254th ACS National Meeting Washington Aug 2017

PubChem as a Biologics Database

Noel O'Boyle and Roger Sayle

NextMove Software

Evan Bolton

PubChem, NCBI-NIH

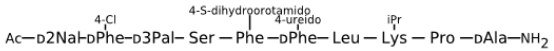
PUBCHEM INTERFACE

Degarelix | Download | Share | Help

Contents

3 Biologic Description

3.1 Biologic Depiction



from PubChem

3.2 Biologic Line Notation

IUPAC Condensed

Ac-D-2Nal-D-Phe(4-Cl)-D-3Pal-Ser-Phe(4-S-dihydroorotamido)-D-Phe(4-ureido)-Leu-Lys(iPr)-Pro-D-Ala-NH2

from PubChem

Sequence

XXXSXXLXPA

from PubChem

https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72

PubChem Classification Browser

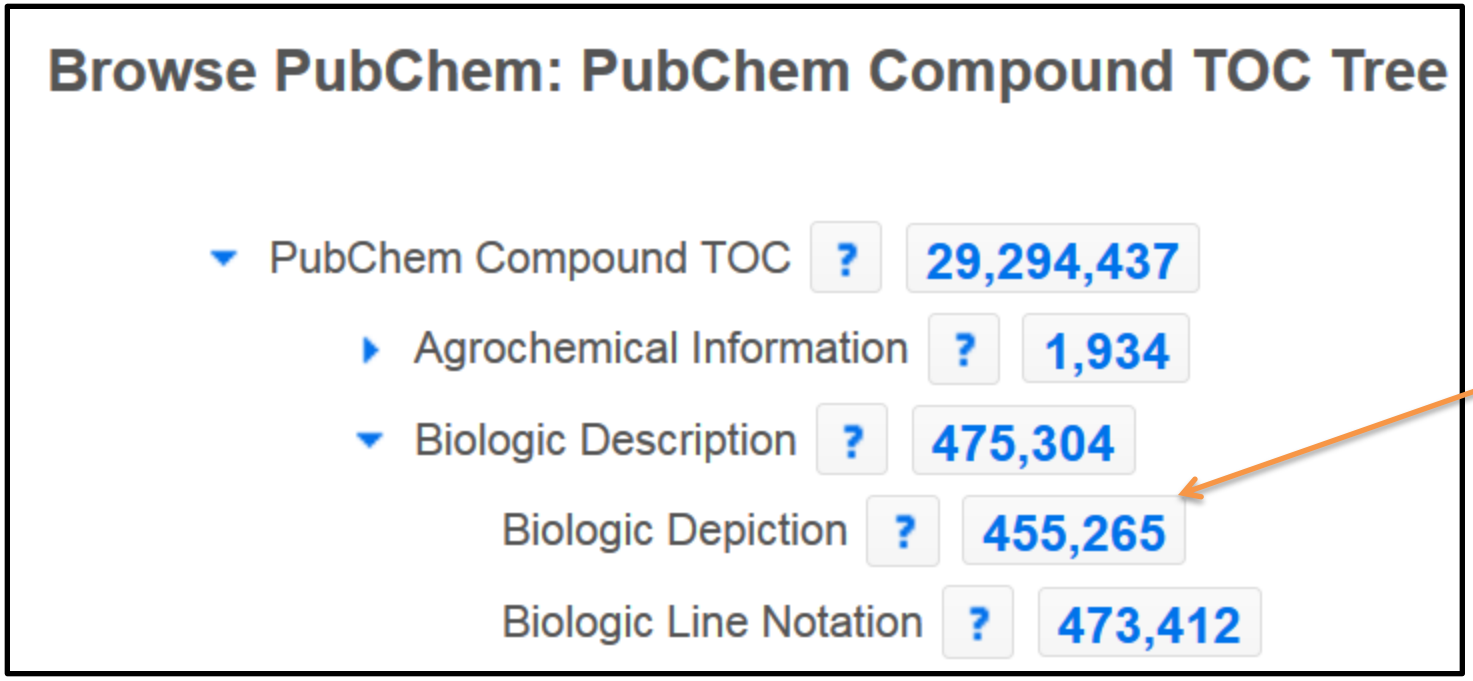
Browse PubChem data using a classification of interest, or search for PubChem records a
Gene Ontology: DNA repair). More...

