REPRESENTATION AND DISPLAY OF NON-STANDARD PEPTIDES USING SEMI-SYSTEMATIC AMINO ACID MONOMER NAMING

ROGER SAYLE

NEXTMOVE SOFTWARE, CAMBRIDGE, UK

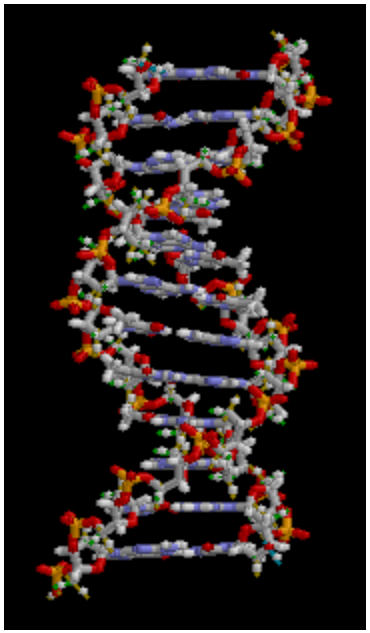
247th ACS National Meeting, Dallas, TX, Wednesday 19th March 2014

THE PRIMORDIAL SOUP

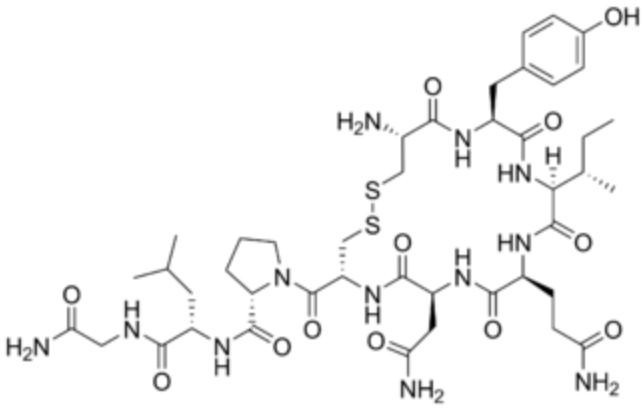
- The vast majority of chemical entities stored in chemical databases and pharmaceutical registration systems are classical “small molecules”.
- For these molecules, the universal cheminformatics “all-heavy-atom” representations work well, providing duplicate checking, substructure search, normalization, depiction, etc.
- Solutions include InChI, SMILES and V2000 mol file.

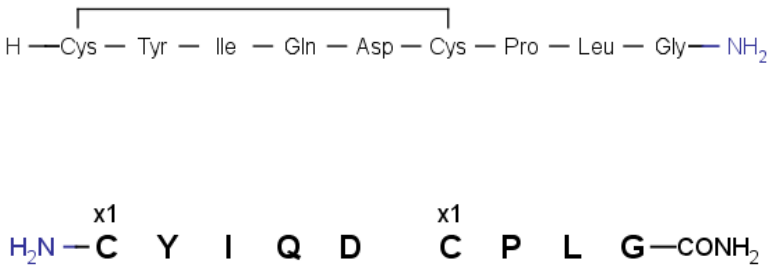
BIOPOLYMERS OF LIFE

- The low Shannon entropy of a small, but scientifically significant, fraction of compounds, **biopolymers**, allow them to represented more succinctly.
- For these frequent subunits (or **monomers**) can be used instead of atoms in connection tables.
- Representations with small monomer sets are easier for scientists to comprehend.



