Ultrafast shape recognition for similarity search in molecular databases

Pedro J. Ballester & W. Graham Richards

Presented by Christian Lessig
lessig@dgp.toronto.edu
Overview

• Objective
Overview

• Objective
Overview

• Objective
Overview

• Objective

Courtesy Ballester & Richards 2007 [1]
Overview

• Objective
Previous Work

• Similarity score based on volume overlap
  – ROCS [2]
• Geometric descriptors
  – Bemis & Kuntz [3], Nilakantant et al. [4]
• Ray tracing reflection histograms
  – Shape Signatures [5]
• Dimensionality reduction
  – EShape3D in MOE
Previous Work

• Similarity score based on volume overlap
• Geometric descriptors
  – Bemis & Kuntz [3], Nilakantan et al. [4]
• Ray tracing reflection histograms
  – Shape Signatures [5] 1.6 y
• Dimensionality reduction
  – EShape3D in MOE 1.2 y
Ultrafast Shape Recognition

- Consider the distances $d_{ki}^k$ between the heavy atoms $A_i$ and reference points $p^k$
Ultrafast Shape Recognition

• Consider the distances $d_i^k$ between the heavy atoms $A_i$ and reference points $p^k$
Ultrafast Shape Recognition

- Consider the distances $d_{i}^{k}$ between the heavy atoms $A_{i}$ and reference points $p^{k}$.
Ultrafast Shape Recognition

• Consider the distances $d_{i}^{k}$ between the heavy atoms $A_{i}$ and reference points $p^{k}$

• Characterise a molecule by the moments $\mu_{k}$ of the distributions $\left\{ d_{i}^{k} \right\}_{i=1}^{N}$
Ultrafast Shape Recognition

- Consider the distances $d_{ik}$ between the heavy atoms $A_i$ and reference points $p_k$
- Characterise a molecule by the moments $\mu_k$ of the distributions $\{d_{ik}\}_{i=1}^N$
  + Translation invariant
  + Independent of the number of atoms in a molecule
  + Can be precomputed
  + Low dimensional
Ultrafast Shape Recognition

• Reference points
  – Centroid (ctd)
  – Atom closest to the centroid (cst)
  – Atom farthest away from centroid (fct)
  – Atom farthest away from fct (ftf)
Ultrafast Shape Recognition

• Reference points (ctd, cst, fct, ftf)
• Moments
  - Mean: measure for molecule size
  - Variance: measure for molecule compactness
  - Skewness: measure for molecule symmetry
Ultrafast Shape Recognition

• Reference points (ctd, cst, fct, ftf)
• Moments (mean, variance, skewness)

⇒ 12 dimensional representation of molecule
Similarity Measure

• Similarity measure $S_{kl}$

$$S_{kl} = \left(1 + \frac{1}{12} \sum_{i=1}^{12} | m_{ki}^k - m_{li}^l | \right)^{-1}$$

with

$$M_k = \{ m_{ki}^k \}_{i=1}^{12} = \{ \mu_1^{ctd}, \mu_2^{ctd}, \mu_3^{ctd}, \mu_1^{cst}, \ldots, \mu_3^{ftf} \}$$
Similarity Measure

• Similarity measure $S_{kl}$

$$S_{kl} = \left( 1 + \frac{1}{12} \sum_{i=1}^{12} | m^k_i - m^l_i | \right)^{-1}$$

with

$$M_k = \{ m^k_i \}_{i=1}^{12} = \{ \mu_1^{ctd}, \mu_2^{ctd}, \mu_3^{ctd}, \mu_1^{cst}, \ldots, \mu_3^{ftf} \}$$
Workflow
Workflow
Workflow
Workflow
Workflow
Experiments

- Database with 2,433,493 compounds
- For each query molecule computation of five most similar molecules in the database
- Evaluation by
  - Visual inspection and comparison with EShape3D
  - Performance comparison
Experiments
Experiments

Courtesy Ballester & Richards 2007 [1]
Experiments
Experiments

USR

Eshape3D
Experiments

Comparisons per Second (Normalized)

ROCS | Shape Signatures | EShape3D | USR
Experiments
Summary

• Drug lead identification using 3D shape
• Low dimensional representation of molecules using distance moments
• Comparison of molecules in the low-dimensional space using $L_1$ norm
• High accuracy for finding similar molecules
• More than three orders of magnitude faster than most efficient previous method
Comments

• Why four reference points?
• Why these reference points?
• Why $L_1$ norm?
• Why only first three moments?
• Why is projection $\mathbb{R}^3 \times N \rightarrow \mathbb{R}^{12}$ distance preserving?
• What are the problems which can result from the non-uniqueness of $\mathbb{R}^3 \times N \rightarrow \mathbb{R}^{12}$?
Comments

• They could not find a molecule where USR did not perform well?

• USR considers always the full molecule.

• No discussion of related work in CAGD, graphics, robotics, image processing, ... [6].

• Use USR as preprocessing step and filter results using chemical properties.
References


Slides available at
www.dgp.toronto.edu/people/lessig/talks/
Moment

• Moments $\mu_k$ of a probability distribution $P(x)$

\[
\mu_k = \left\langle (x - a)^k \right\rangle = \sum_{i=1}^{n} (x - a)^n P(x)
\]

• First moment: mean
• Second moment: variance
• Third moment: skewness
• Fourth moment: kurtosis
Overview

• Objective
  – Identification of novel drug leads

• Premises
  – Template: Known molecule with bioactivity
  – Very large databases with ligands
  – Three-dimensional molecular shape is most discriminating factor for biological activity