Analysis and Prediction of Protein Structure

Jianlin Cheng, PhD

Department of Computer Science
Informatics Institute
University of Missouri, Columbia

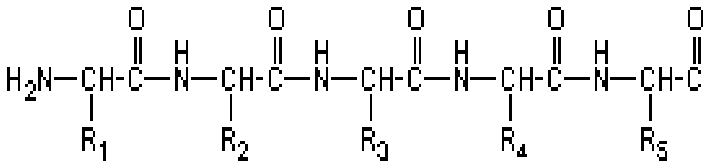
2011

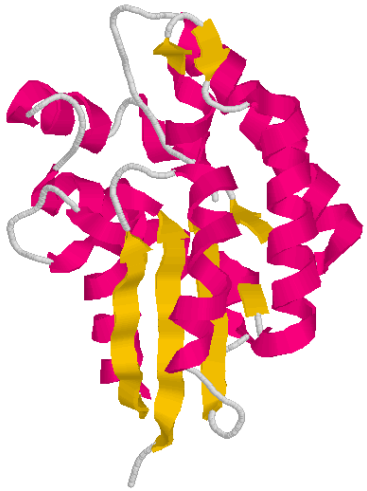
Outline

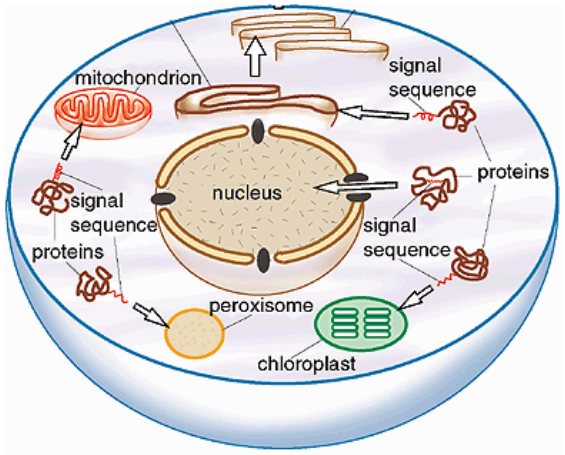
I. **Sequence, Structure, Function Relation**
II. Determination, Storage, Visualization
III. Structure Classification
IV. 1D Prediction
V. 2D Prediction
VI. 3D Prediction
VII. Useful Tools

