

# Introduction to Protein Structure Prediction

BMI/CS 776

[www.biostat.wisc.edu/bmi776/](http://www.biostat.wisc.edu/bmi776/)

Colin Dewey

[cdewey@biostat.wisc.edu](mailto:cdewey@biostat.wisc.edu)

Spring 2015

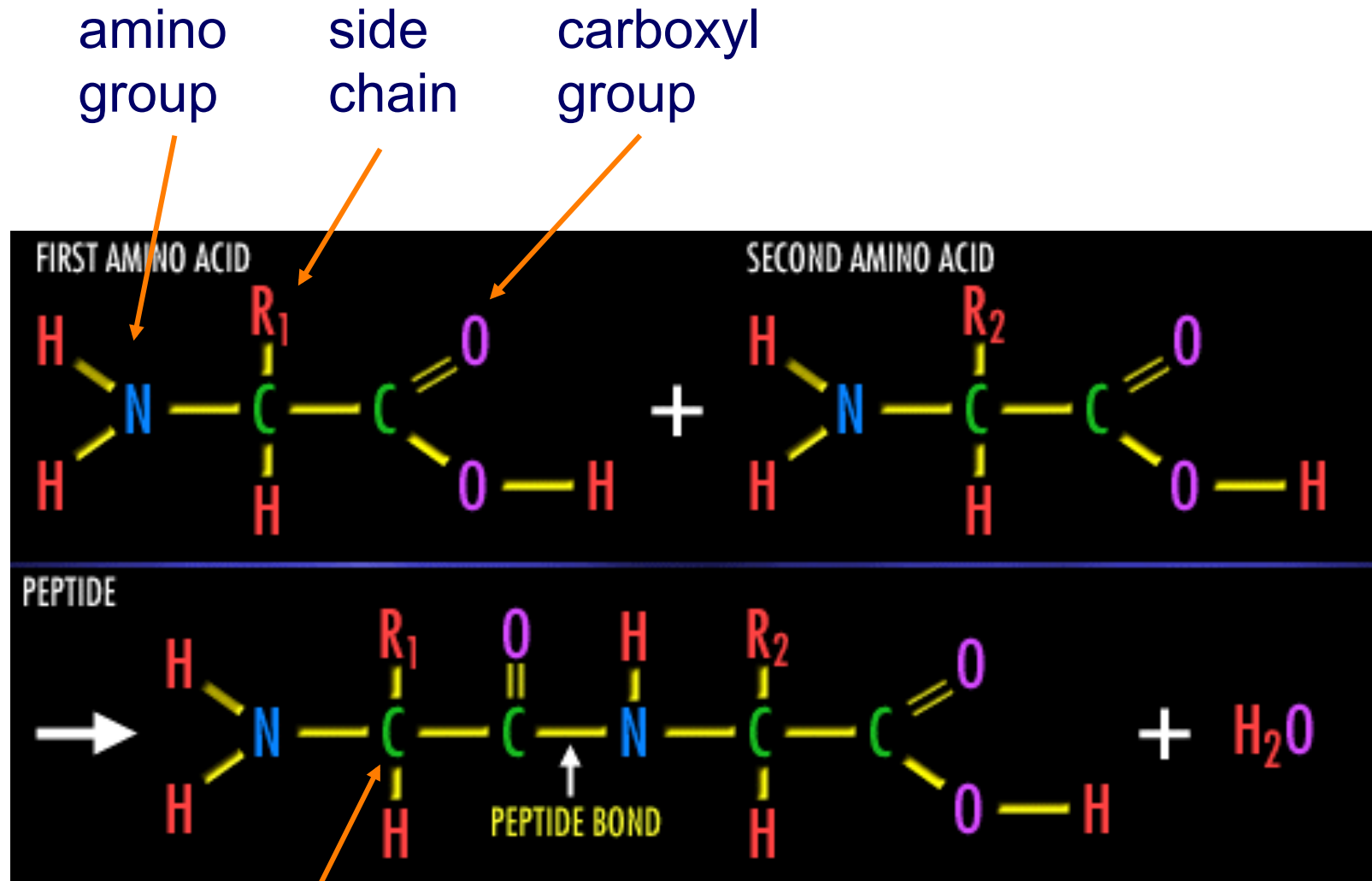
# The Protein Folding Problem

- we know that the function of a protein is determined in large part by its 3D shape (*fold, conformation*)
- can we predict the 3D shape of a protein given only its amino-acid sequence?

