Towards the Automated Deduction of Chemical Reaction Mechanism

(a chemical engineering perspective of high throughput technologies)

Allen Wright

School of Chemical Engineering and Advanced Materials,
University of Newcastle, Newcastle-upon-Tyne, NE1 7RU, England.

Basel 2007
Overview of the batch chemicals business

- Speciality, agrochemical and pharmaceuticals (many companies)
- Large sector of chemicals industry with $9-12bn UK share of $250bn global market
- Tens of thousands of low volume high value products
  - $10k to > $1m per tonne
- Development consumes most of the R&D resource
- Time to market a critical driver

![Chemical R&D Funds Pie Chart]

- Development 59%
- Applied Research 32%
- Basic Research 9%
Key success factors in manufacturing speciality chemicals

- Companies have traditionally relied on developing unique chemical expertise
  - High barrier to new entrants
  - Charge higher margins on specialities produced

- Operate the most cost effective production plant
  - Deters new competitors

But, eventually competitors do catch-up!
Process Development – dynamic resource management

- Retrosynthesis Software
- Synthetic Chemist
- Candidate routes
- Preliminary Decision Criteria
- Ranking Engine
- Interactive users
- Prioritised Workflow
- Decision Criteria
- Preferenced Ordered List

Chemical Engineering, Newcastle University
Differences in education

Chemical engineers
- Thermodynamics
  - Property prediction
- Transport processes
  - Heat & mass transfer
- Fluids
- Dynamics and Control
  - Systems analysis
- Separation processes
  - Distillation, drying, filtration etc
- **Reaction Engineering**

Chemists
- Organic chemistry
  - mechanism and synthesis
- Physical chemistry
  - Kinetics and thermodynamics
  - spectroscopy
- Inorganic
  - Main group, transitional and synthesis
- Analytical
Differences in approach

- **Chemical engineers**
  - Design
  - Mathematical (models)
  - Data users
  - Empirical
  - System (reaction) dynamics
  - Simple synthesis

- **Chemists**
  - Discover
  - Experimental
  - Data providers
  - Theoretical
  - Molecule properties
  - Novel synthesis
Generalised Mass and Energy balances

\[ \text{Accumulation} = \sum \text{rates} \]

\[ V \cdot \sum_{i=1}^{m} \frac{d[i]}{dt} = \sum_{j=1}^{n} \left( \sum_{j=1}^{n} F_j \cdot [i]_{j} - \sum_{k=1}^{p} F_k \cdot [i] + \sum_{l=1}^{q} R_l - [i] \cdot \frac{dV}{dt} \right) \]

- Mass flow rates
- Dilution rate
- Reaction rates
- Heat exchange rate
- Feed enthalpy
- Reaction heats
Reaction engineering: Cement and penicillin
Modern chemistry software tools

DATA

- Chemistry Knowledge
- Substructure Search
- Spectra prediction
- Diversity Search (cluster on props)
- Pattern Analysis (surface props)
- Classification
- Descriptors Generation
- 3D Structures Generation
- Similarity Search

Molecular Networks

Data mining

Accelrys

Chemical Engineering, Newcastle University
Modern chemical engineering software

\[ \ln y_j = \ln \left( \frac{\phi_j x_j}{\Phi_j} \right) + 0.5 \sum_{j=1}^{n} \ln \left( \frac{a_j}{\phi_j} \right) + \sum_{j=1}^{n} \left( \frac{\phi_j \Phi_j}{\sum_{k=1}^{n} \phi_k \Phi_k} \right) \]

ASペン

Process Knowledge

Simulation

Chemical Engineering, Newcastle University
Barrier to collaborative working

Molecular Networks

Data mining

Accelrys

DATA

Chemistry Knowledge

Knowledge Barrier

Process Knowledge

ASPEN

Simulation

Chemical Engineering, Newcastle University
Development and application of enabling technologies

Combining advanced technologies with parallel chemistry, mathematics and engineering; the overall goal being to develop more cost effective chemical processes.

