

the (inverse) protein folding problem — state-of-the-art in computational protein folding and design

Jens Meiler



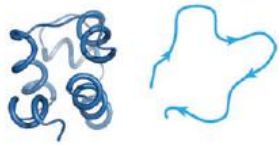
Center for Structural Biology
Departments of Chemistry, Pharmacology, and
Biomedical Informatics



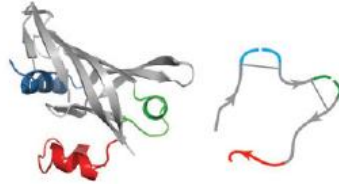
Rosetta: A Unified Framework for Tackling Molecular Modeling



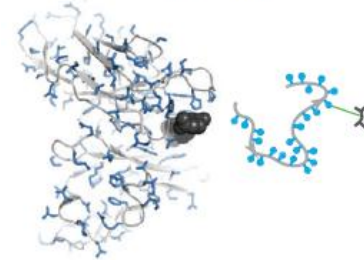
a Protein structure prediction



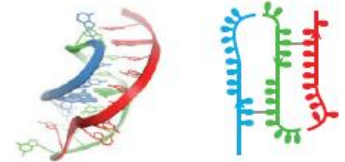
b Loop modeling



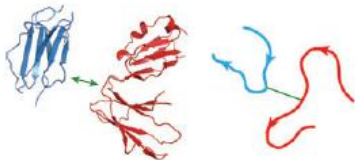
g Small-molecule docking



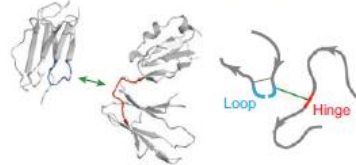
h RNA folding



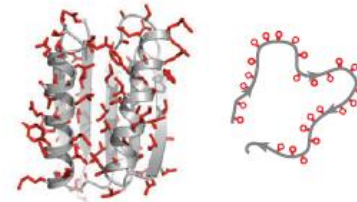
c Protein docking (fully flexible)



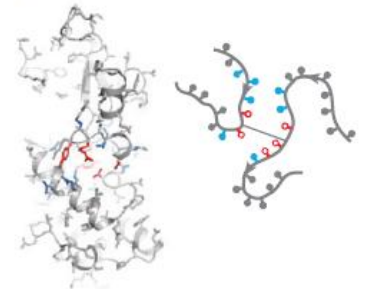
d Protein docking (partly flexible)



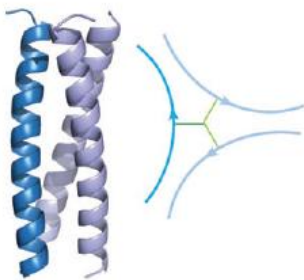
i Protein design



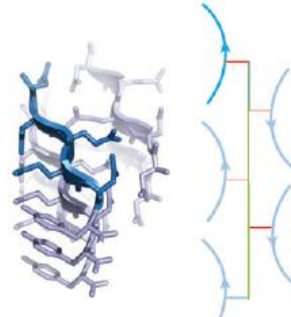
j Protein-protein interface design



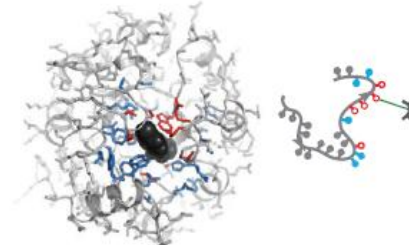
e Symmetric complexes



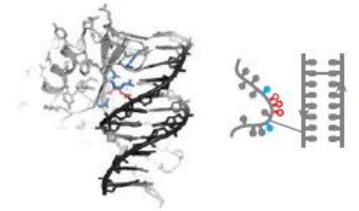
f Fibril modeling



k Enzyme design



l Protein-DNA interface design



Das, R.; Baker, D. "Macromolecular modeling with rosetta" *Annu Rev Biochem* 2008, 77, 363-82.

Rosetta Technology is Developed at more than Ten Laboratories



- Rosetta consists of multiple modules: **protein folding**, **comparative modeling**, **ligand docking**, **protein design**, **antibody/antigen interactions**, etc.
- Rosetta is developed in a consortium of twelve laboratories by around 50 developers
- Rosetta is free for academic use; user guide and tutorials are available
- **PyRosetta** is a python interface that allows integration with Pymol
- **FoldIt** is the better video game for you and your kids
- **Rosetta@home** uses your computer for our research



RosettaCon 2009, Leavenworth, WA, USA

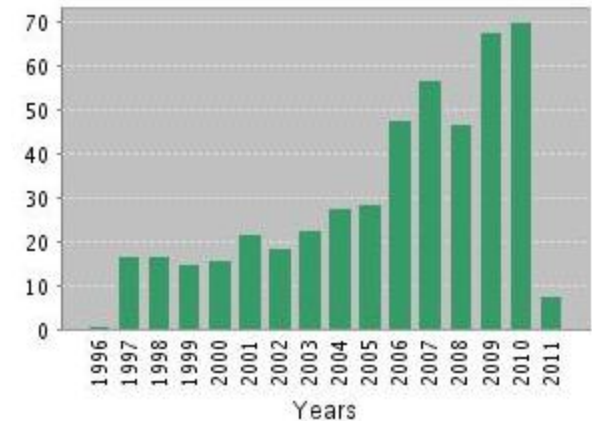
Kaufmann, K. W.; et al. "Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You" *Biochemistry* **2010**.

Rosetta Impact on Field of Protein Prediction and Design over 14 years

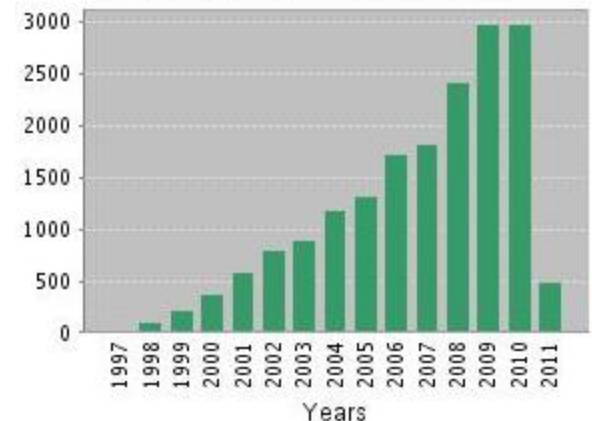


- 495 papers | 18,514 citations | 71 h-index
- Examples of high-impact publications:
 - Simons, K. T.; et al. **"Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions"** J. Mol. Biol. 1997, 268, 209-225. (496 citations)
 - Kuhlman, B.; et al. **"Design of a Novel Globular Protein Fold with Atomic Level Accuracy"** Science 2003, 302, 1364-1368. (440 citations)
 - Gray, J. J.; et al. **"Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations"** J Mol Biol 2003, 331, 281-99. (263 citations)
 - Kortemme, T.; et al. **"A simple physical model for binding energy hot spots in protein-protein complexes"** PNAS 2002, 99, 14116-21. (255 citations)
 - Jiang, L.; et al. **"De novo computational design of retro-aldol enzymes"** Science 2008, 319, 1387-91. (179 citations)

Published Items in Each Year



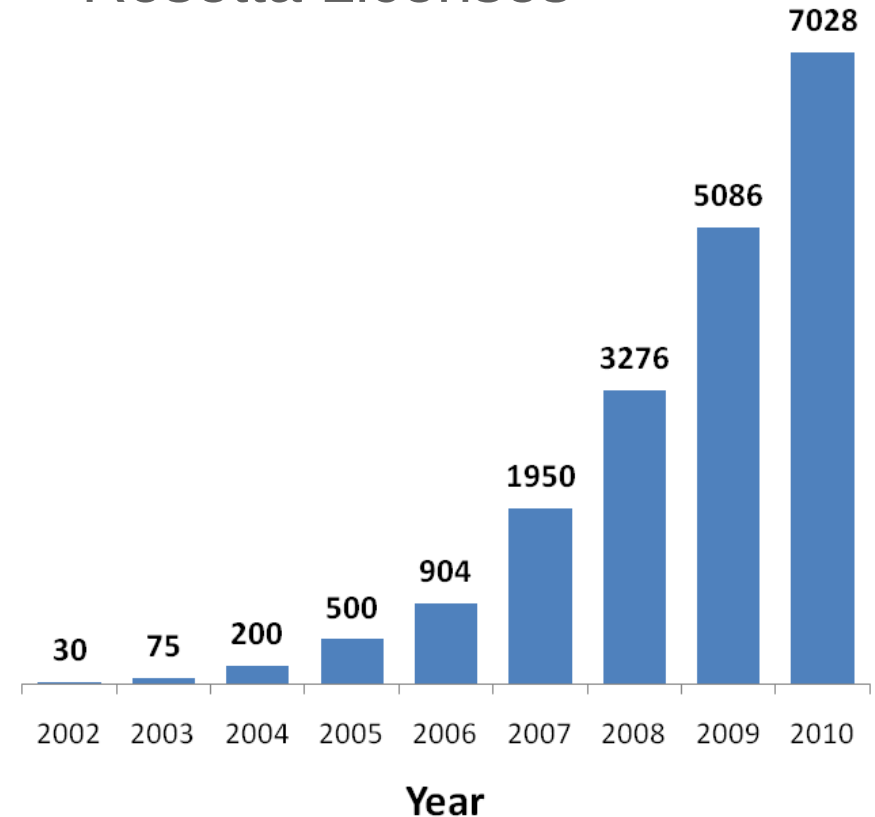
Citations in Each Year



Number of Rosetta Licenses Reached 7000 at the End of 2010



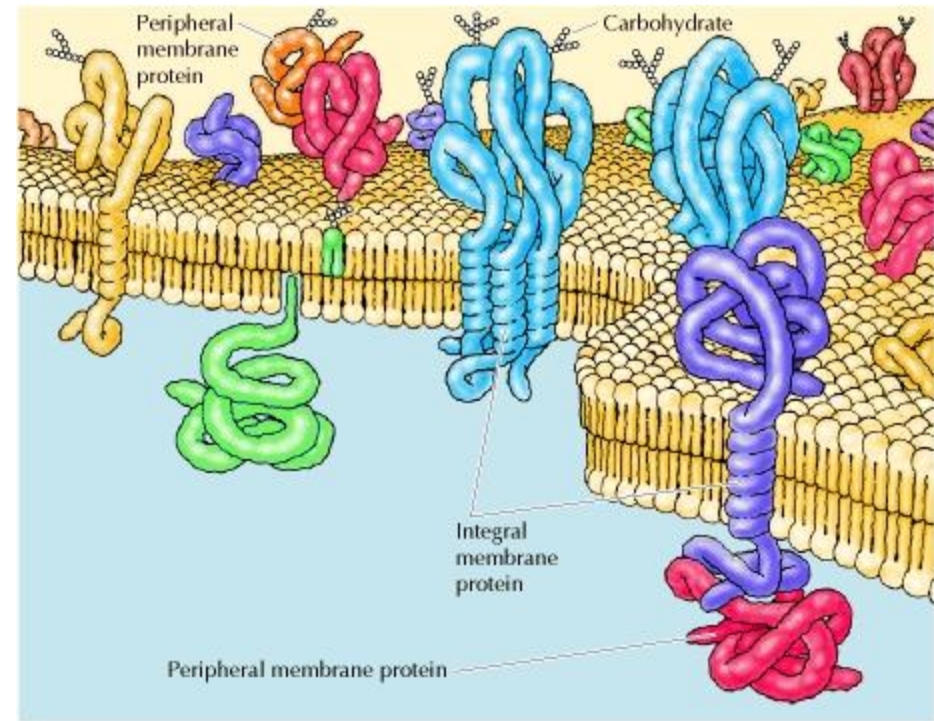
- 249 Organizations in the US
- Licensees in 61 countries
- Cumulative Number of Rosetta Licenses



Integral Membrane Proteins (IMPs) link the Cell to its Environment



- 25% of the human genome are IMPs – around 1000 distinct folds!



www.biopleasure.wordpress.com

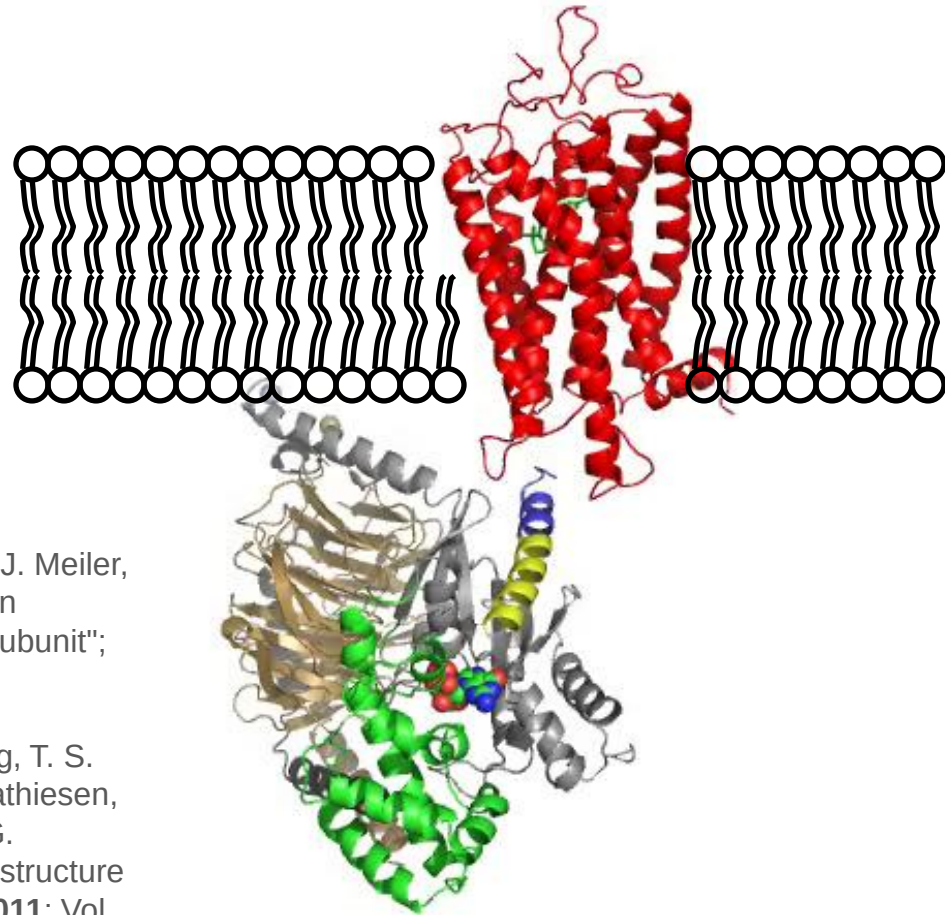
Integral Membrane Proteins (IMPs) link the Cell to its Environment



- 25% of the human genome are IMPs – around 1000 distinct folds!
- 50% of therapeutics target IMPs – G-Protein Coupled Receptors, Ion Channels, Transporter proteins

N. Van Eps, A. M. Preininger, N. Alexander, A. I. Kaya, S. Meier, J. Meiler, H. E. Hamm and W. L. Hubbell; "Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit"; *Proc Natl Acad Sci U S A*; **2011**; Vol. 108 (23): p. 9420-4.

S. G. Rasmussen, B. T. DeVree, Y. Zou, A. C. Kruse, K. Y. Chung, T. S. Kobilka, F. S. Thian, P. S. Chae, E. Pardon, D. Calinski, J. M. Mathiesen, S. T. Shah, J. A. Lyons, M. Caffrey, S. H. Gellman, J. Steyaert, G. Skiniotis, W. I. Weis, R. K. Sunahara and B. K. Kobilka; "Crystal structure of the beta2 adrenergic receptor-Gs protein complex"; *Nature*; **2011**; Vol. 477 (7366): p. 549-55.

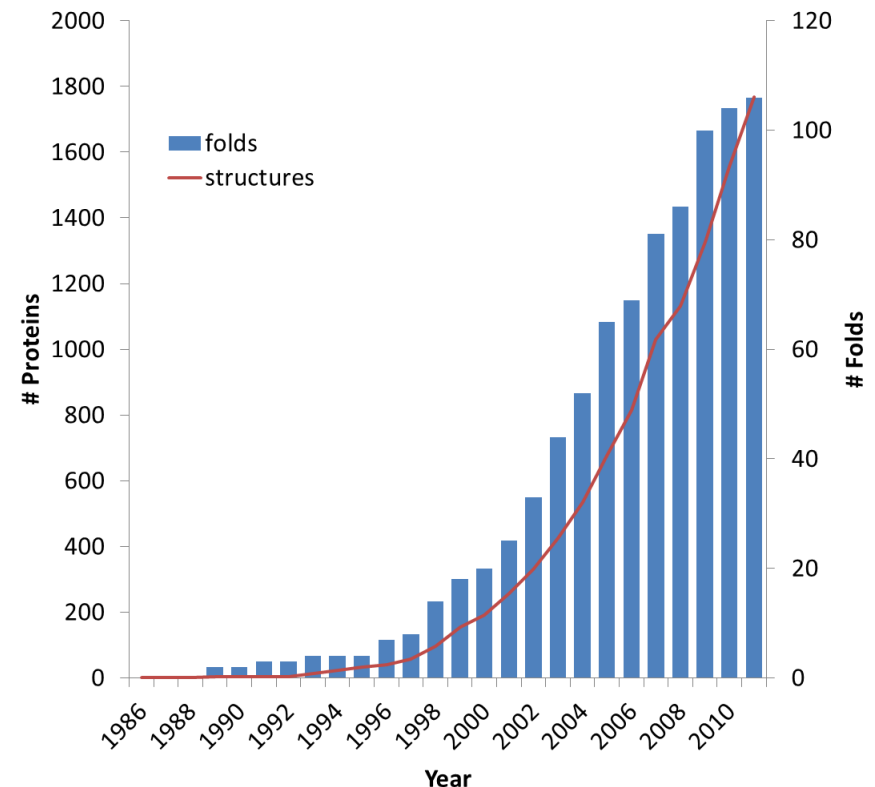


Integral Membrane Proteins (IMPs) link the Cell to its Environment



- 25% of the human genome are IMPs – around 1000 distinct folds!
- 50% of therapeutics target IMPs – G-Protein Coupled Receptors, Ion Channels, Transporter proteins
- But only 2% of protein structures in PDB are IMPs

