

Introduction to Protein Structure Prediction

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2017

Anthony Gitter

gitter@biostat.wisc.edu

These slides, excluding third-party material, are licensed under [CC BY-NC 4.0](https://creativecommons.org/licenses/by-nc/4.0/) by Mark Craven, Colin Dewey, and Anthony Gitter

The Protein Folding Problem

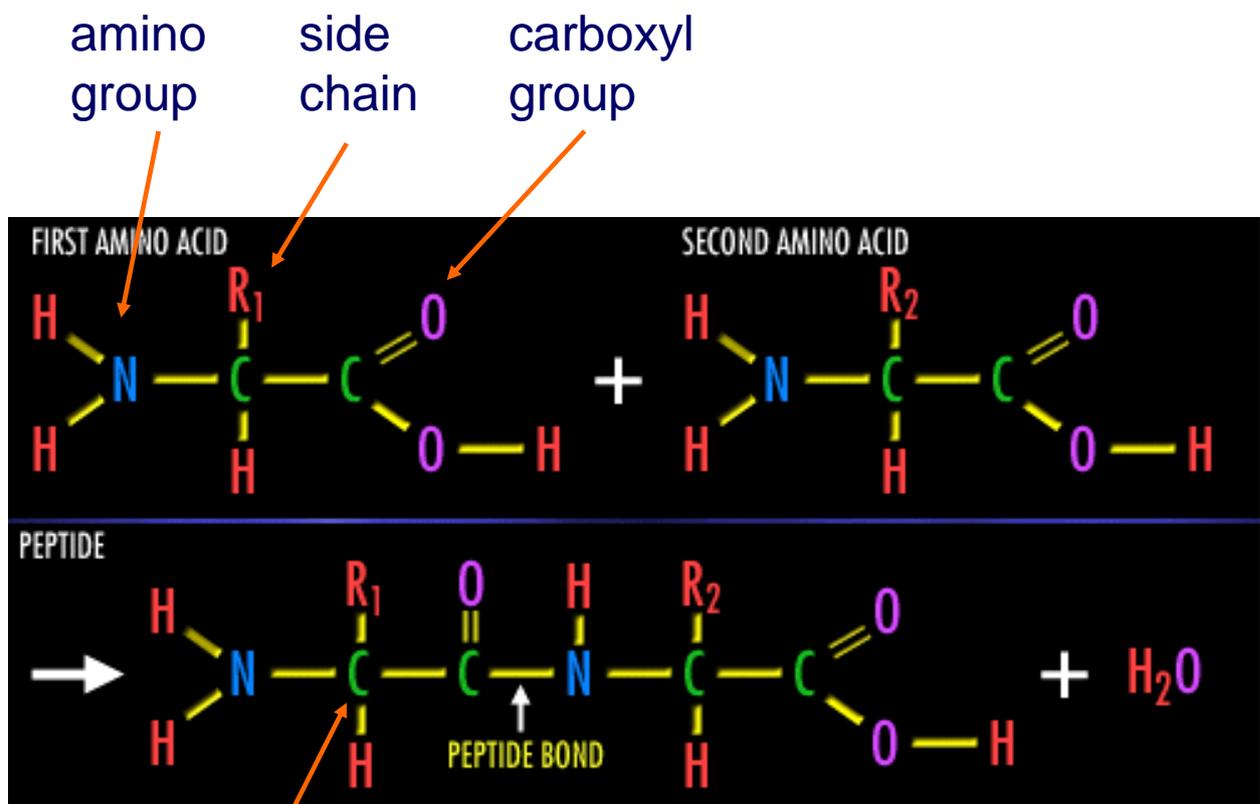
- We know that the function of a protein is determined in large part by its 3D shape (*fold, conformation*)
- Can we predict the 3D shape of a protein given only its amino-acid sequence?