PEPTIDE REPRESENTATION

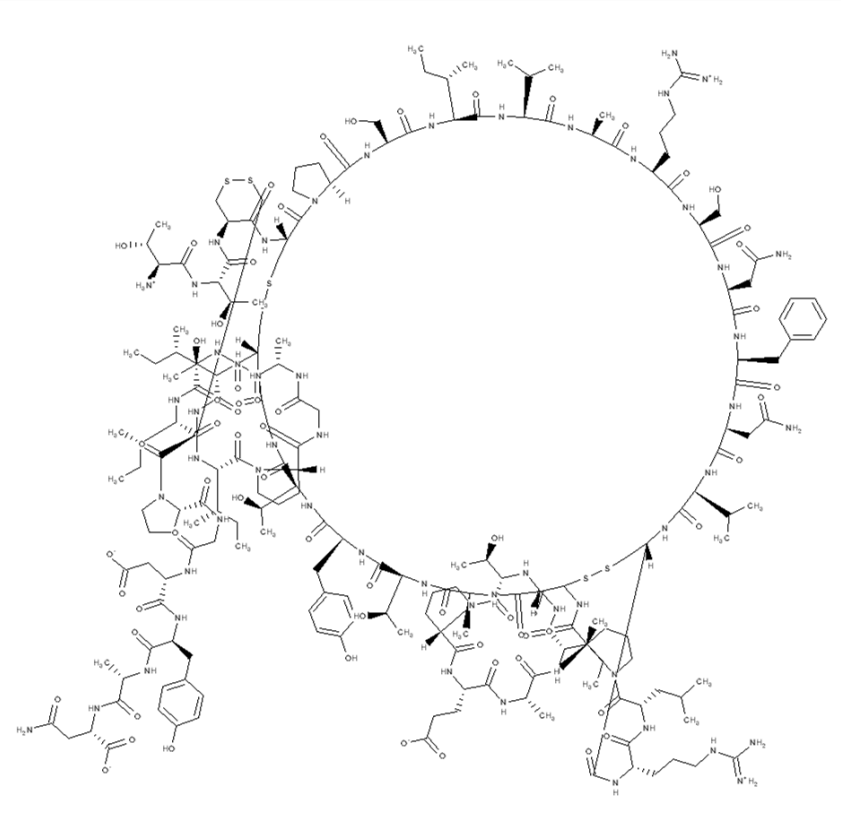


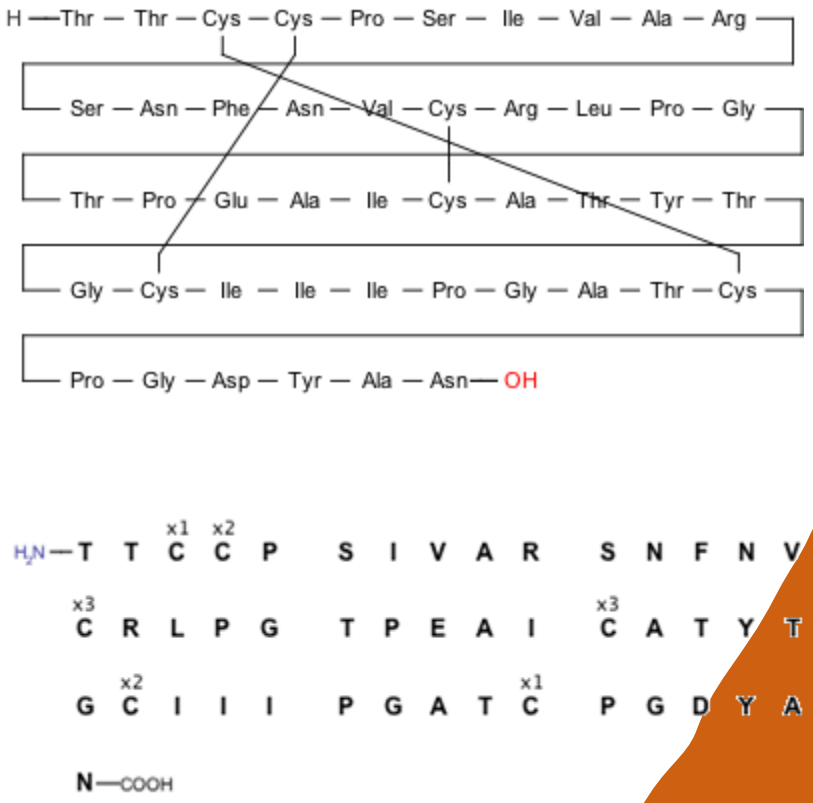


The all-atom depiction (left) is more complicated to interpret than the 3-letter code and 1-letter code forms (right).

Recognizing this is [Asp5]oxytocin is perhaps only possible at the “sequence” level.

