The use of KNIME to support research activity at Lhasa Limited

Data processing through to proof-of-concept implementations

Sam Webb
samuel.webb@lhasalimited.org
Overview

• The Lhasa-KNIME timeline
• Internal KNIME node development
• Use cases
• Example of proof-of-concept developments
• Some use cases acquired from Lhasa employees
  • Opinions are my own
Who is Lhasa?

- Not-for-profit organisation
- Develop software for the prediction of toxicity, metabolism, degradation and supporting databases
- Undertake data sharing initiatives
- ‘Head office’ in Leeds UK

http://www.lhasalimited.org
Who am I?

• Working in the Research Group
• Cheminformatics, data analysis & machine learning
• Develop new libraries / tools to support my research activities
• Share these developments with others at Lhasa
• KNIME makes the sharing easier
  • But if we want heavy visualisation we may choose to prioritise into our internal cheminformatics platform and only provide some functionality in KNIME
Timeline

• 2011
  • A few users within the Research Group
  • Love of meta nodes begins

• 2012
  • Internal KNIME indoctrination training begins
  • KNIME node development starts
    • Integrate our chemical engine
    • First proof-of-concept: black box model interpretation
    • Second proof-of-concept: negative predictions
    • Yay, loops!

• 2013
  • Second push for KNIME training, bigger uptake
  • Evaluated the KNIME server

You can do that in KNIME!

Now implemented in our software.

I’ll help you, but only if we use KNIME
The Lhasa-KNIME timeline

• 2014
  • Hit the 125 internal KNIME nodes
  • Bayesian network proof-of-concept
  • KNIME now popular with scientists
    • Training given to new employees, get them young…
  • Try to avoid using loops
  • Frequent requests: how do I do this not using a java snippet?
  • Attempt at using Pipeline pilot

Getting quicker at making these! But now I know what I did wrong in the first ones

Streaming is a nice concept (but I don’t want to pay for it)

Why doesn’t it store the data in the output!
Timeline

• 2015
  • Jenkins build environment
  • Test workflows
  • Ongoing node development
  • Hosting the next Cheminformatics Special Interest Group meeting
    • May 2015 in Leeds
    • http://www.knime.org/cheminformatics-workshop-may-2015
What do the KNIME users look like now?

- A lot of people KNIME
- 2 developers making nodes
- ~20 users: Research Group, knowledge scientists, database scientists
KNIME INFRASTRUCTURE
Our setup

- File server: sharing workflows
- Web server: running a Wiki
- Build machine: in the works
- Services:
  - Web services
  - Long jobs? ← under investigation
- User
  - Many different versions of KNIME
  - Many different version of plugins
The Wiki

- Documentation of KNIME functionality, example workflows and procedures for common activities

**Usage**

The KNIME node is very crude. It runs the command line configuration for MOPAC which causes MOPAC to load. The row is processed, MOPAC closes and then the next row loads a new instance of MOPAC.

Therefore the node can only really be used for small numbers of structures. For larger datasets you will need to run it overnight.

The node has been configured to process a mop (MOPAC structure format) column. The keywords specified in the dialog will be added in and the complete mop will be run through MOPAC2012.exe. I have set the node to read in the .mop file and store this as a string. A view allows you to easily browse the results. Currently the following values are automatically extracted from the output file and added to the table: energy, HOMO, LUMO, Alpha SOMO, Alpha LUMO, Beta SOMO and Beta LUMO.

Make sure your structures have explicit hydrogens.

**Workflow**

It's best not to run this node when you need your machine. Currently an instance of MOPAC will load on your main monitor above all other windows when a new line is processed.

This workflow will provide you with a table with the energy, HOMO and LUMO of the structures given (if successfully calculated). The .mop output files can be viewed in the File viewer view. On the right of the view you can choose the file (row index) - ID in mop file.

