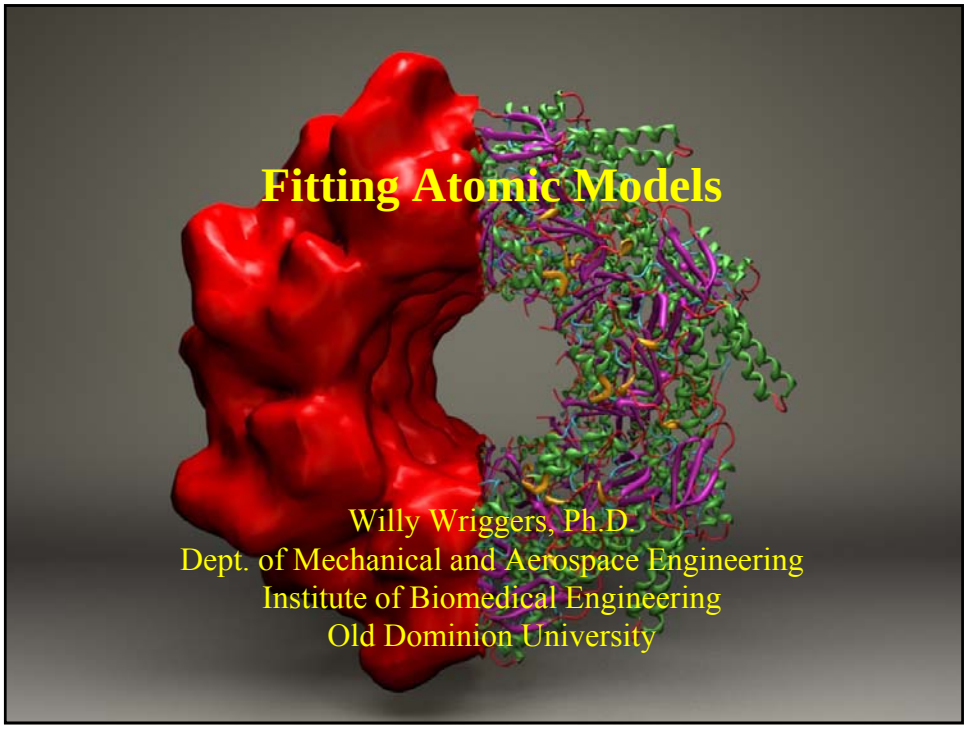
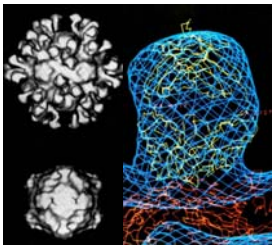
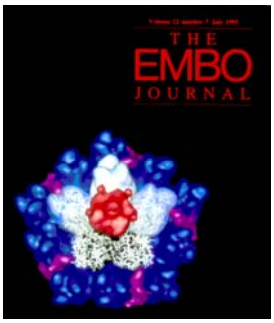

Fitting Atomic Models
Willy Wriggers, Ph.D.
Dept. of Mechanical and Aerospace Engineering
Institute of Biomedical Engineering
Old Dominion University


NEW YORK
PHILADELPHIA
BALTIMORE
WASHINGTON
RICHMOND
WILLIAMSBURG
RALEIGH
Norfolk / Virginia Beach
Virginia Beach
Jefferson Lab
NASA
Norfolk Navy Yard
ODU
Norfolk

24 Years of Fitting Atomic Models



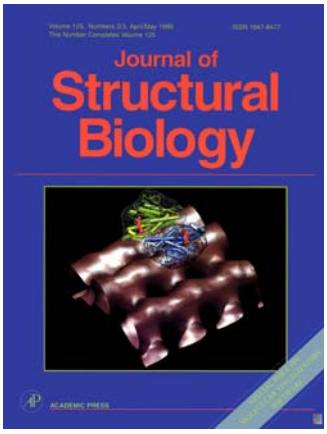


Guoji Wang, Claudine Porta, Zhongguo Chen, Timothy S. Baker, John E. Johnson:
Identification of a Fab interaction footprint site on an icosahedral virus by cryoelectron microscopy and X-ray crystallography.
Nature, 355:275, 1992.

Phoebe L. Stewart, Stephen D. Fuller, Roger M. Burnett:
Difference imaging of adenovirus: bridging the resolution gap between X-ray crystallography and electron microscopy.
EMBO J., 12:2589, 1993.

"At that time, placing an atomic structure into an EM map seemed like a very dangerous idea…"
Phoebe Stewart, 2003

1999: The First Algorithmic Packages

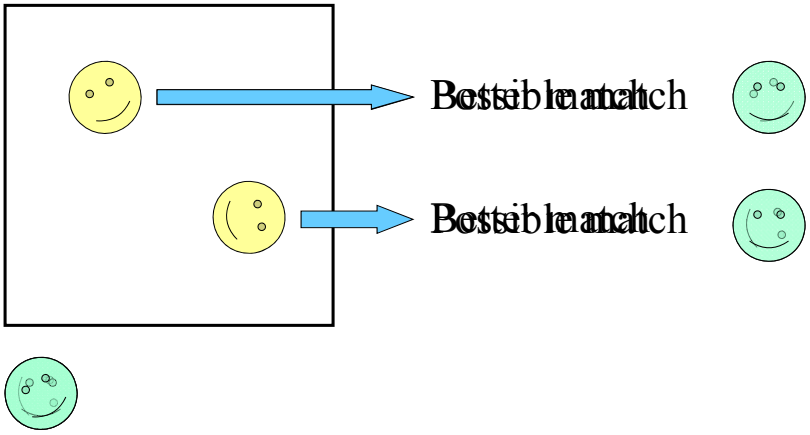


Willy Wriggers, Ronald A. Milligan, and J. Andrew McCammon:
Situs: A Package for Docking Crystal Structures into Low-Resolution Maps from Electron Microscopy.
J. Structural Biology, 125:185, 1999

Niels Volkmann and Dorit Hanein:
Quantitative Fitting of Atomic Models into Observed Densities Derived by Electron Microscopy.
J. Structural Biology, 125:176, 1999

Today:
Dozens of packages available, e.g. Situs, Sculptor, COAN, DockEM, EMFit, DireX, etc… see http://en.wikibooks.org/wiki/Software_Tools_For_Molecular_Microscopy

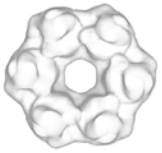
2002: Template "Convolution"



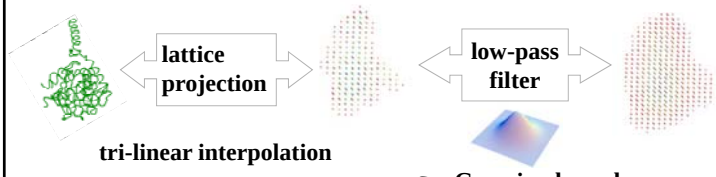
Scoring function: cross-correlation

Template "Convolution"

$\rho_{\mathrm{em}}(\mathbf{r})$ target density on lattice



$\rho_{\mathrm{atomic}}(\mathbf{rR})$ rotated probe molecule density projected to the lattice:

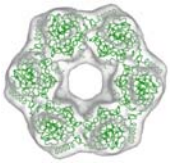


$g \otimes \rho_{\mathrm{atomic}}(\mathbf{rR}) \equiv \rho_{\mathrm{calc}}(\mathbf{r})$

tri-linear interpolation

***g* = Gaussian kernel**

$$C(\mathrm{T}) = \int \rho_{\mathrm{em}}(\mathrm{r}) \cdot \rho_{\mathrm{calc}}(\mathrm{r}+\mathrm{T})\, d^3\mathrm{r}$$



Fitting criterion: e.g. linear cross-correlation,
evaluate for every rotation **R** and translation **T**

Computational Cost

- Three translational degrees of freedom
  - N possible locations
- Three rotational degrees of freedom
  - M possible orientations
- Cost for each cross-correlation calculation
  - N (number of voxels)

→ Total cost: N * M * N = M * N^2

FTM (Fast Translational Matching)

The expression for the cross-correlation is

$$C(\mathrm{T}) = \int \rho_{\mathrm{em}}(\mathrm{r}) \cdot \rho_{\mathrm{calc}}(\mathrm{r}+\mathrm{T})\, d^3\mathrm{r}$$

Using the Fourier Convolution Theorem, we get

$$C(\mathrm{T}) = \mathrm{F}^{-1}\left[\ \mathrm{F}(\rho_{em})^{*} \cdot \mathrm{F}(\rho_{calc})\ \right]$$

Needs to be calculated for every orientation

Needs to be calculated only ONCE

This yields ALL possible translations in one step!

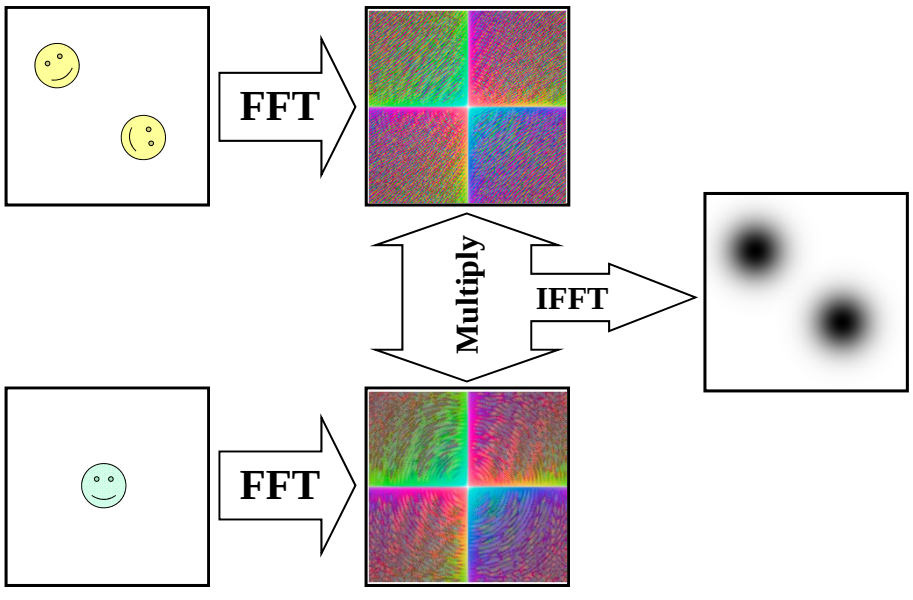
Computational Cost

- Three translational degrees of freedom
  - ~~N possible locations~~
- Cost for each cross-correlation calculation
  - ~~N^3~~
- Three rotational degrees of freedom
  - M possible orientations

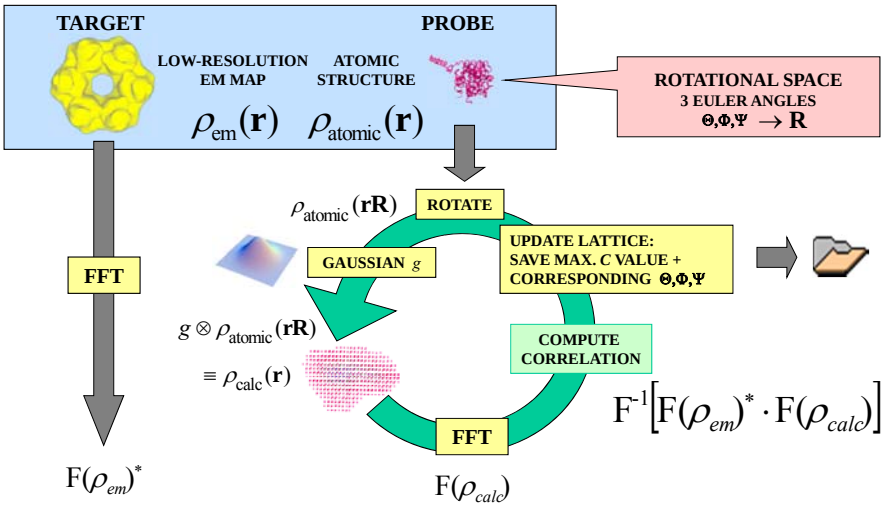
→ Total cost: M * 2 N log(N)

For a 50^3 map this results in a speedup of 4 orders of magnitude!

