SPL format for proteins with structural modifications

**Abstract**

**Key words:** standard, substance, proteins, modified proteins, posttranslational modification, data exchange standard, informatics

1. **Introduction**

Protein-based therapeutic products contribute significantly to the EU and US drug markets. More than 30% of novel drugs approved by the US Food and Drug administration (FDA) in 2015-2016 had a protein as an active ingredient. During the same period FDA approved protein-based biosimilars Zarxio and Inflectra. Proteins accounted for almost 40% of pharmaceutical substances that received INN names in 2016. Hundreds of proteins are in clinical trials for treating cancer, immunological disorders, infections, hematological disorders and other diseases. Most proteins are more complex in structure than conventional drugs and are more difficult to identify or characterize. The bioinformatics approach typically used for characterization of proteins is limited to description of polypeptides consisting of only natural amino acids. However, proteins often have additional structural characteristics also known as modifications. Modifications that happen in result of natural biochemical processes following the translation are known as post-translational modifications. They include enzymatic addition of functional groups (e.g., phosphorylation), sugar molecules (glycosylation), lipids (lipidation), co-factors; formation of disulfide links; cleavage of a signal peptide etc. Modifications can also happen during the translation if a reprogrammed ribosomal synthesis is used. Proteins can by modified in a chemistry lab after isolation from a biological host or they can be chemically synthesized. All these techniques result in insertion of unnatural amino acids in a polypeptide chain or in forming covalent links between amino acids in a chain or between different chains. Modifications that contribute to the biological activity should generally be well characterized and taken into account for unique identification of a therapeutic protein.

This article describes a machine readable format for structural characterization of proteins and modified proteins in medicinal products. The format is an extension of the Health Level Seven (HL7) Structured Product Labeling (SPL) standard adopted in 2004 by FDA for the exchange of health and regulatory product data. It conveys structural characteristics of proteins described in in ISO standard 11238 “Health informatics — Identification of medicinal products — Data elements and structures for the unique identification and exchange of regulated information on substances” (IDMP). Here we provide general description of the format and its applications to use cases from the FDA Substance Registration System and from UNIPROT.

1. **Materials and Methods**
   1. **ISO recommendations for identifying medicinal proteins**

ISO 11238 (<https://www.iso.org/obp/ui/#iso:std:iso:11238:ed-1:v1:en>) is an international standard that describes data elements and structures for the unique identification and exchange of regulated information on substances. It is one of a group of five standards that together provide the basis for the unique identification of medicinal products. Among other substances, ISO 11238 sets forth the principles of identification of proteins and modified proteins in medicinal products. The most important principles for our work are the following:

* Proteins that differ in protein sequence, type of glycosylation, disulfide linkages or glycosylation site shall be defined as separate substances.
* The description of modified proteins shall capture structural changes that result from the modification when a definitive structure is known.
  1. **Structured Product Labeling (SPL) format**
  + **Reference Information Model**
  + **Header**
  + **Structured Body**
  + **Codes and Identifiers**
  + **Validation procedures**
  1. **SPL concept codes in NCI thesaurus**
  2. **SPL format for substances**

The SPL format for substances is a specific subclass of SPL format supporting the exchange of substance information. Concept “substance” is understood as per ISO 11238 standard as any matter of defined composition that has discrete existence, whose origin may be biological, mineral or chemical. SPL format is currently used by FDA for indexing information about substances registered in the FDA Substance Registration System and assigned UNique Ingredient Identifiers (UNIIs). (<http://www.fda.gov/downloads/ForIndustry/DataStandards/StructuredProductLabeling/UCM345939.pdf>)

( <http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII>)

SPL substance index files are available for download at <https://dailymed.nlm.nih.gov/dailymed/spl-resources-all-indexing-files.cfm>

* + 1. **Substance definition**

A substance definition is a set of characteristics that unambiguously describe a substance. One SPL substance index file contains a definition of one main substance. It may include definitions of auxiliary substances that constitute the main substance. Alternatively, it may include references to auxiliary substances defined elsewhere.

* + 1. **Substance identifiers and codes**

The main substance must be associated with one substance code. The XML element <code> is reserved to indicate the code and the code system that assigned the code. The code system identified by OID “2.16.840.1.113883.4.9.”is the FDA Substance Registration System that generates UNIIs.

<code code="2F35K89X9W" codeSystem="2.16.840.1.113883.4.9" />

There might be other codes mapped to the substance. Equivalence mappings declare that the substance is considered equivalent to another description of the substance in a different system.