PROTEIN REPRESENTATION

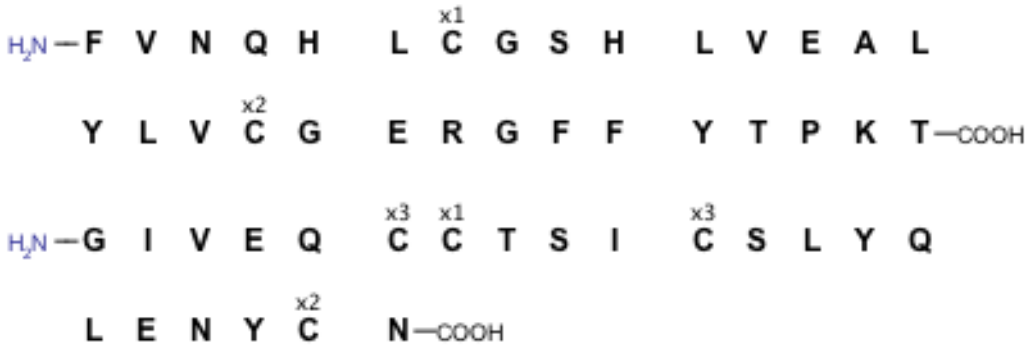




All atom representation vs. 3-letter and 1-letter sequence forms of crambin.

HUMAN INSULIN REPRESENTATIONS

CC[C@H](C)[C@H]1C(=O)N[C@H]2CSSC[C@@H](C(=O)N[C@@H](CSSC[C@@H](C(=O)NCC(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@@H](CSSC[C@H](NC(=O)[C@@H](NC(=O)[C@@H](NC(=O)[C@@H](NC(=O)[C@@H](NC(=O)[C@@H](NC(=O)[C@@H](NC2=O)CO)CC(C)C)CC3=CC=C(C=C3)O)CCC(=O)N)CC(C)C)CCC(=O)O)CC(=O)N)CC4=CC=C(C=C4)O)C(=O)N[C@@H](CC(=O)N)C(=O)O)C(=O)NCC(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](CCCNC(=N)N)C(=O)NCC(=O)N[C@@H](CC5=CC=CC=C5)C(=O)N[C@@H](CC6=CC=CC=C6)C(=O)N[C@@H](CC7=CC=C(C=C7)O)C(=O)N[C@@H]([C@@H](C)O)C(=O)N8CCC[C@H]8C(=O)N[C@@H](CCCCN)C(=O)N[C@@H]([C@@H](C)O)C(=O)O)C(C)C)CC(C)C)CC9=CC=C(C=C9)O)CC(C)C)C)CCC(=O)O)C(C)C)CC(C)C)CC2=CN=CN2)CO)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CC2=CN=CN2)NC(=O)[C@H](CCC(=O)N)NC(=O)[C@H](CC(=O)N)NC(=O)[C@H](C(C)C)NC(=O)[C@H](CC2=CC=CC=C2)N)C(=O)N[C@H](C(=O)N[C@H](C(=O)N1)CO)[C@@H](C)O)NC(=O)[C@H](CCC(=O)N)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](C(C)C)NC(=O)[C@H]([C@@H](C)CC)NC(=O)CN Human Insulin



SEQUENCE BIOINFORMATICS

- Traditional bioinformatics represents sequences as one letter codes, often in FASTA or Uniprot format.
- Four characters for nucleic acids:
  - A,C,G and T for DNA and A,C,G and U for RNA.
- 20 (or 22) characters for proteinogenic amino acids:
  - A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y and V (plus O and U).

- The challenge arises when more than the primary sequence needs to be represented.

3-LETTER CODES TO THE RESCUE?

- The use of 3-letter codes to represent monomers allows for significantly more possible monomers.
- Conveniently common AAs have standard codes:
  - Ala, Arg, Asn, Asp, Cys, Glu, Gln, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr and Val, plus Sec and Pyl.

- These assignments are mandated by IUPAC and IUB recommendations known as 3AA (dated 1983).
- Unfortunately the prevalence of standard amino acids and 3-letter codes has had adverse affects.

PDB 3-LETTER RESIDUE CODES

- wwPDB's components.cif currently lists 17,764 3-letter residue names (digits and capitals).
- These contain not only amino acids, but nucleic acids, carbohydrates, ligands, salts solvents and co-factors.
- Completely different codes are assigned to the L- and D- forms of amino acids.