- #IMP Structures and Folds in the Protein Data Bank

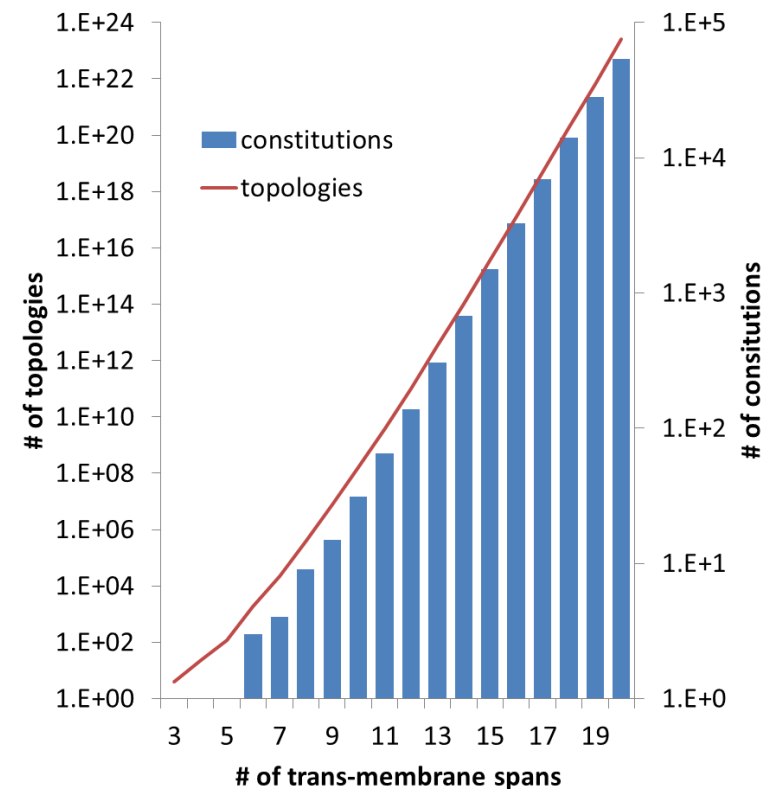


Integral Membrane Proteins (IMPs) link the Cell to its Environment



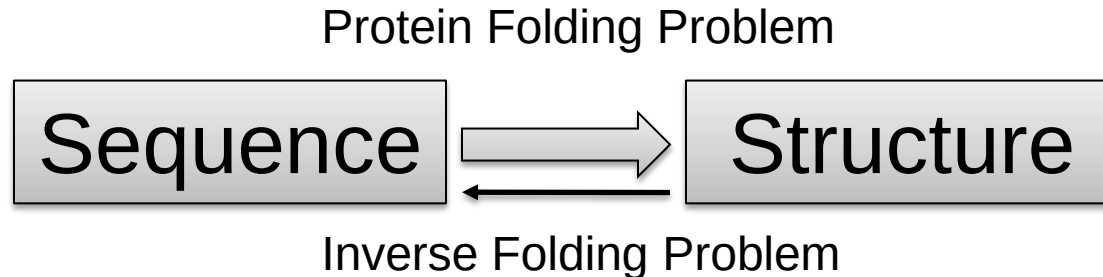
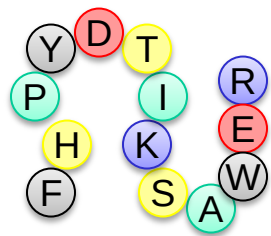
- 25% of the human genome are IMPs – around 1000 distinct folds!
- 50% of therapeutics target IMPs – G-Protein Coupled Receptors, Ion Channels, Transporter proteins
- But only 2% of protein structures in PDB are IMPs
- But only 100 distinct IMP folds have been determined (~10%)

■ #IMP Topologies



(Inverse) Protein Folding Problem

Holy Grail of Comp. Struct. Biology



- 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 $\Rightarrow 10^{200}$ 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 $\Rightarrow 10^{200}$ possible conformations!
- Earth is less than 10^{10} years old.

Outline



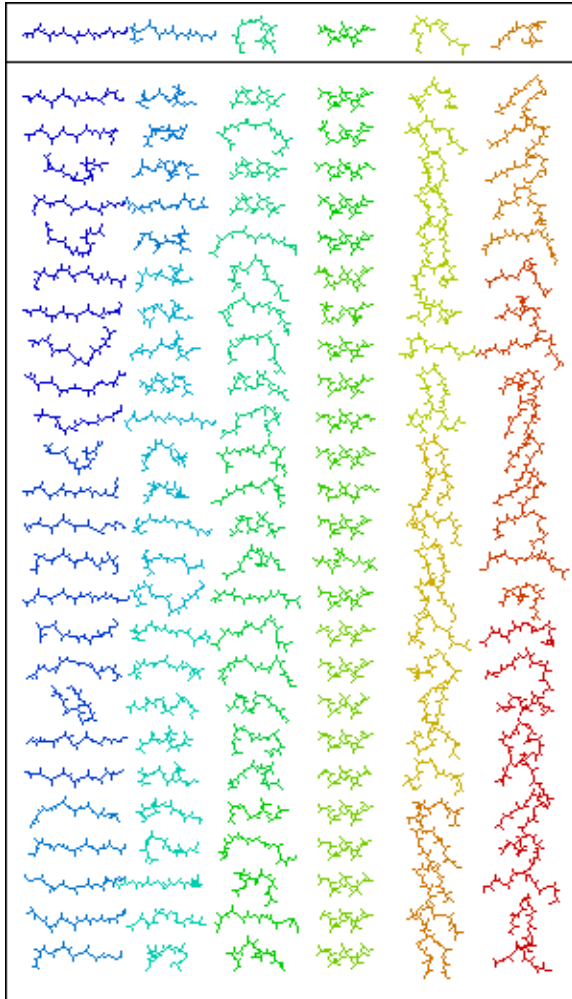
- *De novo* protein folding and the **Critical Assessment of Techniques for Protein Structure Prediction (CASP)**
 - Toward high-resolution *de novo* structure prediction for small proteins
 - Building atomic-detail protein models from sparse and low resolution experimental data
-
- TOP7 – Design of a novel globular protein fold with atomic level accuracy
 - FLR – A computationally designed TIM-barrel protein mimics missing link in evolution of symmetric superfolds

Outline



- *De novo* protein folding and the **Critical Assessment of Techniques for Protein Structure Prediction (CASP)**
 - Toward high-resolution *de novo* structure prediction for small proteins
 - Building atomic-detail protein models from sparse and low resolution experimental data
-
- TOP7 – Design of a novel globular protein fold with atomic level accuracy
 - FLR – A computationally designed TIM-barrel protein mimics missing link in evolution of symmetric superfolds

Local Sequence Bias – Rapid Approximation of Local Interactions

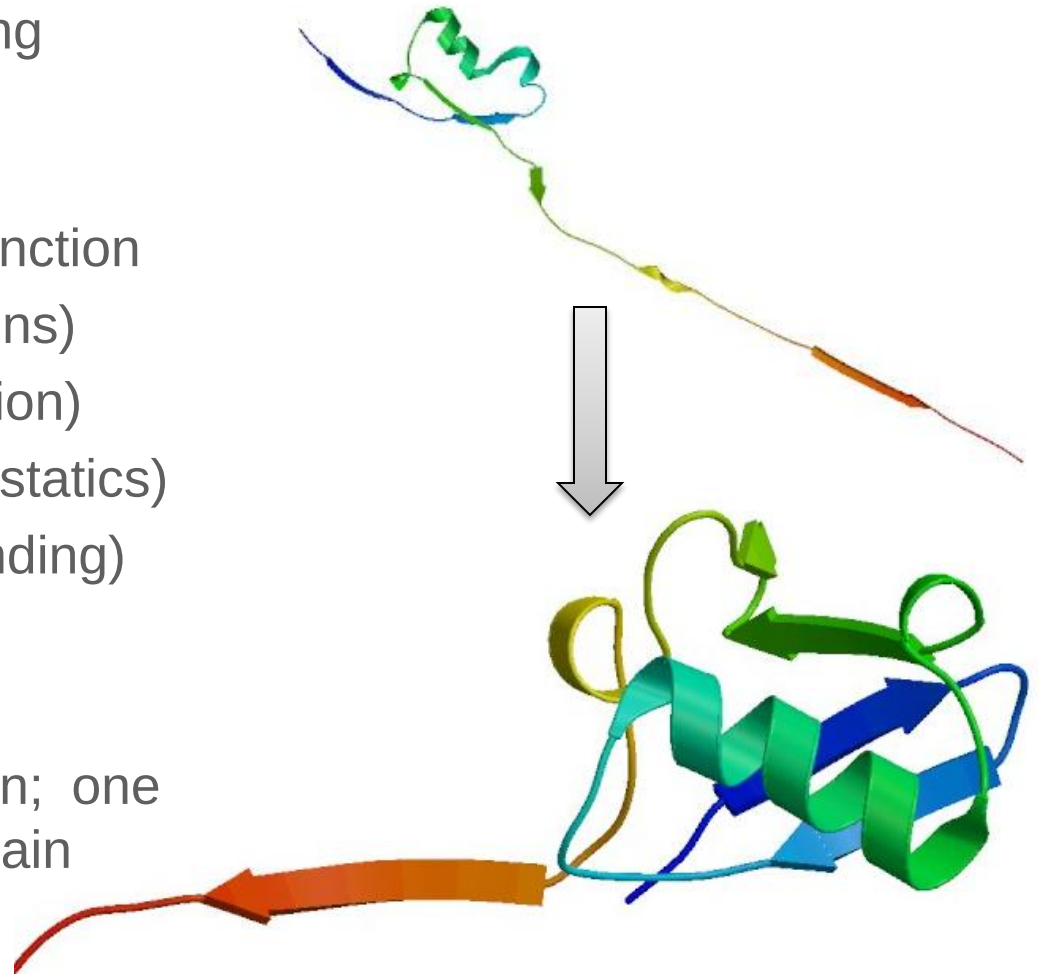


- While not every protein fold is present in the protein databank, all possible conformations of small peptides are!
- Approximate local interactions using the distribution of conformations seen for similar sequences in known protein structures
- For each sequence window, select fragments that represent the conformations sampled during folding

Non-local Interactions Govern Protein Folding Process



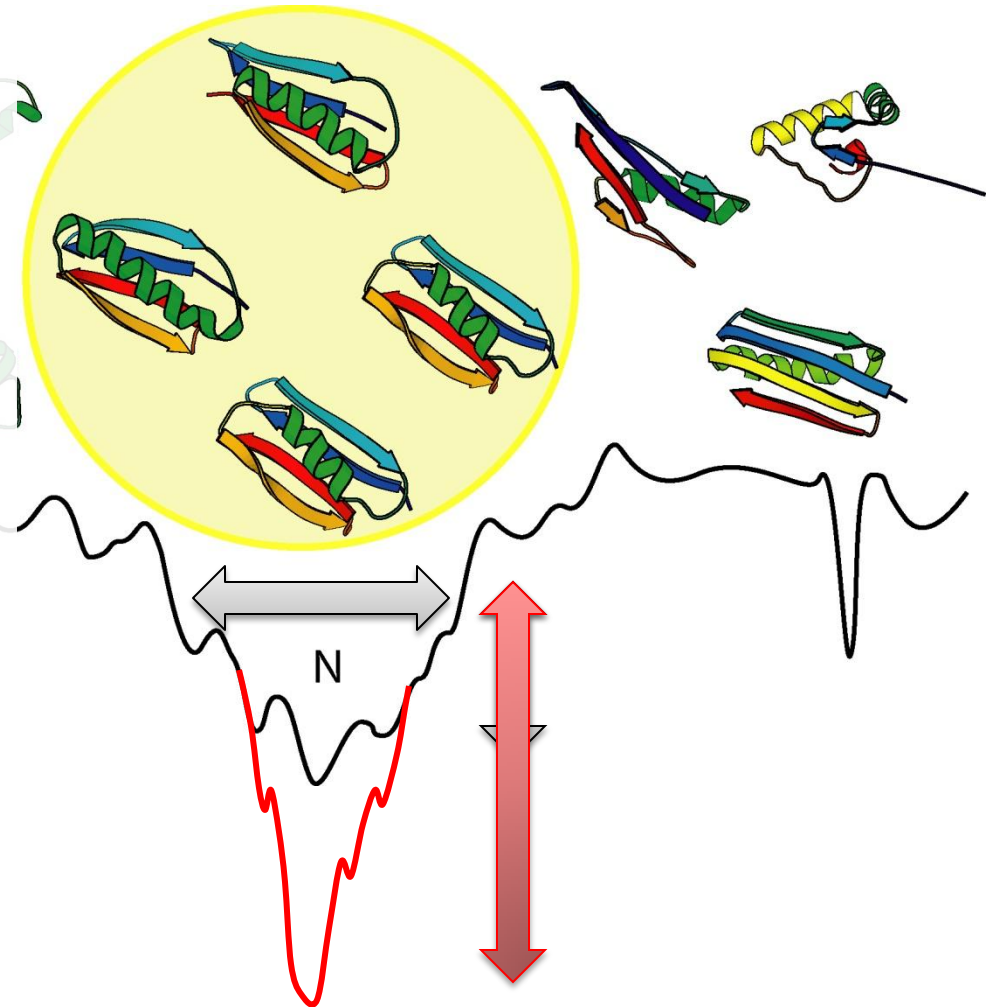
- Monte Carlo simulated annealing assembly of fragments
- Statistically-derived potential function
 - Steric overlap (vdw interactions)
 - Residue environment (solvation)
 - Pairwise interactions (electrostatics)
 - Strand pairing (hydrogen bonding)
 - Compactness (solvation)
- Simplified protein representation;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

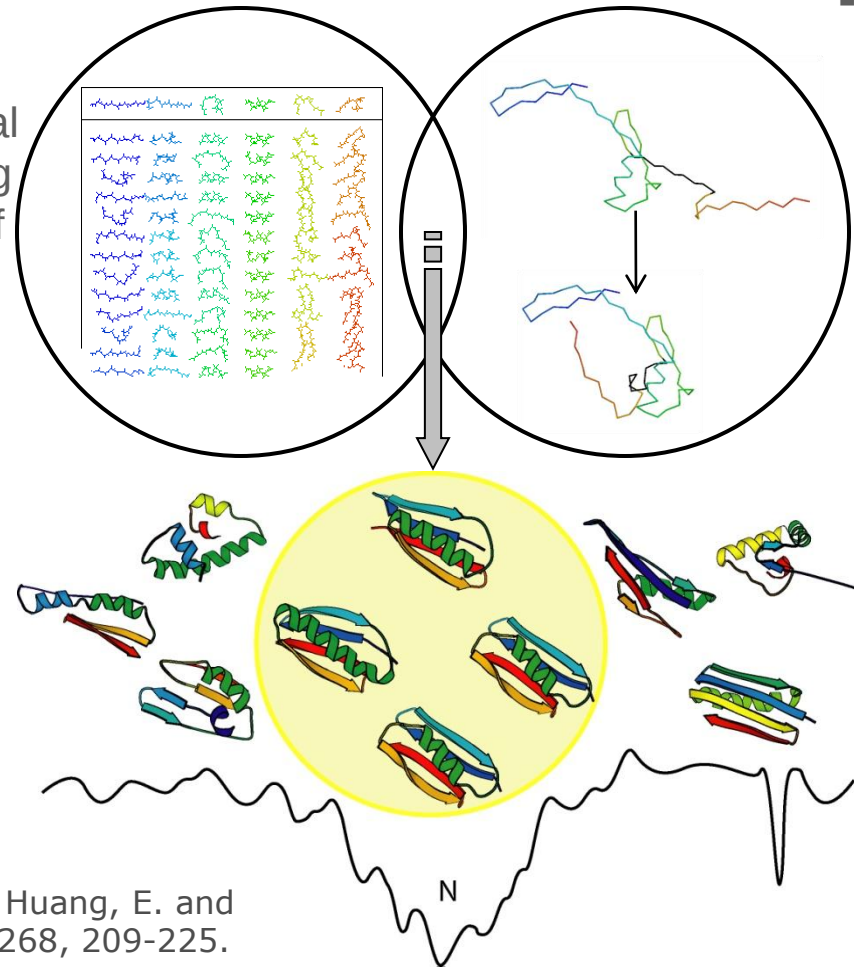


- Local Sequence Bias

- Approximate local interactions using the distribution of conformations seen for similar sequences in known protein structures

- Monte Carlo simulations

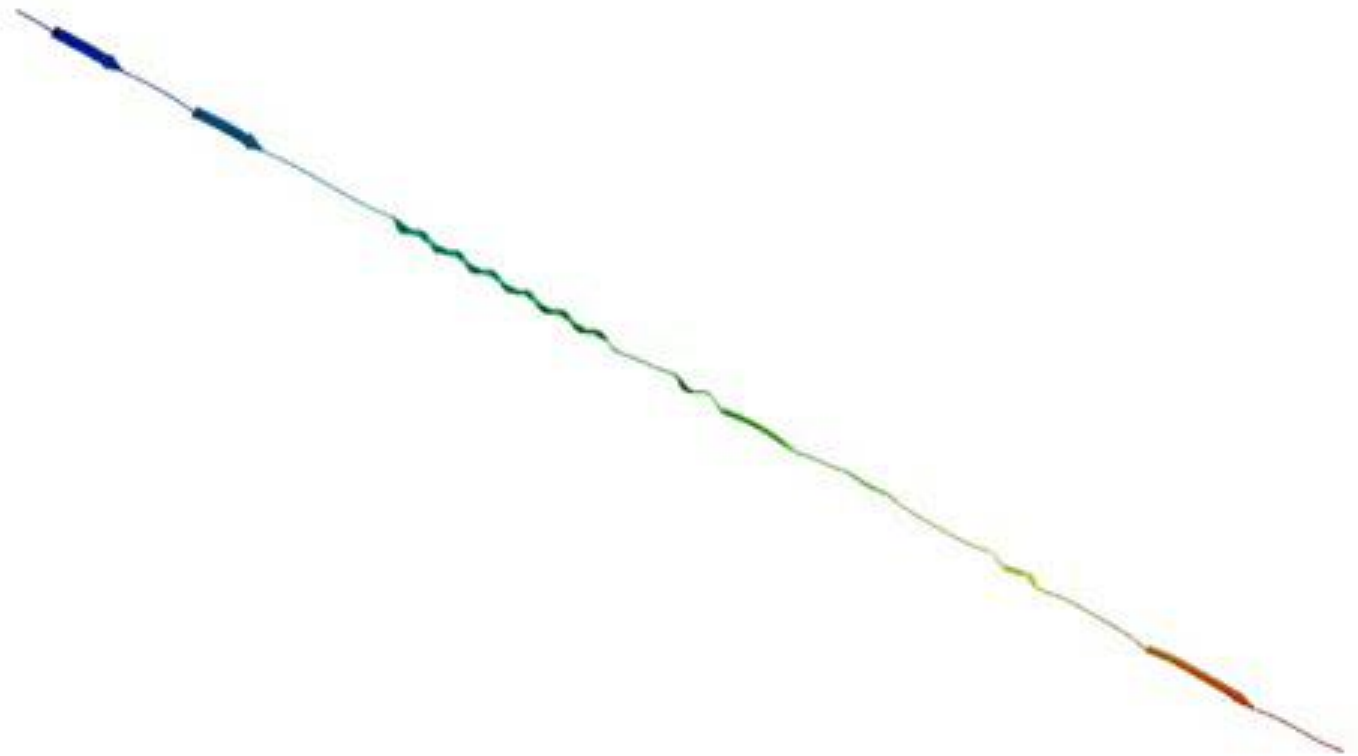
- Select broadest minima using cluster analysis



- Energy evaluation of non-local interactions using knowledge-based energy function

- Steric overlap
- Residue environment
- Pair wise interactions
- Strand pairing
- Compactness
- Secondary Structure Packing

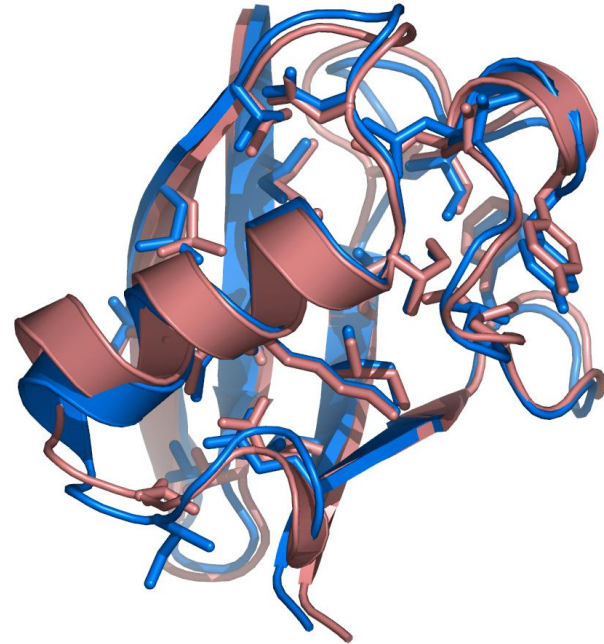
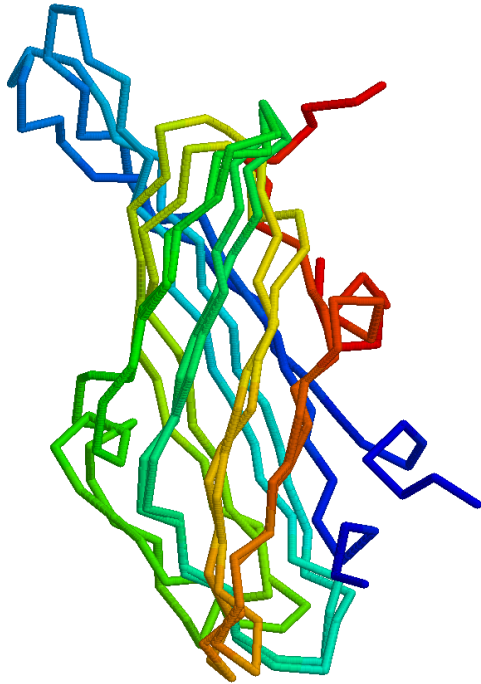
Simons, K. T., Kooperberg, C., Huang, E. and Baker, D. (1997) *J. Mol. Biol.*, 268, 209-225.



