Information Extraction from Chemical Images
Discovery Knowledge & Informatics
April 24th, 2006

Fraunhofer Institut
Algorithmen und Wissenschaftliches Rechnen

Dr. Marc Zimmermann
Available Chemical Information

- Textbooks
- Reports
- Patents
- Databases
- Scientific journals and publications
- Websites
Representations of Chemical Compounds

- Name (trivial, trade, brand, INN, USAN)
- Registration numbers (CAS, NCI, Beilstein)
- Formal description (sum formula, SMILES)
- Chemical nomenclature (IUPAC, CAS, InChI)
- Depictions
Example: Aspirin

- **Name**: Acetylsalicylic acid, Aspirin, Bayer, Colfarit, Dolean PH 8, Duramax, Ecotrin, …
- **CAS**: 50-78-2, **SID**: 35870,
- **Formula**: C9H8O4
- **IUPAC Name**: 2-acetoxylbenzoic acid
- **SMILES**: CC(=O)OC1=CC=CC=C1C(=O)O
- **InChI**: 1.12Beta/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h1H3,2-5H,(H,11,12)
- **Depiction**:
Information Extraction Methods

- Names ⇒ Dictionary based
- Registration numbers ⇒ Databases
- Formal descriptions ⇒ Rule based
- Depictions ⇒ chemical OCR
Representing a Chemical Compound

How much information do you want to include?

- Atoms present
- Connections between atoms
  - bond types
- Isotopes
- Charges
- Stereochemical configuration
Modeling of Chemicals as Graphs

Why use graph theory?

- Established mathematical field
- Graphs can be easily represented in computers
- Existing algorithms for comparison, searching, etc.
- Unlike humans, computers aren’t very good at pattern recognition
A typical example: MDL MOL file (SDF)

For more information on MDL formats, see http://www.mdl.com/downloads/public/ctfile/ctfile.jsp
Disadvantages of Using Graphs

- Many graph algorithms are inherently slow
- Analogy between chemical structures and graphs is not perfect
- Realities of chemical structures cause problems
  - aromaticity
  - stereochemistry
  - tautomerism
  - inorganic compounds
  - macromolecules and polymers
  - incompletely-defined substances
Good News

There is only a limited number of chemical drawing tools
(and these are using templates):

- ChemDraw (CambridgeSoft)
- ChemSketch (ACD)
- ISISdraw (MDL)
- JAVA applets (ChemAxon)
- ...

⇒ Reduced complexity
chemOCR: Reconstruction of Chemical Compounds

1. Document

Pravastatin Sodium

2. Depiction

3. Reconstruction

4. SDF file
Preprocessing Steps

- Page segmentation
- Image extraction
- Image conversion (image restauration, adaptive binarization ...)

Image Viewer
Connected Component Analysis

- Building an image tree
- Using adaptive nested TreeMaps
Component Classification

1. Raster image
2. Extract features
3. Classify as...
   - Single bonds
   - Double bonds
   - Thick chirals
   - Dotted chirals
   - Text
4. Manual curation

Discovery Knowledge & Informatics, April 24th, 2006
Marc Zimmermann
Atomtype Reconstruction

1. Train new characters
2. Define new superatoms
3. Expand superatoms

Need of a chemical intelligent OCR
Vectorization

Fixing vectorization errors using relative neighborhood graphs

Disconnections

Antiparallel double bonds

Fixing bond lengths

Dubious links

Need of a chemical intelligent vectorizer
Graph Matching

- Using a line graph representation
- Searching for subgraph isomorphism
- Database with common fragments
- Decomposition network for fragments
- Recognizing new fragments

Graph matching a solution for mapping bridged ring systems
Manual Curation of Errors

Editing bonds

Reconstruction score

Discovery Knowledge & Informatics, April 24th, 2006
Marc Zimmermann
Post Processing

Workflow plugin technology

- 2D beautify
- File format conversion
- 2D to 3D conversion
- Name generation
- Property calculation / prediction
- …

Input image: /home/marc/workspace/CSR_New/exam...
Number of bond length changes: 2
Number of multiple bonds merged: 6
Reconstruction scoring: 1
SMILES string: CNCC(C(\cc1ccc\cc1)C(F\xF)F)c2ccc2
IUPAC name: lnCh1=1/C17H18F3NO/c1-21-12-11-1...
InChI identifier:
A Real Challenge

- Data set with ~7,600 depictions of natural products
  - to get new scaffolds and super atoms
  - to incorporate the CSR workflow into a grid service
  - to add a database interface

But we need more real training sets…

(i.e. pictures and the solved structure)

current status: ~3,400 fully reconstructed!
Incompletely-defined substances:

- unknown stereochemistry
- unknown attachment position
- unknown repetition
Markush ("Generic") Structures and Reaction Schemes

- shorthand for describing sets of structures with common features
- structures with R-groups
- very important in chemical patents
- can be used to describe combinatorial libraries
- can be used as queries in database searches

![Chemical structure diagram](image)
The Mission: Combination of CSR and Text Mining

Image Analysis / Structure Reconstruction

Text Analysis / Entity Recognition

Reconstruction of Published Chem-, Pharma- and PatentSpace

Discovery Knowledge & Informatics, April 24th, 2006

Marc Zimmermann
The Team (in the order of appearance)

- Marc Zimmermann
- Tanja Fey
- Le Thuy Bui Thi
- Christoph Friedrich
- Yuan Wang
- Maria-Elena Algorri
- Miguel Alvarez
- Wei Wang
CSR can extract chemical depictions from various image sources and convert them into SD-files, which can be further used in nearly all chemical software; it allows for the modification of reconstructed molecules by a structure editor; it maintains the superatom and bond (single, double, triple, or chiral) information; and it accepts user curation and scoring schema to improve its performance.