245th ACS National Meeting, New Orleans, 9th April 2013

ROUNDTRIPPING BETWEEN SMALL-MOLECULE AND BIOPOLYMER REPRESENTATIONS

Noel O'Boyle and Roger Sayle

NextMove Software

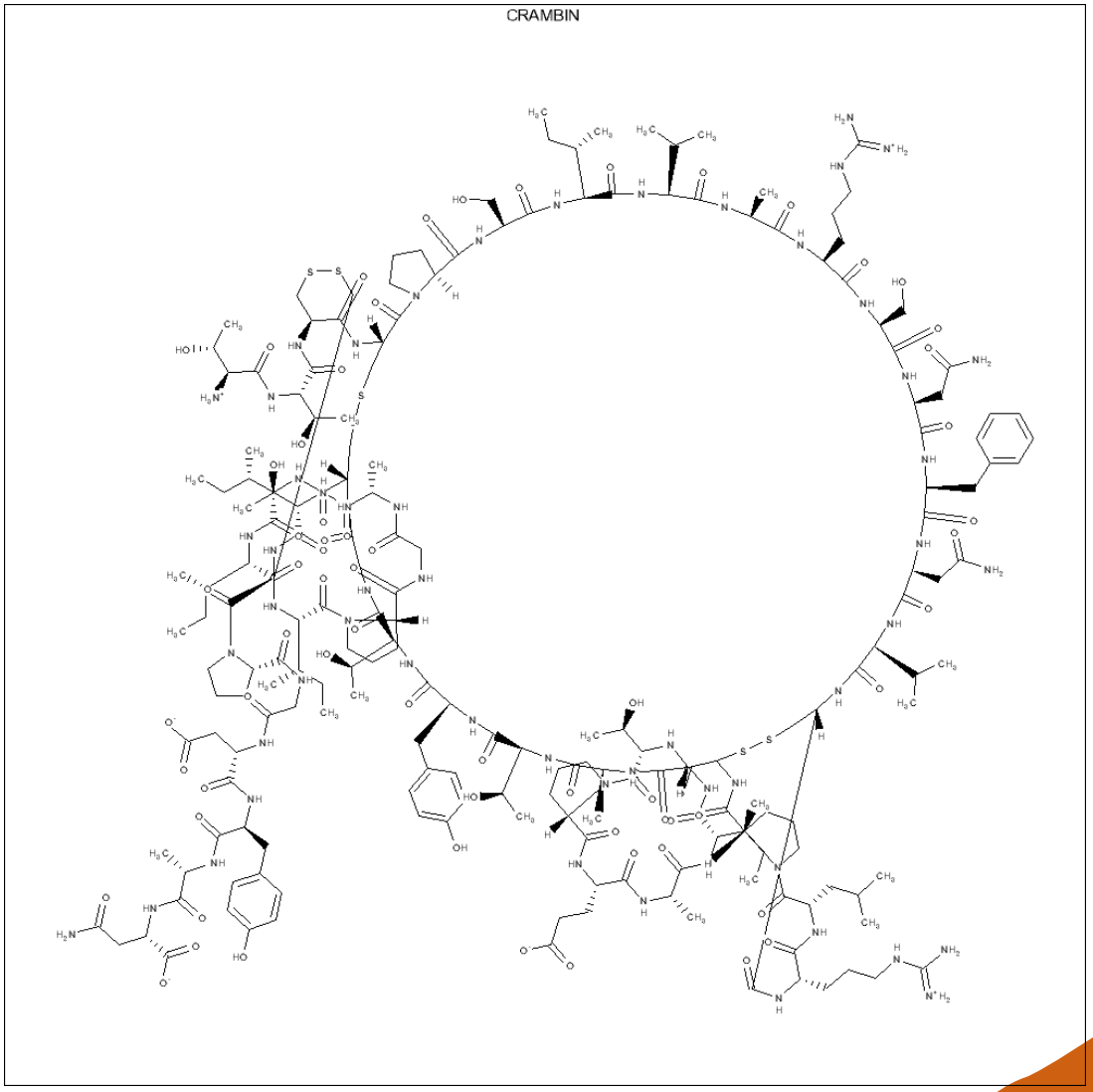
Evan Bolton

PubChem

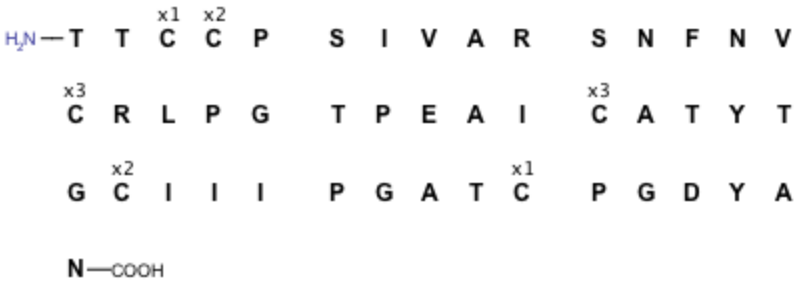
Rank Sales Q4 2012	Trade Name	Name	Type of biologic
1	Abilify	aripiprazole	
2	Nexium	esomeprazole	
3	Crestor	rosuvastatin	
4	Cymbalta	duloxetine	
5	Humira	adalimumab	**Monoclonal antibody**
6	Advair Diskus	fluticasone/salmeterol	
7	Enbrel	etanercept	**Protein attached to monoclonal antibody**
8	Remicade	infliximab	**Monoclonal antibody**
