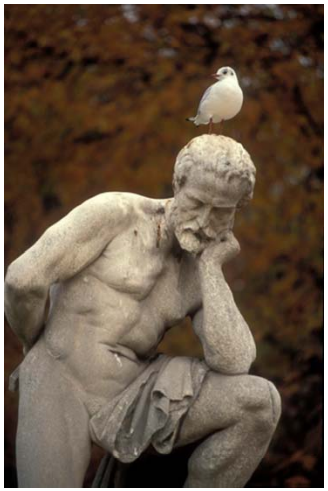
Computational Structural Biology

Jens Meiler

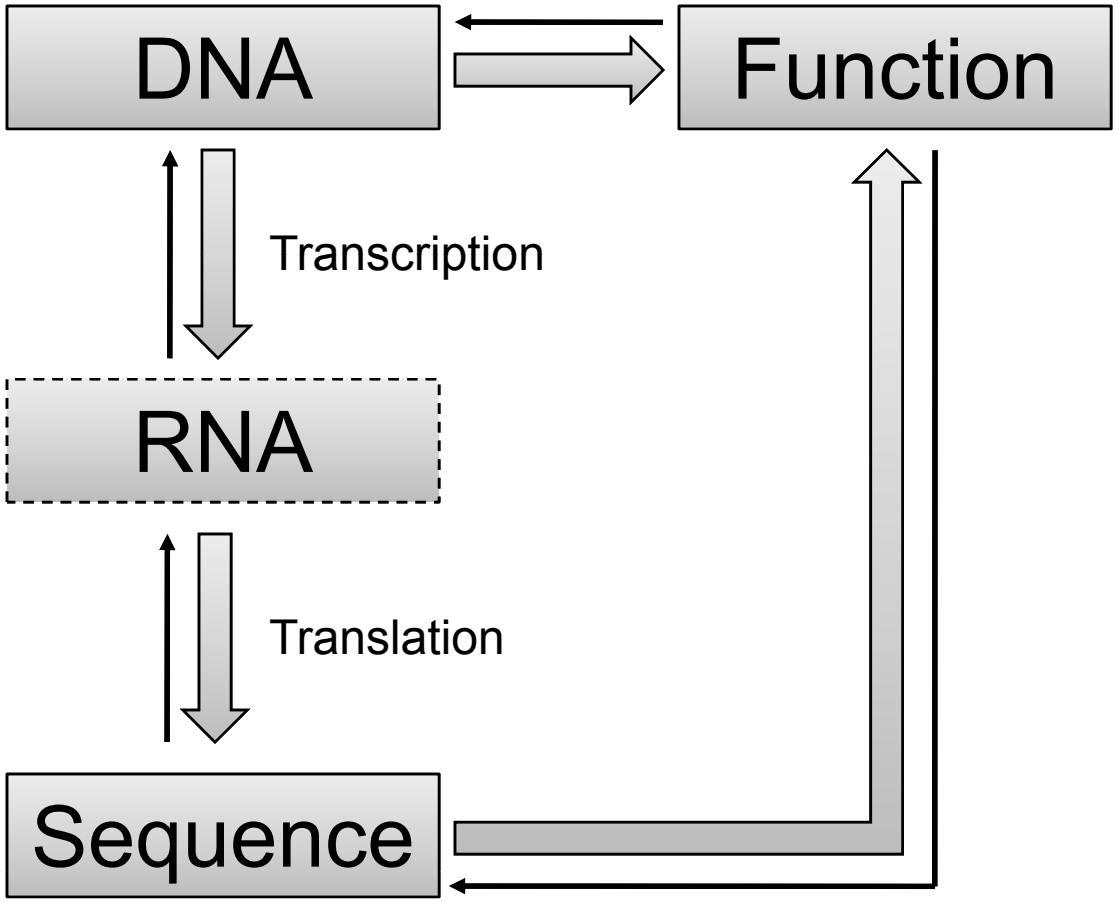
Center for Structural Biology
Departments of Chemistry, Pharmacology, and
Biomedical Informatics

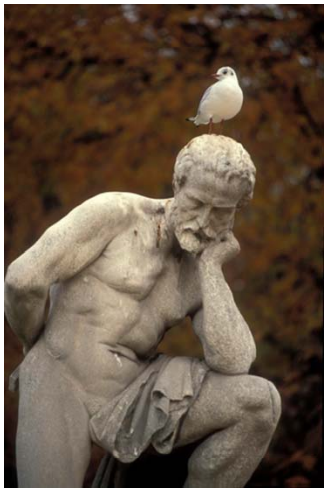
Central Dogma of Evolution

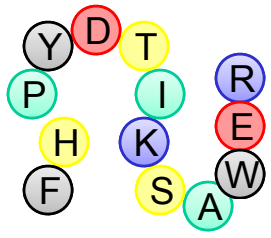
DNA ⇄ Function



Central Dogma of Molecular Biology







Central Dogma of Structural Biology

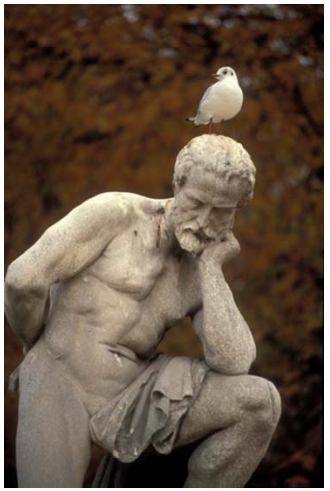
DNA

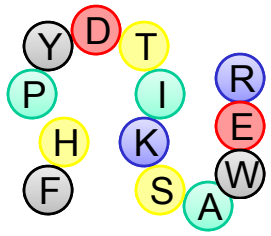
Function

Transcription

RNA

Translation





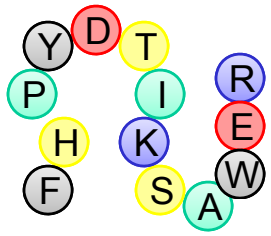
Folding

Sequence

Structure

Protein Sequence and Structure Data

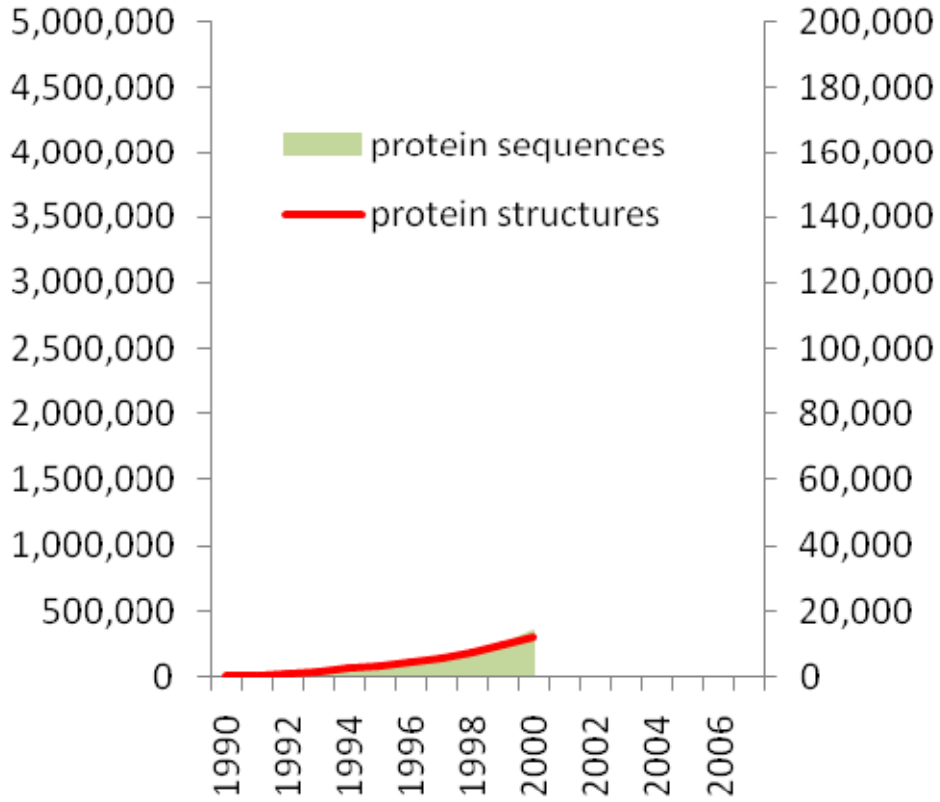
- Genbank
  - ~5,000,000 sequences
- Protein Databank
  - ~40,000 structures

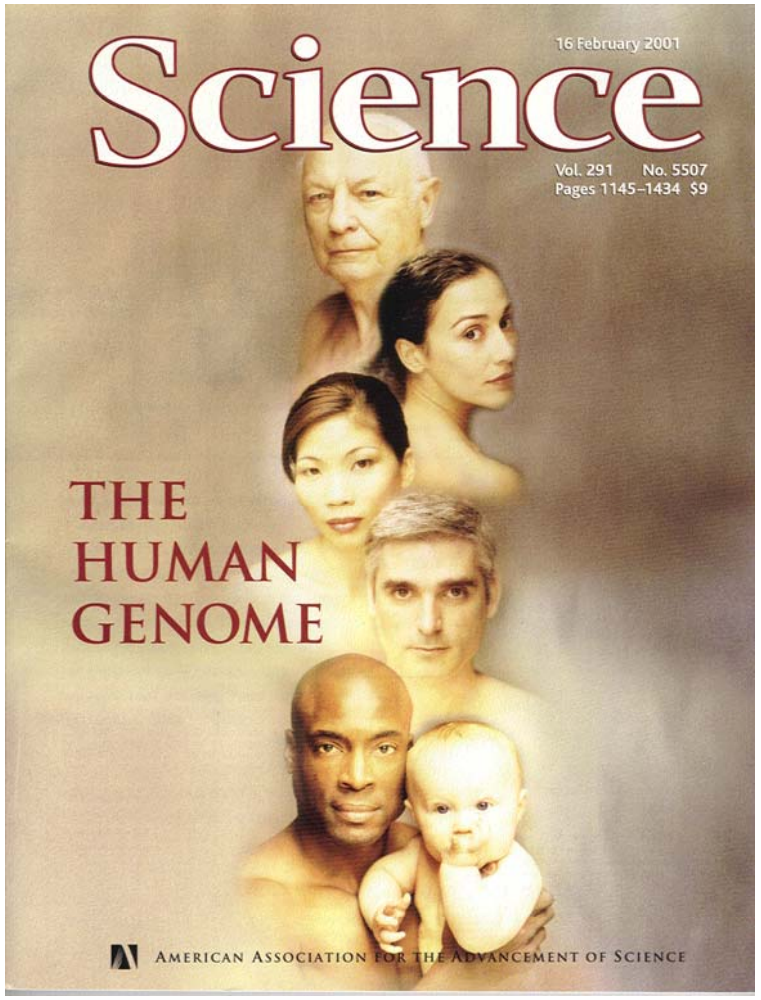


Sequence ⇌ Folding ⇌ Structure

Structural Biology After the Human Genome Project

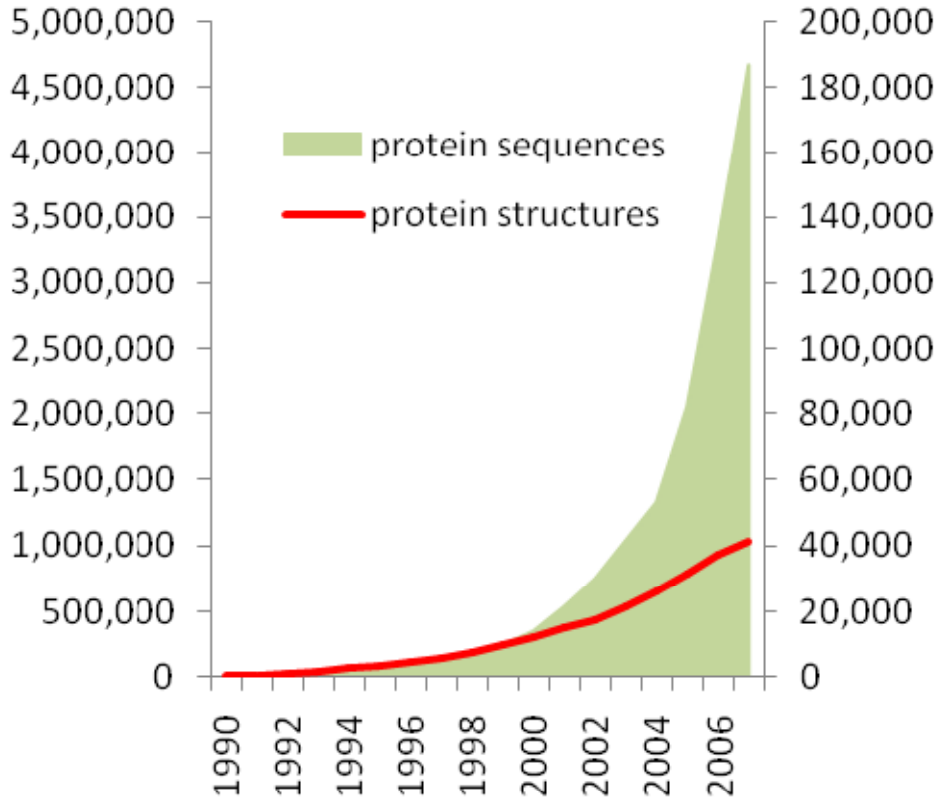
- Sequence versus Structure

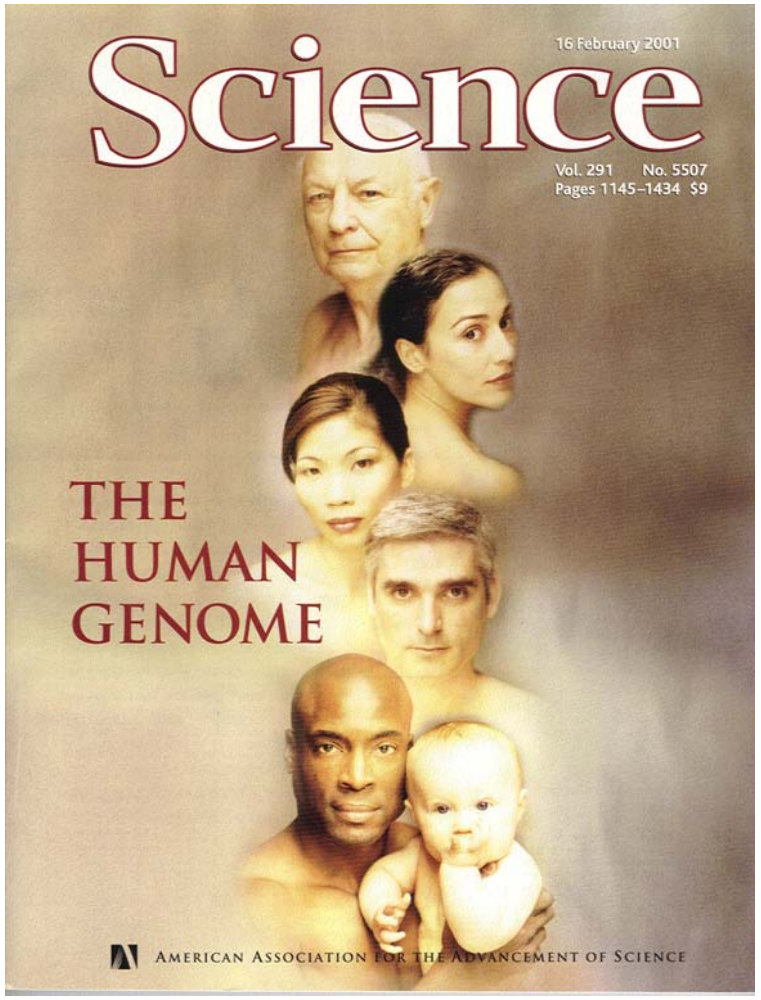




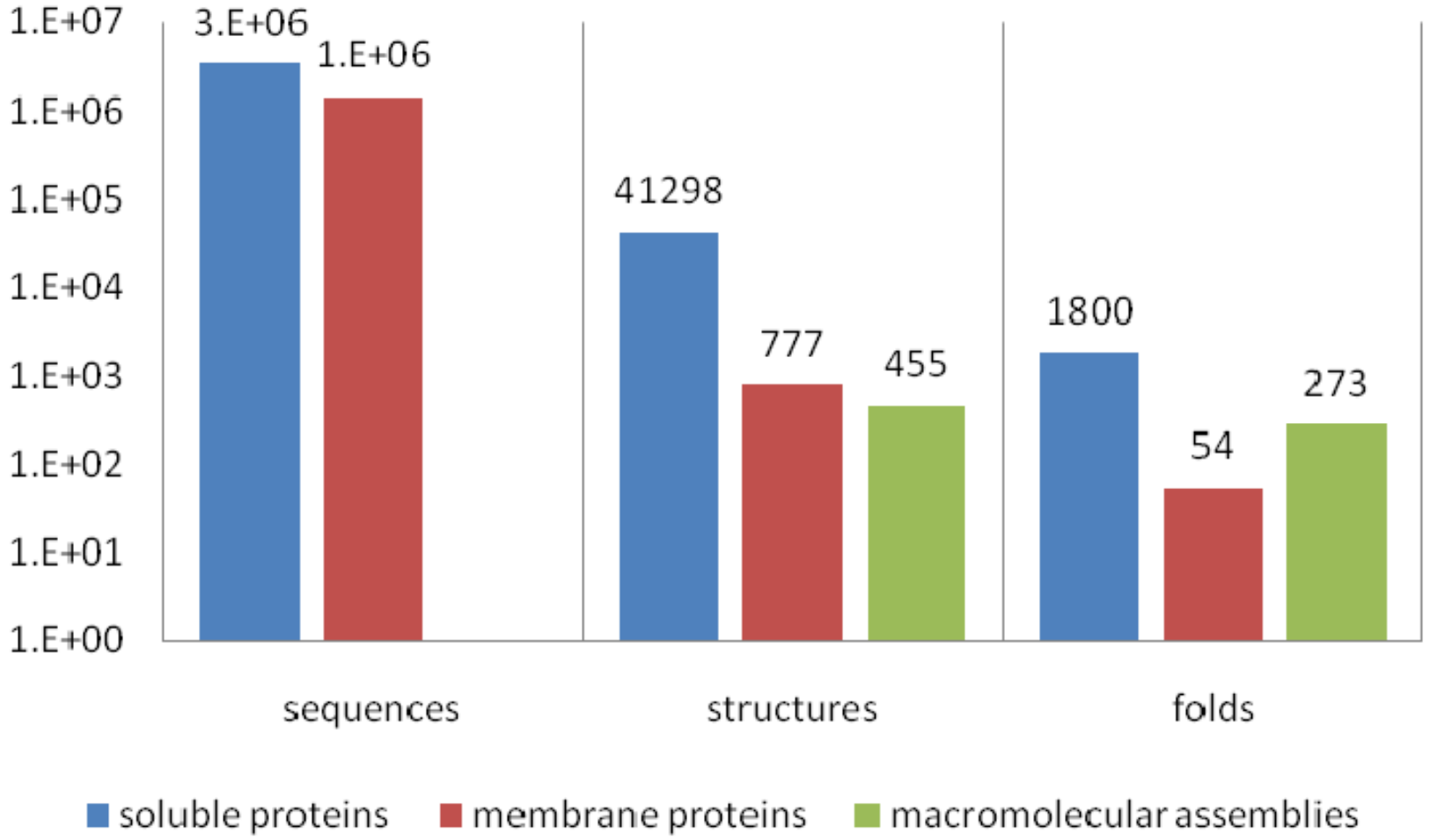
Structural Biology After the Human Genome Project

- Sequence versus Structure

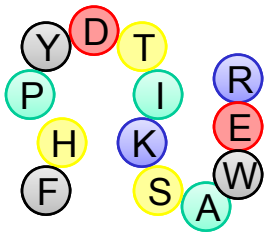




Membrane Proteins and Large Macromolecular Assemblies



(Inverse) Protein Folding Problem
Holy Grail of Comp. Struct. Biology



Protein Folding Problem

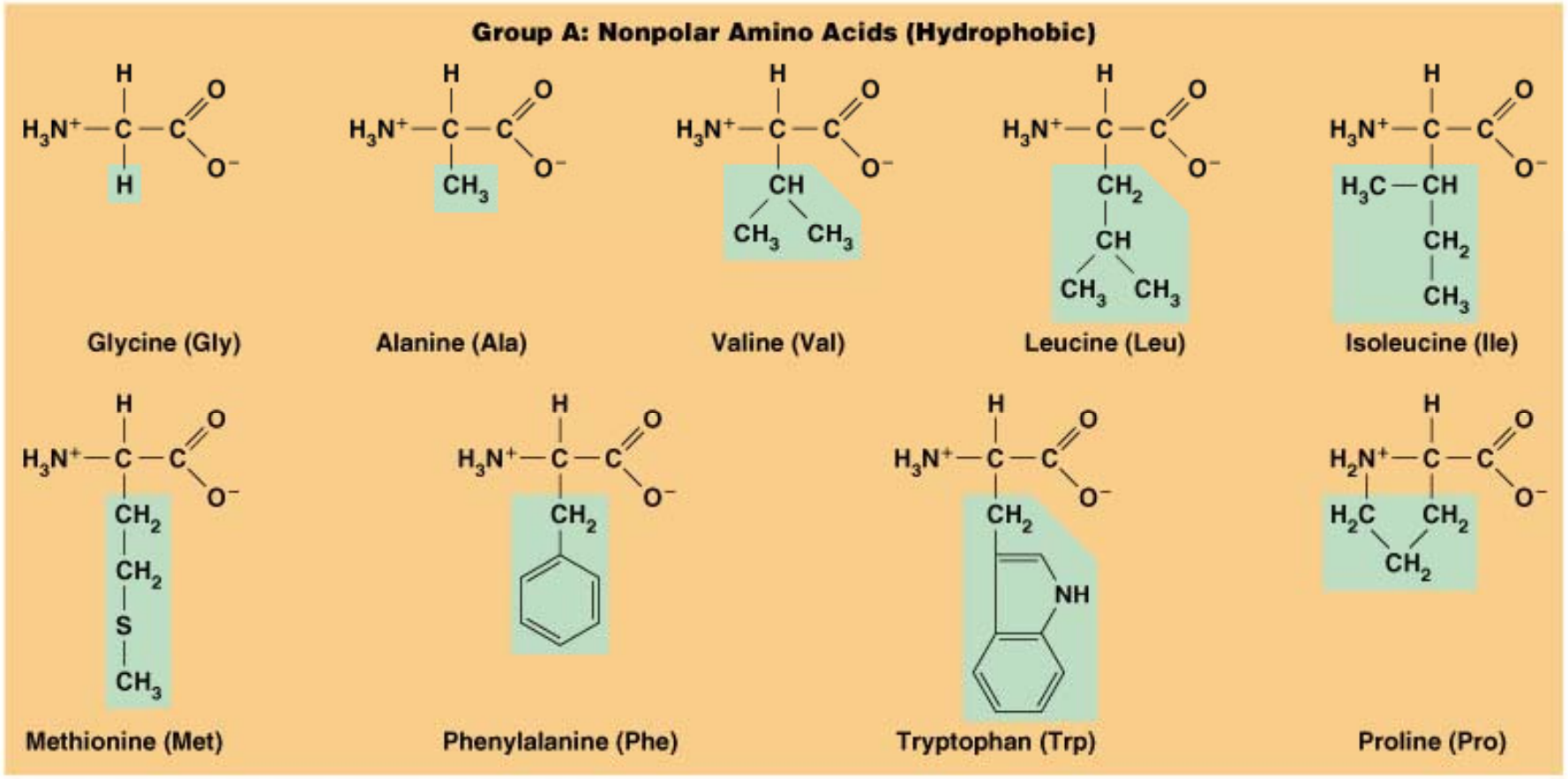
Sequence ⇒ Structure

Inverse Folding Problem

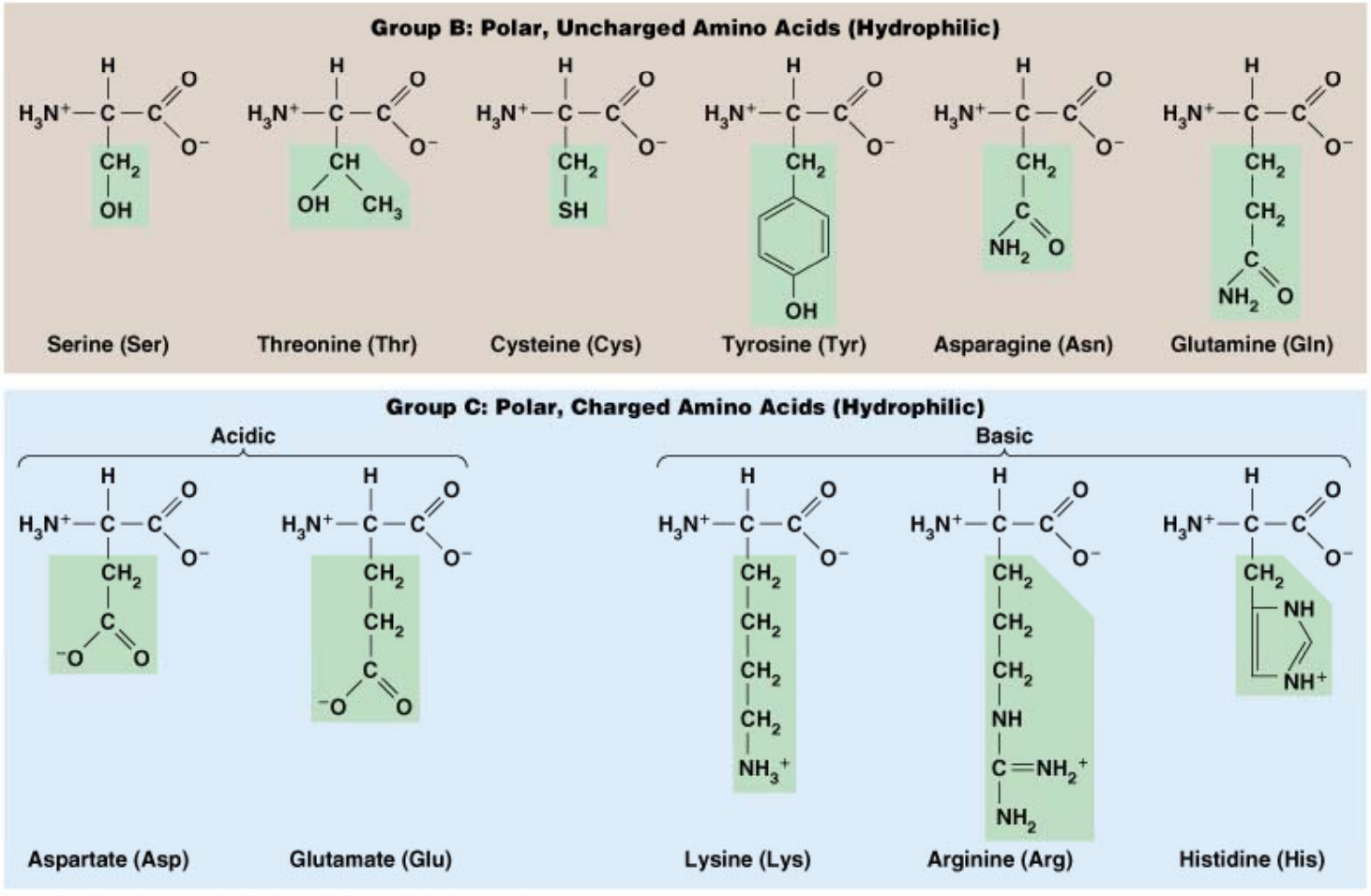
- Given a protein's AA sequence, what is its 3-dimensional fold , and how does it get there?
- Assume 100 conformations for each amino acid in a 100 amino acid protein ⇨ 10^{200} possible conformations!
- Cyrus Levinthal's paradox of protein folding,1968.

- Given a protein fold, which primary sequence(s) fold into it?
- Assume a total of 100 conformations for all 20 natural occurring amino acids side chains in a 100 amino acid protein ⇨ 10^{200} possible conformations!
- Earth is less than 10^{10} years old.

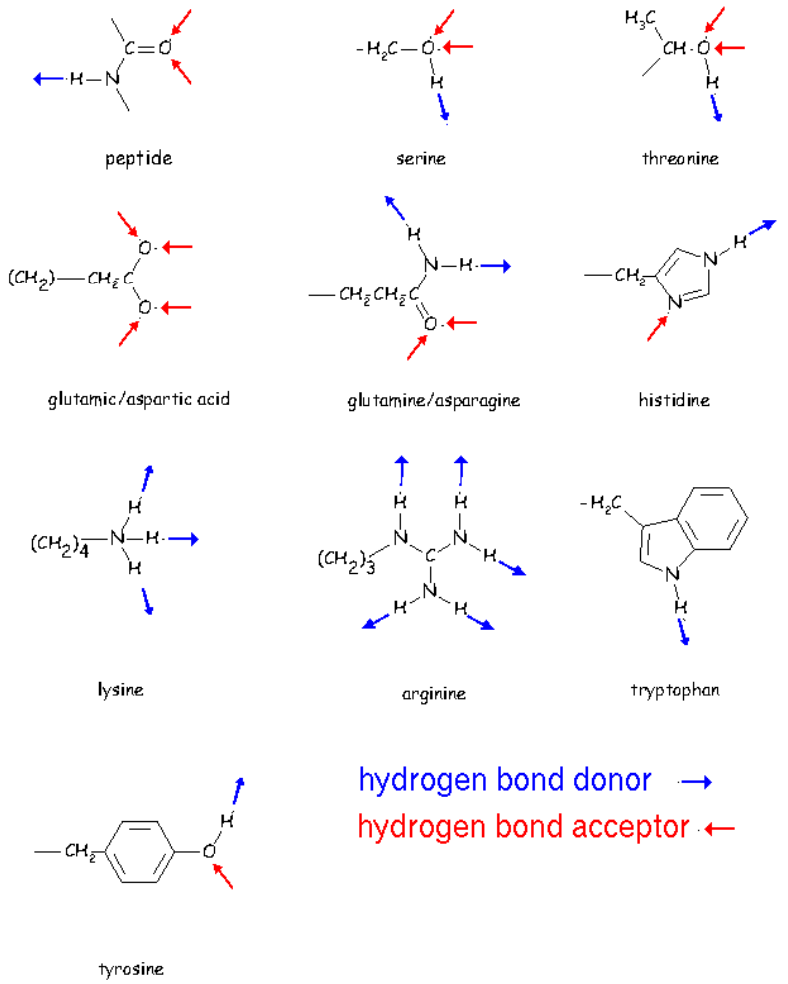
Hydrophobic Amino Acids



Hydrophylic Amino Acids



Hydrogen Bonding Capabilities of Amino Acids



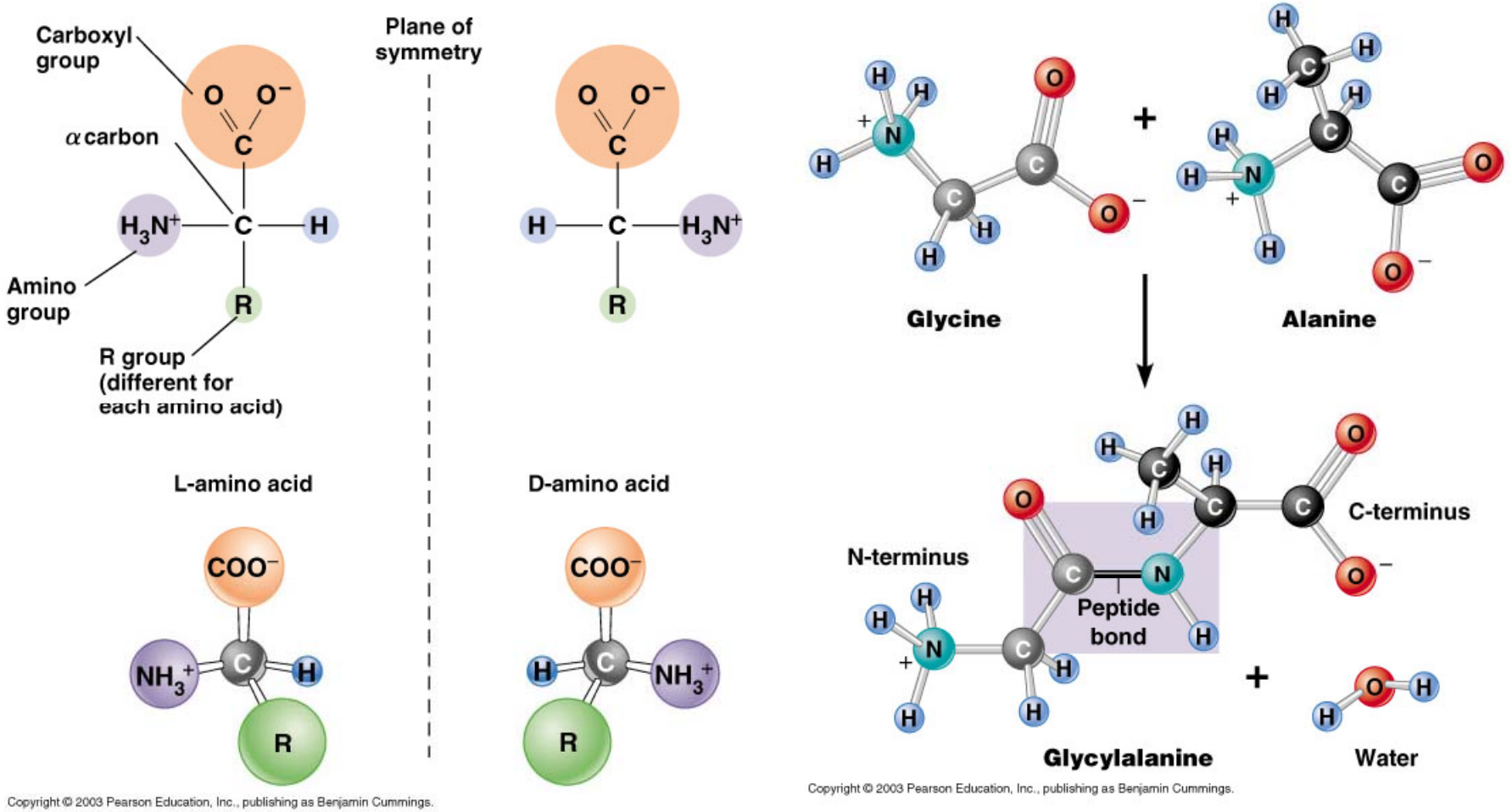
Properties of Amino Acids in Numbers

Table 1 Amino acid parameter sets

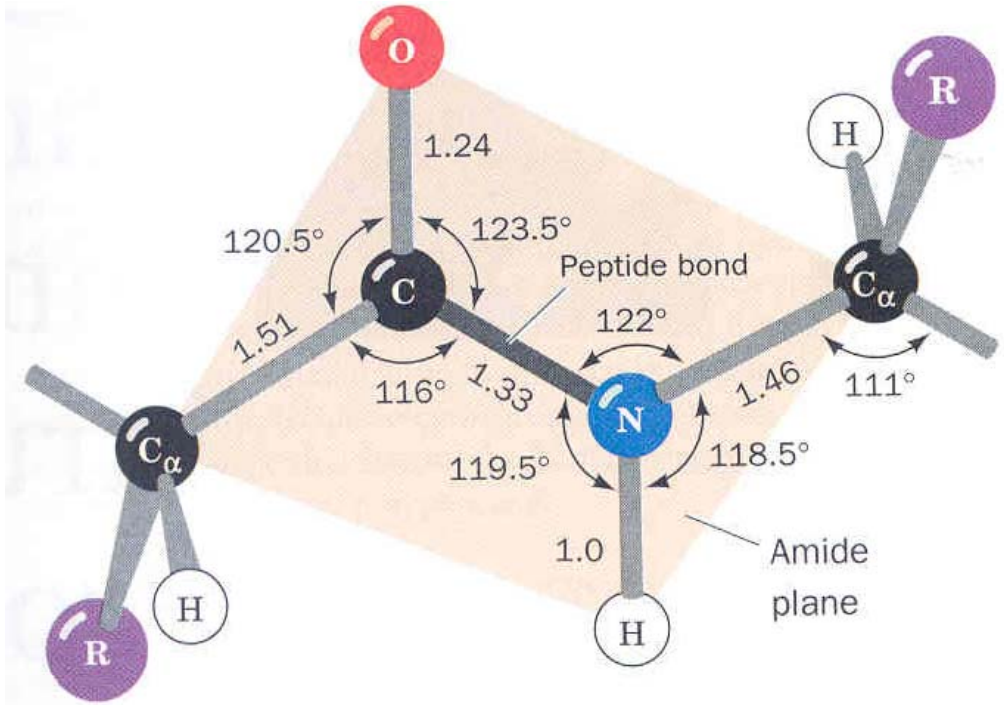
Name	Ξ[a]	α[b]	υ_v [c]	π[d]	I[e]	α[f]	β[g]
ALA	1.28	0.05	1.00	0.31	6.11	0.42	0.23
GLY	0.00	0.00	0.00	0.00	6.07	0.13	0.15
VAL	3.67	0.14	3.00	1.22	6.02	0.27	0.49
LEU	2.59	0.19	4.00	1.70	6.04	0.39	0.31
ILE	4.19	0.19	4.00	1.80	6.04	0.30	0.45
PHE	2.94	0.29	5.89	1.79	5.67	0.30	0.38
TYR	2.94	0.30	6.47	0.96	5.66	0.25	0.41
TRP	3.21	0.41	8.08	2.25	5.94	0.32	0.42
THR	3.03	0.11	2.60	0.26	5.60	0.21	0.36
SER	1.31	0.06	1.60	-0.04	5.70	0.20	0.28
ARG	2.34	0.29	6.13	-1.01	10.74	0.36	0.25
LYS	1.89	0.22	4.77	-0.99	9.99	0.32	0.27
HIS	2.99	0.23	4.66	0.13	7.69	0.27	0.30
ASP	1.60	0.11	2.78	-0.77	2.95	0.25	0.20
GLU	1.56	0.15	3.78	-0.64	3.09	0.42	0.21
ASN	1.60	0.13	2.95	-0.60	6.52	0.21	0.22
GLN	1.56	0.18	3.95	-0.22	5.65	0.36	0.25
MET	2.35	0.22	4.43	1.23	5.71	0.38	0.32
PRO	2.67	0.00	2.72	0.72	6.80	0.13	0.34
CYS	1.77	0.13	2.43	1.54	6.35	0.17	0.41

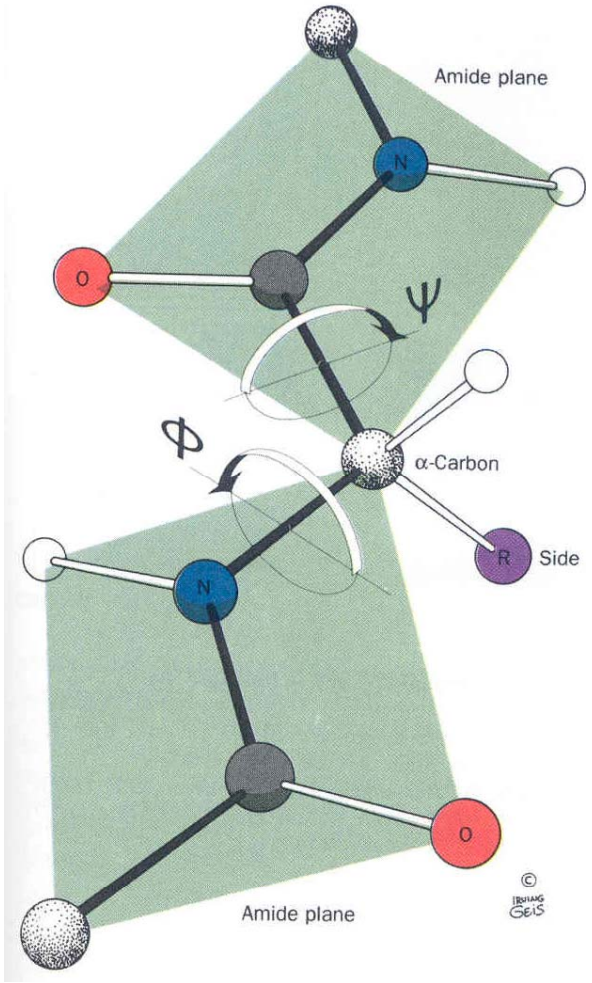
[a] Steric parameter (graph shape index)
[b] Polarizability
[c] Volume (normalized van der Waals volume)
[d] Hydrophobicity
[e] Isoelectric point
[f] Helix probability
[g] Sheet probability

Stereochemistry of amino acids and Planarity of peptide bond

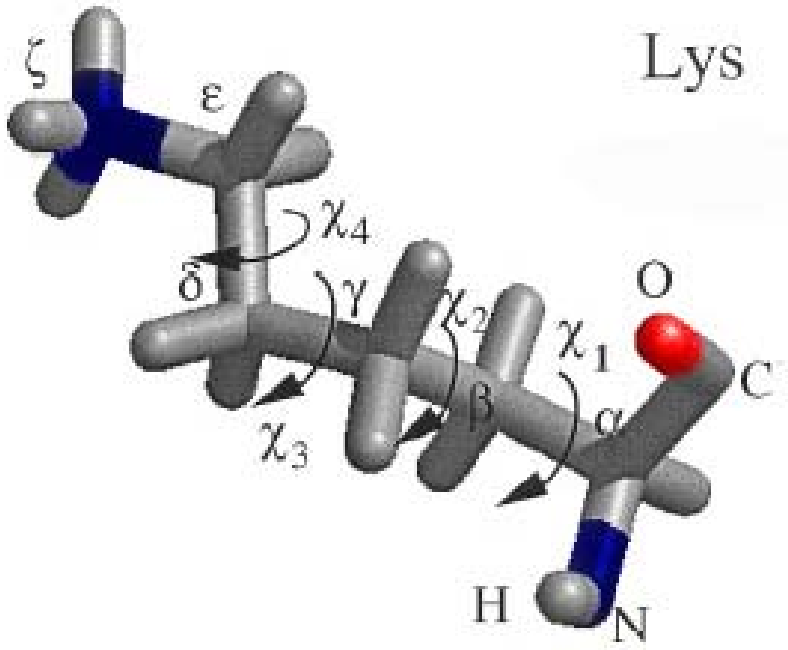


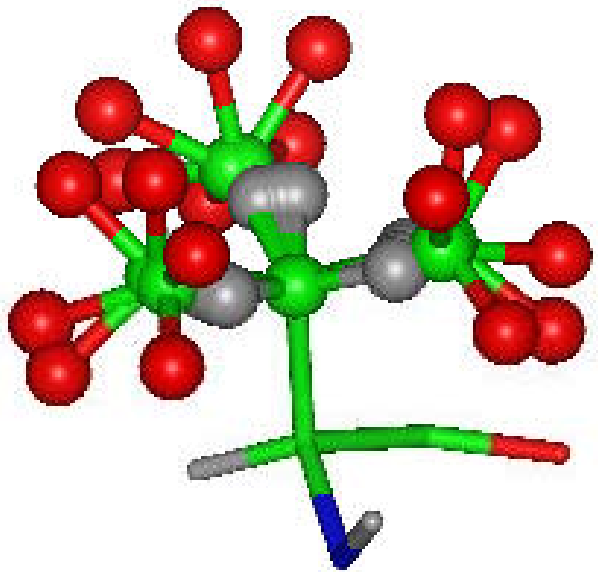
Two Backbone Degrees of Freedom per Amino Acid



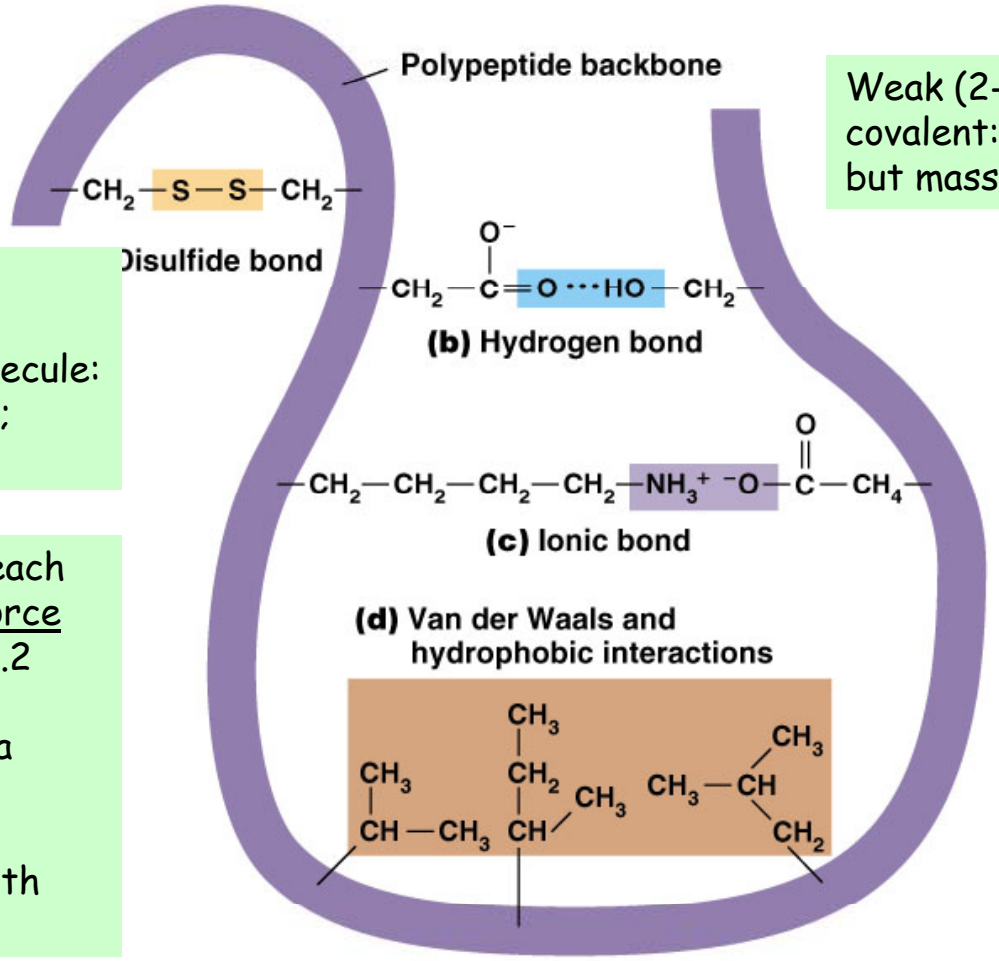


Zero toFour Sidechain Degrees of Freedom – Two on Average





Important bonds for protein folding and stability



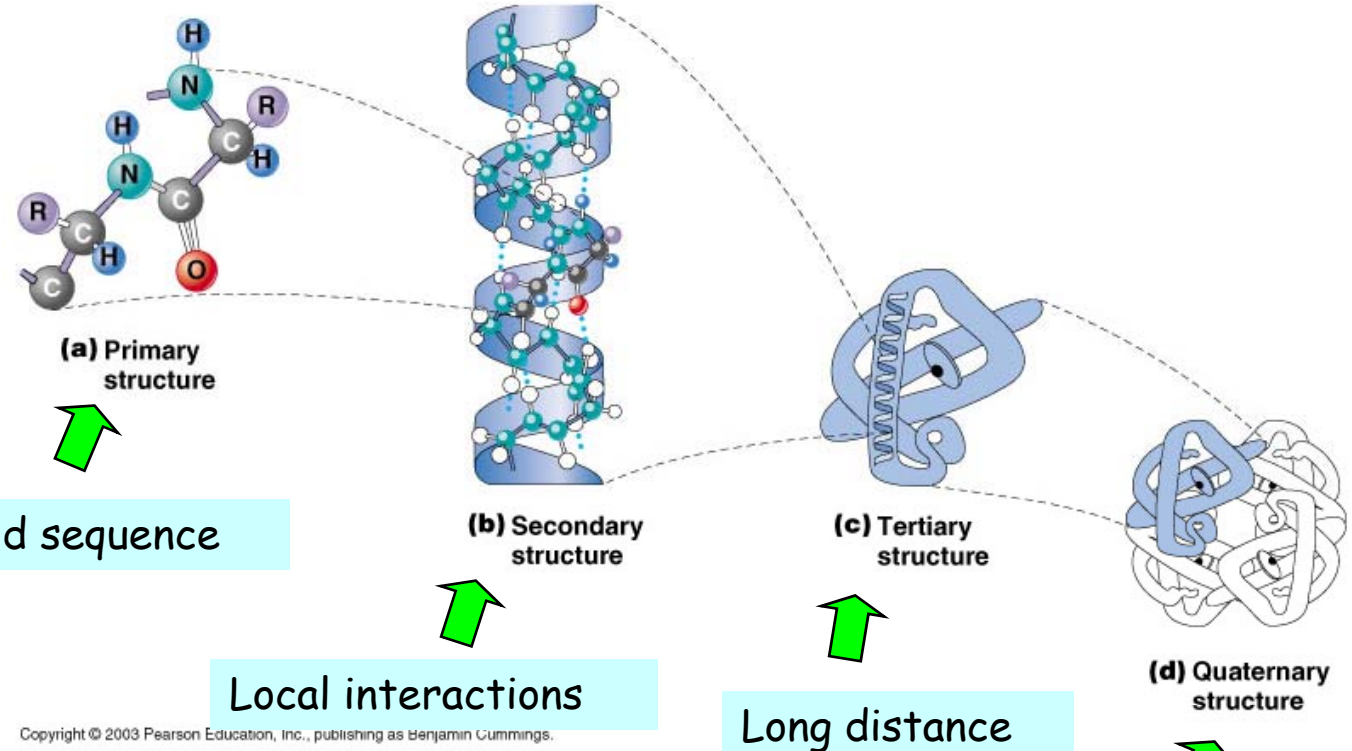
Weak (2-5 kcal/mol vs. covalent: 70-100 kcal/mol), but massive

