Introduction to Protein Structure Prediction

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2016

Anthony Gitter

gitter@biostat.wisc.edu

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

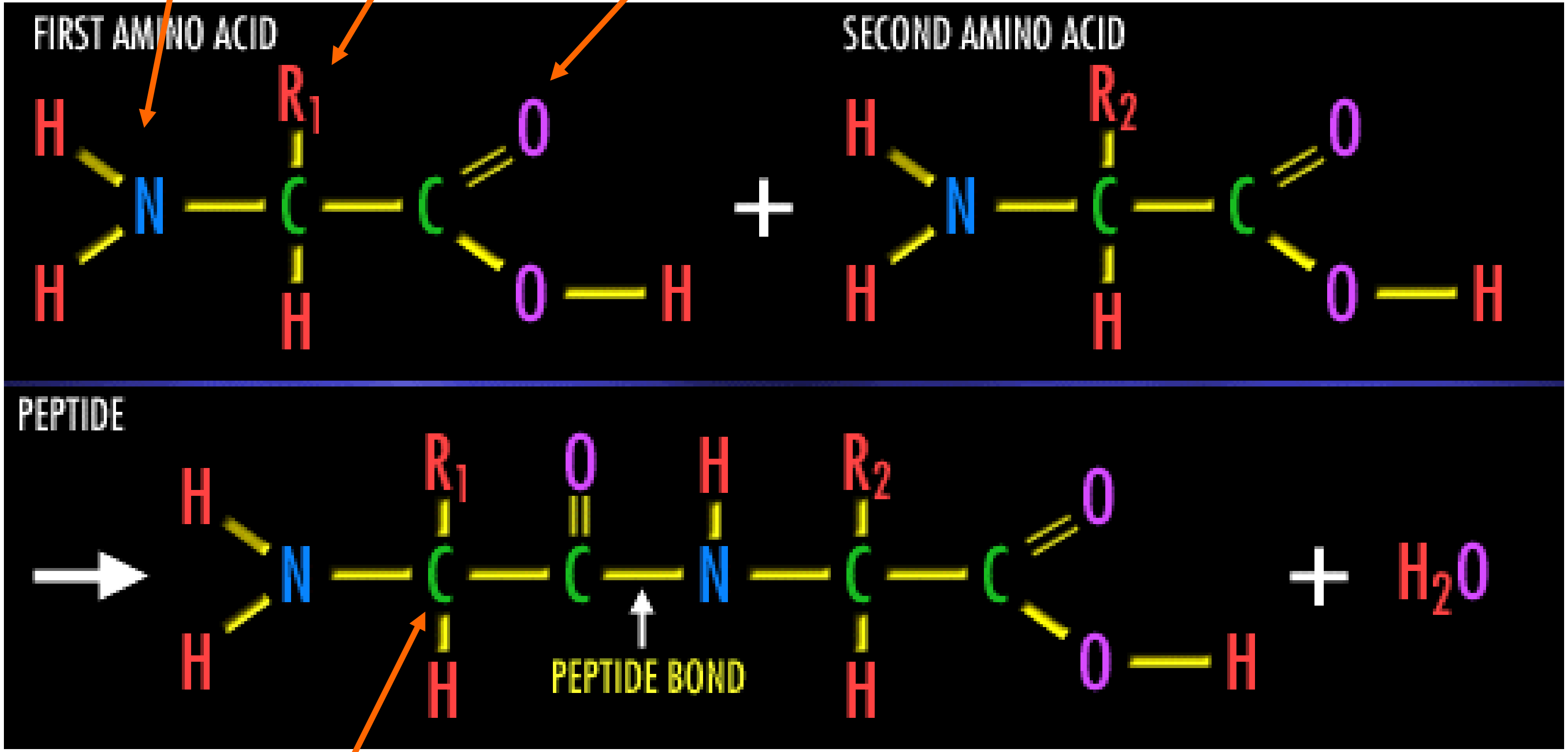
- Proteins are polymers consisting of amino acids linked by *peptide* bonds
- Each amino acid consists of
  - a central carbon atom (α-carbon)
  - an amino group, NH_2
  - a carboxyl group, COOH
  - a side chain
- Differences in side chains distinguish different amino acids

Amino Acids and Peptide Bonds

amino group

side chain

carboxyl group



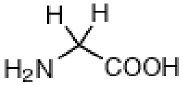
α carbon (common reference point for coordinates of a structure)

