Point Cloud Registration with Anchor Point Matching

Willy Wriggers, Stefan Birmanns
biomachina.org

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.
1999-2007: 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)
Martinetz & Schulten (1993)

Digital Signal Processing,
Speech and Image Compression,
Topology-Representing Network.

Encode data (in $\mathbb{R}^{d\times 3}$) using a finite set $\{w_j\}_{j=1,\ldots,k}$ of codebook vectors. Delaunay triangulation divides $\mathbb{R}^2$ into $k$ Voronoi polyhedra (“receptive fields”):

![Diagram of Delaunay triangulation and Voronoi polyhedra](image)

Minimize encoding distortion error:

$$E = \sum_{i \in \text{atmos. voxels}} \left| V_i - W_{j(i)} \right|^2$$

**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 $k$ of vectors $k$: variability is minimal (“quality” of coarse-grained representation).
Single-Molecule Rigid-Body Docking

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

$$\Delta_I = \left( \frac{1}{k} \sum_{j=1}^{k} \left| w_{I(j)}^{(b)} - w_{j}^{(b)} \right|^2 \right)^{1/2}$$

- Find optimal $I$ by direct enumeration of the $k!$ cases (minimum of $\Delta_I$).

Application Example: Decorated MT

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 ($\Delta_r$).
- Overlap between both data sets:

  Correlation coefficient:

  $$C_{\text{corr}} = \frac{\sum_{x,y} h_{x,y} \cdot I_{x,y}}{\left(\sum_{x,y} h_{x,y}^2\right)^{1/2} \left(\sum_{x,y} I_{x,y}^2\right)^{1/2}}$$

ndc 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.

<table>
<thead>
<tr>
<th>$\Delta_r$</th>
<th>$C_{\text{corr}}$</th>
<th>$I$ (permutation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 4.446</td>
<td>0.053</td>
<td>(6, 3, 2, 1, 5, 4, 7)</td>
</tr>
<tr>
<td>2. 6.407</td>
<td>0.054</td>
<td>(5, 4, 3, 2, 1, 6, 7)</td>
</tr>
<tr>
<td>3. 7.612</td>
<td>0.167</td>
<td>(6, 2, 3, 4, 5, 1, 7)</td>
</tr>
<tr>
<td>4. 7.612</td>
<td>0.167</td>
<td>(6, 2, 3, 4, 5, 1, 7)</td>
</tr>
<tr>
<td>5. 7.088</td>
<td>0.088</td>
<td>(3, 1, 4, 5, 2, 6, 7)</td>
</tr>
<tr>
<td>6. 7.222</td>
<td>0.088</td>
<td>(4, 5, 6, 7, 1, 2, 3)</td>
</tr>
<tr>
<td>7. 9.008</td>
<td>0.092</td>
<td>(1, 6, 4, 5, 2, 3, 7)</td>
</tr>
<tr>
<td>8. 9.111</td>
<td>0.094</td>
<td>(2, 5, 1, 4, 3, 6, 7)</td>
</tr>
<tr>
<td>9. 9.244</td>
<td>0.105</td>
<td>(2, 5, 1, 4, 3, 6, 7)</td>
</tr>
<tr>
<td>10. 8.258</td>
<td>0.081</td>
<td>(1, 6, 4, 5, 2, 3, 7)</td>
</tr>
<tr>
<td>11. 8.546</td>
<td>0.094</td>
<td>(4, 5, 6, 7, 1, 2, 3)</td>
</tr>
<tr>
<td>12. 8.546</td>
<td>0.094</td>
<td>(4, 5, 6, 7, 1, 2, 3)</td>
</tr>
<tr>
<td>13. 8.682</td>
<td>0.087</td>
<td>(3, 1, 2, 4, 5, 6, 7)</td>
</tr>
<tr>
<td>14. 8.682</td>
<td>0.087</td>
<td>(3, 1, 2, 4, 5, 6, 7)</td>
</tr>
<tr>
<td>15. 8.682</td>
<td>0.087</td>
<td>(3, 1, 2, 4, 5, 6, 7)</td>
</tr>
<tr>
<td>16. 8.682</td>
<td>0.087</td>
<td>(3, 1, 2, 4, 5, 6, 7)</td>
</tr>
<tr>
<td>17. 8.682</td>
<td>0.087</td>
<td>(3, 1, 2, 4, 5, 6, 7)</td>
</tr>
<tr>
<td>18. 8.682</td>
<td>0.087</td>
<td>(3, 1, 2, 4, 5, 6, 7)</td>
</tr>
<tr>
<td>19. 8.682</td>
<td>0.087</td>
<td>(3, 1, 2, 4, 5, 6, 7)</td>
</tr>
<tr>
<td>20. 8.682</td>
<td>0.087</td>
<td>(3, 1, 2, 4, 5, 6, 7)</td>
</tr>
</tbody>
</table>

For a fixed $k$, codebook rmsd is more stringent criterion than correlation coefficient!
Performance (I)

Dependence of accuracy on resolution (simulated EM map, automatic assignment of $k$ from $3 \leq k \leq 9$) with Situs qrange tool.

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 $k$?


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.
qrange (Situs 1.x)

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

Limitations:
• Original $k \rightarrow k$ algorithm for limited $3 \leq k \leq 9$ works best for single molecules, not for matching subunits to larger densities.

• *matchpoint* in Situs 2.5 allows $k \rightarrow h \neq k$ matching

Anchor Point Registration: *matchpoint*

Anchor Point Registration
New in Situs 2.5: *matchpoint*

• $k \to h \neq k$ 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.
Clarification of Paradigm Shift in Coarse–Graining Currently Under Way

<table>
<thead>
<tr>
<th>Old style: Situs 1.x, qrange</th>
<th>Current style: Situs 2.5, matchpoint, qplasty</th>
</tr>
</thead>
<tbody>
<tr>
<td>rigid body docking only</td>
<td>rigid-body docking and flexible docking (see below)</td>
</tr>
<tr>
<td>restricted to single molecule matching, i.e. ( h = k )</td>
<td>OK to dock subunits into larger EM maps of assemblies, i.e. ( h &gt; k )</td>
</tr>
<tr>
<td>limited range ( 3 \leq h \leq 9 ) explored automatically by exhaustive enumeration of ( k! ) possibilities</td>
<td>no restriction on number of points (tree pruning in matchpoint handles larger number efficiently)</td>
</tr>
<tr>
<td>number of points typically selected from this limited range by using lowest variability criterion (justification: see slide “Performance II” above)</td>
<td>number of points now typically estimated based on number of resolved features (volume / EM resolution analysis in previous slide)</td>
</tr>
</tbody>
</table>

Take-Home Messages

Rigid body docking precision about one order of magnitude above the nominal EM (and SAXS) resolution

situs.biomachina.org
(UNIX command-line tools)

sculptor.biomachina.org
(GUI-based program)