The oxidization of the sulfhydryl groups of two cystein residues (intramolecule: ribonuclease; intersubunit; dimeric protein insulin)

Weak (3 kcal/mol), affected by pH value

Dipole molecules attract each other by van der Waals force (transient and weak: 0.1-0.2 kcal/mol)
Hydrophobic interaction, a tendency of hydrophobic groups or molecules being excluded from interact with hydrophilic environment

Protein structure depends on amino acid sequence and interactions



Amino acid sequence

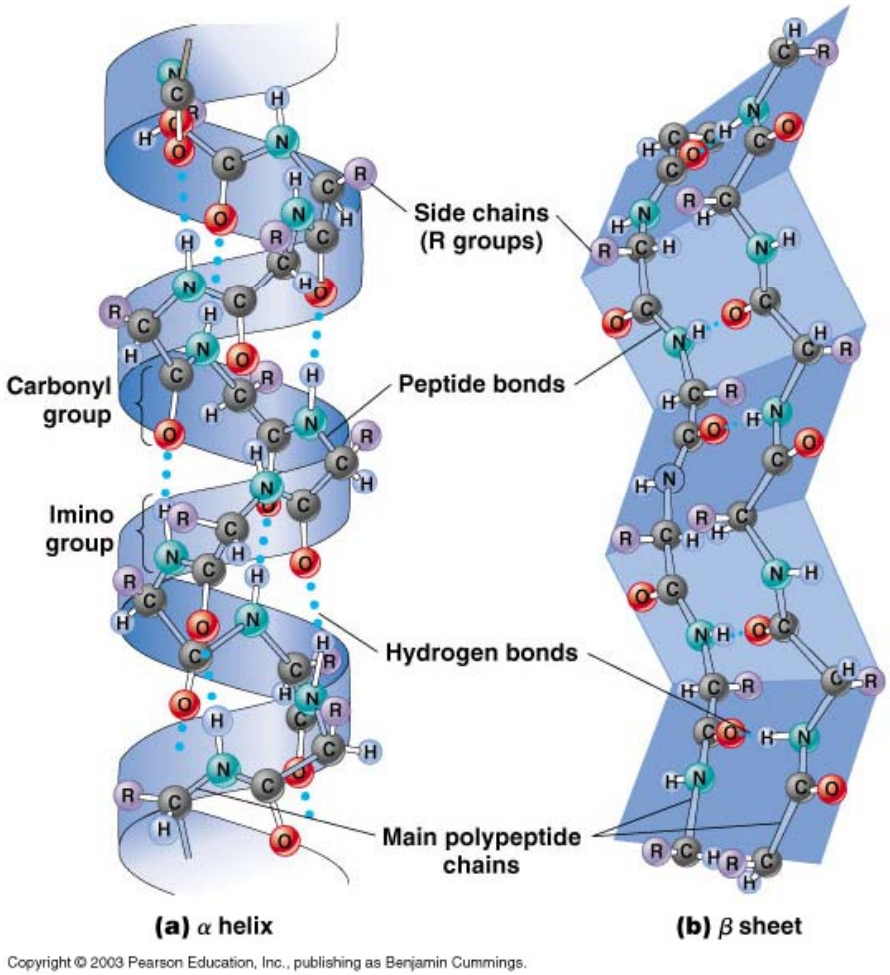
Local interactions

Long distance interactions

Interactions between subunits

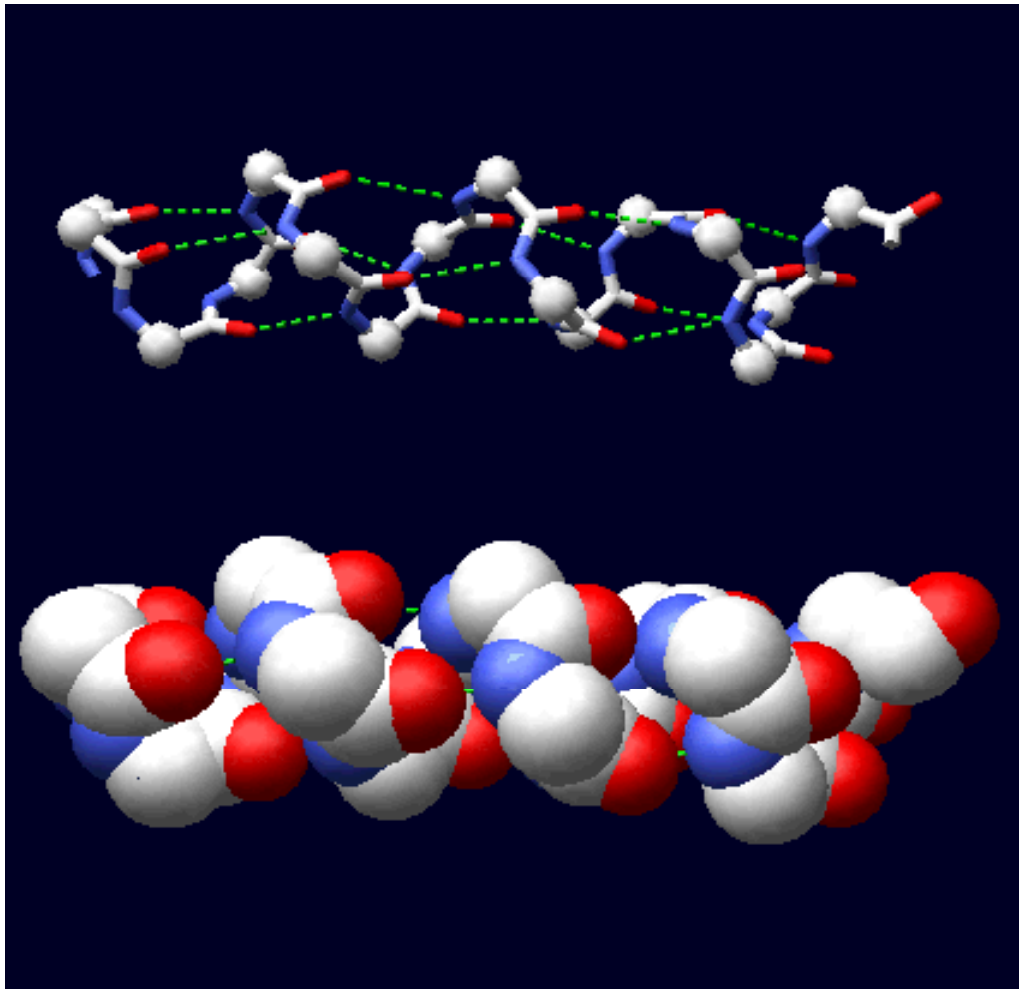
Secondary Structure: Build from Backbone Hydrogen Bonds

- *α-Helix:*
  - *Periodicity = 3.6*
  - *Rise = 1.5Å*
  - *Pitch = 5.4Å*
- *β-Sheet:*
  - *Periodicity = 2*
  - *Translation = 3.4Å*
  - *Distance = 5.4Å*



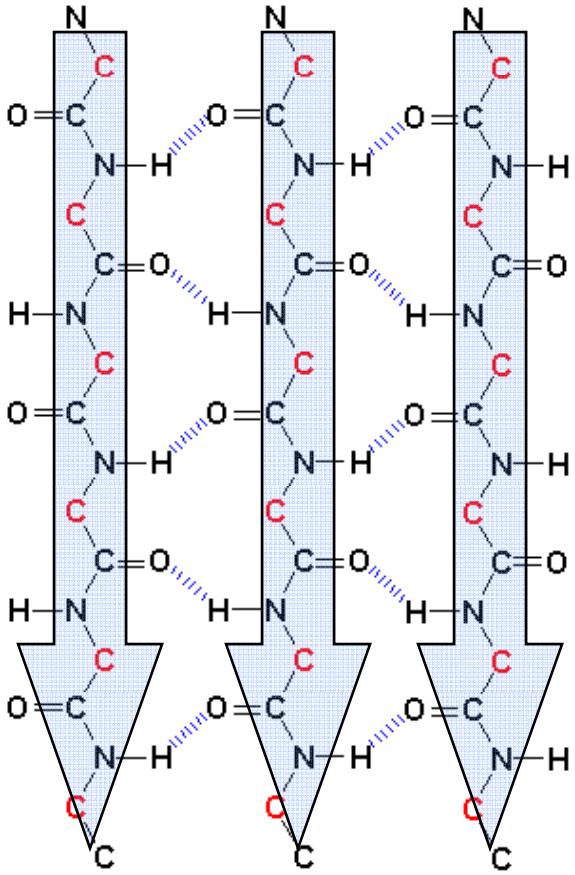
α-Helix

- *Most abundant secondary structure*
- *3.6 amino acids per turn*
- *Hydrogen bond formed between every fourth reside*
- *Average length: 10 amino acids, or 3 turns*
- *Varies from 5 to 40 amino acids*

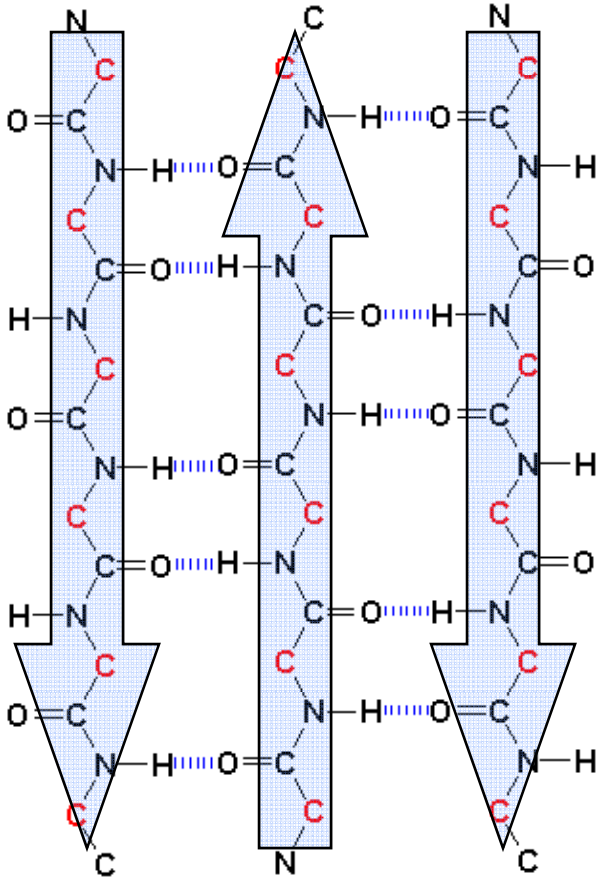


β-Sheet

- *5-10 amino acids in one portion of the chain with another 5-10 farther down the chain*
- *Interacting regions may be adjacent with a short loop, or far apart*

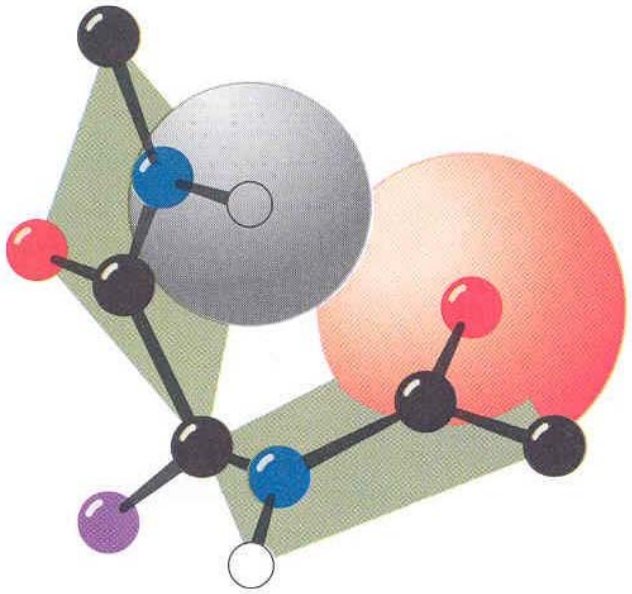


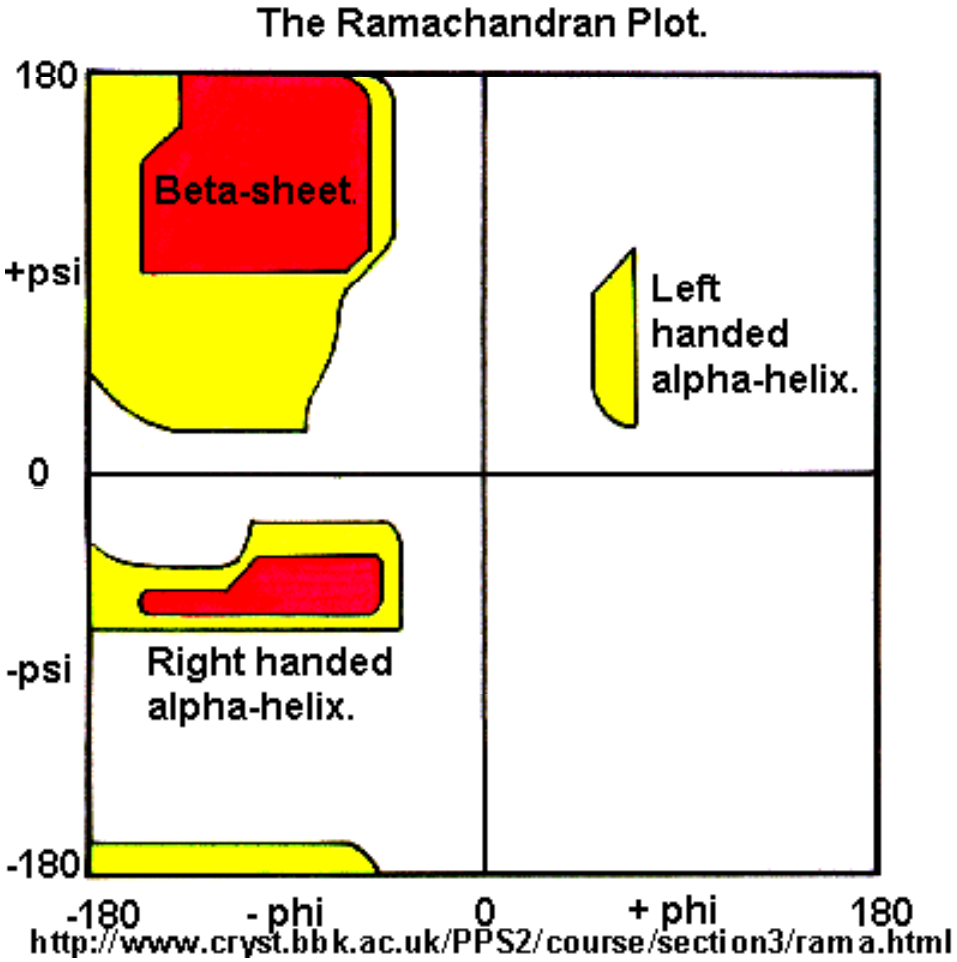
Parallel β-sheet



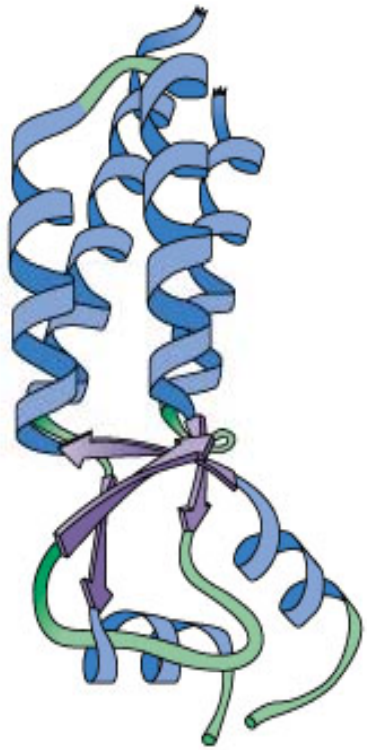
Anti-Parallel β-sheet

Sheet the Ramachandran Plot



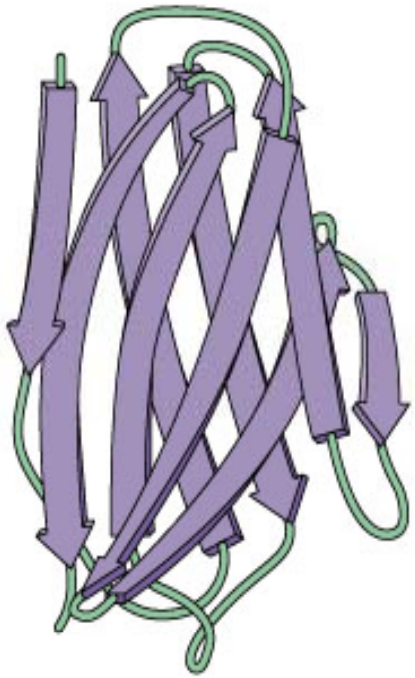


Tertiary structure: Assembly of Secondary Structure in Domains



Tobacco mosaic coat protein

(a) Predominantly α helix

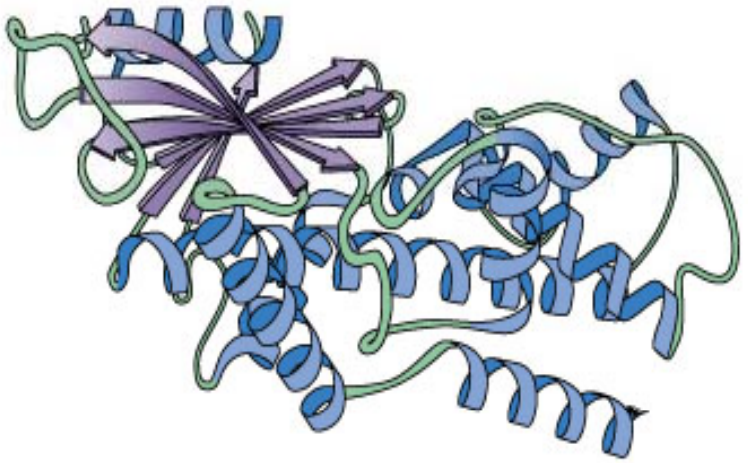


Immunoglobulin, V_2 domain

(b) Predominantly β sheet

Domains

Discrete locally folded units of tertiary structure, often containing regions of alpha helix and beta sheets packed together compactly, typically 50-350 aa in length and usually has a specific function.

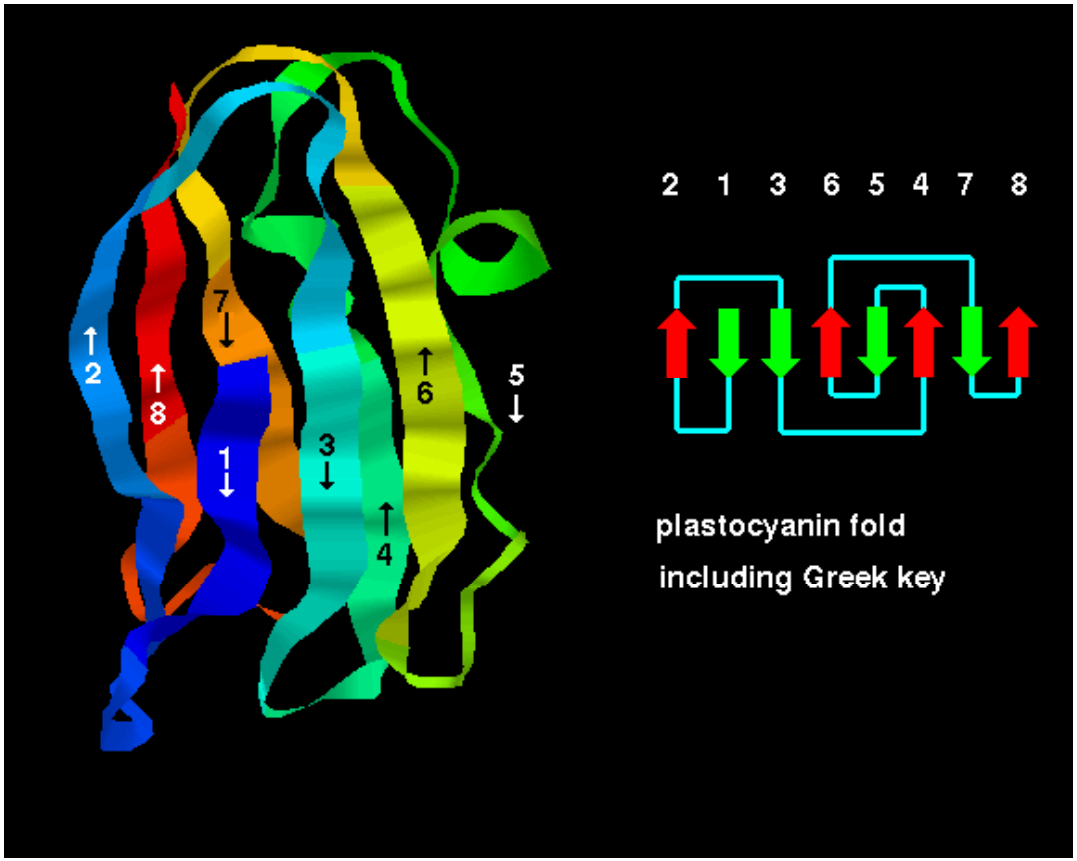


Hexokinase, domain 2

(c) Mixed α helix and β sheet

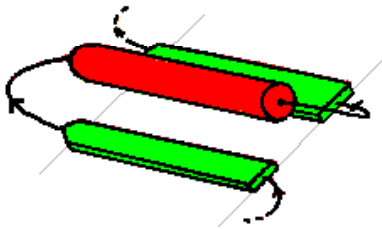
Tertiary Structure: Sheet-Sheet Packing - The Creek Key motif

- *These motifs occur quite frequently in Nature. Theoretical studies suggest that this motif may have originated from one long antiparallel structure with loops in the middle of both strands*



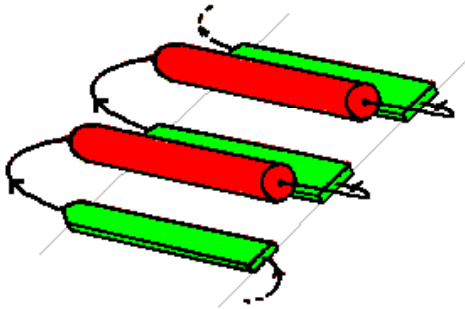
Tertiary Structure: Helix-Sheet Packing – The Rossman Fold

- *The most regular and common domain structures consist of repeating β-α-β units. The outer layer of the structure is composed of α helices packing against a central core of parallel β sheets. These folds are called β/α/β. This motif is always right-handed*



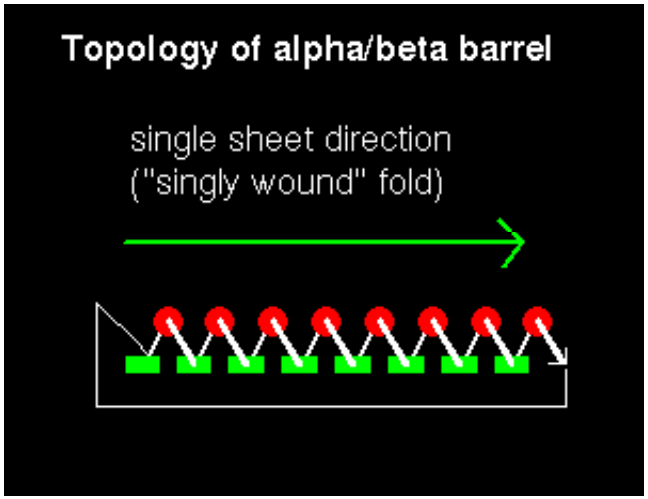
The right–handed beta–alpha–beta unit. The helix lies above the plane of the strands.

The Rossman fold

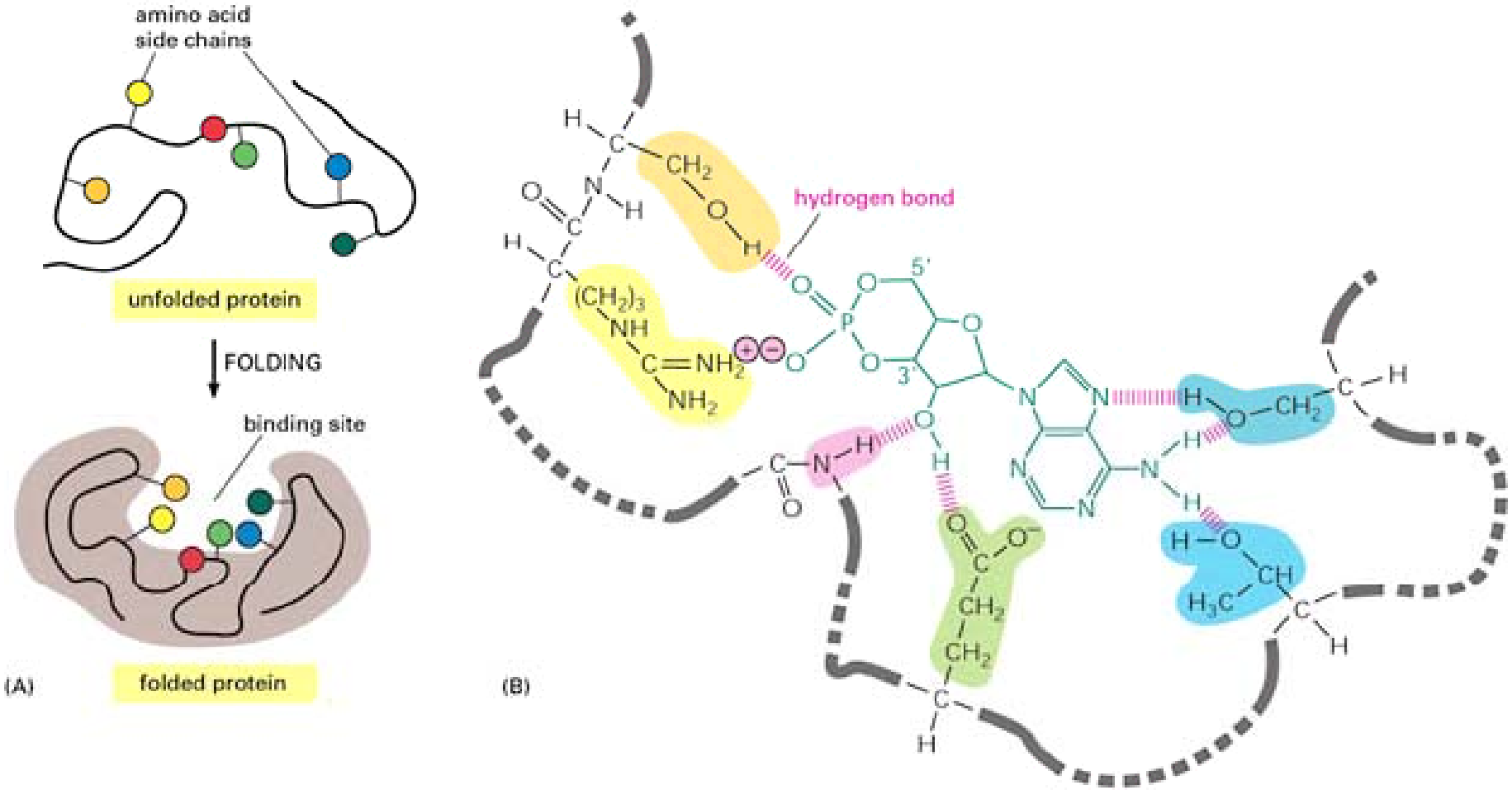


Tertiary Structure: Helix-Sheet Packing – The Barrel Fold

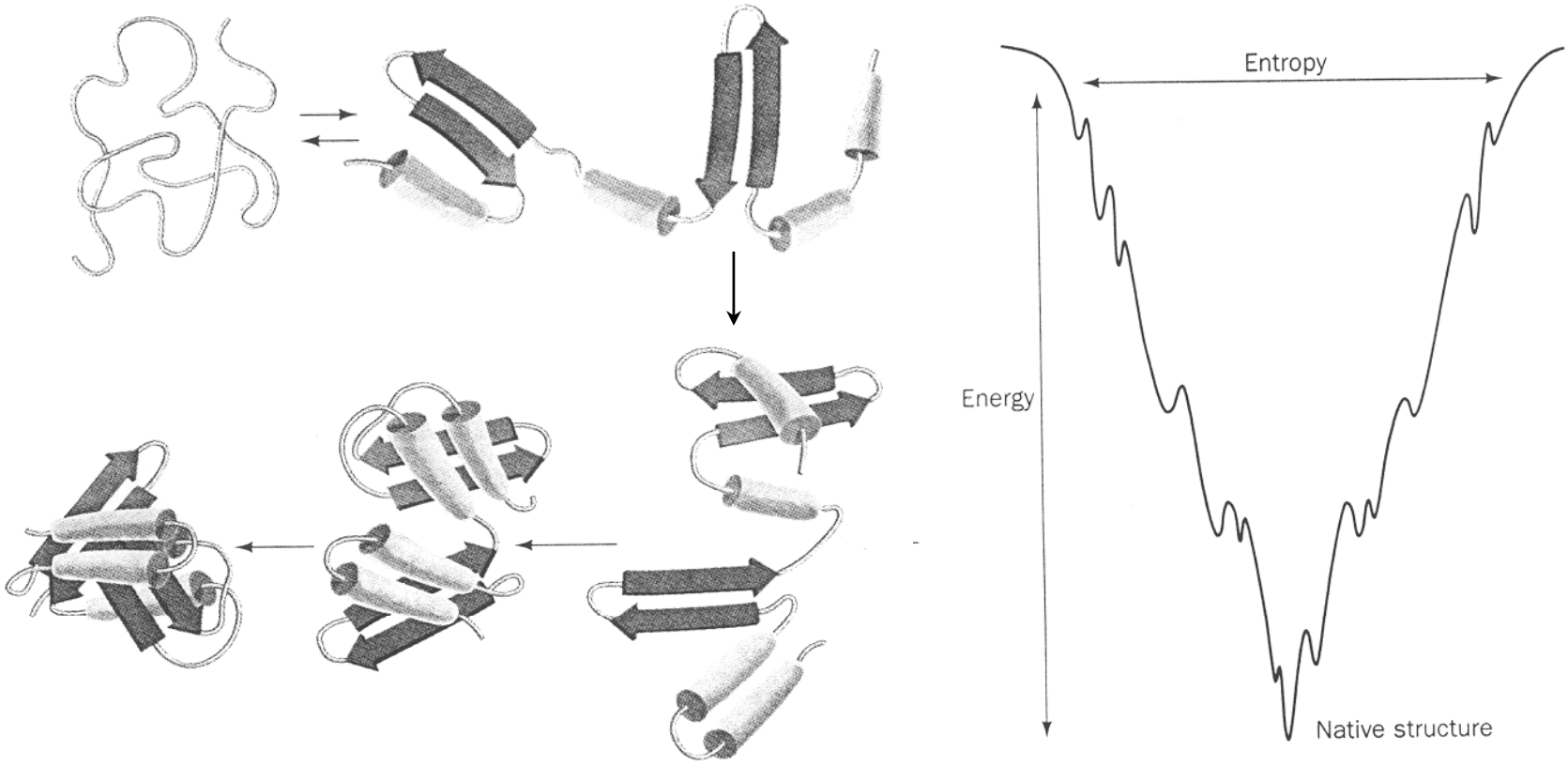
- *This structural motif was first observed in the X-ray structure of triosephosphate isomerase (TIM) so it is also called a TIM fold. (Banner et al., Nature 255, 609-614 (1975).*



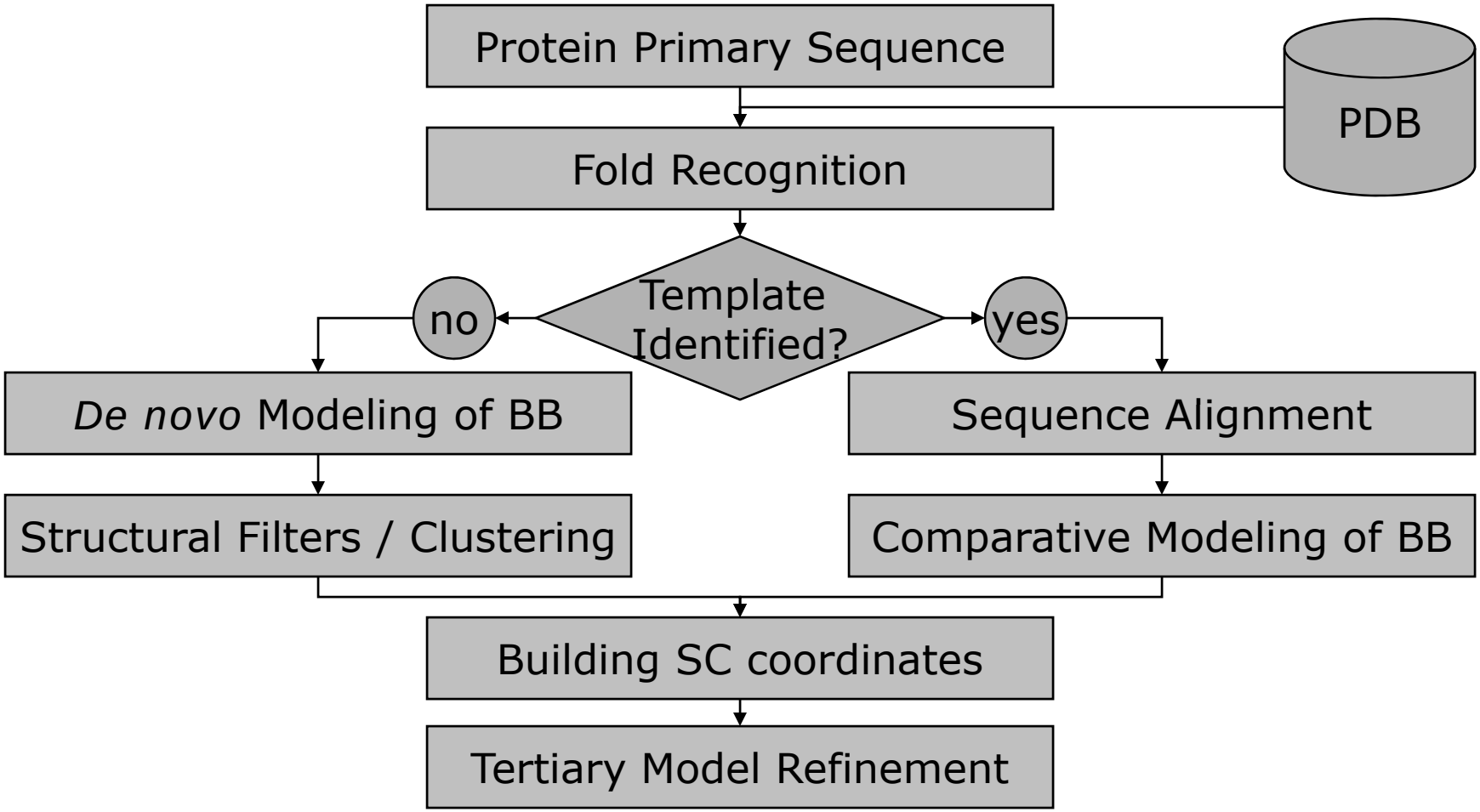
Protein Tertiary Structure is Tied to Function



Protein Folding



General Scheme of Protein Structure Prediction

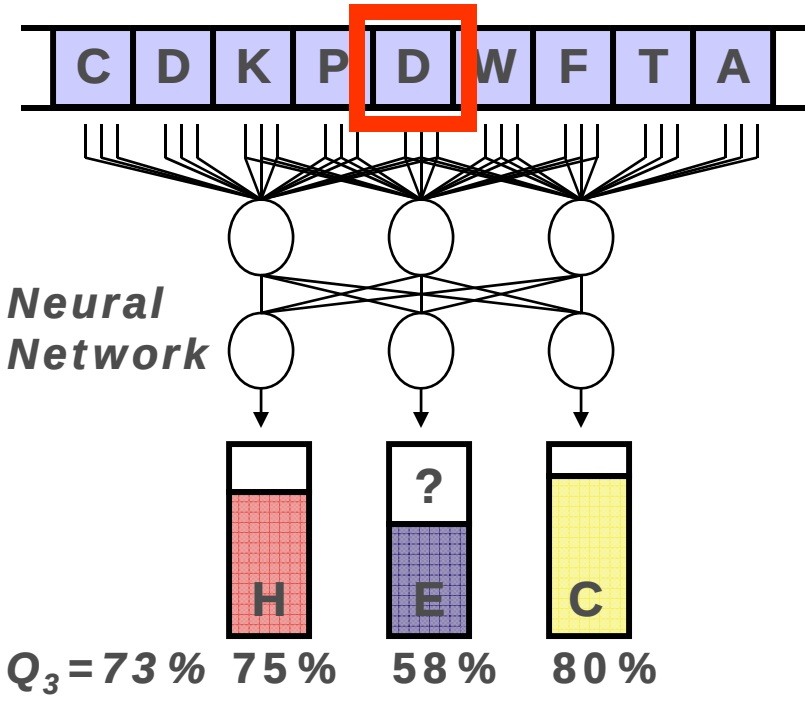


Fold Recognition

- Screens all structures of the PDB to identify possible template for modeling sequence of interest
- Template can be identified by sequence similarity (Homolog) or by structural similarity (Threading)
- Often predicted secondary structure used as input in addition to sequence
- Often multiple methods are applied to arrive at a consensus prediction with increased confidence
- e.g. BioInfo server uses up to 32 other technologies: www.bioinfo.pl

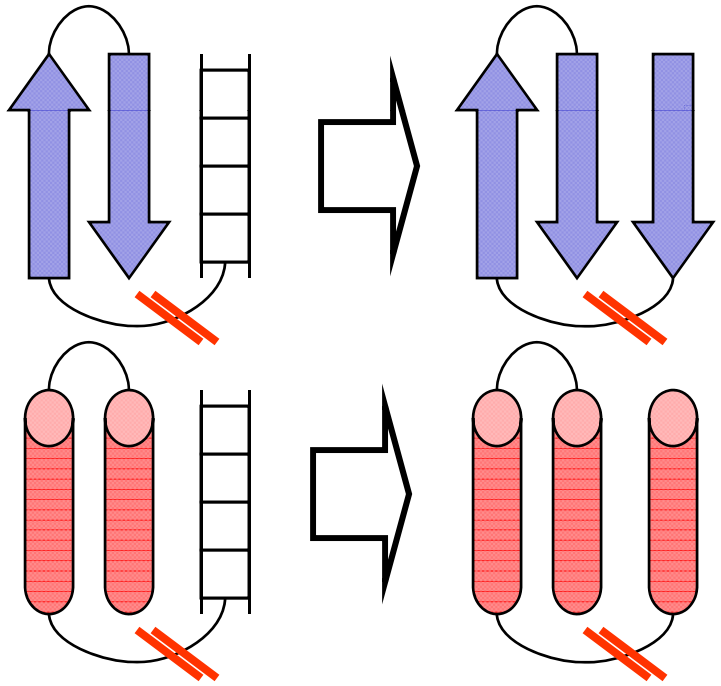
Secondary Structure Prediction

Local influences captured by protein primary structure:



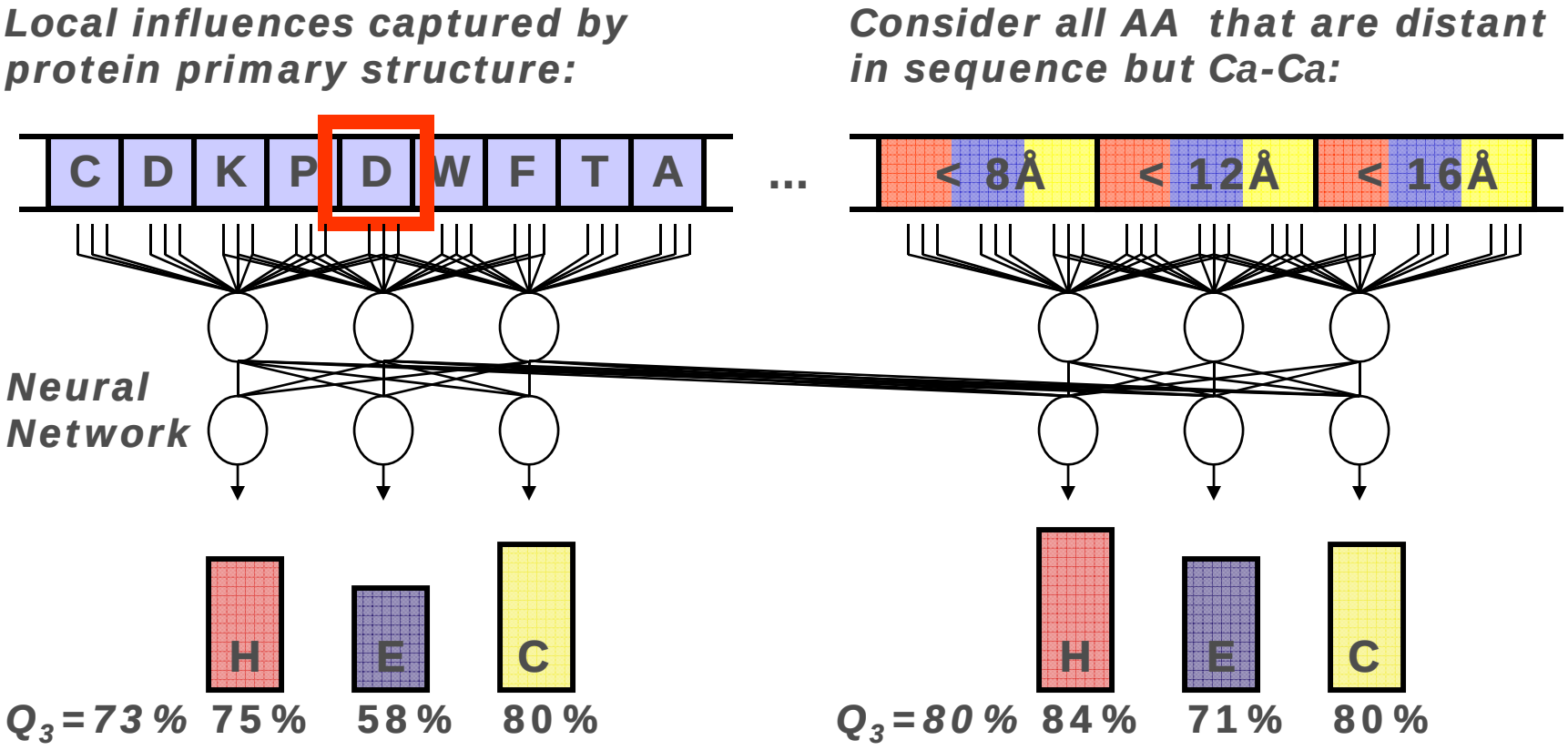
Rost, B. and Sander, C. (1993) *PNAS*, 90, 7558
Jones, D. T. (1999) *J. Mol. Biol.*, 292, 195-202.
Meiler, J., et. al. (2001) *J. Mol. Model.*, 7, 360-369.

Non-local influences captured by protein tertiary structure?!



Meiler, J. and Baker, D. (2003) *PNAS*, 100, 12105.