Select classification

PubChem: PubChem Compound TOC

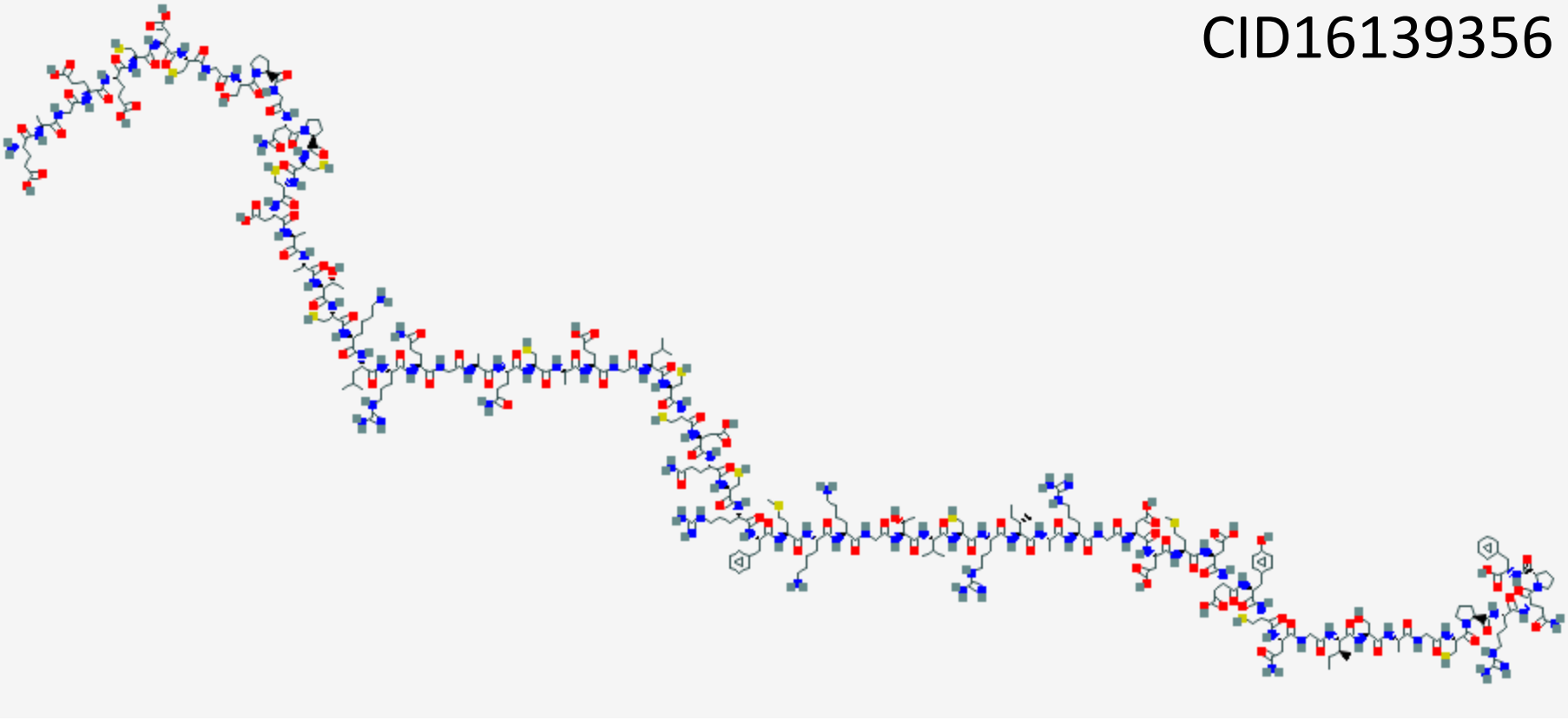
Search selected classificat

Keyword

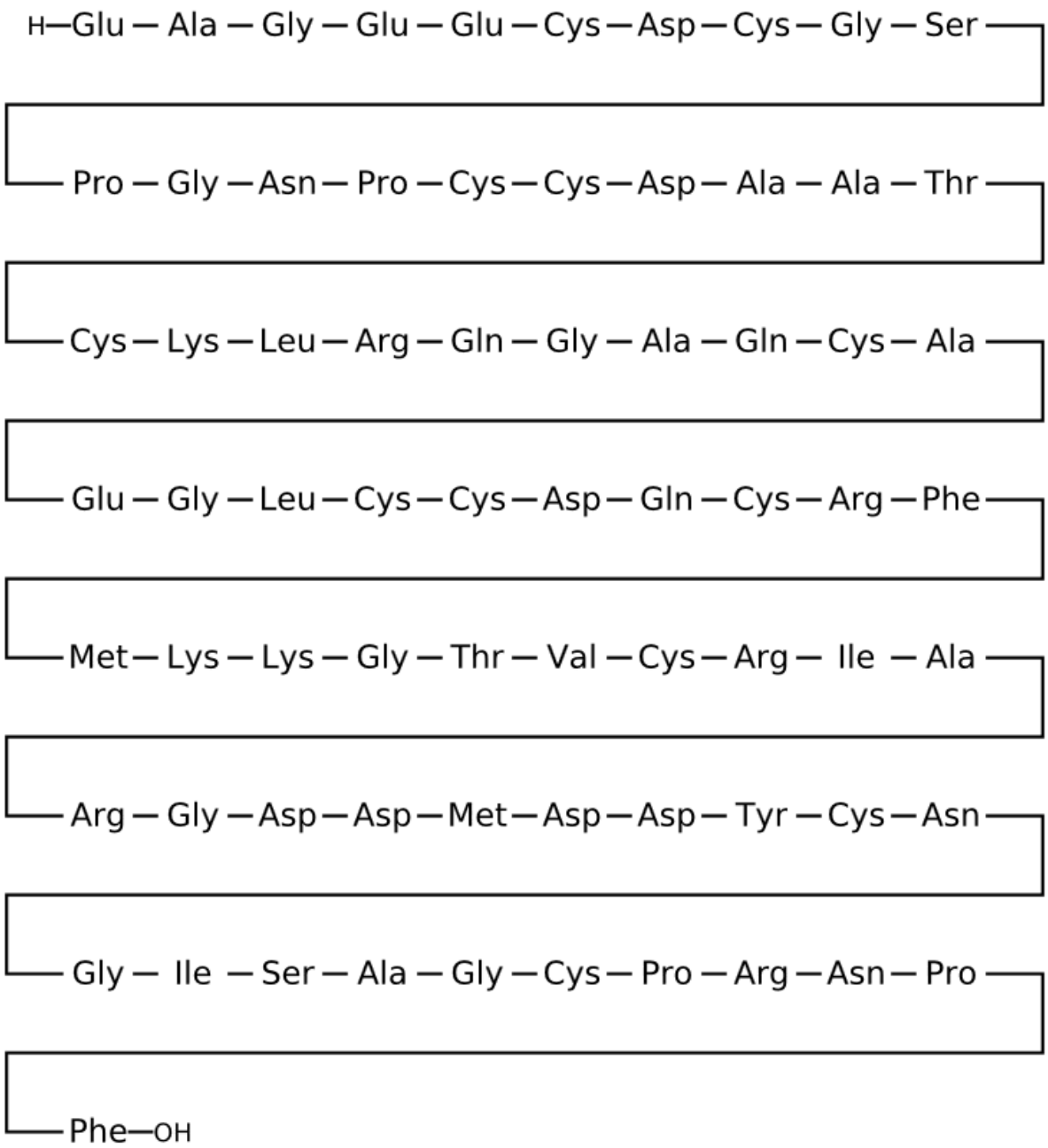


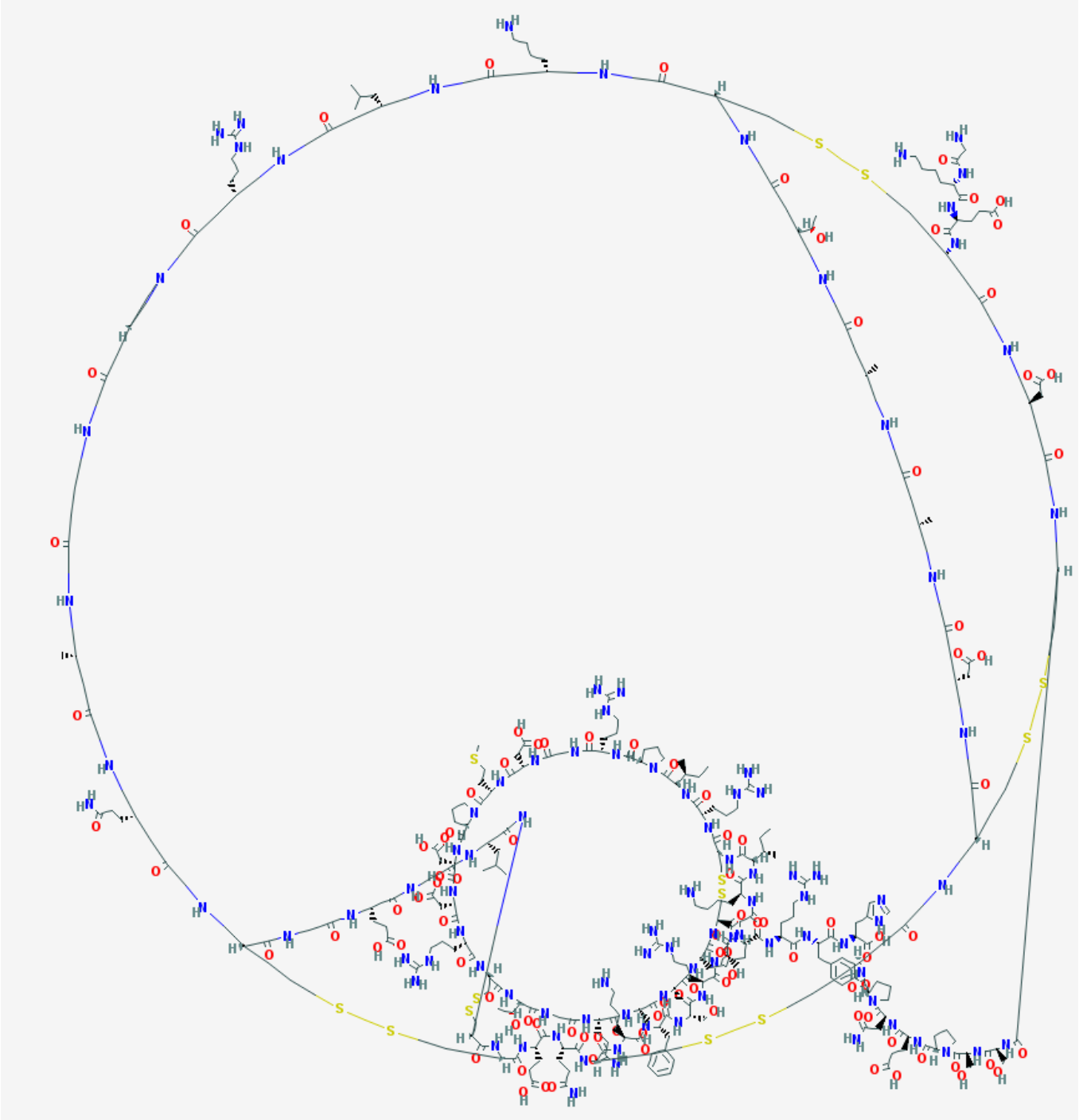
