



256th ACS National Meeting Washington Aug 2018

Building a bridge between
human-readable
and
machine-readable
representations of biopolymers

Noel O'Boyle and Roger Sayle

NextMove Software



INTRODUCTION



Lipidated Peptide Dendrimers Killing Multidrug-Resistant Bacteria

Thissa N. Siriwardena[†], Michaela Stach[†], Runze He^{†‡}, Bee-Ha Gan[†], Sacha Javor[†], Marc Heitz[†], Lan Ma^{‡§}, Xiangju Cai[§], Peng Chen[‡], Dengwen Wei[‡], Hongtao Li[‡], Jun Ma[§], Thilo Köhler[¶], Christian van Delden[¶], Tamis Darbre^{*†} , and Jean-Louis Reymond^{*†} 

[†] Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland

[‡] Shanghai Space Peptides Pharmaceutical Co. Ltd, Shanghai 201210, China

[§] College of Pharmacy, Gansu University of Chinese Medicine, Dingxi East Road 35, Chenguan District, Lanzhou, Gansu Province 730000, China

^{||} Lanzhou Ruibei Pharmaceutical R&D Co., Ltd., Lanzhou, Gansu Province 730000, China

[⊥] Department of General Surgery, Lanzhou General Hospital of Lanzhou Military Region, PLA, 333 South Binhe Road, Qilihe District, Lanzhou, Gansu Province 730046, China

[¶] Department of Microbiology and Molecular Medicine, University of Geneva, CH-1211 Geneva, Switzerland

[°] Service of Infectious Diseases, University Hospital of Geneva, CH-1205 Geneva, Switzerland

