Intro to Structural Prediction

Protein Structure

Primary Sequence
- The amino acid sequence

Secondary Structure
- Alpha Helix, Beta Sheet, Loop, Turn

Tertiary Structure
- How the secondary structures combine

Quarternary Structure
- Combination of tertiary structures
- Multimeric proteins

Secondary Structure

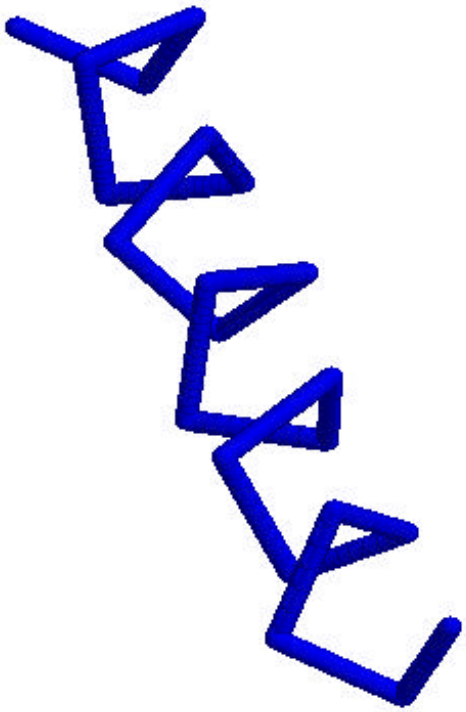
Secondary Structure Analysis

- Alpha Helical Regions
- Beta Sheet Regions
- Turns
- Transmembrane Regions
- Surface Probability
- Internal Domains
- Antigenic Regions

Alpha Helical Regions

3.6 residues per turn
Right Handed

Stabilized by H-bonds
Usually find alanine
Disrupted by proline



Odds of Success

What are the chances of getting the right answer by random chance?

Depends on how many states you have

Helix or Turn	50% chance
Helix, or Sheet, or Turn	33% chance

Most prediction programs are 60-70% accurate

Prediction Methods

ab initio methods
- Physical properties & interactions

Statistics of amino acid distributions
- Chou &Fasman, GOR

Sequence Patterns

Neural Network Methods

Chou-Fasman Methods

Based on observed properties of amino acids to take up α, β or turn configurations

Step 1 : Using sliding window, Look for nucleation centers
- 4 out of 6 residues must be high α helical propensity
- 3 out of 5 residues must be high β strand propensity

Step 2 : Resolve Conflicts between overlapping regions
- Some regions may appear both α and β

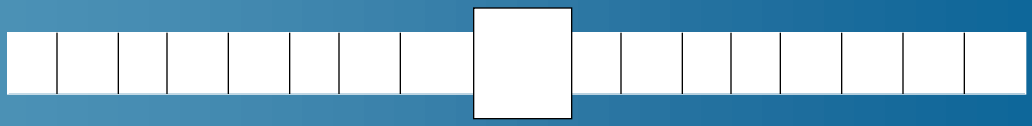
Step 3 : Turns
- Tetrapeptides (each position has a different propensity)

GOR Method

GOR = Garnier, Osguthorpe-Robson

Uses a PSSM, Position Specific Scoring Matrix

Probability of a structure is based on the 8 residues upstream and downstream. Position of residues is important.



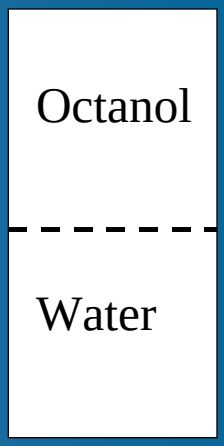
Physical Properties

Hydrophobicity - Fears Water

Hydrophilicity - Likes Water

Hopp-Woods scale

Determines hydrophobicity by seeing whether an amino acid would rather stay in the octanol or water environment.



What does this have to do with protein structure?