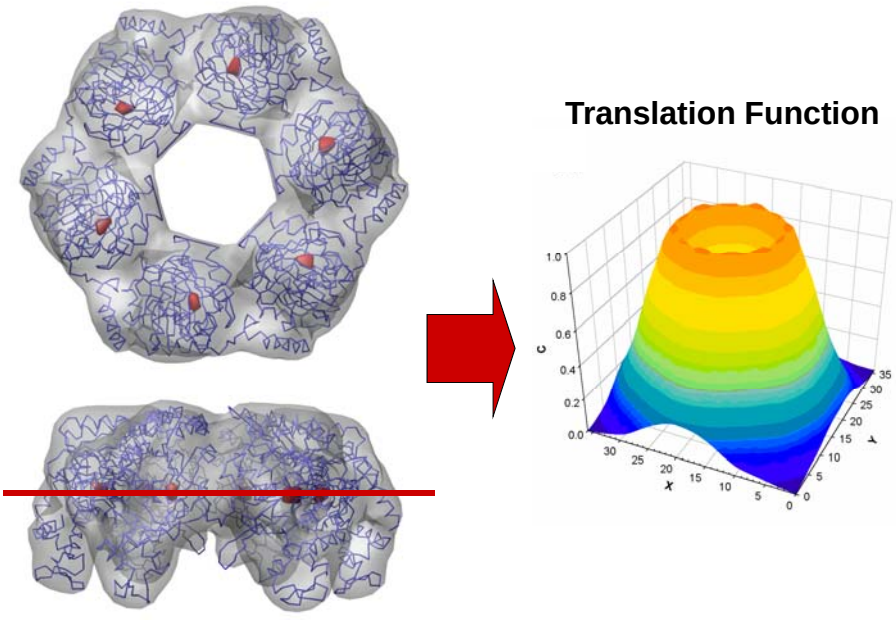
FTM: An Example



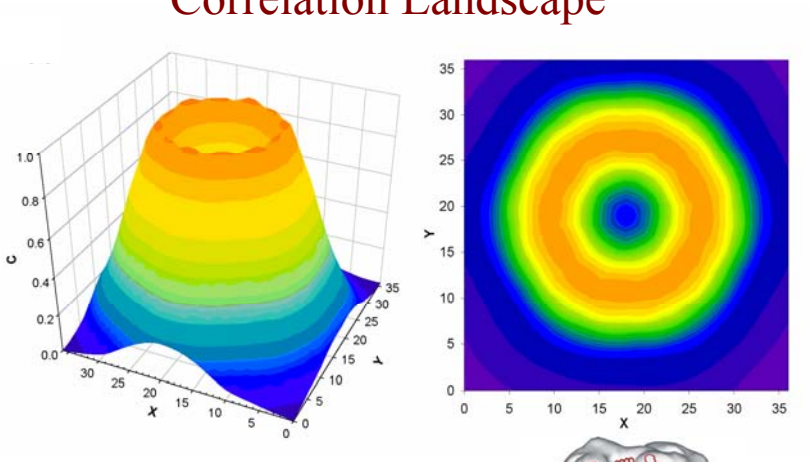
6D Search with FTM

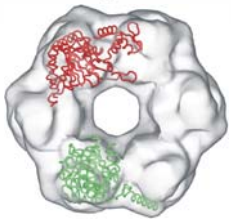


Correlation Landscape



Correlation Landscape





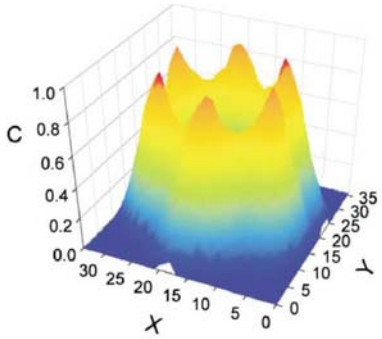
For resolutions below 10Å interior detail is lost and we cannot distinguish between correct and spurious fits

Local Correlation: Density Masking

Renormalize (mask) the correlation locally:

$$C(\mathbf{T}) = \frac{\int_{mask} \rho_{em}(\mathbf{r}) \times \rho_{calc}(\mathbf{r}+\mathbf{T})\ d^3r}{\sqrt{\int_{mask} \rho_{em}^2(\mathbf{r}) d^3r} \sqrt{\int_{mask} \rho_{calc}^2(\mathbf{r})\ d^3r}}$$

$$\text{mask} \rightarrow \rho_{\text{calc}_{l,m,n}} > 0$$



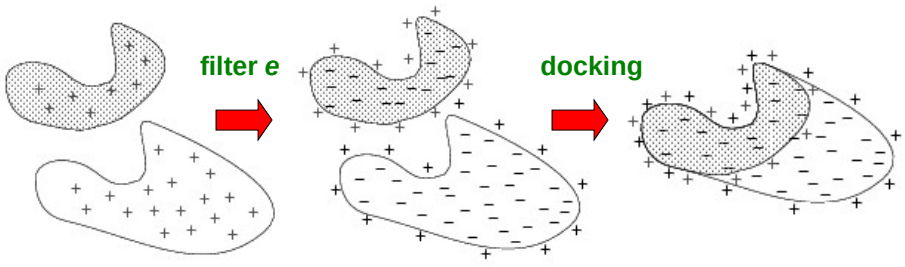
- Extends the reliability of correlation based docking (<15Å)
- Requires approximations to be FFT accelerated

DOCKEM, A.M. Roseman

Density Filtering

Adding surface/contour information

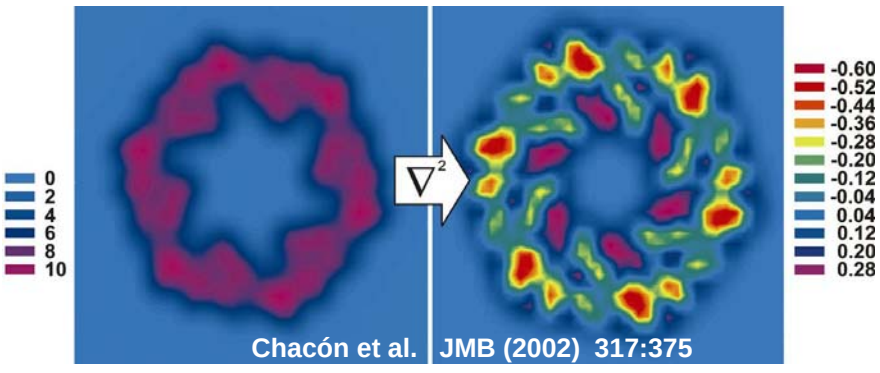
A suitable filter would assign negative values to the interior, positive values to the molecular contour. Both volume and contour matches would provide positive contributions to the correlation criterion:



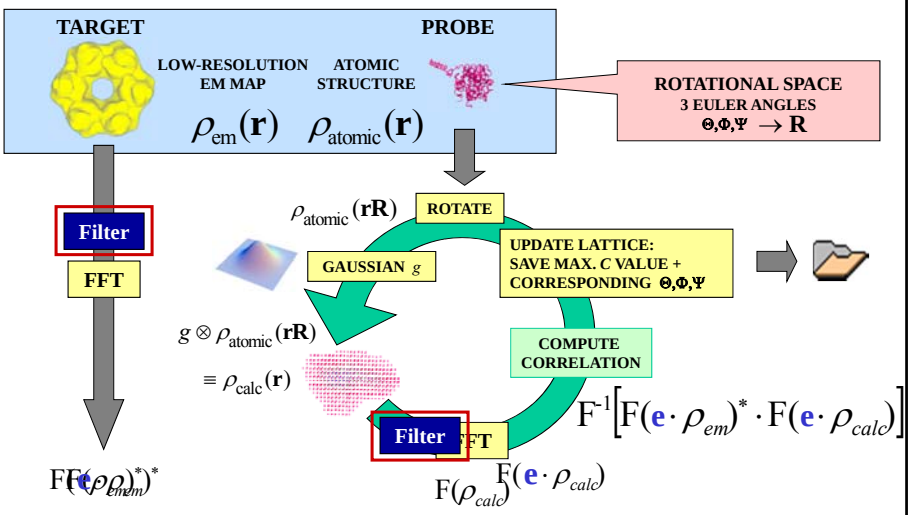
Contour Filter

Laplacian

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

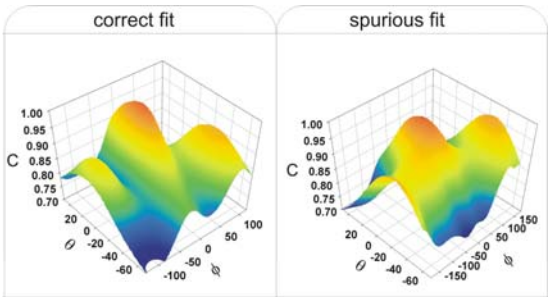


6D Search with FTM and Filtering

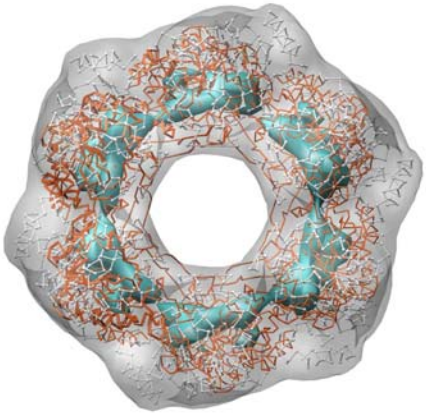


Effect of Filter on Orientation

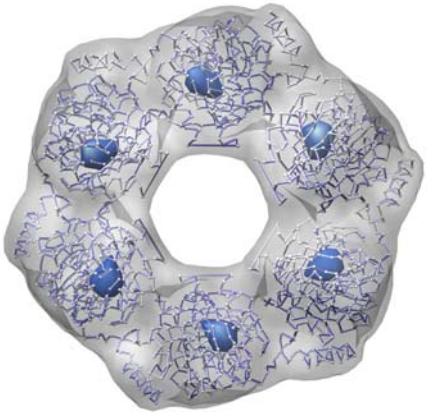
No Filter



Example: RecA Translation Function



Standard cross-correlation

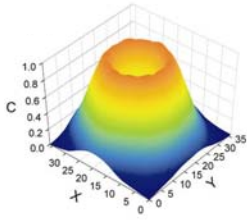


with Laplacian filtering (*colores*)

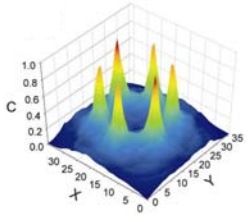
Example: RecA

Grid size 6Å
Resolution 15Å
9˚ steps (30481 rotations)

standard cross-correlation



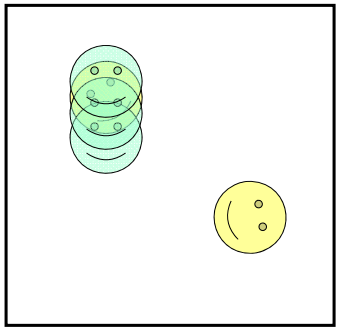
w/ Laplacian filtering (*colores*)



