InChI's core value in the ecology of life science data standards.

***Yulia Borodina**, FDA/Office of Health Informatics*

***Gunther Schadow**, Pragmatic Data, LLC*

Disclaimer

The views and opinions presented here represent those of the speakers and should not be considered to represent advice or guidance on behalf of the Food and Drug Administration.

FDA as InChI trust member (historical notes)

2008: FDA publishes "Guidance for Industry. Indexing Structured Product labeling".
https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM072317.pdf

2012: SPL extension for Substance Indexing begins. InChI is included as "must be present" chemical structure identifier
https://www.fda.gov/downloads/ForIndustry/DataStandards/StructuredProductLabeling/UCM321876.pdf#page=123

2013: FDA adds Substance Indexing to its Indexing Initiative
https://www.fda.gov/downloads/ForIndustry/DataStandards/StructuredProductLabeling/UCM345939.pdf

2014: First Substance Index SPL files are published on SPL site (later on DailyMed) and PubChem.
FDA becomes an official member of InChI trust

Structured Product Labeling (SPL)

*Health Level Seven (HL7) Structured Product Labeling (SPL) is an **ANSI-accredited data exchange standard** which was adopted in 2004 by FDA for the exchange of health and regulatory product and facility data to be used internally and, **in some cases**, also provided to the public.*

Scope (constantly expanding):

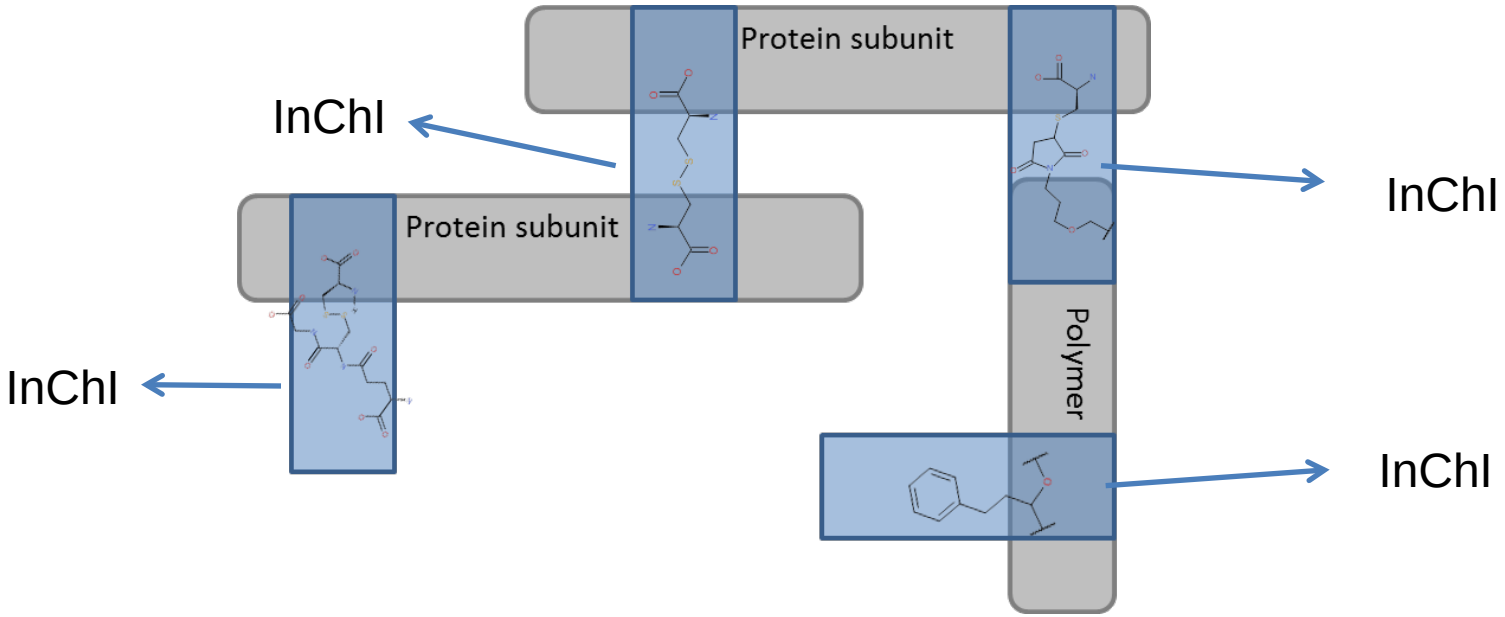
Drug product monographs / labels (FDA)
Pharmacologic Class Indexing of substances
Substance Indexing
Federal regulations about pesticide residue tolerance (EPA)
Identification of Medicinal Products (IDMP) international standard
PQ/CMC Substance / Product Manufacturing and Quality Data