J. Am. Chem. Soc., 2018, 140 (1), pp 423–432


DOI: 10.1021/jacs.7b11037

Publication Date (Web): December 5, 2017

Copyright © 2017 American Chemical Society

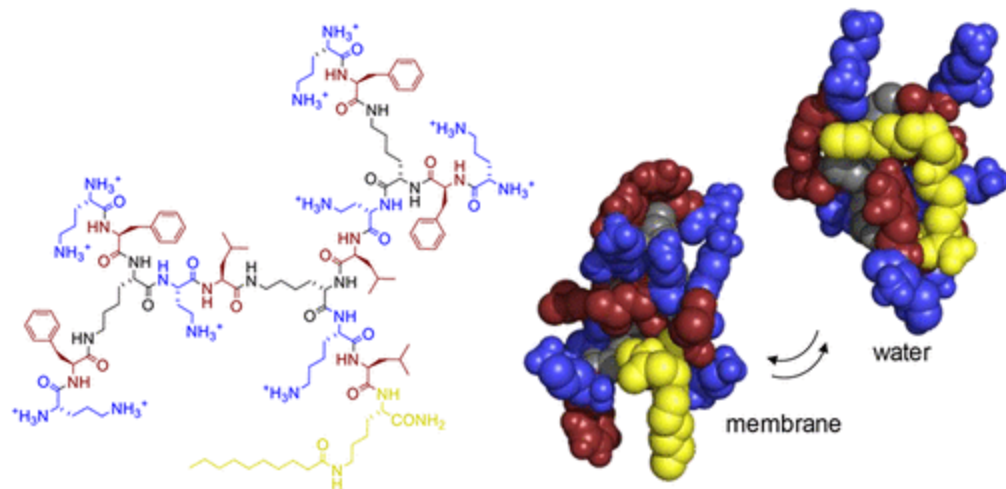
*jean-louis.reymond@dcb.unibe.ch,

*tamis.darbre@dcb.unibe.ch

 Cite this: *J. Am. Chem. Soc.* 140, 1, 423-432

 RIS Citation 

Abstract



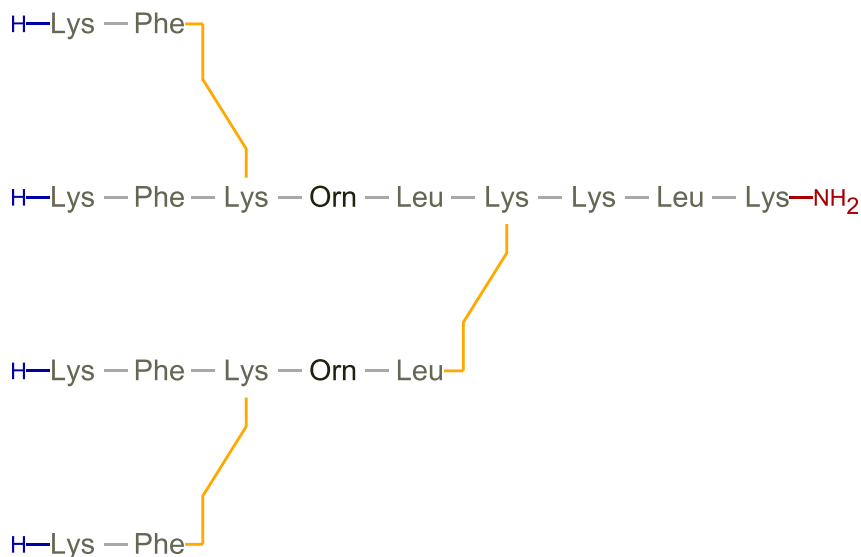
NAMES PREFERRED BY MACHINES

SMILES

```
CC(C)C[C@H](NC(=O)[C@H](CCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](Cc1ccccc1)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](Cc1ccccc1)NC(=O)[C@@H](N)CCCCN)C(=O)NCCCC[C@H](NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](Cc1ccccc1)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](Cc1ccccc1)NC(=O)[C@@H](N)CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](CCCCN)C(N)=O
```

HELM

```
PEPTIDE1{K.F.K.[Orn].L.K.K.L.K.[am]}|PEPTIDE2{K.F.K.[Orn].L}|PEPTIDE3{K.F}|PEPTIDE4{K.F}$PEPTIDE1,PEPTIDE2,6:R3-5:R2|PEPTIDE2,PEPTIDE3,3:R3-2:R2|PEPTIDE1,PEPTIDE4,3:R3-2:R2$$$
```



NAMES PREFERRED BY HUMANS

- IUPAC/IUBMB recommendations from 1983 describe a **three-letter system** for peptides [1]

adrenorphin as H-Tyr-Gly-Gly-Phe-Met-Arg-Arg-Val-NH₂

- L- by default, D-/DL- must be specified
 - Side-chain substitutions like Ser(Ac), Asp(OMe)
 - Terminal modifications like Ac-Tyr-OMe, Me₂-Lys
 - Backbone N substitution like Ala-(Me)Ala
 - Cyclic peptides like cyclo(-Val-Orn-Leu-)
 - Peptide analogs like Ala-[psi](NH-CO)-Ala
- See also Bachem nomenclature guide [2]

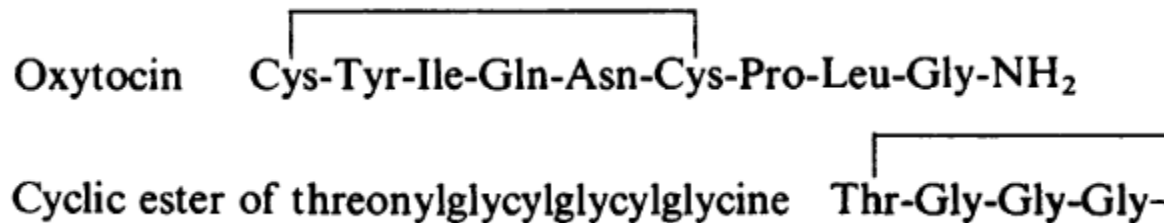
[1] <http://www.sbcs.qmul.ac.uk/iupac/AminoAcid/>

[2] <http://www.bachem.com/service-support/faq/nomenclature/>



DESCRIBING CYCLES WITHOUT LINES

- The recommendations use **drawings of bonds** to indicate heterodetic cyclic peptides



- In practice, people either use free text or **cyclo**
 - **cyclo** only handles simple situations; cannot handle overlapping disulphide bridges for example
- We use **ring closure bonds** like in SMILES [1]
 - H-Cys(1)-Tyr-Ile-Gln-Asn-Cys(1)-Pro-Leu-Gly-NH₂
 - H-Thr(1)-Gly-Gly-Gly-(1)

