Structure Prediction

Biochemistry 530, 2016

David Baker

Principles underlying protein structure prediction

- Physical chemistry
- Evolution

Structure Prediction

I. Secondary structure prediction: prediction of location of helices, sheets, and loops

II. Fold recognition (threading):determine whether a protein sequence is likely to adopt a known fold/structure.

III. Comparative Modeling:prediction of structure based on structure of a closely related homologue

III. Ab initio structure prediction:predict protein tertiary structure de novo.

IV. CASP protein structure prediction competition/experiment.

I. Secondary structure prediction

The basis for secondary structure prediction is that the different amino acid residues occur with different frequencies in helices, sheets, and turns:

Table 6.5 Conformational Preferences of the Amino Acids

Amino acid residue	Preference[a]			α-Helix Preference[b]			Turn Preference		
	α-helix (P_α)	β-strand (P_β)	Reverse turn (P_t)	N-term	Middle	C-term	Type I	Type II	Other
Glu	1.59	0.52	1.01	2.12	1.18	1.21	1.12	0.84	1.06
Ala	1.41	0.72	0.82	1.33	1.60	1.46	0.74	0.94	0.58
Leu	1.34	1.22	0.57	1.05	1.50	1.46	0.61	0.53	0.75
Met	1.30	1.14	0.52	0.75	1.44	1.92	0.66	0.73	0.96
Gln	1.27	0.98	0.84	1.39	1.22	1.24	0.79	1.45	0.96
Lys	1.23	0.69	1.07	0.98	1.05	1.68	0.70	0.73	1.02
Arg	1.21	0.84	0.90	1.26	1.25	1.23	0.88	1.22	1.04
His	1.05	0.80	0.81	0.68	0.97	1.57	0.78	0.64	0.84
									1.00
Val	0.90	1.87	0.41	1.00	1.09	1.08	0.39	0.61	0.48
Ile	1.09	1.67	0.47	0.96	1.31	0.99	0.39	0.43	0.93
Tyr	0.74	1.45	0.76	0.63	0.61	1.00	0.71	0.91	0.97
Cys	0.66	1.40	0.54	0.78	0.66	0.56	1.38	0.99	0.78
Trp	1.02	1.35	0.65	1.20	1.34	0.78	1.35	0.15	0.52
Phe	1.16	1.33	0.59	0.94	1.45	1.20	0.77	0.76	0.53
Thr	0.76	1.17	0.90	0.75	0.87	0.80	1.25	0.67	0.93
Gly	0.43	0.58	1.77	0.60	0.47	0.31	1.14	2.61	1.38
Asn	0.76	0.48	1.34	0.80	0.80	0.75	1.79	0.99	1.37
Pro	0.34	0.31	1.52	0.90	0.19	0.06	0.95	1.80	1.51
Ser	0.57	0.96	1.22	0.67	0.44	0.73	1.47	0.76	1.49
Asp	0.99	0.39	1.24	1.35	1.03	0.67	1.98	0.71	1.28

The Psipred methodology

The best current method, psipred, can be accessed at http://bioinf.cs.ucl.ac.uk/psipred.

Psipred and other state of the art methods use a neural network to extract information from multiple sequence alignments.
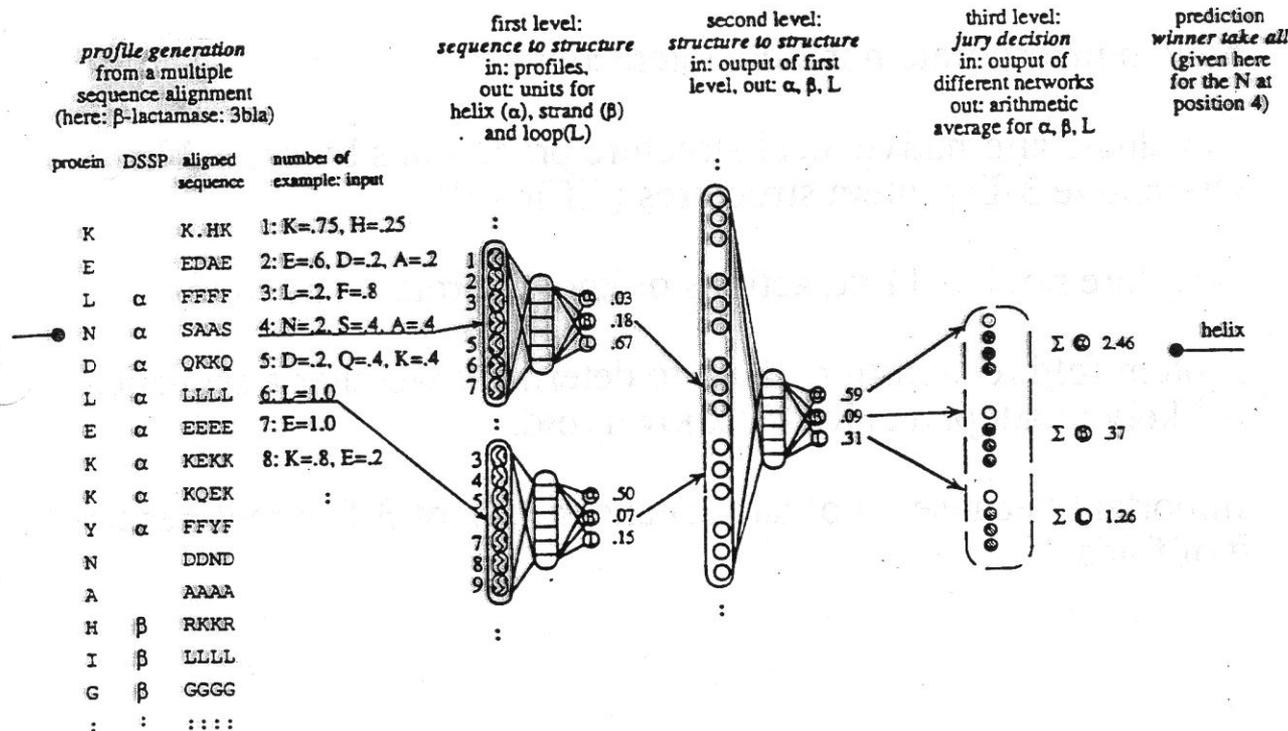

Figure 2. Our network system for secondary structure prediction. Our network system for predicting secondary structure consists of 3 layers: 2 network layers and 1 layer averaging over independently trained networks. ⊖, Basic cell containing 20 + 1 units to code residues at position 1 to w of the input window; here, $w = 7$. ⊖, Hidden units. Circled α, β and L, output units for helix, strand and loop. Stippled circles, output from architectures not shown here. →●, Example: residue N at position 4 predicted to be in helix ●— .

Limits in seconday structure prediction accuracy

- Upper limit to success of secondary structure and solvent accessibility predictions from local sequence information:
  - Non-local interactions play critical roles in stabilizing protein structures
  - Non-local interactions not taken into account in local structure prediction

- How to incorporate interactions?
  - Evaluate alternative local structure predictions by assembling alternative 3D protein structures (difficult!)
  - Explore non-local interactions of known protein structures.

- Protein fold recognition: how to determine whether a sequence is likely to adopt an already known fold.

- Important because # of aa sequences >> # of 3D structures >> # of folds

II Fold recognition

– Match sequence hydrophobicity patterns to solvent accessibility patterns calculated from known structures.
  - Calculate hydrophobicity patterns from aligned sequence sets (higher signal to noise since few conserved hydrophobic residues on surface)
  - Use dynamic programming algorithm to align sequence + solvent accessibility patterns.
– Key insight: reduce 3D structure to 1D solvent accessibility string.

Fold recognition (cont)

- Most successful current fold recognition servers use a combination of sequence and structural information to match sequences with folds:

Query sequence

Sequence profile from msa

Hydrophobicity pattern

Predicted secondary structure

Target structure

Sequence profile from msa

Solvent accessibility pattern

Known secondary structure

Comparative Modeling

- Given a sequence with homology to a protein of known structure, build accurate model

- Four steps:
  - 1) Generate accurate alignment
  - 2) Based on alignment, extract from structure of template either distance constraints or starting coordinate positions
  - 3) Build de novo regions not included in the alignment
  - 4) Refine completed model using evolutionary and/or physical information

Challenges in comparative modeling

- Creating accurate alignments. Particularly for proteins with <20% sequence identity to template. Edge beta strands are particularly difficult. Have to consider multiple alternative templates and alignments.

- Accurate modeling of loops and insertions (are in less deep minima than protein core)

- Modeling systematic shifts in backbone coordinates

Ab initio protein structure prediction

- The "holy grail": we've known for 40 years that structure is determined by amino acid sequence (Anfinsen), but can we predict protein structure from amino acid sequence alone?

First, have to decide how to represent polypeotide chain:
  - 1. All atom:every atom in the protein treated. (very complicated and time intensive)
  - 2. Lattice models (important features left out?)
  - 3. Off lattice, but simplified relative to all atom representation.

- Typically, side chains are represented by 1-2 pseudo atoms and the only degrees of freedom are the backbone torsion angles and ~1 rotation for the side chains.

- Greatly reduces number of interacting groups (# of residues << # of atoms), and numbers of degrees of freedom of chain.

Second, have to choose what energy function to optimize when searching through possible protein conformations

- Sources of information
    - 1. physical chemistry (cf lecture 1)
    - 2. chemical intuition
    - 3. high resolution protein structures
    - 4. Electronic structure calculations (QM).

Finally, once representation and potential functions are chosen, need to search space for low energy states.