Only Laplacian filtering successfully restores the initial position

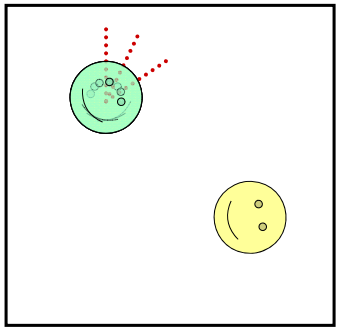
Search Granularity

Translational Granularity



Originates from voxel spacing

Rotational Granularity



Originates from angular sampling

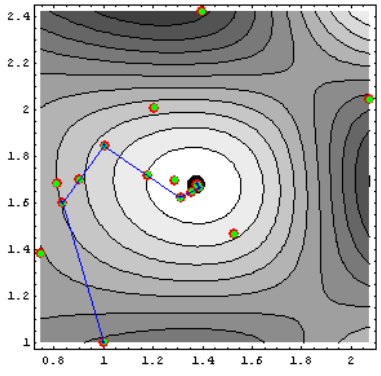
Off-Lattice Refinement

The exhaustive search is limited to a grid of points in the 6D search space

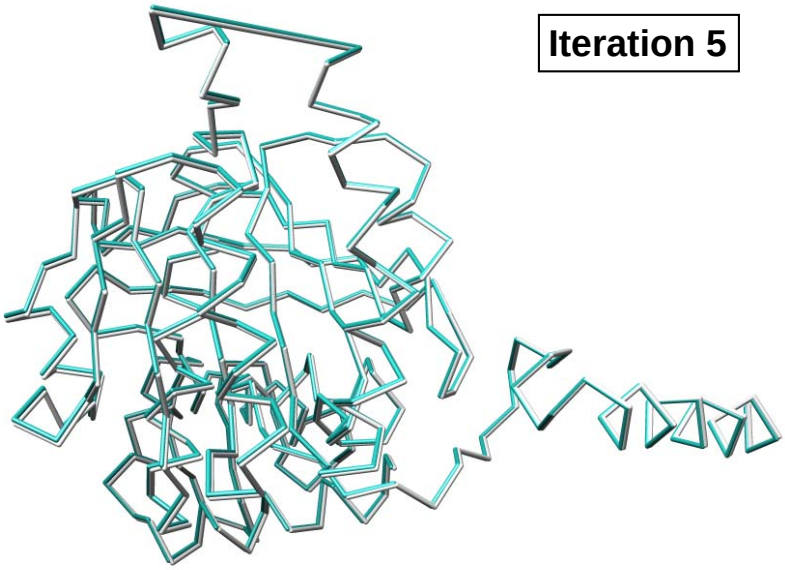
Improve the accuracy

Off-lattice (6D) local maximization of the correlation coefficient

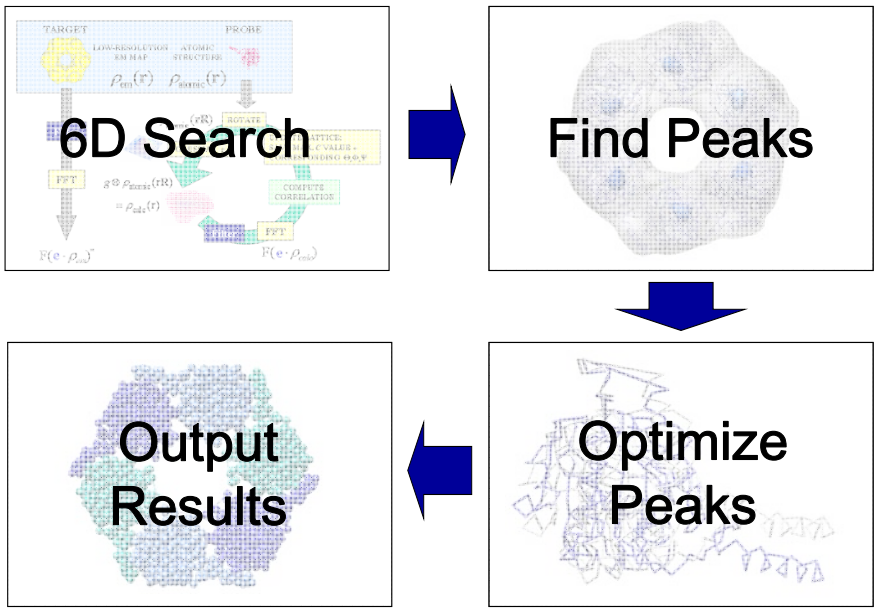
- Powell conjugent gradient method
- included in *colores* by default
- stand-alone tool: *collage*



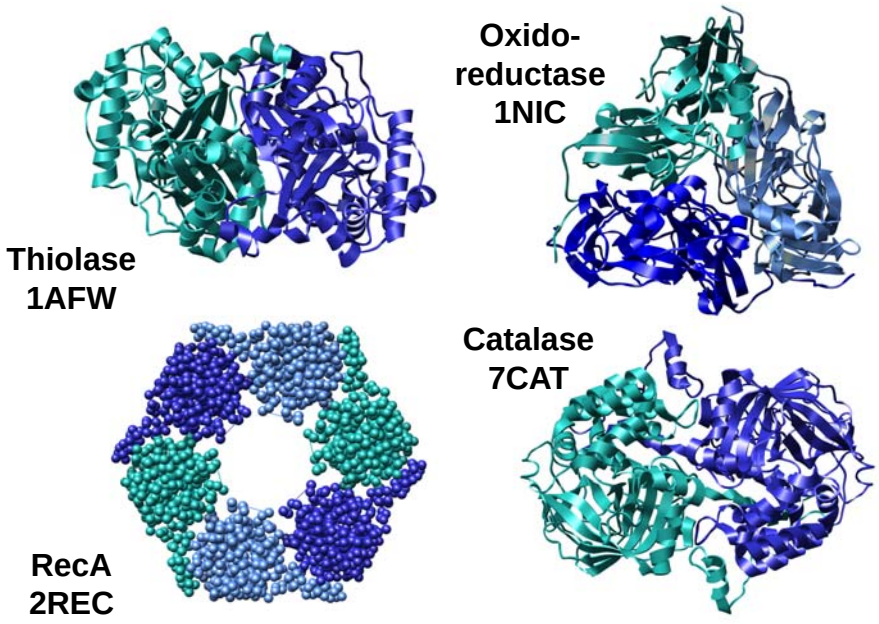
Off-Lattice Refinement: Example



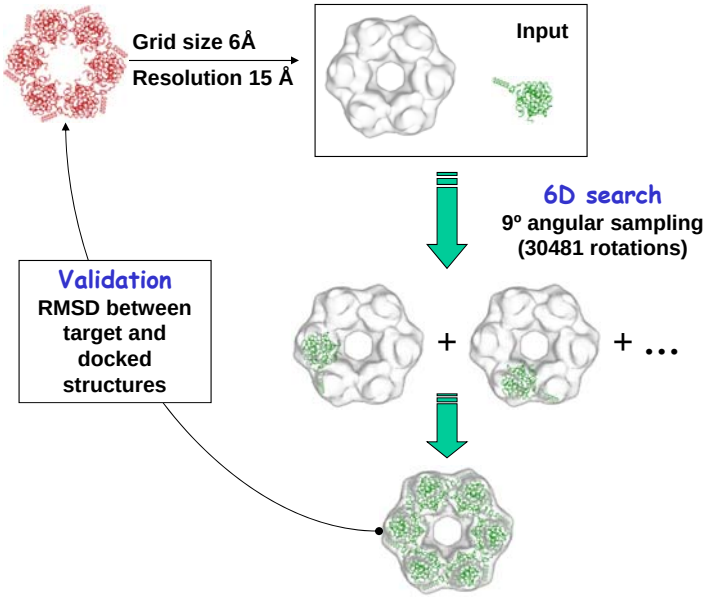
Complete FTM Workflow in *Situs*



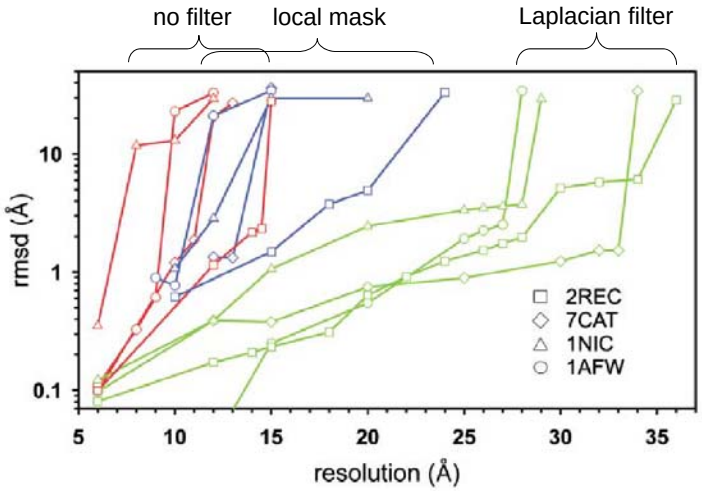
Restoring Various Oligomers



Restoration Tests with Simulated Data



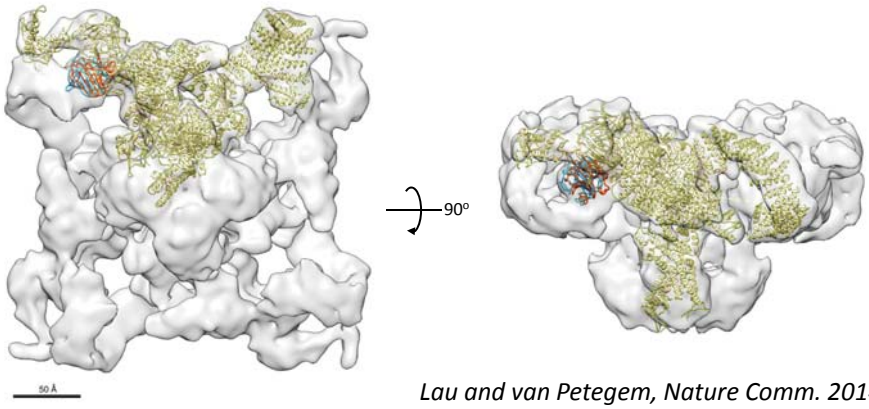
Restoring Various Oligomers



RecA (2REC), thiolase (1AFW), catalase (7CAT), and oxidoreductase (1NIC).

Colores is still Competitive Today

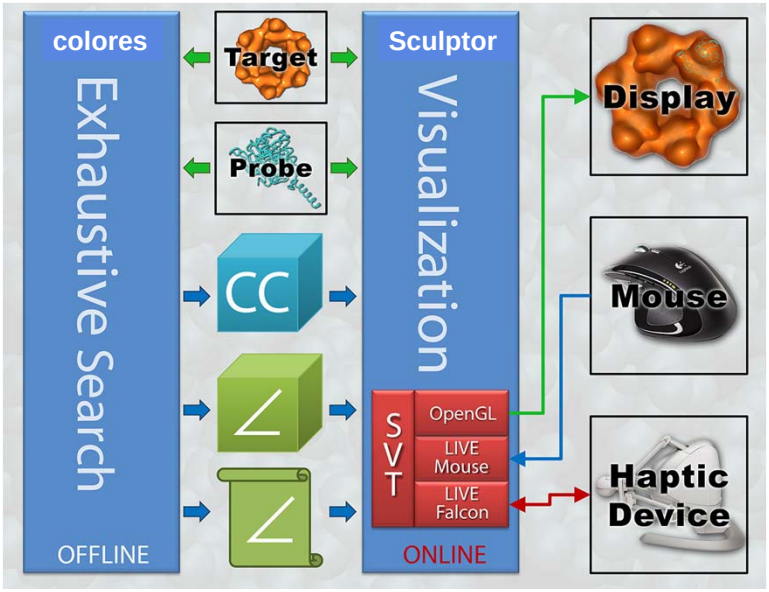
Docking of Models to Maps with Resolution worse than 10 Å



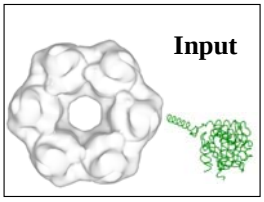
Lau and van Petegem, Nature Comm. 2014

The 50kDa SPRY2 domain can be docked quite precisely into the 10.5 Å map of RYR (EMD5014, EMD1606, EMD1607) using Laplacian filtering by colores. RMSD=2.1 Å vs. a 3.8 Å resolution model solved later.
Wriggers and He, J. Struct. Biol. 192:255 (2015)

Integration with *Sculptor*



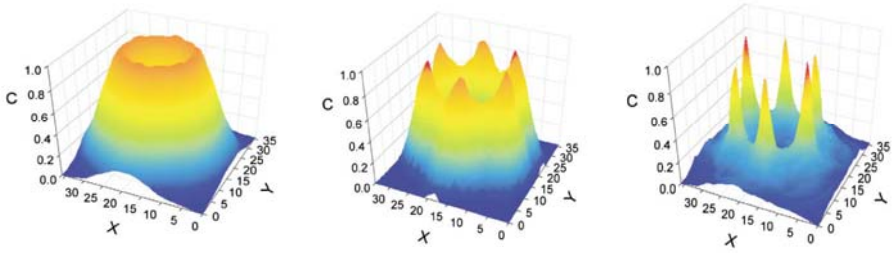
Summary: Correlation Based Matching



6D exhaustive searches:

- Rigid Body
- Fast Translational Matching
- Fast Rotational Matching
- Density Filtering

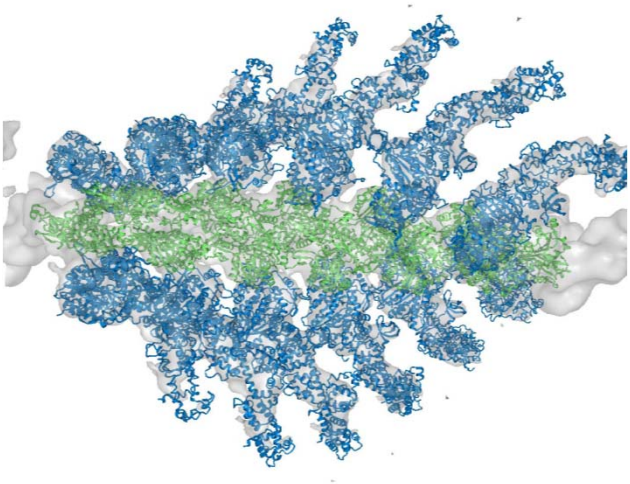
No filter **Local mask** **Laplacian filter**



