NAMING ALGORITHMS FOR DERIVATIVES OF PEPTIDE-LIKE NATURAL PRODUCTS

ROGER SAYLE & NOEL O'BOYLE
NEXTMOVE SOFTWARE, CAMBRIDGE, UK
CHRISTOPHER SOUTHAN,
IUPHAR/BPS GUIDE TO PHARMACOLOGY

250th ACS National Meeting, Boston, MA. Sunday 16th August 2015
30 SECOND OVERVIEW

• This talk describes the development of software for both naming peptides and reading peptide names matching the de facto standard practices currently followed by biochemists.

• Unlike computer representations, like Pisotoia HELM or InChI keys, these names and identifiers match those typically found in the scientific literature and vendor catalogues.

• A significant application of this technology is to check and correct peptide representations in databases.
PROBLEM MOTIVATION

• Maintaining a database of biological activities of mature proteins and peptides presents a significant technical challenge.

• IUPHAR/BPS’ “Guide to Pharmacology” and EBI’s ChEMBL represent current state-of-the-art efforts to capture/represent peptide-like ligands.

• The ligands require more than (FASTA) bioinformatics including disulfide bridging architecture, non-standard amino acids, post-translational modifications, N- and C- terminal modifications etc.
IUPHAR/BPS GTOP PEPTIDES

• The “Guide to Pharmacology” database contains:
  – The common name “oxytocin”
  – The species, e.g. “human”
  – The UniProt ID “P001178”
  – The 1-letter Sequence “CYIQNCPLG”
  – The 3-letter sequence “Cys-Tyr-Ile-…-Leu-Gly-NH2”
  – A text description “Post-translation modification”, e.g. “A disulfide bond is formed between cysteine residues at positions 1 and 5 and the C-terminal glycine is amidated”.
  – Often SMILES and standard InChIKey.
PROBLEM 1: CONSISTENCY

• The challenge with these advanced formats are that the names, three-letter codes and modification descriptions are text-locked, unreadable by software.

Examples of errors and inconsistency:

Ligand #4463: “PheGlnThrSerGluAlaIleLeuPro...”
Ligand #1335: “…Leu-Arg-AlaPro-Leu-Lys…”
Ligand #8263: “val-leu-gln-glu-leu-asn-val-thr-val”
Ligand #5873: “…Pr-oGl-yGl-ySe-rMe-tLy-sLe-u…”
Ligand #3591: “PHQLLRVPro-His-Ala-Gln-Leu...”
Problem 2: Ambiguity

• Ligand #3630 (neuropeptide B29) “(Br)Trp” with note “The n-terminal tryptophan is brominated”.
  – Suggested replacement Trp(6-Br)
• In Ligand #1036, “(Ac)Ala” means N2-acetyl but “(Ac)Lys” means N6-acetyl, in #1188 “Ac-” appears without parenthesis, in Ligand #3853, “AcPhe-” appears without a hyphen...
  – Suggest Ac- at N-term, -N(Ac)Phe infix, Lys(Ac) sidechain
  – This even allows Ac-N(Ac)Lys(Ac)-OH, aka. N(Ac2)Lys(Ac).
Problem 3: Disulfide Bridging

• Capturing the disulfide bridging architecture in the three-letter (condensed) representation allows it to be read/checked for errors.
• This is done in some places but not in others.
• Disulfide bridges are particularly tricky even for the folks at UNIPROT: Annexin I (ligand #1031, P04083) isn’t annotated as disulfide bridged, despite the 3D structure in PDB 1HM6, and the experimental evidence of an intramolecular disulfide described in PubMed 7663390.
TYPES OF PEPTIDE NAME/IDENTIFIER

