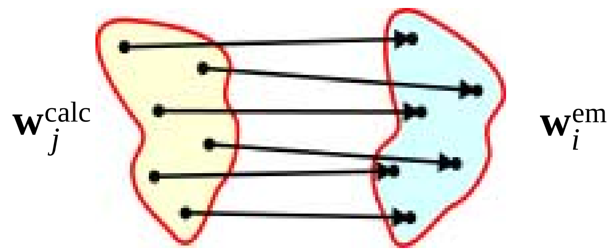
Vector Quantization and Reduced Models

Willy Wriggers, Ph.D.

Department of Molecular Biology
The Scripps Research Institute
10550 N. Torrey Pines Road, Mail TPC6
La Jolla, California, 92037

Situs Modeling Workshop, San Diego, CA, Feb. 3-5, 2003

Reduced Representations of Biomolecular Structure



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

Useful for:

- Rigid-body fitting (today)
- Flexible fitting (today)
- Interactive fitting / force feedback (S. Birmanns, Tu 9AM)
- Building of deformable models (F. Tama, P. Chacon, Tu 10AM)

Vector Quantization

Lloyd (1957)	Digital Signal Processing,
Linde, Buzo, & Gray (1980)	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”):

$$\mathrm{V}_i = \{ v \in \Re^3 \,\|\, \|v - w_i\| \le \|v - w_j\| \,\forall j \}$$

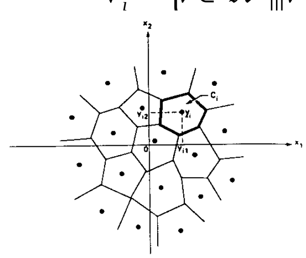
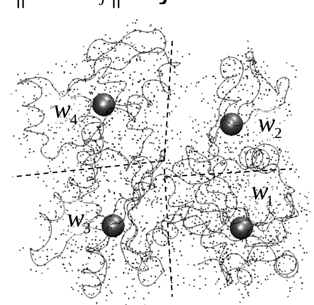
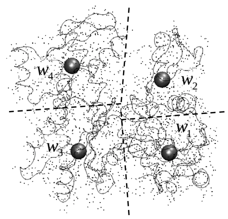


Fig. 3. Partitioning of two-dimensional space (N = 2) into L = 18 cells. All input vectors in cell C_i will be quantized as the code vector y_i. The shapes of the various cells can be very different.



Linde, Buzo, Gray (LBG) Algorithm



Encoding Distortion Error:

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

Lower $E(\{w_j(t)\})$ iteratively: Gradient descent $\forall r:$

$$\Delta w_r(t) \equiv w_r(t) - w_r(t-1) = -\frac{\varepsilon}{2} \cdot \frac{\partial E}{\partial w_r} = \varepsilon \cdot \sum_i \delta_{rj(i)} (v_i - w_r) m_i \ .$$

Inline (Monte Carlo) approach for a sequence $v_i(t)$ selected at random according to weights m_i :

$$\Delta w_r(t) = \tilde{\varepsilon} \cdot \delta_{rj(i)} \cdot (v_i(t) - w_r).$$

How do we avoid getting trapped in the many local minima of E?

Soft-Max Adaptation

Avoid local minima by smoothing of energy function (here: TRN method):

$$\forall r: \ \Delta w_r(t) = \tilde{\varepsilon} \cdot e^{\frac{-s_r}{\lambda}} \cdot (v_i(t) - w_r),$$

Where $s_r(v_i(t), \{w_j\})$ is the closeness rank:

$$\left\| v_i - w_{j0} \right\| \leq \left\| v_i - w_{j1} \right\| \leq \ldots \leq \left\| v_i - w_{j(k-1)} \right\|$$

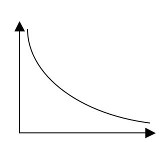
$$s_r = 0 \qquad s_r = 1 \qquad s_r = k-1$$

Note: $\lambda \to 0$: LBG algorithm.
$\lambda \neq 0$: not only "winner" $w_{j(i)}$, also second, third, ... closest are updated.

Can show that this corresponds to stochastic gradient descent on

$$\tilde{E}(\{w_j\}, \lambda) = \sum_{r=1}^{k} e^{\frac{-s_r}{\lambda}} \sum_i \left\| v_i - w_{j(i)} \right\|^2 m_i \ .$$

Note: $\lambda \to 0$: $\tilde{E} \to E$. LBG algorithm.
$\lambda \to \infty$: $\tilde{E}$ parabolic (single minimum).
$\Big\} \Rightarrow \lambda(t)$



Q: How do we know that we have found the global minimum of E?
A: We don't (in general).

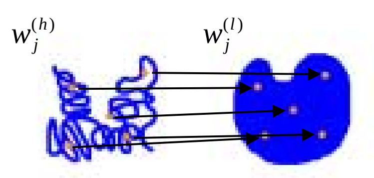
But we can compute the statistical variability of the $\{w_j\}$ by repeating the calculation with different seeds for random number generator.

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.

Single-Molecule Rigid-Body Docking

Xtal structure



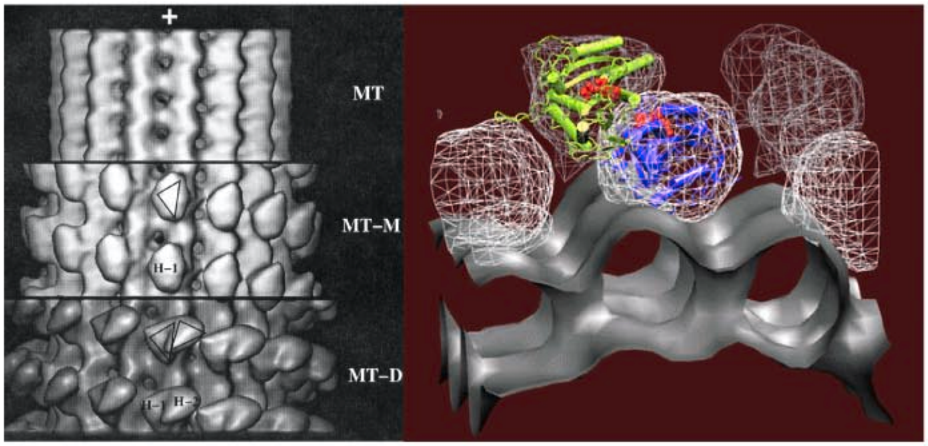
EM low res. data

•Estimate optimum k with variability criterion.
•Index map I: $m \rightarrow n$ ($m, n = 1,\ldots,k$).
• $k! = k\,(k-1)\ldots2$ 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

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

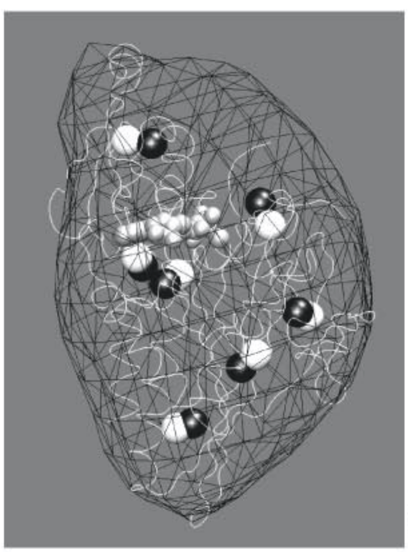
• Find optimal I by direct enumeration of the $k!$ cases (minimum of Δ_I).

