MInChI

a chemical notation for mixtures

Leah McEwen

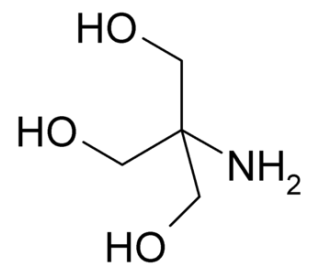
2017-08-16

InChI Workshop @ NIH

Mixtures notation project goals

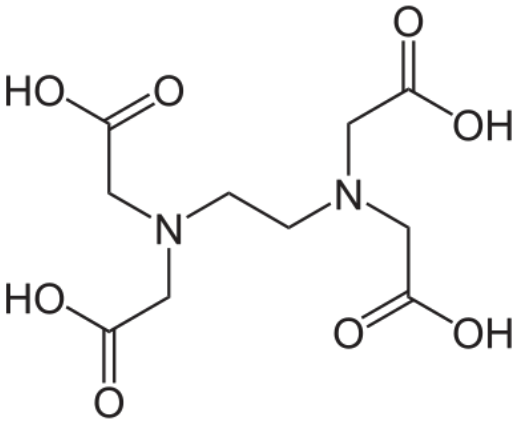
- Develop an unambiguous linear notation for mixed substances that can be hashable and resolvable to unique components
- Articulate what can be said, definitively and in an actionable way, what is known about the chemical composition of a given mixed substance
- Support the practical need to connect data and information on mixtures and individual components
- Support further computation and analysis on properties, composition, etc.

Composition information components

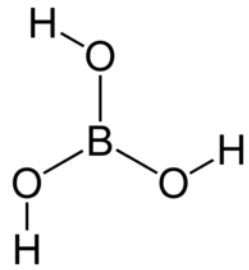


Based on the declared composition at the point of reference:

- Constituent compounds (use standard InChIs)
- Stated concentrations of the constituents
- Other possible relationships, e.g. order, role
- (Other recipe conditions)



➢Use the RInChI code base, an algorithm for concatenating component InChIs and grouping relationships



Multi-component system notation

1.7M t-Butyllithium in Pentane:

MInChI=0.00.1S/
C4H9.Li/c1-4(2)3;/h1-3H3;/q-1;+1
&
C5H12/c1-3-5-4-2/h3-5H2,1-2H3&
/n{1&2}
/g{17mr-1&}

37% wt. Formaldehyde in Water with 10-15% Methanol:

MInChI=0.00.1S/
CH2O/c1-2/h1H2&
CH4O/c1-2/h2H,1H3&
H2O/h1H2
/n{1&2&3}
/g{37wf-2&10-15vf-2&}

- alphabetical order of components
- "**&**" separates components
- "**{}**" mixture groups (e.g., nested)
- "**/n**" indexes components (e.g., order)
- "**/g**" concentration (symbols detailed separately)

Indexing hierarchy

25:24:1 (v/v) Phenol:Chloroform:Isoamyl Alcohol
with 10mM Tris, pH 8.0, and 1 mM EDTA:

MInChI=0.00.1S/
[component InChIs]
/n{{1&3&4}&{2&6}&{5&6}}
/g{{24vp&1vp&25vp}&{1mr-3&}pH8.0&{1mr-2&}}

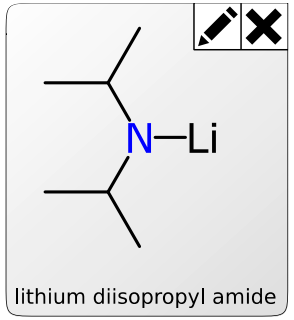
Depicting concentration

Notation	Concentration type	Units
mr	Molarity	mol/L (mol/m^3)
mb	Molality	mol/kg
wv	Weight per Volume	mg/L
wf	Mass Fraction	wt./total wt.
vf	Volume Fraction	vol./total vol.
vp	Volume Proportion	v:v:v
mf	Mole Fraction	mol/total mol
pH	pH	pH

Rules for values expression:

1. scientific notation
   (no percentage signs or ppm)
2. exponent default 0 = 10^0 = 1
3. default NTP
4. min/max values

Desired specificity can be determined downstream to meet local needs.



