Bioinformatics of Proteins

- Atomic Properties
- The Folding Problem
- Structure Alignments
- Structure Prediction

Reza Jacob

4 June 2001

Biochemistry 118Q

Proteins in Bioinformatics

- How do we represent structures for computation?
- How do we compare structures *in silico*?
- How do we classify structures hierarchically?

The Plan

- Apply constraints of chemistry
  - Bond Lengths, Bond Angles, Dihedral (Torsion) Angles
- Place in Coordinate Frame
  - Cartesian, Internal, & Object Based Frames
- Compare Structures with *i* discrete components
  - Root Mean Squared Deviation

Basic Measurements

- Bond Lengths
- Bond Angles
- Dihedral (Torsion) Angles

Bond Length

- Bond Length fixed, given any scenario
- Depends on **type** of bond: single, double, triple, hybridization too
- Depends on which two atoms
- C-H is 1.0 Angstroms, C-C is 1.5 Angstroms
- Bond Length is a function of Spatial Position of the two atoms

Bond Length is Euclidean Distance

For (x1,y1,z1) and (x2,y2,z2),
$d=\{(x1-x2)^{2}+(y1-y2)^{2+(}z1-z2)^{2}\}^{1/2}$

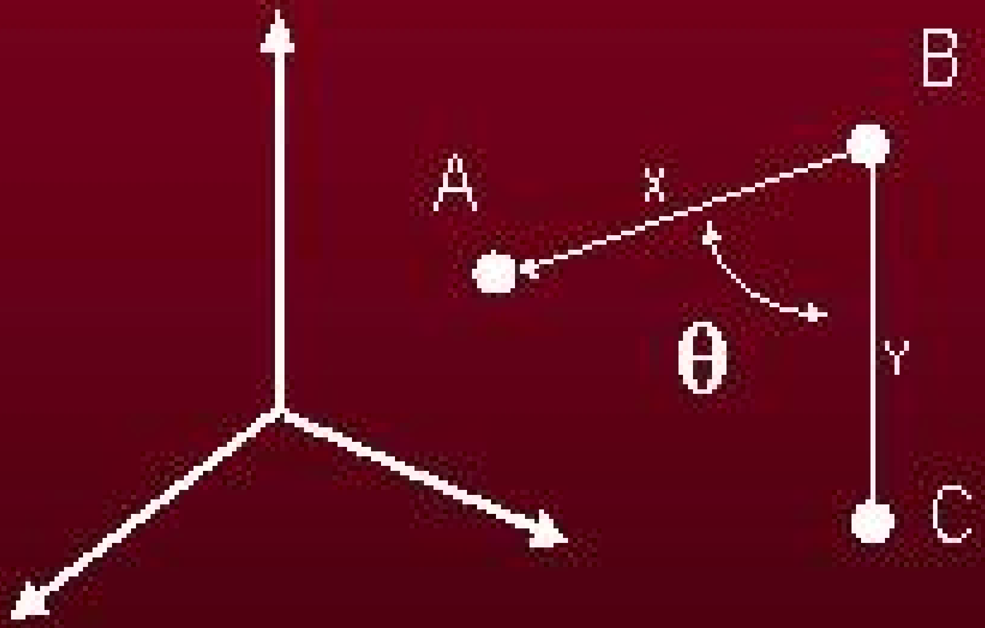
- Some non-covalent distances are also constant in a peptide’s backbone

- C_{alpha}-C_{alpha} distance for consecutive amino acids is constant too because of dihedral constraints

Bond Angles

- Chemistry also fixes Bond Angles
- Depends on types of atoms, hybridization states, and number of lone electron pairs
- Range is 100 degrees to 180 degrees
- Bond Angles is a function of the spatial position of three atoms

Computing Bond Angle



$$X \cdot Y = |X||Y|\cos(\theta)$$

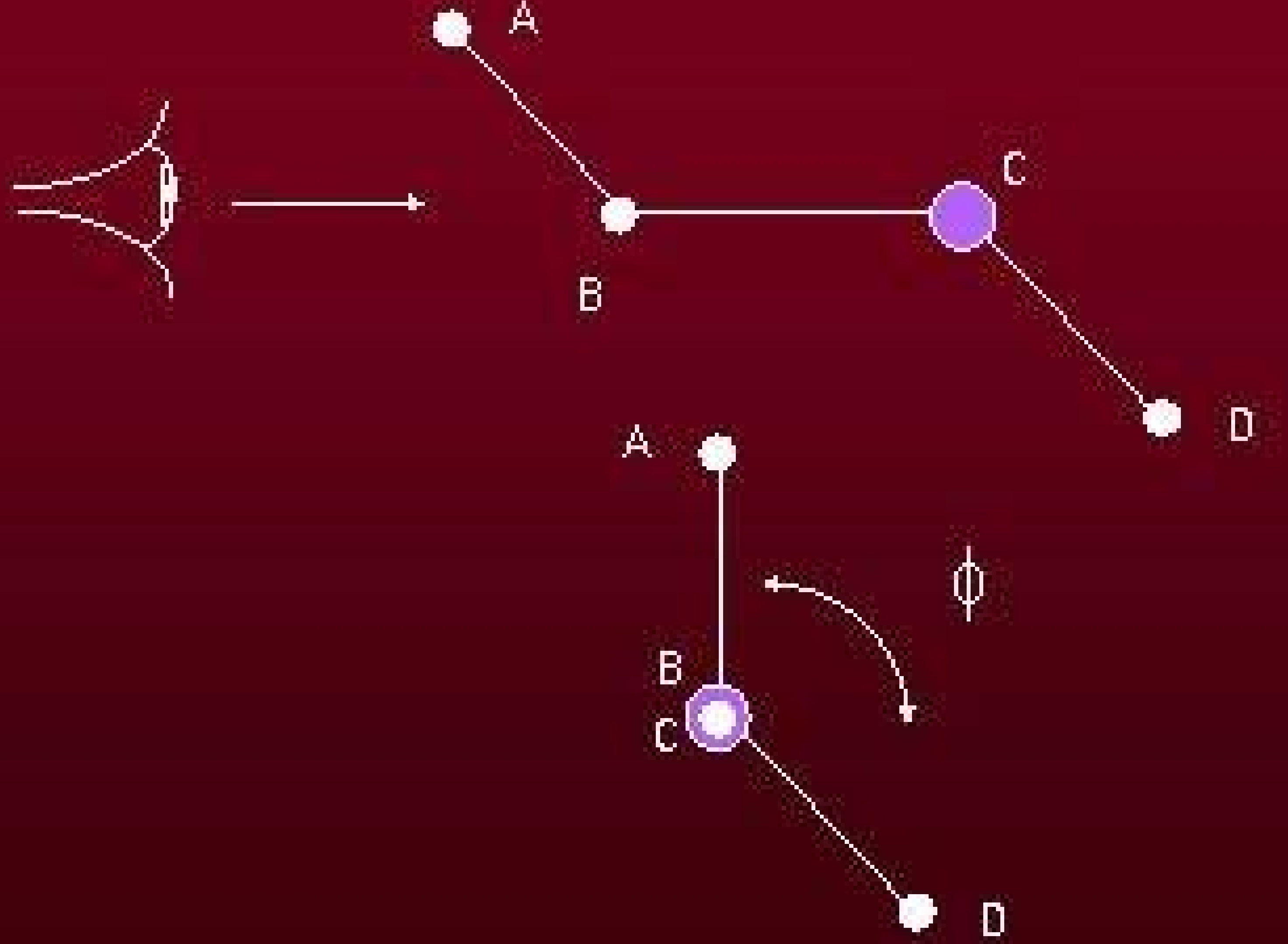
$$\theta = \mathrm{acos}(X \cdot Y / |X||Y|)$$

Angle can be computed by computing the arccosine of the dot product between unit vectors BA and BC.

Dihedral Angles

- These vary
- Range from 0 to 360 in principle
- Common in proteins are ϕ, ψ, ω, & χ
- Dihedral Angles are a function of the spatial position of four atoms in space