Application Example



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

Search for Conformations



Two possible ranking criteria:

- Codebook vector rms deviation (Δ_l).
- Overlap between both data sets:

Voxel-Correlation coefficient:

$$C_{hl} = \frac{\sum_{x,y,z} h_{x,y,z} \cdot l_{x,y,z}}{\left(\sum_{x,y,z} h^2_{x,y,z}\right)^{\frac{1}{2}} \left(\sum_{x,y,z} l^2_{x,y,z}\right)^{\frac{1}{2}}}$$

ncd motor (white, shown with ATP nucleotide) docked to EM map (black) using *k*=7 codebook vectors

Reduced Search Features

Top 20, 7!=5040 possible pairs of codebook vectors.

	Δ_I	C_{hl}	I (permutation)
1.	3.115	0.913	(7,5,1,6,4,2,3)
2.	4.946	0.904	(2,3,5,7,4,6,1)
3.	5.455	0.897	(6,1,3,2,4,7,5)
4.	6.316	0.882	(5,7,4,3,1,2,6)
5.	7.612	0.867	(5,7,1,4,6,3,2)
6.	7.855	0.888	(3,2,4,1,5,6,7)
7.	7.994	0.884	(1,6,4,5,3,7,2)
8.	8.001	0.863	(6,1,4,3,5,2,7)
9.	8.192	0.888	(2,6,4,3,1,7,5)
10.	8.244	0.850	(7,5,6,2,1,3,4)
11.	8.298	0.881	(2,6,7,5,1,3,4)
12.	8.340	0.894	(6,2,4,1,3,5,7)
13.	8.481	0.867	(3,4,6,2,1,5,7)
14.	8.516	0.885	(2,3,4,5,1,7,6)
15.	8.532	0.857	(7,5,4,1,3,6,2)
16.	8.985	0.861	(6,1,5,7,4,3,2)
17.	8.988	0.838	(3,4,5,7,1,2,6)
18.	9.092	0.839	(3,2,5,4,7,1,6)
19.	9.124	0.858	(7,5,3,2,4,1,6)
20.	9.236	0.858	(1,6,5,7,4,2,3)

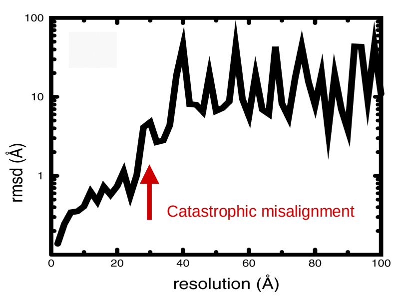
For a fixed k, codebook rmsd is more stringent criterion than correlation coefficient!

Performance (I)

Dependence on experimental EM density threshold (ncd, k=7):

orientations are stable:
+/- 5° variability for +/-50% threshold density variation.
Threshold level can be optimized via radius of gyration of vectors.

Dependence on resolution (simulated EM map, automatic assignment of k from $3 \leq k \leq 9$):



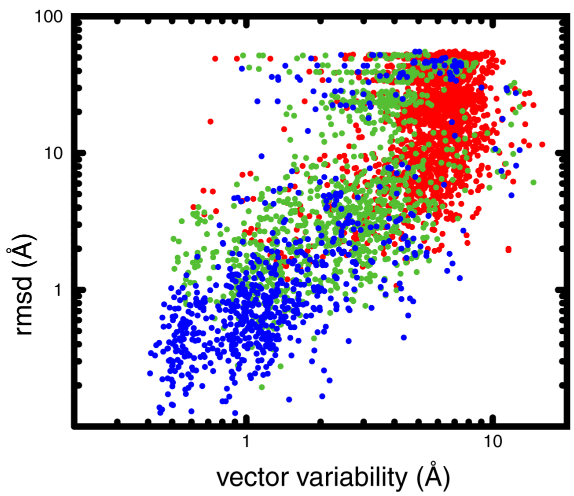
Deviation from start structure (PDB: 1TOP) used to generate simulated EM map.

Accurate matching up to ~30Å

Performance (II)

Is minimum vector variability a suitable choice for optimum k?

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

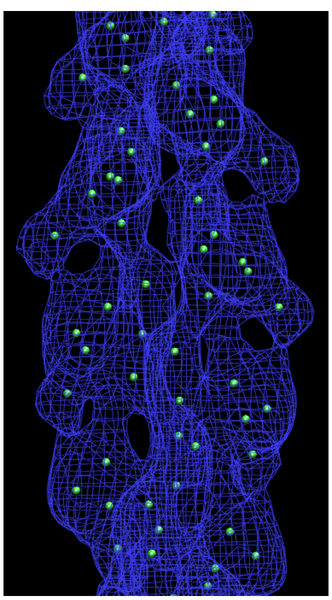


10 test systems, $3 \leq k \leq 9$ simulated EM densities from 2-100Å.

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

Reasonable correlation with actual deviation

No "false positives" for resolution values < 20Å and variability < 1Å.



Performance (III)

Multiple Subunits

Egelman lab: High-resolution reconstructions of F-actin - plant ADF based on single-particle image processing.

Unrestrained vectors fail to distinguish between actin and ADF densities (poor segmentation)

Remedies:

•Skeletons (today)
•Correlation-Based Search (P Chacón, today; J. Kovacs, tomorrow)

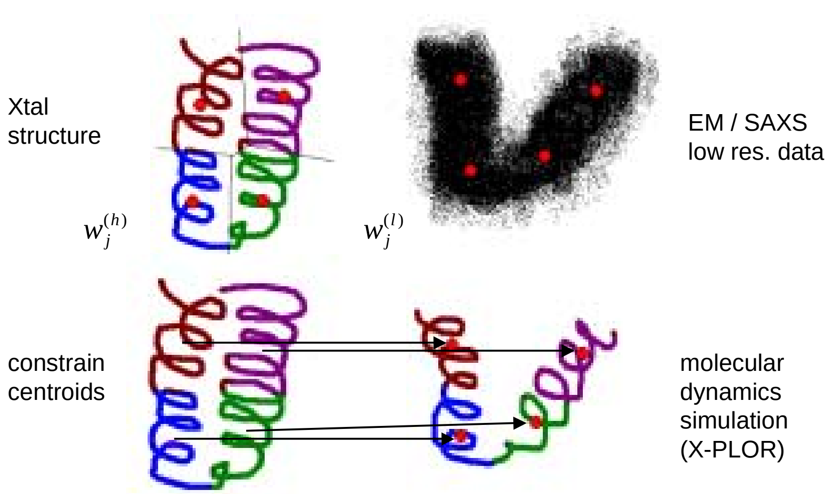
Conclusions (Rigid-Body Fitting)

“Classic” Situs fitting approach, versions 1.0-1.4.

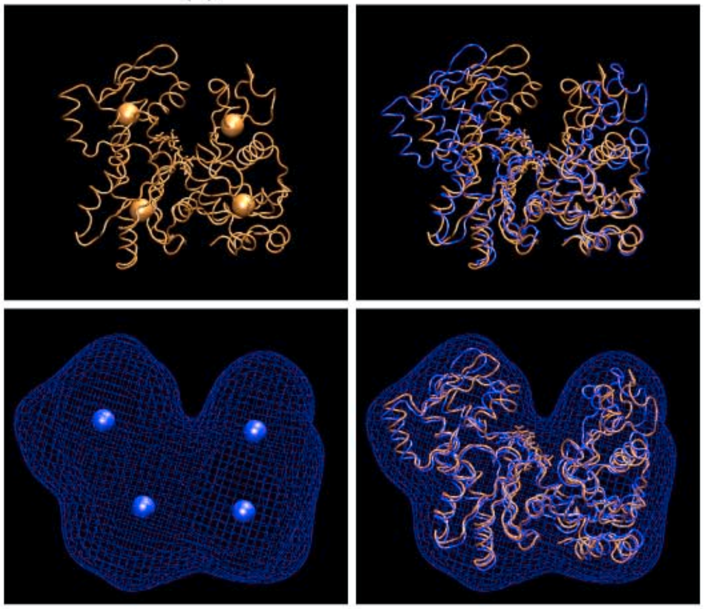
Advantages of vector quantization:
•Fast (seconds of compute time).
•Reduced search is robust.