<asEquivalentSubstance>

<definingSubstance>

<code code="baebd25c-8a81-4ea0-665b-e1cc064f28f6" codeSystem="2.16.840.1.113883.3.2705" />

* + 1. **Moieties**

All structurally defined substances are represented by one or more structural units (moieties.) We do not label a substance as “protein”, “polymer”, “chemical”, etc. The role of a substance is determined by the types of moieties used to describe it. The following types of moieties have been implemented to date:

• Simple chemical

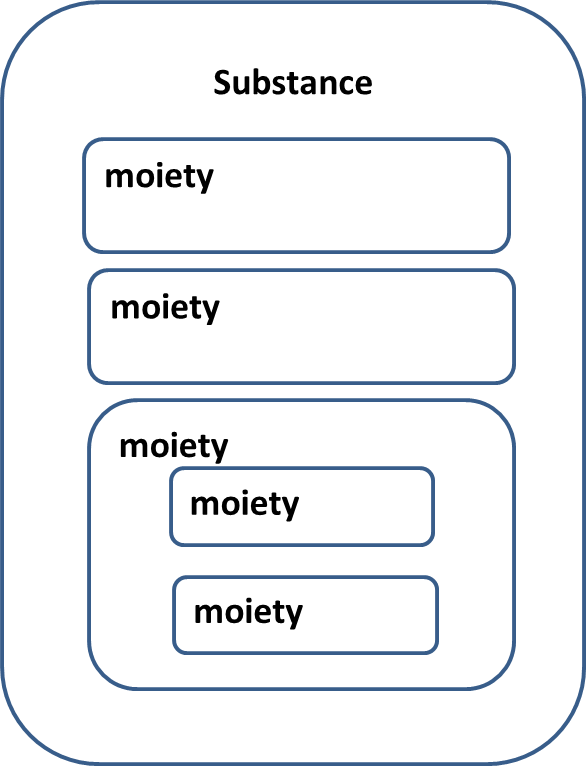
• Mixture component

• Protein subunit

• Structural modification

• Amino acid connection points

The SPL convention allows mixing different types of moieties in one substance. It also allows the use of nested moieties (see Fig…).



Figure… Schematic representation of a substance by moieties

The XML element “moiety” contains or references the definition of the moiety and/or the relationship between the moiety and the substance. The relationship between hierarchical moieties is similar to the relationship between a moiety and a whole substance. The following schema shows how information is organized inside the moiety element.

<moiety>

type of moiety (must be present)

moiety-substance relationship (may be present)

definition of the moiety (may be present)

<partMoiety>

moiety ID (may be present)

reference to an auxiliary substance definition (may be present)

<moiety> (may be present)

Definitions as well as relationships vary depending on the type of the moiety.

* + 1. **Moiety “Protein subunit”**

Concept “protein subunit” has code C118424 in the NCI thesaurus code system and is defined as a single polypeptide chain of a protein. Moiety “Protein subunit” contains or references a definition of a protein subunit. Usually the definition is provided right inside the moiety. A protein subunit is defined by its chemical structure represented using the one-letter amino acid notation. Table1 shows the letters allowed in the amino acid sequence.

**Table 1: Amino Acid Letter Codes.** Capital letters designate the L-configuration. Lower case letters designate the D-configuration. X designates a non-standard amino acid.

|  |  |
| --- | --- |
| **Letter code** | **Amino acid** |
| A (a) | Alanine |
| R (r) | Arginine |
| N (n) | Asparagine |
| D (d) | Aspartic acid |
| B (b) | Asparagine or aspartic acid |
| C (c) | Cysteine |
| E (e) | Glutamic acid |
| Q (q) | Glutamine |
| Z (z) | Glutamine or glutamic acid |
| G (g) | Glycine |
| H (h) | Histidine |
| I (i) | Isoleucine |
| L (l) | Leucine |
| K (k) | Lysine |
| M (m) | Methionine |
| F (f) | Phenylalanine |
| P (p) | Proline |
| S (s) | Serine |
| T (t) | Threonine |
| W (w) | Tryptophan |
| Y (y) | Tyrosine |
| V (v) | Valine |
| X (x) | a non-standard amino acid |

Each protein subunit is assigned an identifier. The identifier is used locally for referencing the subunit.

The relationship between a protein subunit and the substance is described in the element “quantity”. Figure… shows an example of a moiety “protein subunit”.

<moiety xmlns="urn:hl7-org:v3">

<code code="C118424" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="PROTEIN SUBUNIT" />

<quantity>

<numerator value="1" unit="mol" />

<denominator value="1" unit="mol" />

</quantity>

<partMoiety>

<id extension="SU3" root="000dc88f-786a-46fc-882f-44e58192e570" />

</partMoiety>

<subjectOf><characteristic>

<code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="Chemical Structure" />

<value xsi:type="ED" mediaType="application/x-aa-seq" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">DIQLTQSPSSLSASVGDRVTMSCKSSQSVLYSANHKNYLAWYQQKPGKAPKLLIYWASTRESGVPSRFSGSGSGTDFTFTISSLQPEDIATYYCHQYLSSWTFGGGTKVQIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC

</value>

</characteristic></subjectOf>

</moiety>

* + 1. **Moiety “Structural modification”**

Concept “Structural modification” has code C118425 in the NCI thesaurus code system. It is defined as a moiety that modifies the chemical structure of a substance. Each structural modification references the definition of an *auxiliary substance* that modifies the protein structure and describes how the particular instance of the *auxiliary substance* interacts with protein subunits. The XML element <bond> is reserved to describe the interaction. One modification may affect multiple amino acids so multiple bonds are allowed per a modification.

