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,

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

Delaunay triangulation divides $\mathbb{R}^3$ into Voronoi polyhedra (“receptive fields”):

$$V_i = \left\{ v \in \mathbb{R}^3 : \|v - w_j\| \leq \|v - w_i\| \forall j \right\}$$

Fig. 3: Partitioning of two-dimensional space ($n=2$) into $k=16$ cells. All input vectors in cell $C_i$ will be quantized as the code vector $w_j$. The shapes of the various cells can be very different.
Linde, Buzo, Gray (LBG) Algorithm

Encoding Distortion Error:

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

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

\[ \Delta w_r(t) \equiv w_r(t) - w_r(t-1) = -\epsilon \cdot \frac{\partial E}{\partial w_r} = \epsilon \cdot \sum_j \delta_{j(i)}(v_i-w_r)m_i. \]

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

\[ \Delta w_r(t) = \tilde{E} \cdot \delta_{j(i)} \cdot (v_i(t) - w_r). \]

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):

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

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

\[ \left\| v_i - w_{j0} \right\| \leq \left\| v_i - w_{j1} \right\| \leq \cdots \left\| v_i - w_{j(k-1)} \right\| \]

\[ s_r = 0 \quad s_r = 1 \quad s_r = k-1 \]

Note: \( \lambda \rightarrow 0 \): 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

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

Note: \( \lambda \rightarrow 0 : \tilde{E} \rightarrow E \). LBG algorithm.
\( \lambda \rightarrow \infty : \tilde{E} \) parabolic (single minimum). \( \lambda(t) \)
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**

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

$$\Delta_I = \frac{1}{k} \sum_{j=1}^k \left\| w^{(h)}_{I(j)} - w^{(l)}_j \right\|^2$$

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

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

Search for Conformations

Two possible ranking criteria:
- Codebook vector rms deviation ($\Delta_i$).
- Overlap between both data sets:

Voxel-Correlation coefficient:

$$C_M = \frac{\sum_{x,y,z} h_{x,y,z} \cdot I_{x,y,z}}{\sqrt{\left(\sum_{x,y,z} h_{x,y,z}^2\right) \cdot \left(\sum_{x,y,z} I_{x,y,z}^2\right)}}$$

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.

\[
\begin{array}{ccc}
\Delta I & C_{hl} & I \text{ (permutation)} \\
1. & 3.115 & 0.913 (7,5,1,6,4,2,3) \\
2. & 4.946 & 0.904 (2,3,5,7,4,6,1) \\
3. & 5.495 & 0.897 (6,1,3,4,2,7,5) \\
4. & 6.316 & 0.882 (5,7,4,3,1,2,6) \\
5. & 7.612 & 0.867 (5,7,1,4,6,3,2) \\
6. & 7.855 & 0.888 (3,2,4,1,5,6,7) \\
7. & 7.994 & 0.884 (1,6,4,5,3,7,2) \\
8. & 8.001 & 0.863 (6,1,4,3,5,2,7) \\
9. & 8.192 & 0.888 (2,6,4,3,1,7,5) \\
10. & 8.244 & 0.850 (7,1,6,2,1,3,4) \\
11. & 8.298 & 0.881 (2,6,7,5,1,3,4) \\
12. & 8.340 & 0.894 (6,2,4,1,3,7,5) \\
13. & 8.481 & 0.867 (3,4,6,2,1,5,7) \\
14. & 8.516 & 0.885 (2,3,4,6,1,7,5) \\
15. & 8.532 & 0.857 (7,1,5,4,3,6,2) \\
16. & 8.995 & 0.861 (6,1,5,7,4,3,2) \\
17. & 8.998 & 0.839 (3,4,5,7,1,2,6) \\
18. & 9.092 & 0.839 (3,2,5,4,7,1,6) \\
19. & 9.124 & 0.858 (7,5,3,2,4,1,6) \\
20. & 9.236 & 0.858 (1,6,5,7,4,2,3) \\
\end{array}
\]

For a fixed \(k\), codebook \(\text{rmsd}\) is more stringent criterion than correlation coefficient!