Dihedral Angle

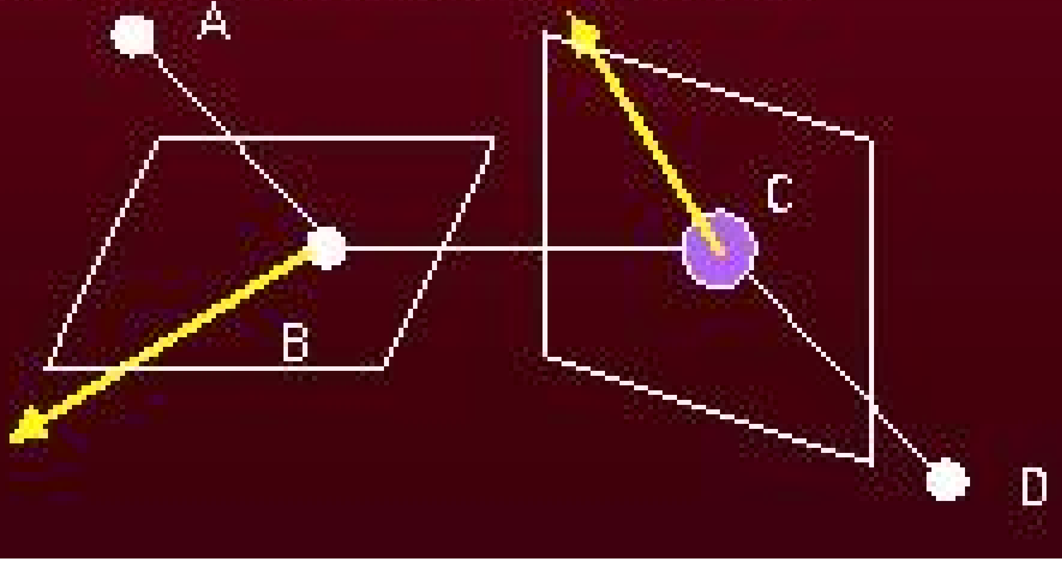


Computing Dihedral Angle

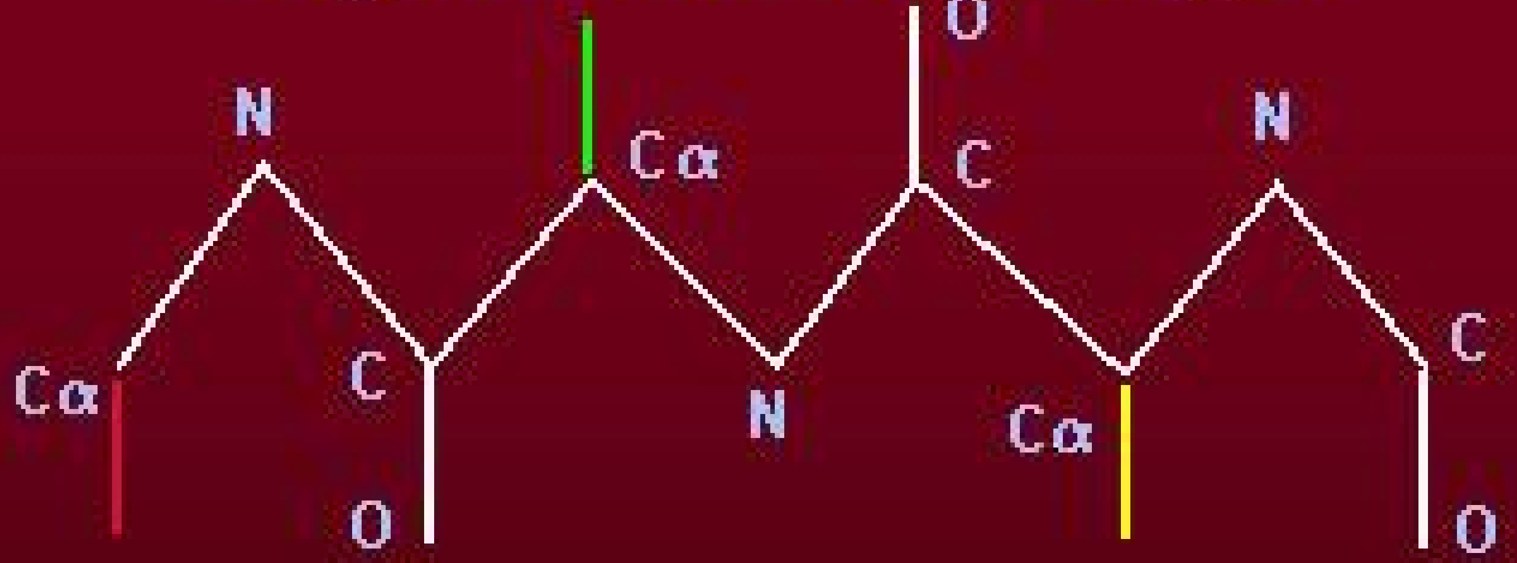
Compute cross-product of BA and CB
Compute cross-product of CB and DC

This produces two vectors perpendicular to the ABC plane and BCD plane.

Angle between those vectors (ala bond angle) is dihedral angle. Need to check if it is positive or negative.



Important Dihedral Angles



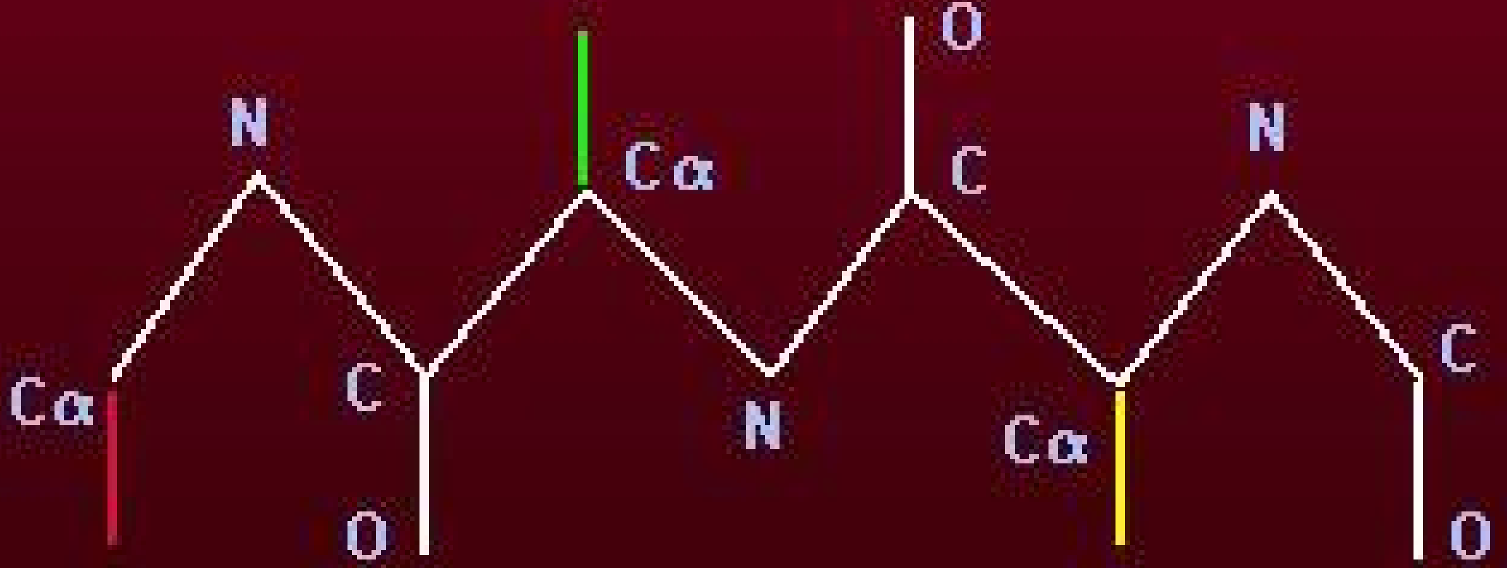
ω (omega) = Cα to Cα

ψ (psi) = N to N

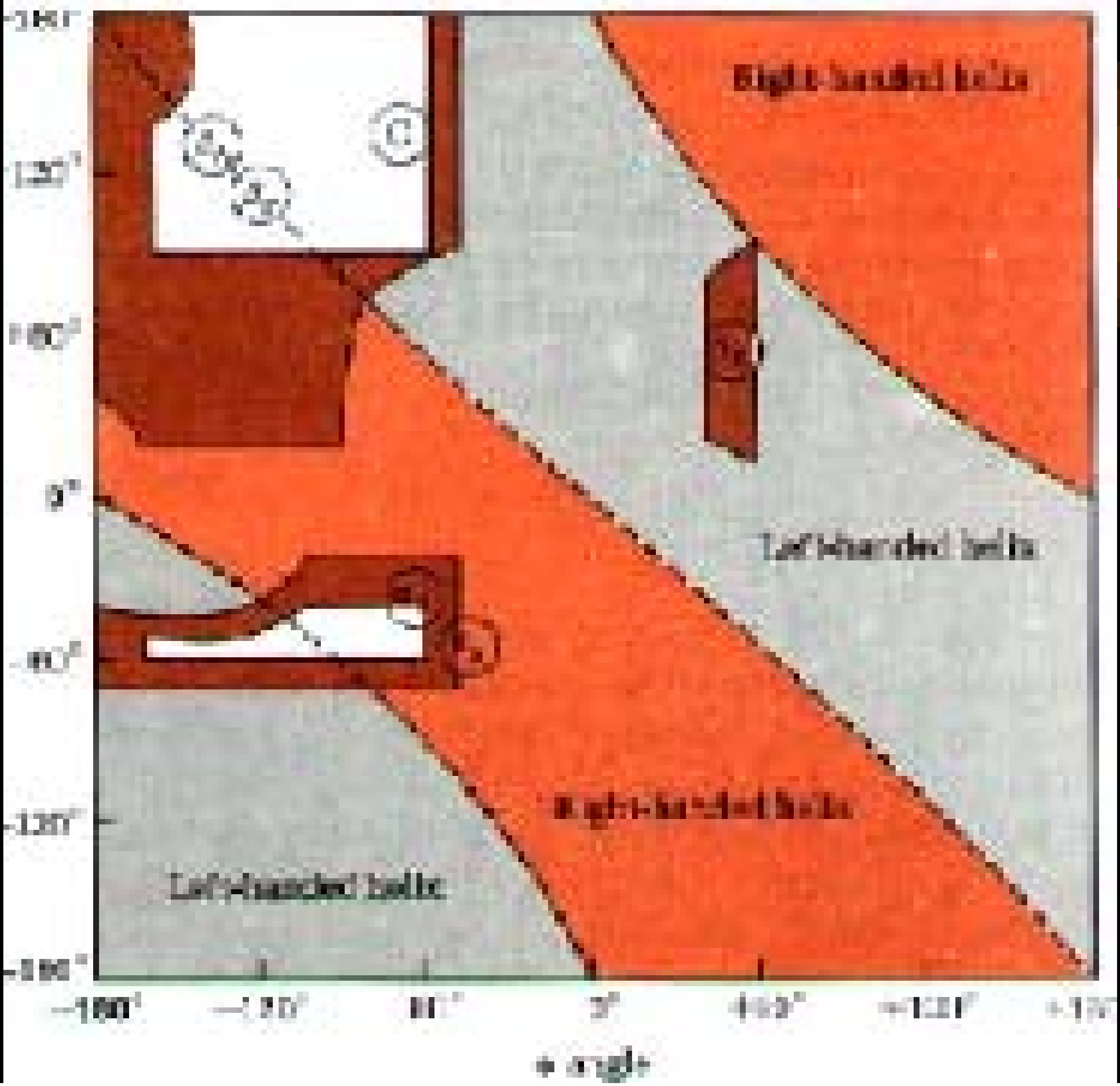
ϕ (phi) = C to C

Omega is constant = 180 (C-N doesn't rotate)
Phi, Psi have range of values (Ca-N, N-C rotate)

The range is restricted by having things not bump into each other.



Ramachandran Plot



Steric constraints restrict possible set of dihedral angles

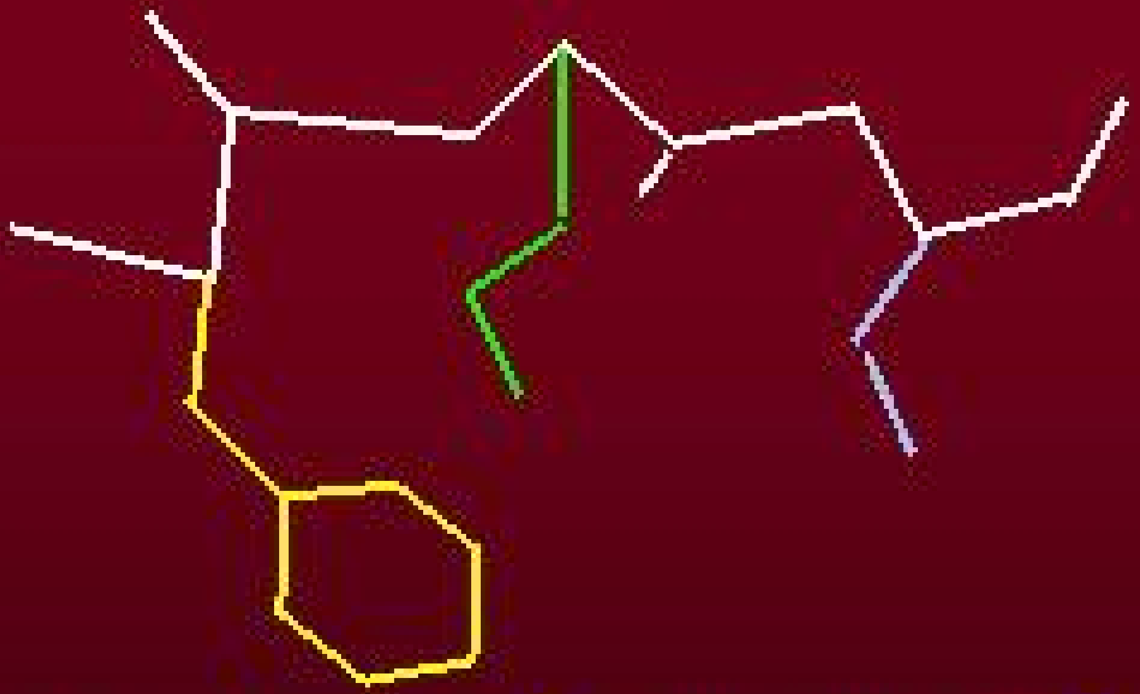
Typical Secondary Structures have known Dihedral Angles

- Alpha Helix
  - Phi=-57 degrees, psi=-47 degrees
- Parallel Beta Strand
  - Phi=-119 degrees, psi=113 degrees
- Antiparallel Beta Strand
  - Phi=-139 degrees, psi=135 degrees

Coordinate Frames

- Cartesian Frame has orthonormal (x,y,z) basis & provides signed lengths for motion along each axis (used in Protein DataBase)
- But since bond lengths and angles are basically constant, why not just specify dihedral angles?
- Leads to internal coordinate frame

Think about advantages



3 Peptide Units = 12 atoms = 36 coordinates OR 6 dihedral angles

3 Sidechains = 12 atoms = 36 coordinates OR 5 dihedral angles

72 Cartesian Coordinates vs. 11 internal coordinates

Disadvantages of Internal Frame?