Click and sort by MW

CID16139356

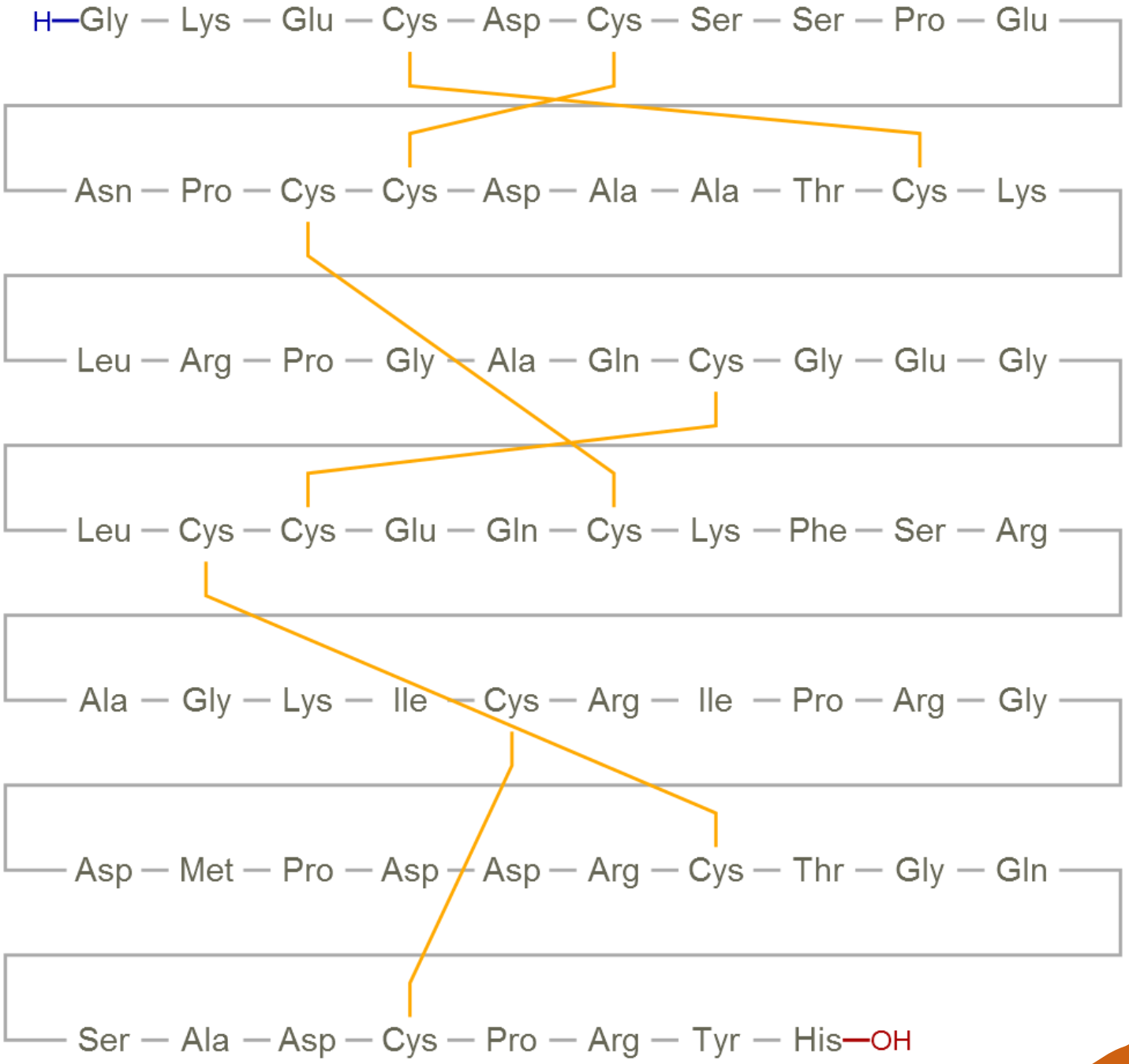


CID16139356

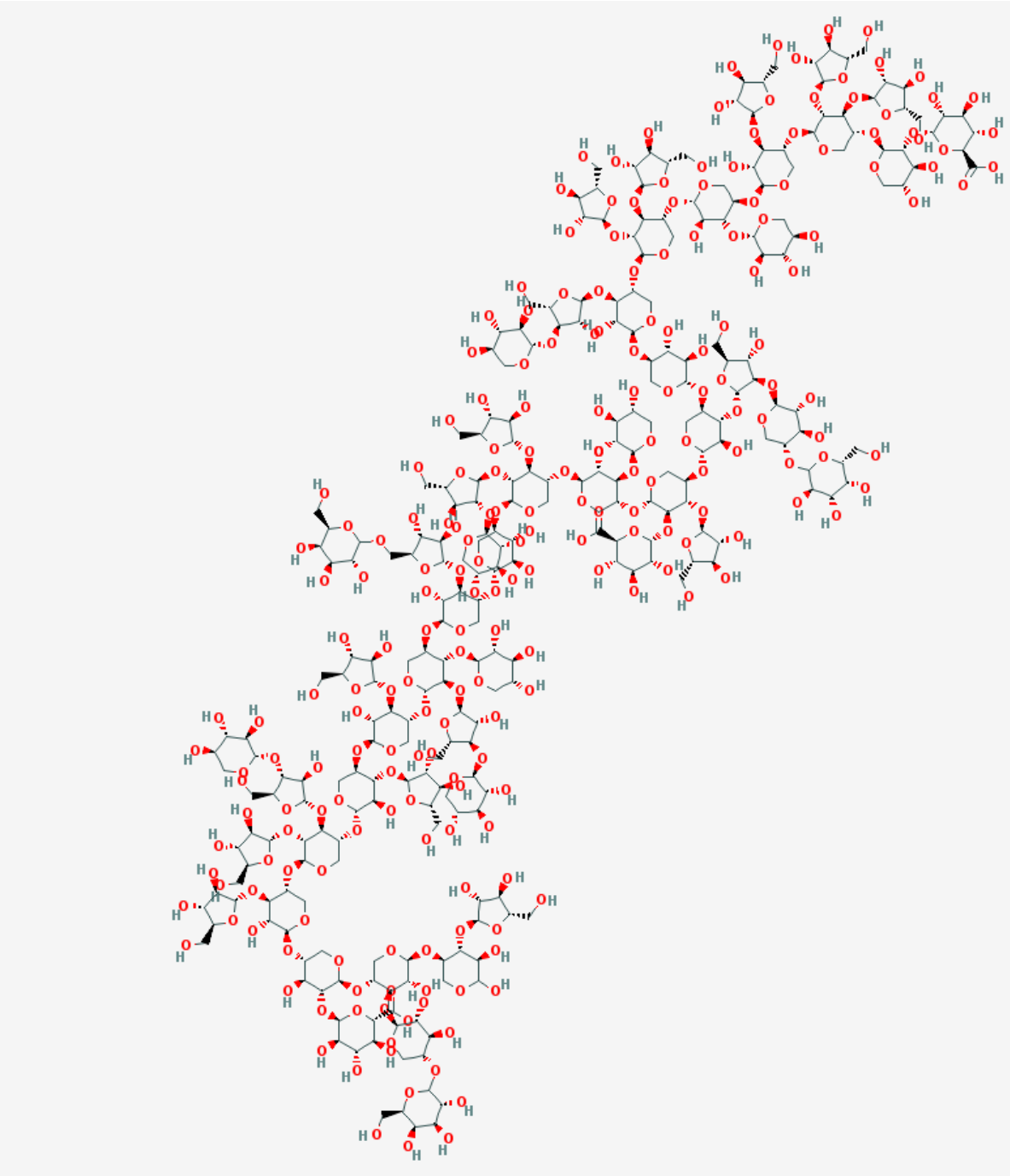




CID56842075
Rhodostomin

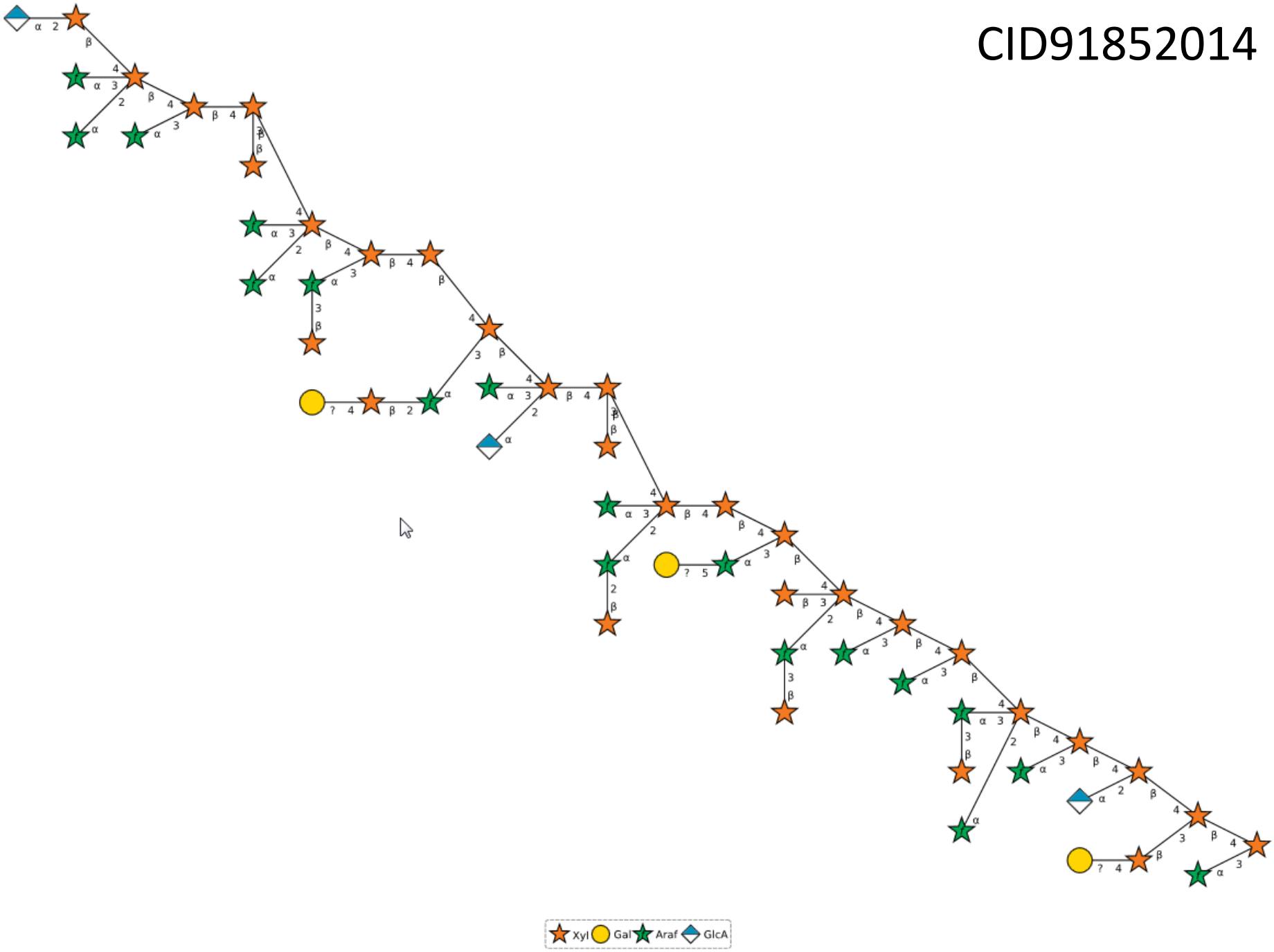