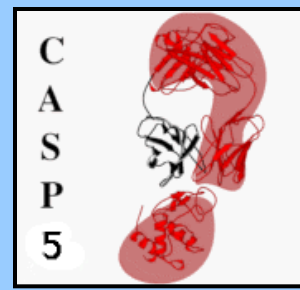
Major Challenges in the Field of Protein Structure Prediction



- 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



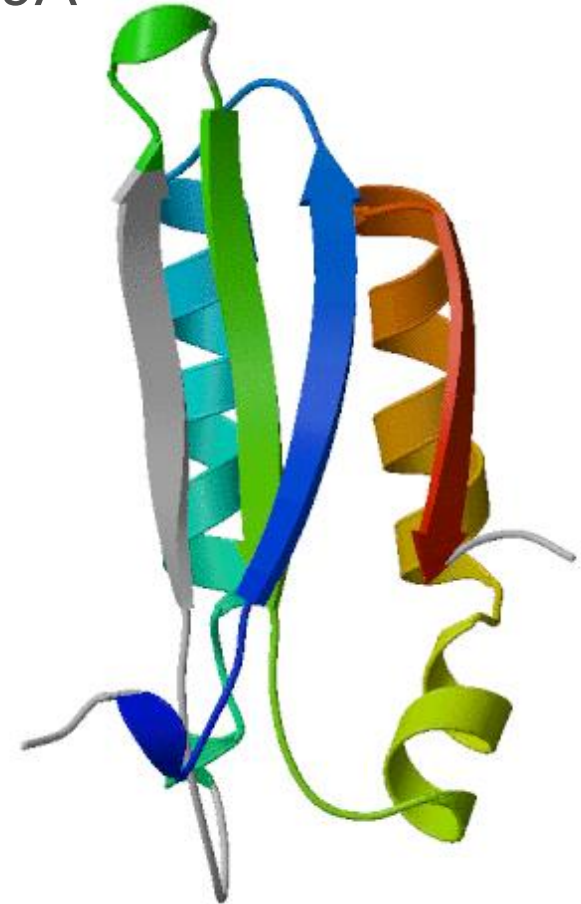
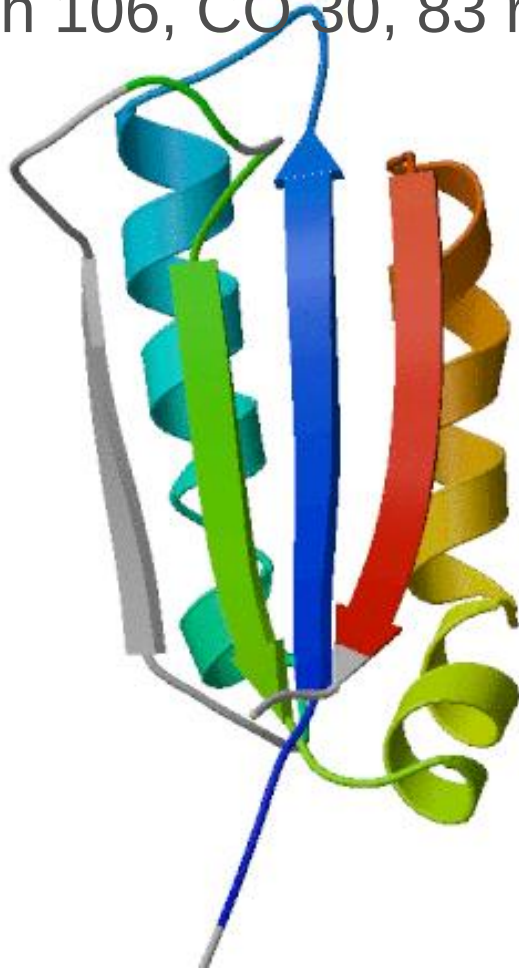
Critical Assessment of Techniques for Protein Structure Prediction



- 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

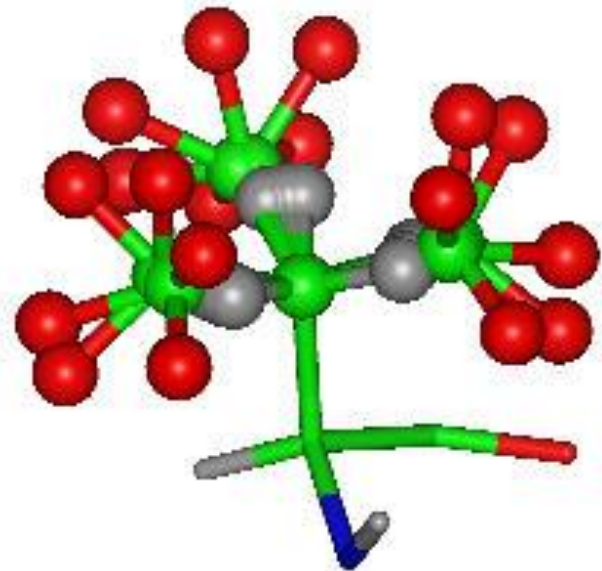
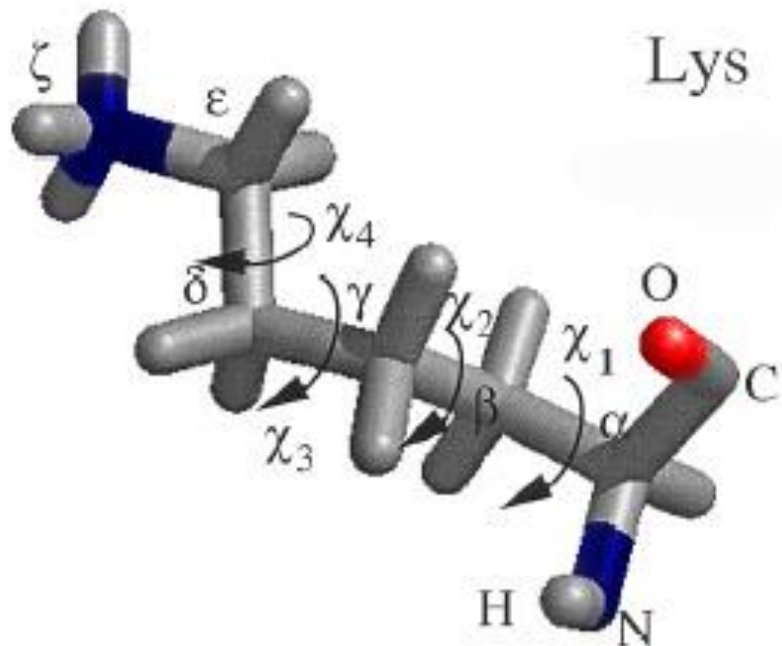


Outline



- *De novo* protein folding and the **Critical Assessment of Techniques for Protein Structure Prediction (CASP)**
 - Toward high-resolution *de novo* structure prediction for small proteins
 - Building atomic-detail protein models from sparse and low resolution experimental data
-
- TOP7 – Design of a novel globular protein fold with atomic level accuracy
 - FLR – A computationally designed TIM-barrel protein mimics missing link in evolution of symmetric superfolds

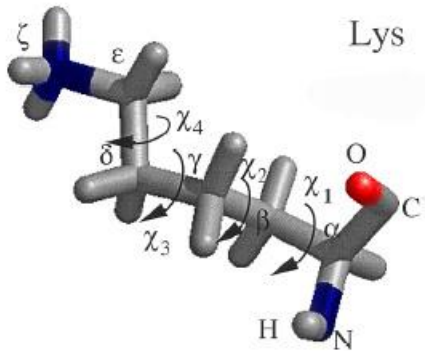
Sidechain Degrees of Freedom



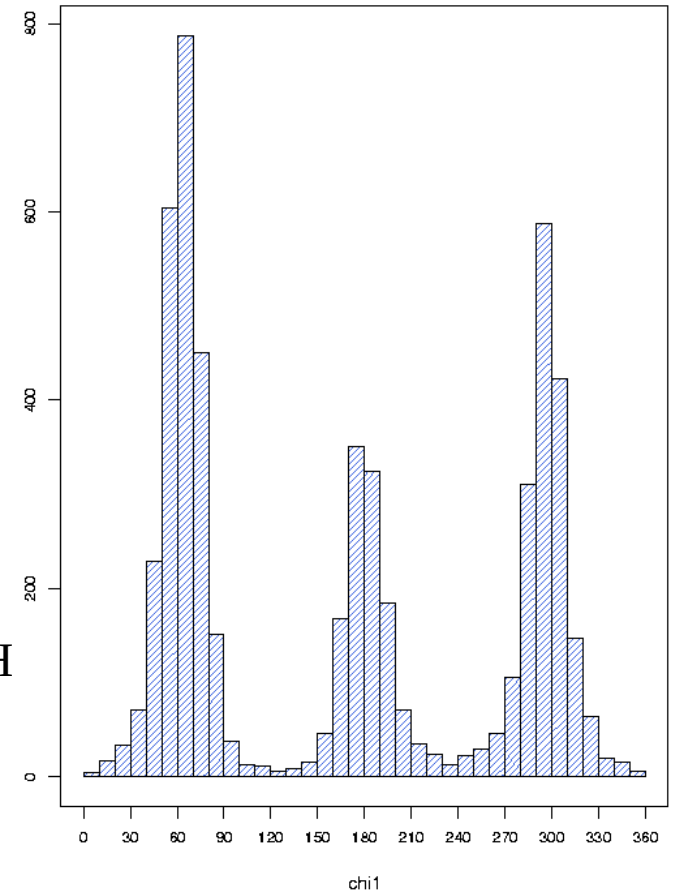
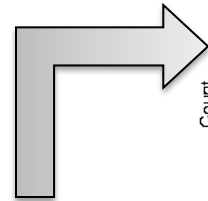
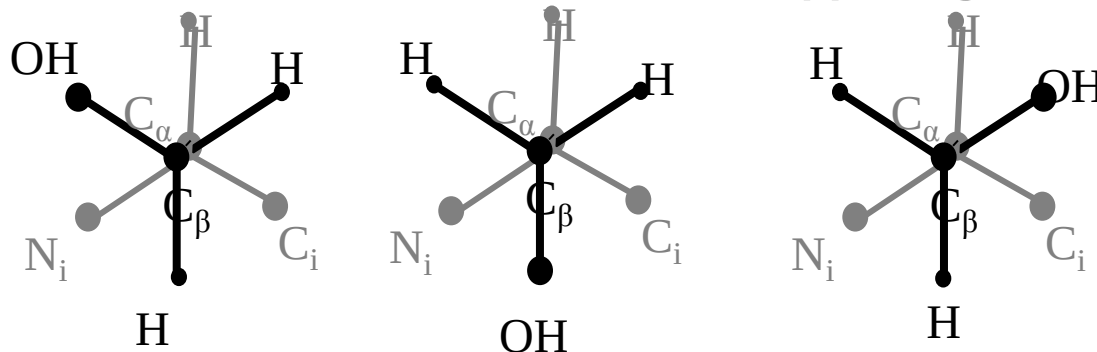
All Likely Side Chain Conformations are Present in the Protein Databank



- 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	$p \chi_1$	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

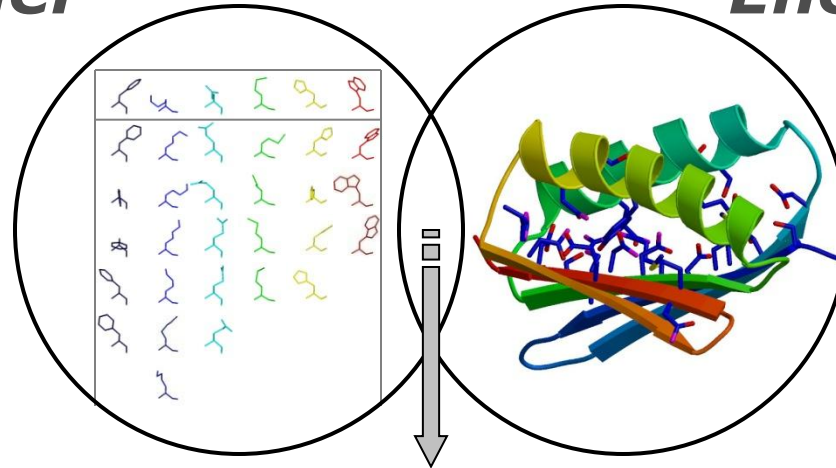
Dunbrack, R. L.; Cohen, F. E. "Bayesian statistical analysis of protein side-chain rotamer preferences." *Protein Sci.* **1997**, *6*, 1661-1681.

Sampling and Scoring for Side Chain Repacking and Design



Local Rotamer Bias

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

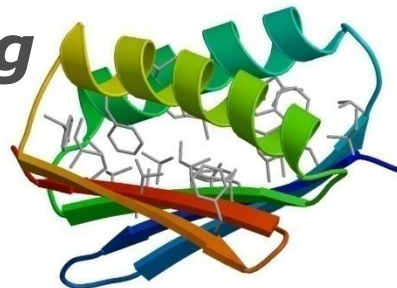


Energy function

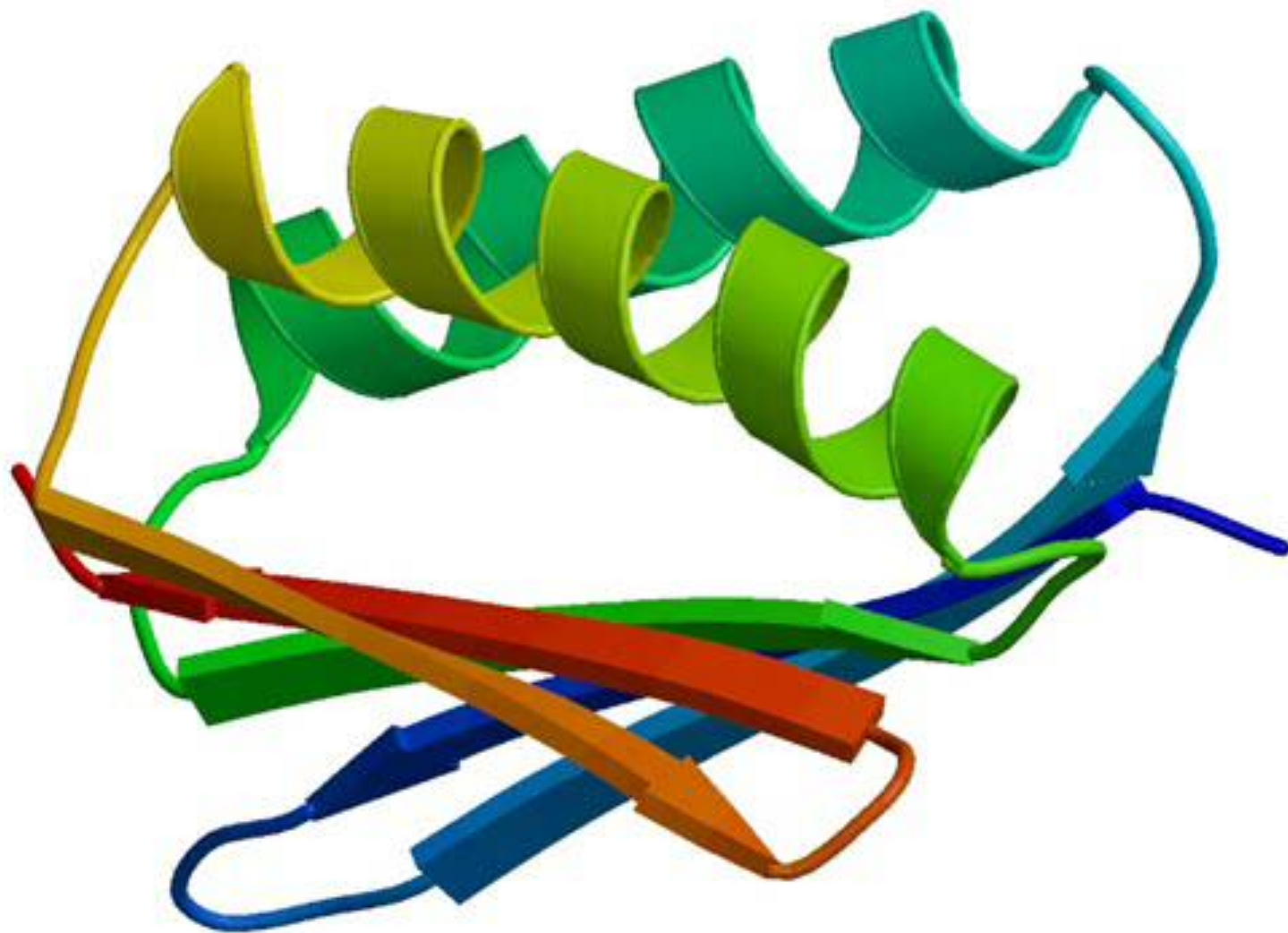
Statistically derived potential function