SPL Substance Indexing

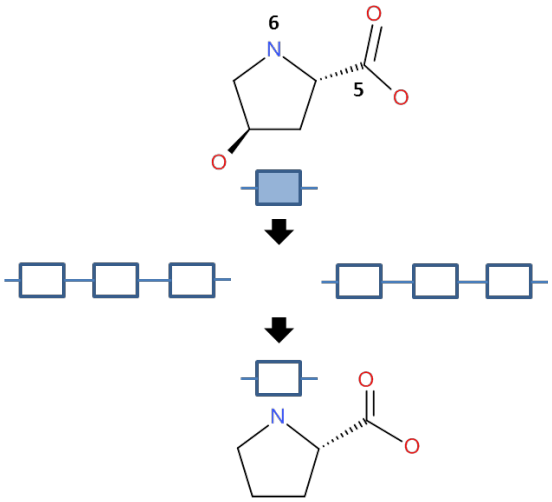
- **Publishing definitions of substances and substance codes (UNIIs)**
- **Substance definition is a set of characteristics that unambiguously defines a substance**
- **SPL model for substances is developing in stages**
- **Definitions of the following substances are available on DailyMed**
  - Chemicals
  - Chemical mixtures
  - Biological organisms
  - Proteins, including proteins with PTMs
- **Work in progress**
  - Polymers
  - Protein-polymeric conjugates

SPL Substance Indexing and InChI

- **Using InChI to unambiguously define chemical moieties in simple and complex substances**
- **Using InChI atom numbering to identify connection points in connected moieties**