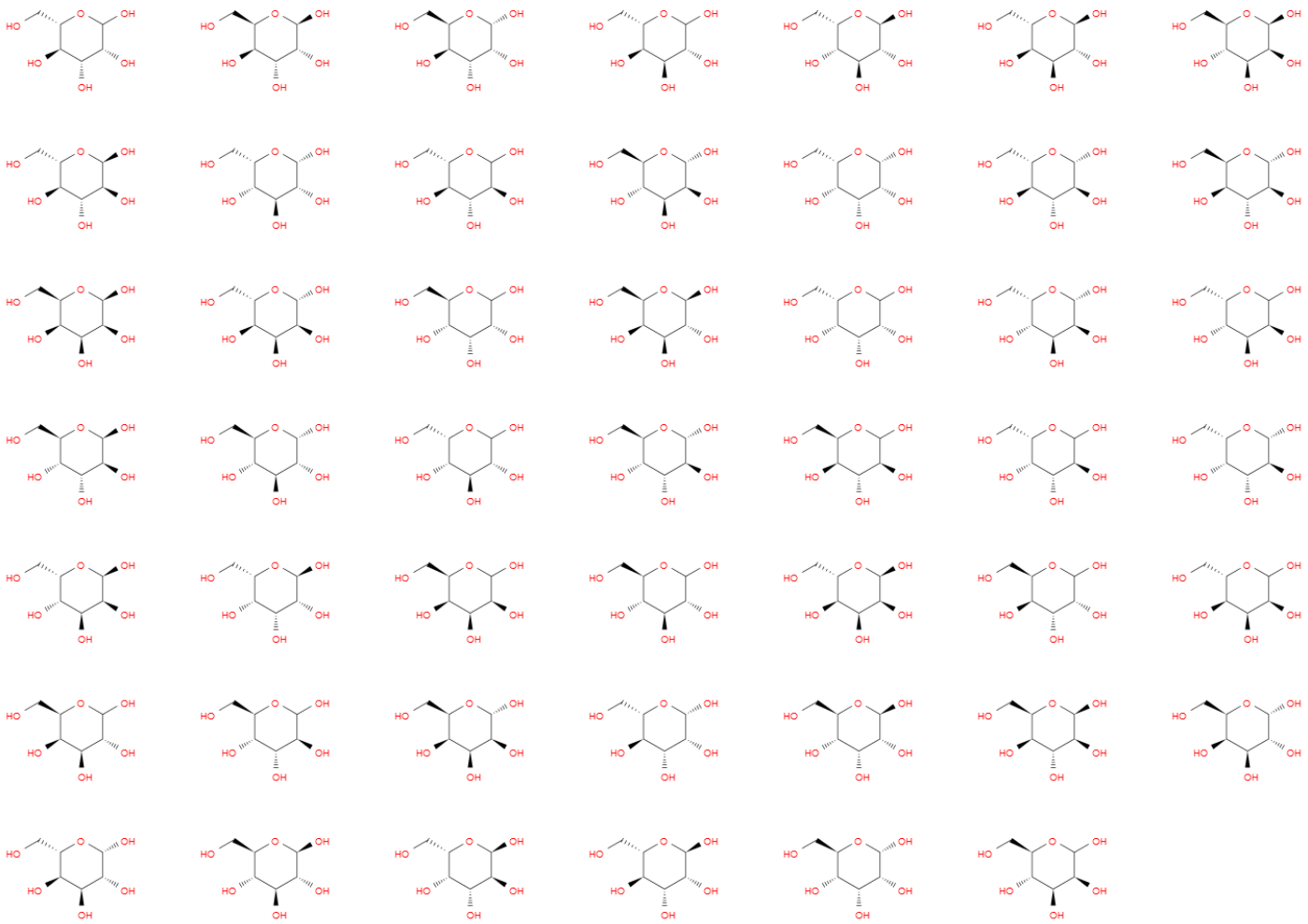
MONOMER DATABASE SOLUTIONS

- Biopolymer systems based on monomer databases and strict 3-letter codes breakdown when faced with the huge number of potential chemical and post-translational modifications.
- At best, a human being can remember 50-100 codes before having to consult “dictionary” to lookup solutions.
- Worse, once beyond the commonly encountered amino acids, biochemists disagree on the meaning of 3-letter codes…

EXAMPLE AMBIGUITIES

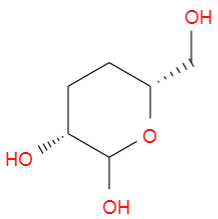
- Cpa
  - Cyano-propionic amino acid
  - Beta-cyclopropylalanine
  - 4-amino-1-carbamoyl-piperidine-4-carboxylic acid
- Hpg
  - 4-hydroxyphenylglycine
  - Homopropargylglycine
- Nty
  - Nortyrosine [Chemical Abstracts Service]
  - N-nitrotyrosine [Pisotoia HELM]