- VDW interaction
- solvation
- hydrogen bonding potential
- pair wise interactions
- rotamer probability

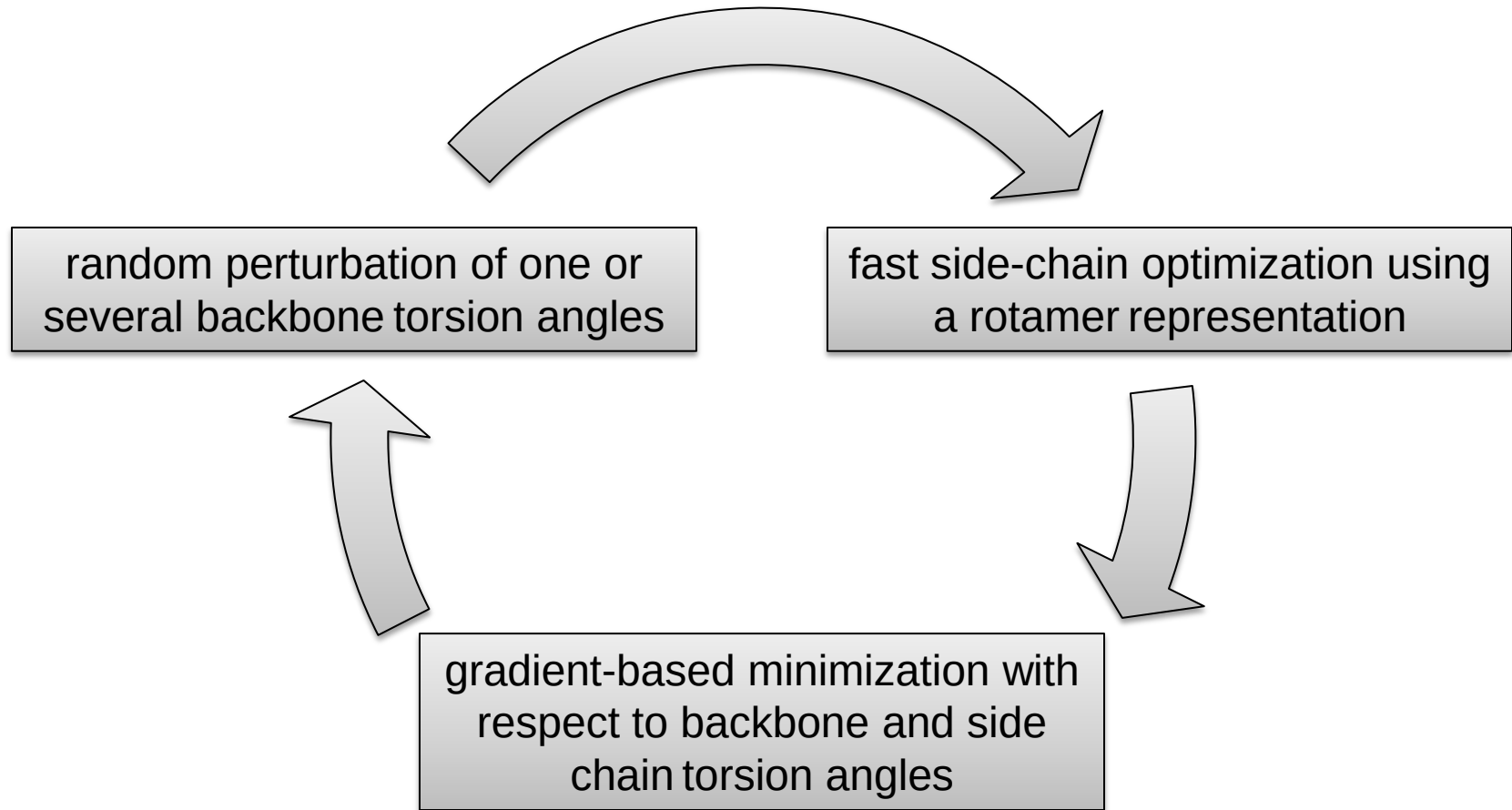
Simulated Annealing Monte Carlo energy minimization



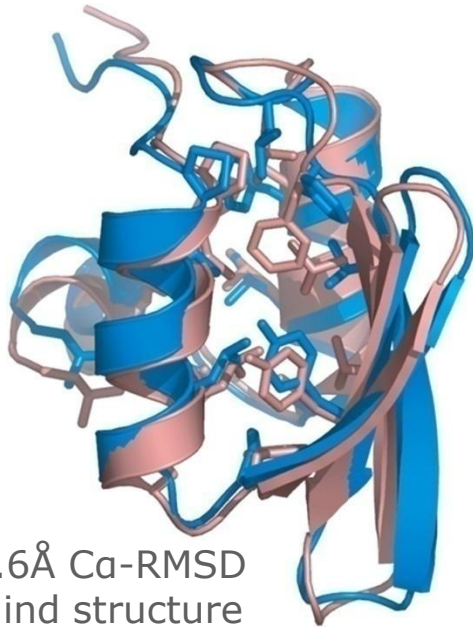
Dahiyat, B. I. and Mayo, S. L. (1997) *Science*, 278, 82-7
Dunbrack, R. L., Jr. and Karplus, M. (1993) *J Mol Biol*, 230, 543-74.
Kuhlman, B., et. al. (2003) *Science*, 302, 1364-1368.



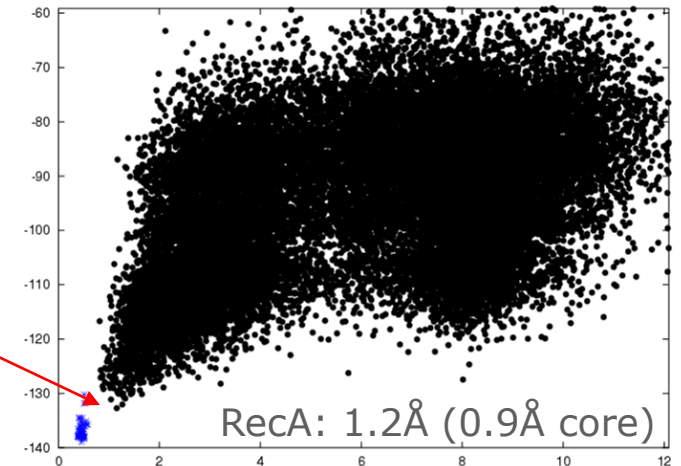
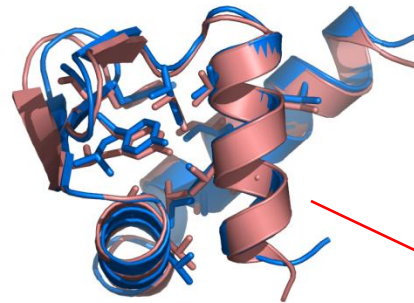
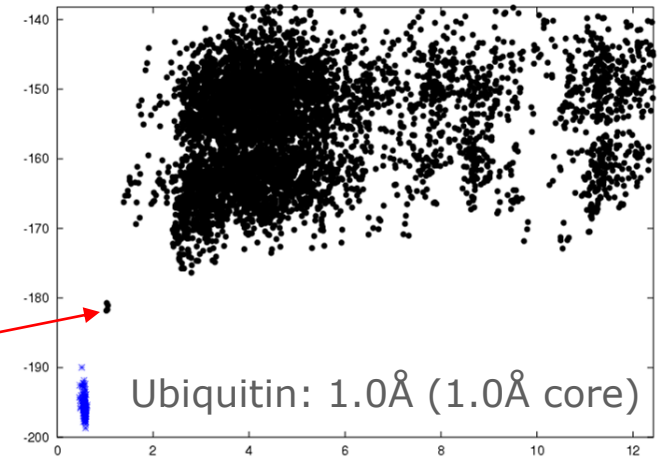
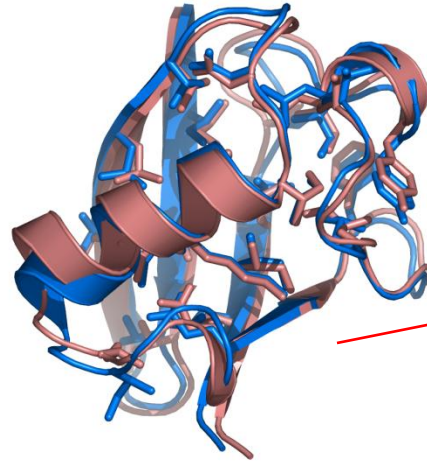
Refinement Cycle with Side Chain Repacking and All Atom Minimization



CASP target T0281 and other Benchmark Examples



1.6Å Ca-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	Ca	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

Outline

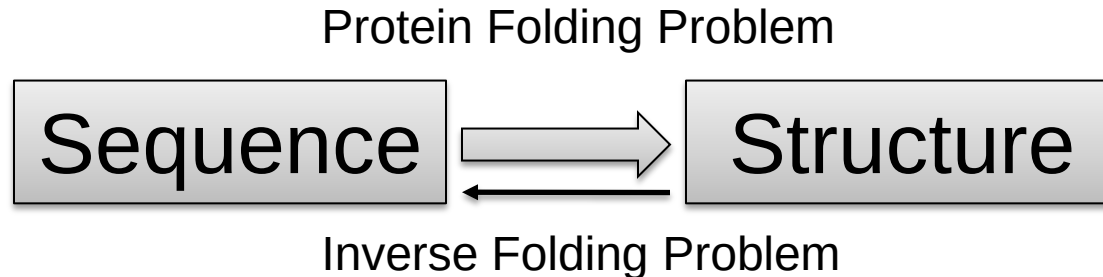
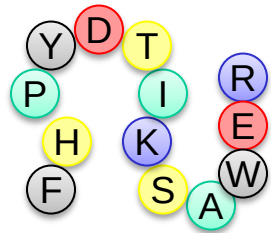


- *De novo* protein folding and the **Critical Assessment of Techniques for Protein Structure Prediction (CASP)**
 - Toward high-resolution *de novo* structure prediction for small proteins
 - Building atomic-detail protein models from sparse and low resolution experimental data
-

- TOP7 – Design of a novel globular protein fold with atomic level accuracy
- FLR – A computationally designed TIM-barrel protein mimics missing link in evolution of symmetric superfolds

(Inverse) Protein Folding Problem

Holy Grail of Comp. Struct. Biology



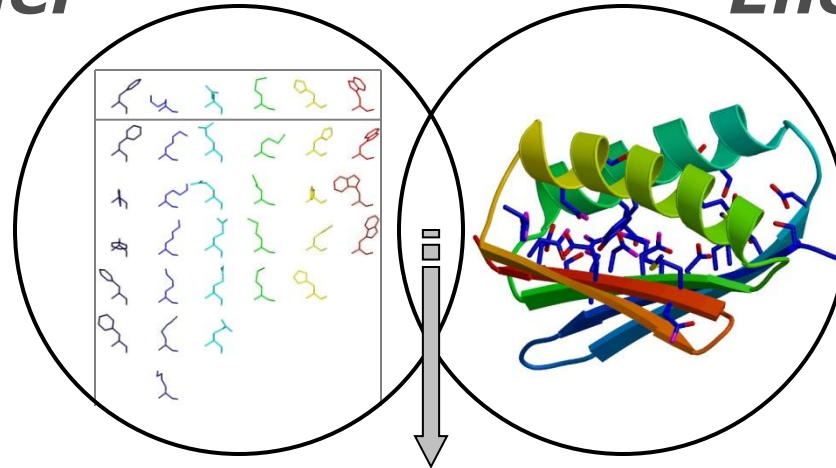
- Given a protein's AA sequence, what is its 3-dimensional fold, and how does it get there?
- Assume 100 conformations for each amino acid in a 100 amino acid protein $\Rightarrow 10^{200}$ possible conformations!
- Cyrus Levinthal's paradox of protein folding, 1968.
- Given a protein fold, which primary 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 $\Rightarrow 10^{200}$ 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 conformations (rotamers) seen in known protein structures

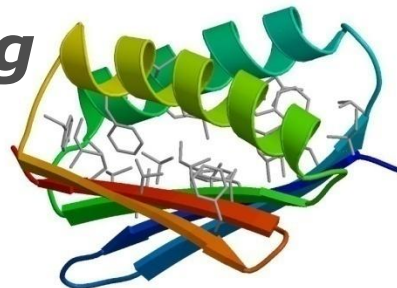


Energy function

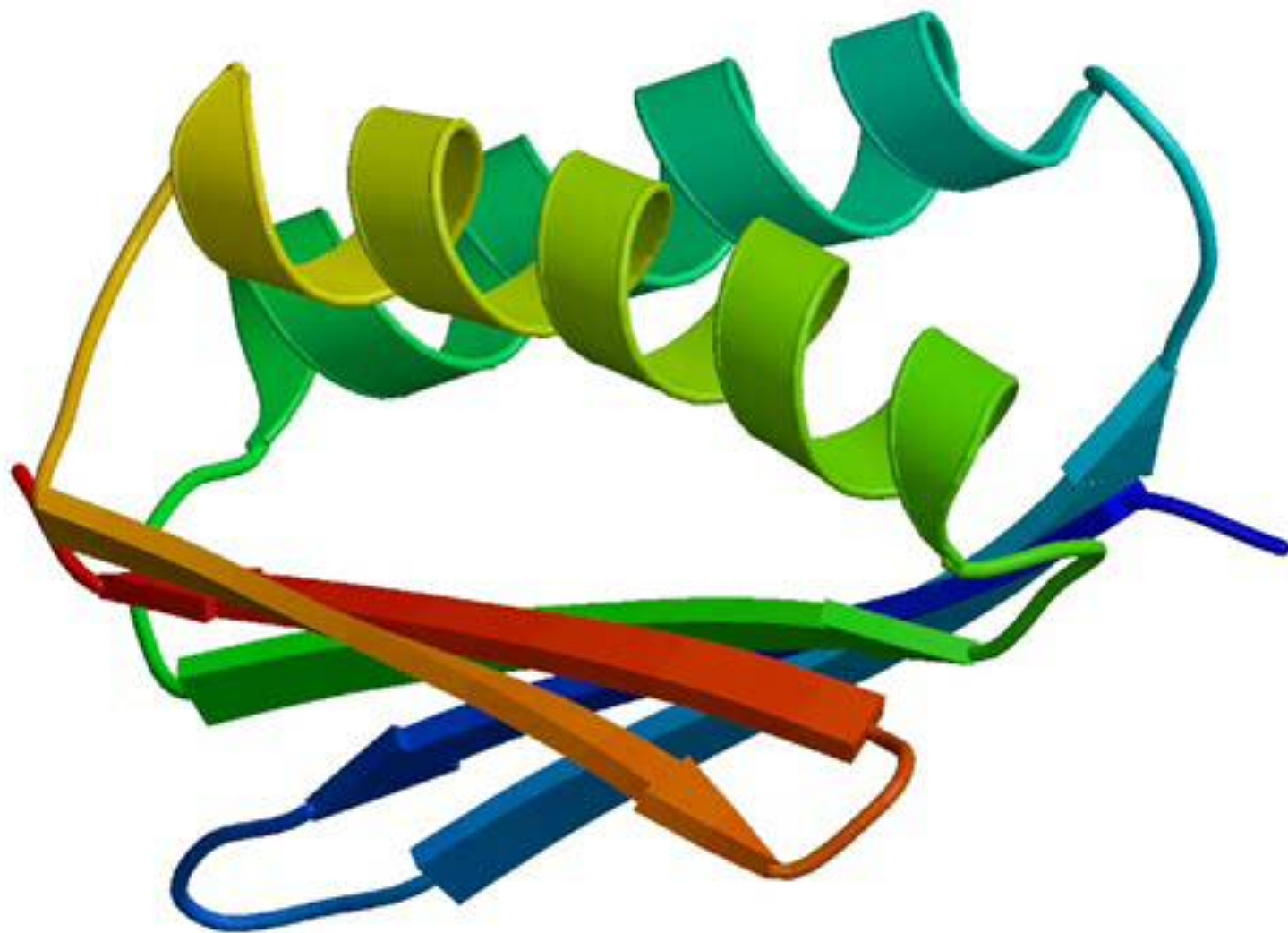
Statistically derived potential function

- VDW interaction
- solvation
- hydrogen bonding potential
- pair wise interactions
- rotamer probability

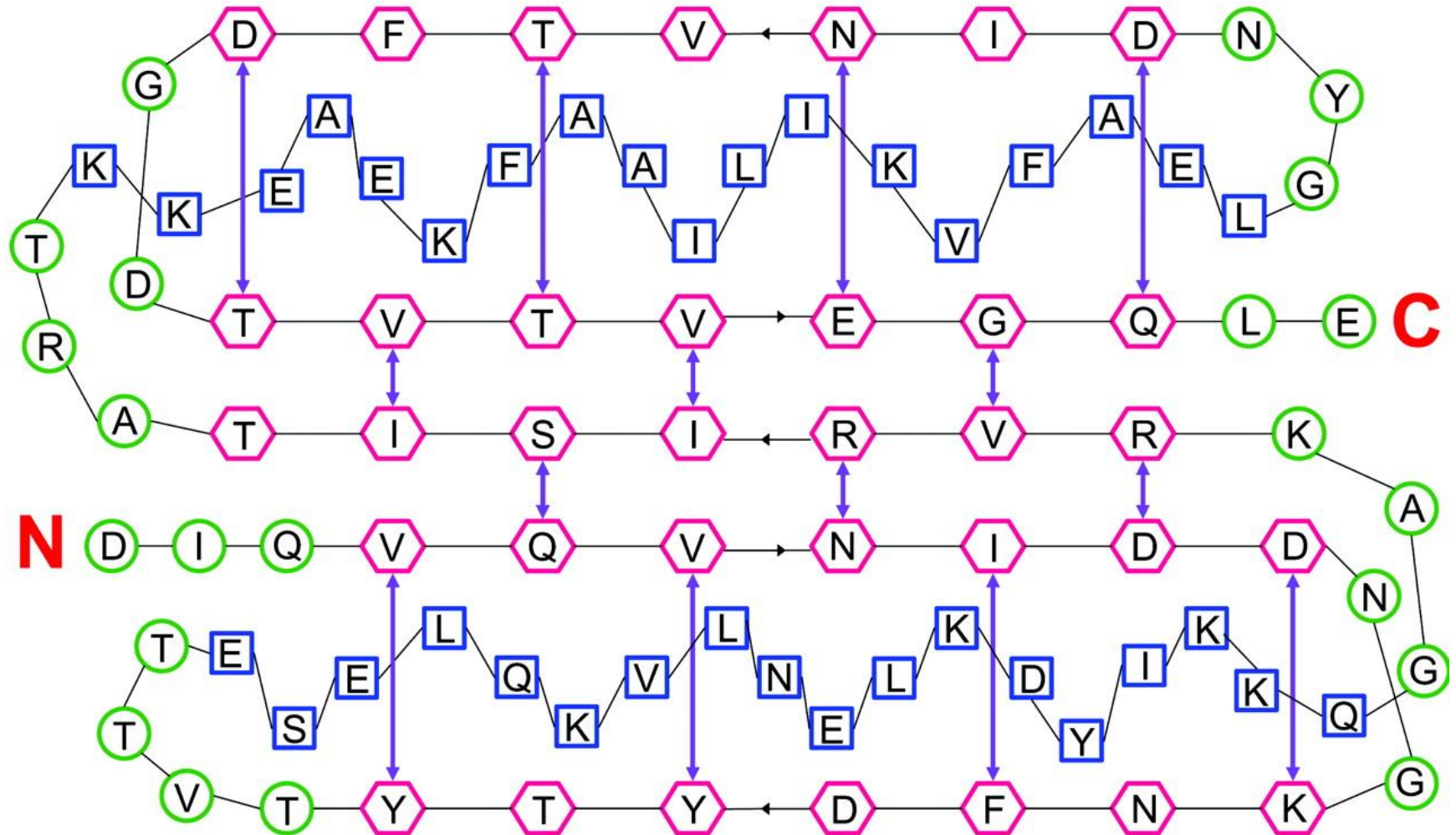
Simulated Annealing Monte Carlo energy minimization