Amino Acid Side Chains

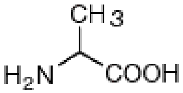
Side chains vary in

- shape
- size
- charge
- polarity

Small

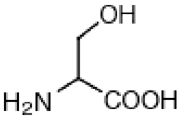


Glycine (Gly, G)
MW: 57.05

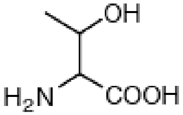


Alanine (Ala, A)
MW: 71.09

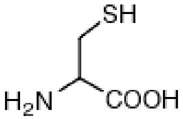
Nucleophilic



Serine (Ser, S)
MW: 87.08, pK_a ~ 16

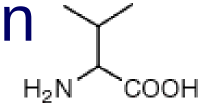


Threonine (Thr, T)
MW: 101.11, pK_a ~ 16

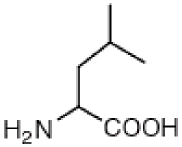


Cysteine (Cys, C)
MW: 103.15, pK_a = 8.35

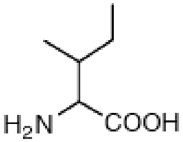
Hydrophobic



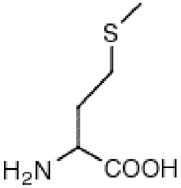
Valine (Val, V)
MW: 99.14



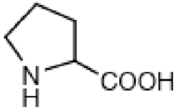
Leucine (Leu, L)
MW: 113.16



Isoleucine (Ile, I)
MW: 113.16

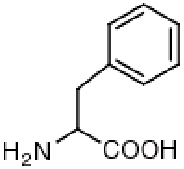


Methionine (Met, M)
MW: 131.19

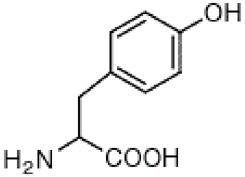


Proline (Pro, P)
MW: 97.12

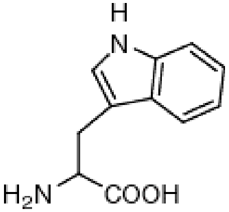
Aromatic



Phenylalanine (Phe, F)
MW: 147.18

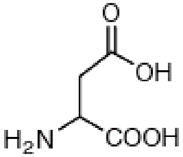


Tyrosine (Tyr, Y)
MW: 163.18

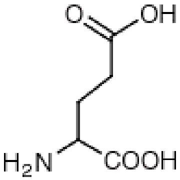


Tryptophan (Trp, W)
MW: 186.21

Acidic

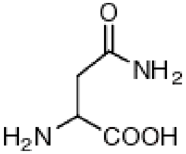


Aspartic Acid (Asp, D)
MW: 115.09, pK_a = 3.9

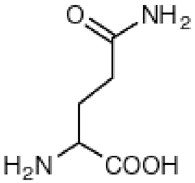


Glutamic Acid (Glu, E)
MW: 129.12, pK_a = 4.07

Amide

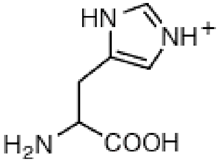


Asparagine (Asn, N)
MW: 114.11

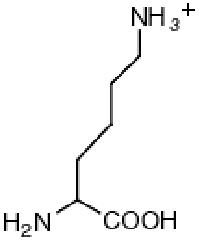


Glutamine (Gln, Q)
MW: 128.14

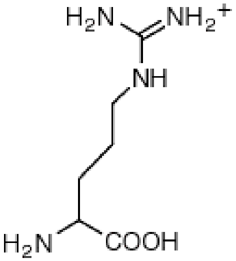
Basic



Histidine (His, H)
MW: 137.14, pK_a = 6.04



Lysine (Lys, K)
MW: 128.17, pK_a = 10.79



Arginine (Arg, R)
MW: 156.19, pK_a = 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.

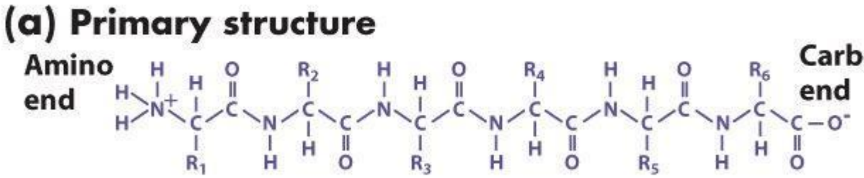
What Determines Conformation?

