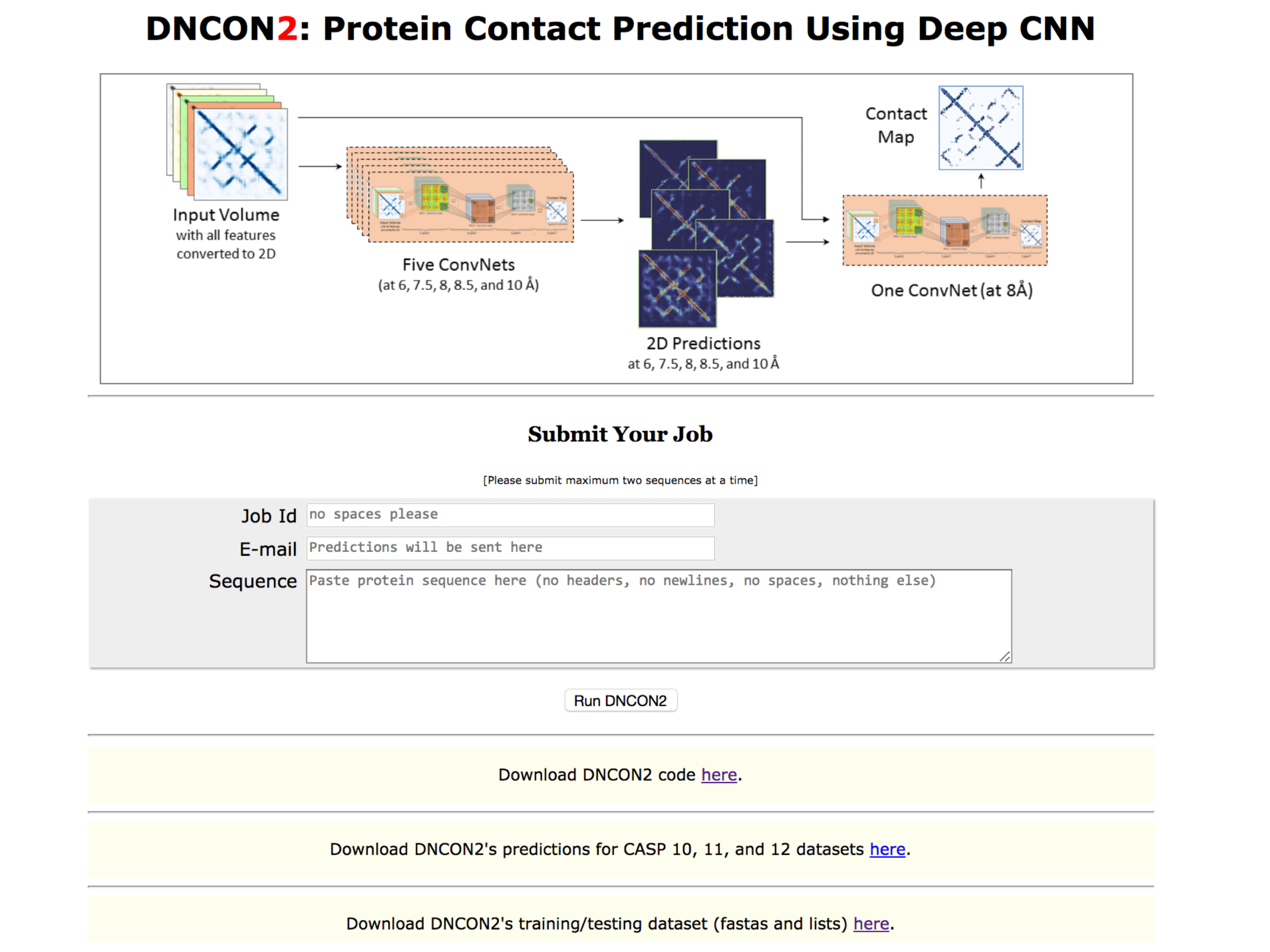


2D Contact Map



Distance Threshold = 8A°

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



DNCON2: Protein Contact Prediction Using Deep CNN

Submit Your Job

[Please submit maximum two sequences at a time]

Job Id: no spaces please

E-mail: Predictions will be sent here

Sequence: Paste protein sequence here (no headers, no newlines, no spaces, nothing else)

Download DNCON2 code here.

Download DNCON2's predictions for CASP 10, 11, and 12 datasets here.

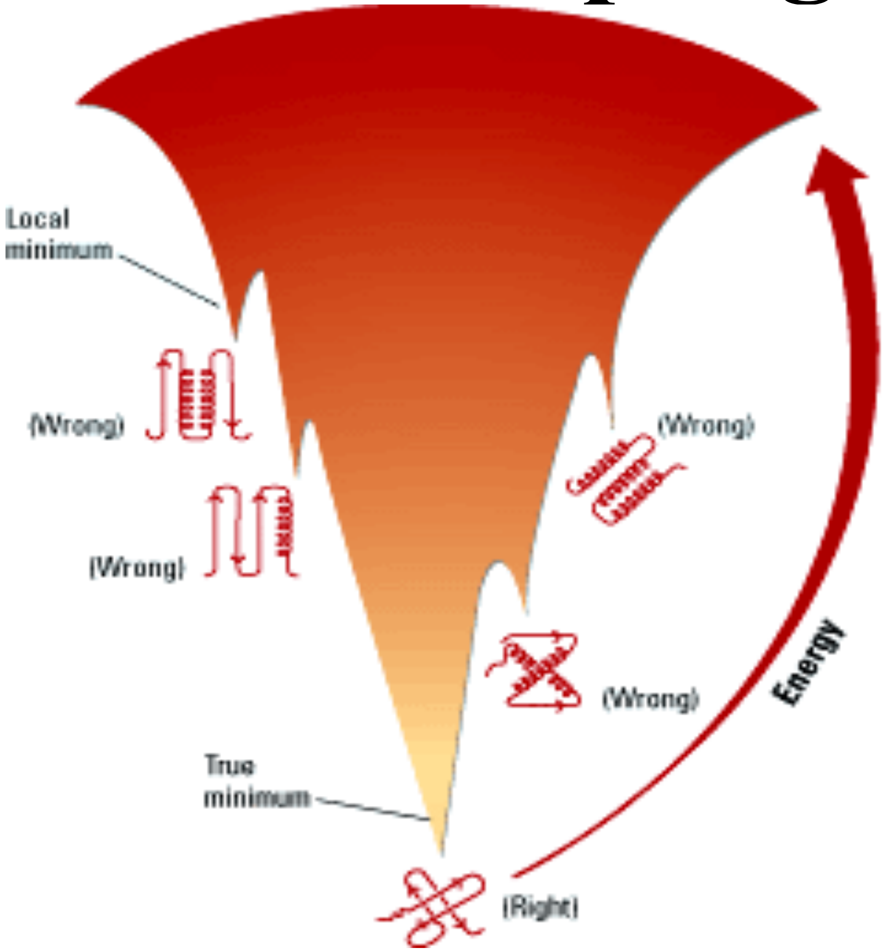
Download DNCON2's training/testing dataset (fastas and lists) here.

Contact Prediction

- PISCOV: http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/
- DNCON2: https://github.com/multicom-toolbox/DNCON2
- DeepCov https://github.com/psipred/DeepCov