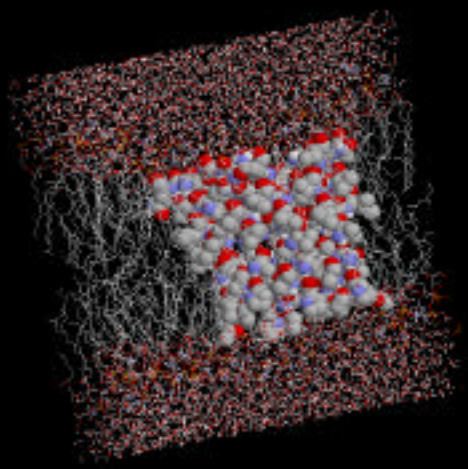
The cores of globular proteins are more hydrophobic

In water soluble proteins, surface residues are hydrophillic

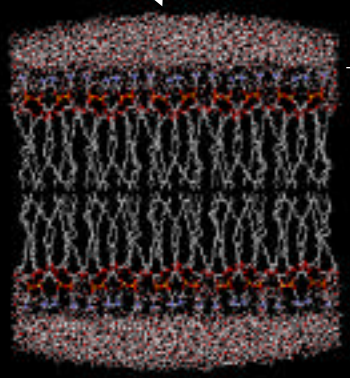
In membrane proteins, regions in contact with membrane are hydrophobic

Membrane Proteins

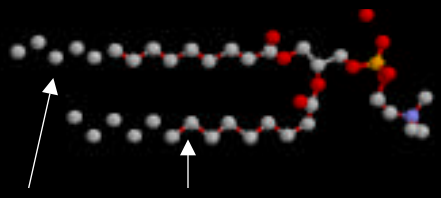
Imbedded Protein



Water Molecules



Typical Lipid Bilayer



Hydropathic Side Chains

Typical Diacylphosphoglyceride

Transmembrane Regions

Step 1: Scan the protein looking for hydrophobic regions

Step 2 : β regions must be 7 to 9 amino acids in length
helical regions must be 20 residues long.

A membrane is 30 Å thick
A 20 amino acid α helix = 31 Å in length, 18Å in diameter
An 8 residue β sheet = 28 Å

Problem - Core of globular proteins is also nonpolar
Software must look for longer regions than usually seen in globular cores (Eisenberg)