<moiety xmlns="urn:hl7-org:v3">

<code code="C118425" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="STRUCTURAL MODIFICATION" />

<partMoiety>

***Reference to an auxiliary substance definition or a concept definition***

<bond>

***Interaction of the auxiliary substance with a protein subunit***

</bond>

</partMoiety>

</moiety>

All post- translational modifications and other variances from the standard amino acid sequence are specified by means of amino acid substitutions with the exception of structurally unspecified modifications which are specified by means of attachments. We use two different types of bonds depending on whether a modification is structurally specific or unspecific.

*Bond “AMINO ACID SUBSTITUTION POINT”* is used to indicate how a structurally specific modification links to the protein. Concept “Amino acid substitution point”has NCIt code C118426 and is defined as the location of amino acid substitution in a protein.

<moiety xmlns="urn:hl7-org:v3">

<code code="C118425" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="STRUCTURAL MODIFICATION" />

<partMoiety>

<code code="RMB44WO89X" codeSystem="f5921d17-60b9-435f-9543-80471ffe7e31" />

<bond>

<code code="C118426" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID SUBSTITUTION POINT" />

<positionNumber value="1" />

<positionNumber value="4" />

<distalMoiety>

<id extension="SU1" root="f5921d17-60b9-435f-9543-80471ffe7e31" />

</distalMoiety>

</bond>

</partMoiety>

</moiety>

Figure… An XML example of moiety “Structural modification” with bond “Amino acid substitution point”. The code and the code system reference the auxiliary substance definition. First positionNumber in the bond refers to a position in the auxiliary substance. Second positionNumber indicates the position of the modification in the protein subunit.

*Bond “STRUCTURAL ATTACHMENT POINT”* is used to indicate the position of a structurally unspecific modification in the protein. Concept “Structural attachment point”hasNCIt code C14050 and is defined as the position where a structural modification is attached.

<moiety xmlns="urn:hl7-org:v3">

<code code="C118425" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="STRUCTURAL MODIFICATION" />

<partMoiety>

<code code="C118431" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="MAMMALIAN TYPE GLYCAN" />

<bond>

<code code="C14050" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="STRUCTURAL ATTACHMENT POINT" />

<positionNumber value="299" />

<distalMoiety>

<id extension="SU1" root="9e48c3e9-da25-46fd-93da-b0256aa1639a" />

</distalMoiety>

</bond>

</partMoiety>

</moiety>

Figure… An XML example of moiety “Structural modification” with bond “Structural attachment point”. The code and the code system reference the concept substance definition. positionNumber indicates the position in the protein subunit where the structure is attached.

* + 1. **Moiety “Simple chemical”**

Moiety “Simple chemical” is the default moiety. No code is specified for this type of moiety. The moiety contains or references a definition of a chemical substance. The definition must include the IUPAC International Chemical Identifier (InChI) and InChIKey. It may also include a structural representation such as a MOLFILE. The structural representation may not have structural characteristics that are not covered by the InChI algorithm. An XML example of a simple chemical moiety is given in Fig… in section “Results”.

* + 1. **Moiety “Mixture component”**

Concept “Mixture component” has code C103243 in the NCI thesaurus code system. It is defined as a substance that occurs as part of a mixture substance. The moiety contains or references a definition of a substance. A moiety of this type may have nested moieties and must have a moiety-substance relationship. An XML example of a mixture component can be found in Fig… in section “Results”.

* + 1. **Moiety “Amino acid connection points”**

In an amino acid substitution one regular amino acid is replaced by an irregular amino acid, or any molecule that fits into the amino acid chain. Such a molecule must be defined by a chemical structure and amino acid connection points. This means, usually, one amino-group must be marked to substitute the amino group of the original amino acid, and one carboxyl group must be marked to substitute the carboxyl group of the original amino acid.