• **Sequence**: CYIQDCPLG
• **Peptide Name**: [Asp5]oxytocin or [5-L-aspartic acid]oxytocin
• **Chemical IUPAC Name**: 2-[(4R,7S,10S,13S,16S,19R)-19-amino-4-[(2S)-2-[[1S]-1-[(2-amino-2-oxo-ethyl)carbamoyl]-3-methylbutyl]carbamoyl]pyrroolidine-1-carbonyl]-10-(3-amino-3-oxo-propyl)-16-[(4-hydroxyphenyl)methyl]13-[(1S)-1-methylpropyl]-6,9,12,15,18-pentaoxo-1,2-dithia-5,8,11,14,17-pentazacycloicos-7-yl]acetic acid
• **Biological IUPAC Name**: L-cysteinyln-L-tyrosyl-L-isoleucyl-L-glutaminyln-L-alpha-aspartyl-L-cysteinyln-L-prolyln-L-leucyl-glycinamide (1->6)-disulfide
• **Condensed**: Cys(1)-Tyr-Ile-Gln-Asp-Cys(1)-Pro-Leu-Gly-NH2
• **Pistoia HELM**: PEPTIDE1{C.Y.I.Q.N.C.P.L.G.[am]}$PEPTIDE1,PEPTIDE1,1:R3-6:R3$$$

250th ACS National Meeting, Boston, MA. Sunday 16th August 2015
Helm Teething Problems

• Pistoia’s HELM notation marks a significant advance over the limitations of one-letter bioinformatics.
• Alas, its original goals didn’t include data exchange, which has only recently been addressed by the extensions of inlineHELM and XHELM [and fixes from NextMove Software for improved interoperability].
• Alas, this still doesn’t address some core limitations:
  – Pistoia Monomer Library: PEPTIDE1{[fmoc].A}$$$$
  – EBI ChEMBL Monomers: PEPTIDE1{[Fmoc_A]}$$$$
IUPAC CONDENSED NAMES (CHEMBL)

• The following names are machine generated

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

CHEMBL501567
CHEMBL500195
CHEMBL438006
CHEMBL500704
CHEMBL439086
CHEMBL507127
CHEMBL1181307
CHEMBL1791047
CHEMBL583516
CHEMBL1976073
PEPTIDE NAMES (CHEMBL)

• The following names are machine generated

  • [15-L-arginine]nociceptin CHEMBL526333
  • [2-4-chloro-L-phenylalanine]neuropeptide S [human] CHEMBL441576
  • [1-L-threonine]cyclosporin A CHEMBL2370014
  • [6-L-tryptophan]sermorelin free acid CHEMBL440438
  • angiotensin II (3-8) CHEMBL261120
  • nociceptin amide CHEMBL389521
  • acetyl-alpha-MSH (4-10) amide CHEMBL410411
  • [2-L-cysteine,13-L-cysteine]neurotensin disulfide CHEMBL3278512
  • myristoyl-[1-L-lysine,4-L-tryptophan]tetrapandin 2 amide CHEMBL3288219
  • [2-(4RS)-thiazolidine-4-carboxylic acid,4-L-proline]endomorphin-2 CHEMBL126611
  • [22-L-serine]kalata B1 CHEMBL1801140
Advanced Peptide Names

- Named peptides imply not only sequence but also N-terminal acetylation, C-terminal amidation and disulfide bridge topology.

- Example derivative naming operations:
  - gastrin (14-17)
  - motilin amide
  - oxytocin free-acid
  - acetyl-oxytocin
  - deacetyl-abarelix
  - oxytocin reduced
  - endothelin-1 (1→3),(11→15)-bis(disulfide)
HOMODETIC CYCLES #1

cyclo[Leu-D-Phe-Pro-Val-Orn-Leu-D-Phe-Pro-Val-Orn]

gramicidin S
Homodetic Cycles #2

cyclo[OAla-Val-D-OVal-D-Val-OAla-Val-D-OVal-D-Val-OAla-Val-D-OVal-D-Val]

valinomycin
AMBIGUOUS/PREFERRED FORMS

- [3-L-isoleucine]lypressin vs. [8-L-lysine]vasotocin
- [2-L-phenylalanine]lypressin vs. [8-L-lysine]phenypresin
- [2-L-phenylalanine]ornipressin vs. [8-L-ornithine]phenypressin
- [3-L-isoleucine]ornipressin vs. [8-L-ornithine]vasotocin
- [4-L-methionine]afamelanotide vs. [7-D-phenylalanine]α-MSH

- [Thr1,Lys2]endomorphin-1 vs. [Trp3,Phe4]tuftsin amide
- [Gln3]thyrotropin-releasing hormone vs. [Pro3]eisenin amide
- [Trp1,Val2]endomorphin-2 vs. [Val2,Phe3]gastrin tetrapeptide
NAMED CYCLIC PEPTIDE DERIVATIVES

- Mutants of named cyclic peptides are identified by comparing against all “rotational” permutations.