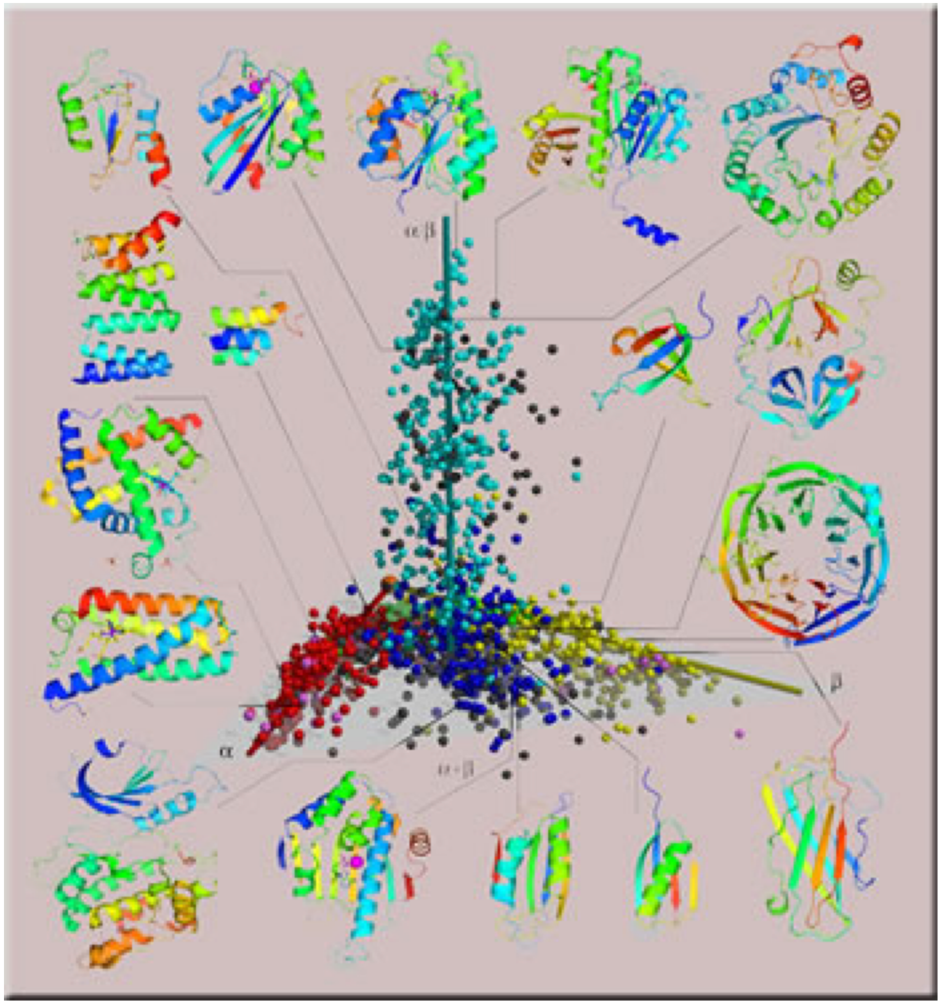
Protein tertiary structure prediction is a space sampling / simuation / optimization problem.

Protein Energy Landscape & Free Sampling



http://pubs.acs.org/subscribe/archive/mdd/v03/i09/html/willis.html

Protein Structure Space & Target Sampling

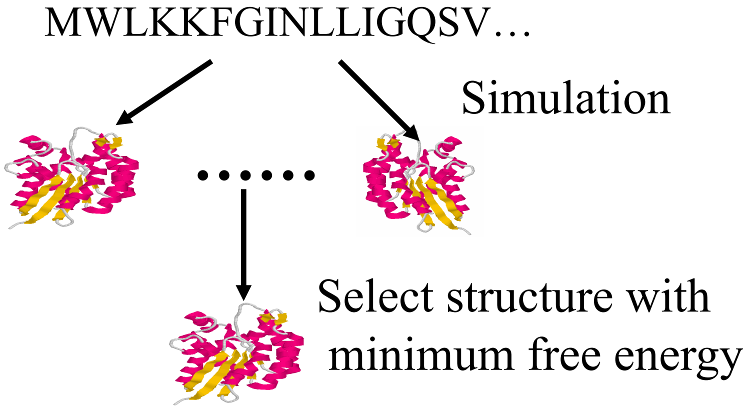


Two Approaches for 3D Structure Prediction

- **Ab Initio Structure Prediction**

Physical force field – protein folding

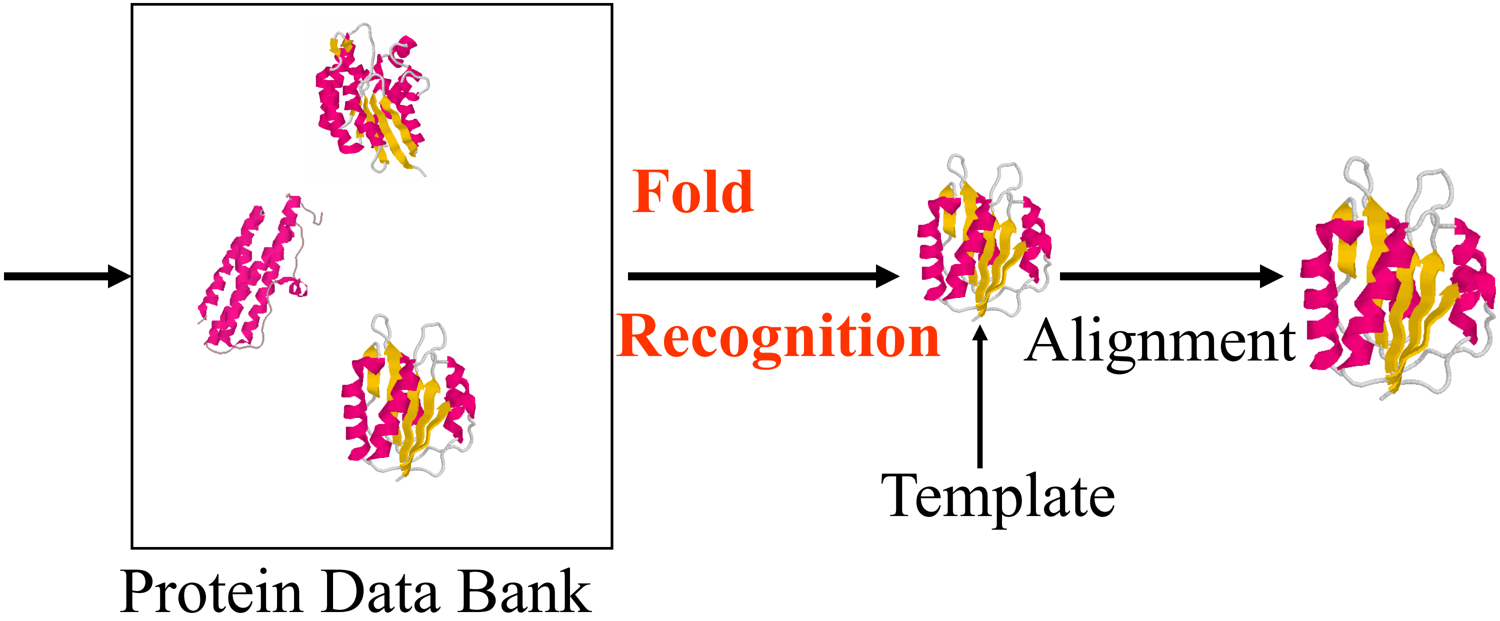
Contact/distance map - reconstruction



- **Template-Based Structure Prediction**

Query protein

MWLKKFGINKH…



Template-Based Structure Prediction ←→ KNN Learning

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

```
ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE
```

Copy
Loop Modeling
Optimization

How to find templates?
How to get alignments?