- Basic computations (like Euclidean distance) are really difficult
- How about objects which aren’t connected?
- Makes algorithms more complex sometimes

Object-Based Coordinate Frame

- Certain part of proteins have less variability, like an alpha helix backbone
- Treat helix backbone as rigid object
- Reduces number of parameters specified

Comparing Structures

- Compare structures A & B
- Need to know which atoms in A correspond to which in B
  - Get this from BLAST
- Need to know position of all atoms
  - Get this from PDB

Comparing Structures

- How closely can two structures be superimposed?
- Need an *objective function* to measure this
- If exactly the same, measure = 0
- If divergent structures, measure is large

The RMSD

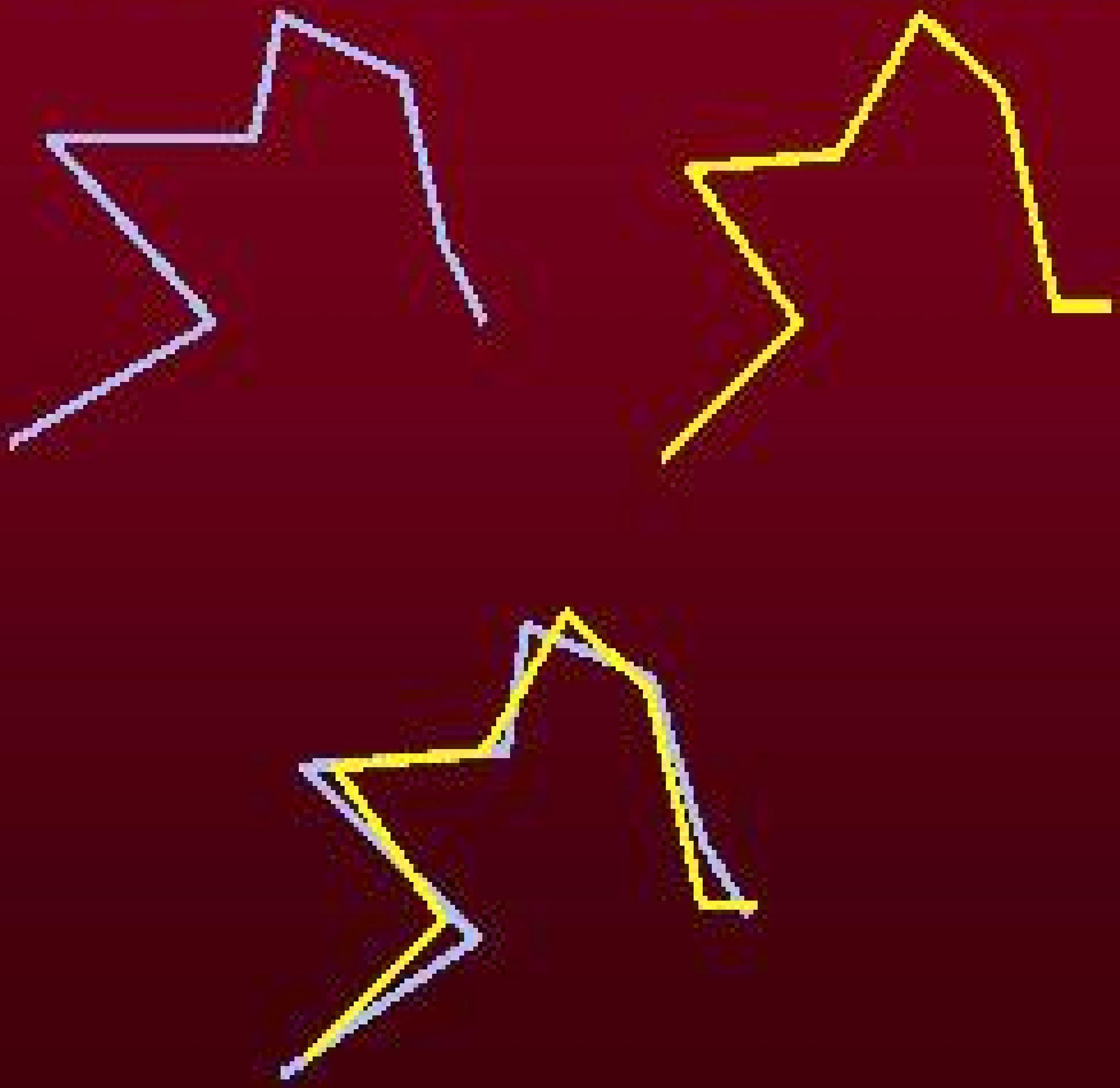
RMSD = root mean squared deviation

$$RMSD = \sqrt{\frac{\sum_i d_i^2}{N}}$$

where *N* is the number of atoms

***d*$_i$** is the distance between two atoms with index i from the two structures

We want the minimum RMSD



RMSD Algorithms

- Greedy search around center of mass for lowest RMSD
  - Superimpose centers of mass
  - Calculate RMSD
  - Rotate slightly
  - Re-calculate RMSD, and chose lowest
- *Method based on translation and rotation matrices*
  - Algorithm based on eigenvectors

Advantages of RMSD

- Nice behavior
  - 0 when identical, falls off continuously
- Easy to compute
- Units are natural (Angstroms)
- Commonly Used
- Similar structures show 1-3 Angstroms RMSD

Disadvantages of RMSD

- All atoms are equally weighed
- Upper bound variable
- Significance cutoff increases as size increases

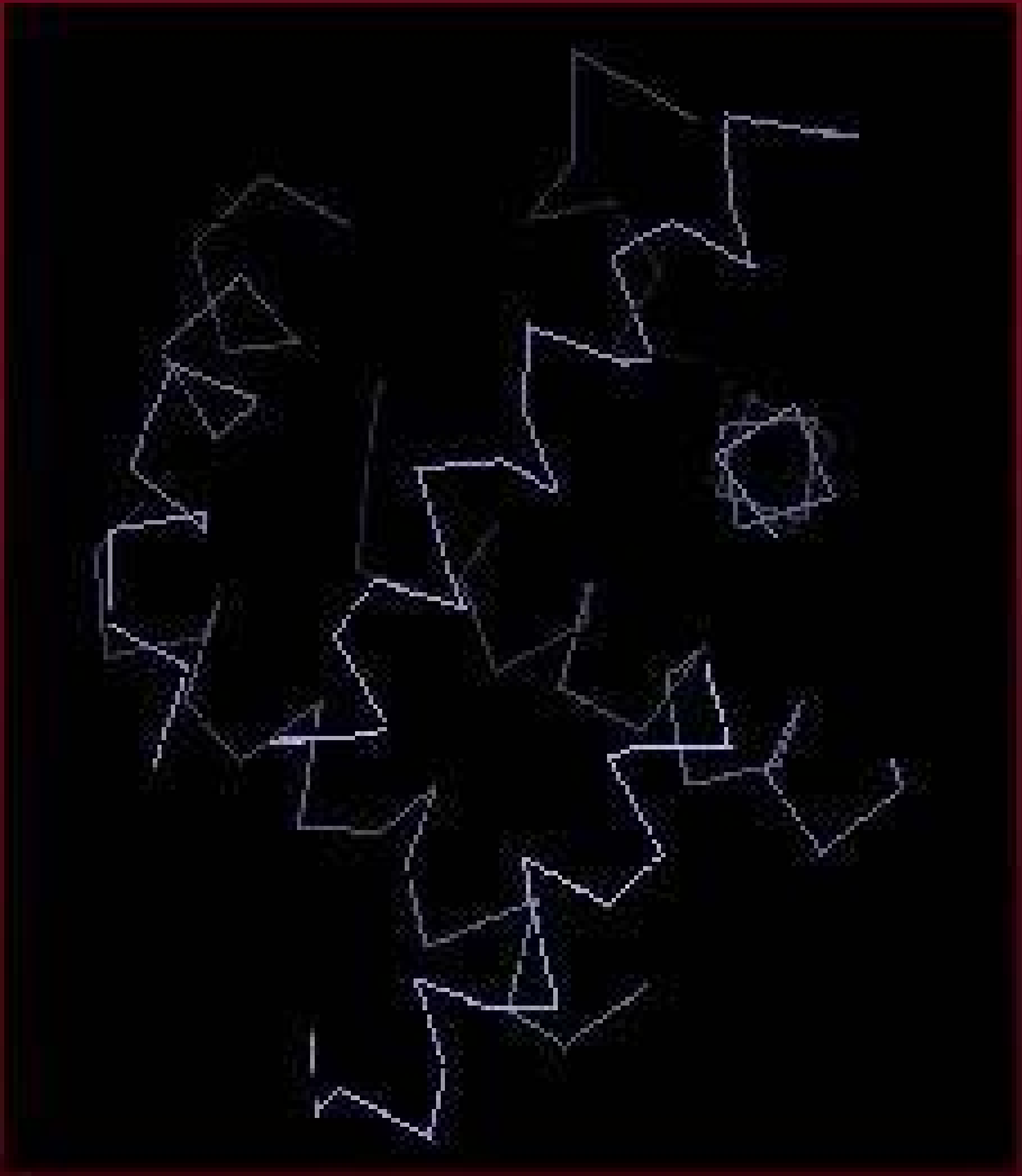
Case Study: Myoglobin Superfamily

- Eight structures involved:
  - Sperm whale myoglobin
  - Sea hare myoglobin
  - Plant leghemoglobin
  - Sea lamprey hemoglobin
  - Human alpha & beta hemoglobin chains
  - Chironomous hemoglobin
  - Bloodworm hemoglobin
- Aligned by hand b/c of low a.a. identity
- 115 common positions