The out, .arc and .mop files can be found in your chosen results directory.
Workflow sharing

- Shared workflow area
- Import directly from the shared workflow area into your KNIME workspace
- In house addition/extension to KNIME
USE CASES
Lhasa use cases: descriptor calculation

• Reduce the number of programs needed to calculate descriptors
• Perform the data analysis where the descriptor calculation happens
• We have a few nodes of our own
• Avoid PaDEL
  • can’t find the source code to fix the issue myself
Lhasa use cases: structure curation

• KNIME has a number of chemical engines accessible: CDK, RDKit, Indigo, ChemAxon and our own

• We have developed structure curation workflows that automatically fix some issues and flag others for manual analysis

• Implemented our own nodes calling out to ChemAxon standardizer and Structure checker
Lhasa use cases: structure curation

Standardiser editor available in KNIME.
Either configure or load in an XML configuration.

Applies standardisation and lists completed tasks
Lhasa use cases: structure curation
Lhasa use cases

• Data processing:
  • Combining datasets: find overlap, compare activities when overlap exists, join in data where no overlap exists…
  • Making overall calls: lots of results for a compound, combine into a single result based on defined rules

• Monitoring:
  • Extracting data from a the database which has been altered identifying review work content

• (Q)SAR
  • Model building, clustering, algorithm development, applicability domains, chemical space investigation…. 
Key benefits of KNIME

• All in one place (nearly)
• Workflow = documentation
  • Only really works when data is present
  • Still need to annotate!
  • Workflows can be overly complex and tricky to understand
• Quickly share new updates to internal code developments with users
  • Doesn’t mean they will actually update though…
• Actions are reversible! No overriding of data, just change a configuration of a node / replace the node!
Issues we’ve had

• People can develop super messy workflows
  • I hide mine behind meta nodes
• Memory can sometimes be an issue
• Reporting can be painful
• Looping in loops can be very slow
• 32 bit requirements on some 3rd party DLL’s
• Getting people to stay up to date
• Remembering to look up what other words a node may go by
  • I have at least on one occasion made a node that already existed…
Node development

• SVN for versioning of our KNIME nodes
• Old build machine for building the update site – plan to switch to a Jenkins build process
  • Thanks to Vernalis for help with this!
• Everyone uses a different version of KNIME and plugins
  • Our version of “have you turned it off and on again is”: “have you updated?”
Workflow and node development

• We now have over 100 nodes
• We also have some Eclipse RCP additions (views)
Chemical engine integration

- We have our own Java based chemical engine and have added some of its functionality to KNIME
  - Structure conversion
  - Inchi generation
  - Structure drawing
  - Structure validation
  - Fingerprint generation
  - Property calculation
  - Similarity calculation
  - Tautomer generation
  - Fragmentation

Overlaps with RDKit, Indigo and CDK
ChemAxon integration

• We licence a number of ChemAxon components
• We develop our own nodes…
• We didn’t want to have to pay to licence the ChemAxon tool twice
• We wrote our own ChemAxon based nodes
BitSet integration

• A lot of our code uses BitSets or SparseBitSets
• The KNIME BitVector wasn’t very well developed in functionality
• We created a number of nodes operating on BitSets and conversion between BitVector and BitSet
• BitVectors are now more developed in functionality and some overlap exists
• Some of this functionality would be better implemented as extensions e.g. new actions for the GroupBy node
BitSet example

- Same feature occurs in multiple locations
- Group on the feature ID and perform an OR operation on the BitSet

We can now highlight all the ACTIVATING atoms
MOPAC integration

- “MOPAC (Molecular Orbital PACkage) is a semiempirical quantum chemistry program based on Dewar and Thiel's NDDO approximation” - http://openmopac.net/
- We wanted to access to some of the values MOPAC can calculate
- Tried to make it easier for people to use
- Made a node / workflow that calls off to MOPAC, parses the output and creates a table of results
- Some issues such as it opening a new instance of MOPAC for each row…
MOPAC example
SmartCyp & WhichCyp