Performance (I)

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

orientations are stable:

\(+/-\ 5^\circ\) 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 \(~30\AA\)
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.

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

Xtal structure

\( w_j^{(h)} \)

EM / SAXS

low res. data

constrain centroids

molecular dynamics simulation
(X-PLOR)
Control (I): Simulation of G-Actin

Control (II): Lactoferrin (1LFG, 1LFH)
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
Note possible overfitting of domain IV!

flexible docking (5 vectors)

flexible docking (10 vectors, variable number per domain)
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:

unrestrained vectors + exp. and meth. uncertainty = distortion

skeleton + distance constraints = less distortion
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 \( C_i \) 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 \( j0 \):
   \( \forall j : T_{j0,j} = C_{j0,j} * (T_{j0,j} + 1) \)
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:
EM map of *E. coli* RNAP: Opalka *et al.*, PNAS 97:617, 2000

Hypothesis: Flexing of RNAP “jaws” encloses DNA
e) 15Å Gaussian kernel convolution (pdbblur)

f) difference map (subtract and vol/pad)

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 (pdbblur)

k) difference map (subtract 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: $F_i(x, y, z) = ax + by + xz + d$

$F_i(w_1) = f_{i,1}$, $F_i(w_2) = f_{i,2}$, $F_i(w_3) = f_{i,3}$, $F_i(w_4) = f_{i,4}$ (similar for $F_y, F_z$).

Cramer’s rule:

$$
\begin{align*}
\begin{vmatrix}
 f_{1,x} & w_{1,y} & w_{1,z} & 1 \\
 f_{2,x} & w_{2,y} & w_{2,z} & 1 \\
 f_{3,x} & w_{3,y} & w_{3,z} & 1 \\
 f_{4,x} & w_{4,y} & w_{4,z} & 1 \\
\end{vmatrix}
\end{align*}
= $$

$$
\begin{align*}
\begin{vmatrix}
 w_{1,x} & f_{1,y} & f_{1,z} & 1 \\
 w_{2,x} & f_{2,y} & f_{2,z} & 1 \\
 w_{3,x} & f_{3,y} & f_{3,z} & 1 \\
 w_{4,x} & f_{4,y} & f_{4,z} & 1 \\
\end{vmatrix}
\end{align*}

D = $$

(ii) Non-Linear Kernel Interpolation

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

Ansatz: $F_i(x, y, z) = a_i + a_{ix}x + a_{iy}y + a_{iz}z + \sum_{j=1}^{k} b_j \cdot U \left[ (w_j - (x, y, z)) \right]$  

$F_i(w_j) = f_{i,j} \text{, } \forall i$ (similar for $F_y, F_z$).

Solve:

$$
L^{-1}(f_{1,i}, \cdots, f_{k,i}, 0, 0, 0, 0) = (b_1, \cdots, b_k, a_i, a_{ix}, a_{iy}, a_{iz})^T,
$$

where

$$
L = \begin{pmatrix}
 Q^T & Q \\
 P & 0 \\
\end{pmatrix},
Q = \begin{pmatrix}
 1 & w_{1,x} & w_{1,y} & w_{1,z} \\
 \vdots & \vdots & \vdots & \vdots \\
 1 & w_{k,x} & w_{k,y} & w_{k,z} \\
\end{pmatrix},
,k \times 4,

P = \begin{pmatrix}
 0 & U(w_{1x}) & \cdots & U(w_{1z}) \\
 U(w_{1x}) & 0 & \cdots & U(w_{2x}) \\
 \vdots & \vdots & \ddots & \vdots \\
 U(w_{x1}) & U(w_{x2}) & \cdots & 0 \\
\end{pmatrix},
,k \times k.
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: $U(r)$

F(x, y)

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

**Bookstein “Thin-Plate” Splines**

**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)