9	Copxone	glatiramer acetate	**Peptide**
10	Neulasta	pegfilgrastim	**PEG attached to protein**
11	Rituxan	rituximab	**Monoclonal antibody**
12	Spiriva	tiotropium bromide	
13	Atripla	emtricitabine/tenofovir/efavirenz	
14	OxyContin	oxycodone	
15	Januvia	sitagliptin	

Source: Drugs.com Statistics, Q4 2012
(http://www.drugs.com/stats/top100/2012/q4/sales)

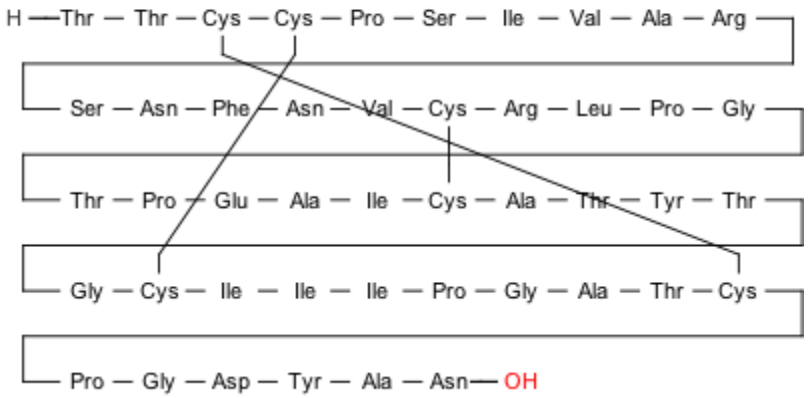
A PICTURE IS WORTH A THOUSAND WORDS...?



...BUT IT DEPENDS ON THE PICTURE



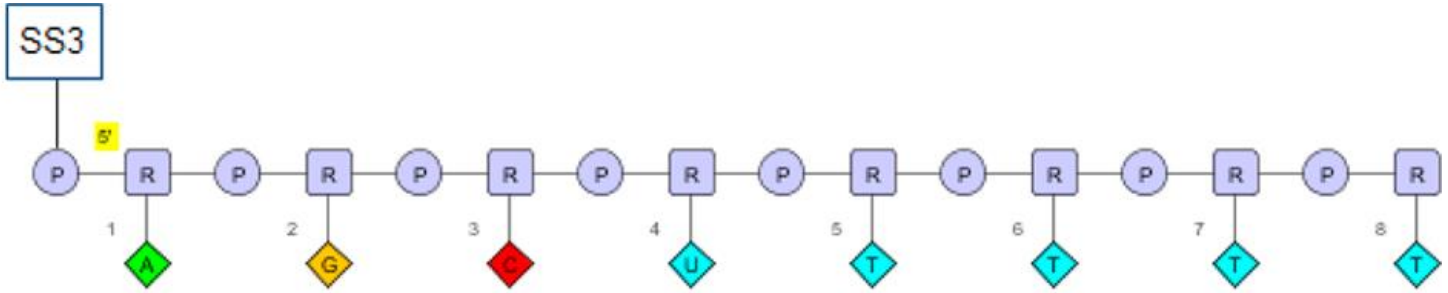
FDA



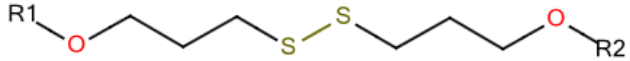
IUPAC

IS IT A FILE FORMAT PROBLEM?