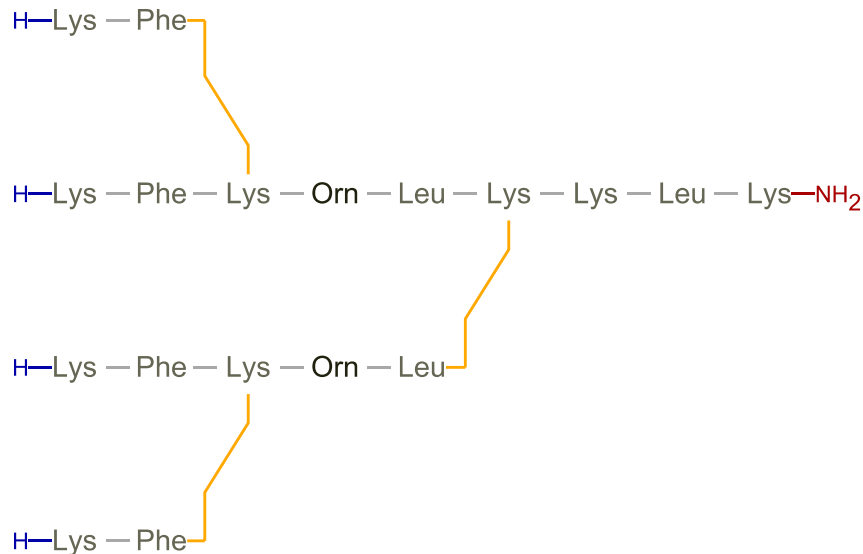
[1] Similar to that described in “Abbreviated nomenclature for cyclic and branched homo- and hetero-detic peptides.” *J. Peptide Res.* **2005**, 65, 550.



NAMES SUITABLE FOR HUMANS *AND* MACHINES

IUPAC Condensed

H-Lys-Phe-(1).H-Lys-Phe-Lys(1)-Orn-Leu-Lys(2)-Lys-Leu-Lys-NH₂.H-Lys-Phe-Lys(3)-Orn-Leu-(2).H-Lys-Phe-(3)



HUMAN-READABLE MONOMER NAMES



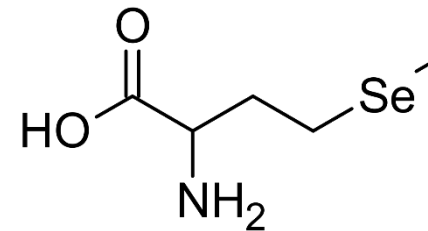
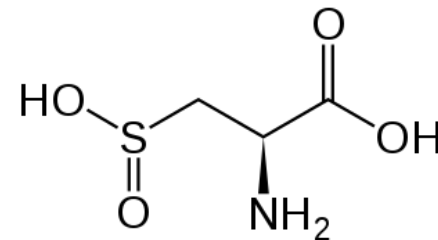
ROGER'S RECOMMENDATIONS [1]

- Don't use a dictionary; create monomers from building blocks and use systematic names
 - Stereo, parent, backbone/sidechain substituents
 - H-Ala-D-N(Bu)Phe(4-Cl)-OH
- Retain widely used 3-letter codes and use substituent abbreviations or line formulae
- Consider aminoacids to have default substitution locants, but have ability to specify
 - Ser(Me), Phe(4-Cl)
- Implicit leaving groups
 - Asp(OMe) – OH for acids, H otherwise



MONOMER NAMES, CHOOSE WISELY

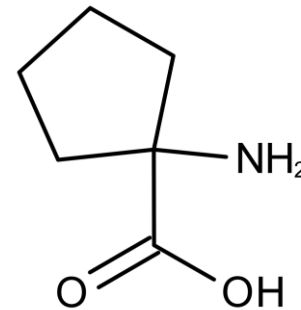
- S&S used some monomer names adapted from PDB monomer codes
 - Sometimes not used in practice in the field
- For cysteine sulfinic acid
 - PDB has CSD, and S&S had Csd
 - Now changed to **Cys(O2H)**
- For selenomethionine
 - PDB has MSE, and S&S had Mse
 - Now changed to **SeMet**



MONOMER NAMES, CHOOSE WISELY

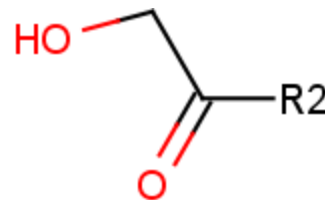
- 1-amino-cyclopentyl carboxylic acid

- HELM 1.0 has **Spg**
- S&S has **Ac5c** (**Ac6c**, etc.)
- Vendors use **Ac5c** but also **Cle**



