**Conversion from SRS to SPL for peptides and proteins**

1. ***Data representation in SRS***
   1. Peptides are represented in three ways:
      1. Chemical structure
      2. XML that has amino acid sequence + modifications
      3. Both i and ii
   2. Proteins such as immunoglobulins are represented by XML that has amino acid sequences + disulfide bonds + modifications
   3. XML is stored in a text field which can be edited manually. No DTD, no schema, or simple control of a well-formed XML is implemented in the current application therefore there might be errors. There used to be multiple “schemas” used for proteins. We’ve cleaned them once but there may still be remains of old “schemas”.
   4. Example of sequence representation in XML

<SUBUNIT\_GROUP>

<SUBUNIT>**1**</SUBUNIT>

<LENGTH>**449**</LENGTH>

<SEQUENCE>**QVQLVQSGAEVKKPGETVKISCKASDYTFTYYGMNWVKQAPGQGLKWMGWIDTTTGEPTYAQKFQGRIAFSLETSASTAYLQIKSLKSEDTATYFCARRGPYNWYFDVWGQGTTVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK**</SEQUENCE>

</SUBUNIT\_GROUP>

* 1. Example of disulfilde bonds representation. Each bond shows which amino-acids are connected. Amino acid position includes a subunit ID and a position in the sequence.

<DISULFIDE\_LINKAGE>**1\_22-1\_96; 1\_146-1\_202; 1\_263-1\_323; 1\_369-1\_427; 2\_22-2\_96; 2\_146-2\_202; 2\_263-2\_323; 2\_369-2\_427; 3\_23-3\_93; 3\_139-3\_199; 4\_23-4\_93; 4\_139-4\_199; 1\_222-3\_219; 2\_222-4\_219; 1\_228-2\_228; 1\_231-2\_231**</DISULFIDE\_LINKAGE>

* 1. Example of a non- specific modification such as glycosylation. Structure of the sugar is not known. Only modification position on the sequence and on the amino acid is known.

<GLYCOSYLATION>

<GLYCOSYLATION\_TYPE/>

<N\_GLYCOSYLATION>**1\_299; 2\_299**</N\_GLYCOSYLATION>

<O\_GLYCOSYLATION/>

<C\_GLYCOSYLATION/>

</GLYCOSYLATION>

* 1. Example of a specific modification (structure of the modified amino acid is known)

<STRUCTURAL\_MODIFICATION\_GROUP>

<RESIDUE\_MODIFIED>**phenylalanine**</RESIDUE\_MODIFIED>

<RESIDUE\_SITE>**1\_6**</RESIDUE\_SITE>

<NUMBER\_OF\_FRAGMENTS\_MOIETIES>**1**</NUMBER\_OF\_FRAGMENTS\_MOIETIES>

<AMOUNT\_TYPE/>

<MOLECULAR\_FRAGMENT\_MOIETY>

<STRUCTURAL\_MODIFICATION\_TYPE/>

<ROLE>**SUBSTITUTION**</ROLE>

<FRAGMENT\_NUMBER>**1**</FRAGMENT\_NUMBER>

<MOLECULAR\_FRAGMENT\_NAME>**4-chloro-D-phenylalanine**</MOLECULAR\_FRAGMENT\_NAME>

<MOLECULAR\_FRAGMENT\_ID/>

<MOLECULAR\_FRAGMENT\_INCHI>**InChI=1S/C9H10ClNO2/c10-7-3-1-6(2-4-7)5-8(11)9(12)13/h1-4,8H,5,11H2,(H,12,13)/t8-/m1/s1**</MOLECULAR\_FRAGMENT\_INCHI>

1. ***Data representation in SPL***
   1. We only have a preliminary template for SPL XML for proteins. All the codes are fake.
   2. Each subunit is represented by a moiety

<moiety>

<code code="C8882-1" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="protein sub-unit"/>

<quantity>

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

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

</quantity>

<partMoiety>

<id extension="SU1" root="b2315501-dc0c-56d0-e044-001185133a64"/>

<subjectOf>

<characteristic>

<code code="X88850-0" displayName="amino acid sequence, complete" codeSystem="2.16.840.1.113883.6.1">

<originalText>**COMPLETE**</originalText>

</code>

<value xsi:type="**ED**" mediaType="application/x-aa-seq">**QIQLQQSGPEVVKPGASVKISCKASGYTFTDYYITWVKQKPGQGLEWIGWIYPGSGNTKYNEKFKGKATLTVDTSSSTAFMQLSSLTSEDTAVYFCANYGNYWFAYWGQGTQVTVSAASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPG**</value>

</characteristic>

</subjectOf>

</moiety>

<moiety>

* 1. Each disulfide bond is represented as a separate bond

<bond>

<code code="C60010" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="disulfide bond"/>

<positionNumber value="22"/>

<positionNumber value="40"/>

<distalMoiety>

<id extension="SU1" root=""/>

</distalMoiety>

</bond>

* 1. A modified amino acids can be represented as a structure (no structure in this example)

<bond>

<code code="C60020" codeSystem="2.16.840.1.113883.3.26.1.1" displayName="amino acid substitution"/>

<positionNumber value="1"/>

<distalMoiety>

<code code="SZB83O1W42" codeSystem="2.16.840.1.113883.4.9"/>

<name>**pyroglutamic acid**</name>

</distalMoiety>

<subjectOf>

<characteristic>

<code code="9999-9" codeSystem="2.16.840.1.113883.6.1" displayName="Chemical Structure"/>

<value xsi:type="**ED**" mediaType="application/x-mdl-molfile"/>

</characteristic>

</subjectOf>

</bond>

<bond>

1. ***SRS to SPL conversion***
   1. Proteins should be converted from one XML representation to another
   2. Peptides that are represented by XML need to be converted into a chemical structure in the case the chemical structure is InChI-able.
   3. Other ideas?

Questions about representing proteins in SPL model

1. A protein may have two or more identical sub-units (moieties).
   1. For chemical substances we don’t allow identical moieties – only distinct moieties.
   2. We should allow identical moieties and not select distinct moieties for proteins, OK?
2. We will use upper case letters for L-amino acids and lower case letters for D-amino acids. OK?
3. If the sequence has a modified amino acid or any other chemical fragment,
   1. it can be designated as X on the sequence: ACDEF**X**FGDACC
   2. or we can do this designation only on the back end for the purpose of calculating the hash code:

*sequence:* ACDEF**K**FGDACC *Sequence used for computing hash:* ACDEF**X**FGDACC

* 1. or we can replace modified amino acids with their InChIs when computing hash. Then we don’t need to X anything:

*sequence:* ACDEF**K**FGDACC *Sequence used for computing hash:* ACDEF**InChI=1S/C7H16N2O2/c1-9-5-3-2-4-6(8)7(10)11/h6,9H,2-5,8H2,1H3,(H,10,11)/t6-/m0/s1**FGDACC

1. Currently we don’t have a way to indicate how the modified fragment connects to the sequence. We assume that most modifications preserve α-amino acid backbone. So, if the modified structure is like this:



, we assume the connection points are like this:



But, there are β-amino acids that have a different backbone:



1. Should we specify the type of modification as “Alpha Amino Acid Substitution”, “Beta Amino Acid Substitution”…?
2. Or should we come up with a “fragment InChI” that indicates connection points?
3. We will represent N-terminal and C-terminal modifications the same way we represent other modifications, OK?
4. We have to include bonds such as Disulfide Bonds or other intra- and inter-sequence linkers into the hash. For disulfide bonds it can be done by indicating bond positions, each concatenation of the same level text should be done in lexicographical order. Example:

Subunit A: EDFCRDFCEDFCEDFCEDFC

Subunit B: ACDACDACDACDACDACD

Intra S-S bonds for A: 4-12; 16-20;

Intra S-S bonds for B: 2-5; 11-14;

Inter S-S bonds A-B: 8-8

The text to be hashed may look like this:

ACDACDACDACDACDACD[2-5|11-14][ EDFCRDFCEDFCEDFCEDFC [8-8]]|

EDFCRDFCEDFCEDFCEDFC[4-12|16-20][ ACDACDACDACDACDACD [8-8]]

Other ideas?