CID56842075
Rhodostomin



CID91852014

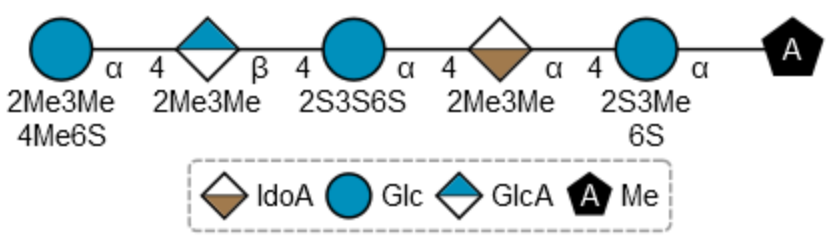
CID91852014

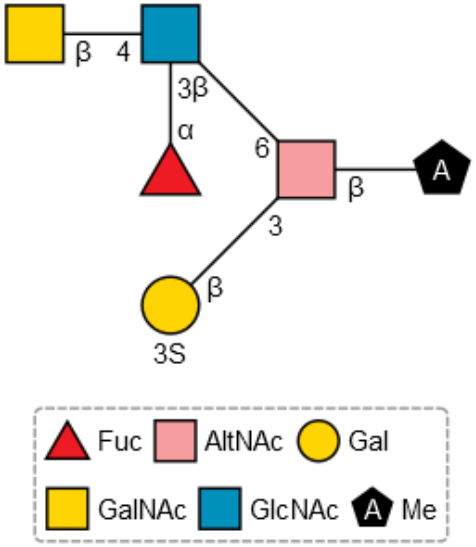


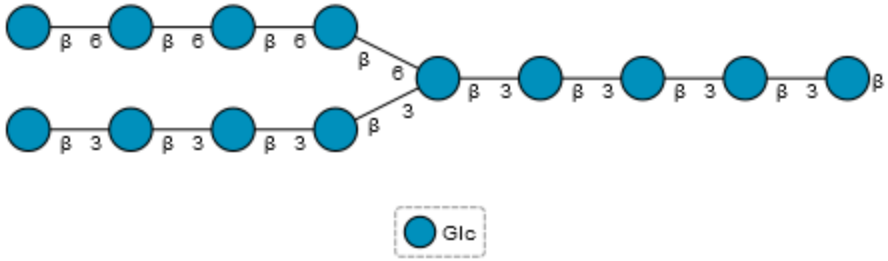
PUBCHEM – A SMALL MOLECULE DATABASE?