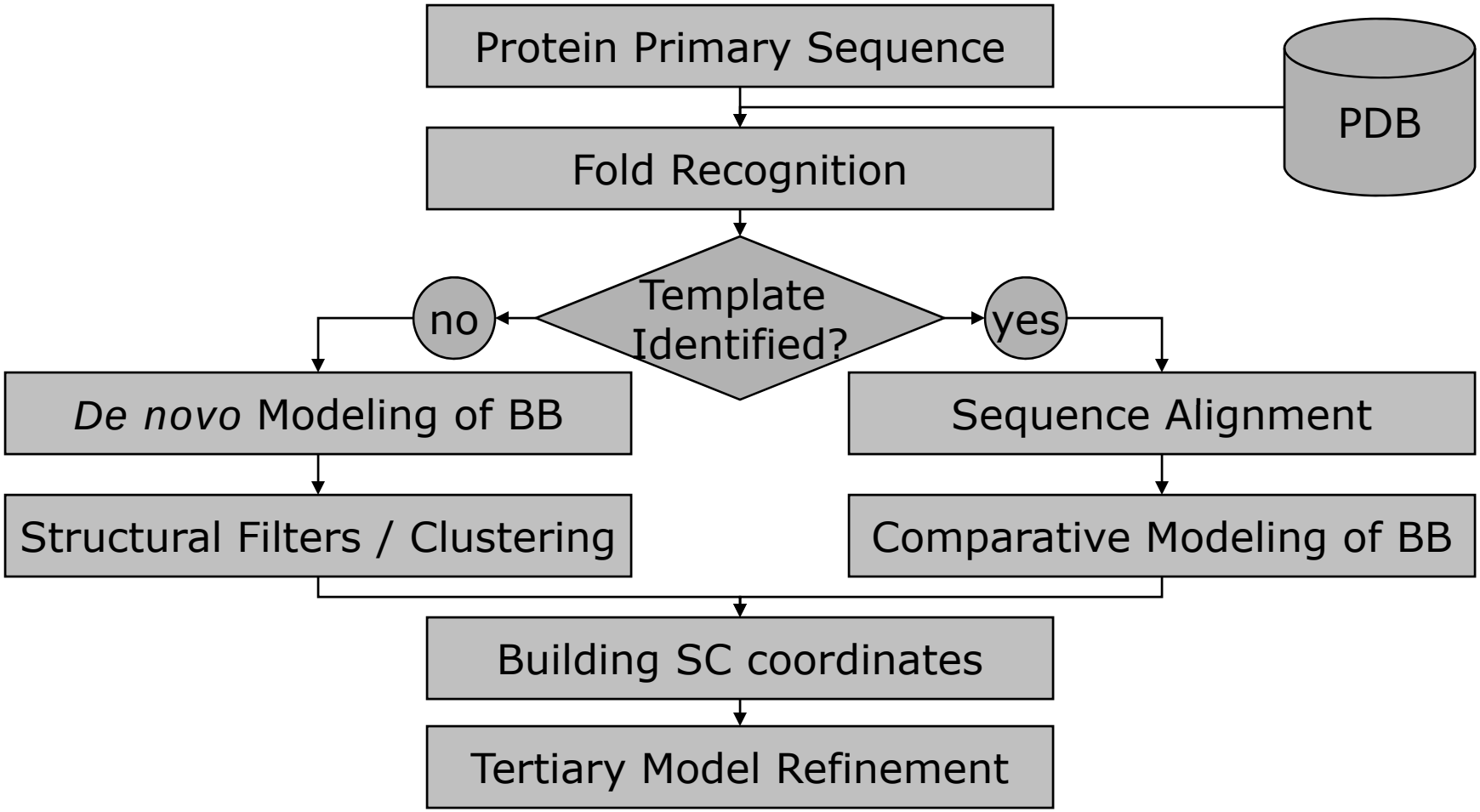
Coupled Prediction of Secondary and Tertiary Structure



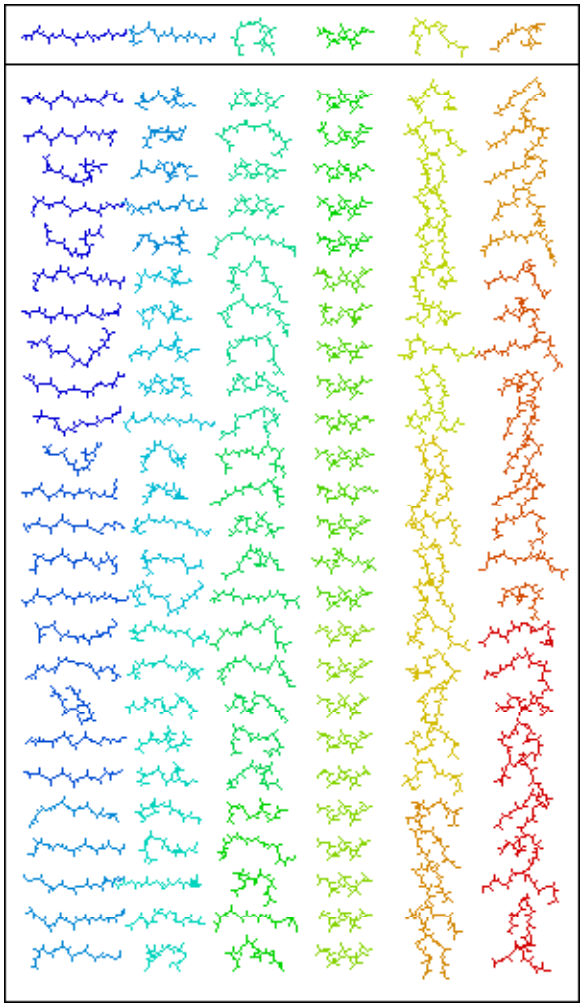
Meiler, J., et. al. (2001) *J. Mol. Model.*, 7, 360-369.

Meiler, J. and Baker, D. (2003) *PNAS*, 100, 12105.

General Scheme of Protein Structure Prediction



Local Sequence Bias – Rapid Approximation of Local Interactions

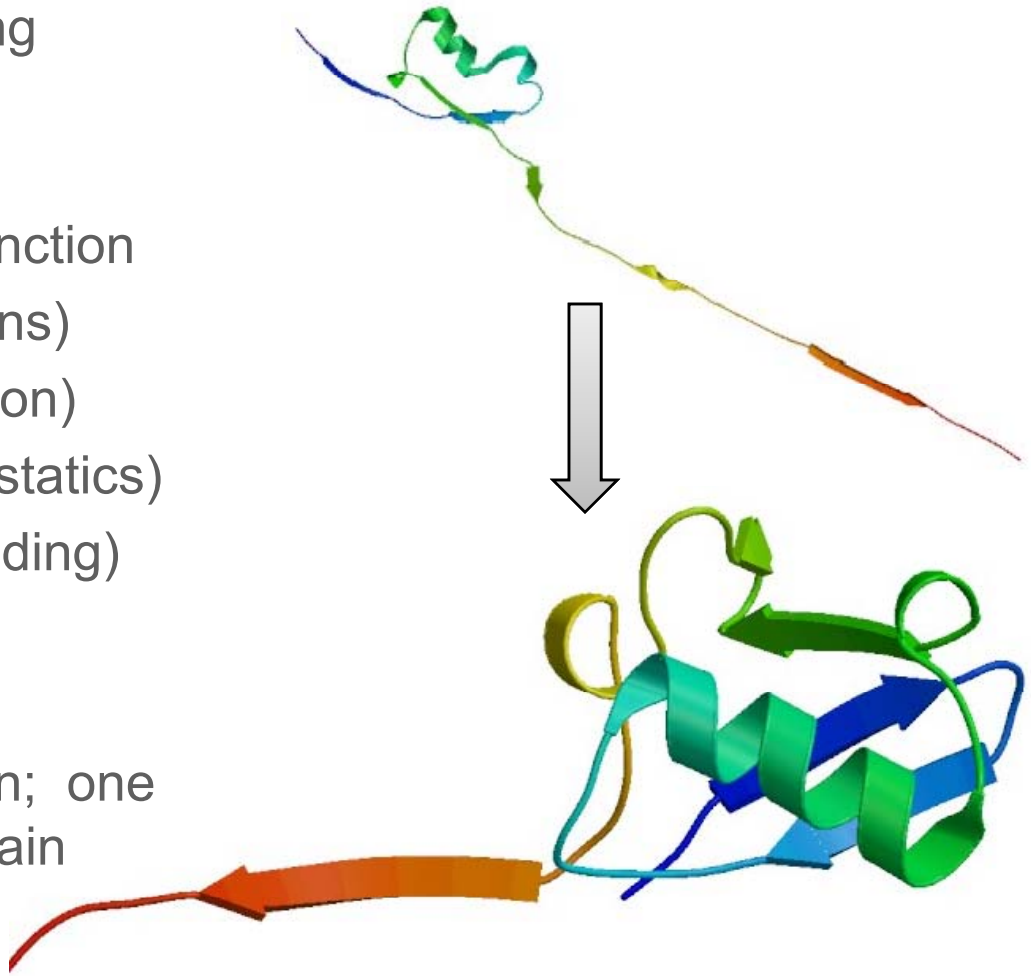


- **While not every protein fold is present in the protein databank, all possible conformations of small peptides are!**

- Approximate local interactions using the distribution of conformations seen for similar sequences in known protein structures
- For each sequence window, select fragments that represent the conformations sampled during folding

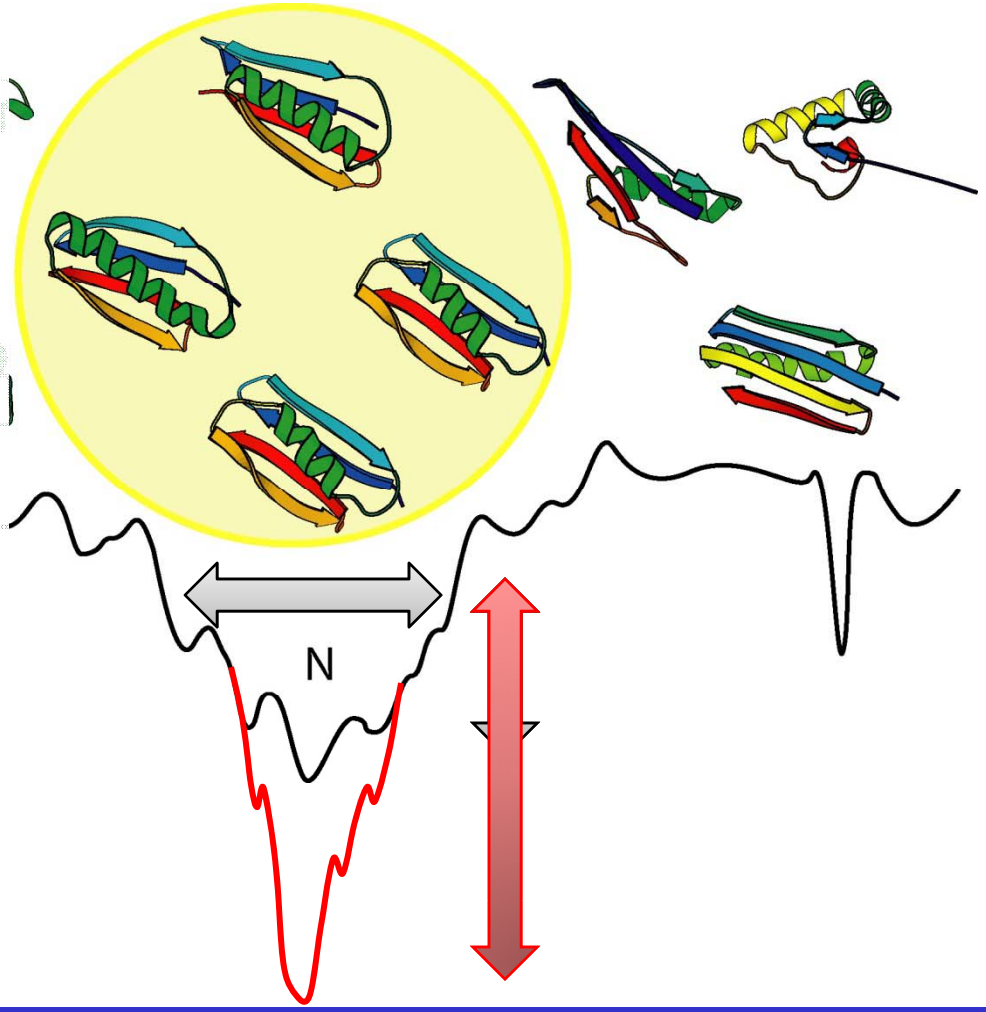
Non-local Interactions Govern Protein Folding Process

- Monte Carlo simulated annealing assembly of fragments
- Statistically-derived potential function
  - Steric overlap (vdw interactions)
  - Residue environment (solvation)
  - Pairwise interactions (electrostatics)
  - Strand pairing (hydrogen bonding)
  - Compactness (solvation)
- Simplified protein representation; one centroid per amino acid side chain



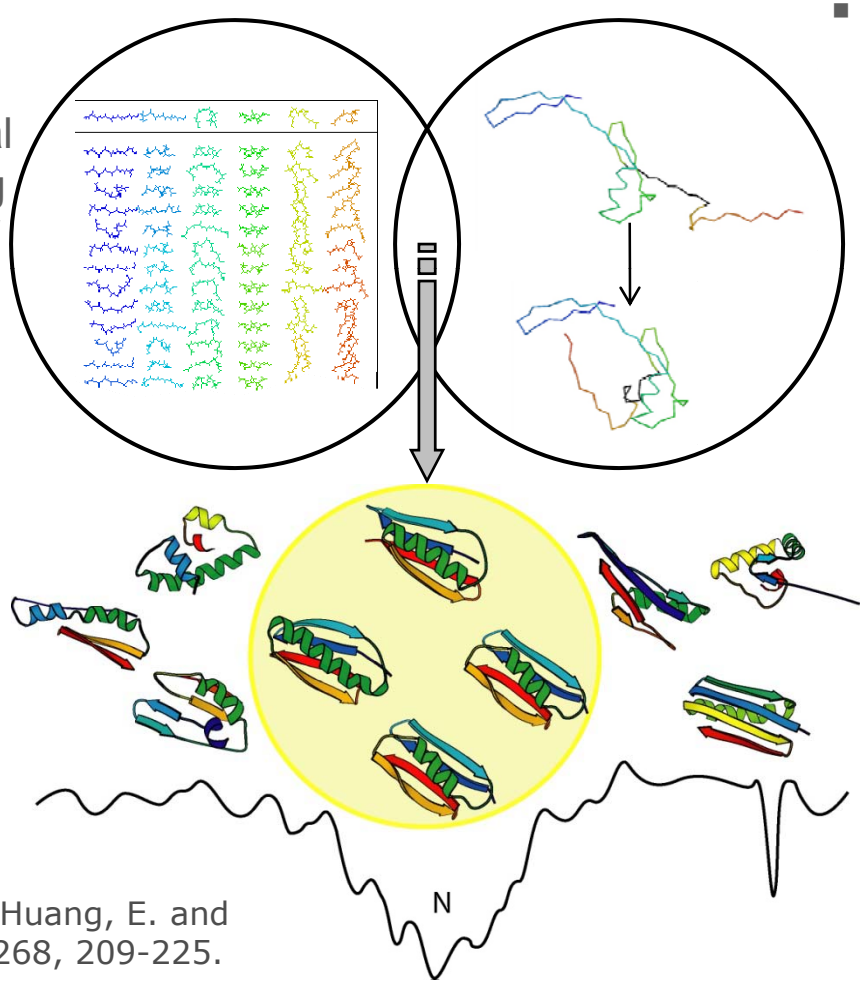
Native-like Protein Models Form Large Clusters

- The free energy minimum corresponds (usually) to the native protein fold
- Its depth is obscured because of the simplified energy approximation
- However, the width of the funnel leading to the free energy minimum of the native protein fold is well preserved



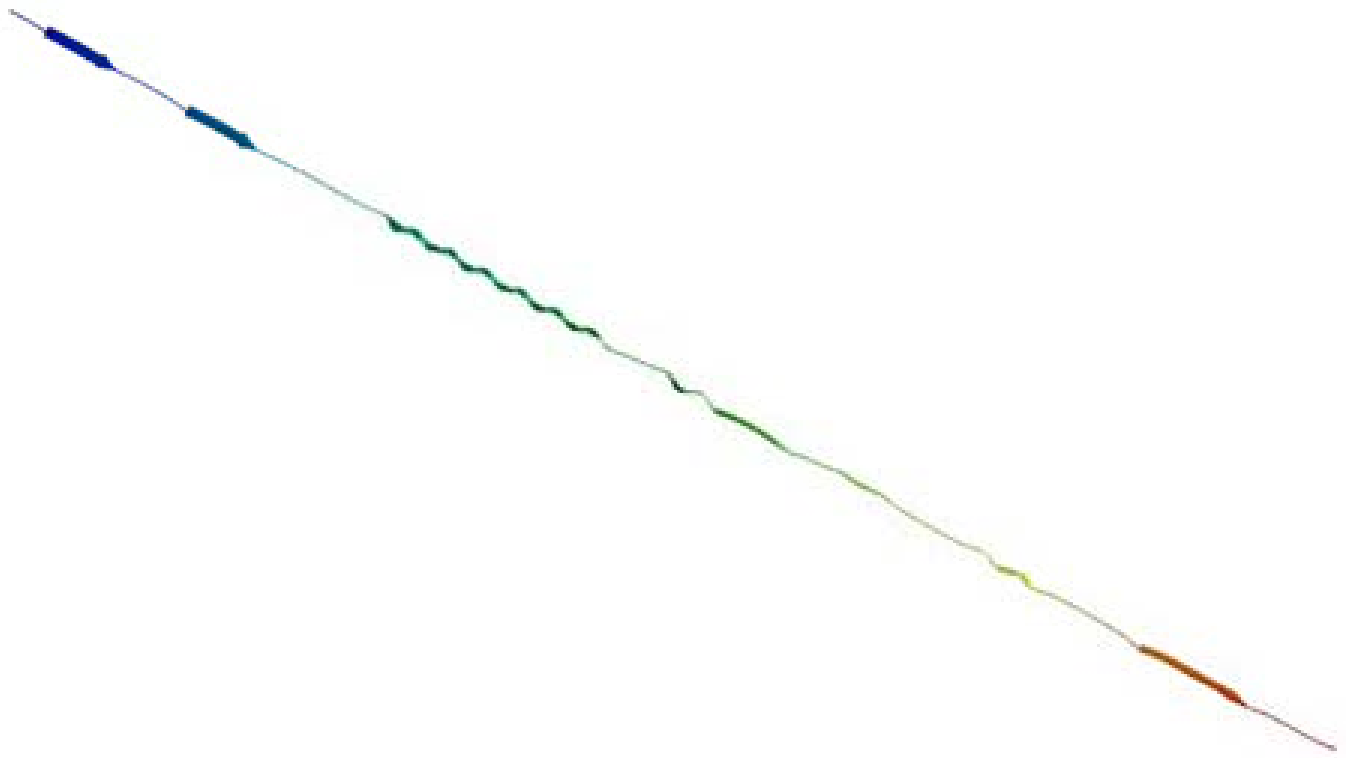
Sampling and Scoring for Protein Folding Simulation

- Local Sequence Bias
  - Approximate local interactions using the distribution of conformations seen for similar sequences in known protein structures
- Monte Carlo simulations
  - Select broadest minima using cluster analysis



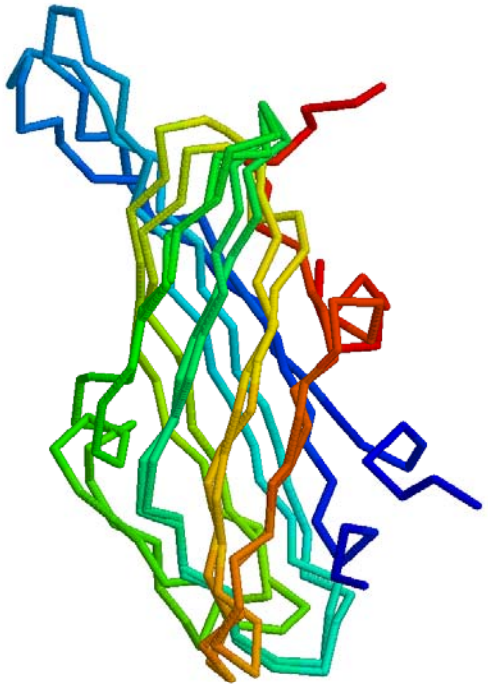
- Energy evaluation of non-local interactions using knowledge-based energy function
  - Steric overlap
  - Residue environment
  - Pair wise interactions
  - Strand pairing
  - Compactness
  - Secondary Structure Packing

Simons, K. T., Kooperberg, C., Huang, E. and Baker, D. (1997) *J. Mol. Biol.*, 268, 209-225.

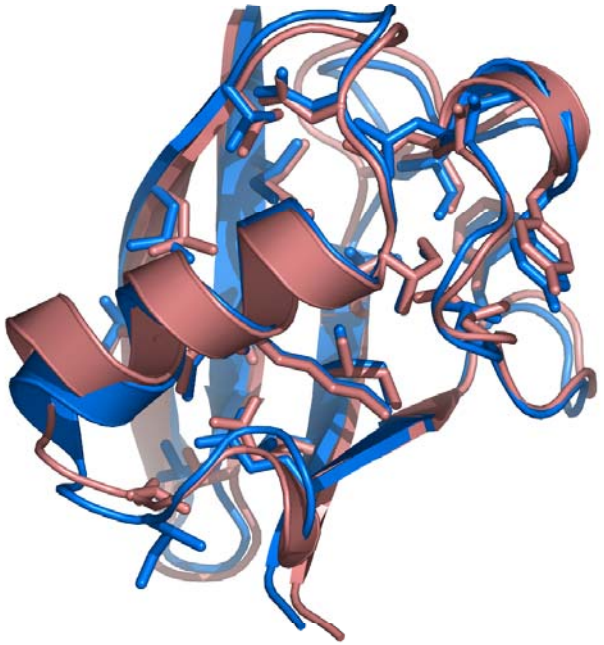


Major Challenges in the Field of Protein Structure Prediction

- Determination of fold for new fold targets of large size, complex topology, or membrane proteins



- Refinement of *de novo* and comparative models to experimental resolution

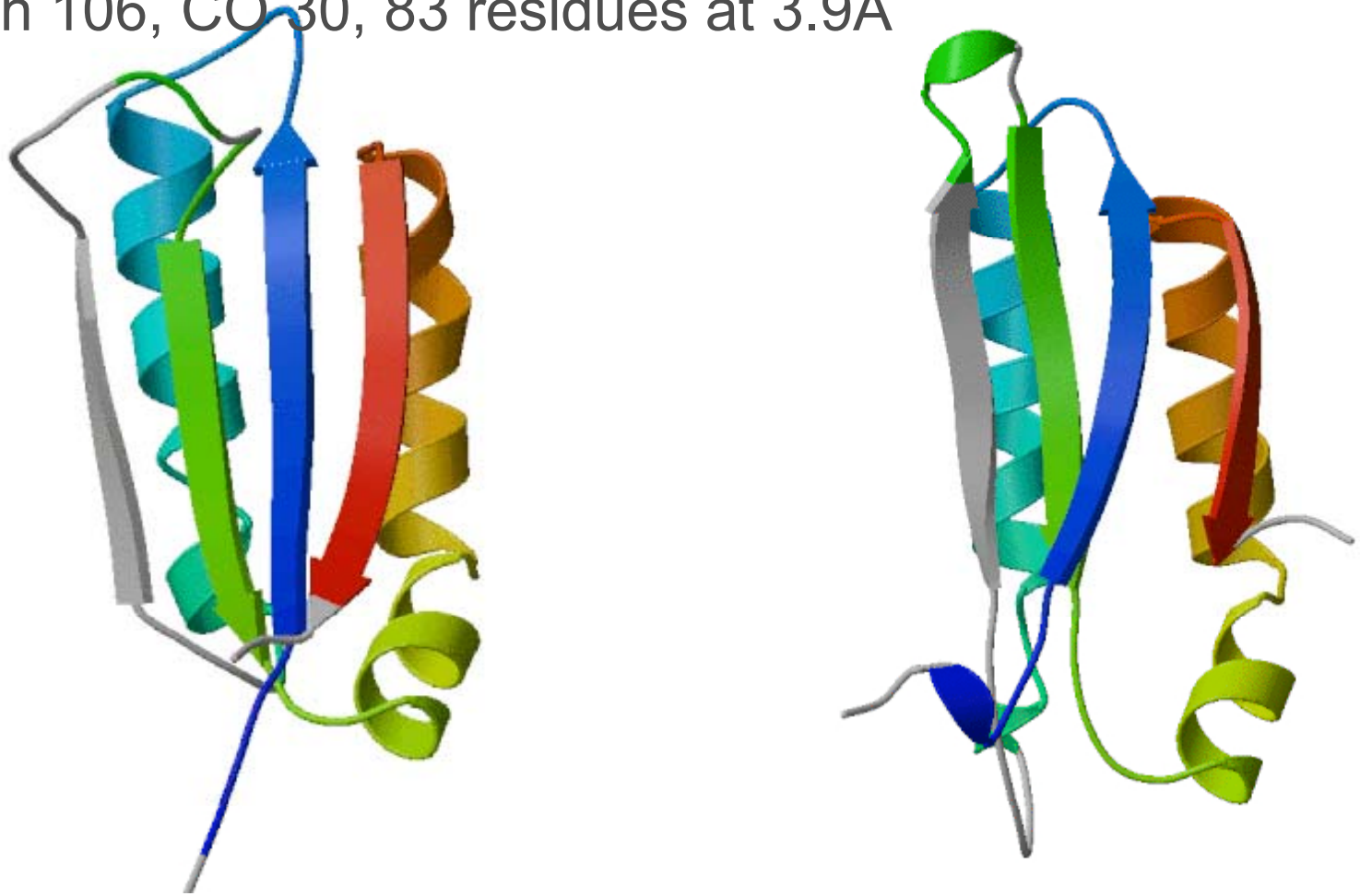


Critical **A**ssessment of Techniques for Protein **S**tructure **P**rediction

- CASP was established in 1994.
- Biyearly experiment to assess progress in the field of Protein Structure Prediction.
- Blind prediction of protein structures from their sequence.
- Sequences of “soon to be solved” proteins are submitted by NMR and X-ray crystallography groups during the summer.
- 200 prediction teams from 24 countries in CASP6 (2004) meet in December in Asilomar to discuss the results.
- 90 targets in CASP6 (2004) in three categories: easy/hard fold recognition and new fold
- Current Review: Moult, J., *Curr Opin Struct Biol*, **2005**.

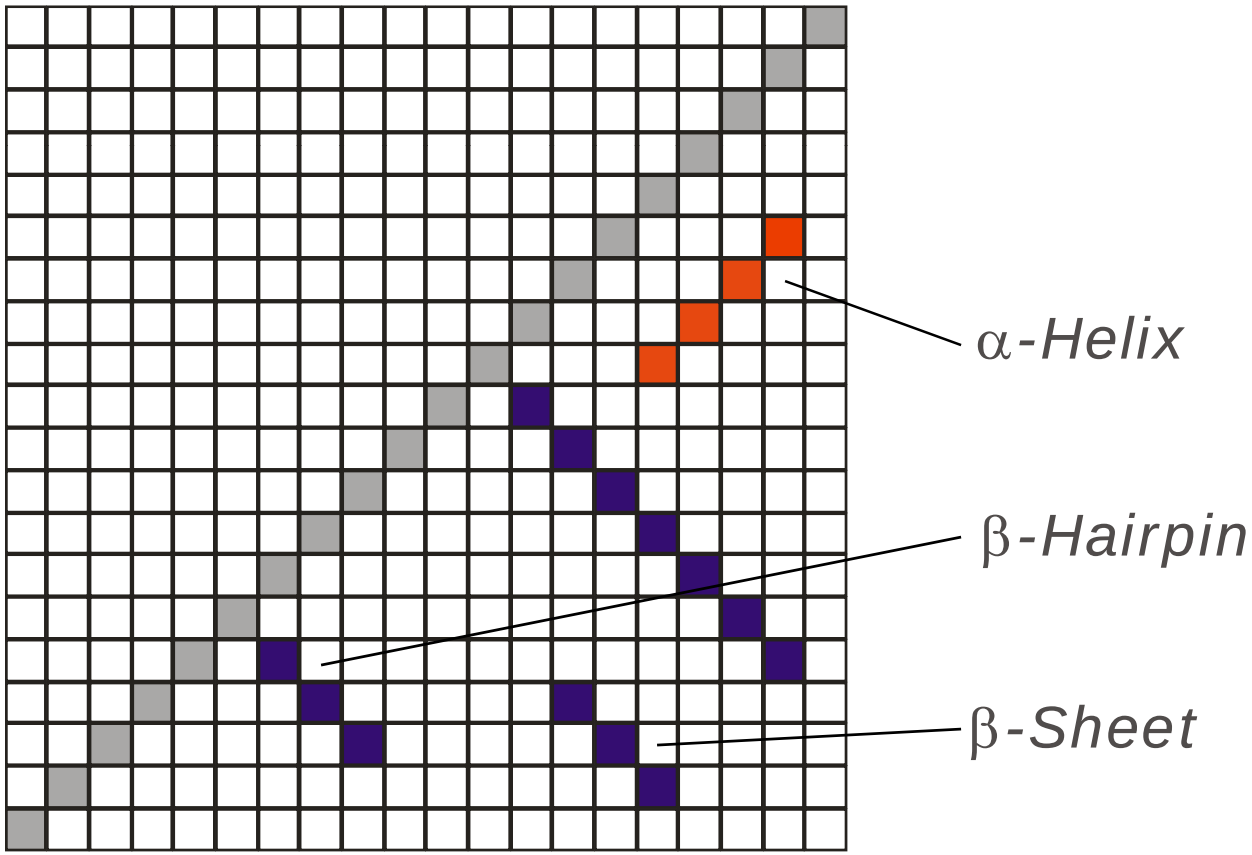
T135 Hypothetical Protein

- Length 106, CO 30, 83 residues at 3.9Å



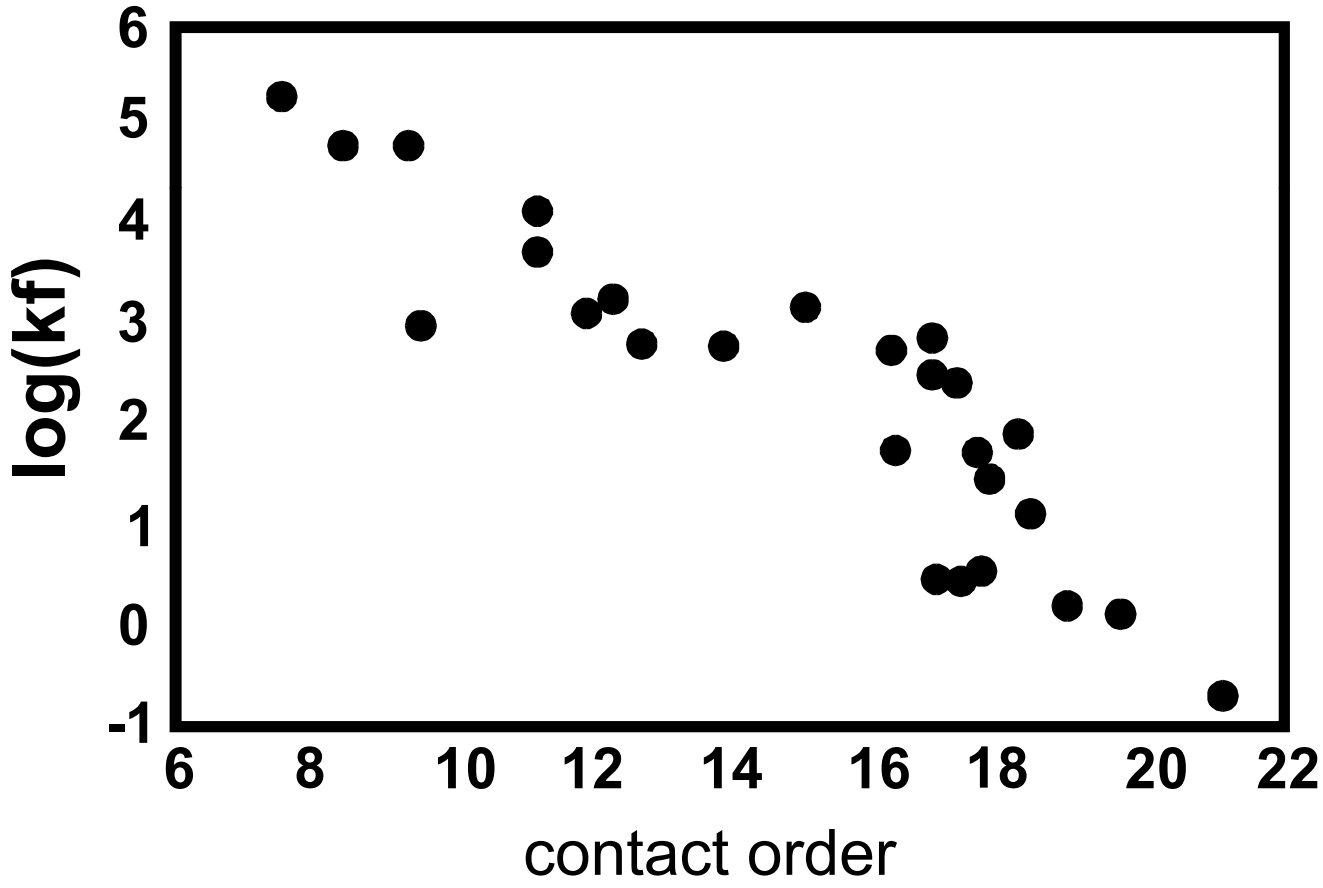
Contact Order

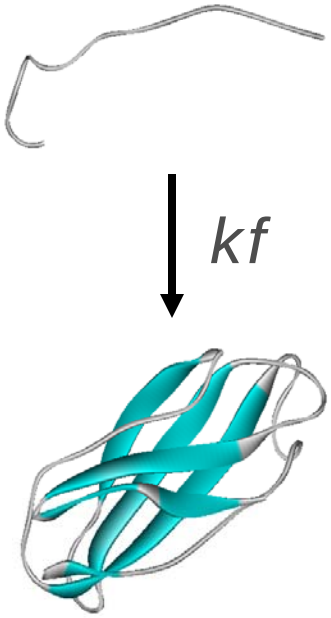
- *average sequence separation of two amino acids that are in contact*



Correlation of Folding Rate and Contact Order

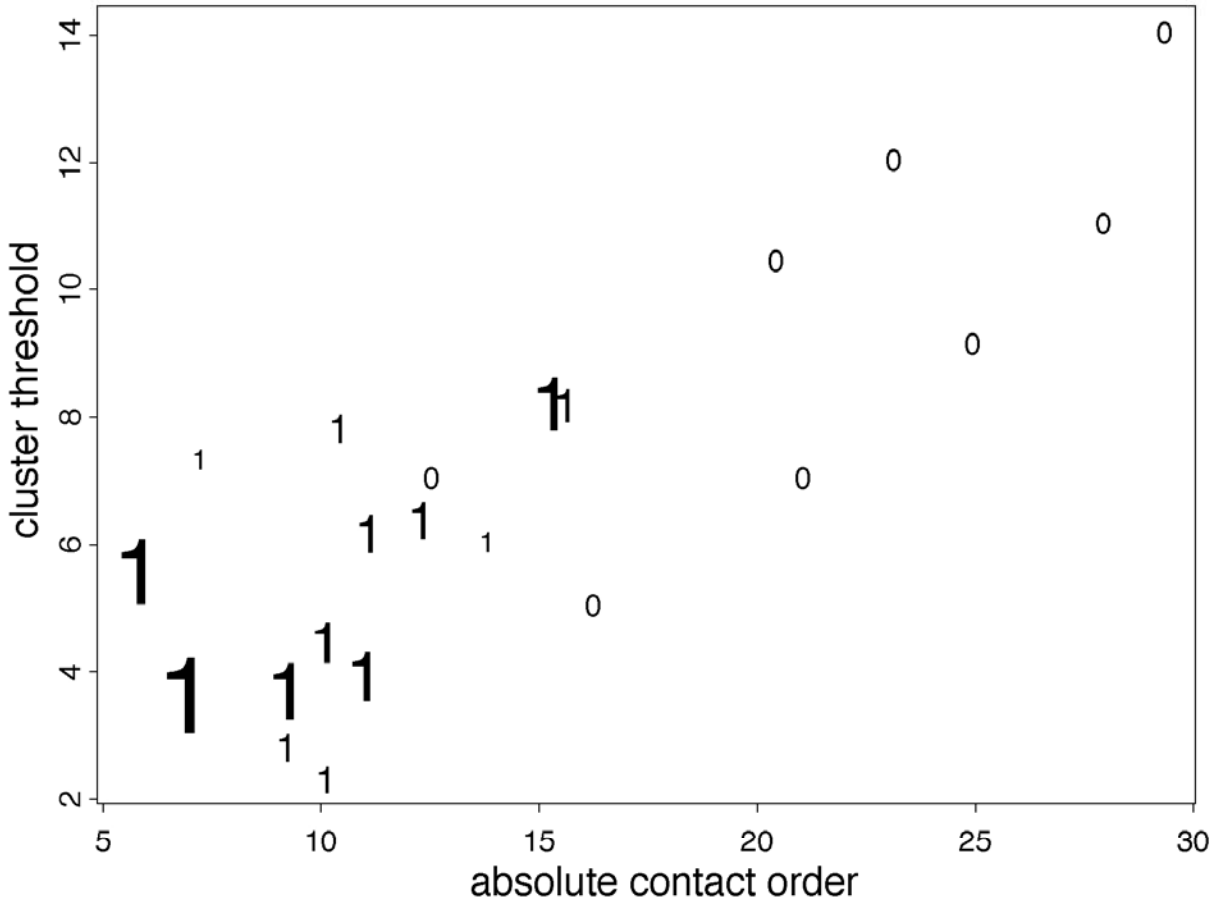
Bonneau, R., Ruczinski, I., Tsai, J. and Baker, D. (2002) *Protein Sci*, 11, 1937-1944.



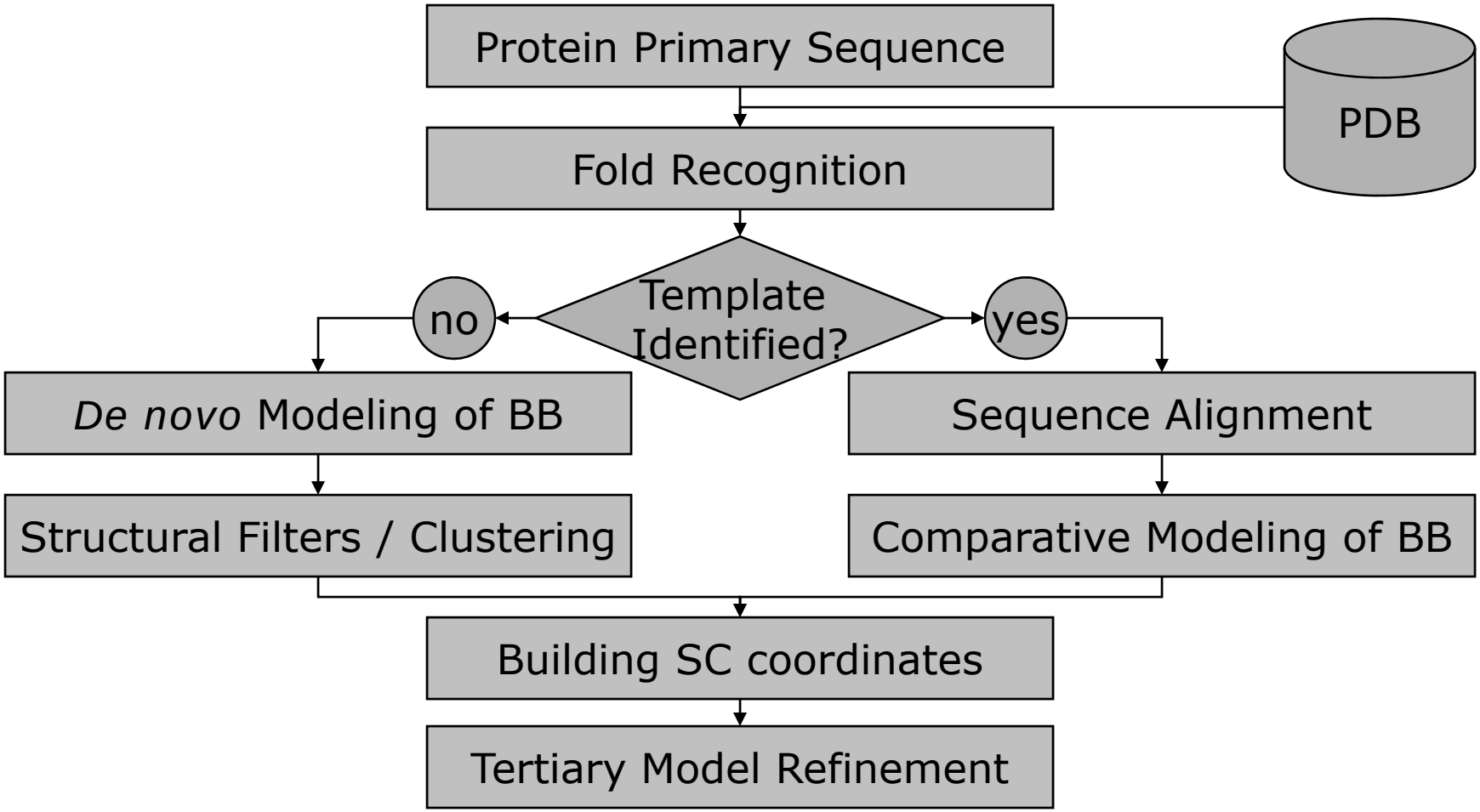


CASP4 results

Bonneau, R., Ruczinski, I., Tsai, J. and Baker, D. (2002) *Protein Sci*, 11, 1937-1944.



General Scheme of Protein Structure Prediction



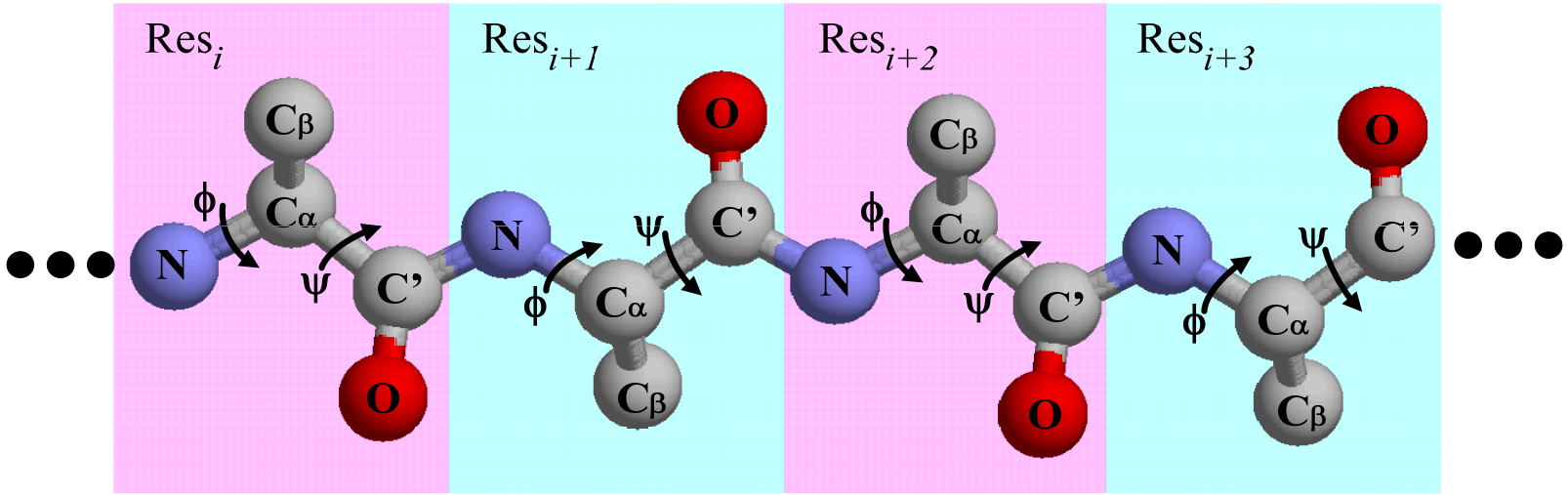
Comparative Modeling – Build Model based on an Existing Structure

- *Homology Modeling*
  Sequence-Sequence Alignment
- *Protein Threading*
  Sequence-Structure Alignment

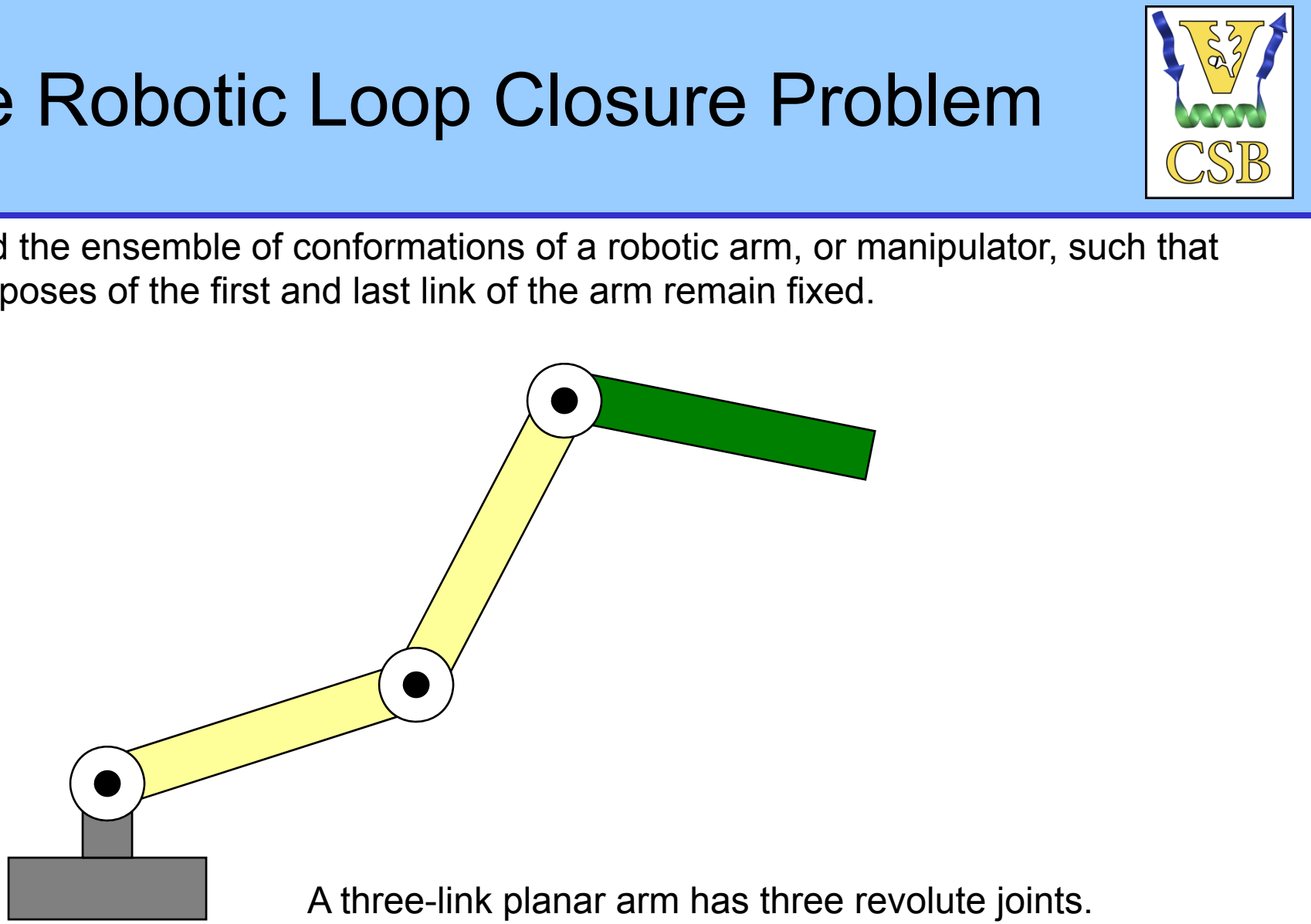
- *Copy Template Coordinates*
- *Rebuild missing Pieces – Loops*
- *Structure Refinement*

Loop Closure Problem

- Input
  - 2 Anchor residues
  - Length of missing fragment
- Output
  - A small number of candidate structures for missing fragment



The Robotic Loop Closure Problem

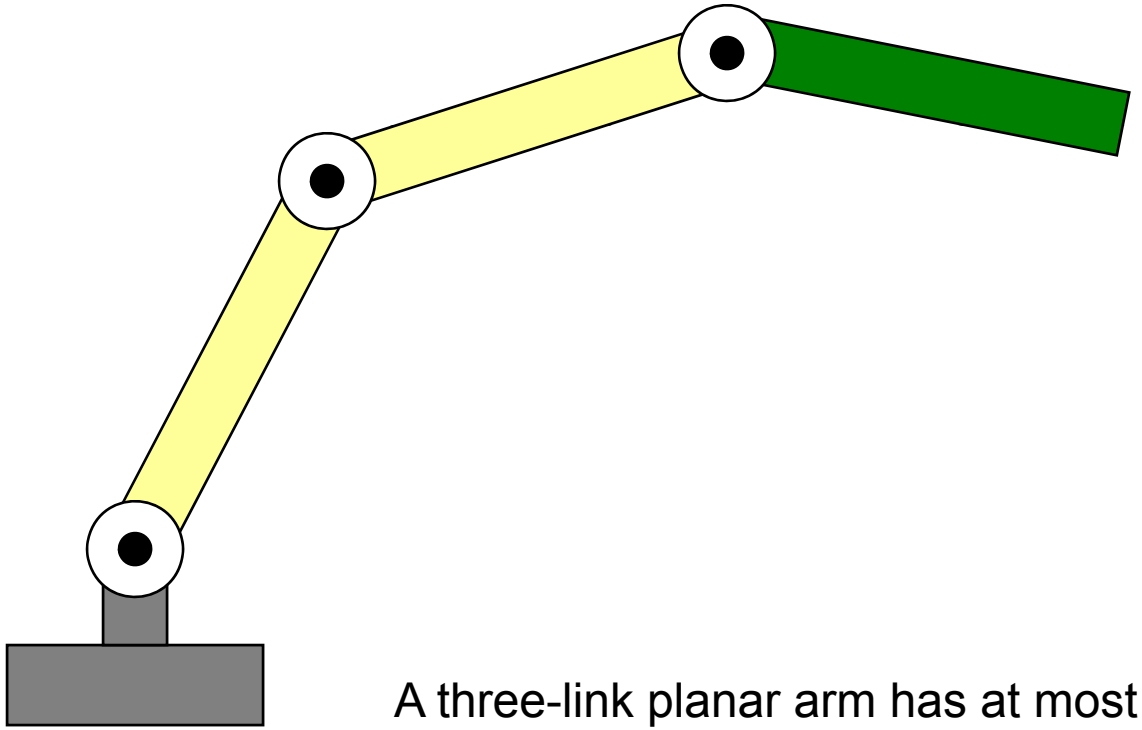


Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.