Using InChI atom numbering to indicate connection points in modified amino acids

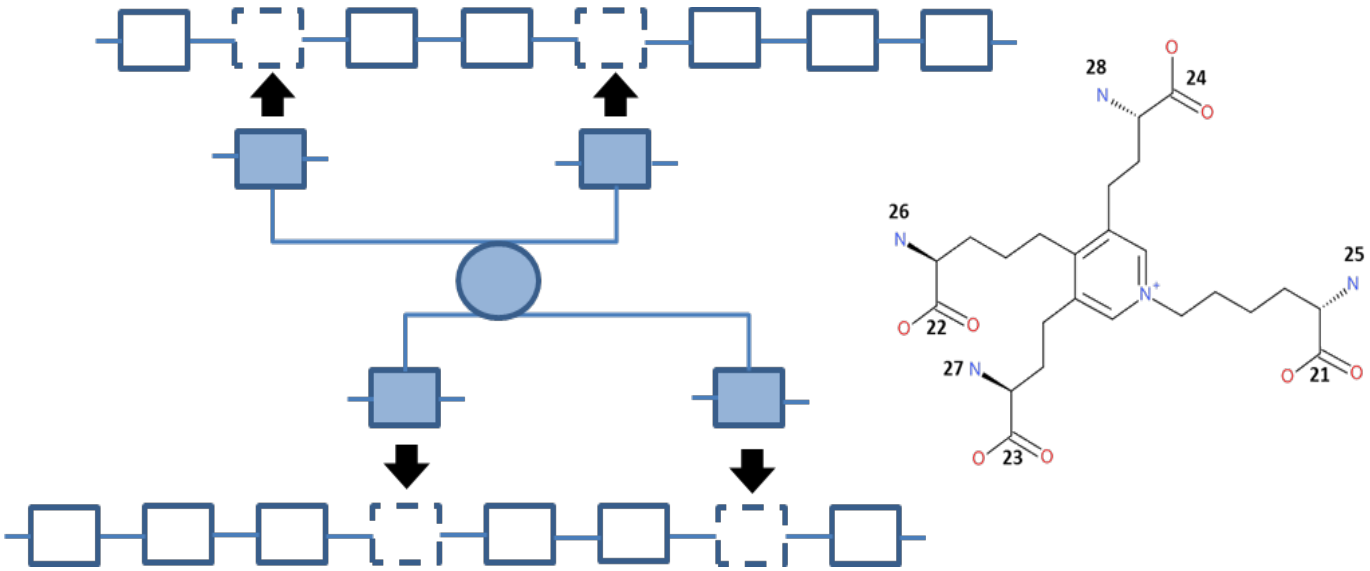


Model of proline – hydroxyproline substitution

InChI=1S/C5H9NO3/c7-3-1-4(5(8)9)6-2-3/h3-4,6-7H,1-2H2,(H,8,9)/t3-,4+/m1/s1

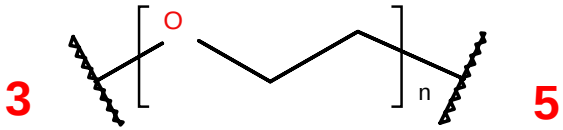
```
<moiety>
    <code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="Amino acid connection points" />
        <positionNumber value="6" />
        <positionNumber value="5" />
</moiety>
```

Using InChI atom numbering to indicate connection points in links



Model of a desmosine link in proteins

Using InChI pseudo atom numbering to indicate connection points in polymeric moieties



Representing polyethylene glycol by its canonical SRU and SRU connection points

InChI=1B/C2H4O/c1-2-3-1/h1-2H2/z101-1-3(1,2,1,3,2,3)

<moiety>

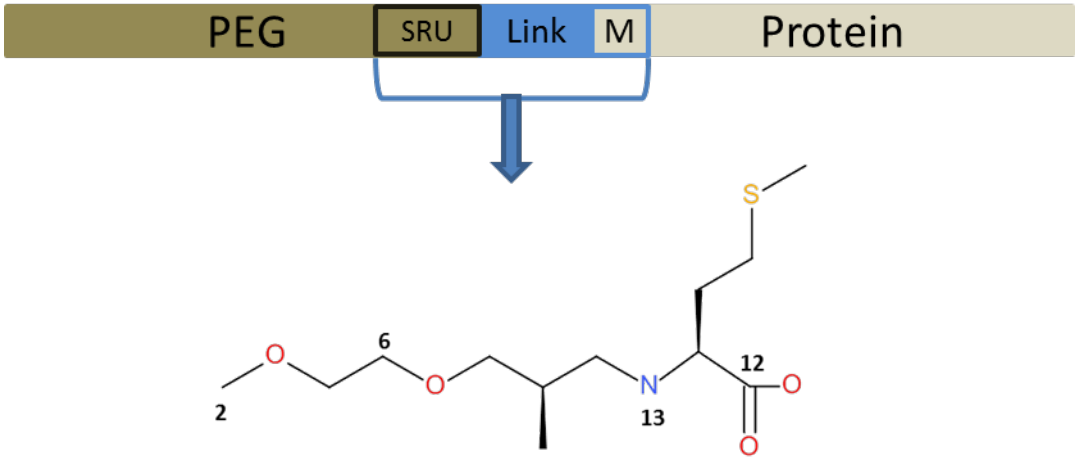
<code code="C132921" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="**Linear structural repeat unit connection points**" />