# Protein Architecture

- proteins are polymers consisting of amino acids linked by *peptide* bonds
- each amino acid consists of
  - a central carbon atom (alpha-carbon)
  - an amino group,  $\text{NH}_2$
  - a carboxyl group,  $\text{COOH}$
  - a side chain
- differences in side chains distinguish different amino acids

# Amino Acids and Peptide Bonds

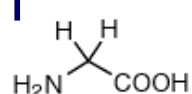


# Amino Acid Side Chains

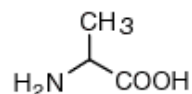
side chains vary in

- shape
- size
- charge
- polarity

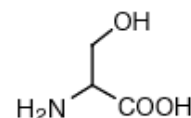
## Small



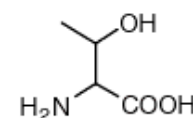
Glycine (Gly, G)  
MW: 57.05



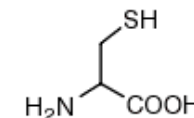
Alanine (Ala, A)  
MW: 71.09



Serine (Ser, S)  
MW: 87.08, pK<sub>a</sub> ~ 16

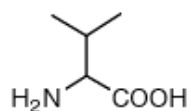


Threonine (Thr, T)  
MW: 101.11, pK<sub>a</sub> ~ 16

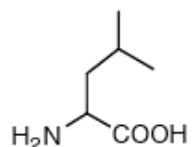


Cysteine (Cys, C)  
MW: 103.15, pK<sub>a</sub> = 8.35

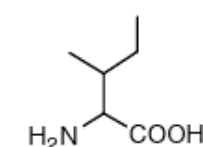
## Hydrophobic



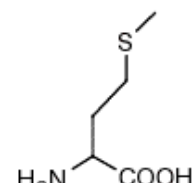
Valine (Val, V)  
MW: 99.14



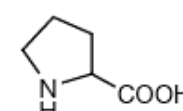
Leucine (Leu, L)  
MW: 113.16



Isoleucine (Ile, I)  
MW: 113.16

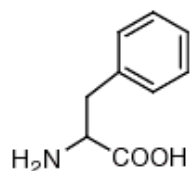


Methionine (Met, M)  
MW: 131.19

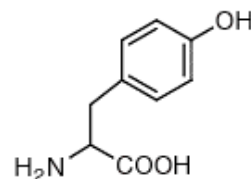


Proline (Pro, P)  
MW: 97.12

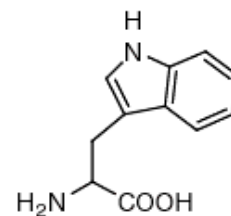
## Aromatic



Phenylalanine (Phe, F)  
MW: 147.18

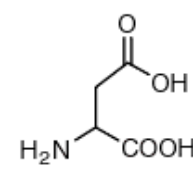


Tyrosine (Tyr, Y)  
MW: 163.18

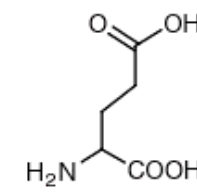


Tryptophan (Trp, W)  
MW: 186.21

## Acidic

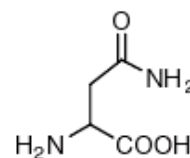


Aspartic Acid (Asp, D)  
MW: 115.09, pK<sub>a</sub> = 3.9

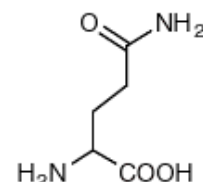


Glutamic Acid (Glu, E)  
MW: 129.12, pK<sub>a</sub> = 4.07