A three-link planar arm has three revolute joints.

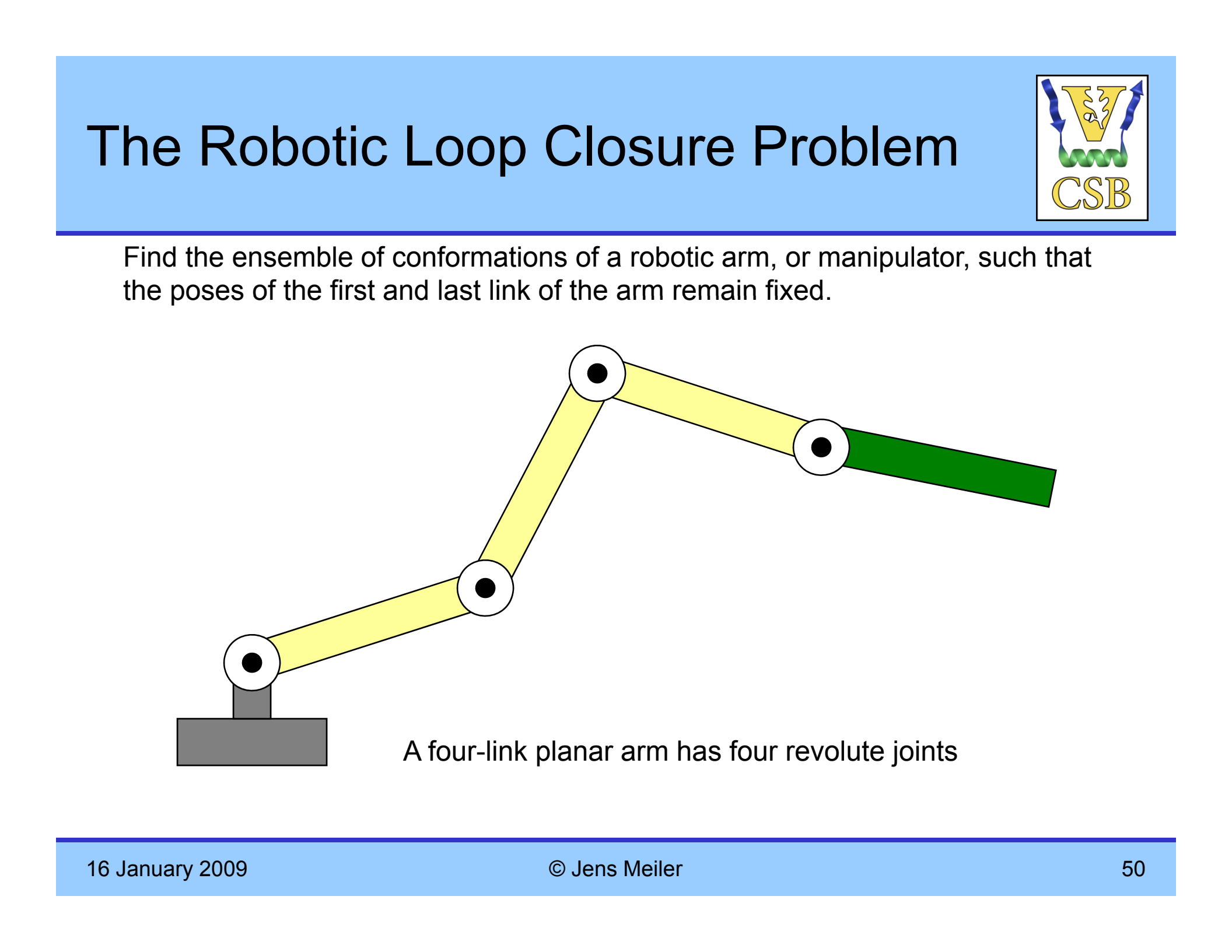
The Robotic Loop Closure Problem

Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



A three-link planar arm has at most two closed loop conformations.

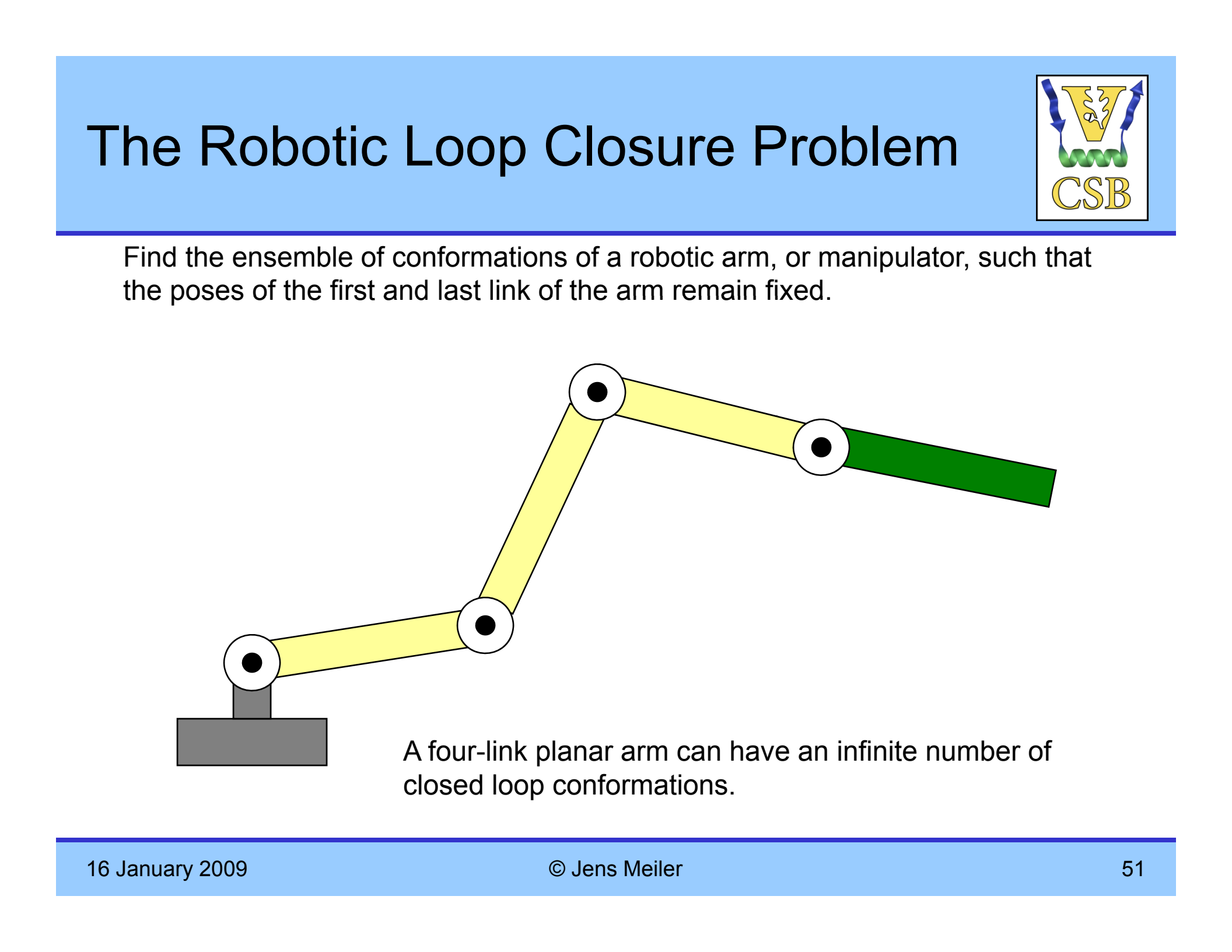
The Robotic Loop Closure Problem



Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.

A four-link planar arm has four revolute joints

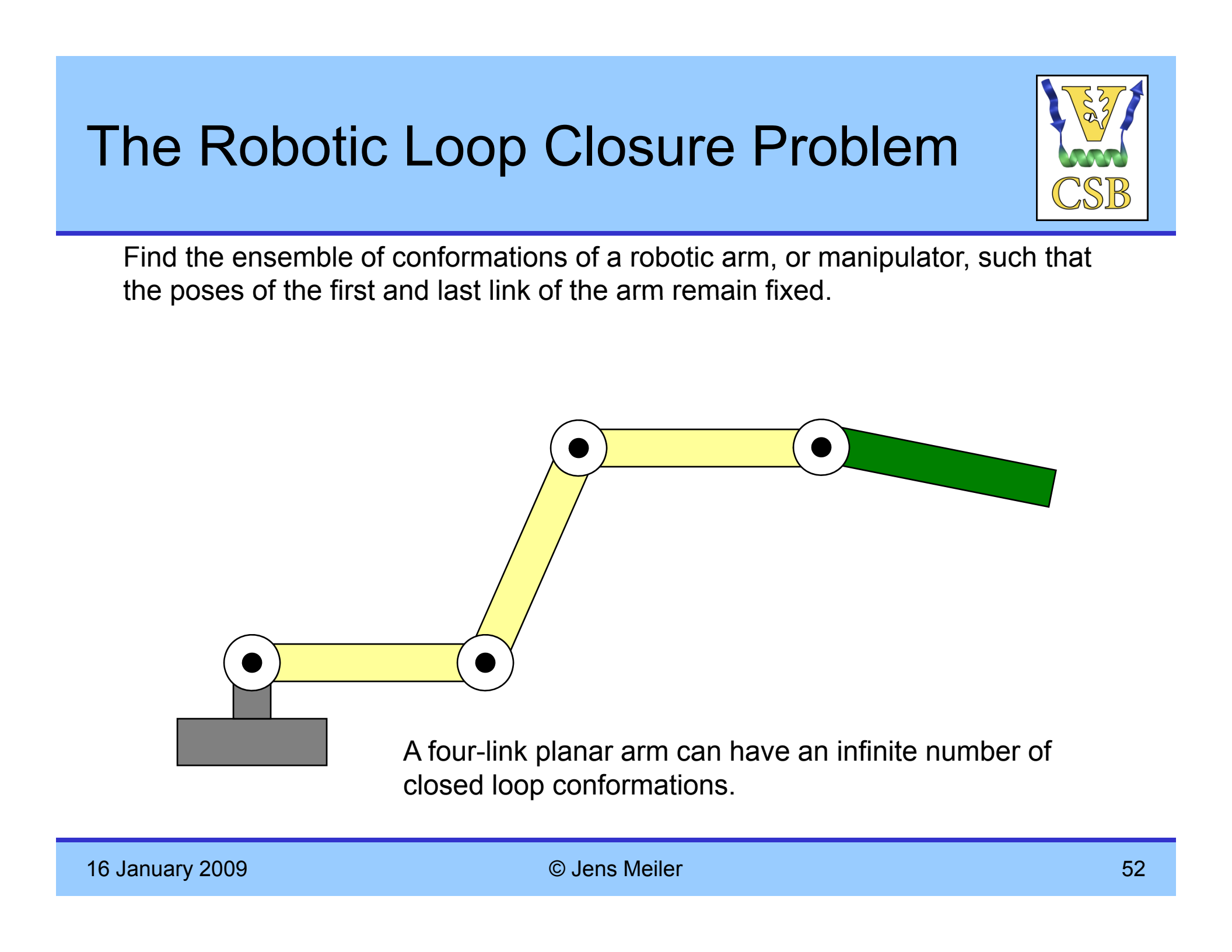
The Robotic Loop Closure Problem



Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.

A four-link planar arm can have an infinite number of closed loop conformations.

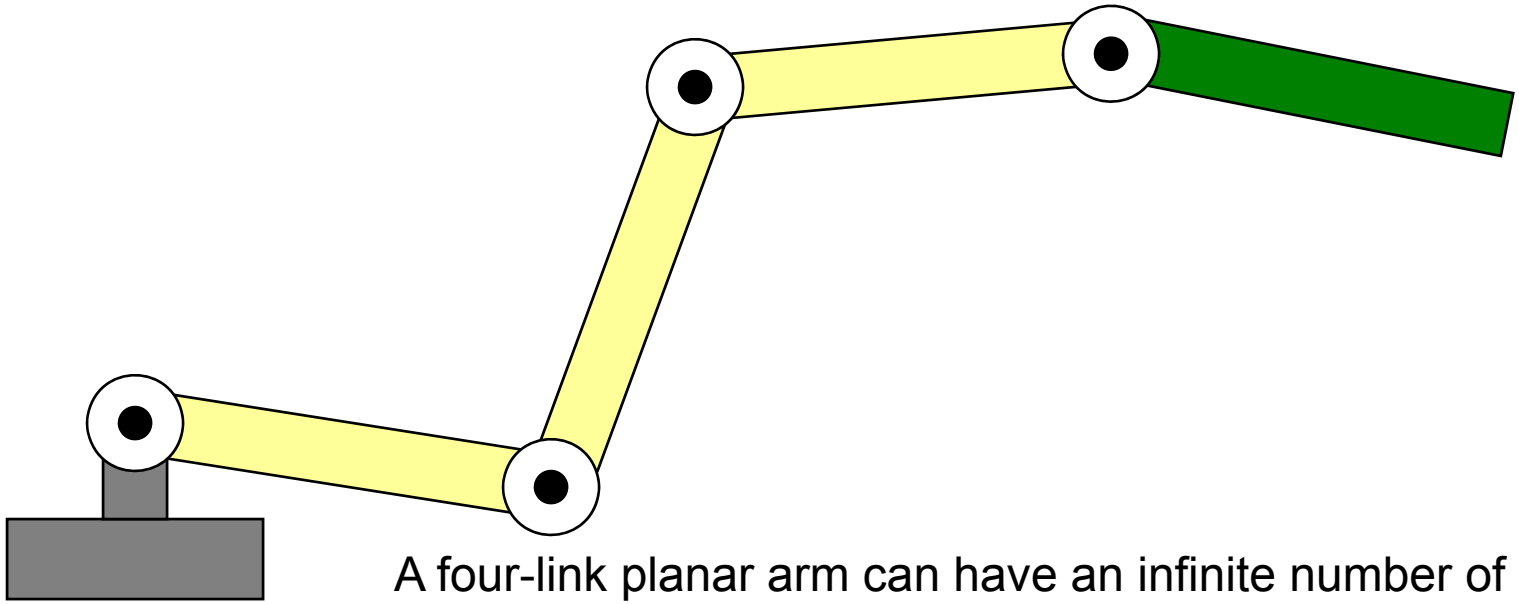
The Robotic Loop Closure Problem



Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.

A four-link planar arm can have an infinite number of closed loop conformations.

The Robotic Loop Closure Problem

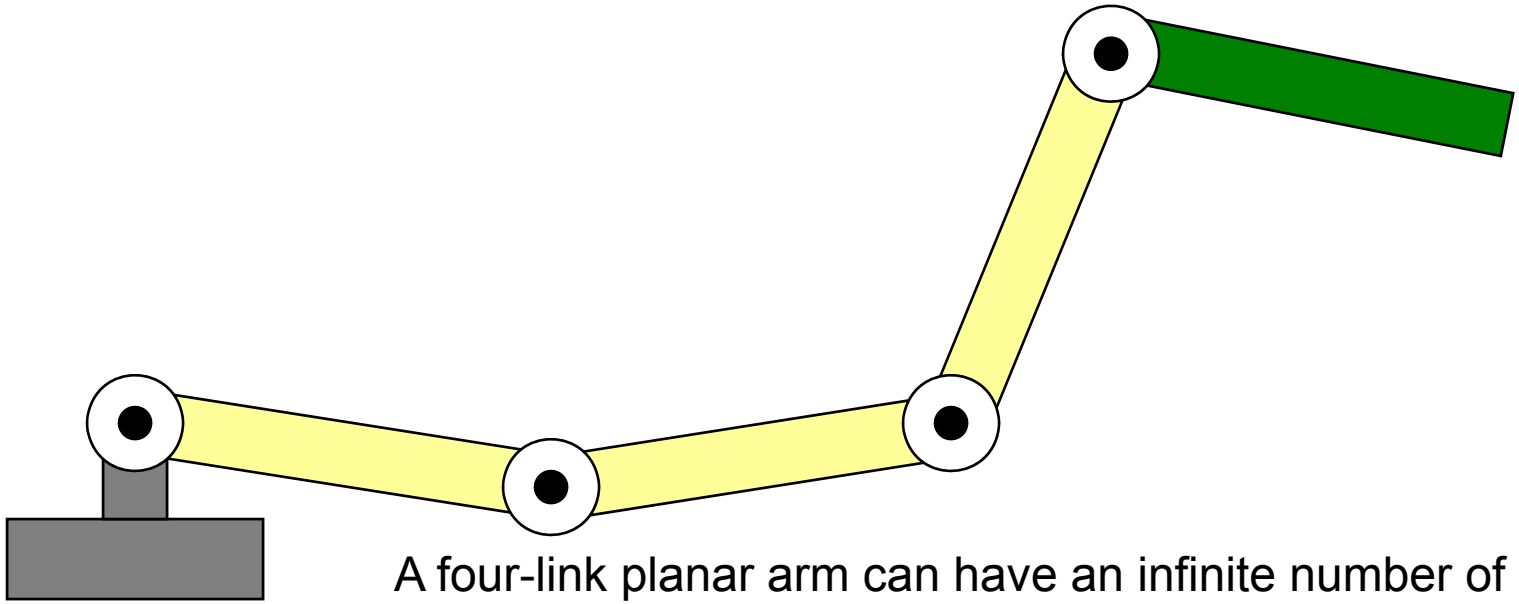


Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.

A four-link planar arm can have an infinite number of closed loop conformations.

The Robotic Loop Closure Problem

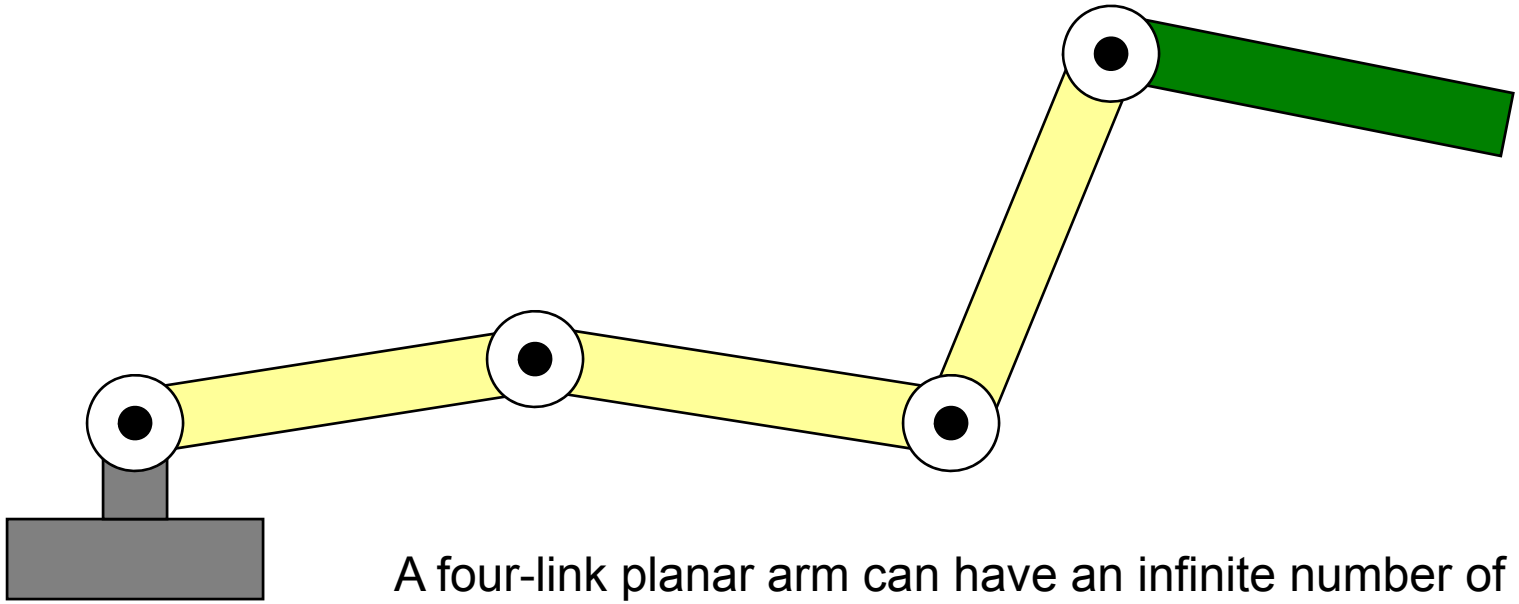
Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.



A four-link planar arm can have an infinite number of closed loop conformations.

The Robotic Loop Closure Problem

Find the ensemble of conformations of a robotic arm, or manipulator, such that the poses of the first and last link of the arm remain fixed.

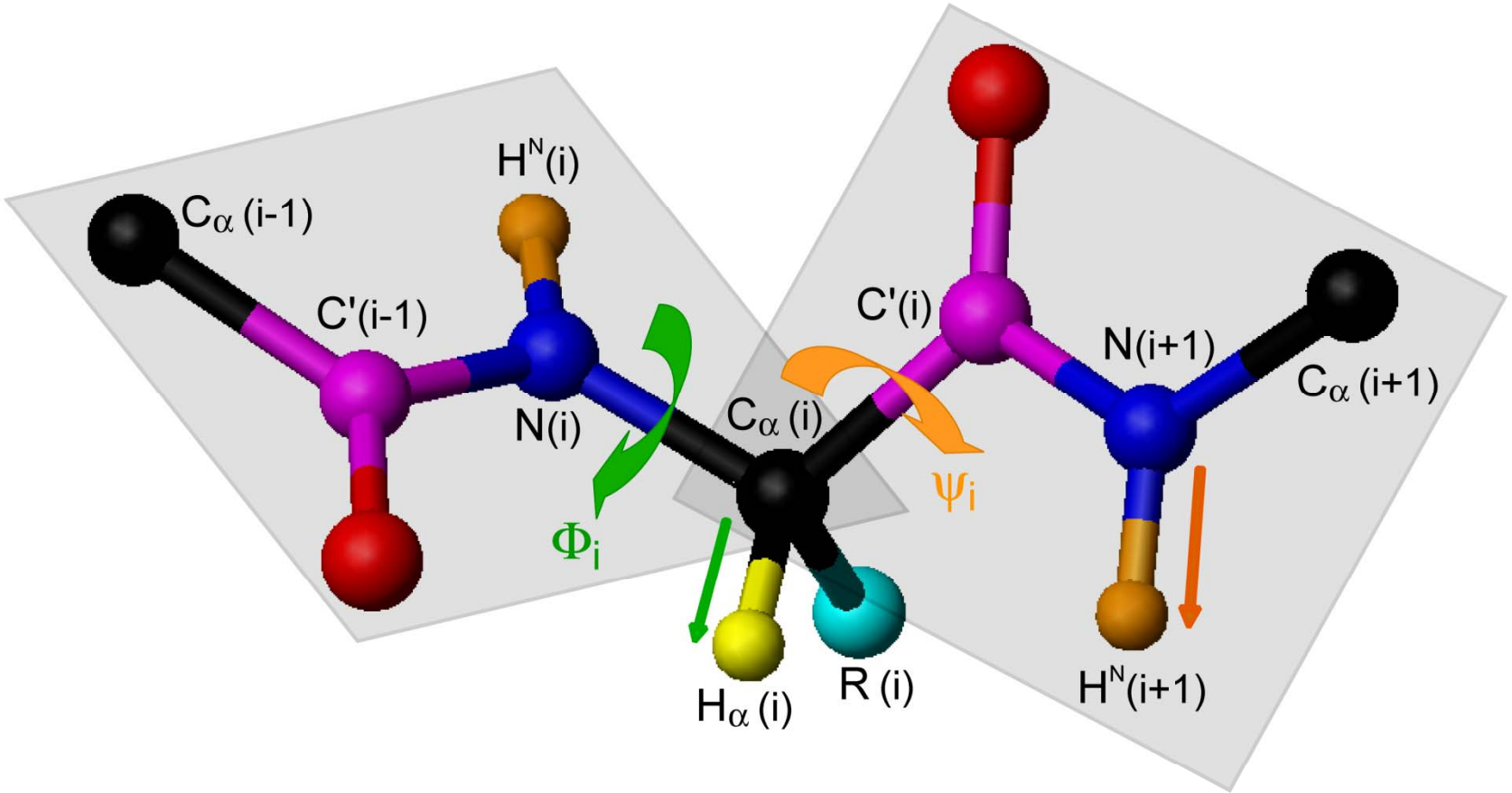


A four-link planar arm can have an infinite number of closed loop conformations.

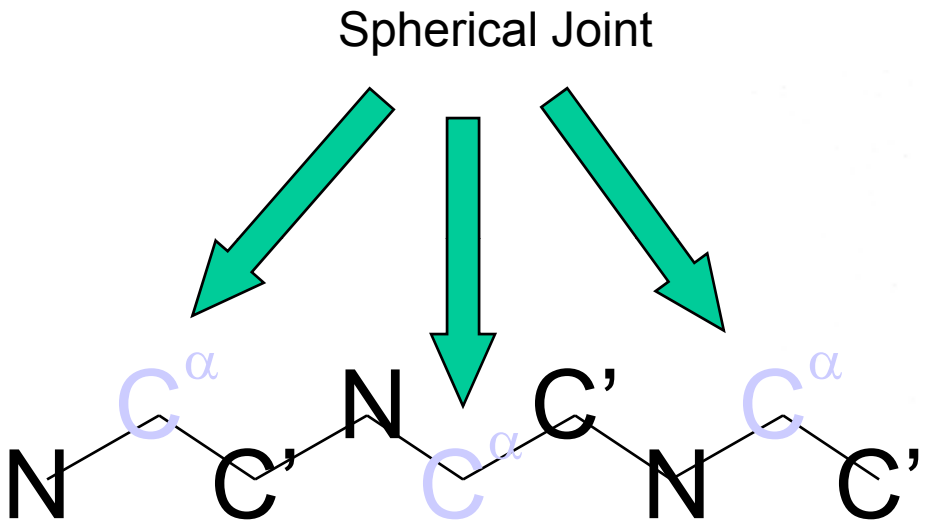
The Molecular Loop Closure Problem and the Degrees of Freedom (DOF)

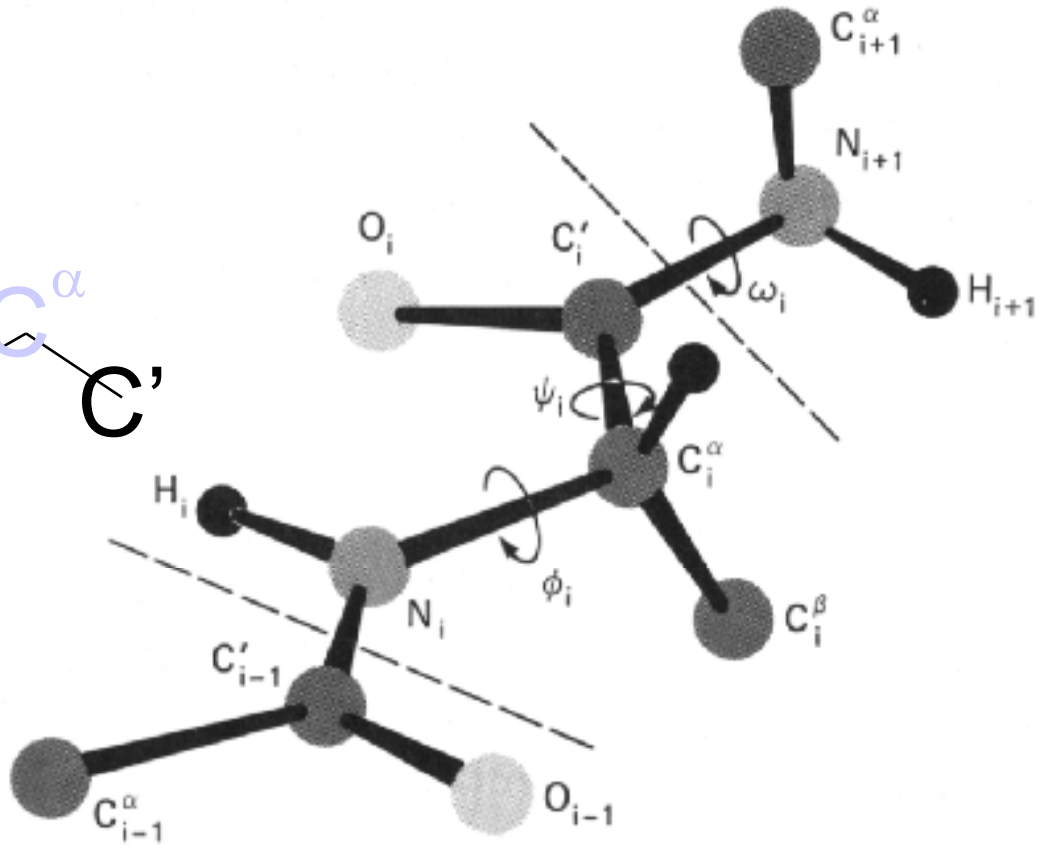
- The molecular loop closure problem is overconstrained for fewer than six DOF, and
- underconstrained for more than six DOF.
- A molecular loop closure problem with more than six DOF has an infinite number of solutions.
- A molecular loop closure problem with six DOF has at most 16 solutions.

DOFs in the Protein Backbone

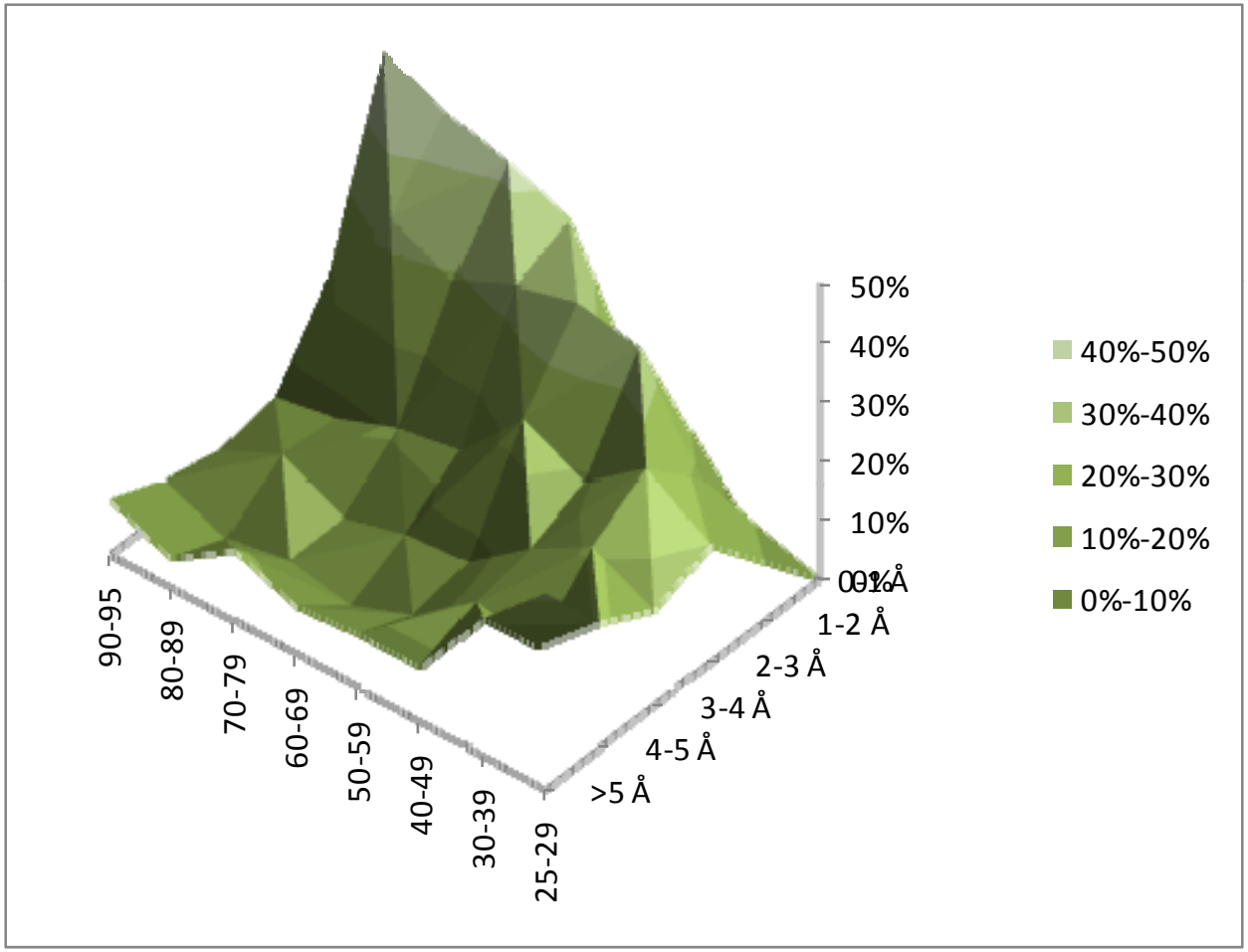


The Tripeptide Problem

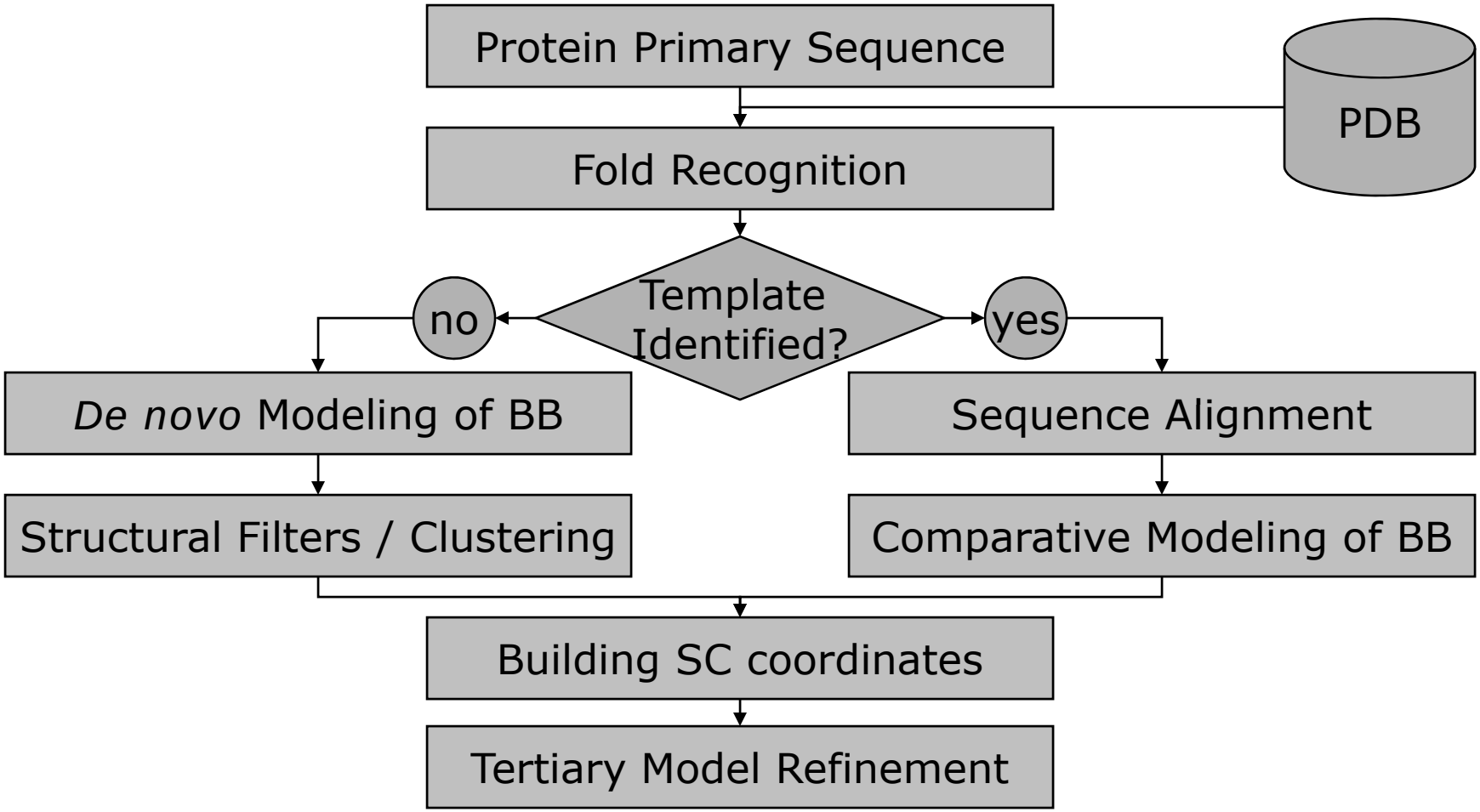




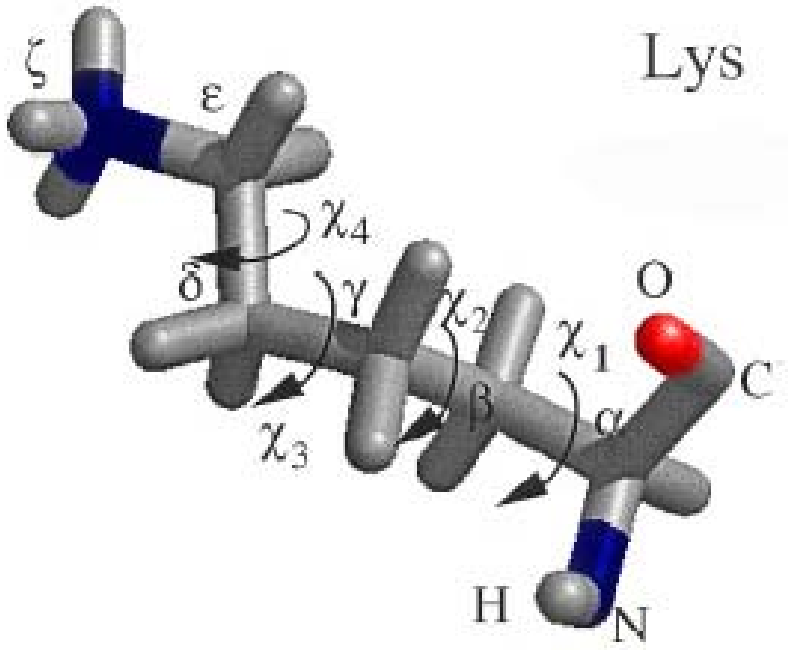
Accuracy of homology models

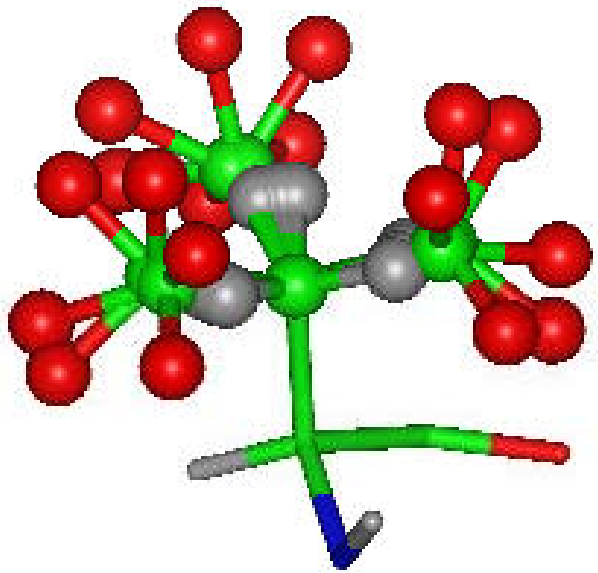


General Scheme of Protein Structure Prediction



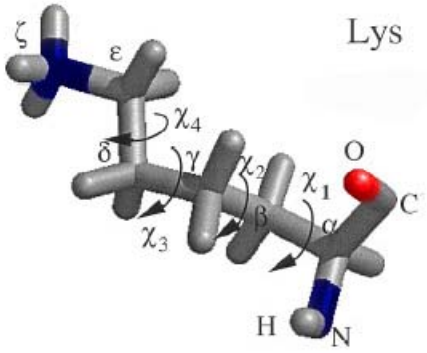
Sidechain Degrees of Freedom

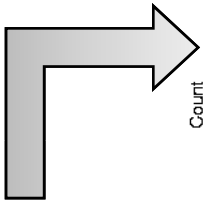


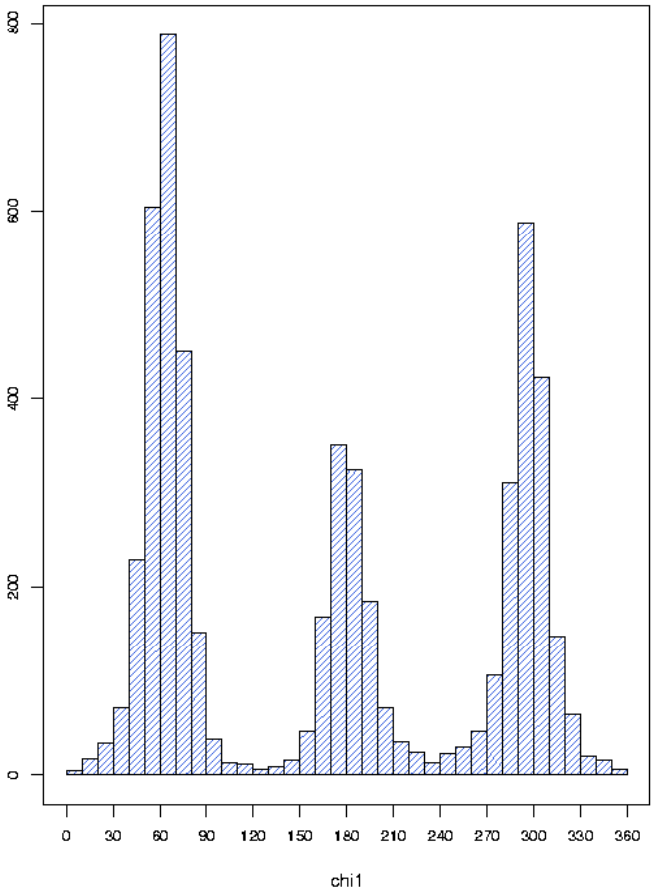


All Likely Side Chain Conformations are Present in the Protein Databank

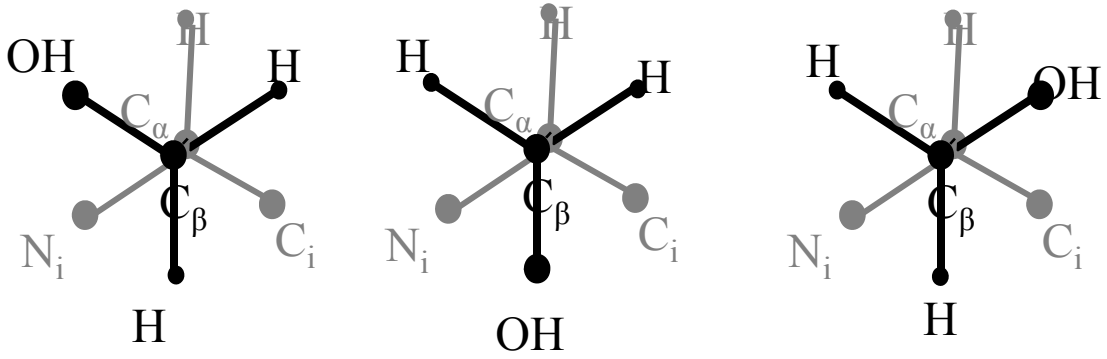
- Lysine has four side chain χ-angles







- Serine has one side chain χ-angle



“Rotamer” Libraries Encompass all Likely Side Chain Conformations

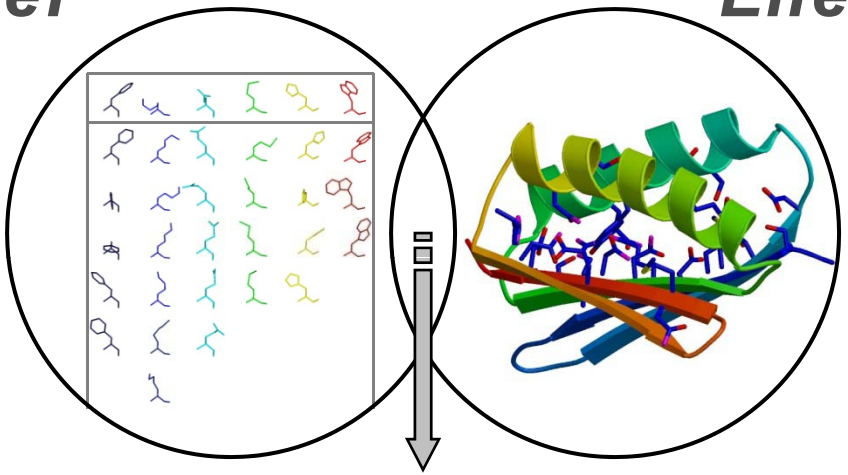
	No. χ_1	No.	p	σ	p|χ_1	σ	χ_1	σ	χ_2	σ
SER 1 0 0 0	4125	4125	46.61	0.43	100.00	0.00	65.0	10.7		
SER 2 0 0 0	2059	2059	23.27	0.37	100.00	0.00	179.6	11.7		
SER 3 0 0 0	2665	2665	30.12	0.40	100.00	0.00	-64.2	11.0		
THR 1 0 0 0	4165	4165	48.38	0.44	100.00	0.00	61.1	8.8		
THR 2 0 0 0	686	686	7.98	0.24	100.00	0.00	-173.3	12.8		
THR 3 0 0 0	3757	3757	43.64	0.44	100.00	0.00	-60.4	8.2		
TRP 1 1 0 0	337	215	9.56	0.51	63.62	2.13	61.7	9.7	-90.9	9.4
TRP 1 2 0 0	337	16	0.74	0.15	4.92	0.96	65.6	7.5	-16.7	40.9
TRP 1 3 0 0	337	106	4.73	0.36	31.47	2.06	59.4	12.0	88.2	10.1
TRP 2 1 0 0	786	359	15.94	0.63	45.64	1.45	-178.4	12.5	-104.1	15.1
TRP 2 2 0 0	786	139	6.19	0.41	17.72	1.11	-175.5	12.4	18.2	31.0
TRP 2 3 0 0	786	288	12.80	0.57	36.63	1.40	179.8	8.8	84.8	9.7
TRP 3 1 0 0	1127	106	4.73	0.36	9.45	0.71	-70.4	13.2	-91.4	15.4
TRP 3 2 0 0	1127	303	13.46	0.59	26.90	1.08	-68.5	9.9	-2.5	26.8
TRP 3 3 0 0	1127	718	31.86	0.80	63.66	1.17	-67.4	11.3	99.8	16.4

Dunbrack, R. L.; Cohen, F. E. "Bayesian statistical analysis of protein side-chain rotamer preferences." *Protein Sci.* ***1997, 6, 1661-1681.***