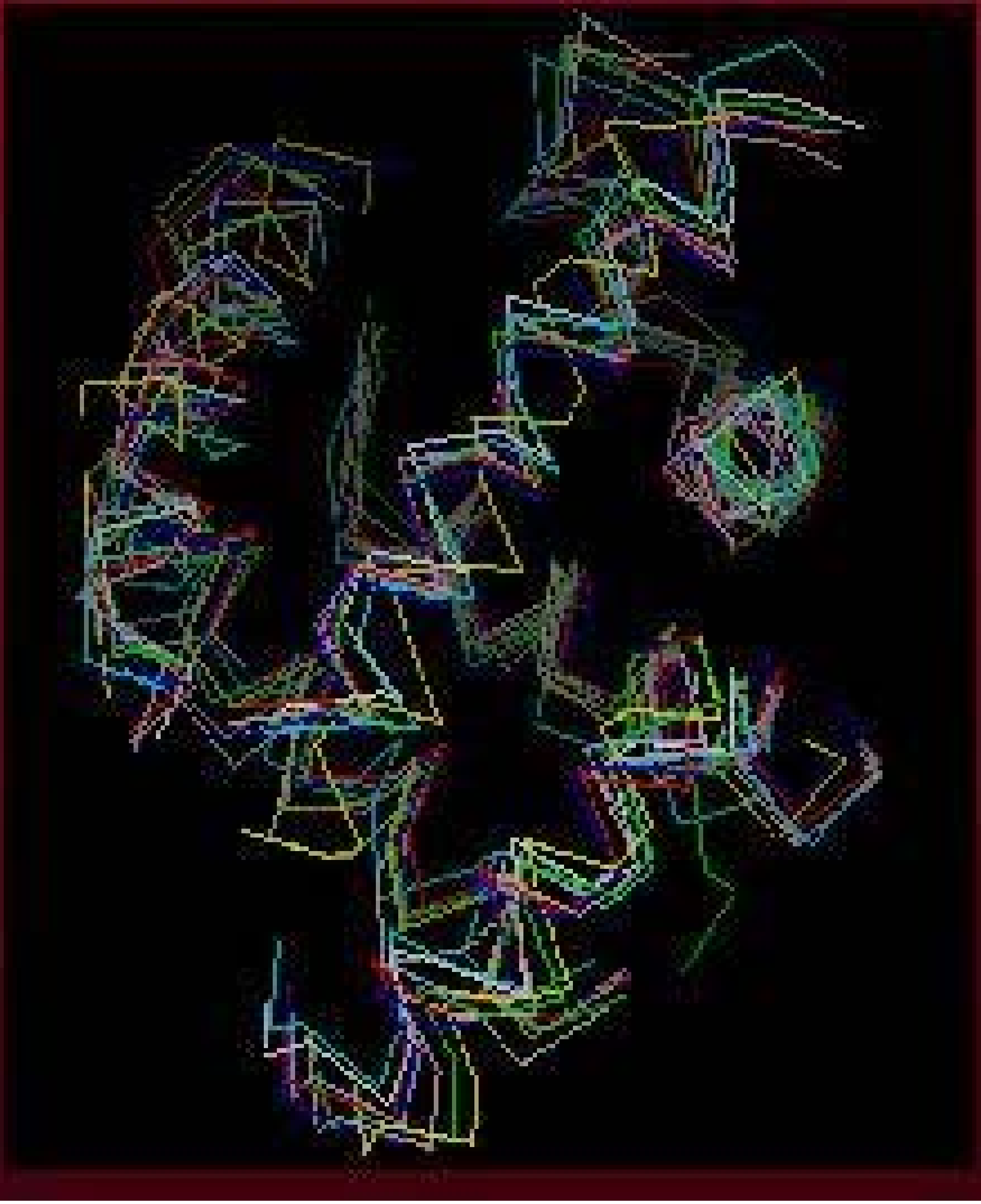
```
012345678-------------AAAAAAAAAAAAAA------AAAAAAAAAAAAAAAAAAAAAA---------AAAAAAAAAAAAAAAAAAAAAAA-
2mm1    .  ----------GLSDGEWQLVLNVWGKVE--ADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKG
1ymb    .  ----------GLSDGEWQQVLNVWGKVE--ADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKKHGTVVLTALGGILKKG
1pmb    A  ----------GLSDGEWQLVLNVWGKVE--ADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGNTVLTALGGILKKG-
1mbs    .  ----------GLSDGEWHLVLNVWGKVE--TDLAGHGQEVLIRLFKSHPETLEKFDKFKHLKSEDDMRRSEDLRKHGNTVLTALGGILKKG-
4mbn    .  ----------VLSEGEWQLVLHVWAKVE--ADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKG-
1myt    .  --------------ADFDAVLKCWGPVE--ADYTTMGGLVLTRLFKEHPETQKLFPKFAGI-AQADIAGNAAISAHGATVLKKLGELLKAKG-
2hhb    A  ----------VLSPADKTNVKAAWGKVG--AHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GSAQVKGHGKKVADALTNAVAHVD
2mhb    A  ----------VLSAADKTNVKAAWSKVG--GHAGEYGAEALERMFLGFPTTKTYFPHF-DLSH-----GSAQVKAHGKKVGDALTLAVGHLD-
1pbx    A  ----------SLSDKDKAAVRALWSKIG--KSADAIGNDALSRMIVVYPQTKTYFSHWPDVTP-----GSPHIKAHGKKVMGGIALAVSKID-
2hhb    B  ---------VHLTPEEKSAVTALWGKV---NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD-
2mhb    B  ---------VQLSGEEKAAVLALWDKV---NEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLD-
2lhb    .  -PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKFKGLTTADQLKKSADVRWHAERIINAVNDAVASMD
1mba    .  ----------SLSAAEADLAGKSWAPVFA-NKNANGLDFLVALFEKFPDSANFFADFKGKSVADIKASPKLRDVSSRIFTRLNEFVNNAAN-
1sdh    A  --PSVYDAAAQLTADVKKDLRDSWKVIGS--DKKGNGVALMTTLFADNQETIGYFKRLGDVSQGMANDKLRGHSITLMYALQNFIDQLDN
1lh1    .  ---------GALTESQAALVKSSWEEFNA-NIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPELQAHAGKVFKLVYEAAIQLEV
1hlb    .  GGTLAIQAQGDLTLAQKKIVRKTWHQLMRN--KTSFVTDVFIRIFAYDPSAQNKFPQMAGMSASQLRSSRQMQAHAIRVSSIMSEYVEELDS
1ith    A  ----------GLTAAQIKAIQDHWFLNIKGCL-QAAADSIFFKYLTAYPGDLAFFHKFSSVPL-YGLRSNPAYKAQTLTVINYLDKVVDALGG
1ecd    .  ------------LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQFAGKDLESIKGTAPFETHANRIVGFFSKIIGEL-
2hbg    .  ----------GLSAAQRQVIAATWKDIAGADNGAGVGKKCLIKFLSAHPQMAAVFGFSGAS--------DPGVAALGAKVLAQIGVAVSHLG-
```

RMS for alpha carbons

- N(N-1)/2 pairwise RMSs computed (N=8)
- Ranged from 1.22 to 3.16 Angstroms
- Average was 2.19 Angstroms



Superimpose all on an average



Conclusions

- Compute bond length, bond angles, dihedral angles
- Work in different coordinate frames
- Use RMSD for structure comparison
- Graphical superimposition can elucidate structural similarities & differences

The Protein Folding Problem

- The Search Space
- Definitions of Energy
- Computing Free Energy
- The Energy Function
- MonteCarlo Methods
- Molecular Dynamics

The Folding Problem

- How does the linear a.a. sequence fold to the 3-D shape off the ribosome?
- And more broadly, how do we get the 3-D structure given a linear a.a. sequence?

The Input Space

- Linear amino acid sequence
- Structure of each amino acid and peptide backbone
  - Lists of atoms, bond lengths, bond angles
  - Ramachandran constraints on dihedral angles
- The media
  - Water and dissolved solutes (salts)

The Output Space

- The 3-D coordinates of the protein in some frame
- Partial Answers:
  - 3-D structure of active site
  - Location in linear sequence of secondary structure
  - Prediction of "*class*" or "*family*" of the protein

Why should we care?

- Sequence ---> Structure ---> Function
- Structure very useful for Drug Design
- Hard to get structures experimentally
  - X-ray crystallography (80%) 1-2 A
  - Nuclear Magnetic Resonance (20%) 1-3 A
  - Cryo Electron Microscopy (<<1%) 7-10 A

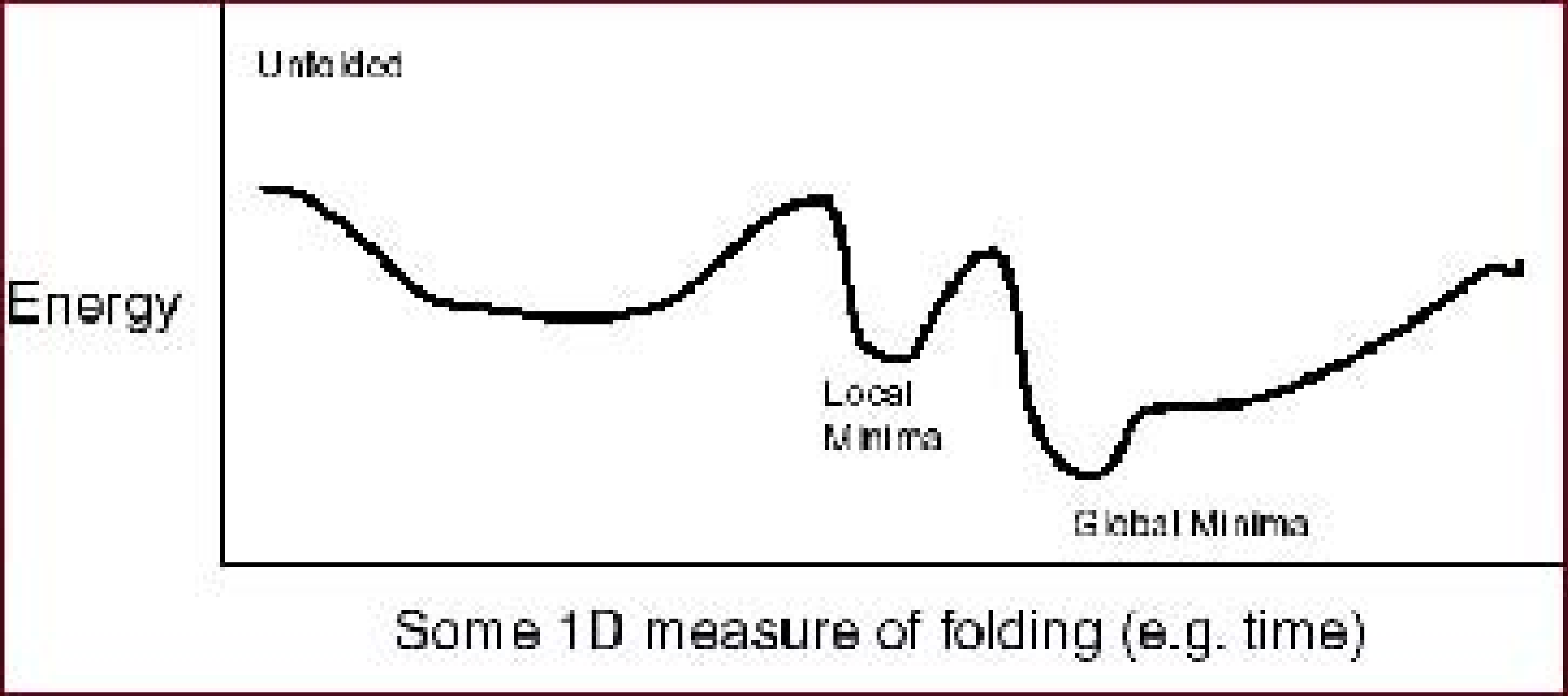
How hard is the problem?