A. Fisher, 2005

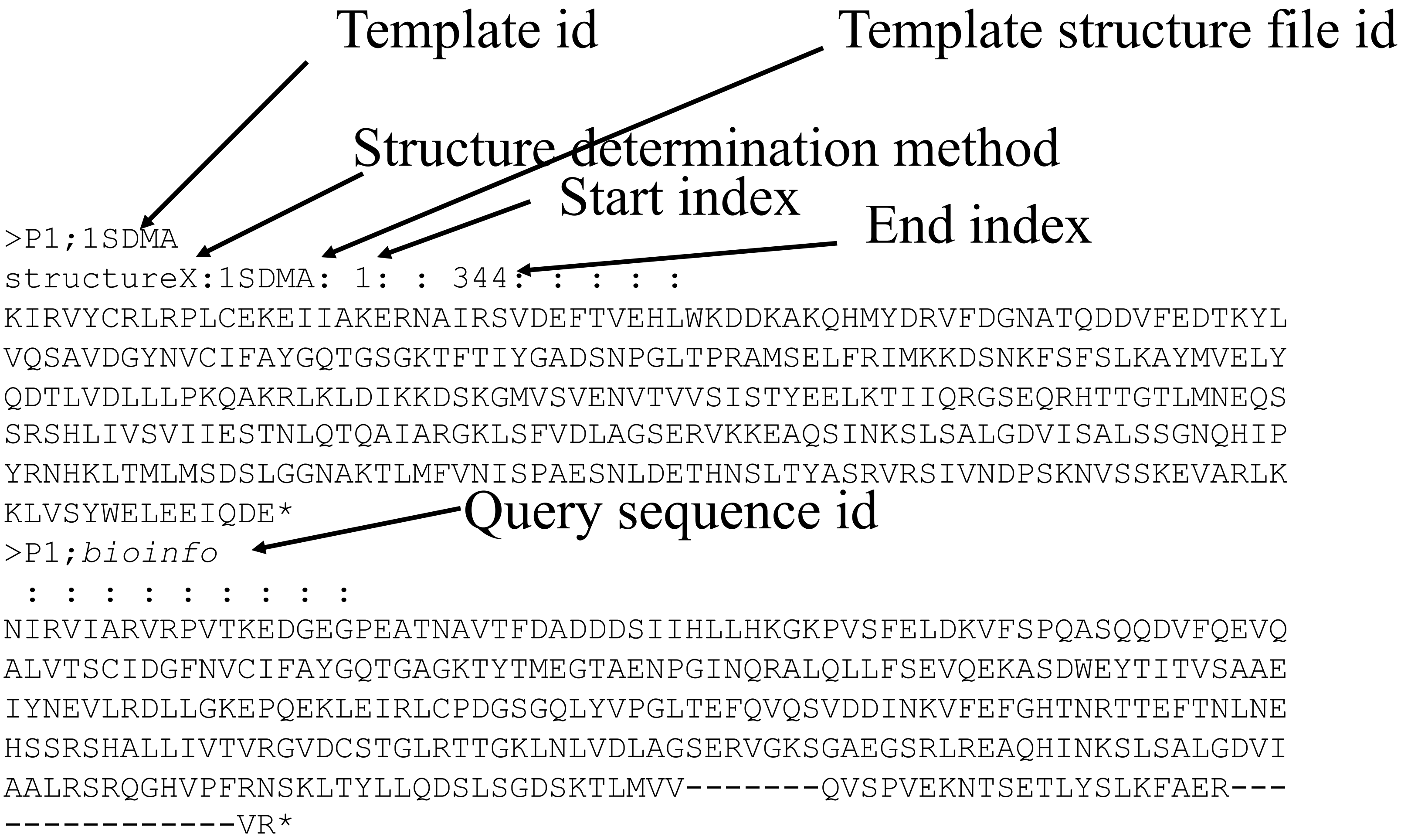
Modeller

- 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.

How to Get Templates and Alignments

- PSI-BLAST
- Hhblits
- Sequence/profile databases curated from the Protein Data Bank (PDB)

An PIR Alignment Example



Template id

Template structure file id

Structure determination method

Start index

End index

Query sequence id

```
>P1;1SDMA
structureX:1SDMA: 1: : 344: : : : :
KIRVYCRLRPLCEKEIIAKERNAIRSVDEFTVEHLWKDDKAKQHMYDRVFDGNATQDDVFEDTKYL
VQSAVDGYNVCIFAYGQTGSGKTFTIYGADSNPGLTPRAMSELFRIMKKDSNKFSFSLKAYMVELY
QDTLVDLLLPKQAKRLKLDIKKDSKGMVSVENVTVVSISTYEELKTIIQRGSEQRHTTGTLMNEQS
SRSHLIVSVIIESTNLQTQAIARGKLSFVDLAGSERVKKEAQSINKSLSALGDVISALSSGNQHIP
YRNHKLTMLMSDSLGGNAKTLMFVNISPAESNLDETHNSLTYASRVRSIVNDPSKNVSSKEVARLK
KLVSYWELEEIQDE*
>P1;bioinfo
 : : : : : : : : :
NIRVIARVRPVTKEDGEGPEATNAVTFDADDDSIIHLLHKGKPVSFELDKVFSPQASQQDVFQEVQ
ALVTSCIDGFNVCIFAYGQTGAGKTYTMEGTAENPGINQRALQLLFSEVQEKASDWEYTITVSAAE
IYNEVLRDLLGKEPQEKLEIRLCPDGSGQLYVPGLTEFQVQSVDDINKVFEFGHTNRTTEFTNLNE
HSSRSHALLIVTVRGVDCSTGLRTTGKLNLVDLAGSERVGKSGAEGSRLREAQHINKSLSALGDVI
AALRSRQGHVPFRNSKLTYLLQDSLSGDSKTLMVV-------QVSPVEKNTSETLYSLKFAER---
------------VR*
```

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

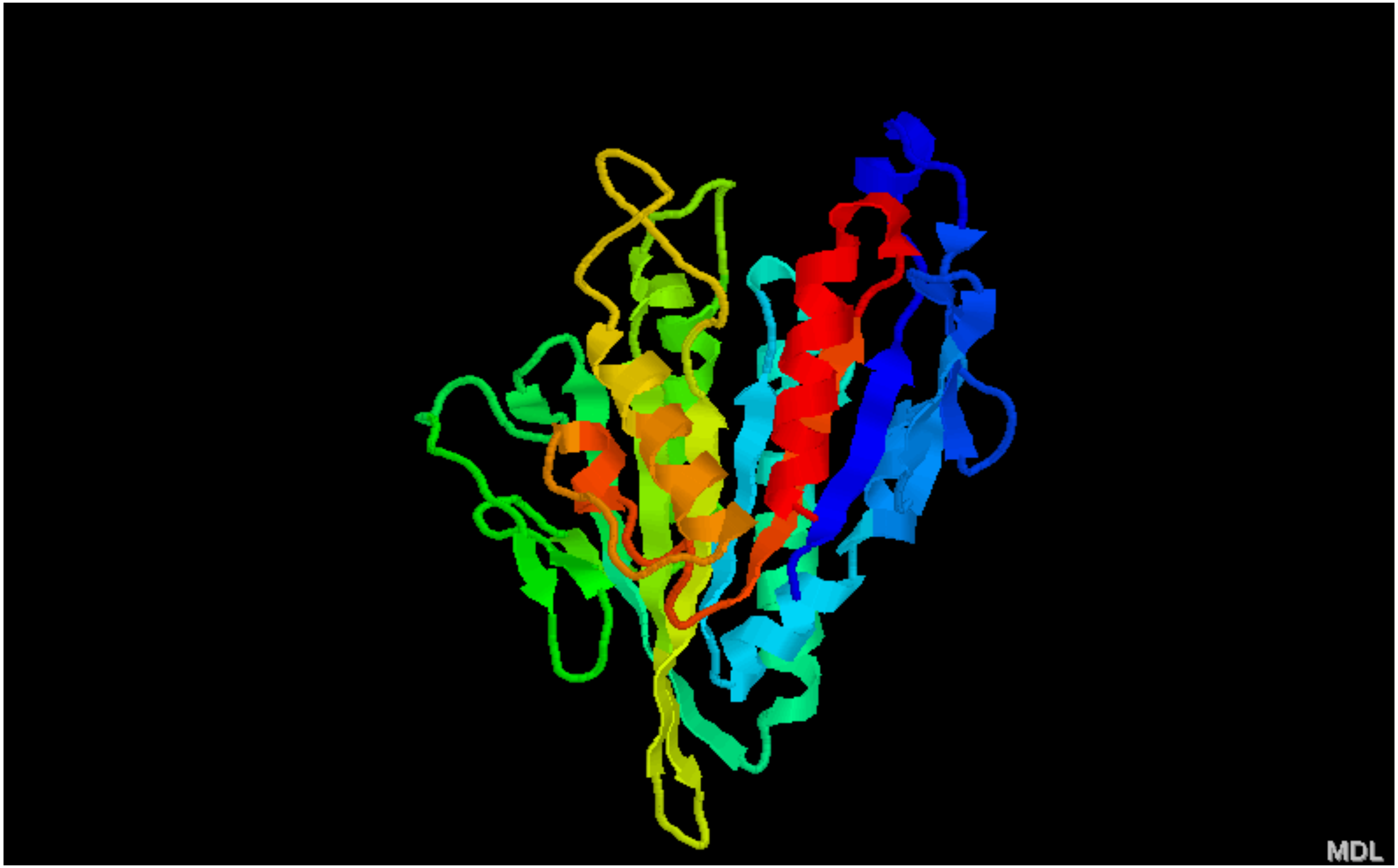
PIR alignment file name

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

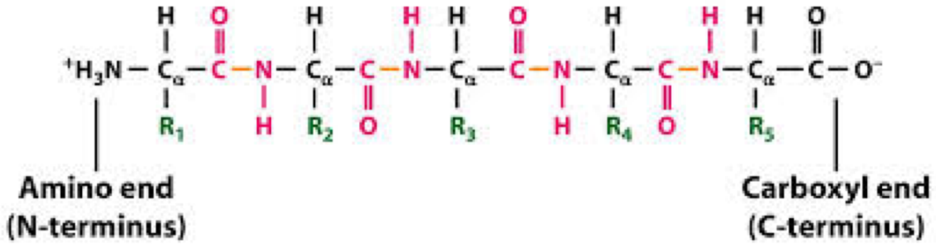


Figure 3-3b
Molecular Cell Biology, Sixth Edition
© 2008 W.H. Freeman and Company