- Simplest procedure—always go downhill (steepest descent)

    Doesn't work (energy surfaces have multiple minima)

Search Algorithms

- Molecular Dynamics: Popular because models actual protein dynamics. But slow because time step has to be very small

- Monte Carlo: Make random perturbation (typically to backbone or sidechain torsion angles), compute energy, and accept if the energy is decreased, and roll the dice if the energy increases. Allows overcoming of barriers

- Monte Carlo Minimization: Same as Monte Carlo, but minimize before computing energy

- Simulated Annealing: MD or MC starting with high temperature and then slowly cooling

- Replica Exchange: Carry out multiple parallel MD or MC trajectories at different temperatures, allowing occasional swaps between trajectories

- Genetic Algorithms: Start with population of conformations. Evolve by iterating between mutation, recombination, and selection

Implementation of insights from experimental folding studies in ROSETTA

1. Local interactions bias but do not uniquely determine conformations sampled by short segments of the chain.

2. Folding occurs when local structure segments oriented so as to bury   hydrophobic residues, pair beta strands, etc.

3. Stability determined by detailed sidechain -sidechain interactions in folded structure.

4. Folding rates are largely determined by contact order of native structure. Short folding   time   low contact order structures.

Rosetta high resolution refinement

- SAMPLING PROTOCOL--Monte Carlo minimization with combinatorial sidechain optimization in torsion space
  - 1) randomly chosen backbone deformation (phi/psi change, fragment insertion, etc.)
  - 2) sidechain repacking (Monte Carlo search through Dunbrack library)
  - 3) gradient-based minimization of energy with respect to torsion angles (DFPmin)
  - 4) acceptance according to standard Metropolis criterion
- POTENTIAL FUNCTION

    Lennard Jones, LK implicit solvation, orientation-dependent hydrogen bonding, PDB derived torsional potential
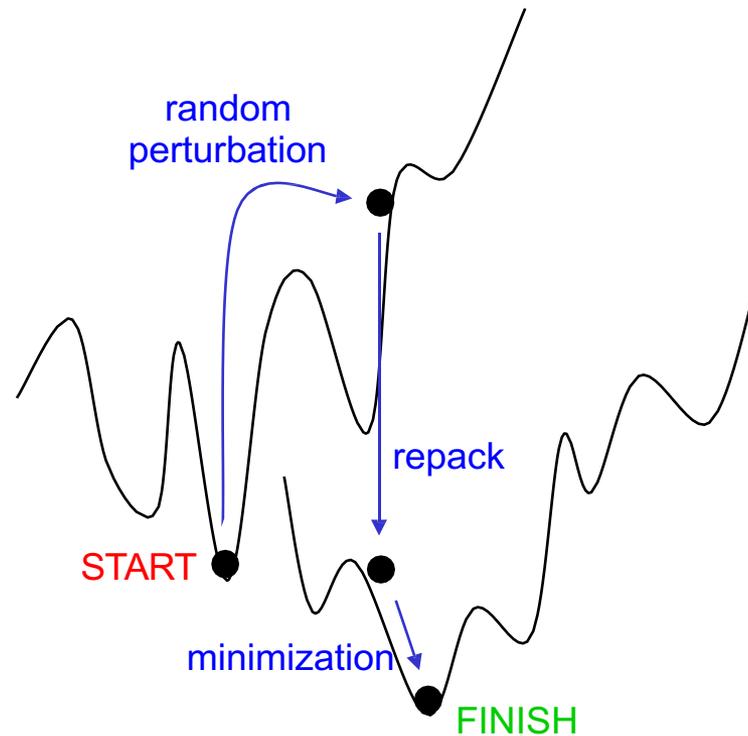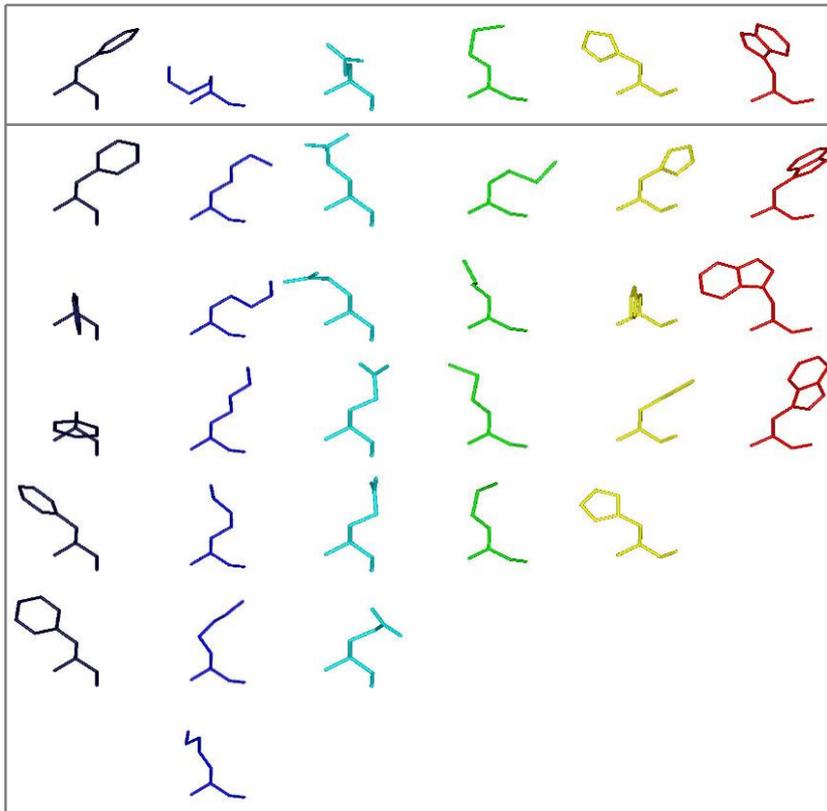
Lowest energy structures sampled on independent trajectories
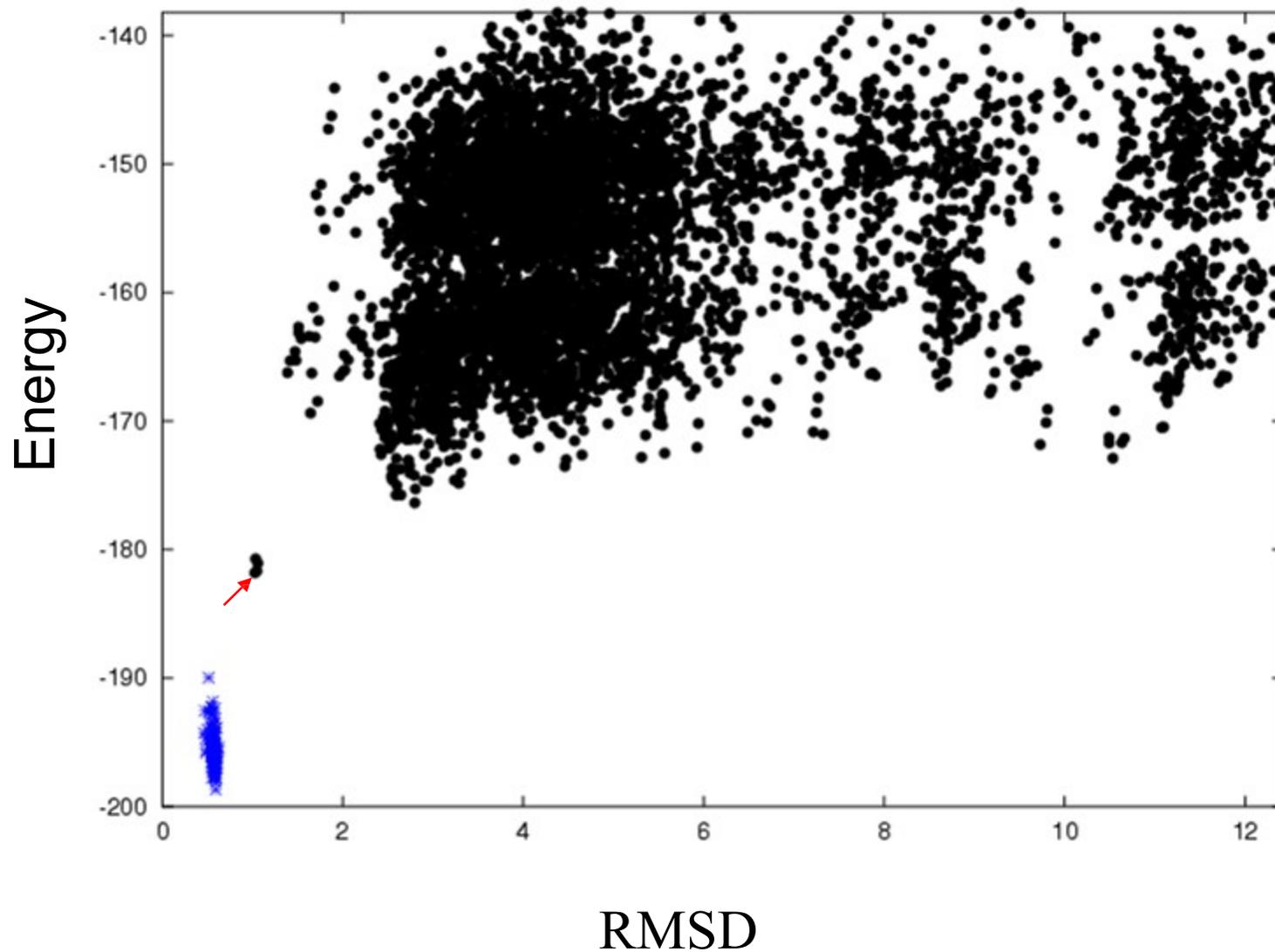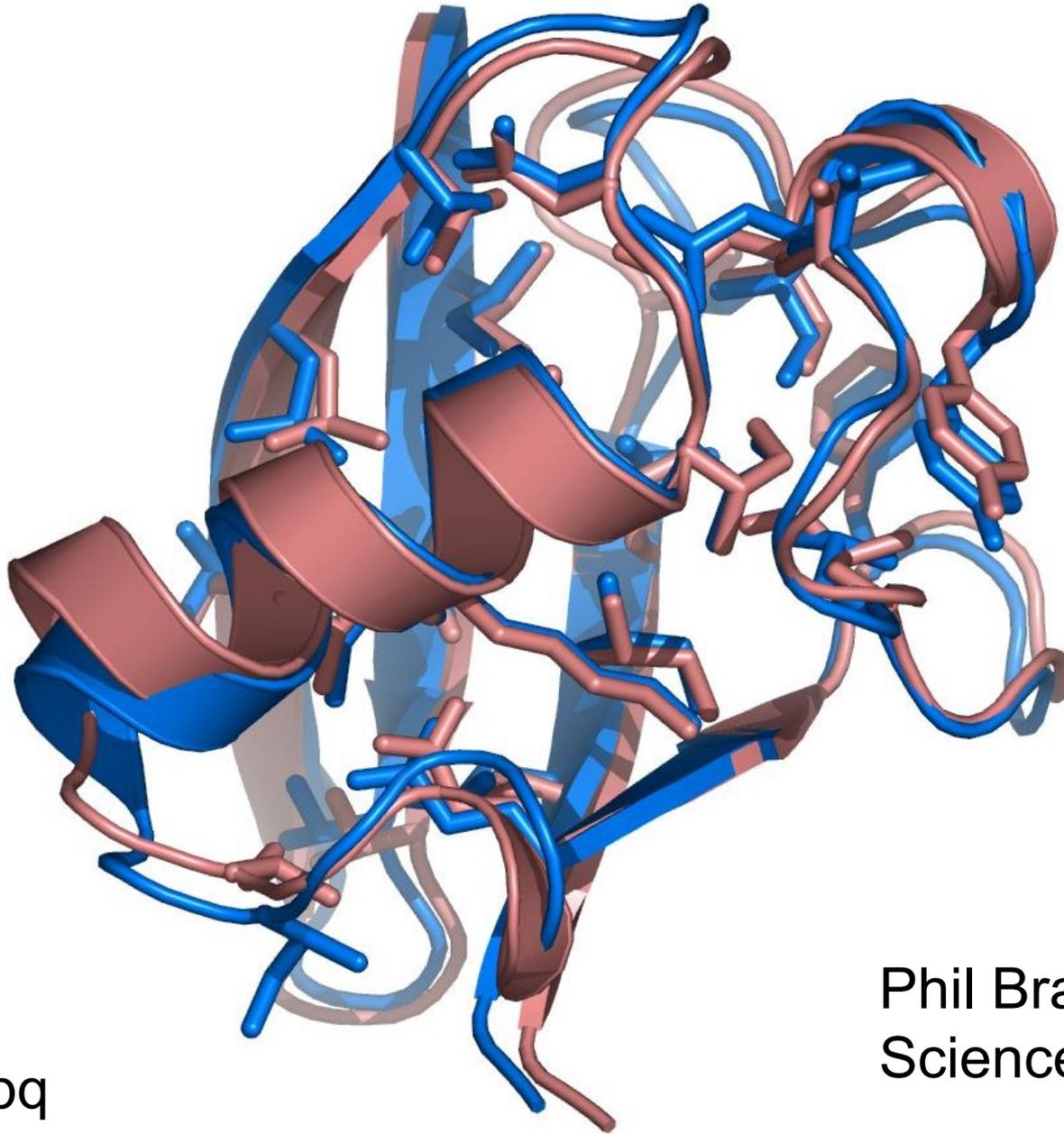
1ubq