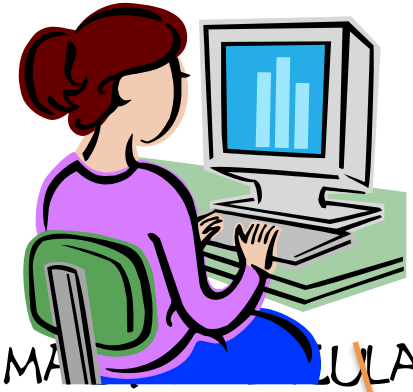
- HELM (Hierarchical editing language for macromolecules)
  - Developed at Pfizer, and promoted by Pistoia Alliance
  - Hierarchical description of monomers and oligomers and connections between them



RNA1{P.R(A)P.R(G)P.R(C)P.R(U)P.R(T)P.R(T)P.R(T)}|CHEM1{SS3}
$RNA1,CHEM1,1:R1-1:R1$$$

ID	Name	Structure	Attachment Points
SS3	Dipropanol Disulfide		R1-H R2-H

Zhang et al. J. Chem. Inf. Model. 2012, 52, 2796.



PFIZER MA[illegible]LAR EDITOR

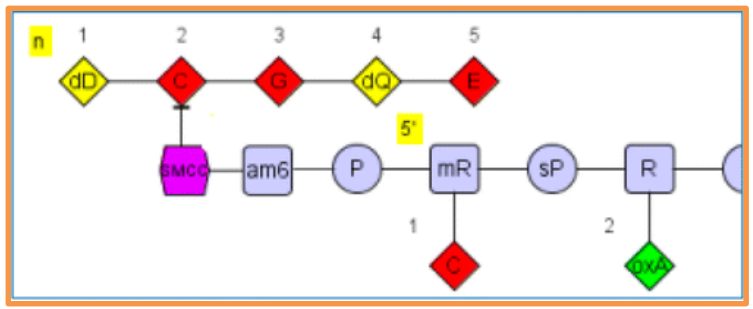
HELM

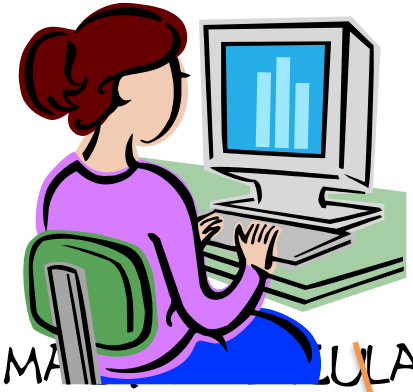
INCHI

SMILES

C[C@H](O1)[C@@H](O)[C@@H](O)[C@H](O)[C@@H]1(O2).C(O)[C@@H](O1)[C@H](O)[C@H](O)...

DEPICTION





DATABASE OF MONOMER DEFINITIONS

PFIZER MA[illegible]LAR EDITOR

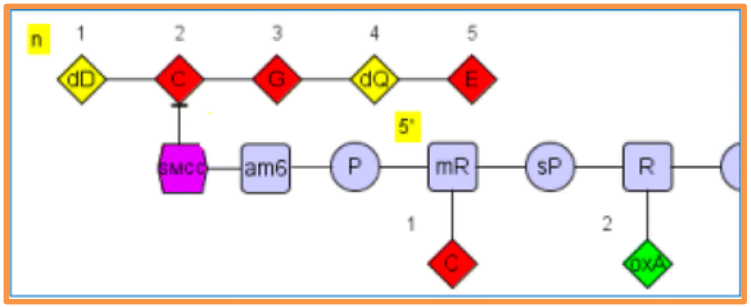
HELM

INCHI

SMILES

C[C@H](O1)[C@@H](O)[C@@H](O)[C@H](O)[C@@H]1(O2).C(O)[C@@H](O1)[C@H](O)[C@H](O)...

DEPICTION



IS IT A FILE FORMAT PROBLEM?

- PLN (Protein line notation)
  - Developed by Biochemfusion
  - Also software to edit and depict
  - Adheres closely to IUPAC recommendations and extended to handle cycles and arbitrary modified residues

H-ASDF-OH.H-CGTY-OH id=P0001 **

- SCSR (Self-contained sequence representation)
  - Developed as part of Accelrys Draw
  - Extension of Mol V3000
  - Can convert between all-atom representation and condensed forms

...OR IS IT A PERCEPTION PROBLEM?

