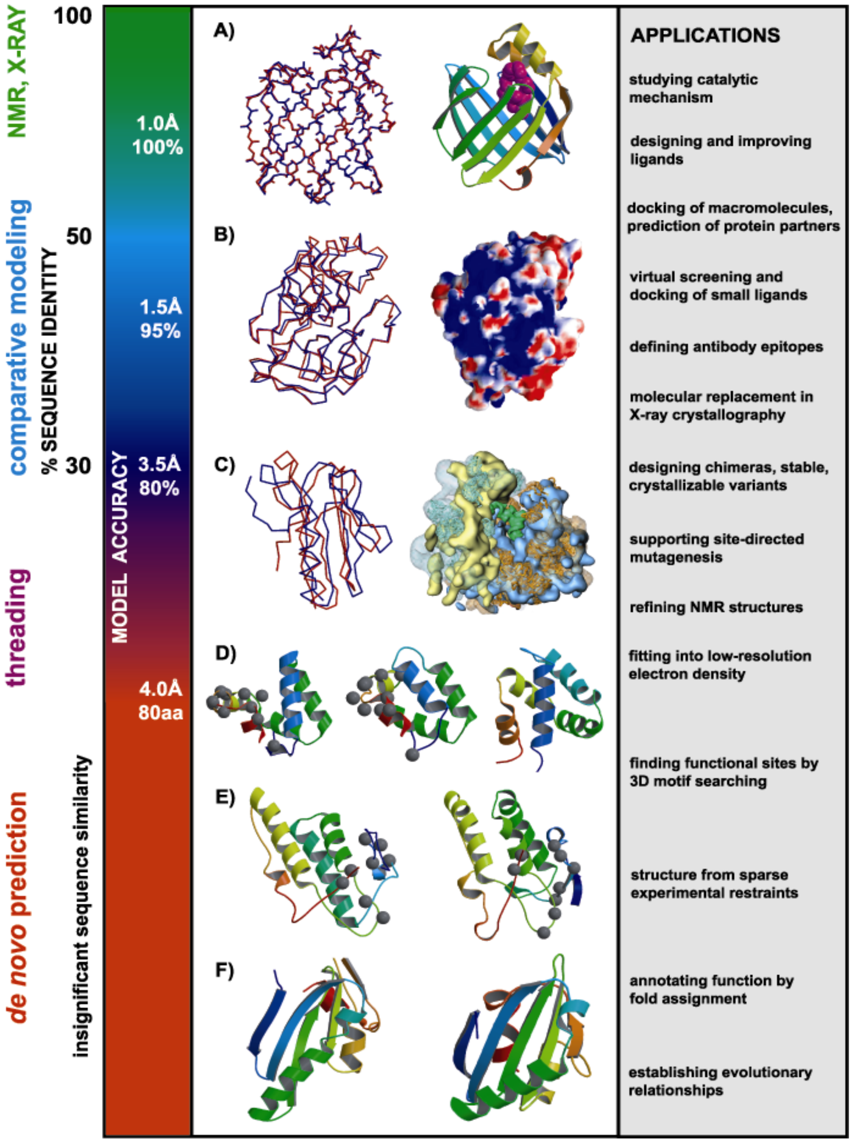
Lecture 10: Fitting and Building Models

1. Model building and fitting into EM maps
2. Comparative and homology modeling
3. Rigid body fitting of atomic models
4. Flexible fitting of atomic models
5. Building models, hybrid methods
6. De novo model building

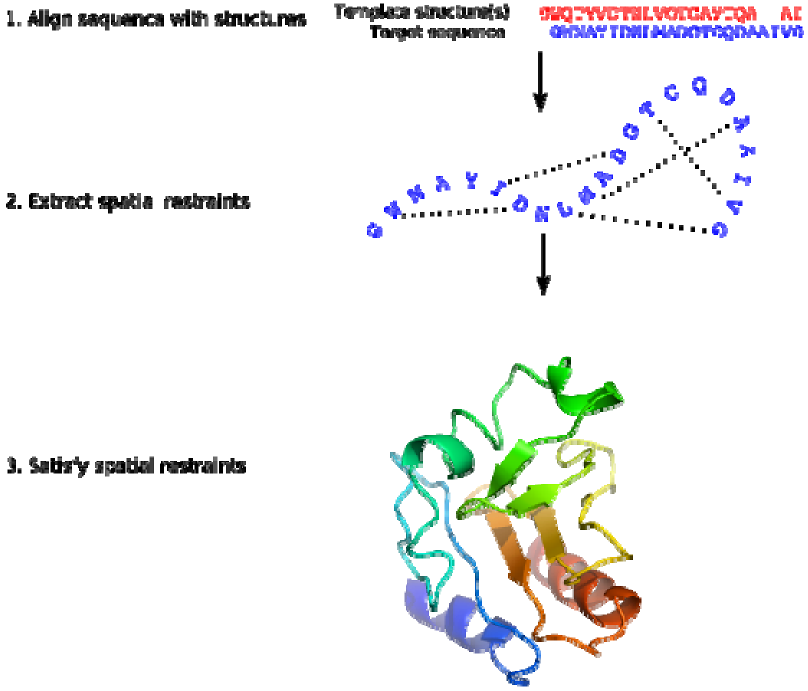
Model Building Approaches


NMR, X-RAY
comparative modeling
threading
de novo prediction
% SEQUENCE IDENTITY
100
50
30
insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4.0Å
80aa
A)
B)
C)
D)
E)
F)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
finding functional sites by 3D motif searching
structure from sparse experimental restraints
annotating function by fold assignment
establishing evolutionary relationships

D. Baker & A. Sali. Science 294, 93, 2001.

Comparative Modeling

- Many more sequences available than structures
- Many applications rely on structural information
- Structure is often more conserved than sequence (evolution preserves function)


1. Align sequence with structures
Template structure(s)
Target sequence
2. Extract spatial restraints
3. Satisfy spatial restraints

1) Assembly of rigid bodies (core, loops, sidechains)

2) Segment matching

3) Satisfaction of spatial restraints

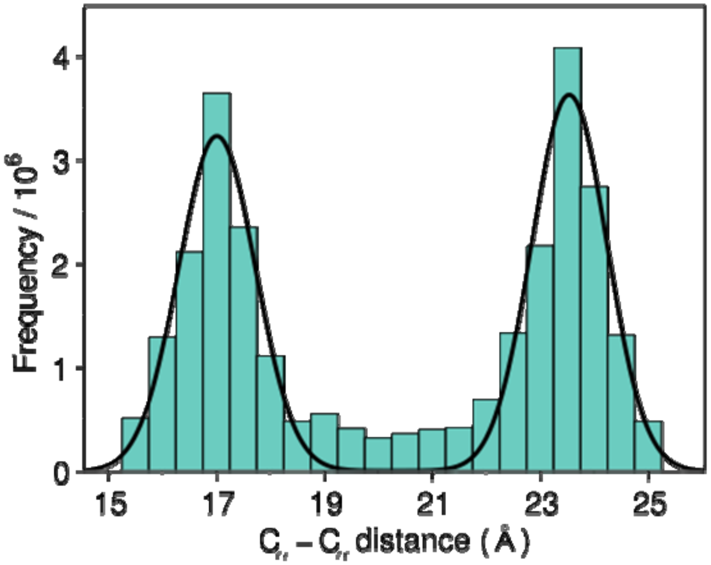
A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

Comparative Modeling

- First, must determine the template structures
  - Simplistically, try to align the target sequence against every known structure's sequence
  - In practice, this is too slow, so heuristics are used (e.g. BLAST)
  - Profile or HMM searches are generally more sensitive in difficult cases (e.g. Modeller's profile.build method, or PSI-BLAST)
  - Could also use threading or other web servers
- Alignment to templates generally uses global dynamic programming
  - Sequence-sequence: relies purely on a matrix of observed residue-residue mutation probabilities ('align')
  - Sequence-structure: gap insertion is penalized within secondary structure (helices etc.) ('align2d')
  - Other features and/or user-defined ('salign') or use an external program

Comparative Modeling

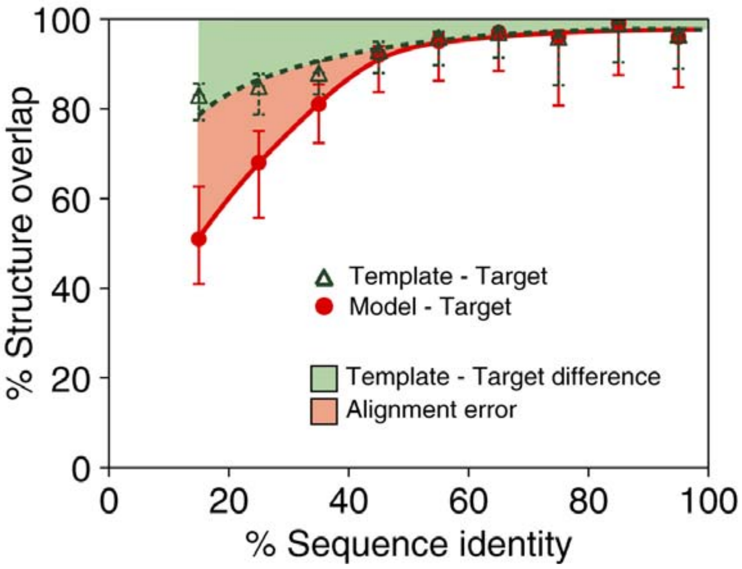
- Spatial restraints incorporate homology information, statistical preferences, and physical knowledge
  - Template Cα- Cα internal distances
  - Backbone dihedrals (φ/ψ)
  - Sidechain dihedrals given residue type of both target and template
  - Force field stereochemistry (bond, angle, dihedral)
  - Statistical potentials
  - Other experimental constraints
  - etc.