Phil Bradley
Science 2005

Highly unrepresentative blind de novo
Rosetta server prediction (CASP9)
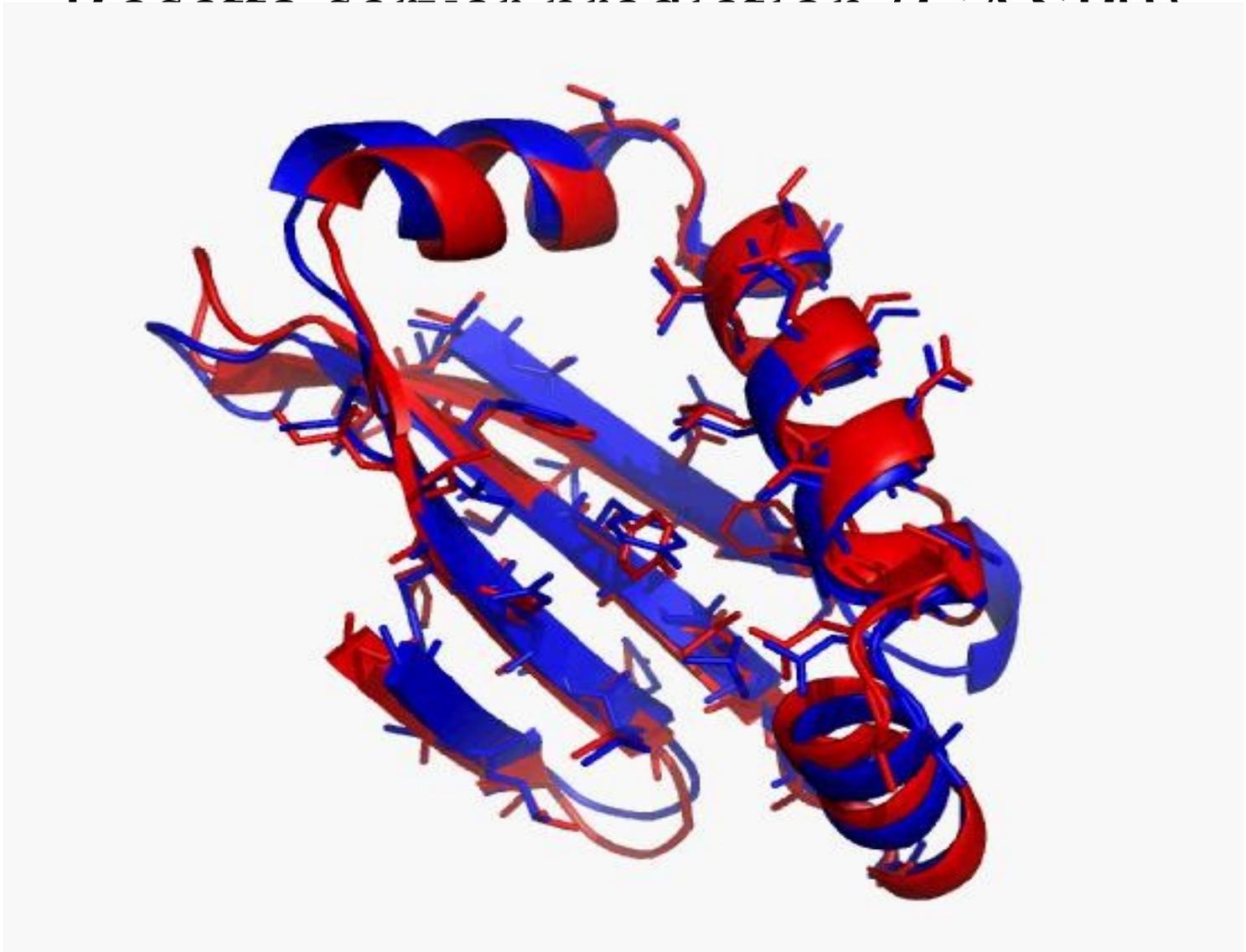
Native free energy gaps recurrent feature of structure prediction problems

- Soluble proteins, multimeric proteins, heterodimers, RNAs, membrane proteins, etc.
- Reflection of very large free energy gaps required for existence of single unique native state
- Prediction possible because (magnitude of actual free energy gap) >> (error in free energy calculation)
- Challenge: how to sample close to native state?

How to find global minimum?

- Smarter algorithms
- Volunteer computing:rosetta@home
- Start closer:comparative modeling
- Use experimental data to limit search
- Collective brain power of game playing humans: http:fold.it

Use experimental data to help locate global minimum

- X-ray diffraction data

- Backbone only NMR data

- Low resolution CryoEM density

- Different from traditional approaches: data guides search, does not specify structure