<positionNumber value="**3**"></positionNumber>

<positionNumber value="**5**"></positionNumber>

</moiety>

Protein-polymeric conjugates

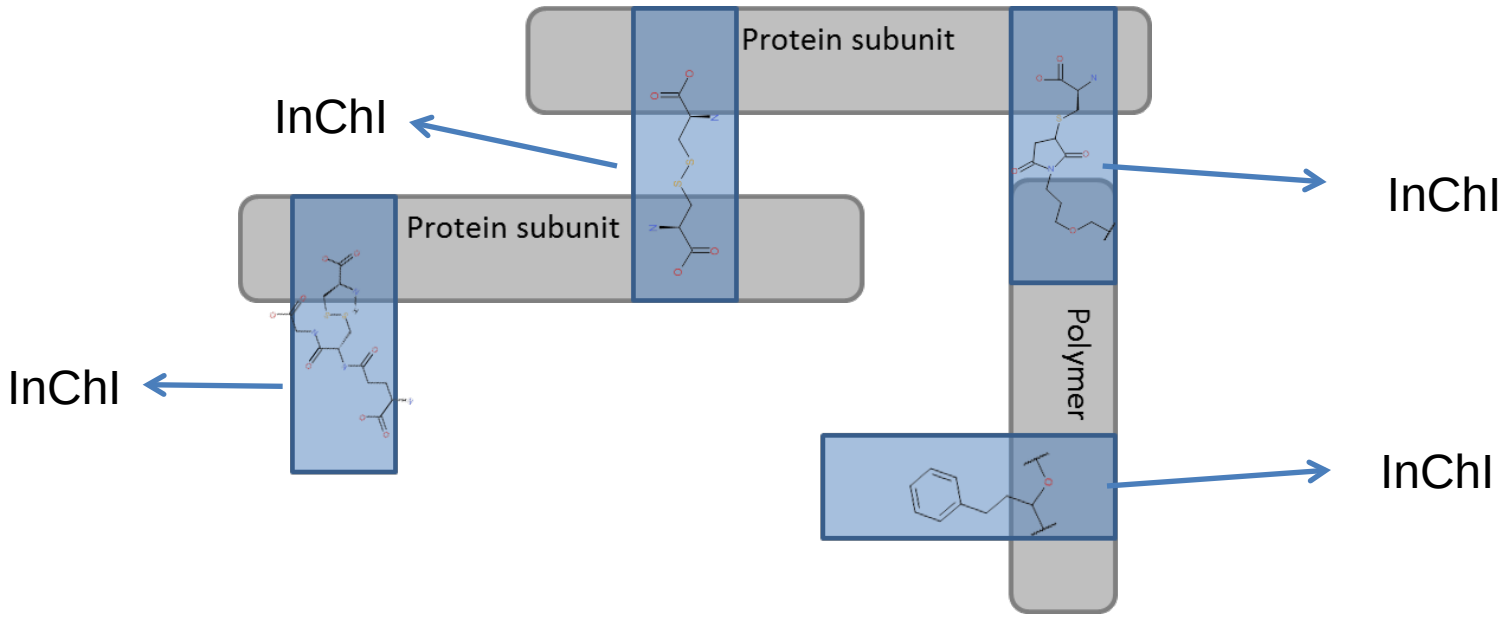


Connection between polymeric and protein moieties

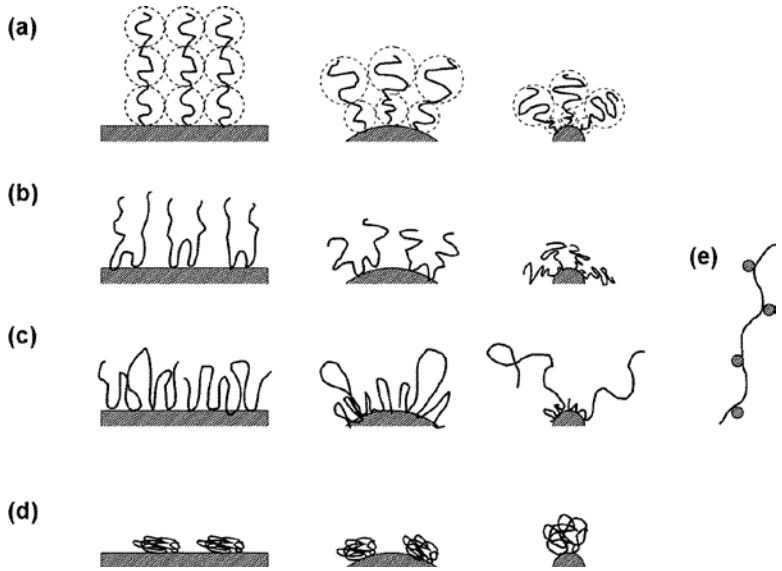
```
<moiety>
    <code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1"
          displayName="Amino acid connection points" />
    <positionNumber value="13" />
    <positionNumber value="12" />
</moiety>
<moiety>
   <code code="C132921" codeSystem="2.16.840.1.113883.3.26.1.1"
         displayName="Linear structural repeat unit connection points" />
     <positionNumber value="2" />
     <positionNumber value="6" />
</moiety>
```

SPL Substance Indexing and InChI

- **Linking between moieties is unambiguously defined, so that a complete molecular model can be recreated**



SPL model takes into consideration the fact that polydisperse substances (polymers, proteins) are NOT just large molecules !



