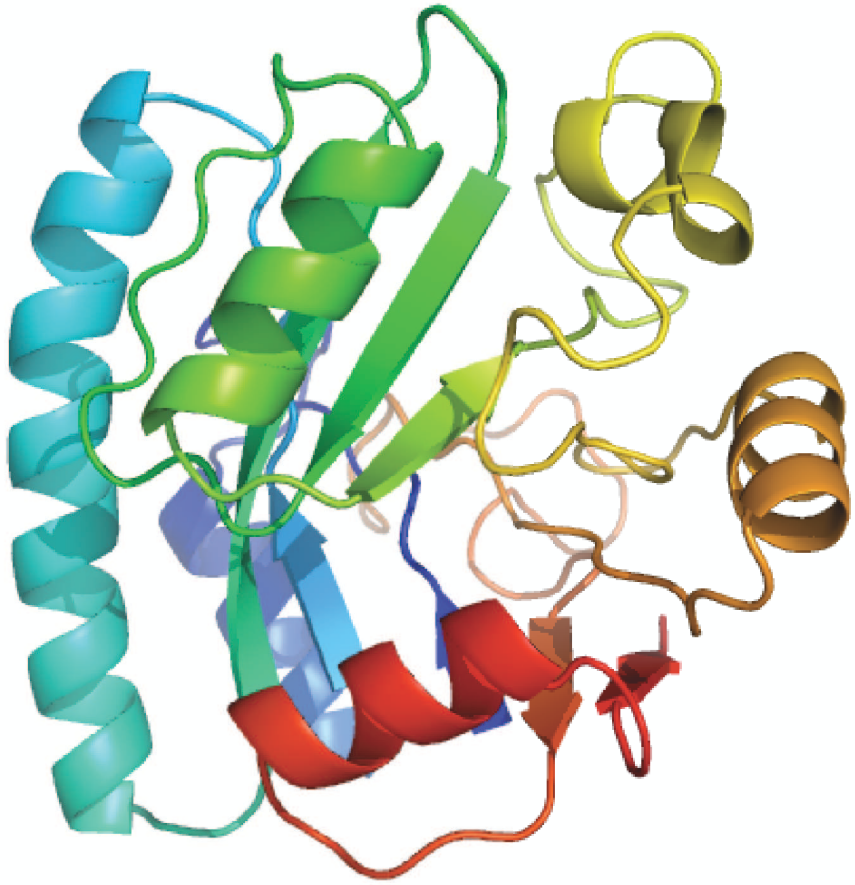
High Resolution Protein Refinement and the X-ray Crystallographic Phase Problem