- HELM 1.0 has **Glc** as a peptide N-terminal modifier

- Representing glycolic acid
- In sugar nomenclature, **Glc** is glucose, the most common monosaccharide



TEXT-MINING HUMAN REPRESENTATIONS

- What three (or more) letter codes do people use?
 - For non-standard aminoacids
 - For sidechain substituents
 - For C- and N- terminal modifications
 - For non-standard connections between aminoacids
- We can answer these questions by **text-mining** PubMed Abstracts or the patent literature
 - Using a grammar for IUPAC condensed notation



MINED THE GAPS

- Text-mining is usually for matching phrases that you know
 - How do you text-mine phrases that you **don't know**?
- **Look at the gaps** between text that is recognised (“entity extension”)
 - Filter for gaps that do not contain space
 - Sort the results by frequency to identify common abbreviations that were missed



Format: Abstract ▾

Send to ▾

Can J Physiol Pharmacol. 1997 Jun;75(6):719-24.

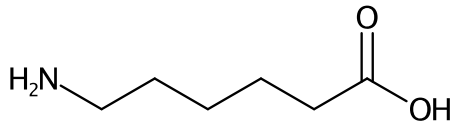
Potent, long-acting bradykinin antagonists for a wide range of applications.

Stewart JM¹, Gera L, Chan DC, Whalley ET, Hanson WL, Zuzack JS.

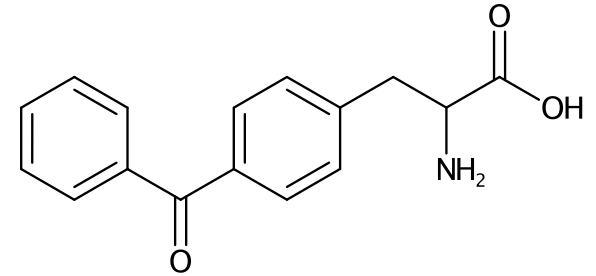
Author information

Abstract

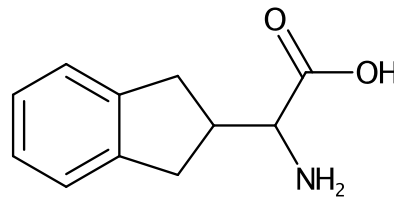
Actions of bradykinin (Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg; BK) are mediated by constitutively expressed B2 receptors (which require the full BK peptide chain) and by B1 receptors (which require BK (1-8) as ligand) that are induced in inflammation. BK has many functions in normal and pathological physiology, including initiation of most, if not all, inflammation. BK also evidently functions as an autocrine stimulant for growth of small cell lung cancer (SCLC). A new group of BK antagonists containing the novel amino acid alpha-(2-indanyl)glycine provides both broad-spectrum and selective antagonists for all these functions. As examples, D-Arg-Arg-Pro-Hyp-Gly-Igl-Ser-D-Igl-Oic-Arg (B9430) is an extremely potent and long-acting antagonist of both B1 and B2 receptors, is stable against endogenous kininase enzymes and is active in various in vivo models, including by intragastric administration. Acylation of B9430 with dehydroepiandrosterone-2-carboxylic acid (Dhq) gives B9562, a highly selective B2 antagonist. In contrast, Lys-Lys-Arg-Pro-Hyp-Gly-Igl-Ser-D-Igl-Oic (B9858) is a highly potent and selective B1 antagonist. The dimer of B9430 linked at the amino terminus with suberimide is a potent selectively cytotoxic agent for SCLC cells. Results with these peptides suggest that a new generation of antiinflammatory and anticancer drugs may be at hand.