Very Hard

- Huge search space
- For a 100 a.a. chain, assume each a.a. can be in either alpha, beta, or coil state (simplification)
- 3^{100}=5 * 10^{47} possible distinct folds
- At 1 fold every 0.10 ps, it takes 10^{27} years
- Universe is 10^{10} years old

Why is the problem hard?

- How do we know when we have the “correct” fold?
- Need to measure interactions between a.a.’s, water, and other molecules
- You are folding proteins right now
- You do it in seconds


Unfolded
Energy
Local
Minima
Global Minima
Some 1D measure of folding (e.g. time)

Sampling the Output Space

- Secondary structure occurs regularly
  - Can form locally, independent of global structure
- Steric constraints eliminate some possibilities
- Maybe a nonrandom search?
  - Local structure can form and induce cascades

Gibbs Free Energy

- $\Delta G = \Delta H - T\Delta S$
- Free Energy=Enthalpic Energy - Entropic Energy

  ΔH = benefits of interactions (negative for folding)

  $T\Delta S$ = costs of imposing order (negative for folding)
- Proteins fold **because** $\Delta H < T\Delta S$
- Usually just by a narrow margin

Entropy

- High entropy means disorder
- $S = k \ln \Omega$, where Ω=# arrangments
- If only 1 state is allowed $\Omega = 1$, and $S=0$
- Often hard to compute by statistical mechanics
- Turn to a more classical approach

Energy

- Total Energy = Potential + Kinetic
- E = U + K
- Use Newtonian physical approximations
  - Atoms and bonds as balls and springs
- Seek energy minima

Writing an Energy Function

- Bond Lengths
- Bond Angles
- Dihedral Angles (Ramachandran constraints)
- Packing term (nature abhors a vacuum)
- Electrostatic interactions

Potential Energy function:

$$U = \sum_{bonds} K_b (b_i - b_o)^2$$

$$+ \sum_{angles} K_\theta (\theta_i - \theta_o)^2$$

$$+ \sum_{dihedrals} K_\phi \left[1 - \cos(n\phi_i + \delta)\right]$$

$$+ \sum_{pairs} \varepsilon \left[\left(\frac{r_o}{r_{ij}} \right)^{12} - 2 \left(\frac{r_o}{r_{ij}} \right)^{6} \right]$$

$$+ \sum_{charges} \frac{q_i q_j}{r_{ij}}$$

The potential energy is a function of all atomic coordinates.

The sum of all these terms creates a function with many local minima and maxima.

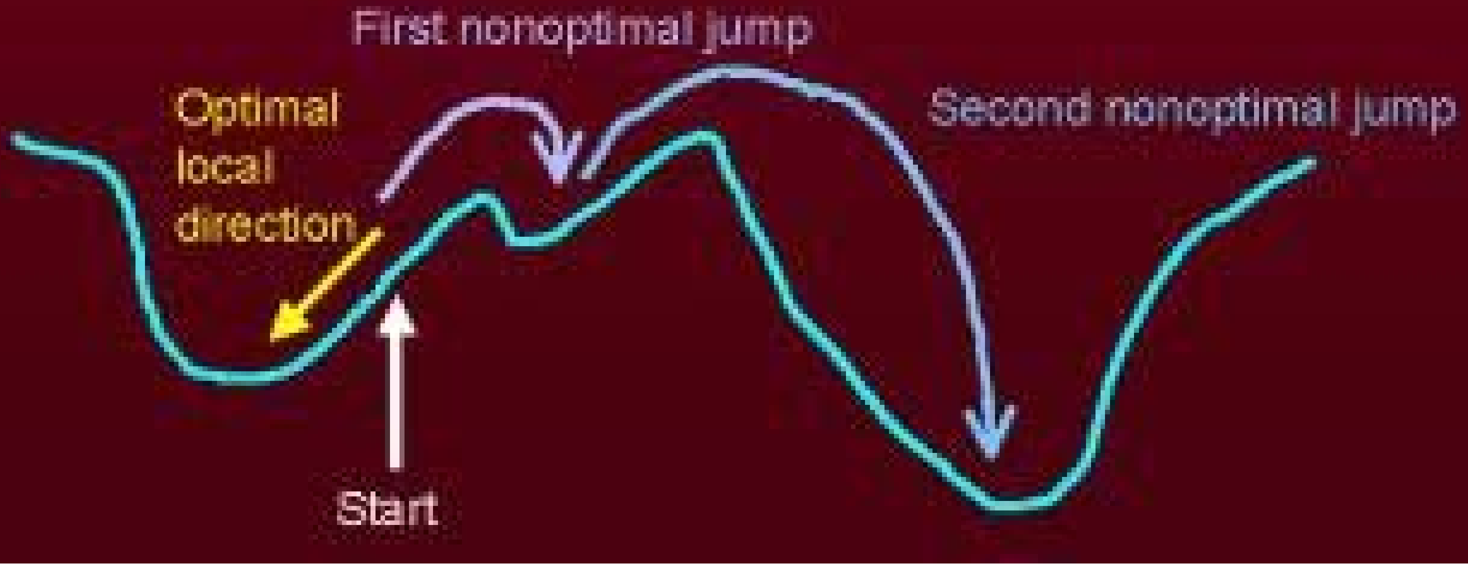
Very hard to sample this function well.

$$U = \sum_{bonds} K_b (b_i - b_o)^2 + \sum_{angles} K_\theta (\theta_i - \theta_o)^2 + \sum_{dihedrals} K_\phi \left[1 - \cos(n\phi_i + \delta)\right] + \sum_{pairs} \varepsilon \left[\left(\frac{r_o}{r_{ij}} \right)^{12} - 2 \left(\frac{r_o}{r_{ij}} \right)^{6} \right] + \sum_{charges} \frac{q_i q_j}{r_{ij}}$$

DALI optimizes with a MonteCarlo (Metropolis, 1953) algorithm.

Basic idea: iterative improvement by random walk through search space, with occasional excursions into "non-optimal" territory.

Only be allowing non-optimal jumps, do you allow yourself to leave local optima.



MonteCarlo Algorithm

- Choose a starting position P
- Evaluate the objective scoring function S
- Perturb the current position (randomly or otherwise) to P’ and compute S’
- If S’<S, let P = P’
- Else let P = P’ with probability $e^{\beta(S'-S)}$
- Loop

Relative Energies

• Hydrogen Bond	-5.0 kcal/mol
• Change in Bond Angle by 10 degrees	+2.0 kcal/mol
• Stretch bond length by 0.1 Angstroms	+2.5 kcal/mol
• Pack two atoms snugly	-0.2 kcal/mol
• Break a bond	+100 kcal/mol
• Bring two +1 charges to 3 Angstroms	+100 kcal/mol

Searching for Global Energy Minima

- Search for atomic coordinates that minimize U
- Generally finds only local minima
- Can use MonteCarlo algorithms,
- Need good (nonrandom) starting structure
- Works well for relaxing perturbations of known structures
- No water, no solutes included

Molecular Dynamics

- F(x,y,z) = -Grad[U(x,y,z)]
- F = m a
- Simulate atomic paths by small linear motions
- To make small motions, need small time step

Time Steps

• Bond stretching	0.01 ps
• Angle bending	0.1 ps
• Rotating methyl group	1.0 ps
• Water tumbling	10 ps
• Protein tumbling in water	10,000 ps
• Chemical Reaction	1,000,000 ps

Need time step = 0.001 ps = 1.00 femtoseconds!!!

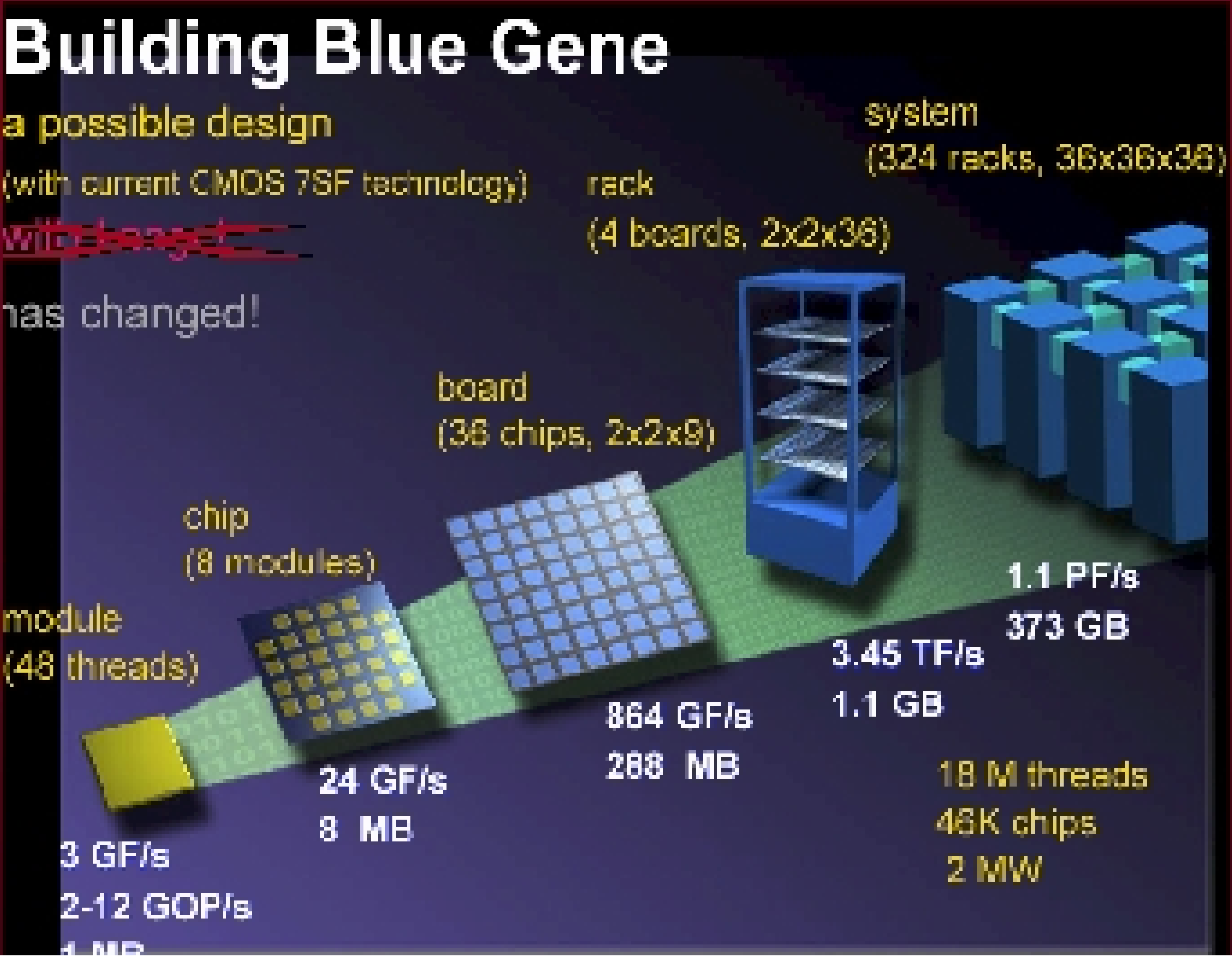
Goals of Molecular Dynamics

- Learn how protein moves in water
- Learn response to perturbation
- Fold proteins *ab initio*
- Run microseconds of simulation