Vatsan Raman
Baker Lab
University of Washington, Seattle

July 22nd 2008 - ROSETTACON

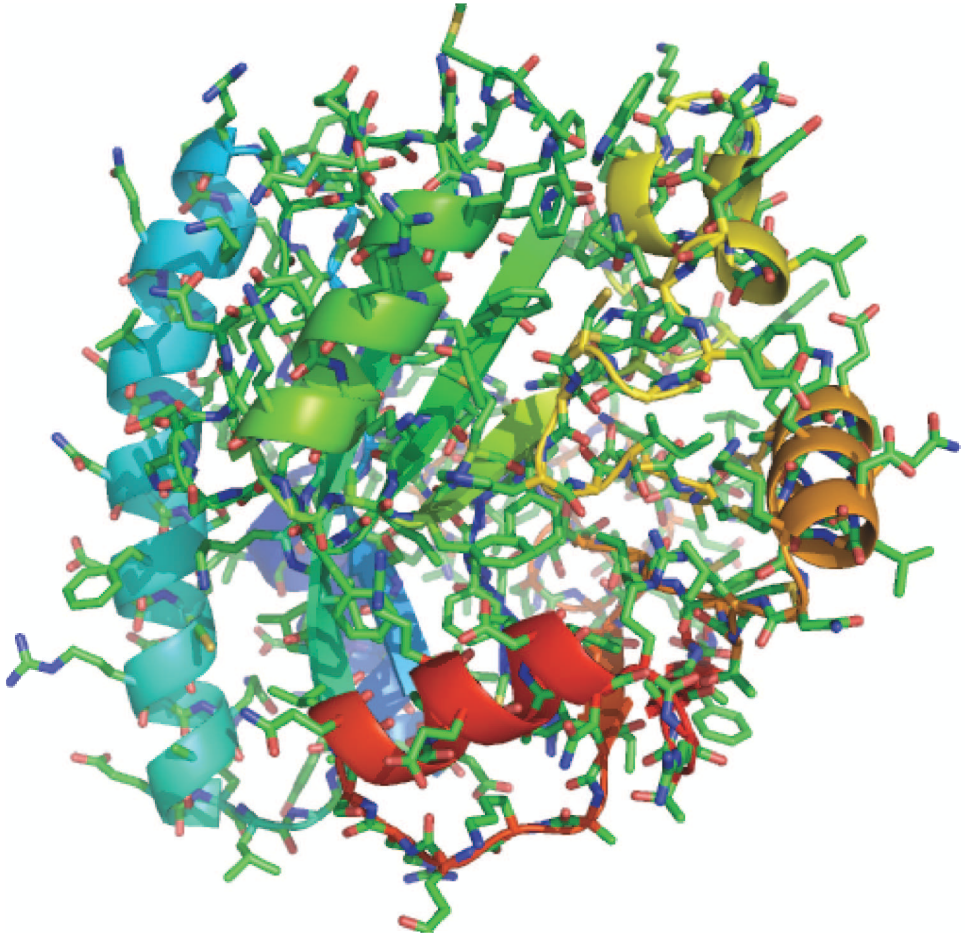
The Protein Folding Problem

GTPDIIVNAQINS
EDENVLDFIIEDE
YYLKKRGVGAHII
KVASSPQLRLLY
KNAYSTVSCGNY
GVLCNLVQNGEY
DLNAIMFNCAEIK
LNKGQMLFQTKI
WR

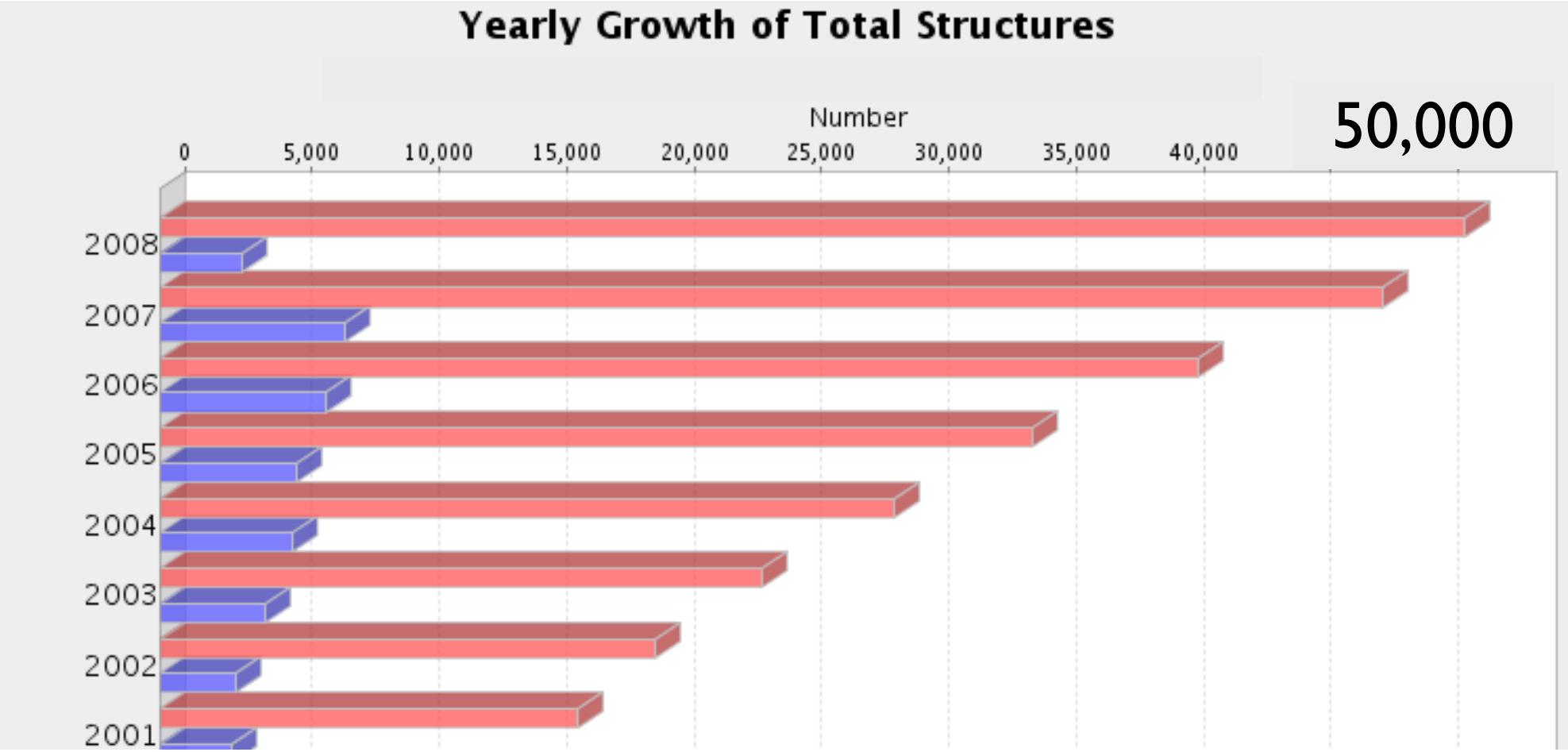


The Protein Folding Problem

GTPDIIVNAQINS
EDENVLDFIIEDE
YYLKKRGVGAHII
KVASSPQLRLLY
KNAYSTVSCGNY
GVLCNLVQNGEY
DLNAIMFNCAEIK
LNKGQMLFQTKI
WR



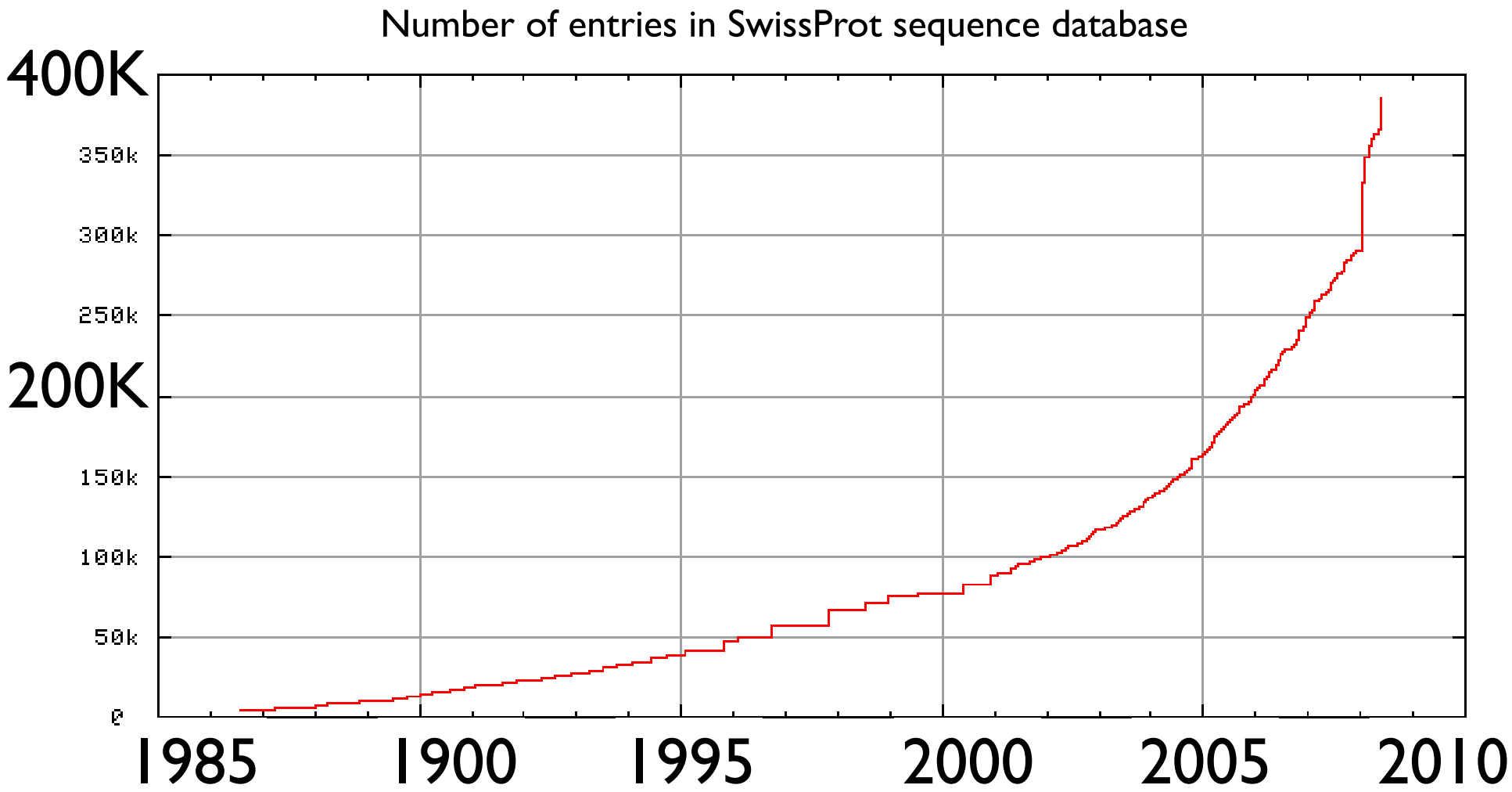
Protein sequences vs Protein structures



~50,000 known structures in PDB

RCSB Protein Data Bank - May 2008

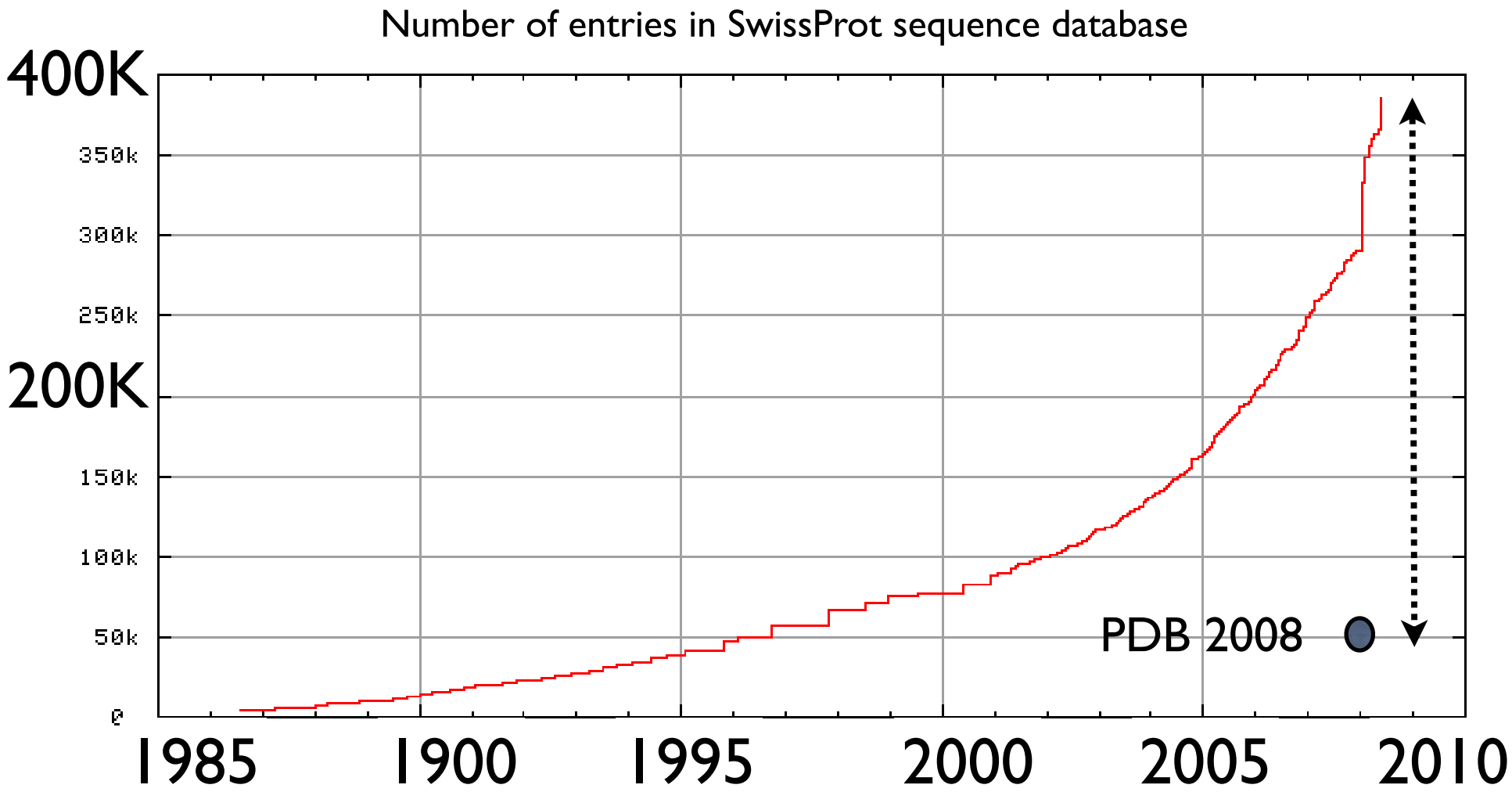
Protein sequences vs Protein structures



~400,000 known protein sequences

UniProtKB/Swiss-Prot protein sequence database - May 2008

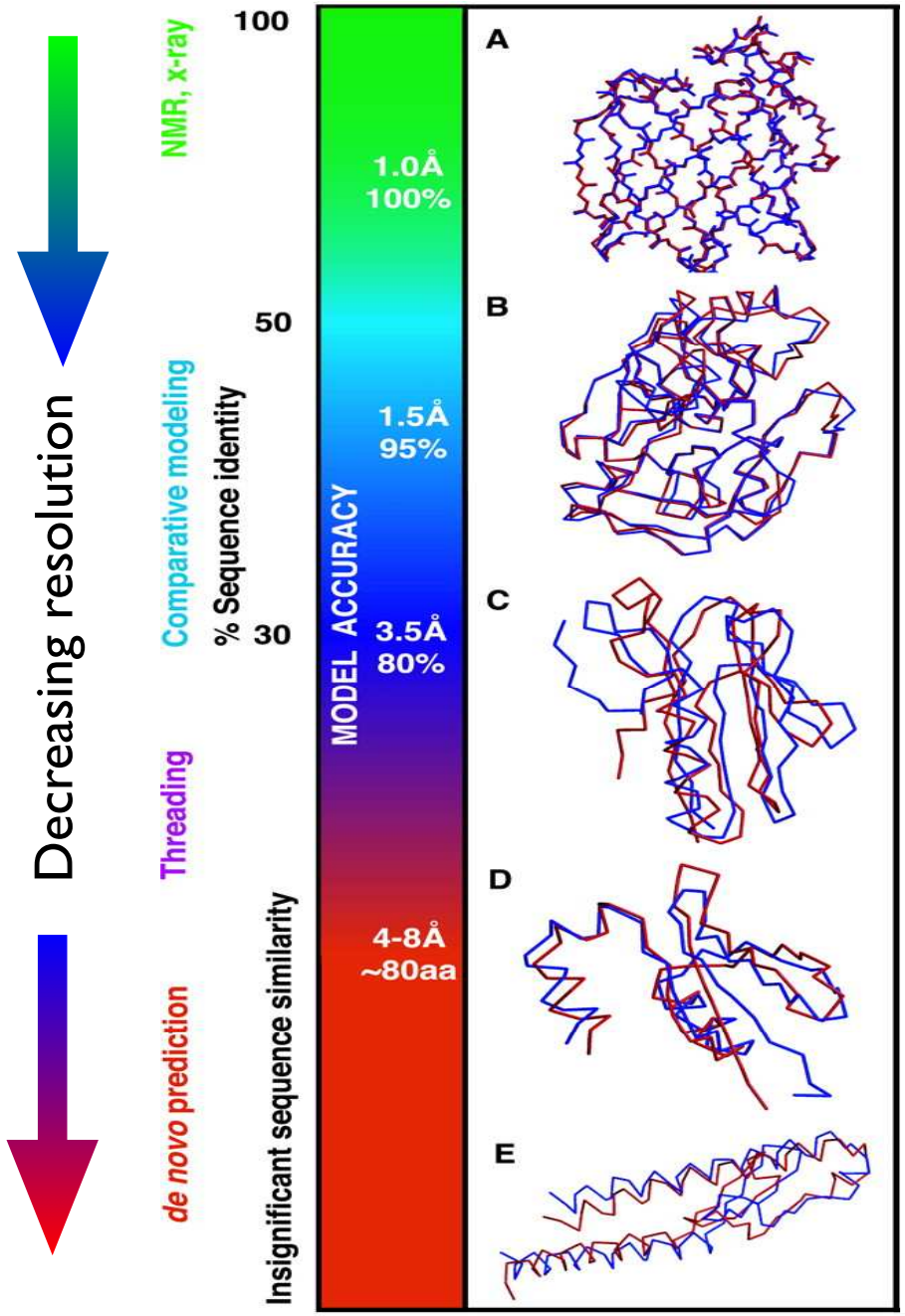
Protein sequences vs Protein structures



~400,000 known protein sequences

UniProtKB/Swiss-Prot protein sequence database - May 2008

Protein Structure Prediction Accuracy and Resolution

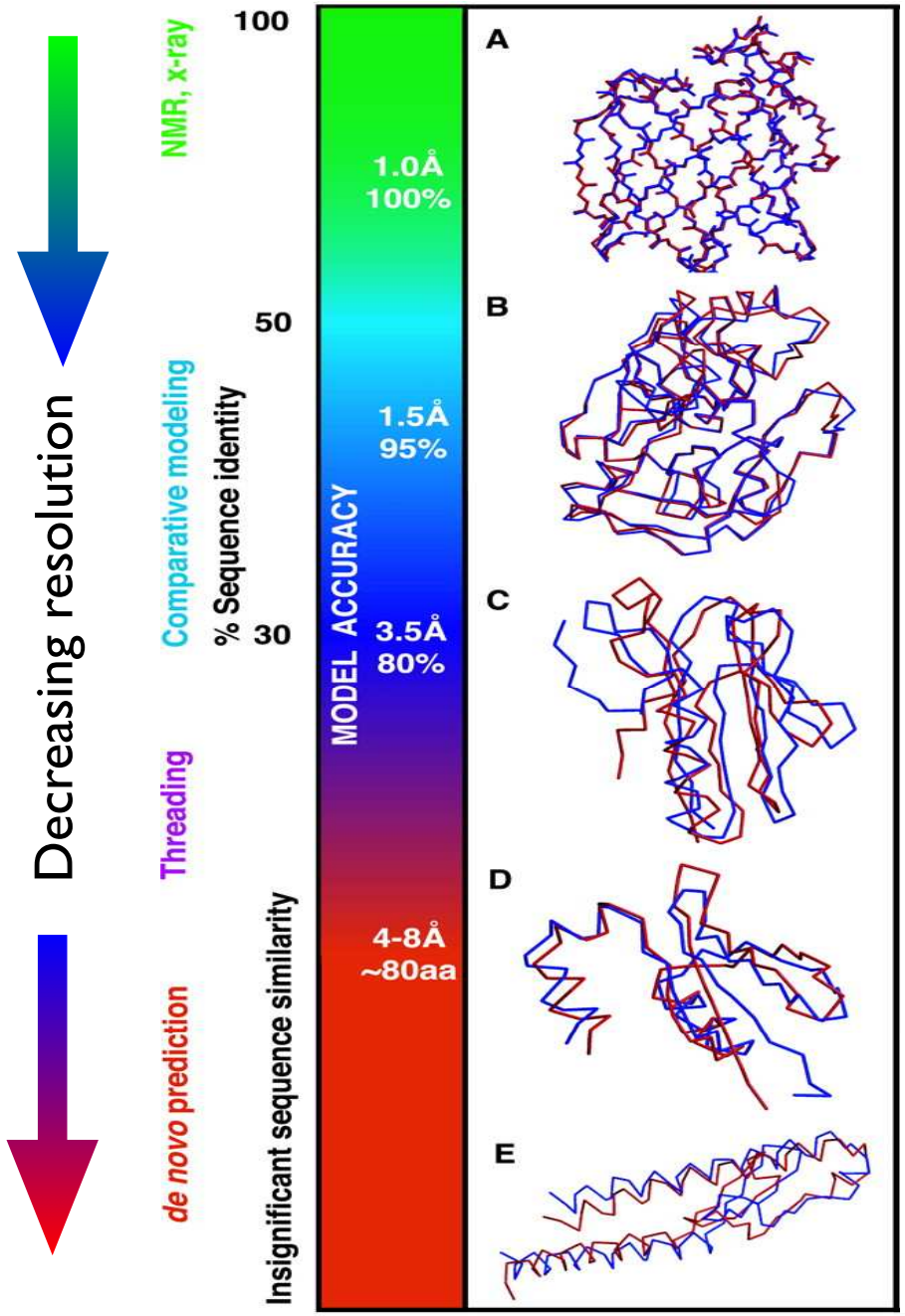


High resolution X-ray structure (Gold Standard)

Increasing accuracy

Baker and Sali *Science* (2001)

Protein Structure Prediction Accuracy and Resolution



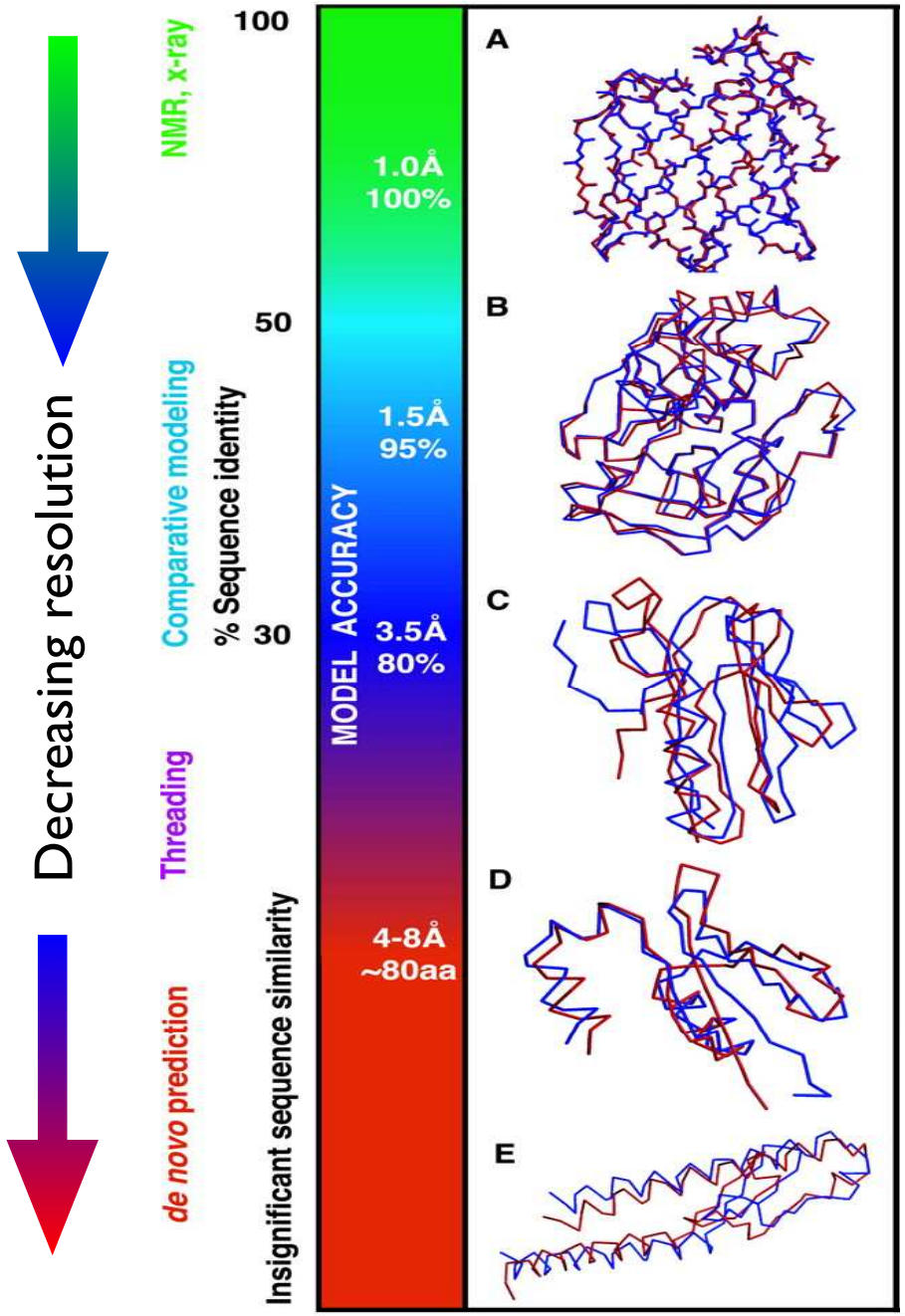
High resolution X-ray structure (Gold Standard)

de-novo models
RMSD to native : > 5 A

Increasing accuracy

Baker and Sali *Science* (2001)

Protein Structure Prediction Accuracy and Resolution



High resolution X-ray structure (Gold Standard)

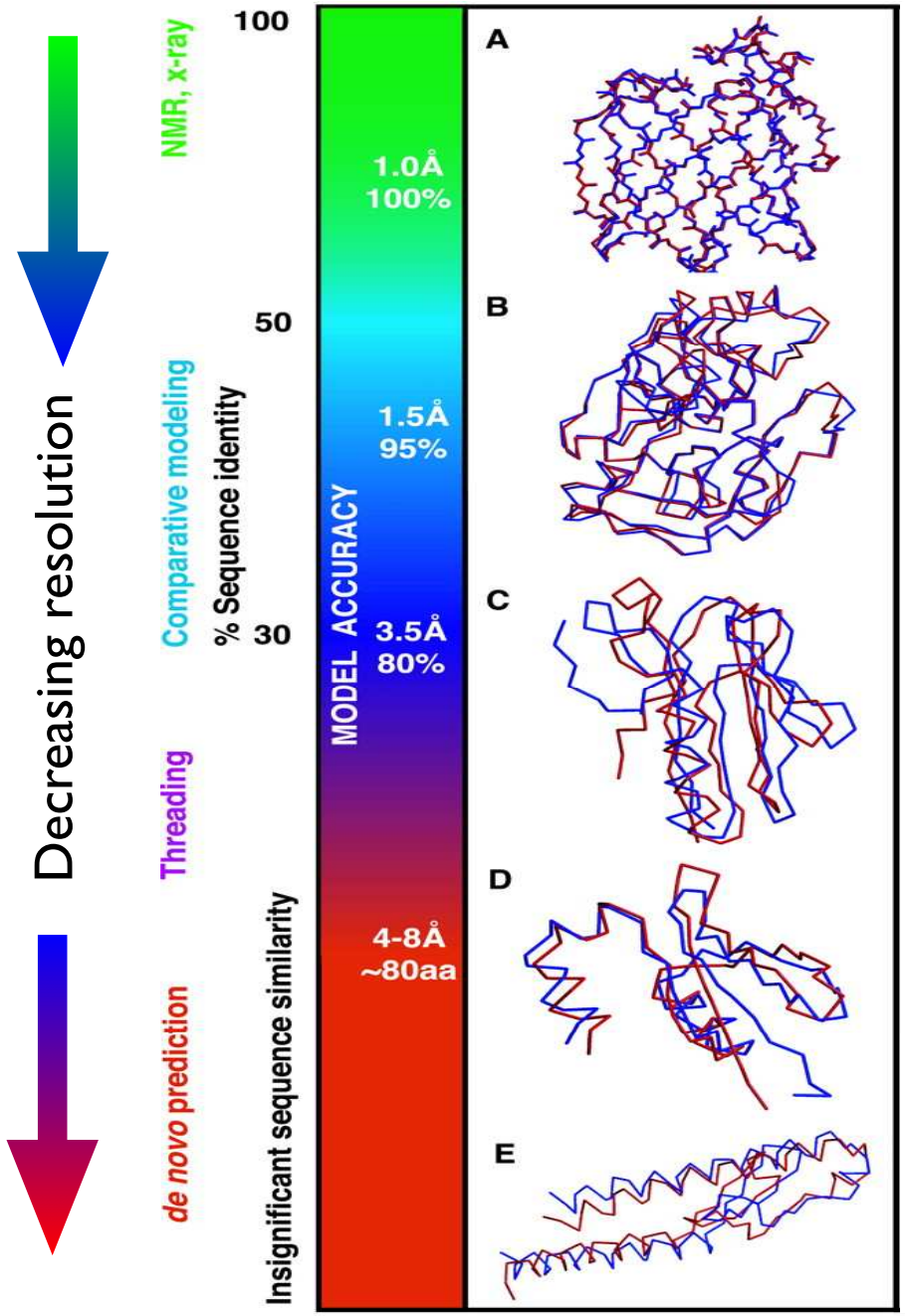
Low Resolution models
RMSD to native : 1-5 A

de-novo models
RMSD to native : > 5 A

Increasing accuracy

Baker and Sali *Science* (2001)

Protein Structure Prediction Accuracy and Resolution



High resolution X-ray structure (Gold Standard)

Low Resolution models
RMSD to native : 1-5 A

de-novo models
RMSD to native : > 5 A

Increasing accuracy

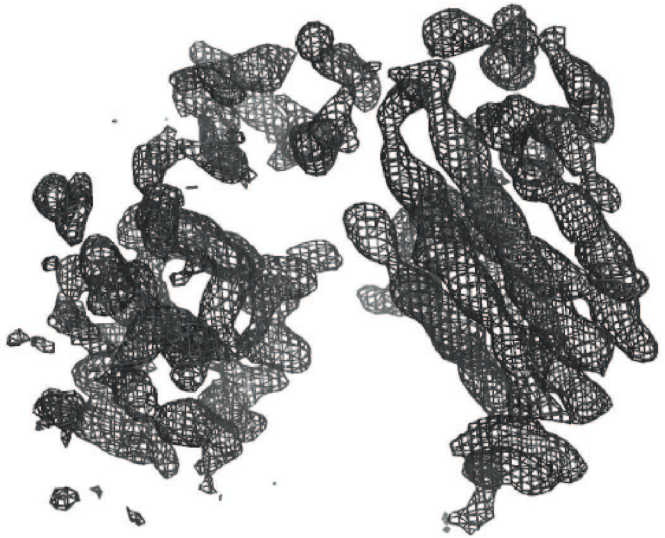
Baker and Sali *Science* (2001)

What are the sources of low resolution models ?

1. Homology models

2. NMR structures

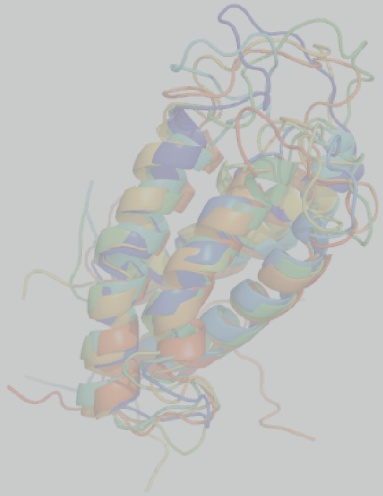
3. Cryo Electron Microscopy Structures



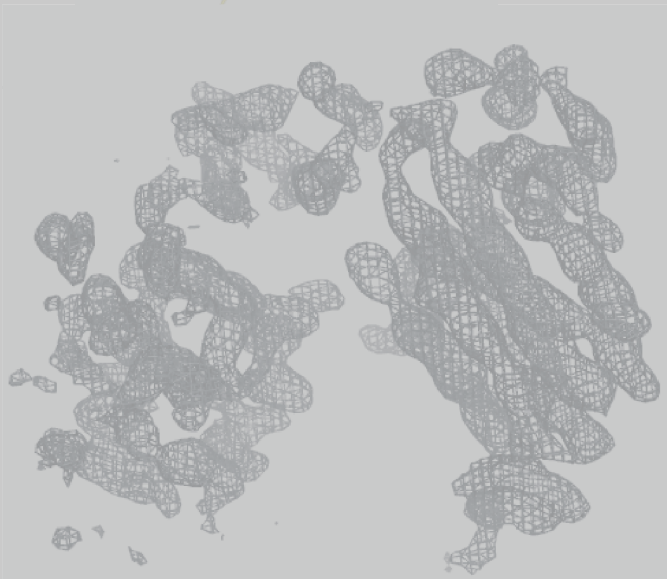
What are the sources of low resolution models ?

1. Homology models

2. NMR structures



3. Cryo Electron Microscopy Structures



What are the sources of low resolution models ?

