Chemical Space: Modeling Exploration & Understanding

Rajarshi Guha

School of Informatics
Indiana University

16th August, 2006
Outline

1. Overview

2. Modeling & Algorithms
   - Aspects of QSAR Modeling
   - Exploring Chemical Space
   - Adding Meaning to Chemical Information

3. Tool Development
   - CDK
   - R
Outline

1. Overview

2. Modeling & Algorithms
   - Aspects of QSAR Modeling
   - Exploring Chemical Space
   - Adding Meaning to Chemical Information

3. Tool Development
   - CDK
   - R
Goals of Cheminformatics Research at IU

- Extend the state of the art in cheminformatics
- Extract chemistry, not just $R^2$, $q^2$ et al.
- Provide intuitive and efficient ways to handle chemical information
- Supply expertise to bench chemists and non-computational chemists
Statistical Modeling of Chemical Information

- QSAR model development
  - Lazy regression
  - Ensemble descriptor selection
  - Wavelet-based spectral descriptors
- Interpretation of QSAR models
- Measuring model applicability
R-NN curves
  - Outlier detection
  - Cluster cardinality
LSH based applications
Ensemble descriptor selection
Interpretation techniques
- Contributions to the CDK
- Packages for R
  - rcdk
  - spe
  - fingerprint
  - spectral clustering
  - rpubchem (to come)
- Automated QSAR pipeline (collaboration with P&G)
- Development of workflows
Outline

1. Overview

2. Modeling & Algorithms
   - Aspects of QSAR Modeling
   - Exploring Chemical Space
   - Adding Meaning to Chemical Information

3. Tool Development
   - CDK
   - R
QSAR Model Interpretation

Why interpret?

- Predictive ability is useful for screening
- Interpretation provides extra value
- Interpretability might be more useful than predictive ability
- Interpretability depends on modeling technique and descriptors involved

Predictive ability is useful for screening

Interpretation provides extra value

Interpretability might be more useful than predictive ability

Interpretability depends on modeling technique and descriptors involved
The trade-offs in interpretation

- Interpretability is usually a trade-off with accuracy
- OLS models are easily interpretable, not always accurate
- CNN models are usually black boxes, more accurate
- Some methods (RF) lie in between

Interpretability vs. Accuracy

Linear Model vs. Non-linear Model
Detailed Interpretation Methods

**OLS Models**
- Build a *good* model
- Run it through PLS
- The X-loadings indicate which descriptors are important
  - magnitude lets us rank them
  - sign lets us indicate the nature of their effects
- Allows us to identify effects of descriptors on a *molecule-wise* basis

**CNN Models**
- Analogous to PLS based interpretations
- *Linearizes* the CNN
- Information is lost
- Resultant interpretations match the corresponding interpretations for an OLS model quite well

The CNN interpretation uses significant approximations and does not make full use of biases.

Develop interpretation techniques for other methods, RF in particular.

Ensemble descriptor selection to choose a subset that is *simultaneously* good for multiple model types.

Use these methods on real datasets.
QSAR Model Applicability

**Model Validation**
- Goal is to test the reliability of the model
- Ensures that the model is not due to chance factors
- Based on dataset used to develop the model

**Model Applicability**
- Goal is to test the applicability of the model to new compounds
- Tells us: The model will predict the activity well (or not)
- Similar to confidence measures
What is Model Applicability

Question?
How will a model perform when faced with molecules that it has not been trained on or validated with?

Aspects
- Similarity to the TSET?
- Can we consider a *global* chemistry space?
- Structural or statistical similarity?
- Quantitative or qualitative?
Define Model Performance

Performance is measured by prediction residuals. The model performs well on a new molecule if it predicts its activity with low residual error.

Correlate ‘X’ With Performance

- ‘X’ could be similarity between a query molecule and the original training set
- ‘X’ could be derived from a cluster membership approach
- Alternatively, predict performance itself

Nearest Neighbor Methods

- Traditional $k$NN methods are simple, fast, intuitive
- Applications in
  - regression & classification
  - diversity analysis
- Can be misleading if the nearest neighbor is far away
- $R$-NN methods may be more suitable
The question . . .