Sequence, Structure and Function

AGCWY……

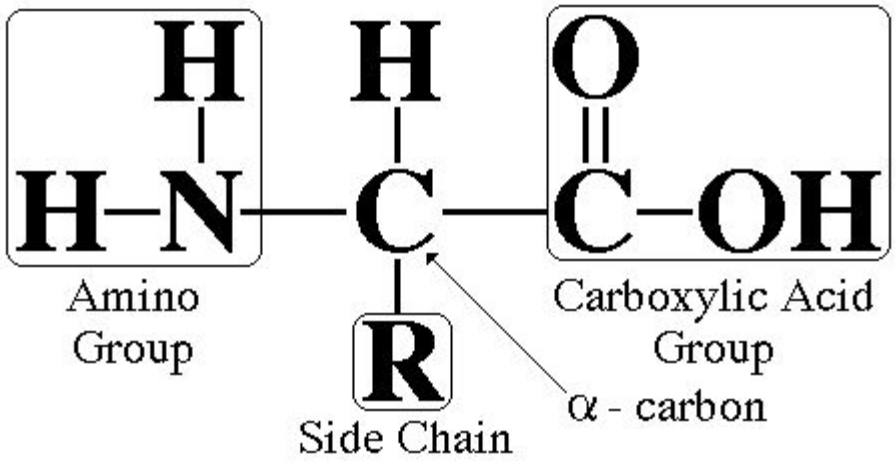






Cell

Amino Acid Structure



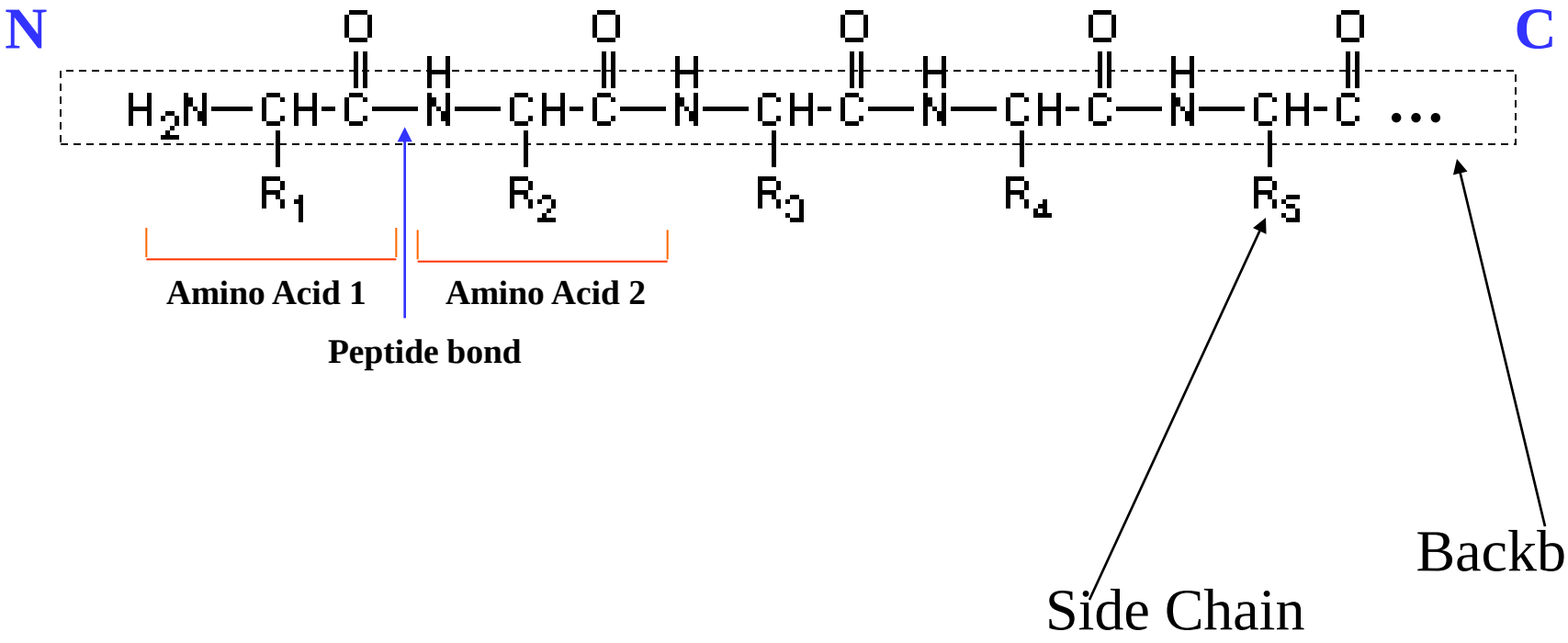
Amino Acids

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

Hydrophilic

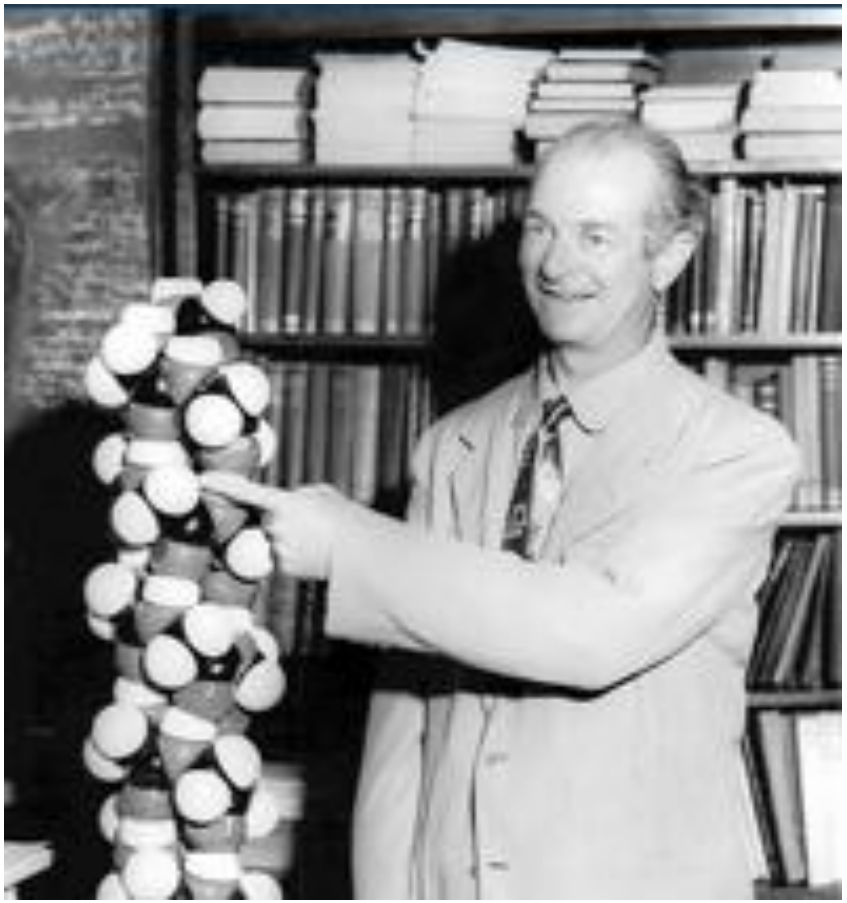
Protein Sequence

A directional sequence of amino acids/residues

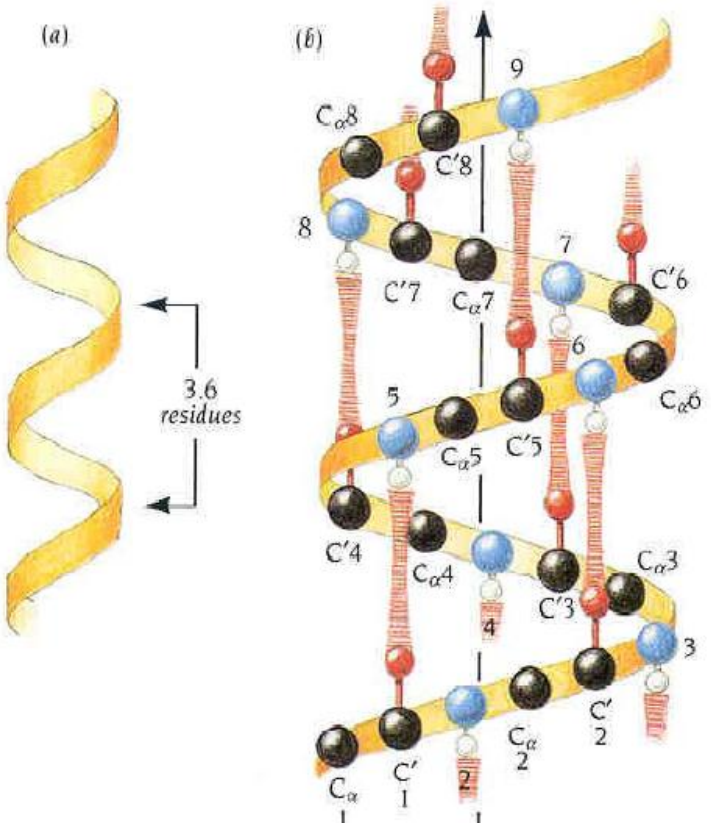


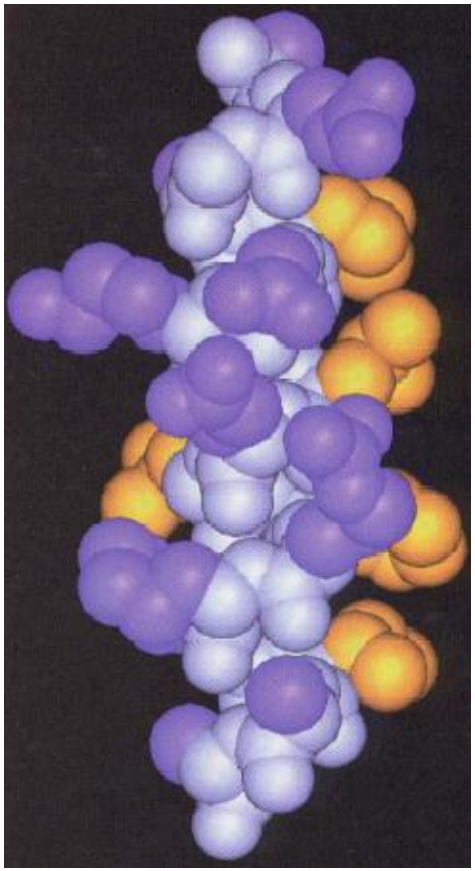
Protein Secondary Structure

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



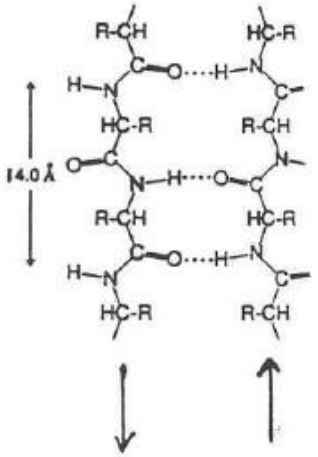
Alpha-Helix

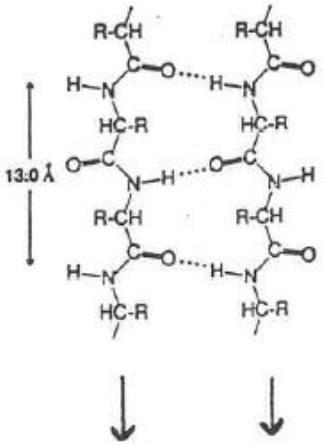




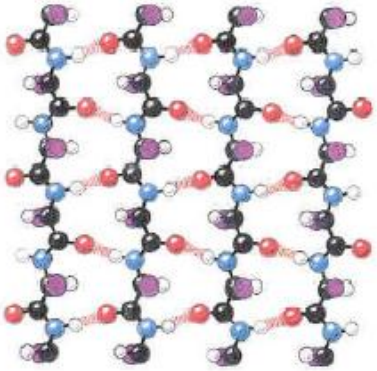
Jurnak, 2003

Beta-Sheet

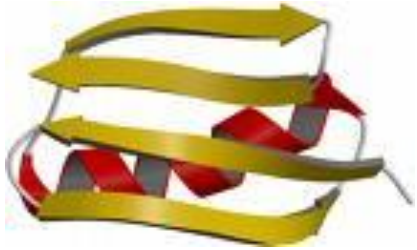




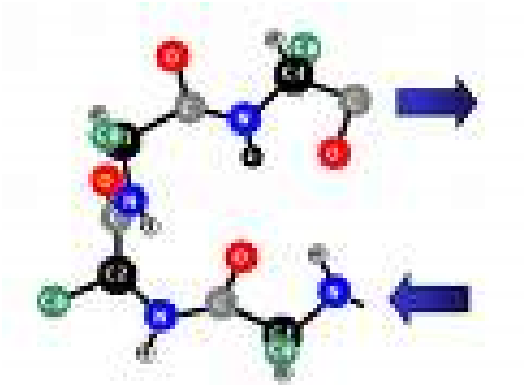
Anti-Parallel



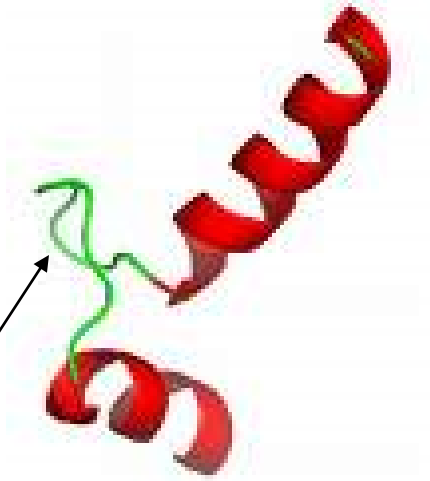
Parallel



Non-Repetitive Secondary Structure



Beta-Turn

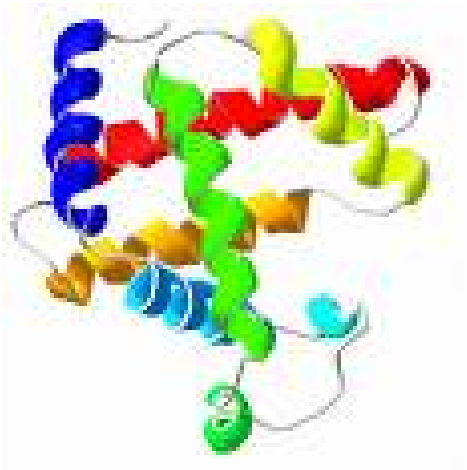


Loop

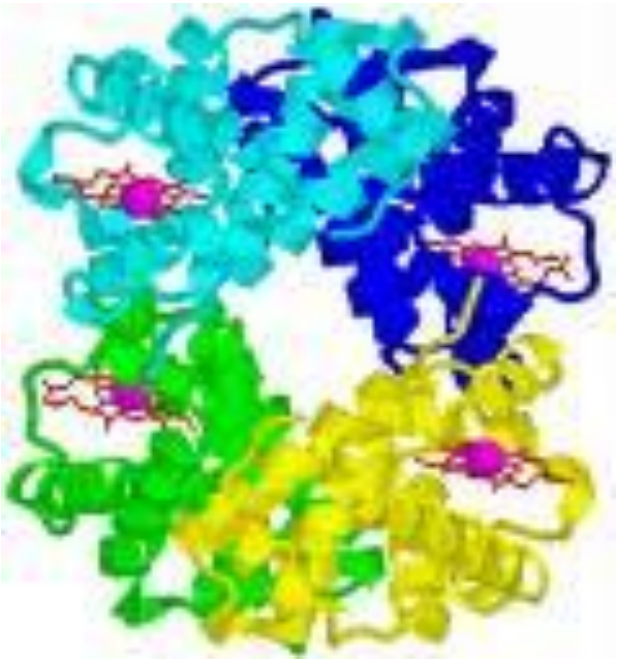
Tertiary Structure

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

Perutz Kendrew



myoglobin

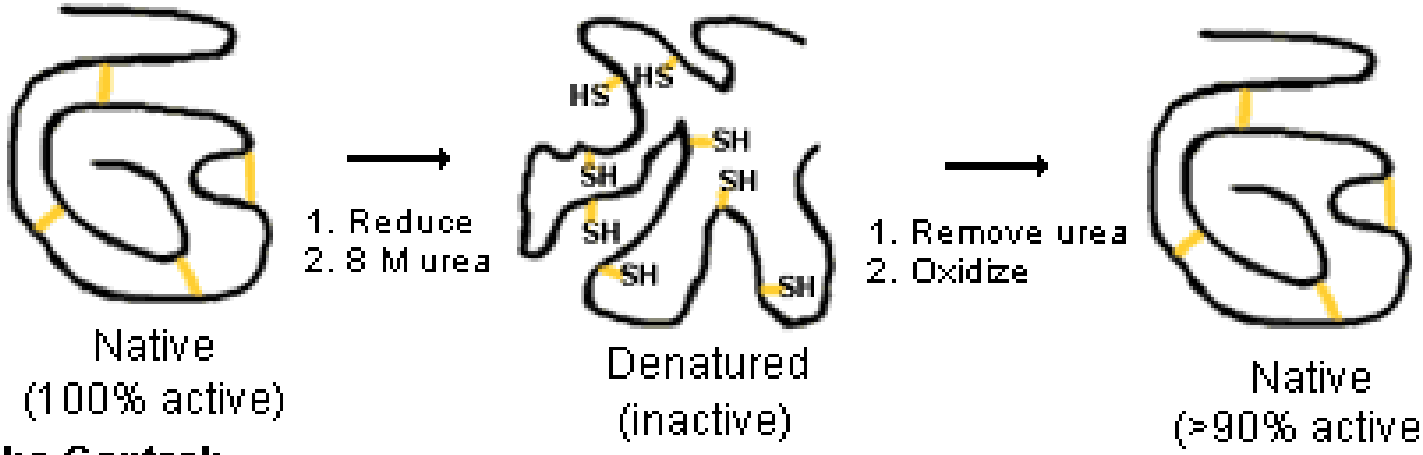


haemoglobin

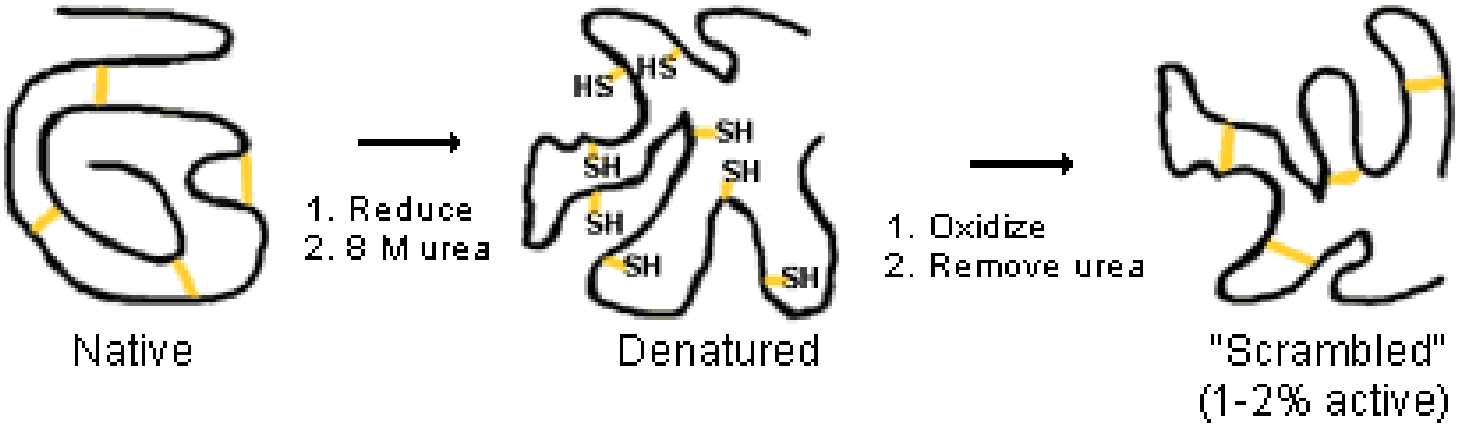
Anfinsen's Folding Experiment

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

The Observation:



The Control:



Protein Folding Movie

- http://www.youtube.com/watch?v=fvBO3TqJ6FE&feature=fvw (Demo)

Outline

I. Sequence, Structure, Function Relation

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

III. Structure Classification

IV. 1D Prediction

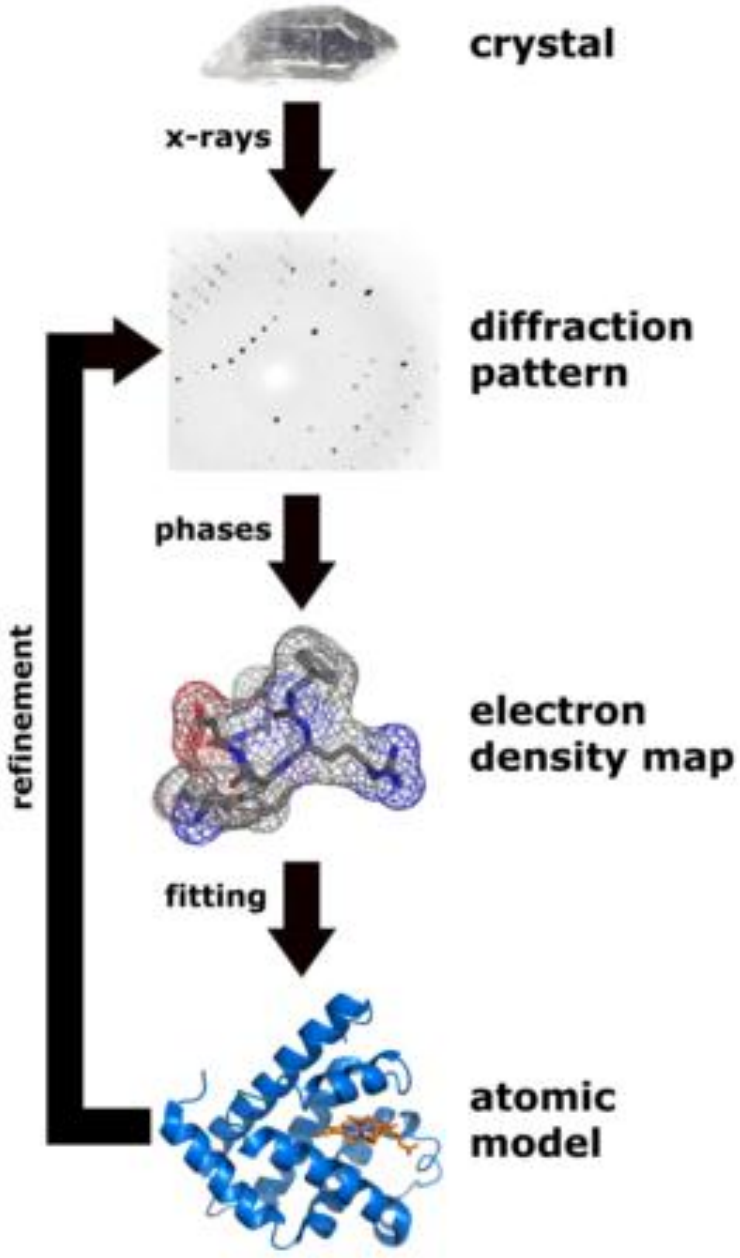
V. 2D Prediction

VI. 3D Prediction

VII. Useful Tools

Protein Structure Determination

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



Wikipedia, the free encyclopedia

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

Wikipedia, the free encyclopedia

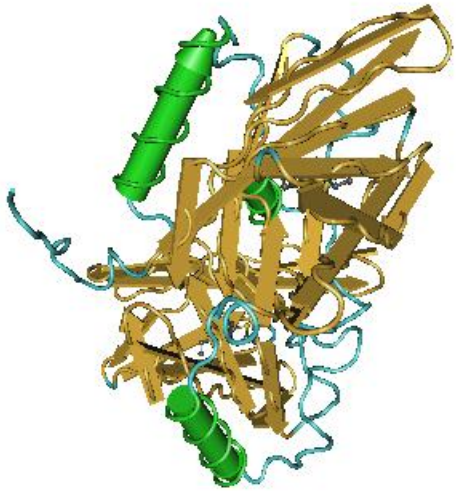
Storage in Protein Data Bank

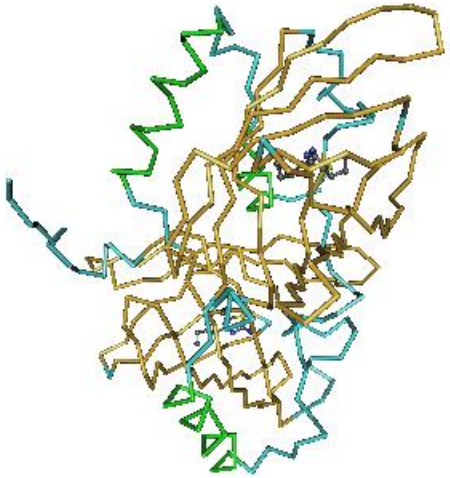
Search protein 1VJG

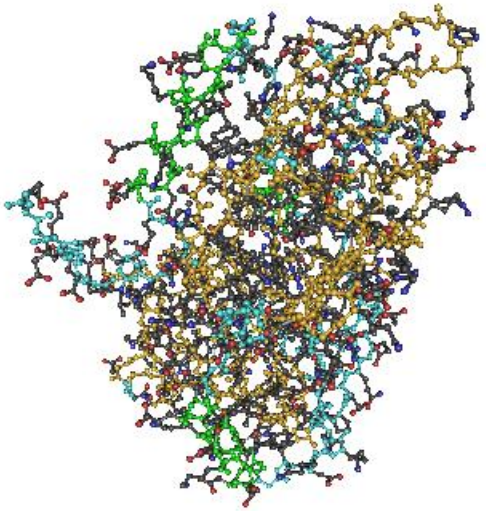
```
SEQRES   1 A   21  GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU
SEQRES   2 A   21  TYR GLN LEU GLU ASN TYR CYS ASN
SEQRES   1 B   29  PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU
SEQRES   2 B   29  ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR
SEQRES   3 B   29  THR PRO LYS
FORMUL   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
```

