

Development of a KNIME Workflow using JChem Nodes for Evotec's Compound Library Enhancement Program



Michael P. Mazanetz, Oliver R. Barker, Ian Berry, Richard Law¹, James Madden, Lester Marrison², Mark Whittaker¹

¹Evotec (UK) Ltd., 114 Milton Park, Abingdon, Oxfordshire OX14 4SA, United Kingdom; ²Evotec India, Thane, India

Aim of The Library Enhancement Program

The Library Enhancement Program aims to constantly enhance Evotec's screening library. The library consists of over 250,000 small molecules designed and managed to provide high quality hits that will reduce downstream costs. Evotec's screening library is differentiated from other screening libraries through quality, diversity and novelty, both in terms of IP and of the biological target being screened. Evotec also provides access to a library of 30,000 fragments for screening using its fragment based drug discovery platform, EVOLution™. New chemotypes are constantly added to the library to ensure that it is diverse and novel with respect to screening – this is the ongoing Library Enhancement Program. In addition we have developed methods to yield a focused library targeting Protein-Protein Interactions. This poster demonstrates how we are proposing to integrate a KNIME [1] workflow into the Library Enhancement Program at Evotec.

Why Use a KNIME Workflow and JChem Nodes?

- Manage and support projects
 - Workflows can be rapidly developed and integrated into current schema
- Integrates to our compound ordering system, EVOsource and our Compound Registration database, CCD. Chemists can utilise web-based functions
- Centralising development; reducing the chance for duplication of work
- Single type of software architecture
 - Language and architecture independence between KNIME node development and workflows
- Multidisciplinary workflows (e.g., structural biology, cheminformatics, data mining, statistics, bioinformatics, etc.)

Library Enhancement Program Process

The Library Enhancement Program can essentially be broken down into 3 main phases. Phase 1 is the design of chemical templates and the enumeration of virtual libraries from commercial reagents. Phase 2 involves the refinement of the reagent lists. Phase 3 involves the clustering and selection of the reagents (based on druglikeness, diversity and, in some cases, target focus) for reaction with the templates. A proposed workflow for each of these Phases is described below.

Phase 1 - Library Enumeration (Figure 1)

- Chemistry designed templates and reaction schemes
- Assessment of reagent access and reaction feasibility
- *In silico* library enumeration
- Return data to computer-aided drug design teams