## Amide

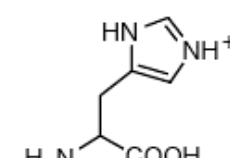


Asparagine (Asn, N)  
MW: 114.11

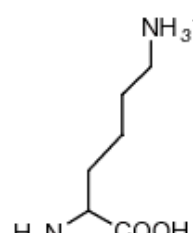


Glutamine (Gln, Q)  
MW: 128.14

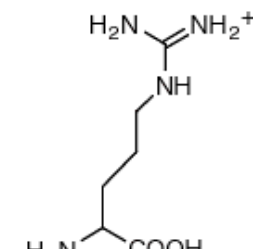
## Basic



Histidine (His, H)  
MW: 137.14, pK<sub>a</sub> = 6.04



Lysine (Lys, K)  
MW: 128.17, pK<sub>a</sub> = 10.79



Arginine (Arg, R)  
MW: 156.19, pK<sub>a</sub> = 12.48

# What Determines Conformation?

- in general, the amino-acid sequence of a protein determines the 3D shape of a protein [Anfinsen et al., 1950s]
- but some qualifications
  - all proteins can be denatured
  - some proteins are inherently *disordered* (i.e. lack a regular structure)
  - some proteins get folding help from *chaperones*
  - there are various mechanisms through which the conformation of a protein can be changed in vivo
    - post-translational modifications such as *phosphorylation*
    - *prions*
    - etc.

# What Determines Conformation?

- Which physical properties of the protein determine its fold?
  - rigidity of the protein backbone
  - interactions among amino acids, including
    - electrostatic interactions
    - van der Waals forces
    - volume constraints
    - hydrogen, disulfide bonds
  - interactions of amino acids with water
    - hydrophobic and hydrophilic residues

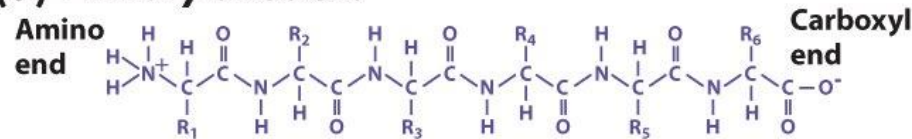
# Levels of Description

- protein structure is often described at four different scales
  - primary structure
  - secondary structure
  - tertiary structure
  - quaternary structure



# Levels of Description

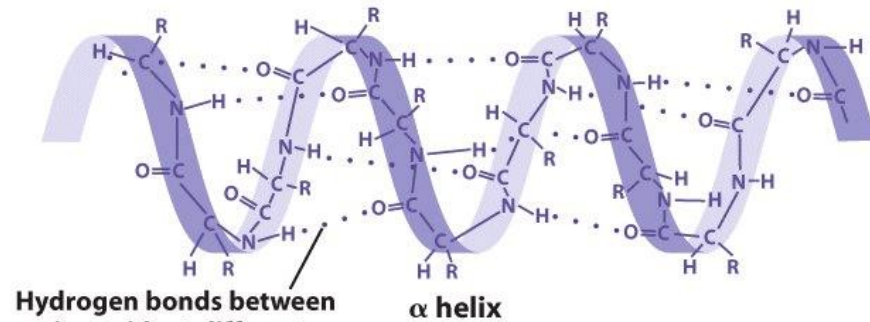
## (a) Primary structure



← the amino acid sequence itself

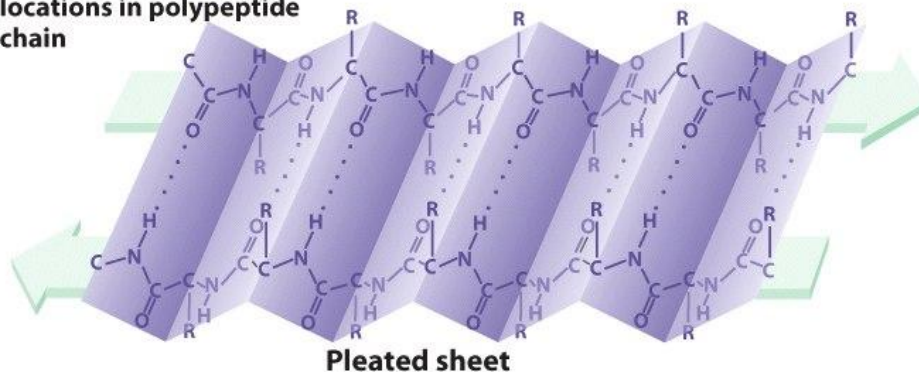
“local” description of structure:  
describes it in terms of certain  
common repeating elements