Concept “Amino acid connection points” has code C118427 in the NCI thesaurus code system. It is defined as the amino acid N and C atoms that participate in forming peptide bonds. Moiety “Amino acid connection points” is reserved for storing the canonical atomic numbers of the N and the C atoms. Atom numbers are normalized using InChI algorithm and therefore are invariant for the same molecule. See figure…

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="9" />

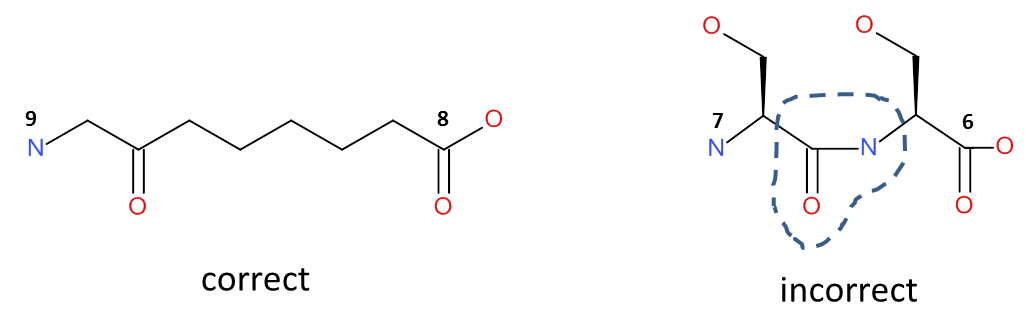
<positionNumber value="8" />

<partMoiety />

</moiety>

Fig… Example of moiety “Amino acid connection points”. First positionNumber indicates the canonical atom number of N of the amino group. Second positionNumber indicates the canonical atom number of C of the carboxyl group.

It is assumed that the molecule doesn’t have other peptide bonds between the amino acid connection points. Figure… shows a correct and an incorrect use of amino acid connection points.



Fig… Correct and incorrect use of amino acid connection points. Dotted line indicates a peptide bond that is not allowed between the connection points.

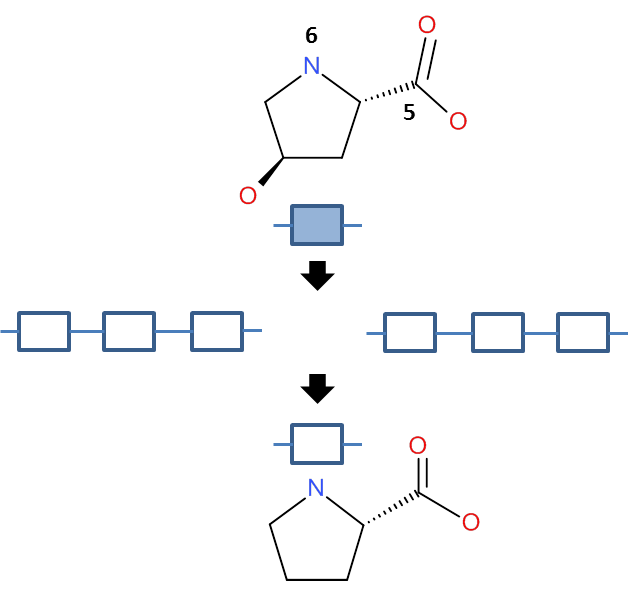
* + 1. **Auxiliary substance**

An auxiliary substance is the substance that is defined once and referenced multiple times from the main substance. Currently, the definitions of auxiliary substances are included in the same SPL substance index file. That provides a self-contained definition of the substance. However, using external definitions is also allowed.

An auxiliary substance that defines a chemical modification of a protein typically contains one moiety “Simple chemical” and one or more moieties “Amino acid connection points”. It may also contain moieties “Mixture component” if a modification is a mixture.

The set of moieties “Amino acid connection points” is a tuple. The order is determined by the positionNumber of N of the amino group. In case the positionNumber of N is not applicable, the order is determined by positionNumber of C of the carboxyl group.

1. **Results**
   1. **Use cases of protein modifications**
      1. **Non-standard amino acid**



Fig… Model of Proline – Hydroxyproline substitution.

The XML below represents the complete definition of the modifying substance Hydroxyproline. It references the Hydroxyproline substance registered in the FDA Substance Registration System via its UNII code “RMB44WO89X”. It provides both chemical structure (MOLFILE) and unique chemical structure identifiers (InChI and InChIkey) of Hydroxyproline molecule. Finally, it shows that N6 and C5 of Hydroxyproline substitute corresponding N and C termini of a standard amino acid.

<identifiedSubstance xmlns="urn:hl7-org:v3">

<id extension="RMB44WO89X" root="f5921d17-60b9-435f-9543-80471ffe7e31" />

<identifiedSubstance>

<code code="RMB44WO89X" codeSystem="f5921d17-60b9-435f-9543-80471ffe7e31" />

<asSpecializedKind>

<generalizedMaterialKind>

<code code="RMB44WO89X" codeSystem="2.16.840.1.113883.4.9" />

</generalizedMaterialKind>

</asSpecializedKind>

<moiety>

<quantity>

<numerator value="1" unit="mol" />

<denominator value="1" unit="mol" />

</quantity>

<partMoiety>

<code code="RMB44WO89X" codeSystem="2.16.840.1.113883.4.9" />

</partMoiety>

<subjectOf>

<characteristic>

<code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="Chemical Structure" />

