Integrating R with the CDK for QSAR modeling

R. Guha and P.C. Jurs

Department of Chemistry
Pennsylvania State University

September 1st, 2005

Outline

1. Introduction
2. CDK & QSAR Modeling
   - Descriptors
   - Working With R
   - An Application
3. Summary

What is the CDK?

The Chemistry Development Kit

- A Java framework for cheminformatics development
- Open-source
- Suitable for experimentation as well as application development

Useful Links

- The CDK home page: http://almost.cubic.uni-koeln.de/cdk/cdk_top/
- The CDK showcase website: http://www.chemistry-development-kit.org/
- Examples, tips and tricks: http://blue.chem.psu.edu/˜rajarshi/code/java/

Steinbeck, C. et al., *J. Chem. Inf. Comput. Sci.*, **2003**, *43*(2), 495–500

Philosophy

- Source is available to the user
- Functionality is not restricted to what the core developers want and user contributions are welcome
- Provide reliable core cheminformatics functionality
- Clear documentation of
  - algorithms
  - data
  - modifications
- Interoperability with other libraries

Capabilities

Representation

- Atoms, Bonds, Molecules
- Collections of molecules (for reactions)

Manipulation

- Tools to manipulate atoms, bonds, molecules
- Reading and writing a variety of molecular formats (including CML2 and InChI)

Support

- Geometry tools
- Graph tools - ring detection, substructure search
- Structure rendering, 2D & 3D coordinate generation
- IUPAC name generation

Current Usage

- NMRShiftDB
- Seneca
- Cheminformatics web services - descriptors, similarity calculator
- Proteochemometrics
- JChemPaint

Steinbeck, C. et al., *J. Chem. Inf. Comput. Sci.*, **2003**, *43*(2), 1733–1739

`http://almost.cubic.uni-koeln.de/jrg/software/seneca/`

`http://blue.chem.psu.edu/~rajarshi/code/java/cdkws.html`

Spjuth, O. et al., *CDK News*, **2**(2), 54–56

`http://almost.cubic.uni-koeln.de/cdk/jcp`

Outline

1. Introduction

2. CDK & QSAR Modeling
    - Descriptors
    - Working With R
    - An Application

3. Summary

QSAR Modeling Requirements

Components

- Statistics
- Cheminformatics

Approaches

- Access cheminformatics functionality from a statistical environment.
- Access statistical functionality from within a cheminformatics environment.

Designing The CDK Descriptor Classes

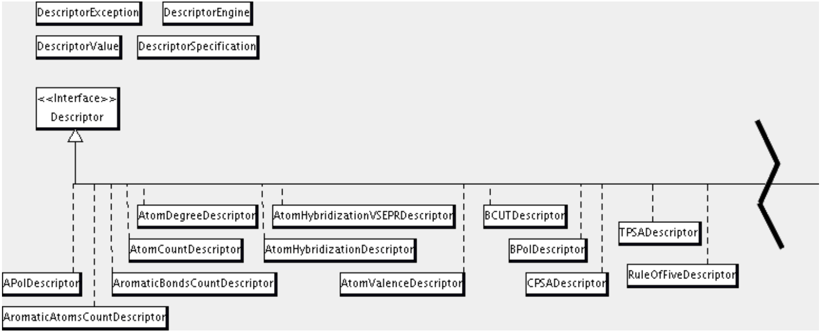
Design Goal

The aim is to allow descriptor calculations to include numerical data as well as provenance. Using CML as an output format, both types of information can be reliably encapsulated

Design Details

- Each descriptor routine has a uniform calling and return interface
- Descriptor specific details can be set via parameters
- Each descriptor has an associated *DescriptorSpecification* object that contains
  - title
  - author
  - reference to a *dictionary* entry which can contain further information

The CDK Descriptor Hierarchy



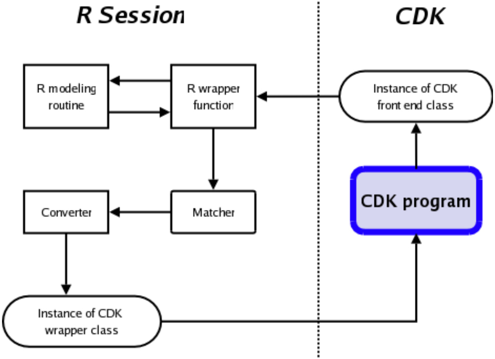
C. Steinbeck et al., *Curr. Pharm. Des.*, in press