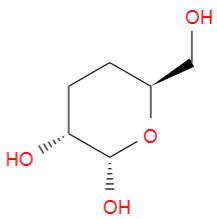
48 HEXOPYRANOSES

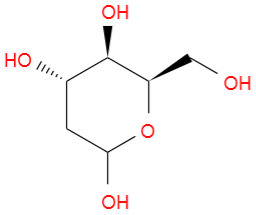


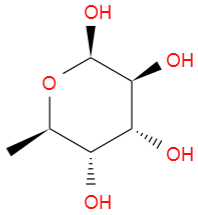
247th ACS National Meeting, Dallas, TX, Wednesday 19th March 2014

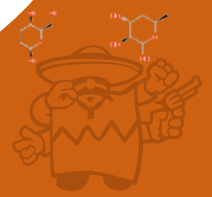
264 DEOXYHEXOPYRANOSES



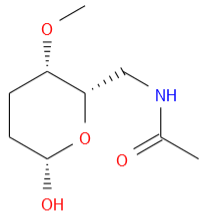


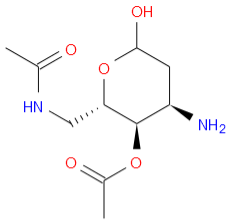


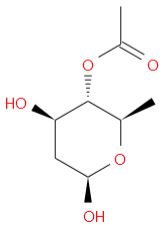


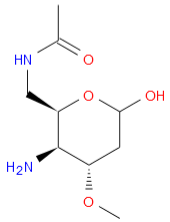


9540 SUBSTITUTED HEXOPYRANOSES









This only considers the four most common substituents: amino, acetyl, acetylamino and methoxy

BITTER SWEET CONCLUSION

- In practice, a glycoinformatics registration system has to handle over 2^{32} monosaccharide monomers.
- This number is clearly too large to assign each its own 3-letter code or to use a monomer database.
- For example, monosaccharidedb.org curates a database of 771 monosaccharides, whereas 936 of the simple monosaccharides on the previous page can be found in PubChem.

THE AA NAMING SOLUTION

- The solution to this problem is the systematic naming of amino acid derivatives.
- This approach is well known to peptide chemists, with these names routinely used in scientific articles, patent applications and supplier catalogues, and easily recognizable by chemists.
- Alas this “de facto” nomenclature, which hasn’t been formalized by IUPAC nor CAS, has been overlooked by the informatics community.

SYSTEMATIC AMINO ACID CODE EXAMPLES

- Examples of systematic codes for representing common non-standard amino acids
  - Thr(tBu), Ser(tBu), Phe(4-Cl), Phe(3-Cl), Phe(4-Ph), Tyr(PO3H2), Cys(Acm), Lys(Boc), Tyr(Bn(2,6-diCl), Tyr(3-F), Lys(Cbz), Glu(OtBu), Tyr(SO3H), Phe(4-CF3), Tyr(tBu), Arg(NO2), Trp(For), Lys(palmitoyl), aThr(tBu), Phe(4-F), Thr(Bn), Ser(Bn), Asp(OtBu), Tyr(decyl), Glu(OBn), Gly(tBu), Phe(4-hexyl), Thr(SO3H), Tyr(Me), Asp(OBn), Tyr(Bn), Phe(4-NO2), Lys(Ac), Nle(Me), Cys(tBu), Ala(cyclopentyl), Cys(StBu), Trp(2-Bn), Ser(PO3H2), Phe(4-I), Arg(Tos), etc.

MONOSACCHARIDE EXAMPLES

- In fact constructing systematic names is the same solution as used by the glycoinformatics community
  - b-D-GlcpNAc, a-D-Neup5Ac, b-D-GlcpA, D-GlcNAc-ol, a-D-GlcpN, a-L-4-en-4-deoxy-thrHexpA, b-D-Galp3Me, b-D-6-deoxy-Glcp, b-D-2,6-deoxy-ribHexp, b-D-GlcpA6Me, a-D-Rhap4N-formyl, b-D-Glcp6Ac, a-D-Neup5Ac9Ac…
- And in RNA informatics
  - P-rGuo, P-Gua-Rib2Me, P-m22Gua-Rib, P-m2Gua-Rib, P-Cyt-Rib2Me, P-m5Cyt-Rib, P-m5Ura-Rib, P-m1Ade-Rib…

RULE #1: USE OF D- AND DL- PREFIXES

- Rather than have separate code forms for D- form (and DL-) amino acids, the widely understood and accepted stereo configuration prefixes “D-”, “L-” (default) and “DL-” should be used.
  - D-Arg
  - DL-Ala
  - D-Ser

RULE #2: RETAIN WIDELY USED 3LC'S.