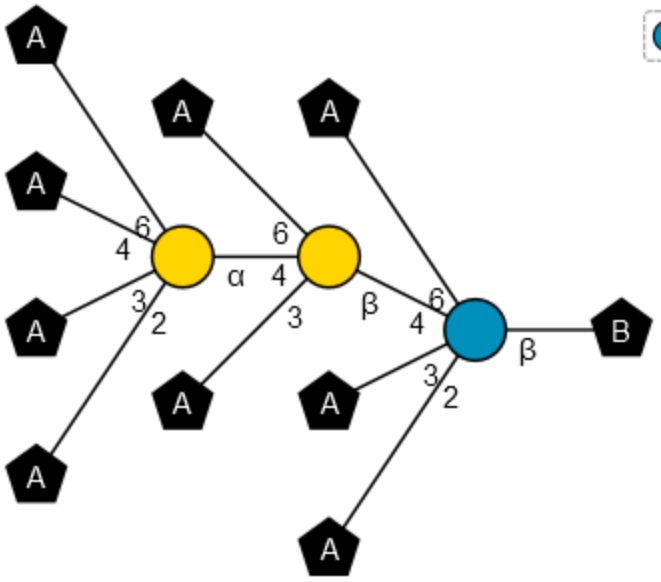
- People don't think of PubChem as a peptide database
  - ~110K X-rays of proteins in PDB
  - ~500K peptides in PubChem
- People don't think of PubChem as a saccharide database
  - ~80K oligosaccharides in GlyTouCan
  - ~67K oligosaccharides in PubChem

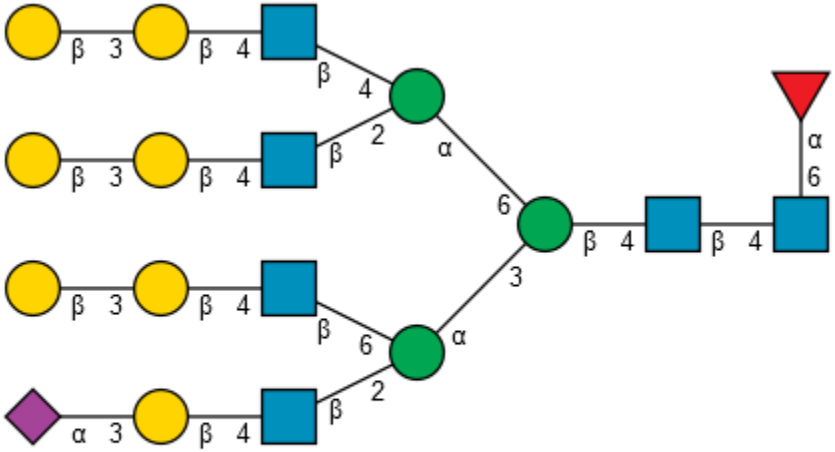
531,618 contain a monosaccharide, of which 66,740 can be depicted