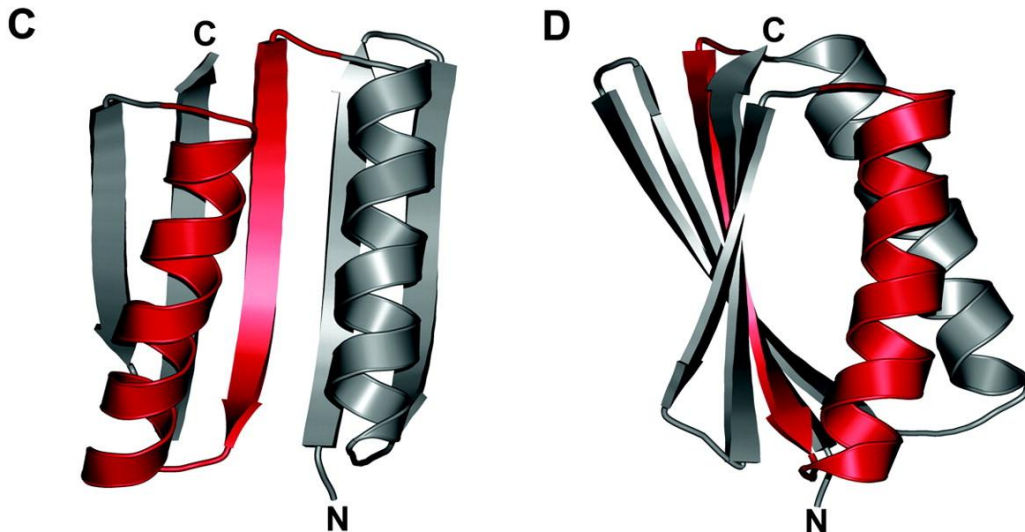
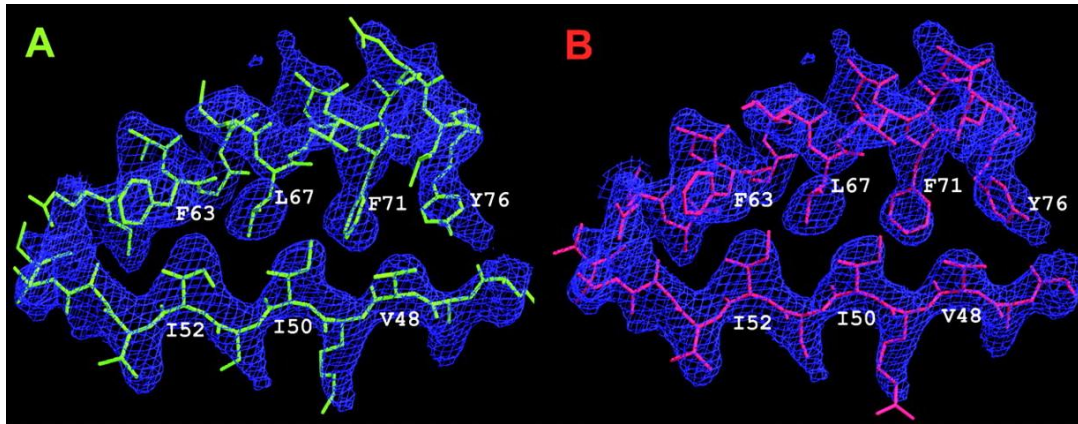
Dahiyat, B. I. and Mayo, S. L. (1997) *Science*, 278, 82-7
Dunbrack, R. L., Jr. and Karplus, M. (1993) *J Mol Biol*, 230, 543-74.
Kuhlman, B., et. al. (2003) *Science*, 302, 1364-1368.



A two-dimensional schematic of the target fold

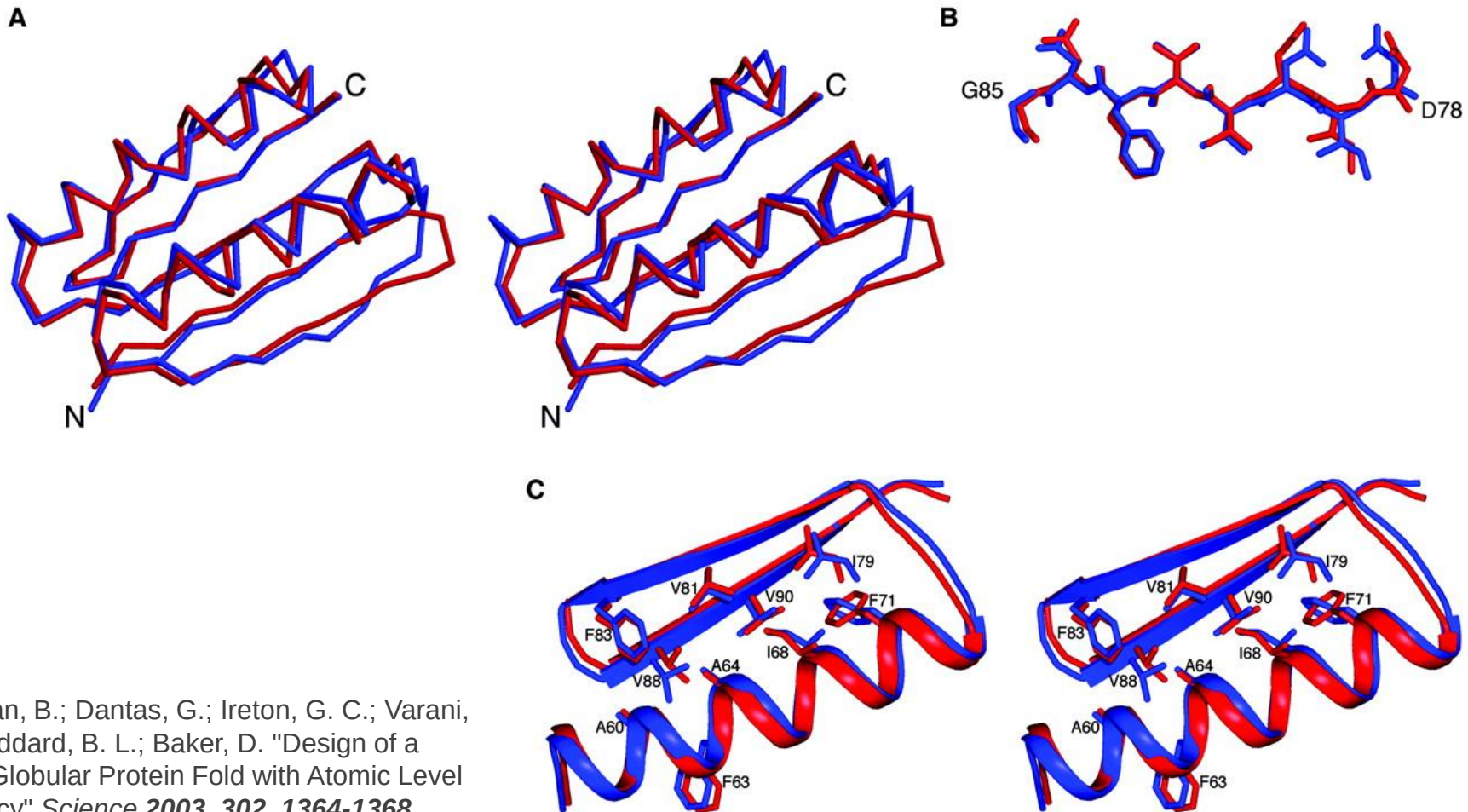


Schematic representation of Top7 residues 46-76



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



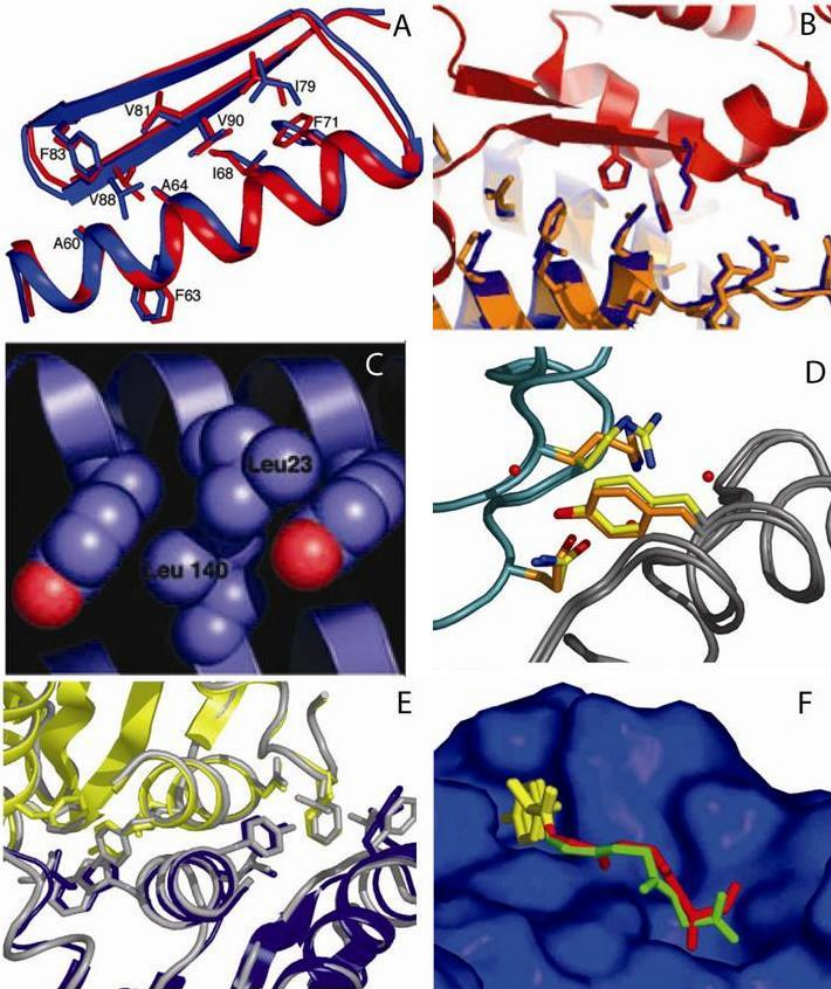
Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. "Design of a Novel Globular Protein Fold with Atomic Level Accuracy" *Science* **2003**, *302*, **1364-1368**.

Protein Design With Rosetta



Comparison of predicted ROSETTA models to solved x-ray structures:

- (A) Representation of the effectively superposable side chains in the core of the designed Top7 protein (backbone RMSD 1.17Å) [Kuhlman].
- (B) Detailed view of Colicin D–Immunity protein D. shows overlay of side-chains of the catalytic residues on the solved crystal structure (0.28Å RMSD) [Schueler-Furman].
- (C) High resolution structure of a ROSETTA thermally stabilized hydrolase enzyme [Korkegian].
- (D) Overlay of the E7_C/Im7_C model (orange sidechains) with the experimentally determined structure (yellow sidechains) [Kortemme].
- (E) 2.4Å crystal structure of the redesigned E-Drel homing endonuclease interface (yellow and blue) superimposed with the computational model (grey) [Kortemme].
- (F) Superposition of p56lck bound glycine containing peptide with solved crystal structure [Sood].



Outline

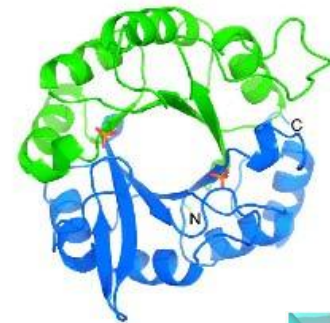
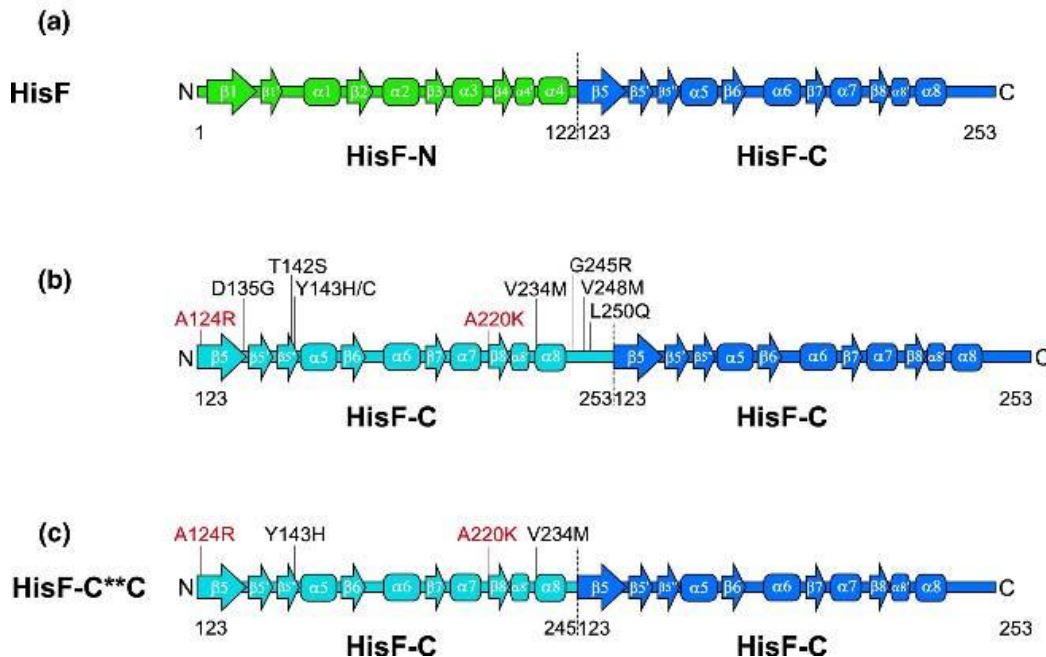


- *De novo* protein folding and the **Critical Assessment of Techniques for Protein Structure Prediction (CASP)**
 - Toward high-resolution *de novo* structure prediction for small proteins
 - Building atomic-detail protein models from sparse and low resolution experimental data
-
- TOP7 – Design of a novel globular protein fold with atomic level accuracy
 - FLR – A computationally designed TIM-barrel protein mimics missing link in evolution of symmetric superfolds

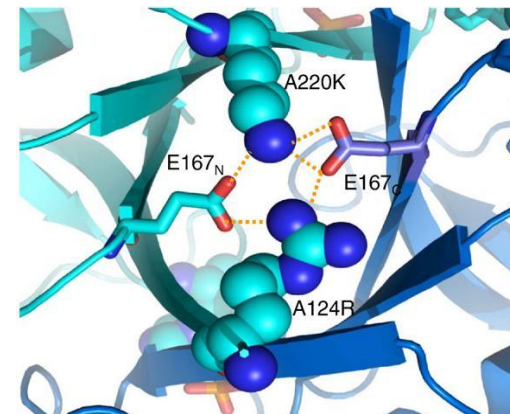
The Structure of Imidazole Glycerol Phosphate Synthase (HisF)



- The TIM-Barrel protein with highest degree of symmetry in primary, secondary, and tertiary structure known
- Symmetric variants created by concatenating N-terminal and C-terminal halves were stabilized and solubilized by directed evolution



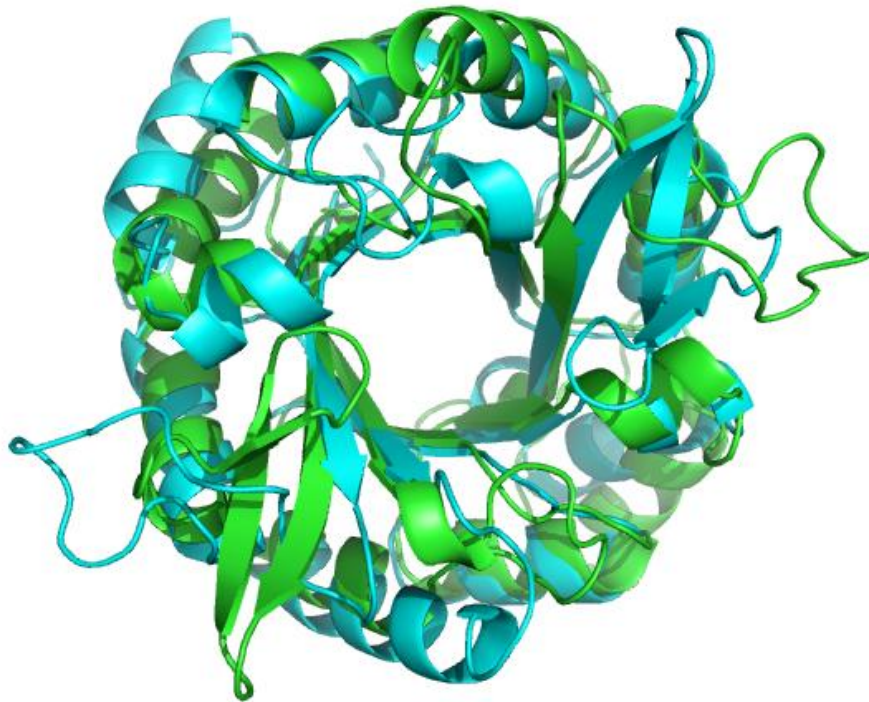
Seitz, T.; Bocola, M.; Claren, J.; Sterner, R. "Stabilisation of a (betaalpha)(8)-Barrel Protein Designed from Identical Half Barrels" *J Mol Biol* **2007**.