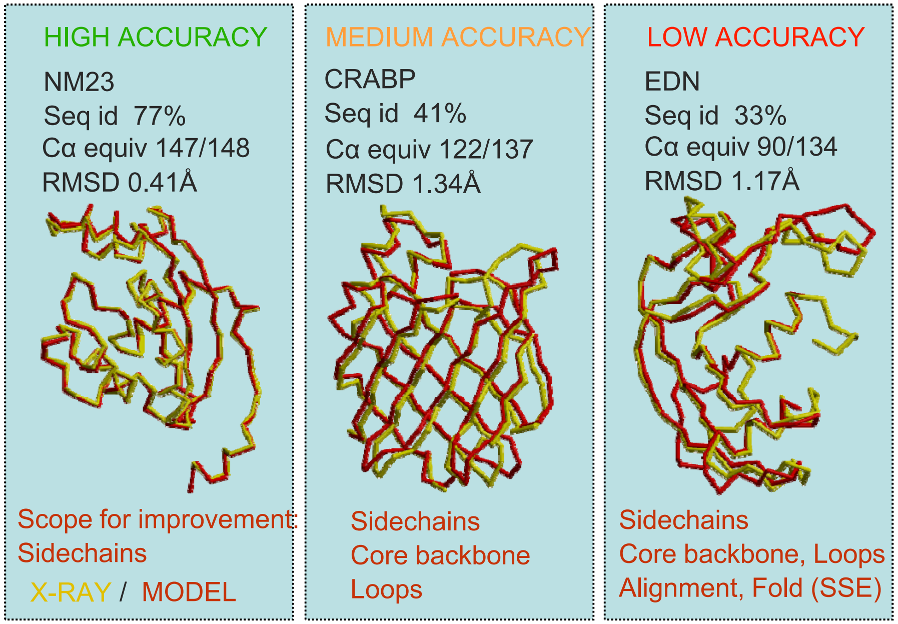
Comparative Modeling

- All information is combined into a single objective function (restraints are converted to an “energy” by taking the negative log)
- Function is optimized by conjugate gradients and simulated annealing molecular dynamics, starting from the target sequence threaded onto template structure(s)

Model Accuracy vs. Sequence Identity



Sánchez, R., Sali, A. PNAS (1998) 95, 13597



HIGH ACCURACY

NM23
Seq id 77%
Cα equiv 147/148
RMSD 0.41Å

Scope for improvement:
Sidechains

X-RAY / MODEL

MEDIUM ACCURACY

CRABP
Seq id 41%
Cα equiv 122/137
RMSD 1.34Å

Sidechains
Core backbone
Loops

LOW ACCURACY

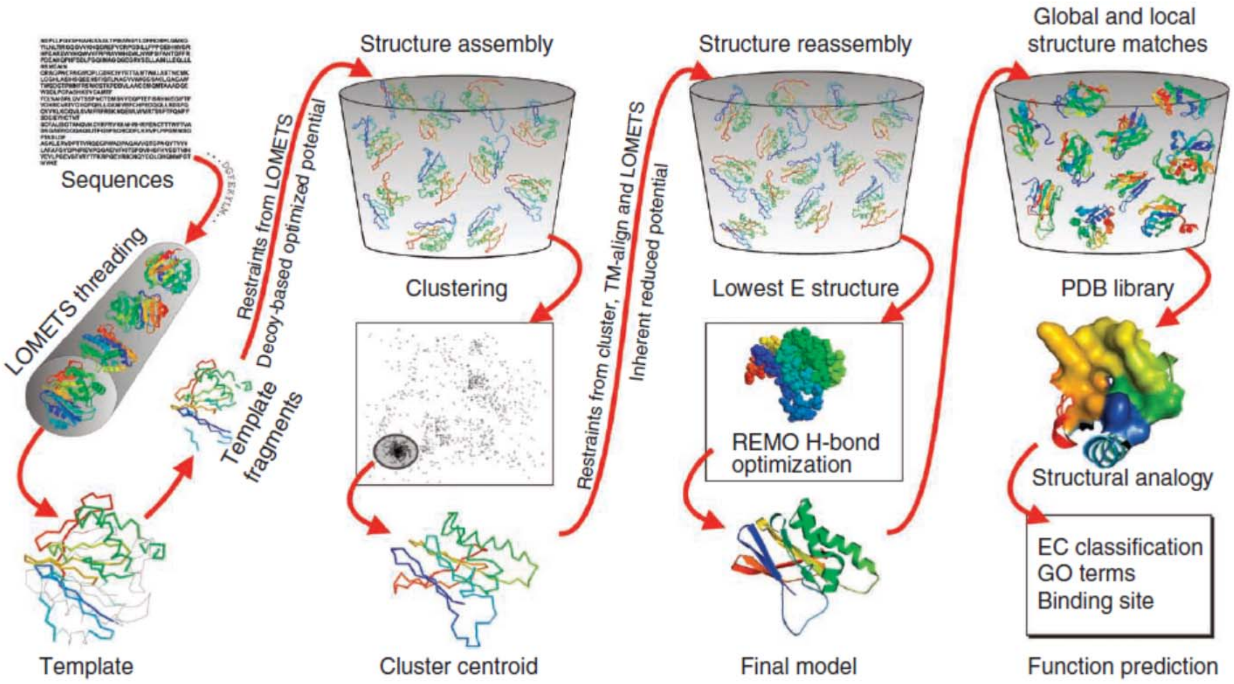
EDN
Seq id 33%
Cα equiv 90/134
RMSD 1.17Å

Sidechains
Core backbone, Loops
Alignment, Fold (SSE)

Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

I-TASSER: Protein Structure Prediction

I-TASSER workflow:



Accuracy estimation:

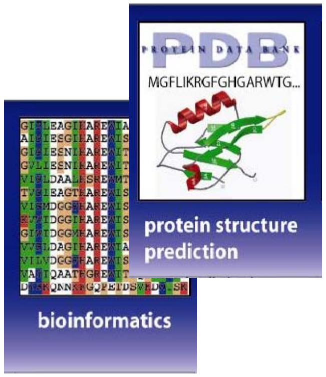
$$\text{C-score} = \ln\left(\frac{M}{M_{\text{tot}}} \times \frac{1}{\langle \text{RMSD} \rangle} \times \frac{1}{7}\sum_{i=1}^{7}\frac{Z(i)}{Z_0(i)}\right)$$

C-score > -1.5
=> 90% correct topology

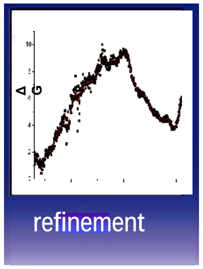
Roy, A. et al. (2010) Nat. Protocols, 5, 725.

Comparative Modeling

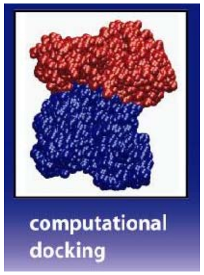
- **Problem:** comparative models are often inaccurate.
- **Solution:** Use cryoEM maps to assess the models by **rigid density fitting.**



- **Problem:** the structures may exhibit conformational changes (induced fit, target-template differences).
- **Solution**: use **flexible fitting** to refine the structures in the map.

