ONE MONOMER, TWO NAMES

THE CASE FOR A CORE SET OF MONOMERS

Noel O'Boyle and Roger Sayle

NextMove Software

HOW MANY AMINO ACIDS PRESENT?

20 common amino acids

Ala, Cys, Lys, Thr

⇩

87 amino acids

Ala, Cys, Hcy, Lys, 2Nal, Ncy, Thr

⇩

1095 including substituents

Thr, Thr(*t*Bu), Thr(Bn), Thr(PO_3H_2)

⇩

3546 including stereo variants, terminal variants, linker variants, α-methylated

Thr, D-Thr, DL-Thr, aThr, Thr-ol, aMeThr

⇩

8125 including N-substituted variants

Thr, Me-Thr, Boc-Thr, Me_2-Thr, Fmoc-N(Me)Thr

HOW MANY MONOSACCHARIDES PRESENT?

113 aldoses, ketoses, aldonic and uronic acids with from 5-9 carbons

AltA, Glc, L-Man, L-Gal, Fru, L-gro-D-glcHept

⇩ **407 including deoxy variants, ring variants**

L-Glcf, Mans, 2-deoxy-D-manHept, 3-deoxy-D-glcOct2ulo-onic

⇩ **971 including anomeric stereo**

a-Man, 3,4-deoxy-a-D-eryHex, b-Tyv

⇩ **7094 including common substituents at non-anomeric positions**

Xylf5Me, a-L-ManNAc3Ac4Ac6Ac, Glc2P3P6P

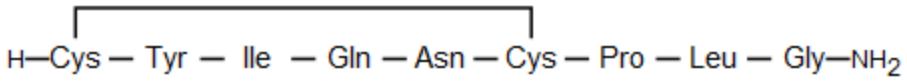
⇩ **26641 including any substituent anywhere**

Bz(-2)[Tos(-3)]Ara4Ac(b)-O-Me, TMS(-4)[TMS(-6)]GlcNAc3Me(a)-O-Me

* This analysis excludes monosaccharides with missing stereochemistry

C-TERMINAL AMIDE

- How should C-terminal amide be represented?
- Consider oxytocin:



- Expected (e.g. from Sugar&Splice or ChEMBL):

```
PEPTIDE1{C.Y.I.Q.N.C.P.L.G.[am]}$PEPTIDE1,PEPTIDE1,1:R
3-6:R3$$$
```

- BioEddie (see [1]) has:

```
PEPTIDE1{C.Y.I.Q.N.C.P.L.G}|CHEM1{[NH2]}$PEPTIDE1,CHEM
1,9:R2-1:R1|PEPTIDE1,PEPTIDE1,6:R3-1:R3$$$V2.0
```

- ChemDraw doesn't recognise the NH2 label

[1] https://chemaxon.com/app/uploads/2018/08/Bridging-the-gap-between-small-molecules-and-biologics-editing.pdf

NUCLEIC ACIDS

- What is the correct representation of deoxyribonucleic acids?

RNA1{[MOE](G).[sP]**[MOE]**([5meC]).[sP][MOE]([5meC]).[sP][MOE](T).[sP][MOE]([5meC]).**[sP][dR](A)**.[sP][dR](G).[sP][dR](T).[sP][dR]([5meC]).[sP][dR](T).[sP][dR](G).[sP][dR]([5meC]).[sP][dR](T).[sP][dR](T).[sP][dR]([5meC]).[sP][MOE](G).[sP][MOE]([5meC]).[sP][MOE](A).[sP][MOE]([5meC]).[sP][MOE]([5meC])}$$$$

- From slide 5 of "HELM update for CPCDS July 2018"

RNA1{[moe](G)[sp].**[moe]**([m5C])[sp].[moe]([m5C])[sp].[moe](T)[sp].[moe]([m5C])[sp].**d(A)[sp]**.d(G)[sp].d(T)[sp].d([m5C])[sp].d(T)[sp].d(G)[sp].d([m5C])[sp].d(T)[sp].d(T)[sp].d([m5C])[sp].[moe](G)[sp].[moe]([m5C])[sp].[moe](A)[sp].[moe]([m5C])[sp].[moe]([m5C])}$$$$V2.0

- ChemDraw 17.1 reads the new version of mipomersen but not the old

ChEMBL HELM: What is a monomer

- In Pistoia's HELM monomer library, Boc is an N-terminal modification, like ac.
- In ChEMBL's HELM monomer library, "Boc_A" is an amino acid.
- This leads to different monomer counts, depending on which monomer library is used

HELM MONOMER DATABASES

- The xHELM database format is an XML file
  - as used by HELM1 and ChEMBL for example
    - chembl_22_1_monomer_library.xml
- But…
  - BIOVIA uses an SDF file
  - HELM web editor uses a JSON file
  - ChemDraw uses a different JSON file

HELM MONOMER DATABASES

ChemDraw 17.1


```
{
        "id": "190",
        "name": "C-Terminal amine",
        "naturalAnalog": "X",
        "polymerType": "PEPTIDE",
        "monomerType": "Backbone",
        "version": "1.0",
        "createdDate": "2017-10-27T14:43:10.6942017-04:00",
        "cdxml": "<?xml version=\"1.0\" encoding=\"UTF-8\" ?><!DOCTYPE CDXML SYSTEM
\"http:\/\/www.cambridgesoft.com\/xml\/cdxml.dtd\" ><CDXML BondLength=\"30\"><colortable><color r=\"1\"
g=\"1\" b=\"1\"\/><color r=\"0\" g=\"0\" b=\"0\"\/><\/colortable><page><fragment id=\"1\"><n id=\"2\"
p=\"0 11.81\" Element=\"7\"\/><n id=\"3\" p=\"20.47 0\" Z=\"1\" NodeType=\"ExternalConnectionPoint\"
ExternalConnectionType=\"Diamond\" ExternalConnectionNum=\"1\"\/><b id=\"5\" B=\"2\"
E=\"3\"\/><\/fragment><\/page><\/CDXML>",
        "monomerAsTargetSubstructureScreen": "35,40,90,237",
        "monomerAsTargetExactStructureScreen": "35,40,90,104,162,237,249",
        "formula": "20FH2N",
        "symbol": "am",
        "topologyName": "Undetermined",
        "stereoName": "Undetermined",
        "attachmentList": [
           {
              "label": "R1",
              "capGroupName": "H"
           }
        ]
    },
```


JSON file, no SMILES or MOL file

Has “am”, but no entry for “ac”