Progress towards “Large Molecule” support within InChI

Evan Bolton, Ph.D. – Program Head of Chemistry

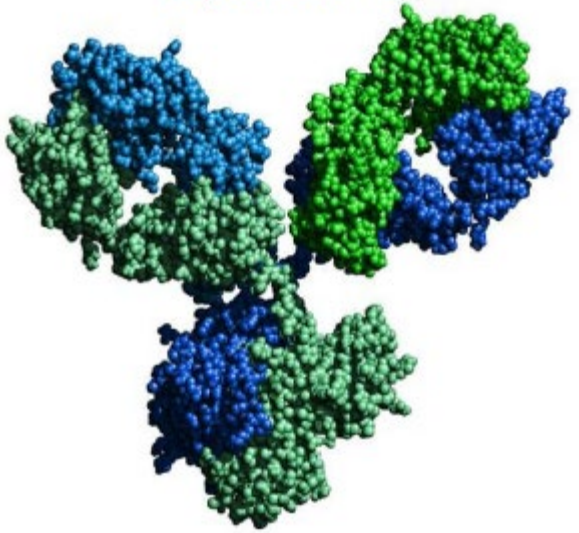
bolton@ncbi.nlm.nih.gov

National Library of Medicine
National Center for Biotechnology Information

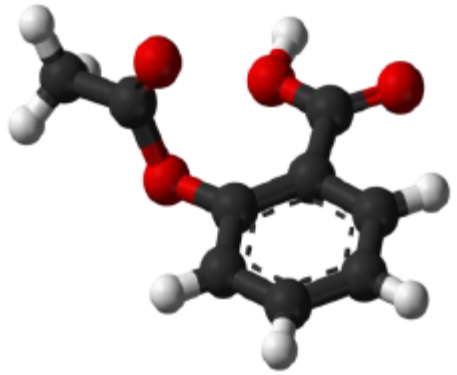
What is a “large molecule”?
(a.k.a., ‘macromolecule’)

- An increasingly important set of chemicals with more than 1,000 atoms and bonds
- Often these are biopolymers
  - amino acids, nucleic acids, carbohydrates (and lipids?)
- Often contain more than known bio-monomers
  - linkers, functionalized, post-translational modifications (PTMs)
- Often containing some degree of uncertainty in its composition
  - unknown sections, variable repeating units, unknown connection points, multiple possible connection points, probabilities of different monomers, mixtures

Large Biologic
Herceptin
~25,000 atoms



Small Molecule Drug / Pharmaceutical
Aspirin
21 atoms

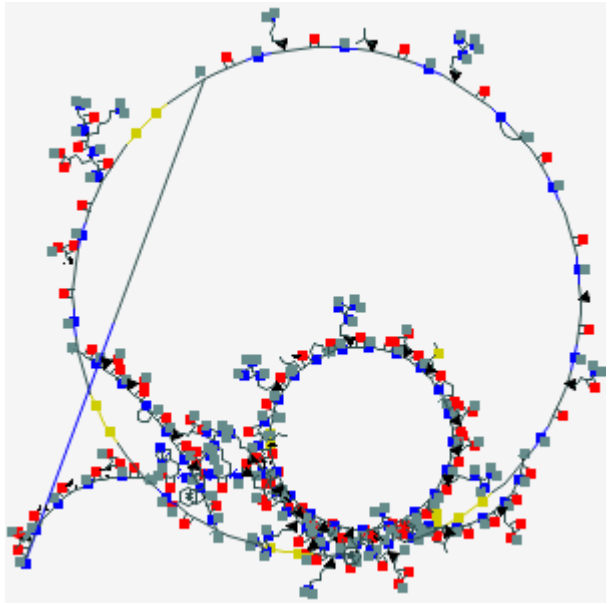


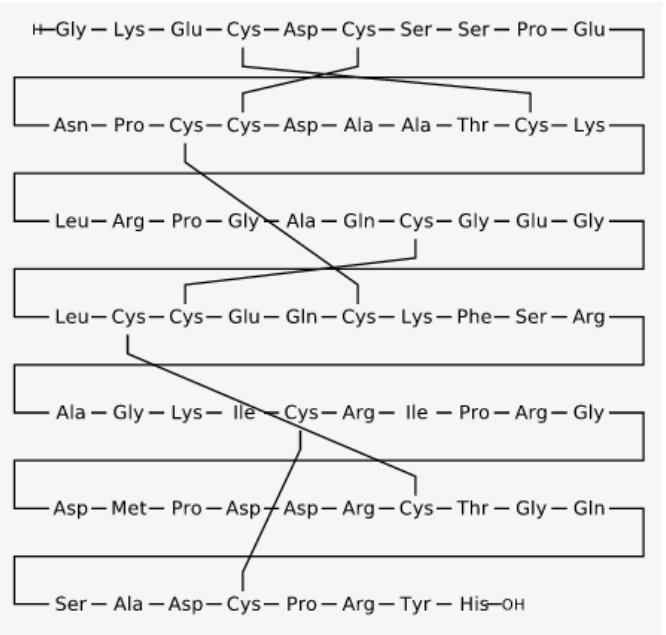
Large Molecule Drug
Human Growth Hormone
~3,000 atoms