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

3

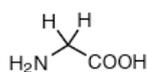
Amino Acids and Peptide Bonds



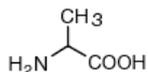
4

Amino Acid Side Chains

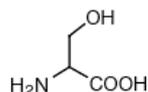
Small



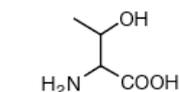
Glycine (Gly, G)
MW: 57.05



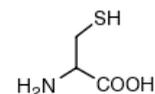
Alanine (Ala, A)
MW: 71.09



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

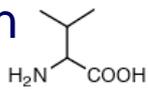


Threonine (Thr, T)
MW: 101.11, pK_a ~ 16

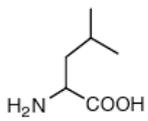


Cysteine (Cys, C)
MW: 103.15, pK_a = 8.35

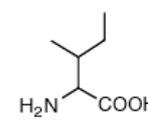
Hydrophobic



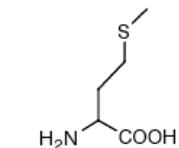
Valine (Val, V)
MW: 99.14



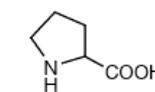
Leucine (Leu, L)
MW: 113.16



Isoleucine (Ile, I)
MW: 113.16

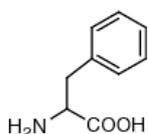


Methionine (Met, M)
MW: 131.19

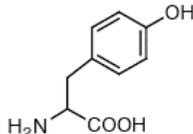


Proline (Pro, P)
MW: 97.12

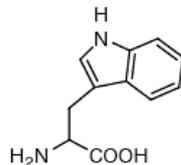
Aromatic



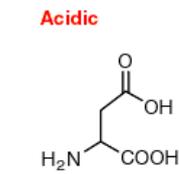
Phenylalanine (Phe, F)
MW: 147.18



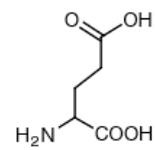
Tyrosine (Tyr, Y)
MW: 163.18



Tryptophan (Trp, W)
MW: 186.21

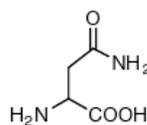


Aspartic Acid (Asp, D)
MW: 115.09, pK_a = 3.9

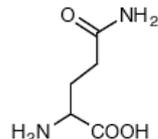


Glutamic Acid (Glu, E)
MW: 129.12, pK_a = 4.07

Amide

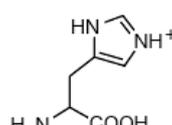


Asparagine (Asn, N)
MW: 114.11

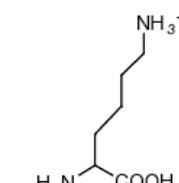


Glutamine (Gln, Q)
MW: 128.14

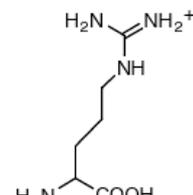
Basic



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

Side chains vary in

- shape
- size
- charge
- polarity

5

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.