Does the variation of nearest neighbor count with radius allow us to characterize the location of a query point in a dataset?

The answer . . . R-NN curves

\[ D_{\text{max}} \leftarrow \text{max pairwise distance} \]
\[ \text{for molecule in dataset do} \]
\[ R \leftarrow 0.01 \times D_{\text{max}} \]
\[ \text{while } R \leq D_{\text{max}} \text{ do} \]
\[ \text{Find NN’s within radius } R \]
\[ \text{Increment } R \]
\[ \text{end while} \]
\[ \text{end for} \]

Plot NN count vs. \( R \)

The question . . .

Does the variation of nearest neighbor count with radius allow us to characterize the location of a query point in a dataset?

The answer . . . $R$-NN curves

\[ D_{\text{max}} \leftarrow \text{max pairwise distance} \]

\[ \text{for molecule in dataset do} \]
\[ \quad R \leftarrow 0.01 \times D_{\text{max}} \]
\[ \quad \text{while } R \leq D_{\text{max}} \text{ do} \]
\[ \quad \quad \text{Find NN’s within radius } R \]
\[ \quad \quad \text{Increment } R \]
\[ \quad \text{end while} \]
\[ \text{end for} \]

Plot NN count vs. $R$

Characterizing $R$-NN Curves
A single plot identifies the location characteristics of all the molecules

**R-NN Curves and Clusters**

*R-NN curves are indicative of the number of clusters*
**R-NN Curves and Clusters**

### Counting the steps
- Essentially a curve matching problem
- All points will not be indicative of the number of clusters
- Not applicable for concentric clusters

### Approaches
- Hausdorff / Fréchet distance
  - requires *canonical* curves
- RMSE from distance matrix
- Slope analysis
R-NN Curves and Their Slopes

Smoothed first derivative of the \( R-NN \) Curves
R-NN Curves and Their Slopes

- Identifying peaks identifies the number of clusters
- Automated picking can identify spurious peaks
Slope Analysis of R-NN Curves

Procedure

```plaintext
for i in molecules do
    Evaluate R-NN curve
    F ← smoothed R-NN curve
    Evaluate F''
    Smooth F''
    N_{root,i} ← no. of roots of F''
end for
N_{cluster} = [\max(N_{root}) + 1]/2
```

Possible improvements

- Sample from the collection of R-NN curves
- Improve handling of concentric clusters
Preliminary Results

- Simulated 2D data
- Predicted $k$, followed by kmeans clustering using $k$
- Investigated similar values of $k$

<table>
<thead>
<tr>
<th>$k$</th>
<th>ASW</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.61</td>
</tr>
<tr>
<td>3</td>
<td>0.74</td>
</tr>
<tr>
<td>5</td>
<td>0.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>ASW</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.44</td>
</tr>
<tr>
<td>2</td>
<td>0.65</td>
</tr>
<tr>
<td>4</td>
<td>0.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>ASW</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>0.48</td>
</tr>
<tr>
<td>5</td>
<td>0.56</td>
</tr>
</tbody>
</table>

ASW - average silhouette width, higher is better; $k$ - number of clusters
What is an ontology?

- Defines a controlled vocabulary (keywords)
- Defines a set of relationships between them
- Allows for
  - meaning to be added to data and algorithms
  - automated inference of relationships
- An example is the Gene Ontology
Ontologies in Chemistry

What’s available?

- No really comprehensive ontology available
- Some work is on to add chemistry semantics to biology-related ontologies

What can we do?

- Start small - focus on one area of chemistry, descriptors
- The CDK currently provides an ontology for implemented descriptors
- Vocabulary includes
  - author
  - literature reference
  - class (molecular, atomic, bond)
  - type (geometric, electronic, ...)
How Can We Use Ontologies?

Allow interoperability of software

- Different program use different naming schemes
- Programs may be open- or close-source
- Annotation via dictionaries allows us to make conclusions regarding descriptors, suggest similar descriptors etc.

