Scalable Partitioning & Exploration of Chemical Spaces Using Geometric Hashing

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

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

28th March, 2006

Outline

Chemical Data Mining

Common tasks

- Clustering
  - library design and selection
  - data partitioning
  - classification
- Virtual screening
- Substructure searching

Problems

- Datasets are large and are getting larger every day
- Feature space has high dimensionality and is usually correlated
- Many common algorithms are of high computational complexity

Chemical Data Mining

Common tasks

- Clustering
  - library design and selection
  - data partitioning
  - classification
- Virtual screening
- Substructure searching

Problems

- Datasets are large and are getting larger every day
- Feature space has high dimensionality and is usually correlated
- Many common algorithms are of high computational complexity

How Can We Efficiently Analyze Chemical Data?

- *Parallelize*
  - Throw more CPU's at the problem
  - Gets the answer but doesn't address the underlying problems
  - Lacks elegance
- *Approximation algorithms*
  - Transforms the problem to a space of lower dimensions
  - Parametric in nature
  - Error bounds of the approximation are desirable
- *Avoid the distance matrix*
  - Many utilize a pairwise distance matrix
  - The distance matrix for 50,000 structures will need 4GB of memory
  - Most distance matrix based algorithms are $O(N^2)$

Guha, R; Jurs, P.C.; "Applications of Spectral Clustering to Chemical Datasets", *Gordon Conference on Computer Aided Drug Design*, Tilton, NH, August 2005

How Can We Efficiently Analyze Chemical Data?

- *Parallelize*
  - Throw more CPU's at the problem
  - Gets the answer but doesn't address the underlying problems
  - Lacks elegance
- *Approximation algorithms*
  - Transforms the problem to a space of lower dimensions
  - Parametric in nature
  - Error bounds of the approximation are desirable
- *Avoid the distance matrix*
  - Many utilize a pairwise distance matrix
  - The distance matrix for 50,000 structures will need 4GB of memory
  - Most distance matrix based algorithms are $O(N^2)$

Guha, R; Jurs, P.C.; "Applications of Spectral Clustering to Chemical Datasets", *Gordon Conference on Computer Aided Drug Design*, Tilton, NH, August 2005

How Can We Efficiently Analyze Chemical Data?

- *Parallelize*
  - Throw more CPU's at the problem
  - Gets the answer but doesn't address the underlying problems
  - Lacks elegance
- *Approximation algorithms*
  - Transforms the problem to a space of lower dimensions
  - Parametric in nature
  - Error bounds of the approximation are desirable
- *Avoid the distance matrix*
  - Many utilize a pairwise distance matrix
  - The distance matrix for 50,000 structures will need 4GB of memory
  - Most distance matrix based algorithms are $O(N^2)$

Guha, R; Jurs, P.C.; "Applications of Spectral Clustering to Chemical Datasets", *Gordon Conference on Computer Aided Drug Design*, Tilton, NH, August 2005

How Can We Efficiently Analyze Chemical Data?

- *Parallelize*
  - Throw more CPU's at the problem
  - Gets the answer but doesn't answer the underlying problems
  - Lacks elegance
- *Approximation algorithms*
  - Transforms the problem to a space of lower dimensions
  - Parametric in nature
  - Error bounds of the approximation are desirable
- *Avoid the distance matrix*
  - Many algorithms utilize a pairwise distance matrix
  - The distance matrix for 50,000 structures will need 4GB of memory
  - Most distance matrix based algorithms are $O(N^2)$

Guha, R; Jurs, P.C.; "Applications of Spectral Clustering to Chemical Datasets", *Gordon Conference on Computer Aided Drug Design*, Tilton, NH, August 2005

Locality Sensitive Hashing

What is it?

Very fast, approximate nearest neighbor detection algorithm

Why is this significant?

- Fast
  - Theoretically sublinear
  - Avoids the distance matrix computation
  - Not hampered by high dimensional data
- Approximate
  - Based on a radius value
  - The nearest neighbors in the radius are reported in a probabilistic manner

Datar, M. *et al.*; "Locality Sensitive Hashing Scheme Based on p-Stable Distributions", in *Proceedings of the 20th Annual Symposium on Computational Geometry*, 2004

Locality Sensitive Hashing

What is it?