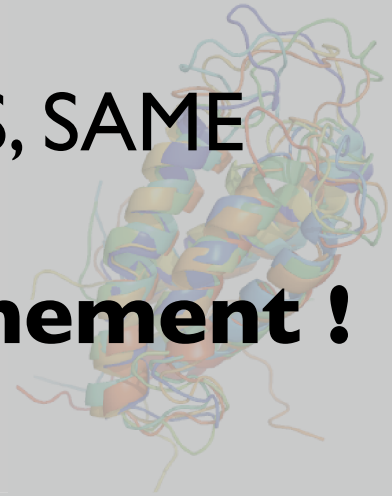
1. Homology models

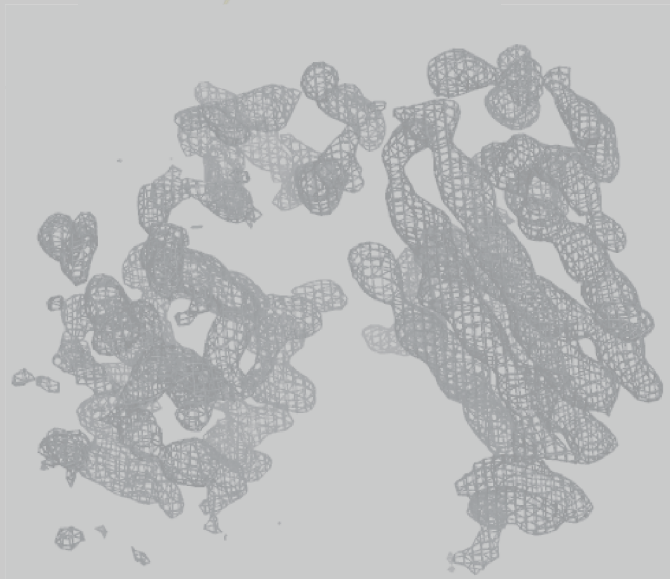
2. NMR structures

3. Cryo Electron Microscopy Structures

DIFFERENT STARTING POINTS, SAME UNDERLYING PROBLEM

High Resolution Refinement !





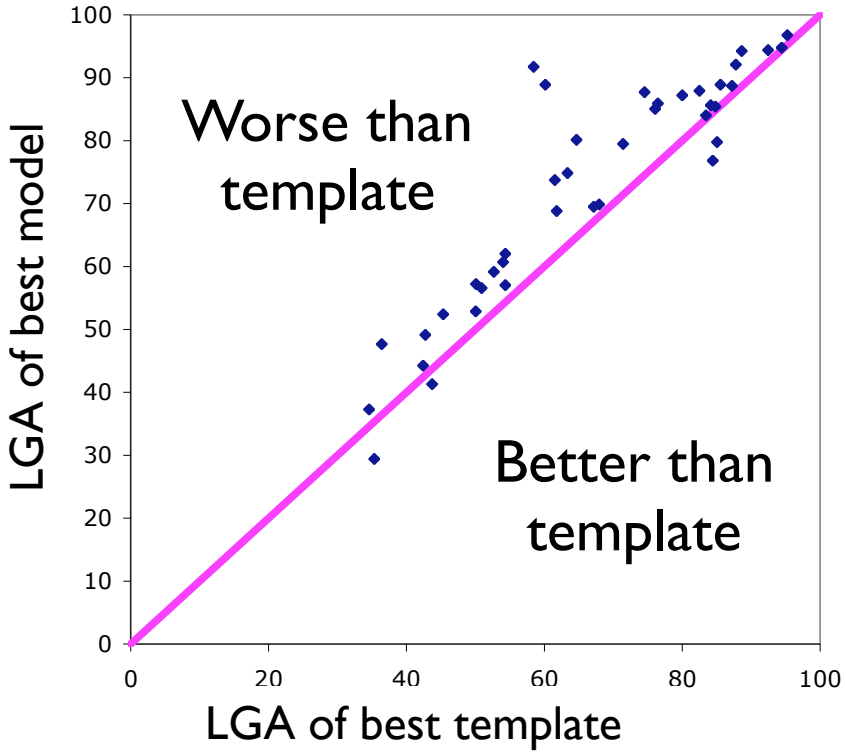
Driving force for innovation in protein structure prediction - CASP

CASP : Critical Assessment of Structure Prediction
double blind prediction of protein structures
First CASP experiment : 1994

Driving force for innovation in protein structure prediction - CASP

CASP : Critical Assessment of Structure Prediction
double blind prediction of protein structures
First CASP experiment : 1994

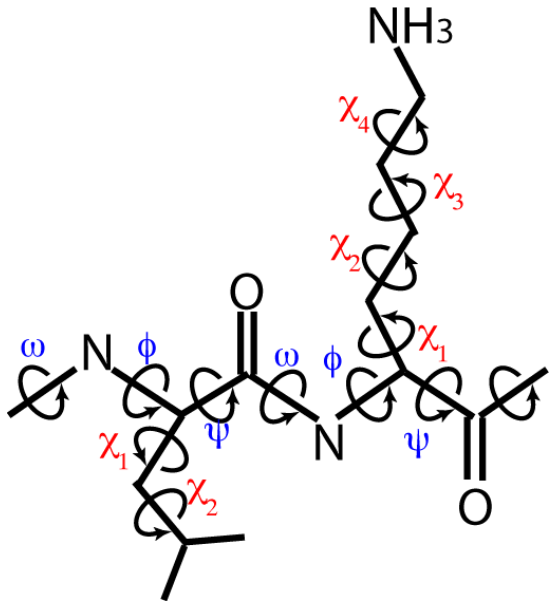
Long-standing in CASP : Improvement over template



Predictions from all groups: from CASP assessors' talk : 2004

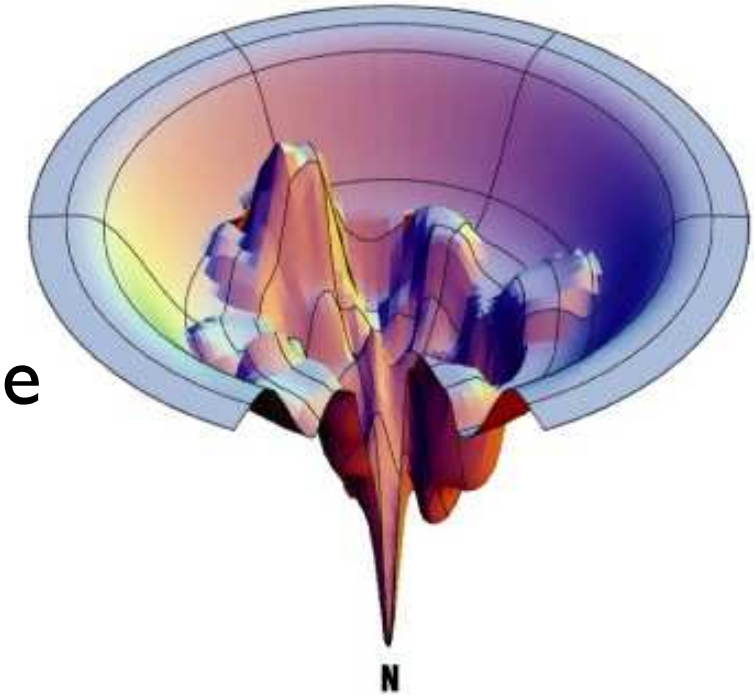
What makes high resolution refinement hard ?

Large number of degrees of freedom



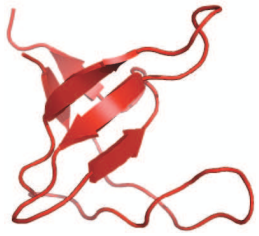
The energy landscape is littered with local minima

Very narrow radius-of-convergence to native structure



Dill et al *Nat. Str. Biol* (1997)

High Resolution Refinement Method



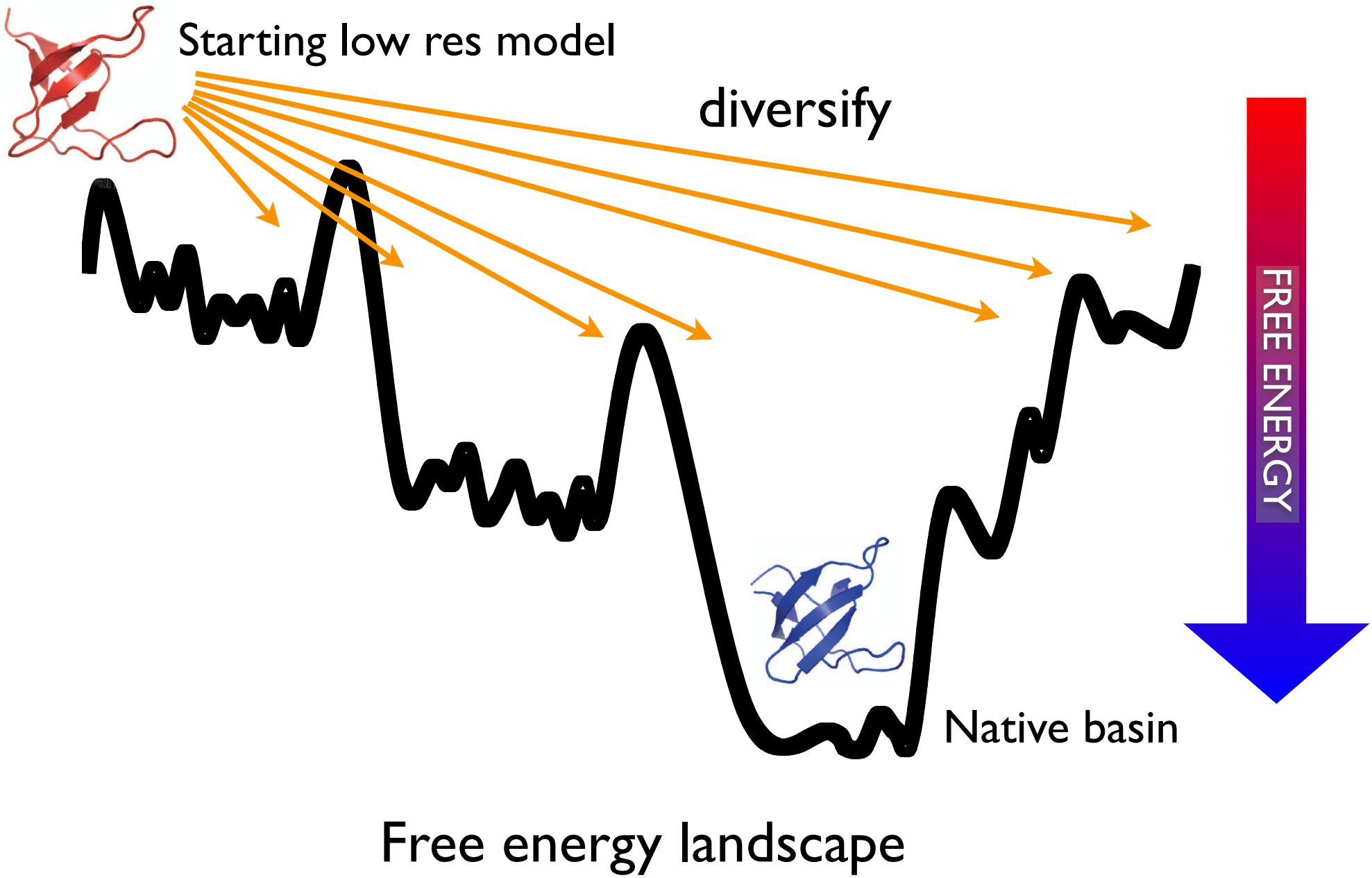
Starting low res model

Native basin

FREE ENERGY

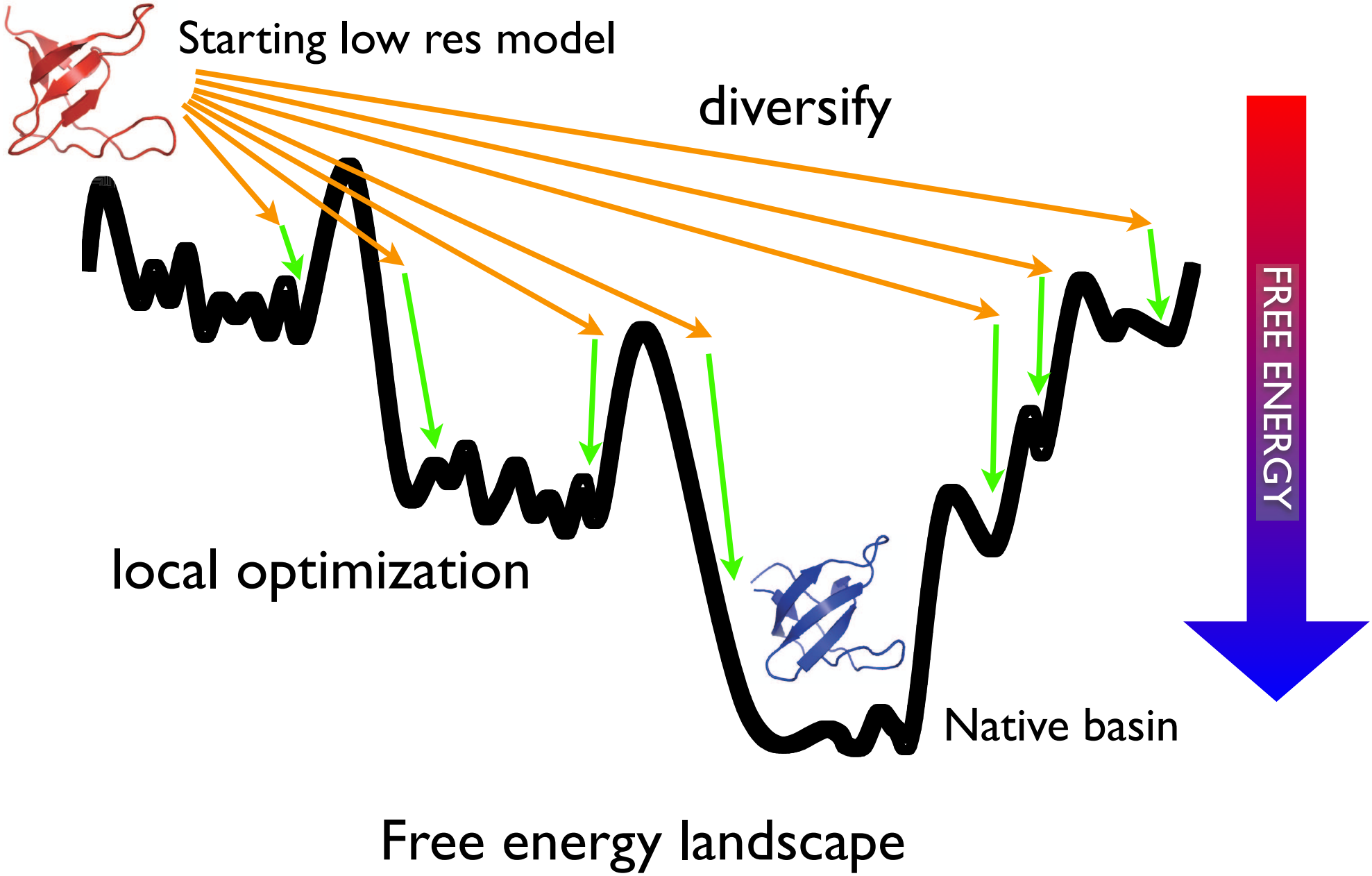
Free energy landscape

High Resolution Refinement Method



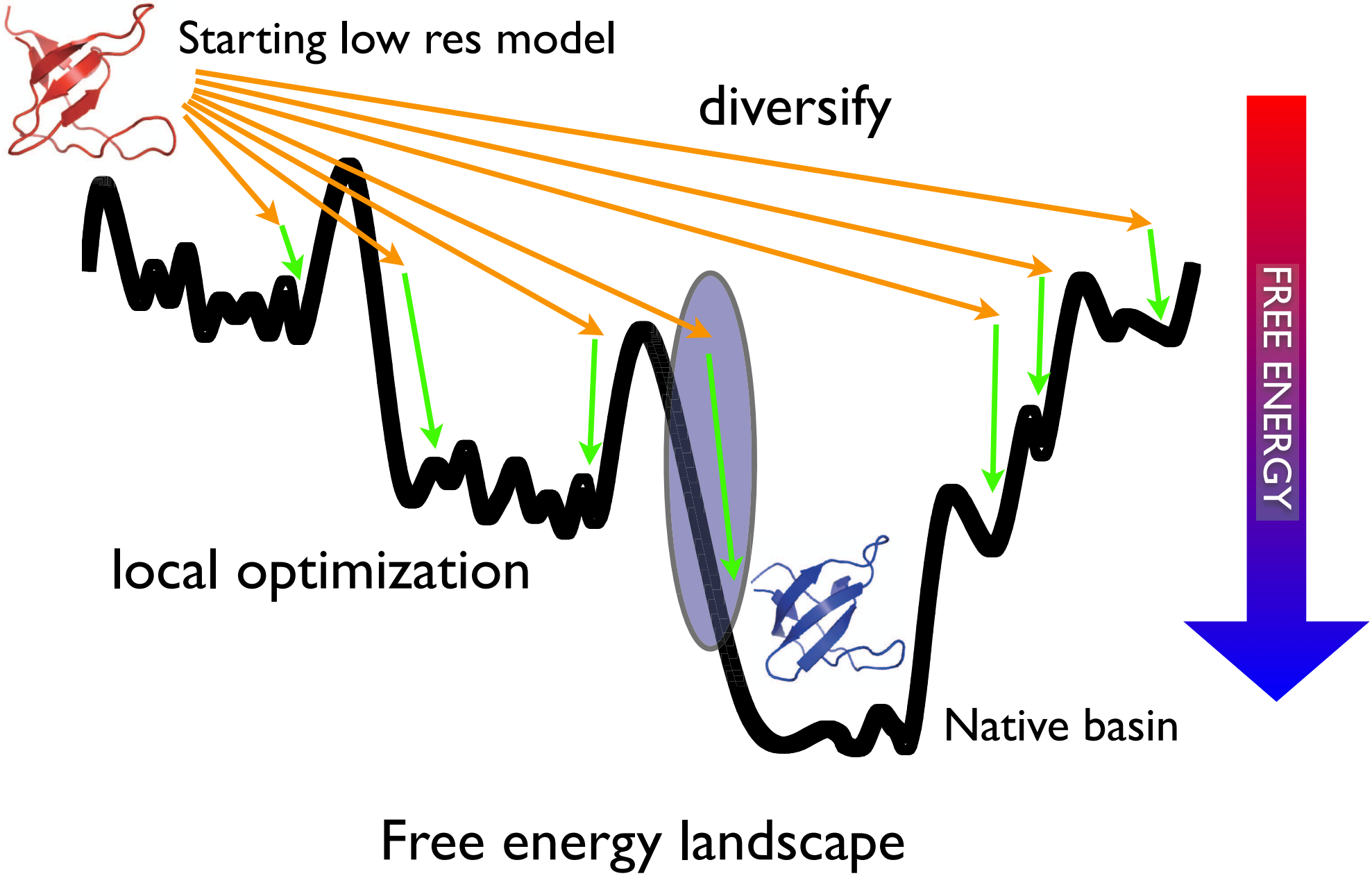
Free energy landscape

High Resolution Refinement Method



Free energy landscape

High Resolution Refinement Method



Free energy landscape

Diversifying Sampling

Diversifying Sampling

How do we diversify ?

Diversifying Sampling

How do we diversify ?