Predicting CYP sites of metabolism

http://www.farma.ku.dk/smartcyp/
Generic nodes: multi column row splitter

- Attribute matching taken from the row filter
- Select multiple columns to apply filter to
- Include / exclude based on matching one or all columns
Generic nodes: model performance

- Similar functionality to the Scorer node
- Calculates various performance metrics for binary classification models
- Can choose multiple prediction columns
Generic nodes: table to HTML

- Convert a table to a single HTML cell
- The String render will render HTML tags
- Select which columns to include
  - StringValue, IntValue, DoubleValue
- Creates a single cell output
Where would you use these nodes?

Here we calculate the performance of the Random Forest with morgan and MACCS fingerprints.

Convert the performance table to HTML and email.

Filter out rows where either model predicts active.
PROOF OF CONCEPTS
Bayesian networks: background

• Had a project investigating combining human experts opinion and data
• Some of this work could be done using R or python libraries
  • They were a bit messy to use and not as convenient
• Core Bayesian algorithms available from SMILE and Genie: https://dslpitt.org/genie/
• Implemented some algorithms in Java to learn the posterior
  • Markov Chain Monte Carlo simulation
  • Approximate Bayesian Computation
Bayesian Networks: where does KNIME come in?

- Using KNIME as a method for allowing non coders to investigate various configurations of experimental codebases
- KNIME nodes developed providing a user interface for the Bayesian Network libraries (in house)
- We would end up doing the data analysis in KNIME anyway
- Can now let users who are less comfortable with command line applications use the code
Bayesian networks

- I’m beginning to dislike loops.
  - If it needs a complex loop maybe I should write a node to do it?
  - Looping in loops can be very slow!
    - One particularly bad loop had a 6 hour runtime, it takes < 5 seconds as its own node. I suspect I made bad choices in the development of the workflow…
- Complex loops may result in multiple points of failure
- We had a parameter grid for the Bayesian network optimisation.
  - We could use existing KNIME nodes to loop over the grid.
  - We could do batch processing!
Bayesian networks: batch processing

• Write a batch script to iterate through the parameters
• The workflow saves a number of files (csv, svg and png) per run

```bash
set WORKFLOW_FILE=workflow.zip
set saveLoc=C:\\container\\current\\bayesian\\skinIrritation\\irrCorr\\experiments\\priorDistributions\\uniform_expert\\
set TRAIN_FILE=saveLoc\train.csv
set PRIOR_FILE=saveLoc\prior.csv
set TEST_FILE=saveLoc\test.csv
set METRO=JAVA
set distance=1.8
set numSamples=1000
set folder=output_method$2_numSamples$6_distance$2
set workflow=saveLoc\folder\\\\workflow\\
C:\\container\\knime\\knime_2.10.1\\knime.exe -destDir=\\\\folder\\ -reset -preferences="prefs.epf" -consoleLog -nosplash -application org.knime.product.KNIME_BATCH_APPLICATION -work
```
Bayesian networks: workflow snippets

Sample from 56 distributions 1000 times

Calculate statistics and density plots
We are creating 1 image per column. We could loop and get the image back in KNIME.

Instead we do the loop in R and save to a temp location.
Bayesian networks: reporting

- Point to a directory and read in the files saved in the batch processing
- Create a PDF report

It would be nice to save a report during a batch process
Bayesian networks: reporting

ABC run output
Location: C:\container\current\bayesian\skinIrritation\rrCorr\expertDistributions\train_test
\noMissing\batch\output_JAVA_1000_0.4
The folder name indicates the setup of the ABC job. output_METHOD_NUMSAMPLES_DISTANCE. So JAVA_1000_1.0 is the JAVA simulation method, with 1000 samples at a distance of 1.0.