Incorporate Experimental Facts

- The part off the ribosome first doesn’t necessarily fold first
- Secondary structure forms rapidly, making problem easier

IBM Blue Gene = Fold a protein



Building Blue Gene

a possible design

(with current CMOS 7SF technology)

~~with budget~~

has changed!

system
(324 racks, 36x36x36)

rack
(4 boards, 2x2x36)

board
(36 chips, 2x2x9)

chip
(8 modules)

module
(48 threads)

3 GF/s
2-12 GOP/s
1 MB

24 GF/s
8 MB

864 GF/s
288 MB

3.45 TF/s
1.1 GB

1.1 PF/s
373 GB

18 M threads
46K chips
2 MW

Structure Alignment

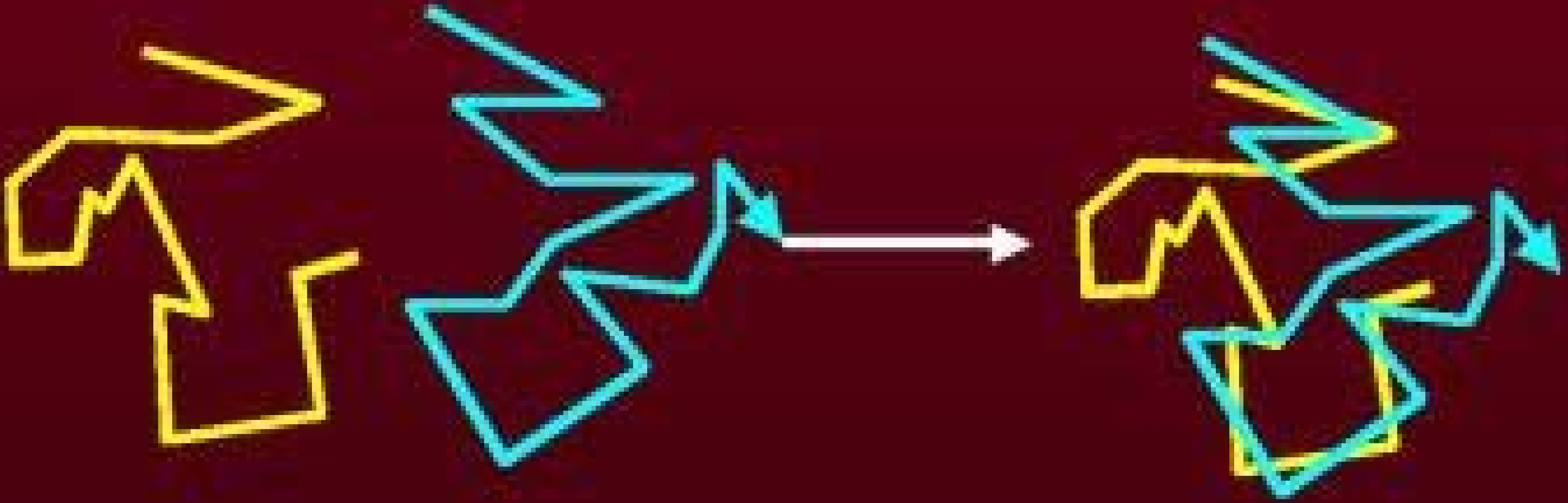
- Fit structure A with *i* elements to B with *j* elements
- Analogy

RMSD	*i* to *i*	BLAST without gaps
Structure Alignment	*i* to *j*	BLAST with gaps

- Use RMS as tool in computing Structure Alignment

Problem

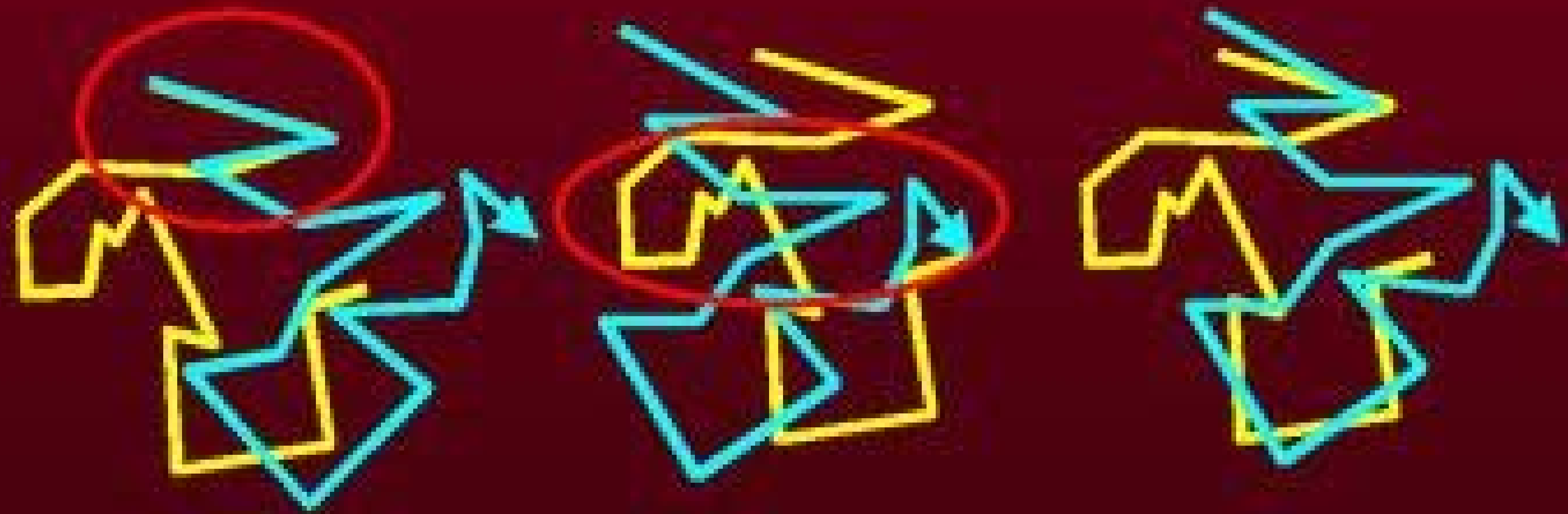
Given a pair of molecular structures, find the correspondences between atoms that leads to the “best” alignment.



Note:

"best" means the most atoms aligned with the lowest RMS.

Tradeoff: few atoms aligned very well vs. lots of atoms aligned not so well.



Criteria for Alignment

- *i* and *j*
- % identity or similarity of aligned a.a.'s
- # of gaps
- Shared active site?

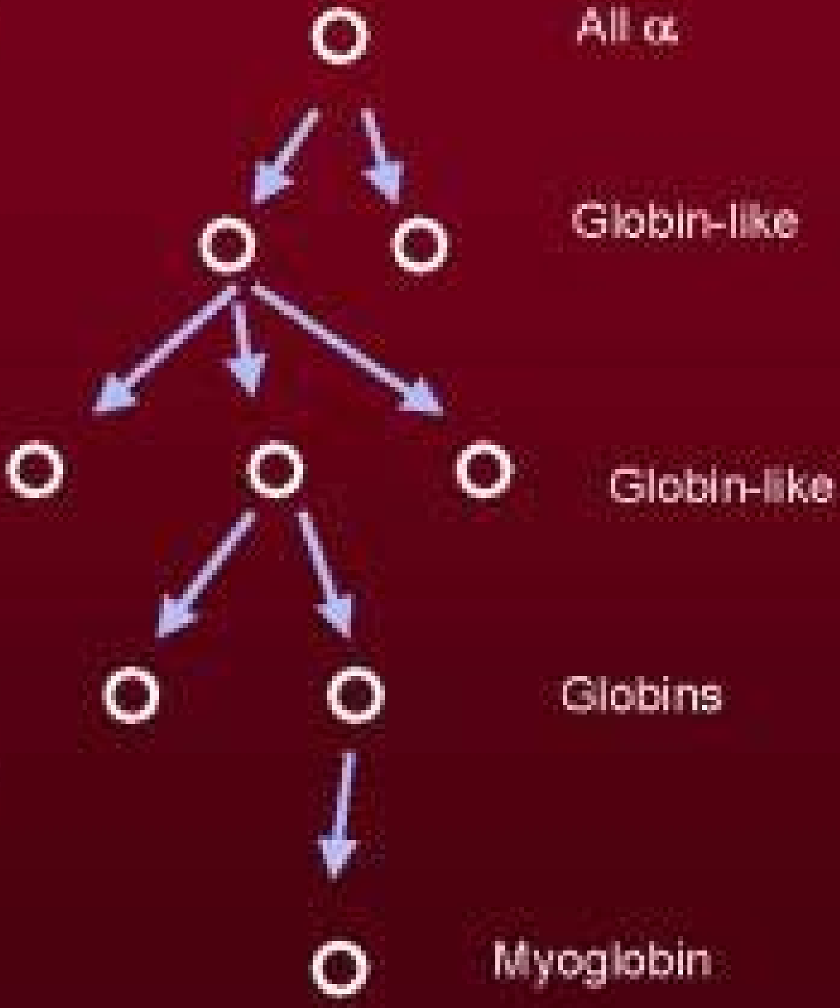
Why bother aligning?