By aggressively rebuilding parts of the structure

Diversifying Sampling

How do we diversify ?

By aggressively rebuilding parts of the structure

Which parts are chosen for rebuilding ?

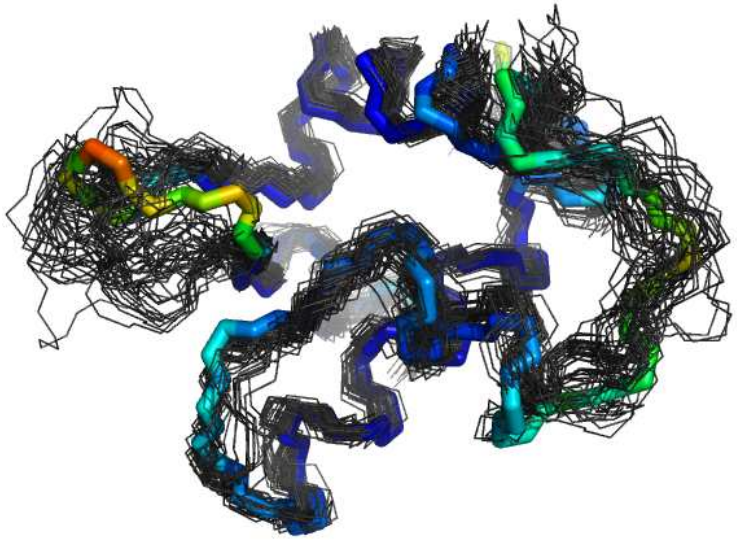
Diversifying Sampling

How do we diversify ?

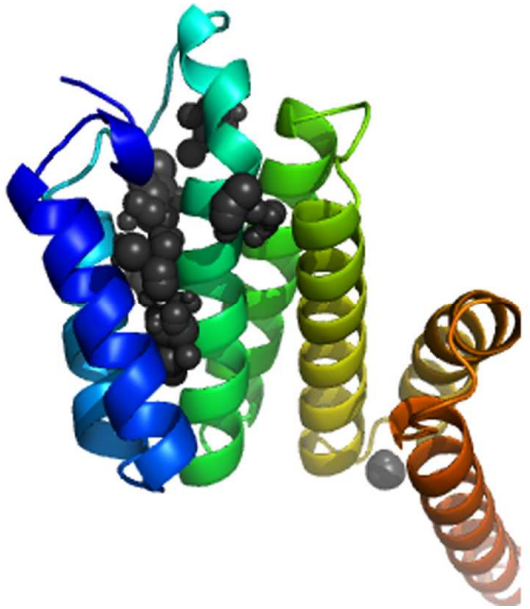
By aggressively rebuilding parts of the structure

Which parts are chosen for rebuilding ?

1. Highly varying regions in the starting structures
2. Poorly packed region in the starting structures



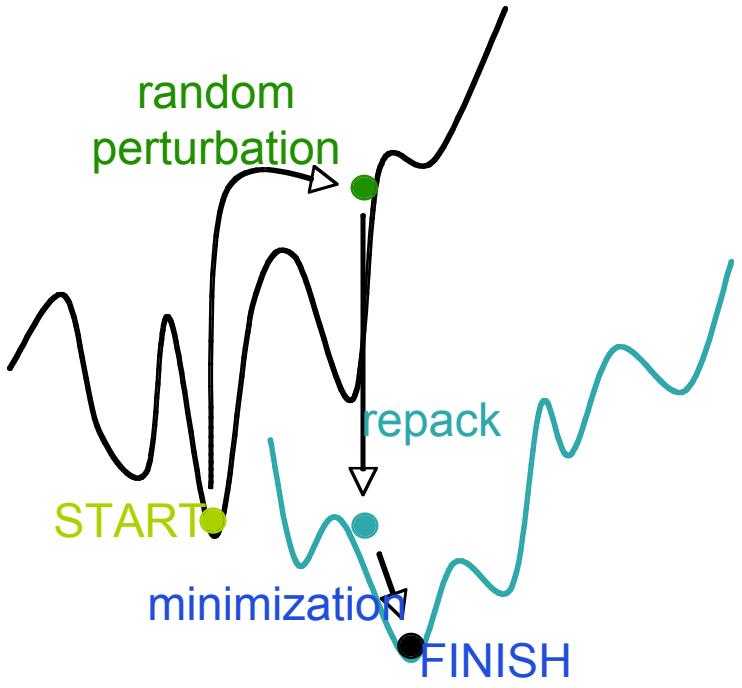
Highly varying regions



Poorly packed regions

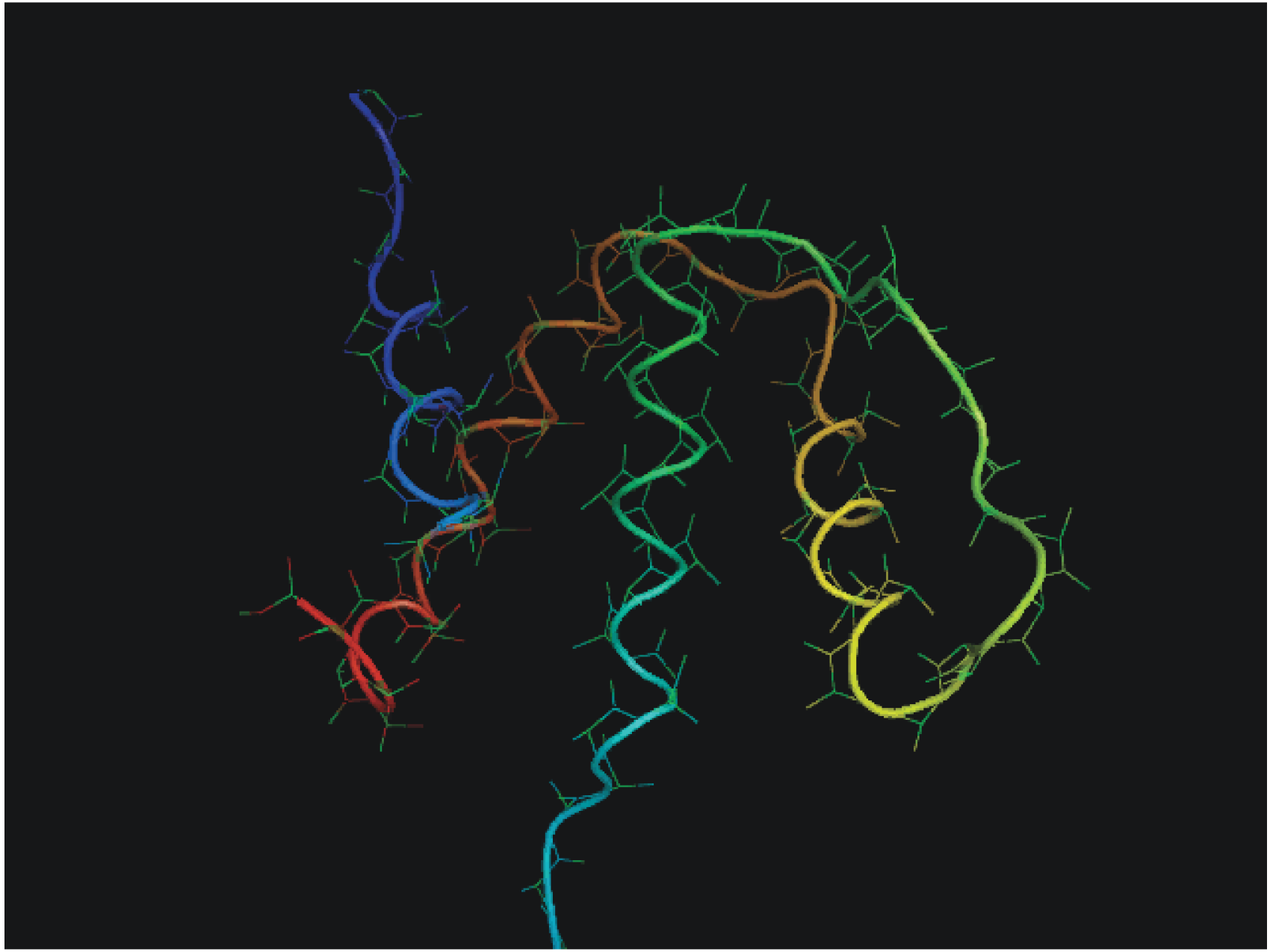
Local Optimization

- Stochastically choose a residue on the protein
- Perturb the backbone at that position
- Rearrange the side chains
- Minimization
- Metropolis Monte-Carlo



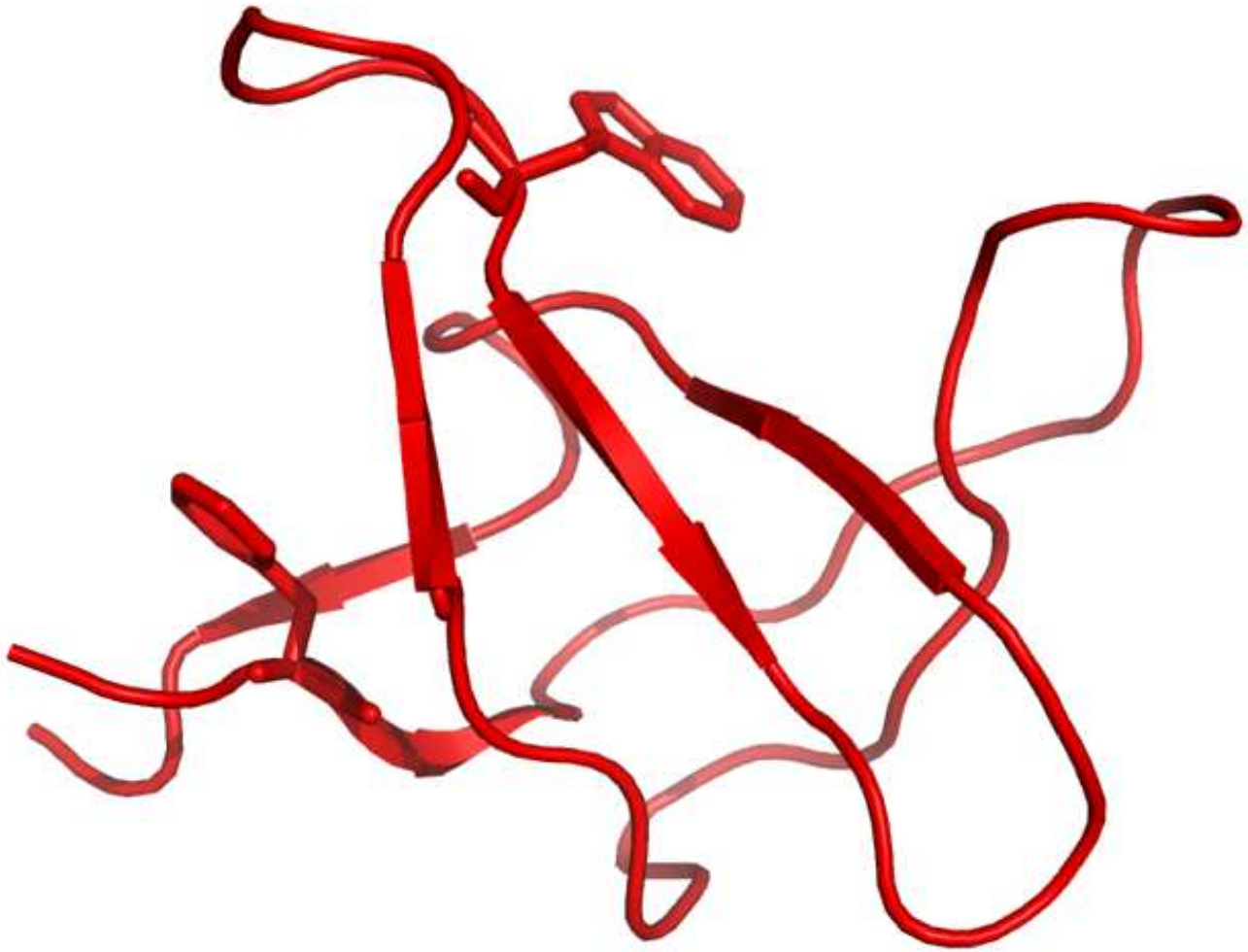
Bradley et al *Science* (2005)

diversification movie



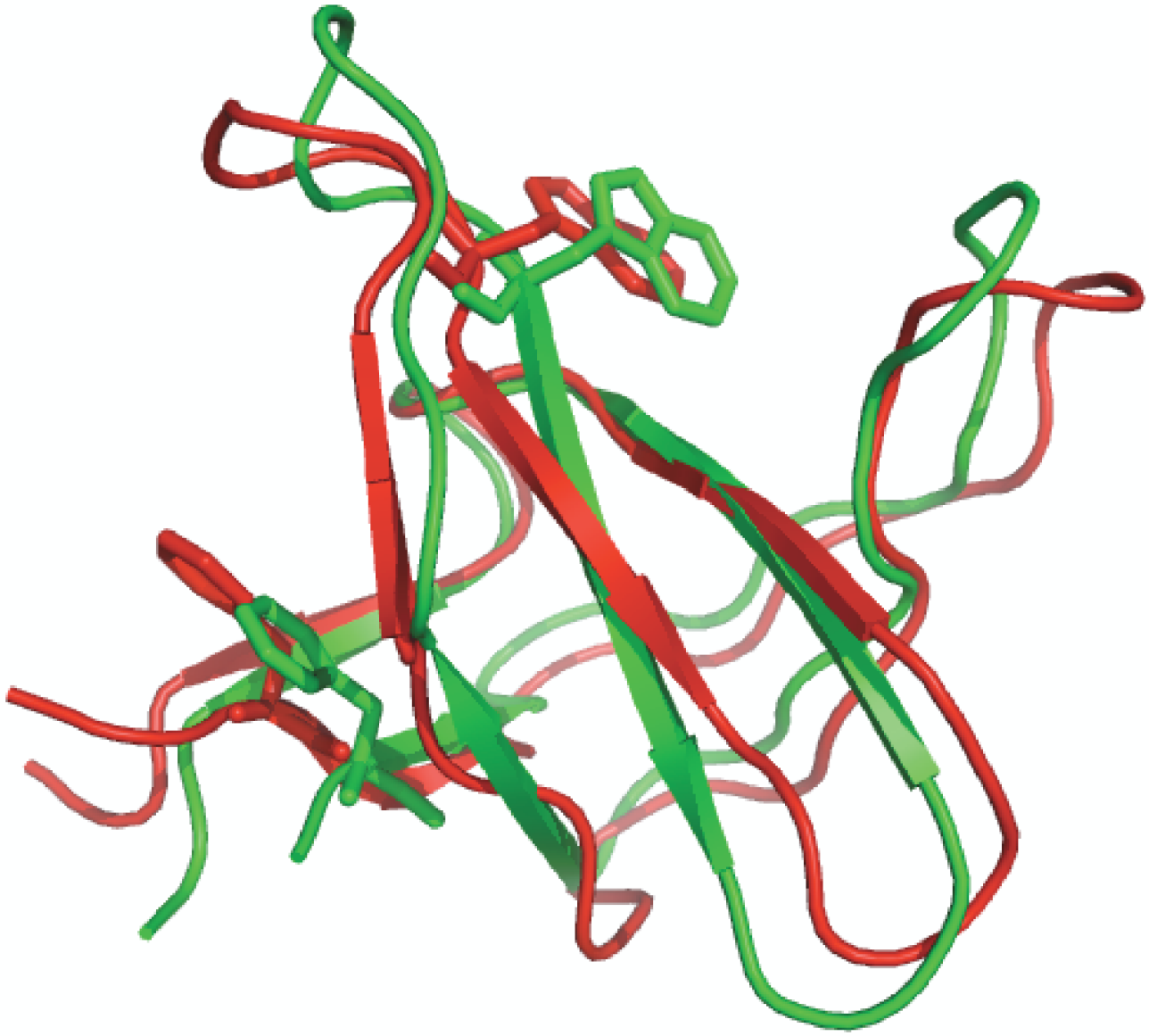
local optimization movie

SH3 domain of ABL Tyr kinase - all beta fold



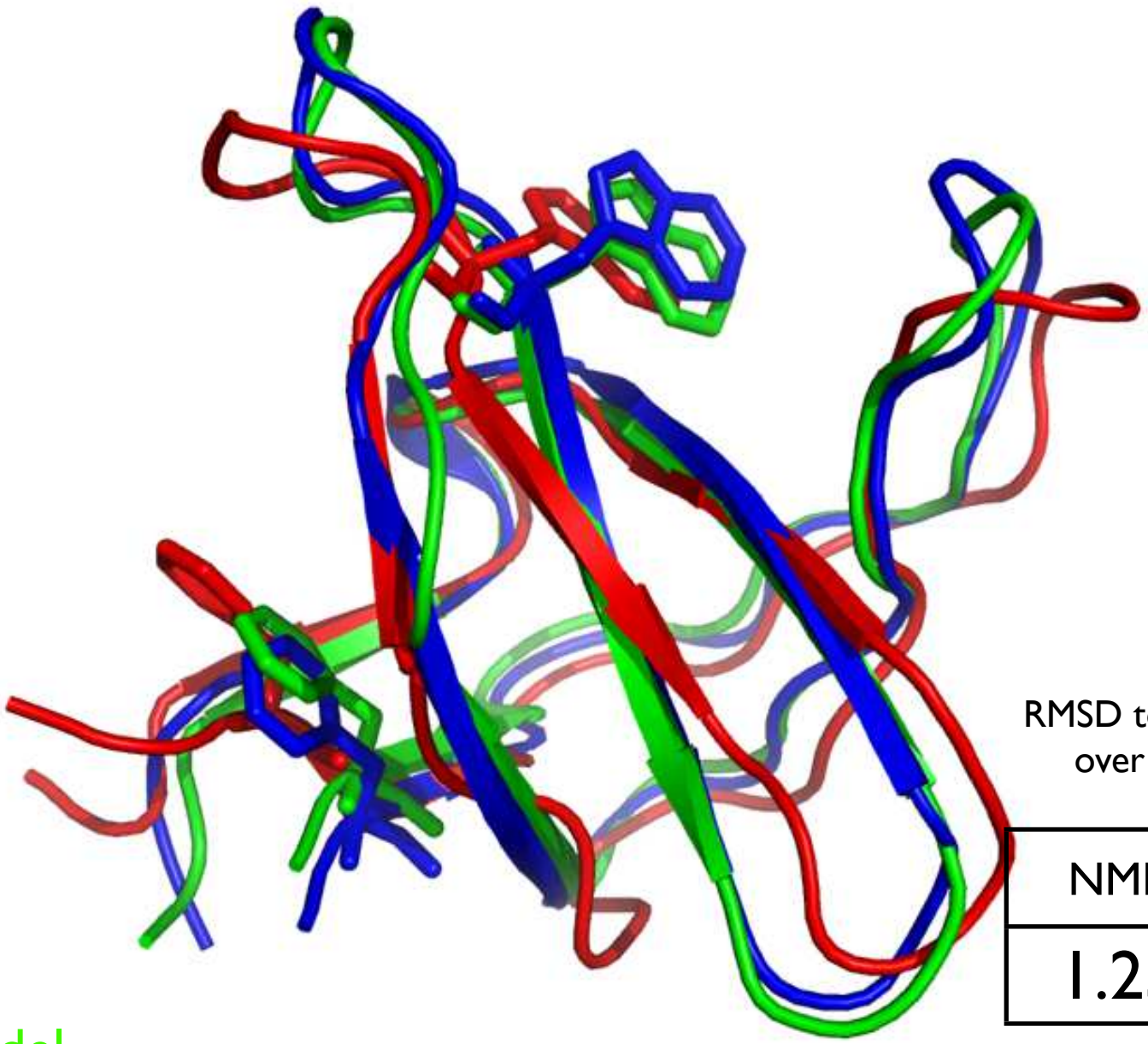
Qian, Raman, Das et al *Nature* (2007)