Strong validation criterion—lower energies in data-constrained calculation
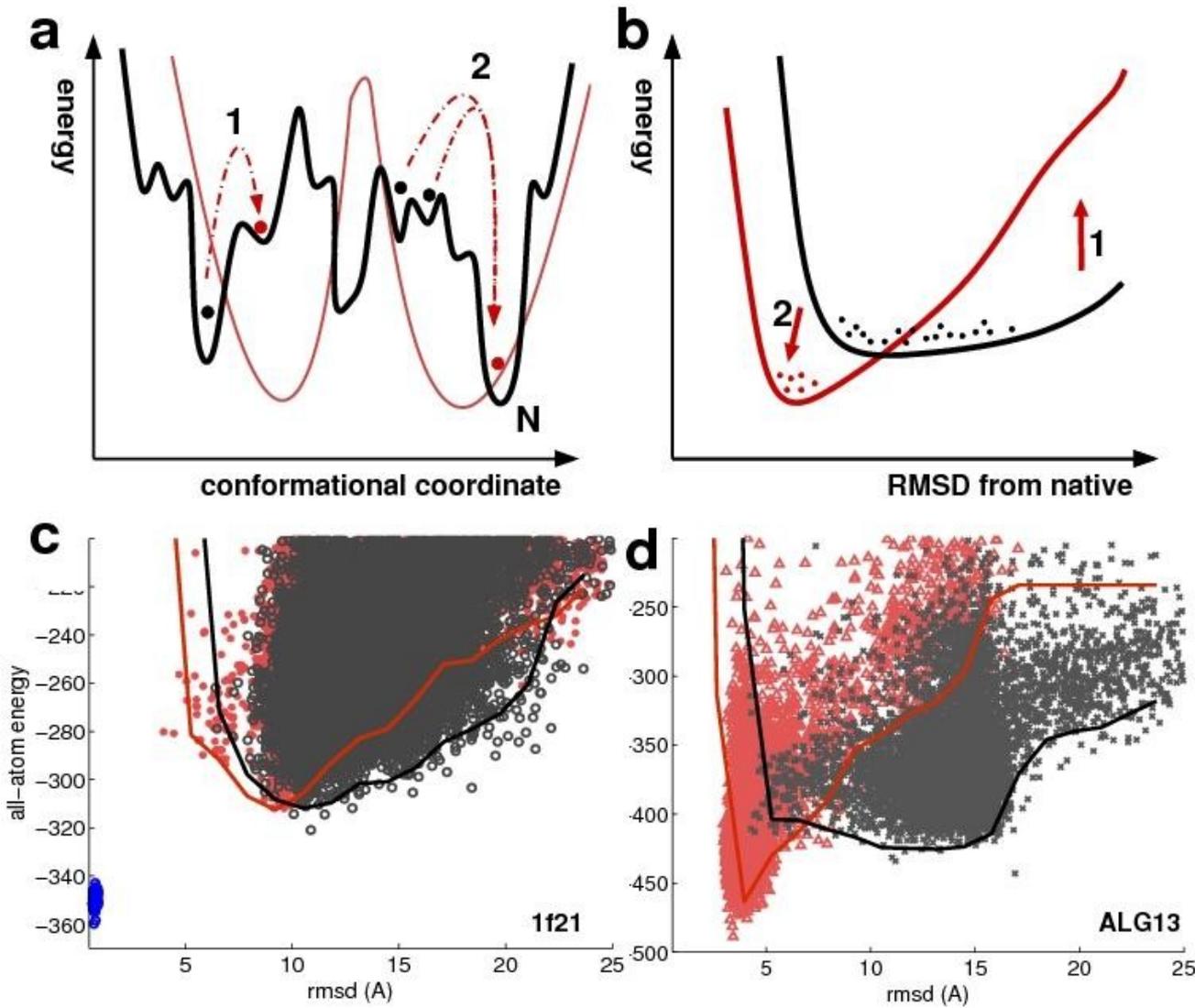
MPMV retroviral protease had resisted crystal structure determination efforts for > 5 years

- Diffraction data collected but no phase information

- Despite extensive efforts, molecular replacement failed with all available templates

- Only known monomeric retroviral protease

- Posted as FoldIt puzzle two months ago

Can you come up with a model that crystallographers can use?

You can see
the variation in
the models
solved by NMR:

http://www.rcsb.org/pdb/explore/explore.do?structureId=1NSO
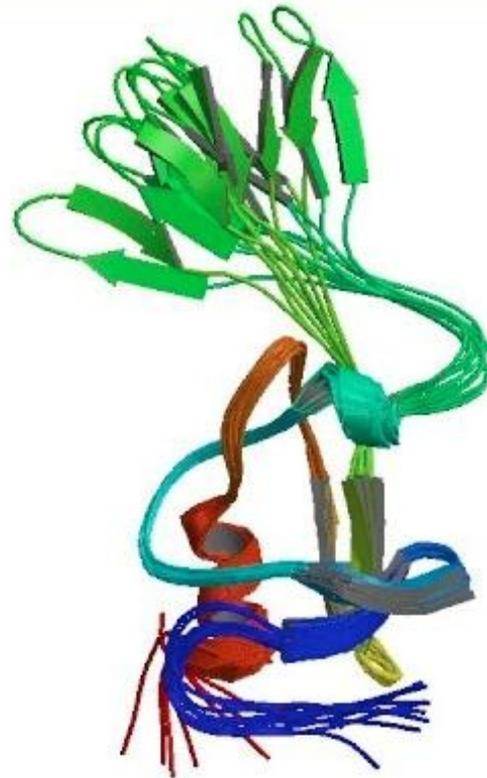
Our hope is that Foldit players can come up with a model that fits the
more recent X-ray crystallographic data better than these NMR
models from 2003. Then we could use that prediction for molecular
replacement (http://en.wikipedia.org/wiki/Molecular_replacement)
and solve this monkey virus protein using X-ray crystallography!

This would be an amazing scientific achievement, as we have been
unable to use Rosetta to solve this particular structure using
molecular replacement, but our lab has been able to do it with other
proteins.

So we are giving you all 10 NMR models as starting structures (every
time you reset the puzzle it will randomly select one of these 10
structures) and all 10 starts are also available in your Template
Reserve in the Alignment Tool (as well as an extended chain
conformation).

Molecular
replacement was
successful with
Mimi's model
and allowed rapid
determination of
the structure
(blue)