- 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

- CASP12/CASP13 TBM targets
- Known templates at CASP12/CASP13 web sites
- Develop a homology-based algorithm / tool to build models from templates (gradient descent algorithm preferred)
- Assess the quality of models
- Implement from scratch
- **Form your group**

Feature Restraints from Template Data

- Given the information (a distance between two amino acids) in template, what can we know about the target?
- Feature constraint is represented as conditional distribution. E.g. P(ca-ca distance in target | ca-ca distance in template, residue type 1, residue type 2, …), P(psi angle of a residue in target | psi angle of an equivalent residue in template, …)

How to quantify the information? Function Fitting from Known Data - Learning

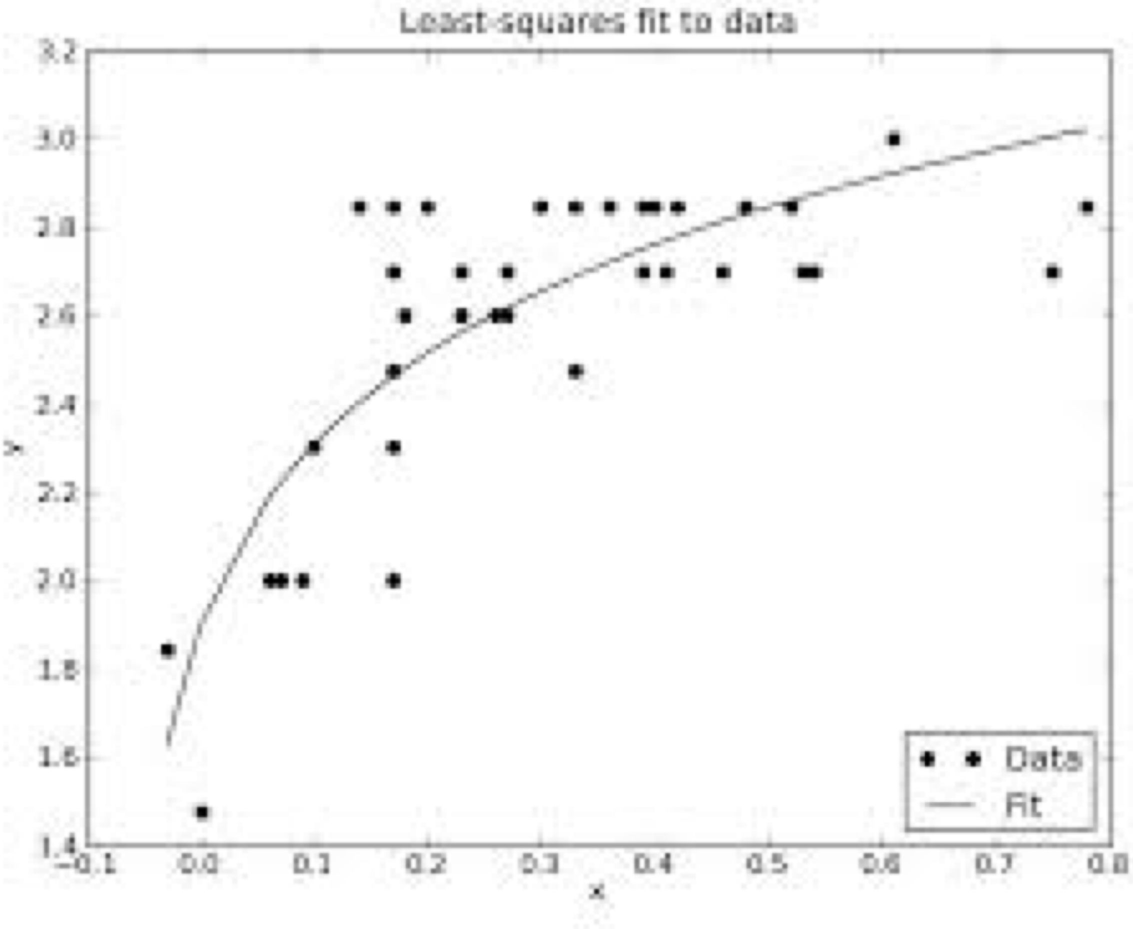
- A probability density function: P(y|x, a,b,c, …)
- Distribution form: normal distribution?
- Estimate the mean and standard deviation?
- Get some known data (template, target structures)
- Fitting algorithm: *Levenberg-Marquardt* algorithm for non-constrained least-squares fitting of a non-linear multidimensional model

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	9.5, 11, 10.8…
…	…	…	…

A database of 17 family alignments including 80 proteins was constructed to obtain feature statistics (**training/fitting**).

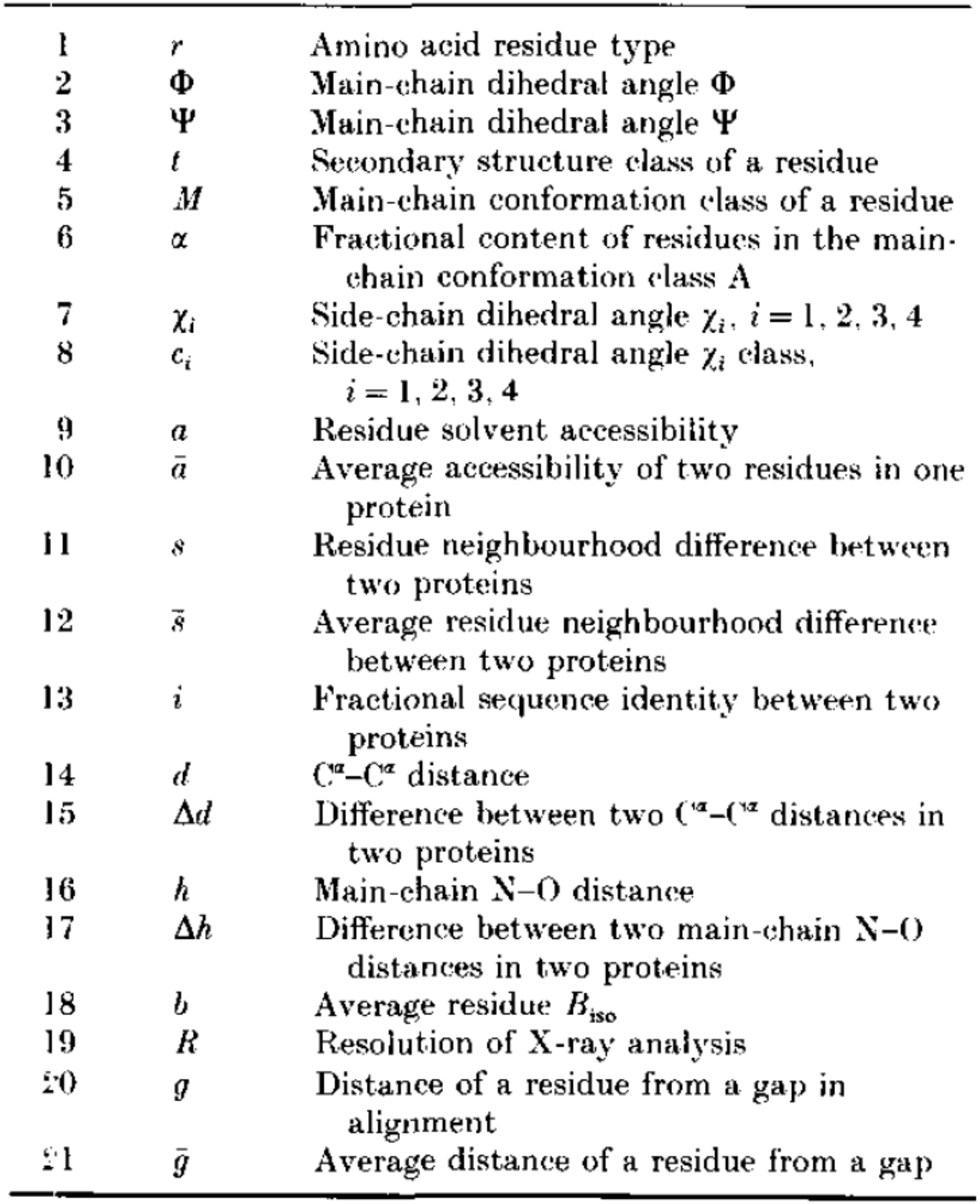
Levenberg-Marquardt algorithm



Calculate mean from the function

Estimate standard deviation

Commons Features



1	r	Amino acid residue type
2	Φ	Main-chain dihedral angle Φ
3	Ψ	Main-chain dihedral angle Ψ
4	t	Secondary structure class of a residue
5	M	Main-chain conformation class of a residue
6	α	Fractional content of residues in the main-chain conformation class A
7	χ_i	Side-chain dihedral angle χ_i, $i = 1, 2, 3, 4$
8	c_i	Side-chain dihedral angle χ_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^{α}–C^{α} distance
15	Δd	Difference between two C^{α}–C^{α} distances in two proteins
16	h	Main-chain N–O distance
17	Δh	Difference between two main-chain N–O distances in two proteins
18	b	Average residue B_{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

Side Chain & Main Chain

- Main-chain and side-chain modeling can be separated or carried out simultaneously
- Many tools model main chain first and then use SCWRL to add side chains in order to simplify the problem.
- All-atom modeling is more complex and time consuming, but can be more accurate sometime.