- 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

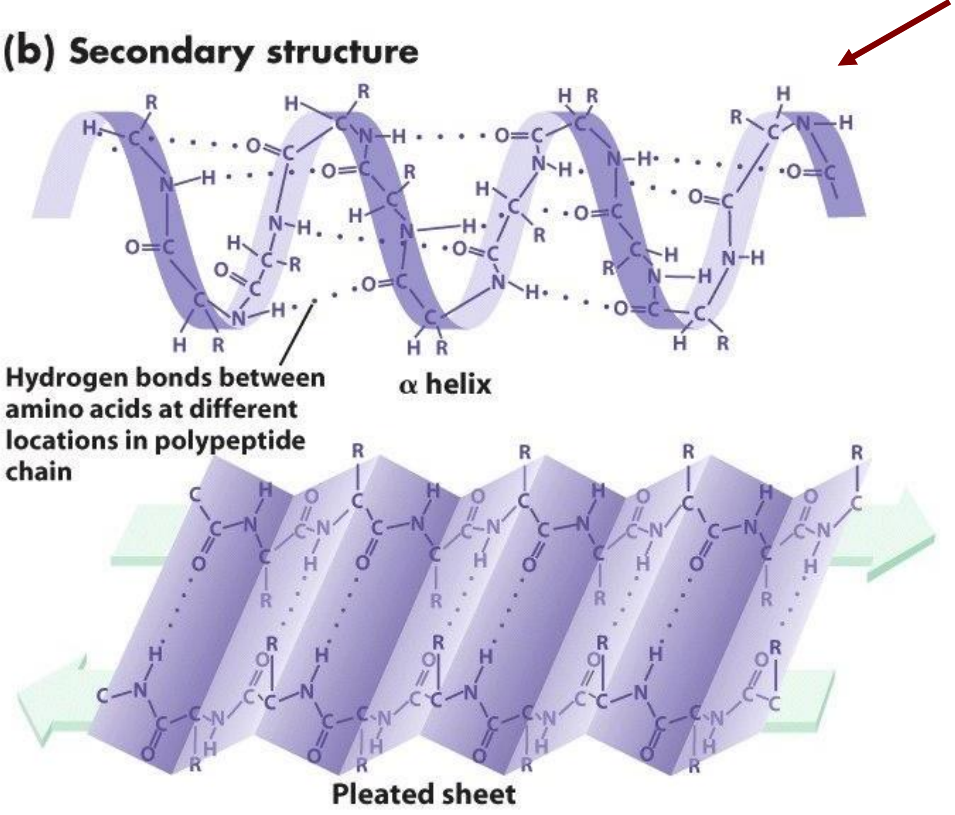
- Protein structure is often described at four different scales
  - primary structure
  - secondary structure
  - tertiary structure
  - quaternary structure

Levels of Description



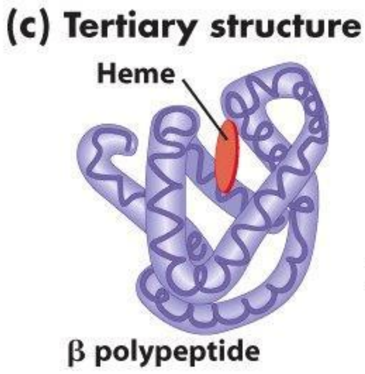
the amino acid sequence itself

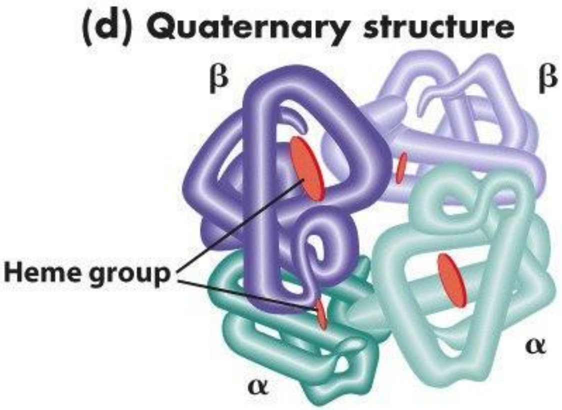
“local” description of structure: describes it in terms of certain common repeating elements


(b) Secondary structure
Hydrogen bonds between amino acids at different locations in polypeptide chain
α helix
Pleated sheet

