EXTRACTION, ANALYSIS, ATOM MAPPING, CLASSIFICATION AND NAMING OF REACTIONS FROM PHARMACEUTICAL ELNS

Roger Sayle[1], Daniel Lowe[1], Noel O'Boyle[1], Mick Kappler[2], Anna Paola Pellicciolli[3], Nick Tomkinson[4] and Daniel Stoffler[5]

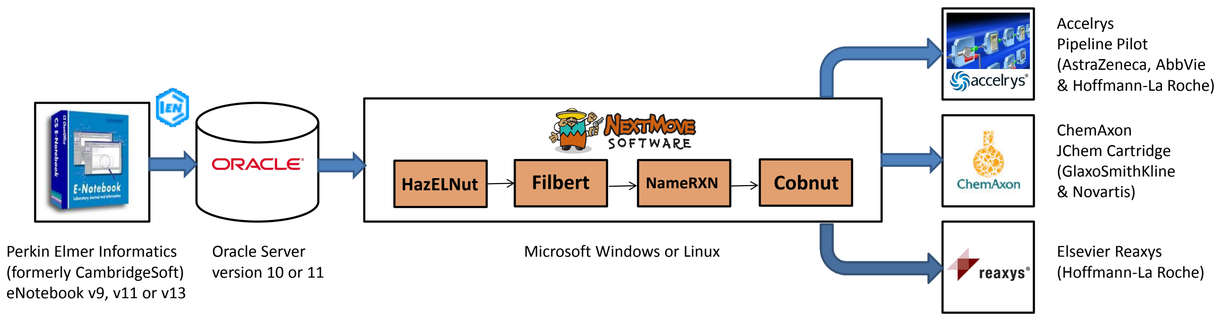
[1] NextMove Software Ltd, Cambridge, UK. [2] Hoffmann-La Roche, Nutley, USA. [3] Novartis NIBR, Basel, Switzerland.
[4] AstraZeneca R&D, Alderley Park, UK. [5] F. Hoffmann-La Roche, Basel, Switzerland.

1. Abstract

Electronic Laboratory Notebooks (ELNs) are widely used in the pharmaceutical industry for recording the details of chemical synthesis experiments. The primary use of this information is often for the capture of intellectual property for future patent filings, however this data can also be used in a number of additional applications, including synthetic accessibility calculations, reaction planning and reaction yield/optimization. Not only does a pharmaceutical ELN capture those classes of reactions suitable for small scale medicinal chemistry, but it is also uniquely a source of information on failed and poor yield reactions; an important source of data rarely found in the scientific literature or commercial reaction databases.

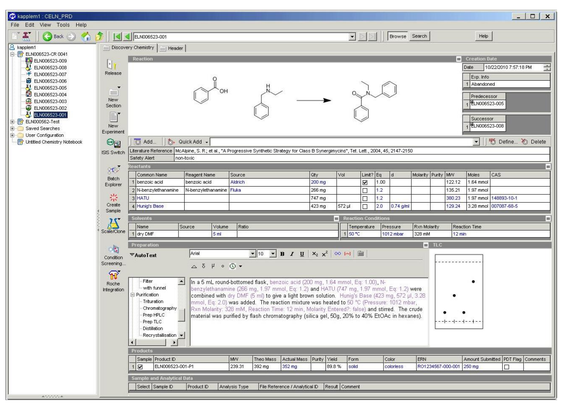
2. Method overview

In this work we describe the use of a suite of programs for extracting reaction information and associated textual and numeric data from the PKI/CambridgeSoft eNotebook ELN, converting and normalizing it to "open" flat files that can be read by third party applications, such as Accelrys/MDL RD files or reaction SMILES.