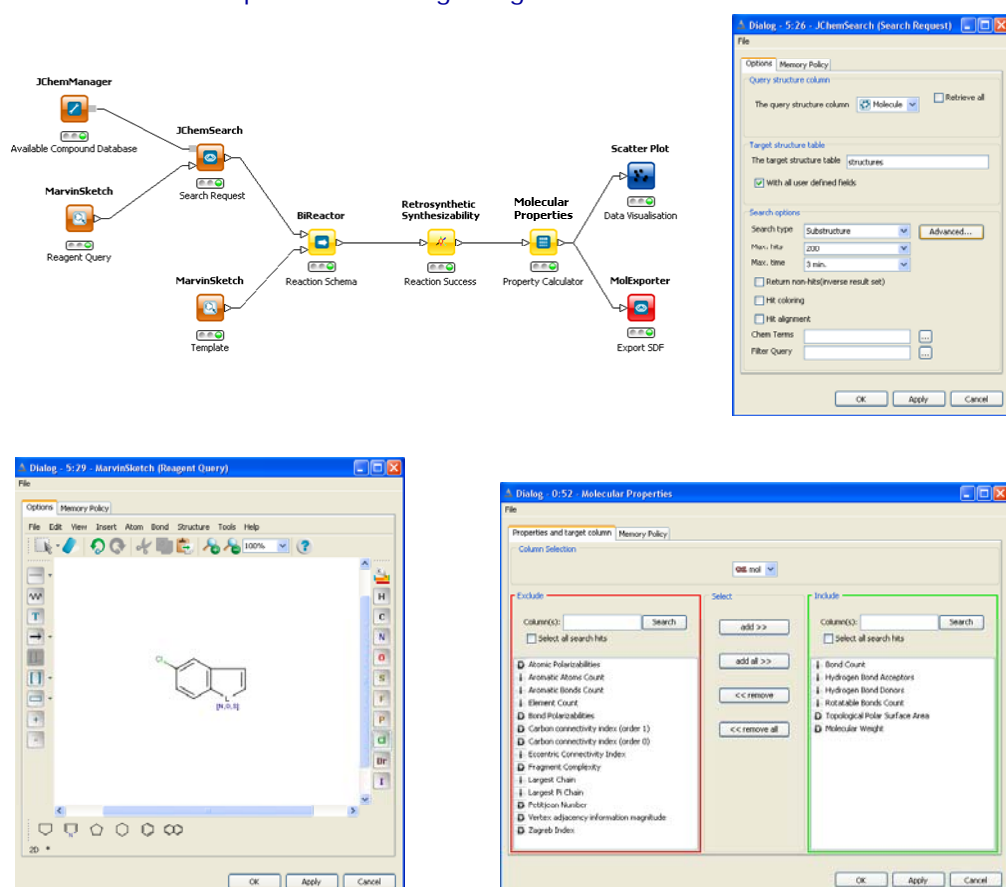


Figure 1
Phase 1: Template and reaction schema design. The JChem Extensions [2] set of KNIME nodes consists of ca. 70 nodes. The nodes include basic functions for drawing reagent and template queries in MarvinSketch, to special functions using ChemAxon's [3] software tools such as Reactor to enumerate virtual libraries. Synthetic feasibility is measured using a MOE node [4]. Data can be extracted from and returned to JChem managed databases, and queries made using JChemSearch. The built-in KNIME chemistry CDK is useful for calculating simple molecular properties.

Phase 2 - Reagent List Refinement

- Reagents converted to SD file formats in 2D using Corina. Filters are applied to screen for duplicates, remove salts, and check tautomeric states (Figure 2)
- Enumeration library is filtered into drug-, lead- and fragment-like diversity classifications based on in-house descriptors (Figure 3)
- Reagents are then ranked according to the number of times it yields a product of a certain diversity classification and those which are either Amber/Red/Green (Figure 3)
- Ranked reagents are then sent to the chemists for reassessment to produce a reduced reagent list

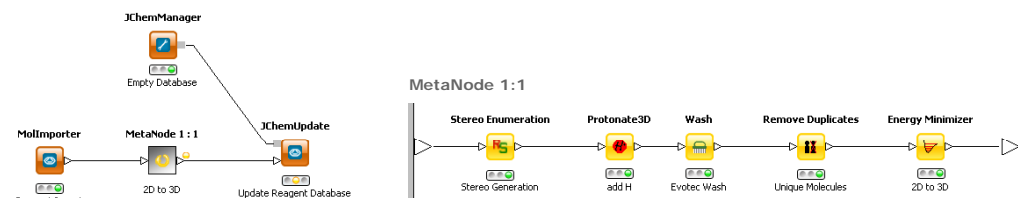


Figure 2
Phase 2: 2D to 3D conversion and reagent refinement. 2D to 3D conversions can be done using the MOE functions wrapped around ChemAxon functionality for reading files and exporting data to databases.