Limitations:
•No estimation of “fitting contrast” near optimum
•Works best for single molecules, not for matching subunits to larger densities.

Flexible Fitting with Molecular Dynamics



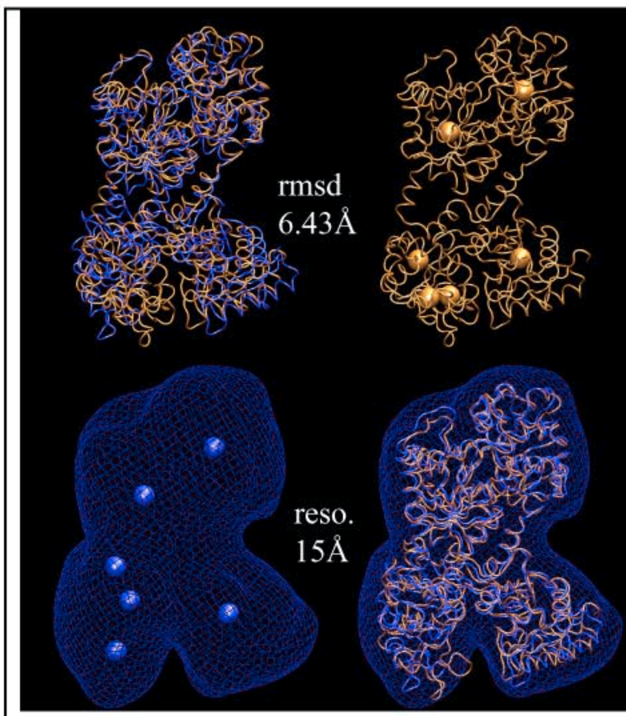
Control (I): Simulation of G-Actin



rmsd
4.36Å

reso.
15Å

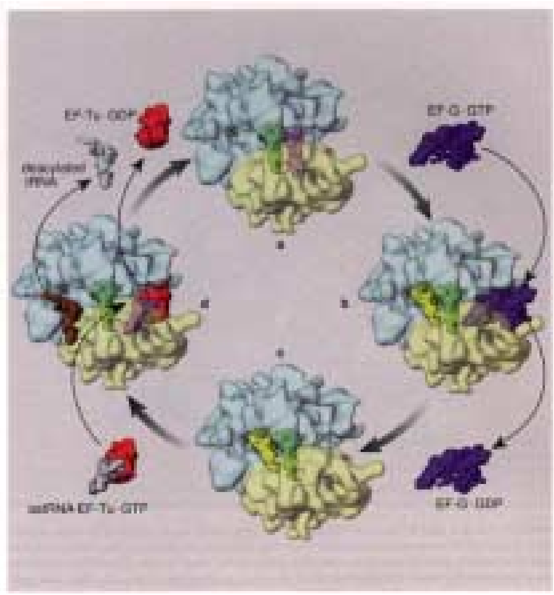
rmsd
1.37Å



Control (II):
Lactoferrin
(1LFG,1LFH)

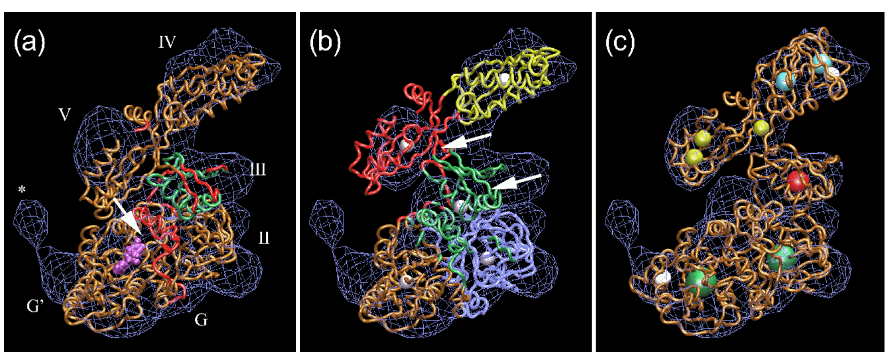
rmsd
2.72Å

Flexible Docking of Elongation Factor G



binding of EF-G
and EF-Tu
to the ribosome

Flexible Docking of Elongation Factor G



rigid-body docking

flexible docking
(5 vectors)

flexible docking
(10 vectors,
variable number per
domain)

see Wriggers *et al*., Biophys. J. (2000) 79:1670-1678.

Note possible overfitting of domain IV!

Stereochemical Quality of Flexible Fitting

1.) Assumption: structure remains locally similar to the initial crystal structure.

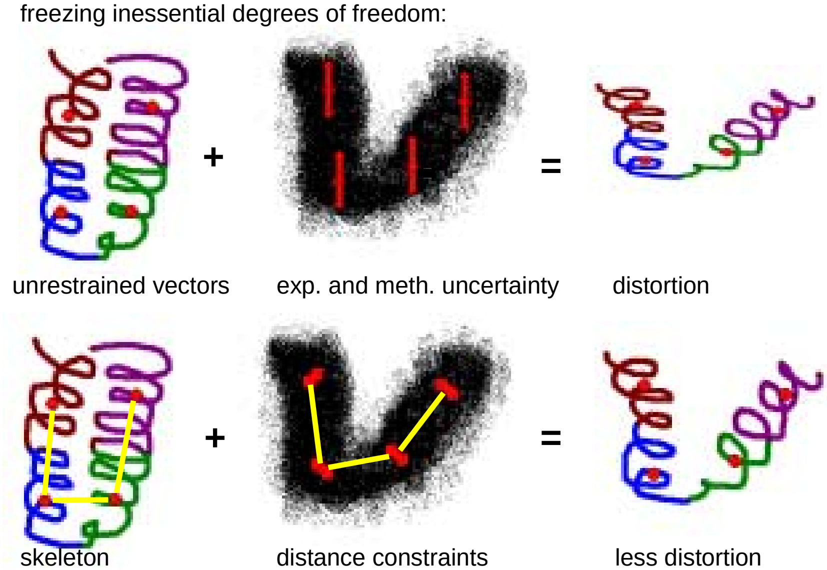
In this case precision: ~10 times above the nominal resolution of the EM map, but it is not known in advance if the assumption holds.