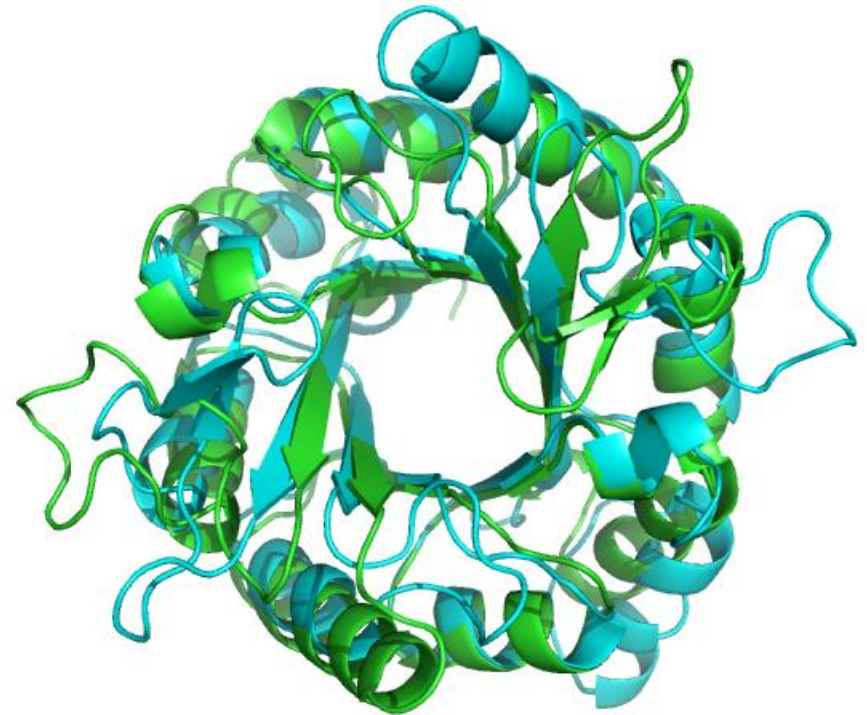
Structural Superimposition of HisF Half Barrels



1thf: 3.40Å over 202 AA

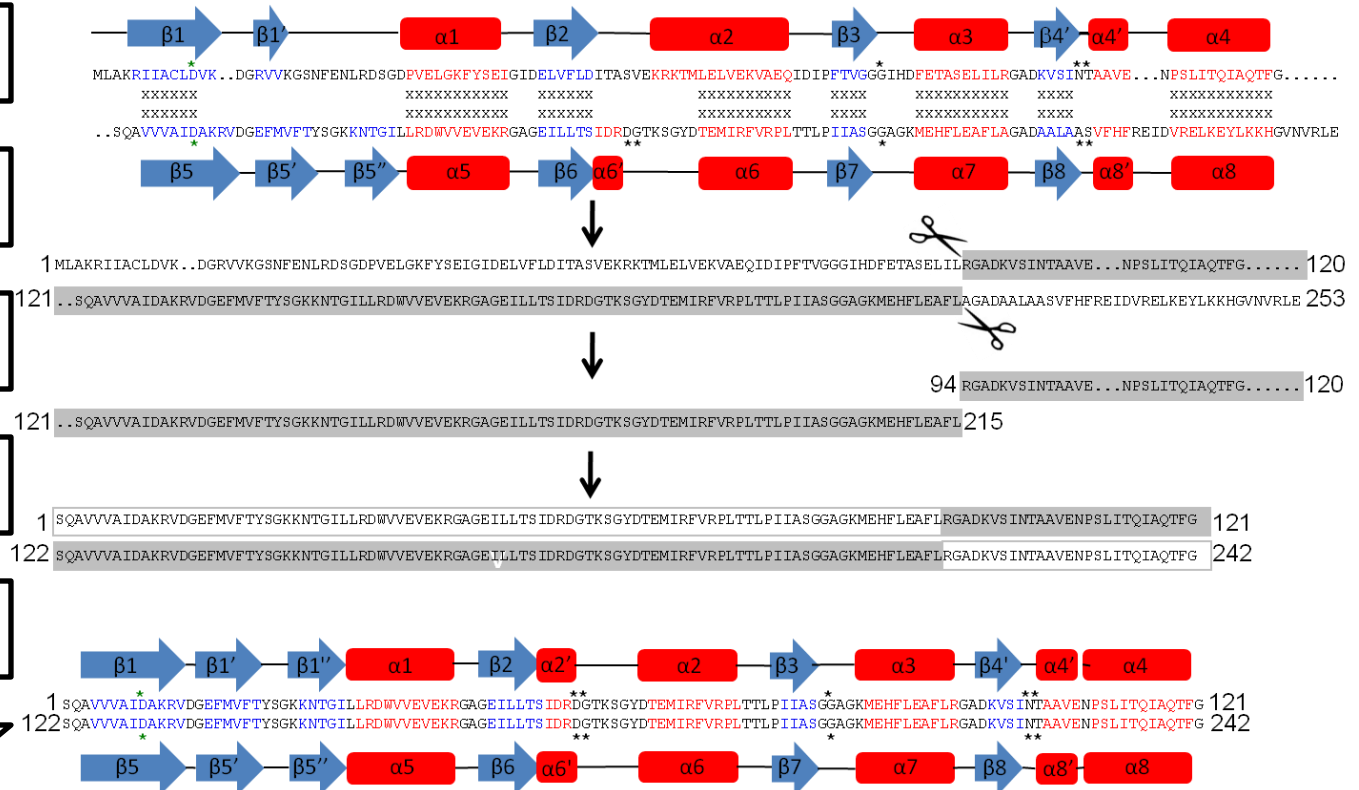
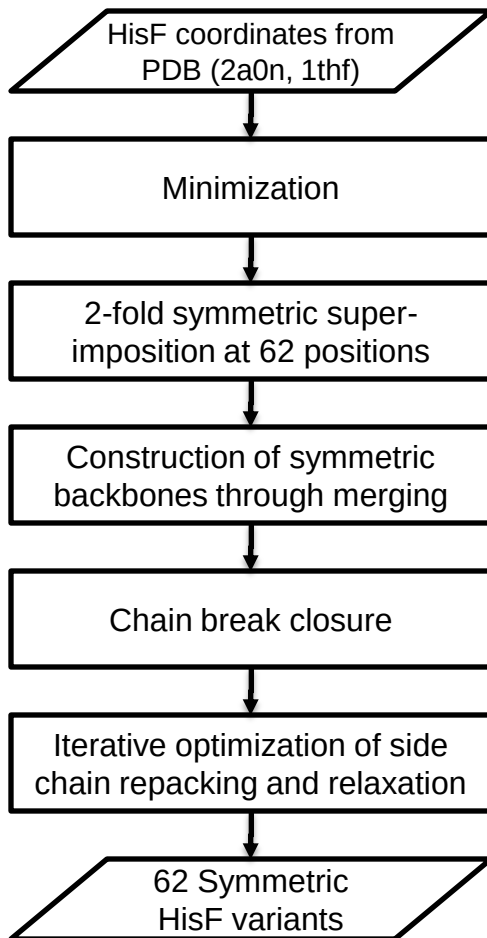


2a0n: 3.02Å over 178 AA

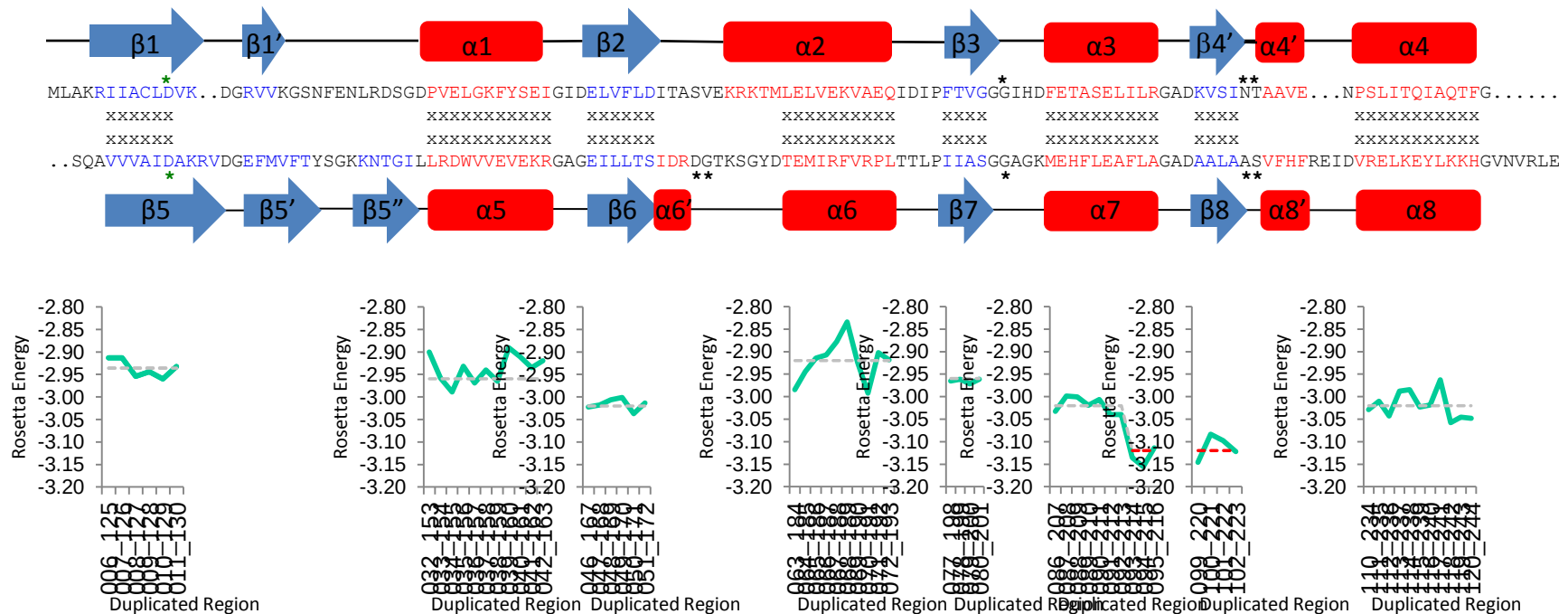


- 62 positions within α -helices and β -strands superimpose spatially with high accuracy $\rightarrow$ 62 symmetric variants

Computational Protocol for Construction of Symmetric Variants

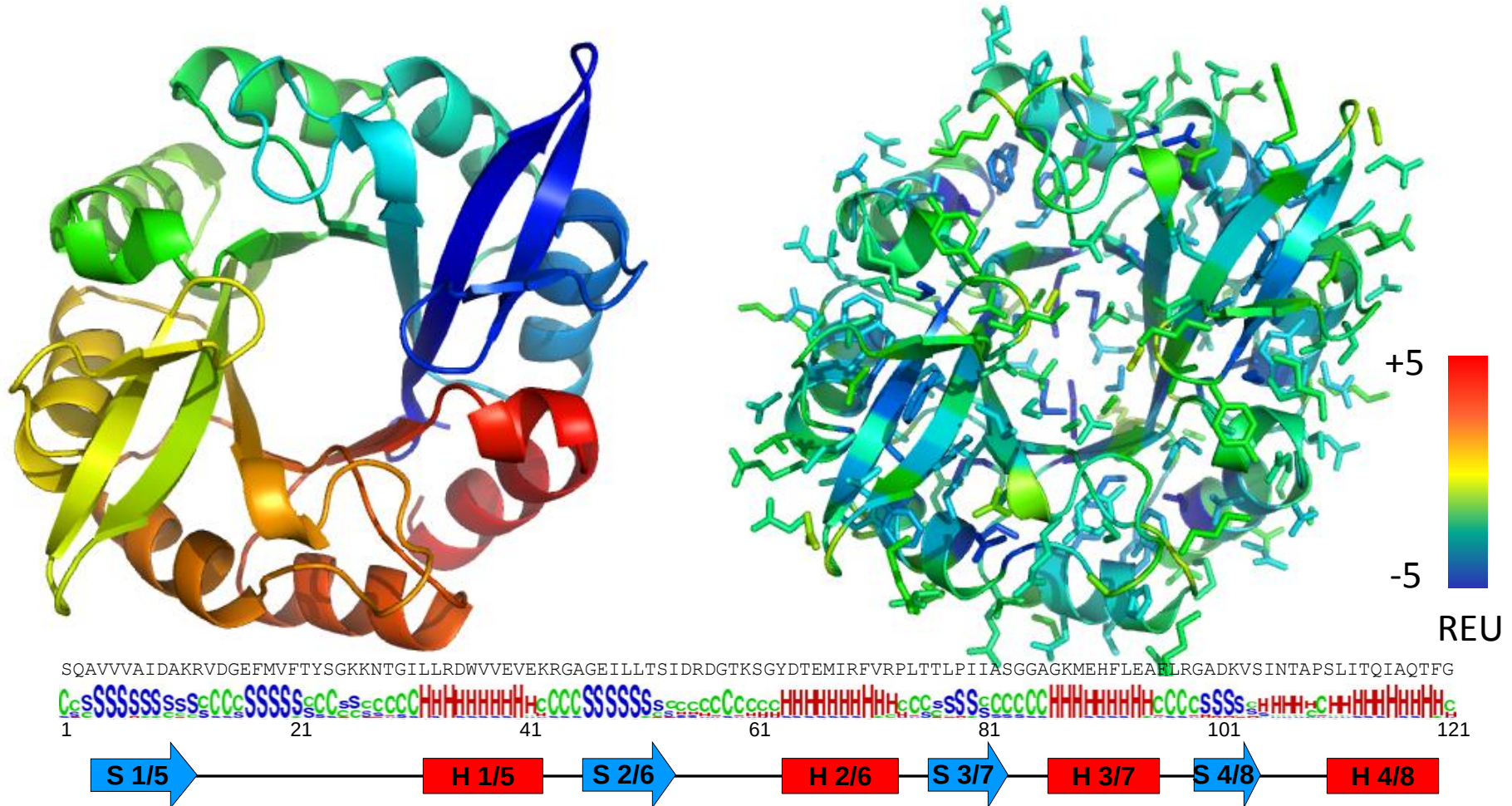


62 Positions Superimpose between HisF Half-Barrels

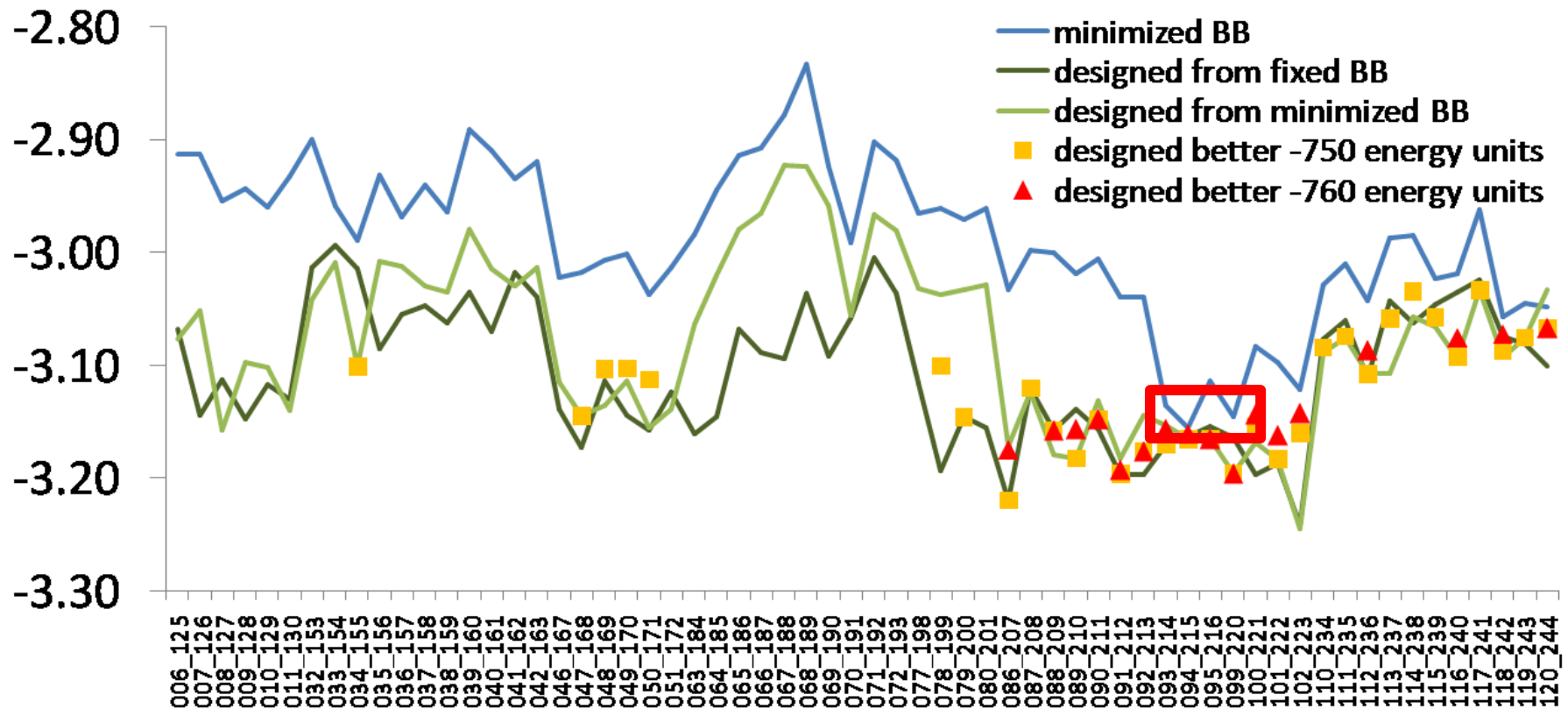


- Superimposed positions map to secondary structure element pairs β1/β5, α1/α5, β2/β6, α2/α6, β3/β7, α3/α7, β4/β8, α4/α8
- Rosetta Energy differs significantly depending on cut point
- Three most stable Variants maintain salt bridge cluster on αβ face