Performance

<table>
<thead>
<tr>
<th>Row ID</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>Sen</th>
<th>Spec</th>
<th>ACC</th>
<th>Cohen's kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>NON_IRRIT</td>
<td>701</td>
<td>318</td>
<td>820</td>
<td>161</td>
<td>0.81</td>
<td>0.72</td>
<td>0.71</td>
<td>54.34%</td>
</tr>
<tr>
<td>CORROSIVE</td>
<td>324</td>
<td>114</td>
<td>1452</td>
<td>110</td>
<td>0.75</td>
<td>0.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRRITANT</td>
<td>393</td>
<td>150</td>
<td>1146</td>
<td>311</td>
<td>0.56</td>
<td>0.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.71</td>
<td>54.34%</td>
</tr>
</tbody>
</table>

Uncertainty
An evaluation of the uncertainty of the predictions has also been performed. The KNIME statistics node was run on a per query basis on the test set. Each query having x predictions based upon each set of approved sampled phi values.

Density plots have then been calculated using ggplot2 (in R) on the entire test set (2,000 Draize class is predicted separately.)

Distributions
Comparison of posterior and prior distributions, the prior represents 1,000 samples from the prior and the posterior is the distribution of the approved samples in the ABC run.
Model interpretation

- Implemented the Similarity maps method by Riniker & Landrum
- [http://www.jcheminf.com/content/5/1/43](http://www.jcheminf.com/content/5/1/43)
- Summary:
  - Assign a contribution to an atom as the difference between the active class probability with the atom vs without the atom
Model interpretation

Table converted to a HTML string to then be rendered in the cell
Model interpretation

• Implemented the Feature combination networks interpretation method (easy to do as we developed it)

• http://www.jcheminf.com/content/6/1/8

• Summary:
  • Elucidate the models behaviour based on fragments not individual atoms
## Model interpretation

<table>
<thead>
<tr>
<th>RowID</th>
<th>Atom</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atoms(0)</td>
<td>0</td>
<td>0.32</td>
</tr>
<tr>
<td>Atoms(1)</td>
<td>1</td>
<td>0.36</td>
</tr>
<tr>
<td>Atoms(2)</td>
<td>2</td>
<td>0.58</td>
</tr>
<tr>
<td>Atoms(3)</td>
<td>3</td>
<td>1.0</td>
</tr>
<tr>
<td>Atoms(4)</td>
<td>4</td>
<td>0.16</td>
</tr>
<tr>
<td>Atoms(5)</td>
<td>5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Similarity maps

1. **Image 1**
   - Atoms(0): 0.32
   - Atoms(1): 0.36
   - Atoms(2): 0.58
   - Atoms(3): 1.0
   - Atoms(4): 0.16
   - Atoms(5): 1.0

2. **Image 2**
   - Atoms(0): 0.4
   - Atoms(1): 0.44
   - Atoms(2): 0.4
   - Atoms(3): 1.0

3. **Image 3**
   - Atoms(0): 0.12
   - Atoms(1): 0.34
   - Atoms(2): 1.0
   - Atoms(3): 0.0
   - Atoms(4): 0.0
   - Atoms(5): 0.0

### Feature combination networks

- Similarity maps
- Feature combination networks
Emerging Pattern mining

- Emerging Pattern mining algorithm has been implemented in a couple of nodes (R. Sherhod ~ now at Vernalis)

Emerging Pattern mining

- Fragment dictionary approach
- Fragment the dataset using a fragmentation approach
  - We use the reduced graph approach developed in house
Emerging Pattern mining: visualisation of EP

- This is where it becomes a bit tricky in KNIME
Emerging Pattern mining: visualisation of EP

- 23 structures contain Bit357
- Bit357 is an acid chloride motif
- We have two lists of RowID’s: one for active support and one for inactive support
- We could make a report
  - Can this be automated? There’s over 100 EP’s!
Derek web service

- Web service (SOAP) available for the expert system for toxicity prediction Derek.

Acknowledgements

• KNIME
• Community contributors
• Forum users
• All my colleagues who shared their use cases