Image credit:
https://www.edisongroup.com/wp-content/uploads/2019/06/1.-Exhibit-Molecular-structures-of-small-molecule-and-biologic-drugs.png

Why InChI for “large molecules”?

- InChI was designed with “small molecules” in mind
- Chemical informatic approaches are increasingly expected to handle entities beyond small molecules
- Canonicalization approaches lacking for large molecules containing more than linear sequences of “monomers”
- Cannot simply increase the allowed count of atoms/bonds
  - Excessive compute time
  - InChI too long
  - Greater chance of InChIKey collisions
- New algorithms or approaches needed to improve utility





“Large molecule” effort aims?

- Help to further validate the limits of the experimental 32K atom/bond limit of InChI
- Provide ‘enhanced’ large molecule optimized capabilities
  - more compact representation, extra layers
  - each layer handling a degree of 'sameness'
- Canonicalize large molecule representation
- Reduce the computation complexity to work with large molecules
  - e.g., “pseudo atoms” for common monomers
- Handle large biopolymer containing molecules with variability

Image credit:
https://www.health.harvard.edu/media/content/images/TargetBullseye_ML1602_ts78818071.png

How will these aims be accomplished?

- DyVInChI - dynamic variable InChI, pseudo atoms
  - Use predefined 'well behaved' monomers, no variation allowed
- Leverage pre-existing file formats
  - Biovia MOL/SDF Self-Contained Sequence Representation (SCSR), ChemAxon extended SMILES, and Pistoia HELM
- Design a layered ‘large molecule’ specification built on-top-of the InChI code base

Image credit:
https://images.saba.com/uploads/blog/what-should-you-do-when-you-dont-accomplish-your-goals/what-should-you-do-when-you-dont-accomplish-your-goals.jpg?s=1696x900

Progress towards large molecules in InChI?

- Max limit of 1K atoms/bonds increased to 32K (v1.0.5)
- Two mini-proposals circulated
- Any atom ('Zz') added (v1.0.6)
- Worked with HELM (Claire Bellamy) to establish a set of 'common' monomers using SugarNSplice (NextMove Software)

Add support for pseudo-atom biopolymer monomer representation within the IUPAC InChI standard

Purpose:

Enable chemically modified biopolymers identity to be rapidly accessed (by comparing InChI/Key strings). Add fc
atom approach
length. Make p
speed to compu

Add support for an 'any' atom within the IUPAC InChI standard.

Purpose:

Add support to InChI for dealing with molecular fragments. Provide InChI canonicalization and ide
utility to fragments, motifs, and other molecular entities that are not a completely defined chemic
structure. It is envisioned that this addition will enable additional enhancements and support of o
forms of variability and mixtures using the IUPAC InChI standard.

Monomer	#CIDs	#ChEMBLs	#PMIDs	#Patent	Exemplar CIDs	Exemplar ChEMBL
Ac3c	2,786	69	1,694	11,936	535 10217360 15925510	CHEMBL1098996 CHEMB
Ac5c	3,850	363	1,492	12,973	2901 10529589 16134986	CHEMBL105833 CHEMBL
Aib	17,653	800	5,300	62,228	2706 10885437 17938449	CHEMBL107381 CHEMBL
bAla	47,056	1,212	29,200	146,391	111 7827858 10412008 21	CHEMBL100038 CHEMBL
D-2Nal	3,767	687	9,549	86,366	119348 10189836 107248	CHEMBL103817 CHEMBL
D-Abu	15,270	47	30,831	21,455	121844 65808522 739271	CHEMBL1198130 CHEMB
D-aIle	2,603	334	11,464	39,217	83171 8574699 10148124	CHEMBL103538 CHEMBL
D-Ala	41,664	2,043	304,028	429,315	53370 9833688 12205382	CHEMBL100202 CHEMBL
D-Arg	7,305	1,601	101,098	218,773	12326 10102548 1090699	CHEMBL100712 CHEMBL
D-Asn	2,610	446	27,674	17,669	78334 6999728 7592257	CHEMBL1162388 CHEMB
D-Asp	5,146	879	61,248	66,574	22880 9811905 10532628	CHEMBL103089 CHEMBL
D-aThr	1,466	295	17,559	43,390	90624 7097851 10433955	CHEMBL1076803 CHEMB
D-Cys	5,889	1,315	290,056	61,323	82333 10395509 1132062	CHEMBL103799 CHEMBL
deamino-Cys	1,581	340	64,403	53,539	8229 10102548 11388106	CHEMBL10962 CHEMBL1

