How InChI helps describe and identify biopharmaceuticals

Yulia Borodina

Office of Health Informatics

InChI symposium Fall 2019 San Diego, CA, August 23 - 24, 2019
Biopharmaceuticals

- Substances extracted from biological organism or synthesized on biological matrices
- Majority are proteins
- Often conjugated with small molecules, polymers or themselves in order to increase bioavailability, immunogenicity or to deliver an unspecific toxin
- Therapy of cancer, rheumatoid arthritis, migraine, asthma etc.
- Cost is very high
<table>
<thead>
<tr>
<th>Clinical Trial</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NCT Number:</strong></td>
</tr>
<tr>
<td><strong>Title:</strong></td>
</tr>
<tr>
<td><strong>Conditions:</strong></td>
</tr>
<tr>
<td><strong>Sponsor:</strong></td>
</tr>
<tr>
<td><strong>Intervention:</strong></td>
</tr>
<tr>
<td><strong>Phases:</strong></td>
</tr>
<tr>
<td><strong>Age Groups:</strong></td>
</tr>
<tr>
<td><strong>Study Results:</strong></td>
</tr>
<tr>
<td><strong>Enrollment:</strong></td>
</tr>
<tr>
<td><strong>Other IDs:</strong></td>
</tr>
</tbody>
</table>
RINDOPEPIMUT described by USAN

STATEMENT ON A NONPROPRIETARY NAME ADOPTED BY THE USAN COUNCIL

USAN (WW-34)  RINDOPEPIMUT

PRONUNCIATION  rin’ doe pep’ i mut

THERAPEUTIC CLAIM  Immunotherapy for glioblastoma multiforme
and potentially other tumor types

CHEMICAL NAME

L-leucyl-L-alpha,-glutamyl-L-alpha,-glutamyl-L-lysyl-L-lysylglycyl-L-asparaginyl-L-tyrosyl-L-
valyl-L-valyl-L-threonyl-L-alpha,-aspartyl-L-histidyl-L-cysteine bicether with N-[4-({3-
mercapto-2,5-dioxo-1-pyrrolidinyl}) methyl][cyclohexyl][carboxyl]-keyhole limpet hemocyanin
(Megastoma cransulata)(1:1)

STRUCTURAL FORMULA

MOLECULAR WEIGHT  450-550 kDa

TRADEMARKS  Rintega™

MANUFACTURER  Cellidex Therapeutics, Inc.

CODE DESIGNATION  CDX-110

CAS REGISTRY NUMBER  1108208-65-6

Hemocyanin
Hemocyanin

Keyhole limpet hemocyanin (KLH) is a very large, copper-containing protein molecule derived from the haemolymph of the inedible mollusc, Megathura crenulata. KLH is a highly immunogenic T-cell dependent antigen that is used increasingly in immunotoxicological studies, particularly in those involving animals.
**We do not use InChI to define primary protein structure**

Primary protein structure is defined by amino acid single letter code

<table>
<thead>
<tr>
<th>Code</th>
<th>Amino acid</th>
<th>Code</th>
<th>Amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>alanine</td>
<td>I</td>
<td>isoleucine</td>
</tr>
<tr>
<td>R</td>
<td>arginine</td>
<td>L</td>
<td>leucine</td>
</tr>
<tr>
<td>N</td>
<td>asparagine</td>
<td>K</td>
<td>Lysine</td>
</tr>
<tr>
<td>D</td>
<td>aspartic acid</td>
<td>M</td>
<td>methionine</td>
</tr>
<tr>
<td>B</td>
<td>asparagine or aspartic acid</td>
<td>F</td>
<td>phenylalanine</td>
</tr>
<tr>
<td>C</td>
<td>cysteine</td>
<td>P</td>
<td>proline</td>
</tr>
<tr>
<td>E</td>
<td>glutamic acid</td>
<td>S</td>
<td>serine</td>
</tr>
<tr>
<td>Q</td>
<td>glutamine</td>
<td>T</td>
<td>threonine</td>
</tr>
<tr>
<td>Z</td>
<td>glutamine or glutamic acid</td>
<td>W</td>
<td>tryptophan</td>
</tr>
<tr>
<td>G</td>
<td>glycine</td>
<td>Y</td>
<td>tyrosine</td>
</tr>
<tr>
<td>H</td>
<td>histidine</td>
<td>V</td>
<td>Valine</td>
</tr>
<tr>
<td>X</td>
<td>( wildcard / placeholder)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We use InChI to define modifications to the primary protein structure

• Modified amino acids
• Links
• Prosthetic groups
Example: definition of a modified amino acid

Lysine linked to Epidermal Growth Factor Receptor variant III

InChI canonical atom numbers for atoms that substitute N of amino group and C of carboxyl group of a natural amino acid
Example: definition of a link

CYSTEINE-GLUTATHIONE DISULFIDE

InChI canonical atom numbers for atoms that substitute N of amino group and C of carboxyl group of a natural amino acid

InChI=1S/C13H22N4O8S2/c14-6(12(22)23)1-2-9(18)17-8(11(21)16-3-10(19)20)5-27-26-4-7(15)13(24)25/h6-8H,1-5,14-15H2,(H,16,21)(H,17,18)(H,19,20)(H,22,23)(H,24,25)/t6-,7-,8-/m0/s1

(N,C) pairs

(14,12)
(15,13)
Example: definition of type 3 copper center

InChI=1S/6C6H9N3O2.2Cu/c6*7-5(6(10)11)1-4-2-8-3-9-4;;/h6*2-3,5H,1,7H2,(H2,8,9,10,11);;/q;;;;2+2/p-4/t6*5-;;/m000000../s1

(N,C) pairs
(7,6)
(18,17)
(29,28)
(40,39)
(51,50)
(62,61)
Identification

- InChI and ordered list of canonical (N,C) pairs provide that every modification is uniquely identified.

- Location of a modification in the protein is specified by positions of substituted amino acids.

- For example, Type 3 copper center position in hemocyanin is (2940, 2959, 2968, 3069, 3073, 3100).

- Complete 2D chemical structure of the modified protein can be reconstructed if needed.
Uncertain modifications are annotated by empirical probabilities

<quantity>
  <numerator xsi:type="URG_PQ" value="30" unit="mol"/>
  <denominator value="170" unit="mol"/>
</quantity>
Data on (some) proteins with UNIIIs are published

- Substance structures in XML format
  https://dailymed.nlm.nih.gov/dailymed/spl-resources-all-indexing-files.cfm

- Substance names in text format
  https://fdasis.nlm.nih.gov/srs/

- Other product-related information in XML format

- SPL Substance Implementation Guide is #14 in
  https://www.fda.gov/media/84201/download
Protein descriptions provided by PubChem are misleading

This is not Aprutumab ixadotin
Protein descriptions provided by PubChem are misleading.
Hemoglobin Glutamer

Red Blood Cells (Bovine Derived) → Hemoglobin → Native HGB
- Tetramer 64 kD
- Dimer 32 kD

Stabilized Tetramer
- 64 kD

Gluteraldehyde Polymer

Hemopure
- Average 250 kD

Clinical Investigator’s Brochure, Version: ISHEM-003.20 Sept05