R-NN Curves: A Method for Diversity Analysis and Cluster Identification

Rajarshi Guha, Debojyoti Dutta, Peter C. Jurs and Ting Chen

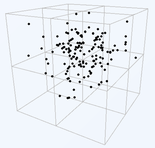
Department of Chemistry, Pennsylvania State University
Department of Computational Biology, University of Southern California

1 Diversity Analysis

- *What is it?*
  - *Understand the distribution of chemical structures in a chemical space as well as with respect to themselves*
- *Why is it useful?*
  - *Compound acquisition*
  - *Lead hopping*
  - *Enhancing local regression techniques*

2 Approaches to Diversity Analysis

Cell based

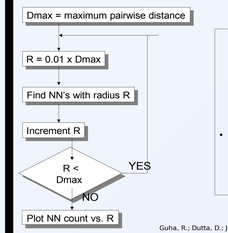


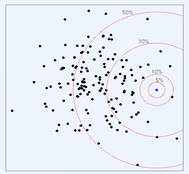
- Not useful for high dimensional data
- Choice of bin size is tricky

Distance based

- Generally requires pair-wise distance calculations
- Can be sped up using kD-trees

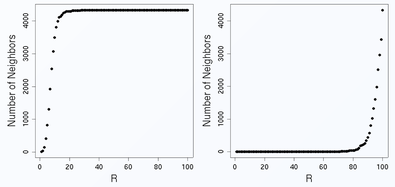
3 Generating R-NN Curves





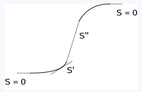
Guha, R.; Dutta, D.; Jurs, P.C.; Chen, T.; *J. Chem. Inf. Model.*, **2006**, *ASAP*

4 R-NN Curve Examples



5 Characterizing R-NN Curves

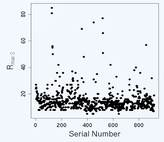
- Rigorous Approach
  - Fit a sigmoid curve and use the fit parameters
- Simpler Approach
  - Determine the value of R where the lower tail transitions to the linear portion of the curve



- Determine slope at varying R
- Slope is maximal on the linearly increasing portion
- Find R for first occurrence of the *maximal* slope - $R_{max(S)}$
- Rapidly performed using a finite difference approach

6 Summarizing R-NN Curves

- Visual inspection is reasonable for a few curves $R_{max(S)}$
- Inspecting hundreds of curves is painful
- Summarize large datasets using $R_{max(S)}$ values



- *Points towards the top are located in sparse regions*
- *Points towards the bottom are located in dense regions*
- *Significant outliers are easily detected*

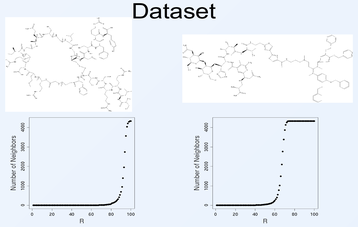
7 R-NN Curves : Kazius Dataset

- Dataset
  - *4337 molecules*
  - *Avg. MW = 240*
  - *Avg. Tanimoto = 0.21*
  - *Calculated MOE descriptors*
- Two significant outliers
- Descriptor space
  - Random 5-descriptor space was selected

- Defined outliers as
  - *R = 0.5 x Dmax*
  - *NN count < 10% of the dataset*
- Resulted in 2 outliers
  - *3904 had 2 NN's*
  - *1861 had 41 NN's*
- ***Subjective procedure***

Kazius, J.; McGuire, R.; Bursi, R.; *J. Med. Chem.*, **2005**, *48*, 312–320

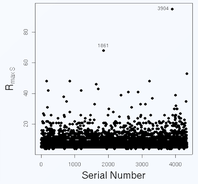
8 Outliers in the Kazius Dataset



9 Outliers in the Kazius Dataset

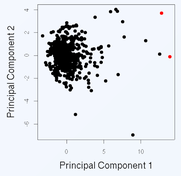
- 1861 & 3904 are immediately identifiable as outliers
- Not many significant outliers

One plot summarizes the location characteristics of all the molecules in the dataset



10 Alternative Methods?

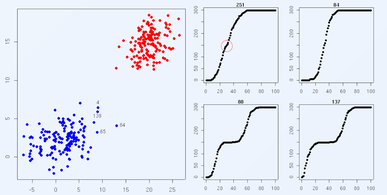
- R-NN curves essentially perform dimension reduction
- Alternative approach is to use PCA
- *Problem*
  - *Eigendecomposition via SVD is O(mn²)*
  - *Difficult to visualize more than 2 or 3 PC's at the same time*



11 R-NN Curves & Clusters?

- When clusters are present R-NN curves in general exhibit a *stepped* appearance
- Counting the number of steps identifies the number of clusters
- Problems
  - Steps may not be distinct
  - All the points in a dataset may not exhibit stepping

12 R-NN Curves for Clustered Data



Simulated Data

13 Counting the Steps

- Essentially a curve matching problem
- Start with a set of *canonical* sigmoid curves
- Match R-NN curves to the canonical set
  - How many R-NN curves do we look at?

- Standardize R-NN Curves
  - Procrustes transformation
- *Matching approaches*
  - Slope analysis
  - Hausdorff / Frechet distance
  - RMSE from distance matrix

14 Curve Matching – Distance Metrics

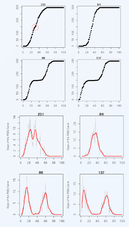
- Use techniques from the field of object recognition
- Hausdorff distance is a measure of the maximum distance between curves represented as *sets of points*
- Frechet distance takes into account the location and *ordering* of the points on the curves

Two dissimilar curves with small Hausdorff distance

Frechet, M., *Sur quelques points du calcul fonctionnel. Rendiconti del Circolo Mathematico di Palermo*, **1906**, *22*, 1–74.
http://www.cim.mcgill.ca/~stephane/cs507/Project.html

15 Curve Matching - Slopes

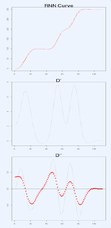
- The slope on a step is a minima
- Plotting slope vs. R generates peaks
  - 1 step : 2 peaks
  - 2 steps : 3 peaks
  - ……
- Identify the peaks
- Automated identification can select spurious peaks



16 Slope Analysis

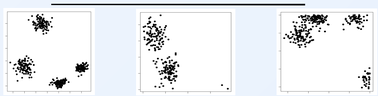
- Generate an R-NN curve and smooth it
- Evaluate the second derivative and smooth it
- Determine the number of roots of the second derivative, N_{root}

$$N_{peak} = [max(N_{root}) + 1]/2$$



17 Preliminary Results

- Simulated 2D data
- Predicted *k*, followed by k-means clustering
- Investigated similar values of *k*



k	ASW
4	0.61
3	0.74
5	0.70

k	ASW
3	0.44
2	0.65
4	0.47

k	ASW
4	0.64
3	0.48
5	0.56

ASW – average silhouette width, higher is better

18 Conclusions

- Applicable to arbitrary dimensions
- Can be extended to handle large datasets
- Summarizing a dataset does not require user-defined parameters
- Simple and intuitive approach to visualizing molecules in a chemical space
- Useful approach to identifying number of clusters
  - A number of techniques for curve matching are available
  - Need to achieve balance between specificity and speed
  - Not necessarily automatic