## (b) Secondary structure



Hydrogen bonds between  
amino acids at different  
locations in polypeptide  
chain

$\alpha$  helix

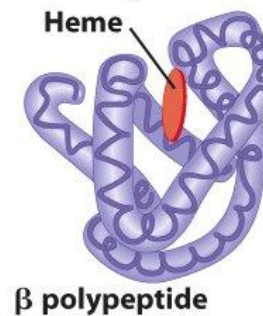


Pleated sheet

3D conformation  
of a polypeptide

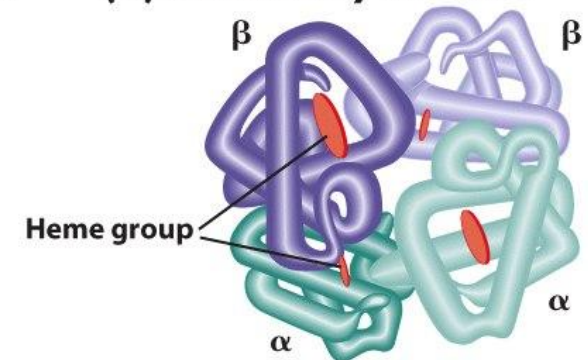
3D conformation  
of a complex of  
polypeptides

## (c) Tertiary structure



$\beta$  polypeptide

## (d) Quaternary structure

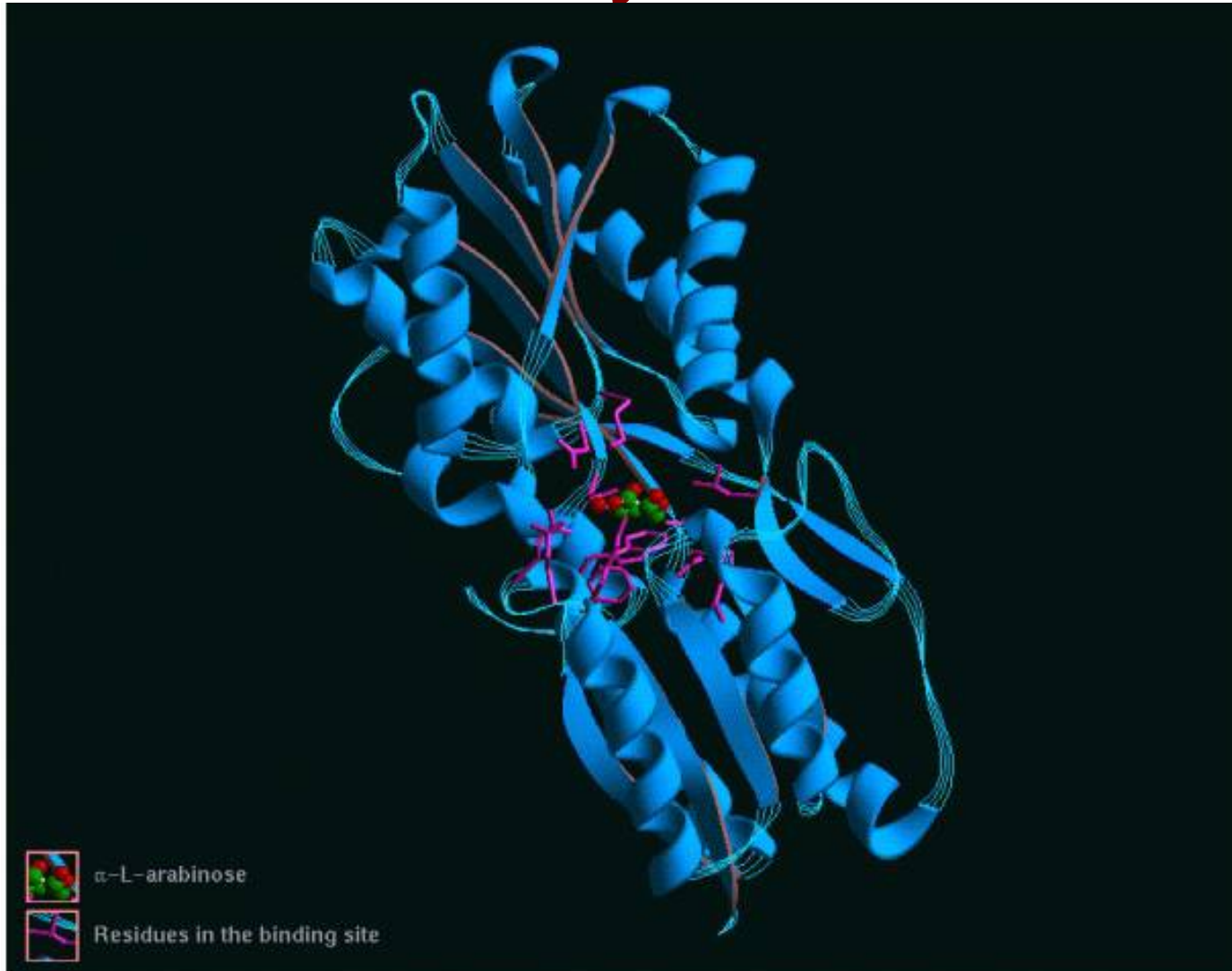


$\alpha$

# Secondary Structure

- secondary structure refers to certain common repeating structures
- it is a “local” description of structure
- two common secondary structures
  - $\alpha$  helices
  - $\beta$  strands/sheets