6

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

7

Levels of Description

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

8

Levels of Description

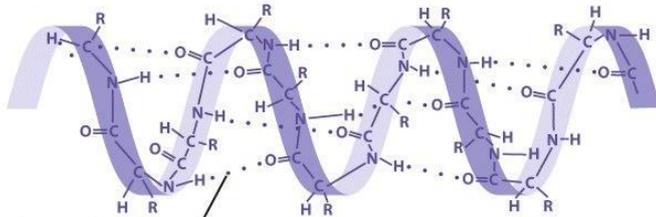
(a) Primary structure



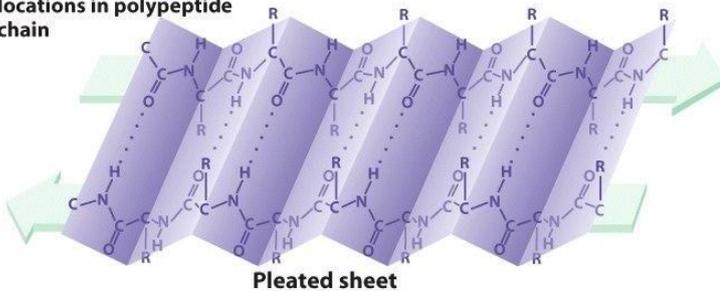
← the amino acid sequence itself

“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

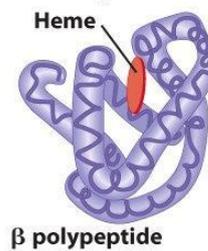


Pleated sheet

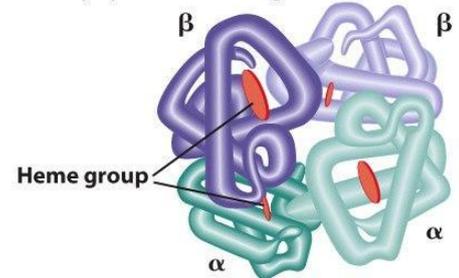
3D conformation
of a polypeptide

3D conformation
of a complex of
polypeptides

(c) Tertiary structure



(d) Quaternary structure



9

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 (pleated sheet on previous slide)
- A third category, called *coil* or *loop*, refers to everything else

Secondary Structure

“Is the neural network an essential tool for the most accurate secondary structure prediction?”

- Burkhard Rost, 1998

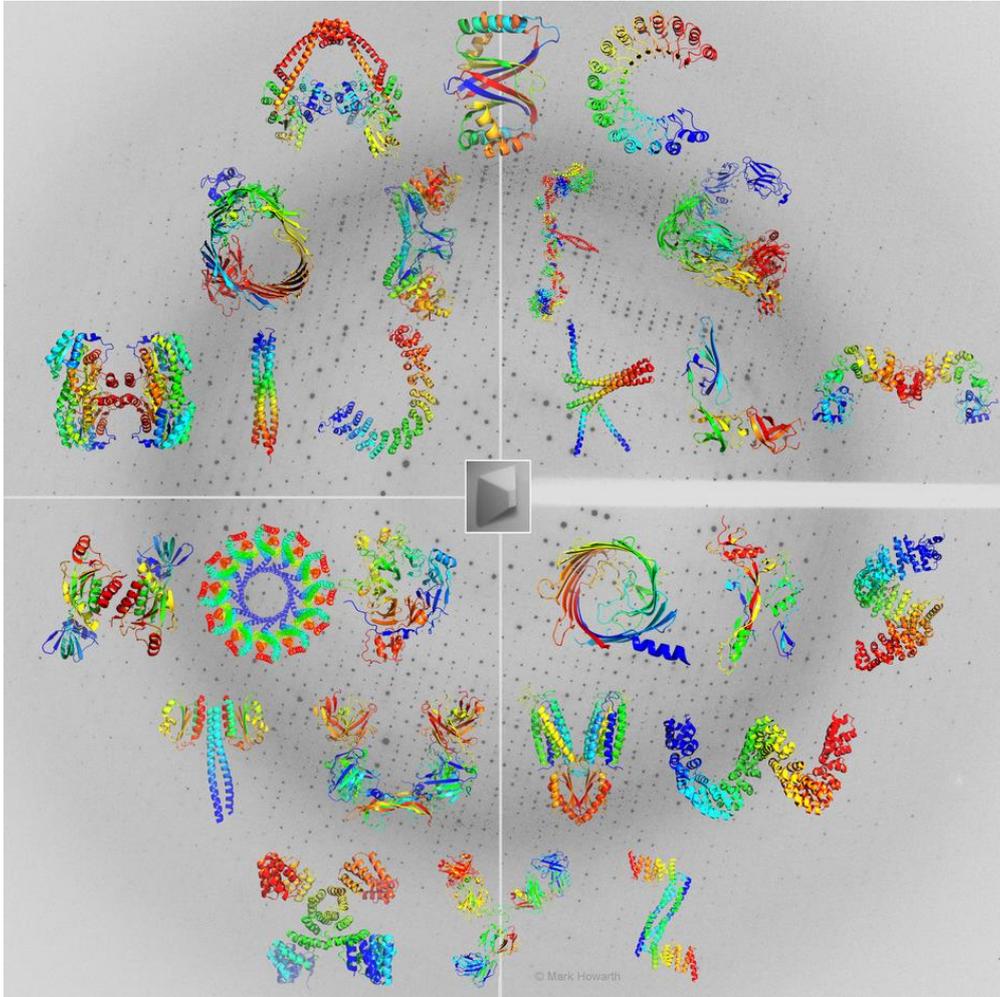
11

Ribbon Diagram Showing Secondary Structures



12

Diversity of Protein Structures



Howarth *Nature*
Structural &
Molecular
Biology 2015

13

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?