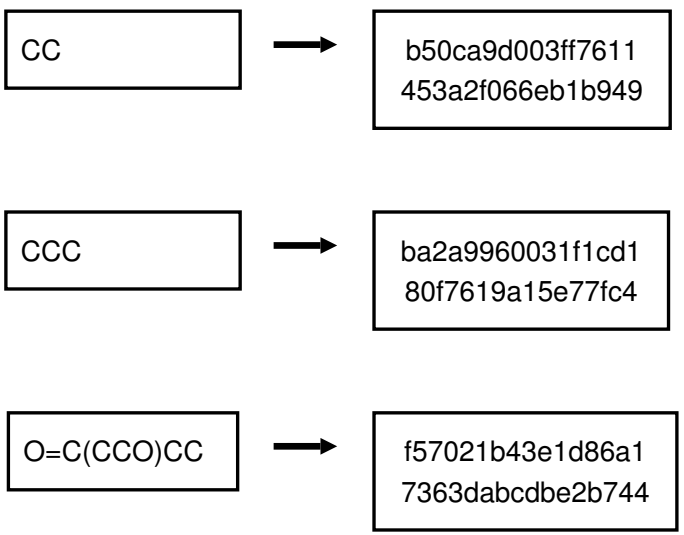
Very fast, approximate nearest neighbor detection algorithm

Why is this significant?

- Fast
  - Theoretically sublinear
  - Avoids the distance matrix computation
  - Not hampered by high dimensional data
- Approximate
  - Based on a radius value
  - The nearest neighbors in the radius are reported in a probabilistic manner

Datar, M. *et al.*; "Locality Sensitive Hashing Scheme Based on p-Stable Distributions", in *Proceedings of the 20th Annual Symposium on Computational Geometry*, 2004

How Does It Work?

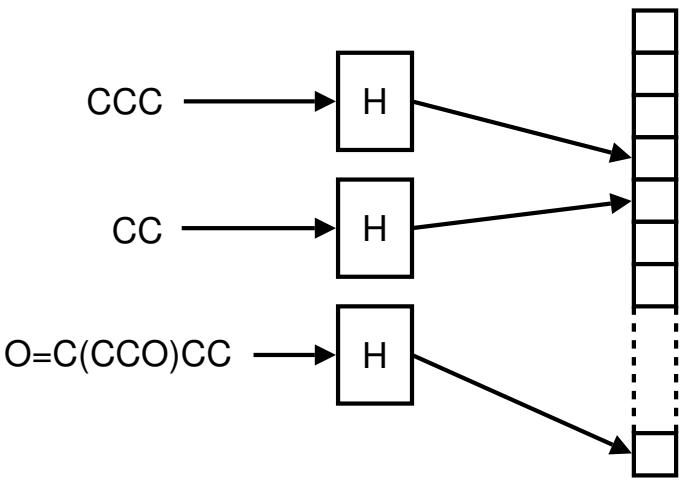


Hash Functions

- A hash function, H, takes an object, O, and tries to convert it to a unique value
- O can be a string or numeric
- A good hash function will avoid too many collisions
- $H(O_1) = H(O_2)$ does not always imply $O_1 = O_2$

Knuth, D. .E., *The Art of Computer Programming*, Vol. 3, Addison Wesley, 1973

How Does It Work?



Hash Tables

- A hash table uses a hash function to convert an object O to a location in the table
- A good hash function should be used so that similar objects do not land in the same location

Knuth, D. .E., *The Art of Computer Programming*, Vol. 3, Addison Wesley, 1973

How Does It Work?

Overview

- High dimensional observations are projected onto *random lines*
- The lines are chopped into equal width segments, h_i
- Observations are assigned a hash value depending on which segment it is projected onto

Locality Sensitive Hash Functions

- If two points q and v are close they will hash to the same value with high probability
- If they are distant they should collide with low probability
- Families of hash functions can be defined using s-stable distributions which ensures the locality property

How Does It Work?

Overview

- High dimensional observations are projected onto *random lines*
- The lines are chopped into equal width segments, h_i
- Observations are assigned a hash value depending on which segment it is projected onto

Locality Sensitive Hash Functions

- If two points q and v are close they will hash to the same value with high probability
- If they are distant they should collide with low probability
- Families of hash functions can be defined using s-stable distributions which ensures the locality property

How Does It Work?

User Parameters

- R - the radius of interest
- k, L - hashing parameters
- p - the probability that $g(q) = g(v)$, where

$$g(q) = (h_1(q), \ldots, h_k(q))$$

- In practice the parameters k and L are estimated empirically
- R and p are specified by the user

What Does All This Mean?

We can use the LSH algorithm to determine the neighbors of a point q that are within a radius R from q, with a probability p

What Is It Good For?

- Rapidly partition a dataset into more manageable chunks
- kNN regression and classification
- Clustering
- Intuitive diversity analysis of chemical spaces

LSH gives us a framework which can be applied to nearest neighbor *type* problems for very large datasets

It can be used as an analysis method in itself as well as perform the role of a preprocessor for subsequent methods

What Is It Good For?