- Existing all-atom representations are fully capable of storing biopolymers
  - PDB files have been used to store biopolymers for some time
  - SMILES and MOL files can store macromolecules
- On-the-fly perception from connection table
  - Perception of biopolymer units and the nature of the connections between those
- Convert perceived structure to any of several all-atom or biopolymer representations

PERCEPTION ENGINE

PDB

MOL

SMILES

IUPAC CONDENSED

HELM

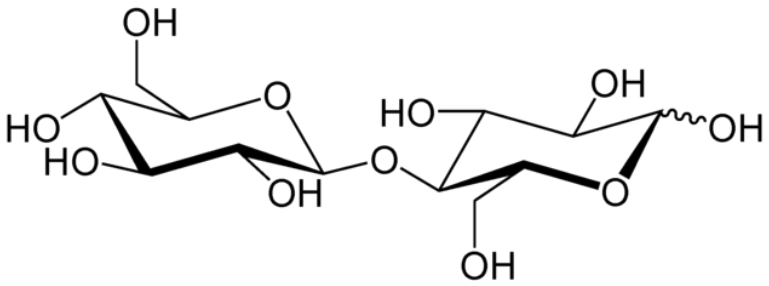
PLN

IUPAC NAME

DEPICTION

SMILES

[C@@H]1([C@H](O)[C@@H](O)[C@H](O)[C@H](O1)CO)O[C@H]1[C@@H]([C@H](C(O)O[C@@H]1CO)O)O



SUGAR & SPLICE

COMMON NAME

cellobiose

IUPAC CONDENSED

Glc(b1-4)Glc

IUPAC NAME

beta-D-gluco-hexopyranosyl-(1->4)-D-gluco-hexopyranose

LINUCS

[][D-Glcp]{
[(4+1)][b-D-Glcp]{}
}

PDB

MARK BOTH AS BGC
ATOMS AS C1-C6, O1-O6

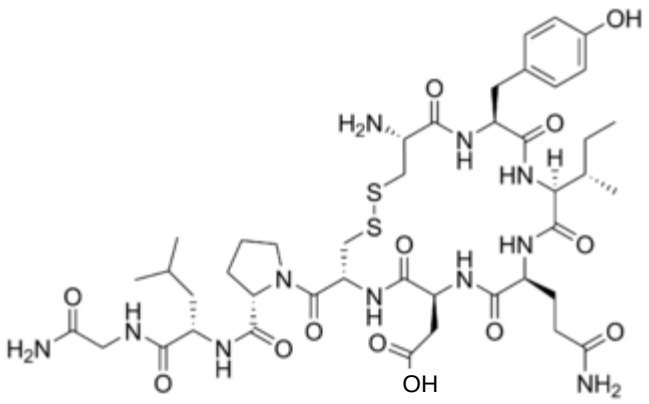
DEPICTION

COMING SOON...

SUGAR & SPLICE

SMILES

```
[C@H]1(CCCN1C(=O)[C@@H]1CSSC[C@@H](C(=O)N[C@@H](Cc2ccc(cc2)O)C(=O)N[C@@H]([C@H](CC)C)C(=O)N[C@@H](CCC(=O)N)C(=O)N[C@@H](CC(=O)O)C(=O)N1)N)C(=O)N[C@@H](CC(C)C)C(=O)NCC(=O)N
```



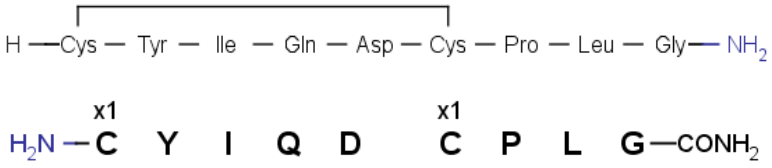
PLN

H-C(1)YIQDC(1)PLG-[NH2]

COMMON NAME

[5-L-aspartic acid]oxytocin

DEPICTIONS



PDB

L-Cys(1)-L-Tyr-L-Ile-L-Gln-L-Asp-L-Cys(1)-L-Pro-L-Leu-Gly-NH2

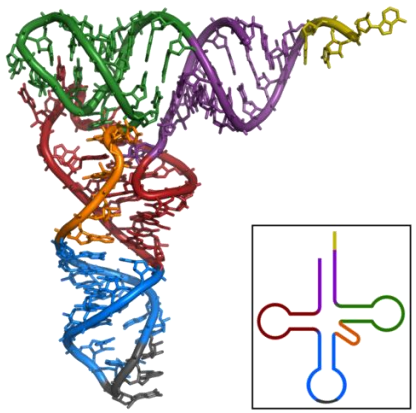
IUPAC CONDENSED

IUPAC NAME

L-cysteinyl-L-tyrosyl-L-isoleucyl-L-glutaminyl-L-alpha-aspartyl-L-cysteinyl-L-prolyl-L-leucyl-glycinamide (1->6)-disulfide

