quasi-Monte Carlo integration and molecular shape fingerprints

John D. MacCuish
Norah E. MacCuish,
Michael Hawrylycz, and Mitch Chapman

ACS San Francisco 2010
COMP
Outline

• Why Shape? Why Shape Fingerprints?
• Prior Work
• quasi-Monte Carlo Integration
• Shape Fingerprint Generation
• Experiment Examples
• Future Work
Why Shape?

• Shape is a factor in binding...
  • 3D QSAR
  • similarity searching
  • compound acquisition (shape diversity)
  • virtual screening...

• Can be confounding -- multiple binding sites, surface binding...

• One more tool in the toolbox...

• Boost to 2D
Why Shape Fingerprints?

- Compute shape comparisons efficiently on a large scale
- Efficient storage
- Simple.
Prior Work

• Shape as a mixture of spherical gaussians (atoms) (Grant and Pickup...’96)
• Fingerprints composed of key shapes, given the above method (Haigh, et al...’05).
• Ultrafast Shape Recognition: Ballester, Richards 2007
• ShaEP, Vainio, et al 2009
Monte Carlo Integration

- Approximate an integral (e.g., a 3D volume) by random sampling
- Approximate volume of odd shapes or manifolds that are analytically difficult or have no closed form solution.
- Better error convergence than grid sampling (Metropolis).
quasi-Monte Carlo Integration (QMC)

- In practice, quasi-randomly generated points have best error convergence in low dimensions.
- Beats uniform random sampling (e.g., pseudo, dart throwing, etc.)
- QMC became popular in early-90s in computer graphics (Shirley, ’91) mid-90s in finance - options/futures pricing (Morokoff, Caflisch, ’95)
Approximate Volume

grid

pseudo-random

quasi-random
(Hammersley sequence
generated points)

Error related to the number of points and the dimension

Error related to just number of points

Error related to just number of points
Approximation Error

• grid -- for error $\varepsilon$, need $1/\varepsilon^d$ grid points -- error convergence exponential in the dimension

• pseudo -- $1/\sqrt{n}$ convergence

• quasi -- $1/n$ convergence in practice in low dimensions

• quasi-random fewest points needed for equivalent approximation error.
Approximating the Volume of a molecule

• E.g., CPK with van der Waals radii -- Set of intersecting spheres

• Find a suitable bounding region for molecules

• Point sample the bounding region and tally up the points either inside or outside of the atom spheres.
Approximating the Volume of a molecule

- Bounding Volume times (Points Inside)/(All Points) \(\approx\) molecular volume.

- Sphere or scalene ellipsoid as a bounding region reduces total overall points.
Volume Percent Error

10240 pseudo-points
95% below 4.5% error

20480 pseudo-points
95% below 3.6% error

10240 quasi-points
95% below 2% error

20480 quasi-points
95% below 1.3% error

1357 Molecules
4961 low energy conformations
100-550 MW flexible, inflexible balloon conformations
relative point error
Fingerprint Generation

Preliminaries

• Need a bounding region that covers the conformational space of the database of small molecule conformations.

• Cube? Sphere? Scalene Ellipsoid.

• Need orientation and alignment.
Fewer points

Don’t need these points

Don’t need these points

Fewer points
Fingerprint Generation

1. Generate a sample quasi-point set bounding region centered and fixed at (0,0,0); Sphere of ~11 Å radius. This is fixed, done once.

2. Mean center atom centers point set of a conformation to be fingerprinted.

3. Find sample points in the atom spheres (or function or your choice, e.g., solvent accessible surface ...)

4. Find the principal axes for a shape using the sampled points and the atom centers with SVD. (Adjust for SVD sign ambiguity (Bro, et al, 2008).)
Fingerprint Generation

5. Rotate atom centers point set to Principal Axes (PA)

6. Find points in PA configuration and create fingerprint with 4 orientations -- points in molecule ‘1’, points out ‘0’

7. Subfingerprints, each the length of the number of points.
Molecule in a Volume of quasi-Random points
Alignment Overlay
Shape Fingerprints

- 4 fingerprints per shape. e.g., $10,240 \times 4 \text{ bits} = 5.12 \text{ Kbytes}$.

- Number and density of points determines resolution, speed of comparison, and storage size.

- Number of bits on is small, typically 3-10% with CPK model. Fingerprints compress significantly.
Fingerprint Similarity

- Choose 1 subFP of Shape FP A and compare it with all 4 subFPs of Shape FP B.
- The largest similarity comparison is chosen.
- Bit difference per subFP is small in a given FP
- So all 4 choose 2 comparisons are not needed -- all-pairs maximums differ by very little if at all.
Fingerprint Similarity

- Inherent slight asymmetry of alignment of point sets of different sizes.

- Aligning B to A’s principal axes, vs aligning A to B’s principal axes.

- Try this on your favorite method...

- Use bit string similarity measure of your choice... Tanimoto, Ochiai (Cosine), ..., Tversky,... Baroni-Urbani/Bush...
Performance

• Fingerprint generation: ~?K conformers per hour on a single ? gigahertz core with ~10K points.

• Fingerprint Tanimoto comparisons: ~?K per hour on a single ? gigahertz core with ~10K points.

• Similarity searching: $M = 8$, $N = ?$, on 8 cores in an hour.

• Clustering: $N = ?$ NXN matrix with 8 cores in 24 hours?
Space!

- 5.12 Kbytes per fingerprint -- ~90% compression
- 2 million conformations = 1+ GBs
- Space and CPU improving all of the time...
FP Experiments

- Clustering simple set of 2D compounds, expanded into a set of low energy, multi-conformations in the thousands.

- Similarity searching with crystal structure and a larger compound database composed of low energy, multi-conformations, 1 million conformations.
Future Work

- Other shape functions
- Other low discrepancy point sets in different distributions -- density where it is needed.
- Speed (time) and compression (space)
- More practical applications
Software Acknowledgments

- Jmol
- Balloon
- OpenBabel
- R
Acknowledgments
(Humans)

- Jeremy Yang (UNM)
- Rajarshi Guha (NIH)