14

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)

15

Prediction in 1D, 2D and 3D

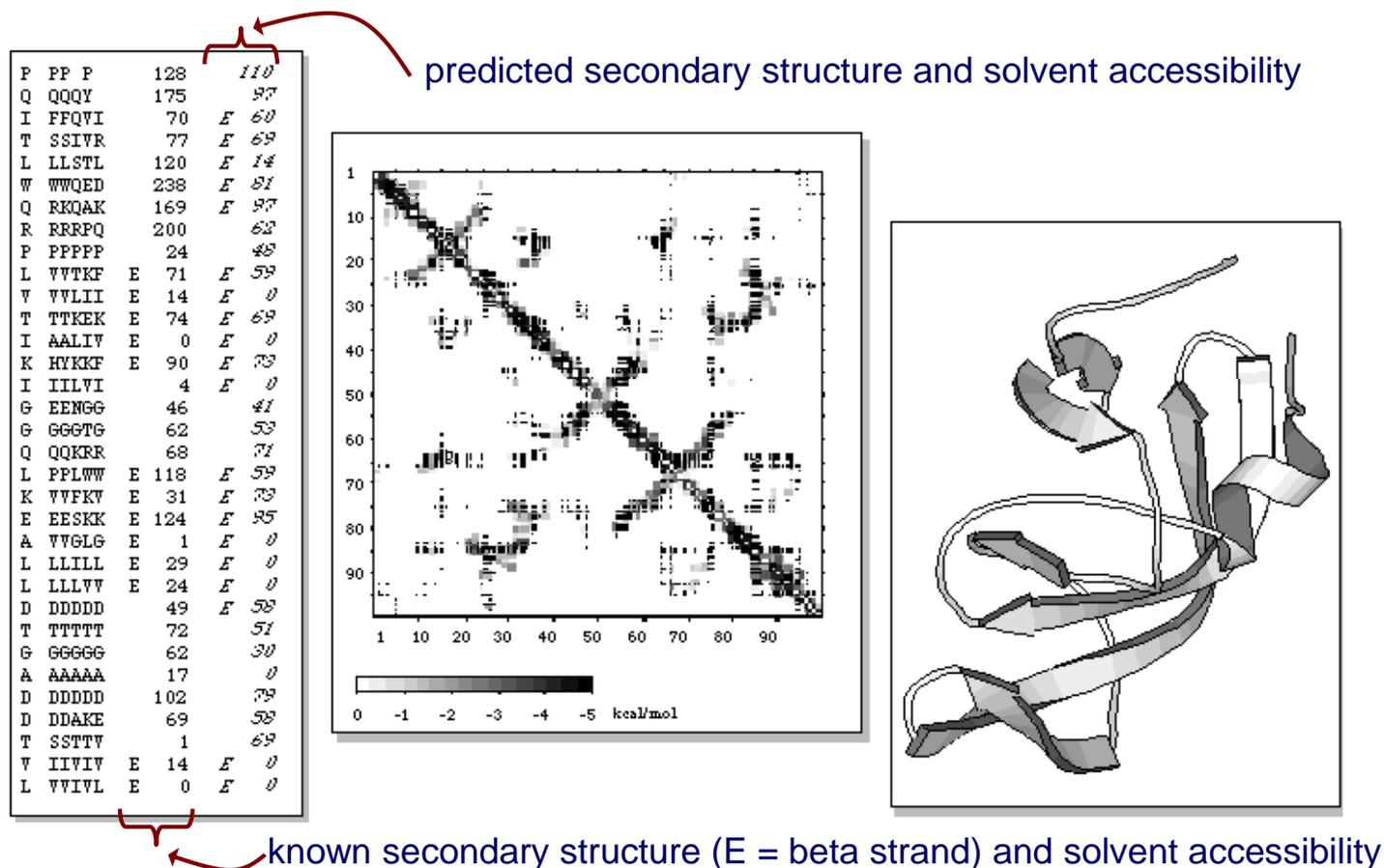
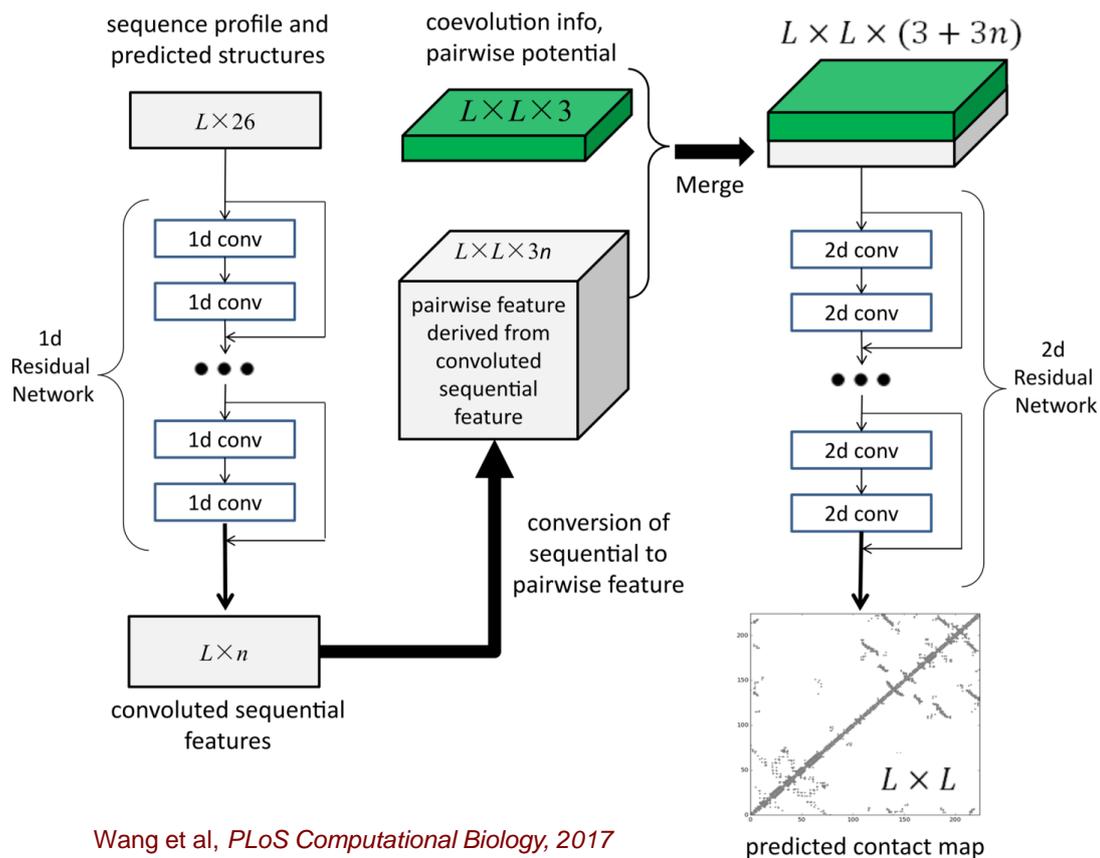