Sampling and Scoring for Side Chain Repacking and Design

Local Rotamer Bias

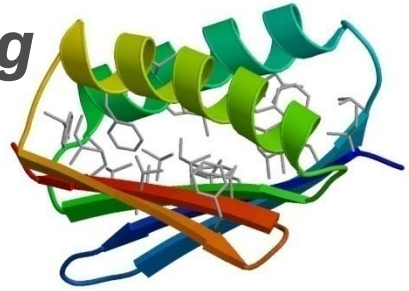
Approximate interactions within sidechain using the distribution of sidechain confor-mations (rotamers) seen in known protein structures



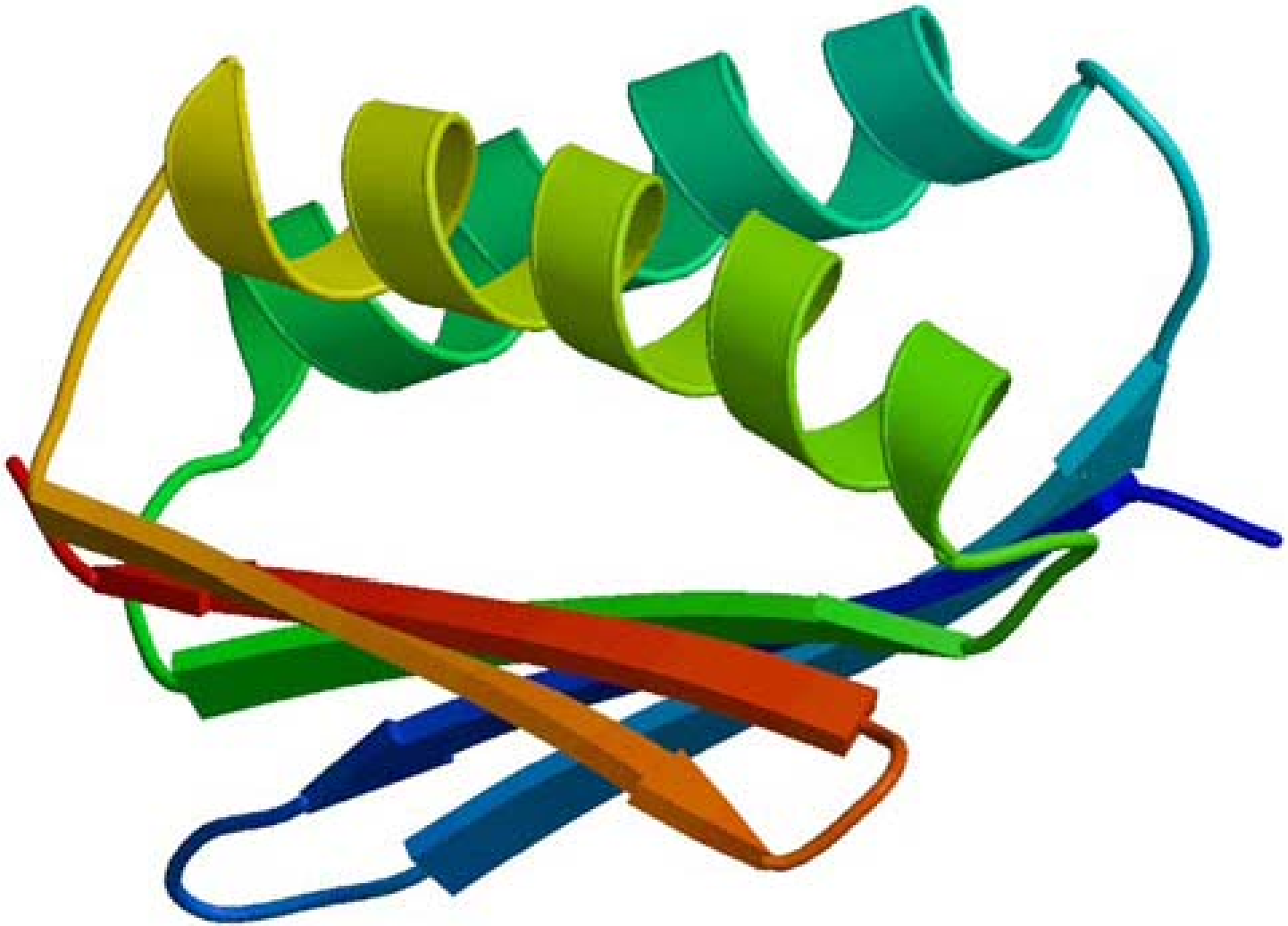
Energy function

Statistically derived potential function
- ***VDW interaction***
- ***solvation***
- ***hydrogen bonding potential***
- ***pair wise interactions***
- ***rotamer probability***

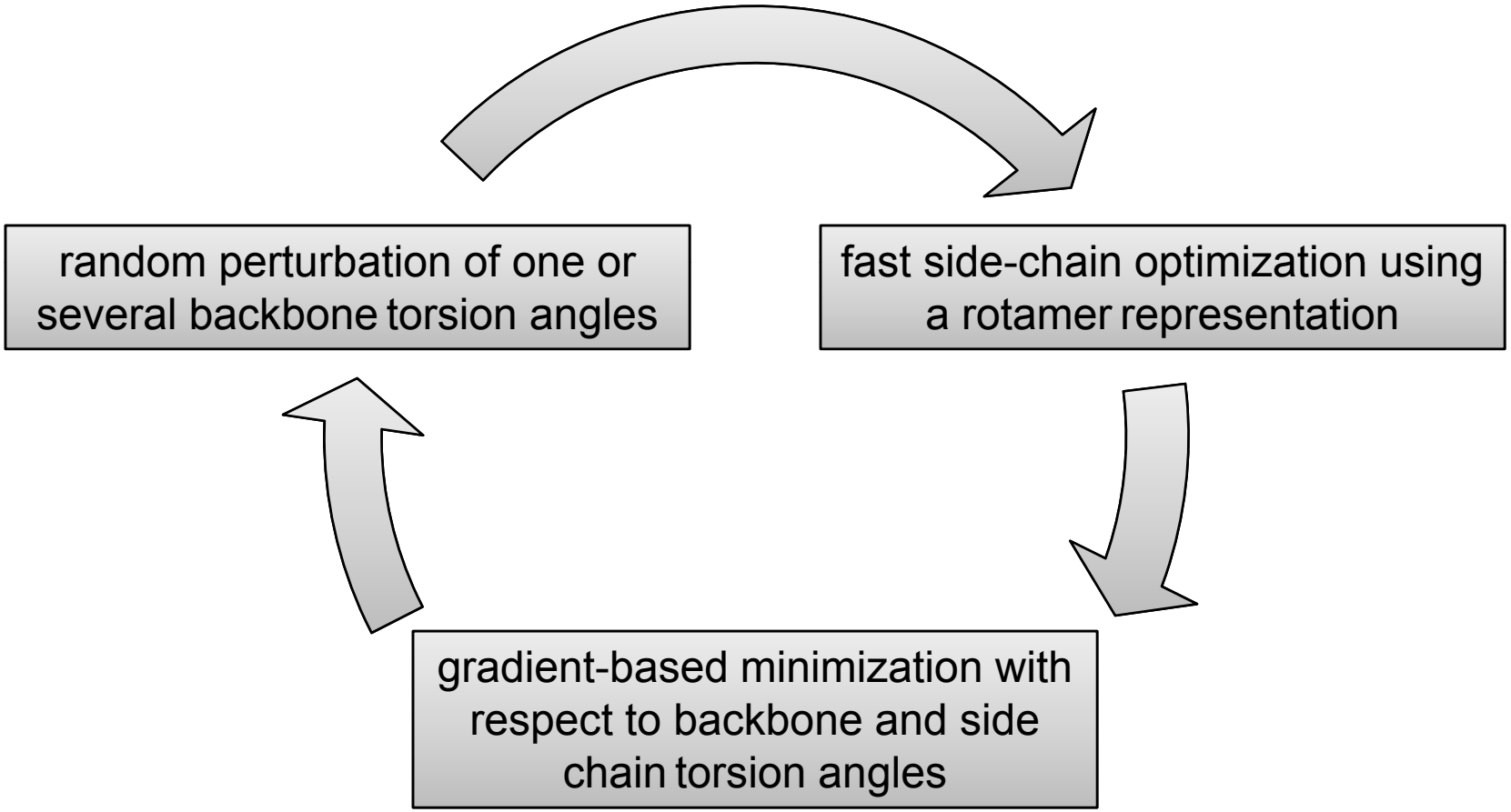
Simulated Annealing Monte Carlo energy minimization



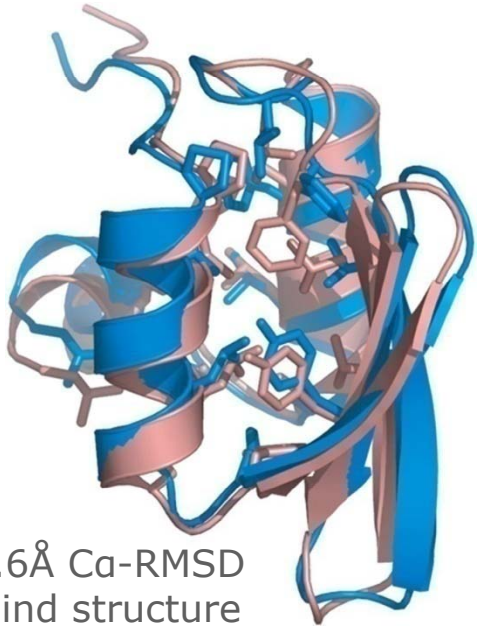
Dahiyat, B. I. and Mayo, S. L. (1997) *Science*, 278, 82-7
Dunbrack, R. L., Jr. and Karplus, M. (1993) *J Mol Biol*, 230, 543-74.
Kuhlman, B., et. al. (2003) *Science*, 302, 1364-1368.



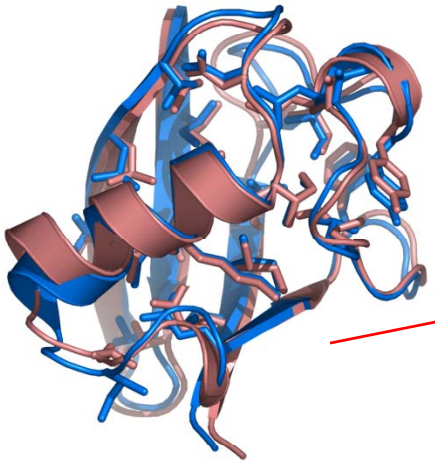
Refinement Cycle with Side Chain Repacking and All Atom Minimization

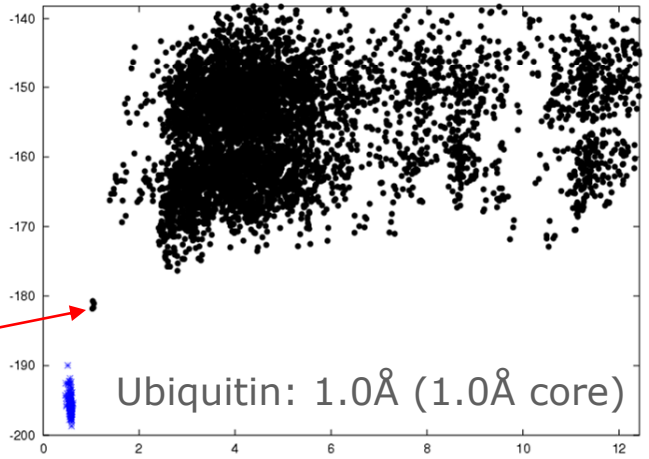


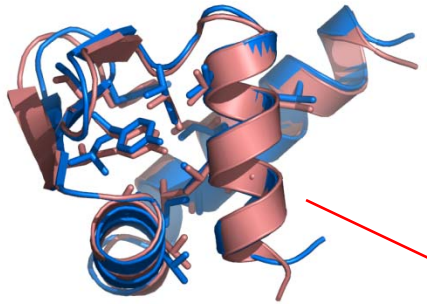
CASP target T0281 and other Benchmark Examples

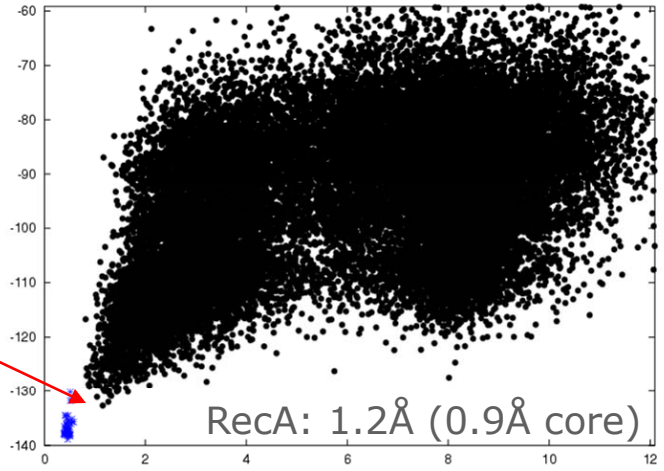


1.6Å Cα-RMSD blind structure prediction for CASP6 target T0281, hypothetical protein from Thermus thermophilus Hb8. Superposition of our submitted model for this target in CASP6 (blue) with the crystal structure (red; PDB code 1whz)







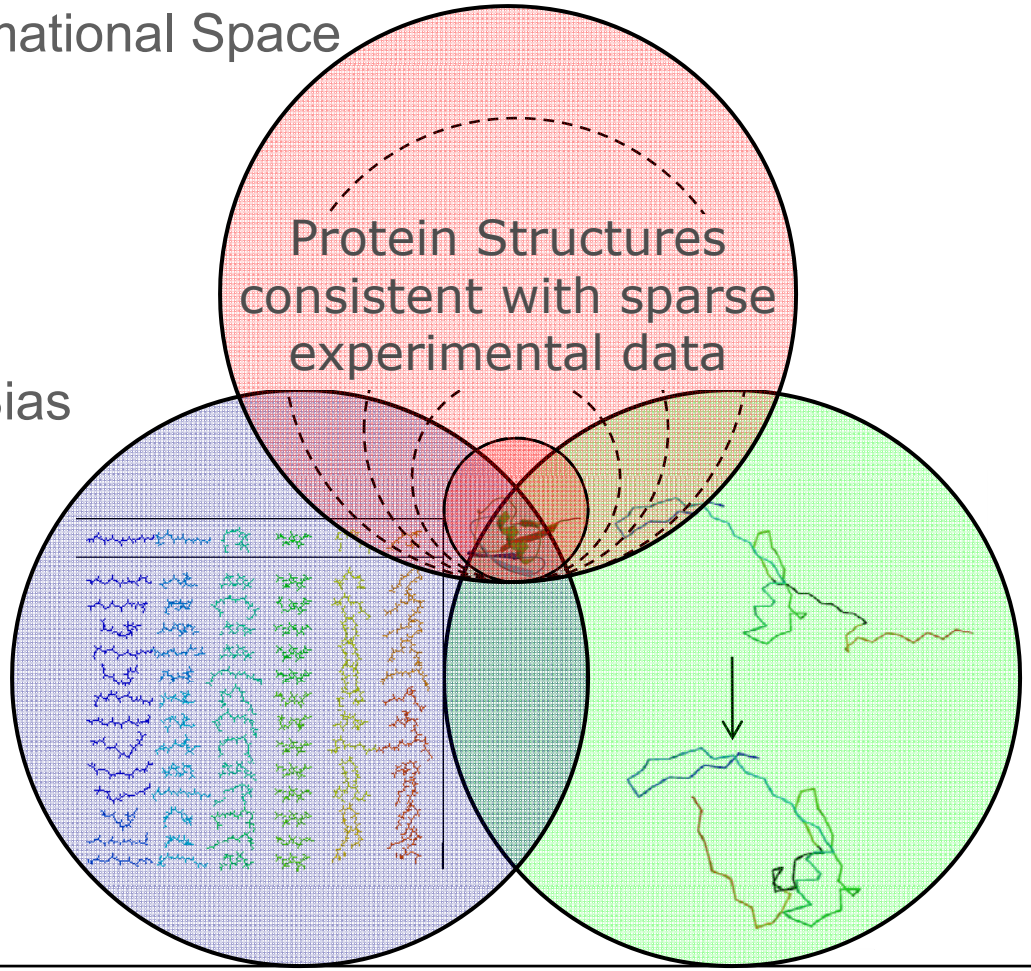


Benchmark results

ID	Length	%α	%β	RMSD Cα	RMSD core	Protein
1b72A	49	69	0	**0.8**	**0.8**	Hox-B1 homeobox protein
1shfA	59	5	40	10.8	8.5	Fyn tyrosine kinase
1tif_	59	22	37	4.1	2.3	IF3-N
2reb_2	60	61	20	**1.2**	**0.9**	RecA
1r69_	61	63	0	**1.2**	**1.5**	434 repressor
1csp_	67	4	53	4.7	4.2	Cold-shock protein
1di2A_	69	46	33	2.6	2.2	RNA binding protein A
1n0uA4	69	43	24	9.9	8.1	Elongation factor 2
1mla_2	70	34	37	8.4	7.3	Malonyl-CoA ACP transacylase
1af7__	72	72	0	10.1	7.9	Cher domain 1
1ogwA_	72	26	33	**1.0**	**1.0**	Ubiquitin
1dcjA_	73	31	27	2.5	2.2	Yhhp
1dtjA_	74	39	27	**1.0**	**0.8**	KH domain of Nova-2
1o2fB_	77	38	27	10.1	8.7	Glucose-permease IIBC
1mkyA3	81	32	24	3.2	3.6	Enga
1tig_	88	35	35	3.5	3.4	IF3-C

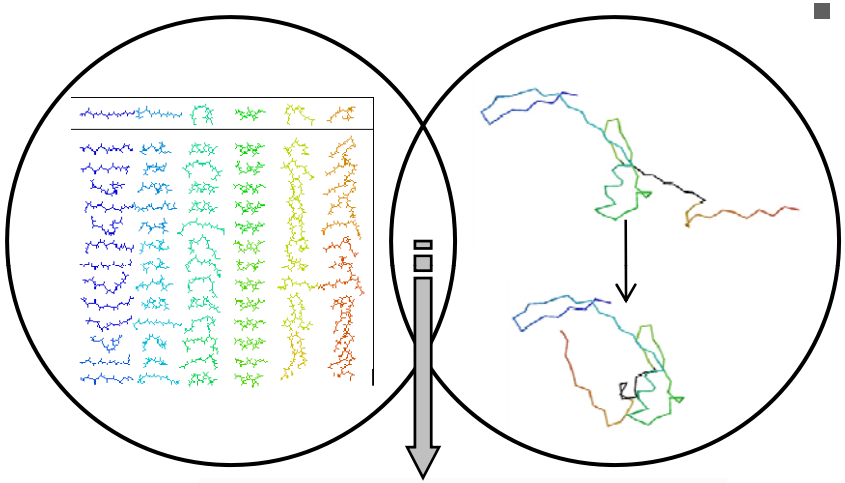
Combining Strengths: Building Accurate Models from Sparse Data

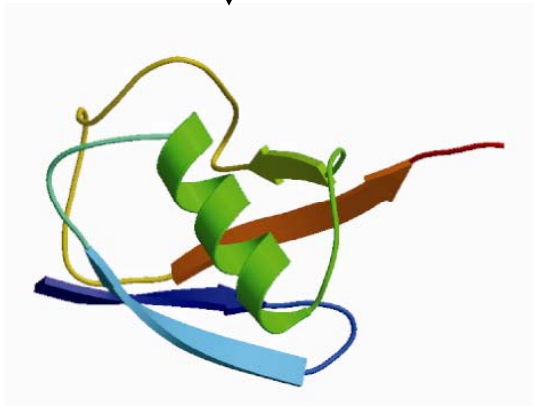
- Complete Conformational Space
- Local Sequence Bias
- Energy Evaluation of non-Local Interactions



RosettaNMR: Usage of CSs, NOEs, and RDCs

- NMR data are used ***in addition*** to the Local Sequence Bias
  - CS derived dihedral angle restrictions (via TALOS)
  - Local NOE distance restraints
  - RDC orientation restraints





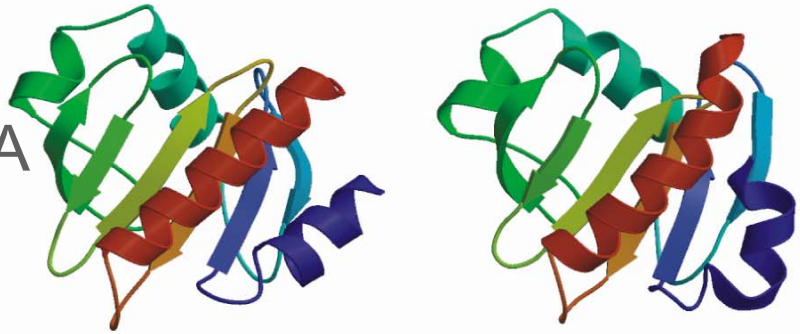
- NMR data are used ***in addition*** to energy evaluation of non-local Interactions
  - Long-range NOE distance restraints
  - RDC orientation restraints

Bowers, P. M.; Strauss, C. E. M.; Baker, D., *J. Biomol. NMR* **2000,** 18, 311-318.
Rohl, C.; Baker, D., *J. Am. Chem. Soc.* **2002,** 124, (11), 2723-2729.

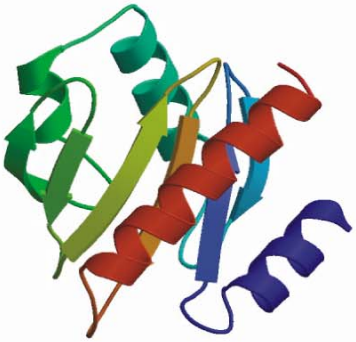
RosettaNMR: High-Resolution from "one" Restraint per Amino Acid

- Profilin I with CSs and 1.01 NOEs/AA

Xray/NMR

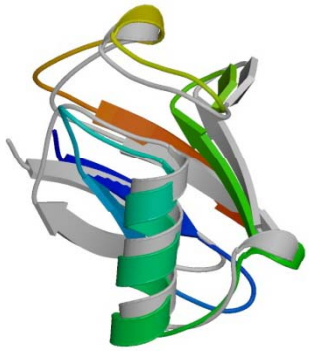


RosettaNMR

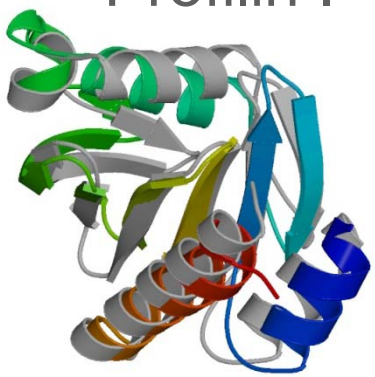


- Ubiquitin with 0.89 RDCs/AA
- Profilin I with CSs and 1.25 RDCs/AA

Ubiquitin

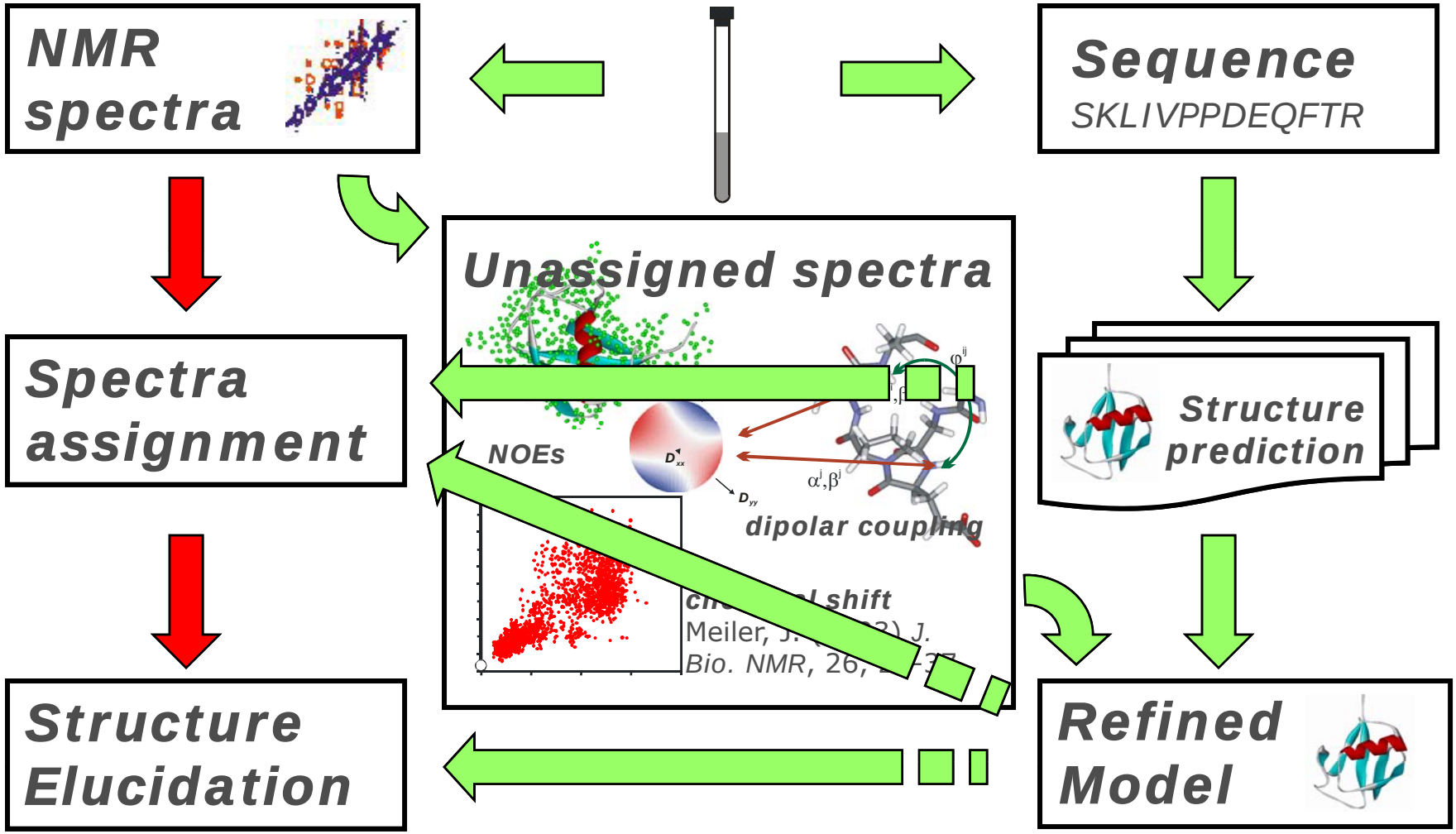


Profilin I

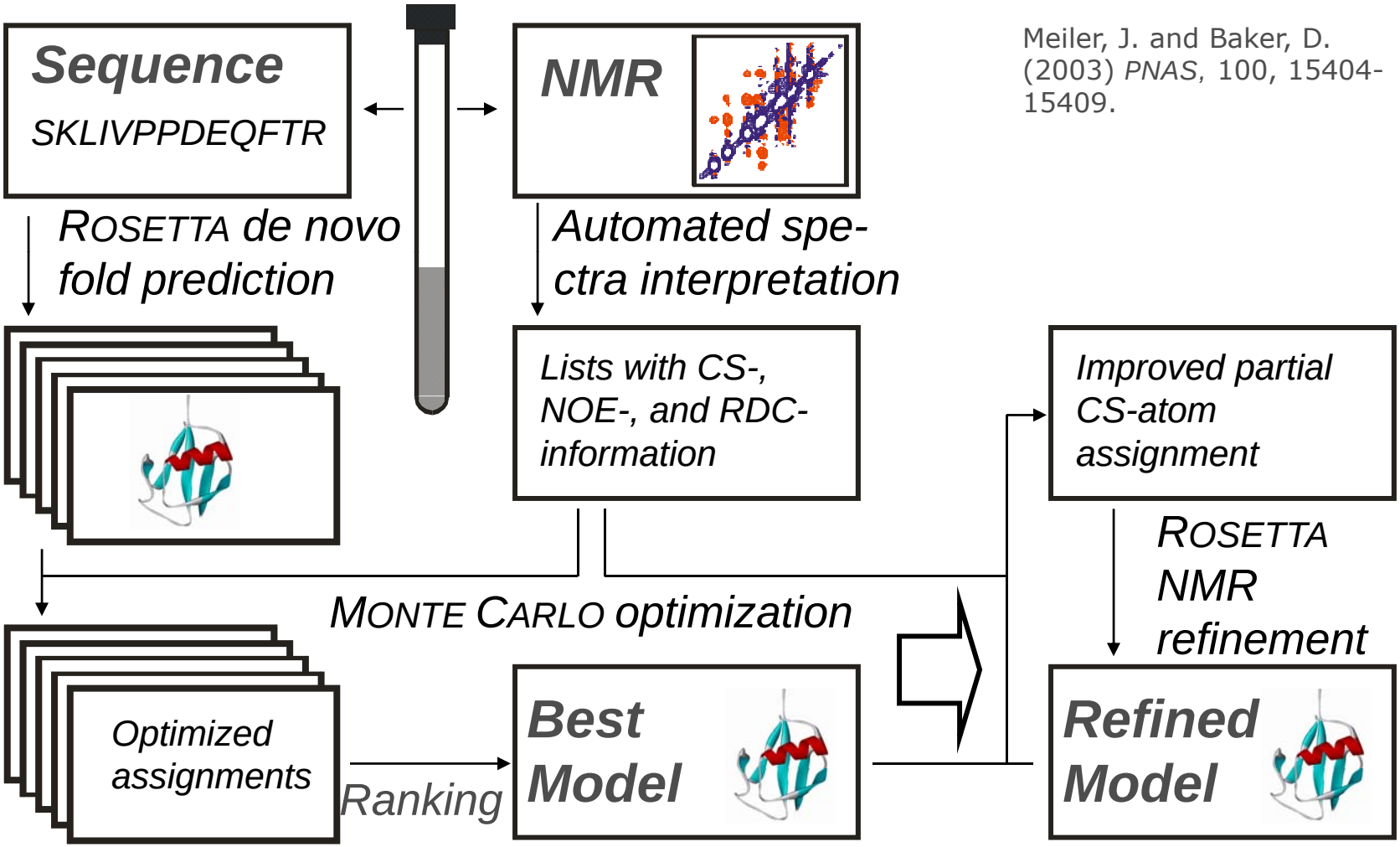


Bowers, P. M.; Strauss, C. E. M.; Baker, D., *J. Biomol. NMR* **2000,** 18, 311-318.
Rohl, C.; Baker, D., *J. Am. Chem. Soc.* **2002,** 124, (11), 2723-2729.

Protein Structure Elucidation with NMR Spectroscopy

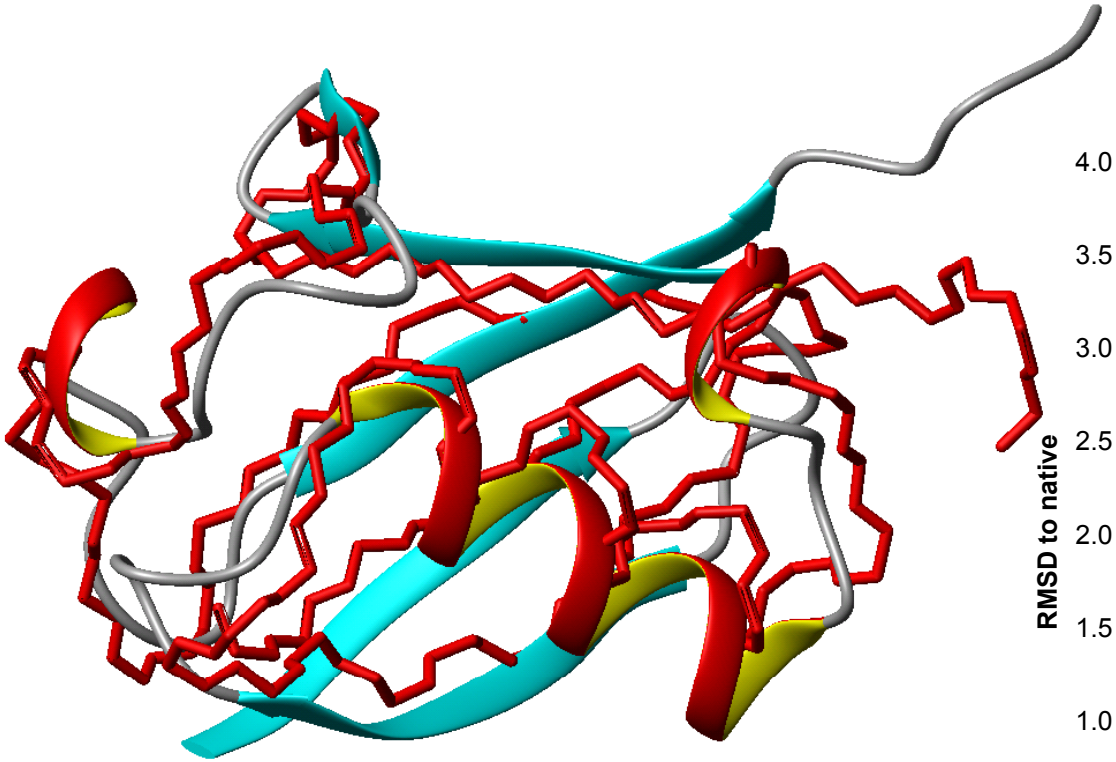


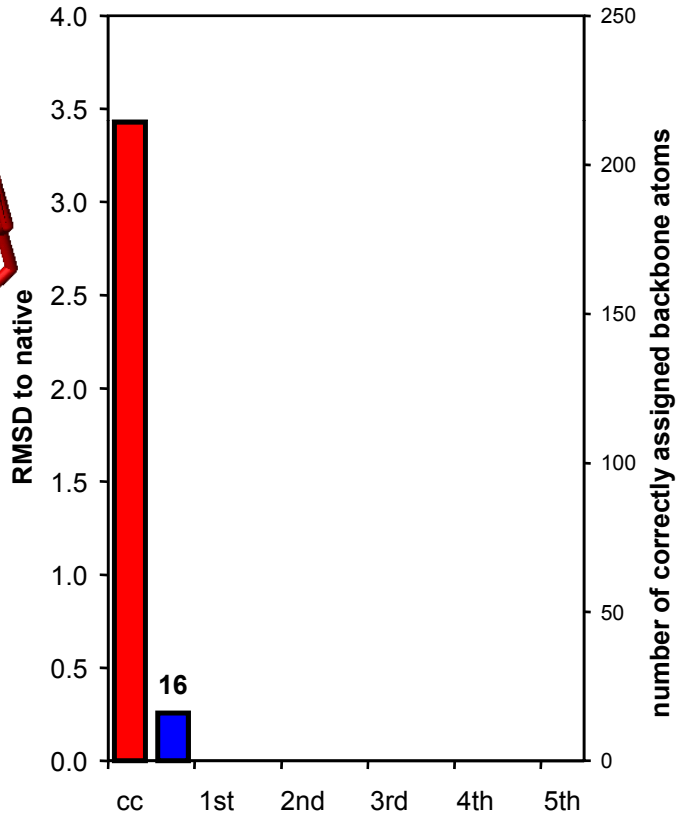
Fold Determination and High-Resolution Model Refinement



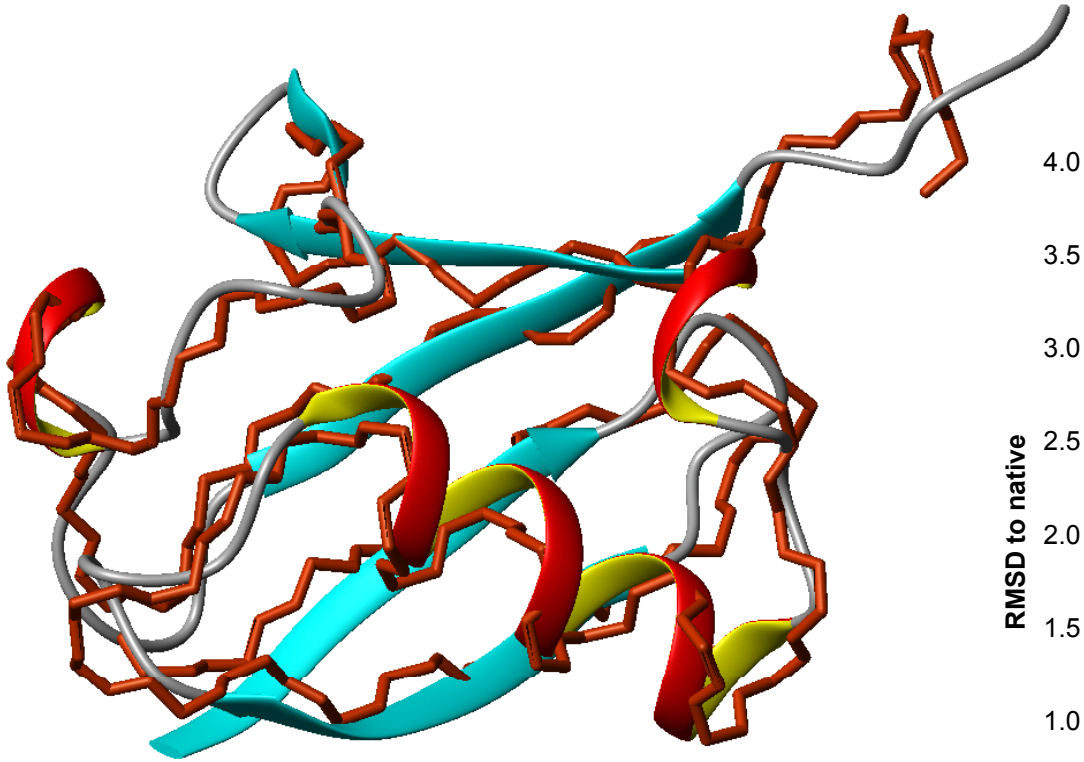
Meiler, J. and Baker, D. (2003) *PNAS*, 100, 15404-15409.