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**

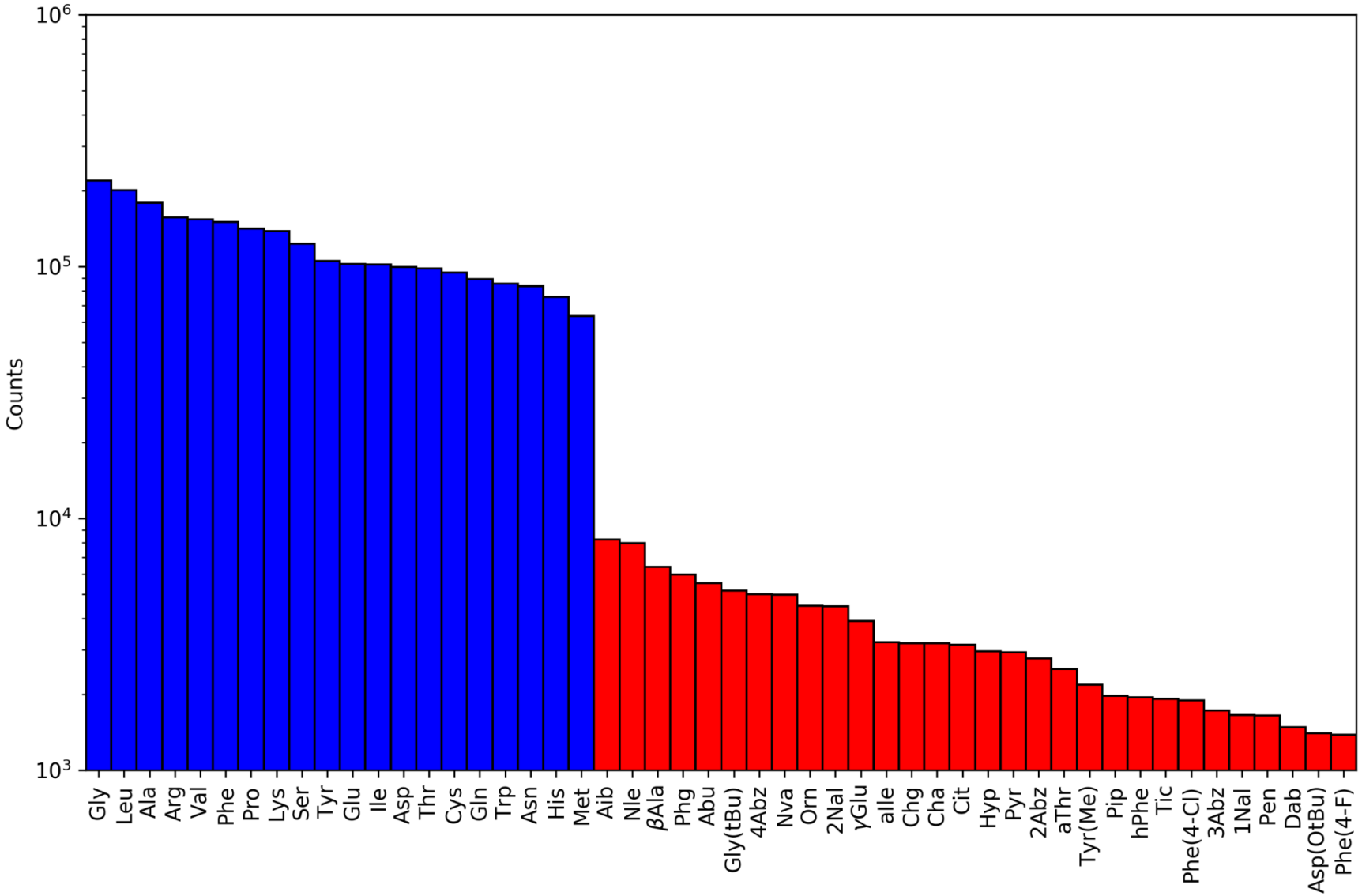
Xylf5Me, a-L-ManNAc3Ac4Ac6Ac, Glc2P3P6P

⇩ **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

AMINO ACIDS IN PUBCHEM STRUCTURES CONTAINING AT LEAST THREE AMINO ACIDS



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

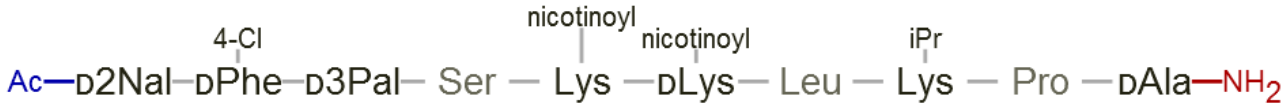
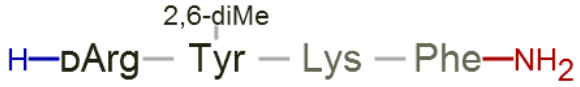
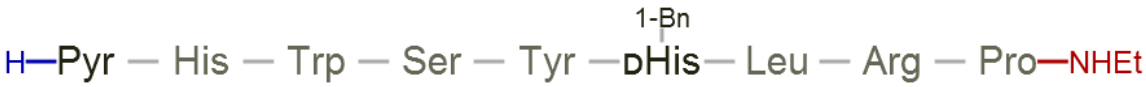
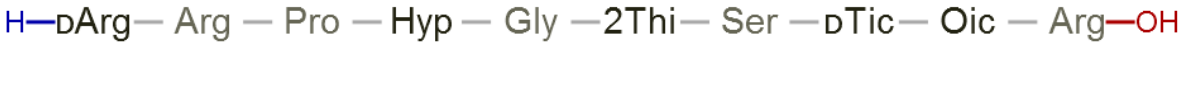
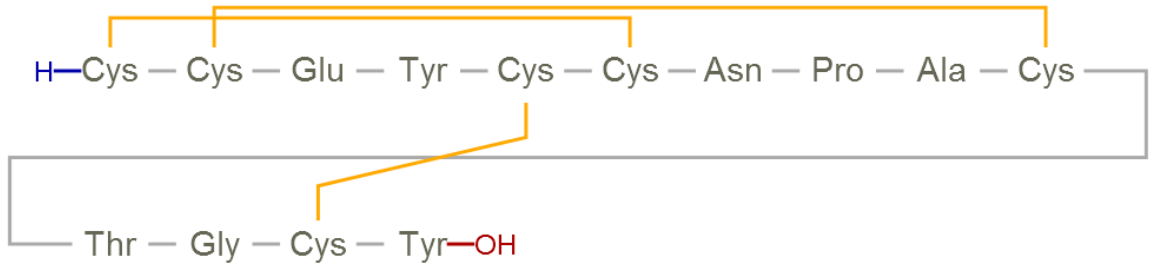
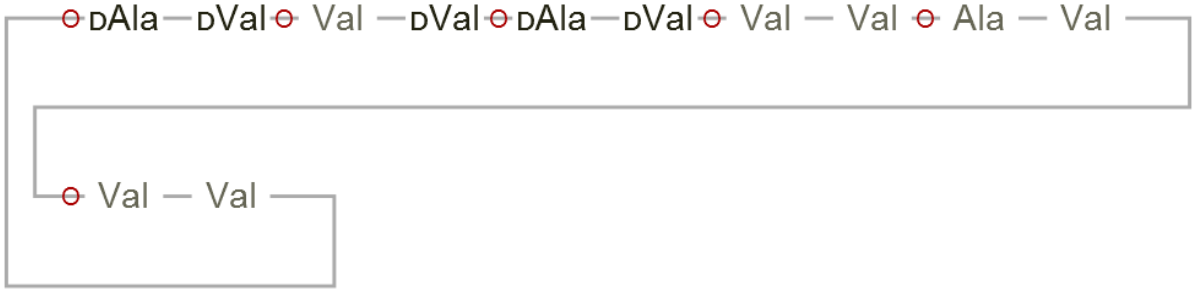
⇩

8125 including N-substituted variants

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

HOW MANY PEPTIDES PRESENT?

- Depends how you count…
- 447,026 have 3 or more amino acids
- 668,229 structures are recognised in their entirety as peptides

iturelix CID16130938	
elamipretide CID11764719	
histrelin CID25077993	
icatibant CID71364	
linaclotide CID16158208	
valinomycin CID5649	

SEQUENCE REPRESENTATION

- Depending on the task or quality of datasource, different sequence representations may be preferred
  - Distinguish between D-/L-/DL- amino acids using upper/lowercase?
    - D-Ala as `a` or `A`
  - Distinguish between sidechain stereo variants?
    - alloThr as `X` or `T`
  - Distinguish between substituted amino acids and their parent?
    - Ser(PO3H2) as `X` or `S`

EXACT SEQUENCE SEARCH

- Given that features of the structure are normalized or ignored
  - Exact sequence search can be used to find similar structures (sequence as hash)
- Create hierarchy of similarity
  - First, those structures with the same sequence, if we normalise as much as possible
  - Then successively discriminate based on stereochemistry, side-chain substitution

SEARCH FOR KEMPTIDE: LRRASLG

LRRASLG		
LRRASLG		
Ac-DL-Leu-DL-Arg-DL-Arg-DL-Ala-DL-Ser-DL-Leu-Gly-OH	78069426	acetyl-kemptide
Ac-Leu-Arg-Arg-Ala-Ser-Leu-Gly-OH	71429096	acetyl-kemptide
H-DL-Leu-DL-Arg-DL-Arg-DL-Ala-DL-Ser-DL-Leu-Gly-NH2	85062657	kemptide amide
H-DL-Leu-DL-Arg-DL-Arg-DL-Ala-DL-Ser-DL-Leu-Gly-OH	100074	kemptide
H-DL-Leu-DL-Arg-DL-Arg-DL-Ala-DL-Ser-DL-Leu-Gly-OH.TFA	118797564	
H-Leu-Arg-Arg-Ala-Ser-Leu-Gly-NH2	9897033	kemptide amide
H-Leu-Arg-Arg-Ala-Ser-Leu-Gly-OH	9962276	kemptide
Unk-Leu-Arg-Arg-Ala-Ser-Leu-Gly-OH	11650926,101224399,101878757	
LRRAXLG		
H-Leu-Arg-Arg-Ala-Ser(PO3H2)-Leu-Gly-NH2	102212089	[Ser(PO3H2)-5]kemptide amide
H-Leu-Arg-Arg-Ala-Ser(PO3H2)-Leu-Gly-OH	13783725	[Ser(PO3H2)-5]kemptide
LRraSLG		
LRraSLG		
H-Leu-Arg-D-Arg-D-Ala-Ser-Leu-Gly-OH	53393688	[D-Arg3,D-Ala4]kemptide
LrRASLG		
LrRASLG		
H-Leu-D-Arg-Arg-Ala-Ser-Leu-Gly-OH	99864041	[D-Arg2]kemptide
lRRASLG		
lRRASLG		
H-D-Leu-Arg-Arg-Ala-Ser-Leu-Gly-OH	99864040	[D-Leu1]kemptide
lrRASLG		
lrRASLG		
H-D-Leu-D-Arg-Arg-Ala-Ser-Leu-Gly-OH	99864042	[D-Leu1,D-Arg2]kemptide

DISULFIDE BRIDGING PATTERNS

- Use a sequence representation to find peptides with different disulfide bridges
  - Does not occur naturally
  - Either errors by depositor, or artificially created
- Convert peptides with at least four cysteines to sequence format and collate
  - 16 cases found with different bridges
  - 12 were erroneous, 4 real

DISULFIDE BRIDGING PATTERNS

ICCNPACGPKYSC

CID11480353

CID101041637

CHEMICAL BIOLOGY & DRUG DESIGN

Explore this journal >

Controlled syntheses of natural and disulfide-mispaired regioisomers of α-conotoxin SI†