grabh
orn

mimi

spvincent

Phaser LLG of optimally aligned model

50
45
40
35
30
25
20
15
10
5
0

0   10000   20000   30000   40000   50000   60000   70000

Model Number
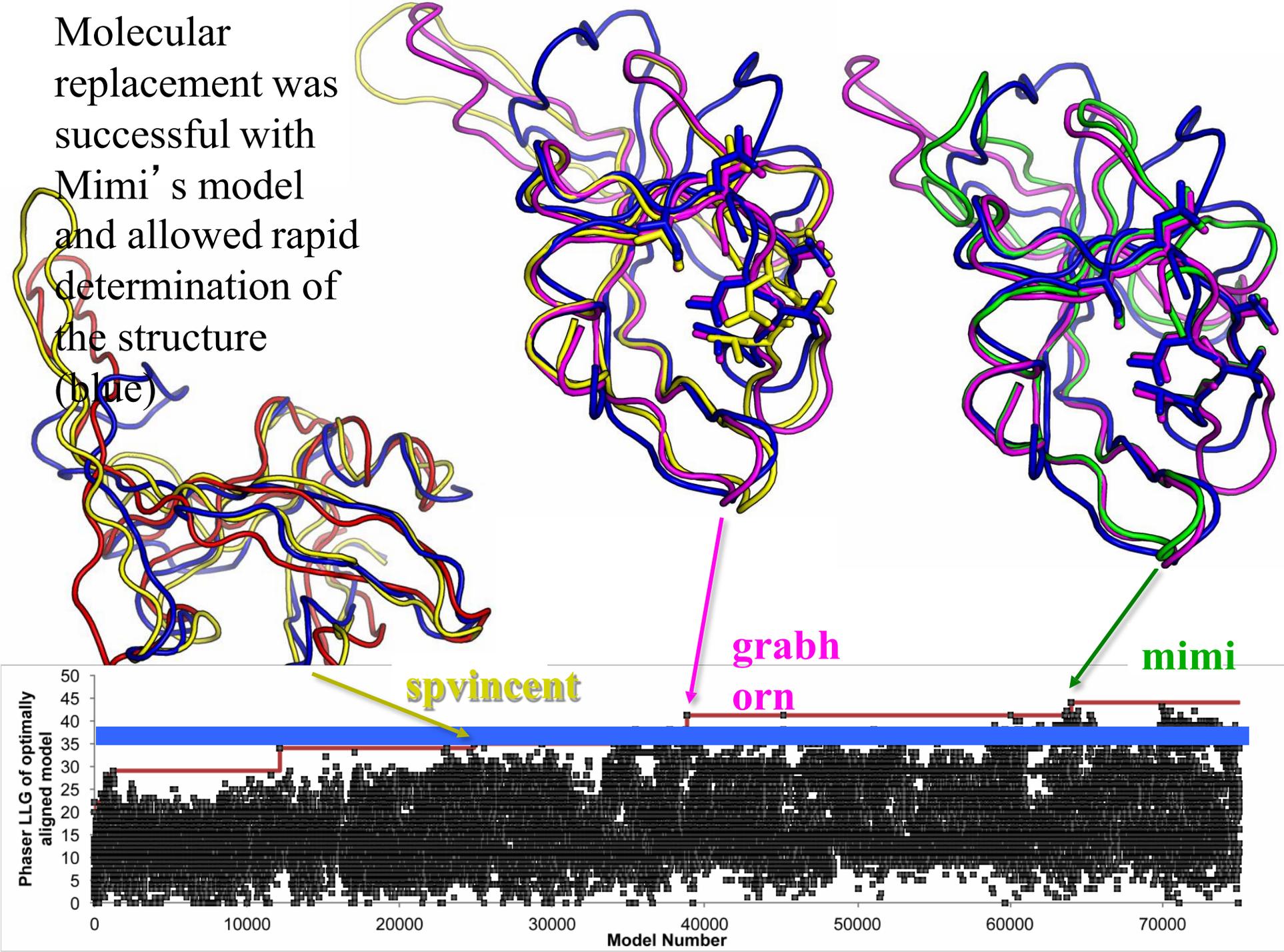
End of lecture

Energy landscapes for 117 proteins
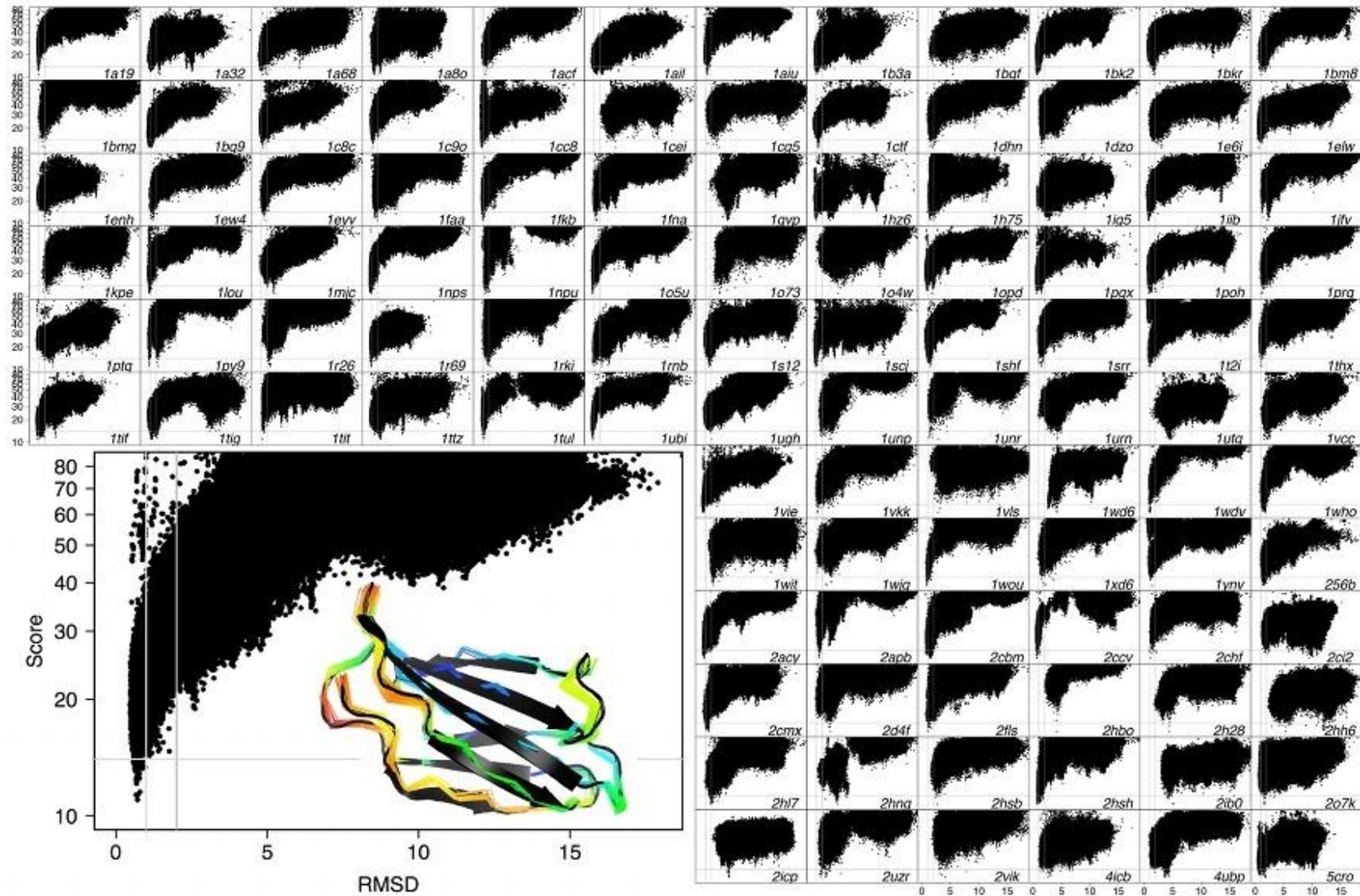
Blind tests of current methods: CASP

- 43-70 new NMR and X-ray structures (unpublished)
- 4000 predictions from 98 different groups
- Types of predictions
  - Homology modeling:predict the structure adopted by a sequence **a** that is related to a sequence **b** with known structure **B**.
  - Fold recognition
  - Ab initio

Native free energy gaps recurrent feature of structure prediction problems

- Soluble proteins, multimeric proteins, heterodimers, RNAs, membrane proteins, etc.
- Reflection of very large free energy gaps required for existence of single unique native state
