ARChem
Route Designer

Antony Williams
On behalf of A. Peter Johnson
University of Leeds / SimBioSys Inc.

Howard Y. Ando (Pfizer Inc.), Zsolt Zsoldos,
Aniko Simon, James Law, Darryl Reid, Yang Liu,
Sing Yoong Khew, Sarah Major
Computer Aided Organic Synthesis Design (CAOSD) assists chemists in finding synthetic routes to target compounds.

The first CAOS system (LHASA) was introduced by EJ Corey nearly 40 years ago.

Since then...
CAOSD systems

- LHASA (Logic and Heuristics Applied to Synthetic Analysis) (E.J. Corey et al.)
- SynChem (Gelenter et al.)
- IGOR (Ugi et al.)
- EROS and WODCA (Gasteiger et al.)
- SynGen (Hendrickson et al.)
- Arthur (commercial: Synthematix)
- Others…
Typical Retrosynthetic Analysis

Retrosynthetic analysis works backward from the target and generates increasingly simple precursors.
Retrosynthetic Analysis versus Reaction Databases

- Reaction databases are popular aids in reaction planning
- Databases are large, highly curated and good tools exist for searching and data-mining
- Compare with “general prediction technology”
  - LogP
  - NMR spectra etc…
CAOSD Systems

- There is little *routine* use of retrosynthetic analysis tools. Why?
- Chemists ARE the knowledge base?
- Reaction database data mining suffices?
- Who trusts computers anyway?

- While reaction databases are very valuable couple “predictions” and reference data
Knowledge base creation

Retrosynthetic analysis is driven by rules describing scope, limitations and structure changes associated with a reaction.

Rules have to be manually encoded.

Only experienced synthetic chemists have the knowledge to create good rules.
Goals of Route Designer

• perform **rule based** retrosynthetic analyses of target molecules back to **readily available materials**

• provide **fully automated** generation of retrosynthetic reaction rules by analysis of a reaction database – avoid **time consuming manual creation**

• provide the user with literature examples of the transformations suggested by the retrosynthetic analysis

• provide a set of alternative routes to a given target
System Design Overview

User input: Molecule

Starting Materials: Aldrich, Acros Lancaster

Reaction DB (MOS, Beilstein, CASREACT etc.)

Automatic Rule Extraction

Route Des. search

Reaction Rules

Output: Reaction pathways

http://www.simbiosys.ca
Starting Materials

- Automatic selection of starting materials from commercially available compounds is important for retrosynthetic analysis.
- With known starting materials as a basis analysis is directed at portions of the target molecule that cannot be made from available starting materials.

Aldrich  
Lancaster  
Acros  
Others  

http://www.simbiosys.ca
Previous works:

Rule extraction from reaction databases


2) Gelernter et al: Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Machine Learning

1) The extraction process converts many reactions into a few rules.  
2) The combinatorial explosion of the retrosynthetic search process is controlled.

Methods of Organic Synthesis (MOS)  
~42k Reactions  
Large reaction DBs OK  
(millions of reactions supported)

Cluster reactions by chemical equivalence

Generic leaving groups reduce the number of rules

4k rules from 47k reactions
Identifying Reaction Cores

- The Core - atoms that undergo “changes” during a reaction
- Atom mappings identify atoms attached to bonds changed, made or broken in the reaction

Extracted Core:
Extension to “non-reacting” atoms

- Initial core is extended to include structural features *essential* for the reaction (difficult process)

- Empirical rules attempt to capture these features
Reaction Core Extension

Reaction

Extracted Core

Extended Core
Generic rule

Nucleofuge (NF) - a leaving group which carries away the bonding electron pair

Reaction Rule
Clustering Cores

- Established approaches (Morgan numbers) are used to identify the reaction core and the entire extended core.
- Clustered by exact matching of the extended cores and different extended cores may be combined.
- Rules specifying bond making and breaking operations are constructed.
Rule Generation Summary

Reaction DB in RDF file format

Esterification examples clustered:

Other examples clustered:

Esterification Reaction Rule:

Some other rule

http://www.simbiosys.ca
RuleGenerationfromMOSDB

- The Methods in Organic Synthesis database contains ca. 42k reactions

- Rule extraction performed on this database gave ~3800 rules
## Reaction Classification

- Automatic classification of MOS gave:

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th># clusters</th>
<th>Active ?</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD</td>
<td>Regular Disconnective</td>
<td>2764</td>
<td>Always as needed</td>
</tr>
<tr>
<td>FGI</td>
<td>Functional Group Interchange</td>
<td>505</td>
<td>needed as needed</td>
</tr>
<tr>
<td>FGA</td>
<td>Functional Group Addition</td>
<td>261</td>
<td>needed as needed</td>
</tr>
<tr>
<td>BFGI</td>
<td>Bond-oriented FGI</td>
<td>238</td>
<td>needed as</td>
</tr>
<tr>
<td>FGR</td>
<td>Functional Group Removal</td>
<td>96</td>
<td>needed</td>
</tr>
<tr>
<td>UPC</td>
<td>Unmapped Product Carbon</td>
<td>41</td>
<td>NO</td>
</tr>
<tr>
<td>DP</td>
<td>Deprotection Reaction</td>
<td>27</td>
<td>NO</td>
</tr>
<tr>
<td>SAP</td>
<td>Simple Alkane Product</td>
<td>16</td>
<td>NO</td>
</tr>
<tr>
<td>BCR</td>
<td>Bare Carbonyl Reaction</td>
<td>13</td>
<td>NO</td>
</tr>
<tr>
<td>UER</td>
<td>Unactivated Educt Reaction</td>
<td>7</td>
<td>NO</td>
</tr>
<tr>
<td>NR</td>
<td>Rearrangement</td>
<td>3</td>
<td>needed</td>
</tr>
</tbody>
</table>