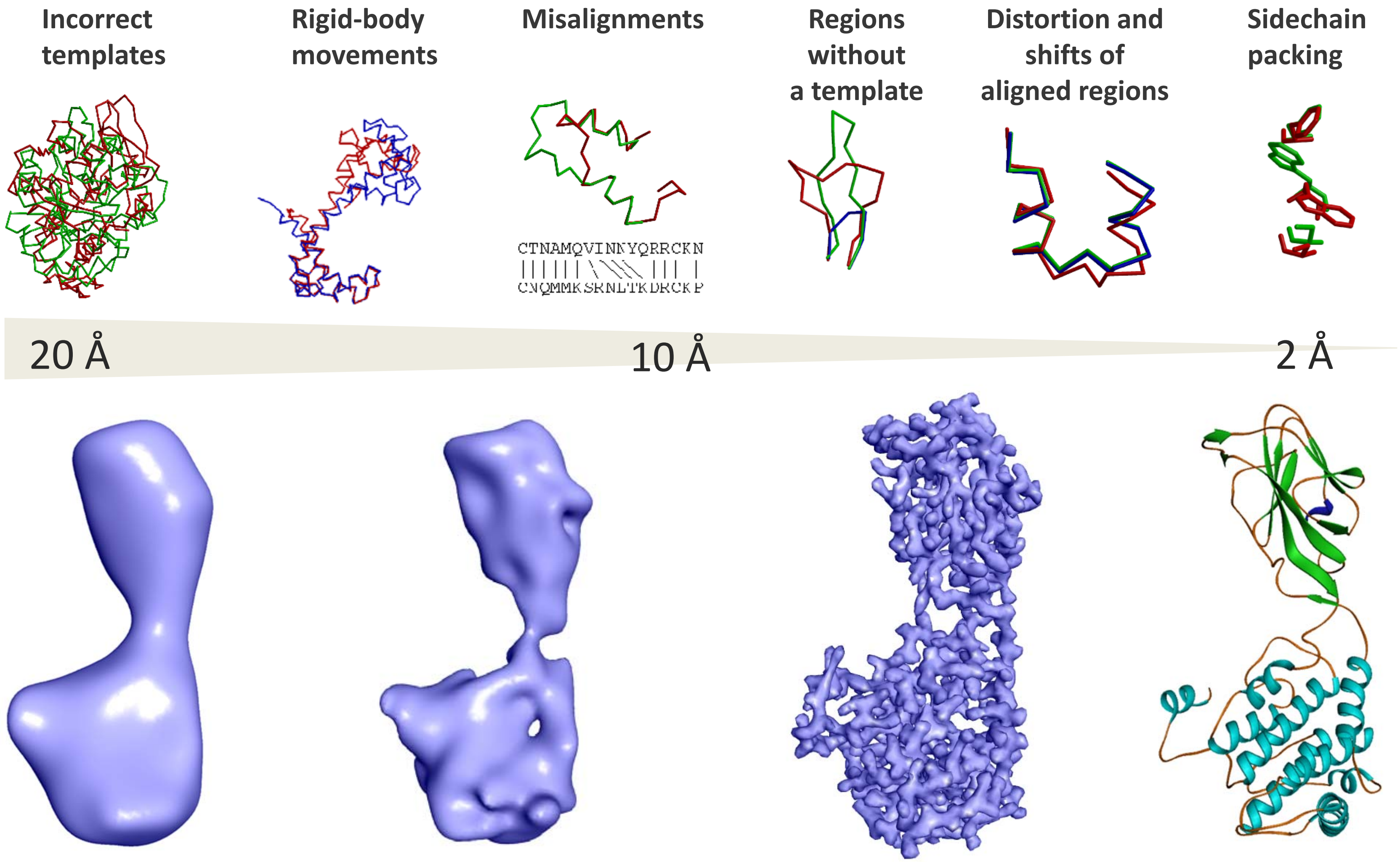


- **Problem:** the resolution of the map can be too low for an unambiguous placement of a component.
- **Solution:** use additional information to determine the **assembly architecture.**



Topf & Sali. *Curr Opin Struct Biol* 2005.

Errors in Comparative Modeling



Incorrect templates

Rigid-body movements

Misalignments

Regions without a template

Distortion and shifts of aligned regions

Sidechain packing

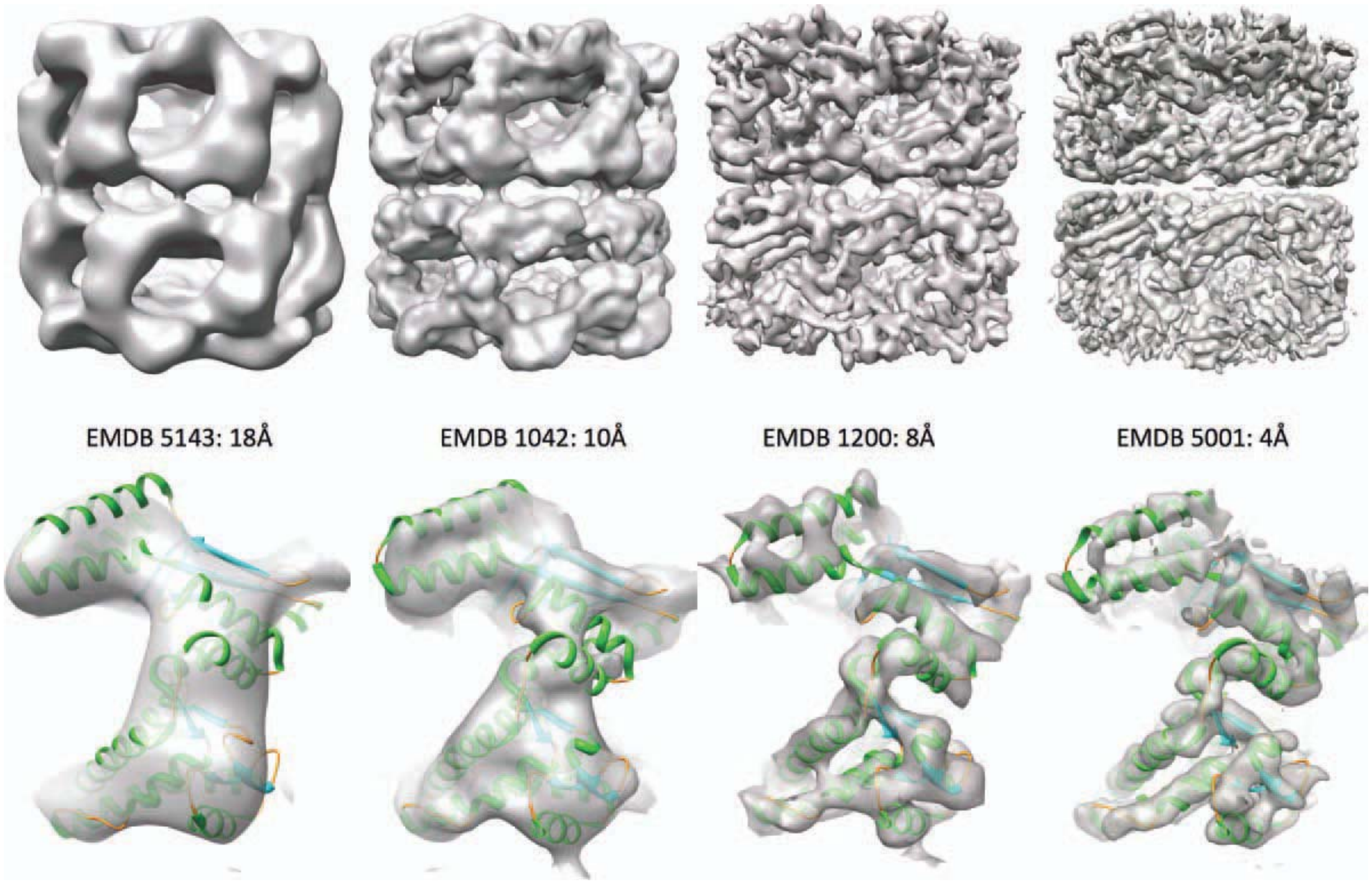
20 Å

10 Å

2 Å

Information & EM-map Resolution

GroEL at different resolutions (levels of detail)



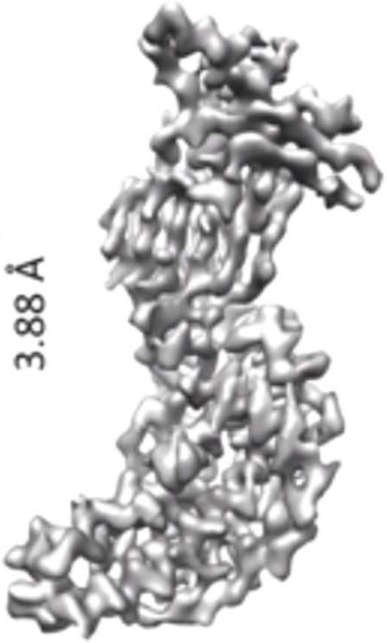
Fitting of known structures
(rigid body fitting)

