The Electronic Representation of Chemical Structures: beyond the low hanging fruit

How Accelrys Plans to Address the Remaining Challenges in Structure Representation and Searching: Chemically Modified Biologics, Non-specific Structures, and Organometallic Compounds

The New Developments in Chemical Information: Best Practice

Keith T Taylor, PhD, BSc, MRSC
Product Manager, Chemistry Foundation
Accelrys Inc, California
The Language of Chemistry

• The Chemical Structure Diagram

• What is it?
  – psilocybin, [3-[2-(dimethylamino)ethyl]-1H-indol-4-yl] dihydrogen phosphate
  – An: indole; phosphate ester; tertiary amine; acid; base
Chemical structure diagrams work for structures that can be described as atoms linked by a definite number of bonds to other atoms:

- Works well for most drug-like structures that contain main group elements
- Second row elements can present difficulties
- Need to accommodate multiple valencies in periods three and higher
- Some interesting problems in period 2

Searching is best with a standardized representation:

- Structure representation conventions are needed
Two styles;

- Connection Table
  - The structure is defined as a table of atom types and bond types that connect to the atom
  - Each atom and bond is given an arbitrary number in a series
  - Relative coordinates for each atom are usually included
  - Molfile is the most common type
    - SDfile is an extension of the molfile

- Line Notations
  - Arbitrary atom is selected and then the structure is described as a sequence of atoms connected by symbols that represent bond types
  - Includes labels to identify ring closures
  - SMILES is the most common type
  - InChI is a line notation
Examples

- Molfile

ACCLDraw06271311182D

8 8 0 0 0 0 0 0 0 0 0999 V2000
15.8192 -4.2369 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
14.6879 -4.6046 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
14.6879 -5.7940 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
13.6627 -6.3940 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
12.6375 -5.7940 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
12.6375 -4.6060 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
13.6627 -4.0181 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
13.6627 -2.8370 0.0000 O   0 0 0 0 0 0 0 0 0 0 0 0

- SMILES

Cc1ccccc1O

Examples

- SMILES

Cc1ccccc1O

- Molfile

ACCLDraw06271311182D

8 8 0 0 0 0 0 0 0 0 0999 V2000
15.8192 -4.2369 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
14.6879 -4.6046 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
14.6879 -5.7940 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
13.6627 -6.3940 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
12.6375 -5.7940 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
12.6375 -4.6060 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
13.6627 -4.0181 0.0000 C   0 0 0 0 0 0 0 0 0 0 0 0
13.6627 -2.8370 0.0000 O   0 0 0 0 0 0 0 0 0 0 0 0

Verbose
Preserves layout

Concise
No layout information
• The structure is treated as a graph
  – Atoms are Nodes
  – Bonds are Edges

• Graph theory is used to match Nodes/Atoms and Edges/Graphs
  – The mechanism for substructure searching

• The structure is canonicalized
  – A unique layout that can be reproduced for any input version
  – A name is generated from the canonical structure
    • NEMA key (Accelrys)
    • InChI Name and Key
    • Canonical SMILES
  – Ideally all approaches would produce the same canonical form but in practice different approaches produce different results
  – Names are therefore algorithm dependent
  – Names are used for exact structure matching
<table>
<thead>
<tr>
<th>Structure</th>
<th>SMILES</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Structure 1" /></td>
<td>\texttt{Cc1cccccc1O}</td>
</tr>
<tr>
<td><img src="image2" alt="Structure 2" /></td>
<td>\texttt{c1c(O)c(C)ccc1}</td>
</tr>
<tr>
<td><img src="image3" alt="Structure 3" /></td>
<td>\texttt{c1ccccc(O)c1C}</td>
</tr>
</tbody>
</table>
Different Representations

- Group 2 elements can give problems
  - Carbon Monoxide – ‘divalent’ carbon
  - Nitric Oxide – ‘divalent’ nitrogen
  - Nitrogroup – ‘pentavalent’ nitrogen

- Tautomers
  - Acetone or prop-1-en-2-ol

- Aromaticity
Define a standard representation and enforce it in your databases

- Accelrys’ Available Chemicals Directory is ubiquitous and most companies have adopted its representation rules