Business intensification

Whole lifecycle information requirements
Development and application of IT systems and business methods to Chemical Process Development

Chemical Engineering, Newcastle University
The conceptual model for BatchCAD
Case study Scale up of reactive distillation

- poor laboratory result – 50% yield
- reaction times longer than 48 hours
- complex process with large number of operating parameters to vary

Solution
  - Determine kinetics,
  - model process
  - Use model for safe optimum scale up
Reactive distillation – 10 reactions

successive substitution of allyl alcohol for ethanol on triethylorthoacetate

Conventional approach requires large number of experiments for optimization

oxy-Claisen rearrangements
20 experiments and 15 days modeling work to take process from laboratory to production

- yield increase from 50% to > 98%
- pilot plant trials designed by simulation
- full scale process designed by simulation
<table>
<thead>
<tr>
<th>Time taken to develop models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>kinetic modeling</td>
</tr>
<tr>
<td>pilot plant model</td>
</tr>
<tr>
<td>production plant model</td>
</tr>
<tr>
<td>TOTAL</td>
</tr>
</tbody>
</table>

*Simulation reduced the need for a large experimental program*

*Consulting costs recovered by minimizing the use of technical resources*
Kinetics as descriptors, a continuum from molecular modelling to plant

- Match kinetics
- DOE
- Experimental conditions
- Experiments
- Yields
- Kinetics
- Rate constants
- Process simulation
- Optimum kinetics
- Development
Case study: a virtual development company

- No manufacturing plant
- No pilot plant
- Limited laboratory facilities

Started 1995 …
A major product of IFF

- Aromachemical Sales
  - world market 1400 tonnes @ $50 per kg (1995)
  - $70 million per annum

- Patented four stage process
  - 2 reaction
  - 2 distillation

- Patent expired
  - Competitors develop ‘me too’ products when patent expires
# The reality of virtual process development

## Conventional development

<table>
<thead>
<tr>
<th>R&amp;D</th>
<th>2 years</th>
<th>Project leader</th>
<th>Chemists</th>
<th>Chemical engineer</th>
<th>Contractors</th>
<th>Chemists</th>
<th>Chemical engineer</th>
<th>HSE + environmental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process development</td>
<td>1 year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time to market</td>
<td>3 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Costs</td>
<td>$750k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Throughput</td>
<td>&lt; 100 tonnes pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Virtual development

<table>
<thead>
<tr>
<th>R&amp;D</th>
<th>1 year</th>
<th>Project leader</th>
<th>Chemists</th>
<th>BatchCAD consultant</th>
<th>Contract manufacturer’s resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time to market</td>
<td>1 year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Costs</td>
<td>$60k</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Throughput</td>
<td>300 tonnes pa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Time to market**: 2 years early
- **92% cost saving**
Kinetics are central to reaction engineering

the kinetic model of the reaction system allows the chemical engineer to use modelling and simulation software and reaction-engineering principles to explore regions of parameter space not investigated experimentally (for scale up, thermal safety and process optimisation studies).
♦ Literature mechanisms are often wrong

Oxazoborilidine - Corey
Complex enzyme substrate reaction:
- $I_1$, $I_2$ are the initial species
- $S_3$ – $S_6$ are substrates
- $E_1$ – $E_6$ are enzymes
- $X_1$ – $X_6$ are fixed (buffered)

Experiment involves randomly varying $I_1$ and $I_2$ and measuring the response of $S_3$ – $S_6$.
Automation of mechanism elucidation

Extending High Throughput Technologies to Chemical Process Development: The automation of mechanism determination

- Allen Wright, Mark Willis (Newcastle Chem Eng)
- Andy Whiting (Durham Chemistry)
- Steve Scott (Leeds Chemistry)
- 4 PostDocs
- 6 PhD students

Workpackages
- Mathematical methods
  Started 2005
- Experimental design
- Synthetic chemistry
- Oscillatory reactions
Chemical Engineering, Newcastle University

Reaction Networks and Reaction Mechanisms

For example

\[ x_1 + x_2 \rightarrow x_4 \quad x_3 \rightarrow x_5 \rightarrow x_6 \]

Chemistry
Chemical reaction networks are composed of a number of elementary reactions; eg addition, elimination, substitution and rearrangement.