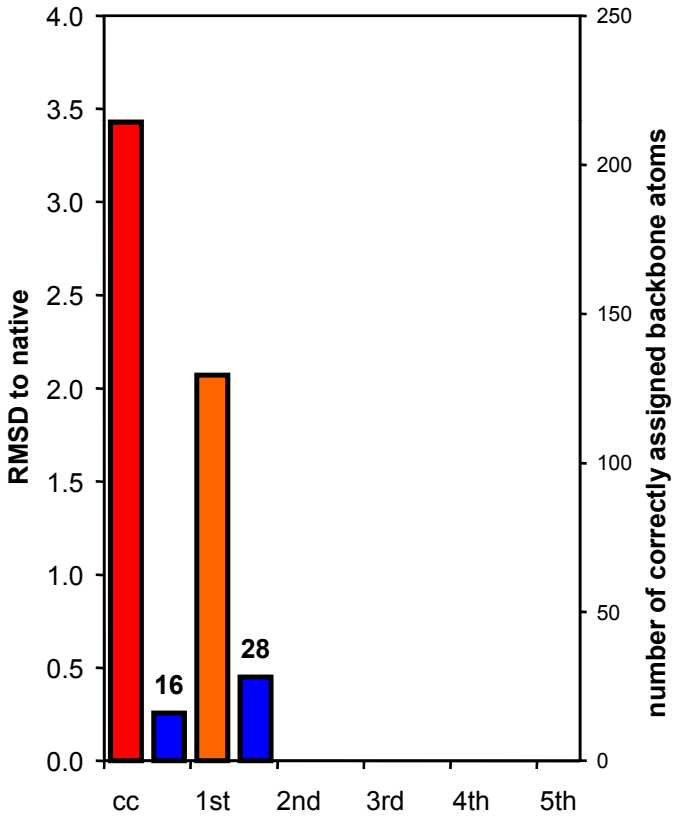
1ubi_ initial model



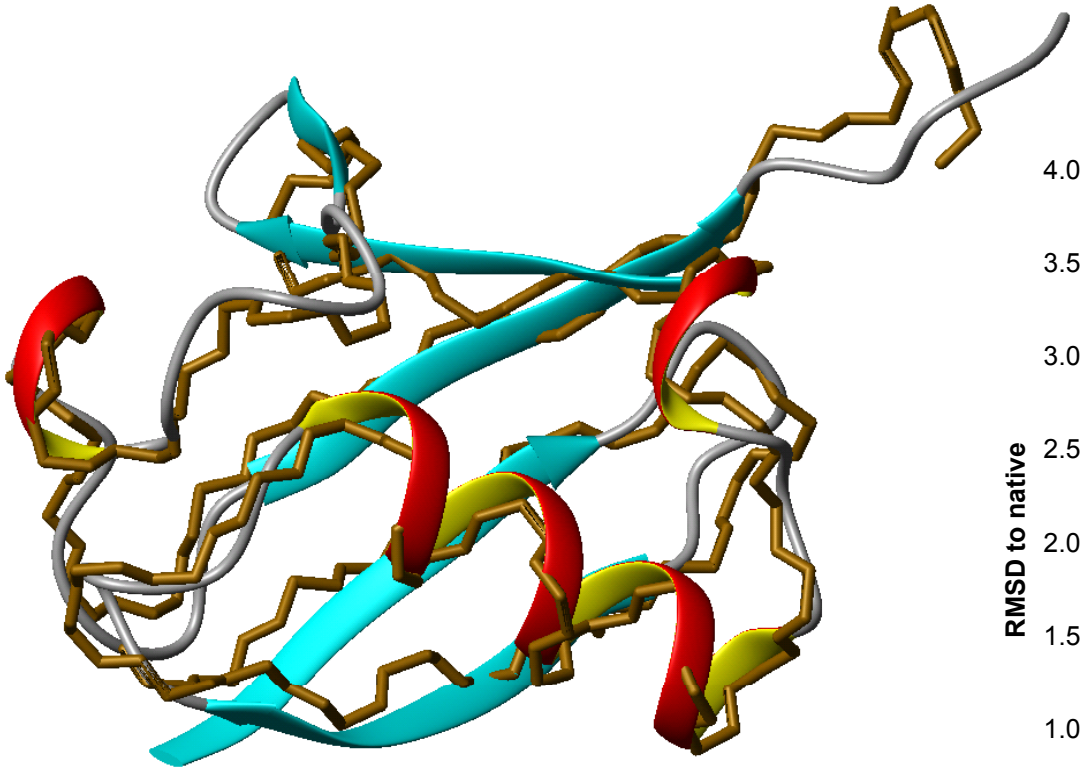


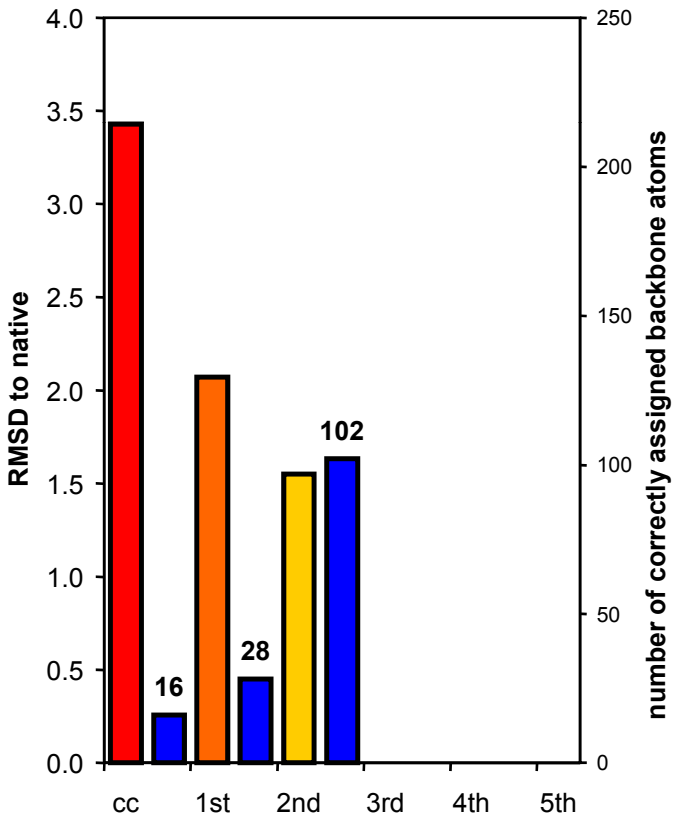
1ubi_ Refinement 1st Step



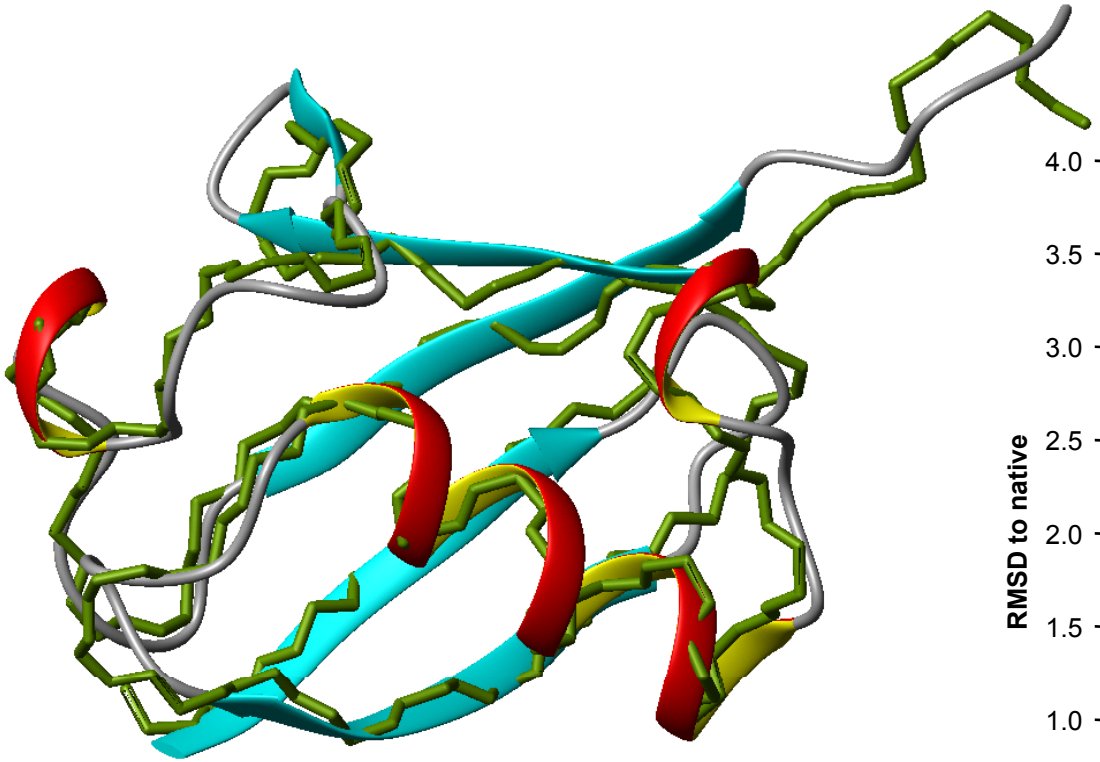


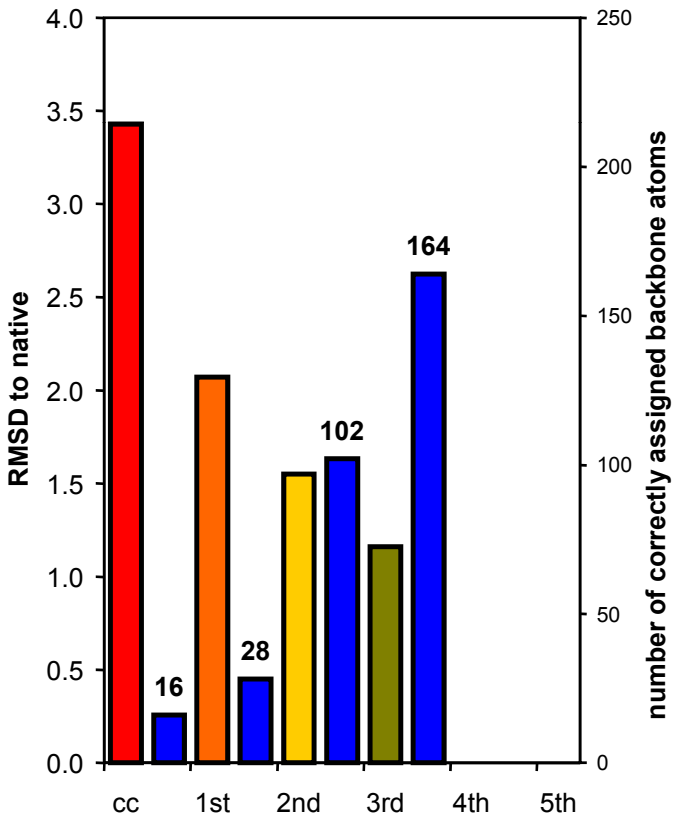
1ubi_ Refinement 2nd Step



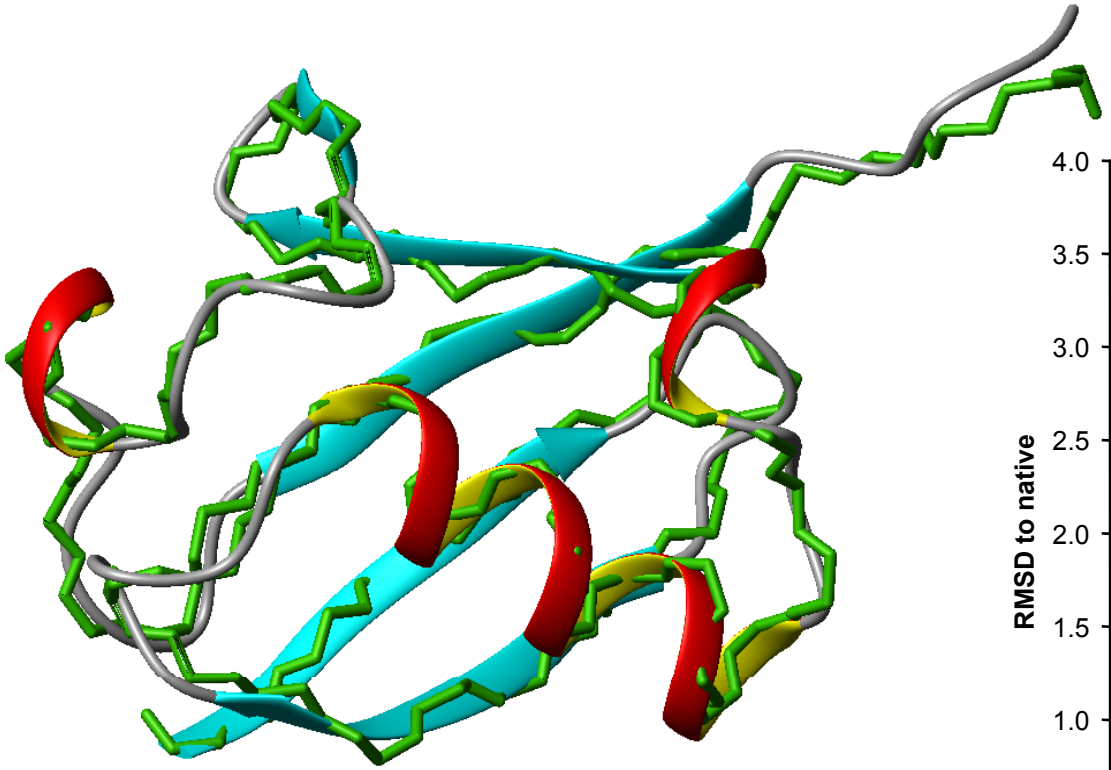


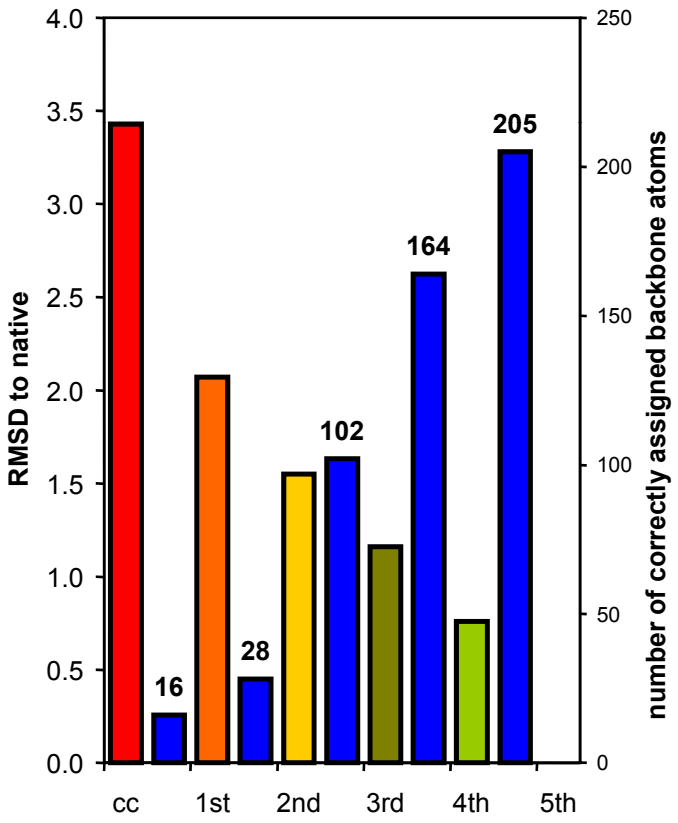
1ubi_ Refinement 3rd Step



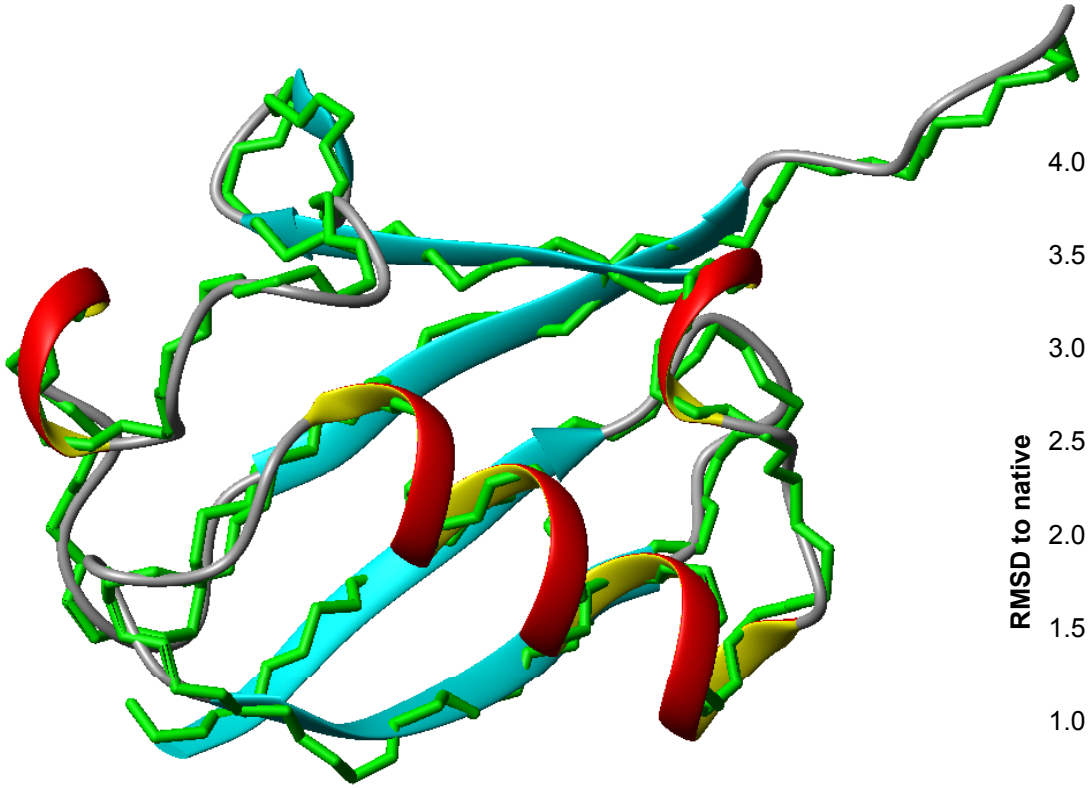


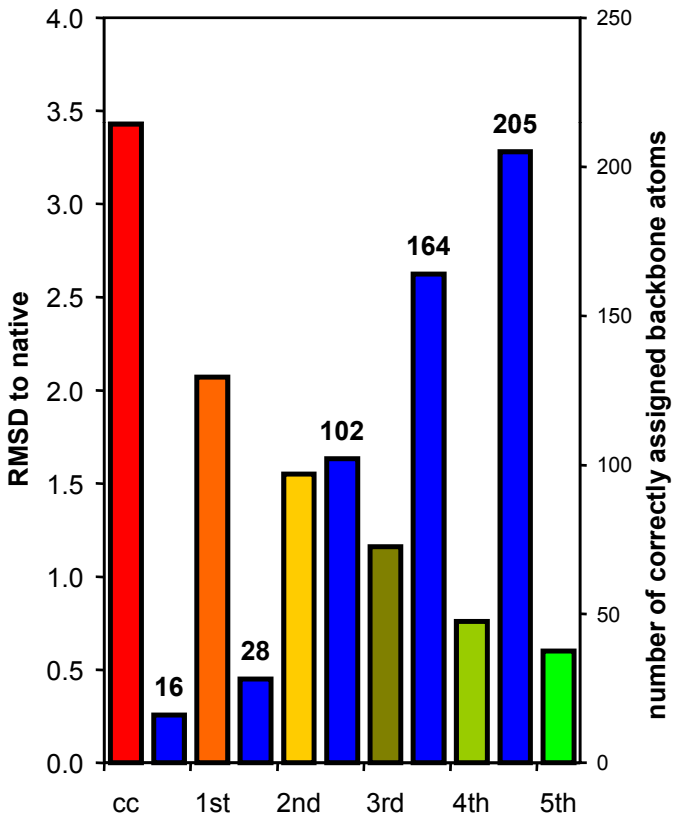
1ubi_ Refinement 4th Step



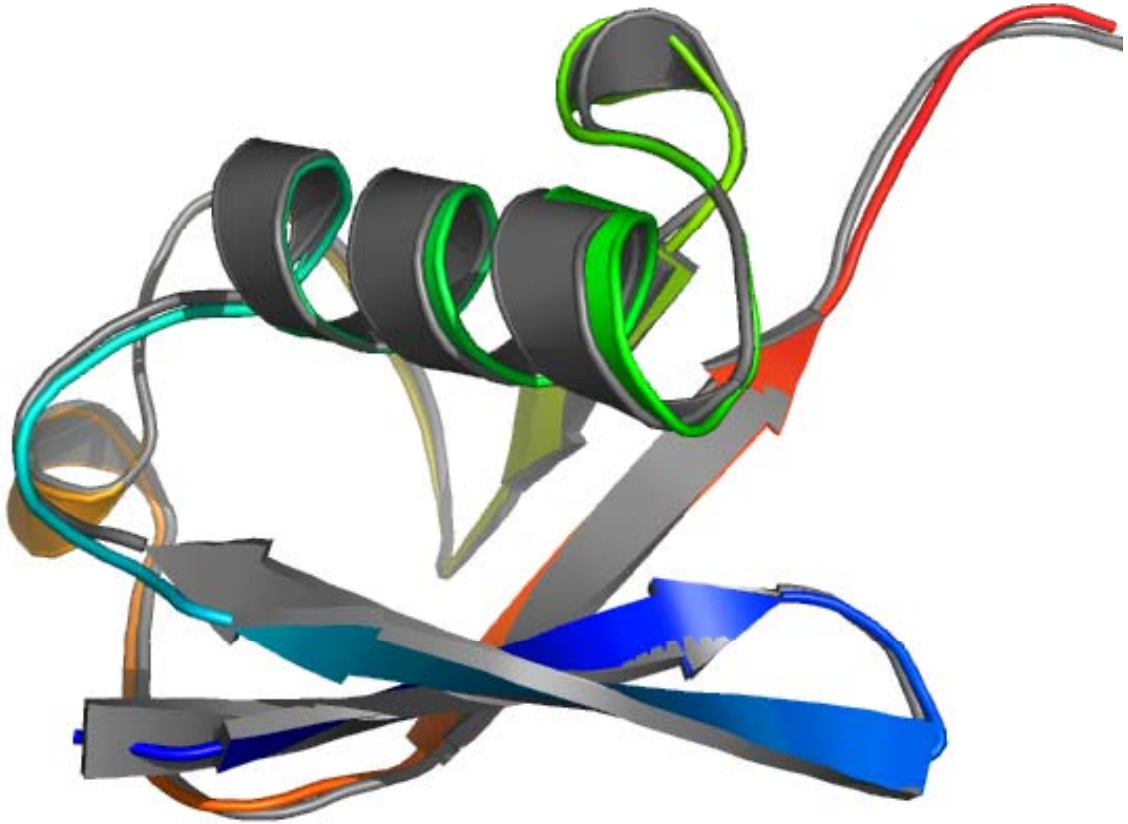


1ubi_ Refinement 5th Step



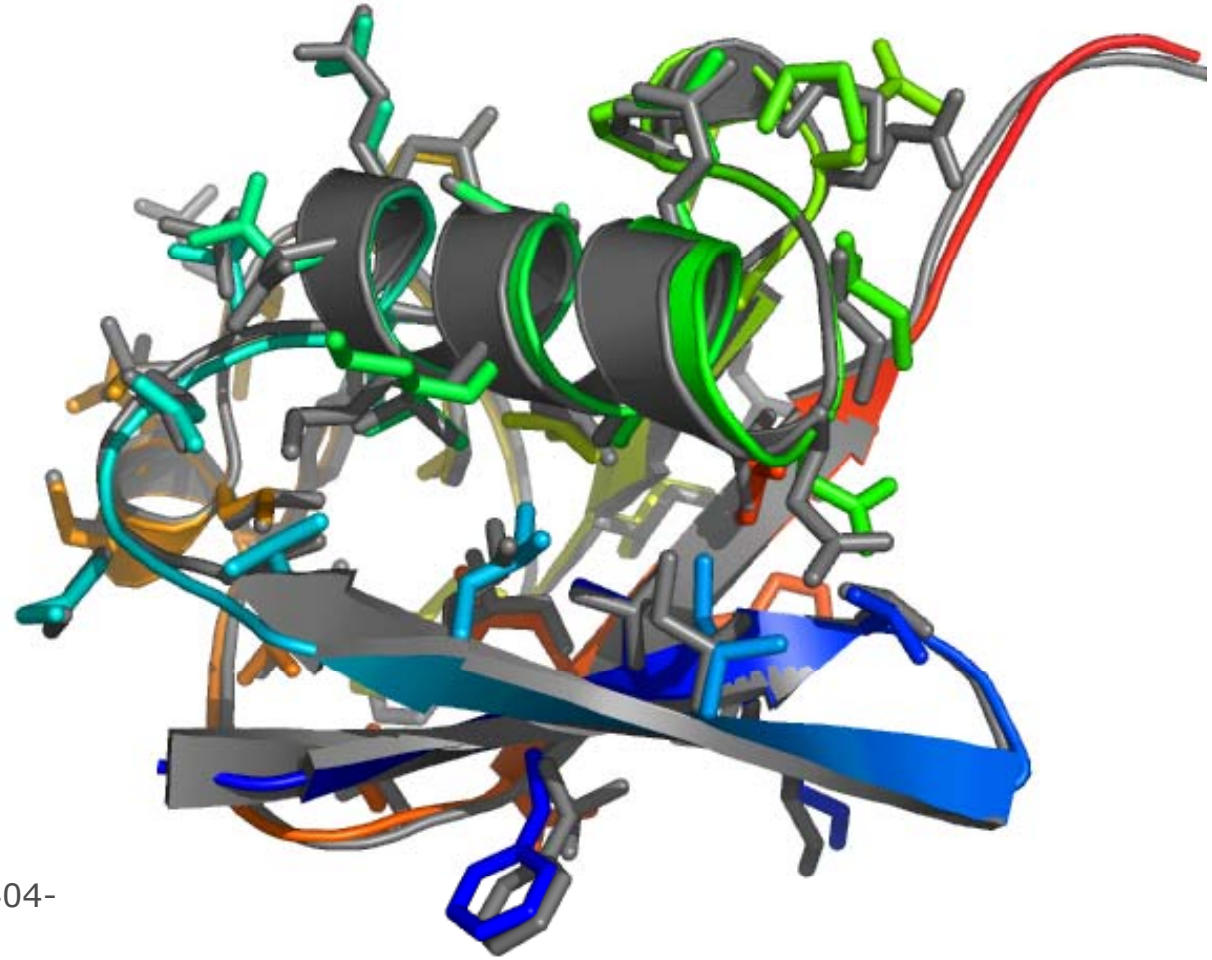


Backbone RMSD is 0.6Å



Meiler, J. and Baker, D. (2003) *PNAS*, 100, 15404-15409.

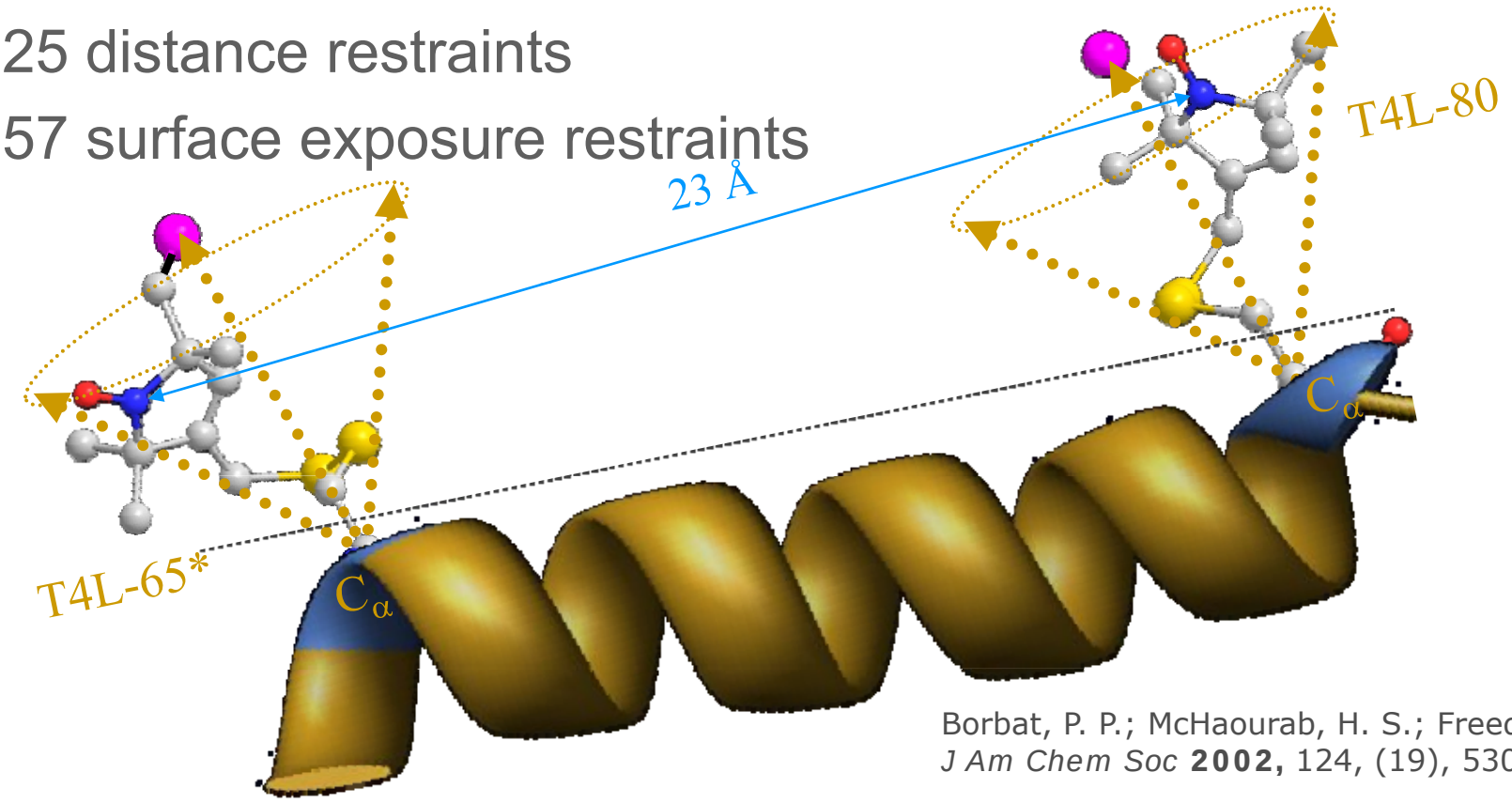
All Core Amino Acids have Correct Side Chain Conformation



Meiler, J. and Baker, D. (2003) *PNAS*, 100, 15404-15409.

Structure Elucidation of Lysozyme from 25 Experimental EPR Distances

- 25 distance restraints
- 57 surface exposure restraints

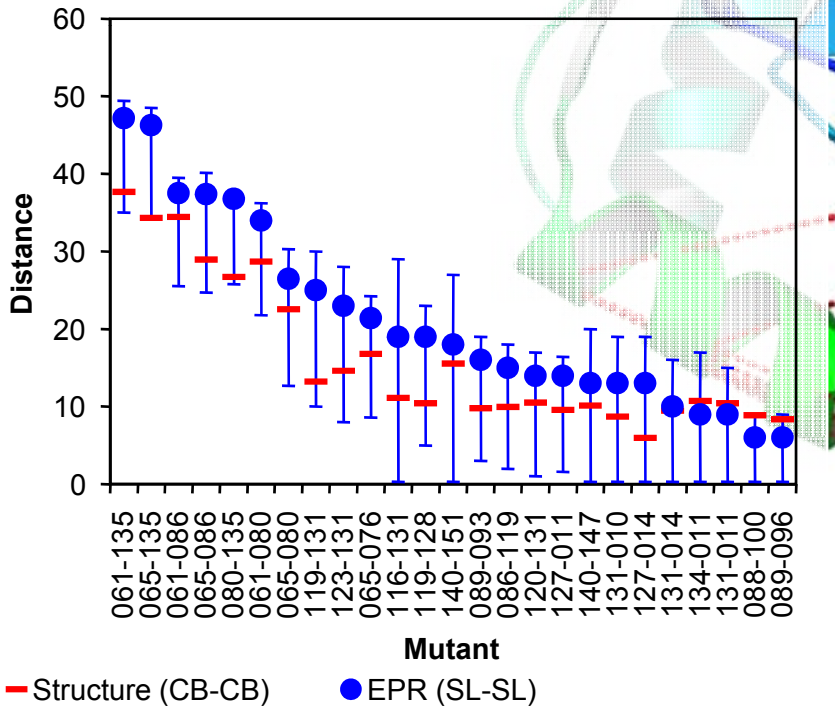


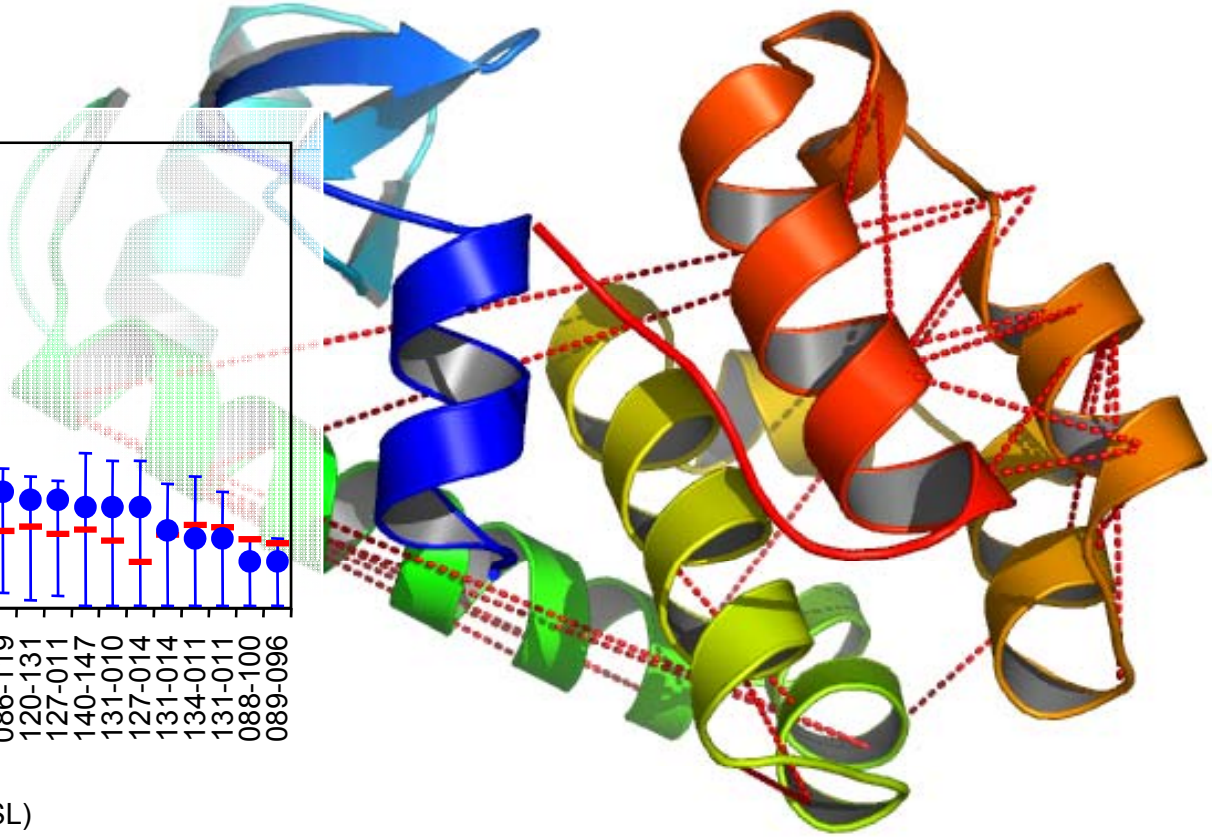
Borbat, P. P.; McHaourab, H. S.; Freed, J. H., *J Am Chem Soc* **2002,** 124, (19), 5304-14.

- Thanks to Hassane Mchaourab and his lab for experimental data

25 Experimental Distance Restraints

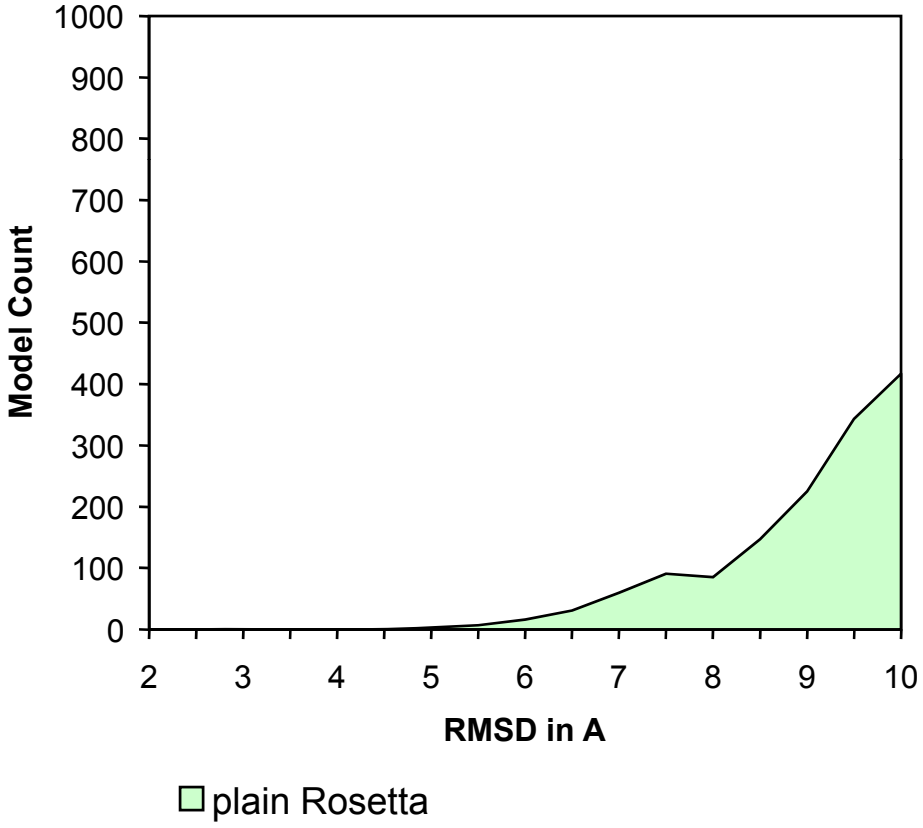
- Restraint: $(d_{SL\text{-}SL}-\sigma_{SL\text{-}SL}\ -10Å) \leq d_{CB\text{-}CB} \leq (d_{SL\text{-}SL}+\sigma_{SL\text{-}SL})$
- Harmonic penalty function



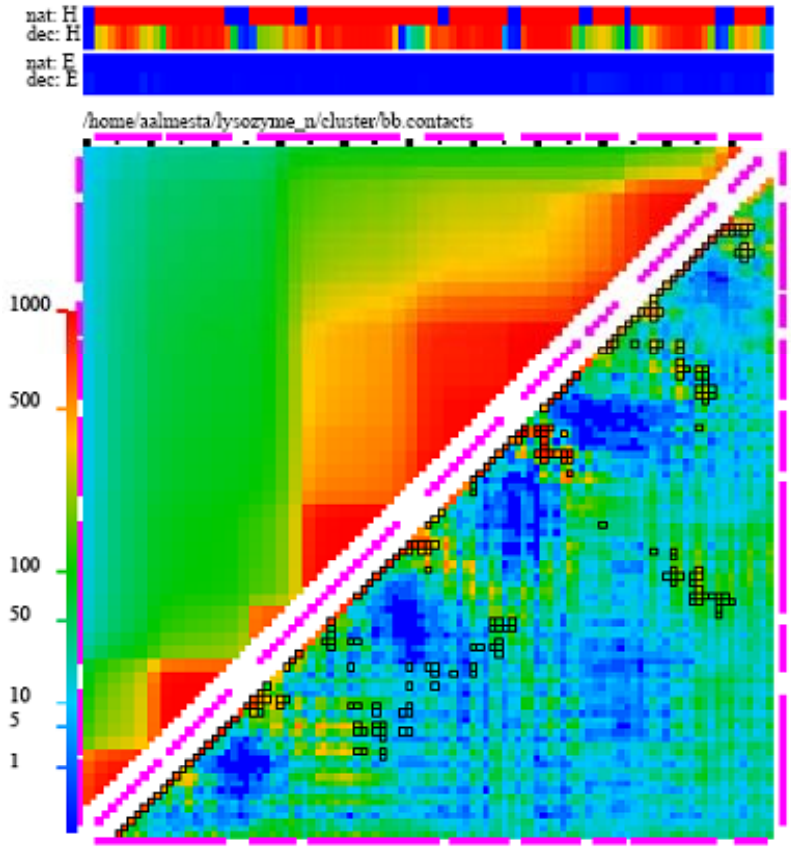


Influence of Experimental Data on Sampling and Model Quality

- RMSD histogram



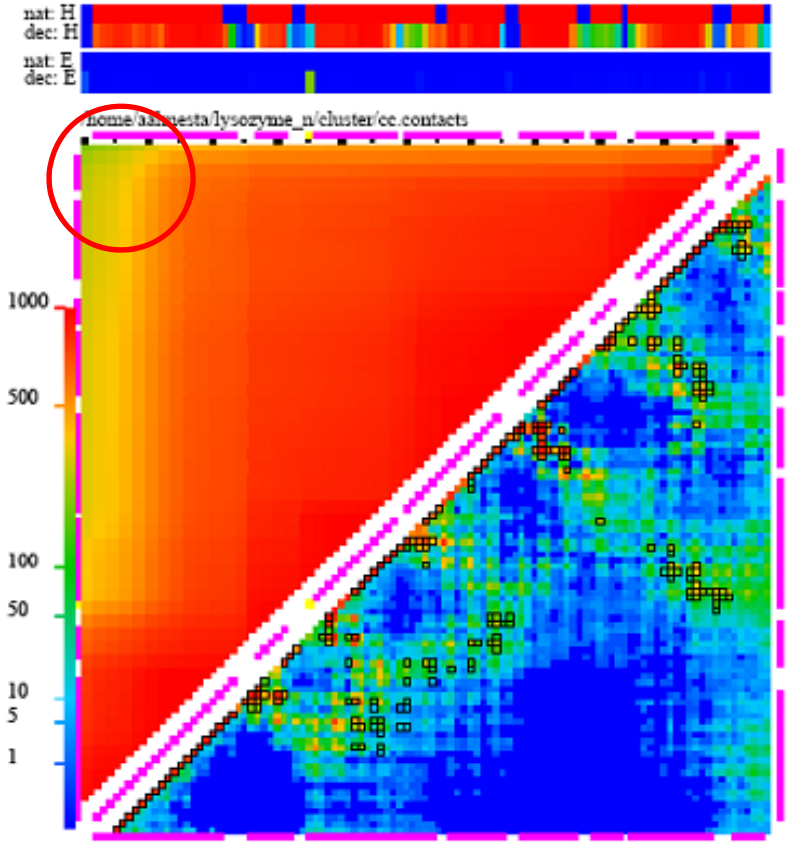
- C



Influence of Experimental Data on Sampling and Model Quality

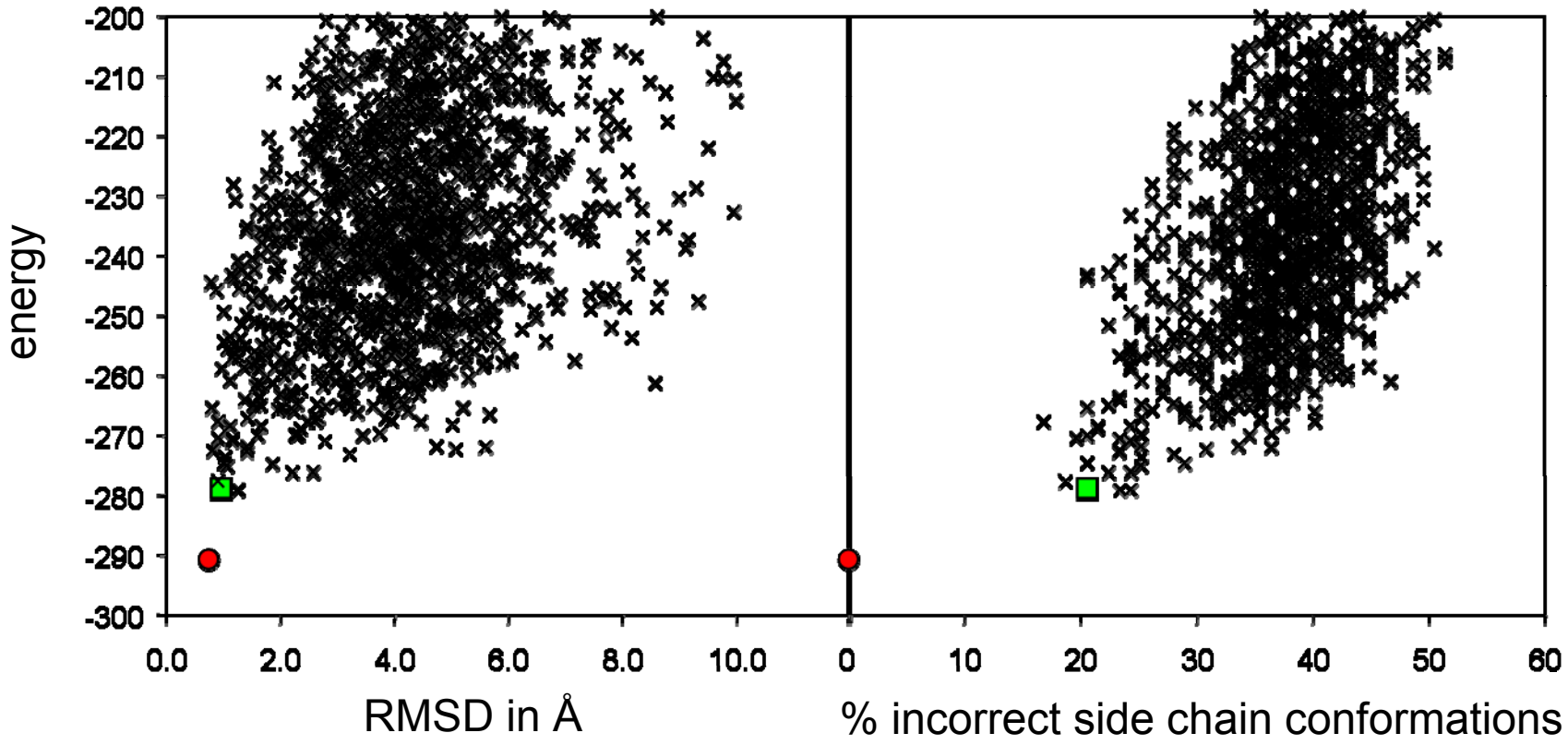
- RMSD histogram

- C

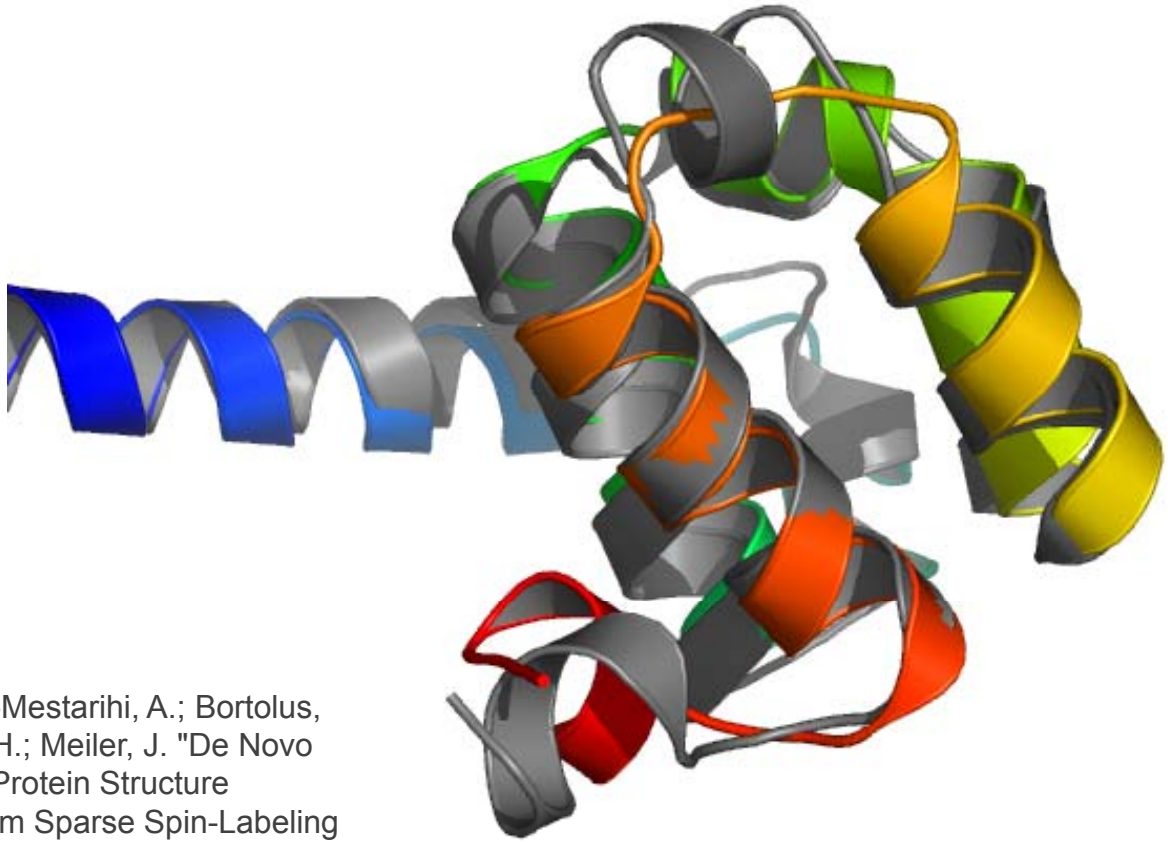


High Resolution Energy Refinement

- Lowest scoring ~11,000 models out of 500,000 were refined

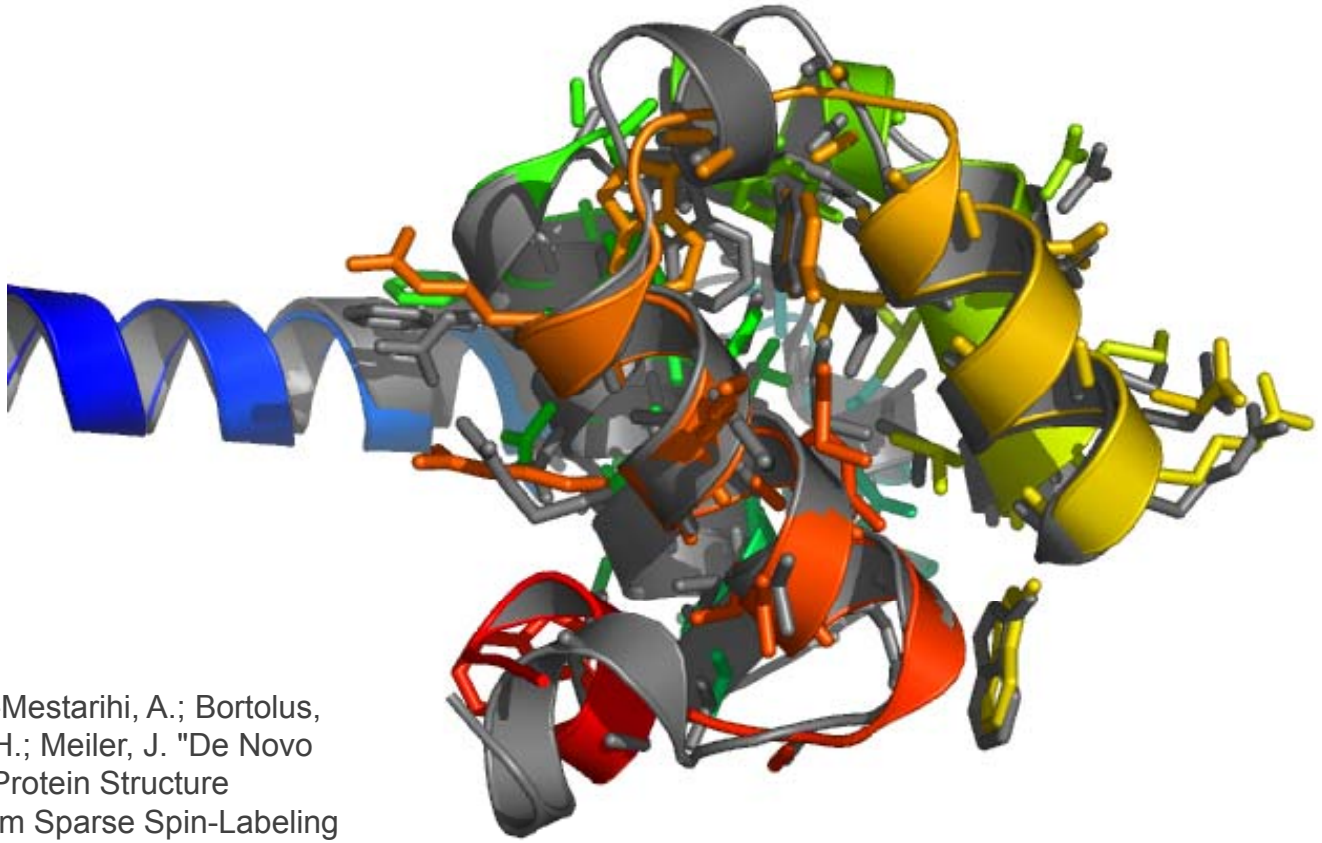


Backbone RMSD is 0.96Å



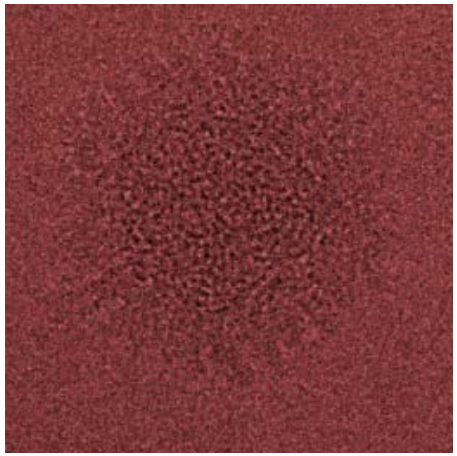
Alexander, N.; Al-Mestarihi, A.; Bortolus, M.; McHaourab, H.; Meiler, J. "De Novo High-Resolution Protein Structure Determination from Sparse Spin-Labeling EPR Data" *Structure* ***2008, 16, 181-95.***

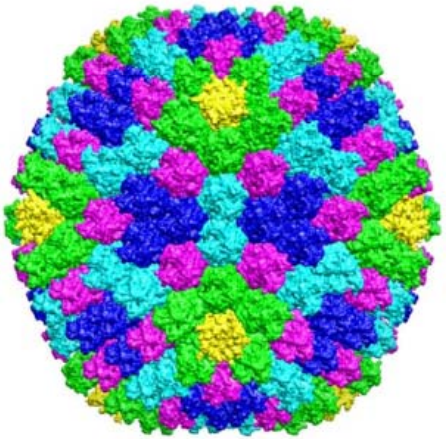
All but Two Core Amino Acids have Correct Side Chain Conformation

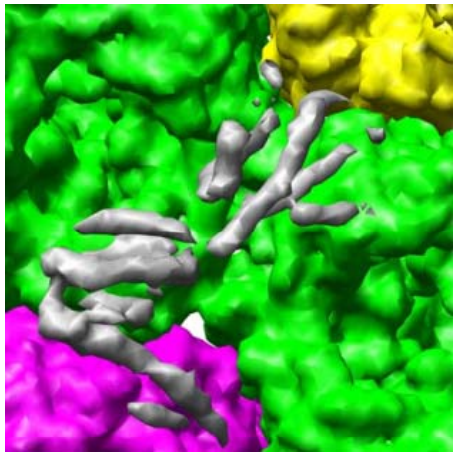


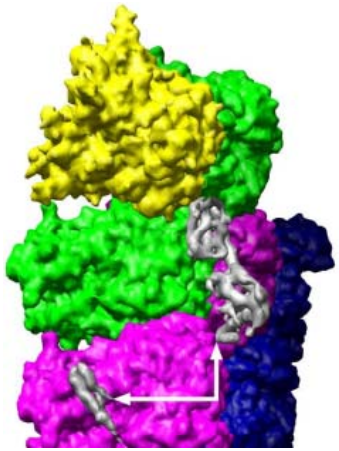
Alexander, N.; Al-Mestarihi, A.; Bortolus, M.; McHaourab, H.; Meiler, J. "De Novo High-Resolution Protein Structure Determination from Sparse Spin-Labeling EPR Data" *Structure* ***2008, 16, 181-95.***

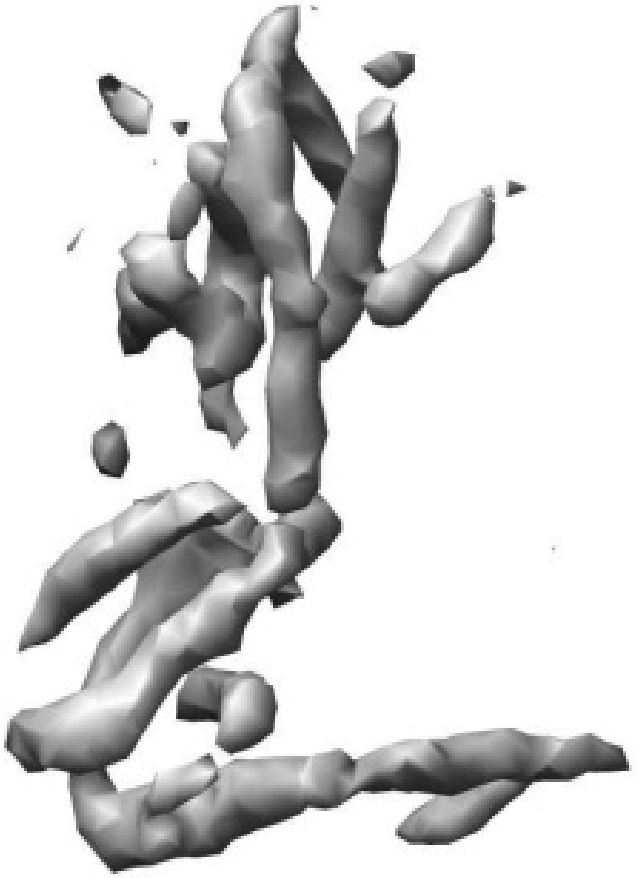
Adenovirus Protein IIIa – A Novel Topology?