3D conformation of a polypeptide

3D conformation of a complex of polypeptides

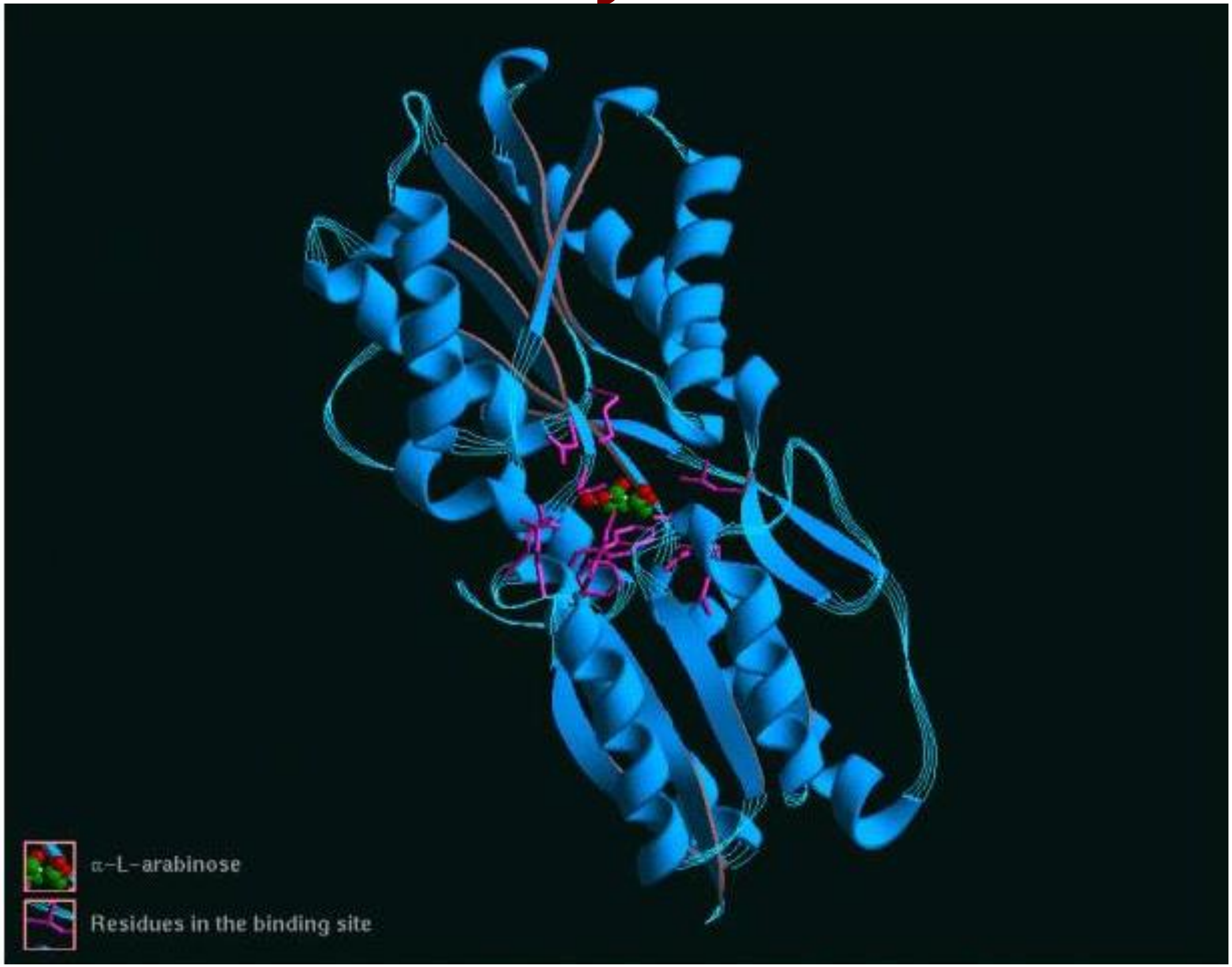




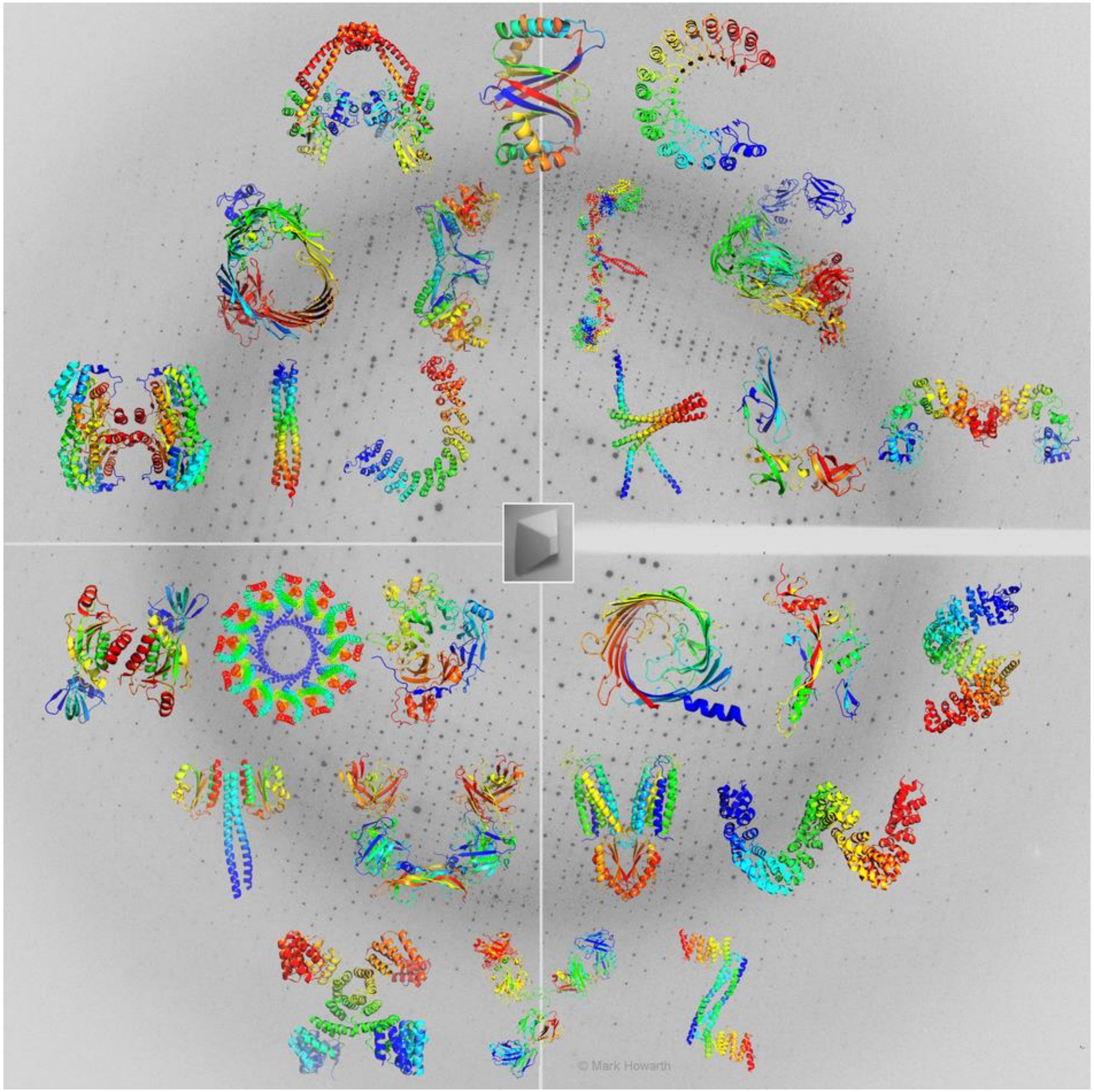
Secondary Structure

- Secondary structure refers to certain common repeating structures
- It is a “local” description of structure
- Two common secondary structure
  - α helices
  - β strands/sheets
- A third category, called *coil* or *loop*, refers to everything else

Ribbon Diagram Showing Secondary Structures



Diversity of Protein Structures



Howarth *Nature Structural & Molecular Biology* 2015

Determining Protein Structures

- Protein structures can be determined experimentally (in most cases) by
  - x-ray crystallography
  - nuclear magnetic resonance (NMR)
  - cryo-electron microscopy (cryo-EM)
- 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?