Antigenic Regions

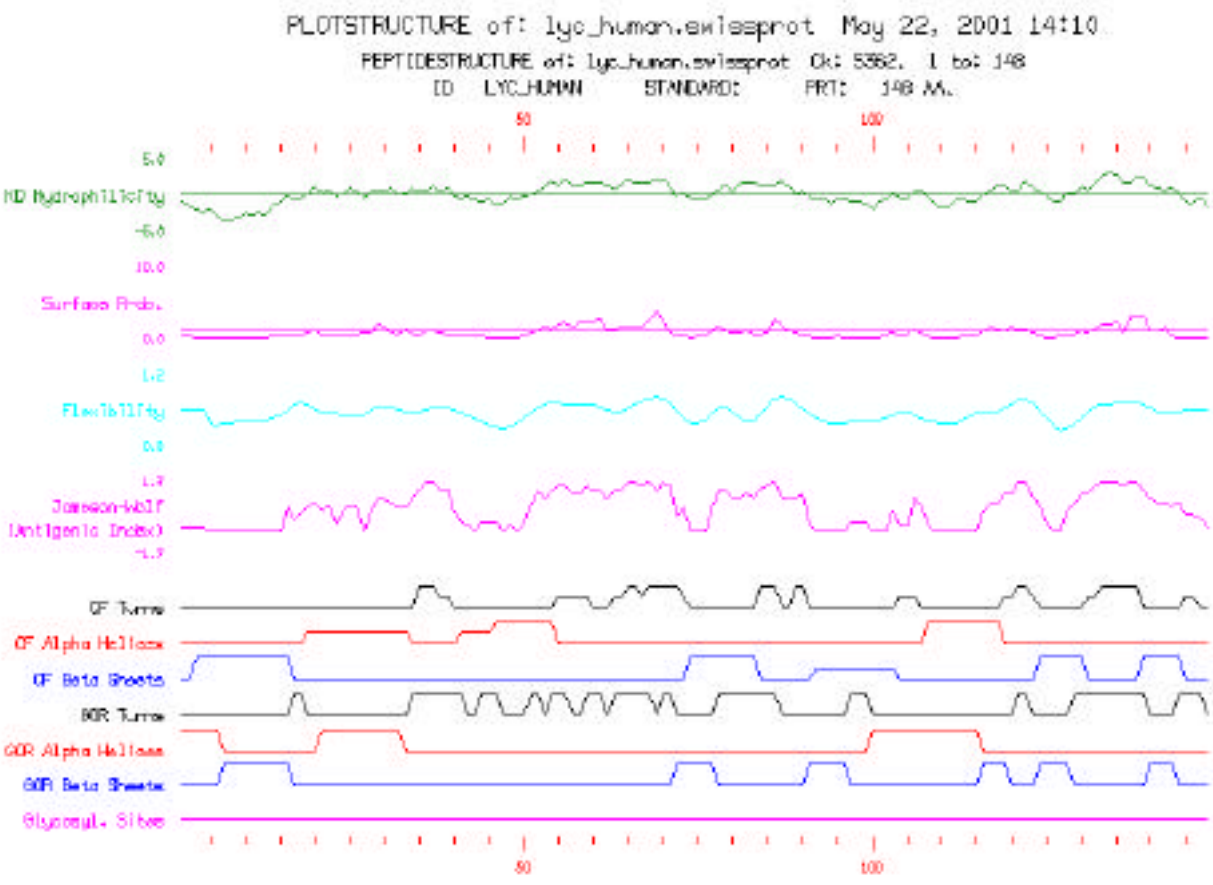
Jameson - Wolf antigenic regions

Combines

- Hydrophilicity, side chain flexibility
- Surface probability
- Turn prediction (CF and GOR)

Regions that score high on all three indices
are regions that will probably induce an antigenic response

Peptide Structure Results



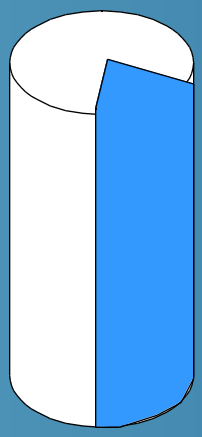
Peptide Structure Results

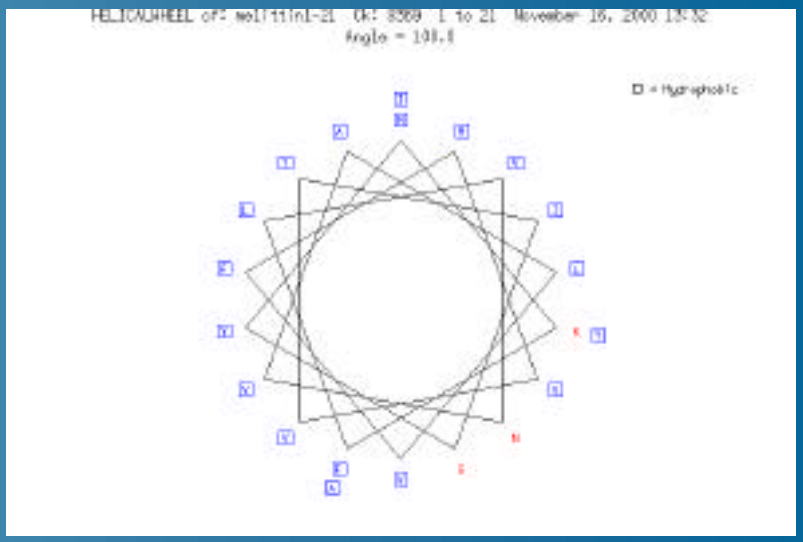
Prediction of human lysozyme secondary structure

Pos	AA	GlycoS	HyPhil	SurfPr	FlexPr	CF-Pred	GORPred	AI-Ind	.
1	K	.	0.100	0.834	1.000	h	H	0.450	
2	V	.	0.980	1.278	1.000	h	H	0.900	
3	F	.	0.400	0.536	1.000	h	H	0.450	
4	E	.	0.843	0.464	1.000	h	H	0.750	
5	R	.	-0.257	0.515	0.980	h	H	-0.300	
6	C	.	0.086	0.601	0.973	h	H	0.300	
7	E	.	1.129	0.680	0.969	h	H	0.600	
8	L	.	0.729	0.501	0.971	h	H	0.600	
9	A	.	-0.457	0.771	0.978	h	H	-0.600	
.	.	.	.	.	.	.	.	.	
16	G	.	1.000	0.599	0.987	.	T	1.000	
17	M	.	0.629	0.479	0.989	T	T	1.400	
18	D	.	0.629	1.139	1.005	T	T	1.700	
19	G	.	1.229	1.139	1.015	T	T	1.700	
20	Y	.	0.529	0.807	1.017	t	T	1.350	