Increasing Fitting Contrast →

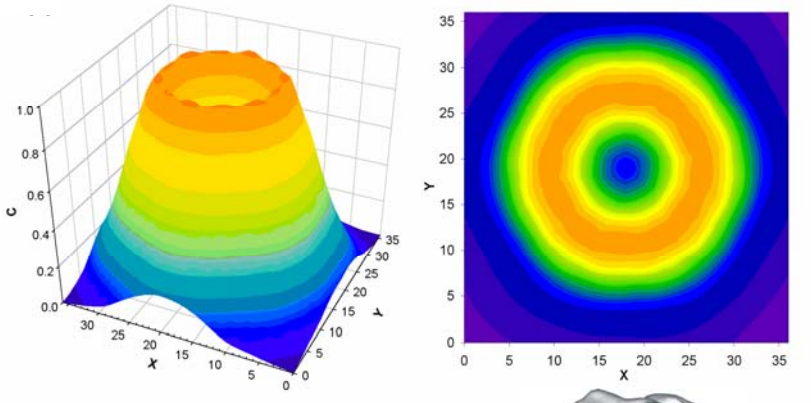
2011: Simultaneous Multi-Fragment Refinement

Actomyosin Example (14Å Resolution)

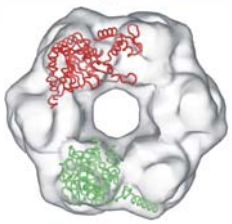


12 G-actin monomers + 12 myosin S1 motors

"One At A Time" Correlation Landscape



For resolutions below 10Å interior detail is lost and we cannot distinguish between correct and spurious fits



Solution Proposed Here: Simultaneous Multi-Fragment Refinement

- Powell conjugent gradient, 6N degrees of freedom
- new stand-alone tool in Situs 2.6: *collage*
- What is new? Fragments see each other (i.e avoid steric clashes) via normalization of cross correlation:

$$C(\mathbf{T}) = \frac{\int \rho_{\text{em}}(\mathbf{r}) \cdot \rho_{\text{calc}}(\mathbf{r}+\mathbf{T})\ \mathrm{d}^3\mathrm{r}}{\sqrt{\int \rho^2_{\text{em}}(\mathbf{r})\mathrm{d}^3\mathrm{r}}\sqrt{\int \rho^2_{\text{calc}}(\mathbf{r})\ \mathrm{d}^3\mathrm{r}}}$$

Birmanns, Rusu & Wriggers, *J. Struct. Biol.*, 173:428, 2011

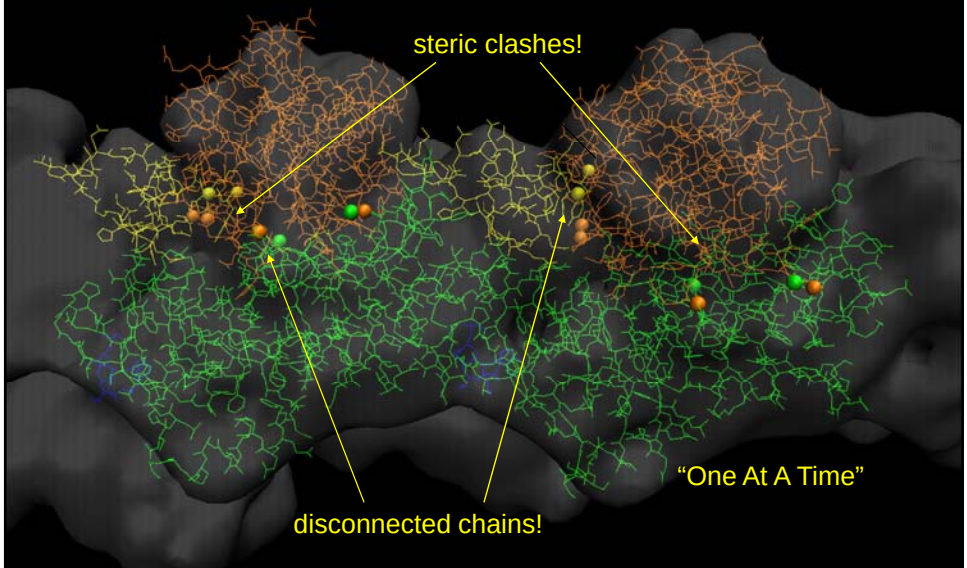
Improved Docking Accuracy

	Actomyosin Complex		GroEL (emd-1080	
Reference Model	F-actin / Actomyosin		PDB Code: 1XCK	
	RMSD (Å)	CC	RMSD (Å)	CC
Reference Model		0.576 / 0.703		0.946
Interactive Peak Search	5.1 / 3.8	0.537 / 0.672	4.3	0.881
Single-Body Refinement	2.6 / 2.1	0.569 / 0.698	1.7	0.945
Multi-Body Refinement	1.8 / 1.4	0.583 / 0.709	1.3	0.950

Table 1: Comparison of various docking approaches for the actomyosin complex and the chaperonin GroEL (subunits include 12 G-actin monomers / 12 myosin S1 for the actomyosin complex, and 14 monomers for GroEL). Root mean square deviation (RMSD) from the references and cross-correlation coefficients (CC) are shown. The CC values of actomyosin correspond to the map shown in Fig. 6 and are systematically lower than those of the GroEL due to filament end effects. The interactive peak search model is equivalent to *colores* before Powell optimization (Chacón and Wriggers, 2002). In the single-body refinement each fragment is fitted independently (equivalent to *colores* after Powell optimization), while in the multi-body refinement all fragments were simultaneously optimized.

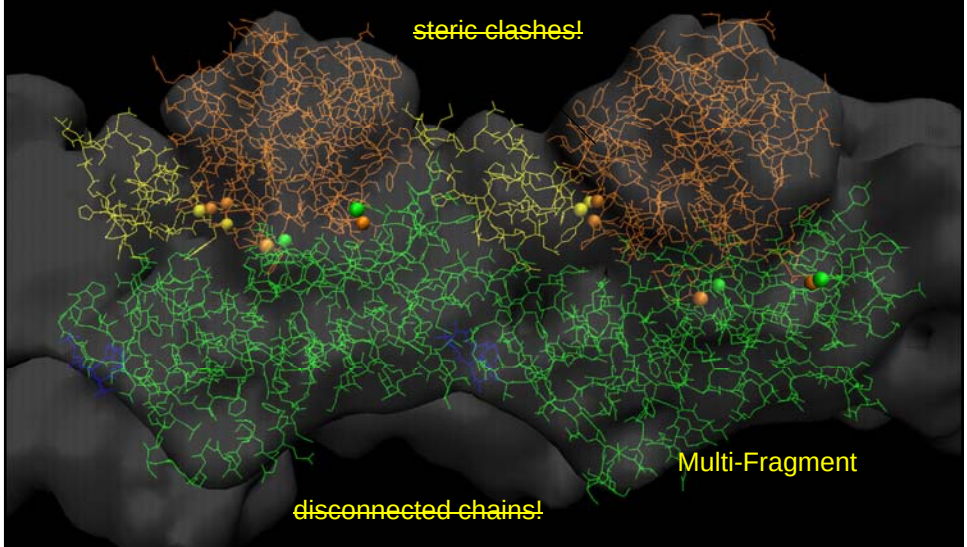
Birmanns, Rusu & Wriggers, *J. Struct. Biol.*, 173:428, 2011

Better Quality Models



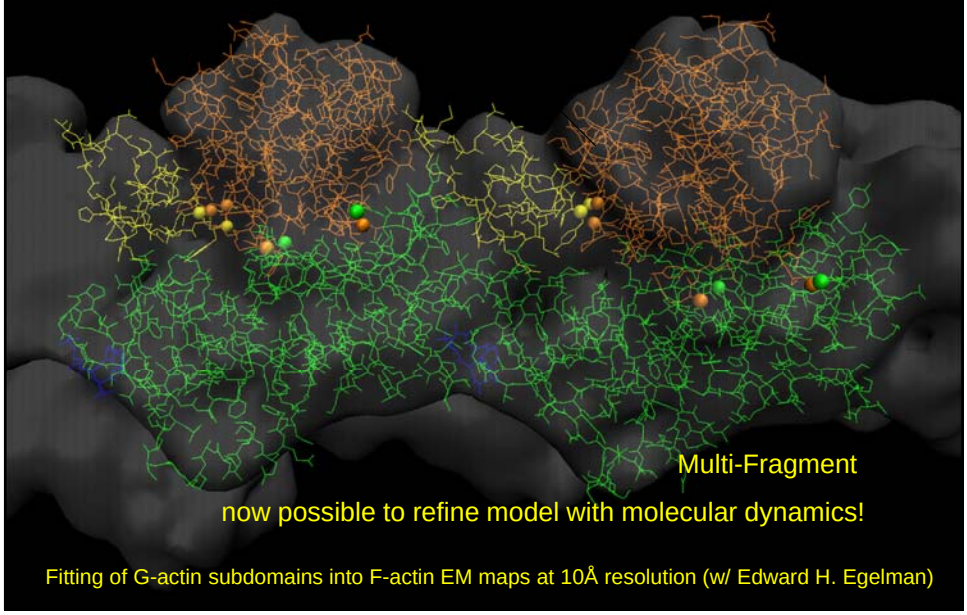
Fitting of G-actin subdomains into F-actin EM maps at 10Å resolution (w/ Edward H. Egelman)

Better Quality Models



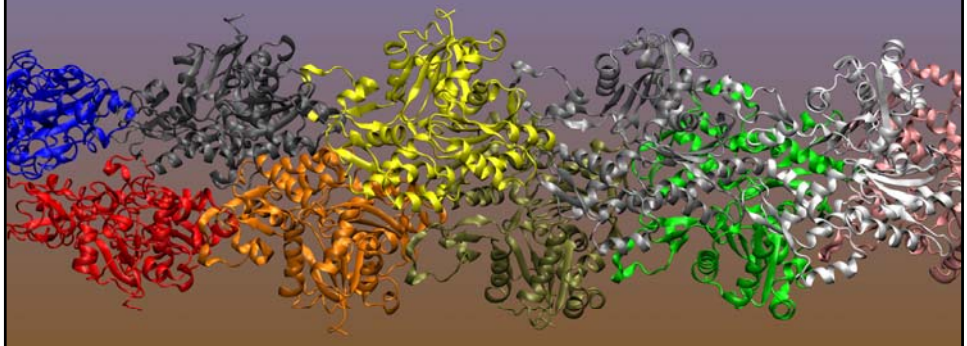
Fitting of G-actin subdomains into F-actin EM maps at 10Å resolution (w/ Edward H. Egelman)

Better Quality Models



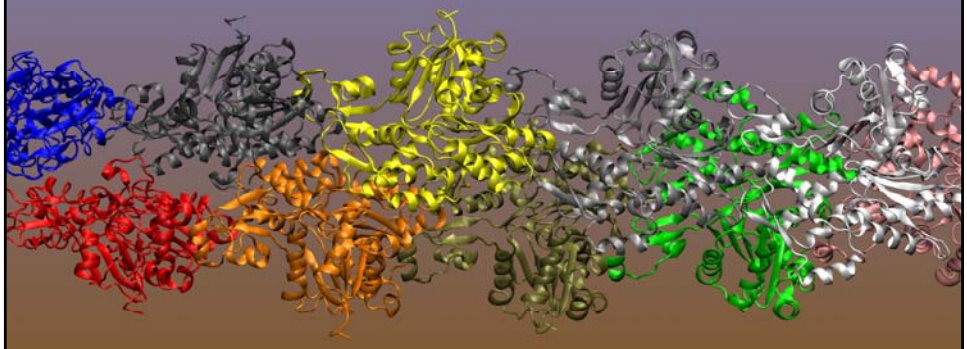
Fitting of G-actin subdomains into F-actin EM maps at 10Å resolution (w/ Edward H. Egelman)

Application: Tracking of F-Actin Variability



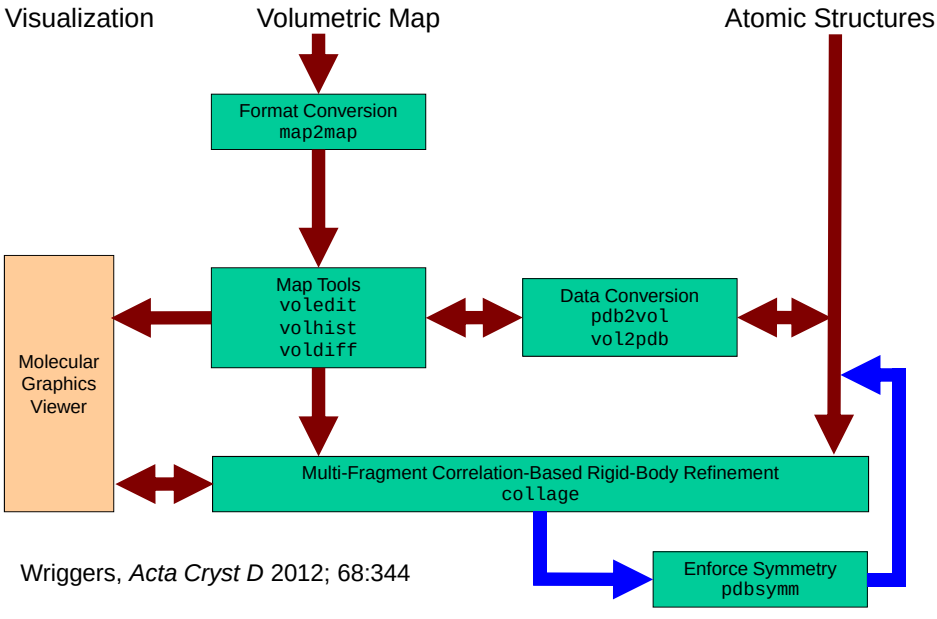
Fitting of G-actin subdomains into F-actin EM maps at 10Å resolution (w/ Edward H. Egelman)