1. Oligopeptides will be converted into a chemical structure
   1. We can either use Mol 2000 limitation (1000 atoms)
   2. or InChI limitation (1024 atoms or ???) to separate oligopeptides from polypeptides.
2. Non-specific modifications (position not known, structure not known) should be represented and included into the hash function differently. Ideas how?

Task: “Develop utility for converting non-modified proteins to SPL format”

Delivery: **By the end of the current contract period**

Subtasks:

1. Get familiar with SRS XML for proteins and SPL XML-template for proteins (Valery, Igor)
   1. Analyze XML examples
   2. Ask questions
2. Adjust SRSUTIL software to
   1. Identify SRS records that have protein data (Valery)
   2. Identify proteins that do not have modifications (Valery)
   3. Convert protein sequences and disulfide bonds to structure (Igor)
   4. Generate InChI string for the structure if possible (Valery)
   5. If InChI string can be generated, create a chemical substance SPL (Valery)
   6. If InChI can’t be generated, convert protein sequences and disulfide bonds to SPL format (Valery)
   7. Compute unique hash code based on sequences and disulfide bonds (Valery)

<SUBSTANCE>

…

<ELEMENT\_TYPE>

<PROTEIN>

<SEQUENCE\_TYPE>COMPLETE</SEQUENCE\_TYPE>

<NUMBER\_OF\_SUBUNITS/>

<SUBUNIT\_GROUP>

<SUBUNIT>1</SUBUNIT>

<LENGTH>723</LENGTH>

**<SEQUENCE>QSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQNMNNAGDKWSAFLKEQSTLAQMYPLQEIQNLTVKLQLQALQQNGSSVLSEDKSKRLNTILNTMSTIYSTGKVCNPDNPQECLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLYEEYVVLKNEMARANHYEDYGDYWRGDYEVNGVDGYDYSRGQLIEDVEHTFEEIKPLYEHLHAYVRAKLMNAYPSYISPIGCLPAHLLGDMWGRFWTNLYSLTVPFGQKPNIDVTDAMVDQAWDAQRIFKEAEKFFVSVGLPNMTQGFWENSMLTDPGNVQKAVCHPTAWDLGKGDFRILMCTKVTMDDFLTAHHEMGHIQYDMAYAAQPFLLRNGANEGFHEAVGEIMSLSAATPKHLKSIGLLSPDFQEDNETEINFLLKQALTIVGTLPFTYMLEKWRWMVFKGEIPKDQWMKKWWEMKREIVGVVEPVPHDETYCDPASLFHVSNDYSFIRYYTRTLYQFQFQEALCQAAKHEGPLHKCDISNSTEAGQKLFNMLRLGKSEPWTLALENVVGAKNMNVRPLLNYFEPLFTWLKDQNKNSFVGWSTDWSPYADQSIKVRISLKSALGDKAYEWNDNEMYLFRSSVAYAMRQYFLKVKNQMILFGEEDVRVANLKPRISFNFFVTAPKNVSDIIPRTEVEKAIRMSRSRINDAFRLNDNSLEFLGIQPTLGPPNQPPVS</SEQUENCE>**

</SUBUNIT\_GROUP>

**<DISULFIDE\_LINKAGE>1\_116-1\_124; 1\_327-344; 1\_513-1\_525</DISULFIDE\_LINKAGE>**

<GLYCOSYLATION>

<GLYCOSYLATION\_TYPE>MAMMALIAN</GLYCOSYLATION\_TYPE>

<N\_GLYCOSYLATION>1\_36; 1\_73; 1\_86; 1\_305; 1\_415; 1\_529; 1\_673</N\_GLYCOSYLATION>

<O\_GLYCOSYLATION/>

<C\_GLYCOSYLATION/>

</GLYCOSYLATION>

<MODIFICATION\_GROUP>

<PHYSICAL\_MODIFICATION\_GROUP>

<ROLE/>

<NUMBER\_OF\_PARAMETERS>0</NUMBER\_OF\_PARAMETERS>

<PARAMETER\_GROUP>

<PARAMETER/>

<AMOUNT>

<AVERAGE/>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</PARAMETER\_GROUP>

</PHYSICAL\_MODIFICATION\_GROUP>

<AGENT\_MODIFICATION\_GROUP>

<AGENT\_MODIFICATION\_TYPE/>

<ROLE/>

<MODIFICATION\_AGENT/>

<MODIFICATION\_AGENT\_ID/>

<MODIFICATION\_PROCESS/>

<AMOUNT\_TYPE/>

<AMOUNT>

<AVERAGE/>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</AGENT\_MODIFICATION\_GROUP>

<STRUCTURAL\_MODIFICATION\_GROUP>

<RESIDUE\_MODIFIED/>

<RESIDUE\_SITE/>

<NUMBER\_OF\_FRAGMENTS\_MOIETIES/>

<AMOUNT\_TYPE/>

<MOLECULAR\_FRAGMENT\_MOIETY>

<STRUCTURAL\_MODIFICATION\_TYPE/>

<ROLE/>

<FRAGMENT\_NUMBER/>

<MOLECULAR\_FRAGMENT\_NAME/>

<MOLECULAR\_FRAGMENT\_ID/>

<FRAGMENT\_CONNECTIVTY/>

<AMOUNT>

<AVERAGE>1</AVERAGE>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</MOLECULAR\_FRAGMENT\_MOIETY>

</STRUCTURAL\_MODIFICATION\_GROUP>

</MODIFICATION\_GROUP>

<PROPERTY\_GROUP>

<PROPERTY\_TYPE/>

<PROPERTY/>

<PROPERTY\_DESCRIPTION>

<SUBSTANCE\_NAME/>

<SUBSTANCE\_ID/>

<SUBSTANCE\_ID/>

<AMOUNT\_TYPE/>

<AMOUNT>

<AVERAGE/>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</PROPERTY\_DESCRIPTION>

</PROPERTY\_GROUP>

<MOLECULAR\_WEIGHT>

<MOLECULAR\_WEIGHT\_TYPE>eSTIMATE</MOLECULAR\_WEIGHT\_TYPE>

<MOLECULAR\_WEIGHT\_METHOD>SDS GEL</MOLECULAR\_WEIGHT\_METHOD>

<AMOUNT>

<AVERAGE>102000</AVERAGE>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</MOLECULAR\_WEIGHT>

<COMMENTS/>

<PUBLIC\_DOMAIN/>

</PROTEIN>

</ELEMENT\_TYPE>

<REFERENCE\_SOURCE\_GROUP>

<PUBLIC\_DOMAIN/>

<REFERENCE\_SOURCE\_TYPE/>

<REFERENCE\_SOURCE\_IDENTIFIER/>

<REFERENCE\_SOURCE\_CITATION/>

</REFERENCE\_SOURCE\_GROUP>

</SINGLE\_SUBSTANCE>

</SUBSTANCE>

**Discuss:**

1. Interface from XML to Igor’s converter
2. Tags that have to be evaluated
3. Timeline
4. String for hashing

Task: “Converting modified proteins to SPL format”

Delivery: **By the end of the current contract period ??-- Most likely not**

Subtasks:

1. Adjust SRSUTIL software to
   1. Identify SRS proteins that have modifications (Valery)
   2. Convert protein sequences and modifications to structure (Igor)
   3. Generate InChI string for the structure if possible (Valery)
   4. If InChI can’t be generated, convert protein sequences and modifications to SPL format (Valery)
   5. Compute unique hash code based on sequences and disulfide bonds (Valery)

<SUBSTANCE>

<ELEMENT\_TYPE>

<PROTEIN>

<SEQUENCE\_TYPE>COMPLETE</SEQUENCE\_TYPE>

<NUMBER\_OF\_SUBUNITS/>

<SUBUNIT\_GROUP>

<SUBUNIT>1</SUBUNIT>

<LENGTH>502</LENGTH>

<SEQUENCE>MDLIPNLAVETWLLLAVSLVLLYLYGTRTHGLFKRLGIPGPTPLPLLGNVLSYRQGLWKFDTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISLAEDEEWKRIRSLLSPTFTSGKLKEMFPIIAQYGDVLVRNLRREAEKGKPVTLKDIFGAYSMDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSLFPKDTINFLSKSVNRMKKSRLNDKQKHRLDFLQLMIDSQNSKETESHKALSDLELAAQSIIFIFAGYETTSSVLSFTLYELATHPDVQQKLQKEIDAVLPNKAPPTYDAVVQMEYLDMVVNETLRLFPVAIRLERTCKKDVEINGVFIPKGSMVVIPTYALHHDPKYWTEPEEFRPERFSKKKDSIDPYIYTPFGTGPRNCIGMRFALMNMKLALIRVLQNFSFKPCKETQIPLKLDTQGLLQPEKPIVLKVDSRDGTLSGE</SEQUENCE>

</SUBUNIT\_GROUP>

<DISULFIDE\_LINKAGE/>

<OTHER\_LINKAGE>

<SITE/>

<LINKAGE\_TYPE/>

</OTHER\_LINKAGE>

<GLYCOSYLATION>

<GLYCOSYLATION\_TYPE/>

<N\_GLYCOSYLATION/>

<O\_GLYCOSYLATION/>

<C\_GLYCOSYLATION/>

</GLYCOSYLATION>

<MODIFICATION\_GROUP>

<PHYSICAL\_MODIFICATION\_GROUP>

<ROLE/>

<NUMBER\_OF\_PARAMETERS/>

<PARAMETER\_GROUP>

<PARAMETER/>

<AMOUNT>

<AVERAGE/>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</PARAMETER\_GROUP>

</PHYSICAL\_MODIFICATION\_GROUP>

<AGENT\_MODIFICATION\_GROUP>

<AGENT\_MODIFICATION\_TYPE/>

<ROLE/>

<MODIFICATION\_AGENT/>

<MODIFICATION\_AGENT\_ID/>

<MODIFICATION\_PROCESS/>

<AMOUNT\_TYPE/>

<AMOUNT>

<AVERAGE/>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</AGENT\_MODIFICATION\_GROUP>

<STRUCTURAL\_MODIFICATION\_GROUP>

<RESIDUE\_MODIFIED>CYS</RESIDUE\_MODIFIED>

<RESIDUE\_SITE>1\_441</RESIDUE\_SITE>

<NUMBER\_OF\_FRAGMENTS\_MOIETIES/>

<AMOUNT\_TYPE/>

<MOLECULAR\_FRAGMENT\_MOIETY>

<STRUCTURAL\_MODIFICATION\_TYPE>Heme binding</STRUCTURAL\_MODIFICATION\_TYPE>

<ROLE/>

<FRAGMENT\_NUMBER/>

<MOLECULAR\_FRAGMENT\_NAME/>

<MOLECULAR\_FRAGMENT\_ID>G0HAF1JK3T</MOLECULAR\_FRAGMENT\_ID>

**<MOLECULAR\_FRAGMENT\_INCHI>InChI=1S/C34H34N4O4.C3H7NO2S.Fe.H2O/c1-7-21-17(3)25-13-26-19(5)23(9-11-33(39)40)31(37-26)16-32-24(10-12-34(41)42)20(6)28(38-32)15-30-22(8-2)18(4)27(36-30)14-29(21)35-25;4-2(1-7)3(5)6;;/h7-8,13-16H,1-2,9-12H2,3-6H3,(H4,35,36,37,38,39,40,41,42);2,7H,1,4H2,(H,5,6);;1H2/q;;+3;/p-3/b25-13-,26-13-,27-14-,28-15-,29-14-,30-15-,31-16-,32-16-;;;/t;2-;;/m.0../s1</MOLECULAR\_FRAGMENT\_INCHI>**

<FRAGMENT\_CONNECTIVTY/>

<AMOUNT>

<AVERAGE/>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</MOLECULAR\_FRAGMENT\_MOIETY>

</STRUCTURAL\_MODIFICATION\_GROUP>

</MODIFICATION\_GROUP>

<PROPERTY\_GROUP>

<PROPERTY\_TYPE/>

<PROPERTY/>

<PROPERTY\_DESCRIPTION>

<SUBSTANCE\_NAME/>

<SUBSTANCE\_ID/>

<SUBSTANCE\_ID/>

<AMOUNT\_TYPE/>

<AMOUNT>

<AVERAGE/>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</PROPERTY\_DESCRIPTION>