Example line notation query (CHEMBL478596)

cyclo[Ala-Gly-Thr-Phe-Val-Tyr]

Reference database line notations:

cyclo[Gly-Thr-Phe-Leu-Tyr-Thr]  dichotomin B

cyclo[Ala-Gly-Thr-Phe-Leu-Tyr]  dichotomin C

Resulting Sugar & Splice peptide name:

[5-L-valine]dichotomin C
LoweR LocaNTs IN CYClic PePTides

- Symmetric cyclic peptides provide an interesting challenge, where substitutions are different locants can potentially be synonymous.

- CHEMBL1934531
  - [3-(4S)-4-amino-L-proline]gramicidin S preferred
  - [8-(4S)-4-amino-L-proline]gramicidin S acceptable

- CHEMBL1934536
  - [3-(4R)-4-amino-L-proline]gramicidin S preferred
  - [8-(4R)-4-amino-L-proline]gramicidin S acceptable
SCALING-UP PROTEIN VARIANT NAMING

• The algorithm described for naming peptides can also be applied to naming arbitrary protein variants.

• Consider the a database of the following 11 peptides:
  – CFFQNCPRG  phenylpressin
  – CFVRNCPTG  annetocin
  – CFWTSCPIG  octopressin
  – CYFQNCPRG  argipressin
  – CYFQNCPKG  lypressin
  – CYFRNCPIG  cephalotocin
  – CYIQNCPLG  oxytocin
  – CYIQNCPPG  prol-oxytocin
  – CYIQNCPRG  vasotocin
  – CYIQSCPIG  seritocin
  – CYISNCPIG  isotocin
These 11 peptides may be efficiently represented and search as a “directed acyclic graph” [38 vs. 99 states]
Using this representation, all 540546 protein sequences in uniprot_sprot, which contains over 192M amino acids, requires 142M states (1.4Gb).

This data structure allows close analogues to be identified much faster than using NCBI blastp.

For example, all 540546 sequences can be queried against this database (i.e. all-against-all) in ~9m30s on a single core on a laptop.

The sequence from PDB 1CRN (crambin 46AA) is canonically named as [L25I]P01542 in 0.002s.
APPLICATION TO PRECISION MEDICINE

• A more realistic example is that sequence of the gene “spastic paraplegia4” with six mutations from OMIM:604277 can be canonically named as [I344K,S362C,N386S,D441G,C448Y,R499C]Q9UBP0

• Run-time for this query is 0.2s.

• By comparison, blastp 2.2.29+ takes about 6s.
  – With default arguments, NCBI blastp run time is 7s.
  – Only 6s with --num_descriptions 1 --num_alignments 1.
30 SECOND SUMMARY

• This talk describes the development of software for both naming peptides and reading peptide names matching the de facto standard practices currently followed by biochemists.

• Unlike computer representations, like Pisotoia HELM or InChI keys, these names and identifiers match those typically found in the scientific literature and vendor catalogues.

• A significant application of this technology is to check and correct peptide representations in databases.
ACKNOWLEDGEMENTS

• Joanna Sharman, IUPhar, Edinburgh University, UK.
• Lisa Sach-Peltason, Hoffmann-La Roche, Basel.
• Joann Prescott-Roy, Novartis, Boston, MA.
• Greg Landrum, Novatis, Basel, Switzerland.
• Evan Bolton, NCBI PubChem project, Bethesda, MD.
• Daniel Lowe, NextMove Software, Cambridge, UK.
• John May, NextMove Software, Cambridge, UK.