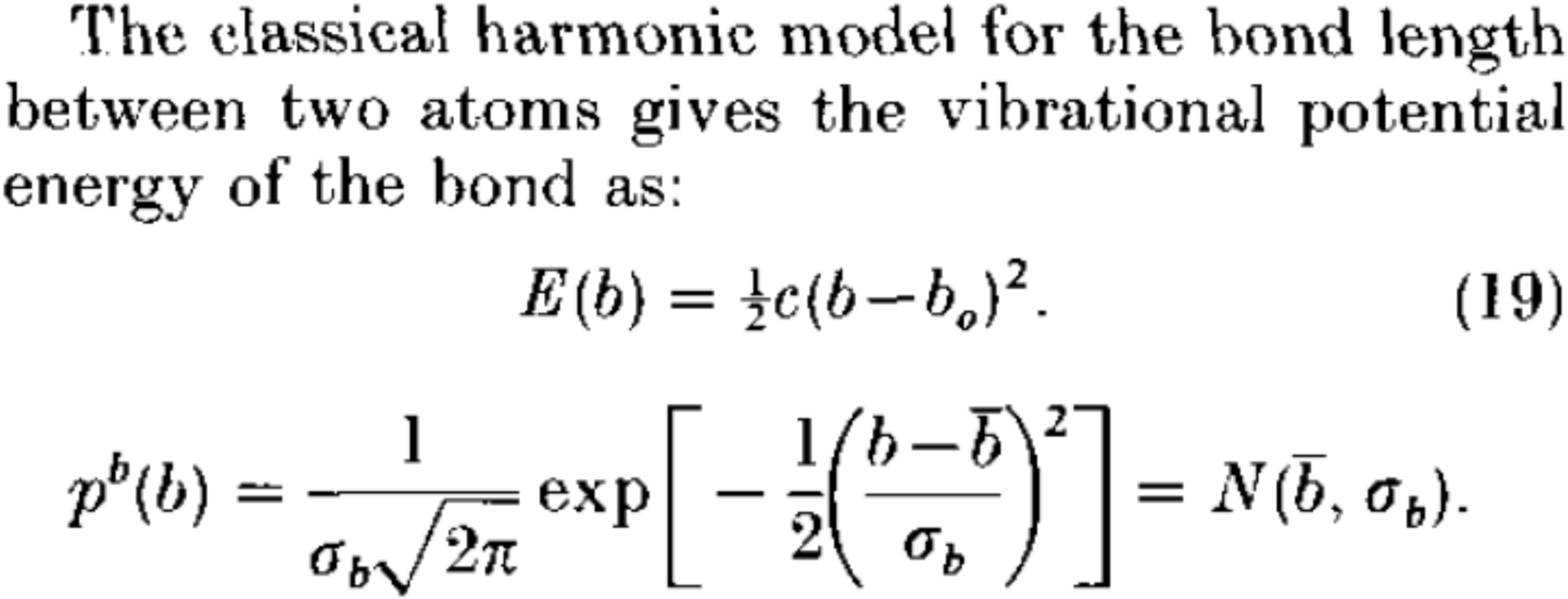
Usefulness of Features

- The most useful pdf is the one that predicts the unknown feature most accurately, measured by the entropy of a pdf.
- Two kinds of features: (1) generic features for all proteins and (2) features specific for the target protein

Stereochemical Restraints (Generic for any protein)

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

Bond Length and Angles (harmoic model)

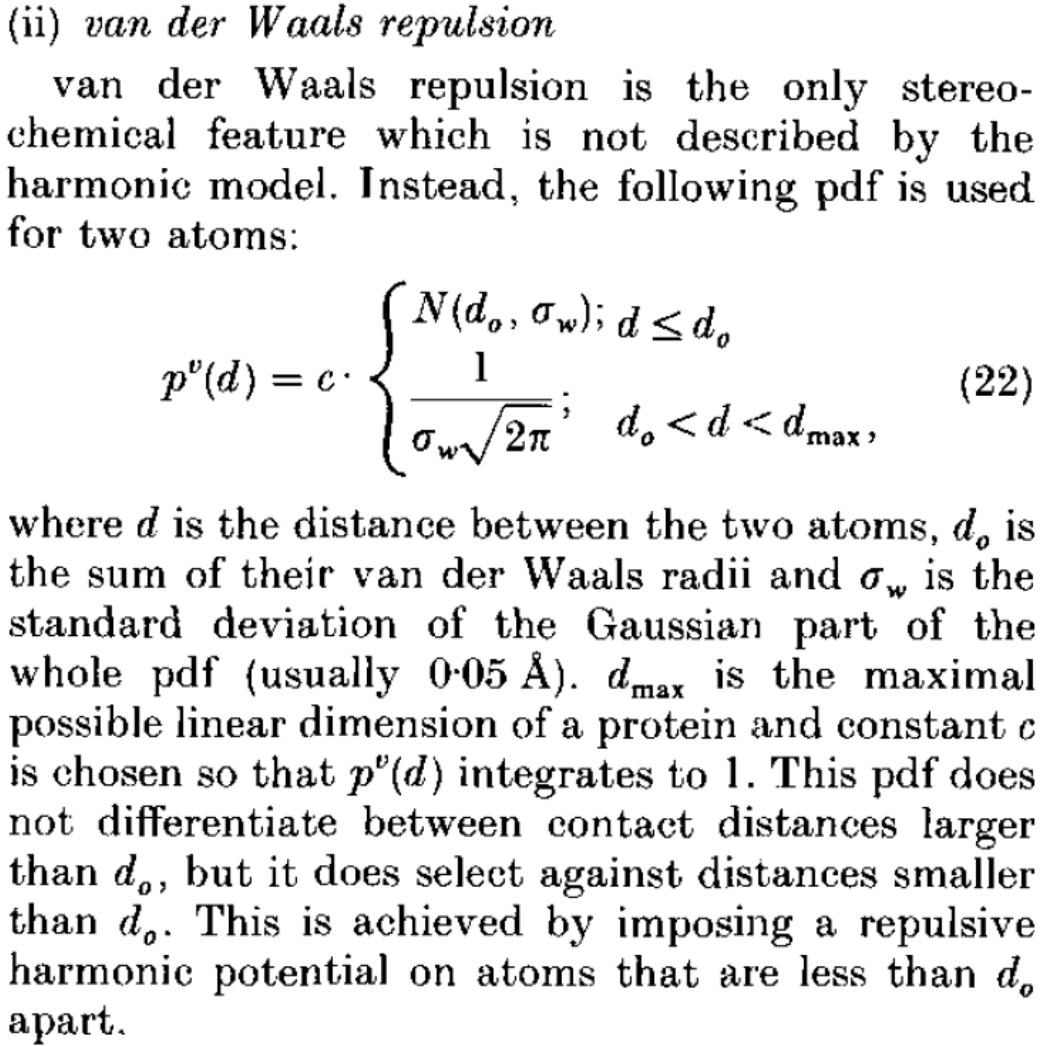


The classical harmonic model for the bond length between two atoms gives the vibrational potential energy of the bond as:

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

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

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:

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

where d is the distance between the two atoms, d_o is the sum of their van der Waals radii and σ_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.