- Universally accepted code/definitions should be used
  - Abu (2-aminobutyric acid), Aib (2-aminoisobutyric acid), aIle (allo-isoleucine), aThe (allo-threonine), bAla (beta-alanine), Cha (beta-cyclohexylalanine), Chg (alpha-cyclohexylglycine), Cit (Citrulline), Dab (2,3-diaminobutyric acid), Dap (2,3-diaminopropionic acid), hArg (homoarginine), Hcy (homocysteine), Hse (homoserine), hPhe (homophenylalanine), Nle (norleucine), Nva (norvaline), Orn (ornithine), Pen (penicillamine), Phg (phenylglycine), Sar (sarcosine), Sec (selenocysteine), 2Thi (thien-2-ylalanine), 3Thi (thien-3-ylalanine), xille (xi-isoleucine), xiThr (xi-threonine), and many more.

RULE #3: MODIFYING PREFIXES

- "N(Me)" and friends are used to specify N-modified variants. "N(Me)Gly" is equivalent to "Sar".
- The "O" prefix is used to specify depsi-peptides that connect via ester linkages, e.g. "Arg-OAla-Ser".
- The "aMe" prefix is used to specify alpha-methyl variants of amino acids. "aMeAla" is same as "Aib".

RULE #4: LINE FORMULAE SUBSTS

- Substitutions are indicated by line formulae
  - Ph phenyl
  - Me methyl
  - Ac acetyl
  - Cl chloro
  - CN cyano
  - Bn benzyl
  - NH2 amino
  - NO2 nitro
  - tBu tert-butyl
  - CF3 trifluoromethyl
  - Tos tosyl
  - Trt trityl
  - Tf triflyl
  - SO3H sulfonic acid
  - OPfp perfluorophenoxy
  - iPr isopropyl
  - For formyl
  - … and many more

RULE #5: DEFAULT SUBSTITUTION

- Many amino acids have default substitution locants
  - Abu ABA.CG
  - Ala ALA.CB
  - Arg ARG.NH2
  - Cys CYS.SG
  - Dab DAB.ND
  - Dap DPP.NG
  - Gly GLY.CA
  - Lys LYS.NZ
  - Ser SER.OG
  - Thr THR.OG1
  - Tyr TYR.OH

RULE #6: SPECIFIED SUBSTITUTION

- Amino acids may be substituted at particular locant using the following syntax:
  - Phe(4-Cl)
  - Phg(4-CH2NH2)
  - Tyr(2,6-diCl-Bn)

- Substitutions may occur at more than one locant:
  - His(2,5-diI)
  - Tyr(2,6-diF)

RULE #7: IMPLICIT LEAVING GROUPS

- For most amino acids, parenthesized substitution implies replaced of hydrogen.
  - Ala(Cl)
- Carboxylic acids assume an OH leaving group, requiring an explicit “oxy” to form esters.
  - Asp(NH2)
  - Asp(OMe)
- Cysteine (and penicillamine) substitute on the sulfur
  - Cys(tBu)
  - Cys(StBu)

THE NEED FOR STANDARDIZATION #1

- For peptide registration, we need a canonical name.
- Preferred forms are most visible in line notations.
  - For carboxy, do we prefer “COOH” or “CO2H”.
  - For acetyl, “Ac” or “COCH3”.
  - For ethylamine, “EtNH2” or “CH2CH2NH2”.
  - For phosphate, “PO3H2” vs. “P”.
- However decomposition also affects preferred forms
  - Gly(Me) is the same as Ala
  - Asp(NH2) is the same as Asn
  - Phe(4-Ph) is the same as 4-Bip

THE NEED FOR STANDARDIZATION #2

- More seriously we need to avoid ambiguous names.

- It is widely acknowledge by peptide chemists that “Ac” means acetyl, rather than Actinium, and connects to the parent at the carbonyl carbon.
- In supplier catalogues for a German vendor, the name “Cys(AcOH)” implies “*CC(=O)O” instead of “*C(=O)CO”.
- Which names should be used, “COCH2OH” etc.?