SH3 domain of ABL Tyr kinase - all beta fold



Qian, Raman, Das et al *Nature* (2007)

SH3 domain of ABL Tyr kinase - all beta fold



NMR

Refined model

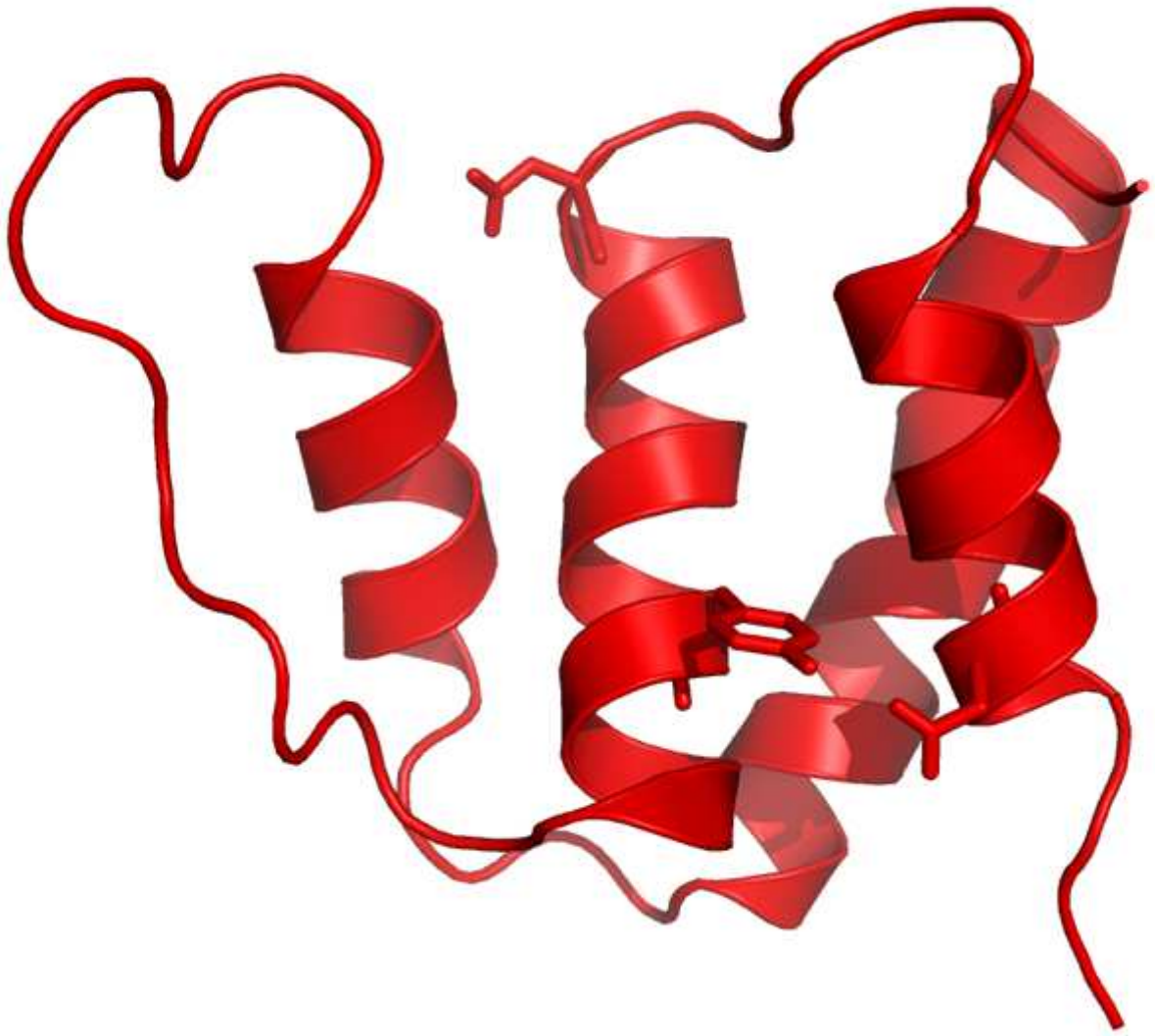
High res X-ray structure

RMSD to Native all-atom over core residues

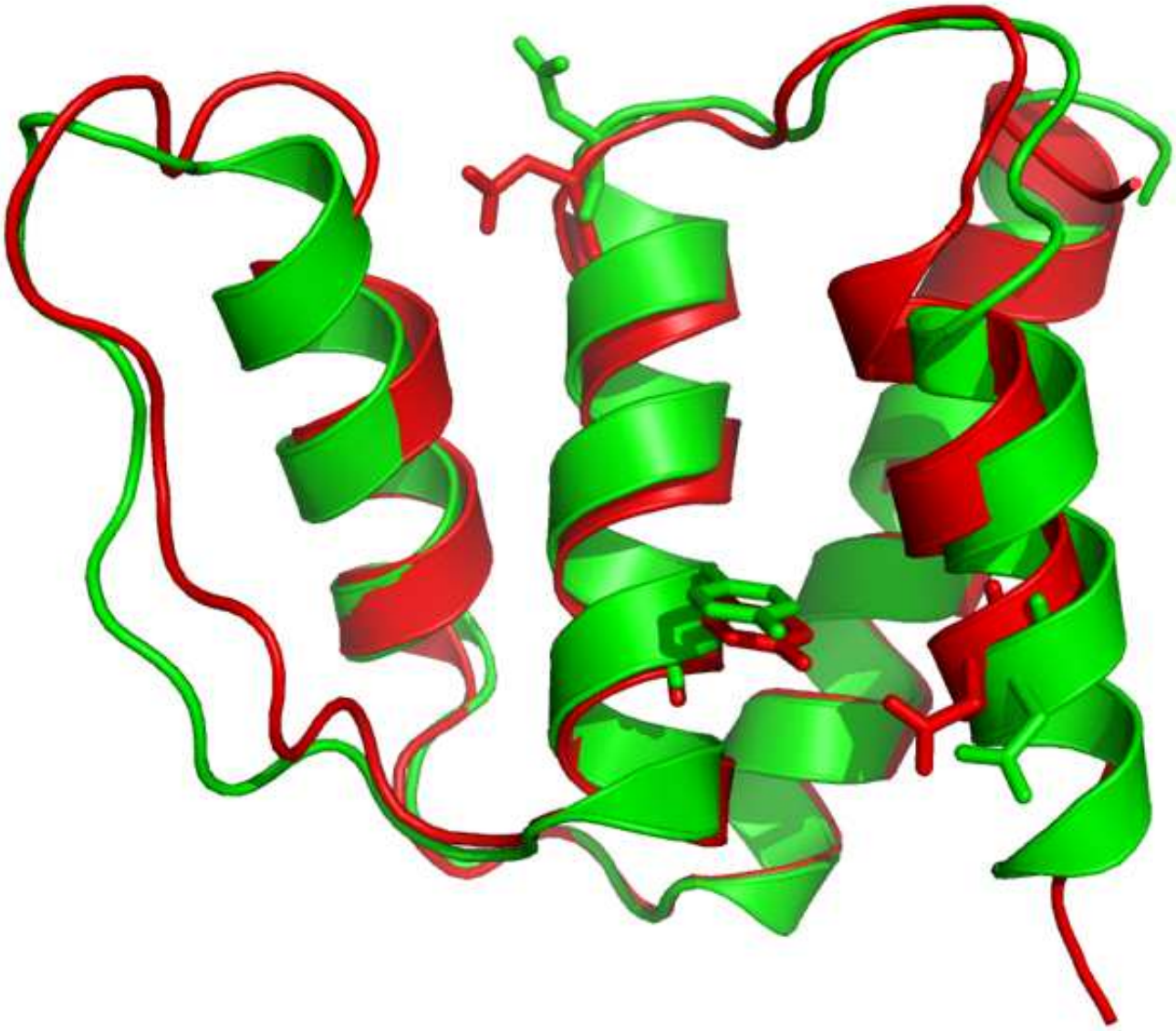
NMR	Refined
1.25	0.78

Qian, Raman, Das et al *Nature* (2007)