2.) The atomic model has many more degrees of freedom than there are independent pieces of information in the EM map. Hence, there is the danger that overfitting distorts the structure

How can overfitting be avoided? Reduce noise by eliminating “inessential” degrees of freedom!…

Skeletons Limit the Effect of Noise:

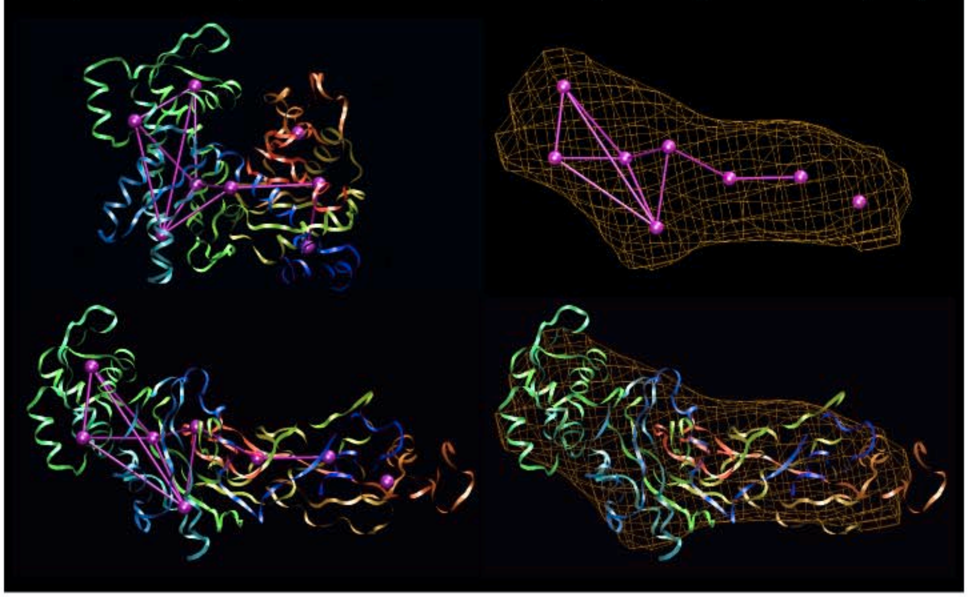
freezing inessential degrees of freedom:



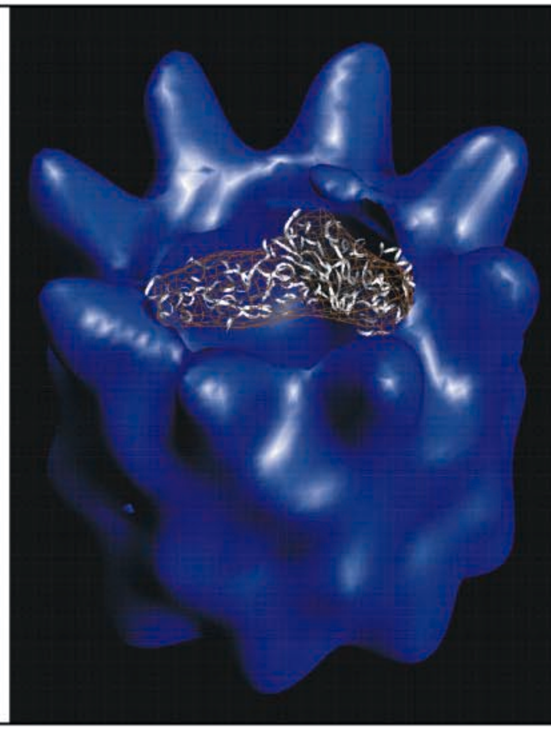
Fitting Skeletons: Motion Capture

Example: Actin-CCT

Valpuesta lab: chaperonin CCT unfolds bound actin (Llorca et al., EMBO J. 19:5971, 2000)



Visualization with Situs and VMD



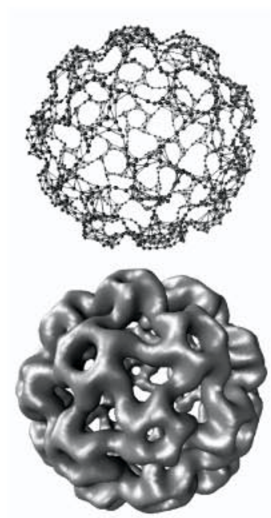
Estimating Adjacency: Competitive Hebb Rule

Implemented after Situs 1.4:

Nearest-neighbor search can be coupled with vector quantization (Martinetz & Schulten, 1993):

Initially, set all connections C_{ij} to zero.
For each VQ adaptation step:

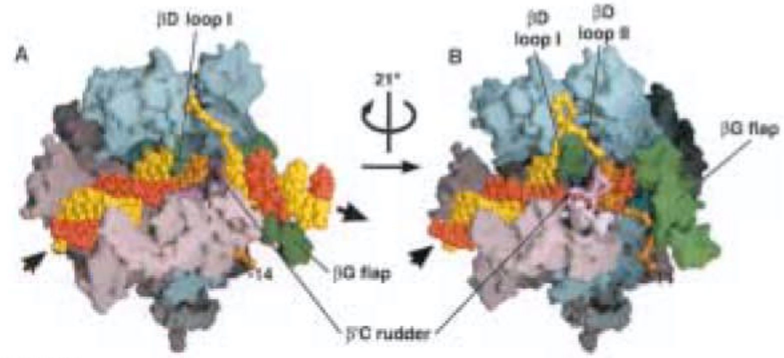
1. Find pair of winning vectors, w_{j0} , w_{j1} .
2. Set $C_{j0,j1} = 1$ (connect) $T_{j0,j1} = 0$ (refresh).
3. Increase the age of all connections of $j0$:
$$\forall j : T_{j0,j} = C_{j0,j} \bullet (T_{j0,j} + 1)$$
4. Remove old connections. If $T_{j0,j1} > T$,
   set $C_{j0,j1} = 0$.
5. Continue with next VQ step.



Cowpea chlorotic mottle virus at 23 Å resolution (1380 vectors).