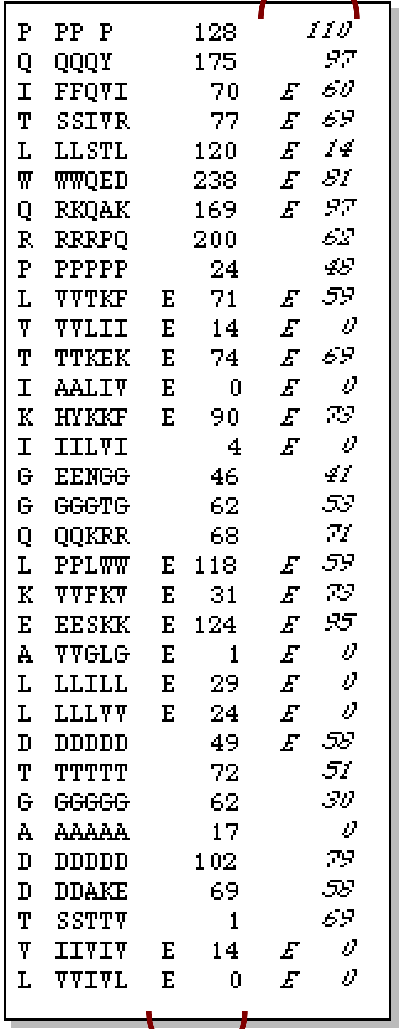
Types of Protein Structure Predictions

- 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

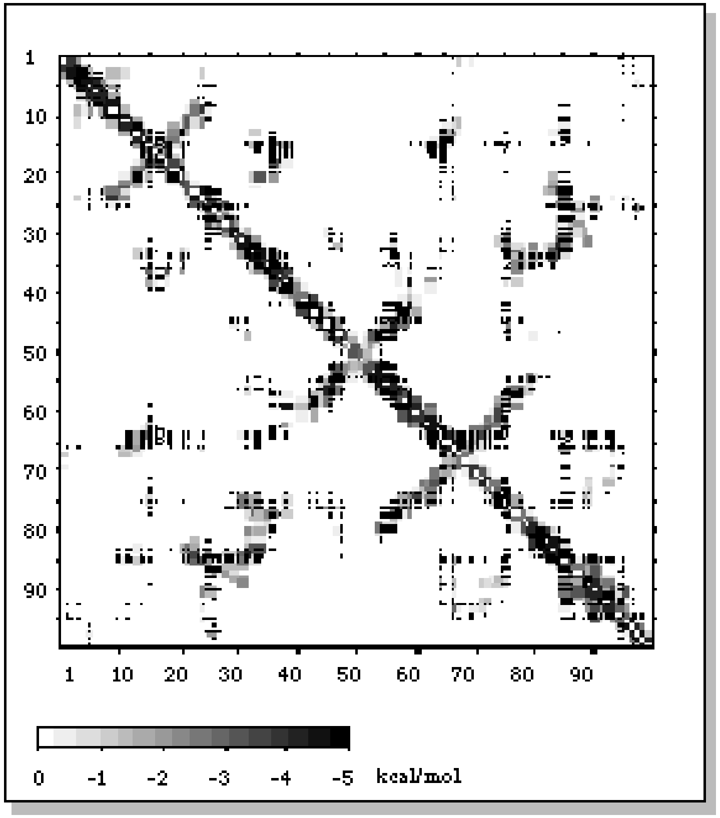
- Prediction in 3D
  - homology modeling
  - fold recognition (e.g. via threading)
  - *ab initio* prediction (e.g. via molecular dynamics)

