Chemical Reaction Databases
Computer-Aided Synthesis Design
Reaction Prediction
Synthetic Feasibility

Dr. Wendy A. Warr
http://www.warr.com
Representation

• rxnfile
• RDfile
• SMILES/SMARTS/SMIRKS
• RInChI

Reaction Queries

A → C

A + B → ?

? → C

Wendy Warr & Associates
Reaction Queries

A \rightarrow ? \rightarrow C

WOH
Reaction Queries

- “Name” reaction (e.g., Diels – Alder)
- Reduction of functional group A in presence of group B
- Stereoselectivity
- etc.
Atom-to-atom Mapping

Query:

"Hit":

Wendy Warr & Associates
Atom-to-atom Mapping
Atom-to-atom mapping

• Automatic mapping is not perfect
• Authors publish incomplete equations
• Takes no account of reaction mechanism
Approaches to Mapping

• Maximum common substructure (MCS)

• Optimization approach
  – Fujita’s imaginary transition state (ITS)
  – Gasteiger ITS
  – Varnek condensed graph of reaction (CGR). Pseudomolecules
    • ISIDA descriptors calculated based on graph
    • similarity search

• Baldi MCS and optimization
MCS Approach

Reaction Database Systems

- MDL’s REACCS
  - later ISIS, Isentris
- CASREACT
  - now in SciFinder
- Beilstein CrossFire
  - superseded by Elsevier’s Reaxys
Reaction Databases

- SPRESI and ChemReact
- Theilheimer
- ChemInform
- Science of Synthesis
- Current Chemical Reactions
- Methods in Organic Synthesis
- Catalysts and Catalysed Reactions
- Organic Syntheses
- Selected Organic Reactions Database
- In-house ELNs

Wendy Warr & Associates
Reaction Classification: Uses (1)

• Teaching similarity of reactions
• Indexing reactions
• Browsing in databases
• Management of large hit lists
• Simplification of query generation
• Linking reactions from different sources
Reaction Classification: Uses (2)

- Access to generic type of information
- Deriving knowledge bases
  - for synthesis design
  - for reaction prediction
- Prediction of new reactions
- Automatic procedures for analysis
- Quality control of databases
- Overlap studies of databases
Reaction Classification

Methods

• Model-driven
  – manual
  – computerized
    • Balaban, Hendrickson, Arens, Zefirov, Fujita
    • Dugundji-Ugi

• Data-driven
Dugundji-Ugi Model
Dugundji-Ugi Model

- WODCA
- EROS
- IGOR
- RAIN
Data-driven Classification

- Goes beyond the reaction center
- Allows sub-classes
- Wilcoxon and Levinson, Blurock, Gelernter, Sello
- InfoChem CLASSIFY
CLASSIFY

• Based on $\text{IC}_{\text{MAP}}$
  – extension of Willett and Funatsu’s work
  – maximum common substructure
  – minimum chemical distance

• Atom hash codes calculated for reaction center
  – uses modified Morgan algorithm

• Sum all hash codes of all reactants and one product → unique Reaction Classification Code (15 digit number)
CLASSIFY

**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
Synthetic Analysis Programs

- Synthesis design (planning)
- Reaction prediction
- Mechanism elucidation
- Synthetic feasibility
Synthesis Planning

- Reaxys Synthesis Planner
- SciFinder SciPlanner
- Chematica
  - Network of 7 million chemicals/reactions
Computer-aided Synthesis Design

- LHASA
  - expert system
  - knowledge base
    - reaction transforms (manual)
  - combinatorial explosion
    - so prune trees using heuristics
    - or user interaction
Computer-aided Synthesis Design

- SECS
- ARChem
- IC$_{SYNTH}$
ARChem

- Rules automatically generated
- Uses large database to verify rules
- Core (reaction center) extended to relevant functionality
- Tries to use reaction mechanism
Computer-aided Synthesis Design

• HORACE
  – mechanistic descriptors
    • inductive effect
    • resonance effect
    • charge distribution etc.
  – topology based on Gelernter classification
  – produces reaction hierarchy
  – extended with Kohonen neural networks
    • Gasteiger and Chen, Funatsu
WODCA and EROS

• WODCA
  – retrosynthesis
  – similarity search in catalogs
  – break strategic bonds
    • charge distribution, and inductive, resonance, and polarizability effects

• EROS knowledge-based system
  – metabolic reactions
  – mass spectrometer reactions
  – with IR, in identification of degradation products
Reaction Prediction

• The reverse of retrosynthesis

• Approaches:
  – simulation of transition states
  – rule-based, expert systems
  – inductive learning methods
IGOR

• Generality of formal techniques
  – can generate new reaction mechanisms
• Dugundji-Ugi model
• Herges predicted and verified new reactions with IGOR
  – and did further work…
Reaction Prediction: More

- Gasteiger (compare WODCA)
- Gasteiger and Chen Kohonen neural networks
- Zefirov’s Symbolic Equations (SYMBEQ)
  - another formal-logical approach
  - can also be used to generate Dugundji-Ugi matrices
• Baldi, Chen *et al.* use multiple approaches:
  – descriptors are MOs and topological and physical attributes (not graph rearrangements)
  – rule-based system Reaction Explorer
  – inductive machine learning
Varnek and Co-workers

• For atom mapping:
  – CGR (pseudomolecules)
  – calculate ISIDA descriptors
  – similarity search

• To model chemical reactivity maybe use ISIDA property-labeled fragment descriptors (IPLF)
Synthetic Feasibility

• Large number of compounds generated by:
  – combinatorial library design
  – de novo design

• Some of them will be hard to make

• CAESA

• SYLVIA
CAESA

• Rule-based system too slow for intermediate structures in de novo design
• Complexity analysis is more practical
• Matches structural motifs in designed structures with those in drugs and starting materials
SYLVIA

- Synthetic complexity score 1-10
- Adds scores from components
  - molecular graph, ring and stereochemistry
  - similarity to starting materials
  - frequency analysis of strategic bonds from reaction databases

Wendy Warr & Associates
Conclusions (1)

• Much research “complete” before 1990
  – but papers on atom-to-atom mapping are still appearing

• Computer-aided synthesis design programs preceded reaction retrieval systems
  – but have never achieved same levels of usage
Conclusions (2)

• Emphasis on “aided”
  – chemist plus machine

• Regio- and stereo-selectivity, interfering functional groups are active fields of research

• Synthetic chemists not interested in reaction prediction?

• In-house systems are using synthetic feasibility

Wendy Warr & Associates