<value xsi:type="ED" mediaType="application/x-mdl-molfile" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

<![CDATA[

-FDASRS- 01051521292D

9 9 0 0 1 0 0 0 0 0999 V2000

6.5634 -6.3611 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0

5.5593 -5.6403 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0

5.7801 -6.3611 0.0000 C 0 0 2 0 0 0 0 0 0 0 0 0

6.7760 -5.6528 0.0000 C 0 0 2 0 0 0 0 0 0 0 0 0

7.4926 -5.2277 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0

6.1718 -5.1902 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0

5.1968 -6.7944 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0

7.4926 -4.6403 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0

8.1843 -5.6361 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0

4 6 1 0 0 0 0

4 5 1 6 0 0 0

8 5 2 0 0 0 0

9 5 1 0 0 0 0

1 4 1 0 0 0 0

1 3 1 0 0 0 0

3 2 1 0 0 0 0

2 6 1 0 0 0 0

3 7 1 1 0 0 0

M END

]]>

</value>

</characteristic>

</subjectOf>

<subjectOf>

<characteristic>

<code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="Chemical Structure InChI" />

<value xsi:type="ED" mediaType="application/x-inchi" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">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</value>

</characteristic>

</subjectOf>

<subjectOf>

<characteristic>

<code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="Chemical Structure InChIKey" />

<value xsi:type="ED" mediaType="application/x-inchi-key" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">PMMYEEVYMWASQN-DMTCNVIQSA-N</value>

</characteristic>

</subjectOf>

</moiety>

<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" />

<partMoiety />

</moiety>

</identifiedSubstance>

</identifiedSubstance>

* + 1. **Terminal modification**

Terminal modifications are modeled same way as middle chain modifications. The position number of amino N is not applicable for N-terminal modifications and the position number of carboxyl C is not applicable for C-terminal modifications. For example, for N-terminus substitution of Glutamine with Pyroglutamate (Fig…) position of amino N is set to NA:

<moiety>

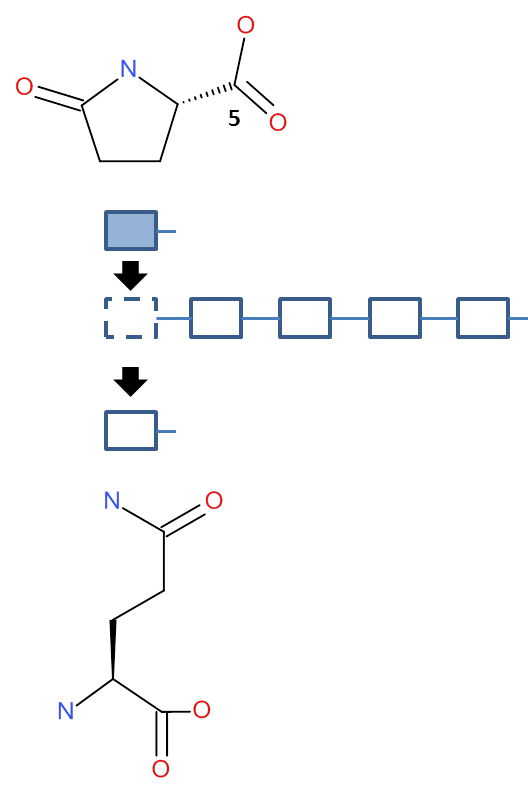
<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber nullFlavor="NA" />