Flexible fitting of
known structures

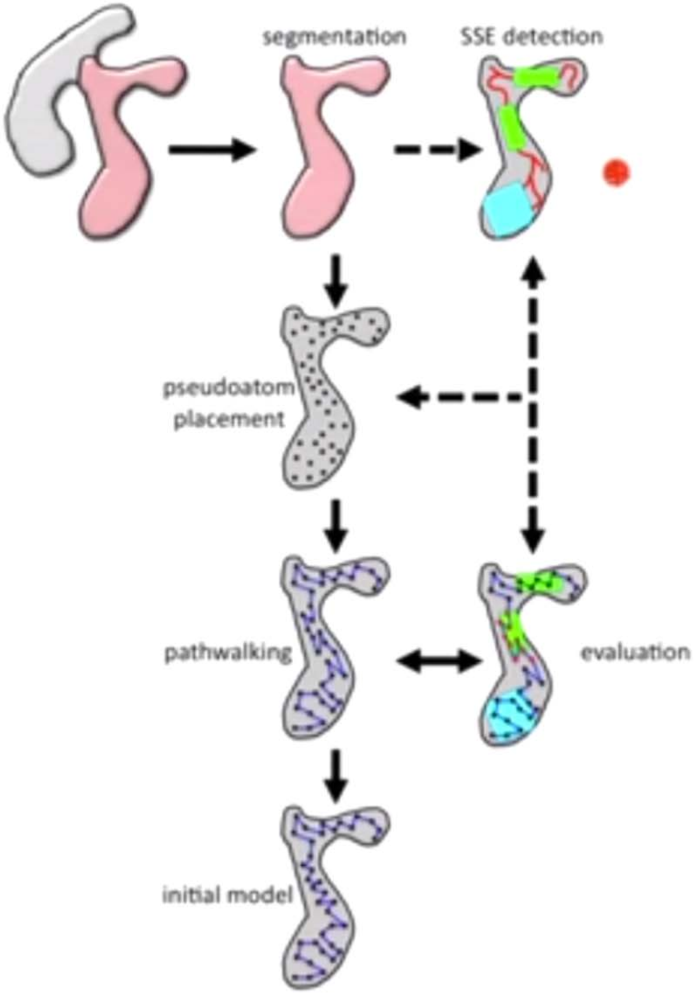
Building of
de novo models

Model Building Approaches

Finding secondary structures and building models



Rotavirus-vp6
3.88 Å

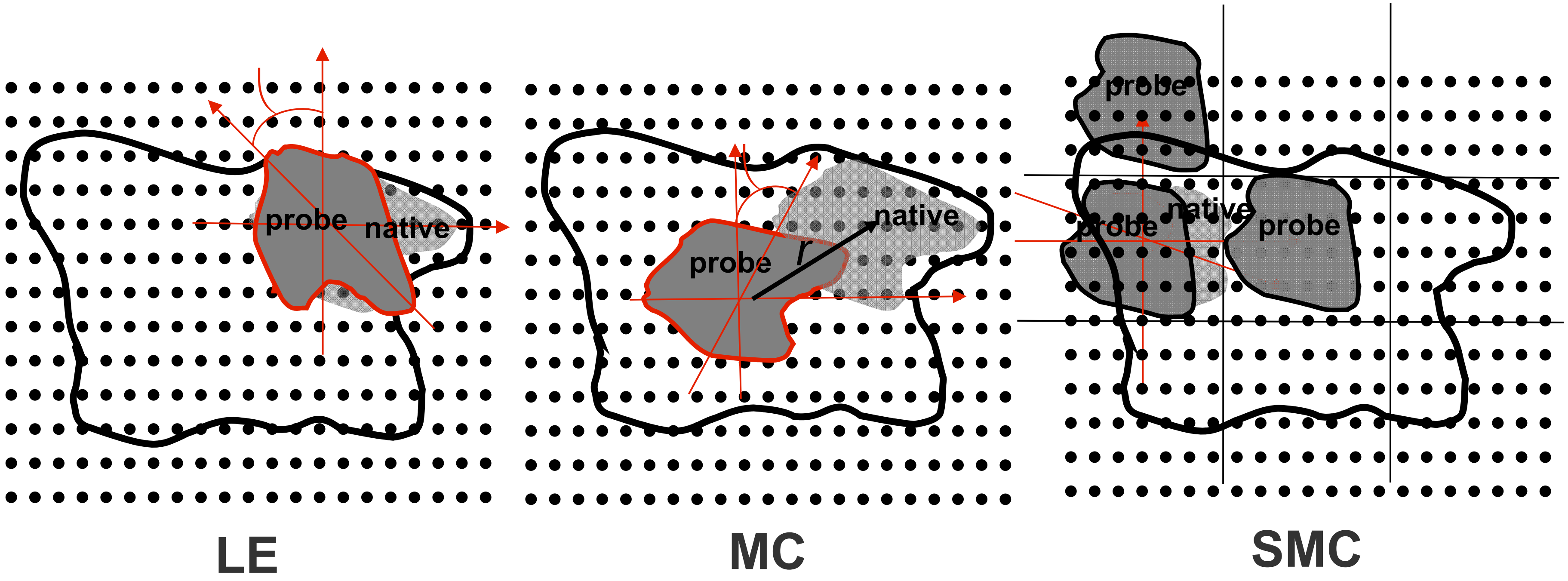


"Pathwalker" Baker et al., Structure 2012

Rigid Body Fitting of Known Structures

$$CC(R_a, r_k) = \sum_{j=1}^{J} \rho^{EM}(r_j)\,\rho^{probe}(R_a r_j + r_k)$$

- LE - Local exhaustive search (rotations only or rotations+translations)
- MC - Monte Carlo in translation, with exhaustive rotation
- SMC - Scanning of the map to find regions with high CC; LE or MC search



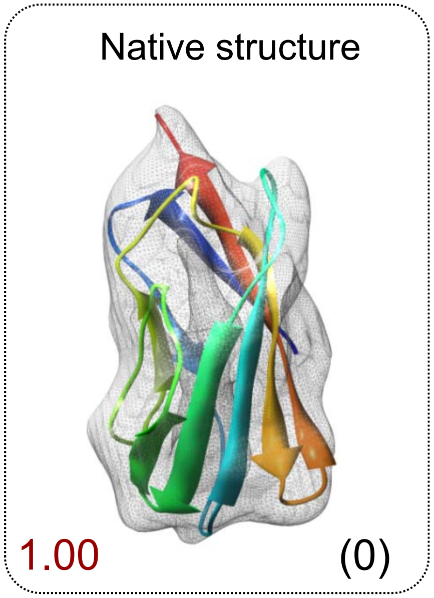
LE

MC

SMC

Topf, Baker, John, Chiu & Sali. *J Struct Biol* 2005.

Rigid Body Fitting of Known Structures


Native structure

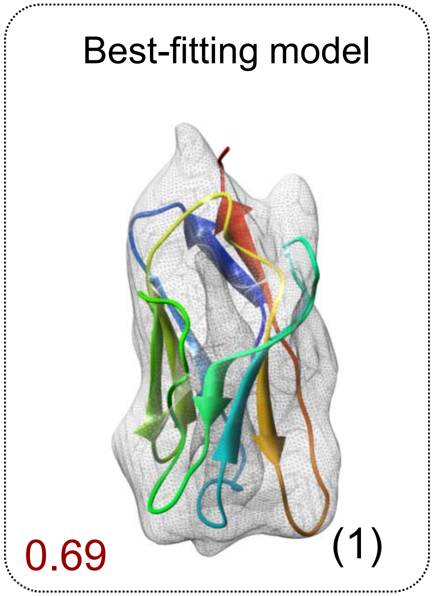
1.00

(0)

1cid:2rhe

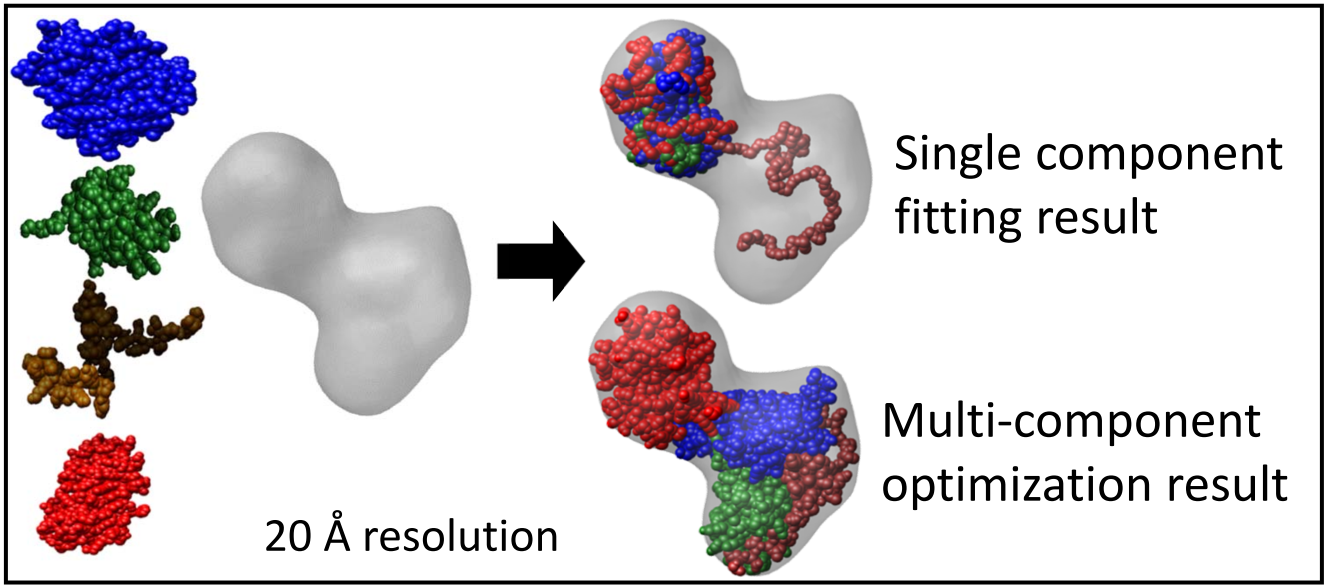
12% seq. identity

10 Å resolution


Best-fitting model

0.69

(1)


20 Å resolution

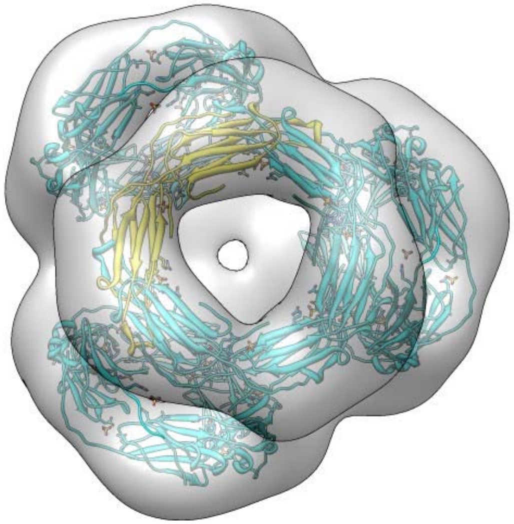
Single component fitting result

Multi-component optimization result

Rigid Body Fitting of Known Structures

Avoiding fitting clashes –> Sequential fitting

- Fit sequentially the three monomers and subtract density.
- Fits each in turn subtracting the other two from the density first.
- Repeat last command to get better convergence.



Avoiding fitting clashes –> Symmetric fitting

- Fit one monomer taking account of clashes of symmetrically placed monomers.
- This optimizes the correlation of the full symmetric assembly by moving only one monomer.
- This avoids clashes because if two monomers overlap they create double density that gives poor correlation with experimental map. Clashes are implicitly avoided and there is no special repulsion introduced.
- Fit command in Chimera"fit #1 #0 res 20 sym true".