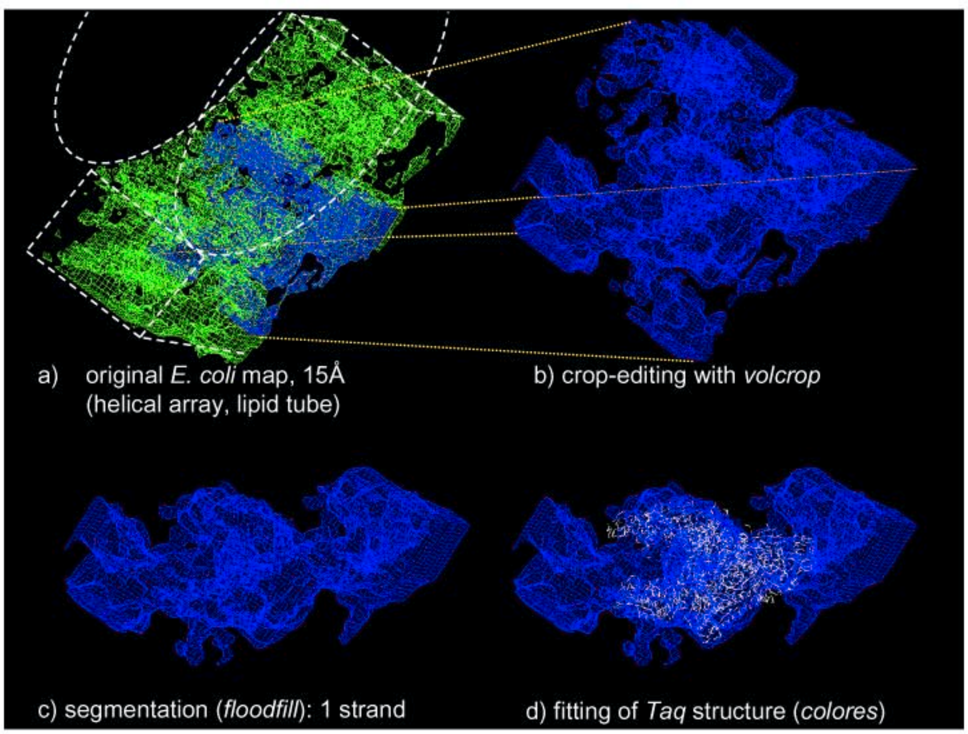
Flexible Fitting of RNAP

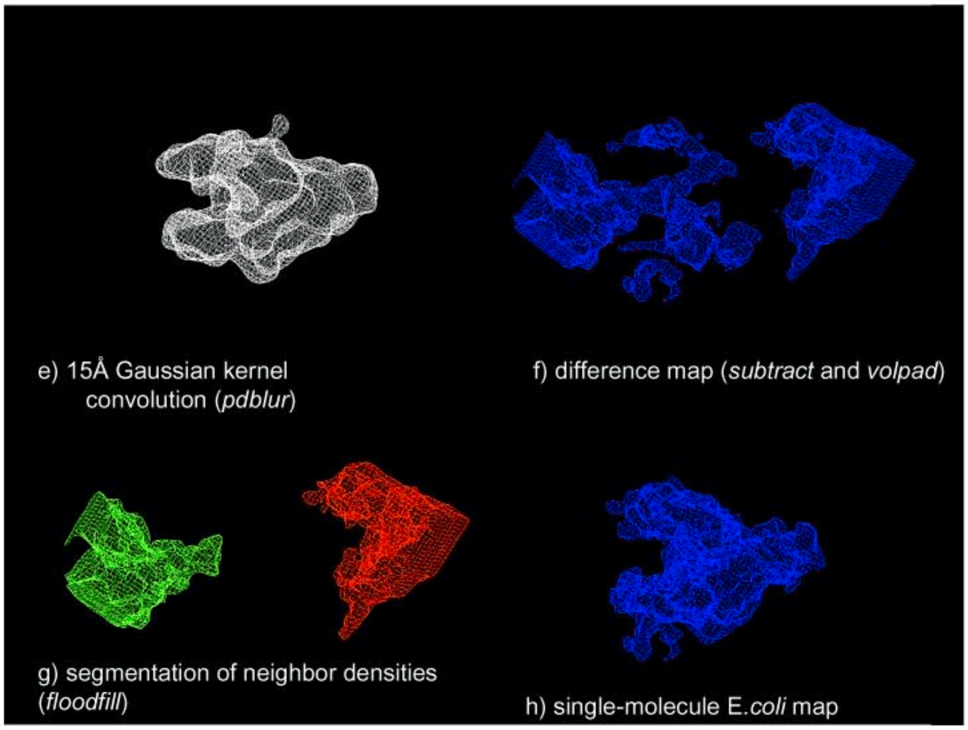
Bacterial RNA Polymerase

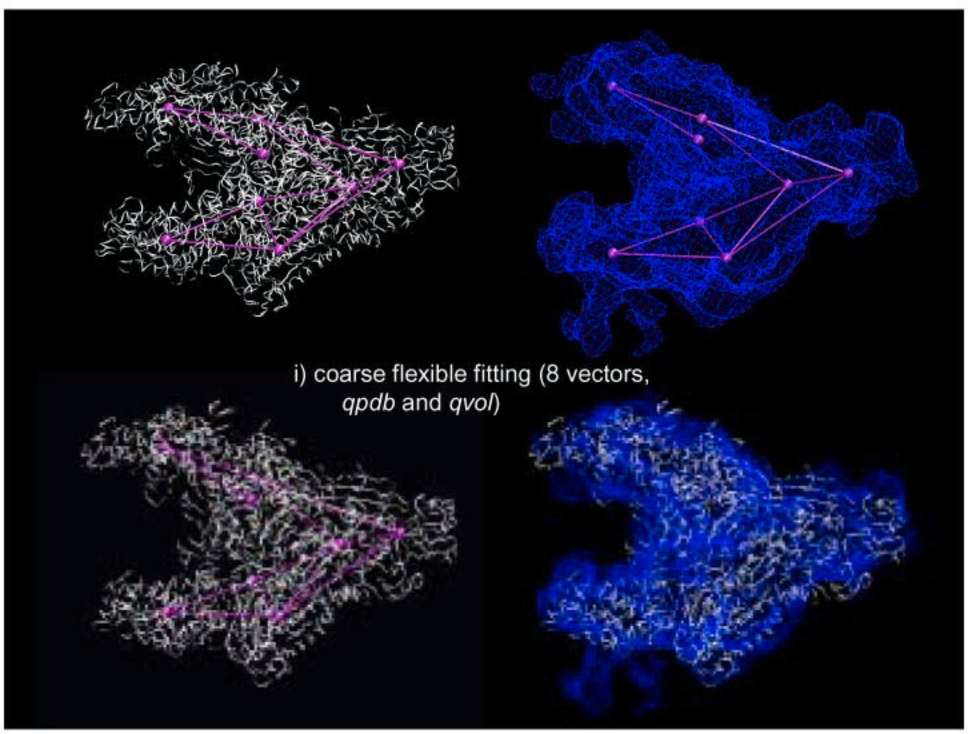


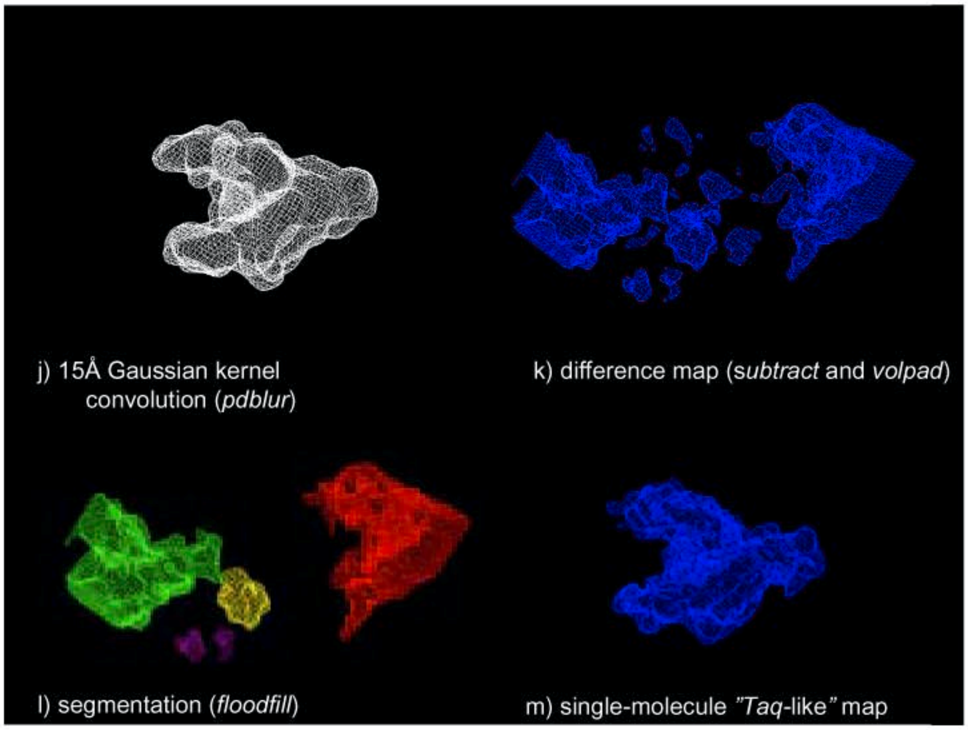
Darst lab:
Crystal Structure of *Thermus aquaticus* RNAP: Zhang *et al.,* Cell 98:811 (1999)
EM map of *E. coli* RNAP: Opalka *et al.,* PNAS 97:617, 2000
Model of Transcription Elongation: Korzheva *et al.,* Science 289:619 (2000)

