The IUPAC Chemical Identifier

__________________________

Steve Stein, Steve Heller, Dmitrii Tchekhovskoi

National Institute of Standards and Technology
Gaithersburg, MD, USA

__________________________

CAS/IUPAC Conference on Chemical Identifiers and XML for Chemistry
Columbus, OH
July 1, 2002
IUPAC & Chemical Identity

• Mission
  – International, open standards for chemical communication

• Printed Media – Nomenclature
  – Human communication
  – Rules for structure to name conversion

• Digital Media – Identifier
  – Computer communication
  – Rules for structure to identifier conversion
    • Freed from restrictions of ‘pronuncibility’
    • Freed from ring index
Chemical Identifiers

• Structures
• Connection Tables
• ‘Trivial’ Names
• Systematic Names
• Index Numbers
Too Many Identifiers

• Structure diagrams
  – various conventions
  – contain ‘too much’ information

• Connection Tables
  – MolFiles, Smiles, ROSDAL, ..

• Pronounceable names
  – IUPAC, CAS, trivial

• Index Numbers
  – EINECS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAF
What kind of Identifier is needed?

• Exactly one Identifier per structure
• Defined by algorithms
• Comprehensive
• Openly available
• Implemented
Requirements

• Different compounds have different identifiers
  – All distinguishing structural information is included

![Diagram showing structural differences between two compounds and their corresponding identifiers IChl - 1 and IChl - 2]
Requirements

• One compound has only one identifier
  – No unnecessary information is included

\[
\text{Same IChI} = \quad \text{Same IChI} = \quad \text{Same IChI} = \quad \text{Same IChI}
\]
IChI Scope
First Version

• Discrete, covalently bonded compounds
  – foundation for other classes

• Isotopes

• Stereochemistry
  – sp3 - tetrahedral
  – Z/E - double bond

• Tautomers
3 Steps to IChl

• ‘Normalize’ Input Structure
  – Implement chemical rules

• ‘Canonicalize’ (label the atoms)
  – Equivalent atoms get the same label

• ‘Serialize’ the Labeled Structure
  – A unique series of bytes
NORMALIZATION
Simplifications

• Ignore ‘Electron Density’
  – Double/triple bonds, Odd-electrons, Charges
  – Still use for Z/E stereo perception

• Free Rotation Around Single Bonds

• Divide IChI into Layers
Ignore Electron Density

• Not required for compound identification
  – Distinguishes ‘excited states’

• Avoids problems
  – Delocalization, aromaticity, zwitterions, …
conventional

redundant

IChI
Münchnones
Assume Free Rotation Around Single Bonds

Ignore Conformation
LAYERS
Divide into ‘Layers’

- Separate ‘Name’ into Fragments by
  - Connectivity
  - Isotopes
  - Stereochemistry
  - Tautomerism
Basic Layer
Simple Connectivity

• Just atoms and their neighbors
  – Ignore everything else

• Robust basic identifier
Isotopes

Treat isotopes as distinct atom types

\[
\begin{array}{c}
\text{H} \\ \\
\text{C} \\
\end{array} 
\quad \neq 
\begin{array}{c}
\text{H} \\
\text{C} \\
\end{array}
\]
Stereochemistry

• Double Bond (Z/E)
  – from coordinates or bonding

• Tetrahedral (sp$^3$)
  – ‘in/out’ bonds or x,y,z coordinates
Varieties of Double Bond Isomers
$\text{sp}^3$ (tetrahedral) stereoisomers

Stereodescriptor needed
Identify Stereogenic Centers

- Speed up processing
- Helpful for chemists
Basic Tautomer Layer

H-migration between 1,3 heteroatoms
Tautomers

L,R = N, O, S, Se, Te
Q = C, N, S, P, ...

also
Electronic Layer

Simply Store Net Charge

\[\text{Possibilities:}
\]
- Neutral
- -1 (anion)
- +1 (radical cation)
- +2 (doubly charged)

Electronic State?
OUTPUT
ICHl Output
9 possible fields

• Basic
  – Isotopic
    • Stereo
    – Stereo
  – Stereo

• Tautomeric
  – Isotopic
    • Stereo
    – Stereo

• Electronic
Possible Output Format

Example: Benzene

Represent atoms as sequence number in formula

C6H6 = C C C C C H H H H H H
tags  1  2  3  4  5  6  7  8  9 10 11 12

Basic Layer:
<basic>C6H6 1-2-7 2-3-8 3-4-9 4-5-10 5-6-11 7-12</basic>
Other Output

• Information Only
  – For user verification
    • Label true stereogenic atoms
    • Identify equivalent atoms

• Warnings
  • Unusual valences
  • Unrecognized input

• ‘Reversibility’ Information
  – Coordinates
  – Electron density
    • Positions of double/triple bonds, charges, odd electrons
PROBLEMS
Two Fundamental Problems

• Chemists
  – Different ways to represent the same thing
  – Different definitions of tautomerism
  – Different guesses

• Chemicals
  – Structures can depend on conditions
  – Tautomers can depend on conditions
When to allow double bond stereoisomerism?

Proposed: If a bond can be single, no Z/E stereo allowed

Considered: Allow users to override default behavior
Drawing Standard Needed?
Bond/No bond
Allow Full ‘Reversibility’?

- Coordinates
  - Structure display

- Original bonds and charges
  - For display and future use

- Original numbering
  - Map to input data
IChI – What can’t it do?

• Discover that two structures with different connectivity represent the same compound
  – Unless they are tautomers

• Predict potential for Z/E isomerism in open shell conjugated networks
  – Cannot predict rotational barriers

• Fix improperly entered data
  – Guarantees wrong IChI for bad data

• Properly treat non-covalent bonding
  – Coordinate bonds

• Represent ‘exotic’ stereochemistry
Version I

• Implement All Normalization Rules – 12/02
• Test against available data sets – 3/03
• Final External Testing and Refinement – 7/03
• Documentation, source, executable – 12/03?

• Open discussions
  – ichi-l@list.rsc.org
Future Extensions

• Organometallics
  – Coordinate bonds

• Other Stereo Forms
  – Non-atom centered
  – Conformations
  – Hydrogen Bonding

• Polymers/Macromolecules

• Compound Classes
  – Markush structures