- Rapidly partition a dataset into more manageable chunks
- kNN regression and classification
- Clustering
- Intuitive diversity analysis of chemical spaces

LSH gives us a framework which can be applied to nearest neighbor *type* problems for very large datasets

It can be used as an analysis method in itself as well as perform the role of a preprocessor for subsequent methods

Datasets & Descriptors

Datasets

		No. of Descriptors	
Dataset	No. of Molecules	Full	Reduced
Kazius	4337	142	20
NCI-AIDS	42613	144	55
NCI-3D	249071	163	53

Kazius, J. et al.; *J. Med. Chem.*, **2005**, *48*, 312-320

`http://cactus.nci.nih.gov/DownLoad/AID2DA99.sdz`

Voigt, J. et al.; *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 702-712

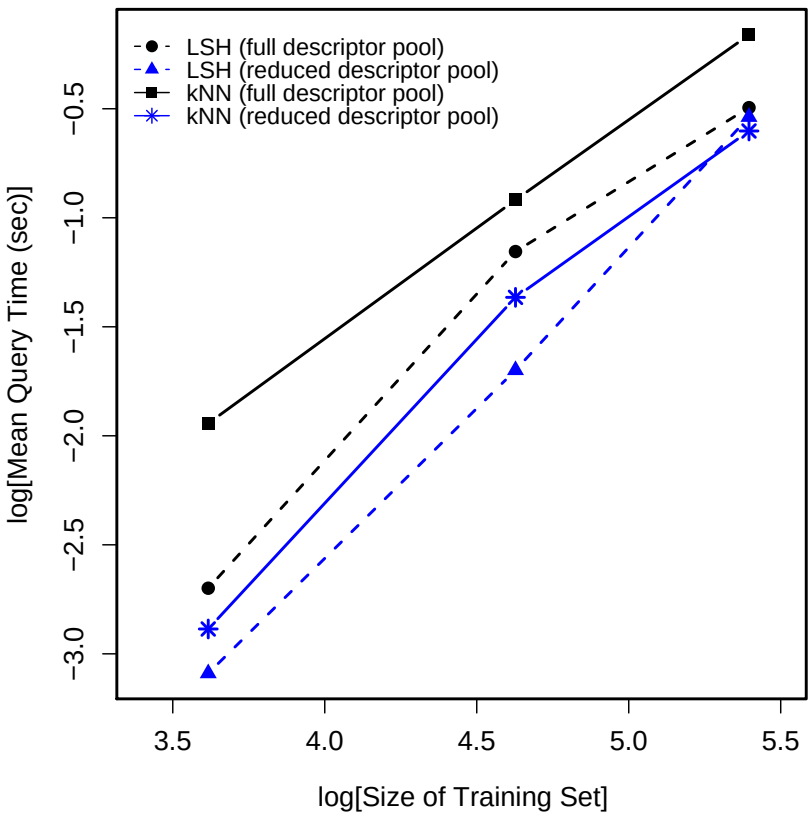
Datasets & Descriptors

Descriptors

- Topological and geometric descriptors were evaluated
- A reduced descriptor pool was also investigated
- No scaling was performed

	Maximum Distance		Mean distance	
Dataset	Full	Reduced	Full	Reduced
Kazius	507252	507144	1793	1656
NCI-AIDS	626517	626438	3712	3490
NCI-3D	14285759	3013170	10810	11784