For example, given an addition reaction between two reactants $x_1$ and $x_2$ to form a product $x_3$:

\[
\frac{d[x_1]}{dt} = -k[x_1][x_2] \quad \frac{d[x_2]}{dt} = -k[x_1][x_2] \quad \frac{d[x_3]}{dt} = +k[x_1][x_2]
\]

Mathematically, finding a useful model of a number of interacting chemical reactions is equivalent to a network inference problem, and so the problem may be redefined as the identification of both the **structure** and **parameters** of a reaction network.
For each species $i$ this change in concentration with respect to time can also be expressed in the form of the following general model

$$\frac{d[x_i]}{dt} = g_i(\varphi, \Theta_i) + \varepsilon$$

$$\varphi = ([x_1], [x_2], [x_3], \ldots)$$

$$\Theta = (\theta_1, \theta_2, \theta_3, \ldots)$$

The choice of the non-linear mapping (i.e. the model structure) largely determines the type of processes that can be modelled.

Many system identification techniques make use of a model structure that can be expressed as the weighted sum of the output of a number of functions (basis functions) of the model inputs

$$g_i(\varphi, \Theta_i) = \sum_{j=1}^{m} \theta_j h_j(\varphi)$$
It is assumed that:

- the reactions are at most bimolecular and the total reaction order is no greater than 2
- the rate equations are of integer order with respect to each species concentration
- there are no terms consisting of purely zero order elements
Specification of the Basis Functions for Model Identification

- These assumptions constrain the right hand side of each ODE to be a linear in the parameters model with quadratic terms

\[ g_i(\varphi, \Theta_i) = \sum_{j=1}^{n} \theta_{0,j,i} [x_j] + \sum_{l=1}^{n} \sum_{j=l}^{n} \theta_{l,j,i} [x_l][x_j] \]

\[ \frac{d[x_i]}{dt} = \theta_{0,1,i} [x_1] + \theta_{0,2,i} [x_2] + \theta_{1,1,i} [x_1]^2 + \theta_{2,2,i} [x_2]^2 + \theta_{1,2,i} [x_1][x_2] + \varepsilon \]

- Reducing the difficult task of simultaneous structure and parameter optimisation to one of just parameter optimisation
Case Study - Reactions with simultaneous gas liquid mass transfer (simulation)

$$\frac{d[x_1]}{dt} = K_h (Sol - [x_i]) - k_1[x_1][x_2] - k_2[x_1][x_3] - k_3[x_1][x_4]$$
$$\frac{d[x_2]}{dt} = -k_1[x_1][x_2]$$
$$\frac{d[x_3]}{dt} = k_1[x_1][x_2] - k_2[x_1][x_3]$$
$$\frac{d[x_4]}{dt} = k_2[x_1][x_3] - k_3[x_1][x_4]$$
$$\frac{d[x_5]}{dt} = k_3[x_1][x_4] - k_4[x_5][x_7]$$
$$\frac{d[x_6]}{dt} = k_3[x_1][x_4]$$
$$\frac{d[x_7]}{dt} = -k_4[x_5][x_7]$$
$$\frac{d[x_8]}{dt} = k_4[x_5][x_7]$$
Approach to automated network mechanism deduction

- Replace all derivatives with the slopes, estimated directly from observed data using splines (or other suitable technique) to generate the differential values for each species

- Fit least square regression models between the postulated models and these differential value

- Least squares regression is preferable as it is not an iterative technique, hence, not computationally costly
  - A further advantage is that informative statistics may be generated
  - Individual t-tests for each term

- Estimate the k values for the models

- Determine likely RHS of each equation from Statistical Significances
  - Terms may be rejected and, hence, be removed from the model

- Rationalise model
Model Complexity Reduction

- Student’s t-statistic may be used to determine whether each model term is statistically significant
- The larger the absolute t-value the higher the likelihood that the term is significant
- The decision process is assisted through use of coefficient p-values
- Terms with relatively high p-values may be rejected and, hence, be removed from the model

\[ t = \frac{\theta_{l, j, i}}{\hat{\sigma}_{l, j, i}} \]
Model Rationalisation

Rationalisation ensures that each ODE is physically consistent with the others in the sense of conservation of mass and adherence to the laws of mass action kinetics. Rationalisation is achieved by:

- exploiting knowledge of the species’ relative molecular masses
- application of the basic rules of reaction chemistry
Chemical Engineering, Newcastle University

Enhanced methods

\[ \hat{x} = [s[x_1], s[x_2], ..., s[x_5]] \]

Number of reactions
number of rate terms

SVD

Best subset regression

Models containing different predictors
and different numbers of predictors

\[ m = 2n + \frac{n(n-1)}{2} \]

Best subset of complexes

Best subset regression

Best combination of complexes

Network
(mechanism)

Chemistry
mechanism

Chemistry
mechanism
Automation of mechanism elucidation

Accelerator platform

Reaction Calorimeter

Design software

Mechanism software

Chemical Mechanism

Physical mechanism

$X_1$ gas

$X_1$ liquid

Chemical Engineering, Newcastle University
Comparison of gas liquid mass transfer coefficients

Stirrer speed rpm

kl (min\(^{-1}\))

Chemical Engineering, Newcastle University
Experimental case study: L-proline catalysed aldol reaction

\[
\begin{align*}
\text{A} & \quad \text{p-nitrobenzaldehyde} \\
\text{B} & \quad \text{Acetone} \\
& \quad \text{Aldol product} \\
\text{C} & \quad \text{Dehydrated product}
\end{align*}
\]

\[(1)\] p-nitrobenzaldehyde + Acetone → Aldol product

\[(2)\] Aldol product → Dehydrated product + H₂O
Overview of approach

Three stage development programme:

- **Stage 1:** Obtaining a preliminary chemical and physical understanding of the system
  - 10 x 13ml reactors Accelerator

- **Stage 2:** Obtaining quantitative understanding of the reactant-catalyst-solvent combinations
  - 16 x 13ml reactors Accelerator

- **Stage 3:** Obtaining time-concentration profile data and mechanistic modelling study
  - 8 x 50ml Accelerator reactors, 1litre reaction calorimeter
Results

Data Acquisition

Eight automated batch experiments were performed in the SLT 106. Five samples were taken from each 50ml reactor, approximately every 25 minutes. Samples were analysed by GC-MS to give time-concentration plots for the aldol product and the benzaldehyde, no dehydrated product was recovered.

The mechanism deduction procedure implied,

\[
\frac{d[Ad]}{dt} \propto (-[Ah])(+[Lp])(+[Ah]^2)
\]
\[
\frac{d[Ah]}{dt} \propto ([Ah])(+[Ah]^2)
\]
\[
\frac{d[Lp]}{dt} \propto ([Ah])(-[Ah]^2)
\]

This indicates that:

- \(Ah\) and \(Ad\) disappear due to \([Ah]\)
- \(Ah\) and \(Ad\) appear due to \([Ah]^2\)
- \(Ad\) appears due to \([Lp]\)
- \(Lp\) appears due to \([Ah]\) and disappears due to \([Ah]^2\)

One possible mechanism this could infer is:

\[
2Ah \rightarrow l_i \quad l_i + Lp \leftrightarrow Ah + l_2 \quad l_2 + Ac \rightarrow Ad + Lp
\]

Which overall gives:

\[
Ah + Ac \rightarrow Ad
\]

in accordance with the literature.
Considerations

- the concentration-time series data is available for all species over the entire duration of each experiment: not any more

- obtaining a reasonable approximation of the first derivative for the concentration of each chemical species: can use \( Q_r \)

- physical mechanisms may be included: fed batch, mass transfer

- improvements by placing either additional constraints on the form of the model structure or through use of constrained optimisation

- use the system to suggest highly targeted additional experiments for the purposes of model discrimination:
  - Accelerator to a Calorimeter

- Software Development: future
Success in the chemicals industry has traditionally relied on:
- Good chemists
- Efficient plant

It’s no longer just about chemistry

Or new engineering design
- Batch to continuous

It’s not just about point solutions
- Software tools, mathematical methods, high throughput

It’s also about information and knowledge and the sharing this knowledge
Acknowledgements

Colleagues and Students at Newcastle

Chemical Engineering, Newcastle University