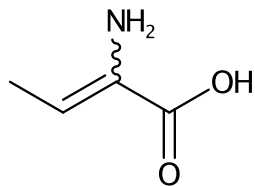
Aca
ε-aminocaproic acid
9 times



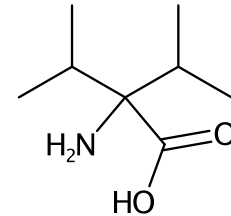
Bpa
p-benzoyl phenylalanine
10 times



Igl
2-indan-2-yl-glycine
19 times



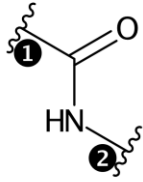
Dhb
dehydrobutryine
21 times



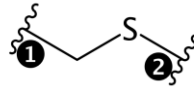
Dpg
α,α-diisopropylglycine
19 times



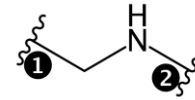
UNRECOGNISED PEPTIDE ANALOGS



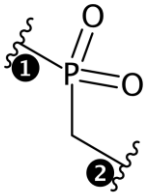
(peptide bond)



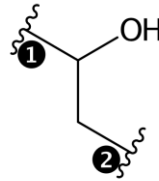
-psi-(CH₂S)-
10 times



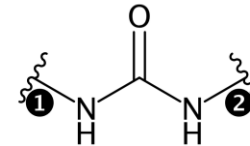
-psi(CH₂NH)-
9 times



-psi(PO₂CH₂)-
7 times



-psi[CH(OH)CH₂]
7 times



-NH-CO-NH-
25 times

E.g. Ala-psi(CH₂NH)-Ala instead of Ala-Ala



UNRECOGNISED N-TERMINUS PREFIXES

- Extract text **preceding** recognised text
 - Must end with a hyphen
 - Stop at the first space

PubMed.gov

PubMed

US National Library of Medicine
National Institutes of Health

Advanced

Format: Abstract

Send to

[Biochem J.](#) 1975 Dec;151(3):527-42.

The substrate specificity of thermomycolase, an extracellular serine proteinase from the thermophilic fungus *Malbranchea pulchella* var. *sulfurea*.

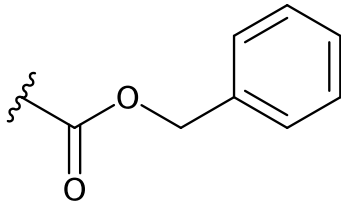
[Stevenson KJ](#), [Gaucher GM](#).

Abstract

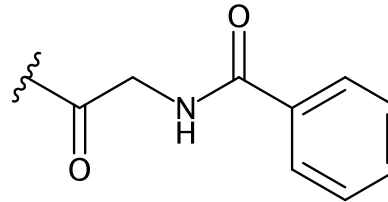
The specificity of thermomycolase toward glucagon and the oxidized A and B chains of insulin was investigated. Extensive digestion of glucagon occurred when conducted at pH 7.0 and 45 degrees C for 40 min, whereas hydrolysis of only three peptide bonds occurred at pH 7.0 and 28 degrees C for 5 min. A similar situation was observed for the oxidized B chain of insulin, which exhibited only a single major cleavage after 5 min at 25 degrees C. No well-defined specificity for particular amino acid residues was evident, but ready hydrolysis of peptide bonds occurred within sequences containing non-polar residues. This endoproteinase must therefore possess an extended hydrophobic binding site for polypeptides. Thermomycolase hydrolysed acetylalanylalanylalanine methyl ester and elastin-Congo Red at 22 and 8.5 degrees C, respectively, rates the rate of porcine elastase respectively. A limited degradation of native collagen and significant hydrolysis of benzyloxycarbonyl-Gly-Pro-Leu-Gly-Pro were suggestive of some collagenase-like activity. No keratinase activity was apparent.

UNRECOGNISED N-TERMINUS PREFIXES

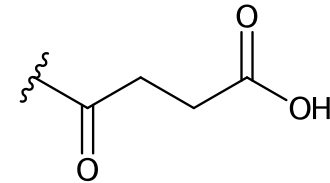
- Extract text **preceding** recognised text
 - Must end with a hyphen
 - Stop at the first space