3. HazELNut: Reaction export from PKI/CambridgeSoft eNotebook ELN

The primary step in exporting a reaction database from an ELN is performed by **HazELNut**, which interprets the chemist's hand-drawn sketch into a connection table. For RD and SD formats, this includes writing the textual, numeric, tabular and molecular data as tagged fields in the output file. The process can work either from XML files exported from the client, or more typically by querying the Oracle server via OCI. A large pharmaceutical ELN can be exported overnight (or transatlantic over a weekend), though incremental export allows a day's or week's experiments to be written in a few minutes. Experimental write-ups (and other rich text) are converted from RTF to text or HTML, superatoms are expanded, labels preserved, and so on. The images below show an ELN reaction, and some of the exported fields as they appear in the corresponding RD file format output.



```
$DTYPE HEADER:EXPERIMENT.HEADER:CREATION.DATE
$DATUM 22-Oct-2010
$DTYPE HEADER:EXPERIMENT.HEADER:CREATION.TIME
$DATUM 19:58:18 -0500
$DTYPE DISCOVERY.CHEMISTRY:REACTANTS(1):CHEMICAL.STRUCTURE
$DATUM c1ccc(cc1)C(=O)O
$DTYPE DISCOVERY.CHEMISTRY:REACTANTS(1):NAME
$DATUM benzoic acid
$DTYPE DISCOVERY.CHEMISTRY:REACTANTS(1):MOLECULAR.WEIGHT:VALUE
$DATUM 122.12
$DTYPE DISCOVERY.CHEMISTRY:PRODUCTS(1):NAME
$DATUM N-benzyl-N-ethylbenzamide
$DTYPE DISCOVERY.CHEMISTRY:PRODUCTS(1):CHEMICAL.STRUCTURE
$DATUM CCN(Cc1ccccc1)C(=O)c2ccccc2
$DTYPE DISCOVERY.CHEMISTRY:PRODUCTS(1):MOLECULAR.FORMULA
$DATUM C16H17NO
$DTYPE DISCOVERY.CHEMISTRY:PRODUCTS(1):MOLECULAR.WEIGHT:VALUE
$DATUM 239.31
$DTYPE DISCOVERY.CHEMISTRY:PRODUCTS(1):YIELD:VALUE
$DATUM 89.8%
$DTYPE DISCOVERY.CHEMISTRY:PREPARATION
$DATUM In a 5 mL round-bottomed flask, benzoic acid (200 mg, 1.64 mmol
N-benzylethanamine (266 mg, 1.97 mmol, Eq: 1.2) and HATU (747 mg, 1.97
mmol, Eq: 1.2) were combined with DMF (5 ml) to give a light brown
solution.  Hunig's Base (423 mg, 572 µl, 3.28 mmol, Eq: 2.0) was added.
The reaction mixture was heated to 50 oC (Pressure: 1012 mbar, Rxn
Molarity: 328 mM, Reaction Time: 12 min, Molarity Entered?: false) and
stirred.  The crude material was purified by flash chromatography
```

4. Filbert: Reaction file format conversion

Whilst **HazELNut** can export reactions in a variety of file formats, its goal is to "dump" the raw data; processing and file format conversion is performed by **Filbert**. This program interconverts MDL RXN, RD and SD file formats, ISIS Sketch, CDXML and reaction SMILES. The roles of agents, catalysts and solvents drawn above and below a reaction arrow can be preserved using ChemAxon's RXN file extensions, stripped or treated as reactants. Co-ordinates can optionally be centered, rescaled or regenerated algorithmically. Atom maps can be removed. Molecules from the reactant/agent tables can be added to the sketch if missing.

5. Cobnut: MDL/Accelrys RDF file format tag stripping/renaming

Typically, ELNs are heavily customized with custom experiment types to capture the data needs of their target scientists. The raw export of this data results in RD files where the field/tag names for yields, volumes, chemist's names, experimental write-ups, etc. vary from reaction to reaction. The utility **Cobnut** was implemented to simplify and speed-up the process of normalizing (renaming) RD tags from different data sources, and stripping out those that aren't required.

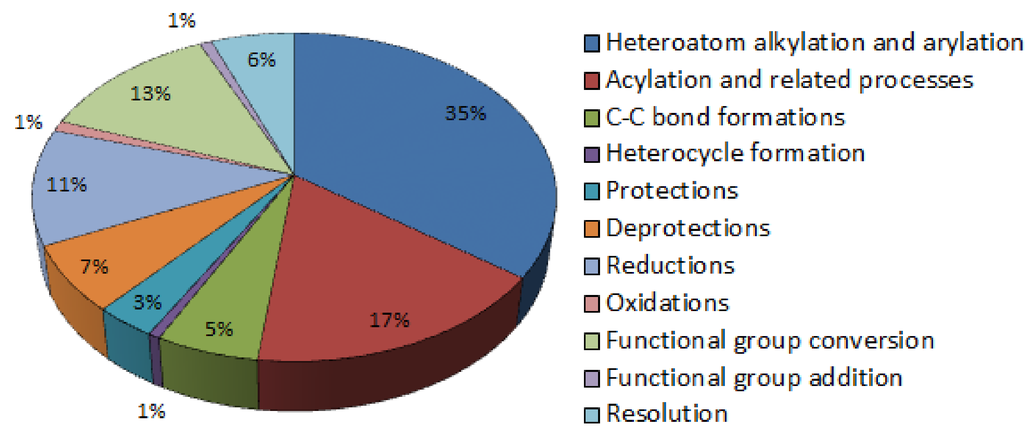
6. NameRXN: Reaction naming and classification