B. Hargittai, G. Barany

First published: December 1999 Full publication history

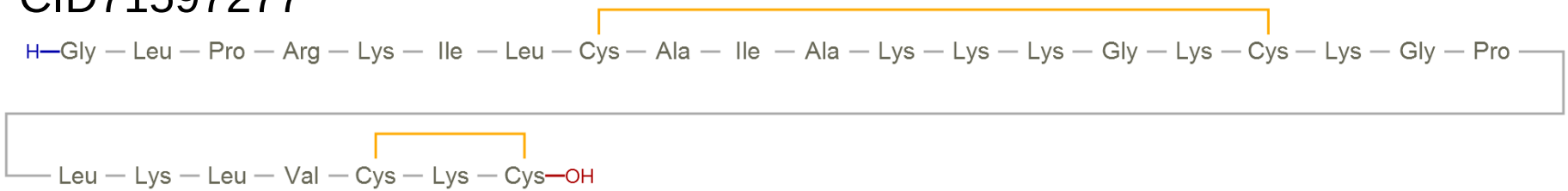
DOI: 10.1034/j.1399-3011.1999.00127.x View/save citation

Deposited by Nikajii

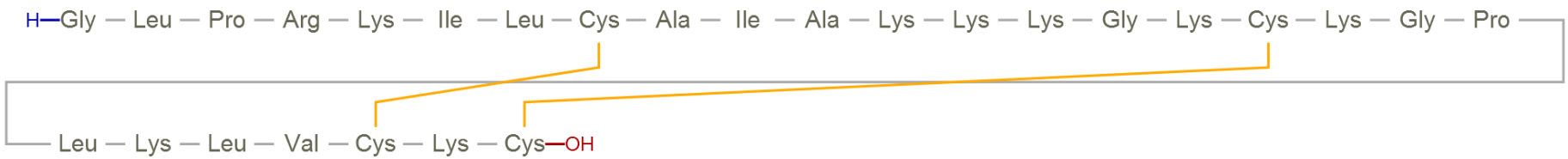
DISULFIDE BRIDGING PATTERNS

GLPRKIL**C**AIAKKKGK**C**KGPLKLV**C**K**C**

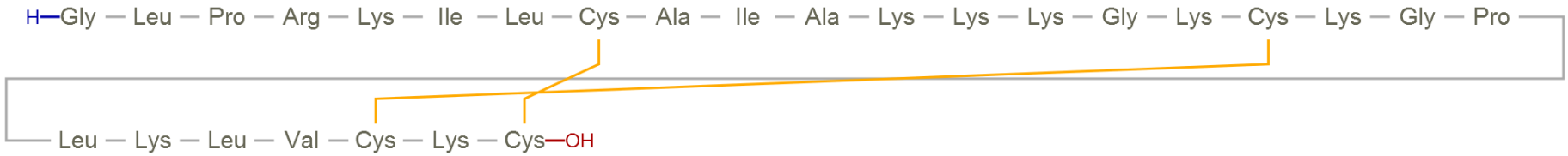
CID71597277



CID71597445



CID71597707

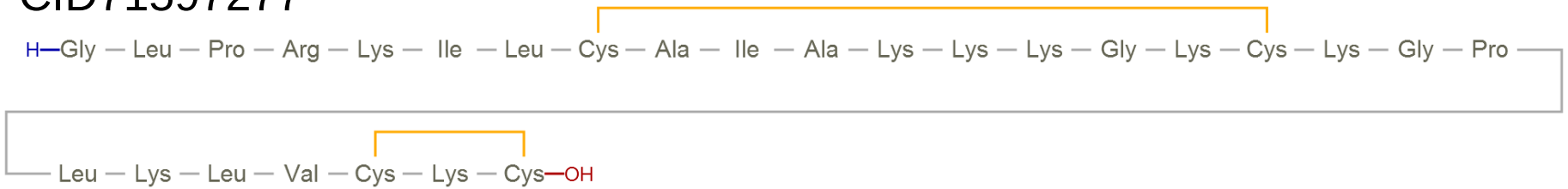


Deposited by NIAID (National Institute of Allergies and Infectious Diseases)

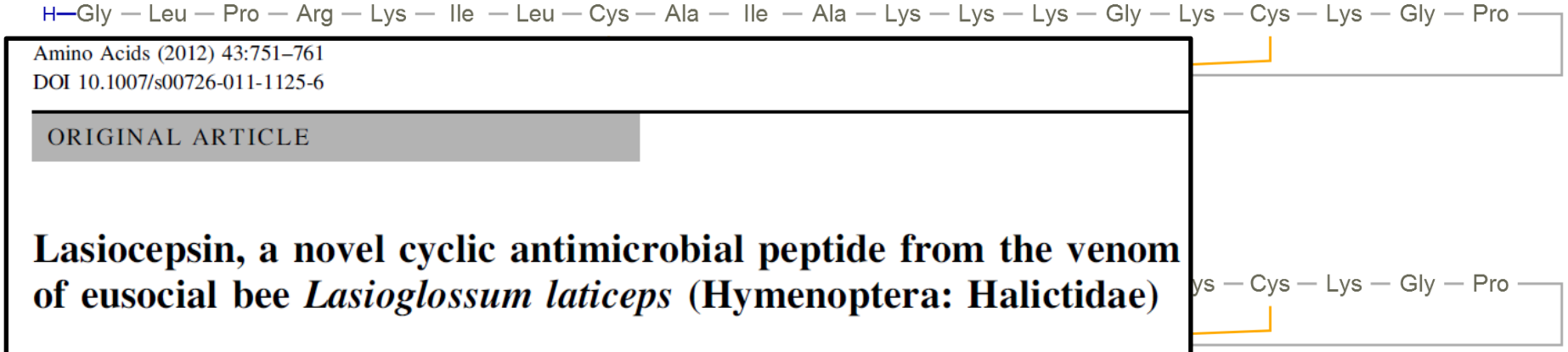
DISULFIDE BRIDGING PATTERNS

GLPRKIL**C**AIAKKKGK**C**KGPLKLV**C**K**C**

CID71597277



CID71597445



Amino Acids (2012) 43:751–761
DOI 10.1007/s00726-011-1125-6

ORIGINAL ARTICLE

Lasiocepsin, a novel cyclic antimicrobial peptide from the venom of eusocial bee *Lasioglossum laticeps* (Hymenoptera: Halictidae)

Lenka Monincová · Jiřina Slaninová · Vladimír Fučík · Oldřich Hovorka · Zdeněk Voburka · Lucie Bednárová · Petr Maloň · Jitka Štokrová · Václav Čeřovský

Deposited by NIAID (National Institute of Allergies and Infectious Diseases)

KNOTTINS

- Peptides with three disulfide bridges, where one threads through the macrocycle formed by the others (see KNOTTIN Database http://knottin.cbs.cnrs.fr)
- Interesting leads for drug discovery
  - Stable fold, sequence tolerant, small

μ-conotoxin KIIIA (CID73350610)

- A necessary (but not sufficient) condition is an arrangement of Cys bridges 123123
  - 90 examples in PubChem

SEQUENCE VARIATION

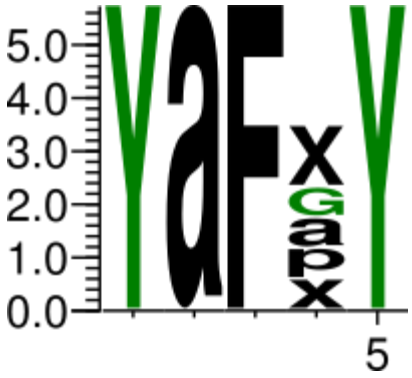
- Which parts of a sequence have seen the most variants?
  - Of interest for drug discovery, activity modulation