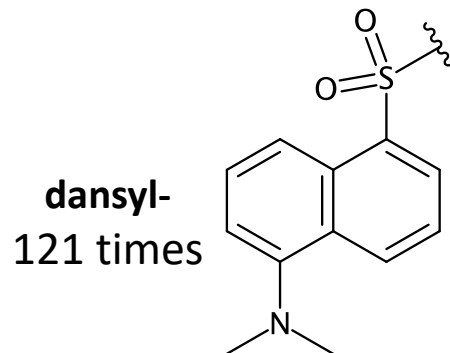
benzyloxycarbonyl-
530 times



hippuryl-
39 times



Suc-/succinyl-
329/167 times

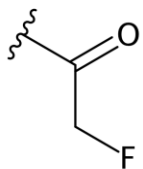


dansyl-
121 times

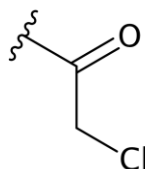


UNRECOGNISED C-TERMINUS SUFFIXES

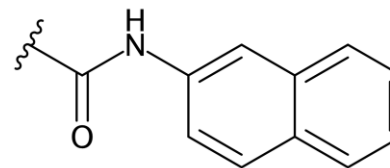
- Extract text **following** any recognised text
 - First pass focused on text beginning with a hyphen up to the first space
 - Later analyses identified common space-separated phrases for esters



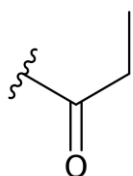
-fluoromethylketone
(and variants)
607 times



-chloromethylketone
(and variants)
270 times



-beta-naphthylamide
(and variants)
193 times

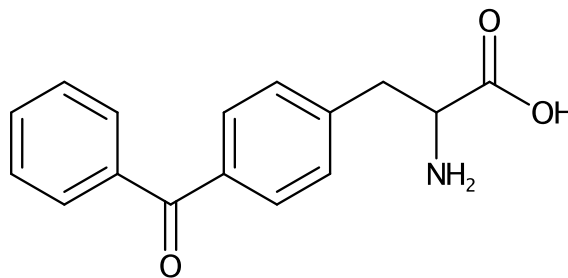


ethyl ester
(already had -OEt)
30 times



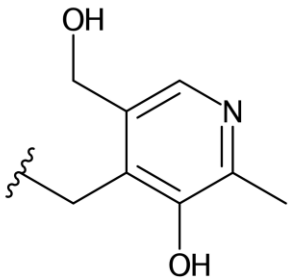
3-LETTER CODES

- “A foolish consistency is the hobgoblin of little minds” – Ralph Waldo Emerson
- No particular reason to stick to 3-letters
 - Eventually leads to ambiguities
 - **Hyp** as hydroxyproline or hypoxanthine
 - **Xan** as xanthen-9-yl or xanthosine
- Unless an abbreviation is very common, favor longer names that are more descriptive
 - **Igl** vs **Gly(indan-2-yl)**
 - **Dhb** vs **Abu(2,3-dehydro)**
 - **Bpa** vs **Phe(4-Bz)**

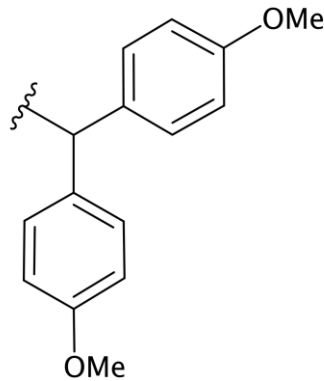


UNRECOGNISED SUBSTITUENTS

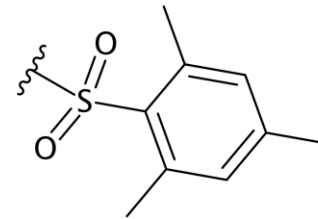
- Search for text containing any aminoacid followed by a **bracketed expression**
 - I used a regular expression



Pxy, e.g. Lys(**Pxy**)
pyridoxyl
7 times



Mbh, e.g. Asn(**Mbh**)
4,4'-dimethoxybenzhydryl
8 times



Mts, e.g. Arg(**Mts**)
mesitylene-2-sulfonyl
7 times



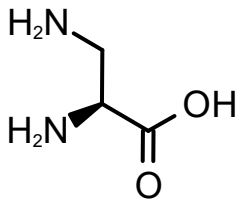
SOME CARE NEEDED

- Not everything that looks like a peptide *is* a peptide
 - Ile-de-France (sheep)
 - Glyol, Leuol and Pheol
 - but not Tyrol, Lysol or Metol
 - Argal and Proal
 - but not Metal, Ileal, Penal or Seral
- Even where it definitely is a peptide, it is necessary to check the details:
 - Is this abbreviation **widely used**?
 - Does it occur in **different** peptides?
 - Is it **unambiguous**?