EXAMPLE PEPTIDE NAMES (CHEMBL)

- The following names are machine generated

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

EXAMPLE PEPTIDE NAMES (PUBCHEM)

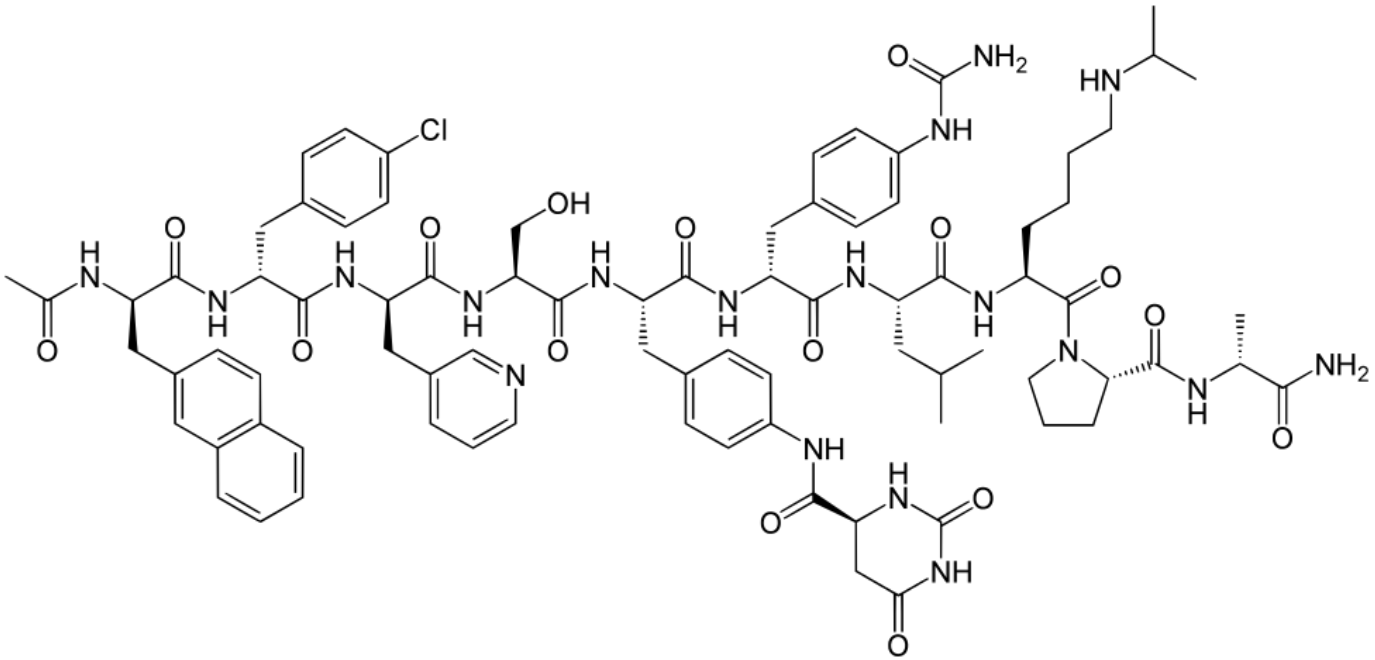
- The following names are existing supplier synonyms

• 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
• Z-D-Val-Lys(Z)-OH	CID5978?
• deamino-Cys-D-Tyr(Et)-Ile-Thr-Asn-Cys-Pro-Orn-Gly-NH2	CID68613
• H-Arg(NO2)-OMe.HCl	CID135193

- Unlike HELM or PLN, these IUPAC condensed line notations (3AA) are already used in publications.

THE BIG WIN: DISPLAY OF PEPTIDES

- One of the largest end-user benefits of systematic naming is in the improved display of compounds:



Structure of Degarelix from wikipedia: http://en.wikipedia.org/wiki/Degarelix

THE BIG WIN: DISPLAY OF PEPTIDES

- The image of Degarelix becomes optimal (IMHO)