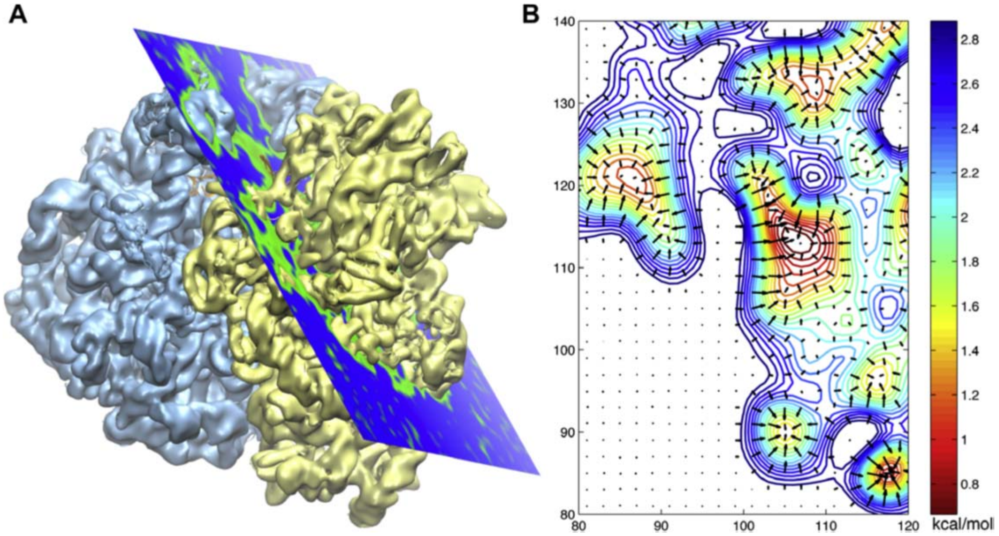
MDFF: Flexible Fitting of Known Structures

Additional potential from the EM map:

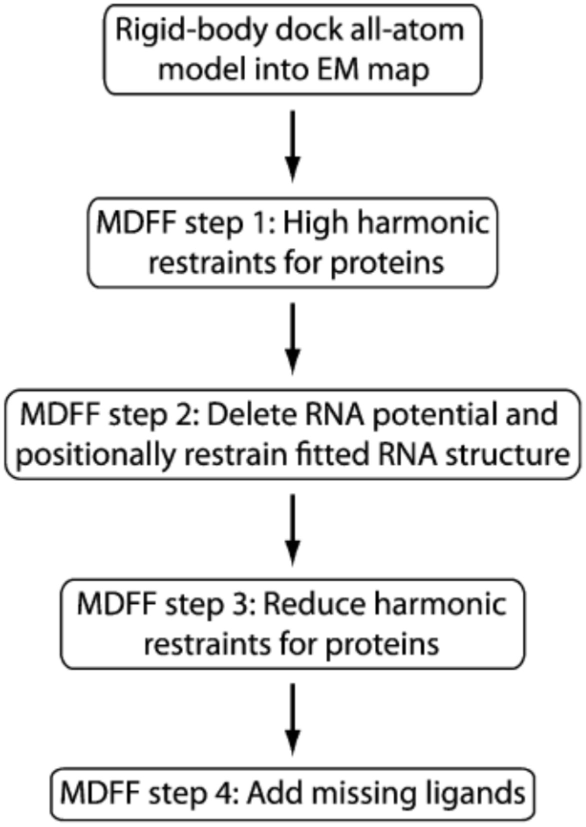
$$U_{\mathrm{EM}}(\mathbf{R}) = \sum_j w_j V_{\mathrm{EM}}(\mathbf{r}_j),$$

$$V_{\mathrm{EM}}(\mathbf{r}) = \begin{cases} \xi \left[1 - \frac{\Phi(\mathbf{r}) - \Phi_{thr}}{\Phi_{max} - \Phi_{thr}} \right] & \text{if } \Phi(\mathbf{r}) \geqslant \Phi_{thr}, \\ \xi & \text{if } \Phi(\mathbf{r}) < \Phi_{thr}. \end{cases}$$

$$\mathbf{f}_i^{\mathrm{EM}} = -\frac{\partial}{\partial \mathbf{r}_i} U_{\mathrm{EM}}(\mathbf{R}) = -w_i \frac{\partial}{\partial \mathbf{r}_i} V_{\mathrm{EM}}(\mathbf{r}_i).$$



Protocol to refine a 6.8-A EM map of the ribosome:



Rigid-body dock all-atom model into EM map

MDFF step 1: High harmonic restraints for proteins

MDFF step 2: Delete RNA potential and positionally restrain fitted RNA structure

MDFF step 3: Reduce harmonic restraints for proteins

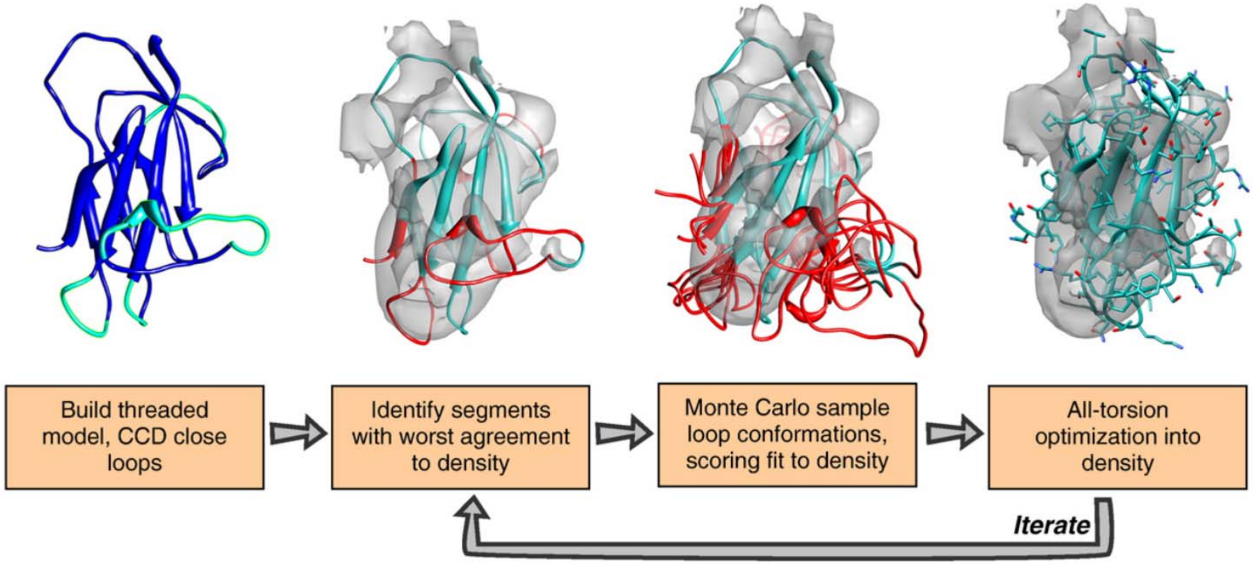
MDFF step 4: Add missing ligands

MDFF: Flexible Fitting of Known Structures

https://youtu.be/_hysNlxDkXw

Rosetta with Low-Resolution Constrains

Rosetta – comparative modeling (EM density at 4-6 A resolution):



Build threaded model, CCD close loops

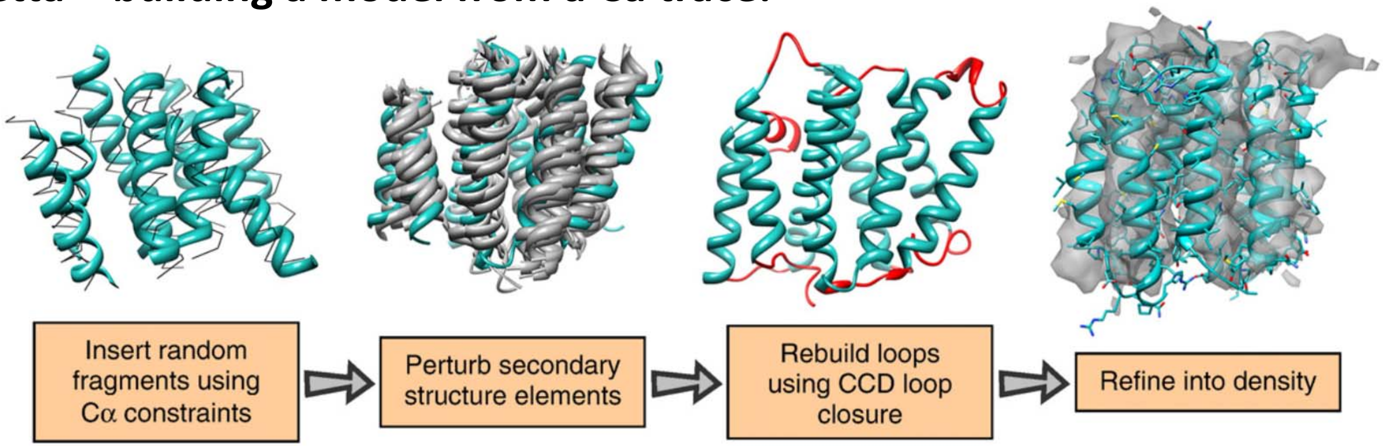
Identify segments with worst agreement to density

