Extracting Knowledge from Reaction Databases: Developments from InfoChem

CICAG meeting 3rd July 2013

Stephanie North, Allyl Consulting Ltd, representing InfoChem in the UK

H. Kraut, H. Matuszczyk, H. Saller, J. Eiblmaier, P. Loew
InfoChem GmbH, Landsberger Strasse 408, Munich, 81241, Germany
Outline

Setting the scene: millions of reactions
- Rapid growth of published reaction databases e.g. SPRESI, 4.2 M reactions
- Electronic Laboratory Notebooks now provide corporate reaction databases
- Reaction knowledge a key driver for chemists’ innovation and decision making processes - but how best to unlock and exploit resource?
- InfoChem’s CLASSIFY technology provides the solution

Today’s presentation
- Introduce InfoChem
- Describe the reaction classification concept
- Show how CLASSIFY works and analyse the results
- Benefits of extending CLASSIFY to corporate reaction databases
InfoChem at a Glance

**Company**
- specializes in chemoinformatics
- founded in 1989
- based in Munich, Germany
- majority owned by Springer Science + Business Media
- subsidiary InfoChimia since 1990

**People**
- 24 full time employees (Munich office)
- 2 representatives in UK
- 3 representatives in Sweden
- 60 freelance abstractors (residing offshore)

**Business Areas**
- Software products
- Projects
- Text/Data mining
- Database building

**Revenues**
Approx. 3.3 million Euros (2012):
- Software products (30%)
- Database building (35%)
- Projects (12%)
- Text/Data mining (23%)
Business Areas

**Software**
- \( \text{IC}_{\text{FSE}}, \text{IC}_{\text{CARTRIDGE}}, \text{IC}_{\text{CHEMDESK}} \)
- \( \text{IC}_{\text{MAP}}, \text{CLASSIFY}, \text{IC}_{\text{NameRXN}} \)
- \( \text{IC}_{\text{SYNTH}}, \text{IC}_{\text{EDIT}}, \text{IC}_{\text{TOOLS}} \)

**Services**
- Project development
- Consulting
- Database building

**Content**
- SPRESI\(^{\text{web}},\) SPRESI\(^{\text{mobile}}\)
- Chemisches Zentralblatt Structural Database

**Text/Data Mining**
- Chemical entity recognition
- Name to structure conversion
- Image to structure conversion
Fundamentals of reaction Classification Concept

• **What does it do?** Organises and analyses large reaction databases according to the type or ‘class’ of transformation

• **How does it work?** Groups similar transformation types together sharing same reaction centres and same environment

• **What are the benefits?** Gain knowledge about the type, quality and diversity of reactions to aid inventive process and decision making
Background: Driving Force for Classification

Status 1988:

- Technology limitations of systems at the time
  - REACCS, ORAC, SynLib handle ca 100K reactions

- InfoChem acquired SPRESI with 1.8 million reactions
  - How could we provide 1.8 million reactions to customers?
  - How to reduce/select reactions so systems could handle?
Concept of Reaction Classification

• Develop a chemically intelligent way to choose an example reaction from a given class of reactions

• Group similar reactions e.g. like chemistry is published

• Select one example transformation to represent the group

• The result is a smaller selective database consisting of one example for each transformation type
CLASSIFY Classification algorithm: how does it work?

- For each reaction, determine and map reaction centers
- Calculate atom hash codes for the reaction centre
- Atoms included depend on extent of reaction centre’s chemical environment considered, resulting in different sized search hit lists (clusters):

0-Sphere (BROAD)
Reaction centers only

1-Sphere (MEDIUM)
Reaction centers plus alpha atoms, excluding hydrogens

2-Sphere (NARROW)
Reaction centers plus beta atoms, excluding hydrogens and consecutive sp\(^3\)-atoms

- The sum of the hash codes from these calculations provides the unique InfoChem Reaction Classification Code. Three codes are assigned to each reaction
Classification results

Large database e.g. SPRESI

Classification results

CLASSIFY

Each reaction processed
ClassCode determined
Cluster according to ClassCodes

Groups of reaction types according to ClassCodes

Extract knowledge from classification

Select one sample for each ClassCode

Smaller representative database e.g. ChemReact

Further analysis of clusters

Codes | # RXNs
-------|------
3257410 82969498 | 13,000
2945604 35478524 | 12,000
2969666 48127528 | 11,000
4257410 82969497 | 10,000
5945604 35478523 | 9,500
6969666 38127527 | 9,000
3257410 72969496 | 8,500
5945604 25478522 | 8,000
4969666 38127526 | 7,000
2257410 72969495 | 6,000
4945604 25478521 | 5,000
3969666 38127524 | 4,000
3257410 72969494 | 3,000
2945604 25478522 | 2,000
2969666 38127523 | 1,000
4257410 72969495 | 500
5348521 83070506 | 90
7560743 05292708 | 25
8671845 16303819 | 1

.............  ....
ClassCodes: representative database creation

- Apply NARROW Classification to SPRESI
- Most frequent ClassCode reduction of a nitro group:

ONE representative sample reaction per ClassCode from SPRESI used to create ChemReact database. Selection criteria: yield, journal, year of publication etc.
### Sample reactions from Classcodes with high populations