- Prediction possible because (magnitude of actual free energy gap) >> (error in free energy calculation)
- Challenge: how to sample close to native state?

Structure modeling in combination with experimental data

- Phase diffraction data with models (ab initio, NMR, homology)
- Higher resolution models starting from low resolution X-ray or cryo EM maps
- Accurate and rapid model generation from limited NMR data
- Rosetta now generalized to model
  - Membrane proteins
  - Protein-protein, protein-DNA and protein-small molecule complexes
  - Amyloid fibrils and other symmetric assemblies
  - RNA

FoldIt players can solve hard refinement problems!

Blue = Native
Red = Foldit Puzzle
Green = Foldit Solution
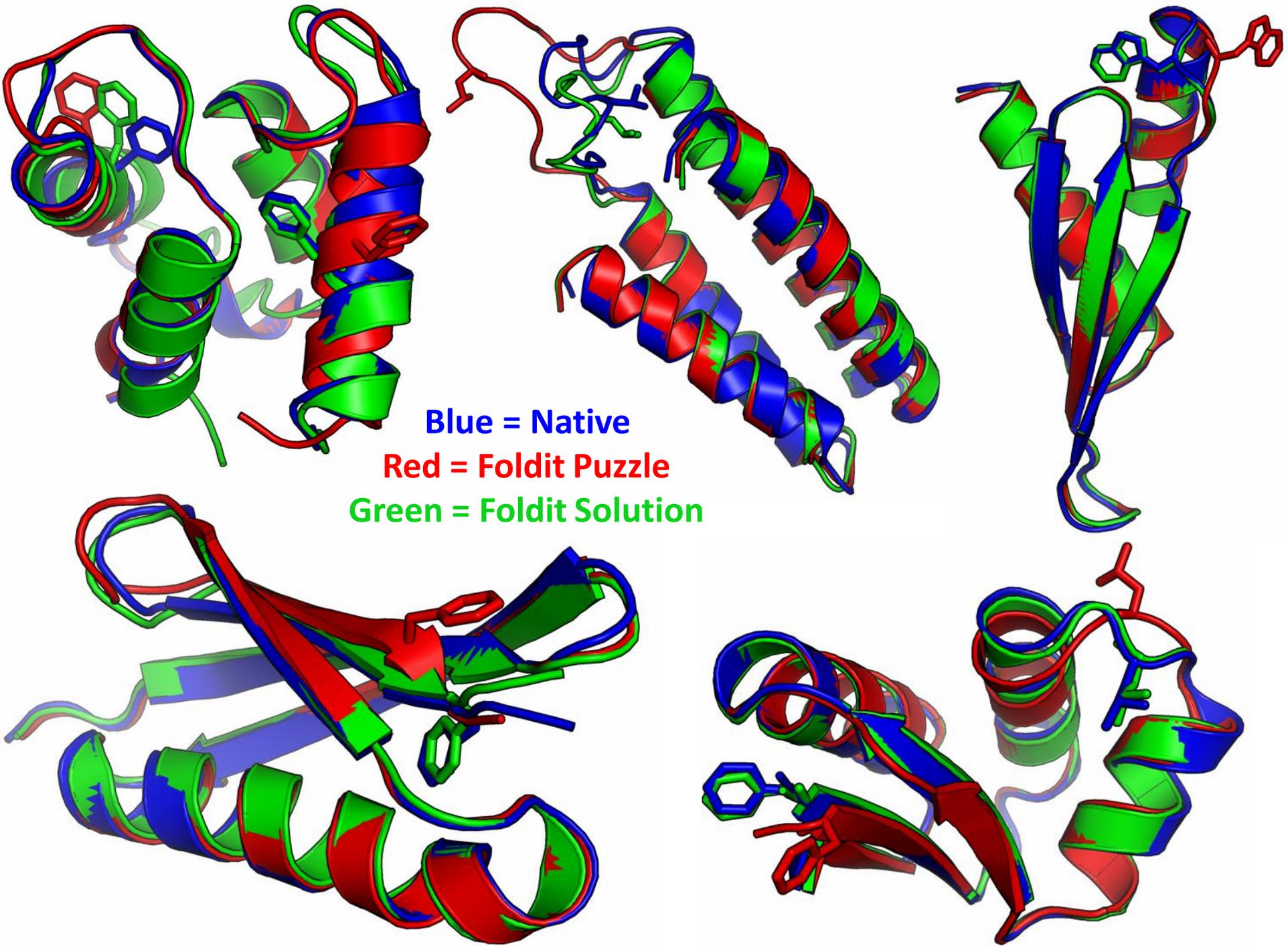
CASP7 target T0283 (112 residues)
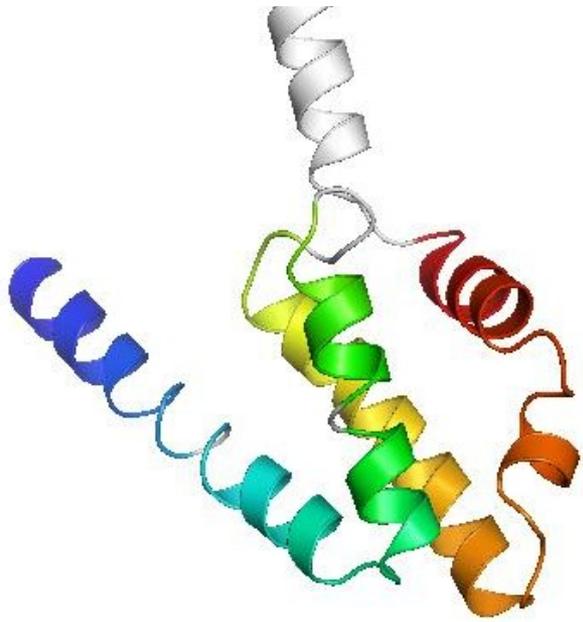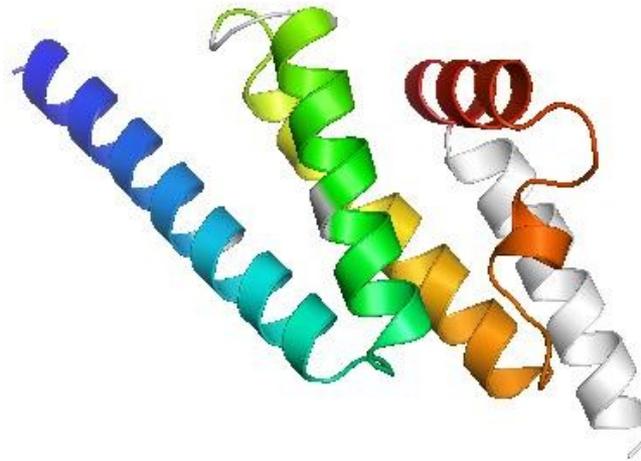
Native   Model 3
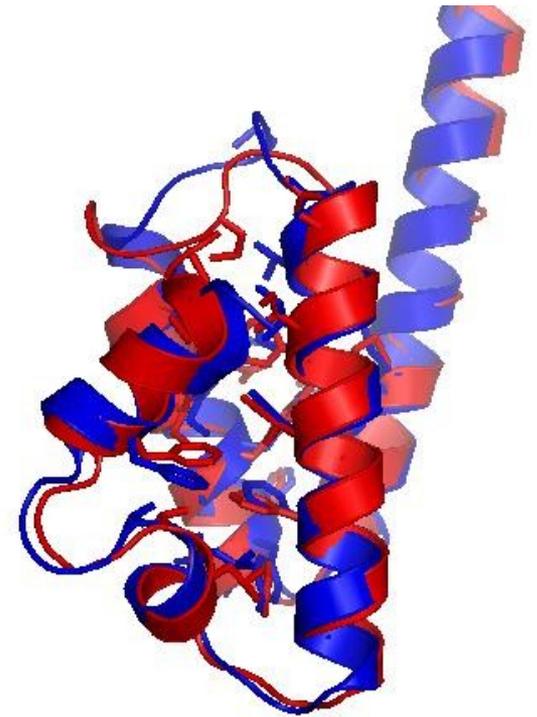
1.40 Å over 90 residues