<positionNumber value="5" />

<partMoiety />

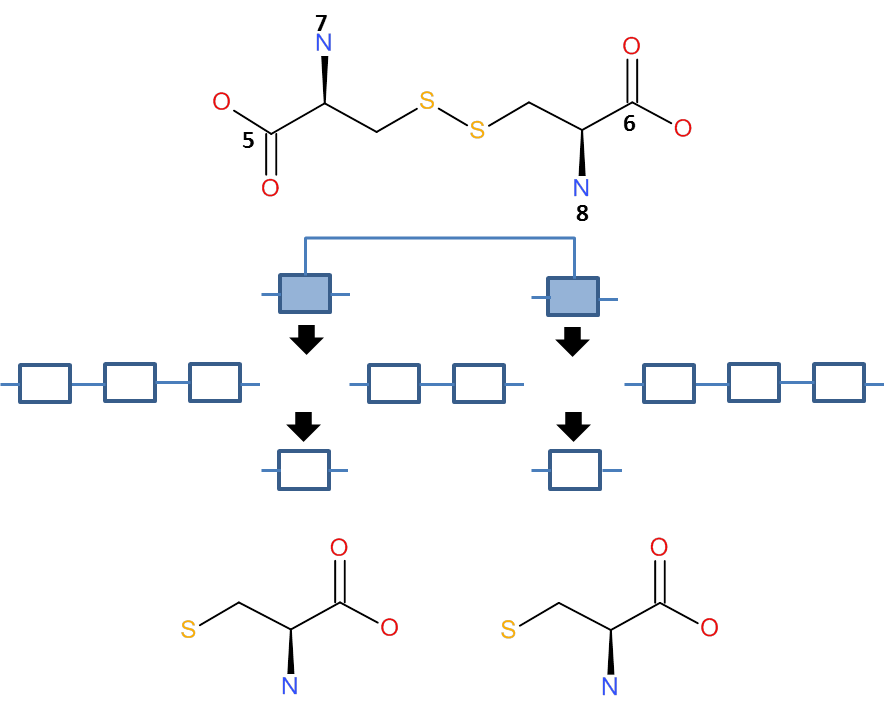
</moiety>



Fig… Model of Glutamine-Pyroglutamate substitution at N-terminus.

* + 1. **Two-site covalent link**

Links are modifications that substitute more than one amino acid in the chain(s). For example, in the disulfide bridge two Cysteine (C) amino acids are replaced by one shared Cystine.



Fig… Model of a Cystine link. Two Cysteine amino acids are substituted by a Cystine.

The Cystine, would be defined as an auxiliary substance with two sets of amino acid connection points:

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="7" />

<positionNumber value="5" />

<partMoiety />

</moiety>

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="8" />

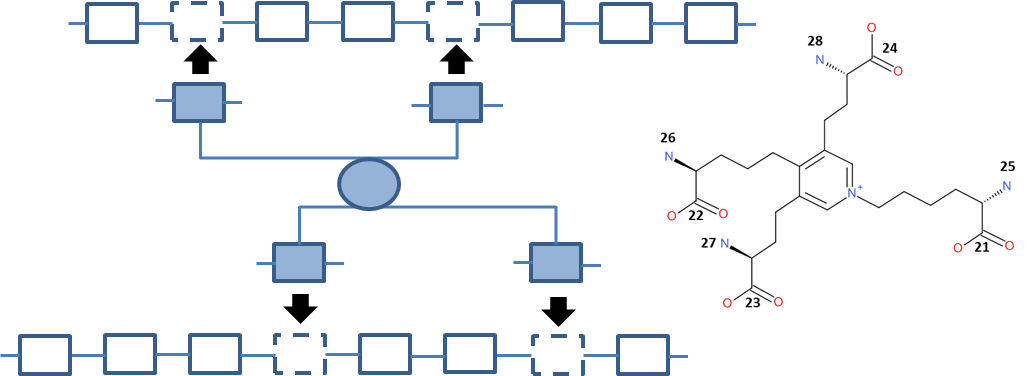
<positionNumber value="6" />

<partMoiety />

</moiety>

* + 1. **Four site covalent link**

Some modifications can cross-link more than two amino acids of a protein. For example, Desmosine links four amino acids (See Fig. …).



Fig… Model of Desmosine link.

Therefore, four moieties “Amino acid connection points” will be used in the definition of modifying substance Desmosine.