Using The CDK Descriptors

```java
AtomContainer ac;

// Instantiate the descriptor
Descriptor descriptor = new BCUTDescriptor();

// Set the parameters
Object[] params = {new Integer(3),new Integer(3)};
descriptor.setParameters(params);

// Get the results
DescriptorValue value = descriptor.calculate(ac);
DoubleArrayResult result = value.getValue();
```

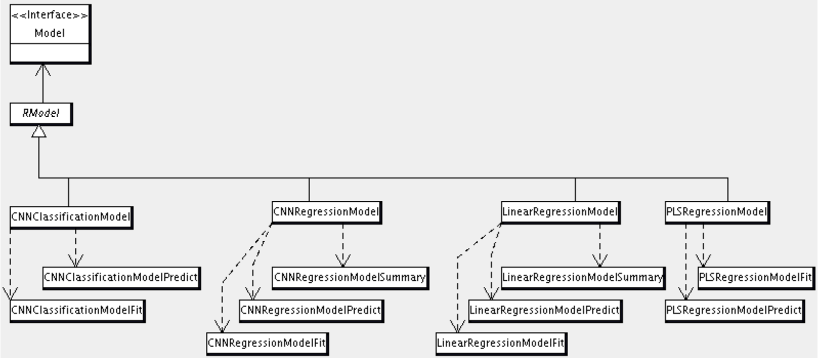
The CDK-R Interface

- Based on the SJava package
- Allows Java code to pass a number of primitives directly to R
- Complex R objects need to have Java wrappers



http://www.omegahat.org/RSJava/

The QSAR Modeling Hierarchy


<<Interface>>
Model
RModel
CNNClassificationModel
CNNRegressionModel
LinearRegressionModel
PLSRegressionModel
CNNClassificationModelPredict
CNNClassificationModelFit
CNNRegressionModelSummary
CNNRegressionModelPredict
CNNRegressionModelFit
LinearRegressionModelSummary
LinearRegressionModelPredict
LinearRegressionModelFit
PLSRegressionModelFit
PLSRegressionModelPredict

Extending the Code

For a model type X . . .

- Create a front-end class: XModel
- Create classes to represent the fitted model and the predictions from that model type: XModelFit & XModelPredict
- Create a class to contain the summary for the model type: XModelSummary
- Provide funtions wrappers on the R side to ensure that the proper Java classes are instantiated when returning the X mode object (or its predictions)

Guha, R., *CDK News*, **2005**, *2*(1), 7–13

Extending the Code

Other Statistical Systems

- The CDK-R hierarchy is a specialization of the *Model* interface
- The *Model* interface specifies two basic functions: **build()** & **predict()**
- Other statistical engines can be easily integrated into the CDK hierarchy
- Future work involves the integration of Weka and Matlab into the CDK QSAR hierarchy

Ian H. Witten et al., "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005

Restrictions

- Embedded R is not multi-threaded
- This implies that in a given Java program all R objects share the same R session
- Since Java garbage collection is not manual, deleting R objects cannot be automated (in general)
- Direct graphical output can be tricky

Guha, R., *CDK News*, **2005**, *2*(1), 7–13

An Application of the CDK-R Interface

Goal

Provide a web-enabled, easily accessible interface for the use of prebuilt QSAR models as well as to provide facilities for building QSAR models with new data

Philosophy

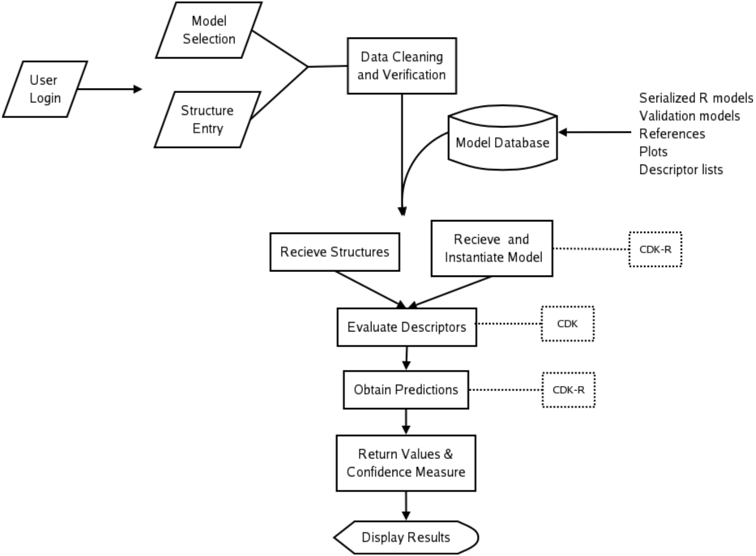
- The use of the application should be black box in nature
- The underlying code and data should be accessible and documented

QSAROnline

Features

- Based on open-source technologies
- Tomcat and Apache provide the web backend
- Spring provides the framework to handle user interaction
- Hibernate provides transparent database access
- CDK and R provide the backend computational facilities

An Application of the CDK-R Interface



QSAROnline

What is Available Now?