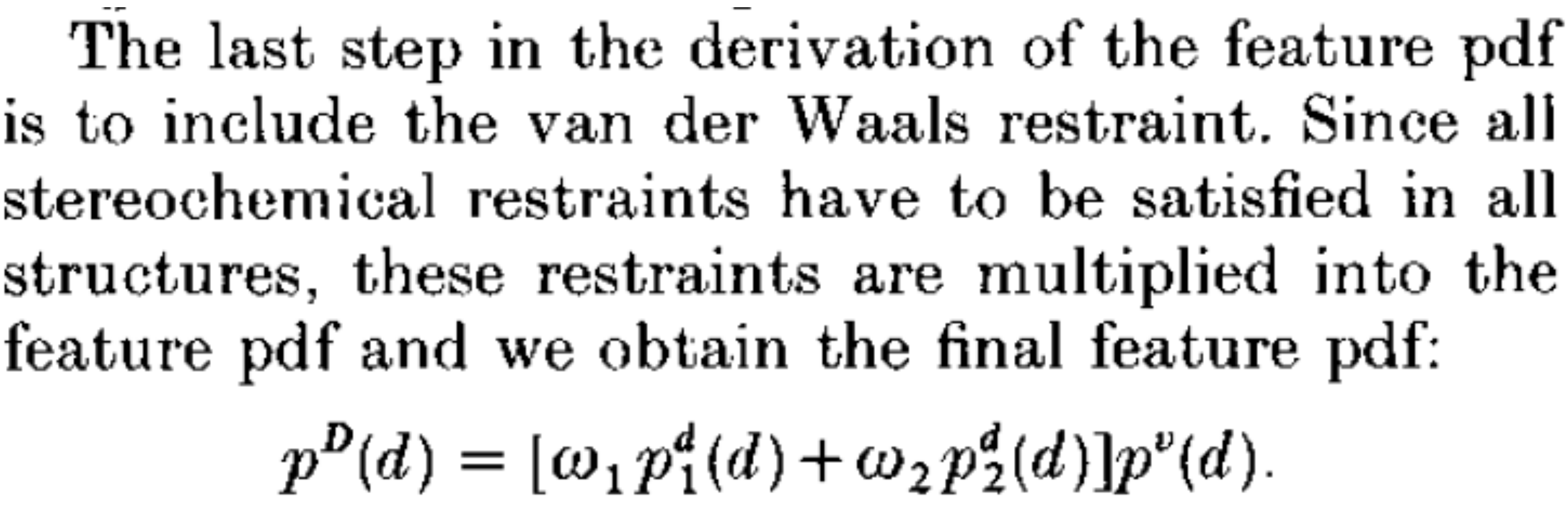
Ca-Ca Distance Features (protein specific)

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

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 individual feature pdfs

- Combine all feature pdfs into a molecular pdf $$P = \prod_i p^F(f_i). \tag{34}$$
- 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 or its logarithm.

Optimization

- Optimize the logarithm of molecular pdf – the objective function F.

$$F = -\ln(P), \qquad (35)$$

- 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.

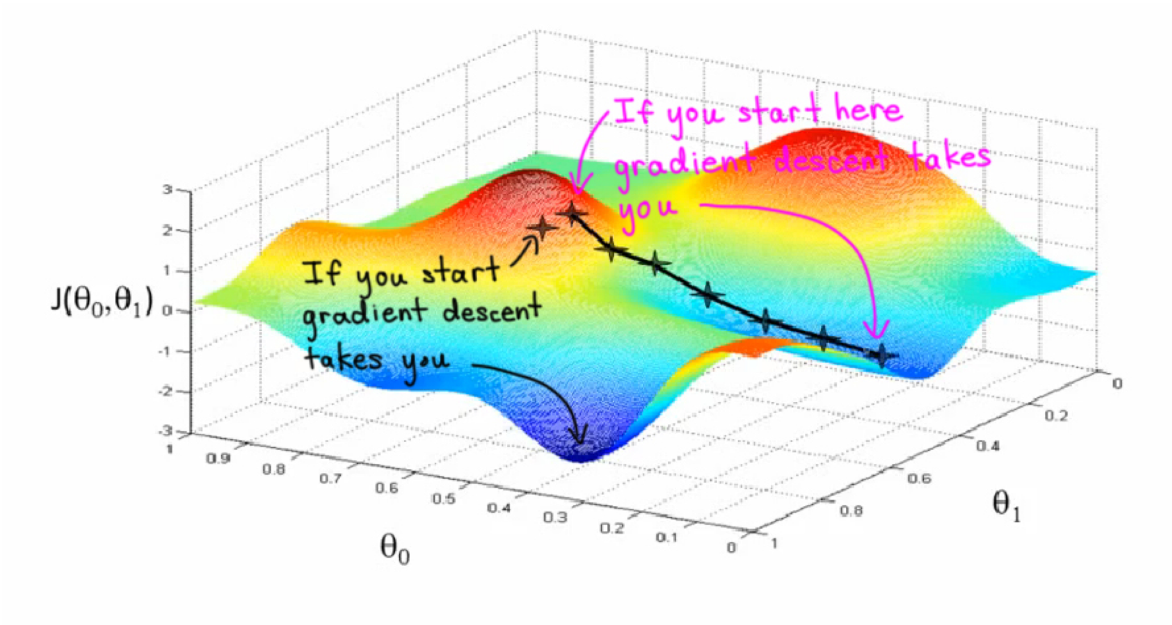
Successive Optimization

- 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 <=   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



Wikipedia

Gradient Descent

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

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

An Example - distance

- Probability of distance obeys normal distribution. -log(P)
- Square of distance error = *f* = (sqrt((x1-x2)^2 + (y1-y2)^2 + (z1-z2)^) – d0)^2
- $\frac{\partial f}{\partial x1}, \frac{\partial f}{\partial y1}, \frac{\partial f}{\partial z1}, \frac{\partial f}{\partial x2}, \frac{\partial f}{\partial y2}, \frac{\partial f}{\partial z2}$
- Partial derivative of angles is more complicated.

Gradient Descent

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

$$X_1^{t+1} = X_1^t - \eta*\Delta X \quad Y_1^{t+1} = Y_1^t - \eta*\Delta Y \quad Z_1^{t+1} = Z_1^t - \eta*\Delta Z$$

$$X_2^{t+1} = X_2^t - \eta*\Delta X \quad Y_2^{t+1} = Y_2^t - \eta*\Delta Y \quad Z_2^{t+1} = Z_2^t - \eta*\Delta Z$$

.

.

.

$$X_N^{t+1} = X_1^t - \eta*\Delta X \quad Y_N^{t+1} = Y_1^t - \eta*\Delta Y \quad Z_N^{t+1} = Z_1^t - \eta*\Delta Z$$

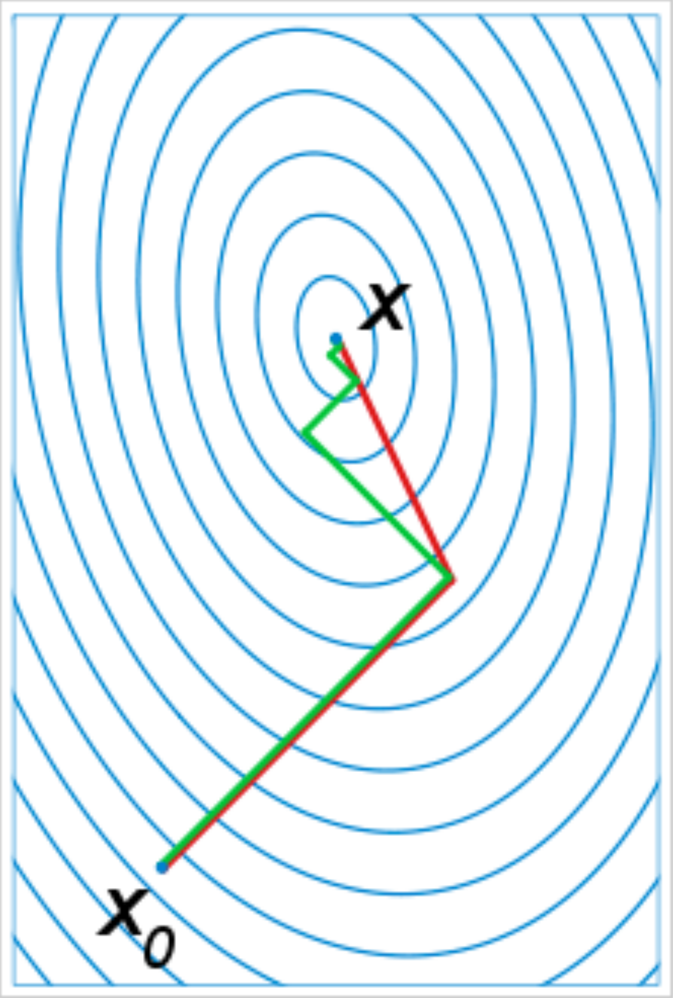
Trieu, Cheng, 2014

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[a]	Feature pdfs[b]	Violations[c]	r.m.s.[d]	r.m.s.[e]
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[f]	919	919	1 (20°)	3·40°	3·40°
van der Waals contacts[g]	531	531	0 (0·2 Å)	0·02 Å	0·02 Å
C^{α}-C^{α} 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_1 dihedral angles	528	176	5 (25°)	8·4°	16·8°
Side-chain χ_2 dihedral angles	264	103	3 (25°)	10·2°	13·0°
Side-chain χ_3 dihedral angles	92	32	2 (25°)	11·9°	48·1°
Side-chain χ_4 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°
cis-Peptides[h]	0	0			

Group Formation

- **Group 1**:
- **Group 2**:
- **Group 3**:
- **Group 4:**

Project 1

- Design and develop a template-based protein structure modeling tool
- Assess its performance on a few TBM targets used in CASP12 or CASP13 benchmark
- Reference programs: (see later slides)

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: a Rosetta function. Rosetta can also create model loops.