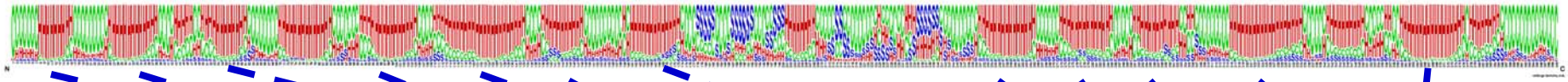




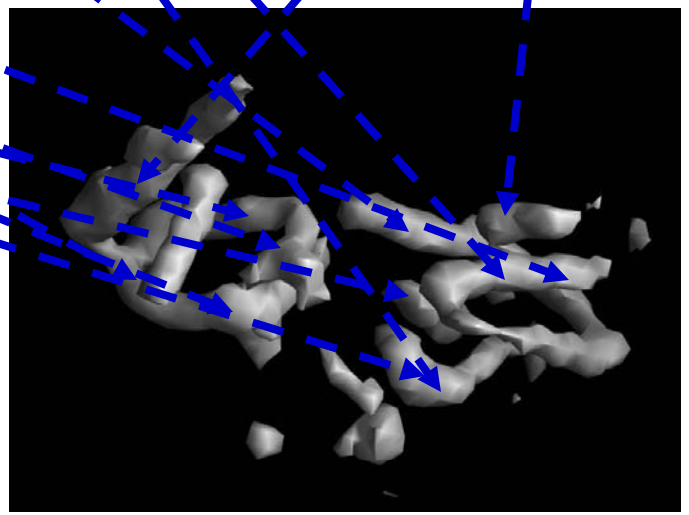
Saban, S. D.; Silvestry, M.; Nemerow, G. R.; Stewart, P. L., *J Virol* ***2006, 80, (24), 12049-59.***

Number of Possible Placements of y Helices in x Density Rods

- Secondary structure prediction with y helices



- EM map with x density rods



$$n = \max(x, y) \quad \| \quad k = \min(x, y)$$

$$\text{permutations} = k!$$

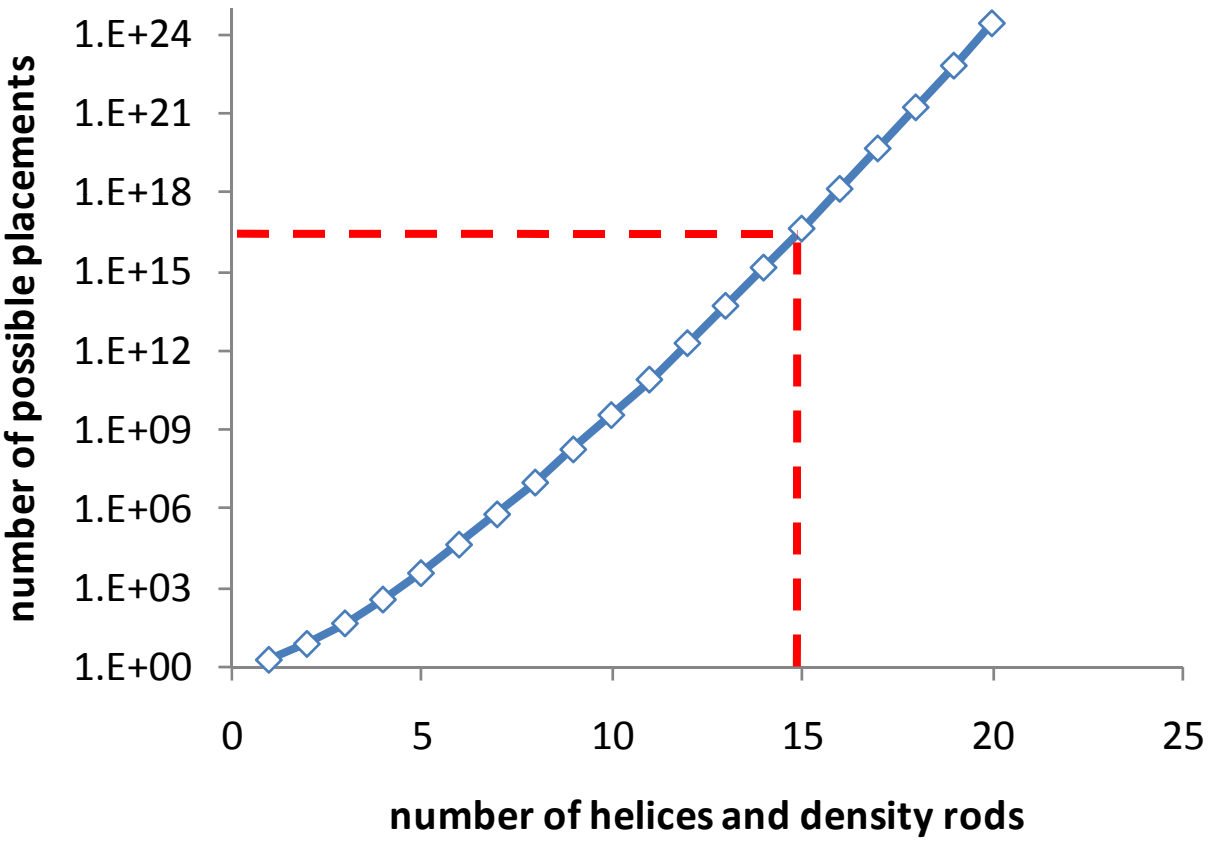
$$\text{orientations} = 2^k$$

$$\text{combinations} = n!/k!(n-k)!$$

$$\text{total} = 2^k\, n!/(n-k)! \quad \| \quad 3 \text{ helices and } 2 \text{ densities: } 2^2 3!/(3-2)! = 24$$

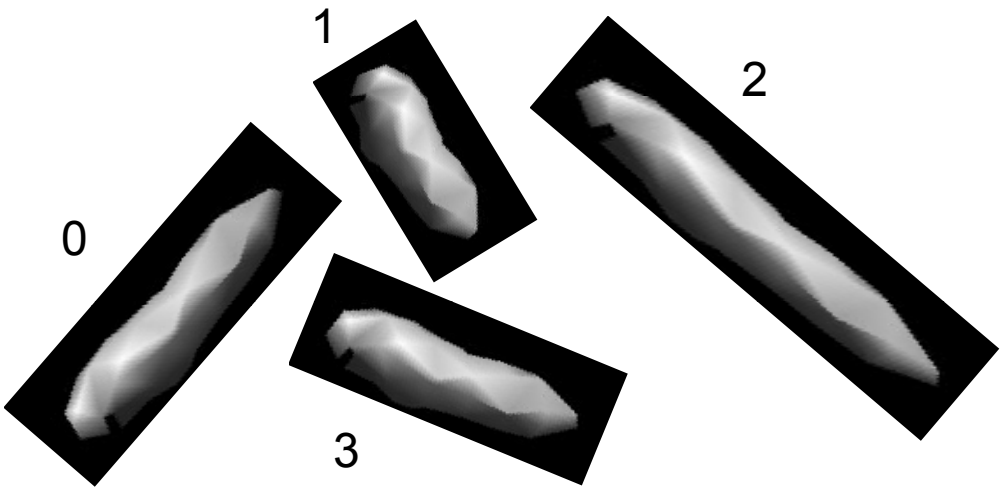
Size of Search Space Grows Exponentially

- About 10^{17} possible placements for protein IIIa
- Sample one per second and you are done in $4 \cdot 10^{9}$ years
- This is about the age of the earth

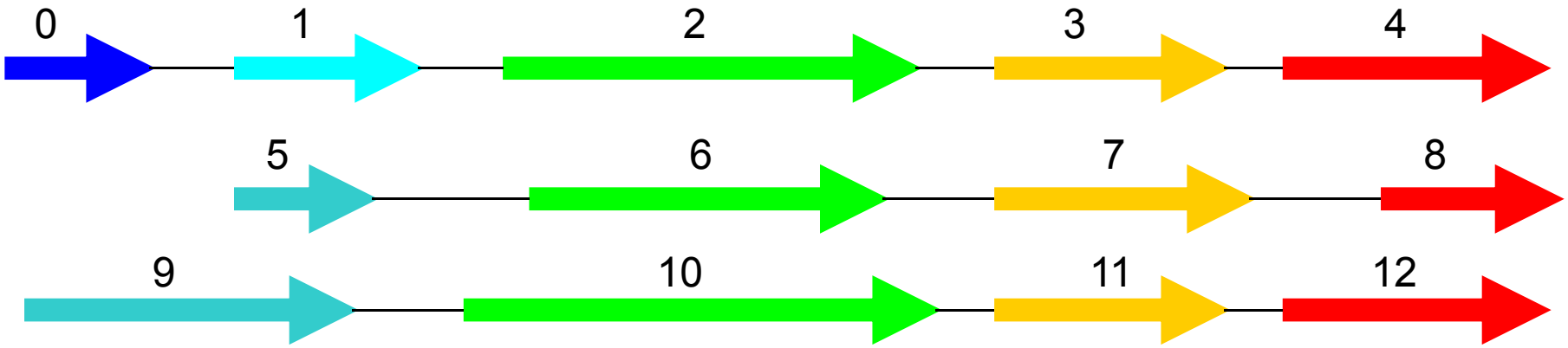


Monte Carlo Assembly Animation Start

- Experimental density

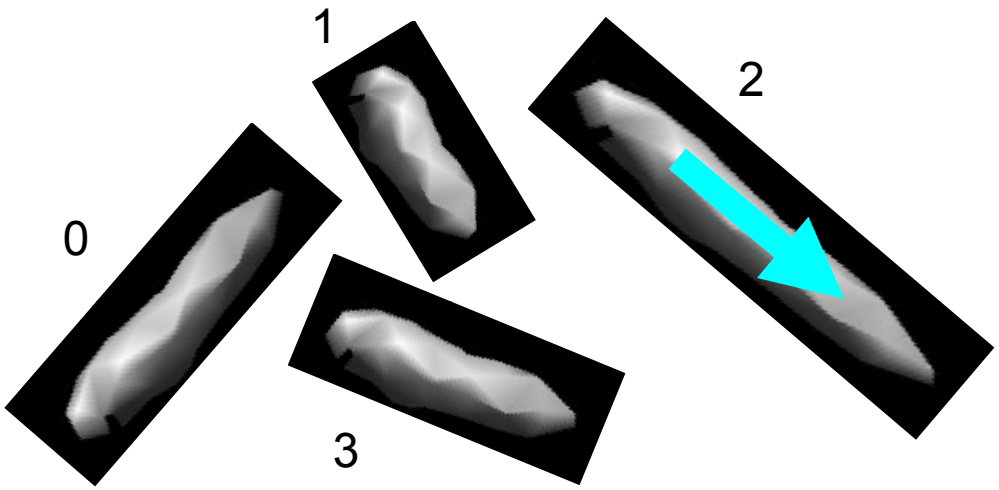


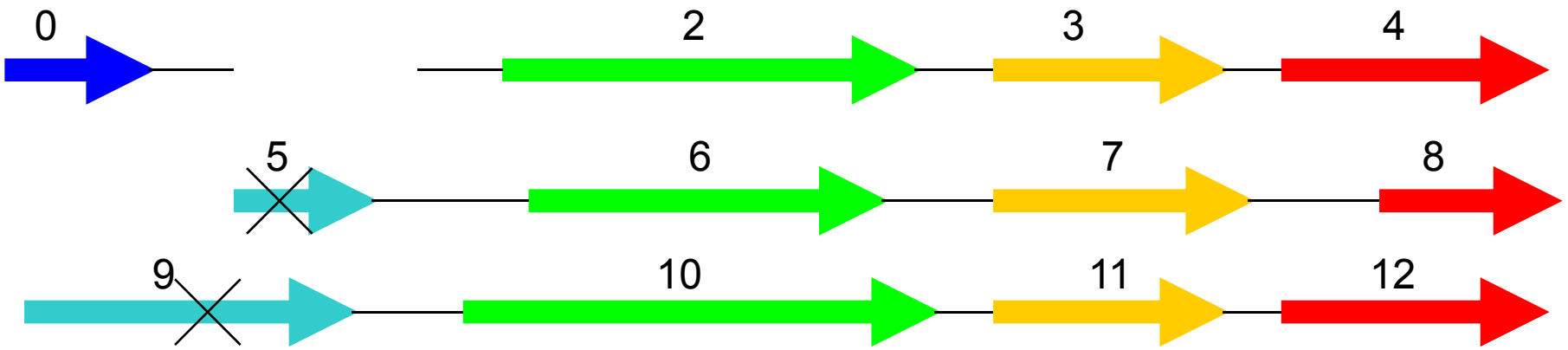
- Secondary Structure Pool



Monte Carlo Assembly Animation
Step 1 :: Score 0

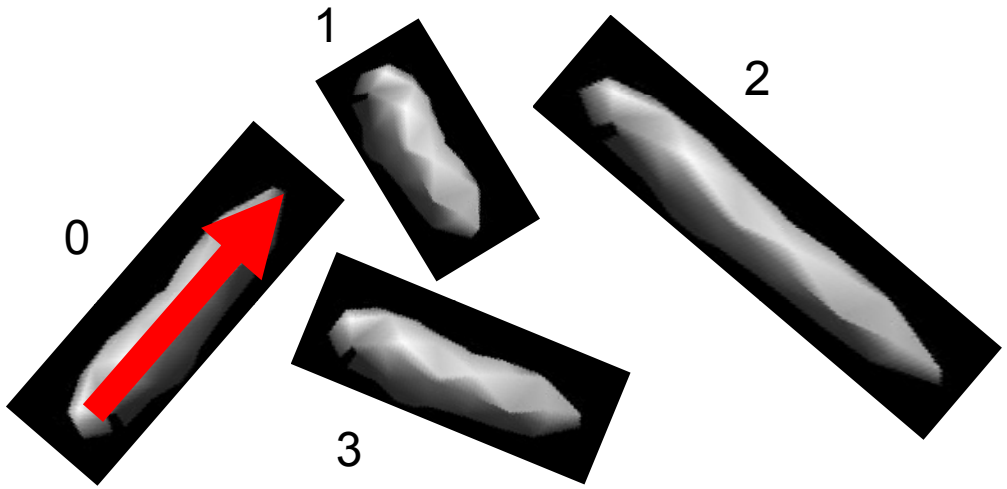
step	1
move	add
length check	✖
loop check	▬
move accepted ?	✖
score after move	0

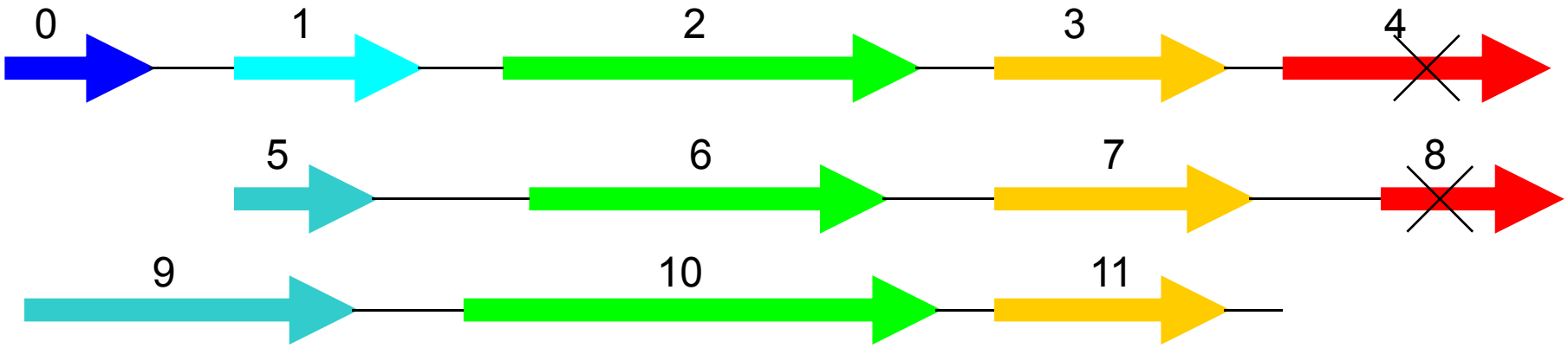




Monte Carlo Assembly Animation
Step 2 :: Score –0.4

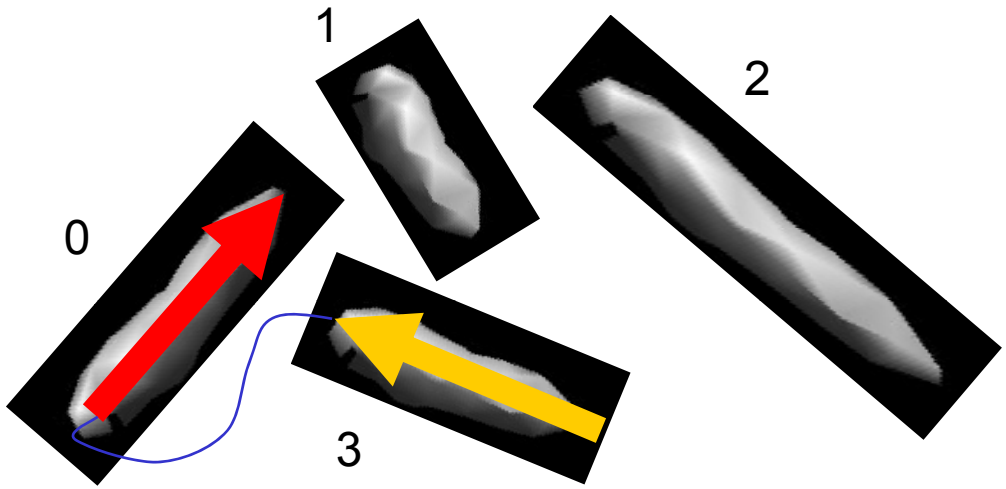
step	2
move	add
length check	✔
loop check	✔
move accepted ?	✔
score after move	- 0.4

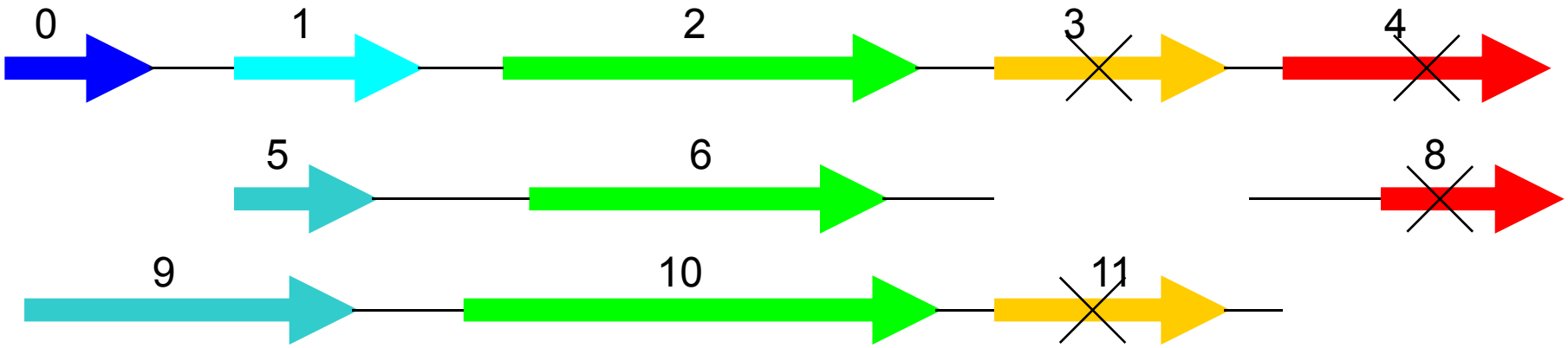




Monte Carlo Assembly Animation
Step 3 :: Score –0.7

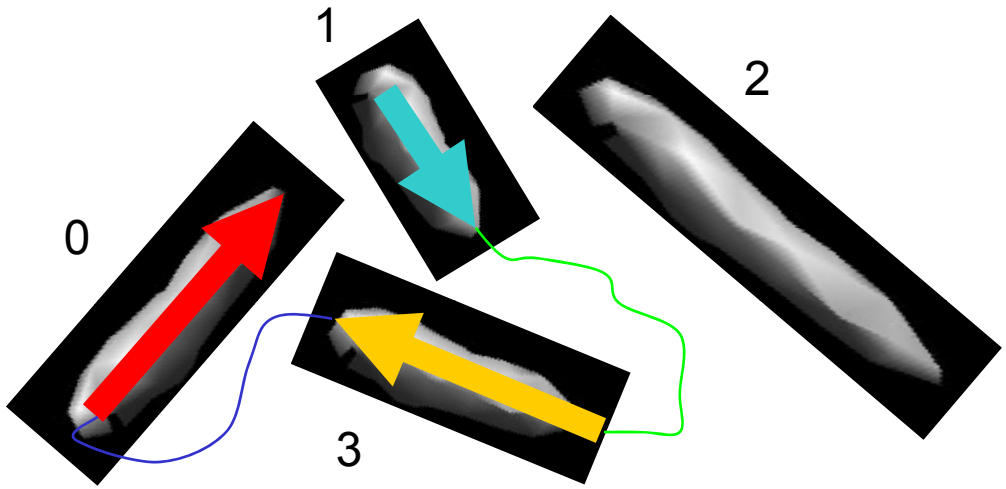
step	3
move	add
length check	✔
loop check	✔
move accepted ?	✔
score after move	- 0.7

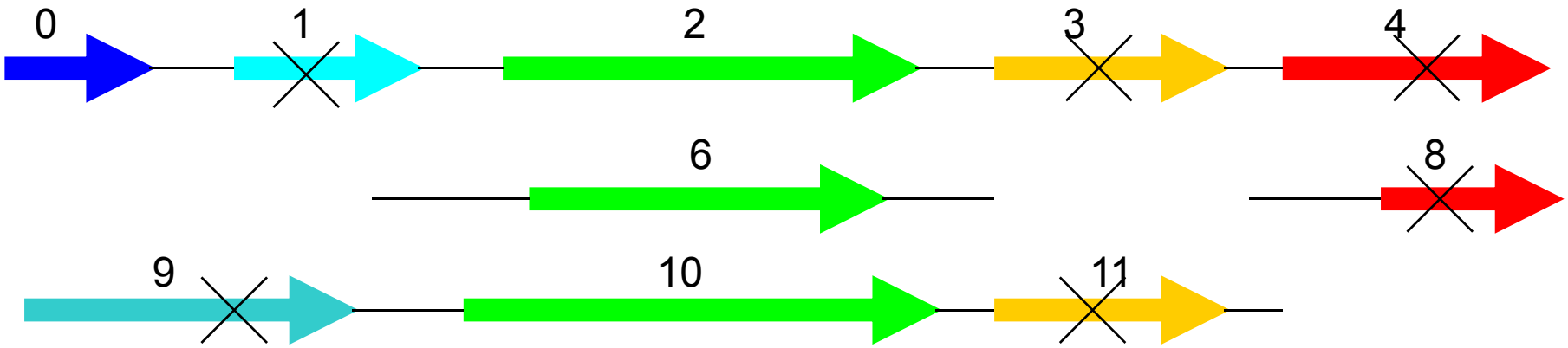




Monte Carlo Assembly Animation
Step 4 :: Score –1.5

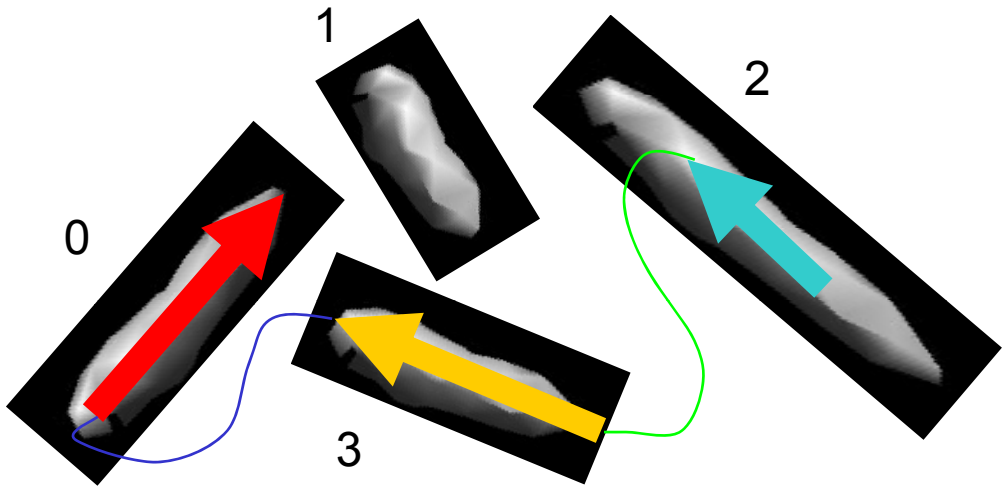
step	4
move	add
length check	✔
loop check	✔
move accepted ?	✔
score after move	- 1.5

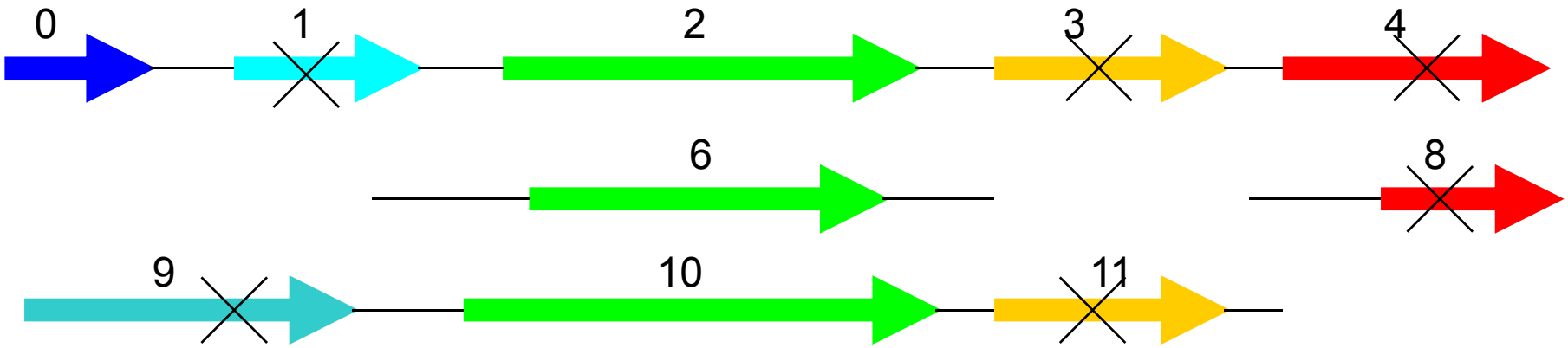




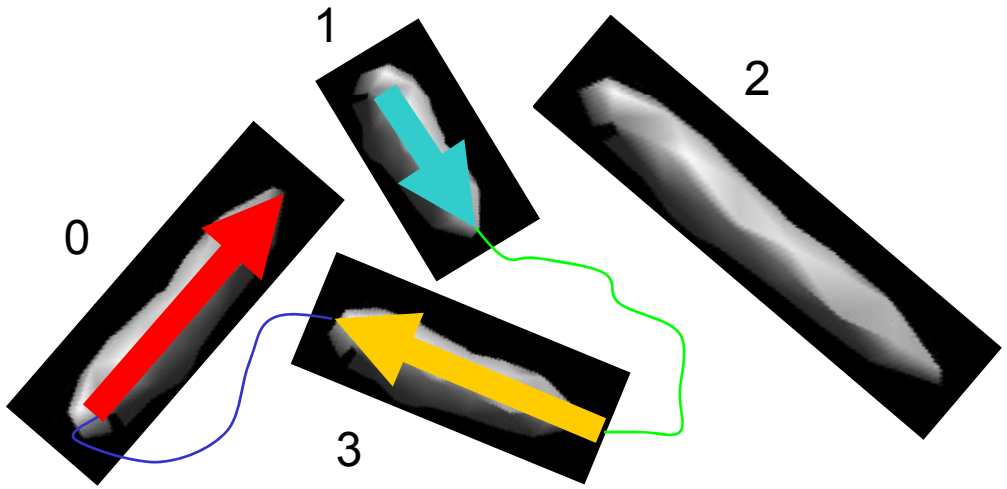
Monte Carlo Assembly Animation
Step 5 :: Score –1.5

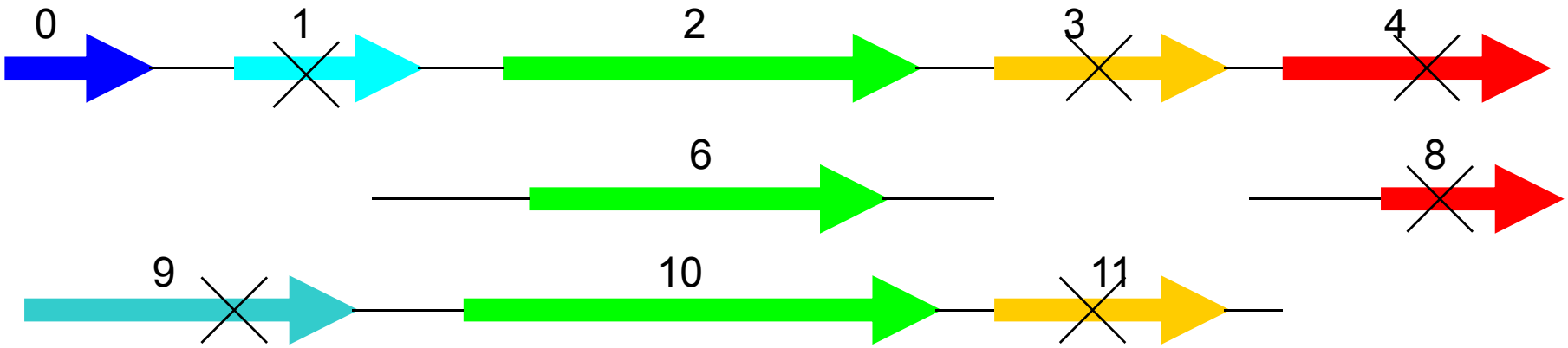
step	5
move	move
length check	✖
loop check	
move accepted ?	✖
score after move	- 1.5





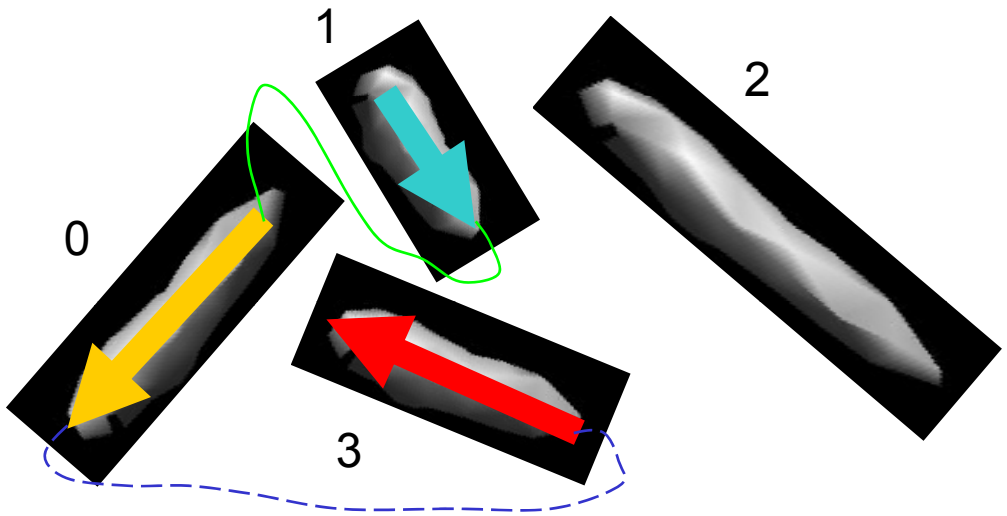
Monte Carlo Assembly Animation
Reject :: Score –1.5

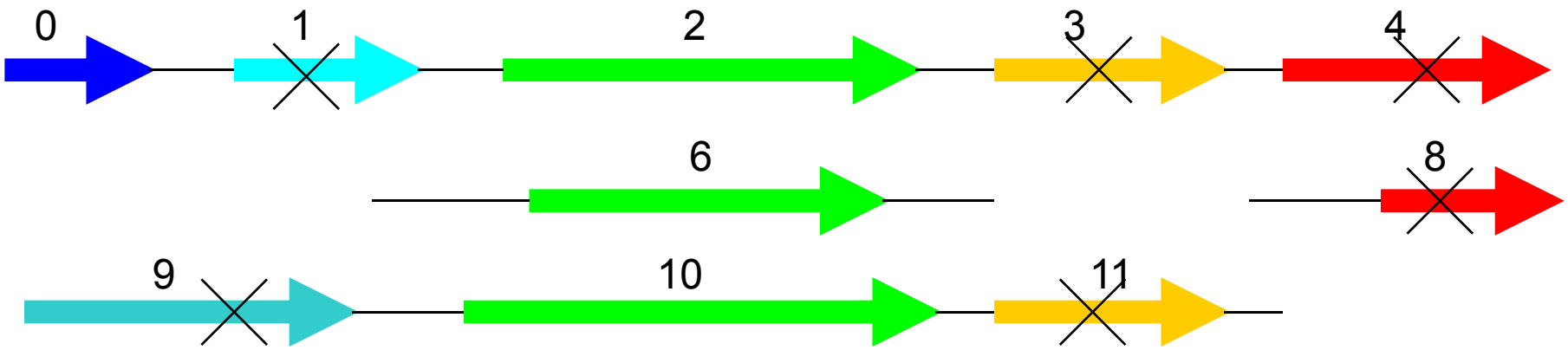




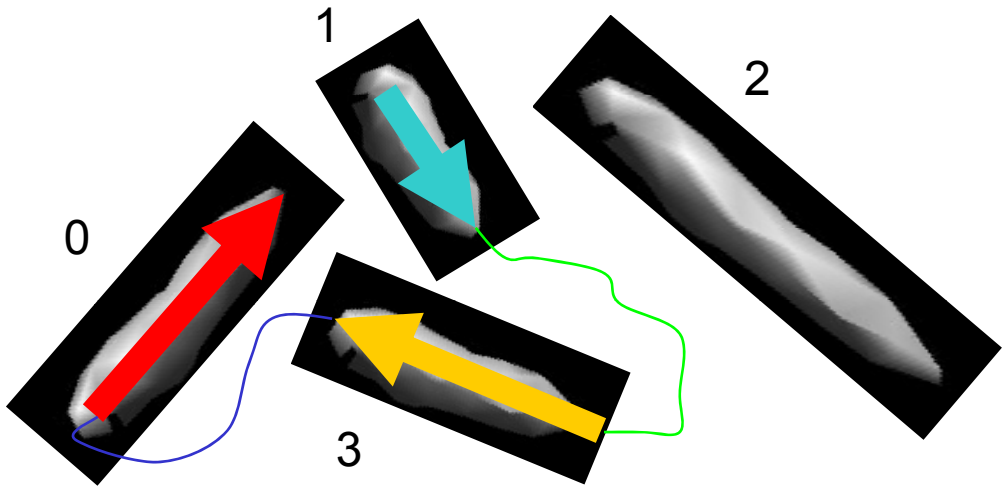
Monte Carlo Assembly Animation
Step 6 :: Score –1.5

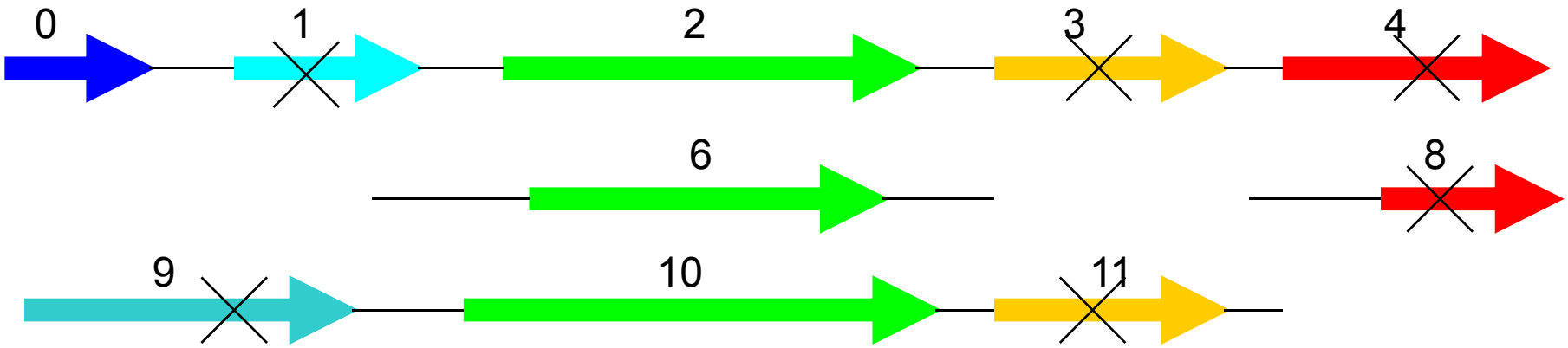
step	6
move	swap
length check	✔
loop check	✖
move accepted ?	✖
score after move	- 1.5





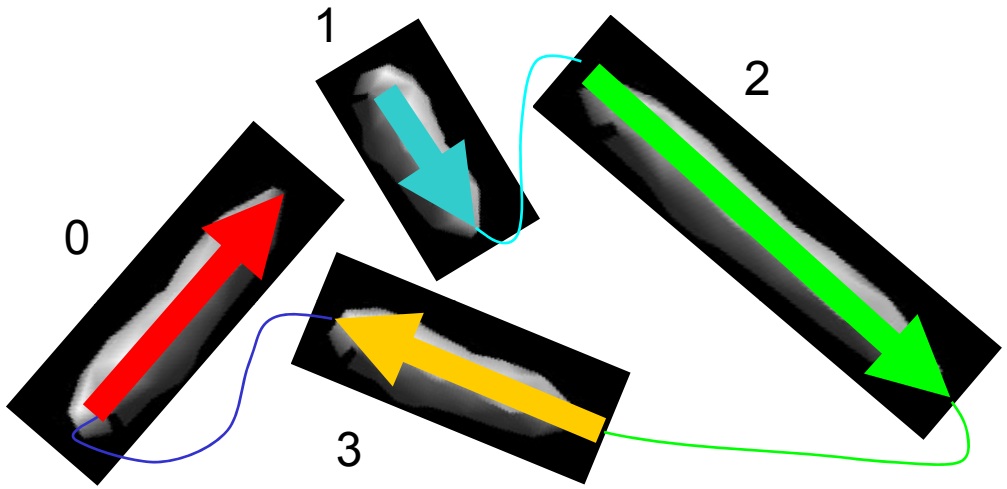
Monte Carlo Assembly Animation
Reject :: Score –1.5

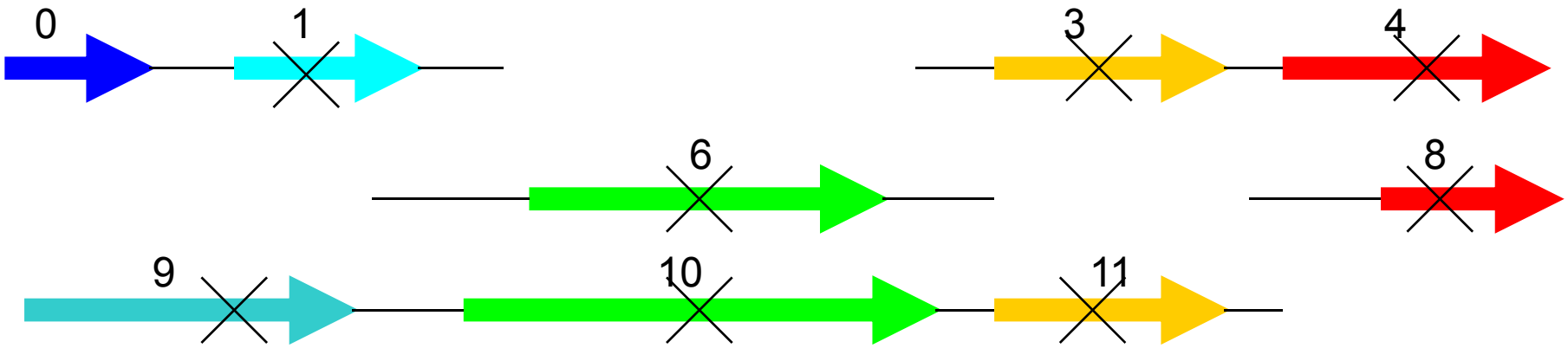




Monte Carlo Assembly Animation
Step 7 :: Score –2.5

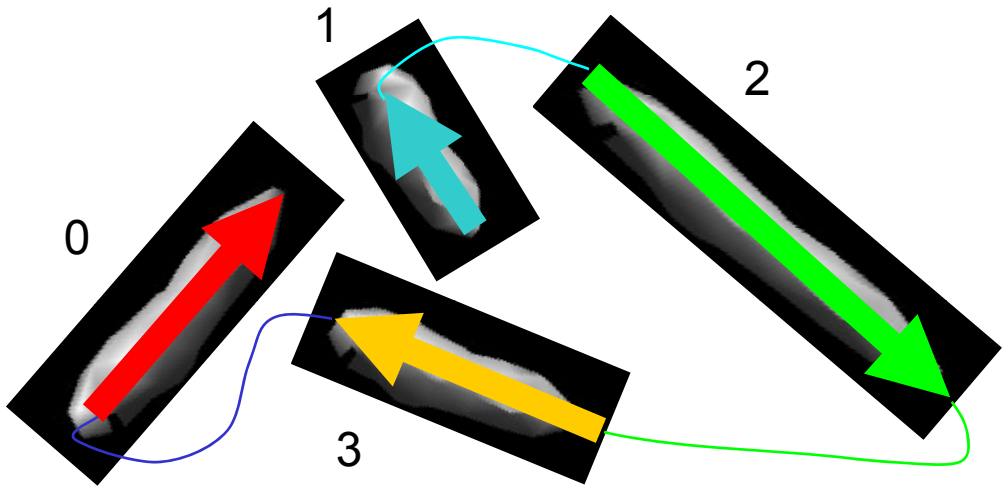
step	7
move	add
length check	✔
loop check	✔
move accepted ?	✔
score after move	- 2.5

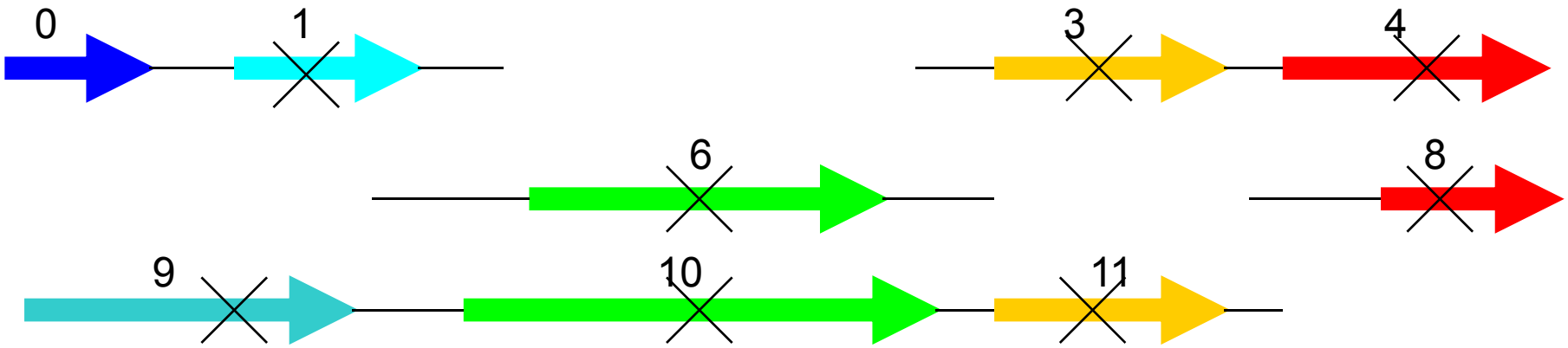




Monte Carlo Assembly Animation
Step 8 :: Score –2.6

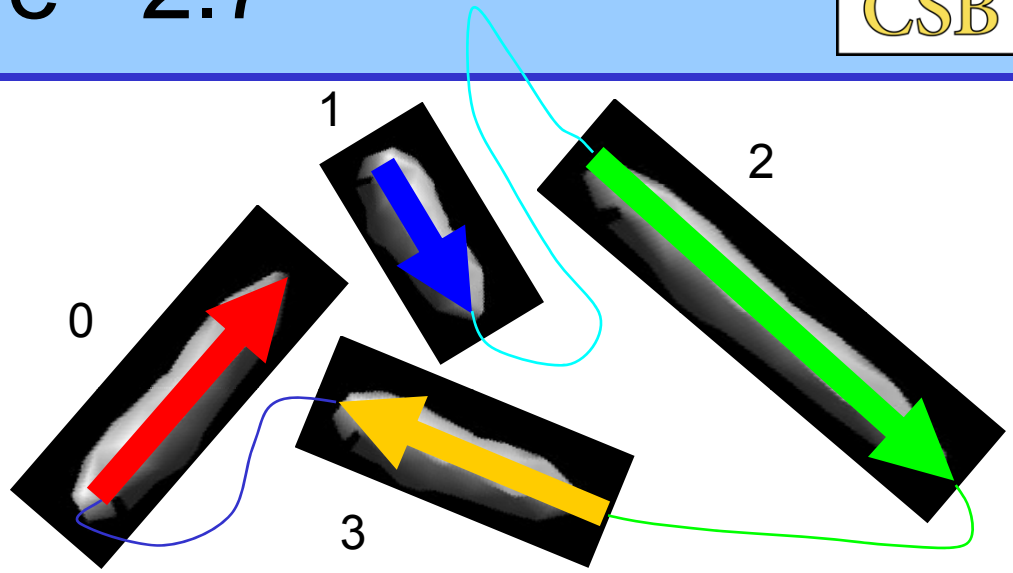
step	8
move	flip
length check	✔
loop check	✔
move accepted ?	✔
score after move	- 2.6