<table>
<thead>
<tr>
<th>ClassCode</th>
<th>Reaction examples in SPRESI (Narrow)</th>
<th>ChemReact sample reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>3257410</td>
<td><img src="image1.png" alt="Reaction 1" /></td>
<td><img src="image2.png" alt="ChemReact 1" /></td>
</tr>
<tr>
<td>82969498</td>
<td><img src="image3.png" alt="Reaction 2" /></td>
<td><img src="image4.png" alt="ChemReact 2" /></td>
</tr>
<tr>
<td>2945604</td>
<td><img src="image5.png" alt="Reaction 3" /></td>
<td><img src="image6.png" alt="ChemReact 3" /></td>
</tr>
<tr>
<td>35478524</td>
<td><img src="image7.png" alt="Reaction 4" /></td>
<td><img src="image8.png" alt="ChemReact 4" /></td>
</tr>
<tr>
<td>2969666</td>
<td><img src="image9.png" alt="Reaction 5" /></td>
<td><img src="image10.png" alt="ChemReact 5" /></td>
</tr>
<tr>
<td>48127528</td>
<td><img src="image11.png" alt="Reaction 6" /></td>
<td><img src="image12.png" alt="ChemReact 6" /></td>
</tr>
<tr>
<td>8,989</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8,947</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis: Frequency distribution of SPRESI ClassCodes (narrow)

Database diversity
- Only 126 class codes > 1,000 example reactions
- First 180 ClassCodes combined contain 348,927 reactions
- Core chemistry within fairly narrow range of reaction types
Classification of SPRESI to ChemReact since 1988

<table>
<thead>
<tr>
<th>Release</th>
<th>SPRESI</th>
<th>ChemReact</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>1,800 K</td>
<td>370 K</td>
</tr>
<tr>
<td>1991</td>
<td>2,300 K</td>
<td>400 K</td>
</tr>
<tr>
<td>2002</td>
<td>3,500 K</td>
<td>450 K</td>
</tr>
<tr>
<td>2012</td>
<td>4,260 K</td>
<td>524 K</td>
</tr>
</tbody>
</table>

- Steady increase in diversity of organic synthesis specific reactions
- Growth of ChemReact not proportional to SPRESI database (overall ca 1.4x vs. 2.4x)
- New reaction references in literature do not necessarily generate new ClassCodes (chemistry)
- Organic synthesis is described by a comparatively small amount of reaction types
Classification for Corporate Reaction Databases

- Organisations are making significant investments in ELNs to capture intellectual property
- ELNs generate large repositories of corporate reactions
- Accessibility of reactions depends on type of ELN

How to benefit and unlock this valuable resource for improved decision making and reaction planning?

- CLASSIFY solution to automatically categorize and analyse corporate reactions
- ICCartridge and Fast Search Engine for storage and fast retrieval of corporate chemical reactions and data
Generation of ELN Classified Reaction database

ELN reactions

Extract reactions

RD Files

CLASSIFY

Oracle Classified reaction database

Integrate with InfoChem Cartridge and Fast Search Engine

RD Files enriched by ClassCodes
Possible Frequency Distribution of a Corporate database

- Strengths
- Possible gap or weakness
- Emerging or obscure chemistry
- Quality analysis of singletons
CLASSIFY unlocks corporate reaction knowledge

Analysis of ClassCodes and frequency distribution provides knowledge about:

- Diversity, strengths and gaps in corporate reactions
- Optimum yields, reagents and conditions for a reaction type
- Low frequency reactions may indicate
  - Emerging valuable new reactions
  - Obscure but useful chemistry not often repeated
- Database Quality
  - Few Singletons (one reaction in a ClassCode) indicate database has a certain quality standard
  - Many singletons may indicate data quality problems
Singleton reactions: deeper look

• New/unique/weird but valid chemistry

\[
\text{Da-Ming Gou and Ching-Shih Chen: Synthesis of L-} \alpha \text{-phosphatidyl-D-myo-inositol 3,4,5-trisphosphate, an important intracellular signalling molecule, J. Chem. Soc., Chem. Commun., 1994, 2125}
\]

• Erroneous or inconsistent reaction or data entry
  • roles, missing reactants, different solvent/catalyst/reagent/reactant definitions

• Inconsistent normalization rules

• Inadequate or incorrect mapping

• Multistep reactions as overall reactions e.g. overall step in 13 step reaction:
Summary of CLASSIFY Benefits

**Diversity** of reaction types identifies strengths and gaps in corporate knowledge
- Does the business need to focus on alternative reaction classes?
- Location of failed internal reactions (how/why) to drive reaction improvement

**Quality**: singleton reaction types may indicate erroneous data
- Drive data quality improvements and business rule compliance

**Improved route design and innovation**
- Maximum value obtained from corporate reaction expertise
- Analysis of reactants and conditions within a reaction class to determine best practice leading to simplification and economy of processes
- Chemists alerted to emerging or novel chemistry, thus speeding discovery
- Enables selection of viable reactions and efficient route maps

**Knowledge sharing**
- Only way to link and compare reaction databases
- Reaction knowledge and expertise shared across organisation
Acknowledgments: InfoChem Team

- Hans Kraut
- Henry Matuszczyk
- Heinz Saller
- Josef Eiblmaier
- Valentina Eigner-Pitto
- Ulf Frieske
- Peter Loew

Look out for publication underway

Thank you