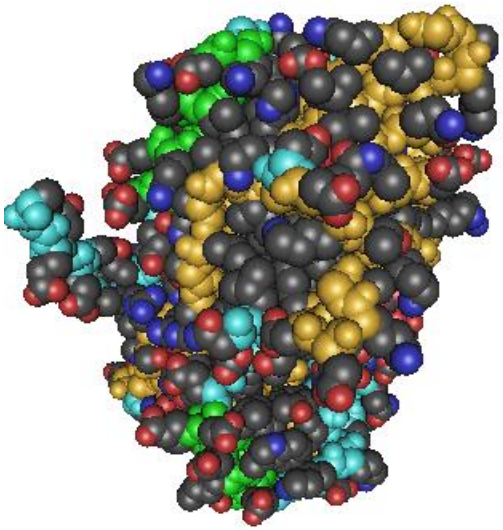
Structure Visualization

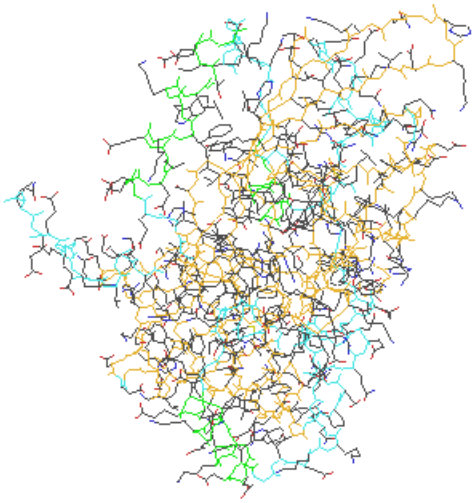
- Rasmol (http://www.umass.edu/microbio/rasmol/getras.htm)
- MDL Chime (plug-in) (http://www.mdl.com/products/framework/chime/)
- Protein Explorer (http://molvis.sdsc.edu/protexpl/frntdoor.htm)
- PyMol











J. Pevsner, 2005

Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, and Comparison

III. Structure Classification

IV. 1D Prediction

V. 2D Prediction

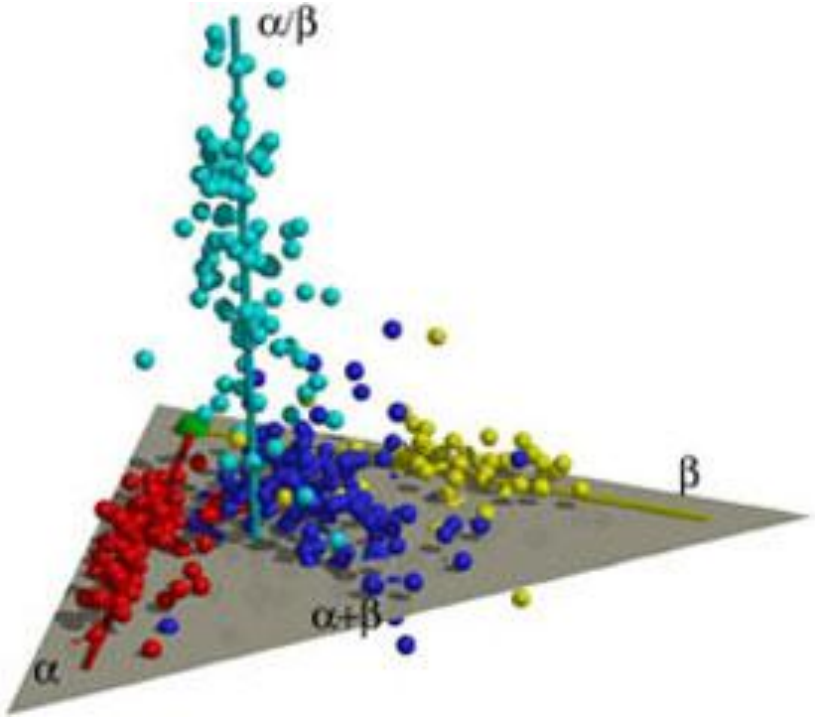
VI. 3D Prediction

VII. Useful Tools

Protein Structure Classification

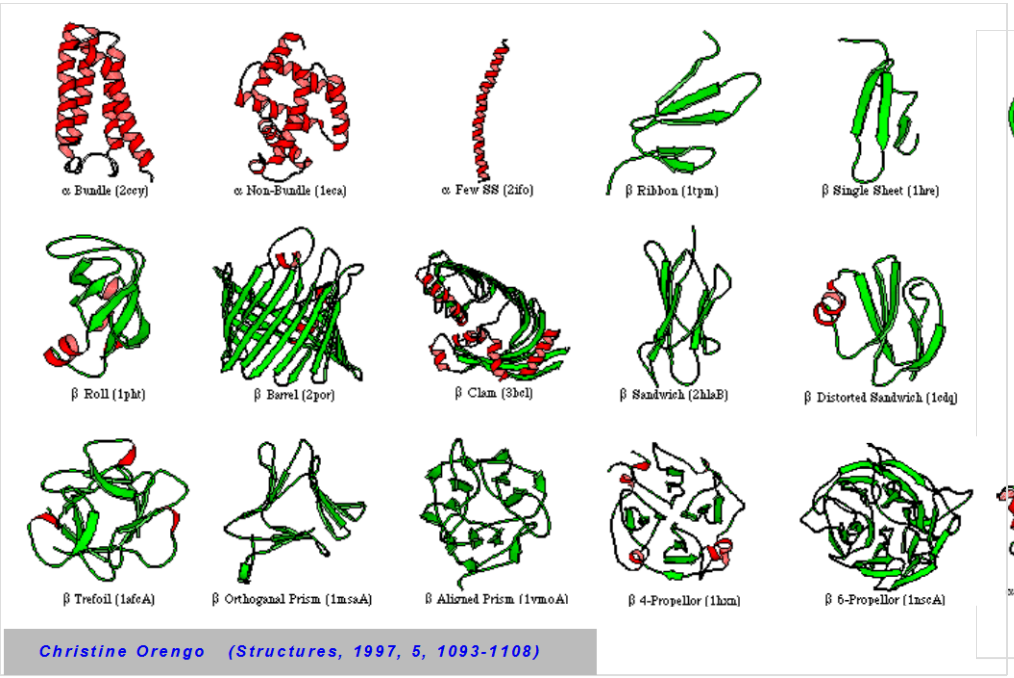
- About 4 million known protein sequences
- About 40,000 known protein structures in PDB
- Many protein structures are similar due to evolutionary relationship
- Many protein structures are similar due to convergent evolution
- Number of unique structure topologies is estimated to be limited (1000 – 1500?)
- Number of protein sequences is huge (20^{300})

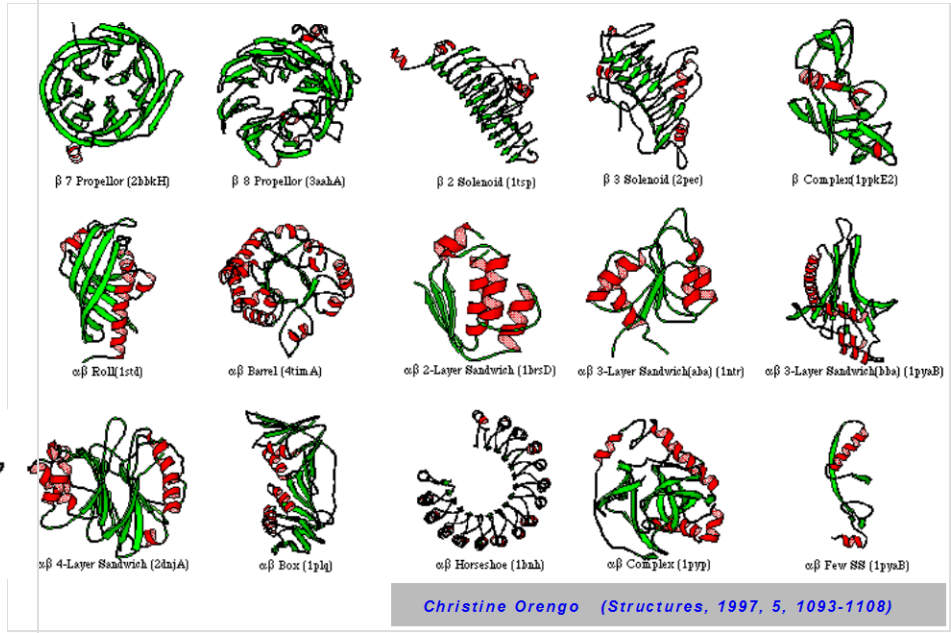
Protein Structure Universe



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

Colors in the universe of protein structures



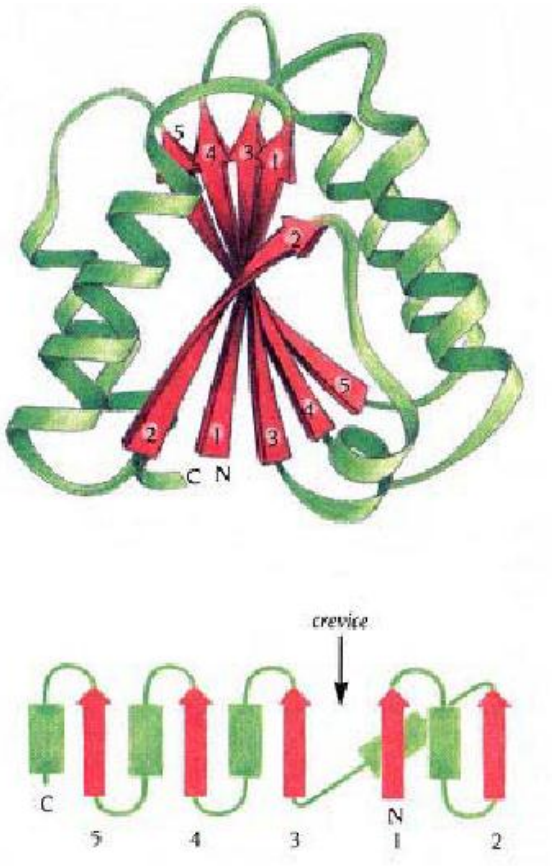


Christine Orengo **1997** *Structures* **5** 1093-1108

B. Rost, 2005

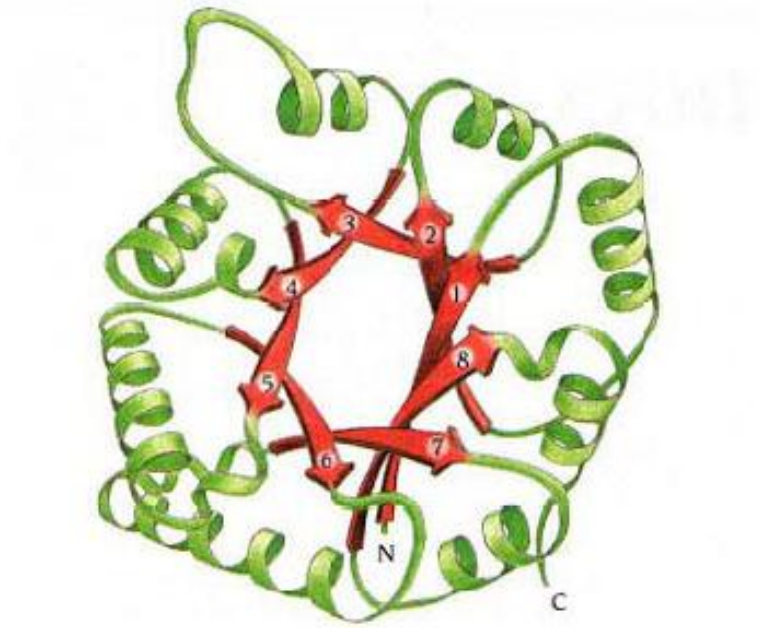
Typical Folds

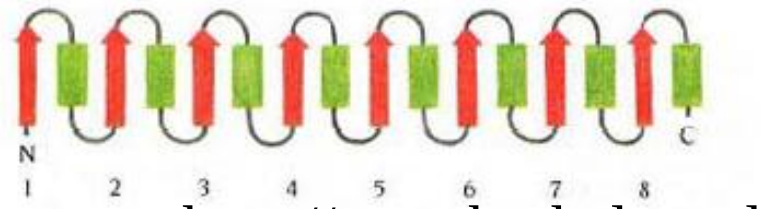
- Fold: connectivity or arrangement of secondary structure elements.
- NAD-binding Rossman fold
- 3 layers, a/b/a, parallel beta-sheet of 3 strands. Order: 321456



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

Another Fold Example: TIM beta-alpha barrel

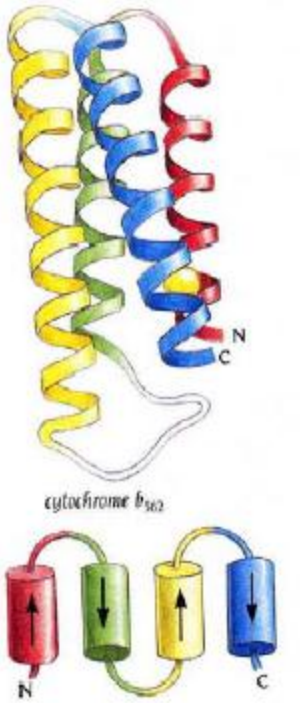




Contains parallel beta-sheet Barrel, closed. 8 strands. Order 1,2,3,4,5,6,7,8.