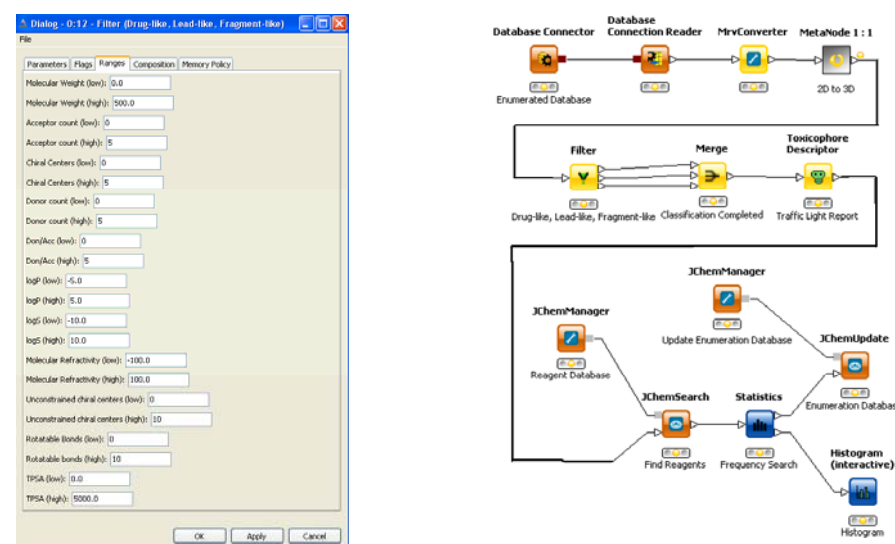


Figure 3
Phase 2: Evotec's property filters can be used to categorise the enumerated compound database. Reagents can then be selected which give return the greatest number of final compounds with the greatest diversity, both structurally and as categorised for each of Evotec's libraries. The results can then be inspected using the reporting functionality in KNIME and potentially exported into an ELN or managed database like Instant JChem.

Phase 3 - Reagent Clustering and Recommendation (Figure 4)

- Fingerprints are generated from the products of the reduced lists and compared to both the Screening Deck and the Screening Supplier Database, EVOsource, to find novel compounds
- The reagents are then clustered and compounds returning high diversity and ranking are prioritised for synthesis in India

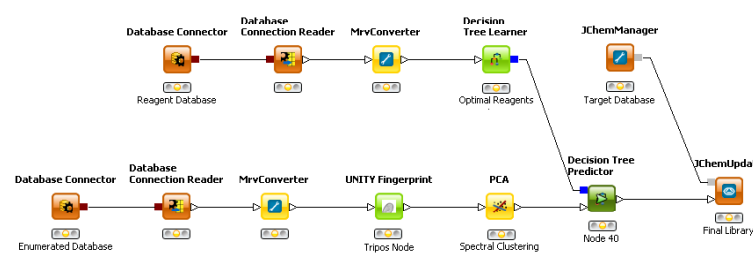
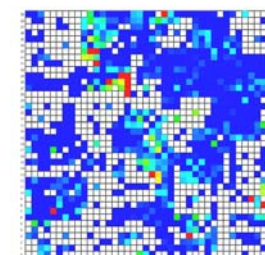


Figure 4
Phase 3: Evotec scripts are used to calculate UNITY fingerprints [5] and cluster the results using a Spectral Clustering algorithm [6]. Data mining tools can be developed to prune the reagent list to return only interesting compounds which would then be used to design a final library for chemical synthesis.

Library Enhancement Program Next Steps

KNIME provides a powerful and flexible workflow to evaluate and expand Evotec's compound libraries. Further work would involve re-analysing the libraries generated with respect to the Kohonen maps describing the focused sets (i.e., kinase, GPCR) in order to determine if other compounds, added to the library since the sets were created, should be added to the sets.



References

- 1) Berthold, M., *et al.*, KNIME: The Konstanz Information Miner, Proceedings Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Freiburg, Germany, Springer-Verlag, 2007.
- 2) JChem Extensions to KNIME ; INFOCOM CORPORATION http://www.infocom.co.jp/index_e.html
- 3) MOE. <http://www.chemcomp.com>
- 4) ChemAxon: <http://www.chemaxon.com>
- 5) TRIPOS: <http://tripos.com/>
- 6) Brewer, M., *J. Chem. Inf. Model.*, **2007**, *47*, 1727-33

Acknowledgements

Thanks goes to the eScience Group at Evotec who have assisted with this work