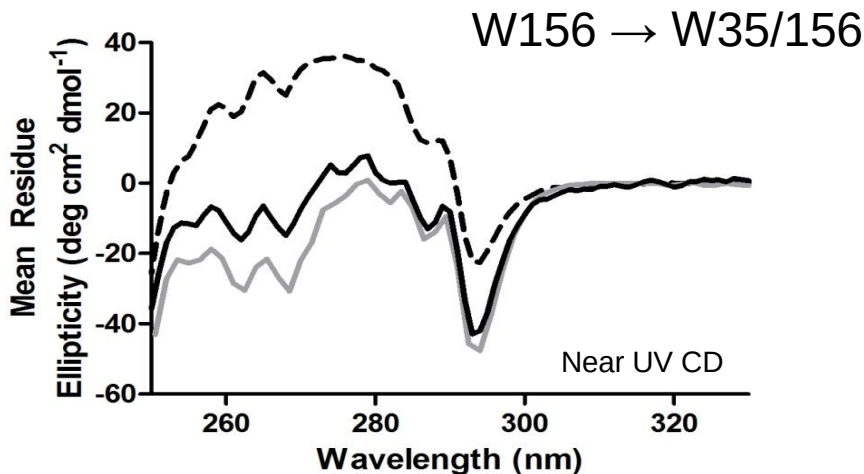
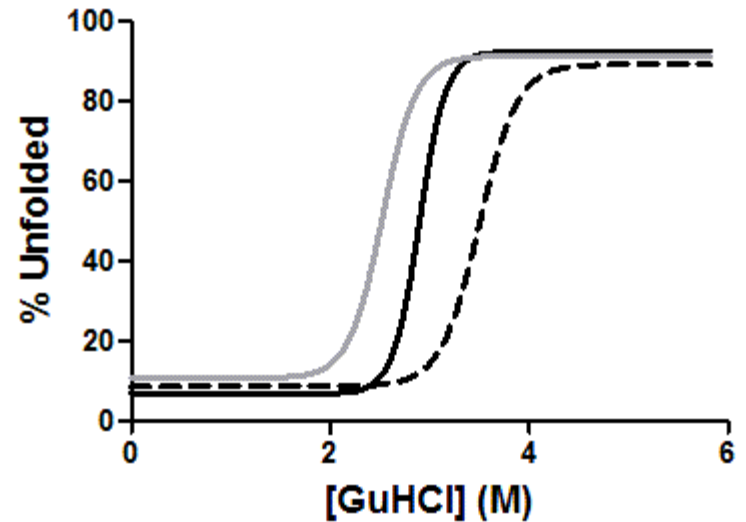
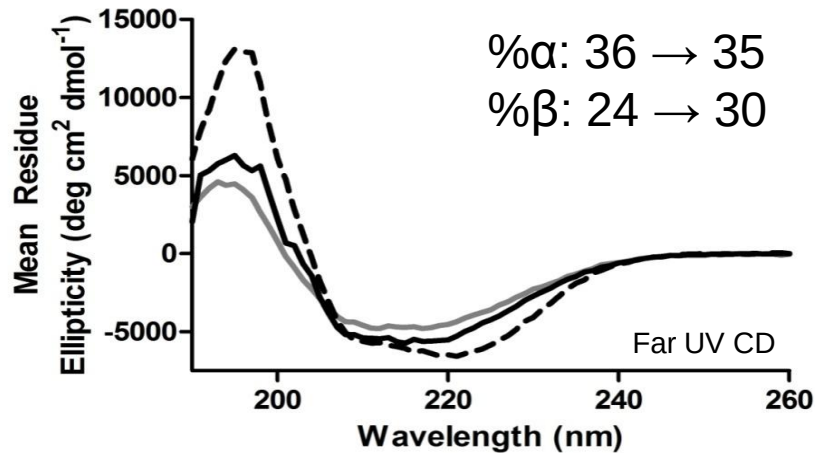
5_1thf_094_215_9.pdb:
 -763.61 REU and -3.16 REU per AA



RosettaDesign with Various Bonuses for Symmetry and Native Sequence



Characterization of monomeric “FLR” and dimeric Half-“FLR”



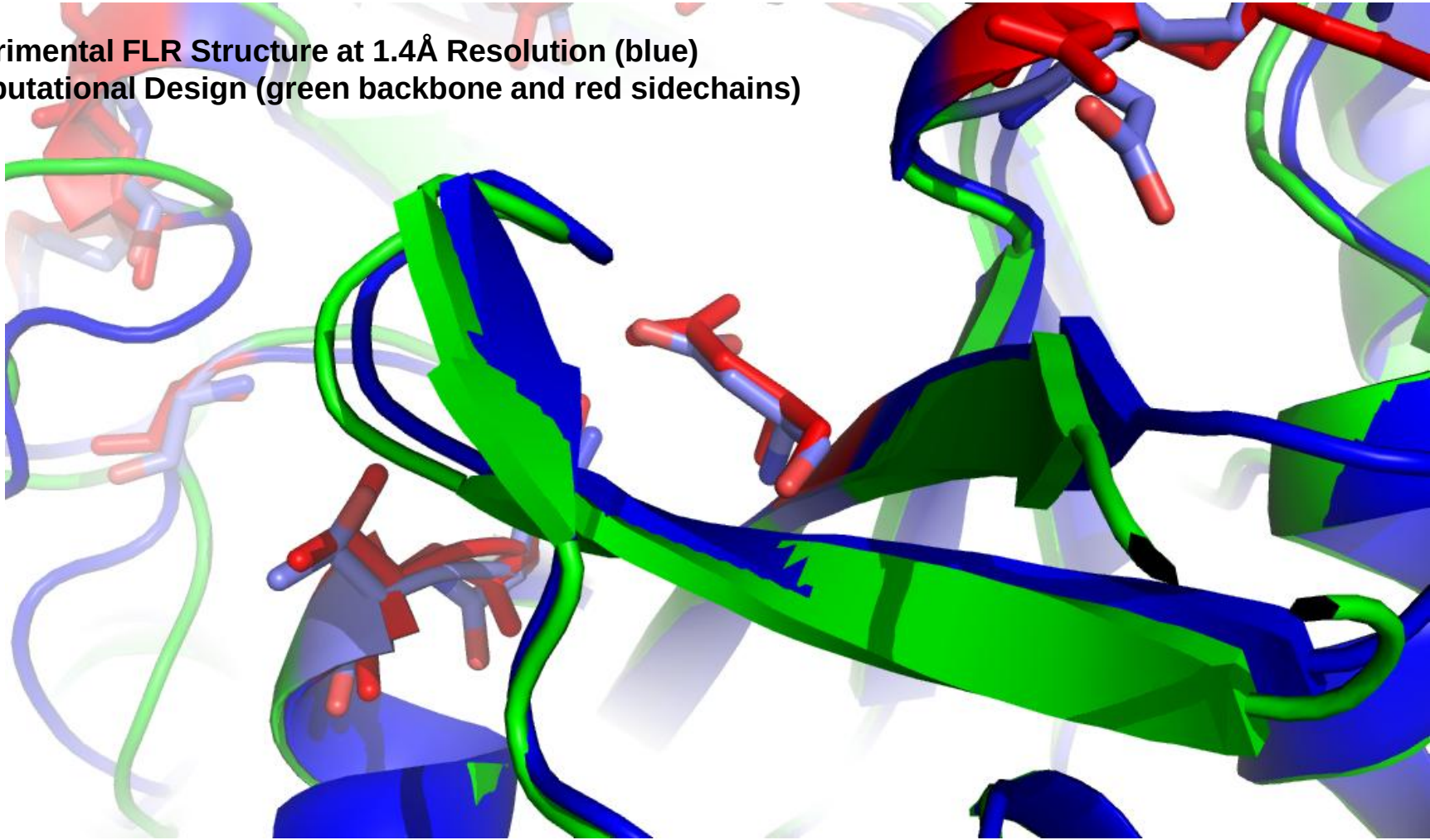
--- hisF
— FLR
— Half-FLR

Half FLR $K_d < 3.5 \times 10^{-8}$ mol

Catalytic Aspartate Residues and Phosphate Binding Sites



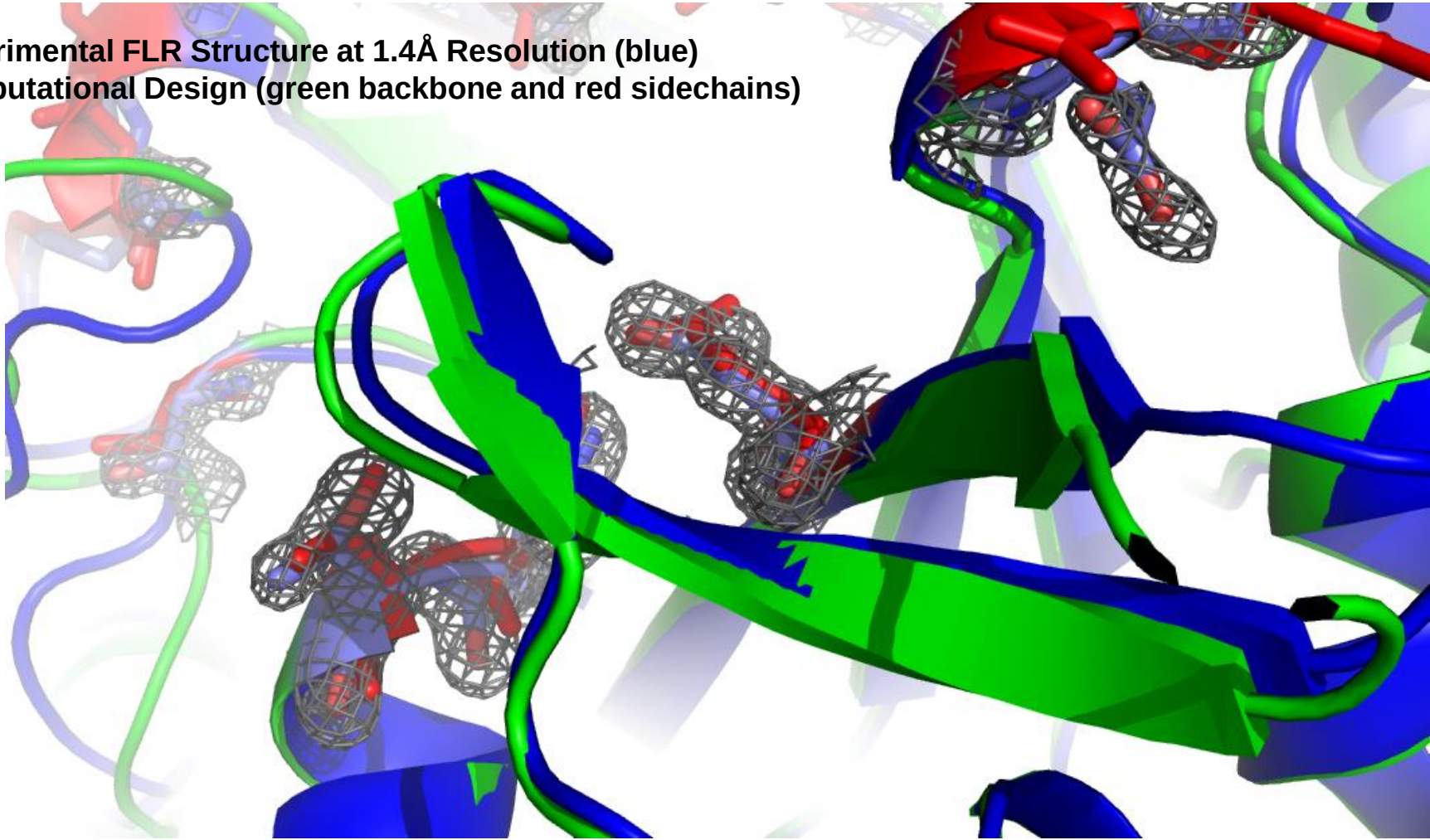
Experimental FLR Structure at 1.4Å Resolution (blue)
Computational Design (green backbone and red sidechains)



Catalytic Aspartate Residues and Phosphate Binding Sites



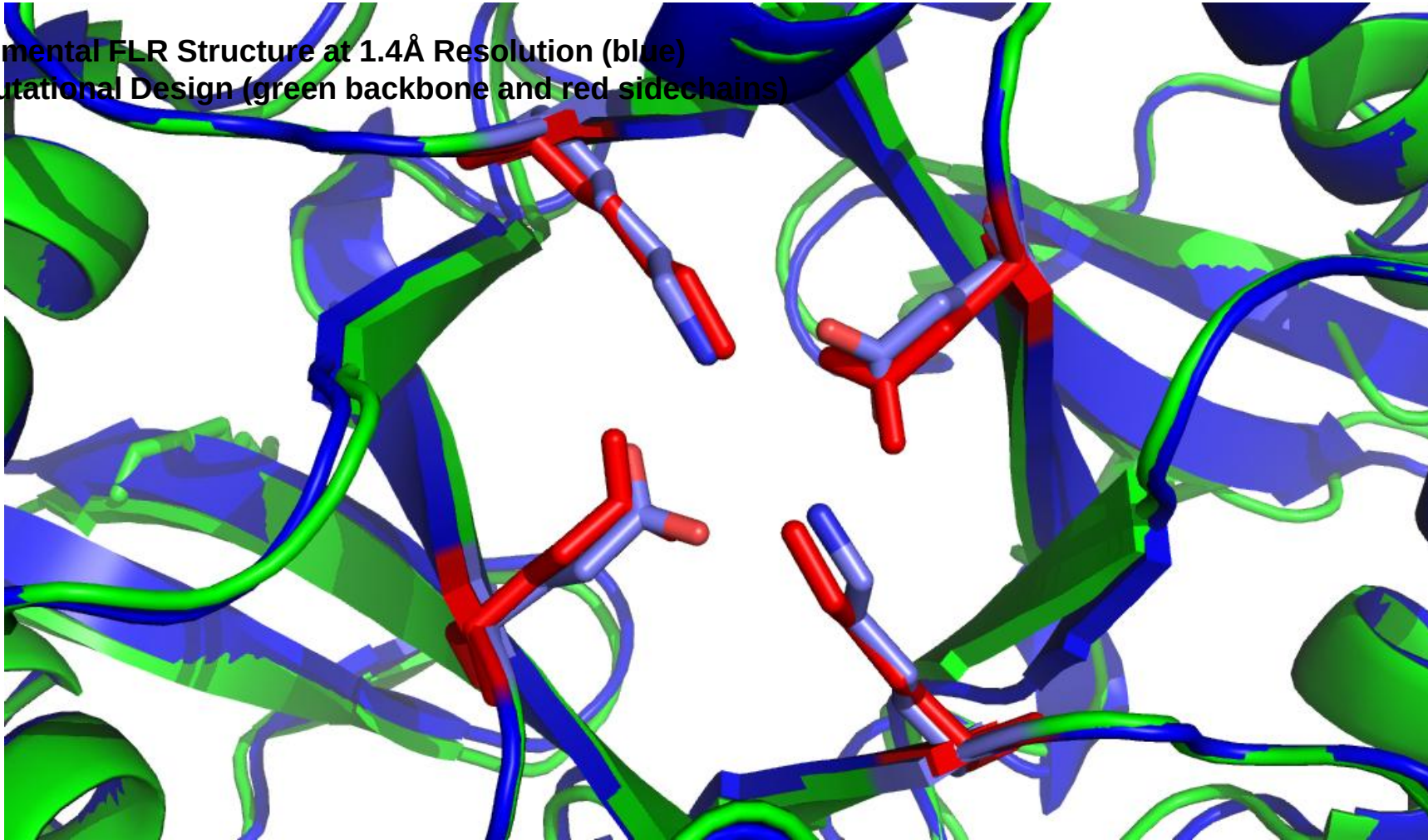
Experimental FLR Structure at 1.4Å Resolution (blue)
Computational Design (green backbone and red sidechains)



The Saltbridge Cluster at the $\alpha\beta$ Face



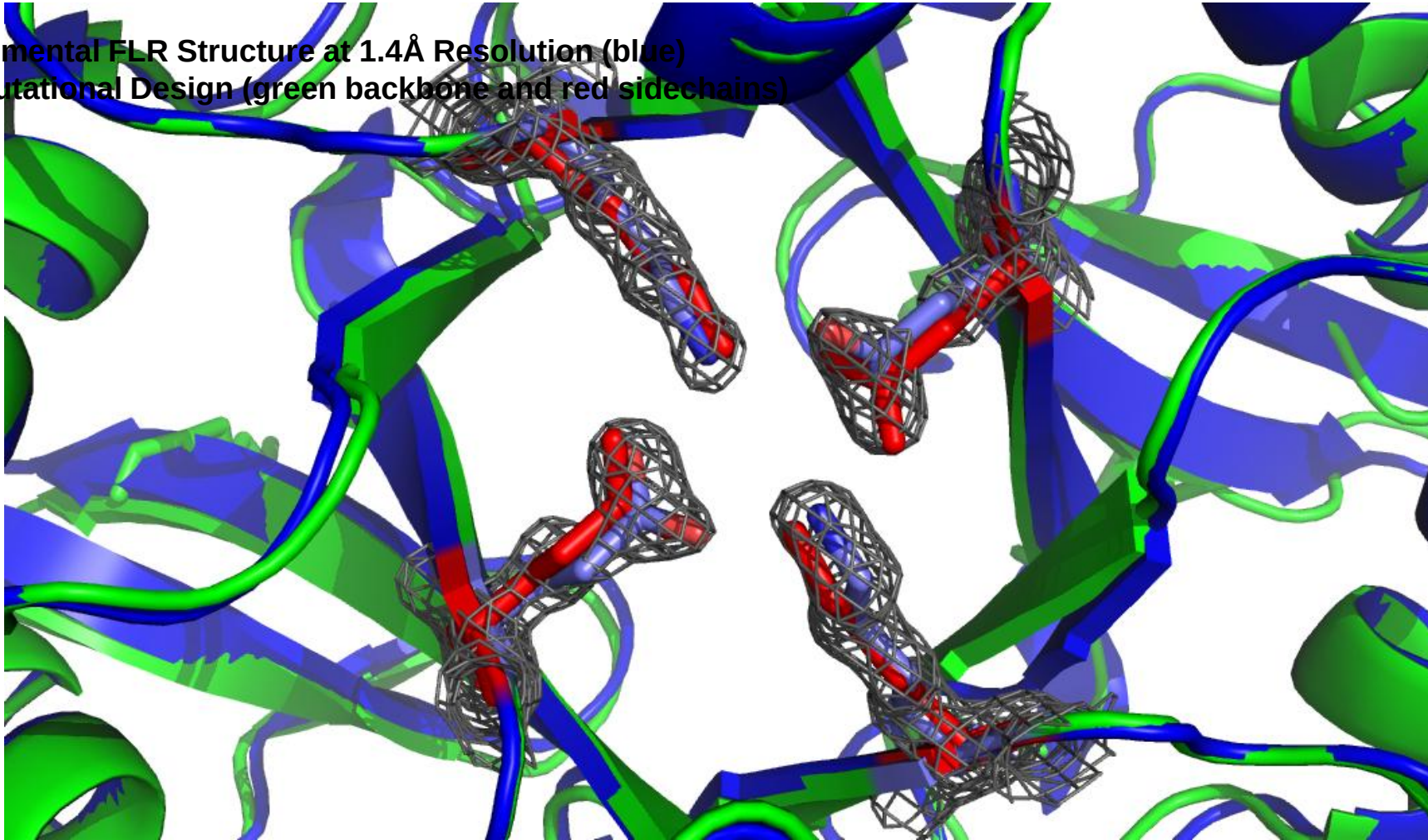
Experimental FLR Structure at 1.4Å Resolution (blue)
Computational Design (green backbone and red sidechains)