- a third category, called *coil* or *loop*, refers to everything else

# Ribbon Diagram Showing Secondary Structures



# Determining Protein Structures

- protein structures can be determined experimentally (in most cases) by
  - x-ray crystallography
  - nuclear magnetic resonance (NMR)
- but this is very expensive and time-consuming
- there is a large sequence-structure gap
  - ≈ 550K protein sequences in SwissProt database
  - ≈ 100K protein structures in PDB database
- key question: can we predict structures by computational means instead?

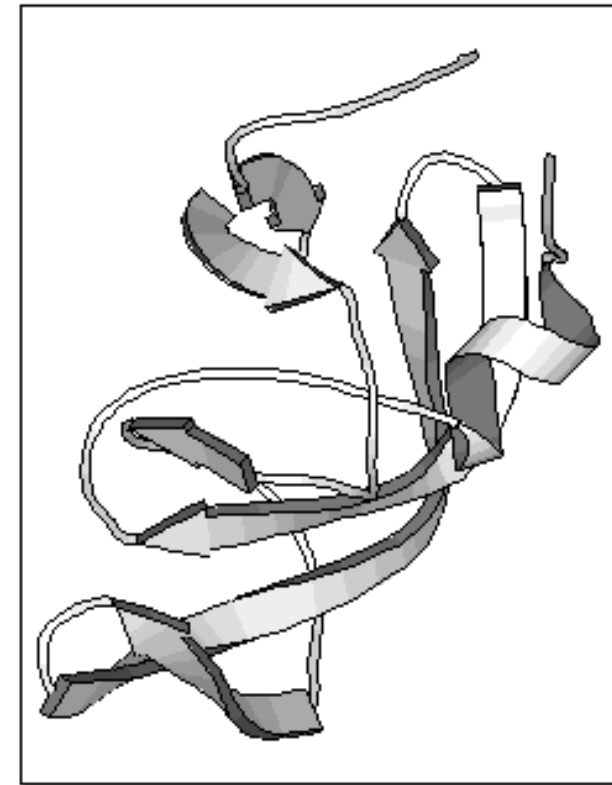
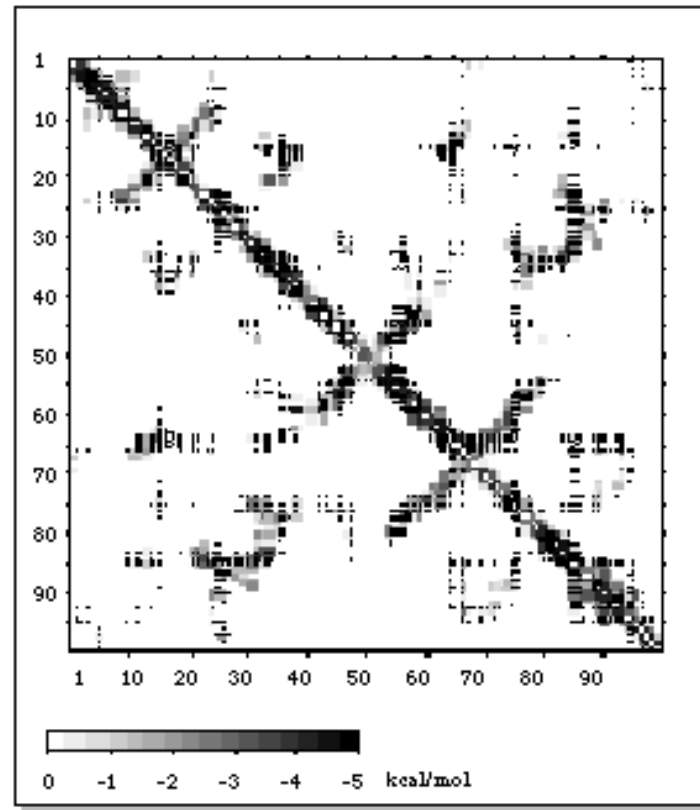
# Types of Protein Structure Predictions

