MIXTURES INCHI OVERVIEW

Alex M. Clark

CDD VAULT®

Complexity Simplified

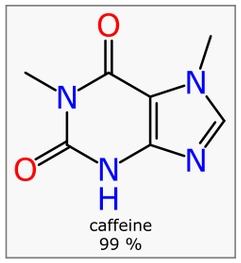
Since Last Meeting

- Paper published in *Journal of Cheminformatics*
  - **11**, article **33** (May 2019)
- iCorps program completed
  - customer research
- Phase 2 grant (NIH/SBIR) reviewed...
  - ... fingers crossed

Previous Work Summary

- Mixtures didn't have any effective standard data format, and urgently needed one
- Defined a simple *Mixfile* format, analogous to *Molfile* (which it encapsulates)
- Open source tool for editing/manipulating, and text-extracted data
- Designed to *feed* the Mixtures InChI (*MInChI*)
- Proof of concept MInChI assembly algorithm

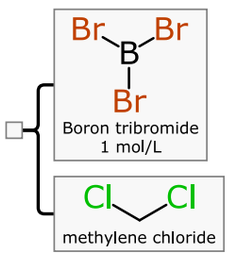
(a)



MInChI=0.00.1S/C7H8N4O2/c1-10-3-8-5-4(10)6(12)11(2)7(13)9-5/h3H,1-2H3,(H,9,13)/n1/g99pp0

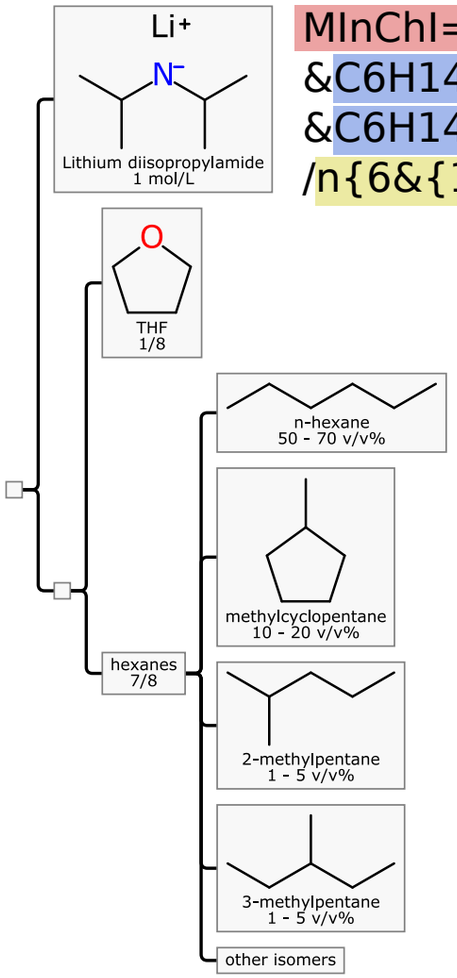
header — structure identifier — indexing — concentration

(b)



MInChI=0.00.1S/BBr3/c2-1(3)4&CH2Cl2/c2-1-3/h1H2/n{1&2}/g{1mr0&}

(c)



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&{3&2&4&5}}}/g{1mr0&{1vp0&{5:7vf-1&1:2vf-1&1:5vf-2&1:5vf-2}7vp0}}

The MInChI

iCorps (NIH)

- 3 colleagues at Collaborative Drug Discovery signed on to do 100 interviews with potential customers
- Learned a lot. Some highlights…
- **Bliss**: most small drug discovery groups don't think it's an informatics problem, unless it's pointed out
  - generally enthusiastic conditional on tooling
  - 1% refused to concede pen & paper

iCorps, ctd

- **Vendors**: best way is to capture mixtures on the way in, purchase → ELN; for ad hoc notebooks, MInChI as a footnote?

- **Value**: besides features like searching & inventory, machine readable mixtures will be noticed during acquisition or filing (e.g. FDA)

- **Properties**: not everyone cares what's in a mixture, but they often care about it's properties, like…

  - **pH**, **density**, **boiling**, **melting**, **vapour pressure**, **solubility**, **stability**, **handling/safety**, …

iCorps, ctd

- **Labelling**: lack of standard format often means that samples are labelled badly or not at all; everyone has a story…
- **Tracking**: in all industries, often nobody really knows what a substance actually is - used as an excuse for lax quality control
- **Biomedical**: formulations, screening samples, natural products (real world samples & synthetic mixing) all tend to be underdefined

iCorps, ctd

- **Software**: informaticians (like us) want a common interchange format
- **Identifier**: obtaining a CASRN for every mixture permutation is inconvenient; could use MInChI instead of a numeric database sequence
- **Markup**: suggestions that someone should grab content from important vendors and mark it all up, without asking permission

Next Phase for CDD

- Assuming that omens check out...

1. Extend *Mixfile* to handle more molecule types (inorganics, ceramics, polymers, variations...)
    - in sync with downstream *MInChI*
2. Major enhancements for text-to-mixture: both public and private collections... much content
3. Integration into ELN software

Acknowledgments

- Leah McEwen & Gerd Blanke
  - and IUPAC/InChI Trust
- Hande Küçük McGinty
  - and Collaborative Drug Discovery

Funding
NIH SBIR