  - Search in-house and external databases with the same query
The Three Rs

- **Representation**
  - A meaningful diagram

- **Registration**
  - Storing a standardized and validated object in a database

- **Retrieval**
  - Use an understandable query to retrieve all the objects in the database that match
Cross-searching of Non Specific Structures

- Imipramine Metabolites
Challenges - Addressed

• Generic Structure
  – Combinatorial Library
  – Patents

• Polymers

• Mixtures with known composition
  – Acetaminophen (paracetamol), aspirin, caffeine, and excipients

• Non-specific structures
  – Natural products
  – Industrial preparations
  – Metabolites

• Biologics
  – Antibody Drug Conjugates

• Organometallics
Generic Structures

- Benzodiazepine library
- Contains 192 structures
Polymer

- Aluminized PET

- Jeffermine ED-2003
• Acetaminophen (paracetamol), aspirin, caffeine, and excipients
Non-specific Structures

- Natural product
- Commercial mixtures
- Metabolites
• Coordination (dative) bonding

• Haptic bonding
Biologics

• Small biologics
  – Up to ~30 residues
  – Use full connectivity
  – Visually confusing

• Depict the individual residues as abbreviations
  – Visually cleaner
  – But underlying structure remains cumbersome

• Large number of stereogenic centers slows down registration and searching
A hybrid approach

- Use pseudoatoms for standard residues
- Embed explicit chemistry for non-standard residues and modifications
- Embed the full structure of standard residues to enable full structure features to be calculated
  - Formula and formula weight

- Much more compact
  - Registration and searching much faster
  - No loss of structural information
  - Structure is portable

- Visually resembles the abbreviated form
Improved depiction of Chemically Modified Biologics

Do not underestimate the scale of depiction work 😊

It is essential for successful adoption
PEGylated peptides
Antibody Drug Conjugates

- Example: Herceptin/Trastuzumab
- Large biologic structure
- Drug attached via a linker to a variable location
- Variable number of attached drug/linker entities
- Structure can be registered and searched using a combination of all the features described

But
- Drawing needs simplification
- Depiction needs improvements
- Representation issue remains
  - How to display a disulfide bridge that may be broken and replaced by the drug payload
  - Account for the disulfide bridges that remain in the formula and formula weight

- In research
What works today

- Herceptin_DM1 Mut Cys
What works today

- Site-specific conjugation to an antibody with an unnatural amino acid glycosylated
Markush Chemically Modified Biologic

- Variable residue
- Variable attachment
• Major focus of Accelrys’ chemical representation development

• Represent parts of the structure as text identifier
  – ALK – represents any alkyl group
  – ...

• Use for registration and searching

• Use for patent searching

• Patents contain all the features described so far (and more)
  – Generic features
    • Defined RGroups
    • Atom lists
    • Generic atoms
  – Homology groups
Currently supported

• Generic features
  – Defined RGroups
  – Atom lists
  – Generic atoms

• Reaxys homology groups
  – Any group       G
  – Acyclic         ACY
  – Carbacyclic     ABC
  – Alkyl           ALK
  – Alkenyl         AEL
  – Alkynyl         AYL
  – Heteroacyclic   AHC
  – Alkoxy          AOX
  – Cyclic          CYC
  – Heterocyclic    CHC
  – Heteroaryl      HAR
  – Carbocyclic     CBC
  – Aryl            ARY
  – Cycloalkyl      CAL
  – Cycloalkenyl    CEL
  – Cyclic (no C)   CXX
Mapping: Homology group screening

- Screen MDDR data set
  - 129,237 structures screened in ~30s
  - No pre-processing

Key:
- Q = Any atom except C & H
- AHC = acyclic chain with a heteroatom
- AOX = alkoxy chain

Hits = 470

Hits = 108

Hits = 45

Hits = 16

Hits = 10
Requirements and stereochemistry covered
- Includes ABSOLUTE, AND and OR centers, and structures with a mixture of types
- Allenes and cumulenes
- Biaryls and any pair of rings with hindered rotation

Work in progress
- Stereochemistry of organometallics
- Helicenes
Summary

• Discrete small molecules are covered

• Biologics well covered but work needed on User Interface (UI) design

• Stereochemistry well covered
  – Organometallics and helicenes need work

• Significant enhancements to homology group handling required
  – Underway