Actually it’s a Helix from Glu 4 to Lys 16 (1LAA)
Turn from 17 to 20

Helical Wheel





Shows which side of helix or sheet is polar or nonpolar
Could indicate which side of helix interacts with membrane

Super Secondary Structure

Zinc Finger - DNA Binding
Helix-Turn-Helix regions - DNA binding
Coiled-Coil - Transmembrane regions
Signal Peptides

Post Translational Modification Sites

- Phosphorylation
- Glycosylation
- Myristoylation (CH_3-$(CH_2)_{12}$-CO_2-N)

Searching the Prosite Database using Motif can discover these

Nucleotide Secondary Structure

Why Is It Important?

RNA

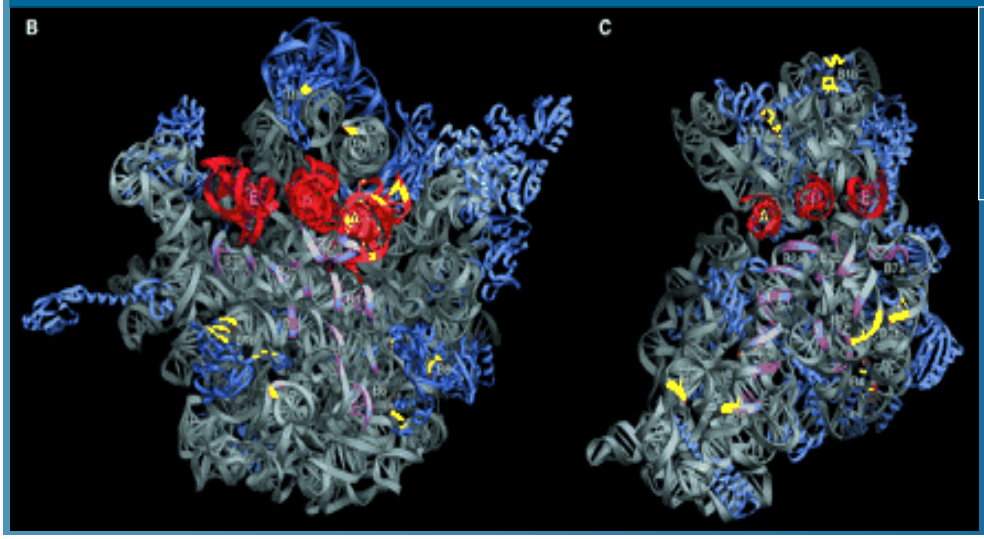
- tRNA
- Ribosomal RNA
- Ribozymes
- snRNA and Splicing (U1, etc)
- RNA World Hypothesis
- Premature Termination sites

DNA

- Translational Start sites
- Transcriptional regulation
- Heat shock promoters
- Telomeres

RNA As Structural Scaffolding

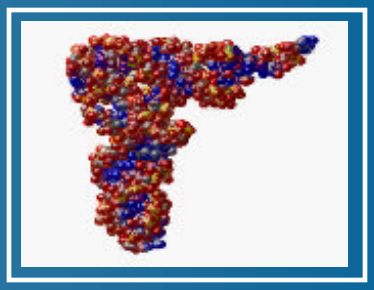
One of the major functions of RNA is as a structural scaffold for molecular machinery. The RNA forms the general shape, and the proteins attach to the RNA to create the functional macromolecule.