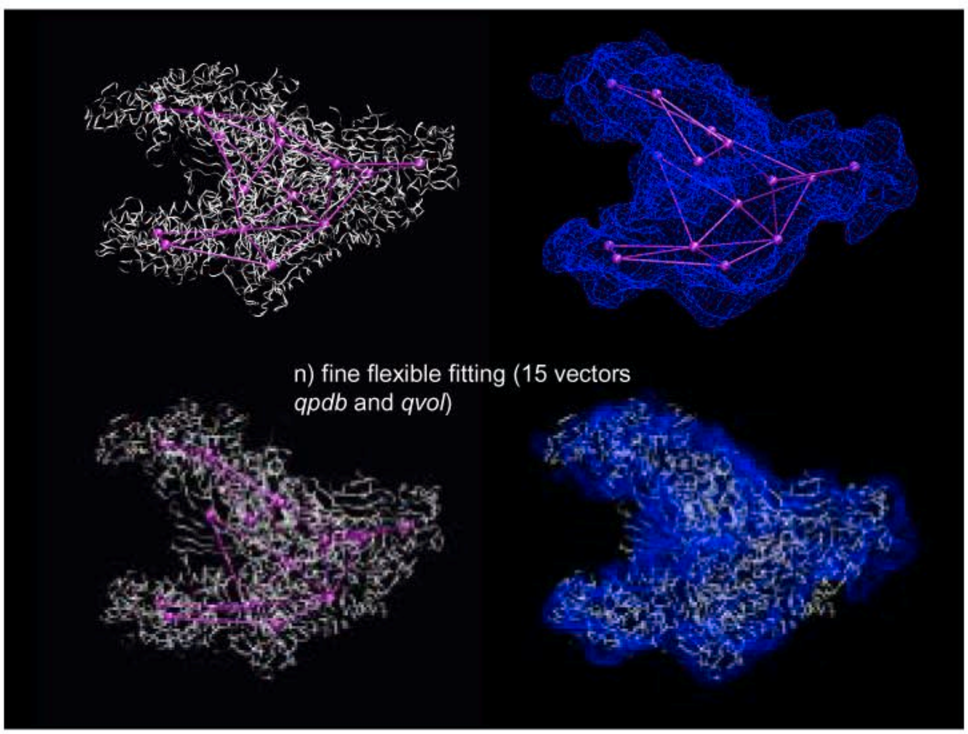
Hypothesis: Flexing of RNAP "jaws" encloses DNA



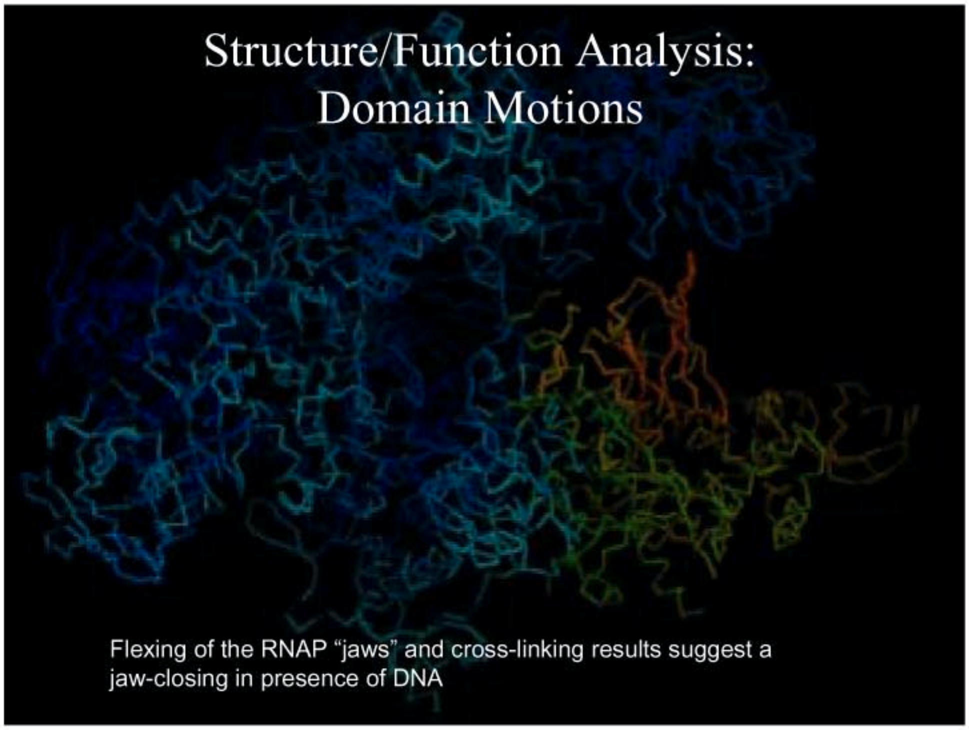

e) 15Å Gaussian kernel
convolution (*pdblur*)
f) difference map (*subtract* and *volpad*)
g) segmentation of neighbor densities
(*floodfill*)
h) single-molecule E.*coli* map


i) coarse flexible fitting (8 vectors,
qpdb and *qvol*)


j) 15Å Gaussian kernel
convolution (*pdblur*)
k) difference map (s*ubtract* and *volpad*)
l) segmentation (*floodfill*)
m) single-molecule *"Taq*-like*"* map


n) fine flexible fitting (15 vectors
qpdb and *qvol*)

Structure/Function Analysis: Domain Motions



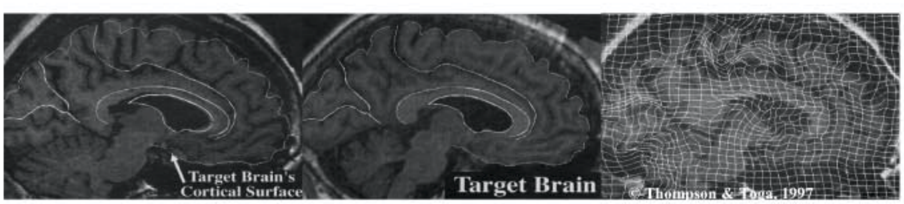
Flexing of the RNAP "jaws" and cross-linking results suggest a jaw-closing in presence of DNA

Molecular Dynamics vs. Interpolation

MD simulation requires an expert user and hours of preparation. We know the codebook vectors, i.e. a sparse estimation of the displacement field. Can we extend the sparse estimate to the full space by an inexpensive interpolation?

Interpolation Pros:

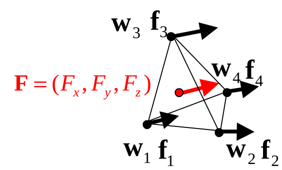
- Ease of use / implementation
- Detailed mass rearrangement plan.
- Linear or nonlinear registration of features
- Used in neuroscience and machine vision:



Cons:
- Validity of physical model?
- Stereochemical (structural) distortions?

(i) Piecewise-Linear Inter- / Extrapolation

For each probe position find 4 closest vectors.



Ansatz:

$$F_x(x,y,z) = ax + by + xz + d$$
$$F_x(\mathbf{w}_1) = f_{1,x},$$
$$F_x(\mathbf{w}_2) = f_{2,x},$$
$$F_x(\mathbf{w}_3) = f_{3,x},$$
$$F_x(\mathbf{w}_4) = f_{4,x} \quad \text{(similar for } F_y, F_z\text{)}.$$

Cramer's rule:

$$a = \frac{\begin{vmatrix} f_{1,x} & w_{1,y} & w_{1,z} & 1 \\ f_{2,x} & w_{2,y} & w_{2,z} & 1 \\ f_{3,x} & w_{3,y} & w_{3,z} & 1 \\ f_{4,x} & w_{4,y} & w_{4,z} & 1 \end{vmatrix}}{D}, \quad b = \frac{\begin{vmatrix} w_{1,x} & f_{1,y} & w_{1,z} & 1 \\ w_{2,x} & f_{2,y} & w_{2,z} & 1 \\ w_{3,x} & f_{3,y} & w_{3,z} & 1 \\ w_{4,x} & f_{4,y} & w_{4,z} & 1 \end{vmatrix}}{D}, \quad \cdots, \quad D = \begin{vmatrix} w_{1,x} & w_{1,y} & w_{1,z} & 1 \\ w_{2,x} & w_{2,y} & w_{2,z} & 1 \\ w_{3,x} & w_{3,y} & w_{3,z} & 1 \\ w_{4,x} & w_{4,y} & w_{4,z} & 1 \end{vmatrix}$$

(ii) Non-Linear Kernel Interpolation

Consider all k vectors and interpolation kernel function $U(r)$.