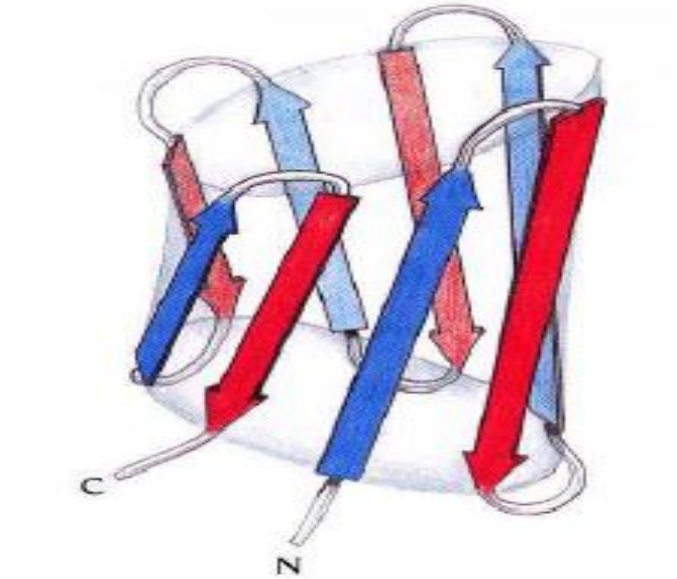
http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1hti;pc=a

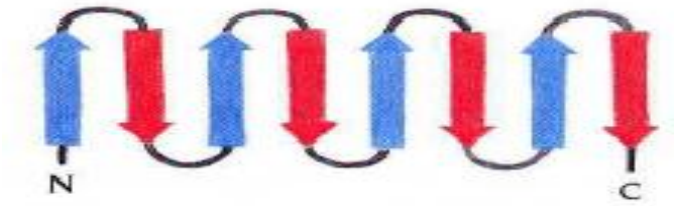
Fold: Helix Bundle (human growth factor)



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

Fold: beta barrel



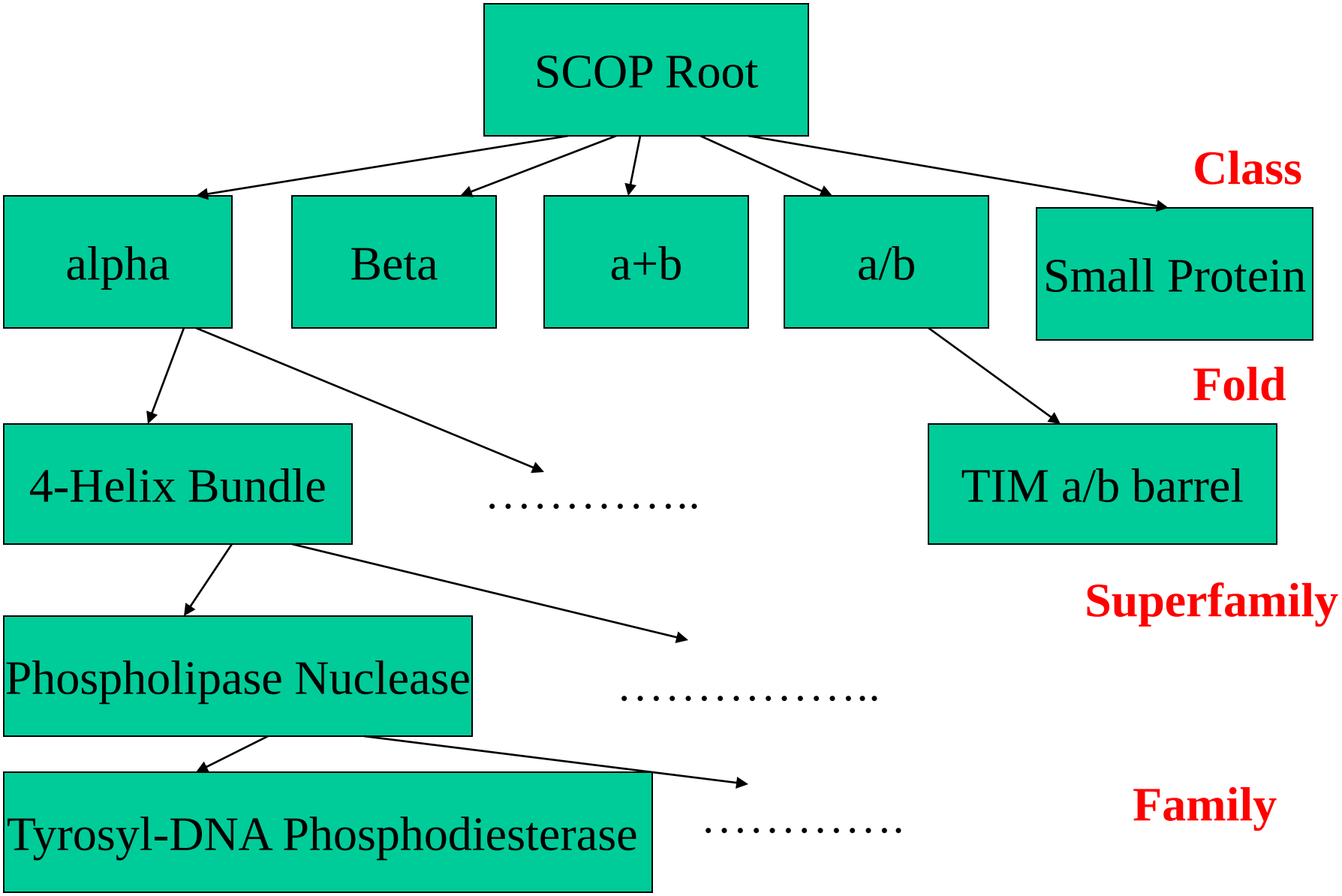


http://scop.berkeley.edu/rsgen.cgi?chime=1;pd=1rbp

Structure Classification Database

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

SCOP Classification Levels



Scop Classification Statistics

SCOP: Structural Classification of Proteins. **1.69** release
25973 PDB Entries (1 Oct 2004). 70859 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	218	376	608
All beta proteins	144	290	560
Alpha and beta proteins (a/b)	136	222	629
Alpha and beta proteins (a+b)	279	409	717
Multi-domain proteins	46	46	61
Membrane and cell surface proteins	47	88	99
Small proteins	75	108	171
Total	945	1539	2845

Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, and Comparison

III. Structure Classification

IV. 1D Prediction

V. 2D Prediction

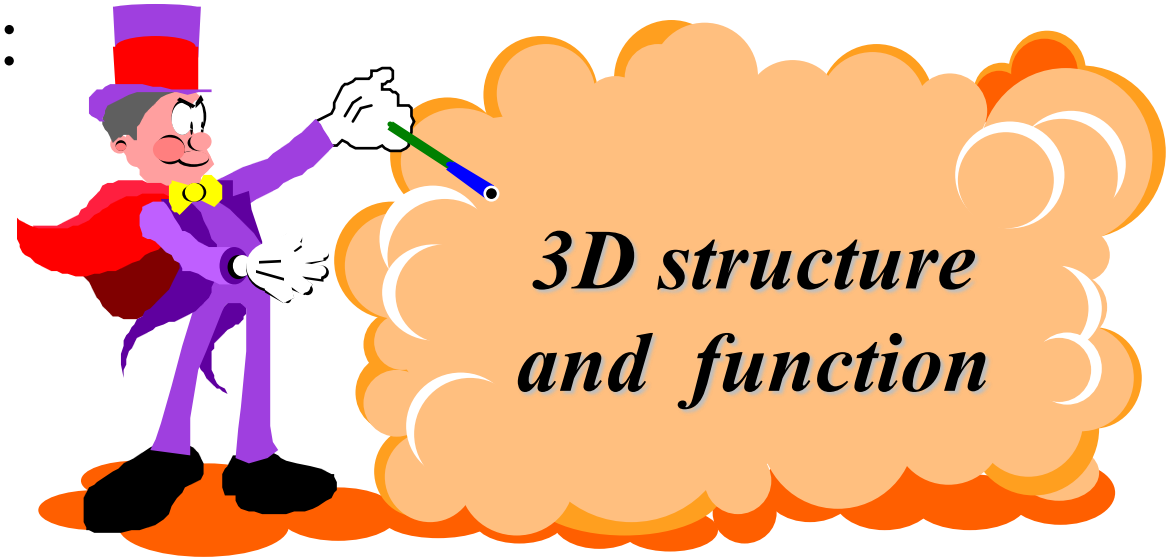
VI. 3D Prediction (emphasis)

VII. Tools and Applications

Goal of structure prediction

- Epstein & Anfinsen, 1961:
  sequence uniquely determines structure

- INPUT: sequence
- OUTPUT:

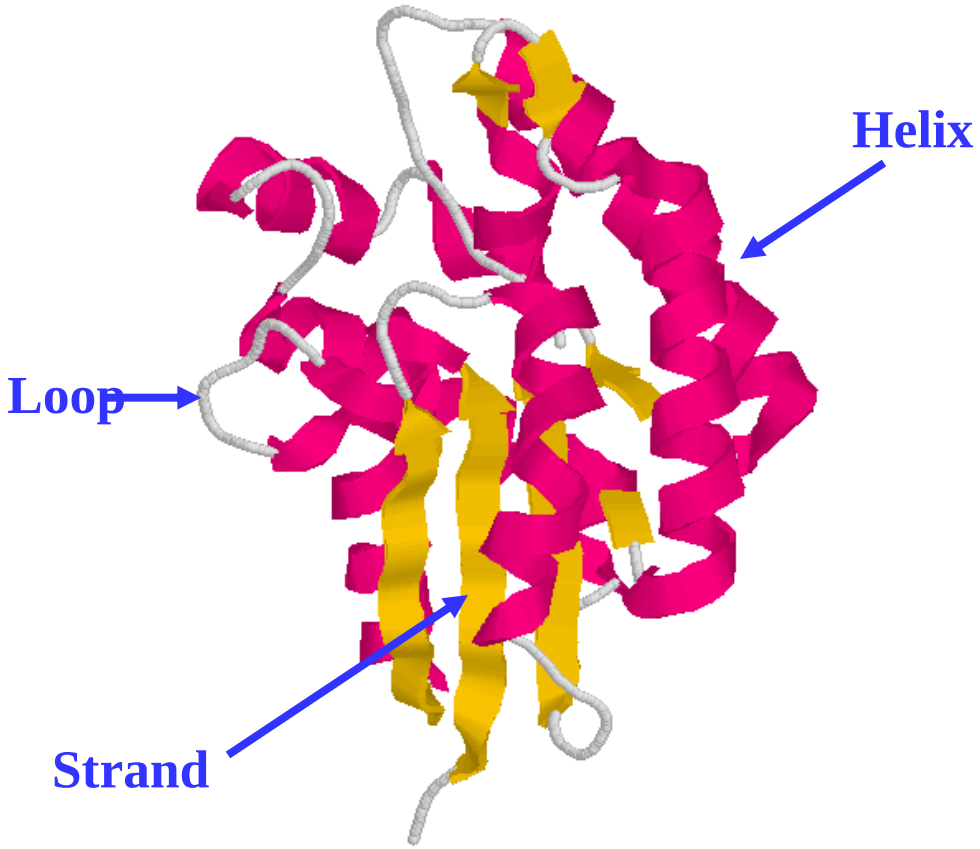


B. Rost, 2005

CASP – Olympics of Protein Structure Prediction

- Critical Assessment of Techniques of Protein Structure Prediction
- 1994,1996,1998,2000,2002,2004,2006
- Blind Test, Independent Evaluation
- CASP7: 100 targets
- CASP8: 120 targets

1D: Secondary Structure Prediction



MWLKKFGINLLIGQSV…

↓

Neural Networks + Alignments

↓

LLLLHHHHHLLLSSSSS…

Secondary Structure Prediction

- Predict secondary structure from sequence information without using any structural information
- Generation I (statistical methods)
- Generation II (statistical window)
- Generation III (evolutionary information + statistical machine learning)

B. Rost, 2005

Secondary Structure Prediction (Generation 1)

Single residues (1. generation)

– Chou-Fasman, GOR 1957-70/80
50-55% accuracy

Secondary structure propensity score of an amino acid AA

Log *P*(AA in Helix) / *P*(AA)
Log *P*(AA in Sheet) / *P*(AA)
Log *P*(AA in Loop) / *P*(AA)

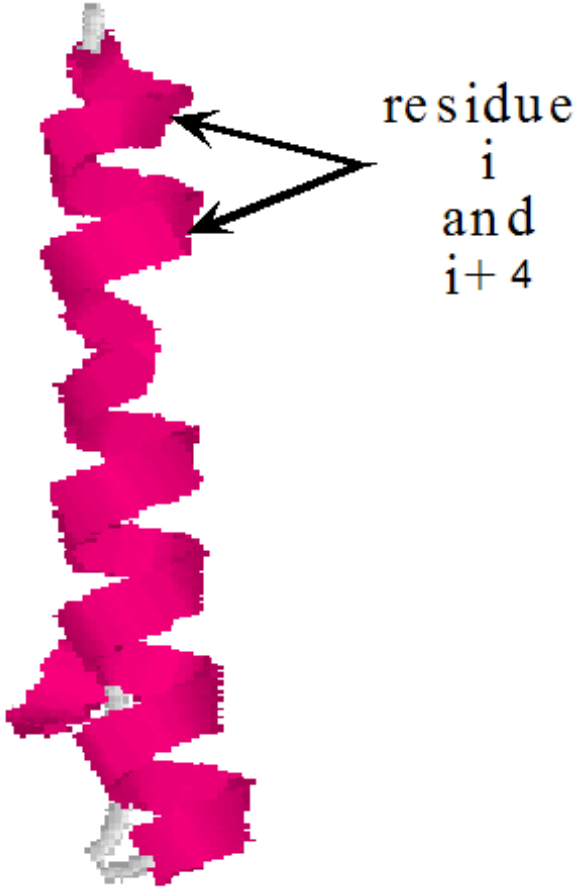
Secondary Structure Prediction (Generation II)

- Segments (window)
- GOR III
- Accuracy: 55-60%
- 1986-1992
- Estimation: max < 65%, Strand: non-local, < 40%