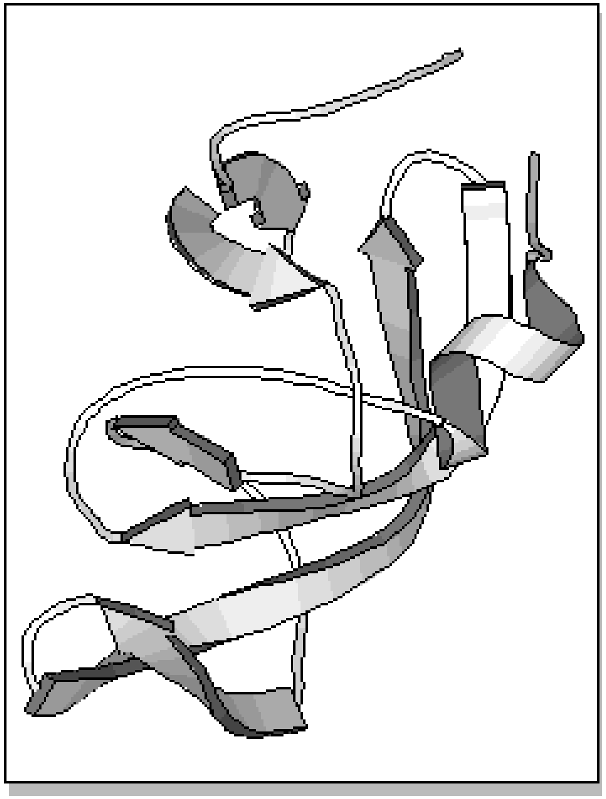
Prediction in 1D, 2D and 3D



P	PP P		128		110
Q	QQQY		175		97
I	FFQVI		70	E	60
T	SSIVR		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	EENGG		46		41
G	GGGTG		62		53
Q	QQKRR		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	LLLVV	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

Figure from B. Rost, "Protein Structure in 1D, 2D, and 3D", *The Encyclopaedia of Computational Chemistry,* 1998

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:

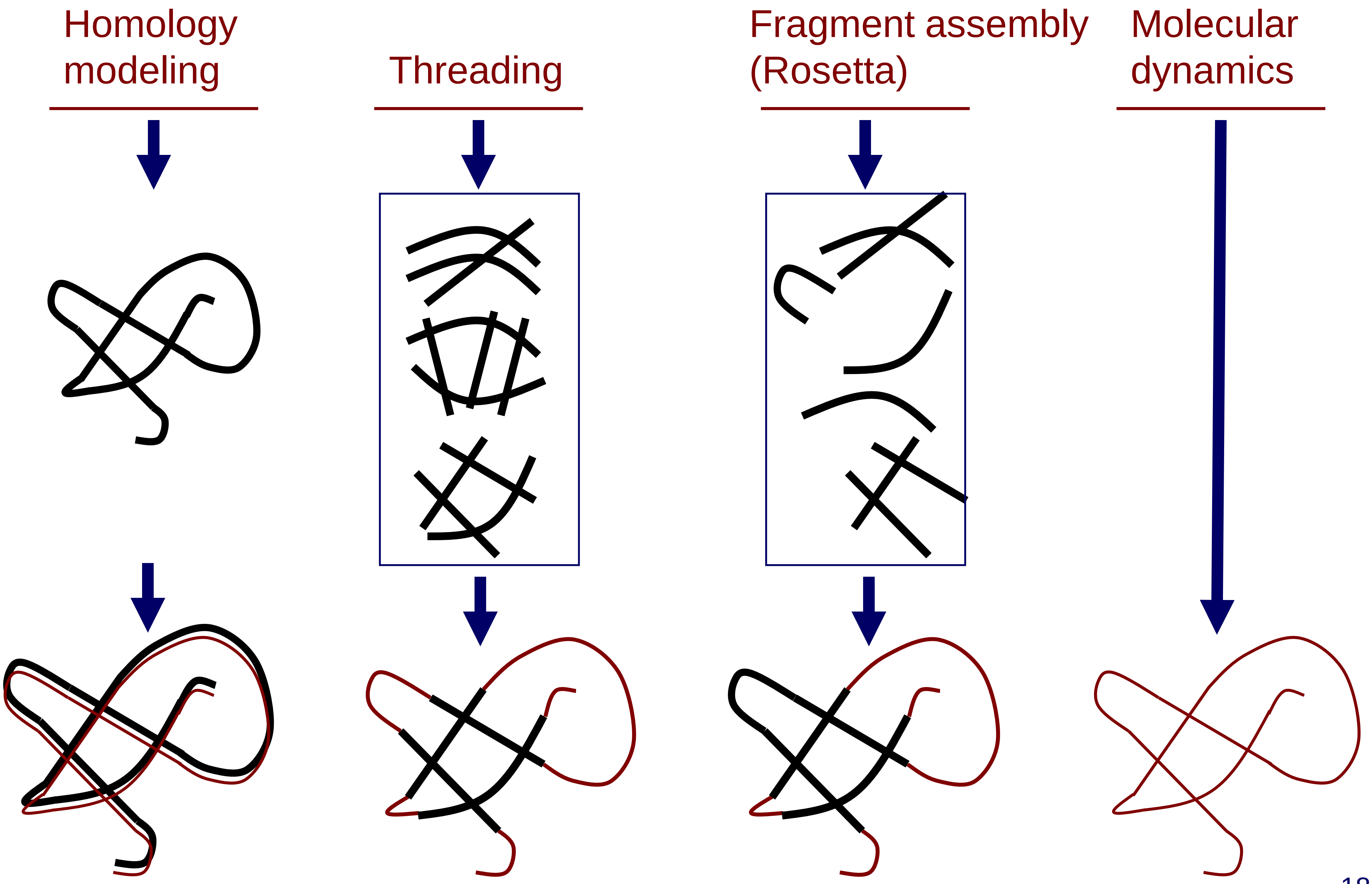
  - find a set of fragments that *Q* can be aligned with in a highly compatible manner
  - return fragment assembly as an approximation to *Q*'s structure