In some cases, can solve phase problem with computed structures

Red: PDB coordinates from crystal structure phased by selenium SAD

Gray: Electron density map, phased by molecular replacement with ab initio Rosetta model

Rhiju Das, Randy Read, Nature 2007
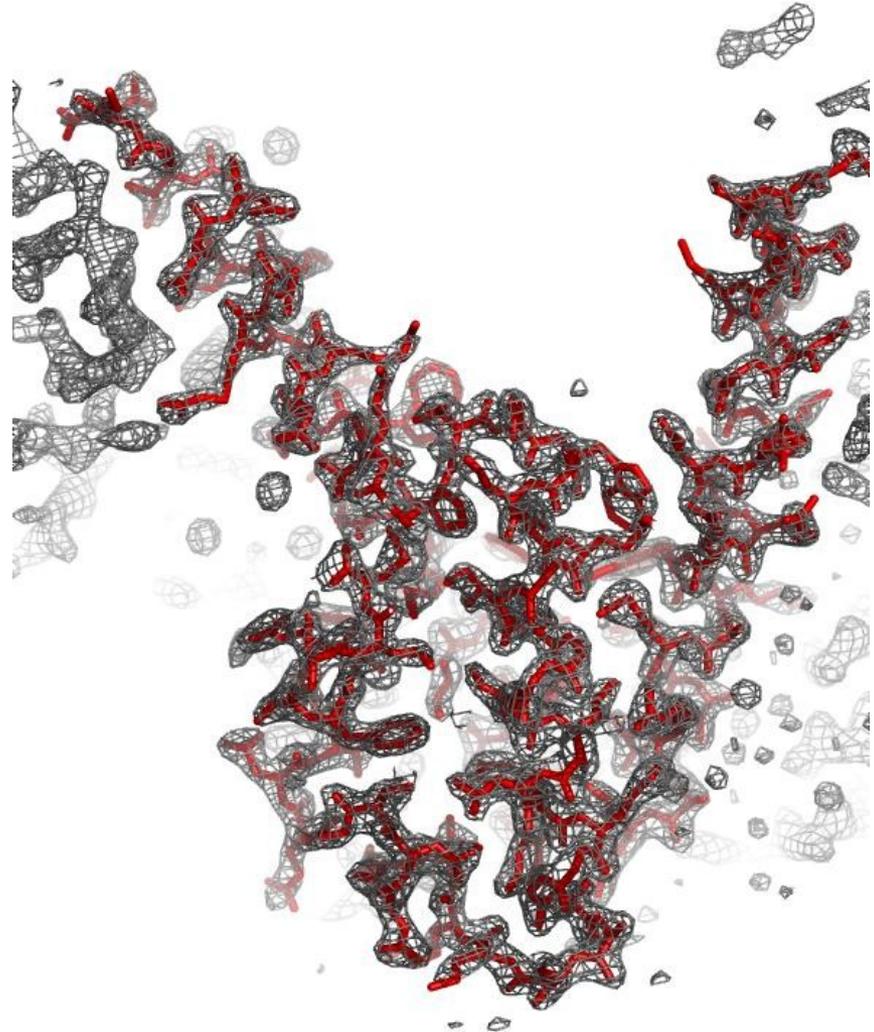
Accurate models from chemical shifts and RDCs: new paradigm for NMR structure determination?
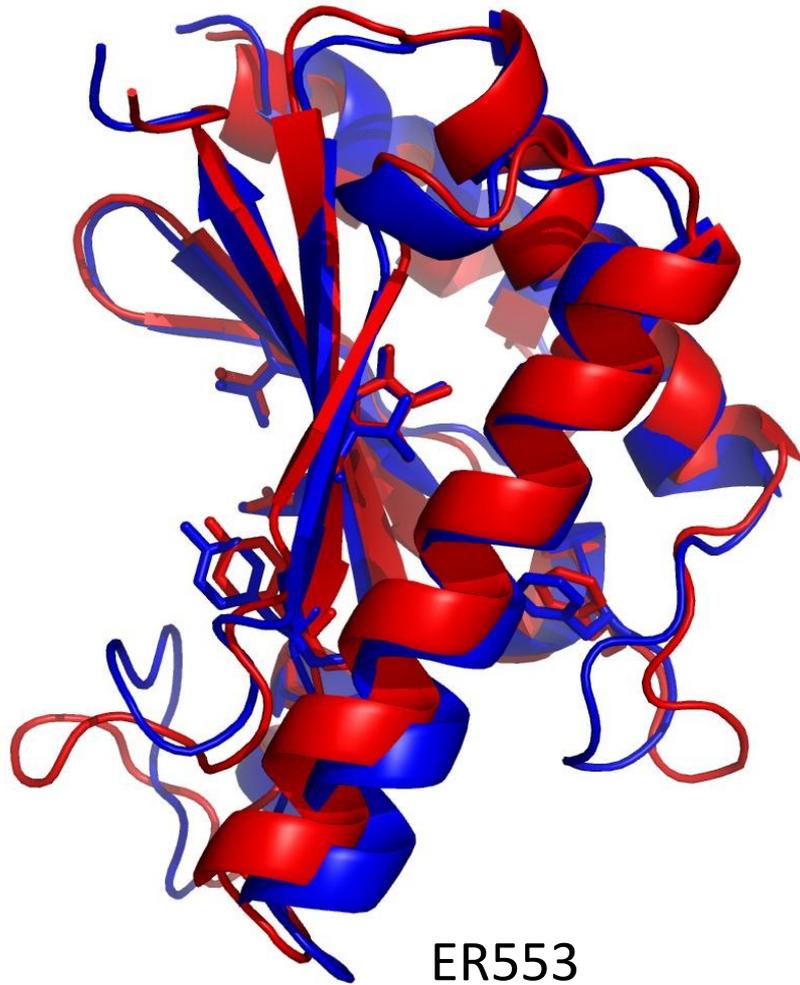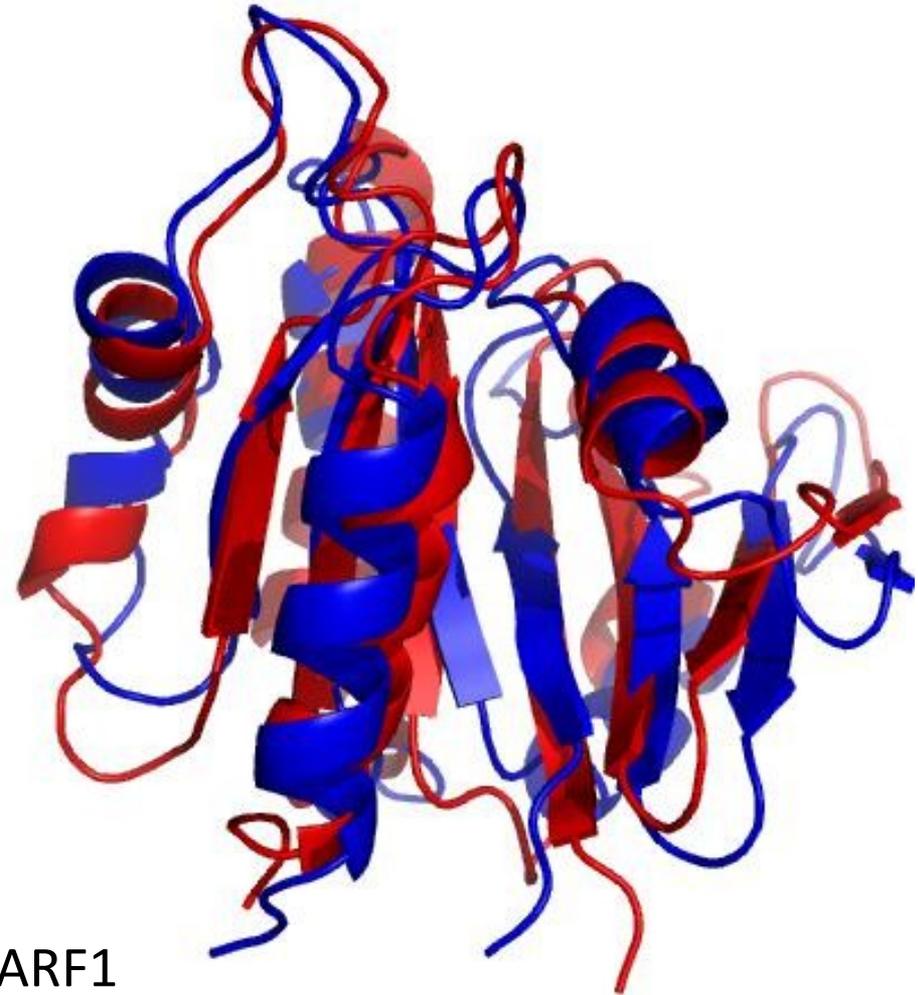
ER553   ARF1
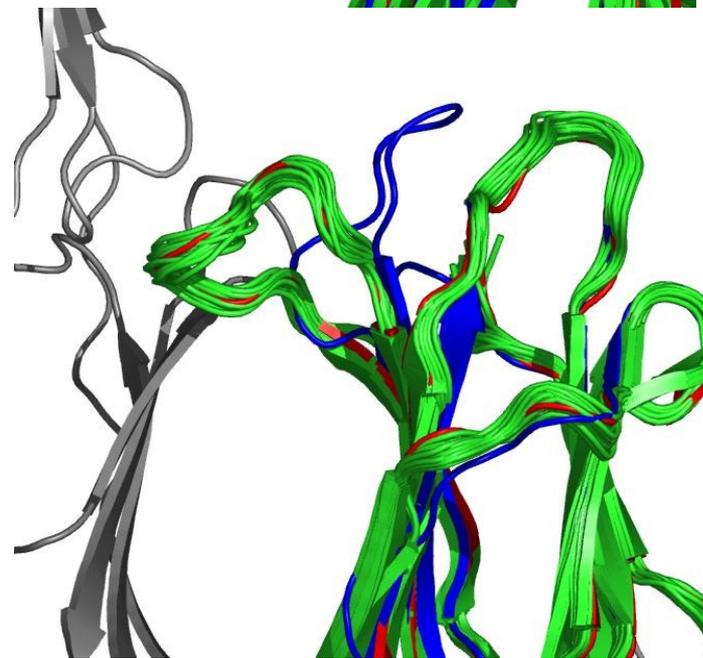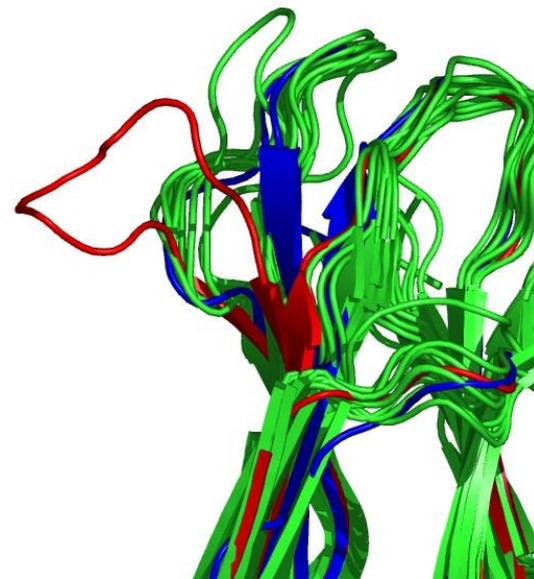**149 aa**   **166 aa**
1.4 Å   **2.6 Å**