</PROPERTY\_GROUP>

<MOLECULAR\_WEIGHT>

<MOLECULAR\_WEIGHT\_TYPE/>

<MOLECULAR\_WEIGHT\_METHOD/>

<AMOUNT>

<AVERAGE/>

<LOW\_LIMIT/>

<HIGH\_LIMIT/>

<UNIT/>

<NON\_NUMERIC\_VALUE/>

</AMOUNT>

</MOLECULAR\_WEIGHT>

<COMMENTS/>

<PUBLIC\_DOMAIN/>

</PROTEIN>

</ELEMENT\_TYPE>

<REFERENCE\_SOURCE\_GROUP>

<PUBLIC\_DOMAIN/>

<REFERENCE\_SOURCE\_TYPE/>

<REFERENCE\_SOURCE\_IDENTIFIER/>

<REFERENCE\_SOURCE\_CITATION/>

</REFERENCE\_SOURCE\_GROUP>

</SINGLE\_SUBSTANCE>

</SUBSTANCE>



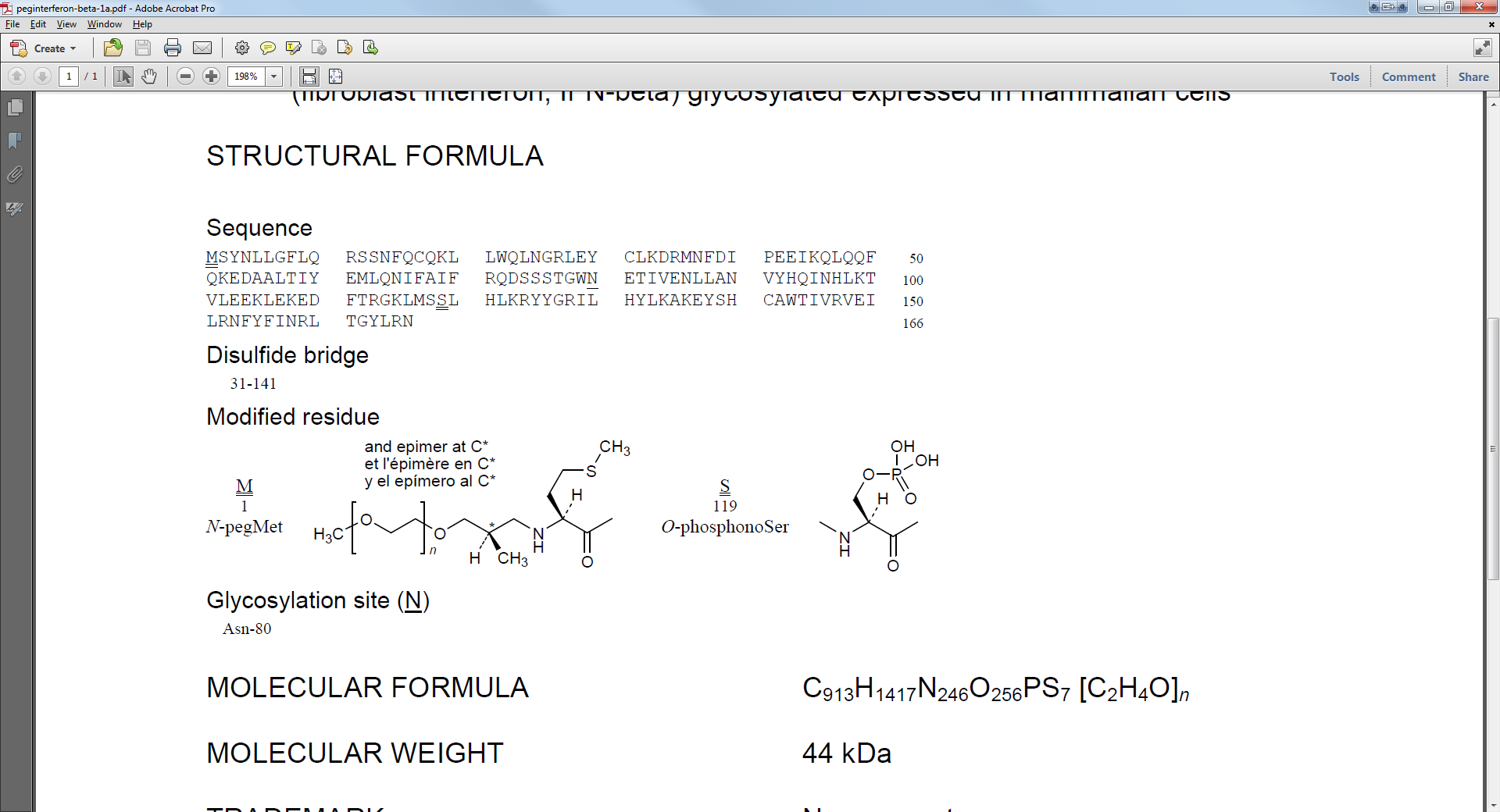
**Discuss:**

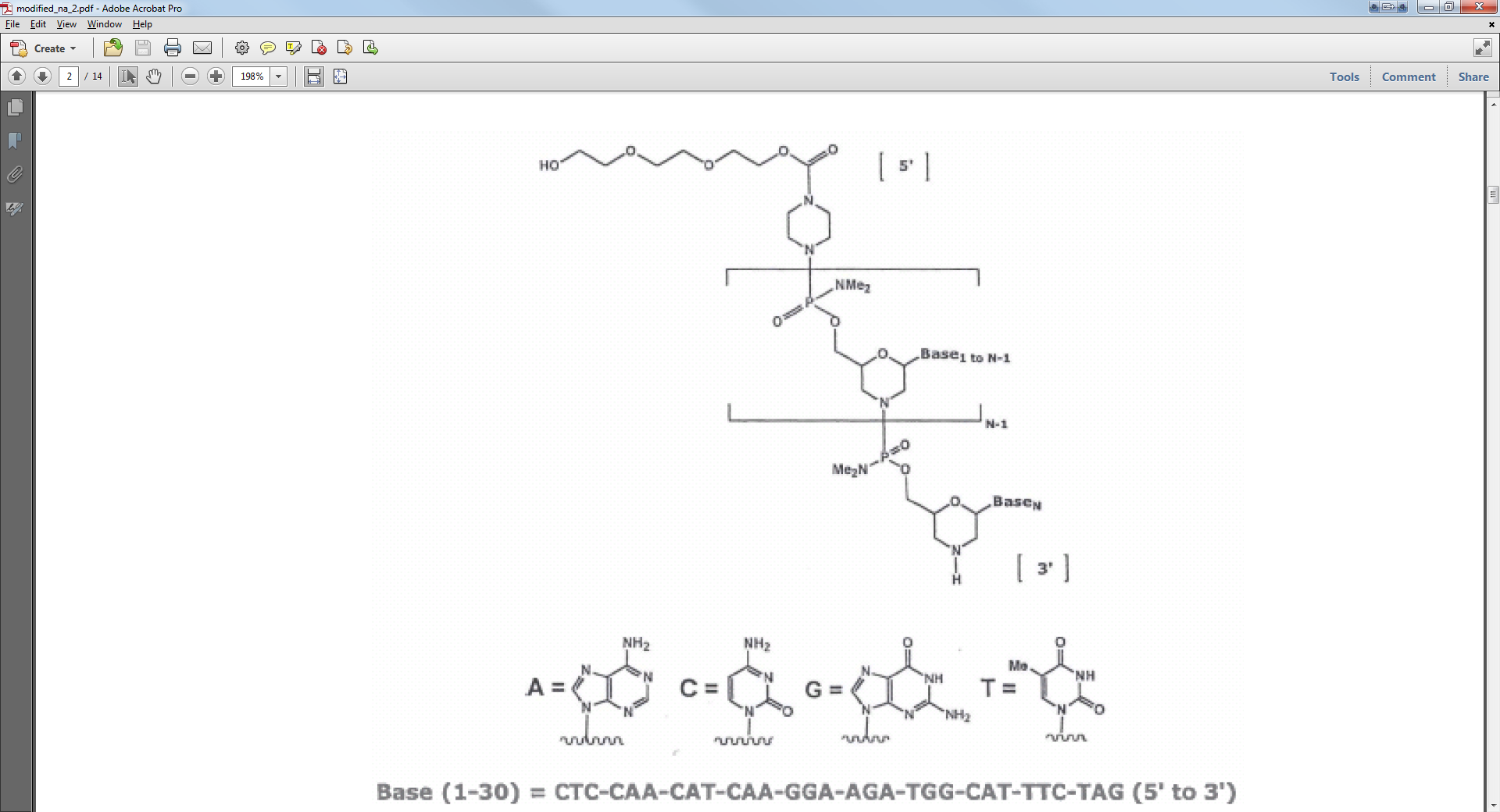
1. We may approach modified proteins independently, but we will also have to deal with nucleic acids and polymers…
2. Interface between XML and Igor’s converter for indicating connection points
3. Fragment InChIs for indicating connection points
4. Connection points for polymers and nucleic acids