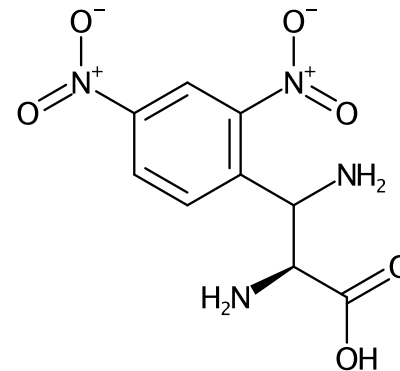


SOME CARE NEEDED

- **Dpa** was found to occur 21 times in the gaps
 - Many times (but not always) in the same peptide
- Inspection of the papers behind the abstracts and googling “Boc-Dpa-OH” indicated **two** potential meanings



diaminopropionic acid
aka **Dap**



3-(2,4-dinitrophenyl)-L-2,3-
diaminopropionyl



HUMAN-READABLE → MACHINE-READABLE

- **16.4K oligopeptides*** textmined from PubMed Abstracts

Tyr-d-Ala-Phe-Gly-Tyr-Pro-Ser-NH(2)

Asp-Thr(P)-Pro-Ala-Lys

Pyr-Gly-Pro-Pro-Ile-Ser-Ile-Asp-Leu-Ser-Leu-Glu-Leu-Leu-Arg-Lys-Met-Ile-Glu-Ile

Gly-Ala-Aib-Pro-Ala-Aib-Aib-Glu

Nle-Leu-Phe-Nle-Tyr-Lys

L-Ala-D-Glu-L-Lys-D-Ala-D-Ala

D-Phe-Cys-Tyr-D-Trp-Orn-Thr-Pen-Thr-NH2

Glu-Asp-Pro-Gln-Gly-Asx-Ala-Ala

Ac-Phe-Leu-Val-His-NH2

Gly-Asx-Glx-Ser-Thr-Cys

Ac-Met-Glu-Glu-Lys-Leu-Lys-Lys-Thr-Lys-Ile-Ile-Phe-Val-Val-Gly-Gly-Pro-Gly-Ser-Gly-Lys-Gly-Thr-Gln-

Cys-Glu-Lys-Ile-Val-Gln-Lys-Tyr-Gly-Tyr-Thr-His-Leu-Ser-Thr-Gly-Asp-Leu-Leu-Arg-Ser-Glu-Val-Ser-Ser-

Gly-Ser-Ala-Arg-Gly-Lys-Lys-Leu-Ser-Glu-Ile-Met-Glu-Lys-Gly-Gln-Leu-Val-Pro-Leu-Glu-Thr-Val-Leu-Asp-

Met-Leu-Arg-Asp-Ala-Met-Val-Ala-Lys-Val-Asn-Thr-Ser-Lys-Gly-Phe-Leu-Ile-Asp-Gly-.....

* Containing at least 5 monomers

HUMAN-READABLE → MACHINE-READABLE

- **16.4K oligopeptides*** textmined from PubMed Abstracts and converted to HELM
 - 2.2K not converted, 1.6K as inline HELM, 12.6K as regular HELM

```
PEPTIDE1{Y.[dA].F.G.Y.P.S.[am]}$$$$
PEPTIDE1{D.[*C(=O)[C@H]([C@@H](C)OP(=O)(O)O)N* |$_R2;;;;;;;;;;;;;_R1$|].P.A.K)}$$$$
PEPTIDE1{[Glp].G.P.P.I.S.I.D.L.S.L.E.L.L.R.K.M.I.E.I)}$$$$
PEPTIDE1{G.A.[Aib].P.A.[Aib].[Aib].E)}$$$$
PEPTIDE1{[Nle].L.F.[Nle].Y.K)}$$$$
PEPTIDE1{A.[dE].K.[dA].[dA]}$$$$
PEPTIDE1{[dF].C.Y.[dW].[Orn].T.[Pen].T.[am]}$$$$
PEPTIDE1{E.D.P.Q.G.(D,N).A.A)}$$$$V2.0
PEPTIDE1{[ac].F.L.V.H.[am]}$$$$
PEPTIDE1{G.(D,N).(E,Q).S.T.C)}$$$$V2.0
PEPTIDE1{[ac].M.E.E.K.L.K.K.T.K.I.I.F.V.V.G.G.P.G.S.G.K.G.T.Q.C.E.K.I.V.Q.K.Y.G.Y.T.H.L.S.T.G.D.L.L.R.S.E.V.S.S
.G.S.A.R.G.K.K.L.S.E.I.M.E.K.G.Q.L.V.P.L.E.T.V.L.D.M.L.R.D.A.M.V.A.K.V.N.T.S.K.G.F.L.I.D.G.Y.P.R.E.V.Q.Q.G.E.
E.F.E.R.R.I.G.Q.P.T.L.L.L.Y.V.D.A.G.P.E.T.M.T.R.R.L.L.K.R.G.E.T.S.G.R.V.D.N.E.E.T.I.K.K.R.L.E.T.Y.Y.K.A.T.E.P.V.I.A.
F.Y.E.K.R.G.I.V.R.K.V.N.A.E.G.S.V.D.E.V.F.S.Q.V.C.T.H.L.D.A.L.K)}$$$$
```