ARSKQPCTWYRE**S**QTCEQRKRTPA

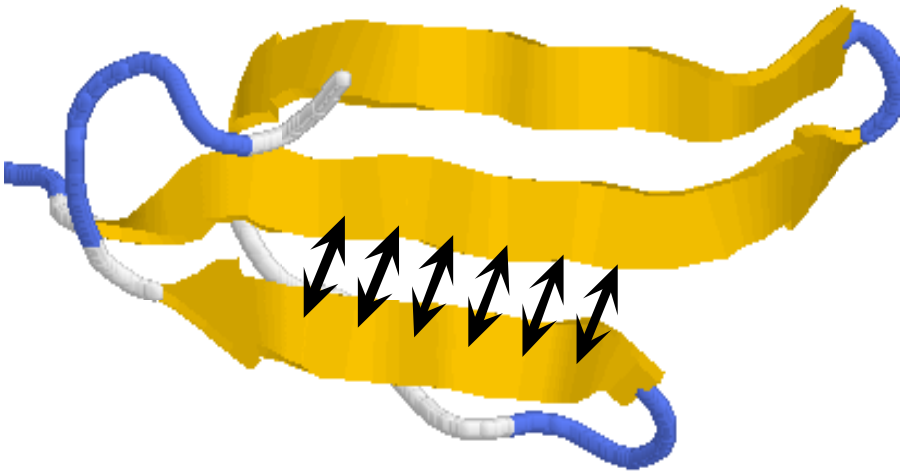
Average propensity scores of a window

Helix formation is local



B. Rost, 2005

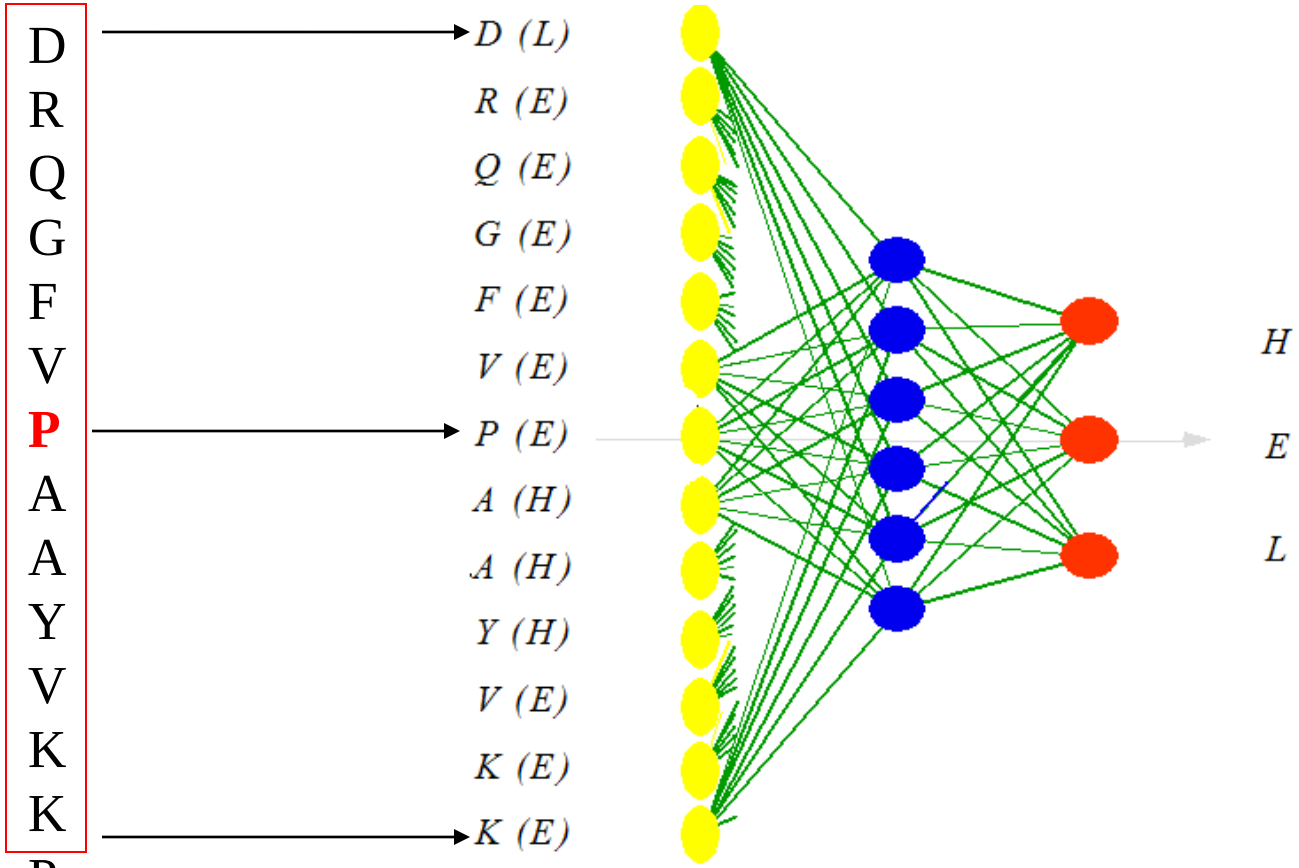
β-sheet formation is NOT local



Erabutoxin β (3ebx)

B. Rost, 2005

Secondary Structure Prediction (Generation III – Neural Network)

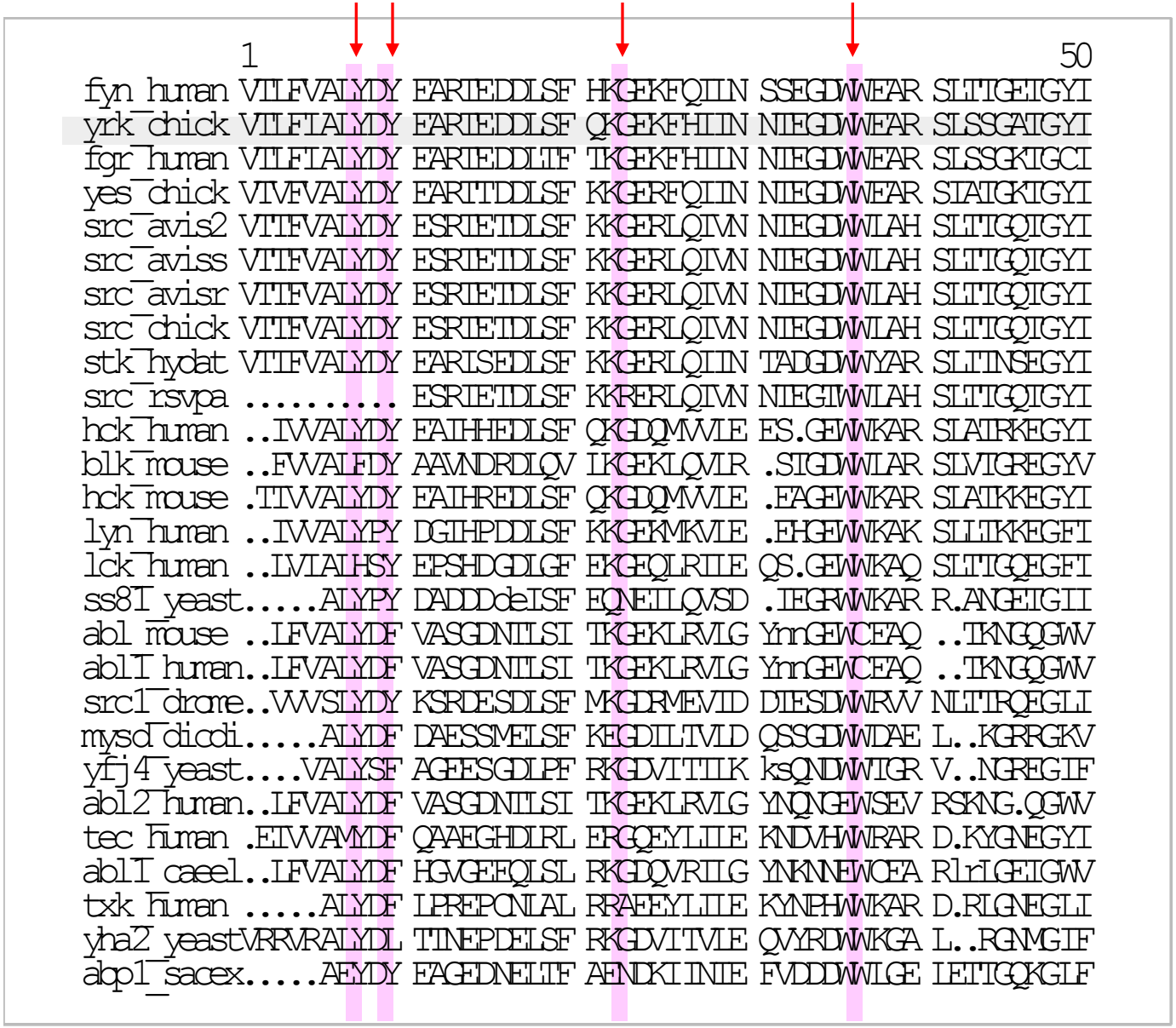


Output:
Prob (H)
Prob (E)
Prob (L)

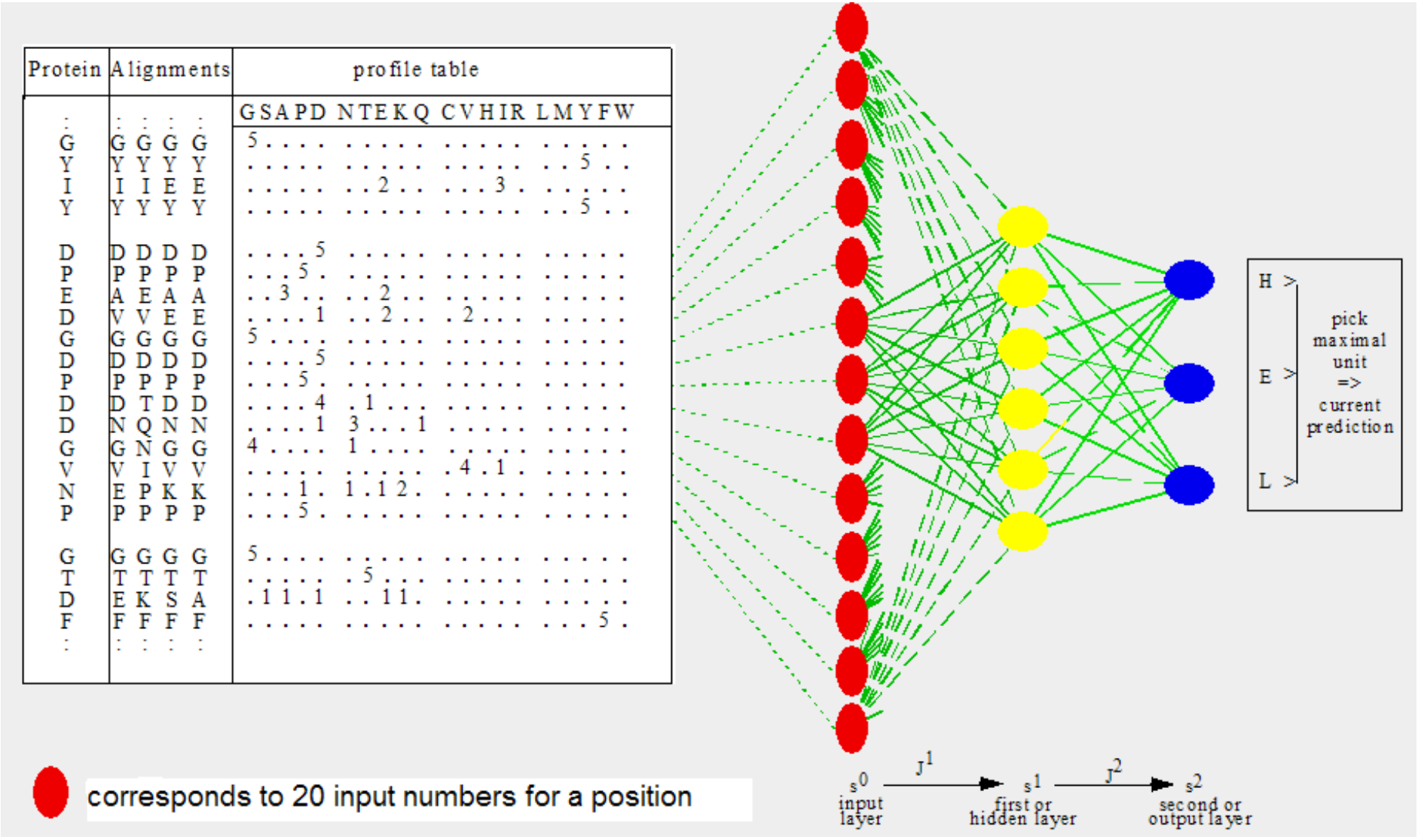
Question 1: how to encode an amino acid?
Question 2: how to train neural networks?
Parameter to decide: window size (51-101)

B. Rost, 2005

Second Breakthrough: **Evolutionary Information - Profile**



B. Rost, 2005



Comments: frequency is often normalized into probability

B. Rost, 2005

Prediction of protein secondary structure

- 1980: 55% simple
- 1990: 60% less simple
- 1993: 70% evolution
- 2000: 76% more evolution
- what is the limit?
  - 88% for proteins of similar structure
  - missing through:
    better definition of secondary structure
    including long-range interactions
  - structural switches
  - chameleon / folding

B. Rost, 2005

Useful Tools

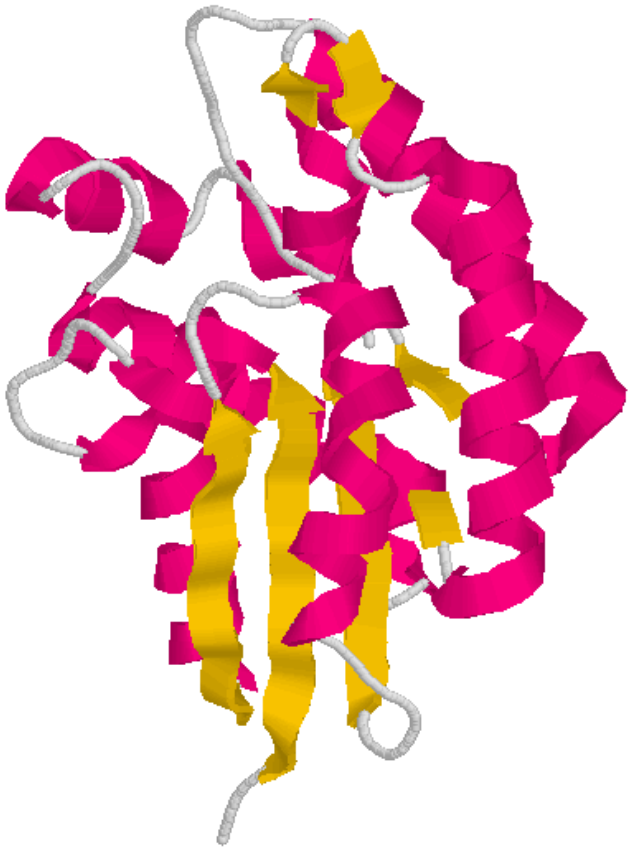
- PSI-PRED (http://bioinf.cs.ucl.ac.uk/psipred/)
- SSpro (http://casp.rnet.missouri.edu/sspro4.html)
- Porter (http://distill.ucd.ie/porter/)
- Prof_PHD (http://cubic.bioc.columbia.edu/predictprotein/)
- SAM (http://www.cse.ucsc.edu/research/compbio/sam.html)

Outline

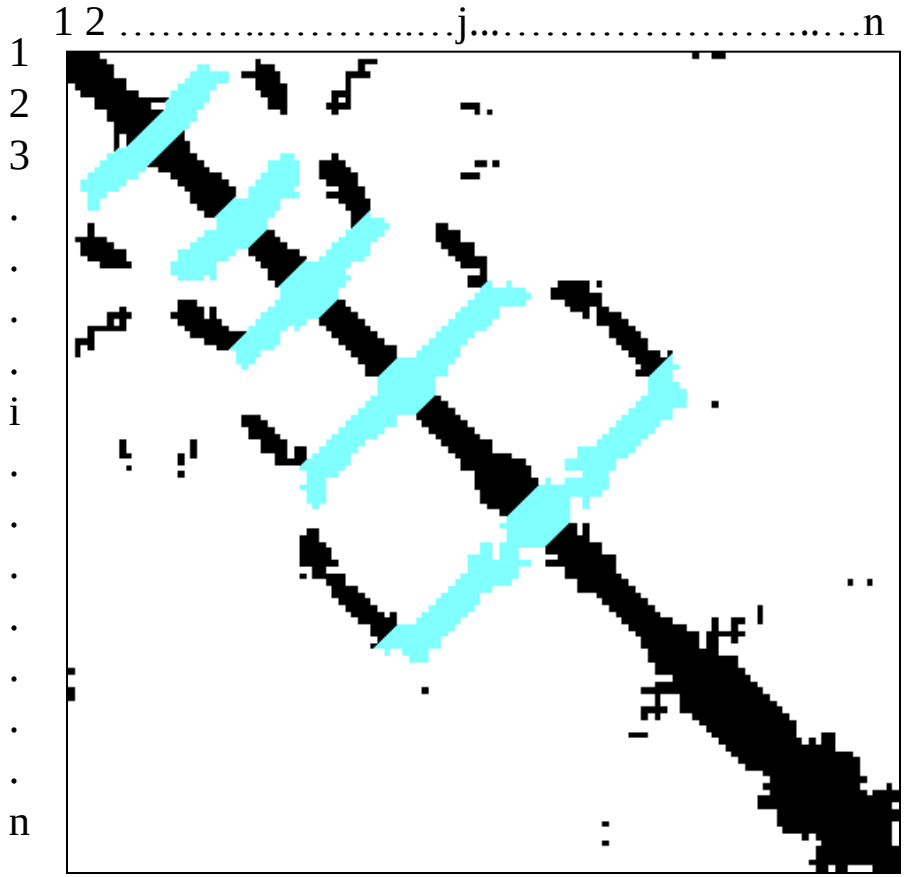
I. Sequence, Structure, Function Relation
II. Determination, Storage, Visualization, and Comparison
III. Structure Classification
IV. 1D Prediction
V. 2D Prediction
VI. 3D Prediction
VII. Tools and Projects