BLUE : Native structure
RED : Rosetta model

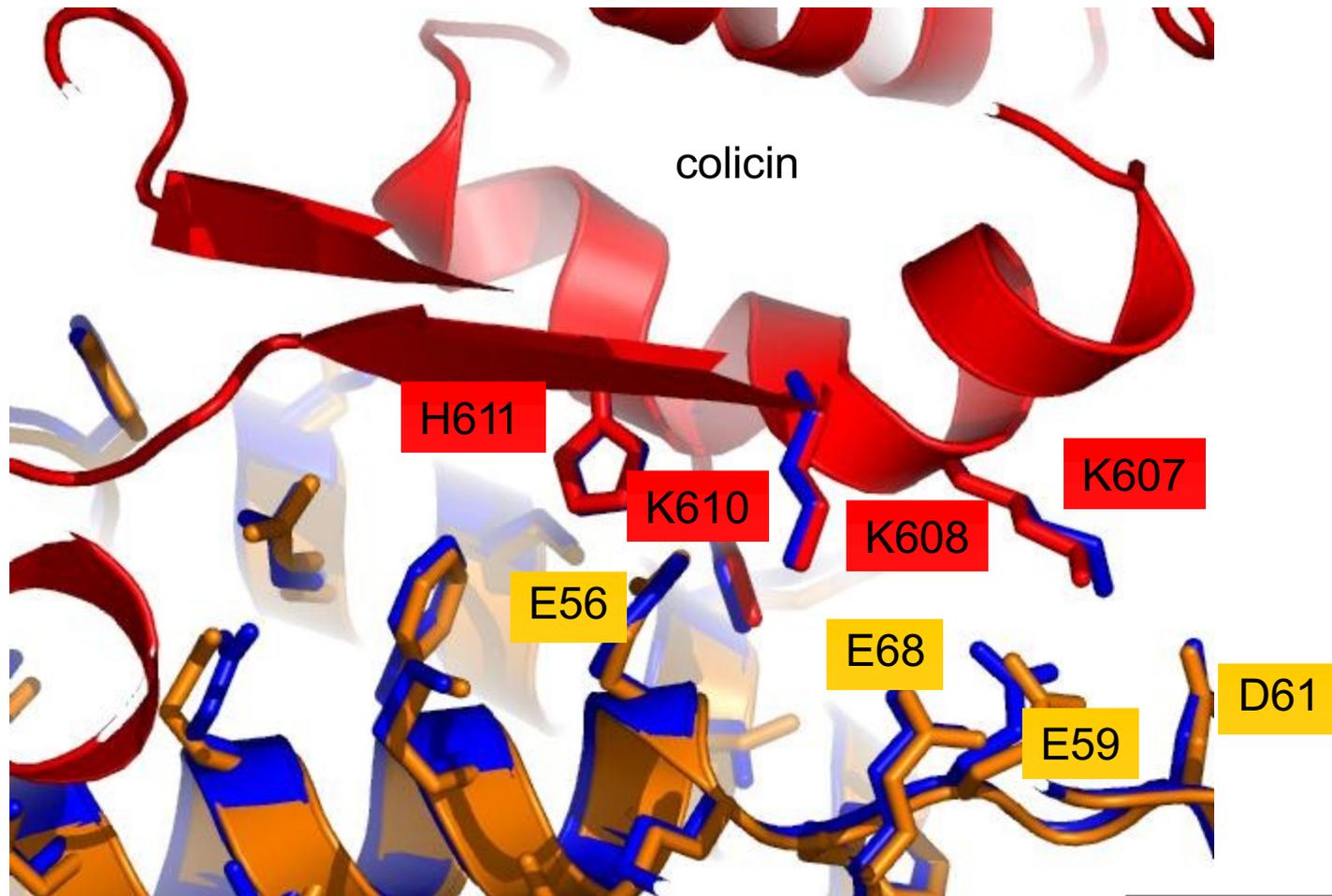# Strong validation criterion—lower energies in data-constrained calculation
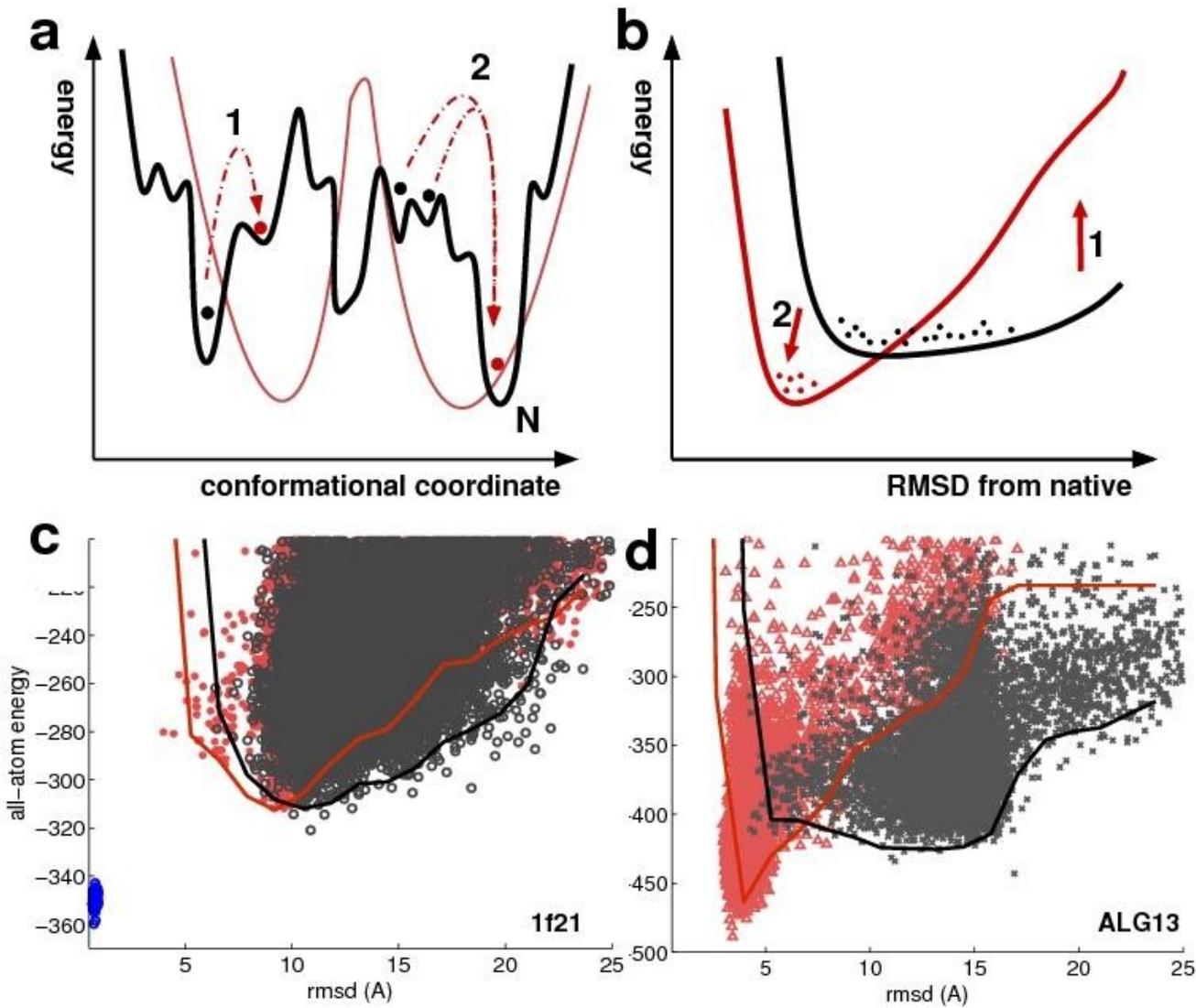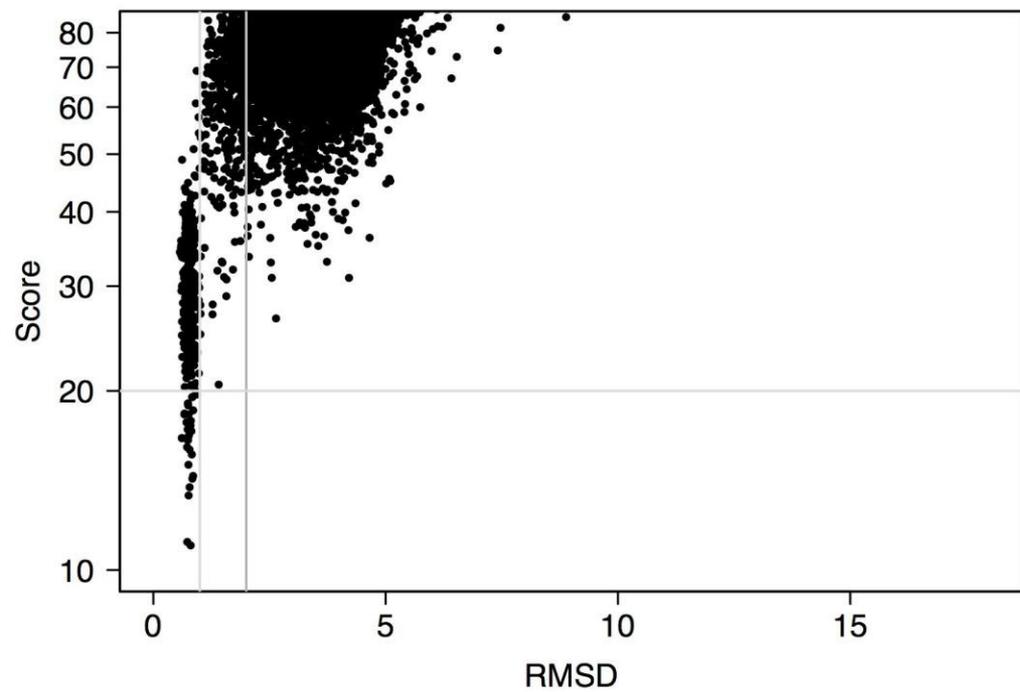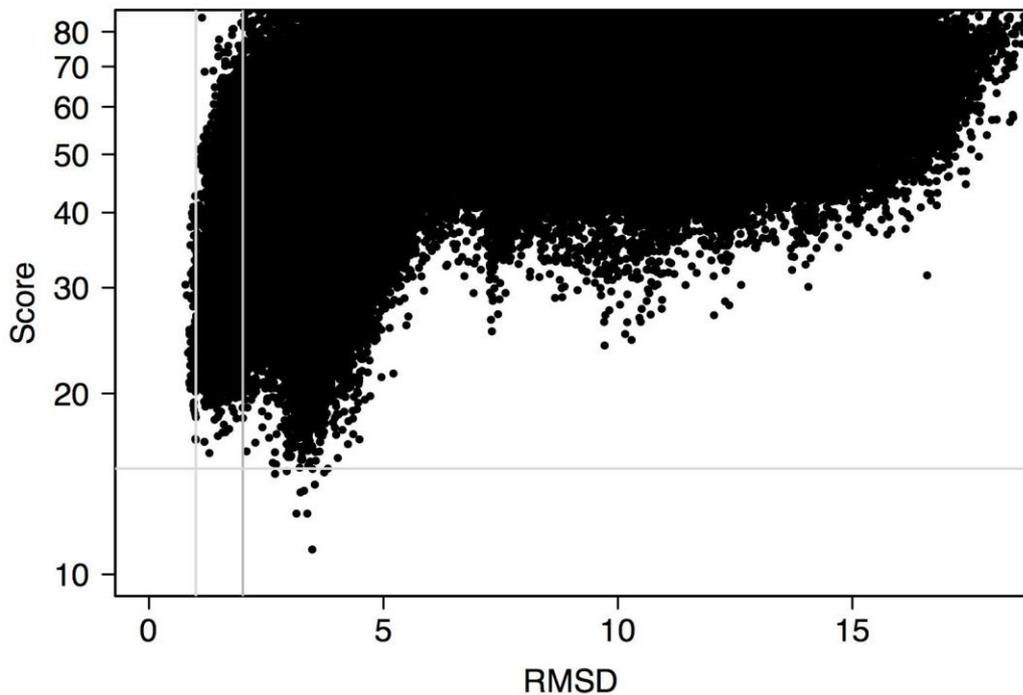
Discrepancies are primarily at crystal contacts!

Protein-protein docking: CAPRI T15 Interface

colicin

H611
K610
K608
K607
E56
E68
E59
D61

immunity protein

red, orange– xray
blue - model

Results

- Homology Modeling
    - three problems:
        1) properly aligning sequence with known structure
        2) remodeling backbone segments with altered structure
        3) repacking the sidechains
    - For 1), psiblast is pretty good.
    - For 2), best to keep the backbone fixed outside of loop regions (current all atom potentials not good enough to let backbone move).
    - 3) is largely solved by rotamer search methods.

- Difficult to improve starting template structure!
- Secondary structure prediction greatly enhanced by multiple sequence information; often quite successful (PsiPred currently the best method, 77% accuracy)
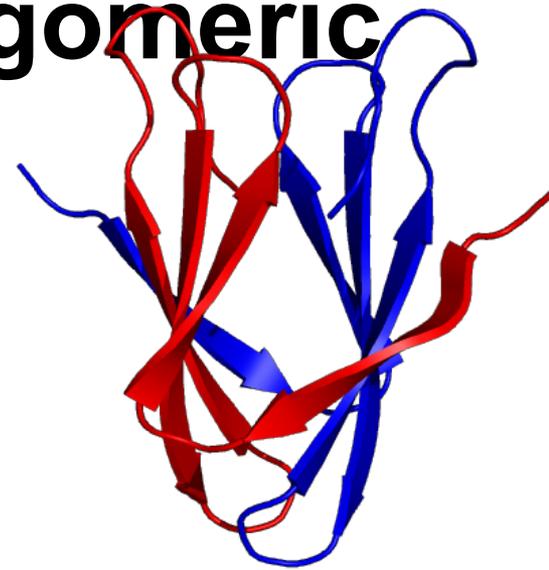
Fold Recognition

Automated web servers do quite well. Best results are with "meta" servers that incorporate results from a variety of different methods and generate significantly more sensitive results than psiblast.
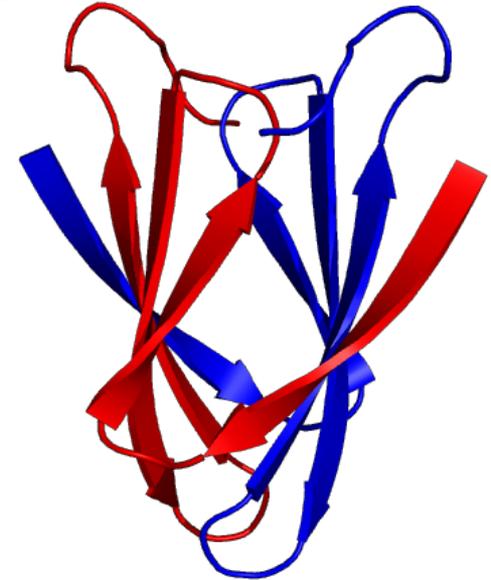
http://bioinfo.pl/LiveBench/

Prediction of homo- oligomeric structures

Sequence

2bti: Model

2bti: Native

Ingemar Andre, Rhiju Das