RNA acts as scaffolding for proteins

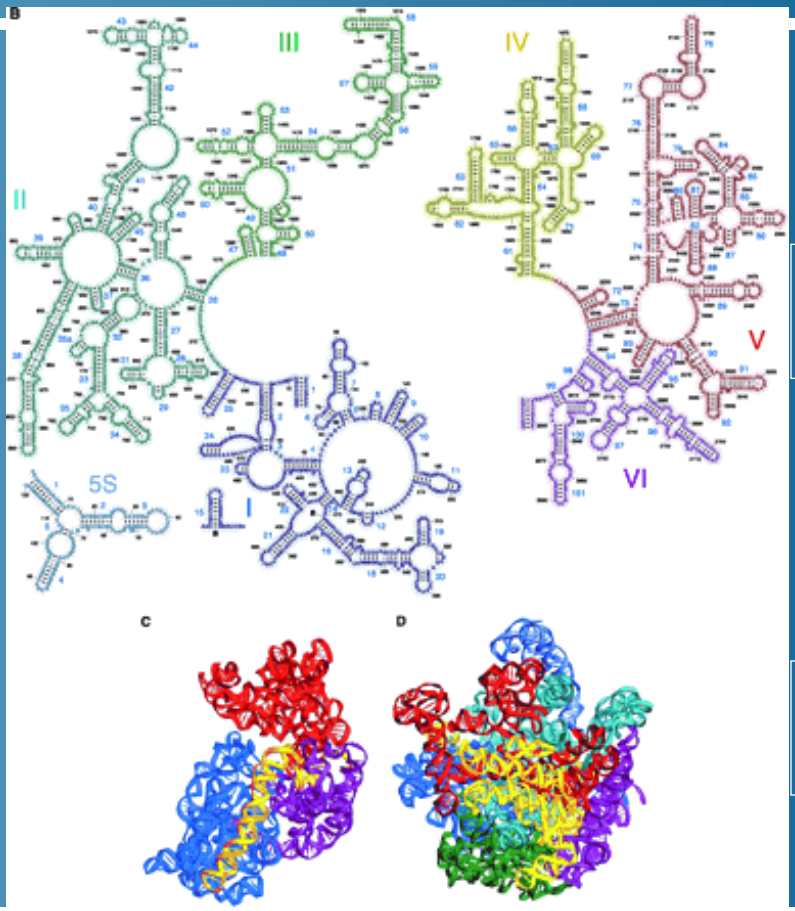
RNA Secondary Structure

AGCGGAUUUAGCUC
AAGUUGGGAGAGCG
ACCAGACUGAAGAU
ACUGGAGGUCCUGU
AGUUCGAUCCACAG
AAAUUCGCACCA



The goal is to take a primary sequence and
Predict the secondary and tertiary structures