One useful form of reaction analysis is to recognize each experiment in the ELN as an instance of a named or known reaction, such as a Diels –Alder cycloaddition, Suzuki coupling or chiral separation. The **NameRXN** program uses a dictionary of SMIRKS-like transformations to annotate reactions in RD, SD, RXN or reaction SMILES format with a name, a reaction category and where applicable an identifier into the Royal Society of Chemistry's RXNO ontology.

The top five reactions in a major pharmaceutical company's ELN were:

	ID	Reaction Name	Reaction Category	RXNO #
#1	1.3.1	Buchwald-Hartwig amination	N-arylation with Ar-X	0000192
#2	7.1	Nitro to amino	Nitro to amine reduction	0000337
#3	1.7.6	Williamson ether synthesis	O-substitution	0000090
#4	2.1.2	Carboxylic acid + amine reaction	N-acylation to amide	
#5	2.2.3	Sulfonamide Shotten-Baumann	N-sulfonylation	0000165

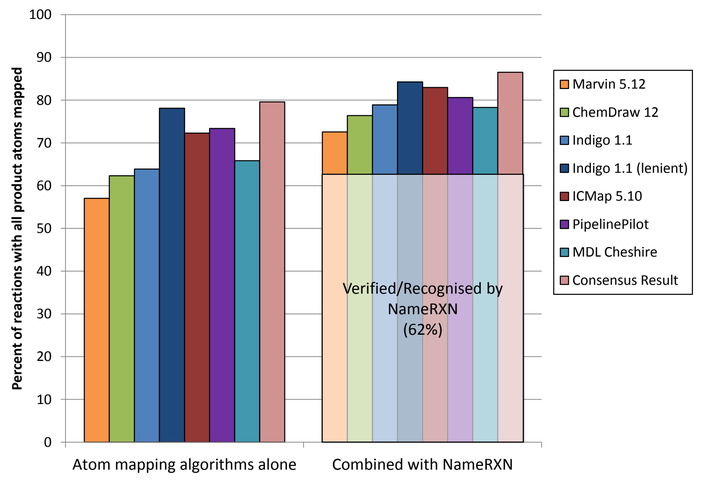
Using the reaction classification categories of Carey et al.[1] the contents of the ELN may be presented as a pie-chart of the kinds of transformations it contains.



7. Third-party atom mapping

One way to investigate the quality of the reaction sketches in an ELN is to apply an automatic atom-atom mapping algorithm, and assess the fraction of unmapped product atoms and average number of carbon-carbon bonds broken. At the Fall 2012 ACS meeting in Philadelphia, we presented an initial comparison of third-party atom mapping software for the purpose of identifying incorrectly drawn reactions.

Here we summarize recent improvements to this approach, using consensus methods to combine the results of multiple atom mapping algorithms, and combining atom-mapping with reaction naming for ELN quality assurance.



8. Summary and future work

The use of the HazELNut suite of tools allows the synthetic chemistry data locked up in corporate ELNs to be exploited in scientifically novel ways. Current efforts include optimization of reaction conditions by plotting yields against catalyst and solvent in SpotFire, and research to reduce the number of steps and application of low yield reaction strategies by identifying/reusing in-house insights/expertise.

9. Bibliography

1. John S. Carey, David Laffan, Colin Thomson and Mike T. Williams, "**Analysis of the Reactions used for the Preparation of Drug Candidate Molecules**", *Organic & Biomolecular Chemistry*, Vol. 4, pp. 2337-2347, 2006.
2. Stephen D. Roughley and Allan M. Jordan, "**The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates**", *Journal of Medicinal Chemistry*, Vol. 54, 3451-3479, 2011.
3. Mikko J. Vainio, Thierry Kogej and Florian Raubacher, "**Automated Recycling of Chemistry for Virtual Screening and Library Design**", *Journal of Chemical Information and Modeling (JCIM)*, Vol. 52, No. 7, pp. 1777-1786, June 2012.

NEXTMOVE SOFTWARE

www.nextmovesoftware.co.uk
www.nextmovesoftware.com

NextMove Software Limited
Innovation Centre (Unit 23)
Cambridge Science Park
Milton Road, Cambridge
England, UK CB4 0EY