- Currently we provide boiling point and melting point models
- Models for other properties are being prepared
- Each model is associated with a measure of validity
- Models also have descriptions including references for data associated with them.

`http://white.chem.psu.edu/`

Confidence Measures

How Do We Provide Confidence in the Predictions?

A classification technique is used to decide whether the model will predict the property for an observation with low or high residual error

Procedure

- Take the TSET residuals and divide them into 2 classes
- Train a CNN classifier on the 2 classes
- The method can be applied to any regression model and provides a probability that a new prediction will be *good* or *bad*

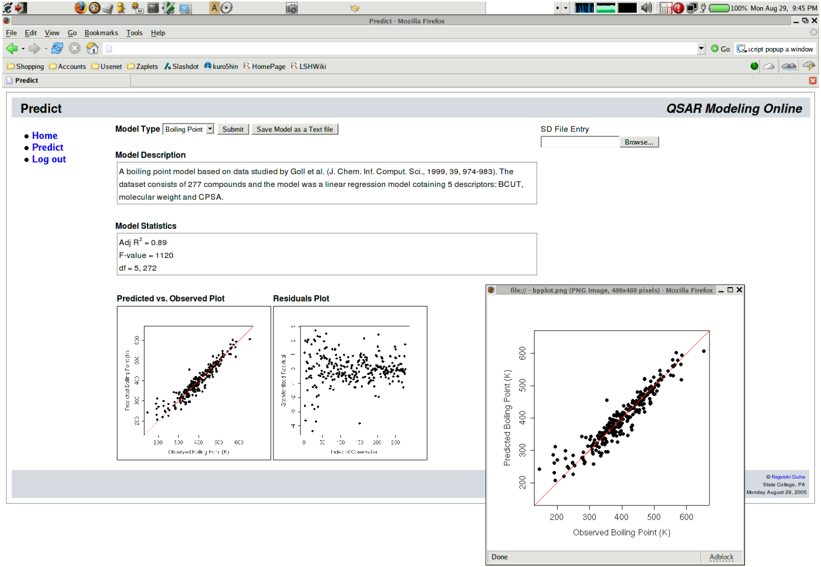
Guha, R. and Jurs, P.C., *J .Chem. Inf. Model.*,**2005**, *45*(1),65–73

QSAROnline

Are the Models & Data Available?

- The data used to build models is obtained from published sources
- The descriptors used to perform modeling are available within the CDK
- The final models are simply serialized R objects and can be downloaded by the user and loaded into a personal R session

QSAROnline



Mon Aug 29, 9:45 PM
Predict - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
script popup a window
Shopping Accounts Usenet Zaplets Slashdot kuro5hin HomePage LSHWiki
Predict
Home
Predict
Log out
Submit
Save Model as a Text file
Browse...
Predicted vs. Observed Plot
Residuals Plot
Predicted Boiling Point (K)
Observed Boiling Point (K)
Standardized Residual
Index of Observation
file:// - bpplot.png (PNG Image, 480x480 pixels) - Mozilla Firefox
200
300
400
500
600
Done
Adblock

Predict — ***QSAR Modeling Online***

Model Type Boiling Point

SD File Entry

Model Description

A boiling point model based on data studied by Goll et al. (J. Chem. Inf. Comput. Sci., 1999, 39, 974-983). The dataset consists of 277 compounds and the model was a linear regression model cotaining 5 descriptors: BCUT, molecular weight and CPSA.

Model Statistics

Adj R^2 = 0.89
F-value = 1120
df = 5, 272

© Rajarshi Guha
State College, PA
Monday August 29, 2005

Outline

1. Introduction

2. CDK & QSAR Modeling
   - Descriptors
   - Working With R
   - An Application

3. Summary

Summary

Future Work

- Larger set of descriptors
- Wrappers for more modeling routines
- Allow users to supply data to build models - difficult!

Conclusions

- The CDK provides a comprehensive platform for cheminformatics projects
- The interface to R provides access to a wide range of statistical functionality
- The open nature of the QSAROnline project allows for user contributions as well as transparency in terms of techniques and data

Acknowledgements

- Nelson Hayes
- C. Steinbeck, E. Willighagen and the CDK community
- The R development team and community