Monte Carlo sample loop conformations, scoring fit to density

All-torsion optimization into density

Iterate

Rosetta - building a model from a Ca trace:



Insert random fragments using Cα constraints

Perturb secondary structure elements

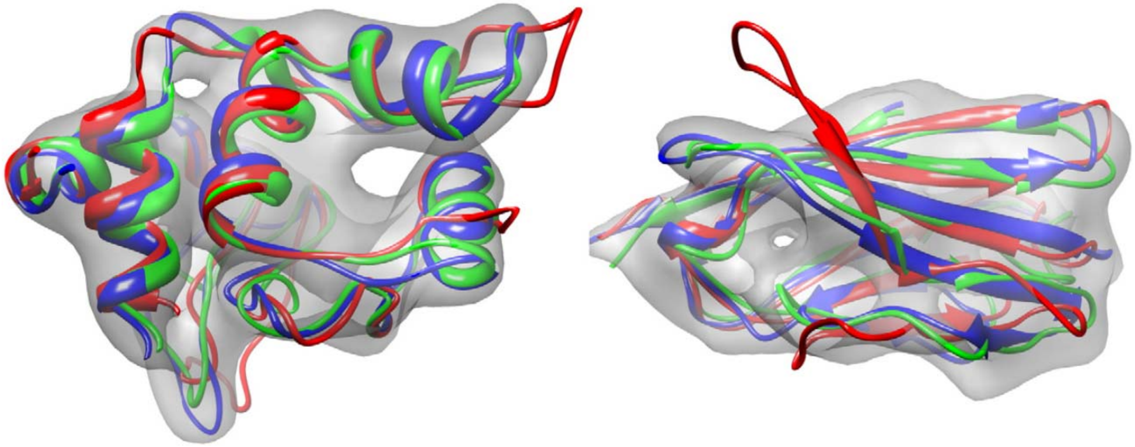
Rebuild loops using CCD loop closure

Refine into density

DiMaio, F. et al. (2009) J.Mol.Biol., 392, 181.

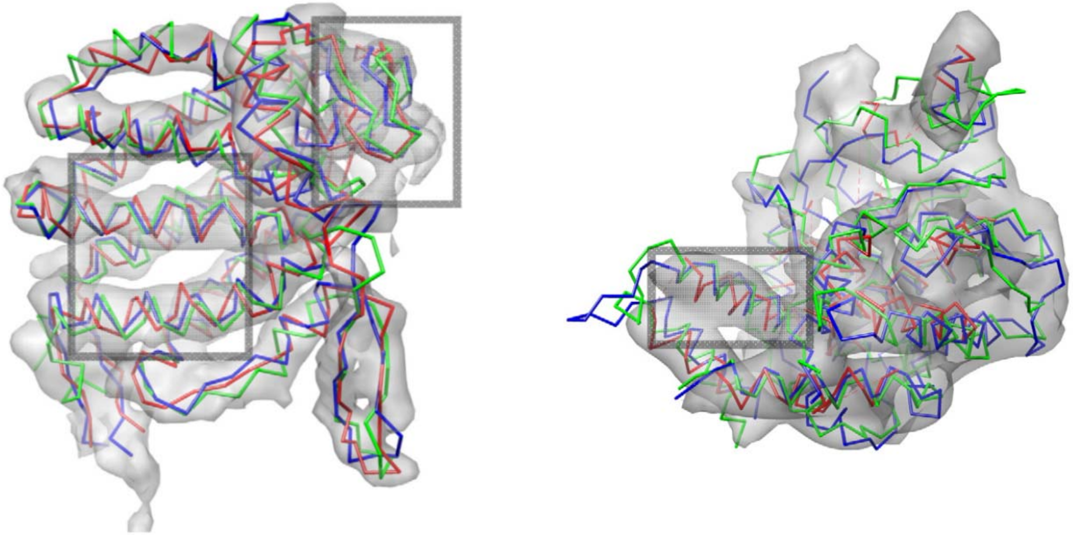
Rosetta with Low-Resolution Constrains

EM density maps at 10 A resolution:

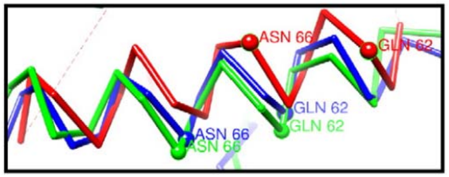


Homology model
Crystal structure
Rosetta model

EM density maps at 4-6 A resolution:



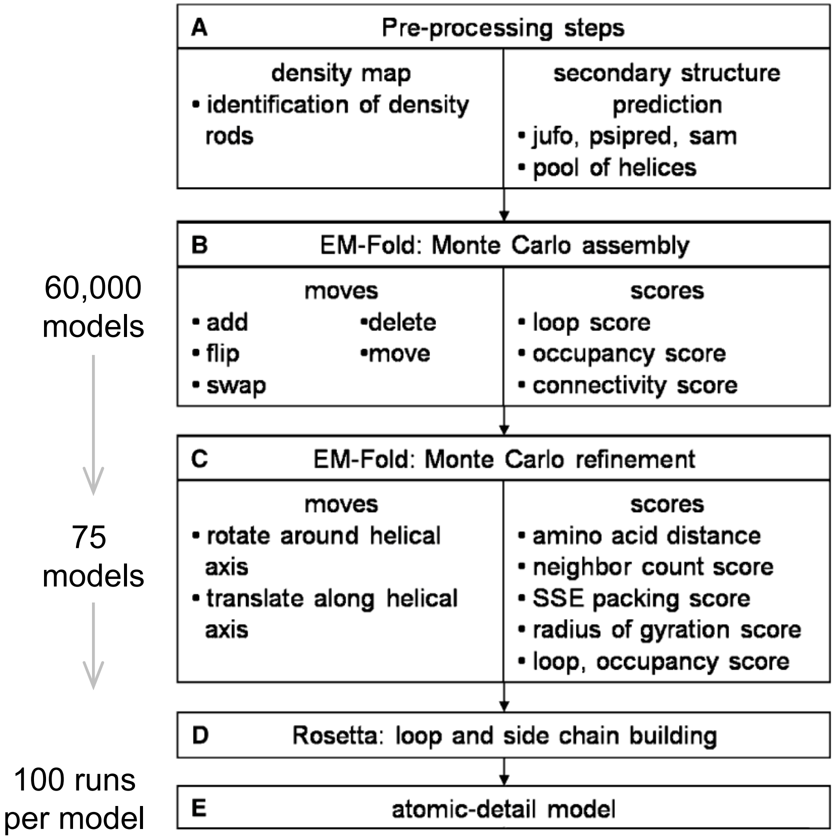
Hand-made model
Crystal structure
Rosetta model



DiMaio, F. et al. (2009) J.Mol.Biol., 392, 181.

EM-Fold: Refinement guided by EM map

Protocol (EM map at 5-7A resolution):



A Pre-processing steps

density map
- identification of density rods

secondary structure prediction
- jufo, psipred, sam
- pool of helices

B EM-Fold: Monte Carlo assembly

moves
- add
- delete
- flip
- move
- swap

scores
- loop score
- occupancy score
- connectivity score

C EM-Fold: Monte Carlo refinement

moves
- rotate around helical axis
- translate along helical axis

scores
- amino acid distance
- neighbor count score
- SSE packing score
- radius of gyration score
- loop, occupancy score

D Rosetta: loop and side chain building

E atomic-detail model

60,000 models

75 models

100 runs per model

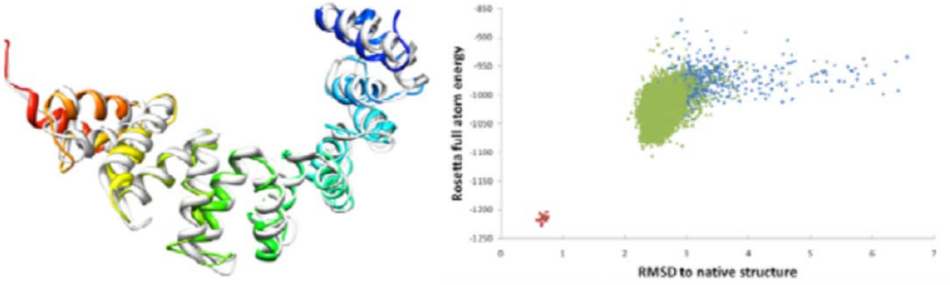
Energy Function terms :

- radius of gyration => increase compactness
- distance between AA pairs => good distance of side chains
- solvation of individual AA => reasonable solvent exposure
- loop distance => proper closure of loops
- pairing of β-strands => proper folding of β-sheets
- packing of secondary structure elements

- connectivity => reasonable placement of SSE
- occupancy => good correspondence with the cryoEM map

Benchmarking using PDB structures with about 300 AA :

- good prediction for 7 of 10 selected proteins (rmsd < 4 Å)
- accuracy is sensitive to the correct prediction of SSE



Example: Final refinement of the helicobacter cysteine-rich protein C

Lindert, S. et al. (2009) Structure, 17, 990.

