Migrating a cheminformatics and reporting platform to KNIME

Serge P. Parel, PhD
KNIME User Group Meeting, Berlin
26th February, 2015
Outline

- Exquiron – Who are we?
- Research informatics infrastructure
  - Why migrate?
  - What to migrate?
  - Who can help?
- Example 1: Reporting of dose-response results
- Example 2: Hit expansion toolbox
- Summary
- Acknowledgments
Exquiron – who are we?

- High throughput screening & hit profiling services
- Independent and privately held company
- 10 scientific staff
- >75 cumulated years of experience in hit identification services
  - Covering pharmaceutical, animal health, consumer care, nutraceutical, and agrochemical research
- 850m² lab & office space
- Located in Reinach, near Basel, Switzerland
Research informatics infrastructure
As of end of 2013

**Chemistry cartridge:**
- ChemAxon JChem on Oracle

**Workflow platform:**
- Accelrys Pipeline Pilot

**Cheminformatics platform:**
- Accelrys PP Chemistry Library

**Biological data warehouse:**
- Exquiron NaviGator database on MS-SQL server
Research informatics infrastructure

Problem

Chemistry cartridge:
• ChemAxon JChem on Oracle

Workflow platform:
• Accelrys Pipeline Pilot

Cheminformatics platform:
• Accelrys PP Chemistry Library

Biological data warehouse:
• Exquiron NaviGator database on MS-SQL server

- Renewal financial terms not suitable for small enterprises with only very few users (compared with pricing of Start-Up package)
- Duplicate functionalities (Accelrys PP Chemistry Library and ChemAxon PP components)
Research informatics infrastructure
Solution?

**Chemistry cartridge:**
- ChemAxon JChem on Oracle

**Workflow platform:**
- Accelrys Pipeline Pilot

**Cheminformatics platform:**
- Accelrys PP Chemistry Library

**Biological data warehouse:**
- Exquiron NaviGator database on MS-SQL server

---

- Simple protocols can be ported in house, but what about more complex workflows?
Research informatics infrastructure
Who can help?

- **Workflow platform:**
  - Accelrys Pipeline Pilot

- **Cheminformatics platform:**
  - Accelrys PP Chemistry Library

- **Workflow platform:**
  - KNIME

- **Cheminformatics platform:**
  - ChemAxon/Infocom nodes

- Ask KNIME if they can help:
  - Yes, but they rely rather on partners for consulting services and support them where needed

- Ask ChemAxon if they can help:
  - Yes, through their Consultancy Services
Scope of the migration

- Lots of short workflows (SD files and lists manipulation, virtual screening,...)
- Dose-response data reporting
  - Input file generation
  - Report generation
- Exhaustive hit expansion toolbox
  - Hit expansion
  - Data fusion and analysis

Done in house

Done over Consultancy with support from KNIME

Done over Consultancy

Done in house
Example 1
Reporting of dose-response experiments

- Typical workflow for a hit finding campaign:
  - Many 100K compounds
  - HTS
  - Confirmation usually in duplicates
  - Dose-response determination assay + counterassay(s)
  - Actives
    - A few 1,000s
  - Confirmed actives
    - A few 100s
  - Hits

Reporting
- Chemists want to see structures and IC50/EC50 values
- Biologists want to see IC50/EC50 values and dose-response curves
Reporting of dose-response experiments
Workflow (I)

Select project and set of experiments

Exquiron data warehouse
- Exp. data points
- IC50 fitting parameters
- Purity

Tag each data point with experiment, concentration and replicate. Tag IC50 parameters by experiment

JChem cartridge or SD File

Additional data (Text file)

SD File
Reporting of dose-response experiments
Workflow (II)

- **SD File**
  - Generate report elements
  - Compound structure
  - Results table
  - Dose-response curves

- **Export as pdf**

Confidential to Cortoso and Exquiron

**EXQ0081338**

<table>
<thead>
<tr>
<th>Experimental data</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Dose (%)</td>
</tr>
<tr>
<td>1.0%</td>
</tr>
<tr>
<td>3.5</td>
</tr>
<tr>
<td>IC50 1st h</td>
</tr>
<tr>
<td>1.2</td>
</tr>
<tr>
<td>IC50 20 h</td>
</tr>
<tr>
<td>0.8</td>
</tr>
</tbody>
</table>