2D: Contact Map Prediction

3D Structure



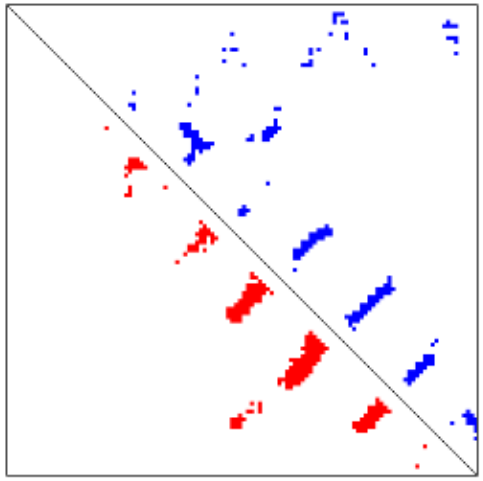
2D Contact Map



Distance Threshold = 8A°

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

Main Contact Prediction Tools



SVMcon: Protein Contact Map Prediction Using Support Vector Machine and a Large Feature Set

(Download SVMcon 1.0 software and source code (Linux version))

Email address (where the prediction will be sent):

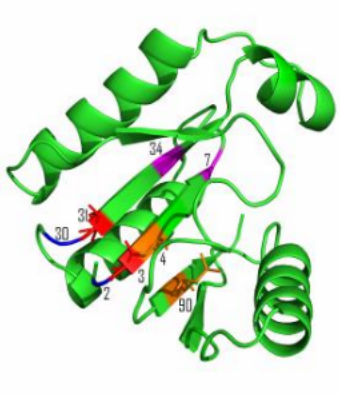
Target Name (required):

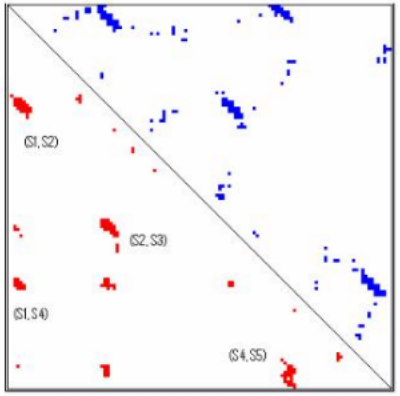
Protein sequence (one plain sequence, no headers, at most 400 residues):

Predict

Cheng and Baldi, BMC Bioinformatics, 2007

Main Contact Prediction Tools





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

Tegge et al., Nucleic Acids Research, 2009

Take home: 2D prediction

- Prediction hard, but stakes are high
- inter-residue
  - Correlated mutations can imply spatial proximity
  - Distinction between different models, no accurate prediction of 3D, yet
- Don’t freak out when accuracy is low
    - 1) how accurate are these prediction methods on average
    - 2) are all important contacts predicted?
- for 5% of the best-predicted contacts prediction accuracy about 50% (sequence separation >= 6)

B. Rost, 2005

Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, and Comparison

III. Structure Classification

IV. 1D Prediction

V. 2D Prediction

VI. 3D Prediction

VII. Useful Tools

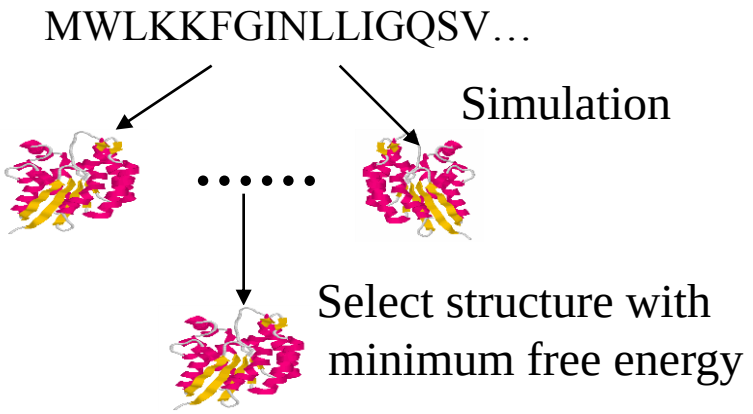
Two Methodologies for 3D Structure Prediction

- AB Initio Method (physical-chemical principles / molecular dynamics, knowledge-based approaches)
- Template-Based Method (knowledge-based approaches)

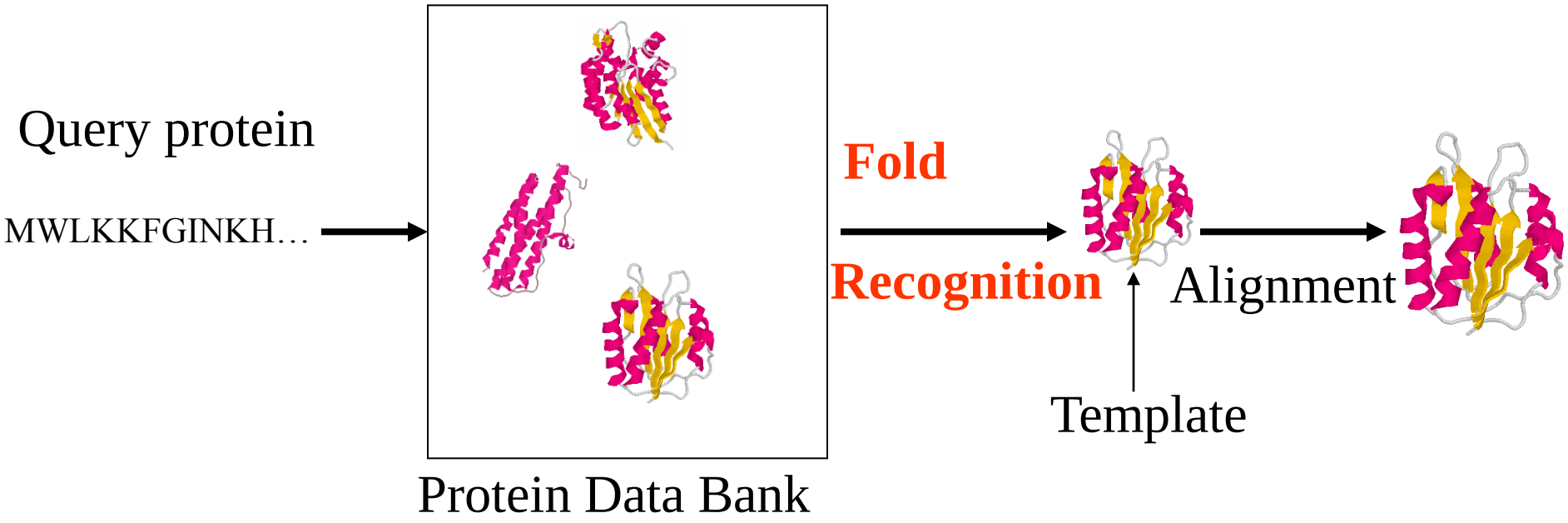
Two Approaches

- **Ab Initio Structure Prediction**

  Physical force field – protein folding
  Contact map - reconstruction



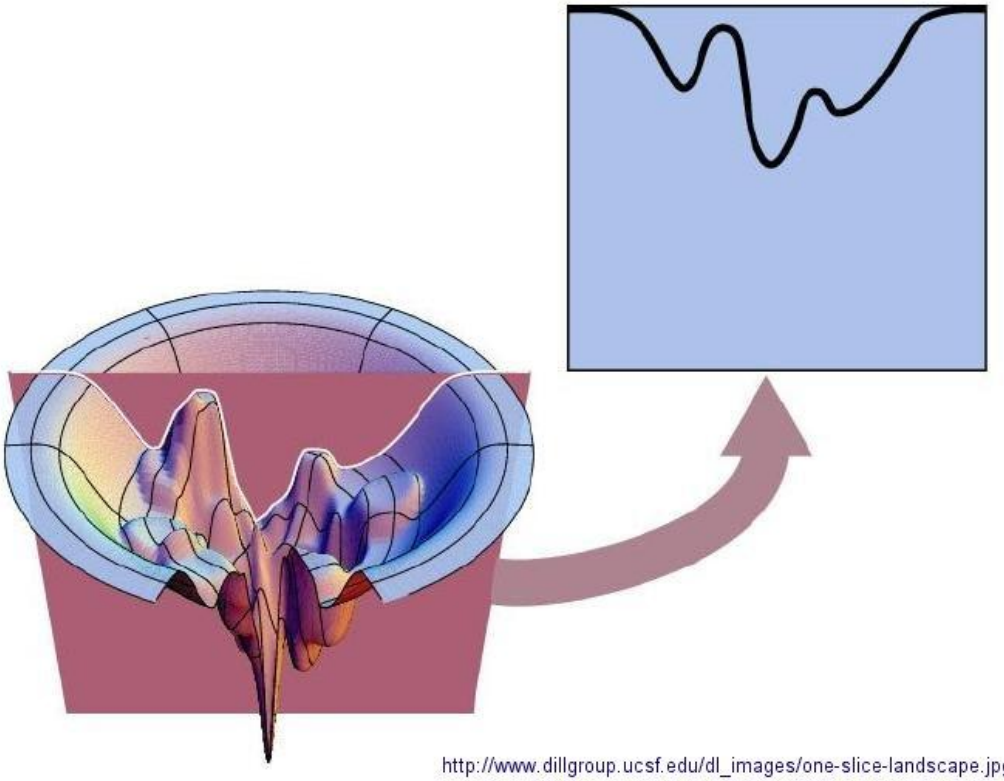
- **Template-Based Structure Prediction**



Ab Initio Structure Prediction

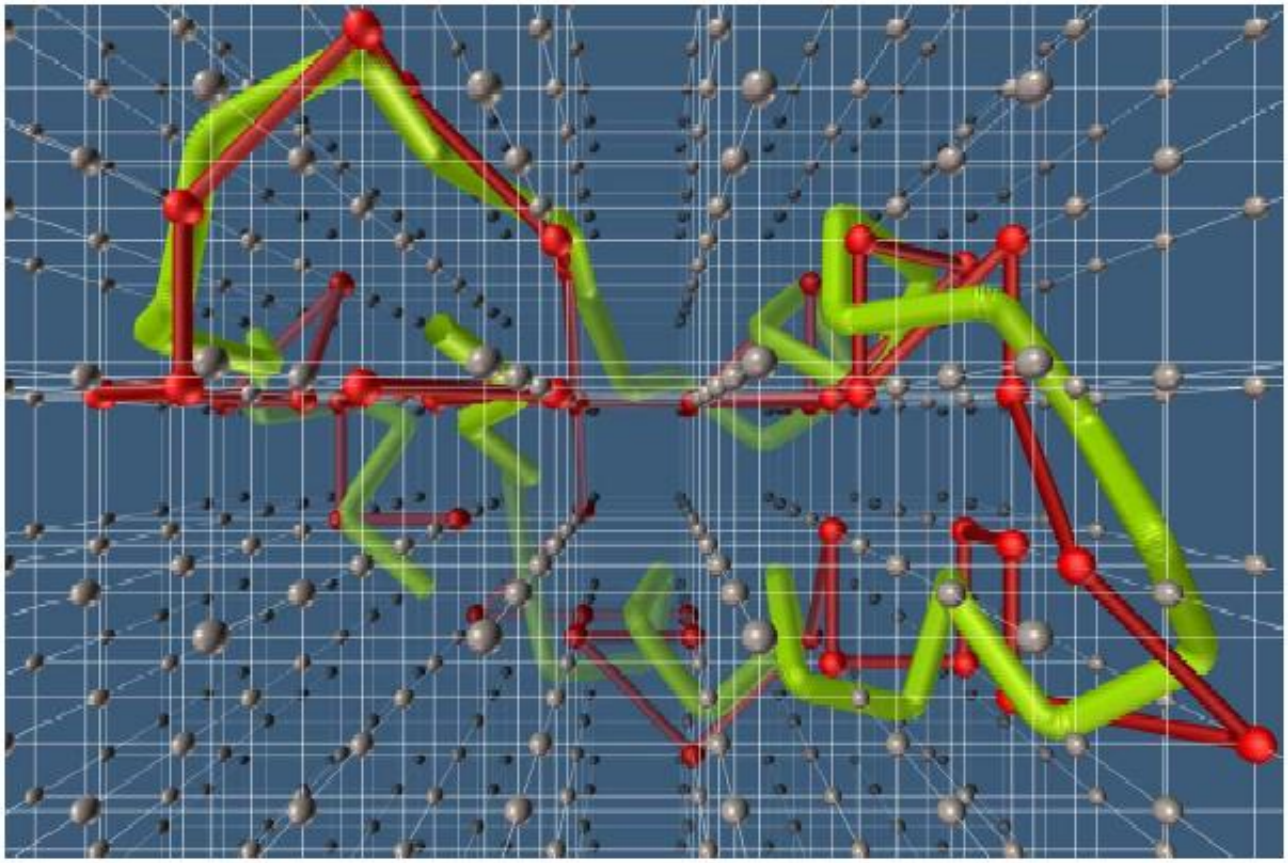
- Energy Function
- Structure Sampling
- Currently, the accuracy of *Ab-Initio* method is low.
- Energy function is not accurate.
- Sampling method can’t find the structure with lowest energy in majority of cases.
- Sampling takes a long time and only is able to sample a small structure space.

Protein Energy Landscape



http://www.dillgroup.ucsf.edu/dl_images/one-slice-landscape.jpg

Markov Chain Monte Carlo Simulation



picture from http://www.scripps.edu/pub/olson-web/people/sanner/html/lat_gallery.html

Some *Ab Initio* Tools

- David Baker's Rosetta (http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/Rosetta/)
- Cheng's MULTICOM
- Xu et al.'s MUFOLD
- Zhang and Skolnick's TASSER

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**.