EM-Fold: Refinement guided by EM map

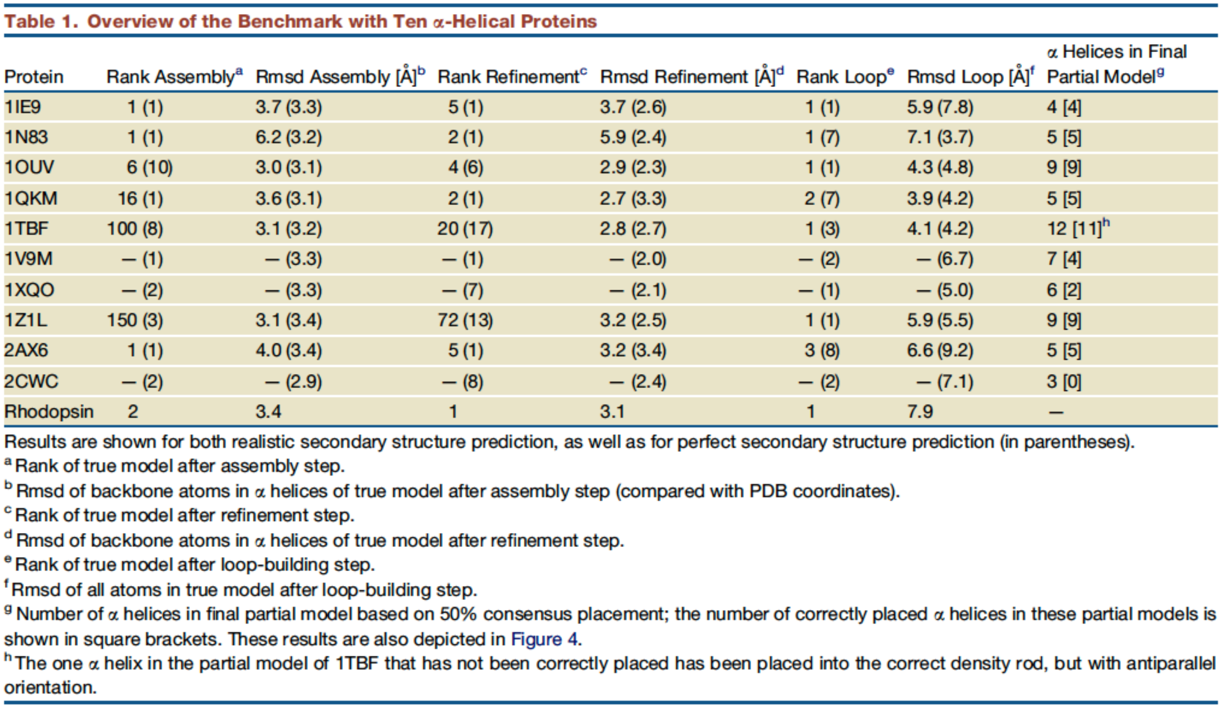


Table 1. Overview of the Benchmark with Ten α-Helical Proteins

Protein	Rank Assembly[a]	Rmsd Assembly [Å][b]	Rank Refinement[c]	Rmsd Refinement [Å][d]	Rank Loop[e]	Rmsd Loop [Å][f]	α Helices in Final Partial Model[g]
1IE9	1 (1)	3.7 (3.3)	5 (1)	3.7 (2.6)	1 (1)	5.9 (7.8)	4 [4]
1N83	1 (1)	6.2 (3.2)	2 (1)	5.9 (2.4)	1 (7)	7.1 (3.7)	5 [5]
1OUV	6 (10)	3.0 (3.1)	4 (6)	2.9 (2.3)	1 (1)	4.3 (4.8)	9 [9]
1QKM	16 (1)	3.6 (3.1)	2 (1)	2.7 (3.3)	2 (7)	3.9 (4.2)	5 [5]
1TBF	100 (8)	3.1 (3.2)	20 (17)	2.8 (2.7)	1 (3)	4.1 (4.2)	12 [11][h]
1V9M	– (1)	– (3.3)	– (1)	– (2.0)	– (2)	– (6.7)	7 [4]
1XQO	– (2)	– (3.3)	– (7)	– (2.1)	– (1)	– (5.0)	6 [2]
1Z1L	150 (3)	3.1 (3.4)	72 (13)	3.2 (2.5)	1 (1)	5.9 (5.5)	9 [9]
2AX6	1 (1)	4.0 (3.4)	5 (1)	3.2 (3.4)	3 (8)	6.6 (9.2)	5 [5]
2CWC	– (2)	– (2.9)	– (8)	– (2.4)	– (2)	– (7.1)	3 [0]
Rhodopsin	2	3.4	1	3.1	1	7.9	–

Results are shown for both realistic secondary structure prediction, as well as for perfect secondary structure prediction (in parentheses).

[a] Rank of true model after assembly step.

[b] Rmsd of backbone atoms in α helices of true model after assembly step (compared with PDB coordinates).

[c] Rank of true model after refinement step.

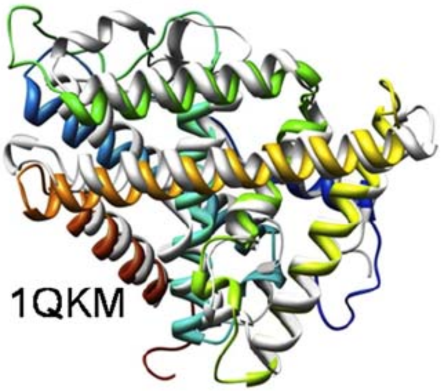
[d] Rmsd of backbone atoms in α helices of true model after refinement step.

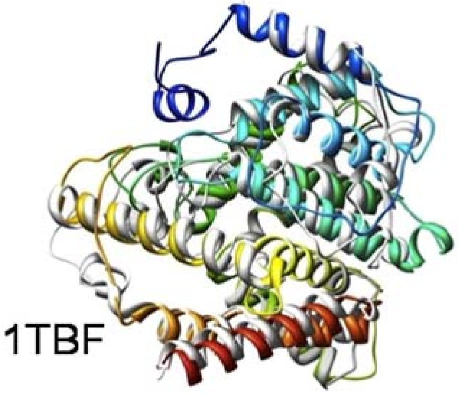
[e] Rank of true model after loop-building step.

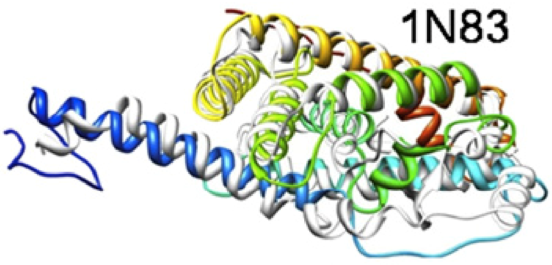
[f] Rmsd of all atoms in true model after loop-building step.

[g] Number of α helices in final partial model based on 50% consensus placement; the number of correctly placed α helices in these partial models is shown in square brackets. These results are also depicted in Figure 4.

[h] The one α helix in the partial model of 1TBF that has not been correctly placed has been placed into the correct density rod, but with antiparallel orientation.



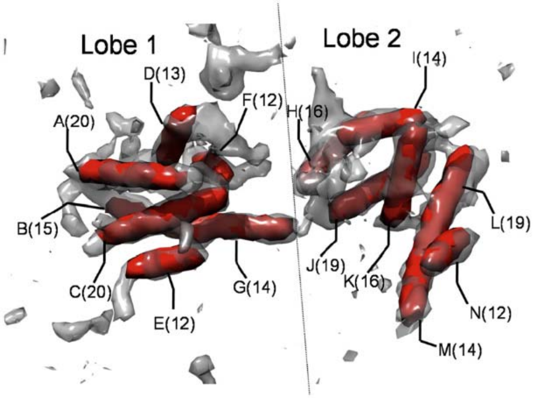




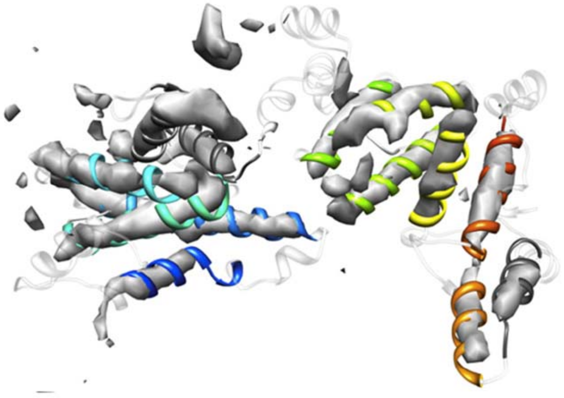
Lindert, S. et al. (2009) Structure, 17, 990.