**Connection points in general**

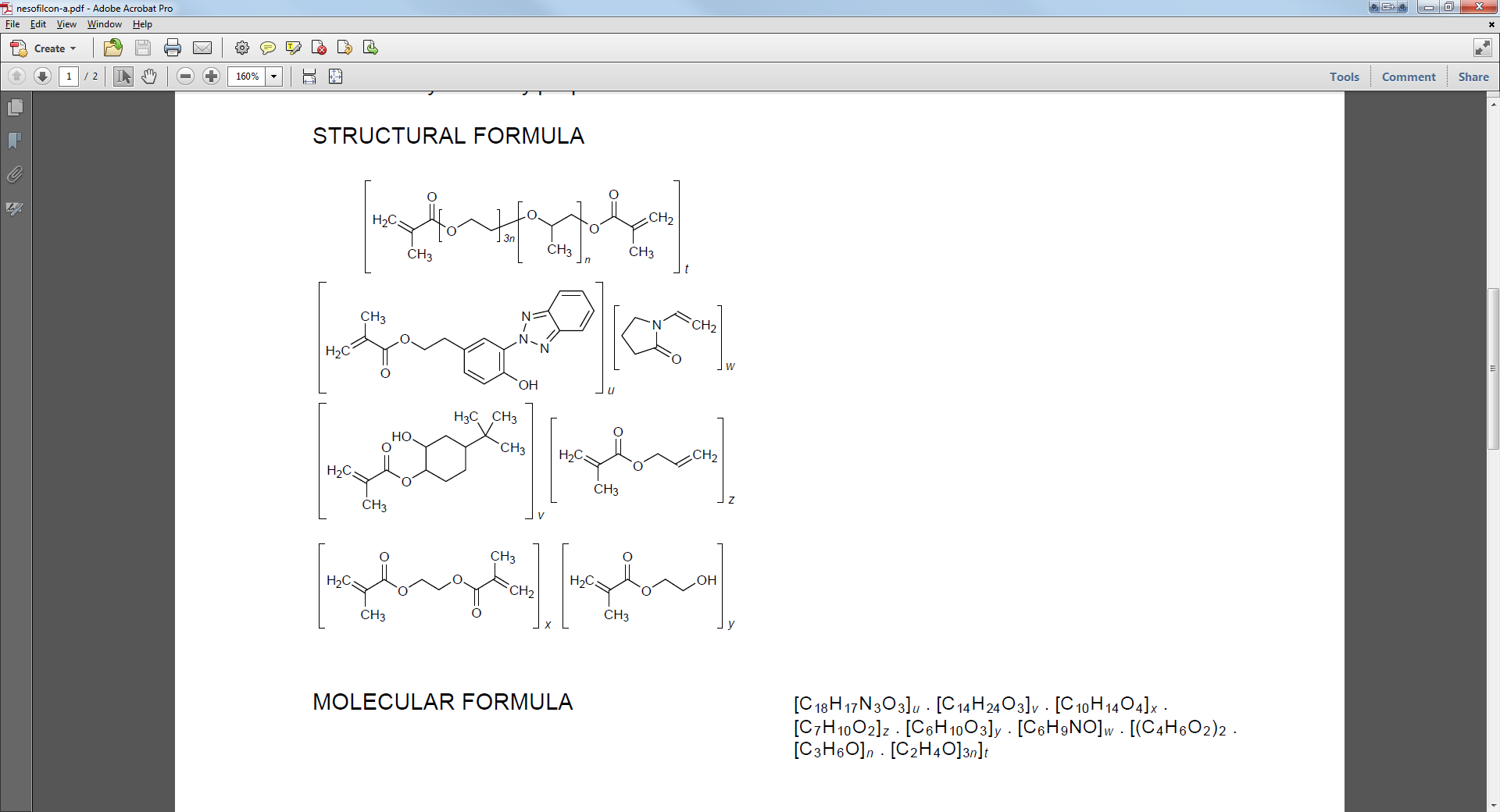
1. We may need connection points for representing modified proteins, nucleic acids and polymers
2. Sometimes it’s difficult to determine where one fragment ends and another begins
   1. For proteins it should always be a peptide bond
   2. For nucleic acids, there are three distinct fragments: base, linker and sugar. Each can be modified
   3. For polymers, there may be one or more monomeric residues. They may connect head-tail or head-head or have multiple connections (network polymers).
3. In order to unambiguously represent fragments, their structure and the numbering of connection points should always be the same – canonical
4. We did experiments in which the numeration of connection points was determined by the structure of the fragment. Fragment InChIs were therefore canonical

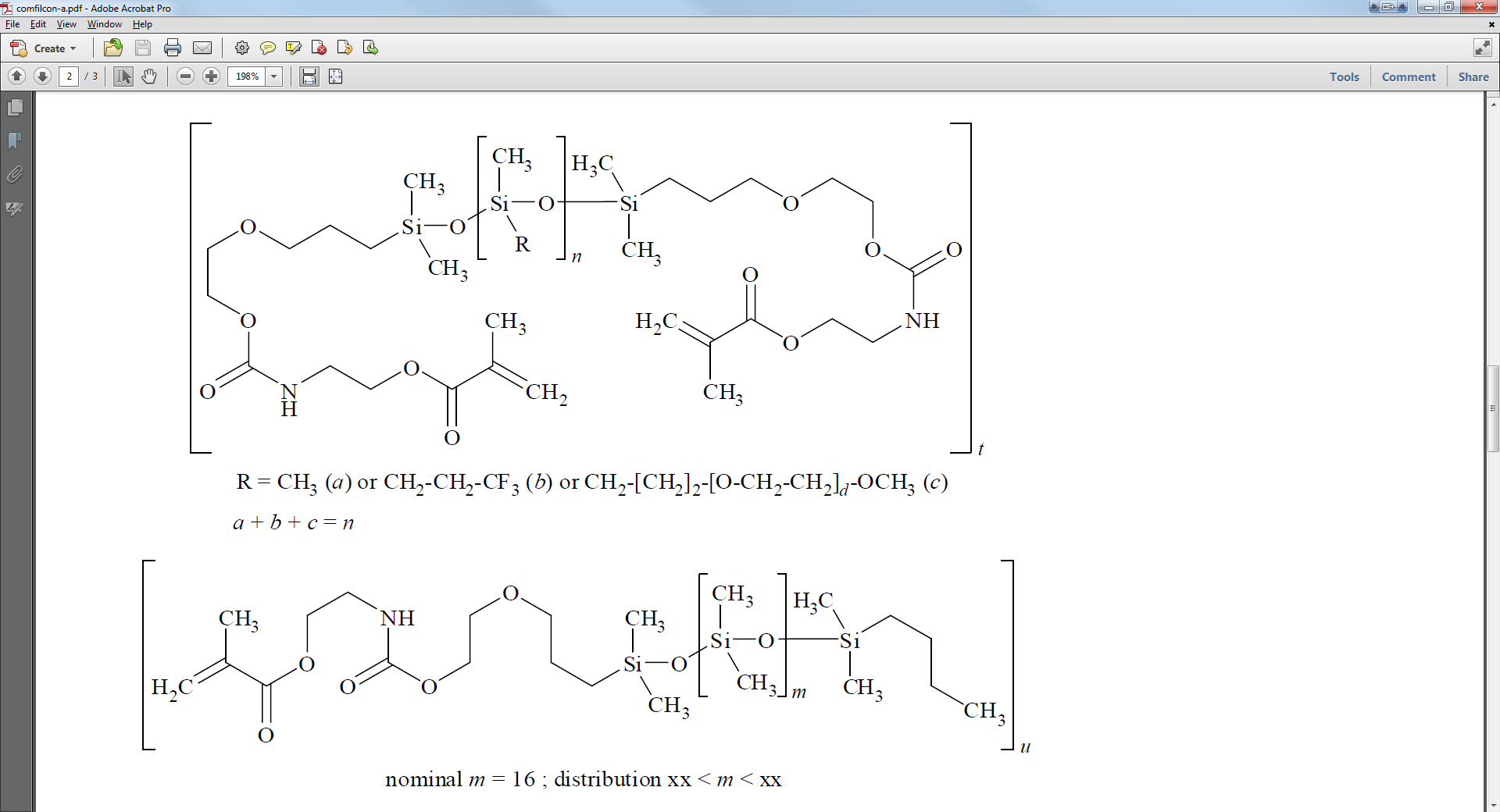
Modified protein:

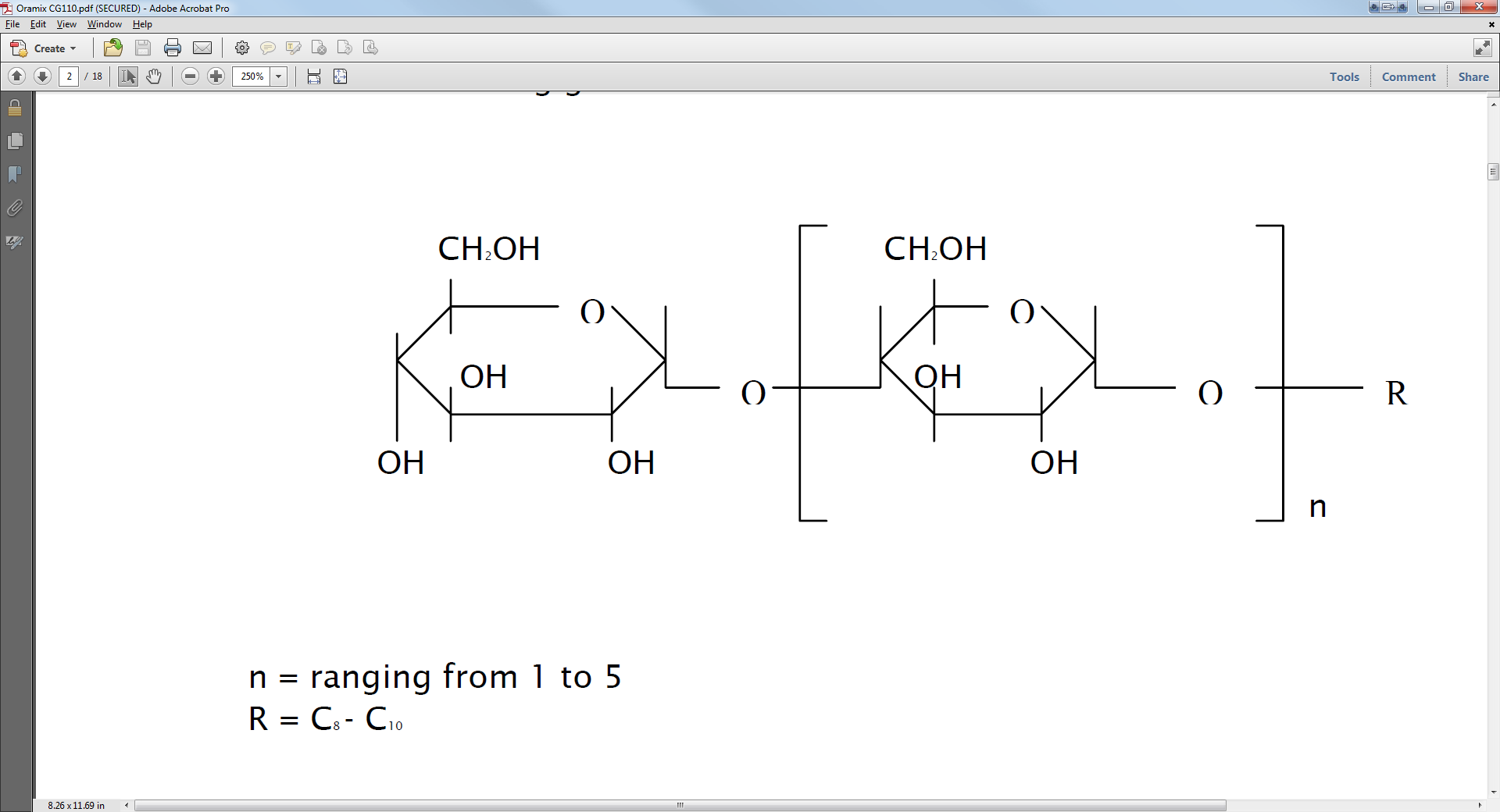


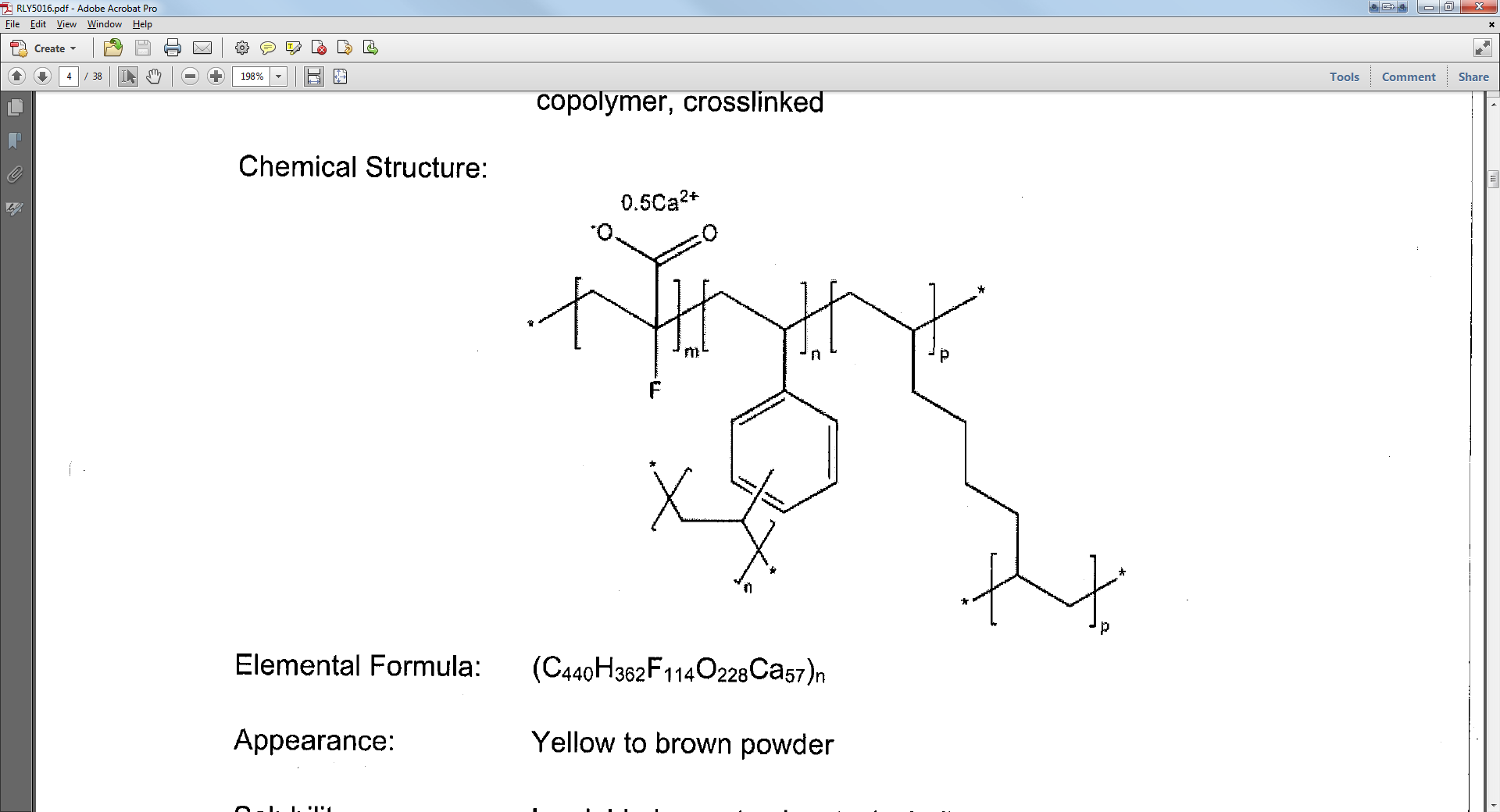
Modified nucleic acid:

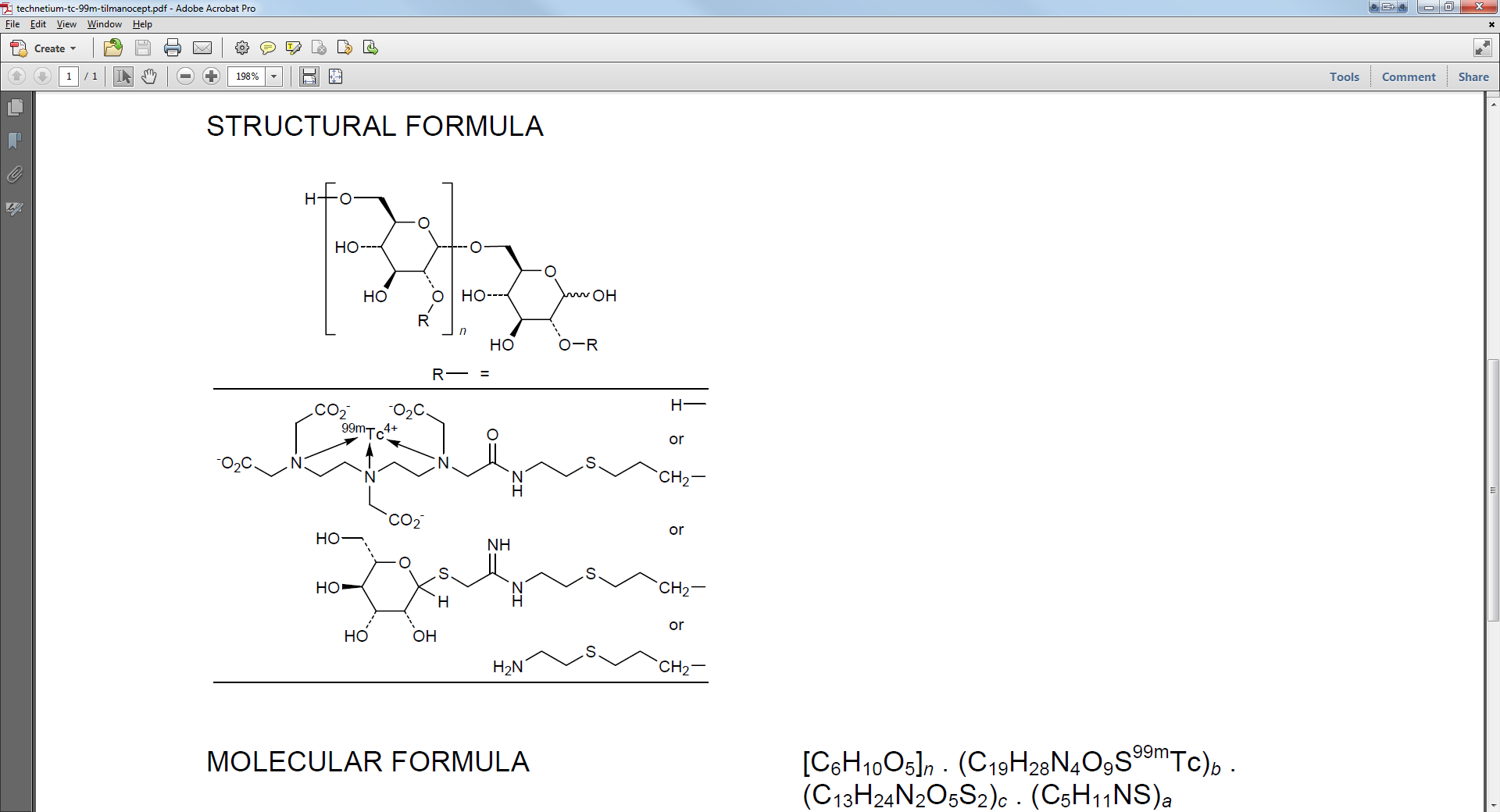
Polymers:

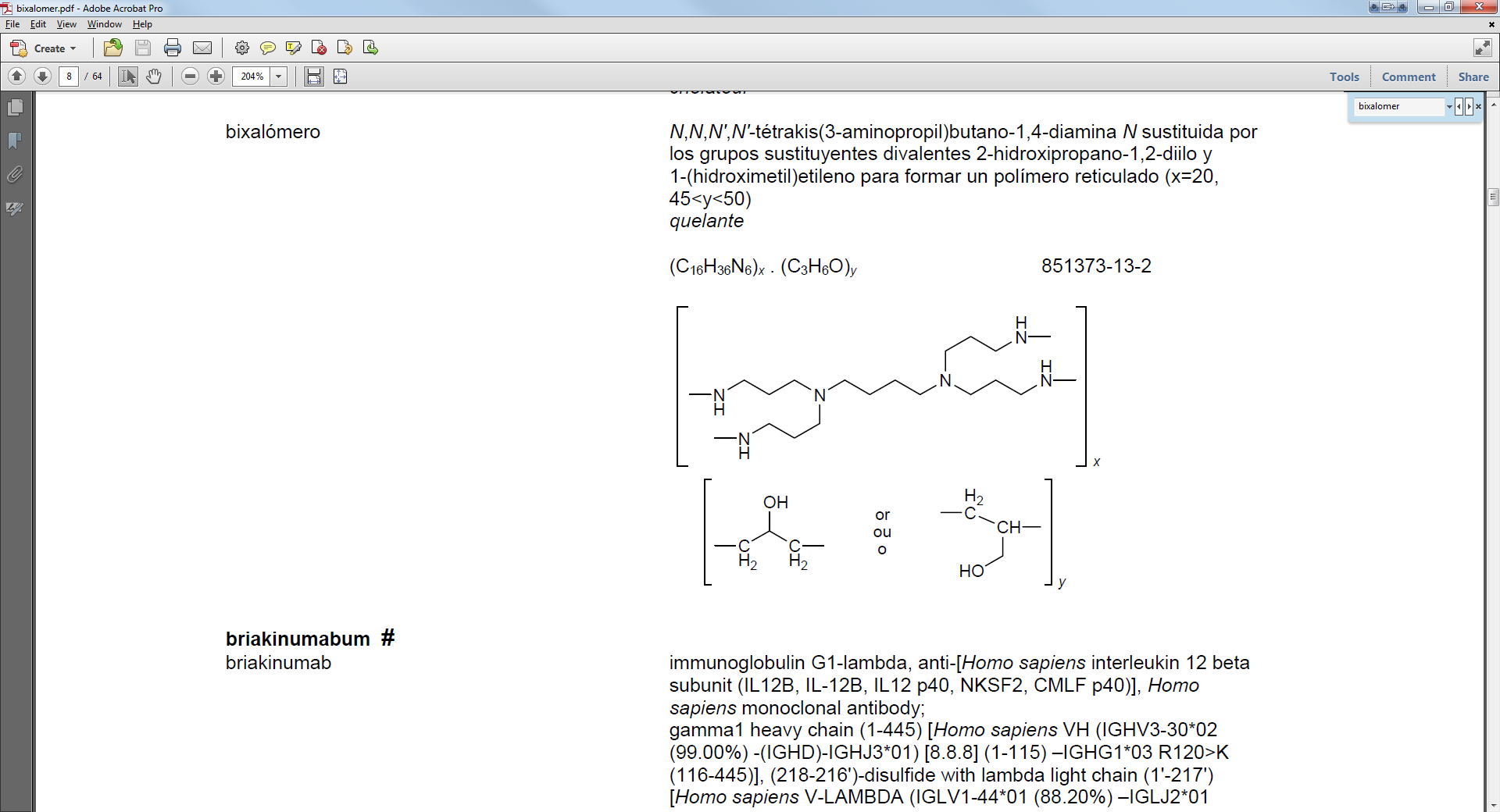












**enfortumabum vedotinum #**

enfortumab vedotin immunoglobulin G1-kappa, anti-[*Homo sapiens* PVRL4 (poliovirus

receptor-related 4, nectin-4, nectin 4, PPR4, LNIR], *Homo sapiens*

monoclonal antibody conjugated to auristatin E;

gamma1 heavy chain (1-447) [*Homo sapiens* VH (IGHV3-48\*02

(98.00%) -(IGHD)-IGHJ6\*01) [8.8.10] (1-117) -IGHG1\*03 (CH1 (118-

215), hinge (216-230), CH2 (231-340), CH3 (341-445), CHS (446-

447)) (118-447)], (220-214')-disulfide with kappa light chain (1'-214')

