InChI for large molecules
Workshop

Supported by:
IUPAC Division VIII InChI subcommittee
NCBI/NLM
InChI Trust

Lister Hill Center Auditorium
National Library of Medicine
Bldg. 38/Lister Hill Center
1st floor Lobby-Auditorium

October 27-28, 2014

Keith T Taylor PhD BSc MRSC
Ladera Consultancy LLC
Sparks, NV
Use cases

• Must be quick. E.g., handle molecules containing up to 15K heavy atoms (1500 residues) in less than one second
• Able to determine the novelty of the chemical entity
• Compare in chemical or sequence-based structures
• Can do a search by search engine (e.g., Google)
• Different input formats yield same result (PDB, HELM, SCSR, SMILES, FASTA, MOL/SDF, etc.)
• Can be converted back into output format (PDB, HELM, SCSR, SMILES, FASTA, MOL/SDF, etc.)
• Can handle undefined attachment points of chemical entities (e.g., 1-4 vs. 1-6 in carbohydrates) and variable/undefined stereochemistry (e.g., alpha/beta) and ring open/close variants
Use cases

• Can handle a range of attachments at a defined set of possible locations (e.g., 3 entities with 5 potential places to go)
• Can handle payloads, mutated and modified residues beyond that handled by FASTA
• Be able to group identifiers by sequence
• Handle stereo center variation (L vs. D) for a large number (up to max supported residues)
• Consider arbitrary limit on molecule size (although may have performance implications)
• Be able to retain original sequence information even if chemically modified to be something else (e.g., covalent bonding modification such as cyclization of peptide side chains, etc.)
Use Cases

• Be able to represent complex connectivity with metals, e.g., {cysteine} S-Fe clusters

• Be able to handle peptide/saccharide complexes within a larger complex system, e.g., biological interesting molecules dictionary (BIRD – 1000 cases) .. E.g, be able to handle saccharide cases.

• Handle representation of non-standard polymers found in PTMs, peptides, saccharides, chromophores cases

• Consider generic polymer handling (e.g., undefined overall chemical structure but known components or connection points .. no arbitrary restrictions)

• *
Use cases

- Ensemble molecule with distributions of moieties (e.g., variably described molecule mixture that contains a range of molecular entities that are attached {2-4 of X attached, where X might be a peptide chain})
- Capturing oxidation state of metals complexed with proteins or in nanoparticles
- Must handle well defined large molecules
- Can handle RNA/DNA (nucleic acids) and other biopolymer types that are well defined
- Ability to handle well-defined quat-structure (non-covalently bound, e.g., hemoglobin but not insulin)
- Attempt to preserve stoichiometry of the moieties in question
Use cases

- Ability to ignore hydration from chemistry/sequence description
- Ignore polymorphs (except if stoichiometry is different, do not ignore)
- Consider PEG-ylation aspects (e.g., of proteins and peptides)
- Ability to cover most biopharmaceuticals that are marketed drugs (as-is possible)
- Must be able to handle drugs like defibrotide, heparin
- Handle lipid nanoparticles (e.g., lipidsomes)
- Can handle isotopes (consider cases of variable isotopic enrichment)
High level use cases

• Chemically Modified Biologics exhibit many challenges in chemical representation
  • Size
  • Variable substitution sites
  • Variable substitution loading
  • Hydrogen bonding
  • Presence of heavy metals
Biopolymer testing with InChI v1.05

Keith T Taylor PhD BSc MRSC
Ladera Consultancy LLC
Sparks, NV
Background

- Initial releases of InChI were limited to 1024 heavy atoms
- Many biopolymers of interest contain more than 1024 heavy atoms
- v1.05 removes this limitation and enables InChI keys to be calculated for large structures
- This presentation summarizes initial work with large structures using a pre-release version of the software
- The Winchi-1.exe was used to calculate the InChI keys
- Filgrastim sequence was used as the basis for most of the experiments
Limitations

• Structures must be in molfile format
  • V2000 and v3000 formats are accepted
    • v3000 is required for large structures

• The Self Contained Sequence Representation (SCSR) is not supported yet

• Sgroups are not supported and must be removed before presentation to the InChI code
  • Many biopolymer structures contain Sgroup features by default
  • Removal can be achieved programmatically or by editing the molfile in a text editor
Large structure

- Filgrastim

```
  1  M T P L G P A S S L  P Q S F L L  K C L E  Q V R K I Q G D G A  A L Q E K L C A T Y
  20 41  K L C H P E E L V L  L G H S L G I P W A  P L S S C P S Q A L  Q L A G C L S Q L H
  121 161  Q M E E L G M A P A  L Q P T Q G A M P A  F A S A F Q R R A G  G V L V A S H L Q S
  161 201  F L E V S Y R V L R  H L A Q P
```