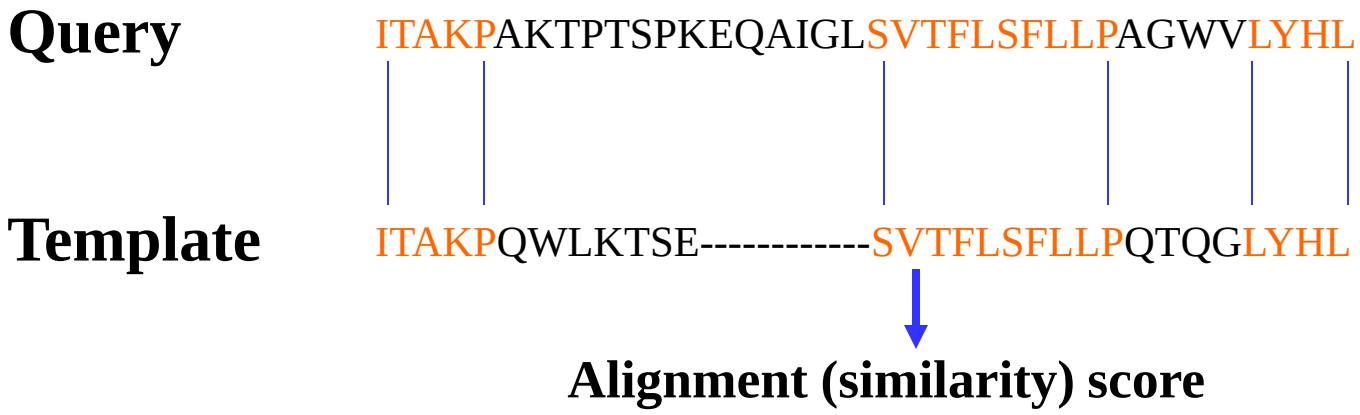
Template Identification

- Sequence alignment (BLAST, Smith-Waterman algorithm)
- Sequence profile alignment (PSI-BLAST, HMM (SAM-T02, Hammer)
- Profile-Profile alignment (HHSearch, Sparks, Compass, FFAS, BASIC)
- Sequence-Structure Alignment (3D-1D, FUGUE, mGenThreader, Raptor)
- Consensus (Pcons, 3D-Jury)
- Machine Learning Information Retrieval Approach (FOLDpro)

Classic Fold Recognition Approaches

Sequence - Sequence Alignment

(Needleman and Wunsch, 1970. Smith and Waterman, 1981)

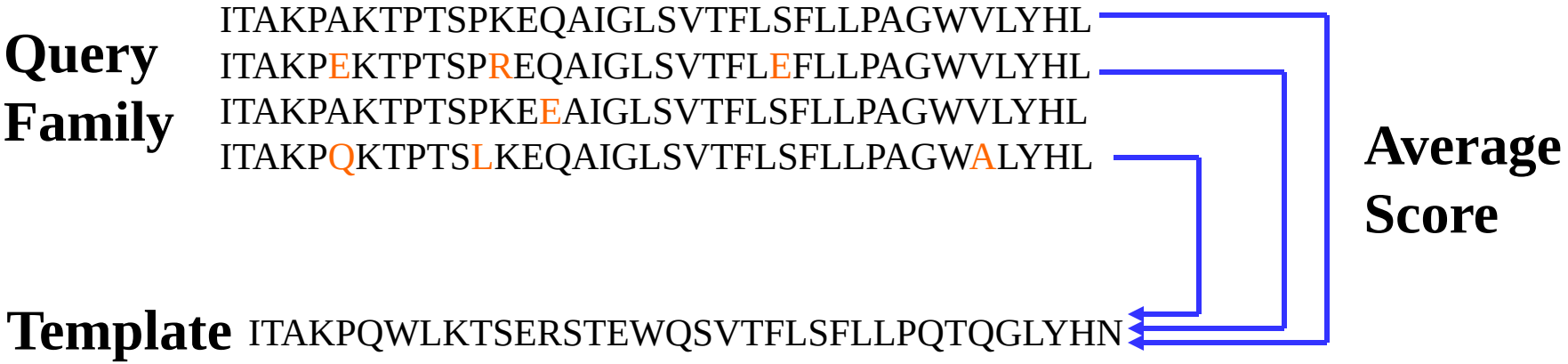


Works for >40% sequence identity
(Close homologs in protein family)

Classic Fold Recognition Approaches

Profile - Sequence Alignment

(Altschul et al., 1997)



More sensitive for distant homologs in superfamily.
(> 25% identity)

Classic Fold Recognition Approaches

Profile - Sequence Alignment

(Altschul et al., 1997)

Query Family

```
12.....n
ITAKPAKTPTSPKEQAIGLSVTFLSFLLPAGWVLYHL
ITAKPEKTPTSPREQAIGLSVTFLEFLLPAGWVLYHL
ITAKPAKTPTSPKEEAIGLSVTFLSFLLPAGWVLYHL
ITAKPQKTPTSLKEQAIGLSVTFLSFLLPAGWALYHL
```

	1	2	…	n
A	0.4			
C	0.1			
…				
W	0.5			

Position Specific Scoring Matrix Or Hidden Markov Model

Template ITAKPQWLKTSERSTEWQSVTFLSFLLPQTQGLYHN

More sensitive for distant homologs in superfamily.
(> 25% identity)

Classic Fold Recognition Approaches

Profile - Profile Alignment

(Rychlewski et al., 2000)

Query Family

ITAKPAKTPTSPKEQAIGLSVTFLSFLLPAGWVLYHL
ITAKPEKTPTSPREQAIGLSVTFLEFLLPAGWVLYHL
ILAKPAKTPTSPKEEAIGLSVTFLSFLLPAGWVLYHL
ITAKPQKTPTSLKEQAIGLSVTFLSFLLPAGWALYHL

	1	2	…	n
A	0.1			
C	0.4			
…				
W	0.5			

Template Family

ITAKPQWLKTSERSTEWQSVTFLSFLLPQTQGLYHN
IPARPQWLKTSKRSTEWQSVTFLSFLLPYTQGLYHN
IGAKPQWLWTSERSTEWHSVTFLSFLLPQTQGLYHM

	1	2	…	m
A	0.3			
C	0.5			
…				
W	0.2			

More sensitive for very distant homologs.
(> 15% identity)

Classic Fold Recognition Approaches

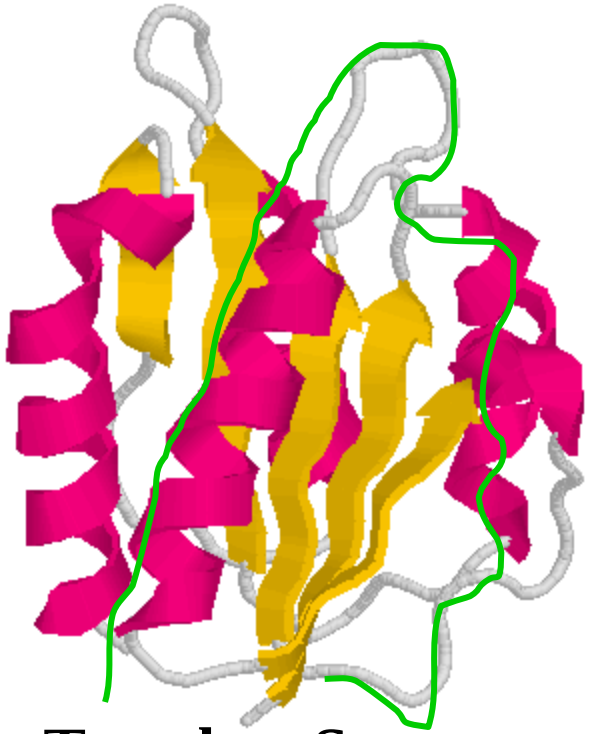
Sequence - Structure Alignment (Threading)

(Bowie et al., 1991. Jones et al., 1992. Godzik, Skolnick, 1992. Lathrop, 1994)

Query

MWLKKFGINLLIGQS….

Fit

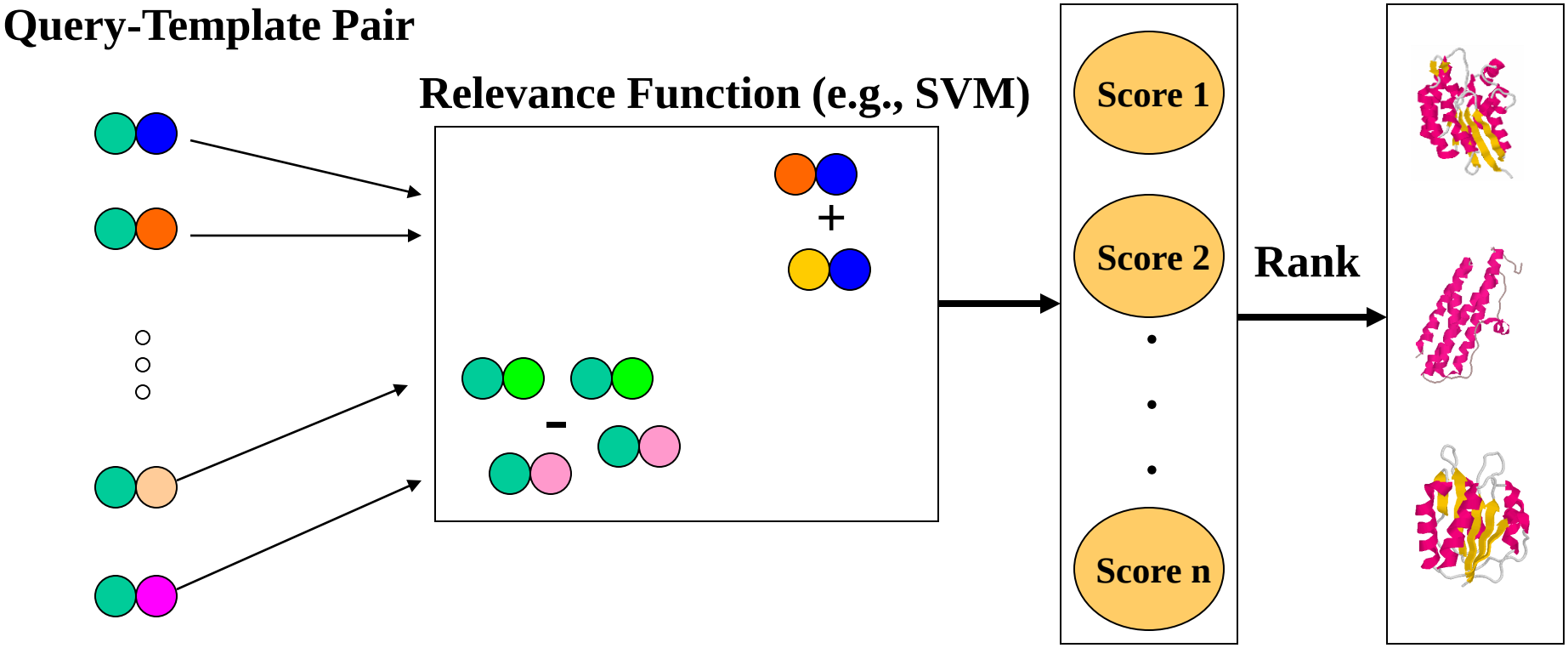


Template Structure

Fitness Score

Useful for recognizing similar folds without sequence similarity. (no evolutionary relationship)

Machine Learning Information Retrieval Framework



Cheng and Baldi. Bioinformatics, 2006

Pairwise Feature Extraction

- **Sequence / Family Information Features**

  Cosine, correlation, and Gaussian kernel
- **Sequence – Sequence Alignment Features**

  Palign, ClustalW
- **Sequence – Profile Alignment Features**

  PSI-BLAST, IMPALA, HMMer, RPS-BLAST
- **Profile – Profile Alignment Features**

  ClustalW, HHSearch, Lobster, Compass, PRC-HMM
- **Structural Features**

  Secondary structure, solvent accessibility, contact map, beta-sheet topology

Query-Template Alignment

- Most fold recognition methods are some kind of specialized alignment methods. So they generate alignments.
- For similar sequences, PSI-BLAST alignment is ok. For distantly related sequences, profile-profile alignment methods seem to be better (HHSearch, COMPASS, LOBSTER (COACH), CLUSTALW, T-Coffee, and so on).

Model Generation

- Modeller (http://www.salilab.org/modeller/download_installation.html)
- Swiss-Model (http://swissmodel.expasy.org//SWISS-MODEL.html)
- **Segmod/ENCAD (csb.stanford.edu/levitt/segmod/)**

TARGET

TEMPLATE

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

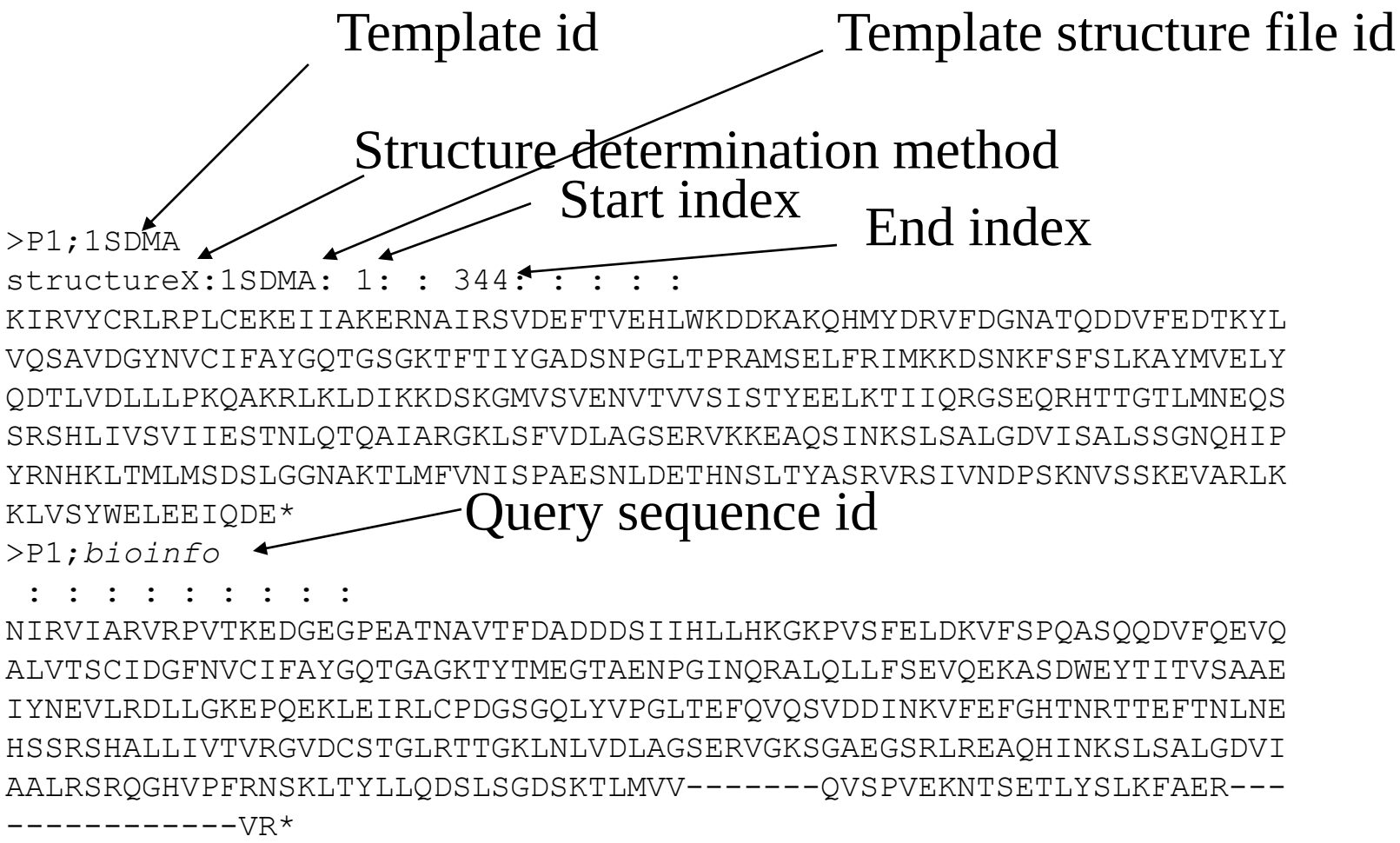
ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

A. Fisher, 2005

How to use 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.
- See project step 5-8 for more details.