The Saltbridge Cluster at the $\alpha\beta$ Face



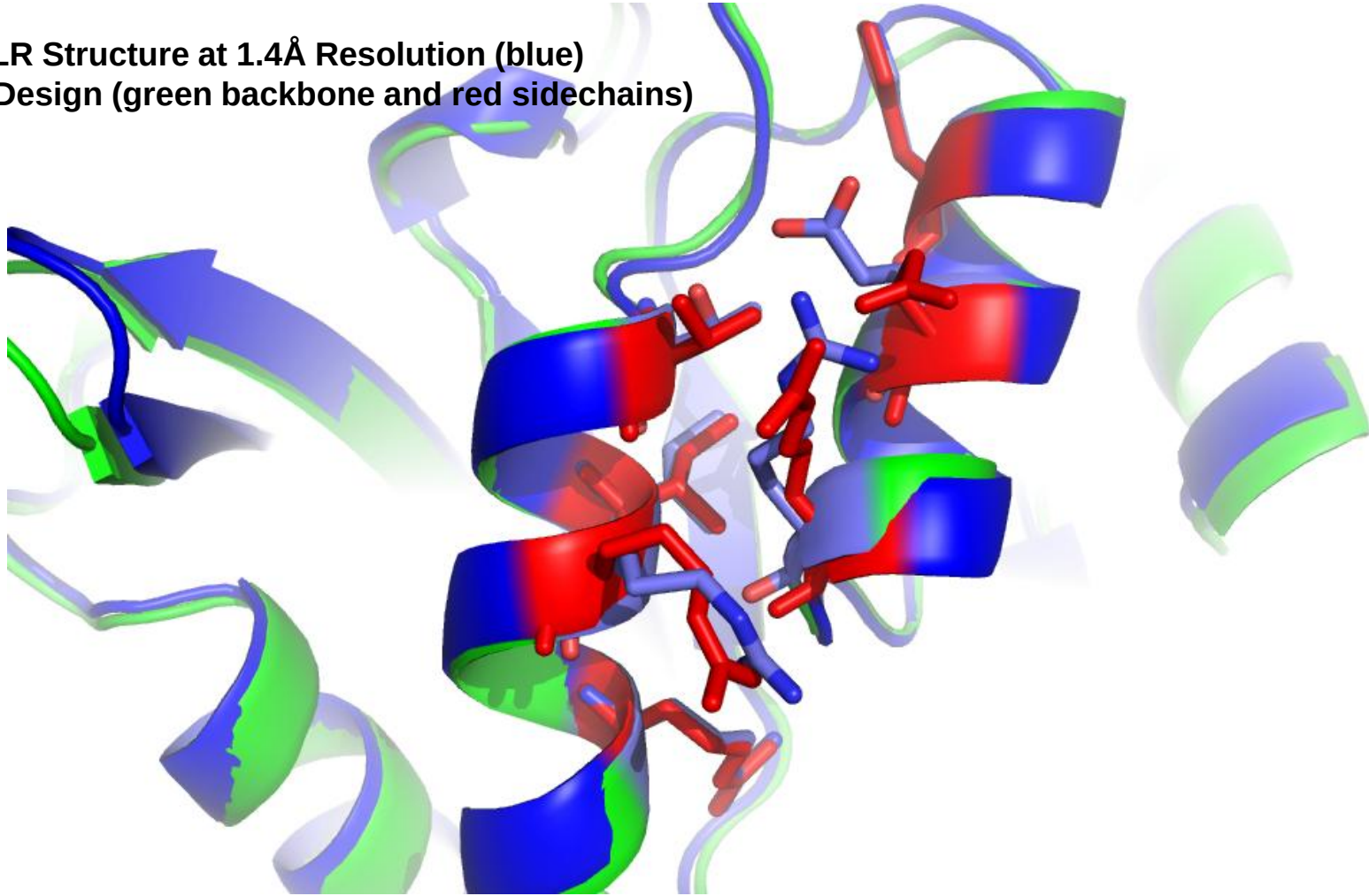
Experimental FLR Structure at 1.4Å Resolution (blue)
Computational Design (green backbone and red sidechains)



Helix 2 – Helix 3 Interface Contains the Cut Site for Duplication



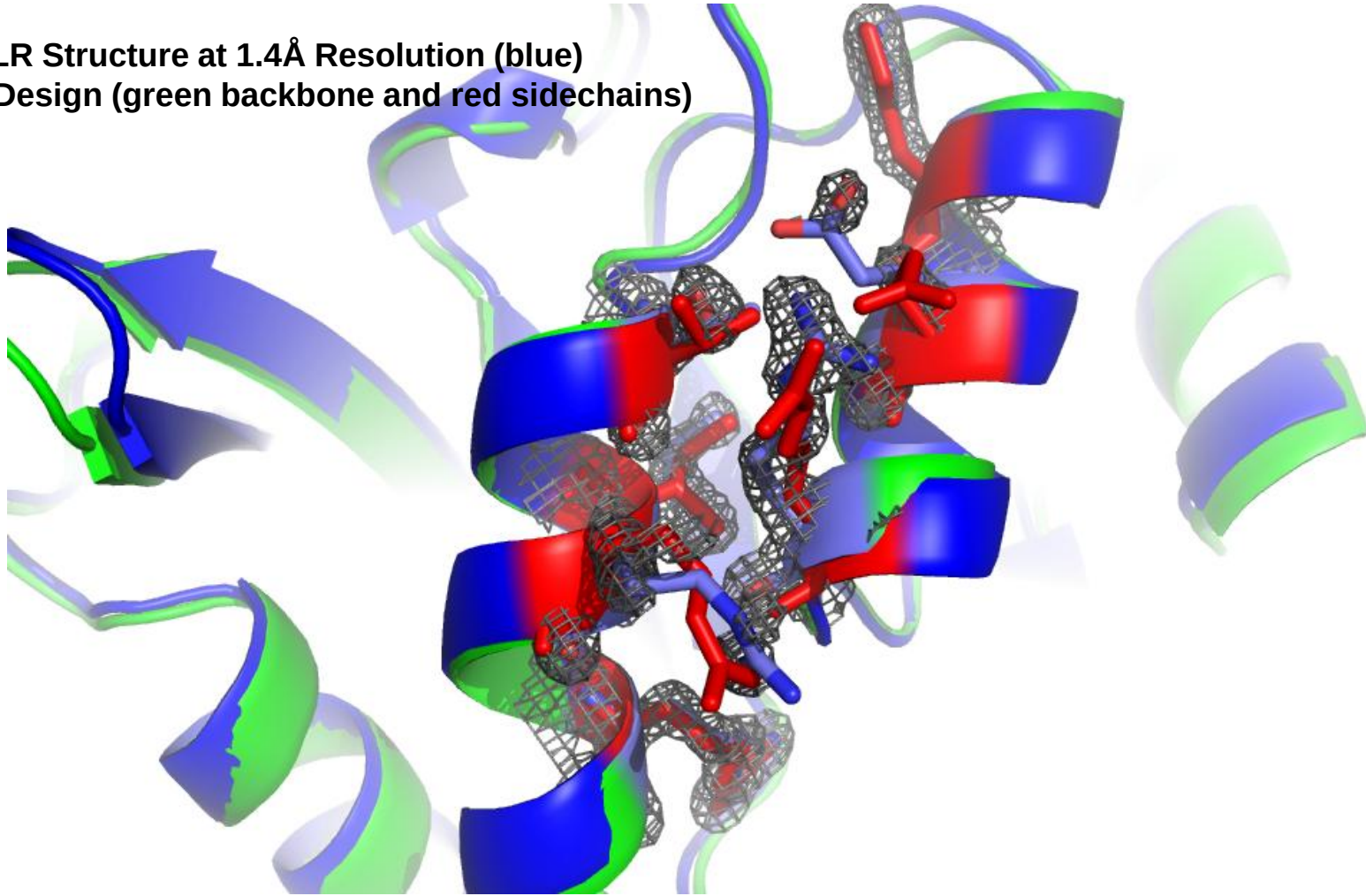
Experimental FLR Structure at 1.4Å Resolution (blue)
Computational Design (green backbone and red sidechains)



Helix 2 – Helix 3 Interface Contains the Cut Site for Duplication



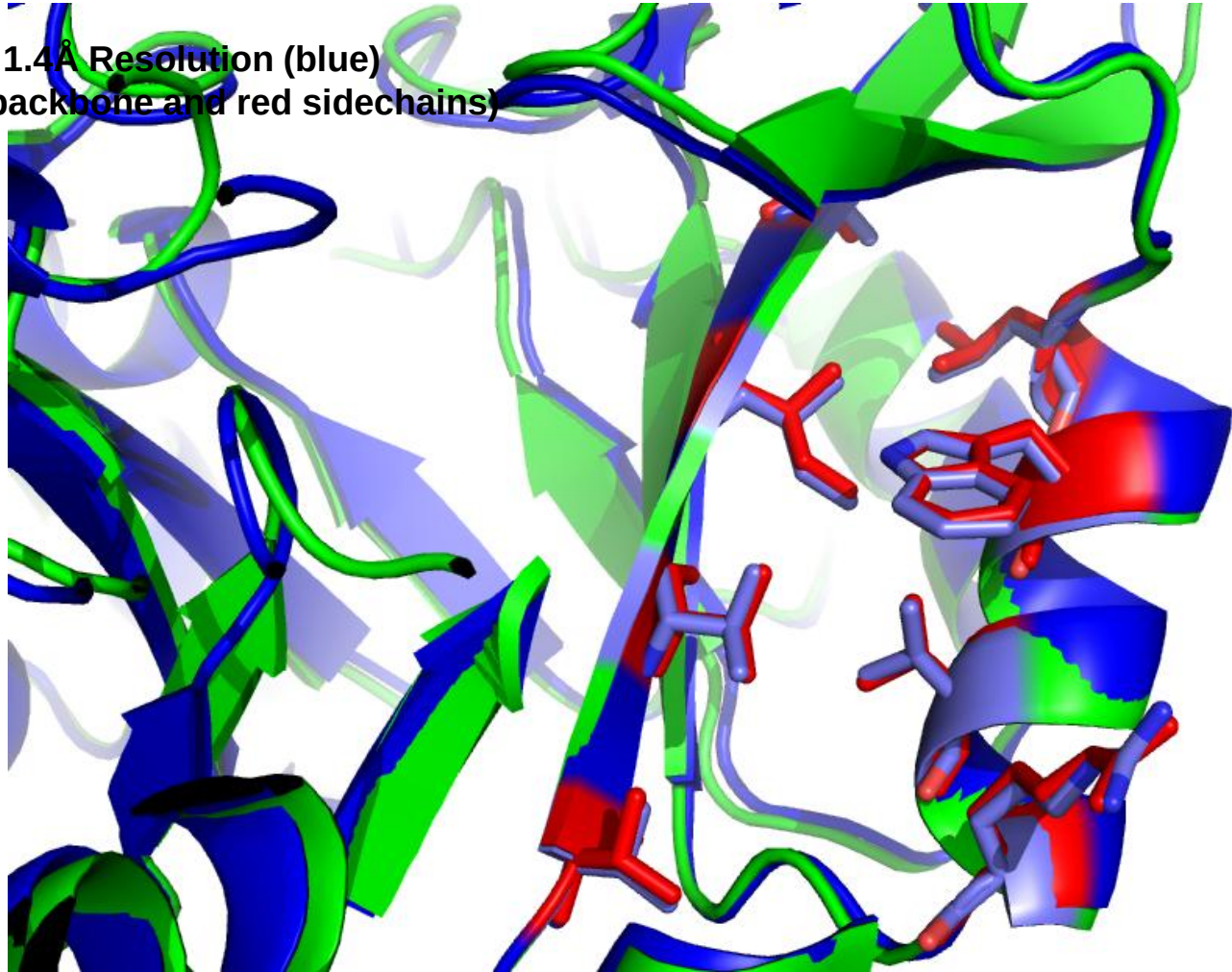
Experimental FLR Structure at 1.4Å Resolution (blue)
Computational Design (green backbone and red sidechains)



Helix 1 – Strand 1 Interface



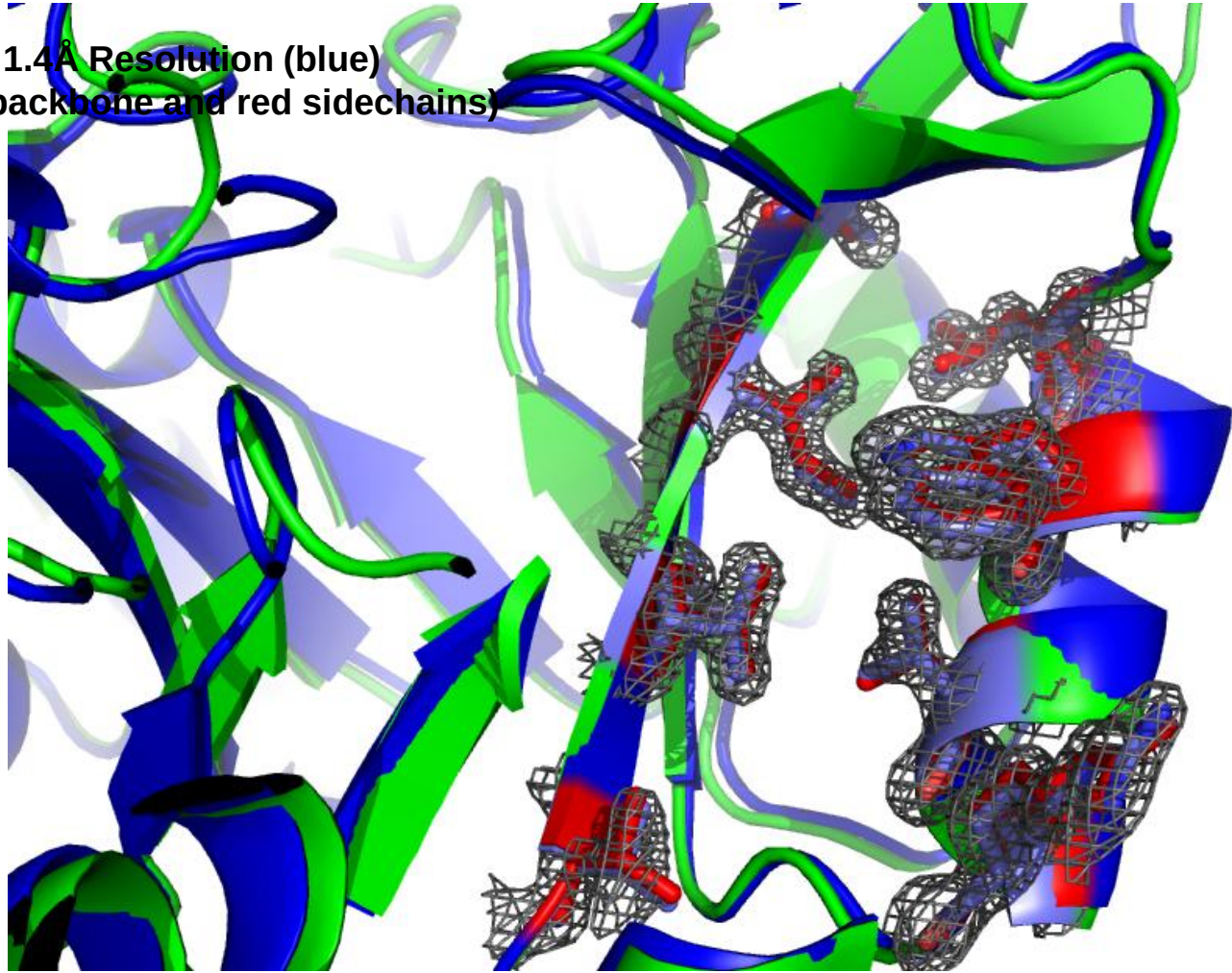
Experimental FLR Structure at 1.4Å Resolution (blue)
Computational Design (green backbone and red sidechains)



Helix 1 – Strand 1 Interface



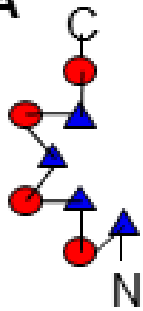
Experimental FLR Structure at 1.4Å Resolution (blue)
Computational Design (green backbone and red sidechains)



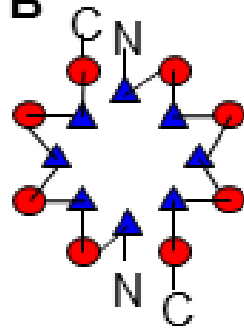
FLR and half-FLR Mimic Missing Link in Evolution



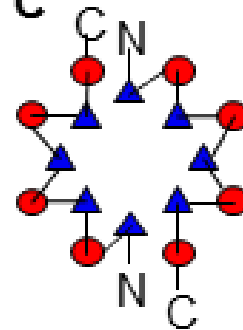
Structure: A



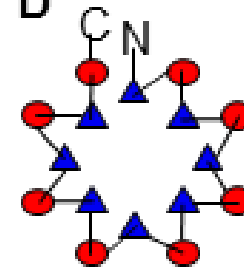
B



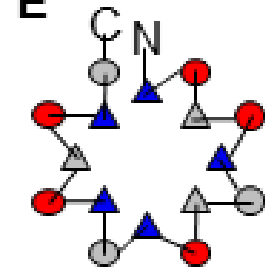
C



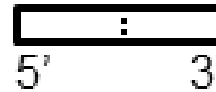
D



E



Gene:



Stable small
protein
domain

Symmetric
homodimer
evolves
through
interface
mutations

Gene
Duplication

Gene Fusion
concatenates
spatially
proximal N-
and C-termini

Sequence
symmetry
vanishes
through
acquisition of
new functions

Conclusion



- *De novo* protein structure prediction works for proteins up to ~150 amino acids.
 - Examples for refinement to experimental resolution of *de novo* models and detection of the native energy minimum for proteins of up to ~80 amino acids.
 - High resolution structure determination is feasible from *very* sparse / low resolution experimental data.
-
- *De novo* design of protein topologies is feasible for small proteins of ~100 amino acids.
 - A computationally designed sequence-symmetric variant of HisF mimics missing link in evolution of symmetric superfolds.

The ACCRE Cluster – 1500 Processors at Your Service



Acknowledgements



- Amanda Duran
- Brian Weiner
- Brittany Allison
- Carie Fortenberry
- Daniel Putnam
- David Nannemann
- Gordon Lemmon
- Greg Sliwoski
- Jeffrey Mendenhall
- Jessica Finn
- Jordan Willis
- Kelli Kazmier



- Liz Dong
- Mariusz Butkiewicz
- Nathan Alexander
- Pedro Teixeira
- Sam DeLuca
- Stephanie DeLuca
- Sandeep Kothiwale
- Soumya Ganguly
- Sten Heinze
- Steven Combs
- Will Lowe
- Yan Xia

- VU Collaborators:
Walter Chazin, Jim Crowe,
Heidi Hamm, Randy Blakely,
Jeff Conn, Steve Fesik, Craig
Lindsley, Seva Gurevich,
Hassane Mchaourab, Dave
Weaver, Ambra Pozzi, Chuck
Sanders

- NIH NIGMS 1R01GM099842
- NIH NIGMS 1R01GM080403
- NIH NIMH 1R01MH090192
- NSF MCB CAREER 0742762
- DTRA HDTRA1-10-0067
- NIH NIAID HHSN 272200900047C

