PEPTIDE LINE NOTATIONS FOR BIOLOGICS REGISTRATION AND PATENT FILINGS

Roger Sayle, Daniel Lowe and Noel O'Boyle
NextMove Software Ltd, Cambridge, UK.

1. Introduction

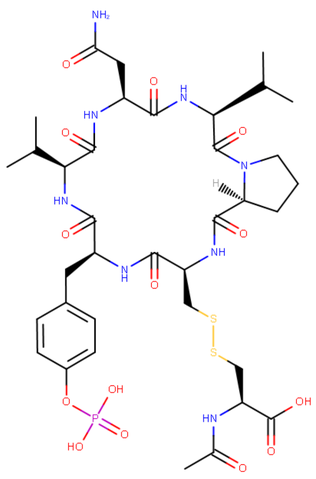
The increasing importance of biotherapeutics ("biologics") to the pharmaceutical industry presents challenges to traditional chemical registration systems, blurring the distinctions between small molecules and biopolymers. The formats and depictions used to describe small molecules may be inappropriate for biopolymers such as polypeptides, DNA and RNA sequences, carbohydrates, glycoproteins, antibodies and antibody drug conjugates. In this poster, we consider the specific task of assigning unique identifiers to peptides and peptide-like drug structures.

2. Line Notations

For small molecules, the utility of canonical line notations such as SMILES and InChI in chemical registration is well established. Likewise for proteinogenic peptide and simple nucleic acid sequences, the one letter codes of bioinformatics and FASTA format are sufficient. Between these two extremes there is less consensus. Amongst the line notations proposed for representing peptidic structures are HELM (Hierarchical Editing Language for Macromolecules) from the Pistoia Alliance and Pfizer, PLN from Biochem Fusion and CHORTLES from Daylight.

A common snag with these proposed solutions is that these artificial notations don't match the "de facto" nomenclature used by biochemists in journal articles, vendor catalogues or patent filings. In much the same way, that "blue book" IUPAC nomenclature is universally used by chemists to describe small molecules, the three letter codes of IUPAC's 3AA standard and "IUPAC condensed" nomenclature are universal in peptide chemistry.

All-atom representations



SMILES

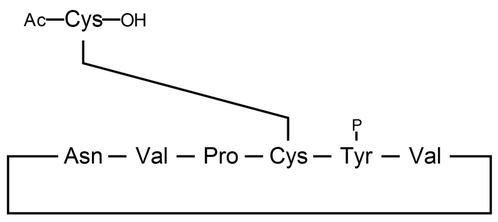
CC(C)[C@H]1C(=O)N[C@H](C(=O)N[C@H](C(=O)N2CCC[C@H]2C(=O)N[C@H](C(=O)N[C@H](C(=O)N1)CC3=CC=C(C=C3)OP(=O)(O)O)CSSC[C@@H](C(=O)O)NC(=O)C)C(C)C)CC(=O)N

Superatom representations

IUPAC Condensed [1]

cyclo[Asn-Val-Pro-Cys(1)-Tyr(P)-Val].Ac-Cys(1)-OH

IUPAC depiction [2]



HELM [3]

PEPTIDE1{[ac].C}|PEPTIDE2{N.V.P.C}|CHEM1{*N[C@@H](Cc1ccc (cc1)OP(=O)(O)O)C(=O)*|$_R1;;;;;;;;;;;;;;;;;;;;;;_R2$|}|PEPTIDE3{V} $PEPTIDE1,PEPTIDE2,2:R3-4:R3|PEPTIDE2,CHEM1,4:R2-1:R1| CHEM1,PEPTIDE3,1:R2-1:R1|PEPTIDE3,PEPTIDE2,1:R2-1:R1$$$

3. IUPAC Condensed Nomenclature and Semi-systematic Monomer Naming

Adoption of IUPAC's recommendations for peptide naming helps solve a number of technical issues, including the tricky one of (portable) non-standard amino acid names. Although the set of three-letter codes used for the 20 standard amino acids are universally accepted, the assignment of codes to non-standard amino acids differs between authors/computer systems.