An PIR Alignment Example




```
>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

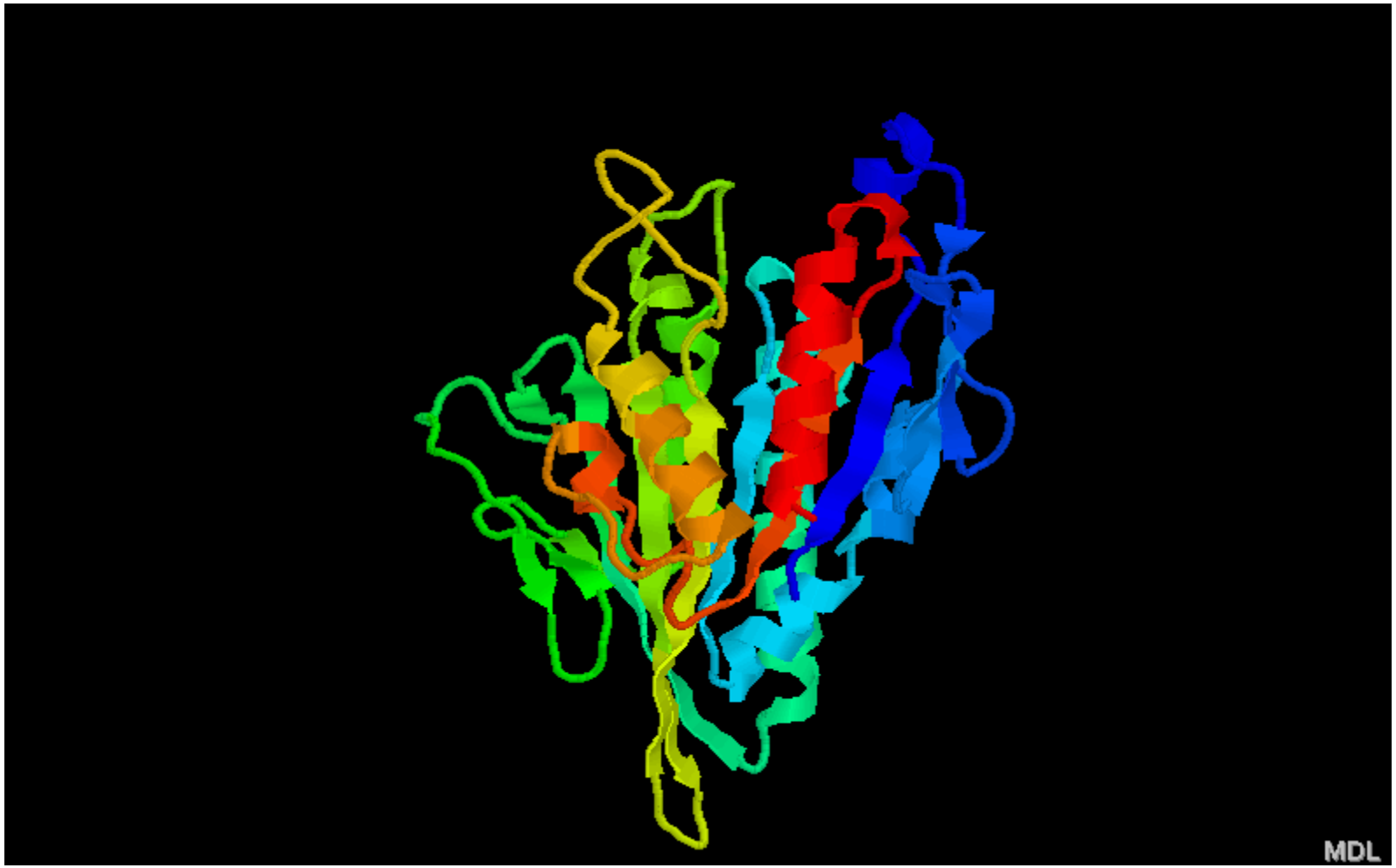
PIR alignment file name

Template structure file id

Query sequence id

Output Example

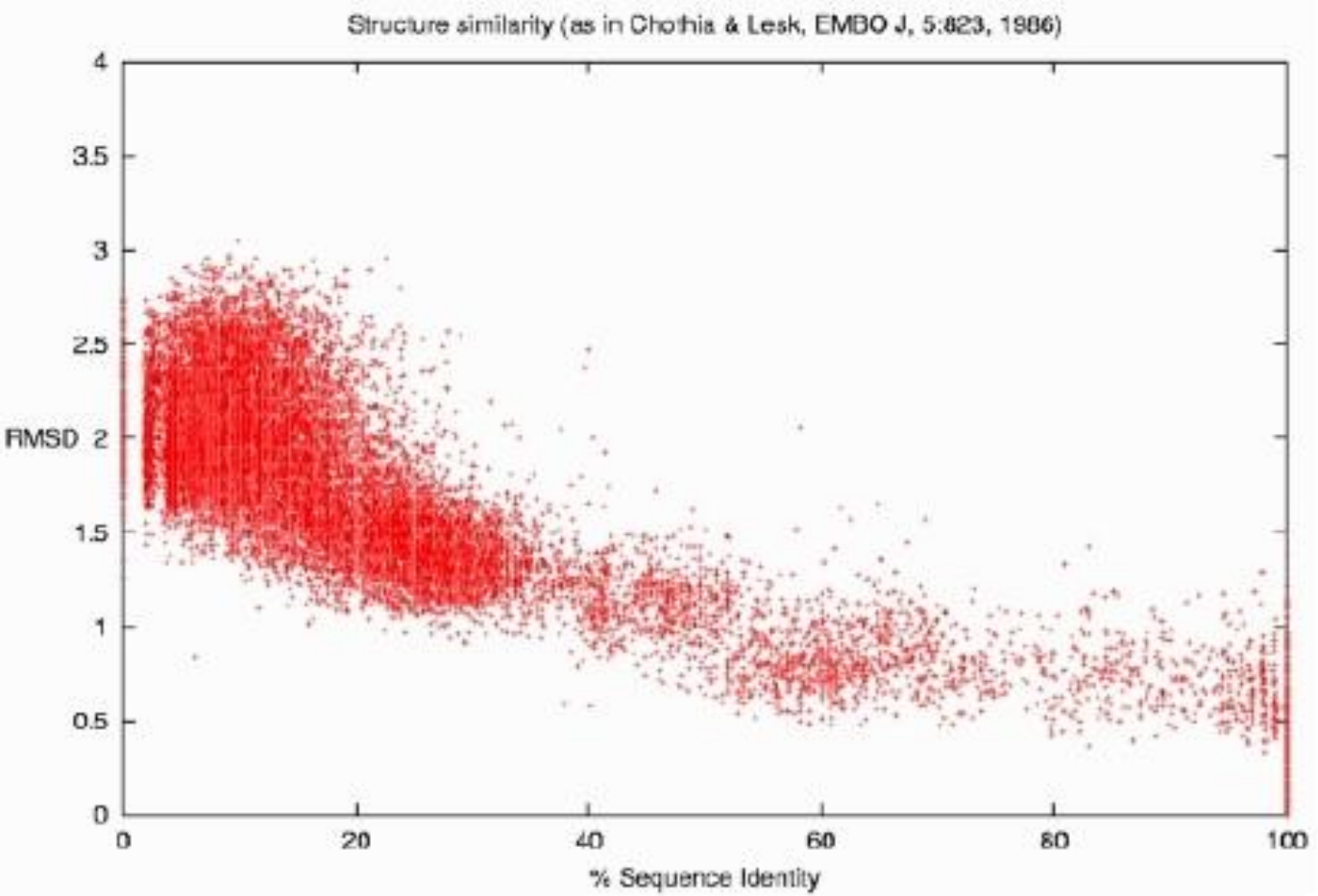
Command: mod8v2 bioinfo.py



Model Evaluation

- Prosa (http://www.came.sbg.ac.at/Services/prosa.html)
- Verify-3D (http://nihserver.mbi.ucla.edu/Verify_3D/)
- ProCheck (http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html)
- ModelEvaluator: Machine learning approach (Zheng et al., Proteins, 2008)
- APOLLO (Zheng et al., Bioinformatics, 2011)

Sequence Identity and Alignment Quality in Structure Prediction



Superimpose
-> RMSD

%Sequence Identity: percent of identical residues in alignment
RMSD: square root of average distance between predicted structure and native structure.

Outline

I. Sequence, Structure, Function Relation

II. Determination, Storage, Visualization, and Comparison

III. Structure Classification

IV. 1D Prediction

V. 2D Prediction

VI. 3D Prediction (emphasis)

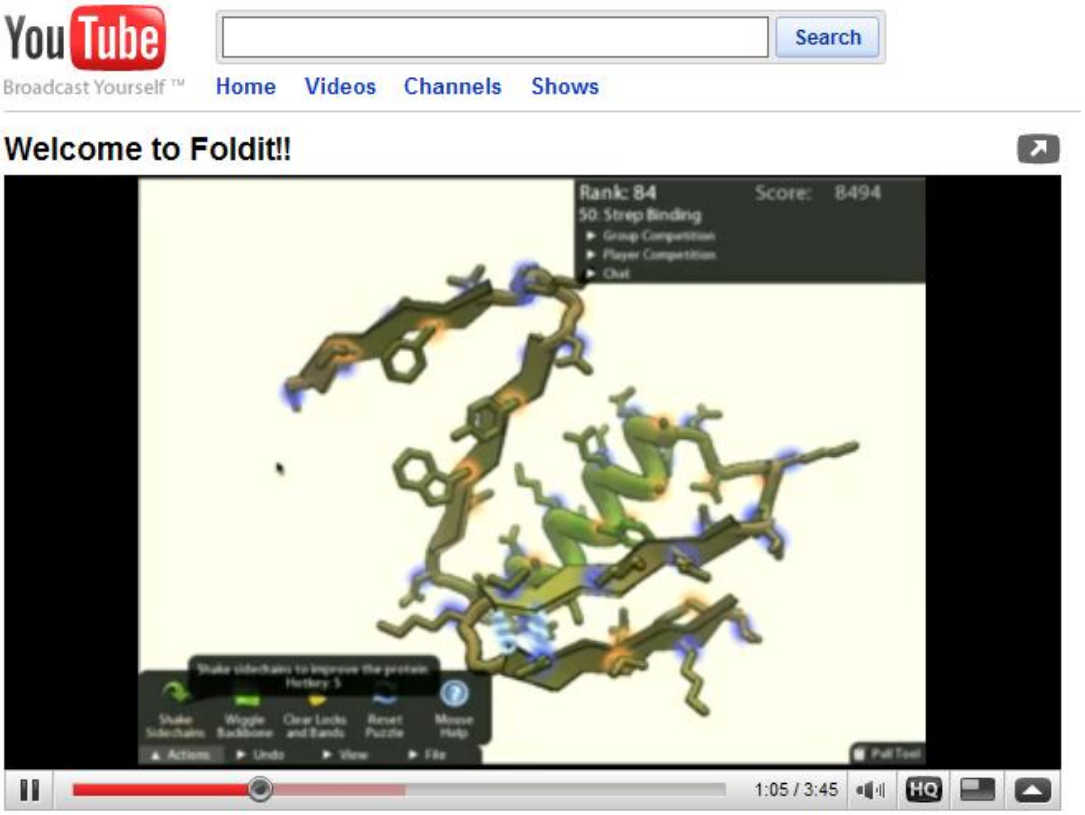
VII. Tools and Applications

3D Structure Prediction Tools

- **I-TASSER: http://zhang.bioinformatics.ku.edu/I-TASSER/**
- **MULTICOM: http://casp.rnet.missouri.edu/multicom.html**
- **Sparks (http://phyyz4.med.buffalo.edu/hzhou/anonymous-fold-sp3.html)**
- **HHpred (http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred)**
- **Robetta (http://robetta.bakerlab.org/)**
- FUGUE (http://www-cryst.bioc.cam.ac.uk/%7Efugue/prfsearch.html)
- FOLDpro (http://mine5.ics.uci.edu:1026/foldpro.html)
- SAM (http://www.cse.ucsc.edu/research/compbio/sam.html)
- 3D-PSSM (http://www.sbg.bio.ic.ac.uk/3dpssm/)
- mGenThreader (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html)
- 3D-Jury (http://bioinfo.pl/Meta/)
- FFAS (http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl)
- PCONS (http://pcons.net/)
- Phyre (http://www.sbg.bio.ic.ac.uk/~phyre/)

Protein Folding Game: FoldIt

- Video:
  http://www.youtube.com/watch?v=lGYJyur4FUA



Fun: Movie Demo

http://www.youtube.com/watch?v=E0J9H3Yxjec&feature=related

Assignment (Review one paper)

- Y. Zhang. Protein structure prediction: when is it useful? Current Opinion in Structural Biology, vol. 19, 145-155, 2009.
- Review the paper and write one page summary
- Due Sept. 9, 2011