Ansatz:

$$F_x(x,y,z) = a_1 + a_x x + a_y y + a_z z + \sum_{k=1}^{k} b_i \cdot U\left(\left|\mathbf{w}_i - (x,y,z)\right|\right)$$
$$F_x(\mathbf{w}_i) = f_{i,x}, \ \forall i \quad \text{(similar for } F_y, F_z\text{)}.$$

Solve :

$$\mathrm{L}^{-1}(f_{1,x},\cdots,f_{k,x},0,0,0,0) = (b_1,\cdots,b_k,a_1,a_x,a_y,a_z)^{\mathrm{T}},$$

$$\text{where} \quad \mathbf{L} = \left(\begin{array}{c|c} \mathbf{P} & \mathbf{Q} \\ \hline \mathbf{Q}^{\mathrm{T}} & \mathbf{0} \end{array}\right), \quad \mathbf{Q} = \begin{pmatrix} 1 & w_{1,x} & w_{1,y} & w_{1,z} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & w_{k,x} & w_{k,y} & w_{k,z} \end{pmatrix}, \ k \times 4,$$

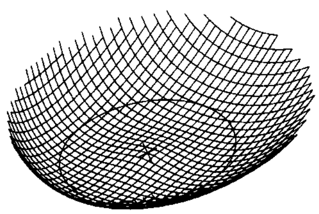
$$\mathbf{P} = \begin{pmatrix} 0 & U(w_{12}) & \cdots & U(w_{1k}) \\ U(w_{21}) & 0 & \cdots & U(w_{2k}) \\ \cdots & \cdots & \cdots & \cdots \\ U(w_{k1}) & U(w_{k2}) & \cdots & 0 \end{pmatrix}, \ k \times k.$$

Bookstein "Thin-Plate" Splines

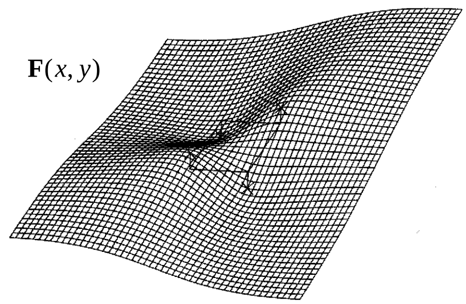
• kernel function $U(r)$ is principal solution of biharmonic equation that arises in elasticity theory of thin plates: $$\Delta^2 U(r) = \nabla^4 U(r) = \delta(r).$$