Nty: nortyrosine (Chemical Abstracts) vs. N-nitrotyrosine (HELM).
Cpa: cyanopropionic amido acid vs. β-cyclopropylalanine.
Hpg: 4-hydroxyphenylglycine vs. Homopropargylglycine.

IUPAC's most obvious guidance on amino acid naming is to proscribe the use of the D-, L- (default) and DL- prefixes to specify stereochemistry. Without this, some three-letter monomer naming schemes, such as the one used by wwPDB, require separate codes for enantiomers of each amino acid.

Far less widely implemented in software are the guidelines for constructing non-standard amino acid monomer names by using short-hand line formulae. 3AA recommends the use of "Ac", "Me", "Et" and "Ph" as condensed forms of "acetyl", "methyl", "ethyl" and "phenyl" respectively. These can then be used with appropriate parenthesis and optional locants to construct systematic monomer names, such as Ser(Ac), Cys(Et) and so on.

Obviously, three letter codes can't be extended to handle all 125K unique amino acid variations in PubChem, but systematic names, such as N(Me)Ser(tBuOH), can represent a significant fraction without a chemist having to look up a name in a large reference table or index.

4. Machine-generated (Sugar & Splice) Examples

H-Cys-Pro-Trp-His-Leu-Leu-Pro-Phe-Cys-OH	CHEMBL501567
H-Tyr-Pro-Phe-Phe-OtBu	CHEMBL500195
cyclo[Ala-Tyr-Val-Orn-Leu-D-Phe-Pro-Phe-D-Phe-Asn]	CHEMBL438006
H-Nle(Et)-Tyr-Pro-Trp-Phe-NH2	CHEMBL500704
H-DL-hPhe-Val-Met-Tyr(PO3H2)-Asn-Leu-Gly-Glu-OH	CHEMBL439086
cyclo[Phe-D-Trp-Tyr(Me)-D-Pro]	CHEMBL507127
H-D-Pyr-D-Leu-pyrrolidide	CHEMBL1181307
Ac-DL-Phe-aThr-Leu-Asp-Ala-Asp-DL-Phe(4-Cl)-OH	CHEMBL1791047
H-D-Cys(1)-D-Asp-Gly-Tyr(3-NO2)-Gly-Hyp-Asp-D-Cys(1)-NH2	CHEMBL583516
Boc-Tyr-Tyr(3-Br)-OMe	CHEMBL1976073

5. Literature Examples

Boc-Asp(OtBu)-Pro-OH	CID57383532
Tyr-D-Ala-Gly-NMePhe	CID45483974
Tyr-D-Pro-Gly-Trp-NMeNle-Asp-Phe-NH2	CID11693641
Ac-Cys-Ile-Tyr-Lys-Phe(4-Cl)-Tyr	CID44412001
deamino-Cys-D-Tyr(Et)-Ile-Thr-Asn-Cys-Pro-Orn-Gly-NH2	CID68613
H-Arg(NO2)-OMe.HCl	CID135193

6. Drug Examples

LHRH (GnRH) Agonists

Pyr-His-Trp-Ser-Tyr-**D-Ser(tBu)**-Leu-Arg-Pro-**NHEt**	Buserelin
Pyr-His-Trp-Ser-Tyr-**Gly**-Leu-Arg-Pro-**Gly-NH2**	Gonadorelin
Pyr-His-Trp-Ser-Tyr-**D-Ser(tBu)**-Leu-Arg-Pro-**AzGly-NH2**	Goserelin
Pyr-His-Trp-Ser-Tyr-**D-His(1-Bn)**-Leu-Arg-Pro-**NHEt**	Histrelin
Pyr-His-Trp-Ser-Tyr-**D-Leu**-Leu-Arg-Pro-**NHEt**	Leuprolide
Pyr-His-Trp-Ser-Tyr-**2Nal**-Leu-Arg-Pro-**Gly-NH2**	Nafarelin
Pyr-His-Trp-Ser-Tyr-**D-Trp**-Leu-Arg-Pro-**Gly-NH2**	Triptorelin