Goal of RNA Folding



Secondary Structure

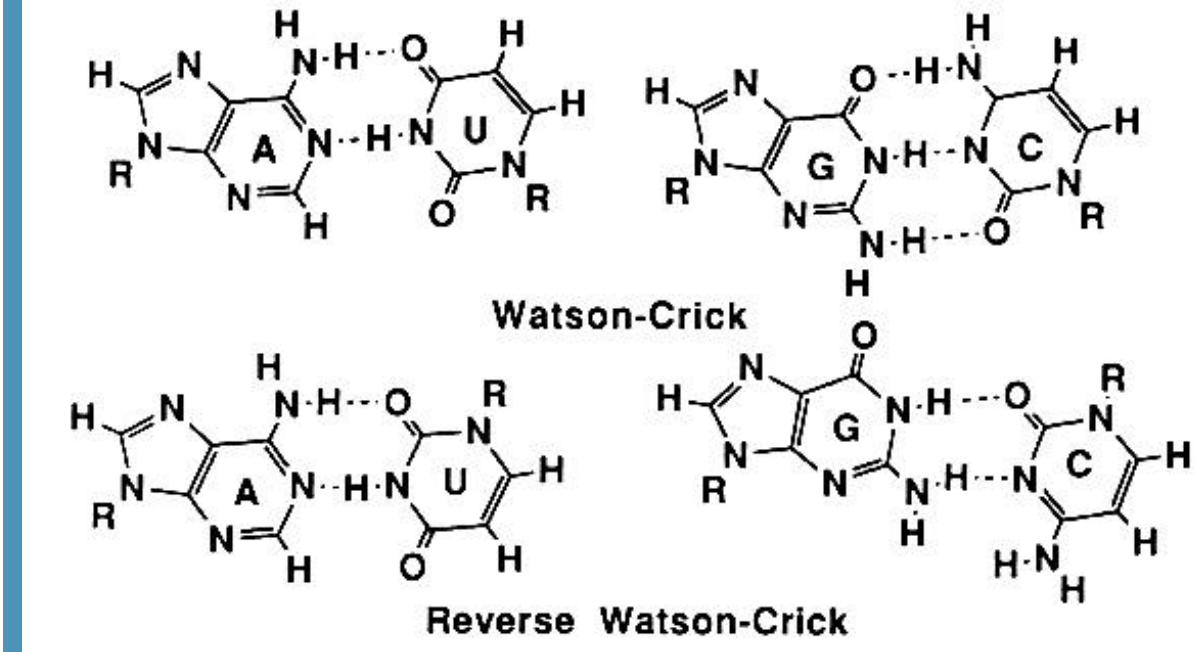
Tertiary Structure

Difficulties in Predicting Structure

The software can predict 2 dimensional interactions but cannot predict 3 dimensional folding. There is a lot more going on than just Watson-Crick H-Bonding

- Non standard base pairing
- Pseudoknots
- Post-transcriptional modification
- Protein interactions
- Metal ion interactions
- Base Stacking
- Folding direction 5'-> 3'

Non-Standard Base Pairing



Note : Only 2 Hydrogen Bonds

Non-Standard Base Pairing

```
REMARK 102 BASES   G A   4 AND   U A  69 ARE MISPAIRED.
REMARK 102 BASES 2MG A  10 AND   G A  45 ARE MISPAIRED.
REMARK 102 BASES H2U A  16 AND   U A  59 ARE MISPAIRED.
REMARK 102 BASES   G A  18 AND PSU A  55 ARE MISPAIRED.
REMARK 102 BASES   G A  22 AND 7MG A  46 ARE MISPAIRED.
REMARK 102 BASES   G A  24 AND   G A  45 ARE MISPAIRED.
REMARK 102 BASES M2G A  26 AND   A A  44 ARE MISPAIRED.
REMARK 103
REMARK 103 THERE ARE NON-WATSON-CRICK HYDROGEN BONDS BETWEEN THE
REMARK 103 FOLLOWING ATOMS:
REMARK 103 N1   G A 4   AND O2   U A 69
REMARK 103 O6   G A 4   AND N3   U A 69
REMARK 103 O2   U A 8   AND N6   A A 14
REMARK 103 N3   U A 8   AND N7   A A 14
REMARK 103 O6  2MG A 10  AND N2   G A 45
REMARK 103 N2   G A 15  AND N3   C A 48
REMARK 103 N1   G A 15  AND O2   C A 48
REMARK 103 N3  H2U A 16  AND O2   U A 59
REMARK 103 O2  H2U A 16  AND N3   U A 59
REMARK 103 N1   G A 18  AND O4  PSU A 55
REMARK 103 O6   G A 22  AND N2  7MG A 46
REMARK 103 N7   G A 22  AND N1  7MG A 46
REMARK 103 O6   G A 24  AND N2   G A 45
REMARK 103 N1  M2G A 26  AND N1   A A 44
```

Entry in PDB Database for a tRNA structure

Zuker's MFold

Step 1: Find all regions that can base pair

Step 2: Calculate lowest energy structure

Lowest Energy is composed of

- Base Pairing
- Stacking
- Loop Destabilizing Energies (unpaired regions)

Structural Prediction Methods

ab initio - from first principles
Calculating the structure using physical interactions

Homology Modeling - Searching for structures with similar sequence homology and trying to superimpose your sequence on the known structure.

Threading (or fingerprinting) - searching a library of folds for sequences that are similar to your query sequence. Build up final structure by combining folds

ab initio methods

This method should work the best because it is simulating the same physical interactions that proteins actually use to fold.

Problems :

1: We don’t know all the physical properties that cause a protein to fold

2. Number of possible interactions is HUGE!

3. Only works for small peptides

Threading Methods

What do you do if sequence similarity is weak?
Look for small regions of similarity with existing structures.

Step 1: Compare sequence against a library of known folds

Step 2: Align the query sequence to each fold

Step 3: Choose the fold with the best score

Step 4: Combine folds together to get complete sequence

Threading Problems

Threading methods can lead to discontinuous structures where no folds can be found

Small folds are found, but they are not connected because the “connections” do not exist in the database of folds.

Homology Modeling

Step 1: Search Protein Database for sequences that may be similar to your sequence.