• 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: $U(r)$

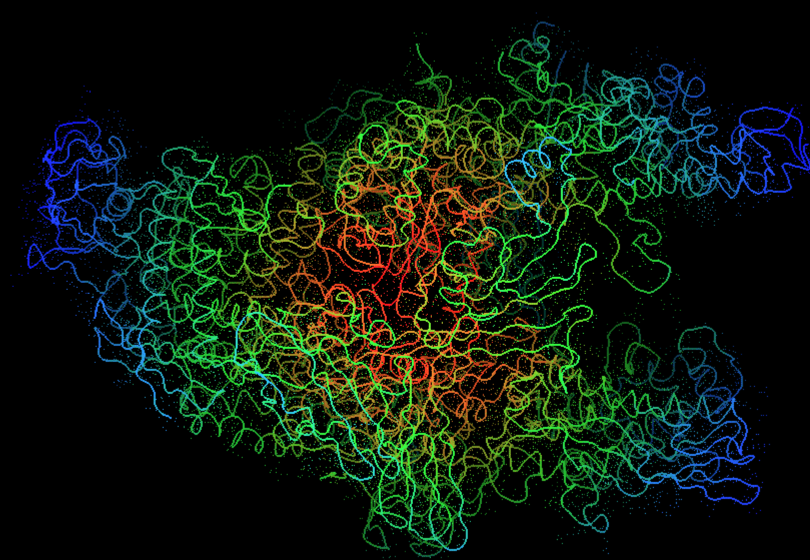


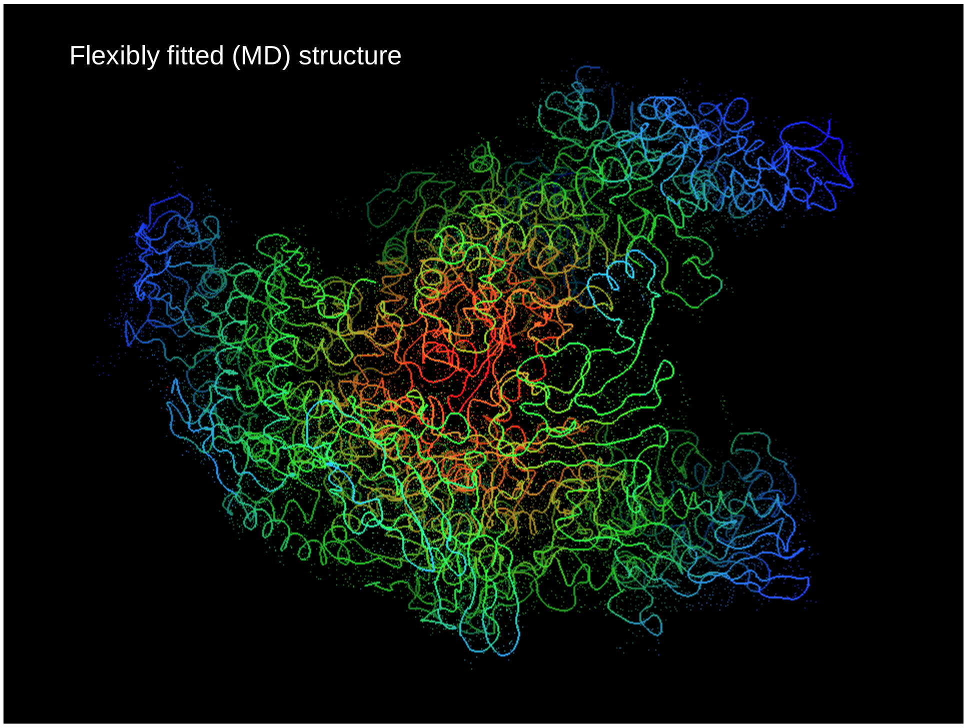
$\mathbf{F}(x, y)$

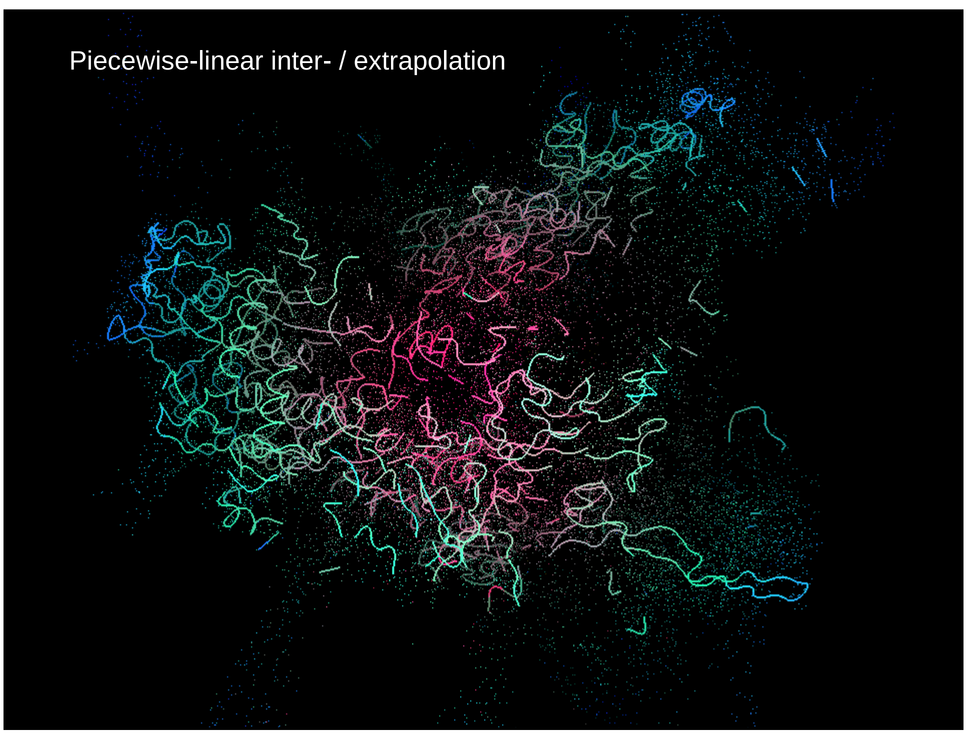


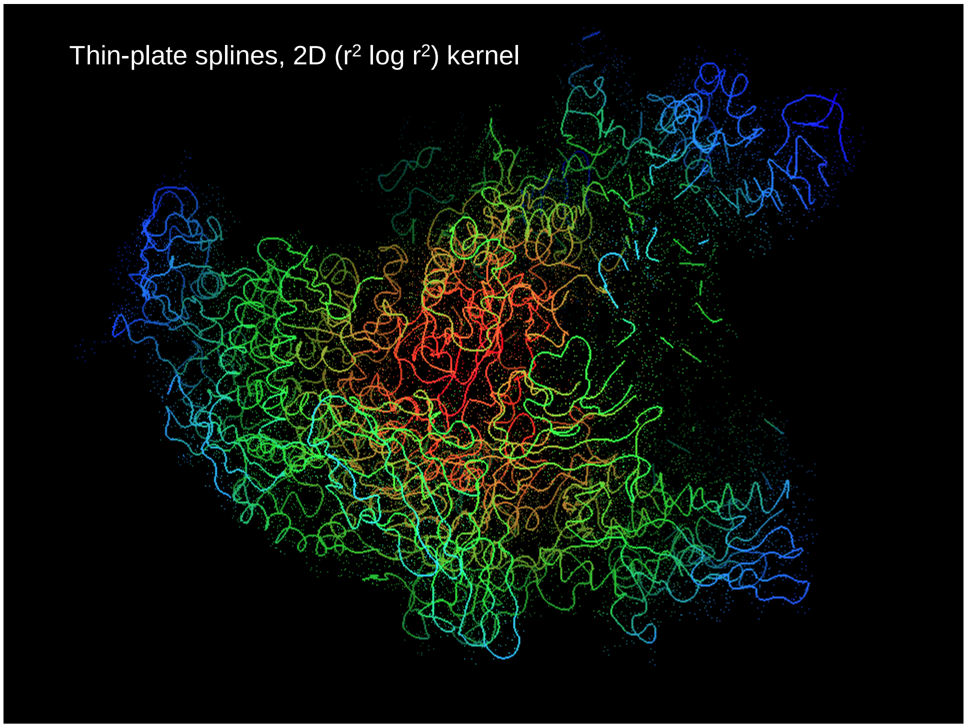
• we are interested mainly in 3D case but will also consider 2D (differentiable).

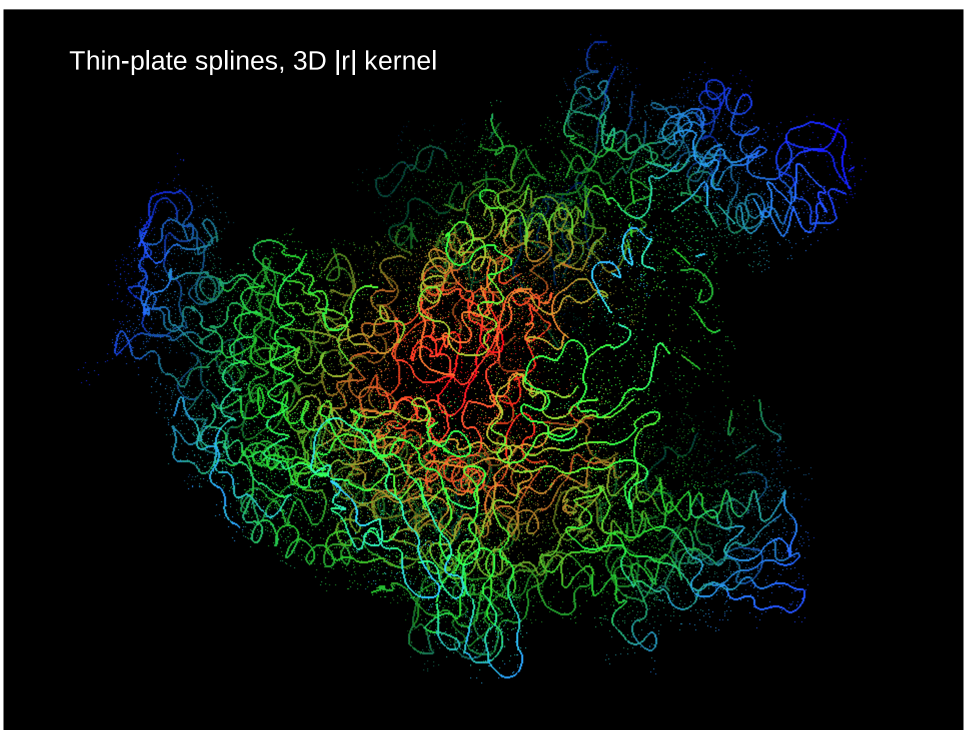
Taq RNAP x-tal structure




Flexibly fitted (MD) structure


Piecewise-linear inter- / extrapolation


Thin-plate splines, 2D ($r^2 \log r^2$) kernel


Thin-plate splines, 3D |r| kernel

Summary

Reduced (vector quantization) representations are useful for a variety of applications:

- Rigid-body docking.
- (Fast computation of forces and torques for haptic devices - S. Birmanns).
- Flexible fitting with molecular dynamics.
- Estimation of displacement vector fields.

(Non-linear) Interpolation is a viable alternative to MD in flexible fitting if stereochemical quality is optimized after morphing.

Interpolation allows displacements of vectors to be interpolated to full space, useful in Normal Modes Analysis (F. Tama, P. Chacón).

Availability: Situs 2.2

Acknowledgements

Pablo Chacón, Julio Kovacs, Stefan Birmanns.

Klaus Schulten, UIUC
Ron Milligan, TSRI
Edward H. Egelman, U Virginia
Joachim Frank, Wadsworth Center
Seth Darst, Rockefeller U

La Jolla Interfaces in Science Program
Grants from NIH to W.W. and Charles Brooks, III (MMTSB)