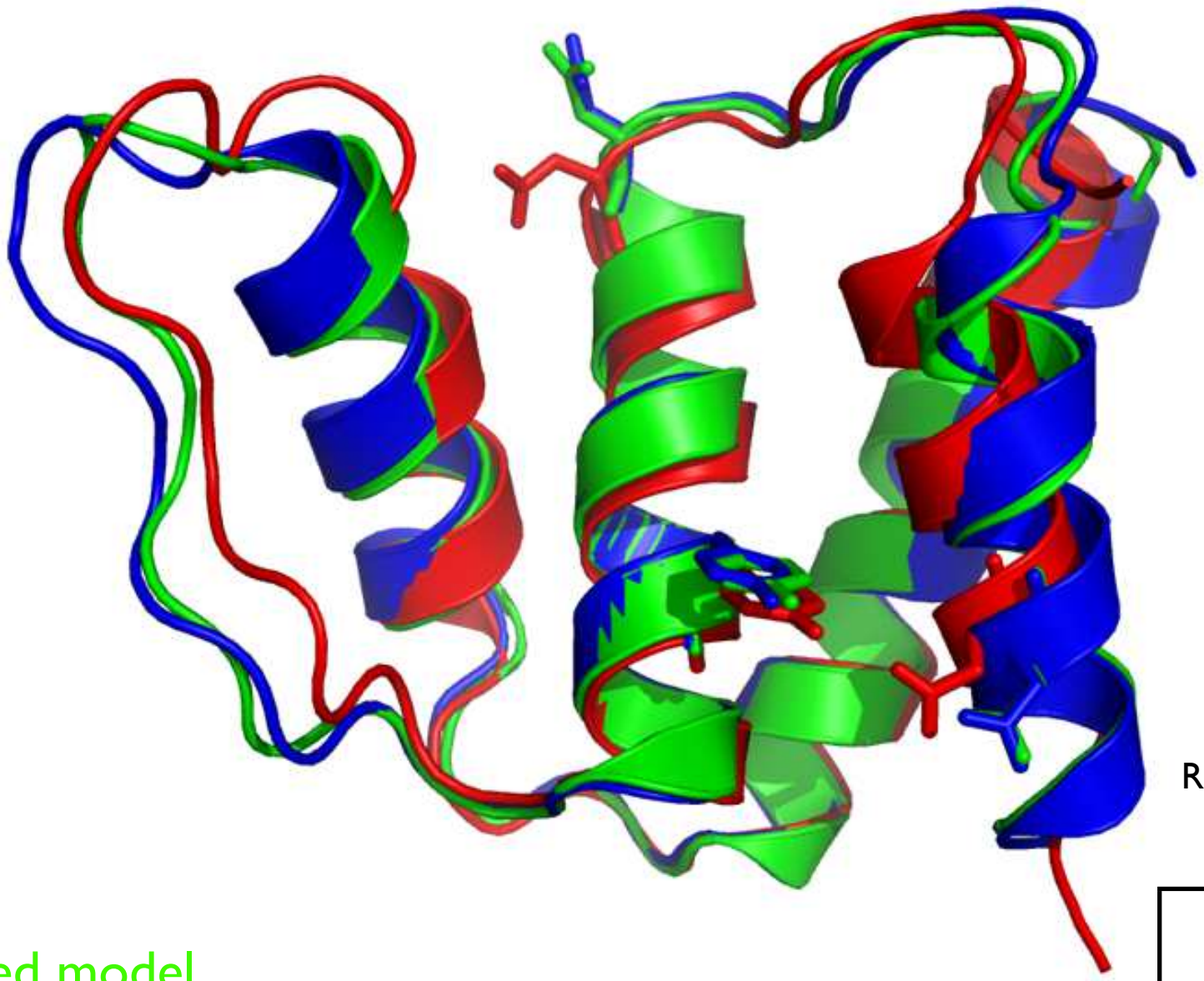
Acyl-CoA inhibitor - all alpha fold



Acyl-CoA inhibitor - all alpha fold



Acyl-CoA inhibitor - all alpha fold

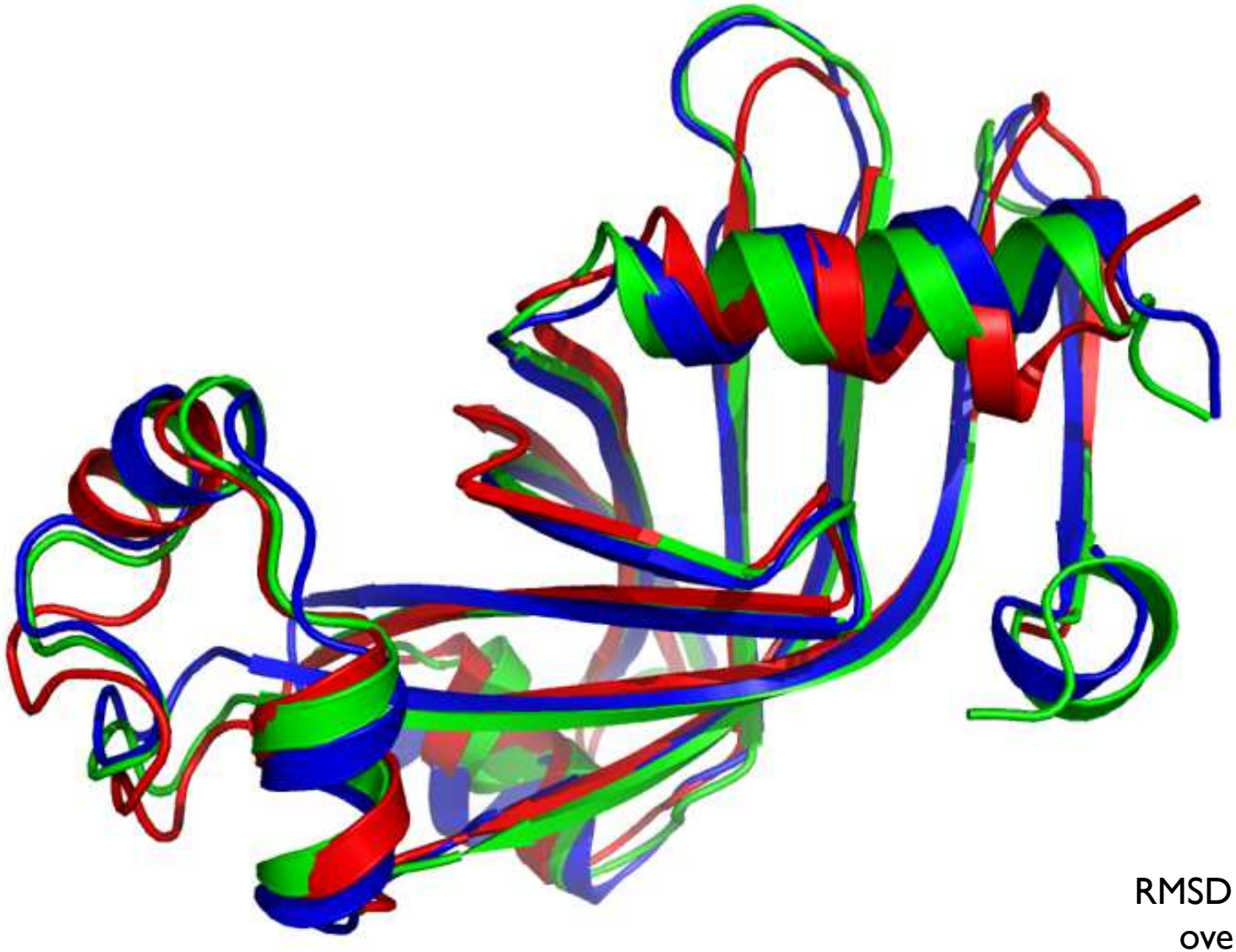


NMR
Refined model
High res X-ray structure

RMSD to native all-atom over core residues

NMR	Refined
1.58	0.91

CASP7 target T380

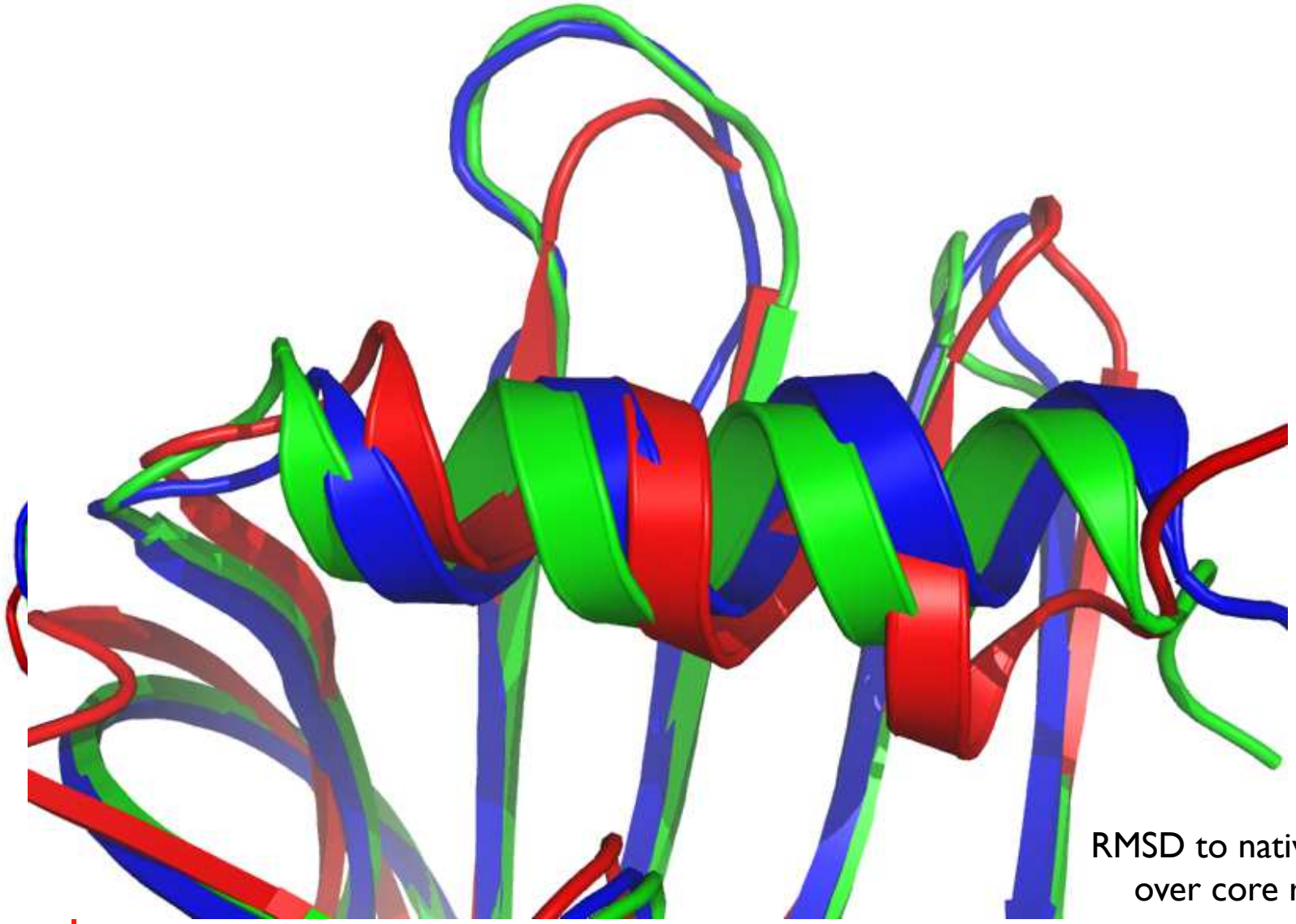


Best template
Refined model
High res X-ray structure

RMSD to native all-atom
over core residues

Template	Refined
2.33	1.41

CASP7 target T380

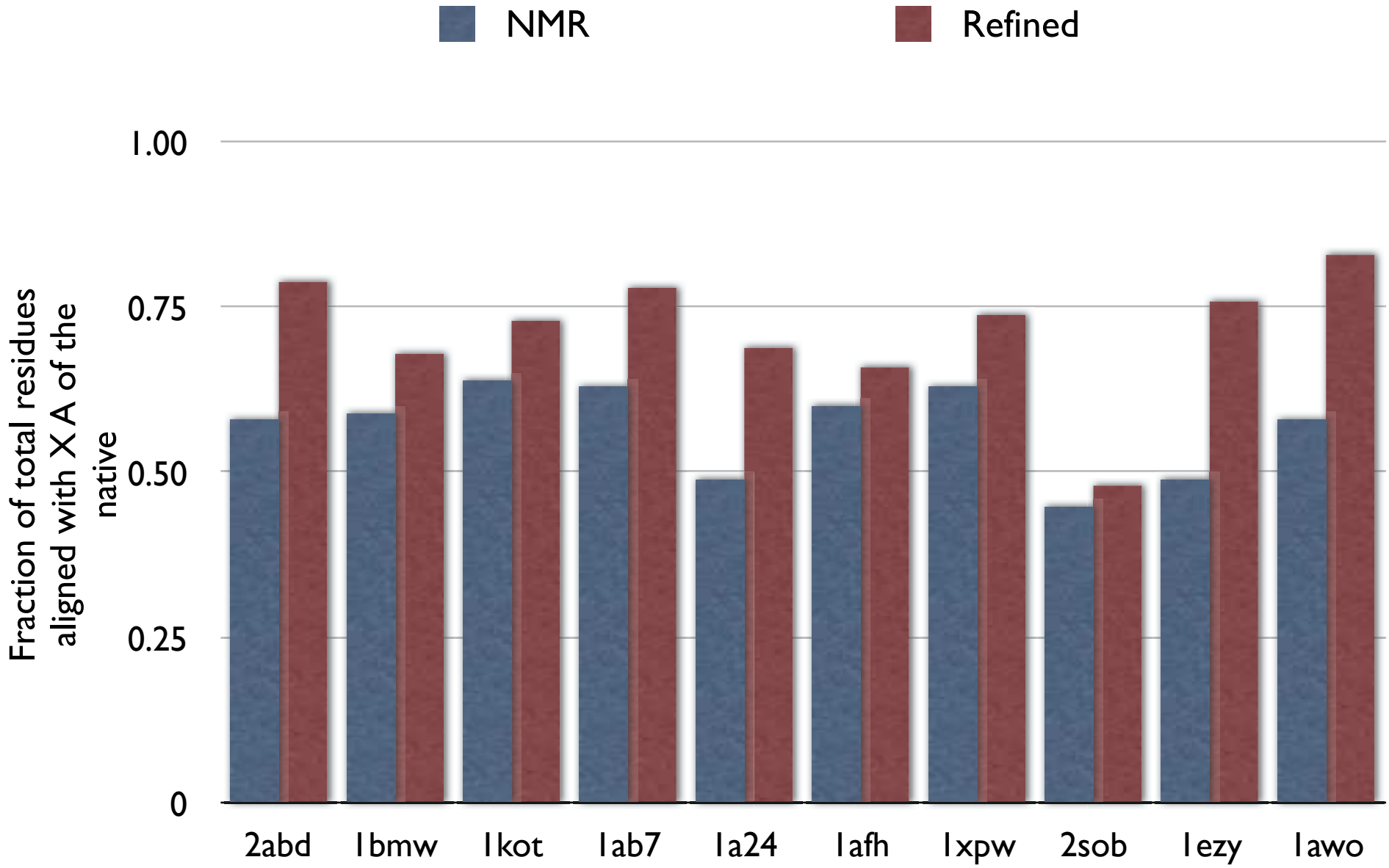


Best template
Refined model
High res X-ray structure

RMSD to native all-atom over core residues

Template	Refined
2.33	1.41

Results summary - NMR refinement



MOLPROBITY score - model quality

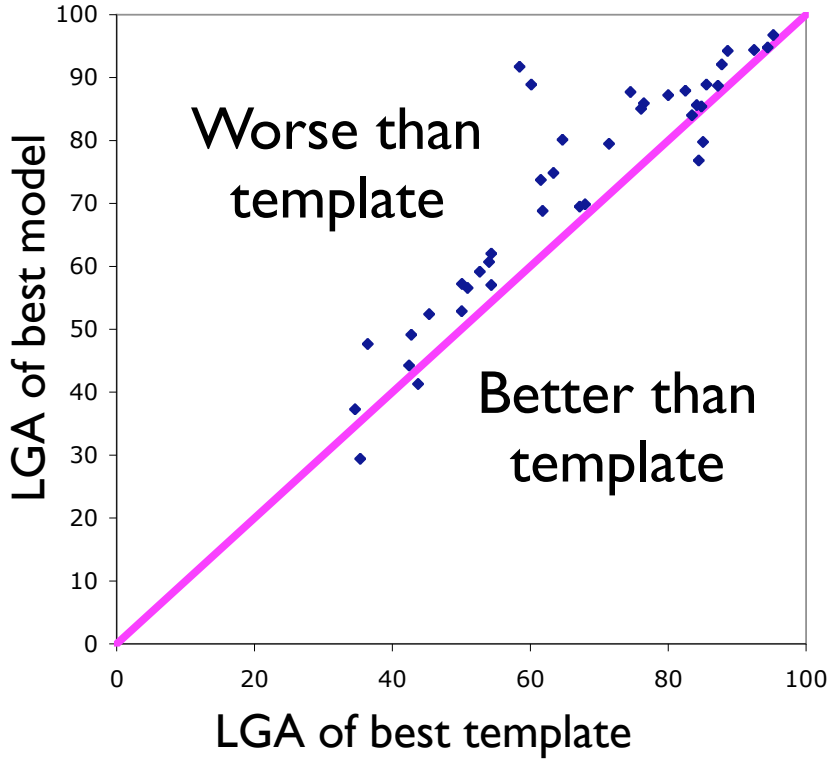
PDB ID	% rotamer outliers		% ramachandran outliers	
	NMR	Refined	NMR	Refined
2abd	19	0	9	0
1bmw	30	0	9	0
1kot	25	1	4	0
1ab7	22	0	4	0
1a24	47	2	7	0
1afh	14	0	2	0
1xpw	6	0	2	0
2sob	29	0	10	0
1ezy	8	0	0	0
1awo	13	0	1	0

Davis et al *NAR* (2007)

Driving force for innovation in protein structure prediction - CASP

CASP : Critical Assessment of Structure Prediction
double blind prediction of protein structures
First CASP experiment : 1994

Long-standing in CASP : Improvement over template

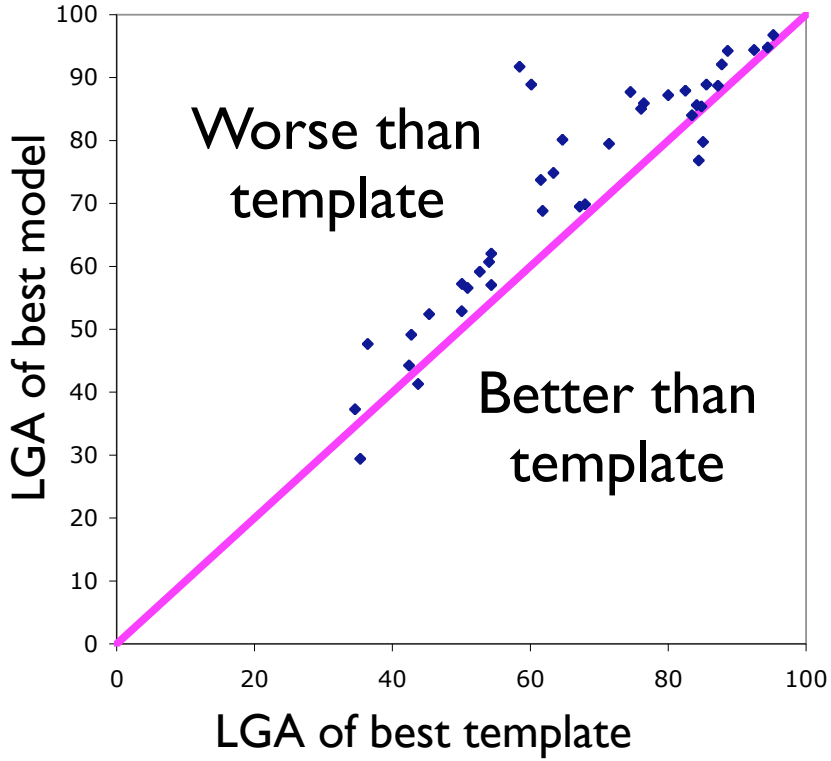


Predictions from all groups: from CASP assessors' talk : 2004

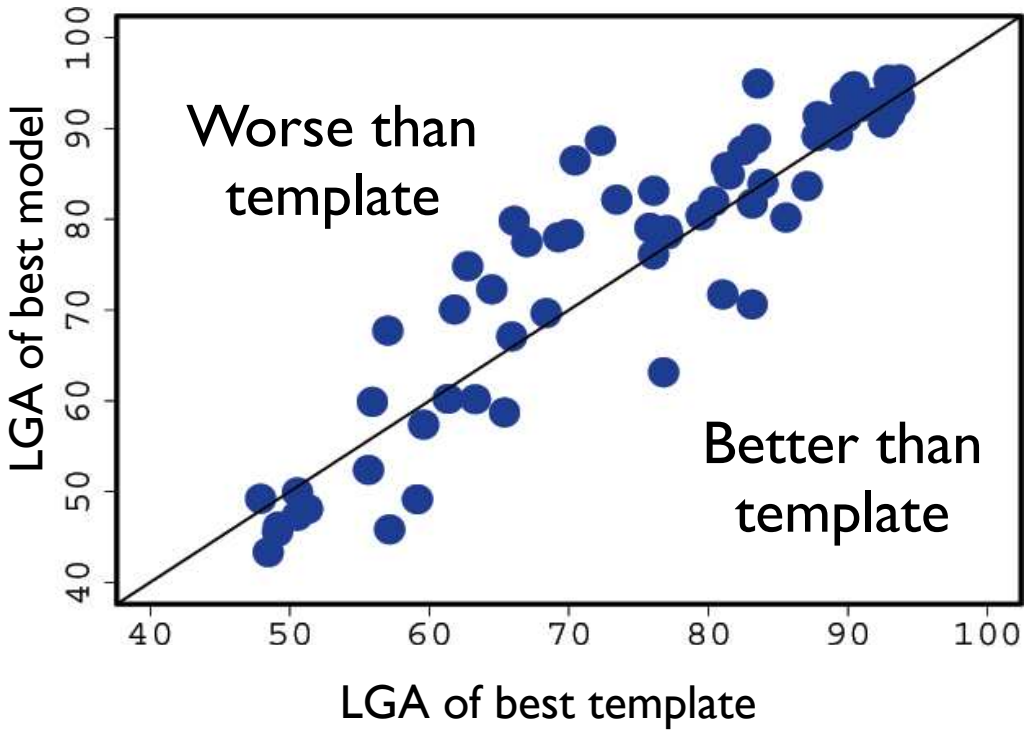
Driving force for innovation in protein structure prediction - CASP

CASP : Critical Assessment of Structure Prediction
double blind prediction of protein structures
First CASP experiment : 1994

Long-standing in CASP : Improvement over template



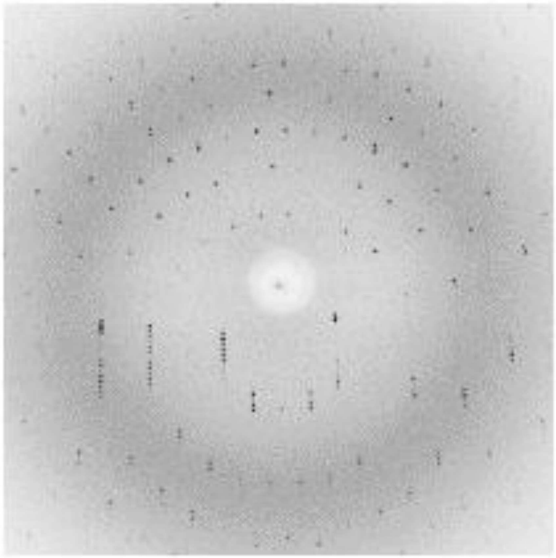
Predictions from all groups: from CASP assessors' talk : 2004



Baker group predictions 2006

The X-ray crystallographic phase problem

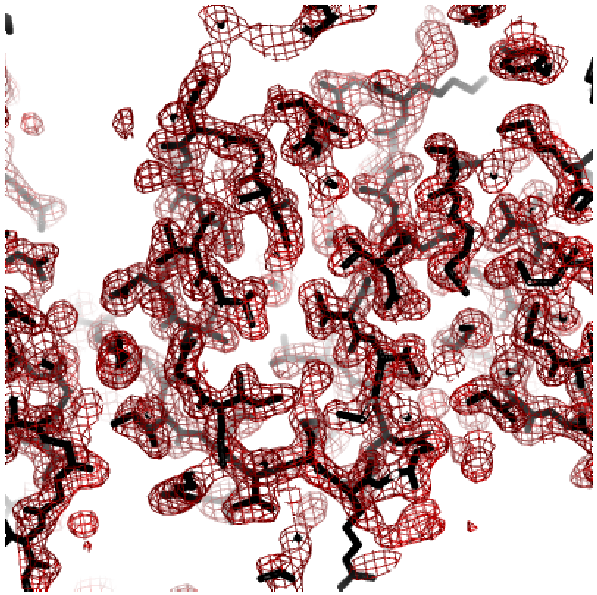
Protein crystal



Diffraction map

+

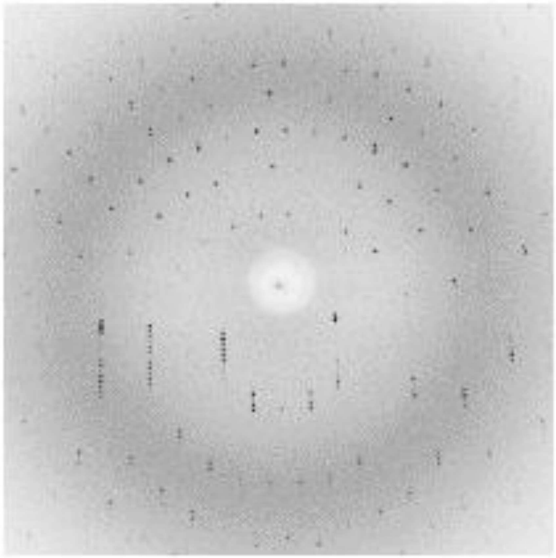
Phases



Electron density map

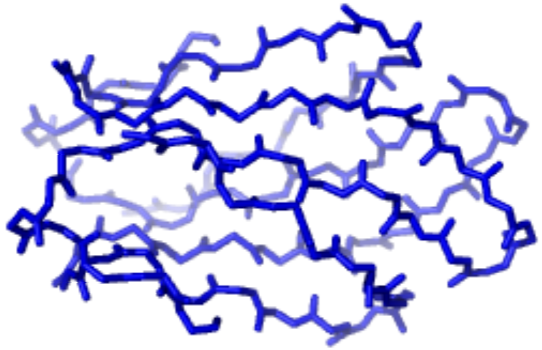
The X-ray crystallographic phase problem

Protein crystal

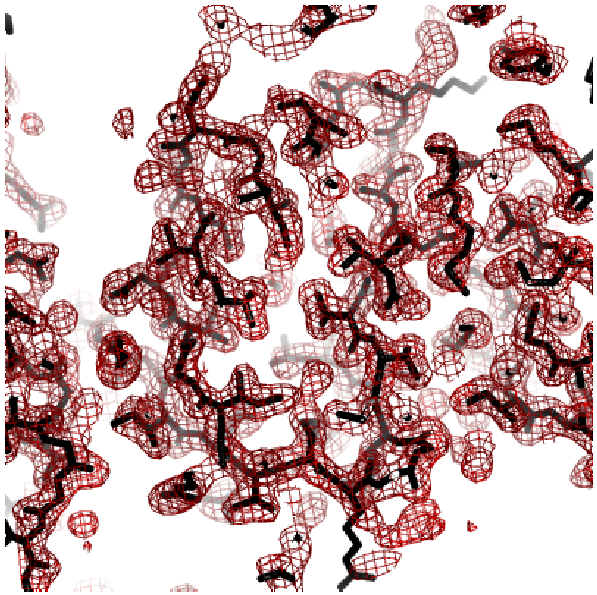


Diffraction map

+



Molecular replacement model



Electron density map

When is the protein folding problem solved ?

"There is an obvious method of evaluation that will allow any structure prediction method to be assessed.