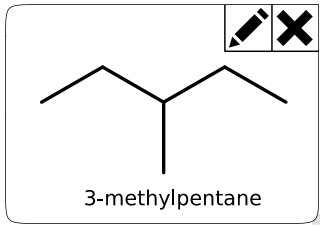
1.0M lithium diisopropyl amide in THF/hexanes

1/8th

solvent

hexanes

7/8th



Generating data…

[Formatting input next slide…]

1.0 M lithium diisopropyl amide in THF/hexanes :

MInChI=0.00.1S/
C4H8O/c1-2-4-5-3-1/h1-4H2&
C6H12/c1-6-4-2-3-5-6/h6H,2-5H2,1H3&
C6H14/c1-3-5-6-4-2/h3-6H2,1-2H3&
C6H14/c1-4-5-6(2)3/h6H,4-5H2,1-3H3&
C6H14/c1-4-6(3)5-2/h6H,4-5H2,1-3H3&
C6H14N.Li/c1-5(2)7-6(3)4;/h5-6H,1-4H3;/q-1;+1
/n{6&{1&{2&3&4&5&nc}}
/g{1mr&{1vp&{1-2vf-1&5-7vf-1&1-5vf-2&1-5vf-2&}7vp}}

```
"mixture": "1.0 M lithium diisopropyl amide in THF/hexanes",
"source": "http://www.sigmaaldrich.com/catalog/product/aldrich/774766",
"contents":
[
        {
                "name": "lithium diisopropylamide",
                "synonyms": ["LDA", "(iPr)2N.Li"],
                "formula": "C6H14LiN",
                "pubchem": "2724682",
                "chemspider": "2006804",
                "inchi": "InChI=1S/C6H14N.Li/c1-5(2)7-6(3)4;/h5-6H,1-4H3;/q-1;+1",
                "smiles": "[Li+].CC(C)[N-]C(C)C",
                "molfile": ". . .",
                "concentration_type": "molarity",
                "quantity": 1.0,
                "units": "mol/L"
        },
        {
                "group": "solvent",
                "contents":
                [
                        {
                                "name": "tetrahydrofuran",
                                "synonyms": ["THF"],
                                "formula": "C4H8O",
                                "pubchem": "8028",
                                "chemspider": "7737",
                                "inchi": "InChI=1S/C4H8O/c1-2-4-5-3-1/h1-4H2",
                                "smiles": "C1CCOC1",
                                "molfile": "...",
                                "concentration_type": "volume proportion",
                                "quantity": [1(1:7)],
                                "units": "v:v"
                        },
                        {
                                "group": "hexanes",
                                "concentration_type": "volume proportion",
                                "quantity": [7(1:7)],
                                "units": "v:v",
                                "contents":
                                [
                                        {
                                                "name": "n-hexane",
                                                "formula": "C6H14",
                                                "pubchem": "8058",
                                                "chemspider": "7767",
                                                "inchi": "InChI=1S/C6H14/c1-3-5-6-4-2/h3-6H2,1-2H3",
                                                "smiles": "CCCCCC",
                                                "molfile": "...",
                                                "concentration_type": "volume fraction"
                                                "quantity": [5X10^-1,7X10-1]
                                                "units": "v/v"
                                        },
```

Minimum Information Requirements:

- Constituent InChIs
- Min/Max quantity values
- Units
- (Concentration type)

Formatting input

JSON, SDF, others?

Embracing ambiguity

- Concentration ranges, as discussed
- Enantiomers, mixture of absolute stereo with each /m flag
- Tautomers, use ‘n’ layer to indicate >1 ‘copies’
- Ambiguous connectivity, e.g. positional isomers
  - Can accommodate characterized forms (e.g., Xylenes)
  - Need an approach to manage uncharacterized isomers…
- Ambiguous chemistry, i.e. not molecularly characterized
  - Often some type of natural product – “Jojoba Oil”
  - Consider some type of generic notation, e.g. ‘*nc*’
- Possible unspecified components, e.g. impurities
  - Do not notate

InChI=0.0.1S/C2H6ClNO/c1-2(3,4)5/h5H,4H2,1H3/t2-/m0/s1
&
C2H6ClNO/c1-2(3,4)5/h5H,4H2,1H3/t2/m1/s1
/n{1&2}

InChI=0.00.1S/C4H3ClN2O/c5-3-1-6-2-4(8)7-3/h1-2H,(H,7,8)
/n{1&1}

Hashing & searching

- MInChIKeys
  - Long-MInChIKey
    - Consists of the InChIKeys of each of the components
  - Short-MInChIKey
    - Fixed length representation of MInChI
    - First portion is hash of all components
- Parsing components enables such uses as:
  - Barcode to MInChI can resolve to components then associated w/ hazards
  - Confirm same/qualified similar entities notebook to notebook: a/components, b/ratios

MInChI Break-Out, Thurs 10:30am, Track 1

<u>Project Team</u>

- **Gerd Blanke**, StructurePendium Technologies GmbH, DEU
- **Alex Clark**, Collaborative Drug Discovery, CAN
- **John Duffus**, Edinburgh Centre for Toxicology, GBR
- **Richard Hartshorn**, University of Canterbury, NZL
- **Chris Jakober**, University of California, USA
- **Jon LaRue**, MilliporeSigma, USA
- **Leah McEwen**, Cornell University, USA, *Chair*
- **Andrey Yerin**, ACD/Labs, RUS