- Looked for variants of a sequence that are one substitution away from a known peptide
  - Required strict matching of the sequence to minimize ‘mutations’ due to errors

PubChem

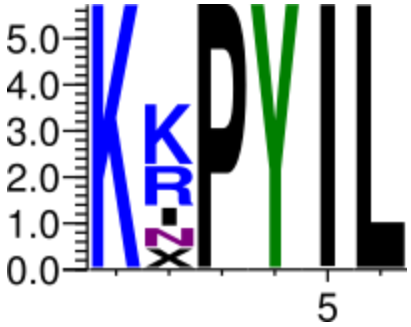
Casokefamide

YaFaY



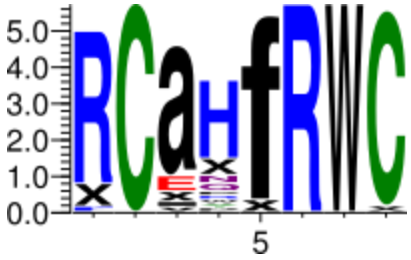
Neuromedin N

KIPYIL



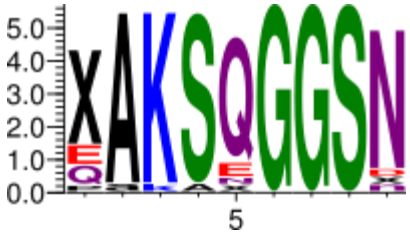
Setmelanotide

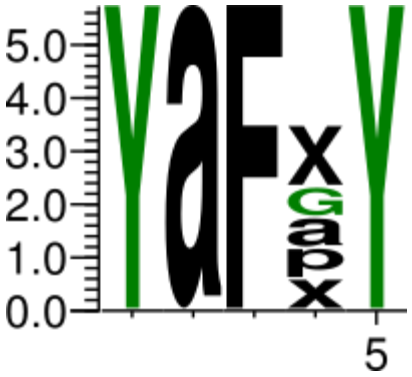
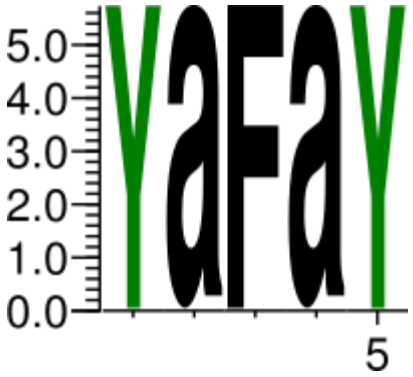
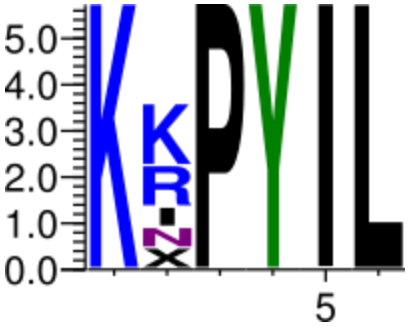
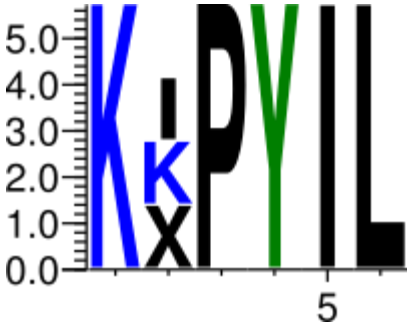
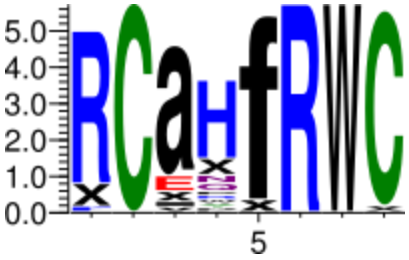
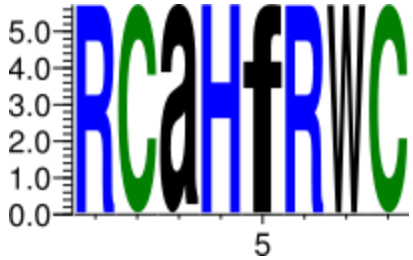
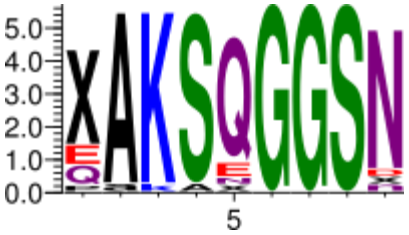
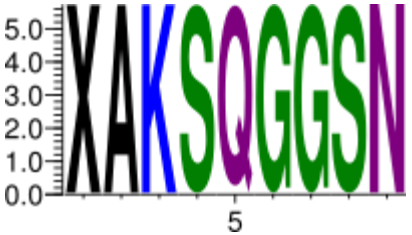
RCaHfRWC



Thymulin

XAKSQGGSN



	PubChem	*ChEMBL*
Casokefamide YaFaY		
Neuromedin N KIPYIL		
Setmelanotide RCaHfRWC		
Thymulin XAKSQGGSN		

ARE PUBCHEM PEPTIDES VARIANTS OF KNOWN PEPTIDES?

- Hypothesis: observed peptides are close variants of a small number of known peptides
- 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)
  - Cbz-cholecystokinin octapeptide (2-7) amide
  - [Ile1,Ser2,Ser8]cyphokinin

SUMMARY

- PubChem is a rich source of information on oligopeptides and oligosaccharides
  - Often heavily modified, rather than natural
- Due to chemical modifications, we need to think in terms of 10s of thousands of monomers
- Sequence representations act as a key:
  - To collate similar peptides
  - To find sequence variants, sites of variation
  - To find disulfide bridge variants, knottins

ACKNOWLEDGEMENTS

- Paul Thiessen PubChem

SUGAR & SPLICE

http://nextmovesoftware.com

http://nextmovesoftware.com/blog

noel@nextmovesoftware.com

Rank Sales 2016	Trade Name	Name	Type of biologic
1	Humira	adalimumab	**Monoclonal antibody**
2	Harvoni	ledipasvir/sofosbuvir	
3	Enbrel	etanercept	**Protein attached to monoclonal antibody**
4	Rituxan	rituximab	**Monoclonal antibody**
5	Remicade	infliximab	**Monoclonal antibody**
6	Revlimid	lenalidomide	
7	Avastin	bevacizumab	**Monoclonal antibody**
8	Herceptin	trastuzumab	**Monoclonal antibody**
9	Lantus	insulin glargine	**Protein**
10	Prevnar 13	Pneumococcal vaccine	**Polysaccharides attached to carrier protein**
11	Xarelto	rivaroxaban	
12	Eylea	aflibercept	**Protein attached to monoclonal antibody**
13	Lyrica	pregabalin	
14	Neulasta	pegfilgrastim	**PEG attached to protein**
15	Advair Diskus	fluticasone/salmeterol	

Source: Genetic Engineering & Biotechnology News
http://www.genengnews.com/the-lists/the-top-15-best-selling-drugs-of-2016/77900868

REPRESENTATIONS AT VARIOUS LEVELS

- All-atom
- Hydrogen-suppressed graphs
  - How to handle non-standard valencies, cycles
- Consider monomer as superatom
  - How to handle cycles, monomer variants such as modifications and opposite stereo

MOST COMMON UNRECOGNISED SUBSTITUENTS ON SUGARS

- Increased number of perceived oligosaccharides from 51,273 to 66,740