From Corkscrew to Swiss Army Knife

The evolving role of KNIME at DNS

Tim Parrott & Brock Luty
September 16, 2016
To create drugs that enhance Long Term Memory (LTM) in Humans
Chemical Library Design Tool

Select Reactants
Enumerate Products
Calculate Properties
Analyze & Filter
Export Library Plan
KNIME to the rescue!
More please!
Tool Overview

- **Spotfire Export**
- **Selection & Configuration Panel**
- **Custom Nodes**
- **DNS Nodes**
  - DNS Nodes
  - Bulk Loader Nodes
  - Database Nodes
  - SQL Select By Row
- **Loop Nodes**
- **Reactor Nodes**
  - BiReactor
  - TriReactor
  - UniReactor
- **Reserved System Nodes**
  - Batch Details
  - Cluster
  - Deduplication
  - Diversity Elements
  - Export For Spotfire
  - Filter
  - Launch Spotfire
  - Library Plan Publisher
  - PMI Calc
  - RN Calc
  - ROCs Calc
  - Remote CalcP
  - Remote Protocol Executor
  - Remove Additive
  - Stereocchemistry Code Assignment
  - DNS Nodes
  - Database Nodes
  - Loop Nodes
  - Reactor Nodes
  - Reserved System Nodes
  - Batch Details
  - Cluster
  - Deduplication
  - Diversity Elements
  - Export For Spotfire
  - Filter
  - Launch Spotfire
  - Library Plan Publisher
  - PMI Calc
  - RN Calc
  - ROCs Calc
  - Remote CalcP
  - Remote Protocol Executor
  - Remove Additive
  - Stereocchemistry Code Assignment
  - DNS Nodes
  - Database Nodes
  - Loop Nodes
  - Reactor Nodes
  - Reserved System Nodes
  - Batch Details
  - Cluster
  - Deduplication
  - Diversity Elements
  - Export For Spotfire
  - Filter
  - Launch Spotfire
  - Library Plan Publisher
  - PMI Calc
  - RN Calc
  - ROCs Calc
  - Remote CalcP
  - Remote Protocol Executor
  - Remove Additive
  - Stereocchemistry Code Assignment
  - Structure Lookup
  - Structure Search
  - VIDA Launcher
Spotfire Selections returned to KNIME

New nodes with selected products & reactants appear in KNIME
Configuration Panel
Pausing Local Execution

- Ready to Execute
- Executing – can’t be paused yet
- Executing – can be paused
- Local Execution has been paused

BiReactor

Ureas from isocyanates and amines

BiReactor

Ureas from isocyanates and amines

Ureas from isocyanates and amines

Done - fetching results

Ureas from isocyanates and amines
Different User Groups

Structure Guided Chemistry
- Never had Pipeline Pilot
- KNIME Library Design Tool since June 2012
- Rxn enumeration & property calculations

Information Management
- Heavy reliance on Pipeline Pilot
- Data curation, ETL operations and ELN archiving
- Early KNIME adoption

Computational Chemistry
- Heavy reliance on Pipeline Pilot
- Working with large datasets
- Searching, clustering, modeling
Gaps

• Workflow Sharing

• Familiarity / hands-on knowledge

• Specific data processing nodes

• Handling large data sets

• Database Operations

Shared Repository (KNIME Server)

Training

Custom nodes, linked metanodes

Custom nodes, special user accts
Metanodes

- Add Element Counts
- CAS Validator
- Current Screening Deck Structures
- Find Novel Molecules
- First Occurrence Filter
- Merge Molecules
- Remove Duplicate Molecules
- Unique Value Filter

Dialog - 0:0 - First Occurrence Filter

Select the field to filter:
- Field name:
  - Structure ID
  - Structure
  - Assay Result
  - Assay Date
  - Molecular Weight
  - Assay Comment

OK Cancel
Things we hated that have now been fixed
Things we should have asked for

• Input list of values and look up additional data based on the input

• Common process for IM workflows

• Pipeline Pilot has an SQL *Select for Each Data* component

• Multiple ways to do this in KNIME, none satisfactory
SQL Select by Row usage

```
select cs.id as CMPD_STRUCTURE_ID
from compound.cmpd_structure cs
join batch.parent bp
on bp.id = cs.parent_id
where bp.corp_name = ?
and cs.active = 1;
```
Don’t ask and ye shall receive

KNIME 3.2
Things we’d like

- Ability to run bits of Oracle PL/SQL
- Ability to select a nextval from an Oracle sequence
- Support for BLOBs as a flow variable type
Where we are now

- Custom nodes
- Usage logging
- Graphical browsing in configuration panel
- Pausable nodes
- Two-way communication with Spotfire
- Shared workflows
- Linked Metanodes
- Services
- Protocol Publisher
What lies beneath …

• We chose to use KNIME Server for sharing protocols and Meta-nodes
• But we already had a computational services system in place with plans to continue to extend
• Solution: Integrate KNIME into our existing infrastructure
• KNIME both as a consumer of Services and as a provider of computations that can be run as services
DNS Computational Services

- Accessible by any client application in DNS
  - Spotfire, Vida (3D Protein-Ligand Design Tool), many custom applications including command-line tools
  - Accessible in KNIME via Custom nodes and a generic Remote Protocol Executor
Protocol Publisher

- Allows “Publishing” of new Protocols (aka Services) that are immediately seen by the desktop applications
  - Protocol Meta-Data: Name, Author, Description, Date, Notebook, URL, Thumbnail
  - Executable: Shell Script, Python, PERL, Groovy, BAT, ChemAxon Chemical Terms, ...
    - KNIME Wizard: Load zip of an existing KNIME Workflow which automatically gets wrapped and is usable as a service by KNIME or other tools
  - Executable Meta-Data: Inputs, Outputs, ...

![Publisher](image1.png)
![Scripts](image2.png)
![KNIME “Wizard”](image3.png)
...and at the darkest depths

- DNS Computational Services
  - Automatic conversion of input files (e.g. SDF to SMILES)
  - Automatic “chunking” of parallelizable algorithms
  - Automatic routing of jobs / chunks to the appropriate hardware (e.g. Linux, Windows, GPU, ...) based on historical metrics for a given protocol and the job requirements (memory, CPUs, etc)
  - Every “Protocol” has self-describing inputs and outputs so that clients can inspect to build inputs widgets or prepare for output results
  - Computational services uses a small amount of meta-data so that the same API is used for all services and clients do not have to be recompiled with each new service
  - Marting (ETL system) built on-top of computational protocols which can use KNIME and/or other (e.g. shells, SQL, ...)
Acknowledgments

**KNIME Development**
loki der quaeler

**Computational Services / Marting Development**
loki der quaeler
Kenny Leung
Karen Do
Francis Van Aeken

**Management**
Ron Blanford
Tim Tully

**Users**
Structure-Guided Chemistry
Computational Chemistry
Information Management