Utilize expertise from chemists

- Some descriptors are useful for certain properties
- Chemists who use models can *tag* descriptors as useful for a property
- Over time the tagging can be indicative of the utility of certain descriptors
Outline

1 Overview

2 Modeling & Algorithms
   - Aspects of QSAR Modeling
   - Exploring Chemical Space
   - Adding Meaning to Chemical Information

3 Tool Development
   - CDK
   - R
The Chemistry Development Kit

What does it do?

- Reads multiple file formats
- Calculates molecular descriptors (in progress)
  - Topologicals (Chi, Kappa, ...)
  - Geometric (Gravitational indices, MI, ...)
  - Hybrid (CPSA)
  - Global (BCUT, WHIM, RDF, ...)
- Provides access to statistical engines (R & Weka)
- Fingerprint calculation, 2D structure diagrams
- Kabsch alignment
- 3D coordinates and force fields (in progress)

Contributions

Main areas of contributions

- Descriptor framework
- QSAR modeling framework (interface to R)
- Web service functionality

Other areas include

- Numerical surface areas
- Rigid alignment
- Descriptor ontology
- Build system
- QA, debugging, support

http://almost.cubic.uni-koeln.de/cdk/cdk.top
What is R?

- An open-source statistical \textit{environment} for
  - model development
  - algorithm prototyping

- Open source version of Splus
  - Splus code runs (mostly) unchanged on R

- Provides a wide array of mathematical and statistical functionality
  - Linear models (OLS, robust regression, GLM, PLS)
  - Neural networks, random forests, SOM, SVM
  - Clustering methods (kmeans, agnes, pam, ...)
  - Optimization routines
  - Database interfaces

- When working with chemical data it would be nice to have access to cheminformatics functionality inside R
Cheminformatics and R

What is R?

- An open-source statistical environment for
  - model development
  - algorithm prototyping
- Open source version of Splus
  - Splus code runs (mostly) unchanged on R
- Provides a wide array of mathematical and statistical functionality
  - Linear models (OLS, robust regression, GLM, PLS)
  - Neural networks, random forests, SOM, SVM
  - Clustering methods (kmeans, agnes, pam, ...)
  - Optimization routines
  - Database interfaces
- When working with chemical data it would be nice to have access to cheminformatics functionality inside R
Some Cheminformatics Related Packages

Connecting R and the CDK
- R can access Java code via the SJava package
- This allows us to use CDK functionality within R
- The rcdk package provides user friendly wrappers to CDK classes and methods
- The result is that we can stay inside R and handle molecules directly

Fingerprints
- The fingerprint package handles binary fingerprint data from CDK and MOE
- Calculates similarity matrices using the Tanimoto metric
- Converts binary fingerprints to Euclidean vectors

Guha, R., CDK News, 2005, 2, 2–6
R Miscellanea

R as a Web Service

- A number of packages are available to access R via the web
- An effort is also underway to provide an explicit SOAP interface
- Very simple to access a remote R process via RServe
  - Currently under investigation as our statistical backend for workflows
Summary

Broadly focused on 2 areas . . .

- **Modeling**
  - Predictive model development
  - Model interpretation and applicability

- **Algorithm development**
  - Exploring nearest neighbor methods
  - Descriptor selection and interpretation
  - Dictionaries and ontologies

Underlying motivation is the extraction of chemistry from the numbers and making it available

Collaborations . . .

- More brains are useful
- Always on the lookout for data
Summary

Broadly focused on 2 areas . . .

- **Modeling**
  - Predictive model development
  - Model interpretation and applicability

- **Algorithm development**
  - Exploring nearest neighbor methods
  - Descriptor selection and interpretation
  - Dictionaries and ontologies

- Underlying motivation is the extraction of *chemistry* from the numbers and making it available

Collaborations . . .

- More brains are useful
- Always on the lookout for data
Broadly focused on 2 areas . . .

- Modeling
  - Predictive model development
  - Model interpretation and applicability
- Algorithm development
  - Exploring nearest neighbor methods
  - Descriptor selection and interpretation
  - Dictionaries and ontologies

Underlying motivation is the extraction of chemistry from the numbers and making it available

Collaborations . . .

- More brains are useful
- Always on the lookout for data