- prediction in 1D
  - secondary structure
  - solvent accessibility (which residues are exposed to water, which are buried)
  - transmembrane helices (which residues span membranes)
- prediction in 2D
  - inter-residue/strand contacts
- prediction in 3D
  - homology modeling
  - fold recognition (e.g. via threading)
  - *ab initio* prediction (e.g. via molecular dynamics)

# Prediction in 1D, 2D and 3D

P	PP P	128	110
Q	QQQY	175	97
I	FFQVI	70	E 60
T	SSIIVR	77	E 69
L	LLSTL	120	E 14
W	WWQED	238	E 81
Q	RKQAK	169	E 97
R	RRRPQ	200	62
P	PPPPP	24	48
L	VVTKF	E 71	E 59
V	VVLII	E 14	E 0
T	TTKEK	E 74	E 69
I	AALIV	E 0	E 0
K	HYKKF	E 90	E 73
I	IILVI	4	E 0
G	EENG	46	41
G	GGGTG	62	53
Q	QKKRR	68	71
L	PPLWW	E 118	E 59
K	VVFKV	E 31	E 73
E	EESKK	E 124	E 95
A	VVGLG	E 1	E 0
L	LLILL	E 29	E 0
L	LLLIV	E 24	E 0
D	DDDDD	49	E 58
T	TTTTT	72	51
G	GGGGG	62	30
A	AAAAA	17	0
D	DDDDD	102	79
D	DDAKE	69	58
T	SSTTV	1	69
V	IIVIV	E 14	E 0
L	VVIVL	E 0	E 0

predicted secondary structure and solvent accessibility



known secondary structure (E = beta strand) and solvent accessibility

# Prediction in 3D

- ***homology modeling***

given: a query sequence  $Q$ , a database of protein structures  
do:

- find protein  $P$  such that
  - structure of  $P$  is known
  - $P$  has high sequence similarity to  $Q$
- return  $P$ 's structure as an approximation to  $Q$ 's structure

- ***fold recognition*** (threading)

given: a query sequence  $Q$ , a database of known folds  
do:

- find fold  $F$  such that  $Q$  can be aligned with  $F$  in a highly compatible manner
- return  $F$  as an approximation to  $Q$ 's structure