Timing Results - *Effect of Dataset Size*

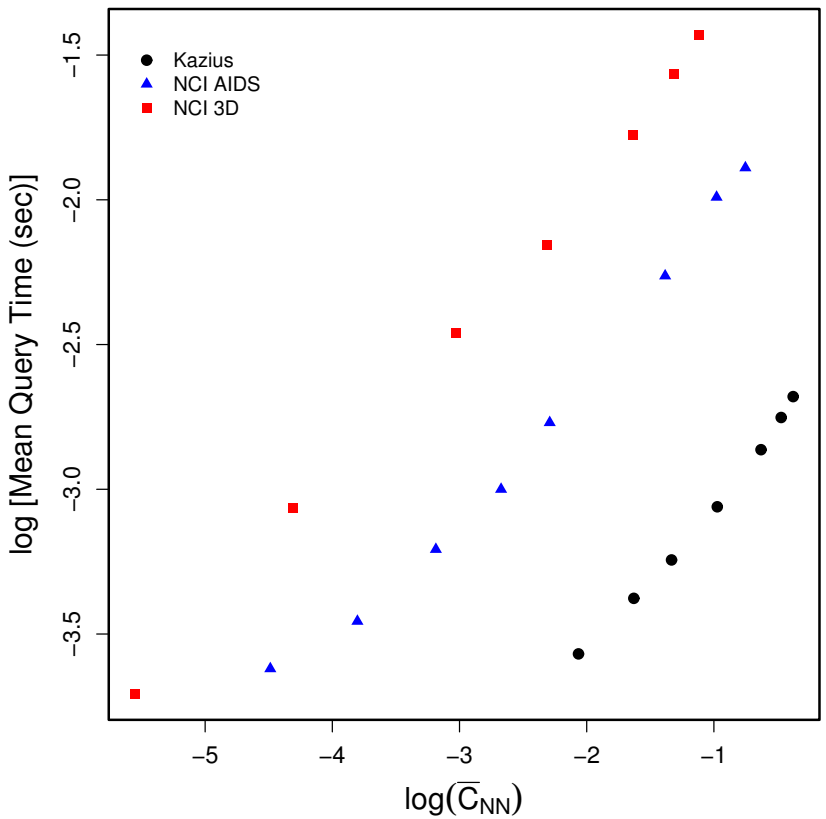


Mean Query Times

- For each dataset the first 200 points were used as the query set
- For the LSH runs, $R = 0.01 \times D_{max}$
- kNN runs were performed with $k = 200$

Dutta, D.; Guha, R.; Jurs, P.C.; Chen, T., *J. Chem. Inf. Model.*, **2006**, *46*, 321-333

Timing Results - *Effect of NN Count*

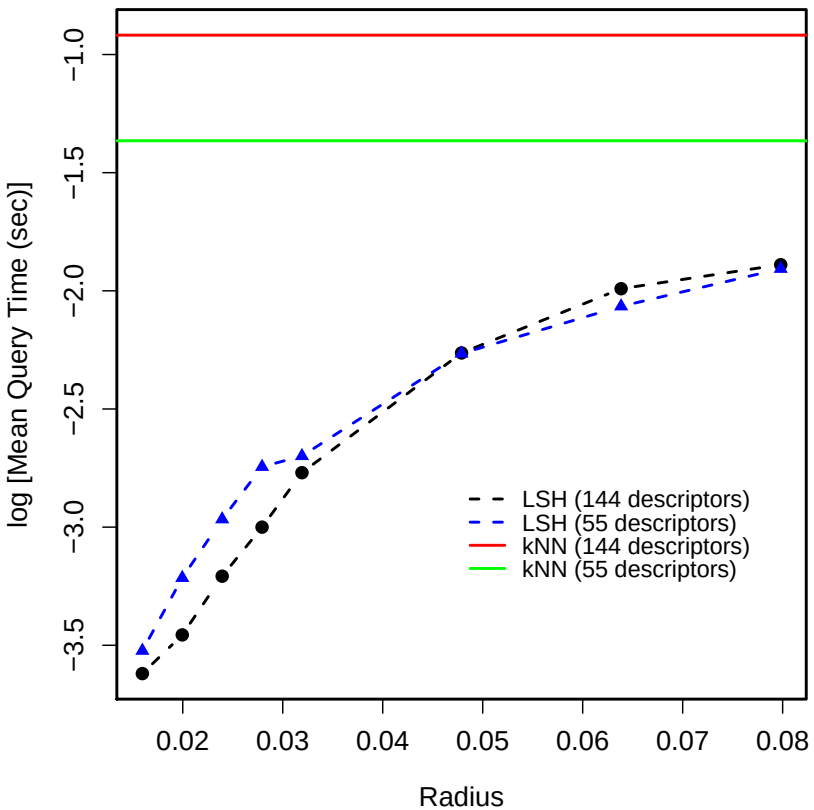


Normalized Query Counts

- $\bar{C}_{NN} = \frac{1}{N_t}\frac{1}{N_q}\sum_{i=1}^{N_q} N_{NN,i}$
- Memory access is significant for large datasets
- C_{NN} takes into account the number of neighbors with respect to the size of the dataset

Dutta, D.; Guha, R.; Jurs, P.C.; Chen, T., *J. Chem. Inf. Model.*, **2006**, *46*, 321–333