**EXQ0081383**

<table>
<thead>
<tr>
<th>Experimental data</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Dose (%)</td>
</tr>
<tr>
<td>1.0%</td>
</tr>
<tr>
<td>3.5</td>
</tr>
<tr>
<td>IC50 1st h</td>
</tr>
<tr>
<td>1.2</td>
</tr>
<tr>
<td>IC50 20 h</td>
</tr>
<tr>
<td>0.8</td>
</tr>
</tbody>
</table>

© copyright 2015 Exquiron Biotech AG
Reporting of dose-response experiments
SD file generation (I)

- Concentration may be different for each experiment
- Number of replicates per concentration may vary
- Datapoint at a given concentration could be invalid
- Number of experiments to report may vary
  - Fields for SD file cannot be defined *a priori*

**Database table**

<table>
<thead>
<tr>
<th>exq_id</th>
<th>exp_id</th>
<th>Conc</th>
<th>Valid</th>
<th>Activity(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXQ0012345</td>
<td>2344</td>
<td>90</td>
<td>TRUE</td>
<td>6.411991119</td>
</tr>
<tr>
<td>EXQ0012345</td>
<td>2344</td>
<td>30</td>
<td>TRUE</td>
<td>41.63282013</td>
</tr>
<tr>
<td>EXQ0012345</td>
<td>2344</td>
<td>10</td>
<td>TRUE</td>
<td>74.8677597</td>
</tr>
<tr>
<td>EXQ0012345</td>
<td>2344</td>
<td>3.333333254</td>
<td>FALSE</td>
<td>83.46128082</td>
</tr>
<tr>
<td>EXQ0012345</td>
<td>2344</td>
<td>1.111111164</td>
<td>TRUE</td>
<td>87.6084137</td>
</tr>
<tr>
<td>EXQ0012345</td>
<td>2344</td>
<td>0.370370358</td>
<td>TRUE</td>
<td>93.19058228</td>
</tr>
</tbody>
</table>

**SD file**

```
M  END
> <exq_id>
EXQ0012345
>   <Activity_0.37037_1_S1>
   93.19058227539062
>   <Activity_1.11111_1_S1>
   87.60841369628906
>   <Activity_3.33333_1_S1_invalid>
   83.4612808227539
>   <Activity_10.0_1_S1>
   74.86775970458984
>   <Activity_30.0_1_S1>
   41.63282012939453
>   <Activity_90.0_1_S1>
   6.41199119384766
```
Get data from database:
- SELECT statement
- Each datapoint is a separate row

Crunch data:
- Through Group By’s, pivoting and some string manipulation
  - All new columns/column names generated at once through pivoting
- For each compound, one single row with each datapoint is a separate column

Finalize:
- Merge structures, data points and regression parameters.
- Write SD file
Reporting of dose-response experiments

Report generation

- Read and parse SD file
- Generate report information as tables
  - Compound ID
  - Structure
  - Results table
- Generate DR curves based on logistic regression parameters (using Math Formula nodes)
- Feed DR curves and experimental data points to R and use *ggplot2* to generate the DR plot, return it as picture
- Send all elements to the KNIME Report Designer (BIRT)
- Combine to a report with header, footer, logo...
- Generate pdf file
Reporting of dose-response experiments

Result

Old platform

KNIME

Confidential to Contoso and Exquiron

EXQ0081338

Experimental data

<table>
<thead>
<tr>
<th>Part (%)</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC50 (pIC50)</td>
<td>5.6</td>
</tr>
<tr>
<td>IC50 1 Hill slope</td>
<td>1.5</td>
</tr>
<tr>
<td>IC50 2 Hill slope</td>
<td>1.7</td>
</tr>
</tbody>
</table>

EXQ0081338

Experimental Results

| Purity (%) | 100 |
| IC50 (µM, S1) | 2.91 |
| Hill slope (S1) | 1.47 |
| IC50 (µM, S2) | 2.75 |
| Hill slope (S2) | 1.7 |

Confidential to Exquiron and Contoso

EXQ0081339

Experimental data

<table>
<thead>
<tr>
<th>Part (%)</th>
<th>69</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC50 (pIC50)</td>
<td>8.2</td>
</tr>
<tr>
<td>IC50 1 Hill slope</td>
<td>1.1</td>
</tr>
<tr>
<td>IC50 2 Hill slope</td>
<td>n.d.</td>
</tr>
</tbody>
</table>

EXQ0081339

Experimental Results

| Purity (%) | 69 |
| IC50 (µM, S1) | 60.89 |
| Hill slope (S1) | 1.12 |
| IC50 (µM, S2) | n.a. |
| Hill slope (S2) | n.a. |
What about portability?