Application: Tracking of F-Actin Variability



Fitting of G-actin subdomains into F-actin EM maps at 10Å resolution (w/ Edward H. Egelman)

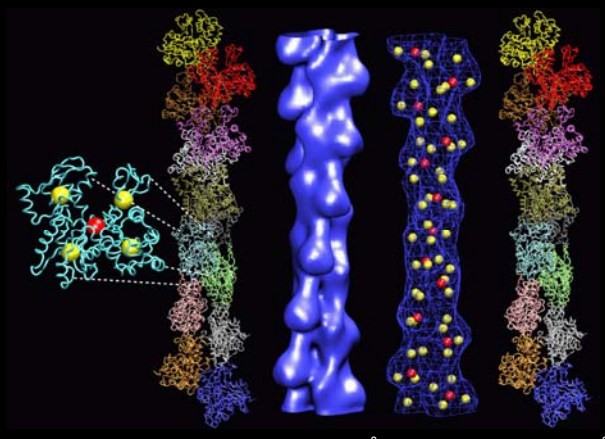
Application of Symmetry



Wriggers, *Acta Cryst D* 2012; 68:344

Break/Questions

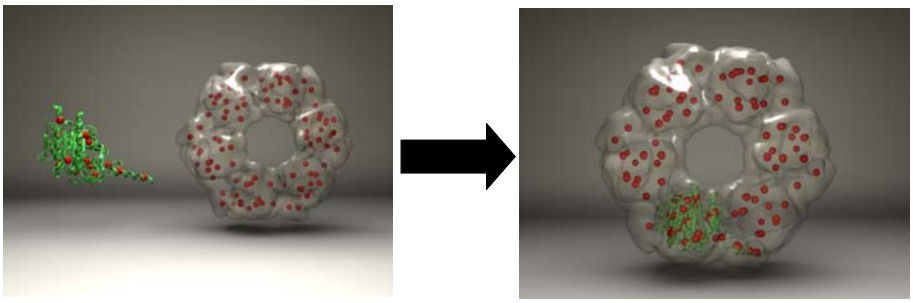
1998: “Simulated Markers”



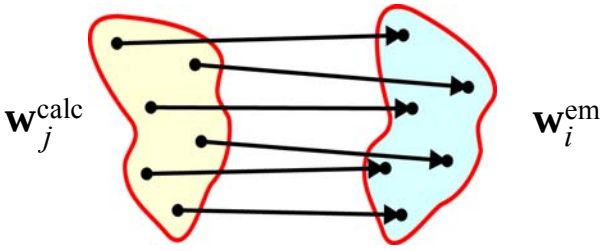
Actin filament: Reconstruction from EM data at 20Å resolution rmsd: 1.1Å

Willy Wriggers, Ronald A. Milligan, Klaus Schulten, and J. Andrew McCammon:
Self-Organizing Neural Networks Bridge the Biomolecular Resolution Gap.
J. Mol. Biol., 284:1247, 1998

1999-2009: Fast "Point Cloud" Fitting



Coarse-Grained Representations of Biomolecular Structure



Feature points (fiducials, landmarks), reduce complexity of search space

Useful for:

- Rigid-body fitting
- Flexible fitting
- Interactive fitting / force feedback
- Building of deformable models

Vector Quantization

Lloyd (1957), Linde, Buzo, & Gray (1980) — Digital Signal Processing, Speech and Image Compression.
Martinetz & Schulten (1993) — Topology-Representing Network.

Encode data (in $\Re^{d=3}$) using a finite set $\{w_j\}$ (j=1,…,k) of *codebook vectors*.

Delaunay triangulation divides $\Re^3$ into k *Voronoi polyhedra* ("receptive fields"):

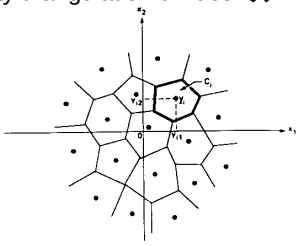
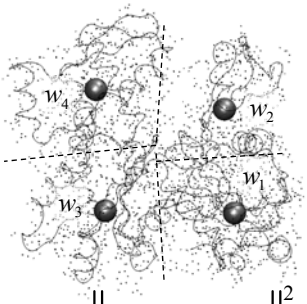


Fig. 3. Partitioning of two-dimensional space (N = 2) into L = 18 cells. All input vectors in cell C_i will be quantized as the code vector y_i. The shapes of the various cells can be very different.



Minimize encoding distortion error: $E = \sum_{i\,(\text{atoms, voxels})} \left\| v_i - w_{j(i)} \right\|^2 m_i$

Convergence and Variability

Q: How do we know that we have found the global minimum of E?

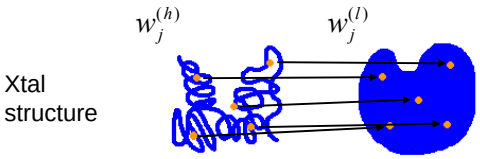
A: We don't (in general).

But we can compute the statistical variability of the $\{w_j\}$ by repeating the calculation with different seeds for random number generator.

Codebook vector variability arises due to:
- statistical uncertainty,
- spread of local minima.

A small variability indicates good convergence behavior.
Optimum choice of # of vectors k: variability is minimal ("quality" of coarse-grained representation).

Single-Molecule Rigid-Body Docking



Xtal structure

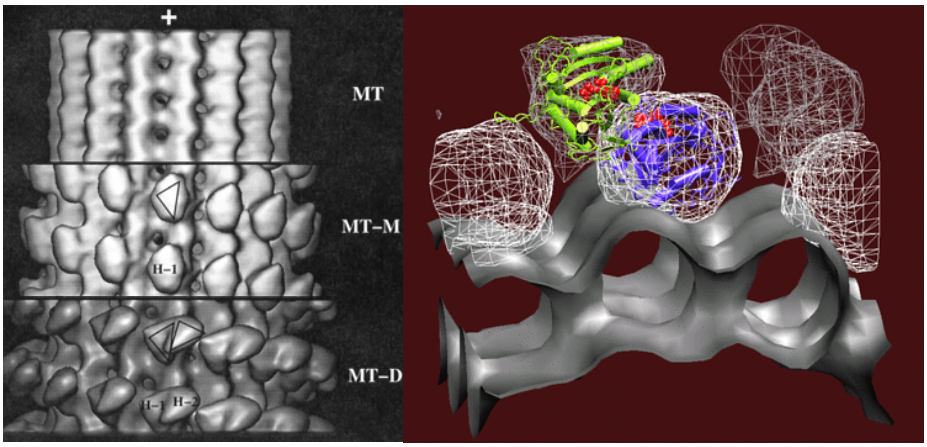
EM low res. data

•Estimate optimum k with variability criterion.
•Index map I: $m \rightarrow n$ ($m, n = 1,\ldots,k$).
• $k! = k\,(k\text{-}1)\ldots 2$ possible combinations.
• For each index map I perform a least squares fit of the $w_{I(j)}^{(h)}$ to the $w_j^{(l)}$.
• Quality of I: residual rms deviation

$$\Delta_I = \sqrt{\frac{1}{k}\sum_{j=1}^{k}\left\| w_{I(j)}^{(h)} - w_j^{(l)} \right\|^2}$$

• Find optimal I by direct enumeration of the $k!$ cases (minimum of Δ_I).

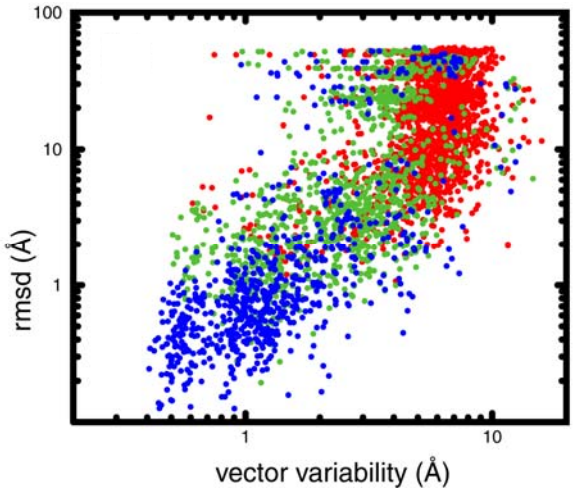
Application Example



ncd monomer and dimer-decorated microtubules (Milligan *et al.*, 1997)
ncd monomer crystal structure (Fletterick *et al.*, 1996,1998)

Precicion vs. Accuracy

Wriggers & Birmanns, J. Struct. Biol 133, 193-202 (2001)

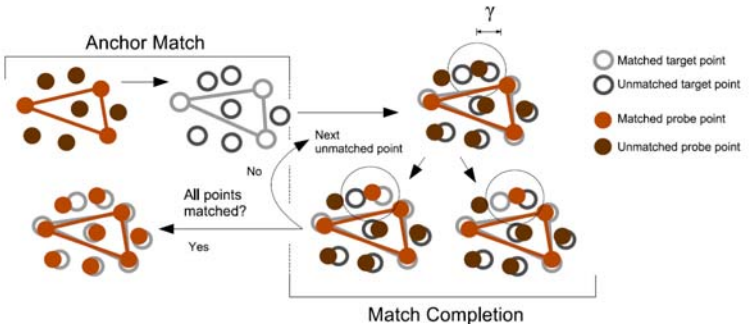


10 test systems, $3 \le k \le 9$ simulated EM densities from 2-100Å.

2-20Å (reliable fitting)
22-50Å (borderline)
52-100Å (mismatches)

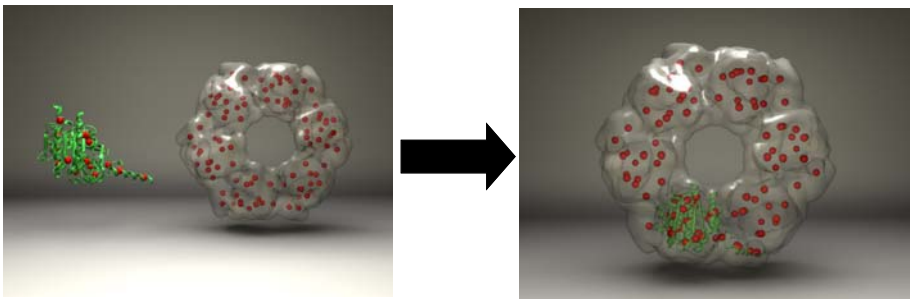
Precision of fiducial markers correlates with actual docking accuracy.

Anchor Point Registration: *matchpoint*



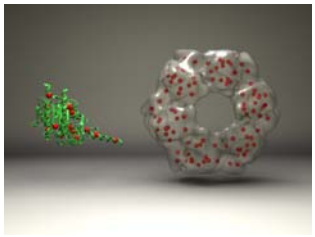
Birmanns & Wriggers *J. Struct. Biol.* (2007) 157:271

Anchor Point Registration: *matchpoint*



- $k \rightarrow h \neq k$ matching
- number of points k (atomic), h (EM) now determined by desired level of detail, not "variability criterion". k and h should give similar point density and are dependent on volume of atomic structure and EM map

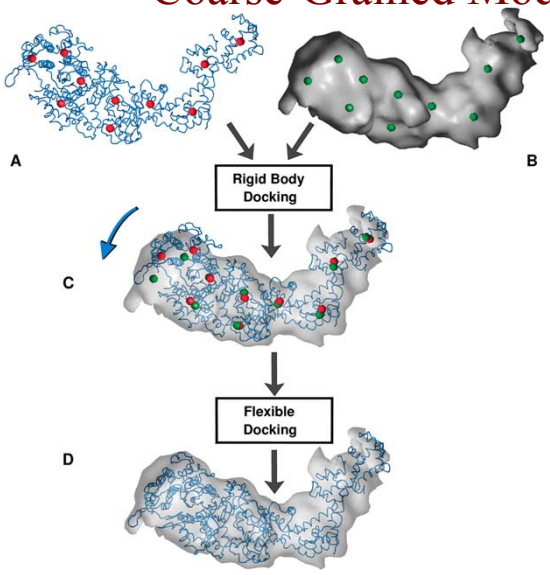
How to Determine Number of Points?



Also relevant for flexible fitting (below)!