- ModLoop a web server for loop modeling based on Modeller
- Add side chains to main chain – SCWRL
- An open source template-based modeling tool - MTMG

Modeller

- https://salilab.org/modeller/
- A widely used, well-documented template-based modeling tool

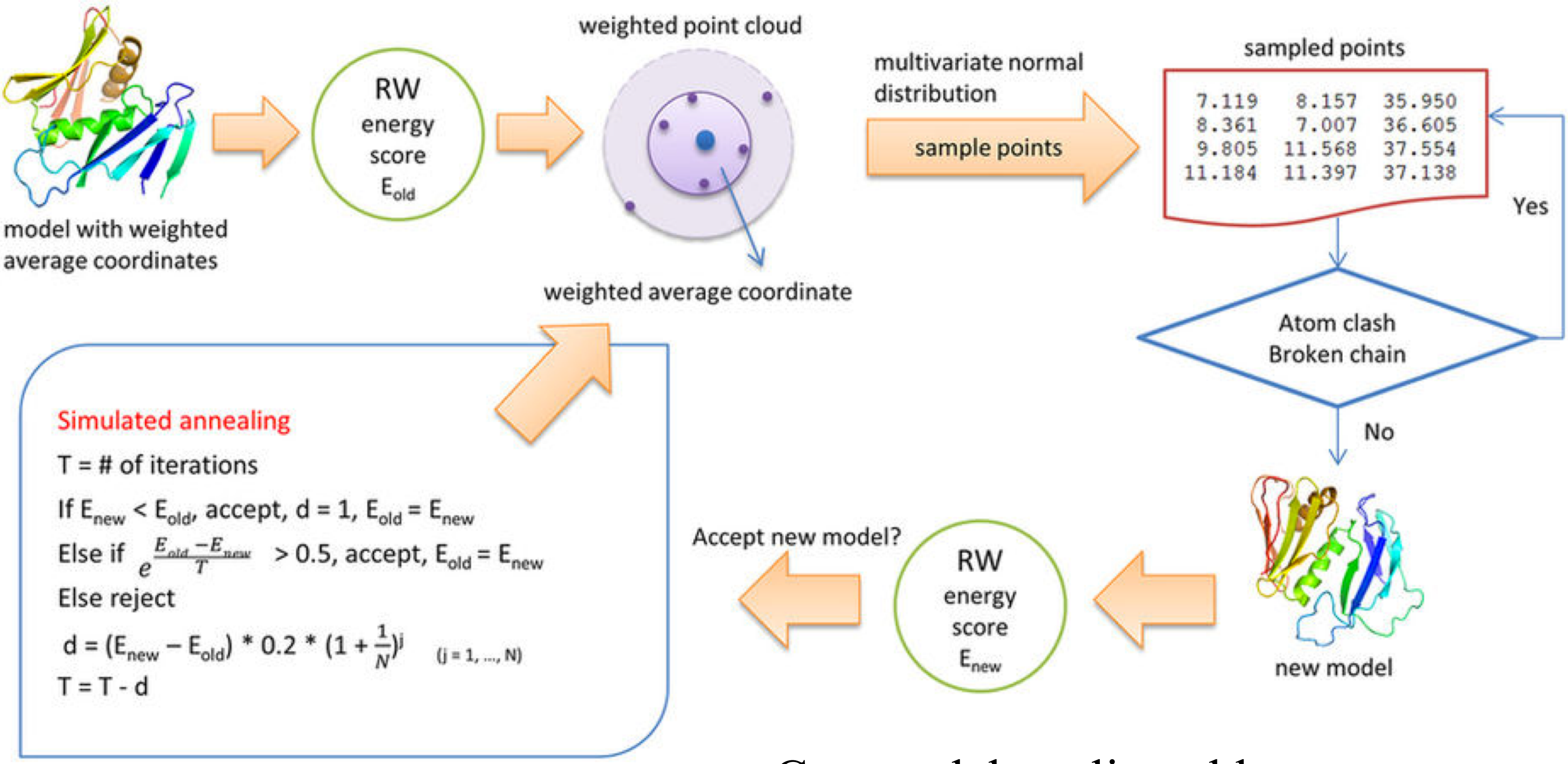
Integrative Modeling Platform

- IMP: https://integrativemodeling.org
- It implements all kinds of optimization methods including gradient descent. (you may refer to some source code there)

MTMG

- A stochastic point cloud sampling method for template-based protein comparative modeling. Scientific Reports, 2016.
- Source code is available: http://sysbio.rnet.missouri.edu/multicom_toolbox/tools.html

Workflow of MTMG


model with weighted average coordinates
RW energy score E_{old}
weighted point cloud
weighted average coordinate
multivariate normal distribution
sample points
sampled points
Atom clash
Broken chain
Yes
No
new model
RW energy score E_{new}
Accept new model?

```
 7.119    8.157   35.950
 8.361    7.007   36.605
 9.805   11.568   37.554
11.184   11.397   37.138
```

Simulated annealing

T = # of iterations

If $E_{new} < E_{old}$, accept, d = 1, $E_{old} = E_{new}$

Else if $e^{\frac{E_{old} - E_{new}}{T}} > 0.5$, accept, $E_{old} = E_{new}$

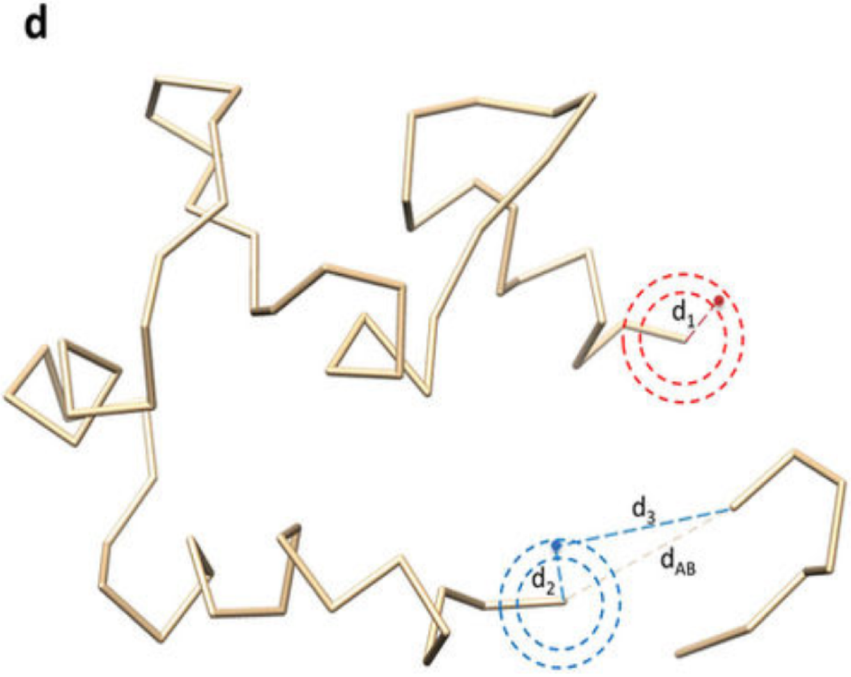
Else reject

$d = (E_{new} - E_{old}) * 0.2 * (1 + \frac{1}{N})^j$ (j = 1, ..., N)