It is simply to demand that the method produce a model that can be used to solve the corresponding crystal structure by ***molecular replacement***"

- Gregory A. Petsko, Genome Biology (2000)

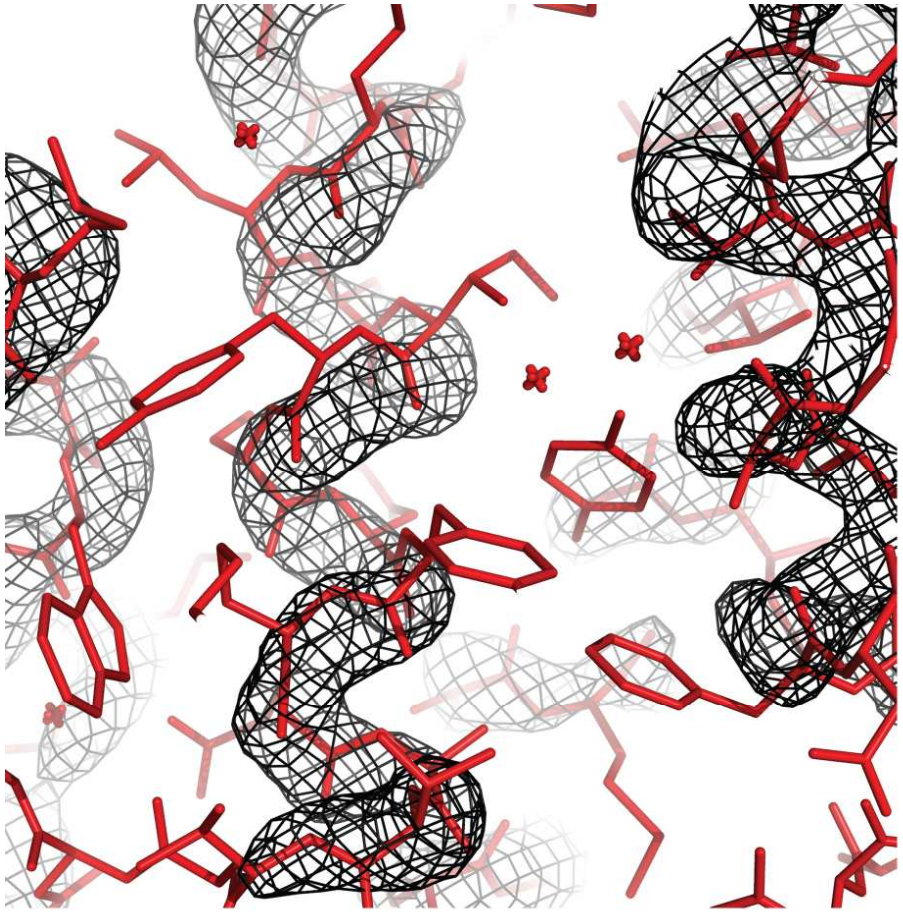
Do we pass this stringent test ?

NMR : Acyl-CoA inhibitor

RED STICKS : High resolution X-ray structure Qian, Raman, Das et al *Nature* (2007)

Do we pass this stringent test ?

NMR : Acyl-CoA inhibitor

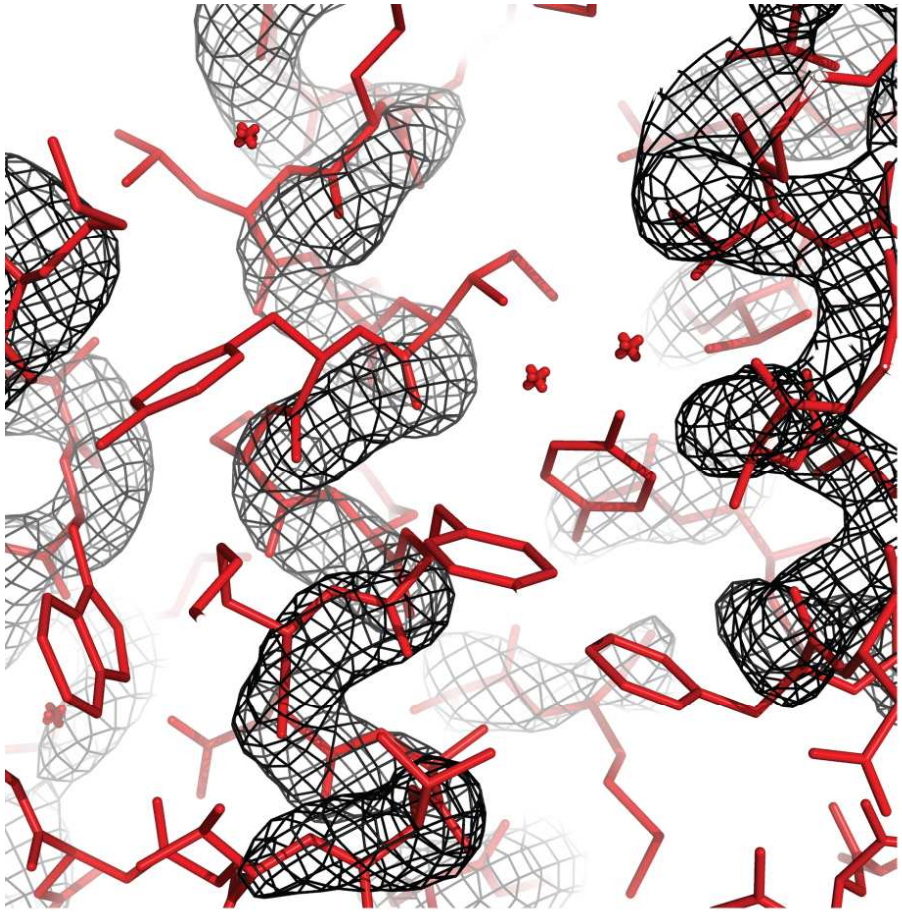


Electron density map from
starting NMR model

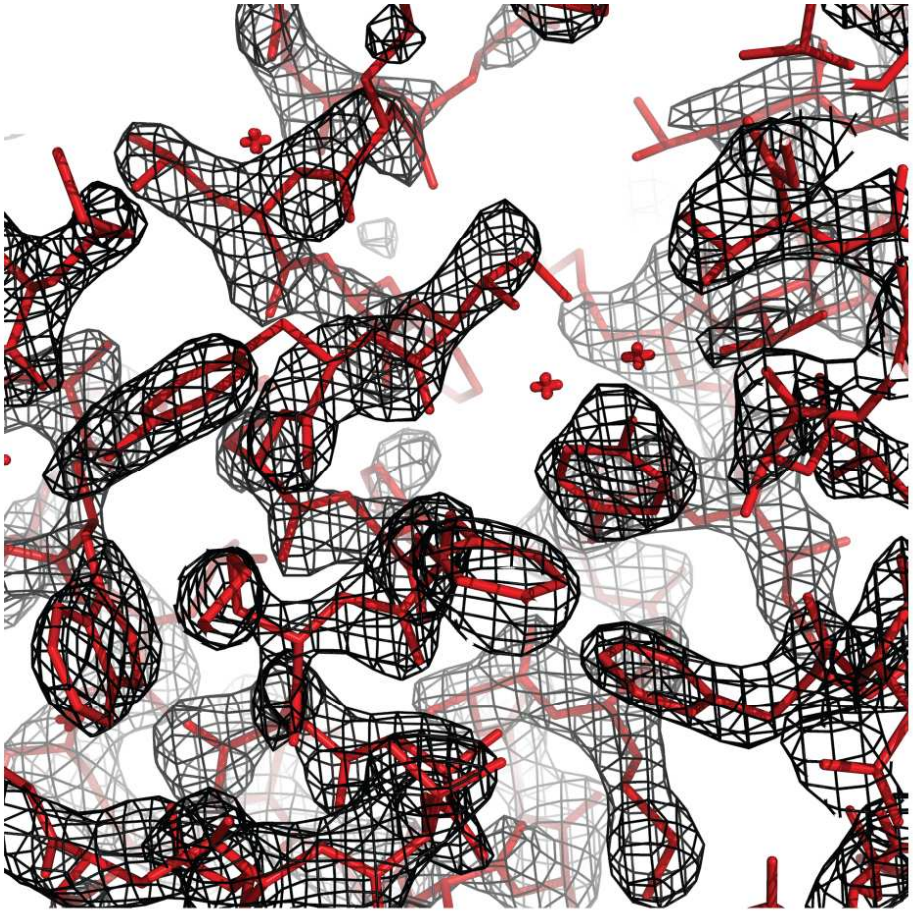
RED STICKS : High resolution X-ray structure Qian, Raman, Das et al *Nature* (2007)

Do we pass this stringent test ?

NMR : Acyl-CoA inhibitor



Electron density map from starting NMR model



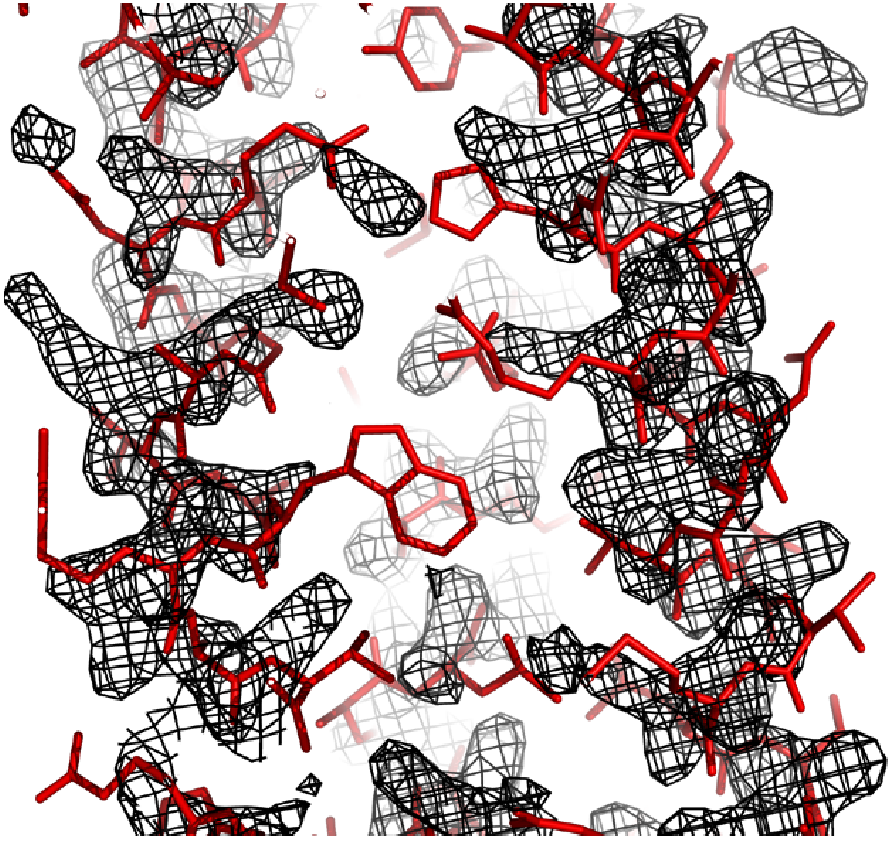
Electron density map from refined model

RED STICKS : High resolution X-ray structure

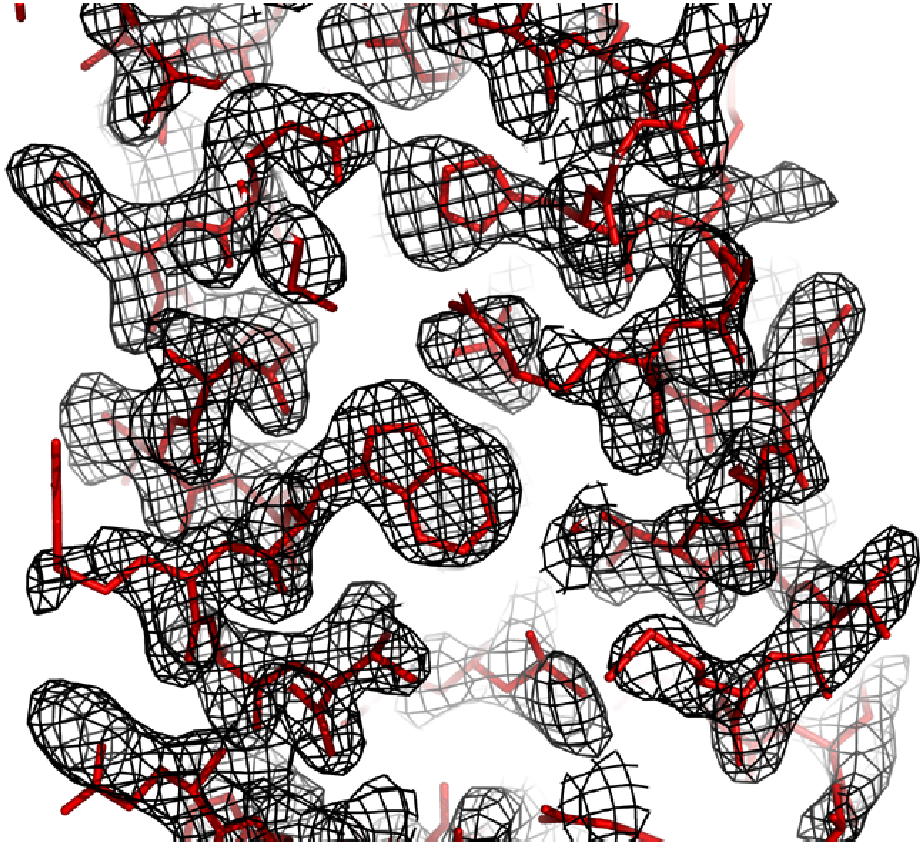
Qian, Raman, Das et al *Nature* (2007)

Do we pass this stringent test ?

Homology model : CASP target T385



Electron density map from starting template



Electron density map from refined model

Bin Qian

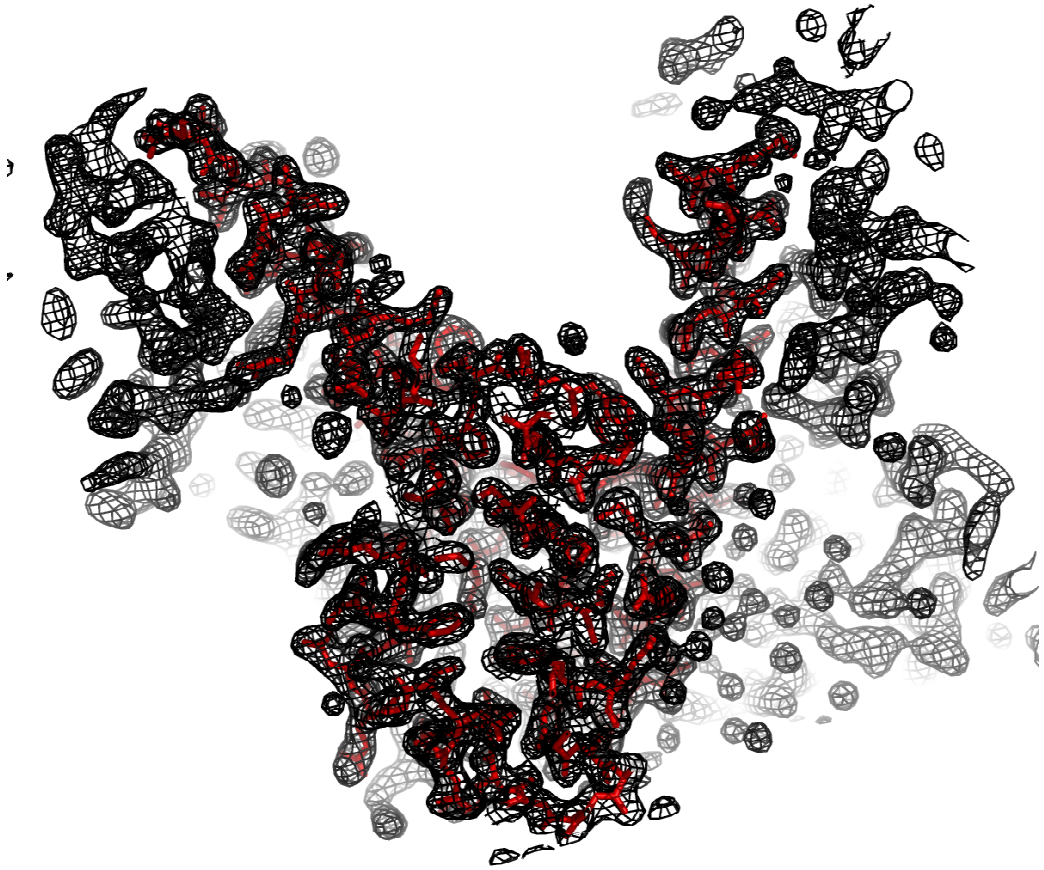
RED STICKS : High resolution X-ray structure

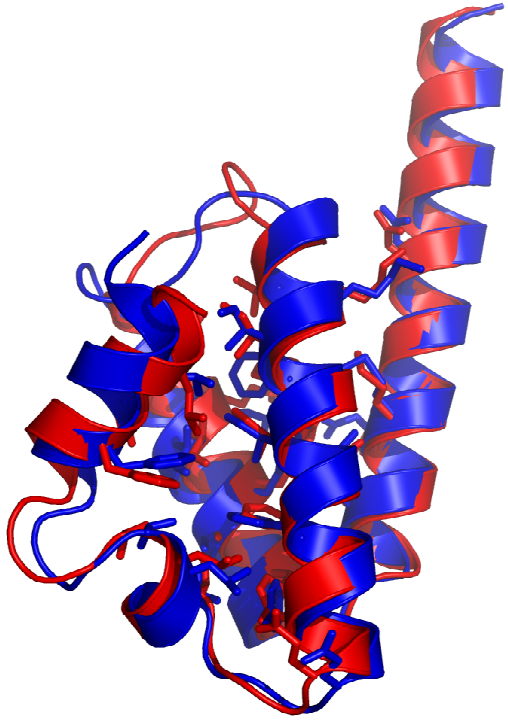
McCoy et al *J. Appl. Cryst.* (2007)

Molecular replacement with an ab-initio model ?

Molecular replacement with an ab-initio model ?