Vasopressin Agonists

Deamino-Cys(1)-**Tyr**-Phe-Gln-Asn-Cys(1)-Pro-**D-Arg**-Gly-NH2	Desmopressin
Cys(1)-**Phe**-Phe-Gln-Asn-Cys(1)-Pro-**Lys**-Gly-NH2	Felypressin
Cys(1)-**Tyr**-Phe-Gln-Asn-Cys(1)-Pro-**Orn**-Gly-NH2	Ornipressin
Gly-Gly-Gly-Cys(1)-**Tyr**-Phe-Gln-Asn-Cys(1)-Pro-**Lys**-Gly-NH2	Terlipressin

Somatostatin Agonists

2Nal-Cys(1)-Tyr-D-Trp-Lys-Val-Cys(1)-Thr-NH2	Lanreotide
D-Phe-Cys(1)-Phe-D-Trp-Lys-Thr-Cys(1)-Thr-ol	Octreotide
D-Phe-Cys(1)-Tyr-D-Trp-Lys-Val-Cys(1)-Trp-NH2	Vapreotide
cyclo[Lys-Tyr(Bn)-Phe-Pro(4-OCONHEtNH2)-Phg-D-Trp]	Pasireotide

LHRH (GnRH) Antagonists

Ac-D-2Nal-D-Phe(4-Cl)-3Pal-Ser-N(Me)Tyr-D-Asn-Leu-Lys(iPr)-Pro-D-Ala-NH2	Abarelix
Ac-D-2Nal-D-Phe(4-Cl)-3Pal-Ser-D-Cit-Leu-Arg-Pro-D-Ala-NH2	Cetrorelix
Ac-D-2Nal-D-Phe(4-Cl)-3Pal-Ser-Tyr-D-hArg(N,N'-diEt)-Leu-hArg(N,N'-diEt)-Pro-D-Ala-NH2	Ganirelix
Ac-D-2Nal-D-Phe(4-Cl)-3Pal-Ser-Phe(dihydroorotamido)-D-Phe(ureido)-Leu-Lys(iPr)-Pro-D-Ala-NH2	Degarelix

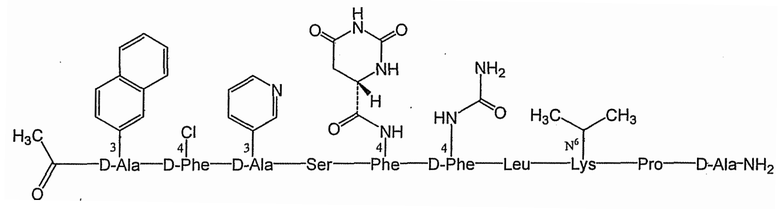


Image credit: WO2011066386A1

10. Conclusions

Many of the problems of non-standard peptide and peptidic compound naming have already been solved by peptide chemists, alas these are rarely implemented.

11. Acknowledgements

The authors would like to thank Lisa Sach-Peltason of Hoffmann-La Roche, Basel and Evan Bolton of the Pubchem project at the NCBI for their scientific guidance.

12. Bibliography

1. IUPAC-IUB Joint Commission on Biochemical Nomenclature. **Nomenclature and Symbolism for Amino Acids and Peptides.** *Pure Appl. Chem.* **1984**, *56*, 595.
2. IUPAC-IUB Joint Commission on Biochemical Nomenclature. **Nomenclature of Cyclic Peptides. 2004**. *Provisional Recommendations.*
3. Zhang T, Li H, Xi H, Stanton RV, Rotstein SH. **HELM: A Hierarchical Notation Language for complex biomolecule structure representation.** *J. Chem. Inf. Model.* **2012**, *52*, 2796.

NextMove Software
www.nextmovesoftware.co.uk
www.nextmovesoftware.com

NextMove Software Limited
Innovation Centre (Unit 23)
Cambridge Science Park
Milton Road, Cambridge
England CB4 0EY