Figure from B. Rost, "Protein Structure in 1D, 2D, and 3D", *The Encyclopaedia of Computational Chemistry*, 1998

16

State-of-the-art in Contact Map Prediction



17

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

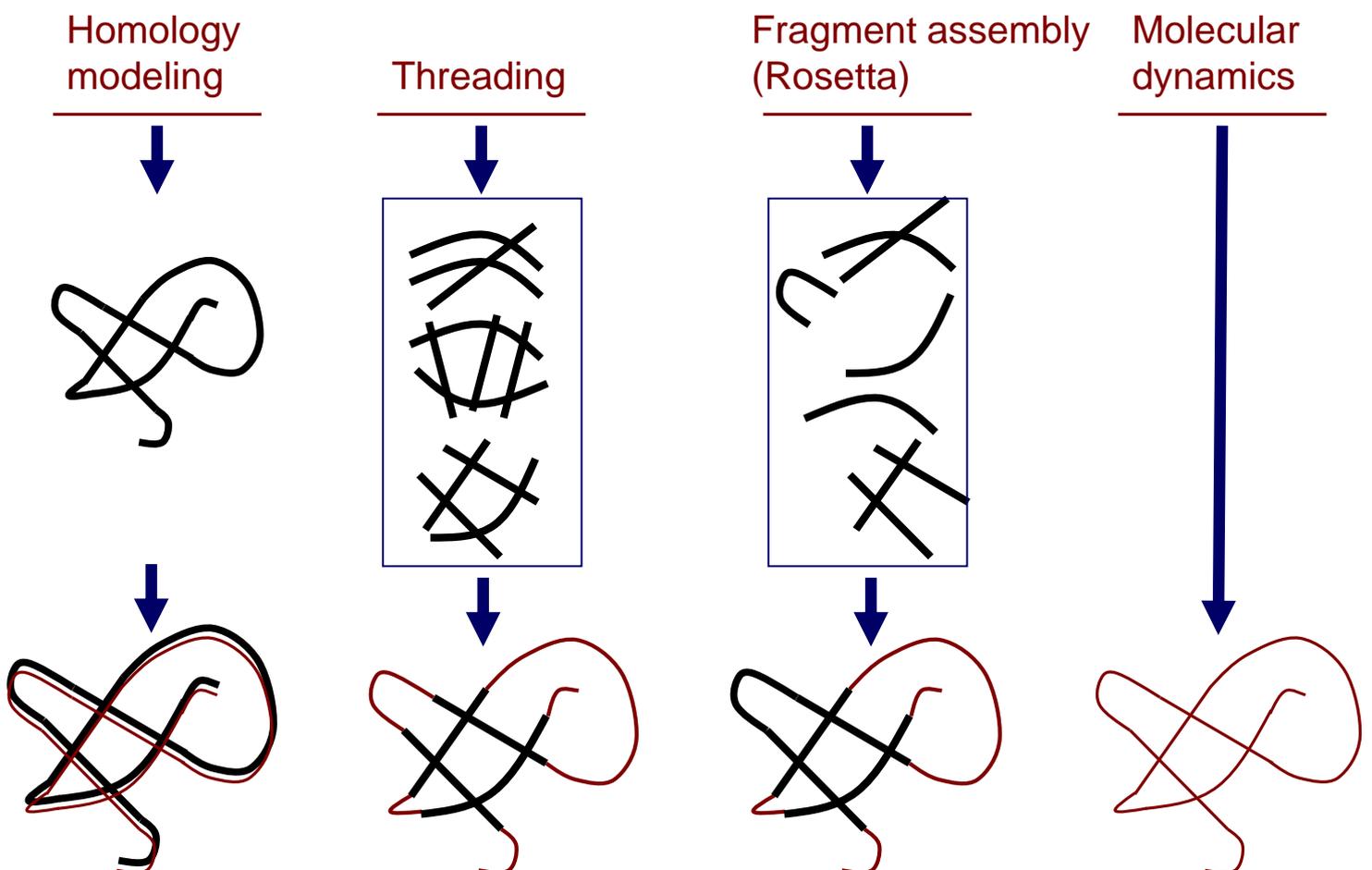
18

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

19

Prediction in 3D



20

“Citizen science”

- Folding@home

<http://folding.stanford.edu>



Molecular dynamics simulations

- Rosetta@home

<http://boinc.bakerlab.org>

Structure prediction



Volunteer/distributed computing

21

Foldit

The screenshot shows the Foldit game interface. On the left is a 3D model of a protein structure, primarily green with some blue and orange highlights. On the right is a competition leaderboard for "48: Pro Peptide". The leaderboard shows a "Group Competition" table and a "Player Competition" table. At the bottom left is a toolbar with icons for "Shake Sidechains", "Wiggle Backbone", "Clear Locks and Bands", "Reset Puzzle", and "Mouse Help". A tooltip for "Shake sidechains" is visible, stating "Shake sidechains to improve the protein. Hotkey: S". At the bottom right is a "Pull Tool" button.

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

Player Competition	
Player Name	Score
16 psen	9098
17 kathleen	9092
18 verzat82	9091
19 dakkarres	9091
20 ccarrico	9066
21 mbjorkegren	9048
22 sslckerson	9038

<http://fold.it/>

22