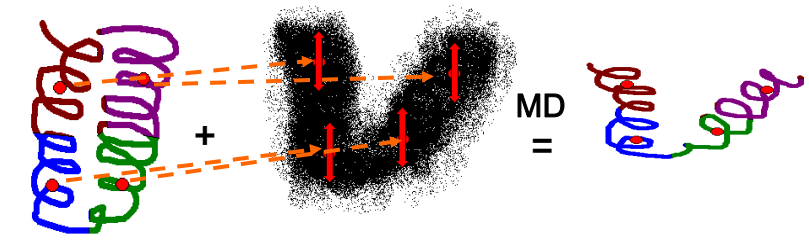
- Divide volume of EM map by volume of a "resolution element" (cube with dimension of numeric resolution value in Å).
- This gives the (maximum) number of resolved spatial features in the map.
- To avoid overfitting, we typically pick 50% of that maximum number for h.
- k is then h times the ratio of atomic to EM volume (yielding same point density, i.e. level of detail, as EM coarse graining).
- note that spatial resolution of coarse grained model scales with cubic root of number of points, so order of magnitude estimate for number of EM points h is sufficient, but k/h must closely reflect the atomic to EM volume ratio to be consistent.

2000-2008: Flexible Fitting using Coarse-Grained Models

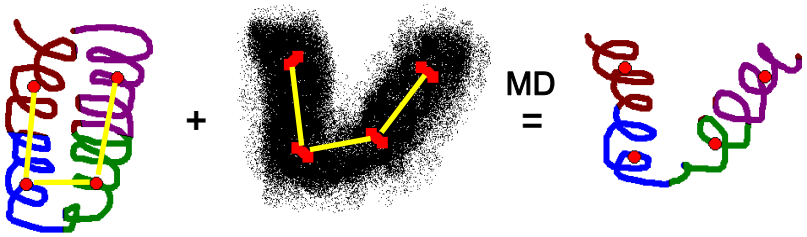


See above discussion on determining number of points in A,B!

Flexible Registration

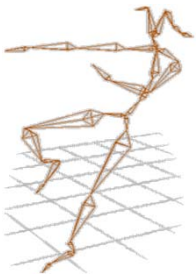


unrestrained markers exp. and meth. uncertainty distortion



skeleton distance constraints less distortion

Motion Capture

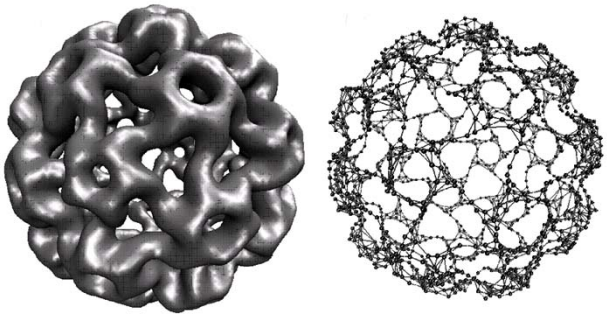



Motion Capture Network

Topology Representing Neural Network
(Martinetz and Schulten, 1993)

+

SHAKE Distance Constraints
(van Gunsteren, 1977)



Wriggers et al., Neurocomputing (2004) 56:365

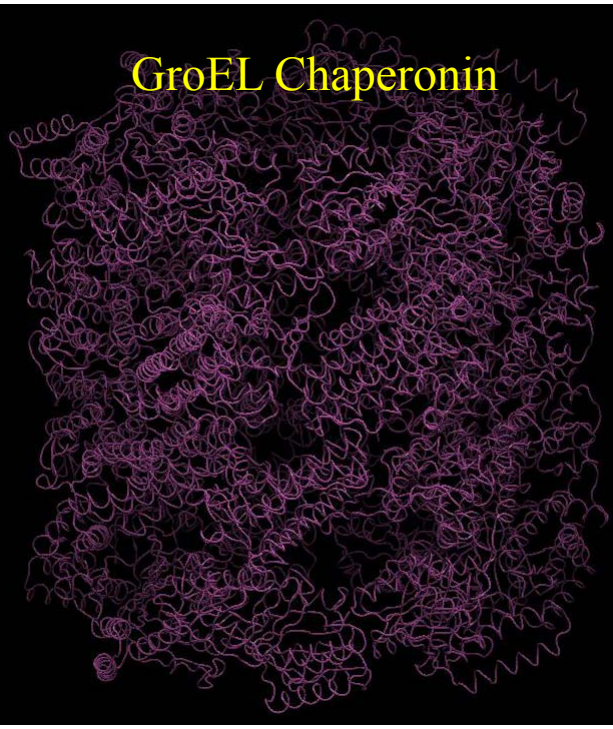
GroEL Chaperonin

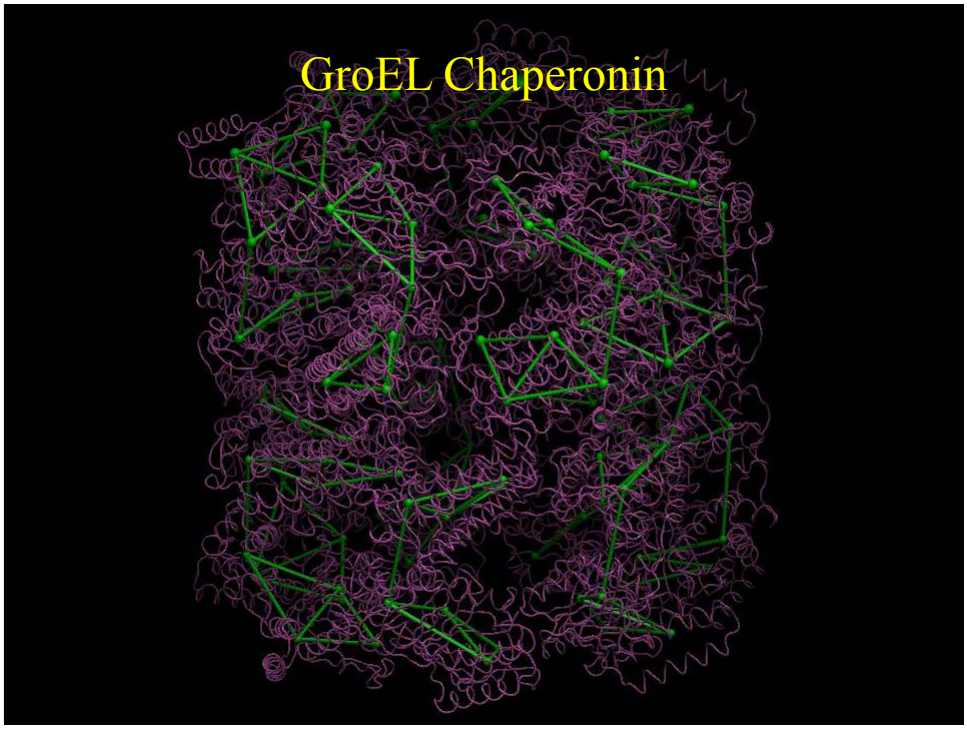
Dalia Segal,
Sharon Wolf,
Amnon Horovitz,
Weizmann
Institute, Israel

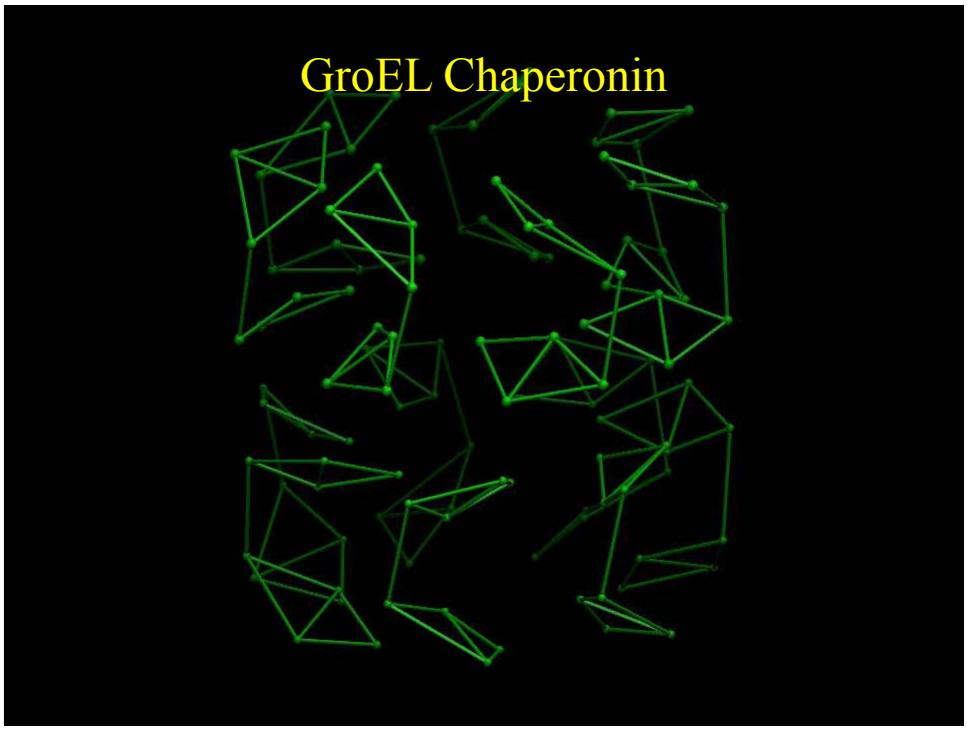
resolution ~14Å
wild type
(Sabil et al.)