CASP7 target T0283





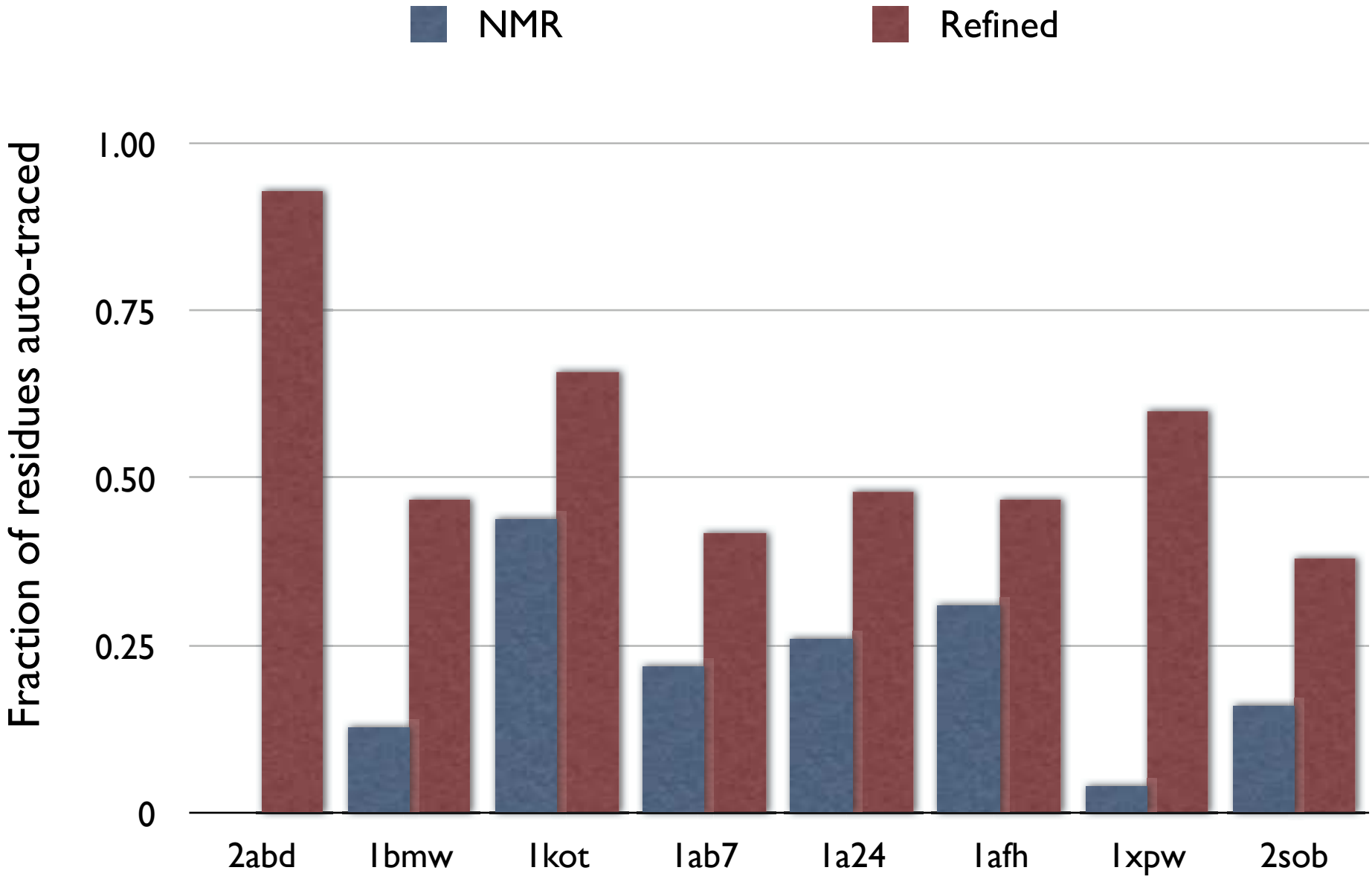
1.4 A RMSD

RED : PDB coordinates from crystal structure solved by experimental phasing

BLACK : Electron density map using phases from ab-initio Rosetta model

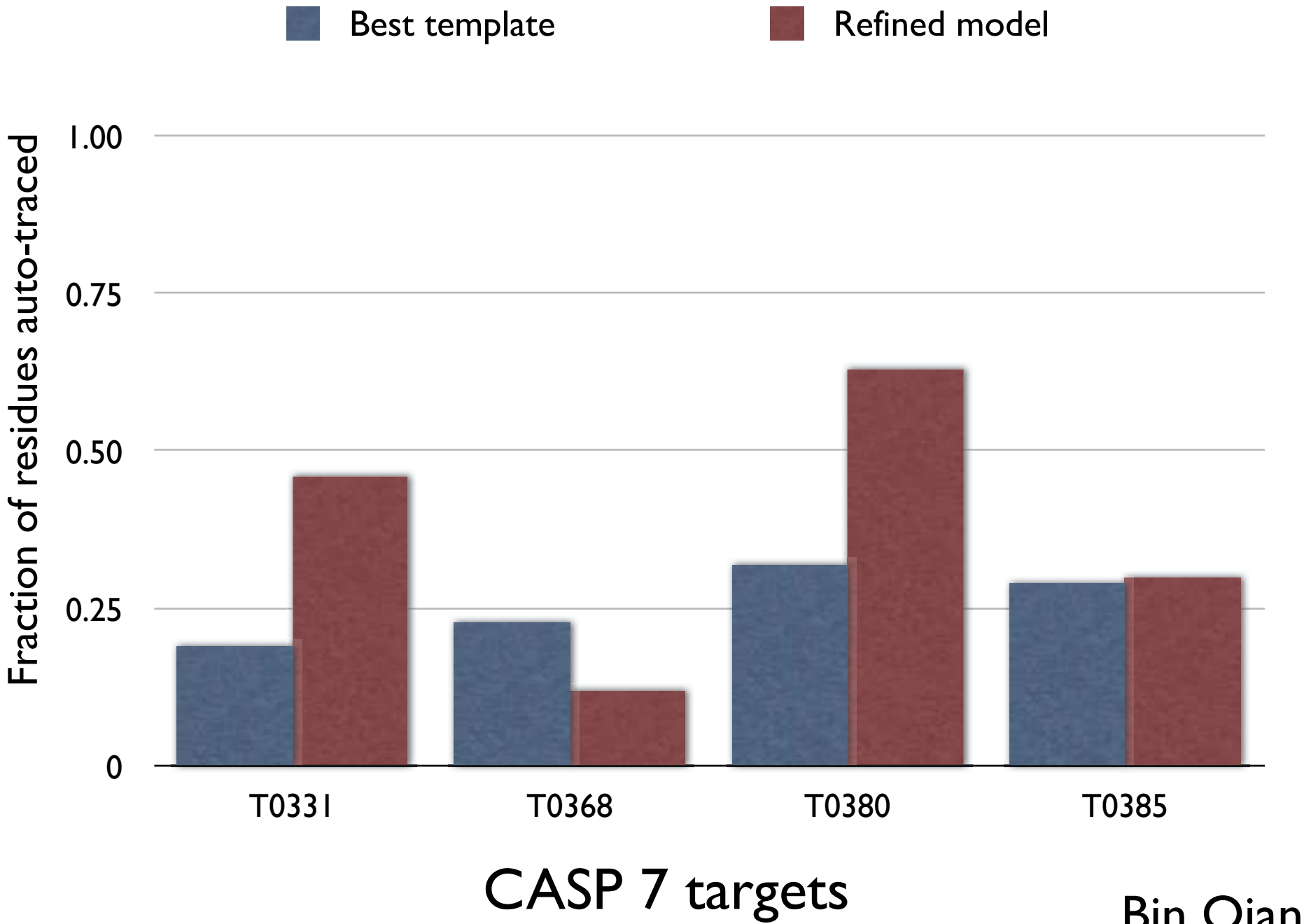
Rhiju Das

Results summary - NMR



Terwilliger *Acta Cryst.D* (2003)
Joosten et al *Acta Cryst.D* (2008)

Results summary - Comparative Modeling



Bin Qian

Conclusions

- Significant advance - the problem is far from being fully solved
- Conformation space sampling - the limiting factor
- Computational methods + limited experimental data = faster, accurate models
- Structural Genomics - covering protein fold - refining homology models will be all important

Acknowledgments

David Baker

STRUCTURE PREDICTION TEAM

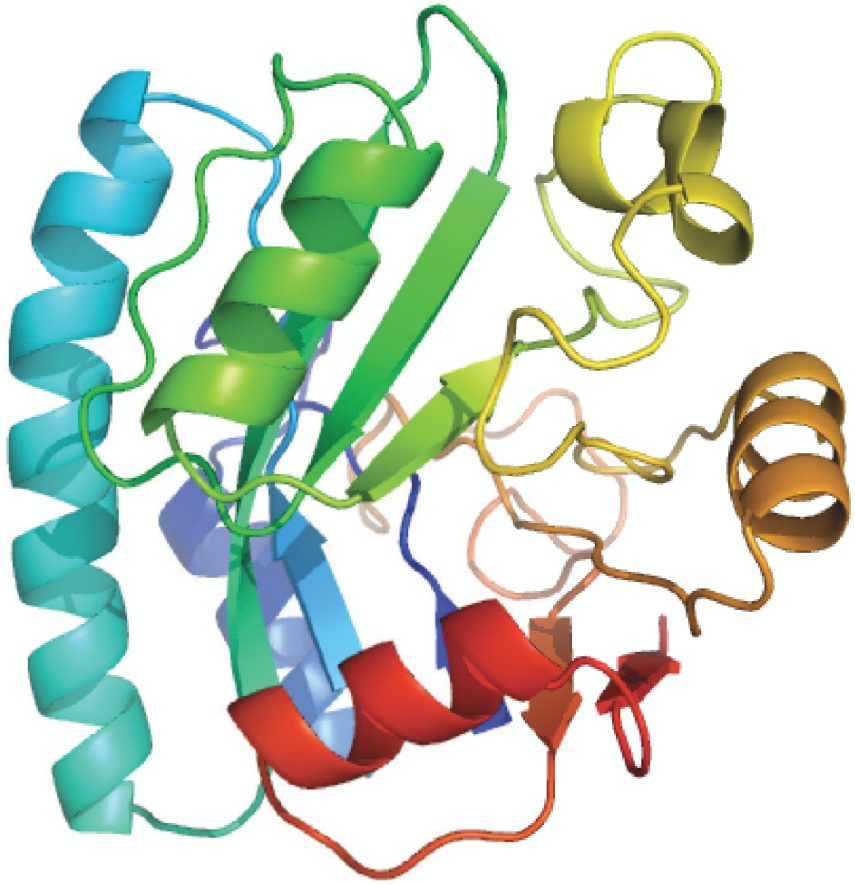
Bin Qian
Rhiju Das
Phil Bradley
Chu Wang
James Thompson
Rob Vernon
Will Sheffler
Liz Kellogg
Frank DiMaio
Oliver Lange
Mike Tyka
David Kim
.... rest of the Baker lab

COMPUTING RESOURCES

- Rosetta@HOME distributed computing volunteers
- IBM Blue Gene T. J. Watson Research Center, Yorktown Heights, NY
- Argonne Leadership Computing Facility, Argonne National Lab, Chicago, IL
- San Diego Supercomputer Center, San Diego, CA
- National Center for Supercomputing Applications, Urbana, IL

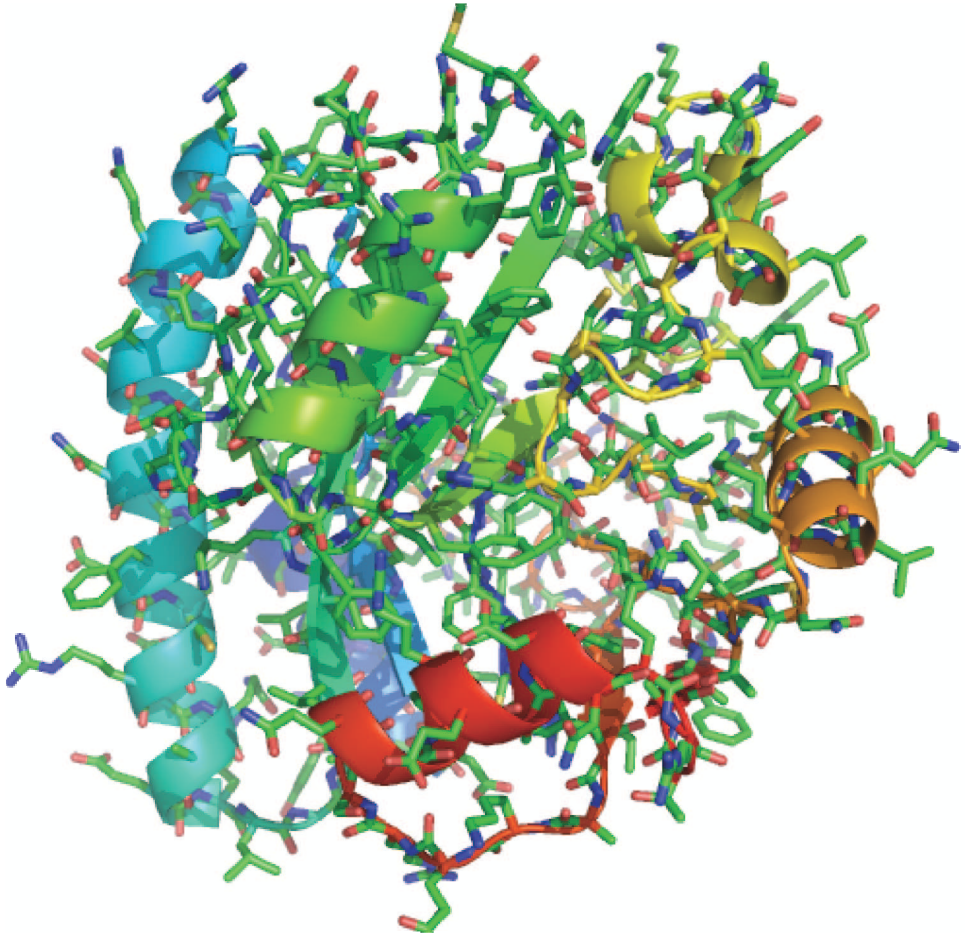
The Protein Folding Problem

GTPDIIVNAQINS
EDENVLDFIIEDE
YYLKKRGVGAHII
KVASSPQLRLLY
KNAYSTVSCGNY
GVLCNLVQNGEY
DLNAIMFNCAEIK
LNKGQMLFQTKI
WR



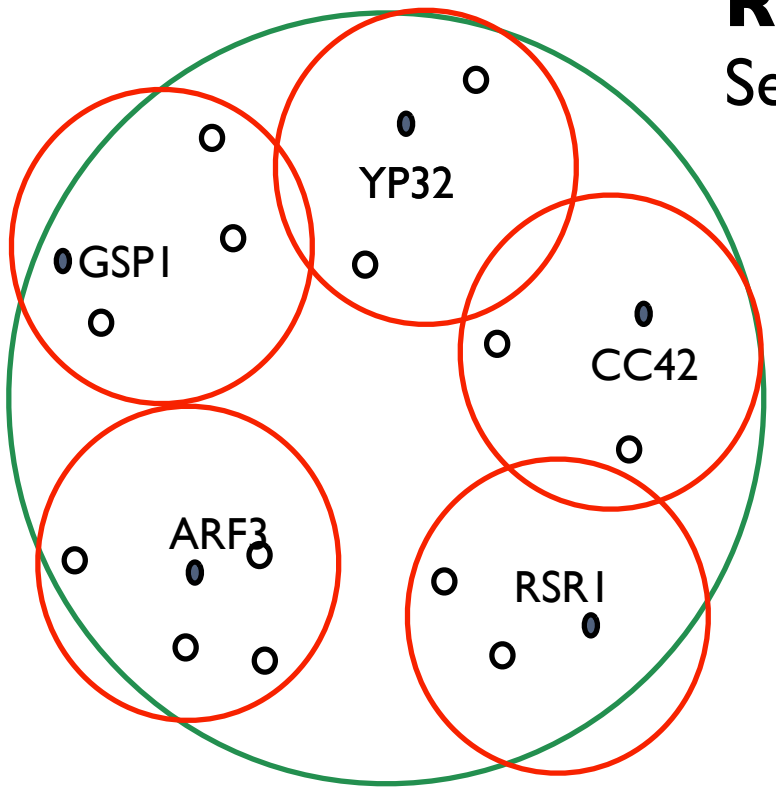
The Protein Folding Problem

GTPDIIVNAQINS
EDENVLDFIIEDE
YYLKKRGVGAHII
KVASSPQLRLLY
KNAYSTVSCGNY
GVLCNLVQNGEY
DLNAIMFNCAEIK
LNKGQMLFQTKI
WR



Solve one, get them all !

- Evolutionarily related proteins - generally same protein fold
- One structure per sequence family or subfamily - rest by HOMOLOGY MODELING
- Use known structure as template



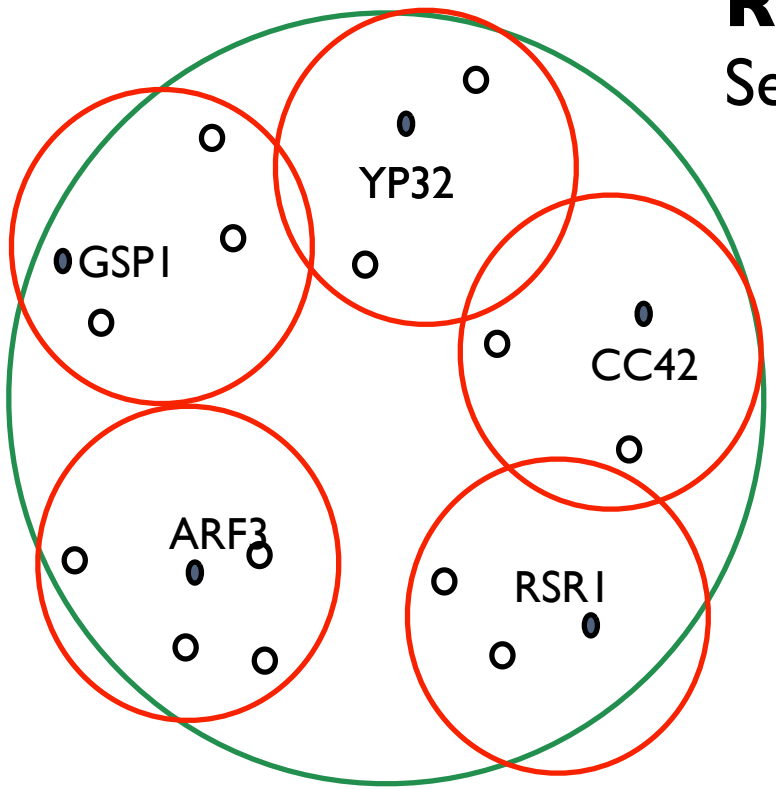
Ras family in yeast

Sequence family and sub-family

Vitkup et al *Nat. Str. Mol. Biol.* (2001), Sellers et al *Proteins* (2008)

Solve one, get them all !

- Evolutionarily related proteins - generally same protein fold
- One structure per sequence family or subfamily - rest by HOMOLOGY MODELING
- Use known structure as template



Ras family in yeast

Sequence family and sub-family

~100 sequences can be modeled using every solved structure as template !

Vitkup et al *Nat. Str. Mol. Biol.* (2001), Sellers et al *Proteins* (2008)

NMR structures and homology models

	NMR structure	Homology model
resolution	generally low (insufficient data)	generally low (depending on seq. similarity)
starting model	ensemble (multiple structures satisfying data)	ensemble (multiple homologs to the target seq.)
core packing	generally poor	generally poor
distance to high res. crystal structure	1-3 A	2-5 A