- InChIKey=KOKXRWZWQJXBOP-NJDFSSKJBA-N
With disulfide bridges

- InChIKey=MMCZGSMNPYTOPN-NJDFSSKJBA-N

- InChIKey=MEMBSBQMAVSGHQ-NJDFSSKJBA-N
Cyclized

• InChIKey=IZNXXFOUFDSLAX-VBNFVGOYBA-N

• InChIKey=IZNXXFOUFDSLAX-VBNFVGOYBA-N
Multiple cyclizations

- InChIKey=AQUGLJGKXYTOSD-VBNFVGOYBA-N
Reversed sequence

- InChIKey=YFXNVYXMKDIHRN-VBNFVGOYBA-N
Filgrastim Lys10-D form

- InChIKey=KOKXRWZWQJXBOP-FNWNWACTBA-N
Synthetic Erythropoietin

• InChIKey=XJBDLLBKVUYAKW-WAXLMBMOBA-D

• PEGylated at K23 and K125
• Acylated at C88 and C106
Polynucleotide - 1

• InChIKey=NELTZQNSFHRPGO-AZBJDUHQBA-N

• Calculation time: ~6s
• Molecular Formula: $C_{2894}H_{3649}N_{1147}O_{1791}P_{300}$
Polynucleotide - 2

- InChIKey=IHDBOWIRPNDUCX-YUQJSOSJBA-N
- Calculation time: ~38s
- Molecular Formula: $C_{5782}H_{7305}N_{2255}O_{3602}P_{600}$
**Polynucleotide - 3**

- **InChIKey=**

- **Calculation timeout at ~125s**
- **Molecular Formula:** $C_{8693}H_{10989}N_{3325}O_{5420}P_{900}$
Myosin-1

- InChIKey=BBJMARUZQDWUQG-PZLOAVSTBA-N
- Calculation time: ~94s
- Molecular Formula: C$_{9725}$H$_{15816}$N$_{2748}$O$_{3100}$S$_{72}$
Trastuzumab dimer

- InChIKey=VRBUFPXQWJVPLO-JNJMYDJTBA-N
- Single arbitrary stereocenter inverted
  - InChIKey=VRBUFPXQWJVPLO-RCEINSQCBA-N
- Calculation time: ~27s
- Molecular Formula: C_{6460}H_{9972}N_{1724}O_{2014}S_{44}
Summary

• InChI v1.05 can generate InChI strings and keys from large structures
  • InChI strings are unwieldy

• All calculations were done using the winchi-1 application
  • Convenient to use but not the most efficient method for calculating InChI strings and keys

• Calculation time for Filgrastim related peptides and the synthetic erythropoietin were not perceptible using the winchi-1 program

• Processing time for large structures needs to be improved

• A large polynucleotide timed out but a polypeptide of similar size did not

• Myosin-1, presented as a linear peptide, took ~94s to process whereas Trastuzumab took ~27s

• Canonicalization may be an area of weakness

• Trastuzumab stereoisomers are differentiated
  • An arbitrary stereocenter in Trastuzumab was inverted
  • Processing time unchanged
  • Different InChI key was produced
Next Steps

Keith T Taylor PhD BSc MRSC
Ladera Consultancy LLC
Sparks, NV
The mertansine is conjugated to the trastuzumab through a maleimidocaproyl (MC) linker which bonds at the maleimide to the 4-thiovaleric acid terminus of the mertansine side chain and forms an amide bond between the carboxyl group of the linker and a lysine basic amine of the trastuzumab. Trastuzumab has 88 lysines (and 32 cysteines). As a result, trastuzumab emtansine is highly heterogeneous, containing dozens of different molecules containing from 0 to 8 mertansine units per trastuzumab, with an average mertansine/trastuzumab ratio of 3.4.
Suggestions

- Remove intolerance of Sgroup data in molfiles
- Support HELM-2 and SCSR as input formats
- Investigate performance issues
  - Canonicalization
  - Timeouts
- Enhance InChI data model to support
  - Variable substitution
  - Variable loading
  - Hydrogen bonds
  - Organometallic bonding
- Remove arbitrary limits
  - In particular maximum atom limit
Question

• Does InChI need to be a rigorous (valence-bond) representation of the structure?
• Is reproducible sufficient
Proposal

• Extend format with extra layers
  • Base InChI correlates to unsubstituted substance
  • Variable substructure
  • Loading variation – 1 to n
  • Position of loading
  • Use new flag to identify that InChI contains variable substituents and variable loading

• InChI key may need third section to contain variability information