EM-Fold: Refinement guided by EM map

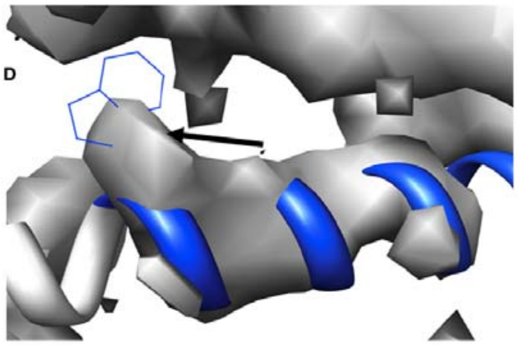
Application to the adenovirus protein IIIa



6.8-Å map of the N-term. of protein IIIa
- 400 AA, predicted 68% α-helical
- identified 14 rod-like densities



A partial model after EM-fold analysis
- 11 confidently placed α-helices
- 3 α-helices and loops are ambiguous

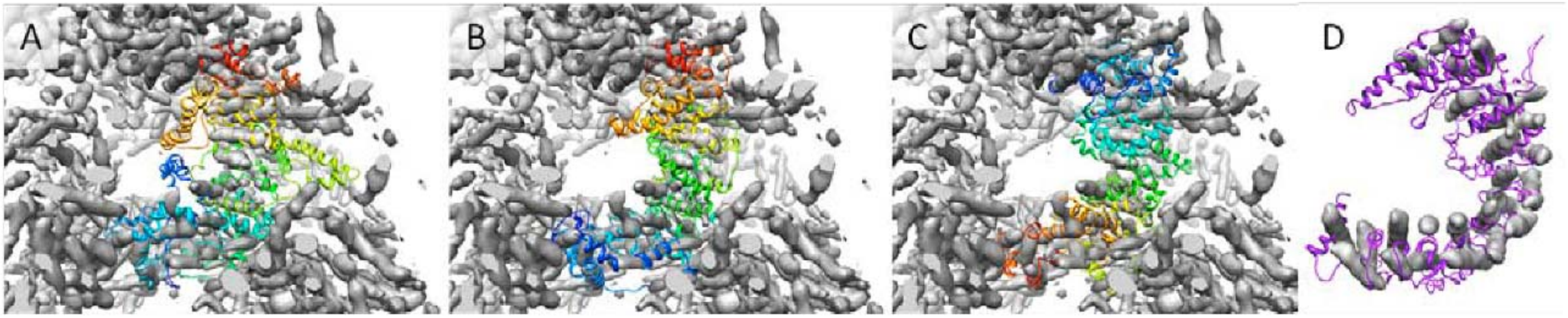


Validation of the model
- density bump at the location of Trp27
- match of Tyr in other two helices

Lindert, S. et al. (2009) Structure, 17, 990.

Application to a domain of DNA-PK catalytic subunit (4128 AA, 135 helices)

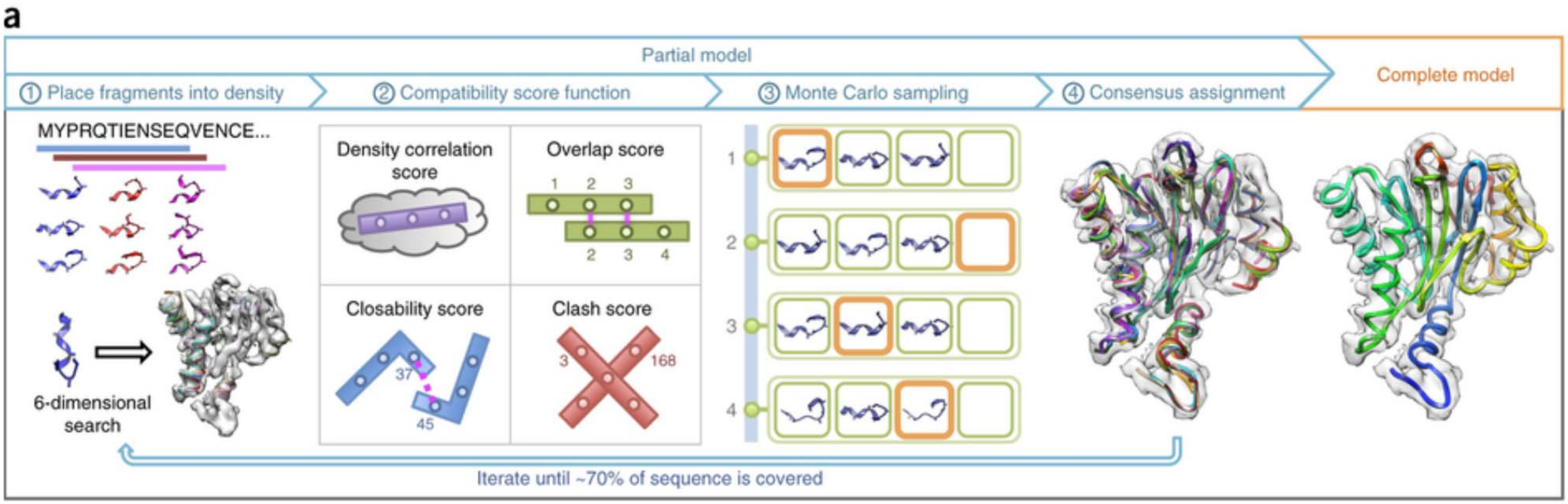
- EM-fold applied only to the heat repeat motive with 25 density rods



model #1 model #2 model #3 homolog

Lindert, S. et al. (2011) Microsc. Microanal. 17 (Supp 2)

De Novo Model Building

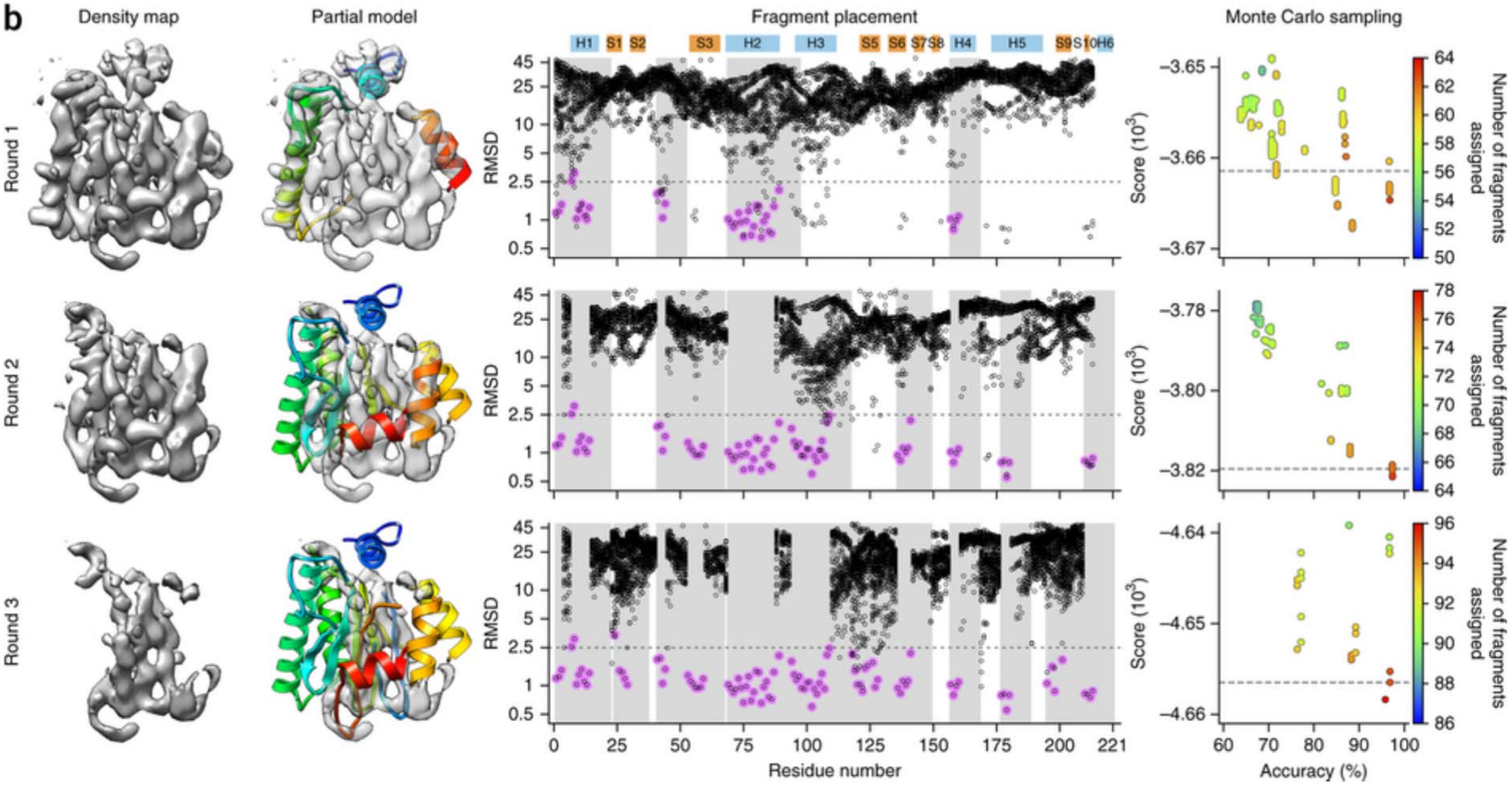


1. Matching fragments into EM density
2. Evaluating sets of compatible fragmets (score$_{\text{total}}$)
3. Simulated annealing with MC sampling
4. Iterative assembly of models
5. Completing models with RossetaCM
6. Model building with Buccaneer

$$\begin{aligned}\text{score}_{\text{total}}(\boldsymbol{F}) = w_{\text{dens}} &\sum_{f_i \in \boldsymbol{F}} \text{score}_{\text{dens}}(f_i)\\ &+ w_{\text{overlap}} \sum_{f_i, f_j \in \boldsymbol{F}} \text{score}_{\text{overlap}}(f_i, f_j)\\ &+ w_{\text{close}} \sum_{f_i, f_j \in \boldsymbol{F}} \text{score}_{\text{close}}(f_i, f_j)\\ &+ w_{\text{clash}} \sum_{f_i, f_j \in \boldsymbol{F}} \text{score}_{\text{clash}}(f_i, f_j)\end{aligned}$$

Wang, R. et al. (2015) Nature Methods, 12, 335

De Novo Model Building



Wang, R. et al. (2015) Nature Methods, 12, 335