Timing Results *(NCI-AIDS Dataset)*

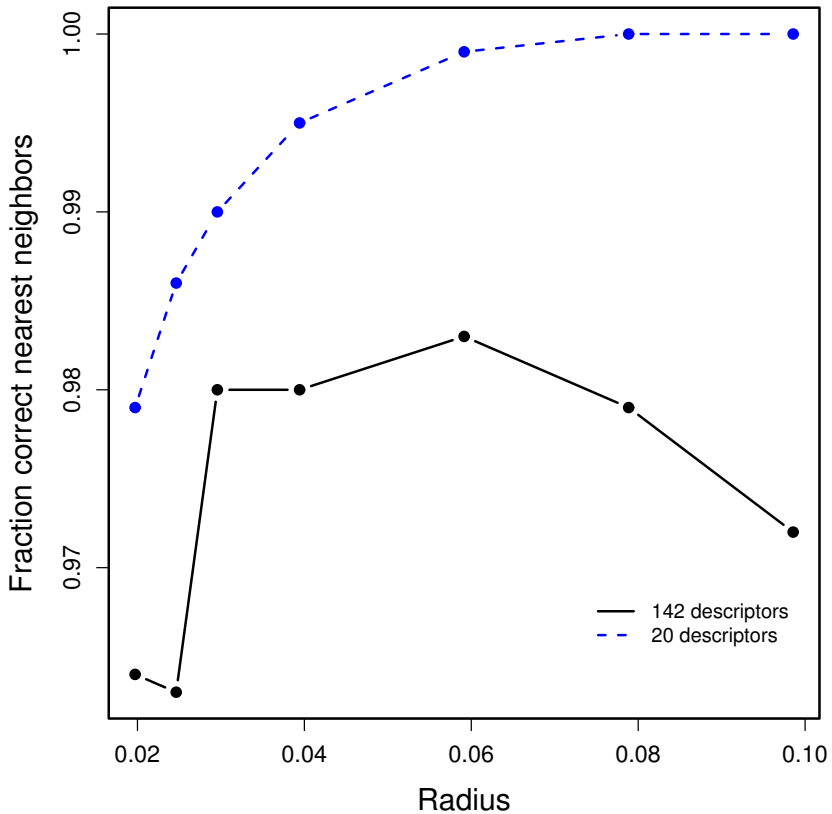


Influence of Radius

- kNN results were obtained with $k = 200$
- For the highest radius a few thousand neighbors were detected
- Query times for LSH are *at least* an order of magnitude lower than kNN
- Highlights the use of LSH as a partitioning tool

Dutta, D.; Guha, R.; Jurs, P.C.; Chen, T., *J. Chem. Inf. Model.*, **2006**, *46*, 321–333

Accuracy Results *(Kazius Dataset)*



Number of correct R-NN's detected versus radius, for the Kazius dataset

- Accuracy was obtained with respect to a linear scan
- Over all datasets studied, accuracy was never lower than 94%
- By increasing the probability parameter we can ensure higher accuracy (at the cost of time efficiency)

Dutta, D.; Guha, R.; Jurs, P.C.; Chen, T., *J. Chem. Inf. Model.*, **2006**, *46*, 321–333

Using LSH for Diversity Analysis - *Dataset Sparsity*

The *R-NN Curve* Algorithm

$D_{max} \leftarrow$ max. pairwise distance
R $\leftarrow 0.01 \times D_{max}$
while $R \leq D_{max}$ **do**
  Find NN's within radius R
  increment R
end while

- Depends on the descriptor space
- NN counts *vs.* radius plots are discriminatory

The R-NN curve for an observation from a dense region of the descriptor space

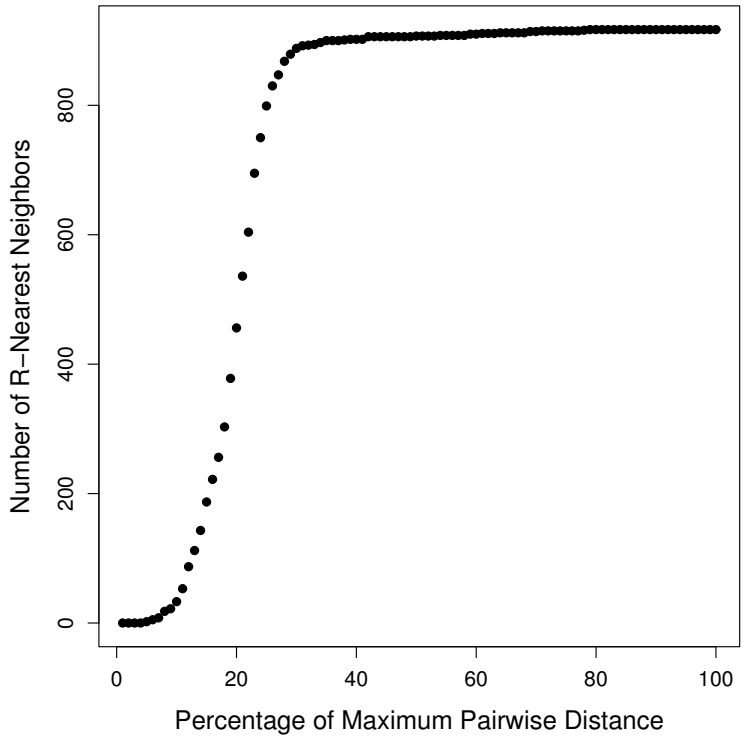
Guha, R., Dutta, D., Jurs, P.C. and Chen, T., *submitted*

Using LSH for Diversity Analysis - *Dataset Sparsity*

The *R-NN Curve* Algorithm

$D_{max} \leftarrow$ max. pairwise distance
R $\leftarrow 0.01 \times D_{max}$
while $R \leq D_{max}$ **do**
  Find NN's within radius R
  increment R
end while