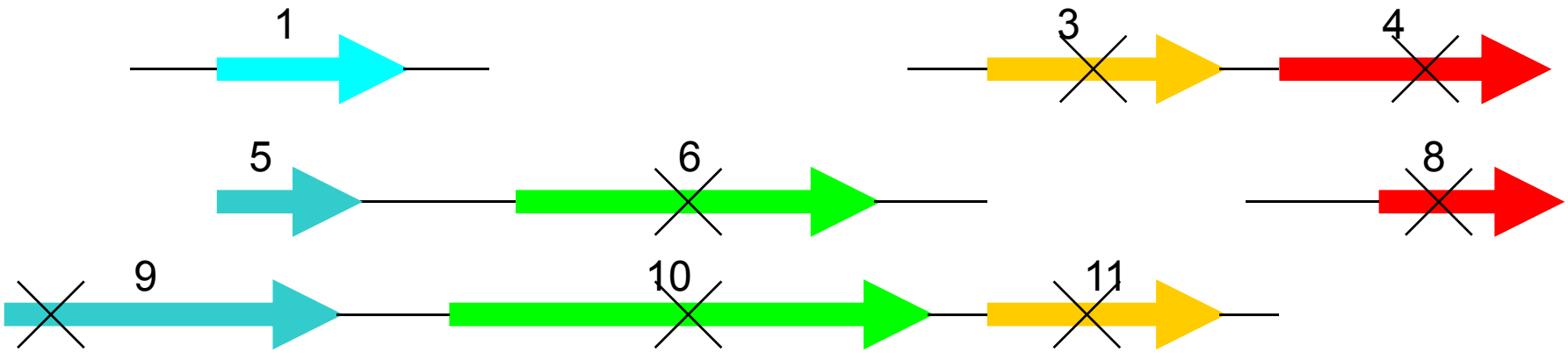


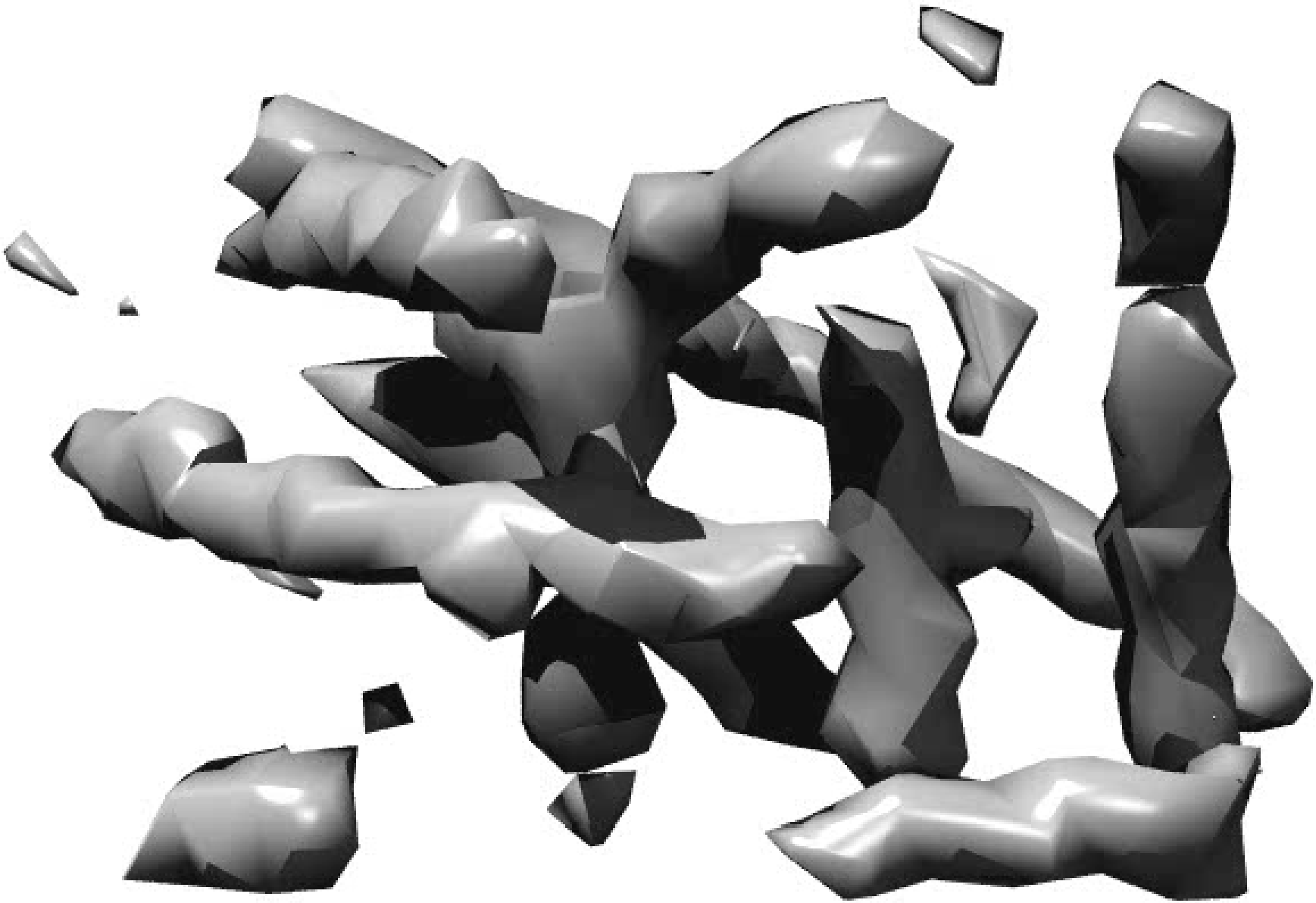


Monte Carlo Assembly Animation
Final Model :: Score –2.7

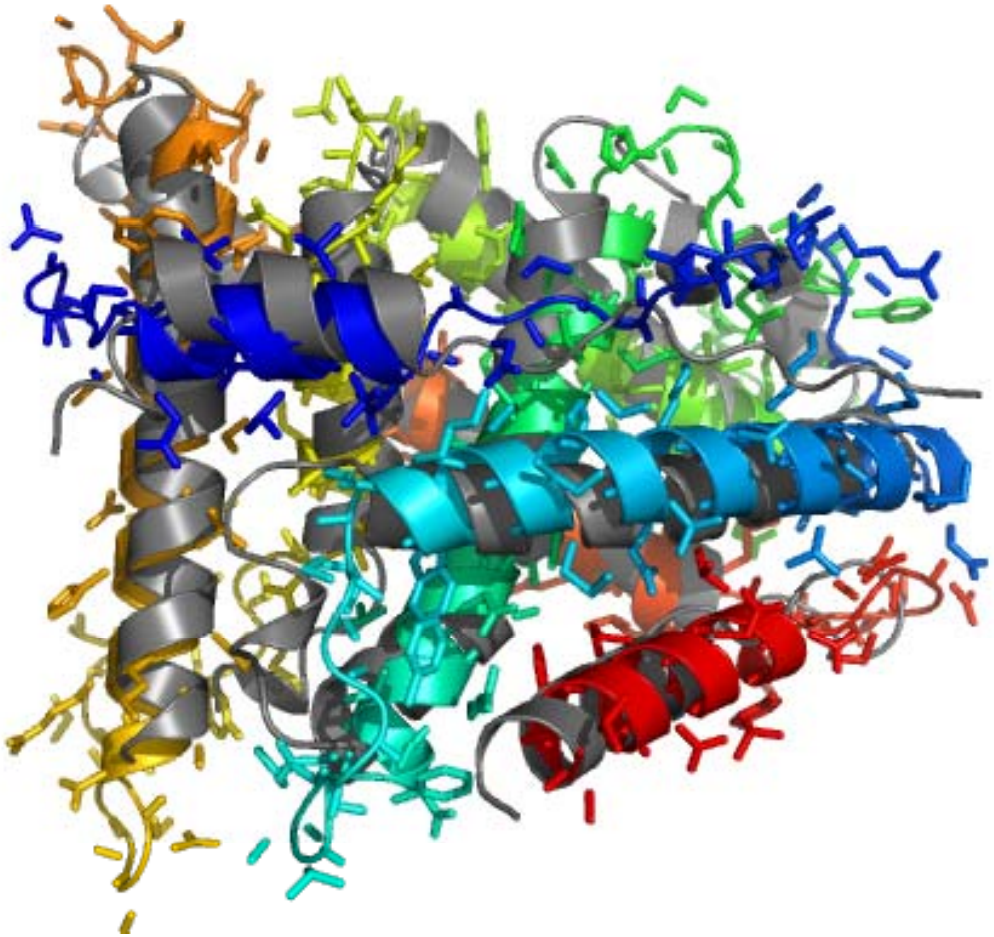
step	9
move	swap
length check	✔
loop check	✔
move accepted ?	✔
score after move	- 2.7







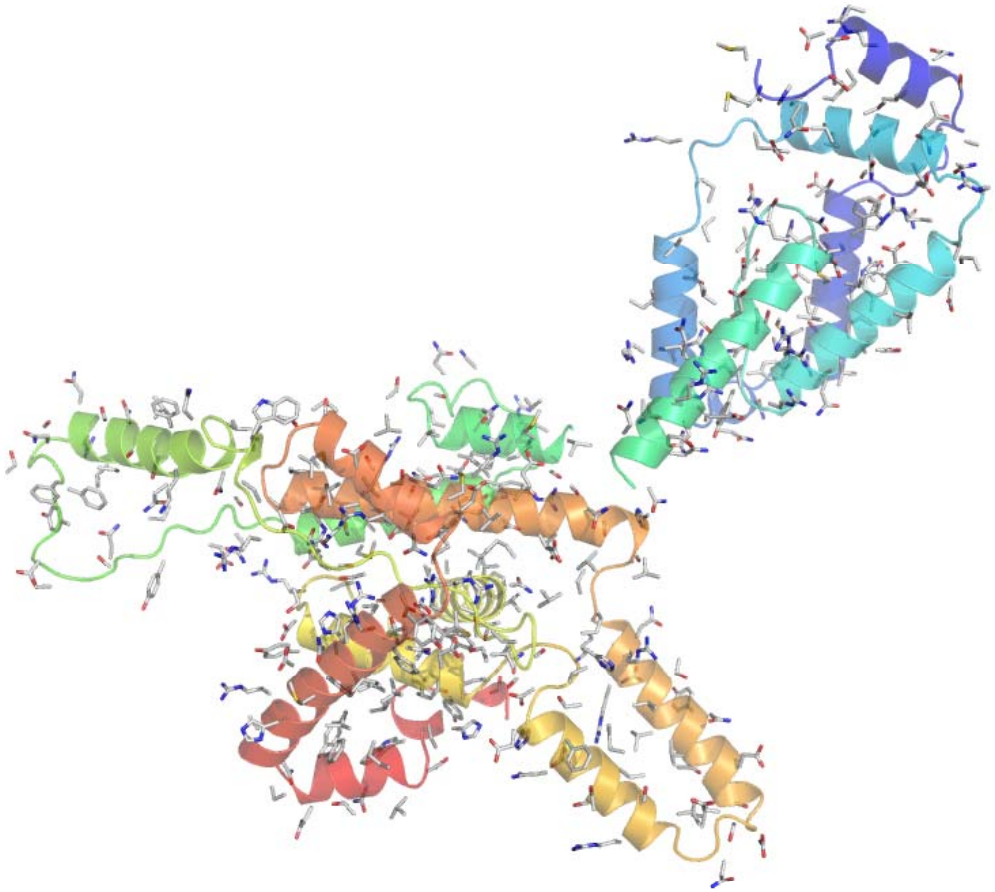
1QKM Loop Building, Side Chain Placement, and Refinement



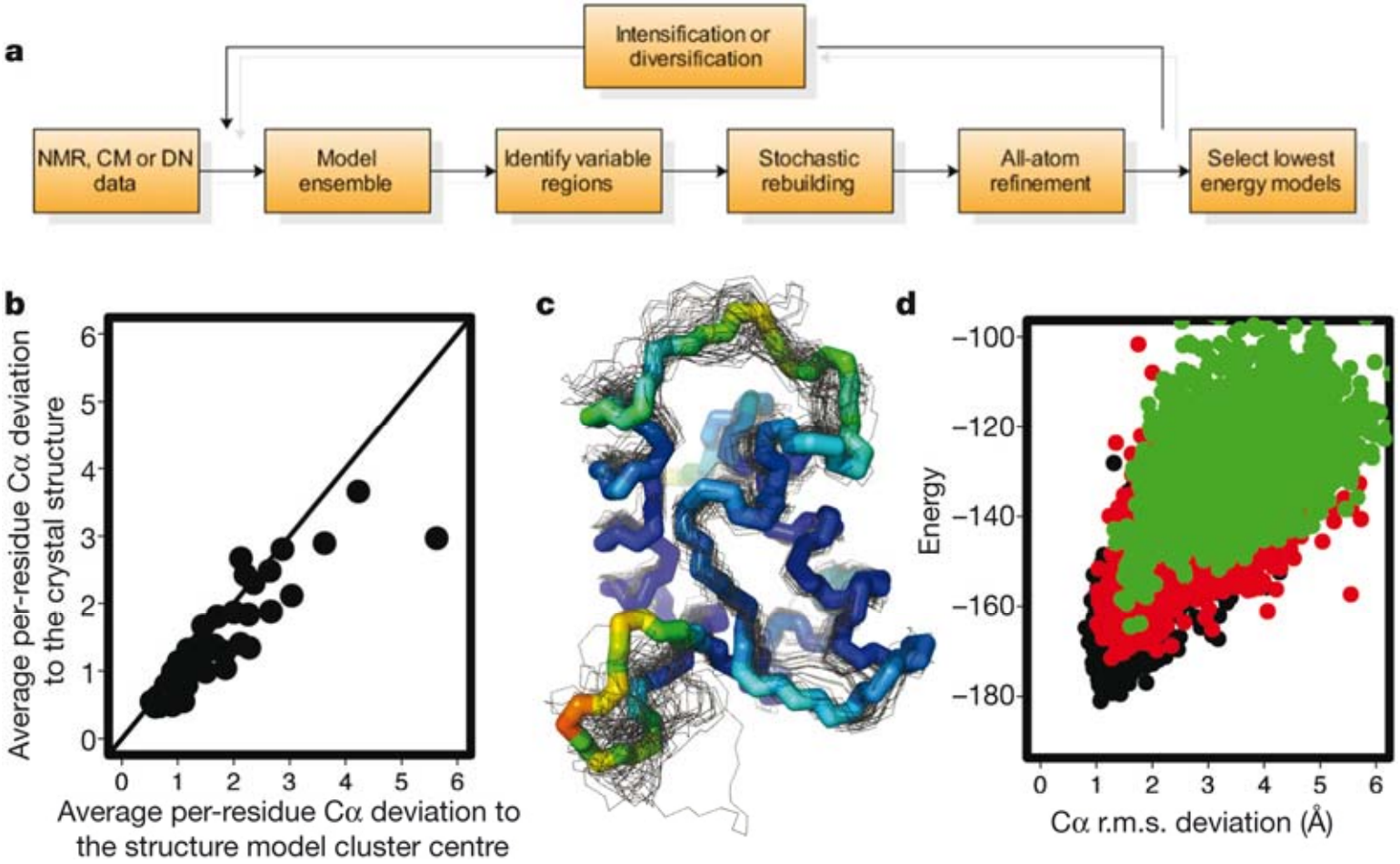
1QKM:

# residues:	**255**
# α-helices:	**8**
%α-helical:	**66**
Score Rank:	**1**
RMSD:	**3.88Å**

Preliminary Model For Adenovirus Protein IIIa



Overview of the Rebuilding-and-Refinement Method.



Qian, B.; Raman, S.; Das, R.; Bradley, P.; McCoy, A. J.; Read, R. J.; Baker, D. "High-resolution structure prediction and the crystallographic phase problem" *Nature* ***2007.***

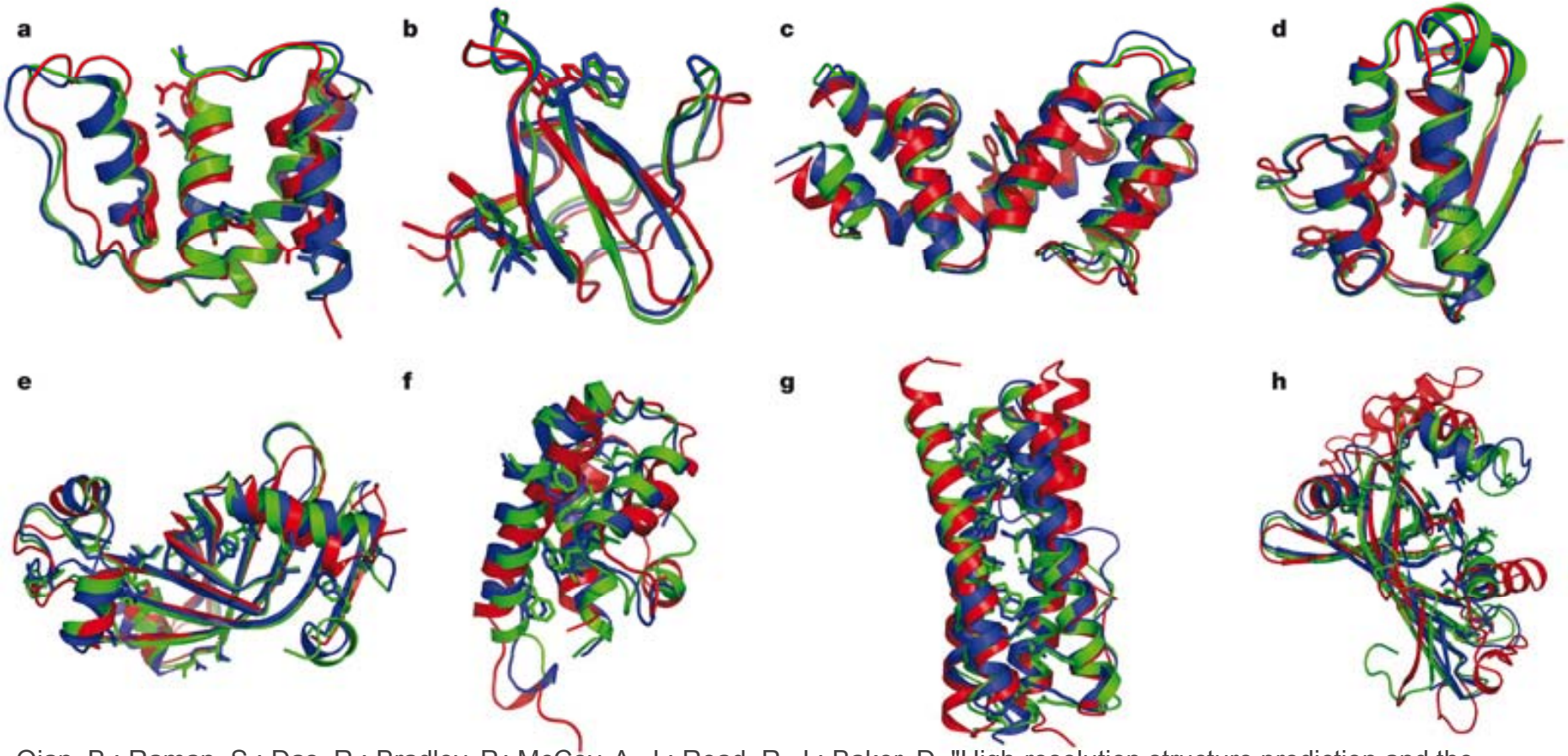
Improvement of model accuracy and molecular replacement

Table 1 | Improvement of model accuracy and molecular replacement by a rebuilding and refinement protocol

	X-ray structure	Starting model*	Length (*n*)†	Sequence identity to best template (%)‡	GDT-HA§		TFZ‖ in molecular replacement		Auto-traced residues (backbone, side chain)¶	
					Best template	Refined model	Best template	Refined model	Best template	Refined model
NMR	1hb6	2abd	86	N/A	0.58	0.79	4.1	11.3	12, 0	80, 80
	1who	1bmw	94	N/A	0.59	0.68	5.7	8.3	25, 12	47, 44
	1gnu	1kot	119	N/A	0.64	0.73	6.6	10.6	62, 53	82, 78
	1a19	1ab7	89 (2)	N/A	0.63	0.78	3.7	8.8	31, 20	48, 37
							4.5	12.5	14, 0	44, 35
	1fvk	1a24	189 (2)	N/A	0.49	0.69	3.4	6.9	66, 50	97, 91
							4.3	12.4	55, 43	85, 68
	1mzl	1afh	93	N/A	0.60	0.66	4.6	5.1	36, 29	58, 44
	1tvg	1xpw	143	N/A	0.63	0.74	4.3	6.7	15, 6	103, 86
	2snm	2sob	97	N/A	0.45	0.48	3.8	4.8	17, 16	43, 37
	1agr	1ezy	129	N/A	0.49	0.76	N/A#		N/A#	
	1abq	1awo	56	N/A	0.58	0.83	N/A☆		N/A☆	
CM	2hhz (T0331)	1ty9A	149	14.5	0.49	0.58	5.4	8.8	28, 24	68, 63
	2hr2 (T0368)	2c21C	158 (6)	14.8	0.57	0.67	6.0	5.4	37, 37	20, 14
	2hq7 (T0380)	2fhqA	145 (2)	25.4	0.58	0.69	4.4	6.6	47, 23	92, 83
							4.6	14.2	30, 17	60, 59
	2ib0 (T0385)	1jgcB	170 (2)	7.8	0.62	0.69	5.1	7.9	63, 37	56, 56
							5.8	15.5	50, 2	52, 52
	2hi0 (T0329_D2)	1rqlA	92 (2)	8.8	0.52	0.67	N/A#		N/A#	
	2hcf (T0330_D2)	1lvhB	75	14.1	0.51	0.65	N/A#		N/A#	
	2hi6 (T0357)**	1aco	132	8.4	0.45	0.52	N/A**		N/A**	
DN	2hh6 (T0283)	2b2j	112	3.6	0.22	0.64	5.4	9.0	26, 12	112, 112

Qian, B.; Raman, S.; Das, R.; Bradley, P.; McCoy, A. J.; Read, R. J.; Baker, D. "High-resolution structure prediction and the crystallographic phase problem" *Nature* ***2007.***

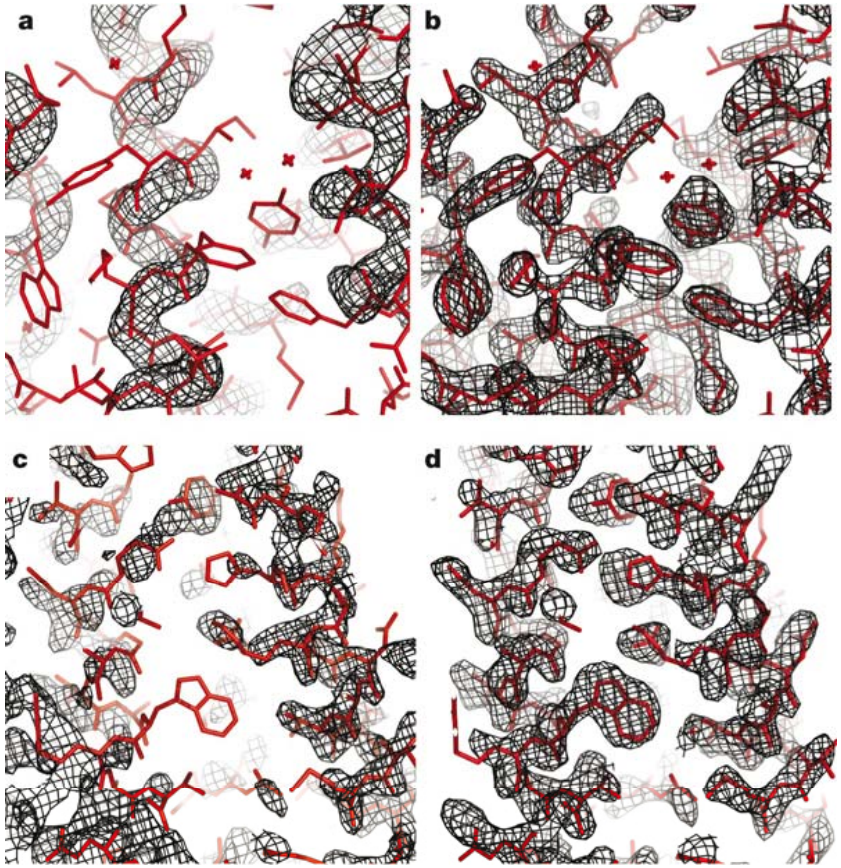
Refinement of NMR Structures (a-d) and Comparative Models (e-h)



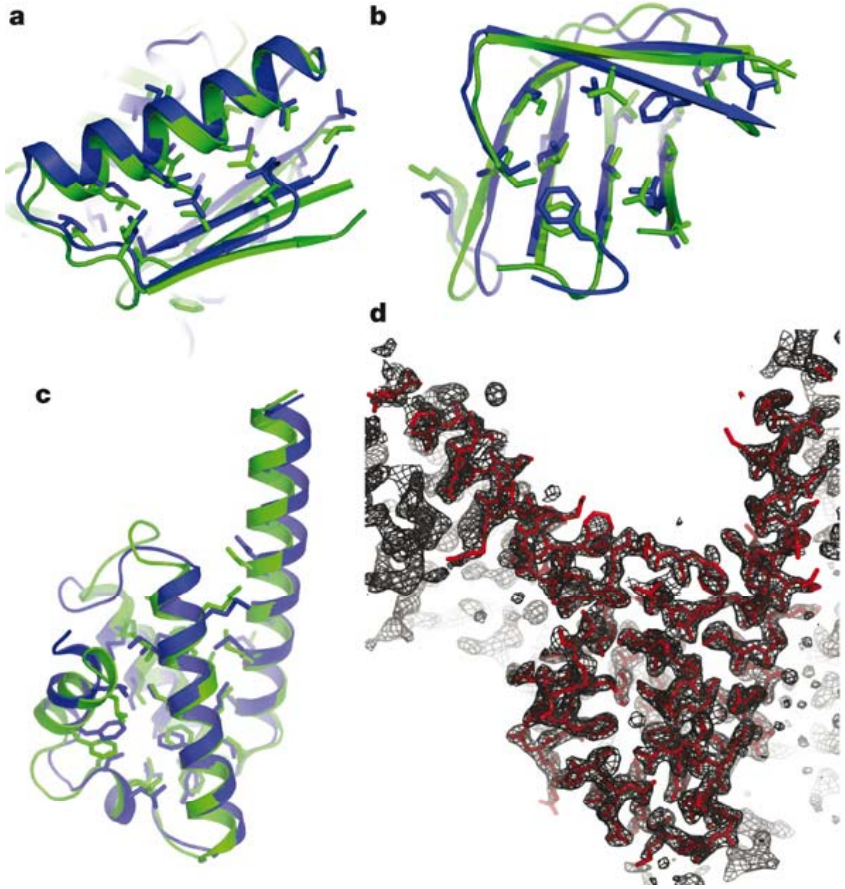
Qian, B.; Raman, S.; Das, R.; Bradley, P.; McCoy, A. J.; Read, R. J.; Baker, D. "High-resolution structure prediction and the crystallographic phase problem" *Nature* ***2007.***
native crystal structure (blue), template/NMR structure (red), and the refined model (green)

Phasing by refined NMR structures, comparative and *de novo* models

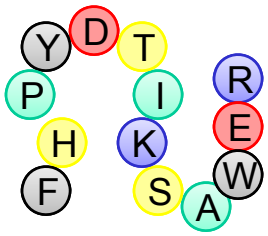
- NMR (a⇨b) / CM (c⇨d)



- *de novo* model (c⇨d)



(Inverse) Protein Folding Problem
Holy Grail of Comp. Struct. Biology



Protein Folding Problem

Sequence ⇒ Structure

Inverse Folding Problem

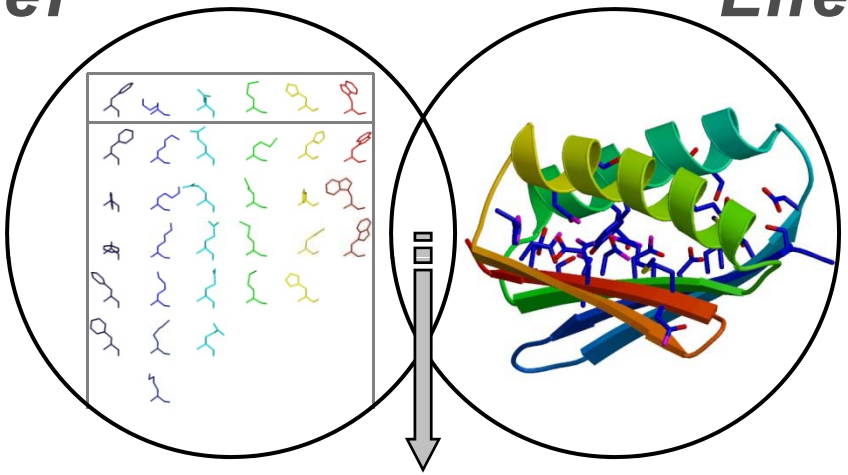
- Given a protein's AA sequence, what is its 3-dimensional fold , and how does it get there?
- Assume 100 conformations for each amino acid in a 100 amino acid protein ⇨ 10^{200} possible conformations!
- Cyrus Levinthal's paradox of protein folding,1968.

- Given a protein fold, which primary sequence(s) fold into it?
- Assume a total of 100 conformations for all 20 natural occurring amino acids side chains in a 100 amino acid protein ⇨ 10^{200} possible conformations!
- Earth is less than 10^{10} years old.

Sampling and Scoring for Side Chain Repacking and Design

Local Rotamer Bias

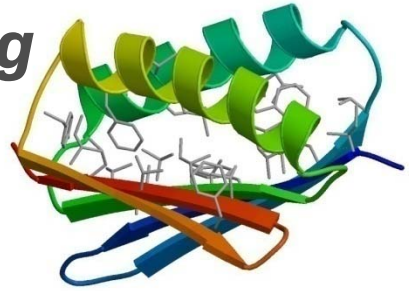
Approximate interactions within sidechain using the distribution of sidechain confor-mations (rotamers) seen in known protein structures



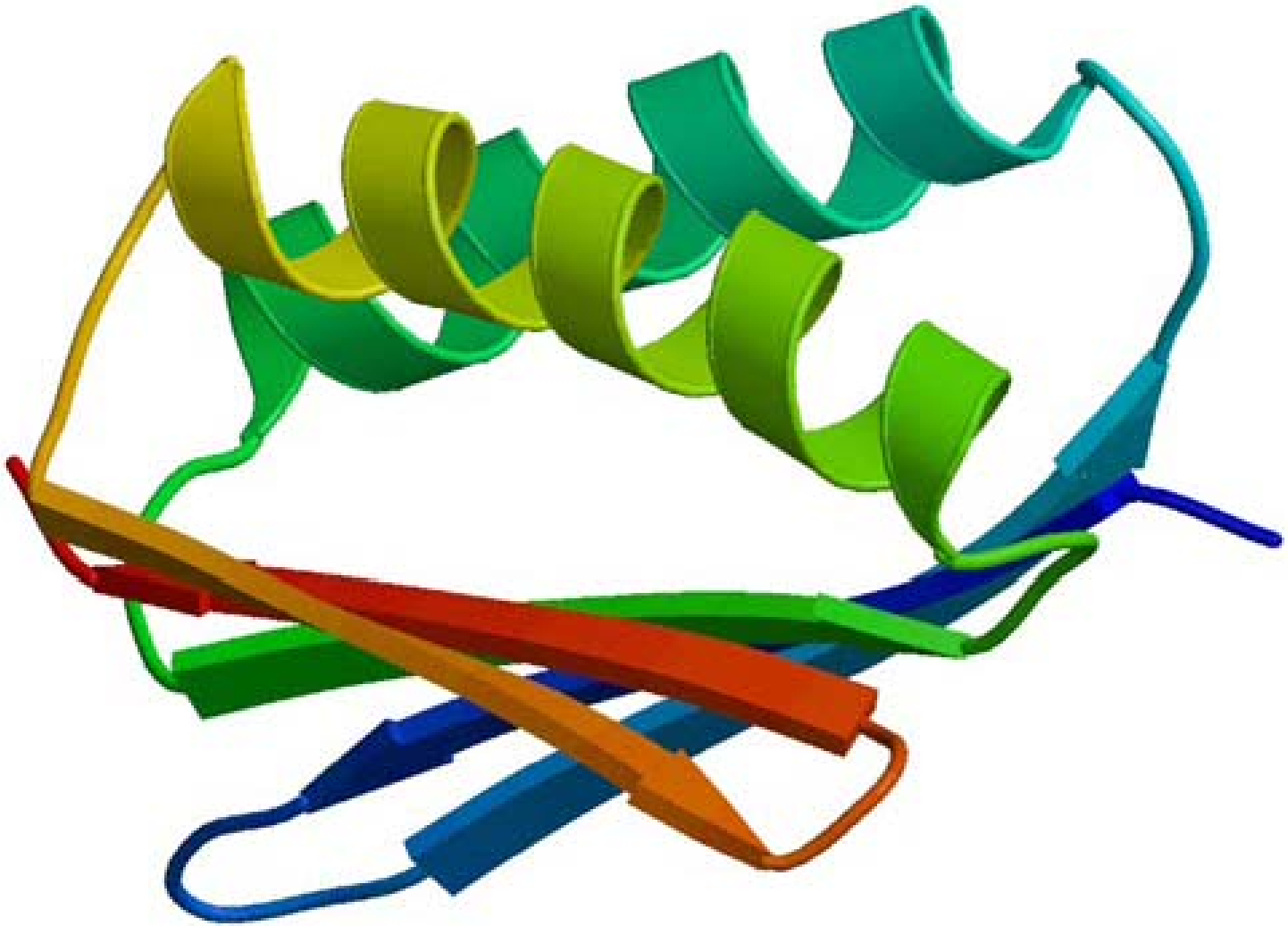
Energy function

Statistically derived potential function
- ***VDW interaction***
- ***solvation***
- ***hydrogen bonding potential***
- ***pair wise interactions***
- ***rotamer probability***

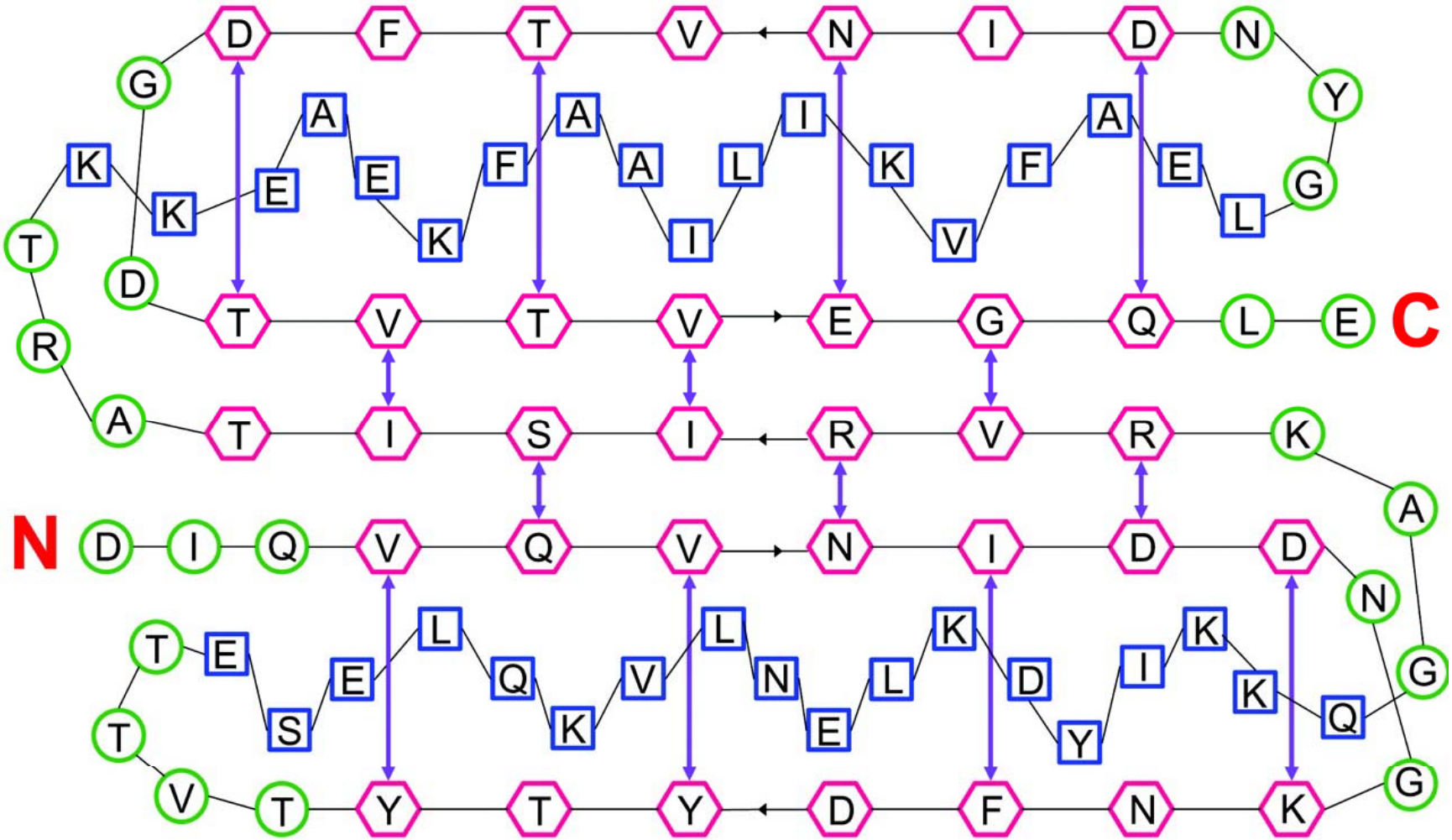
Simulated Annealing Monte Carlo energy minimization



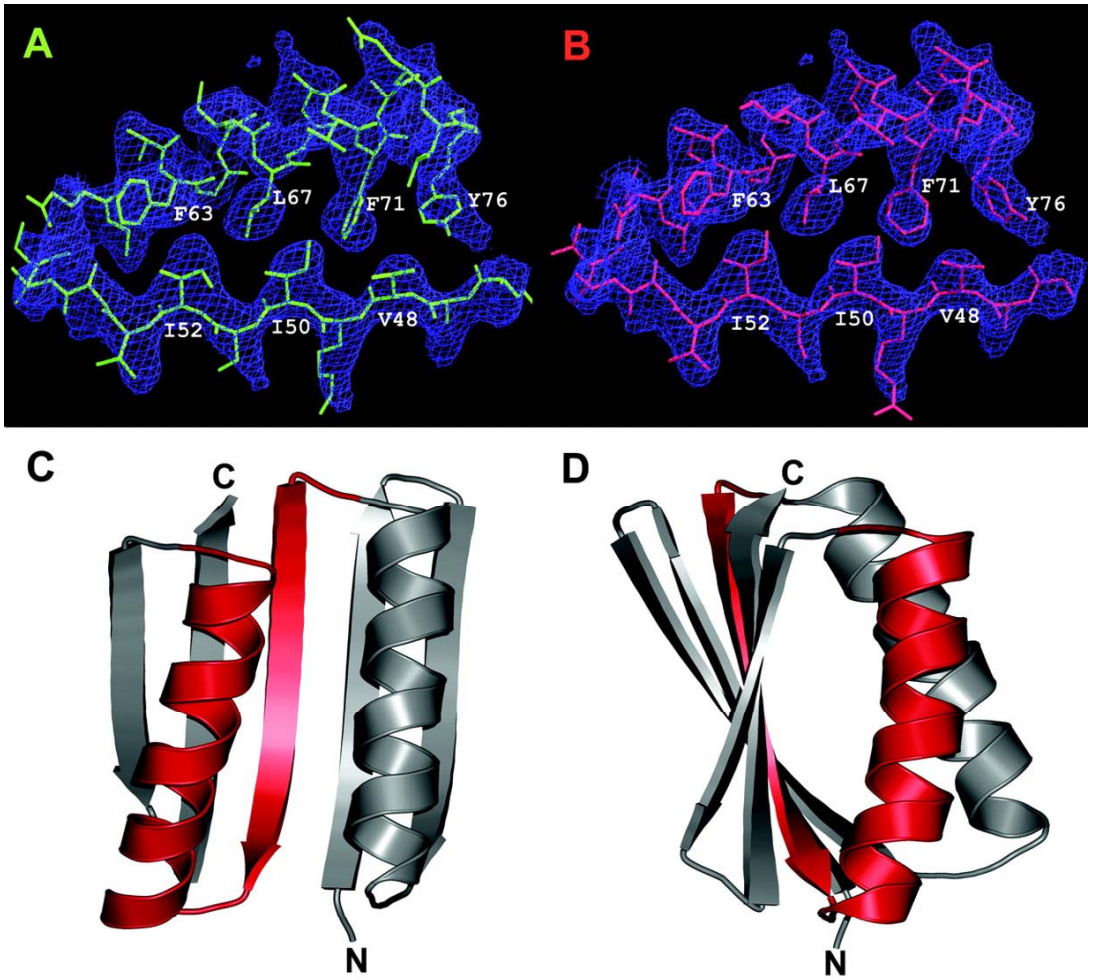
Dahiyat, B. I. and Mayo, S. L. (1997) *Science*, 278, 82-7
Dunbrack, R. L., Jr. and Karplus, M. (1993) *J Mol Biol*, 230, 543-74.
Kuhlman, B., et. al. (2003) *Science*, 302, 1364-1368.



A two-dimensional schematic of the target fold

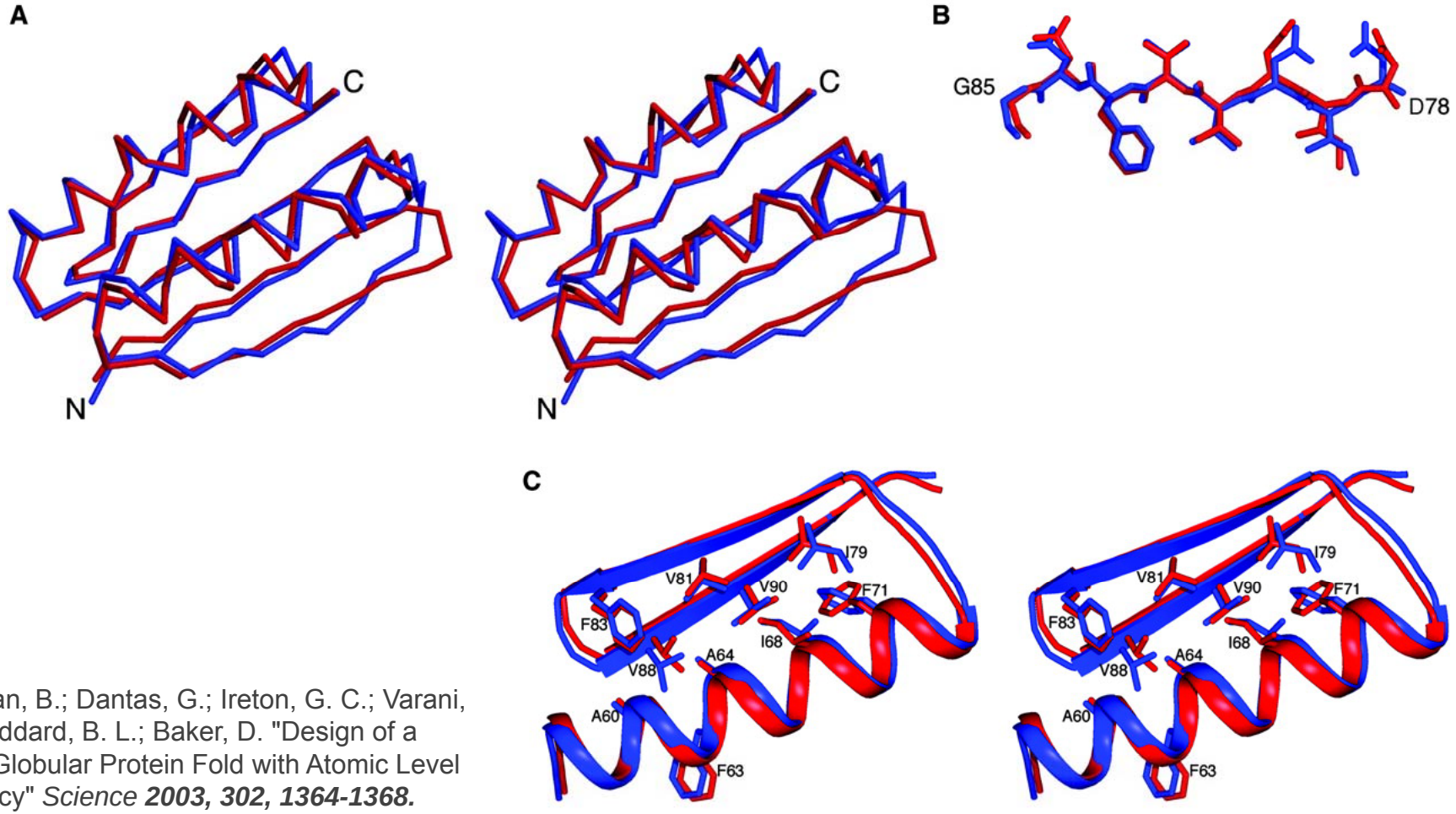


Schematic representation of Top7 residues 46-76



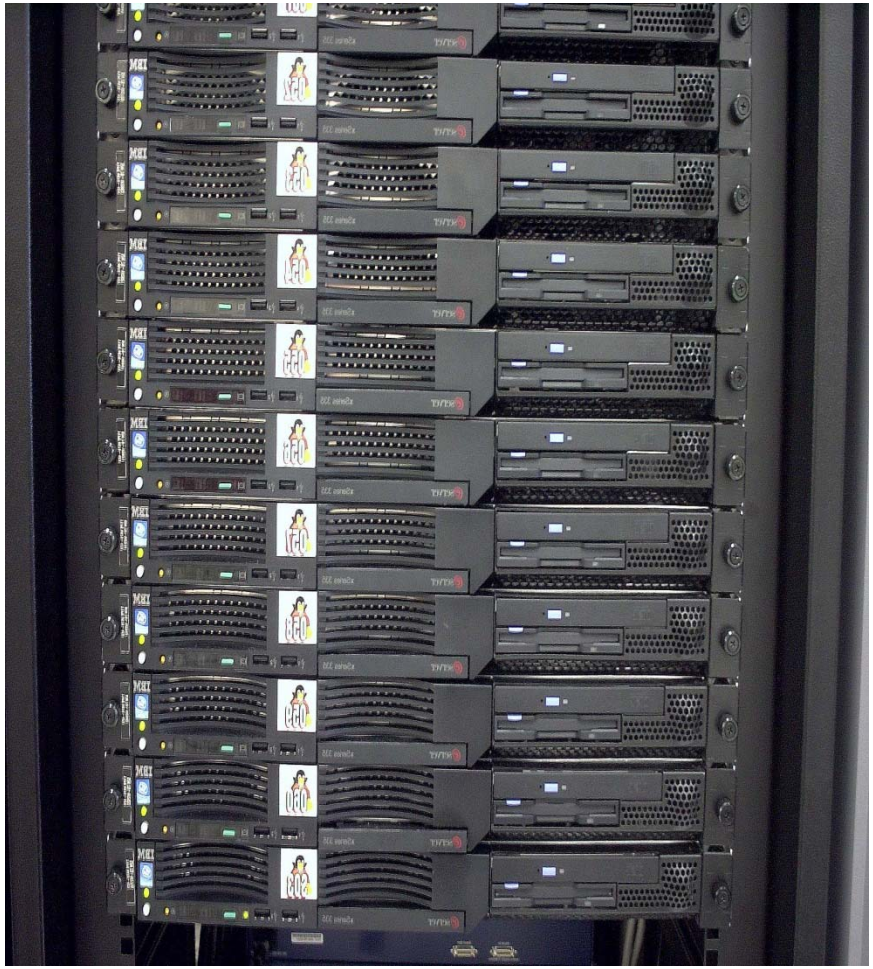
Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. "Design of a Novel Globular Protein Fold with Atomic Level Accuracy" *Science* ***2003, 302, 1364-1368.***

Designed Model (blue) and X-ray Structure (red) of Top7



Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. "Design of a Novel Globular Protein Fold with Atomic Level Accuracy" *Science* ***2003, 302, 1364-1368.***

The ACCRE Cluster – 1500 Processors at Your Service



Protein Folding is …
… KINDERLEICHT

- Says Jonas (3 months)