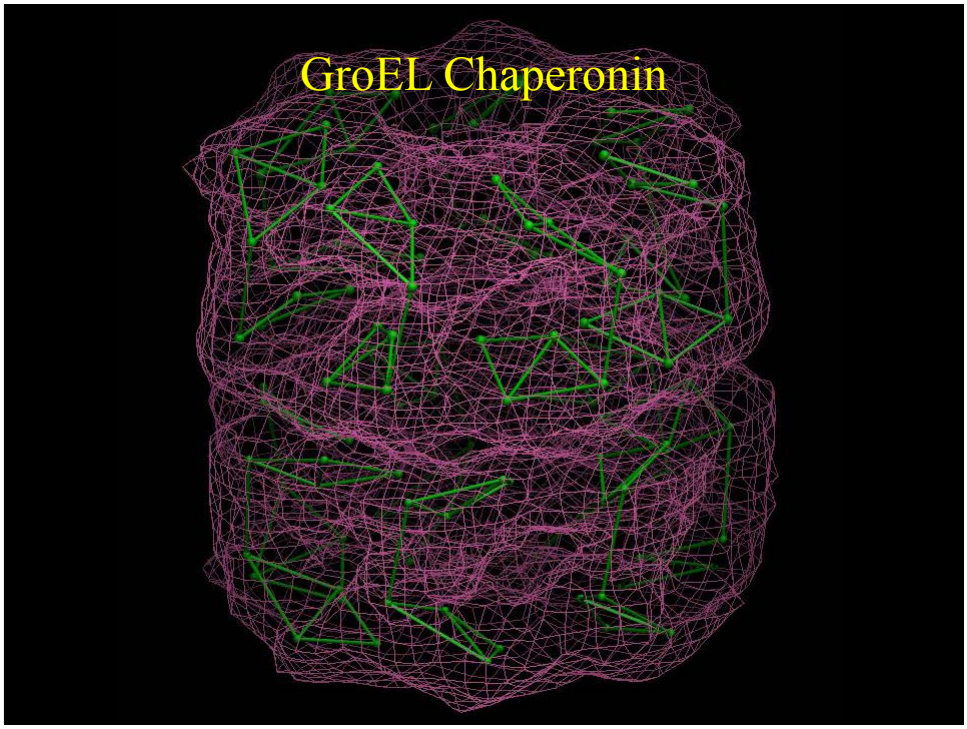
& mutant

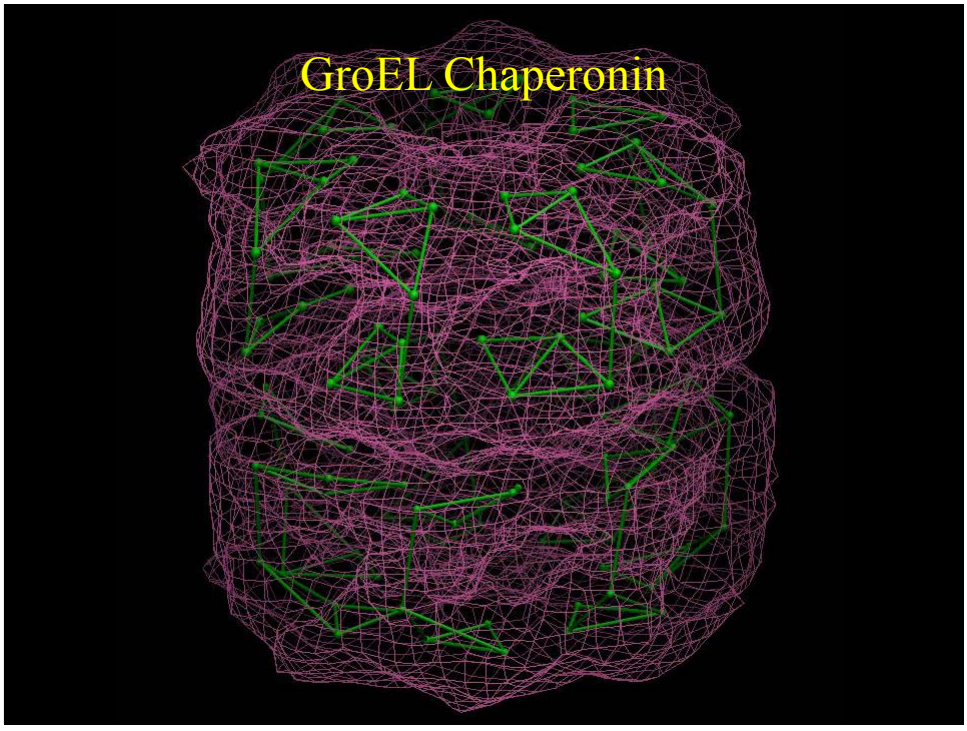
GroEL Chaperonin

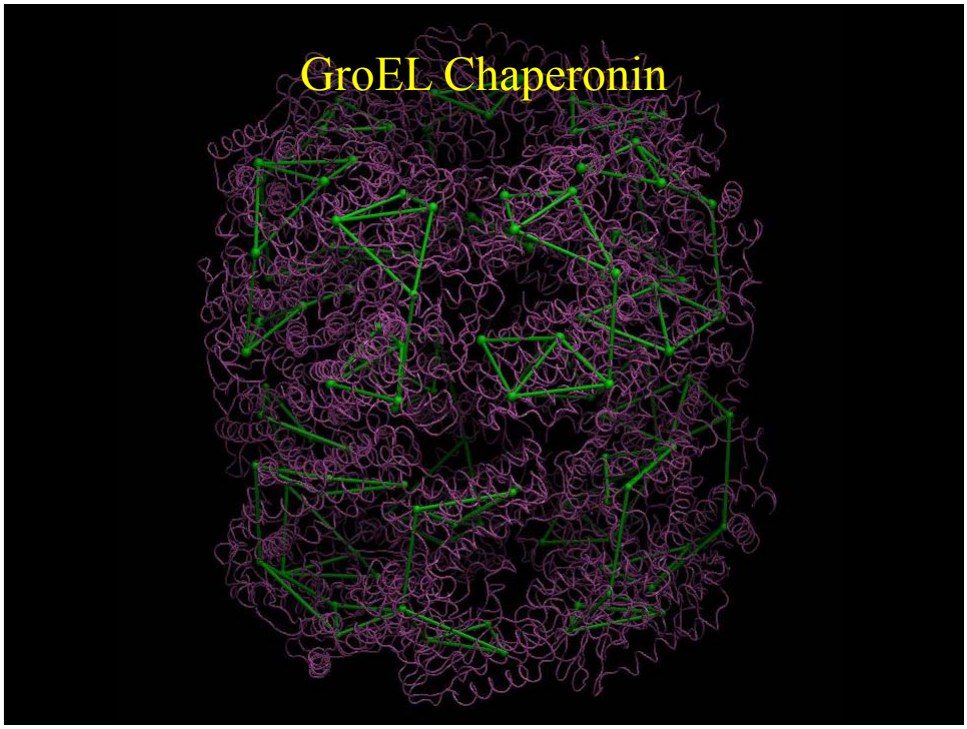


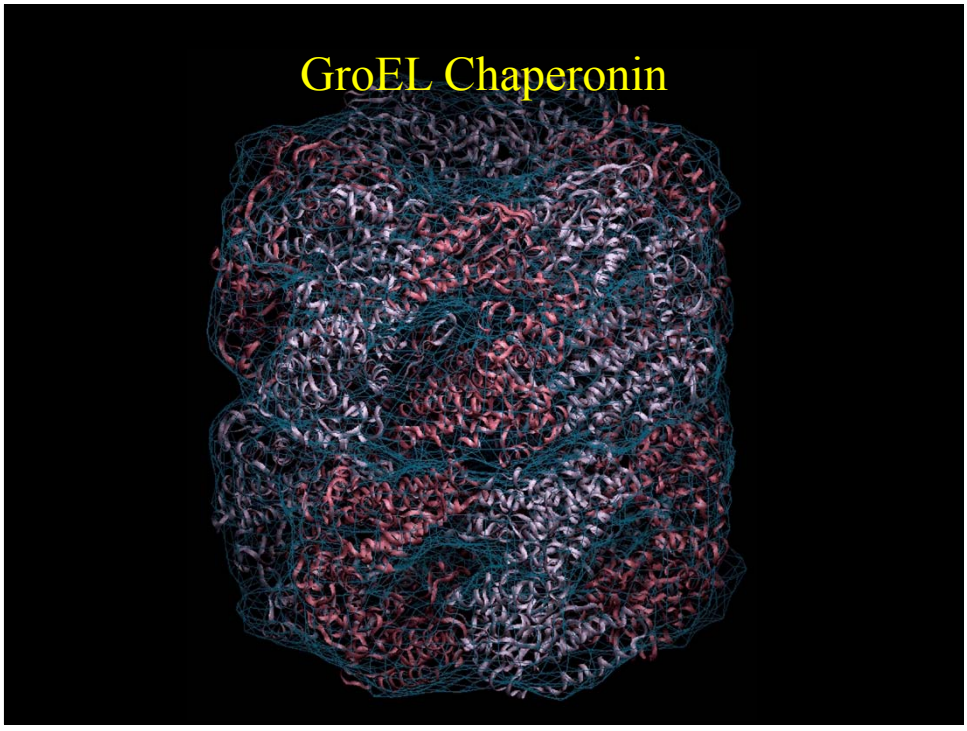

GroEL Chaperonin


GroEL Chaperonin

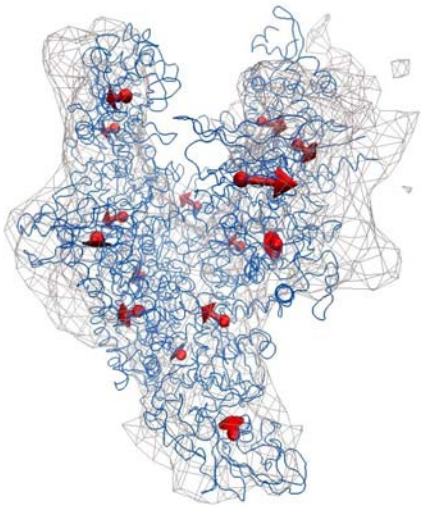

GroEL Chaperonin


GroEL Chaperonin


GroEL Chaperonin


GroEL Chaperonin

What Information is Used in Flexing?



Displacements

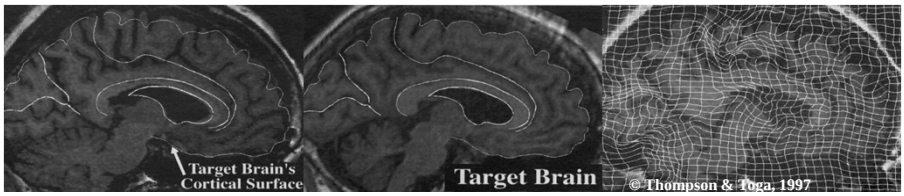
Molecular Dynamics

Molecular Dynamics vs. Interpolation

MD simulation such as in MDFF requires an expert user and hours of preparation. We know a sparse estimation of the displacement field at markers. Can we extend the sparse estimate to the full space by an inexpensive interpolation?

Interpolation Pros:

- Ease of use / implementation
- Detailed mass rearrangement plan.
- Linear or nonlinear registration of features
- Used in neuroscience and machine vision:



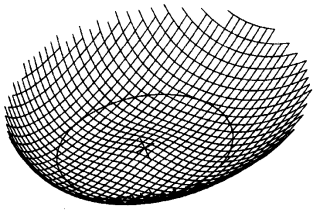
Bookstein "Thin-Plate" Splines

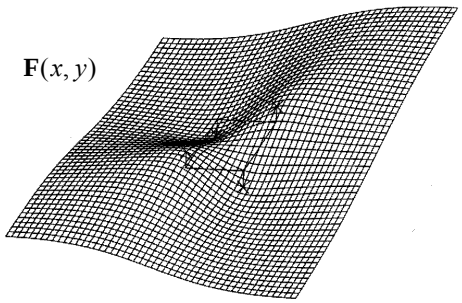
• Interpolation kernel function $U(r)$ is principal solution of biharmonic equation that arises in elasticity theory of thin plates:

$$\Delta^2 U(r) = \nabla^4 U(r) = \delta(r).$$

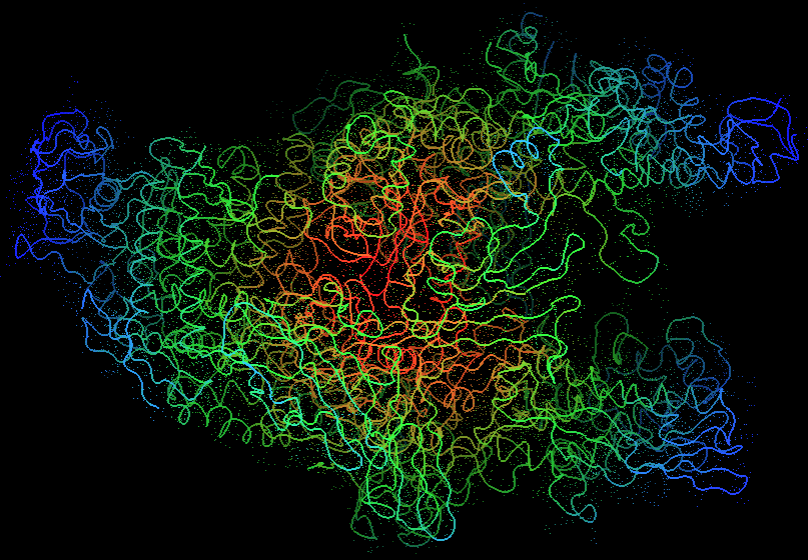
• variational principle: $U(r)$ minimizes the bending energy (not shown).
• 1D: $U(r) = |r^3|$ (cubic spline)
• 2D: $U(r) = r^2 \log r^2$
• 3D: $U(r) = |r|$