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="25" />

<positionNumber value="21" />

<partMoiety />

</moiety>

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="26" />

<positionNumber value="22" />

<partMoiety />

</moiety>

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="27" />

<positionNumber value="23" />

<partMoiety />

</moiety>

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="28" />

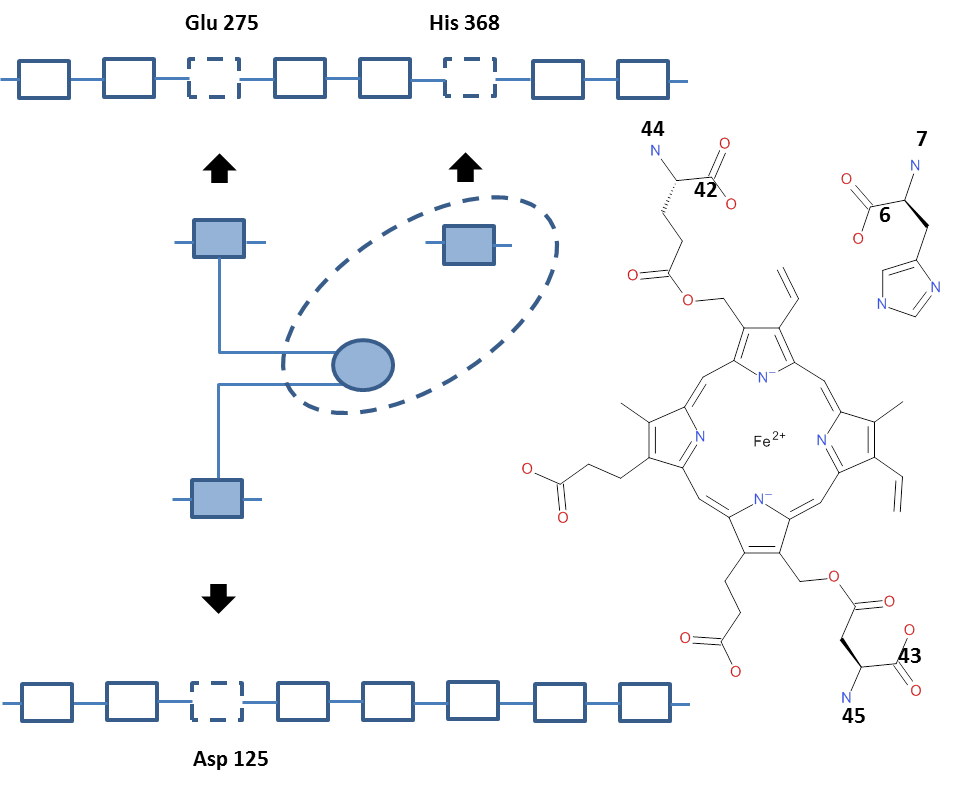
<positionNumber value="24" />

<partMoiety />

</moiety>

* + 1. **Non-covalent link**

Oftentimes proteins form electrostatic interactions or coordinate bonds with ligands or metal atoms. This type of interactions is typical for enzymes. It is possible to apply our approach to describe the active center of an enzyme as accurately as modern chemoinformatics allows one to describe ionic and coordinate interactions. For example, we can represent the heme binding site of Lactoperoxidase (see Fig…) that involves two covalently bound amino acids (Asp 125, Glu 275) and one amino acid (His 368) coordinated with iron(II) of the heme. All three are present in the same structure and have unique atom numbering. The metal-organic complex is represented as recommended in the FDA Substance Registration System Standard Operating Procedure.



Fig… Model of the heme binding site of Lactoperoxidase (Bovine). Binding sites are taken from UNIPROT (P80025, signal peptide removed). The structure of the heme is hypothesized based on literature.

The entire binding site is represented as a single structural modification with three pairs of connection points:

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="7" />

<positionNumber value="6" />

<partMoiety />

</moiety>

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="44" />

<positionNumber value="42" />

<partMoiety />

</moiety>

<moiety>

<code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID CONNECTION POINTS" />

<positionNumber value="45" />

<positionNumber value="43" />

<partMoiety />

</moiety>

* + 1. **Unspecified structure (Glycosylation)**

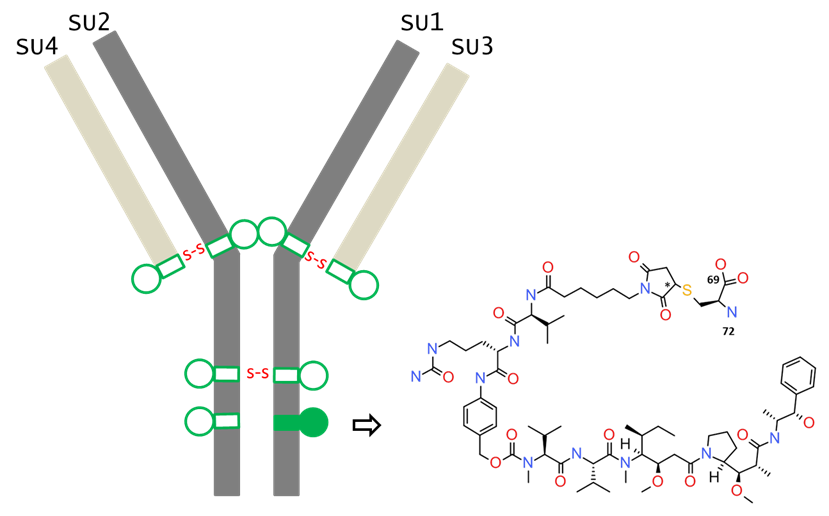
Proteins often undergo variable post-translational modifications either within the cell or bioreactor. Such variable modifications result in microheterogeneity of structures. ISO IDMP 11238 standard suggests ignoring microheterogeneity in most cases but leaves space for capturing some microheterogeneity essential for the functionality of the protein. For glycosylation, the position of the modification is typically specified but not the chemical structure of a glycan. Therefore SPL references the concept definition instead of the definition of a particular molecule.

Table… Concepts of glycans defined in NCI thesaurus

|  |  |
| --- | --- |
| Type of glycan | NCIt code |
| HUMAN TYPE GLYCAN | C118428 |
| AVIAN TYPE GLYCAN | C118429 |
| BACTERIAL TYPE GLYCAN | C118430 |
| MAMMALIAN TYPE GLYCAN | C118431 |
| FUNGAL TYPE GLYCAN | C118432 |
| PLANT TYPE GLYCAN | C128564 |

* + 1. **Uncertain structure**

Sometimes it is important to explicitly represent microheterogeneity of structures. For example, it may be important to indicate that an antibody-drug conjugate can have variable stereochemistry. (See Fig…). We model such situations by defining auxiliary substances as mixtures.