- ***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

Folding@home

Rosetta@home

Protein Folding, Design, and Docking

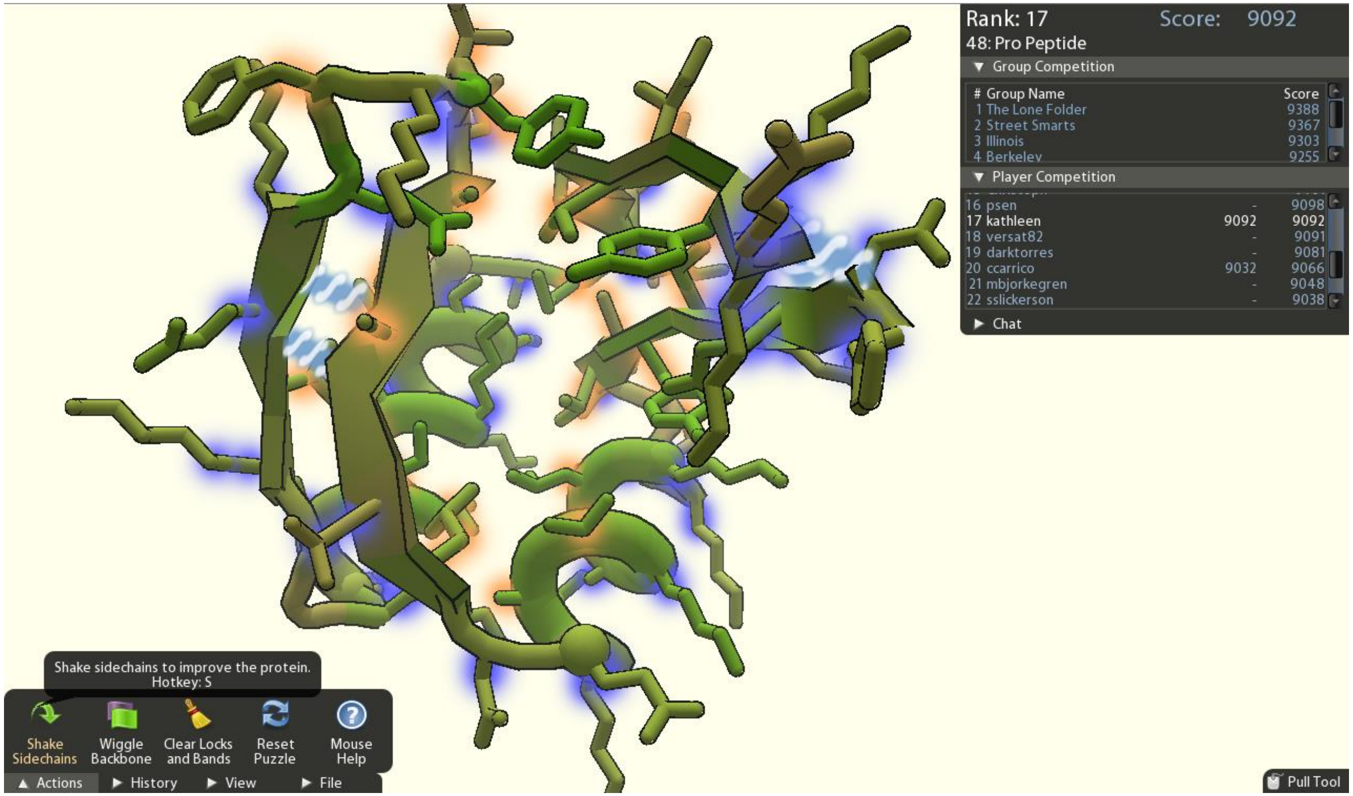
- Rosetta@home

http://boinc.bakerlab.org

Structure prediction

Volunteer/distributed computing

Foldit



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

Shake sidechains to improve the protein.
Hotkey: S

http://fold.it/