- Depends on the descriptor space
- NN counts *vs.* radius plots are discriminatory



The R-NN curve for an observation from a dense region of the descriptor space

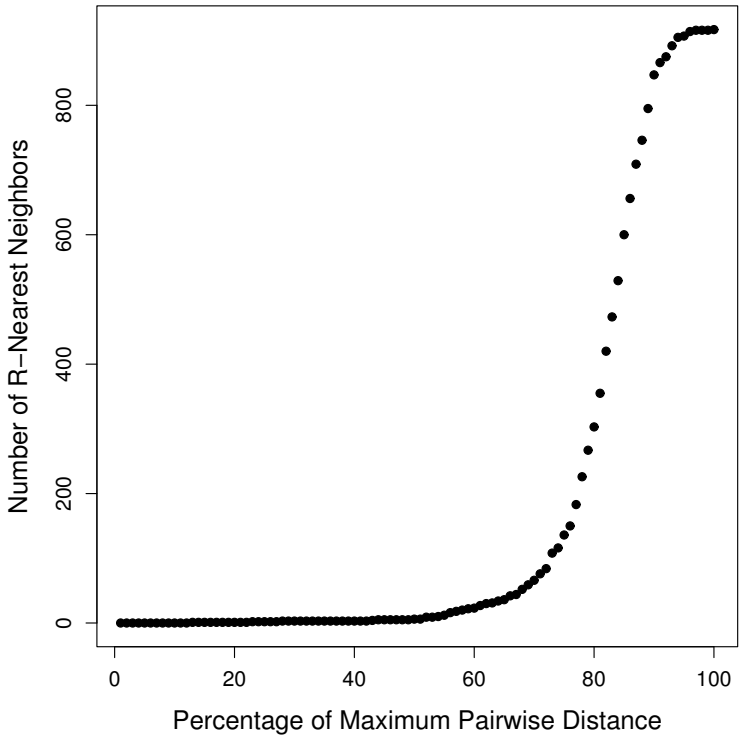
Guha, R., Dutta, D., Jurs, P.C. and Chen, T., *submitted*

Using LSH for Diversity Analysis - Dataset Sparsity

The *R-NN Curve* Algorithm

$D_{max} \leftarrow$ max. pairwise distance
R $\leftarrow 0.01 \times D_{max}$
while $R \leq D_{max}$ **do**
  Find NN's within radius R
  increment R
end while

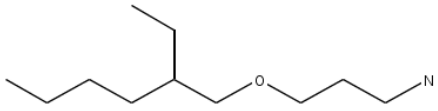
- Depends on the descriptor space
- NN counts *vs.* radius plots are discriminatory



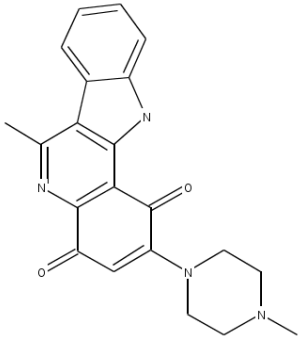
The R-NN curve for an observation from a sparse region of the descriptor space

Guha, R., Dutta, D., Jurs, P.C. and Chen, T., *submitted*

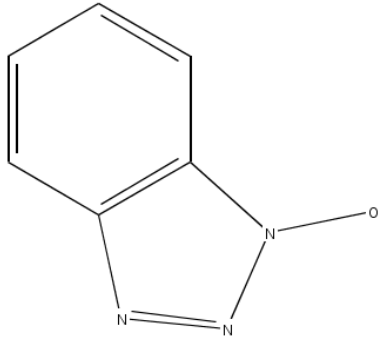
Using LSH for Diversity Analysis



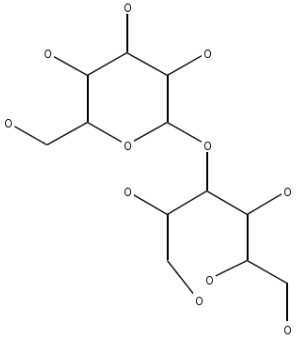
2170 (4281)



98 (4281)