**Total:** 3971
Non-disconnective transformations such as FGI and FGA must either

- lead to an available starting material
- change functionality so that a disconnective transform can be triggered
- uncontrolled use of FGI / FGA would give a combinatorial explosion
Top 5 Clusters (MOS example)

<table>
<thead>
<tr>
<th>D</th>
<th>#Ex</th>
<th>Act</th>
<th>Cmts</th>
<th>Cat</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>6325</td>
<td>290</td>
<td>Yes</td>
<td>0</td>
<td>RD</td>
<td><img src="image1" alt="Reaction 1" /></td>
</tr>
<tr>
<td>6327</td>
<td>260</td>
<td>Yes</td>
<td>0</td>
<td>FGI</td>
<td><img src="image2" alt="Reaction 2" /></td>
</tr>
<tr>
<td>6341</td>
<td>259</td>
<td>Yes</td>
<td>0</td>
<td>FGI</td>
<td><img src="image3" alt="Reaction 3" /></td>
</tr>
<tr>
<td>6340</td>
<td>172</td>
<td>Yes</td>
<td>0</td>
<td>FGI</td>
<td><img src="image4" alt="Reaction 4" /></td>
</tr>
<tr>
<td>6336</td>
<td>170</td>
<td>Yes</td>
<td>0</td>
<td>RD</td>
<td><img src="image5" alt="Reaction 5" /></td>
</tr>
</tbody>
</table>
In the absence of frequent user intervention the combinatorial nature of this approach prohibits very broad searches (many alternative reactions) as well as very deep searches (many steps between starting materials and products).
Exhaustive Search

- Control the combinatorial explosion!!!
  - Fundamental transformations (esterification, amide formation ...) applied at first stage to break up target
  - User selectable search constraints
    - Depth limit for search
    - Minimum number of examples for a rule to be used
    - Strategic disconnections / preserved bonds
  - FGI/FGA (eg CH2OH ==> CO2H) restricted to cases triggering subsequent disconnective transforms or finding starting materials
  - Unstable functionality generated **must** be removed at next step (some organometallics or acyl chlorides)
Preserve and Target Bonds

- **Preserve** selected bonds and **Target** other bonds during analysis
Ordering of Results

- Prioritization of solutions
  - Disconnective transforms before FGI/FGA
  - Minimize wastage (atom efficient reactions)
  - Starting material coverage
  - Prefer thoroughly explored chemistry (based on example count)
  - The more bonds broken in the retrosynthetic transformation the better
Real World Examples

- Tested on hundreds of examples including drugs, natural products, publications

- An example:
  - Zatosetron, a potent, selective and long acting 5HT receptor antagonist from Lilly used in the treatment of nausea and emesis associated with oncolytic drugs
  - < 5 mins on standard PC,
Zatosteron

- The route shown is virtually identical to the published route for Zatosetron (Robertson et al. J. Med. Chem., 1992; Vol. 35, pp 310-319)

- Minor differences include:
  - using the 3-bromo-2-methylpropene rather than the chloro version
  - the aminotropene was not found in our starting material database but the tropinone precursor was
How Fast? How Complex?

- Retrosynthetic analyses from minutes to hours based on complexity and constraints
- System is based on construction of skeletal connections not on stereochemistry
- Does not take into account conditions – temperature, pressure etc
- Estimates yields based on database reactions but they are ESTIMATES!
How Fast? How Complex?

- Very intuitive and fast to initiate request
- Can be expanded with other catalogs of starting materials easily
- “Training” via cluster analysis is not difficult but is not an everyday task either
- Can be tested using an online platform
- Proven application at a number of companies already
Under Development

- Indicate **preferred** starting materials to bias the search
- Improve clustering to fully capture chemical constraints including better regioselectivity and stereoselectivity – target is much smaller rule set
- Deal with interfering functional groups
- Order search results to reflect chemists’ preferences
Interfering functionality

- Compatible functionality is detected through comparison with reaction example databases
- Possible interfering functionality is inferred
- Search result rank is marginally weighted against interfering functional groups
Rule set optimizations

- Promotion of heteroaromatic rules using lower example threshold than other rules
- 300k rules generated from the Beilstein database gave >50k rules with heteroaromatic relevance
- Initial results show dramatic performance improvements
- Future extension to other rule categories is under investigation
Route Designer summary

- “Predictions” and reference data are proven approaches – extend to retrosynthetic analysis

- Route Designer already provides good routes for a variety of targets AND they are “predictions”

- Any updates in reaction databases can be easily incorporated into the rule base – “user training”

- Starting materials database can be enhanced and extended easily – there are 10s of catalog vendors which can be added
Acknowledgements

**Pfizer**
- Robert Wade
- Robert Dugger
- James Gage
- Peter Wuts
- David Entwistle
- Inaki Morano
- Richard Nugent
- Bryan Li
- Julian Smith

**Accelrys**
- Rob Brown
- Eric Jamois

**Symyx-MDL**
- Maurizio Bronzetti
- Jochen Tannemann

http://www.simbiosys.ca