Fig… Drug molecule vedotin attached to a cysteine of a monoclonal antibody. Star indicates an atom with variable stereochemistry (epimeric center). “s-s” indicates disulfide bonds. A solid green rectangle-with-circle shape indicates a cysteine-vedotin complex. The corresponding white shapes with green borders indicate potential cysteine-vedotin complexes.

Moieties of type “Mixture component” should be used to represent a potential modification. Element <quantity> should be used to account for relative amount of the modification in the mixture. The following example indicates a one-to-one ratio of two epimers of vedotin-cystein conjugate.

<identifiedSubstance xmlns="urn:hl7-org:v3">

<id extension="vedotin-cysteine" root="2.16.840.1.113883.4.9" />

<identifiedSubstance>

<code code=" vedotin-cysteine " codeSystem="2.16.840.1.113883.4.9" />

<moiety>

<code displayName="mixture component" codeSystem="2.16.840.1.113883.3.26.1.1" code="C103243" />

<quantity>

<numerator value="1" unit="1" />

<denominator value="1" unit="1" />

</quantity>

<partMoiety >

***CHEMICAL STRUCTURE and AMINO ACID CONNECTION POINTS for EPIMER 1 should be described here***

</partMoiety >

</moiety>

<moiety>

<code displayName="mixture component" codeSystem="2.16.840.1.113883.3.26.1.1" code="C103243" />

<quantity>

<numerator value="1" unit="1" />

<denominator value="1" unit="1" />

</quantity>

<partMoiety >

***CHEMICAL STRUCTURE and AMINO ACID CONNECTION POINTS for EPIMER 2 should be described here***

</partMoiety >

</moiety>

</identifiedSubstance>

</identifiedSubstance>

This approach is not limited to variability in stereochemistry. We can use it for modeling more complex variations of structures such as glycosylation, provided the structures of the potential glycans are known.

* + 1. **Structural modification with uncertain position**

When a protein undergoes a non-enzymatic modification, the position of the modification may be uncertain. The number of modified amino acids can be partly controlled by amount of reagents added to the reaction. However, it is not always possible to control which positions in the chain will be modified. This uncertainty is modeled in SPL by using statistical amounts. For example, in lifastuzumab vedotin the monoclonal antibody lifastuzumab is linked to 3-4 molecules of cytotoxic agent vedotin via thioether conjugation with cysteine. Eight cysteine residues can be potentially modified: 219(SU3), 219(SU4), 223(SU1), 223(SU2)), 229(SU1), 229(SU2), 232(SU1), 232(SU2) (see Fig..).

Moiety-substance relationship <quantity> must be used to indicate the statistical chance of the modification. For lifastuzumab vedotin, the statistical chance of substitution of cysteine 232 with cisteine-vedotin complex would be represented as follows:

<moiety xmlns="urn:hl7-org:v3">

<code code="C118425" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="STRUCTURAL MODIFICATION" />

<quantity>

<numerator xsi:type="URG\_PQ" value="">

<low value="3" unit="1"/>

<high value="4" unit="1"/>

</numerator>

<denominator value="8" unit="1"/>

</quantity>

<partMoiety>

<code code=" 6603L01WUR " codeSystem="f5921d17-60b9-435f-9543-80471ffe7e31" />

<bond>

<code code="C118426" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="AMINO ACID SUBSTITUTION POINT" />

<positionNumber value="1" />

<positionNumber value="232" />

<distalMoiety>

<id extension="SU1" root="f5921d17-60b9-435f-9543-80471ffe7e31" />

</distalMoiety>

</bond>

</partMoiety>

</moiety>

1. **Notes**

* Our informational model assumes low frequency of modified or non-standard amino acids in a polypeptide chain. Significant increase in amount of modified or non-standard amino acids might lead to impracticality of this approach.
* Synthetic peptides with irregular (random) amino acid sequence cannot be described using this approach. They should be described as polymers of amino acids and not as structurally modified proteins.
* We don’t currently take into account atom equivalence when we number connection points

References:

1. Wiltschi B. Expressed Protein Modifications: Making Synthetic Proteins
2. Bradley L. et al. Chemical Synthesis of Proteins
3. Bradley L. Nilsson et al. Protein Assembly to Mine the Human Genome
4. Thomas J. Magliery. Unnatural Protein Engineering: Producing Proteins with Unnatural Amino Acids.
5. Dimitrov DS. Therapeutic proteins.
6. Nicholas Sioutosa et al. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information
7. Tracey D. Rae and Harold M. Goff. Lactoperoxidase Heme Structure Characterized by

Paramagnetic Proton NMR Spectroscopy

1. <http://crdd.osdd.net/raghava/thpdb/stats.php>
2. <http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugInnovation/default.htm>
3. <http://www.fda.gov/downloads/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/ucm127743.pdf>