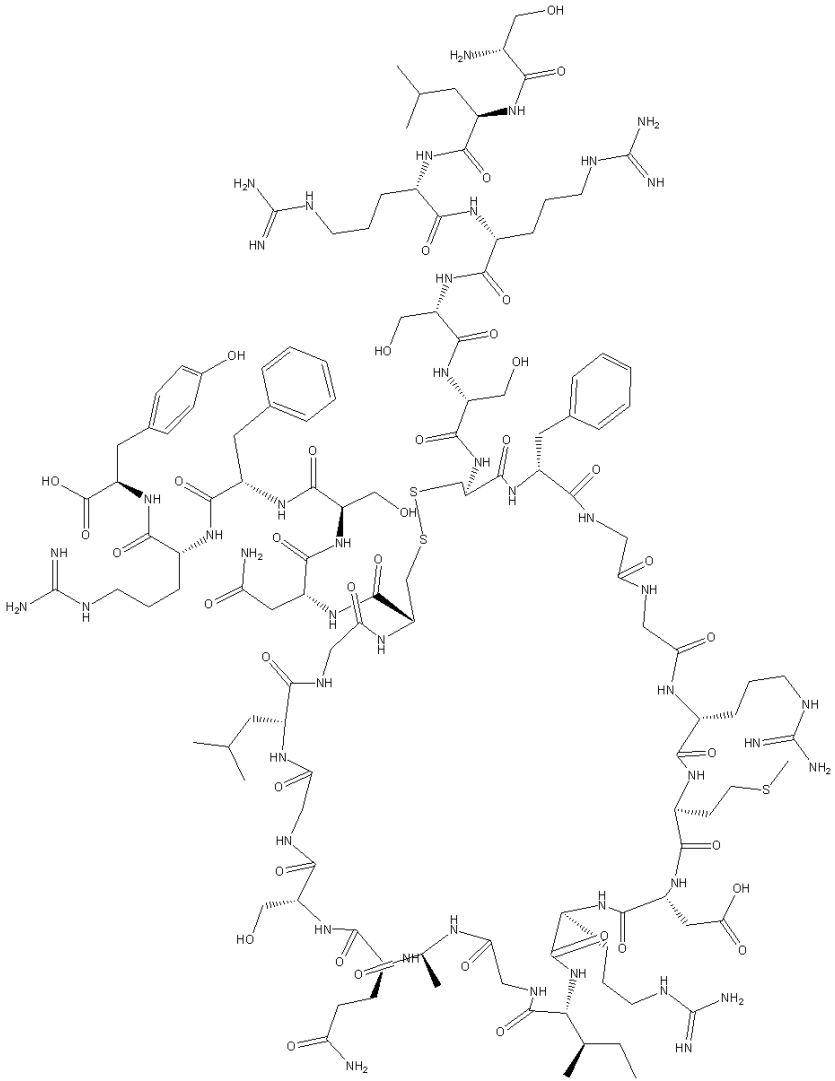
4224 (4279)



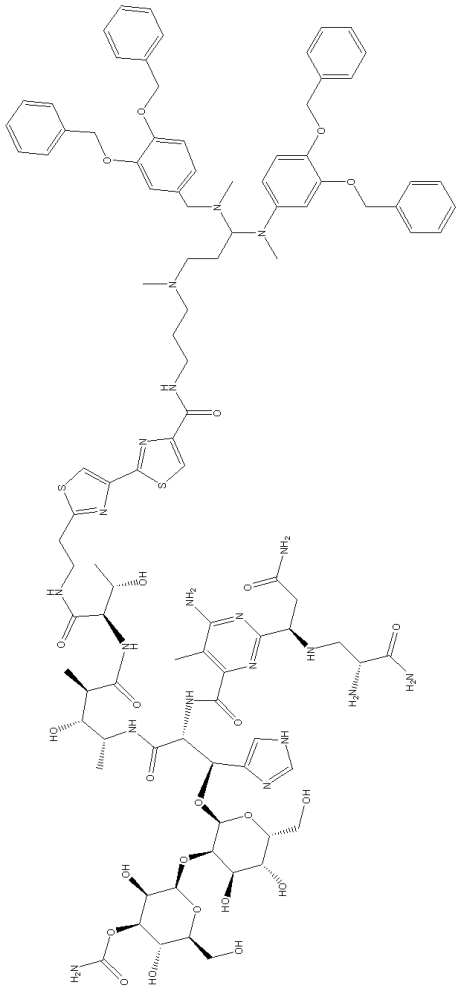
3679 (4279)

- Kazius dataset
- Representative compounds from the dense region
- R = 0.1% of D_{max}