T = T - d

Can model unaligned loops

Handle Gaps

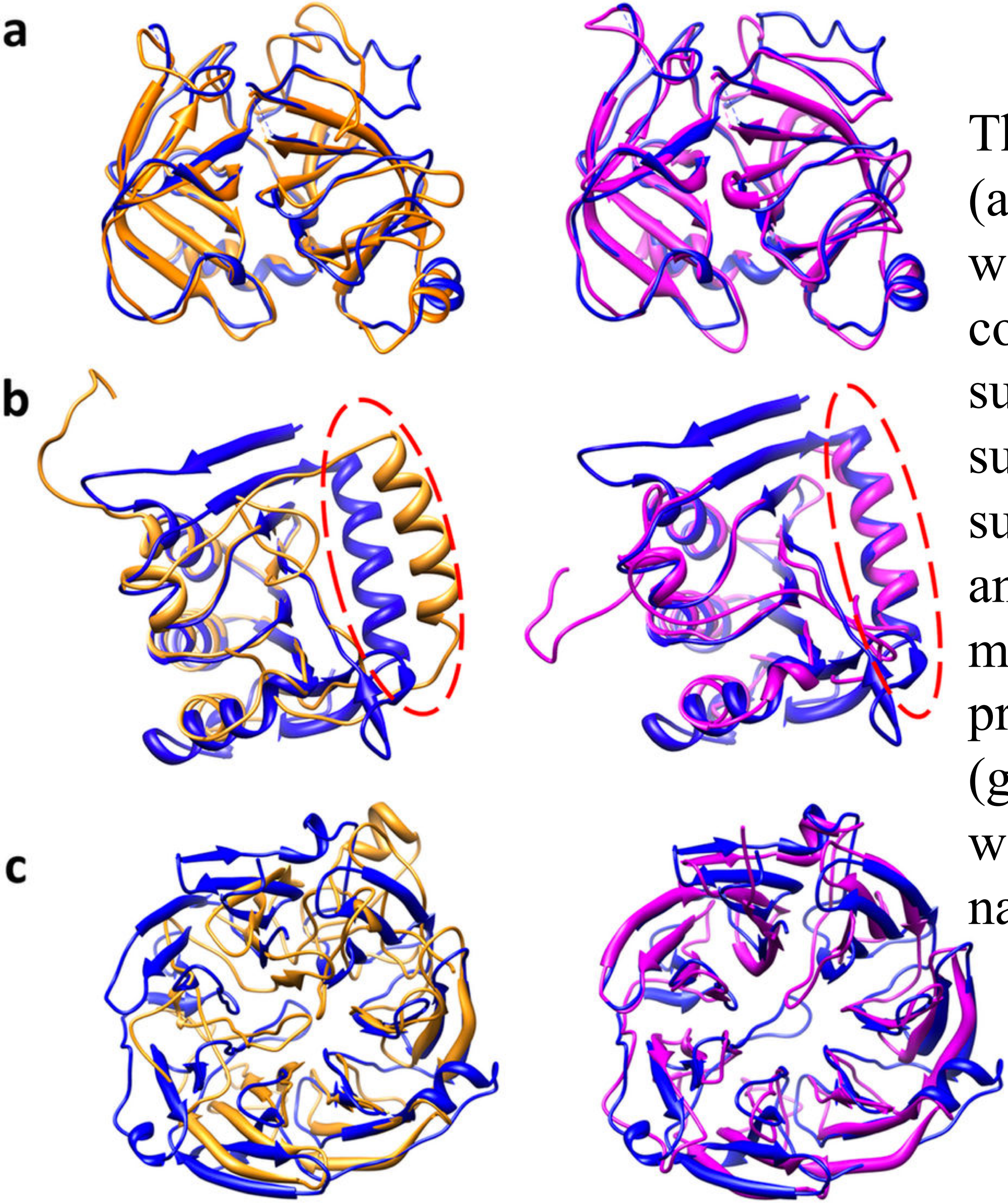


Sampling points for gaps. The radius of the outside circle is 4.5 Å, and the radius of the inner circle is 3.5 Å.

The sampling algorithm randomly samples point between the two circles. In the region circled by red, the gap is at the N-terminal.

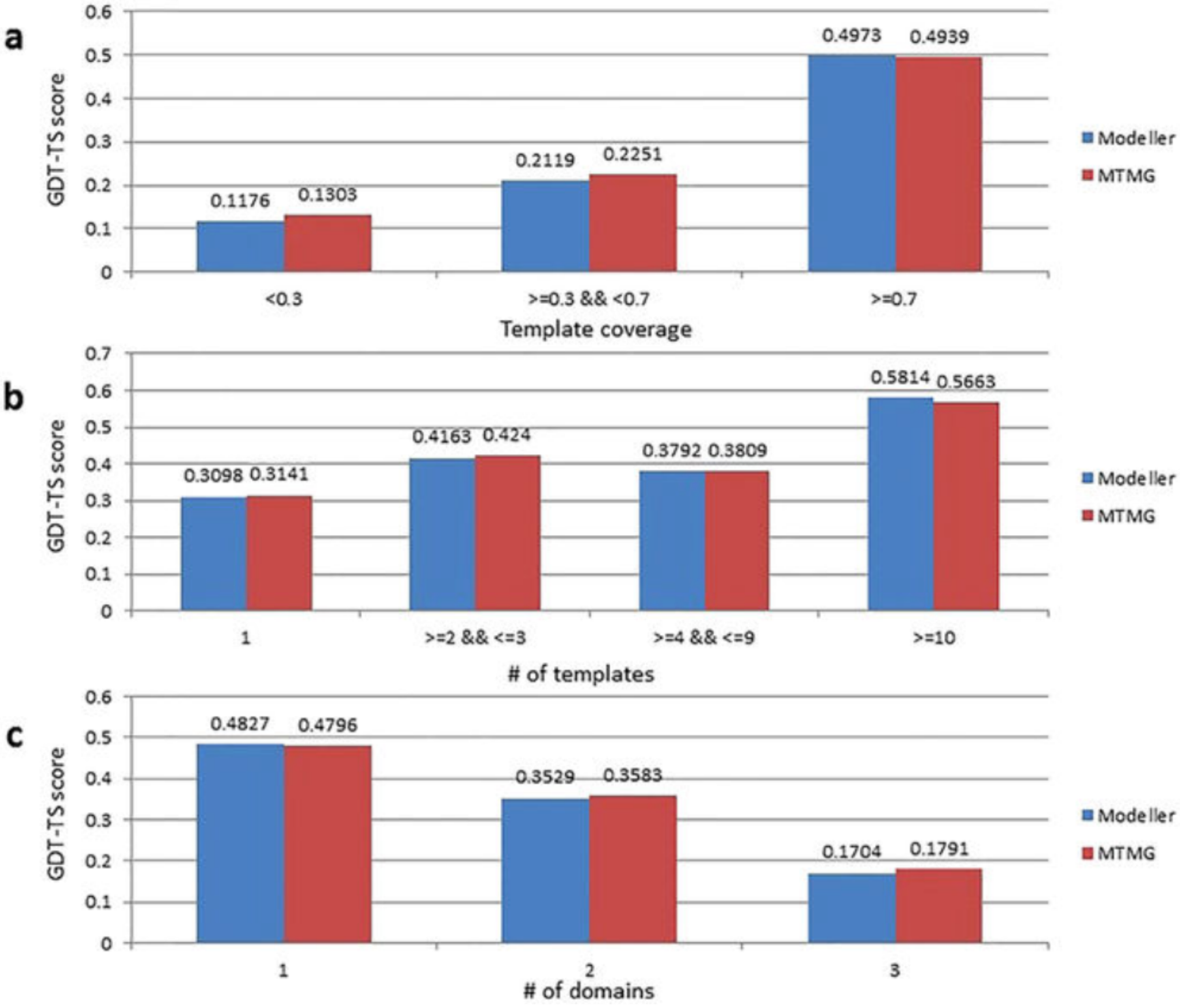
The distance d1 between an accepted sampled point and the first covered residue is between 3.5 Å and 4.5 Å.

In the region circled by blue, the three-residue gap is in the middle, and the distance between the two ends of the gap (dAB) is 8.2 Å. The distance d2 between an accepted sampled point and the last covered residue before the gap is between 3.5 Å and 4.5 Å. The distance d3 between an accepted sampled point and the first covered residue after the gap is between 4.1 Å and 11.4 Å.



Three examples illustrating (a) the successful template weighting and combination, (b) the successful template superposition, and (c) the successful domain division and combination of our method. The models predicted by Modeller (gold) and MTMG (purple) were superposed with the native structure (blue).

Figure 5: Comparison of GDT-TS score between the MTMG models and the Modeller models from three aspects on CASP11 targets.



(a) MTMG performed better than Modeller on targets with <0.7 template coverage. **(b)** MTMG performs better than Modeller on targets covered by <10 templates. **(c)** MTMG performs better than Modeller on targets containing multiple domains.

Key Milestones of Project 1

- Class discussion on Feb. 20
- Presentation of your plan on Feb. 25
- Presentation of your results on Mar. 6