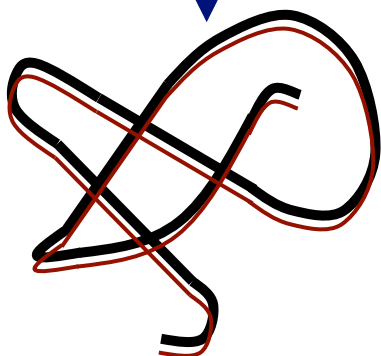
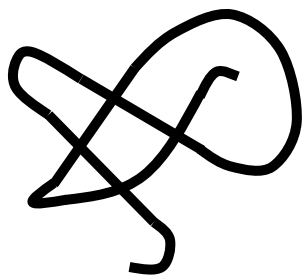
# Prediction in 3D

- ***“fragment assembly”*** (Rosetta)  
given: a query sequence Q, a database of structure fragments  
do:
  - find a set of fragments that Q can be aligned with in a highly compatible manner
  - return fragment assembly as an approximation to Q's structure
- ***molecular dynamics***  
given: a query sequence Q  
do: use laws of Physics to simulate folding of Q

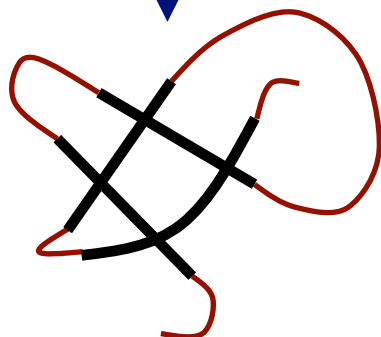
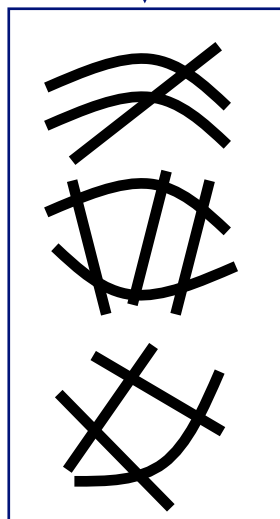


# Prediction in 3D

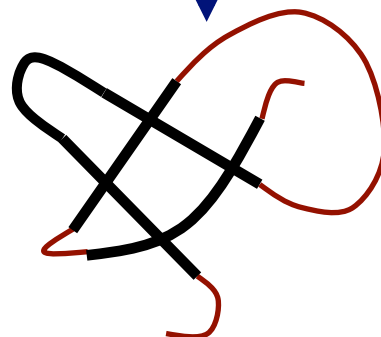
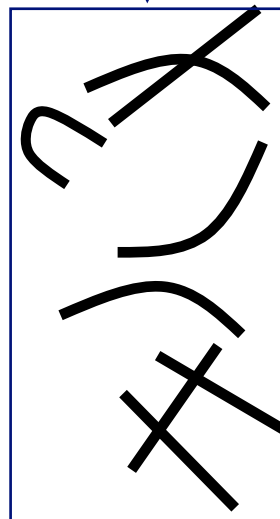
homology  
modeling



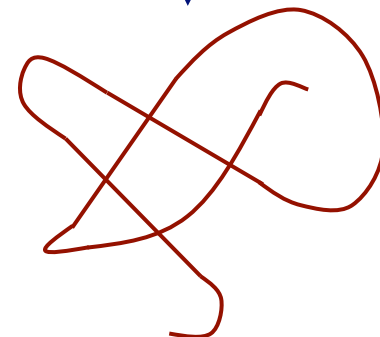
threading



fragment assembly  
(Rosetta)