Using LSH for Diversity Analysis - *Outliers*

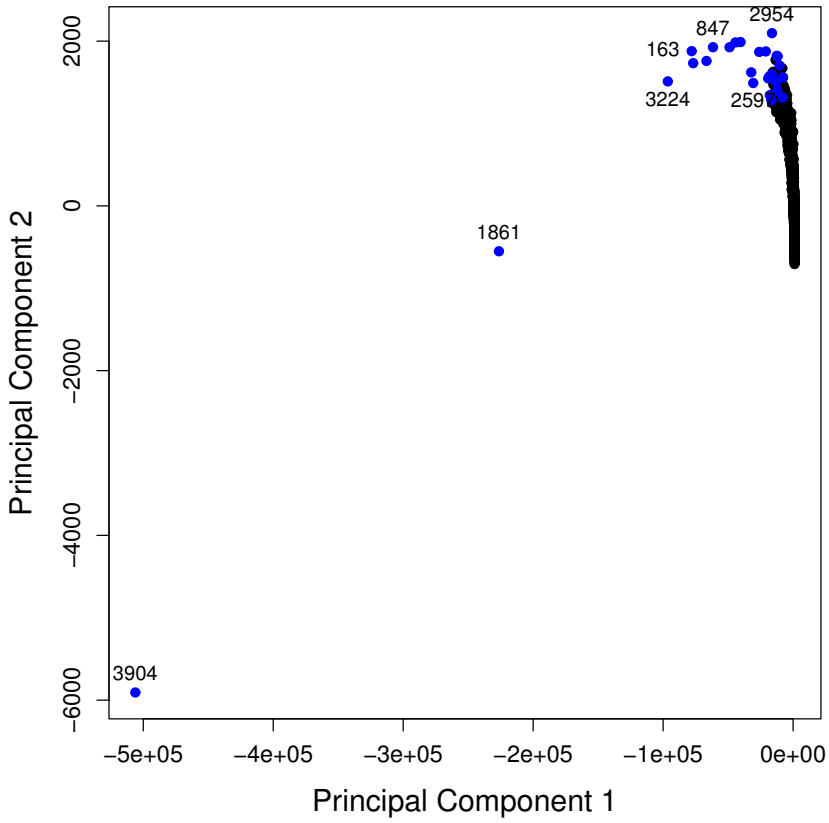


3904

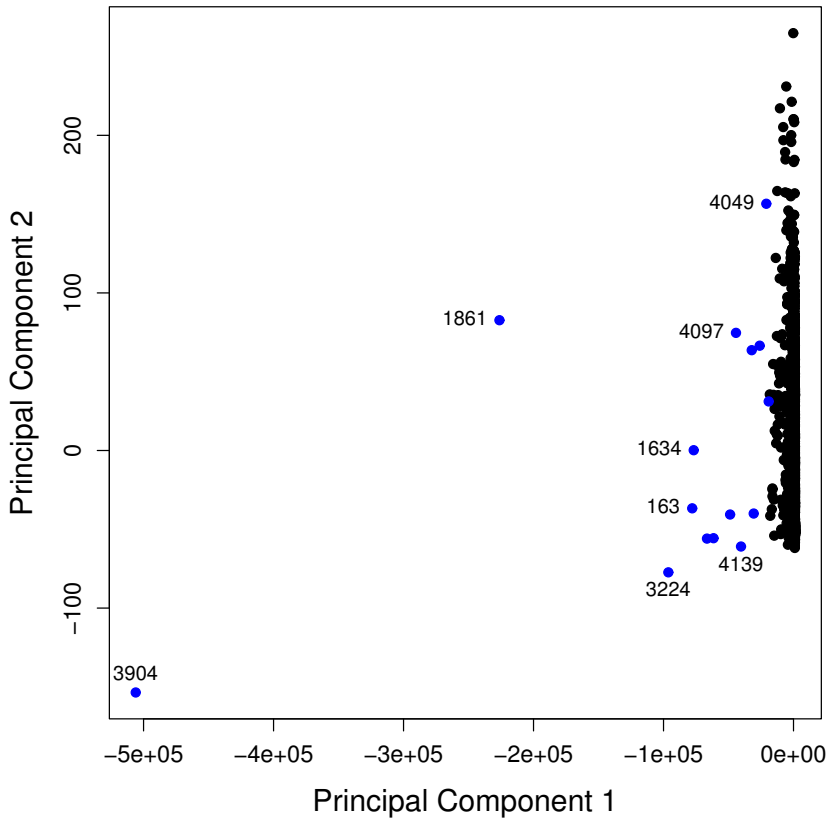


1861

Using LSH for Diversity Analysis - *Comparison*



Whole descriptor pool



Reduced descriptor pool

Future Work

- Is data reduction really necessary?
- How will scaling affect the performance and accuracy of the LSH algorithm?
- Quantification of sparsity for descriptor spaces as a whole
- Use nearest neighbor modeling as a component in a screening protocol
- Approximate substructure searches

Summary

Conclusions

- In practice LSH outperforms traditional kNN by up to 3 orders of magnitude
- Though the LSH is approximate by design, it exhibits high accuracy
- Can be used for a variety of preprocessing and analytical tasks

Acknowledgements

- Prof. Pyotr Indyk
- A. Andoni

Summary

Conclusions

- In practice LSH outperforms traditional kNN by up to 3 orders of magnitude
- Though the LSH is approximate by design, it exhibits high accuracy
- Can be used for a variety of preprocessing and analytical tasks

Acknowledgements

- Prof. Pyotr Indyk
- A. Andoni