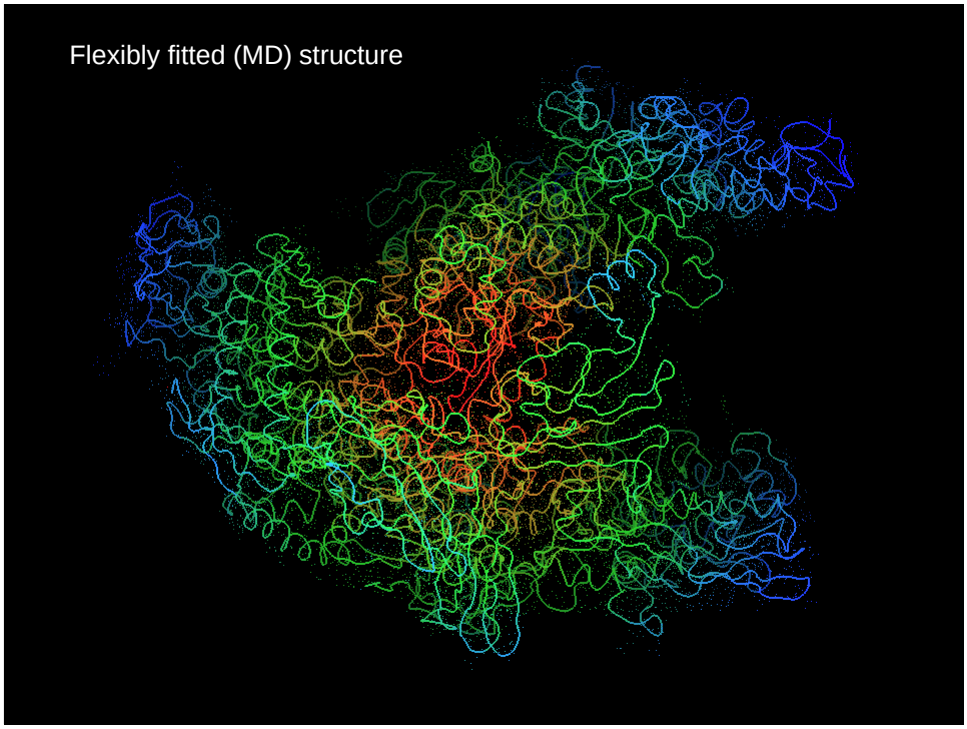
2D: $U(r)$ $\mathbf{F}(x, y)$

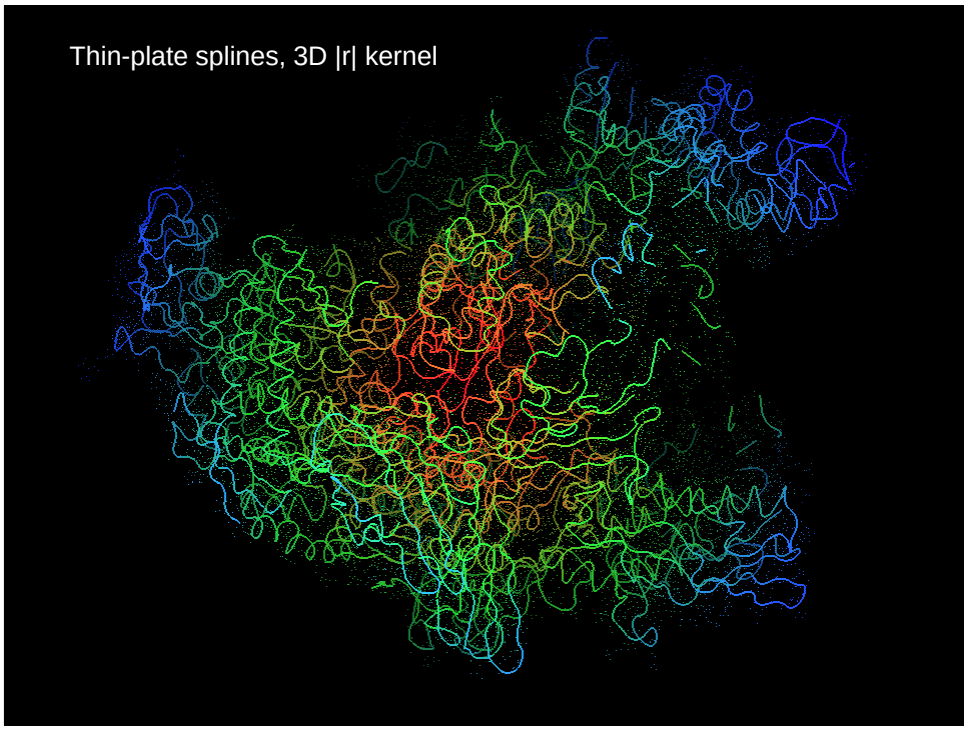




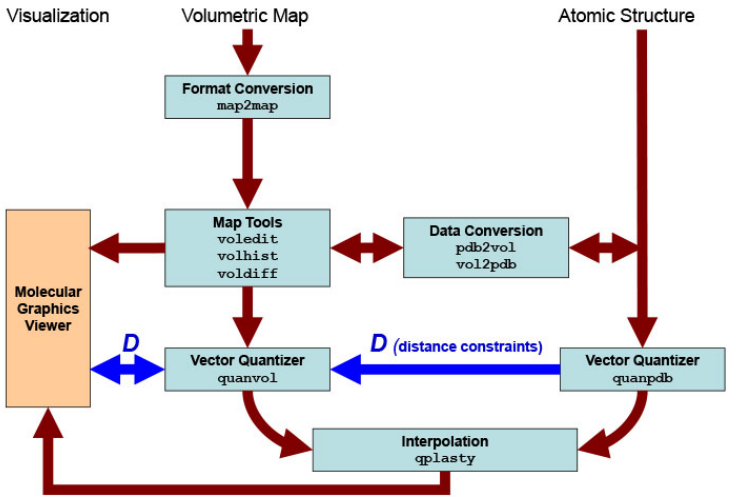
Taq RNAP x-tal structure




Flexibly fitted (MD) structure


Thin-plate splines, 3D |r| kernel

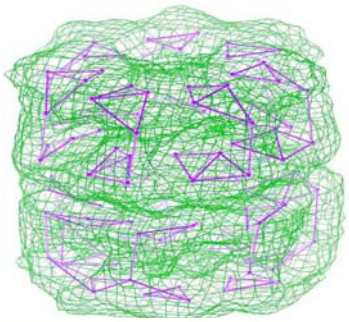
Simplified Flexing Option (*Situs* & *Sculptor*)

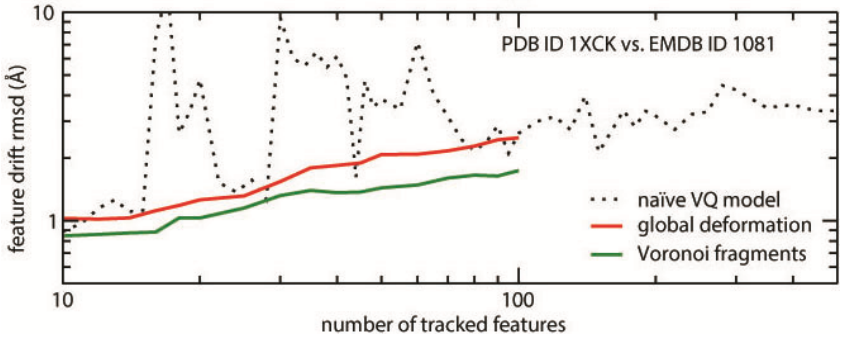


Mirabela Rusu, Stefan Birmanns, and Willy Wriggers.
Biomolecular Pleiomorphism Probed by Spatial Interpolation of Coarse Models.
Bioinformatics, 2008, 24:2460-2466.

Validation of Fitting Using Tracked Fiducials

Wriggers & He, *J Struct Biol* (2015) 192:255



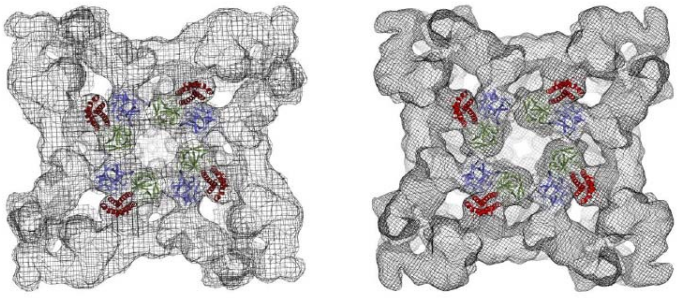


Best Practices for Fitted Model Validation

Wriggers & He, *J Struct Biol* (2015) 192:255

1. Use Different CryoEM Maps

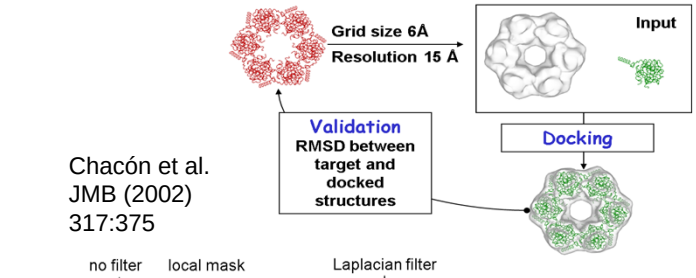
Filip Van Petegem: 3 domains of RyR solved with Xtal + 9.6Å
RyR EM map (Tung et al., *Nature* 468:585-588, 2010)



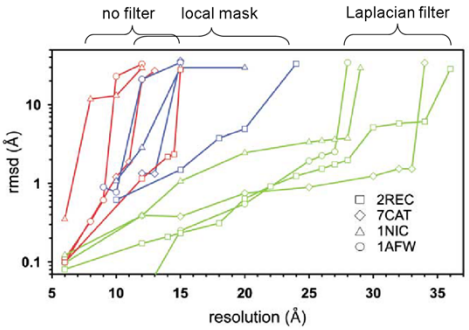
14Å vs. 10Å

Limitation: Comparison with older data may tell you more about problems with earlier maps (or structures) than about reliability of docking to new map.

2. Use Simulated Maps



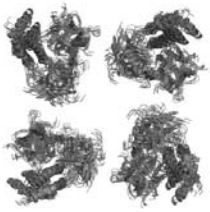
Chacón et al. JMB (2002) 317:375



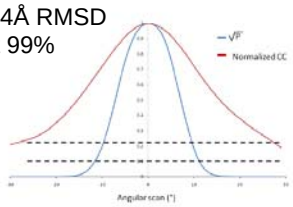
Limitation: simulated maps overly optimistic, only validate docking technique, not quality of structures.

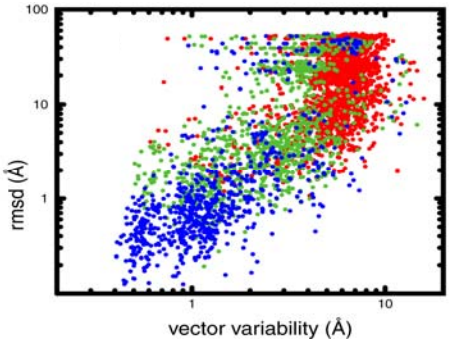
3. Statistical Confidence Analysis

Fisher Z-Transform of CC (Volkmann), see Tung et al., *Nature* 468:585-588, 2010



4.4Å RMSD at 99%



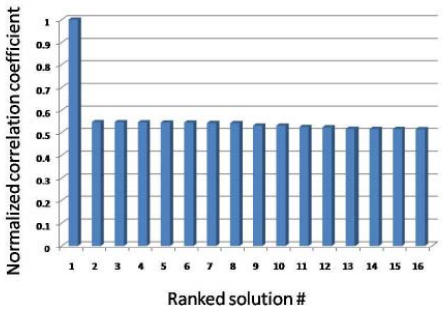


Precision of fiducial markers correlates with docking accuracy (Wriggers & Birmanns, J. Struct. Biol 133:193, 2001)

Limitation: intrinsic statistics of docking criterion (precision) is not always a reliable predictor of accuracy (distance from true structure)

4. "Docking Contrast"

Ranking of Situs (colores) results by correlation coefficient



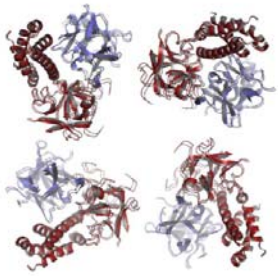
(Tung et al., *Nature*, 468:585-588, 2010)

Limitations:

•in case of low-resolution maps the "docking contrast" may be quite low
•sometimes sub-optimal solutions are correct based on other knowledge (effect of induced fit, steric clashes etc)
•use of Laplacian filter, single vs. multi-body docking, etc will give different contrast profile

5. Use Different Modeling Strategies

•Docking of parts recapitulates assembly
•Use of Laplacian filter
•Use multiple programs (Situs vs. ADP-EM: identical results)



1.26Å RMSD

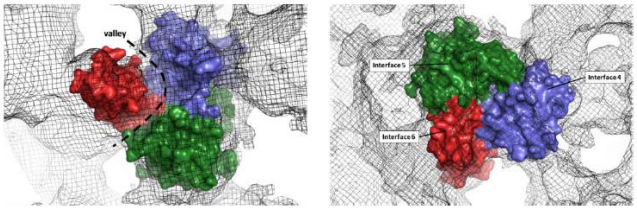
"The results indicate the RyR1ABC structure is at the size limit for standard cross-correlation-based docking into the current 9.6-Å cryoEM map, and that Laplacian filtering is absolutely required for docking of any smaller units".

(Tung et al., *Nature*, 468:585-588, 2010)

Limitation: How to resolve different outcomes in such meta-analysis?

6. Compare with Existing Knowledge

- Distribution of Disease Mutants
- Surface Features
- Labels
- Complementary biophysical or biochemical techniques



(Tung et al., *Nature*, 468:585-588, 2010)

Limitations: None. This is the most reliable "validation strategy" because it is independent of the data and modeling work flow.

Take-Home Messages

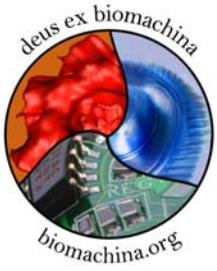
Flexible/rigid body docking precision about one order of magnitude above the nominal EM (or tomography) resolution

Our Fitting Software:
situs.biomachina.org (UNIX command-line tools)
sculptor.biomachina.org (GUI-based program)

Other Software:
http://en.wikibooks.org/wiki/Software_Tools_For_Molecular_Microscopy

Acknowledgements

Pablo Chacón
Jochen Heyd
Julio Kovacs
Yao Cong
Mirabela Rusu
Manuel Wahle
Stefan Birmanns
Zbigniew Starosolski



http://biomachina.org

Collaborators:
Maik Boltes & Herwig Zilken (Forschungszentrum Jülich)
Vitold Galkin and Edward Egelman (U Virginia)
Seth Darst (Rockefeller University)
Dalia Segal, Sharon Wolf, Amnon Horovitz (Weizmann Institute)
Alexander Rigort (MPI Martinsried)
Lorenzo Alamo and Raúl Padrón (HHMI, Venezuela)