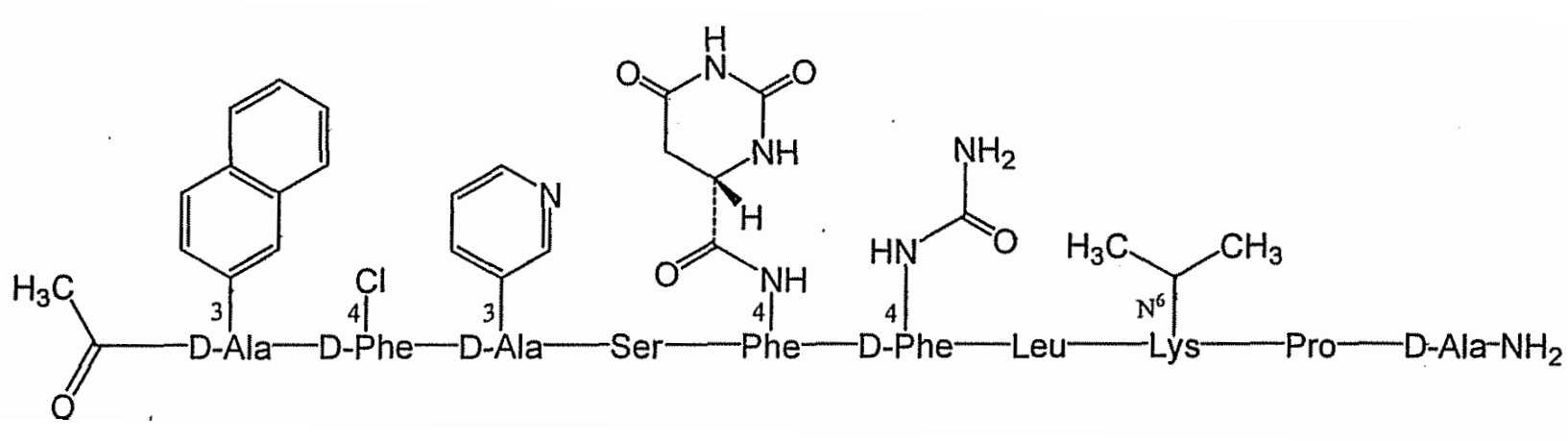
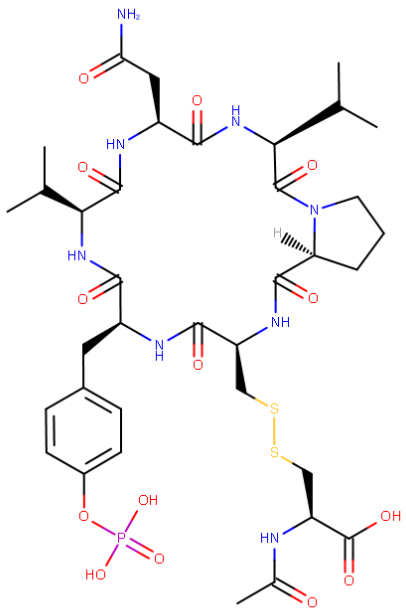
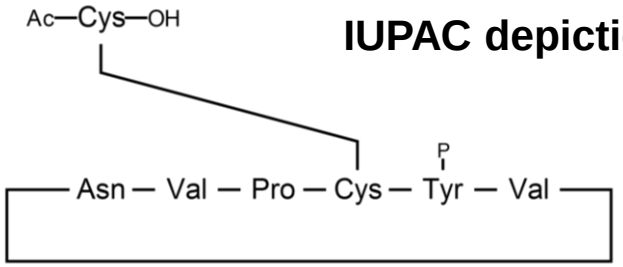


Image credit: WO2011066386A1

A REAL EXAMPLE: CID10033747





IUPAC depiction

IUPAC Condensed

cyclo[Asn-Val-Pro-Cys(1)-Tyr(PO3H2)-Val].
Ac-Cys(1)-OH.

HELM

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:R34:R3|
PEPTIDE2,CHEM1,4:R2-1:R1|
CHEM1,PEPTIDE3,1:R21:R1|
PEPTIDE3,PEPTIDE2,1:R2-1:R1$$$

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

CONCLUSIONS

- The use of 3-letter codes to represent non-standard amino acids doesn't scale to peptide registration systems.
- Adopting the systematic rules already frequently used by peptide chemists looks to solve a number of amino acid nomenclature challenges faced by PDB, HELM and the pharmaceutical industry.
- This presentation/proposal formalizes some of the syntax/rules for constructing systematic codes for non-standard amino acids.

ACKNOWLEDGEMENTS

- Lisa Sach-Peltason, Hoffmann-La Roche, Basel.
- Evan Bolton, NCBI PubChem project, Bethesda, MD.
- Noel O'Boyle, NextMove Software, Cambridge, UK.
- Daniel Lowe, NextMove Software, Cambridge, UK.