- Can these workflows easily be customized by a KNIME newbie for other types of reporting?

YES !!!

Reporting dose-response data as pictures (with file name identical to compound name)

Report results of reversibility of enzyme inhibition analysis
Example 2
Hit expansion workflow

- Purpose of hit expansion (*aka* SAR expansion, SAR by catalogue):
  - To consolidate and expand knowledge on active regions in chemical space
  - To find derivatives of active scaffolds to establish or broaden SAR information
  - To expand hit series by scaffold hopping
  - Strengthen knowledge base for direct use in H2L campaign
Hit Expansion workflow

Workflow


Applied to bridge gaps in chemical space
Hit Expansion workflow
Example (ChemBL VEGFR2 dataset)

A: ChemBLID = CHEMBL187565, Ki = 16 \mu M

Amide \rightarrow\text{Oxadiazole}

T_c = 0.09 \text{ (based on FCFP6 FPs)}

B: ChemBLID = CHEMBL191705, Ki = 13.9 \mu M

1st level Turbosim search

C: ChemBLID = CHEMBL190872, Ki = 210 \text{ nM}

2nd level Turbosim search

D: ChemBLID = CHEMBL190872, Ki = 210 \text{ nM}

Hit expansion workflow

To do

- Port all functionalities of the workflow to KNIME using the ChemAxon/Infocom nodes (with the exception of the pharmacophore graphs)
- Improve and optimize workflow for increased execution speed (original Turbosim implementation very slow due to the large number of similarity searches performed)

➤ Assignment more than fulfilled
Hit expansion workflow
Target database preparation

- Workflow combines similarity searches using FCFP, ECFC and MACCS keys
- Workflow also performs various substructure searches
- Originally, all fingerprints (FP) were pre-calculated and stored with their corresponding structure into a PP cache
  - Still a lot of «overhead» since structures are not needed to run similarity searches (FP already calculated) and FP are not needed when running substructures searches, but everything is carried through the whole workflow
Hit expansion workflow
Target database preparation

After calculation, FPs are stored into a KNIME table file, and the structures are stored separately into a local JChem database, then load only what is needed where you need it.
Hit expansion workflow

Ring generalization

- **Idea:** create substructure queries with generalized ring structures (somewhat the opposite of Bemis-Murcko scaffolds)

- **Implementation:**
  - Loop through each atom, set to atom type ‘any’ if it is in a ring
  - Loop through each bond, set to bond type ‘any’ if it is in a ring

Implemented in KNIME by calling the ChemAxon API in a Java Snippet
Hit expansion workflow
Bioisostere enumeration

- Bioisosteric transformations are applied to generate new virtual template structures
- This step makes use of the Wagener and Lommerse dataset (*J. Chem. Inf. Model.* **2006**, 46, 677) available as MDL .rxn files

- But there is no bioisosteric transformation node available under KNIME
Bioisosteric transformations were implemented as a KNIME metanode using a loop iterating through each reaction file and applying it to the incoming template with a ChemAxon/Infocom UniReactor node.

Works well, but requires a strong atom-to-atom mapping and explicite H-atoms in the .rxn file (which was not the case under PP).
Summary after year 1

- KNIME combined with the ChemAxon/Infocom cheminformatics nodes is an efficient and cost-effective alternative to PipelinePilot
  - 😊 State-of-the-art reporting platform (WYSIWYG and Drag&Drop)
  - 😊 Fingerprint-based similarity search orders of magnitude faster
  - 😊 Required timelines could be maintained (expiration of licenses for old platform)
  - 😊 Workflows have been optimized and streamlined

😊 RAM memory requirements high (at least for the Desktop version of KNIME when dealing with large SD files, prob. less relevant for the Server version). 2 GB assigned to KNIME is probably the minimum

😊 Loops can sometimes be quite slow

😊 Currently no KNIME implementation for the Discengine Pharmacophore Graphs, but alternatives under evaluation
Acknowledgements

- Aaron Hart
- Attila Szabó
- Mihály ‘Medzi’ Medzihradszky
- Andreas Bergner
Thank you!

For more information contact
Serge P. Parel
Exquiron Biotech AG
Kägenstrasse 17
4153 Reinach
Switzerland
E-mail: sparel@exquiron.com
www.exquiron.com