Step 2: Align similar structures to create a core structure

Step 3: Align query sequence to the “core” to create backbone

Step 4: Analyze for known physical constraints

Step 5: Place side chains and close loops if necessary

Step 6: Energy minimization

Which Method To Use

Small changes, use Force Field methods
> 30% Similarity, use Homology Modeling
20-40% similarity, use threading
< 20% similarity, use *ab initio*

Most software programs will pick the best method based on their initial assessment of the similarity of your sequence to other sequences in the protein structural database.

Similar Structures can come from different sequences

Sequence similarity indicates
Structural similarity

Sequence dissimilarity does not
Mean structural dissimilarity.

Compare Structural Neighbors for a PDB entry

Number of Folds is Finite

Although the total number of proteins is huge.

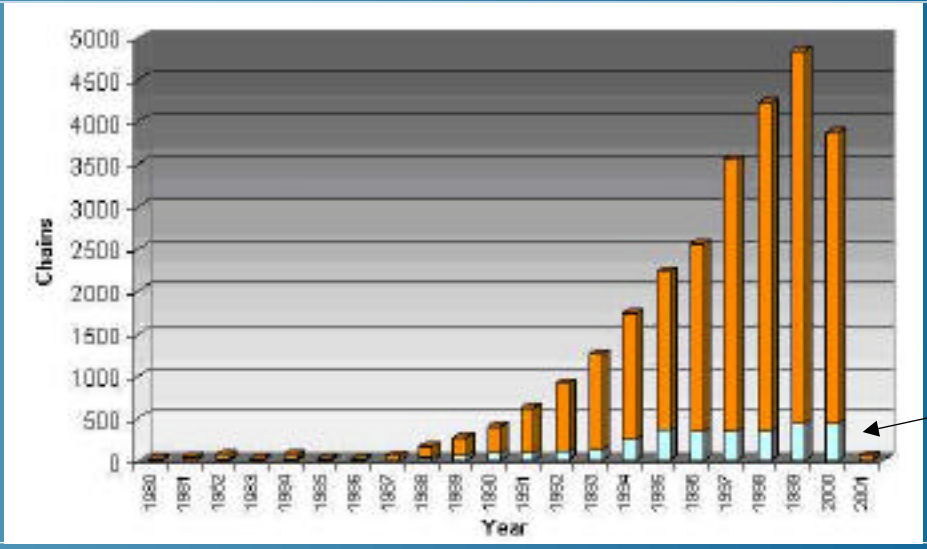
For example
There are 20^{100} different proteins of length 100

The number of possible folds they can create
Is relatively small. Perhaps less than 10,000

You may not have to obtain the structure of every protein to understand the basic structural building blocks.

Discovery of New Folds

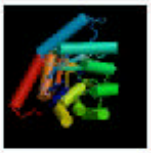
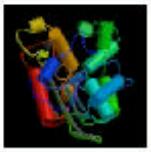
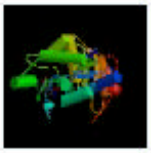
Most new protein structures have common folds



New Folds

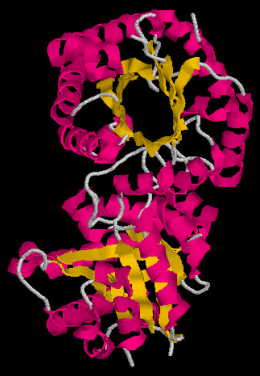
If the number of potential folds is finite, while the number of possible protein sequences is essentially infinite, it may be possible to assign any sequence to a folding pattern.

Folds Can Have Many Functions

Fold		Functions
beta/alpha (TIM)-barrel d1aj2__:3.1 (1.39) A/B		16
alpha/beta-Hydrolases d1cvl__:3.56 (1.39) A/B		9
P-loop containing nucleotide triphosphate hydrolases d1dai__:3.29 (1.39) A/B		6
Ferredoxin-like d2awO__:4.34 (1.39) A+B		6
NAD(P)-binding Rossmann-fold domains d1eny__:3.22 (1.39) A/B		6

1AL7-Oxidase

1TIM-Isomerase



1B5T-Reductase

2TPS-Synthetase

Available Software

- Insight
  - Homology
  - Modeler
- Sybyl
  - Matchmaker
  - Composer
- Swiss Model