HELM collaboration to find a ‘common’ set of biopolymer monomers

Including PubChem and ChEMBL

- Create a list of ‘common’ biopolymer monomers
- Detect and prioritize biopolymers for inclusion in terms of:
  - Found in PubChem chemical records (frequently found)
  - Found in ChEMBL (measure of drug discovery importance)
  - Found in PubMed articles (measure of biomedical importance)
  - Found in patent records (measure of general utility)
- Examined amino acids, nucleic acids, and carbohydrates/glycans
- Future work is to formulate these into a ‘pseudo atom’ scheme
  - Compress the molecule graph when the molecule is ‘large’ and when it is primarily comprised of biopolymer monomers
  - Explore use of the InChI algorithm to canonicalize the graph of pseudo atoms

Variability handling proposal

Add support for a dynamic variability representation within the IUPAC InChI standard (DyVinchi)

Purpose:

Add formal support to IUPAC InChI for dealing with chemical structure variability using a three-layer approach built on top of the InChI with well-defined semantics that defines the molecular fragments, their variable attachment points, and their variable connectivity.

Dynamic Variability InChI – (Dy)Vinchi

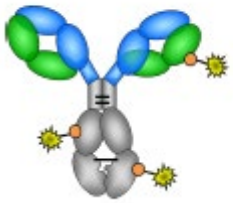
1. Define entities
   - in non-attached state
   - addressable by order
2. Define attachments
   - heavy atom only
   - single valence only
3. Define variability
   - addressable by order
   - * spawns entity, 0=unkn

- Allows multiple attachments per entity
- Can nest variability
- Can provide counts
- Can provide ratios
- Handles Markush, variable connection, variable loading

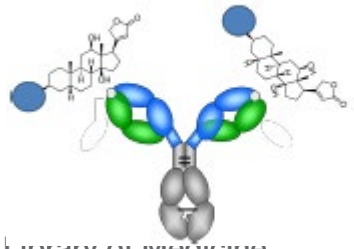
Biopoloymer Representation

HELM Line Notation Extension - Ambiguity

Connected monomer type is known, position unknown



Connection partner/monomer is undefined



Configuring Databases for Biopolyr

This section describes changes that you might nee
biopolymers using the BIOVIA Draw Sequence Too

Condensed Representation of Biopol

Biopolymer residues can be represented abbrevia
or ctab) underneath. The full structure conventior
sequence.

The full structure convention works well for oligon

DyVinchi Specifics

- Define Entities
  - InChI with "[..]"
    [InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H]
  - Order determines ordinal
- Layered separator "|"
- Define attachments
  - [#,[#,#]][#,#]
  - First # is entity
  - Second # is atom ID in InChI, can be a "[..]" list
  - Atom ID "0" means undefined
- Define variability
  - Two parts .. Attachment definition .. Attachment to what
  - Attachment definition can be nested with "[..]" list
  - Simplest is [#,#]
  - First number is entity
  - Second # is atom ID
  - Every # can be nested
  - Atom ID can include a ratio with ":" separator
    - E.g., "[2:1]" atom 2, 1 part

What is needed?

- Volunteers to contribute:
  - Expertise
  - Use cases

bolton@ncbi.nlm.nih.gov

You can help improve chemical informatics

Special thanks

- Igor Pletnev
- IUPAC InChI “Large Molecule” working group (including the former head Keith Taylor)
- NextMove Software (Roger Sayle, John May), and alumni (Noel O’Boyle and Daniel Lowe)
- HELM Pistoia Alliance collaborators (including Claire Bellamy and Anna Gaulton)
- IUPAC InChI subcommittees, InChI Trust, IUPAC CPCDS (Leah McEwen), various scientific workshops, discussions, and communications
- All PubChem Contributors and Collaborators

- This research was supported by the Intramural Research Program of the NIH, National Library of Medicine