[*Homo sapiens* V-KAPPA (IGKV1-12\*01 (96.80%) -IGKJ4\*01) [6.3.9]

(1'-107') -IGKC\*01 (108'-214')]; dimer (226-226'':229-229'')-

bisdisulfide; conjugated, on an average of 3 to 4 cysteinyl, to

monomethylauristatin E (MMAE), via a cleavable maleimidecaproylvalyl-

citrullinyl-*p-*aminobenzylcarbamate (mc-val-cit-PABC) linker

For the *vedotin* part, please refer to the document *"INN for*

*pharmaceutical substances: Names for radicals, groups and*

*others"\*.*

*immunomodulator, antineoplastic*

enfortumab védotine immunoglobuline G1-kappa, anti-[*Homo sapiens* PVRL4 (membre 4

de la famille du récepteur du poliovirus, nectine-4, nectine 4, PPR4,

LNIR], *Homo sapiens* anticorps monoclonal conjugué à l'auristatine

E;

chaîne lourde gamma1 (1-447) [*Homo sapiens* VH (IGHV3-48\*02

(98.00%) -(IGHD)-IGHJ6\*01) [8.8.10] (1-117) -IGHG1\*03 (CH1 (118-

215), charnière (216-230), CH2 (231-340), CH3 (341-445), CHS

(446-447)) (118-447)], (220-214')-disulfure avec la chaîne légère

kappa (1'-214') [*Homo sapiens* V- KAPPA (IGKV1-12\*01 (96.80%) -

IGKJ4\*01) [6.3.9] (1'-107') -IGKC\*01 (108'-214')]; dimère (226-

226'':229-229'')-bisdisulfure; conjugué, sur 3 à 4 cystéinyl en

moyenne, au monométhylauristatine E (MMAE), via un linker clivable

maléimidecaproyl-valyl-citrullinyl-*p-*aminobenzylcarbamate (mc-valcit-

PABC)

Pour la partie *védotine*, veuillez-vous référer au document *"INN for*

*pharmaceutical substances: Names for radicals, groups and*

*others"\*.*

*immunomodulateur, antinéoplasique*

enfortumab vedotina inmunoglobulina G1-kappa, anti-[PVRL4 de *Homo sapiens* (miembro

4 de la familia del receptor de poliovirus, nectina-4, nectina 4, PPR4,

LNIR], anticuerpo monoclonal de *Homo sapiens* conjugado con

auristatina E;

cadena pesada gamma1 (1-447) [*Homo sapiens* VH (IGHV3-48\*02

(98.00%) -(IGHD)-IGHJ6\*01) [8.8.10] (1-117) -IGHG1\*03 (CH1 (118-

215), bisagra(216-230), CH2 (231-340), CH3 (341-445), CHS (446-

447)) (118-447)], (220-214')-disulfuro con la cadena ligera kappa (1'-

214') [*Homo sapiens* V-KAPPA (IGKV1-12\*01 (96.80%) -IGKJ4\*01)

[6.3.9] (1'-107') -IGKC\*01 (108'-214')]; dímero (226-226'':229-229'')-

bisdisulfuro; conjugado, en 3- 4 restos cisteinil por término medio,

con monometilauristatinea E (MMAE), mediante un conector

escindible maleimidocaproil-valil-citrulinil-*p-*aminobencilcarbamato

(mc-val-cit-PABC)

La información sobre la *vedotina*, la encontrarán en el documento

*"INN for pharmaceutical substances: Names for radicals, groups and*

*others"\*.*

*inmunomodulador, antineoplásico*

WHO Drug Information, Vol. 27, No. 2, 2013 Proposed INN: List **109**

157

1346452-25-2

Heavy chain / Chaîne lourde / Cadena pesada

EVQLVESGGG LVQPGGSLRL SCAASGFTFS SYNMNWVRQA PGKGLEWVSY 50

ISSSSSTIYY ADSVKGRFTI SRDNAKNSLS LQMNSLRDED TAVYYCARAY 100

YYGMDVWGQG TTVTVSSAST KGPSVFPLAP SSKSTSGGTA ALGCLVKDYF 150

PEPVTVSWNS GALTSGVHTF PAVLQSSGLY SLSSVVTVPS SSLGTQTYIC 200

NVNHKPSNTK VDKRVEPKSC DKTHTCPPCP APELLGGPSV FLFPPKPKDT 250

LMISRTPEVT CVVVDVSHED PEVKFNWYVD GVEVHNAKTK PREEQYNSTY 300

RVVSVLTVLH QDWLNGKEYK CKVSNKALPA PIEKTISKAK GQPREPQVYT 350

LPPSREEMTK NQVSLTCLVK GFYPSDIAVE WESNGQPENN YKTTPPVLDS 400

DGSFFLYSKL TVDKSRWQQG NVFSCSVMHE ALHNHYTQKS LSLSPGK 447

Light chain / Chaîne légère / Cadena ligera

DIQMTQSPSS VSASVGDRVT ITCRASQGIS GWLAWYQQKP GKAPKFLIYA 50

ASTLQSGVPS RFSGSGSGTD FTLTISSLQP EDFATYYCQQ ANSFPPTFGG 100

GTKVEIKRTV AAPSVFIFPP SDEQLKSGTA SVVCLLNNFY PREAKVQWKV 150

DNALQSGNSQ ESVTEQDSKD STYSLSSTLT LSKADYEKHK VYACEVTHQG 200

LSSPVTKSFN RGEC 214

Disulfide bridges location / Position des ponts disulfure / Posiciones de los puentes disulfuro

Intra-H (C23-C104) 22-96 144-200 261-321 367-425

22''-96'' 144''-200'' 261''-321'' 367''-425''

Intra-L (C23-C104) 23'-88' 134'-194'

23'''-88''' 134'''-194'''

Inter-H-L (h 5-CL 126) \* 220-214' 220''-214'''

Inter-H-H (h 11, h 14) \* 226-226'' 229-229''

\*Two or three of the inter-chain disulfide bridges are not present, an average of 3 to 4

cysteinyl being conjugated each to a drug linker.

\*Deux ou trois des ponts disulfures inter-chaînes ne sont pas présents, 3 à 4 cystéinyl en

moyenne étant chacun conjugué à un linker-principe actif.

\*Faltan dos o tres puentes disulfuro inter-catenarios, una media de 3 a 4 cisteinil está

conjugada a conectores de principio activo.

N-glycosylation sites / Sites de N-glycosylation / Posiciones de N-glicosilación

H CH2 N84.4:

297, 297''

For the vedotin part, please refer to the document "INN for pharmaceutical substances:

Names for radicals, groups and others"\*

Pour la partie védotine, veuillez vous référer au document "INN for pharmaceutical

substances: Names for radicals, groups and others"\*.

Para la fracción vedotina, se pueden dirigir al documento "INN for pharmaceutical

substances: Names for radicals, groups and others"\*