molecular  
dynamics



# “Citizen science”

- Folding@home

<http://folding.stanford.edu>

Molecular dynamics simulations



- Rosetta@home

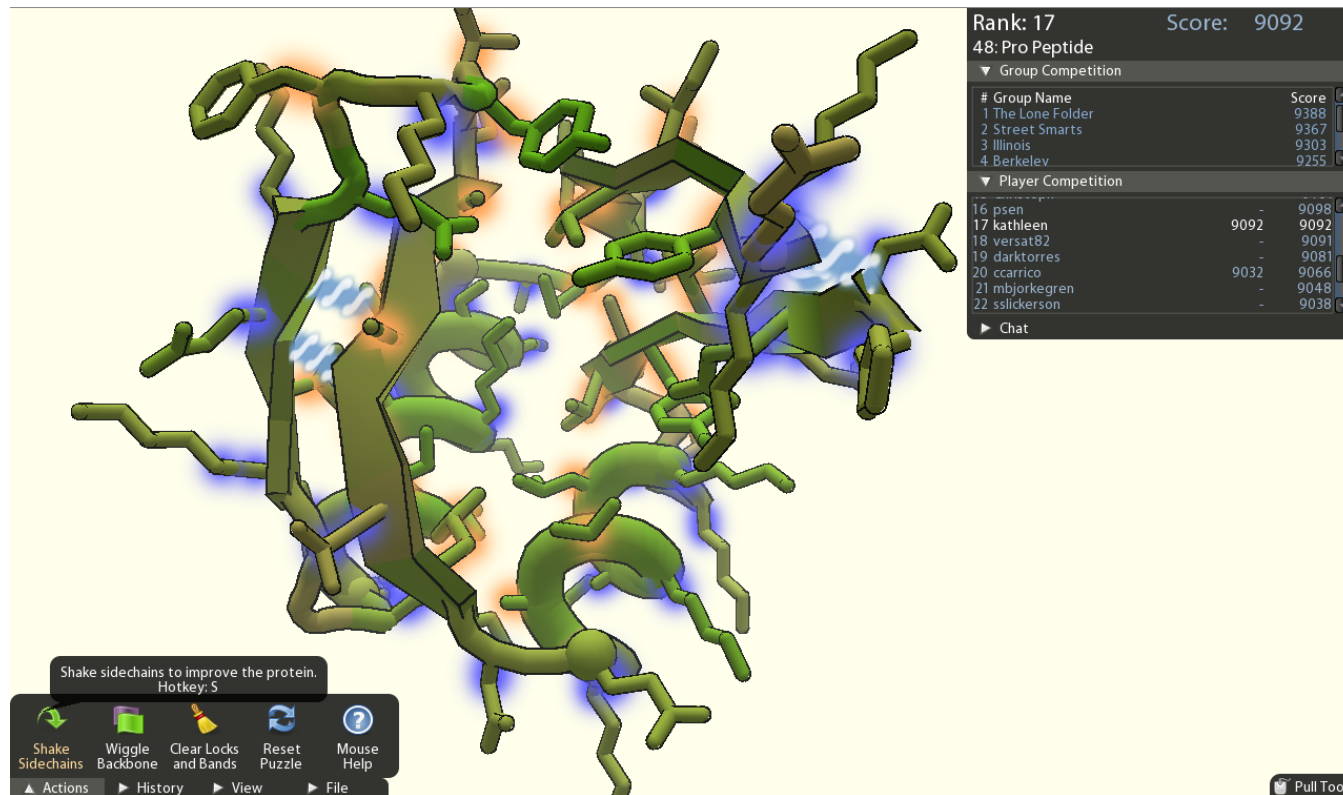
<http://boinc.bakerlab.org>

structure prediction



Volunteer/distributed computing

# Foldit



The screenshot displays the Foldit game interface. On the left, a complex protein structure is shown in a green stick representation, with blue and orange highlights indicating specific regions. A tooltip above the structure reads: "Shake sidechains to improve the protein. Hotkey: S". Below the structure is a toolbar with icons for "Shake Sidechains", "Wiggle Backbone", "Clear Locks and Bands", "Reset Puzzle", and "Mouse Help". At the bottom left, a menu bar includes "Actions", "History", "View", and "File". On the right side, a panel displays the player's rank and score, along with a competition leaderboard.

Rank: 17      Score: 9092  
48: Pro Peptide

▼ Group Competition

#	Group Name	Score
1	The Lone Folder	9388
2	Street Smarts	9367
3	Illinois	9303
4	Berkeley	9255

▼ Player Competition

16	psen	-	9098
17	kathleen	9092	9092
18	versat82	-	9091
19	darktorres	-	9081
20	ccarrico	9032	9066
21	mbjorkegren	-	9048
22	sslickerson	-	9038

► Chat

Pull Tool

<http://fold.it/>