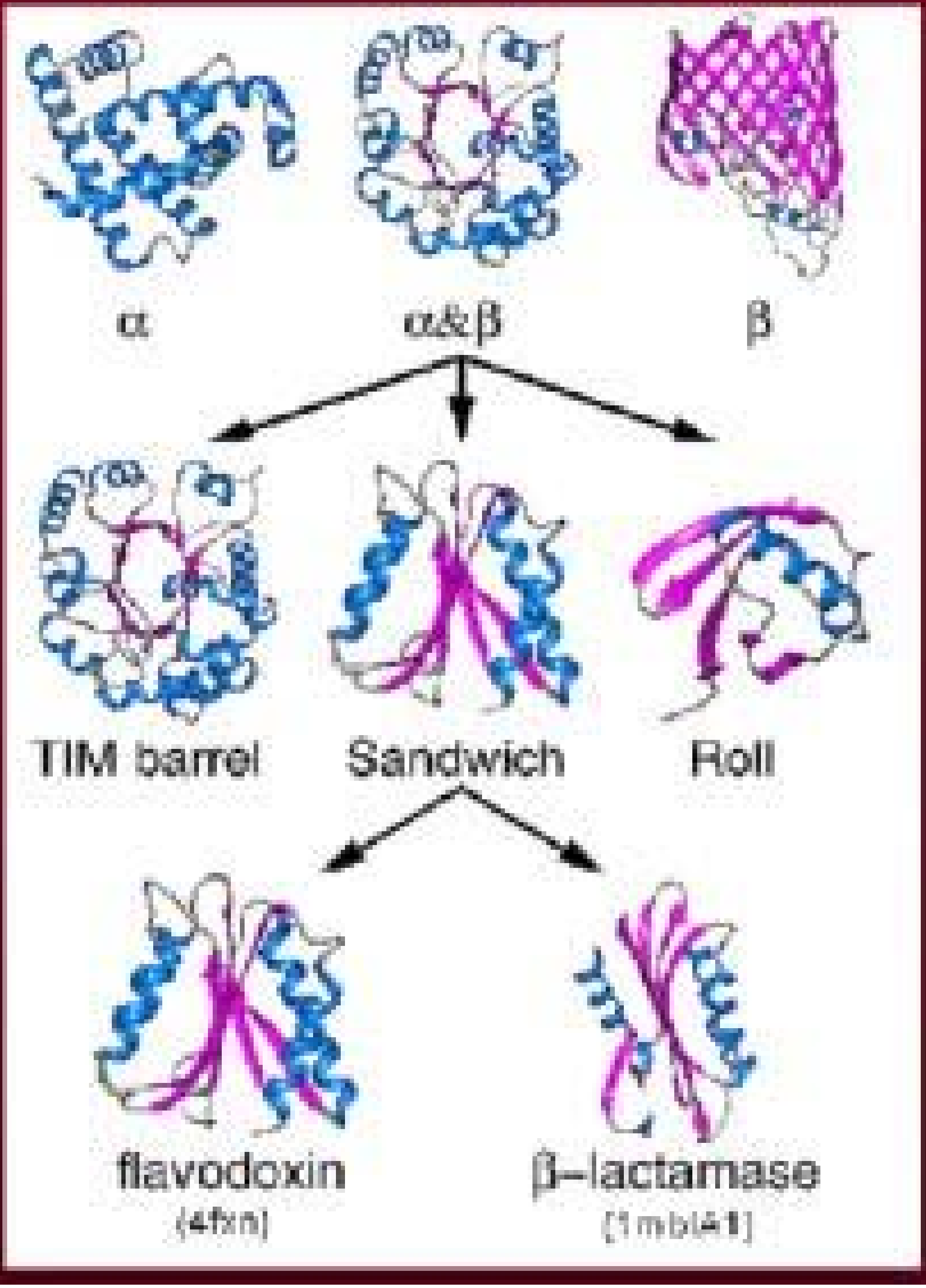
- As a check on sequence searches (BLAST)
- Make a hierarchy of classification of proteins
  - http://scop.stanford.edu
  - Alexei Murzin (manual) or Algorithmically
- Evaluate common ancestry

Manual Clustering of Structures

(see Structural Classification of Proteins = SCOP at http://scop.stanford.edu)

- **Class**
  - similar 2° structure
  - all α, all β, $\alpha + \beta$, α/β
- **Fold**
  - major structural similarity
  - similar arrangement of 2°
- **Superfamily** (topology)
  - probable common ancestry
- **Family**
  - clear evolutionary relationship
  - sequence similarity > 25%
- **Individual Protein**




α
α&β
β
TIM barrel
Sandwich
Roll
flavodoxin
(4fxn)
β–lactamase
[1mblA1]

Algorithms

- STRUCTAL (Levitt, Subbiah, Gerstein)
- DALI (Holm, Sander)
- LOCK (Singh, Brutlag)

Folding vs.. Prediction

- Folding gets to 3-D structure by simulating physical principles
  - Energy minimization
  - Molecular Dynamics
- Prediction gets to 3-D structure using statistical, theoretical, and/or empirical info
  - Just get structure, doesn’t matter how

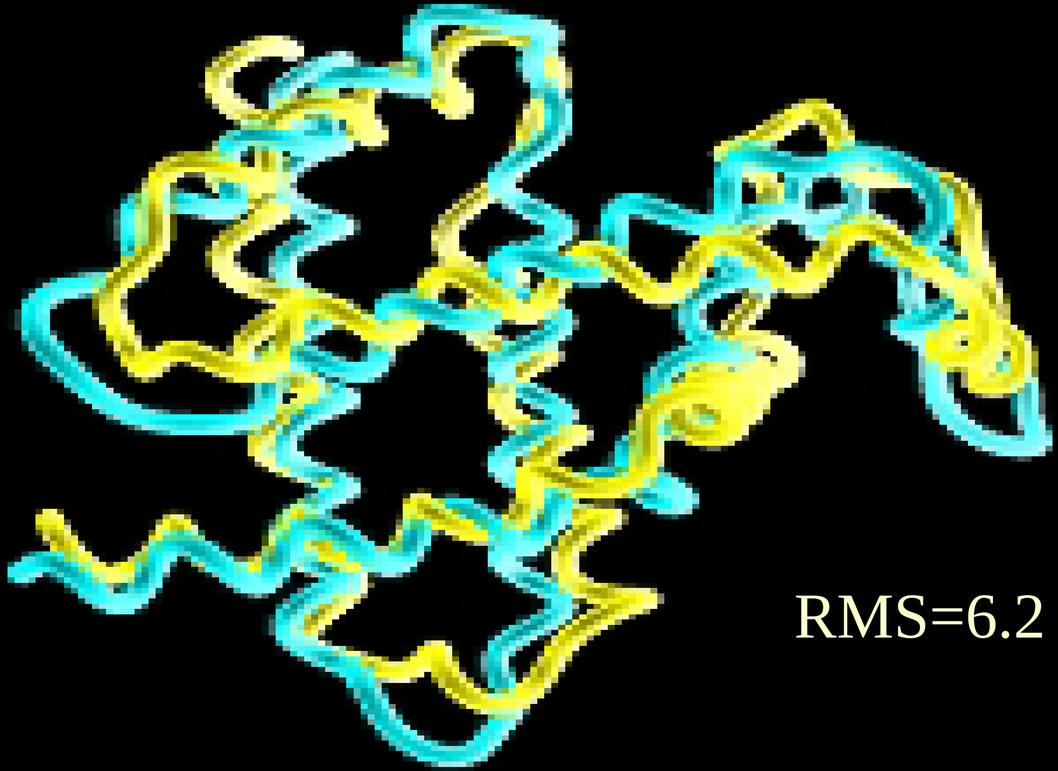
Asilomar Contest

- Started 1994 and runs biannually
- Conference near Monterey
- “Meeting on Critical Assessment of Techniques for Protein Structure Prediction (CASP)
  - Homology Modeling (>25% sequence identity)
  - Fold Recognition (20-25% sequence identity)
  - *Ab initio* prediction (no homology)

The Players

- Experimentalists - gets structure empirically
- Predictors download sequence and minimal info
- Assessors use RMS, alignment to evaluate results of predictors algorithms

Evaluation



RMS=6.2 Angstroms

Homology Modeling

- Goal is final 3-D structure
- >70% homology works great
- PSI-BLAST helps a lot
- Energetic relaxation doesn’t help without a good guess

Fold Recognition

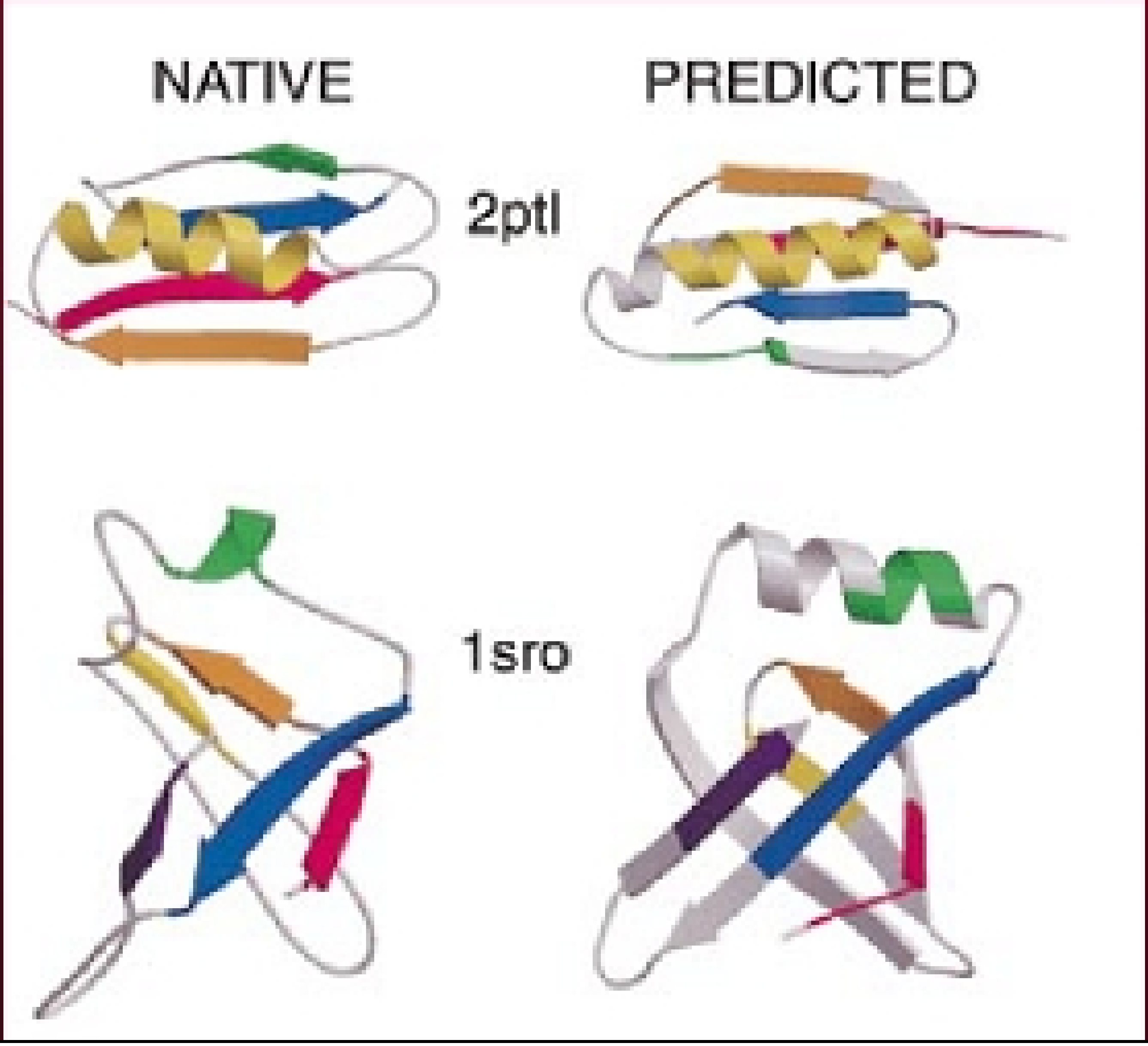
- Goal is to map regions of linear sequence to known folds in PDB
- Worked surprising;y well in 1994
  - Keeps getting a bit better
- Evaluate on RMS, electrostatics, hydrophobic burial, H-bonds, energetics
- Every Predictor got at least one right

