

# CICC Quarterly Report

Rajarshi Guha

School of Informatics  
Indiana University

6<sup>th</sup> October, 2006

- ▶ Circumvents NIH restrictions on using their CGI/SOAP interfaces
- ▶ Allows us to add extra information
- ▶ Will link up with OSCAR derived PubMed database

- ▶ The *Compounds* table which contains most of the fields from PubChem. Excludes
  - ▶ 2D coordinates
  - ▶ Some precalculated fields
- ▶ The *Substance* table
- ▶ The *Compound-Substance* associations table
- ▶ Synonyms for compounds and substances
- ▶ Updated monthly (in progress)

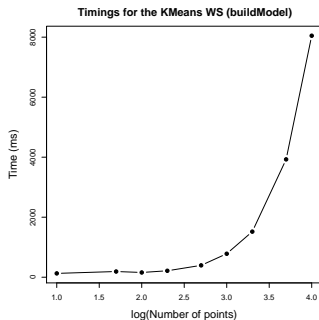
# Value Additions

- ▶ SLogP and MR values from U. Michigan
- ▶ Thermodynamic properties, 3D coordinates (in progress)
  - ▶ Needs to be coordinated with Varuna
- ▶ Fingerprints for compounds
  - ▶ Allows for SMARTS based searches
- ▶ Mode of access
  - ▶ Currently, raw SQL
  - ▶ Web services?
  - ▶ What queries will the WS's be designed for?

- ▶ R web service infrastructure is in place
  - ▶ Currently being populated with actual services
- ▶ Location
  - ▶ Web services on my machine
  - ▶ Rserve on the gridfarm
- ▶ Following services are available
  - ▶ Regression using OLS, CNN and RF
  - ▶ Clustering using *k*-means
  - ▶ Classification using LDA
  - ▶ Scatter plots, histograms (returns byte[])

- ▶ Services try and implement all the parameters for the individual R methods
- ▶ *Simplified* services are also available
  - ▶ They use the R defaults
  - ▶ Allows us to (usually) supply just X & Y data
- ▶ I/O uses Java primitives (`int`, `double`, `String`)
- ▶ However, arrays (`[]` & `[][]`) are also used
  - ▶ 1D arrays can be handled by non-Axis clients
  - ▶ Don't know about the case of 2D arrays

- ▶ All services handle VOTable data, via a URL
- ▶ The whole document is loaded into memory.
- ▶ R works in-memory - so all the data has to go into memory at one point anyway
- ▶ Might be better to shift the load from the JVM onto R?



Needs `-Xmx512M` for Tomcat.  
Clustering time is insignificant

# Enhancements & Issues

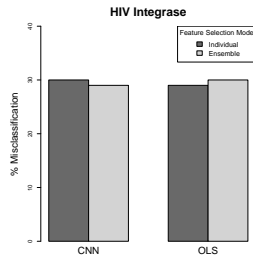
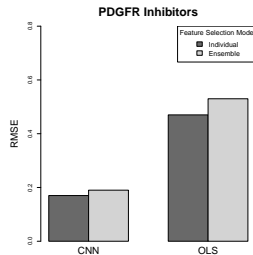
- ▶ Providing predictive model services is trivial
  - ▶ Generating input features is the tricky part
  - ▶ We can continue implementing descriptors
  - ▶ Or we can try and provide services for commercial code
- ▶ State must be maintained between WS method calls
  - ▶ Achieved via serialized R objects
- ▶ Each model is a file on disk and is loaded for each query
- ▶ Makes model storage easy. But how long do the *scratch* models remain for?
- ▶ Bugs/feature requests can be submitted via the CICC Sourceforge page

- ▶ Working on updating CDK services to be in sync with the CDK SVN
- ▶ Adding some new services
  - ▶ TPSA and other fundamental services
- ▶ Collecting links to WSDL from other properties
  - ▶ U. Cologne (NMRShiftDB)
  - ▶ VCC Lab (ALogPS)
  - ▶ NCI (in progress)

- ▶ Implementing more ADAPT descriptors into the CDK
- ▶ R packages
  - ▶ Updated `rcdk`, allows easier access to the CDK from R
  - ▶ Implemented `rpubchem`, allows access to PubChem compound and bio-assay data from within R
  - ▶ Paper submitted to *J .Stat. Soft.*
  - ▶ Packages available on CRAN

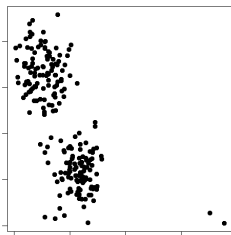
# Algorithms - Ensemble Descriptor Selection

- ▶ Talk presented at Fall ACS
- ▶ Determines *ensemble* features
- ▶ Little degradation in accuracy compared to individually optimized models
- ▶ Write up will be submitted

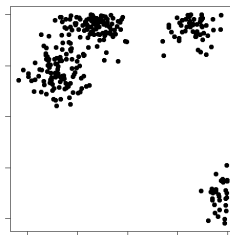


# Algorithm - $R$ -NN Curves & Cluster Counts

- ▶ Poster presented at Fall ACS
- ▶ Correctly detects the *natural* number of clusters
- ▶ Seems to be more reliable than other measures
- ▶ In the final stages of experiments



$N_{cluster} = 3$



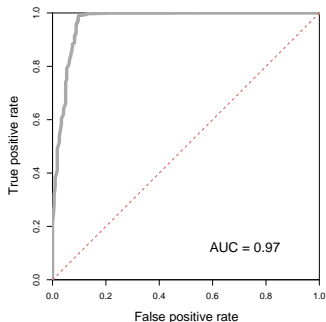
$N_{cluster} = 4$

# Collaboration with Scripps, FL

- ▶ Analysis of tox datasets (LeadScope/MLSCN)
- ▶ Currently have a predictive RF model using BCUTS
  - ▶ Unbalanced problem, used oversampling
  - ▶ No significant feature selection performed
  - ▶ Decent performance within species
  - ▶ Not impressive across species

	Non-toxic	Toxic
Non-toxic	1477	7
Toxic	21	35

Mouse, oral, error rate = 1.82%



# Future Directions

- ▶ Analysis of structural fragments
- ▶ Chemical space of the datasets using BCUTS and ECFP
- ▶ Improving cross-species predictive ability
- ▶ Alternative assay suggestions
- ▶ OSCAR & patents