They are MIXTURES of heterogeneous molecules

Polydisperse substances in SPL

- **Mixtures are represented by moieties of type "mixture component" with a quantity on each component.**
  - numerator amount of component per denominator amount of total mixture
  - stoichiometric (number ratio), mol ratio (mol per mol), or mass ratio (g per g)
  - can accommodate undefined or uncertain amount (e.g., 2-5 units of component per 1 unit of mixture)

- **Proteins are represented by protein subunits and modifications with a quantity on each component**

- **Polymers are represented by Structural Repeat Units (SRUs) and modifications with a quantity on each component**

InChI have limited application for polydisperse substances

- **We don't use InChI for identification of polydisperse substances**
  - InChI's chemical layers are inadequate for representing proteins with posttranslational modification. Using layers of kind <amino acid sequence>/<modifications> would be more practical but would require a complete restructuring of InChI algorithm
  - InChI for polymers don't convey polymerization degree necessary for identifications of polymers
  - InChI for mixtures don't exist
- **We use/intend to use InChI for identification of moieties of polydisperse substances**
  - Modified amino acids
  - SRUs
  - Mixture components

Identifier Creation and Hashing

- **To guarantee that the same substance is identified by only one code, we compute a substance definition hash code.**
  - Similar to InChI-KEY
- **Hashing compositions by hashing the hashes of the**
  - Details are still work in progress
  - Requires fully specified normalization of sub-moieties
    - E.g. replacing modified AAs in peptide chains with “X”
  - Canonical ordering of multiple sub-moieties
    - E.g. alphabetic order of peptide chains, followed by modifications.
  - Developers made certain ad-hoc decisions
- **Only with multiple independent implementations can we prove that our hashing is robust.**
  - A similar issue exists with the InChI development, it is only one code base, is there a reproducible specification?

Eco-system of standards

- **Specialty content standards (e.g. MPEG, InChI) allow us to describe specific things in ways never before possible.**
  - Should avoid specifying too much composition
    - That dilutes the core value and happened with MPEG, MOLFILE, etc.
- **General composition standards (e.g. email, SPL) allow us to combine information to higher order structures.**
  - Should not re-inventing the wheel of special content types
    - Instead allow embedding of special content types
- **Some cross-over is needed**
  - InChI content cannot just exist as a “black box”
    - We need atom numbers to point inside InChI structures.
    - We might create InChI profiles which restrict certain features.
  - InChI can still exercise influence on how it will be used.
    - InChI or IUPAC could develop conceptual recommendations on how InChI could be logically combined.

Recommendations to InChI

- **Don't compete with more complex standards such as SPL.**
- **Do publish specification of InChI that others can implement.**
  - Only an test with multiple independent implementations can prove the specification is fully explicit and unambiguous.
- **Do not try to identify mixtures by extending InChI.**
- **Don't try to identify proteins by extending InChI.**
- **Don't try to identify all polymers by extending InChI.**
- **InChI and IUPAC need not lose control over these matters**
  - May publish recommendations on the use of InChI in higher sturctures
  - May publish conceptual models of mixtures, proteins, and polymers.
  - Identifiers for polydisperse substances don't have to be InChI! If IUPAC wants to create identifiers for polydisperse substances, they should
    - Use a standard terminology
    - Use layers that make sense for polydisperse substances
    - Use an input format that is more appropriate for polydisperse substances, such as SPL

Next Steps, would InChI be interested?

- **… in participating with composition standards and content database developers to standardize:**
- **The right use of mixture notations?**
  - Covalent bonds, ionic bonds that dissociate in normal solution …
  - When and when not to use the period in InChI layers 1 and 2.
  - The conceptual model of mixtures and dissociating moieties.
- **Fully defining real life proteins?**
  - Our proteins SPL model could help other protein data resources
  - Pharmaceutical Industry, Proteomics,
    - UNIPROT, e.g., contains some PTM information but is unable to represent complete real functioning proteins.
- **Fully defining polymers?**
- **These could be organized as IUPAC projects.**