Ab initio Prediction

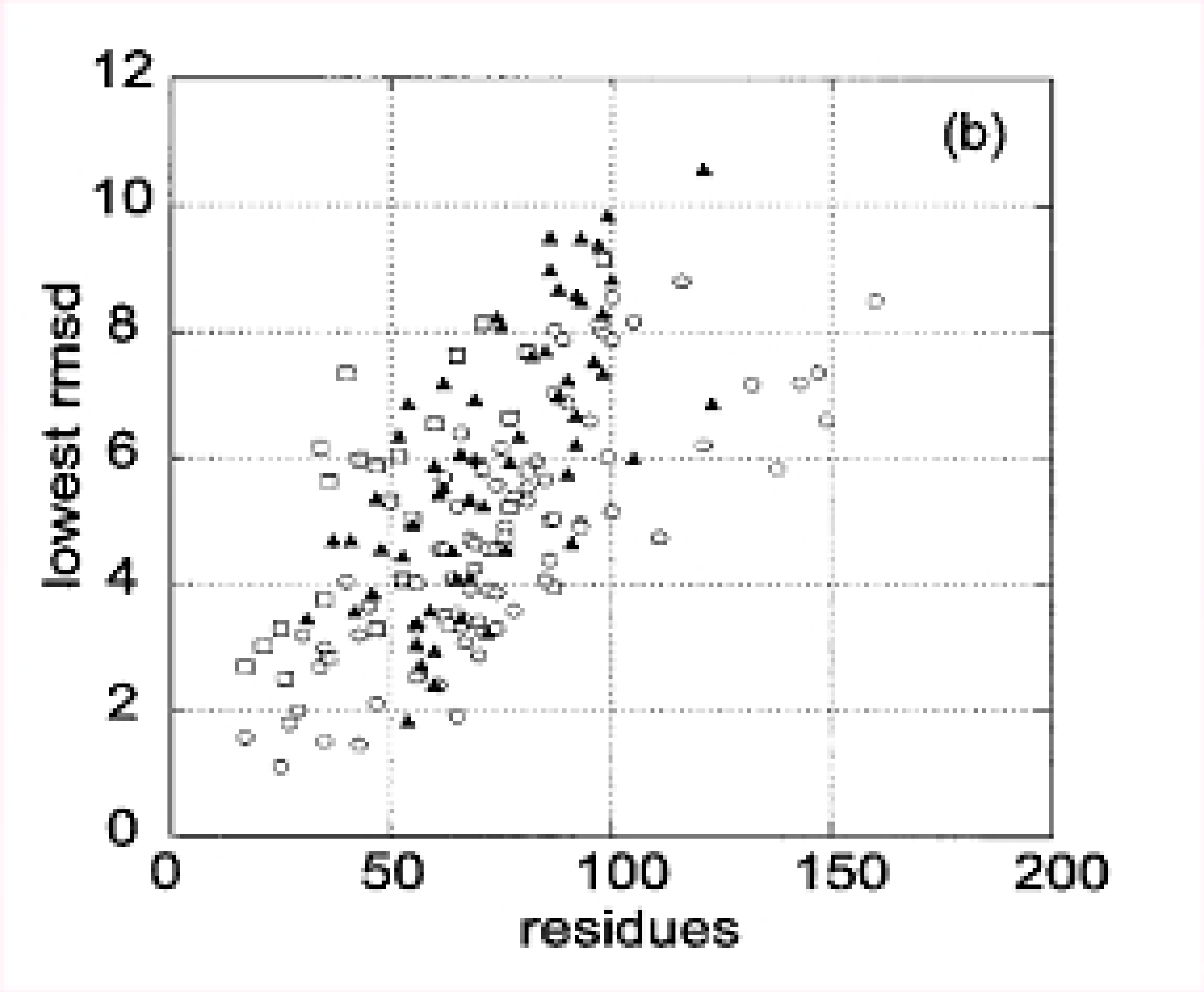
- Goal is secondary and/or 3-D structure
- Secondary
  - 66-77% correct
  - Errors not tolerable, need better techniques
- 3-D Structure
  - Rosetta Method

Rosetta Method

- Break target into 9 a.a. stretches
- Search PDB for that stretch of 9
- Align 9 to best match in PDB
- Steal structure around 9 from PDB
- Shift frame by 1 in linear sequence
- Loop
- Create thousands of structures and average


NATIVE
PREDICTED
2ptl
1sro

Performance of Rosetta Method



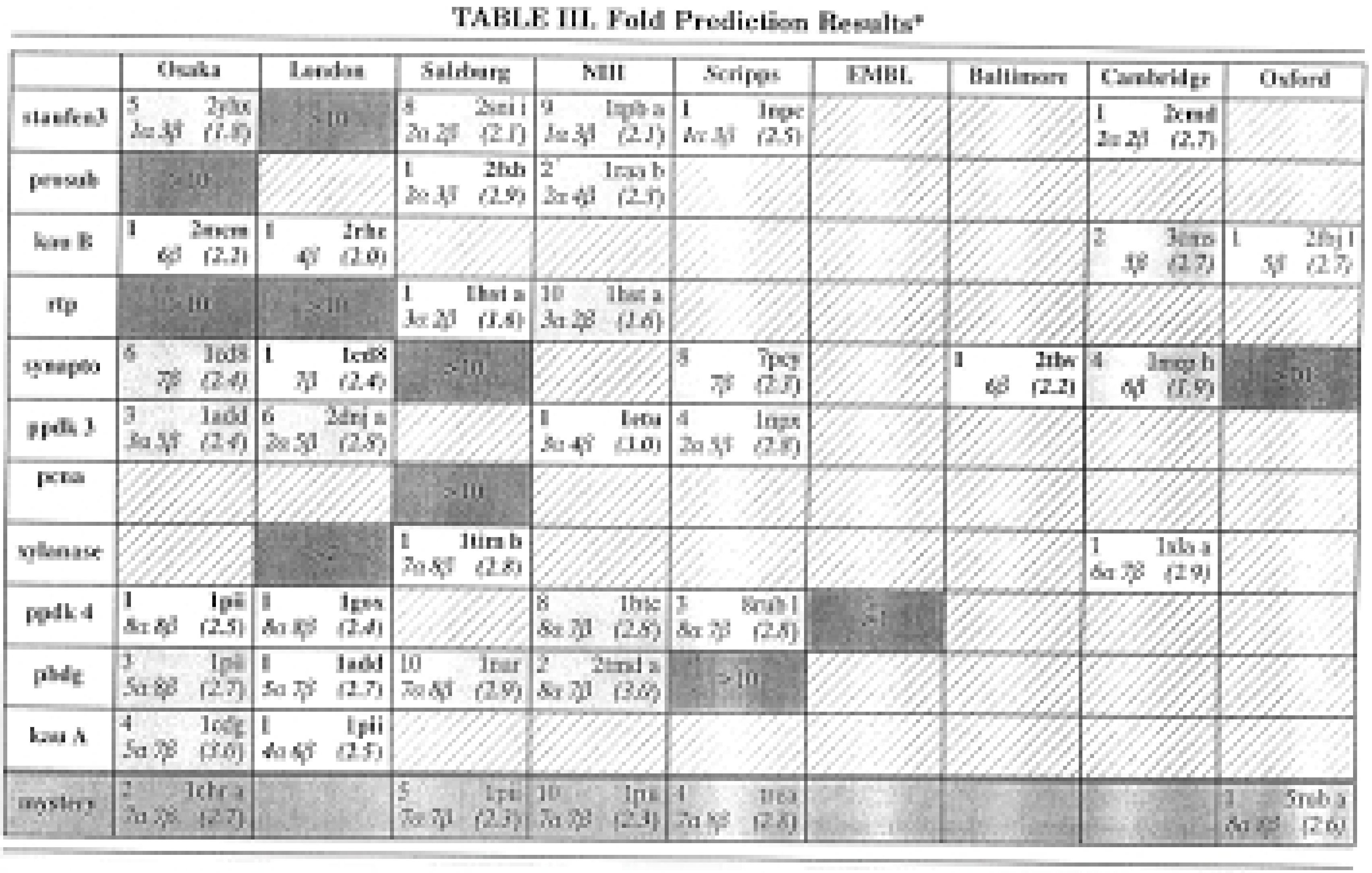


TABLE III. Fold Prediction Results*

	Osaka	London	Salzburg	NIH	Scripps	EMBL	Baltimore	Cambridge	Oxford
staufen3	5 2ylx 3α 3β (1.8)	>10	8 2sni i 2α 2β (2.1)	9 1gpb a 3α 3β (2.1)	1 1ege 4α 3β (2.5)	▨	▨	1 2cnd 2α 2β (2.7)	▨
presub	>10	▨	1 2fsh 2α 3β (2.9)	2 1rca b 2α 4β (2.5)	▨	▨	▨	▨	▨
kin B	1 2mcm 4β (2.2)	1 2rhe 4β (2.9)	▨	▨	▨	▨	▨	2 3tmo 3β (2.7)	1 2fbj l 5β (2.7)
rtp	>10	>10	1 1bst a 3α 2β (1.8)	10 1bst a 3α 2β (1.6)	▨	▨	▨	▨	▨
synapto	6 1td8 7β (2.4)	1 1cd8 7β (2.4)	>10	▨	8 7pcy 7β (2.7)	▨	1 2tbv 6β (2.2)	4 1mup b 6β (1.9)	>10
ppdk 3	3 1add 3α 3β (2.4)	6 2dnj a 2α 3β (2.9)	▨	1 1etu 3α 4β (1.0)	4 1ogs 2α 3β (2.8)	▨	▨	▨	▨
pcna	▨	▨	>10	▨	▨	▨	▨	▨	▨
xylanase	▨	?	1 1tim b 7α 8β (2.8)	▨	▨	▨	▨	1 1xis a 8α 7β (2.9)	▨
ppdk 4	1 1pii 8α 8β (2.5)	1 1gox 8α 8β (2.4)	▨	8 1btc 8α 7β (2.8)	3 8rub l 8α 7β (2.8)	>10	▨	▨	▨
pbdg	3 1pii 5α 8β (2.7)	1 1add 5α 7β (2.7)	10 1nar 7α 8β (2.9)	2 2tmd a 8α 7β (3.6)	>10	▨	▨	▨	▨
kin A	4 1cdg 5α 7β (3.6)	1 1pii 4α 6β (2.5)	▨	▨	▨	▨	▨	▨	▨
mystery	2 1chr a 7α 7β (2.7)		5 1pii 7α 7β (2.3)	10 1pii 7α 7β (2.3)	4 1tca 7α 8β (2.8)				1 5rub a 8α 8β (2.6)

Predictors are along top row. Target sequences along first column. Dark grey means bad prediction, light gray pretty good, white very good. Hatched means no prediction. Upper left corner shows rank of best answer among list submitted by predictors (also shows fold used to make prediction, shift error and general protein class)

Acknowledgments

- Doug Brutlag
- Russ B. Altman
- Mike Levitt
- Amit P. Singh
- Tommy Liu