* Containing at least 5 monomers

HUMAN-READABLE PEPTIDE NAMES



IUPAC NAMES FOR PEPTIDES



(S)-N-((S)-1-((2-amino-2-oxoethyl)amino)-4-methyl-1-oxopentan-2-yl)-1-((4R,7S,10S,13S,16S,19R)-19-amino-7-(2-amino-2-oxoethyl)-10-(3-amino-3-oxopropyl)-16-(4-hydroxybenzyl)-13-isobutyl-6,9,12,15,18-pentaoxo-1,2-dithia-5,8,11,14,17-pentaazacycloicosane-4-carbonyl)pyrrolidine-2-carboxamide

(2S)-1-[(4R,7S,10S,13S,16S,19R)-19-amino-7-(2-amino-2-oxo-ethyl)-10-(3-amino-3-oxo-propyl)-16-[(4-hydroxyphenyl)methyl]-13-isobutyl-6,9,12,15,18-pentaoxo-1,2-dithia-5,8,11,14,17-pentazacycloicosane-4-carbonyl]-N-[(1S)-1-[(2-amino-2-oxo-ethyl)carbamoylethyl]-3-methyl-butyl]pyrrolidine-2-carboxamide

(2S)-2-{{{(2S)-1-[(4R,7S,10S,13S,16S,19R)-19-amino-10-(2-carbamoylethyl)-7-(carbamoylmethyl)-16-[(4-hydroxyphenyl)methyl]-13-(2-methylpropyl)-6,9,12,15,18-pentaoxo-1,2-dithia-5,8,11,14,17-pentazacycloicosane-4-carbonyl]pyrrolidin-2-yl]formamido}-N-(carbamoylmethyl)-4-methylpentanamide

L-cysteinyl-L-tyrosyl-L-leucyl-L-glutaminy-L-asparagyl-L-cysteinyl-L-prolyl-L-leucyl-glycinamide (1->6)-disulphide

[Leu3]oxytocin



NAMES THAT SHOW RELATIONSHIPS

- Often better to describe a structure as a **delta** (or modification) of a known structure
 - Earlier, I showed Bpa versus Phe(4-Bz); reduces cognitive load; similarity of Phe(4-X) follows intuitively
- If we apply this to whole peptides:
 - Name a peptide as a delta from a **reference set** of peptides – e.g. ‘known peptides’, or an in-house database
- Modification nomenclature described in 1983 IUPAC-IUBMB recommendations



ANALYSIS OF PUBCHEM

- Curated database of oligopeptides of biological interest (currently 452 entries)
- 10.5% of the 170,708 peptides of length 5 or greater in PubChem can be named as variants of these
 - **argipressin (1-8)** vs H-Cys(1)-Tyr-Phe-Gln-Asn-Cys(1)-Pro-Arg-OH
 - **Cbz-cholecystokinin octapeptide (2-7) amide** vs Cbz-Tyr(SO₃H)-Met-Gly-Trp-Met-Asp-NH₂
 - **[Ile1,Ser2,Ser8]cyphokinin** vs H-Ile-Ser-Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg-OH



IN CONCLUSION

- Machines can bridge the gap to humans by reading/writing **human-readable representations** of biopolymers
- Appropriate monomer names can be found by **mining** the literature

TOOLS USED

- Textmining using LeadMine
- Biologic representations using Sugar&Splice

THANKS!