Cc1cn(c(=O)[nH]c1=O)[C@H]2[C@@H]([C@@H]([C@H](O2)CO[P@@](=O)([O-])O[C@@H]3[C@H](O[C@H]([C@@H]3O)n4cnc5c4nc([nH]c5=O)N)CO[P@@](=O)([O-])O[C@@H]6[C@H](O[C@H]([C@@H]6O)n7ccc(=O)[nH]c7=O)CO[P@@](=O)([O-])O[C@@H]8[C@H](O[C@H]([C@@H]8O)n9cnc1c9nc([nH]c1=O)N)CO[P@@](=O)([O-])O[C@@H]1[C@H]… **(4984 characters)**

SMILES



tRNA(Phe)

SUGAR & SPLICE

PDB

IUPAC NAME

…-5'-cytidylyl)-5-methyl-5'-cytidylyl)-5'-uridylyl)-5'-guanylyl)-5'-uridylyl)-5'-guanylyl)-5-methyl-5'-uridylyl)-1-uracil-5-yl-1-deoxy-beta-D-ribofuranos-O5-yl)(hydroxy)phosphoryl)-5'-cytidylyl)-5'-guanylyl)-1-(6-imino-1-methyl-purin-9-yl)-1-deoxy-beta-D-ribofuranos-O5-yl)(hydroxy)phosphoryl)-5'-uridylyl)-5'-cytidylyl)-5'-cytidylyl…

IUPAC CONDENSED

P-rGuo-P-rCyd-P-rGuo-P-rGuo-P-rAdo-P-rUrd-P-rUrd-P-rUrd-P-rAdo-P-m2Gua-Rib-P-rCyd-P-rUrd-P-rCyd-P-rAdo-P-rGuo-P-rUrd-P-rUrd-P-rGuo-P-rGuo-P-rGuo-P-rAdo-P-rGuo-P-rAdo-P-rGuo-P-rCyd-P-m22Gua-Rib-P-rCyd-P-rCyd-P-rAdo-P-rGuo-P-rAdo-P-**Cyt-Rib2Me**-P-rUrd-P-Gua-Rib2Me-P-rAdo-P-rAdo-P-**rWyb**-P-rAdo-P-rPrd-P-m5Cyt-Rib-P-rUrd-P-rGuo-P-rGuo-P-rAdo-P-rGuo-P-m7Gua-Rib-P-rUrd-P-rCyd-P-**m5Cyt-Rib**-P-rUrd-P-rGuo-P-rUrd-P-rGuo-P-m5Ura-Rib-P-rPrd-P-rCyd-P-rGuo-P-m1Ade-Rib-P-rUrd-P-rCyd-P-rCyd-P-rAdo-P-rCyd-P-rAdo-P-rGuo-P-rAdo-P-rAdo-P-rUrd-P-rUrd-P-rCyd-P-rGuo-P-rCyd-P-rAdo-P-rCyd-P-rCyd-P-rAdo

IN FAVOR OF PERCEPTION

- Cost
  - A new file format means a new compound registry system
- Ability to roundtrip
- Flexibility
  - New and unusual monomer? Will be faithfully stored in the MOL file
- Interchange of data
  - SMILES and MOL files are already a de-facto standard
  - HELM requires the (all-atom) monomer definitions
- Analysis
  - Can combine tools for small-molecule analysis with those that take as input biopolymer formats
- File-format lock-in
  - Difficult to migrate if you base your registry system on a particular file format

CHAINS PERCEPTION ALGORITHM

- The input is a connection table
- Typical sources include SMILES, MOL and PDB files
- As an option, graph connectivity *only* can be used, allowing molecules without bond orders to be handled (e.g. from PDB files)

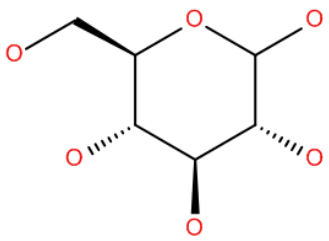
CHAINS PERCEPTION ALGORITHM

- Identify nucleic acid backbone
- Recognize sidechains
- Identify peptide backbone
- Recognize sidechains
- Identify oligosaccharide backbone
- Recognize monosaccharides
- Label unidentified connected components as “Unknown” monomers
- Add connections between monomers

CHAINS PERCEPTION ALGORITHM

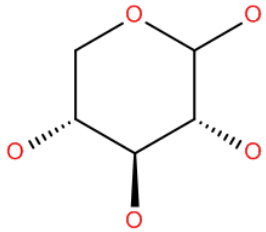
- Identify nucleic acid backbone
- Recognize sidechains
- Identify peptide backbone
- Recognize sidechains
- Identify oligosaccharide backbone
- Recognize monosaccharides
- Label unidentified connected components as “Unknown” monomers
- Add connections between monomers

BIOLOGICALLY-IMPORTANT MONOSACCHARIDE CLASSES



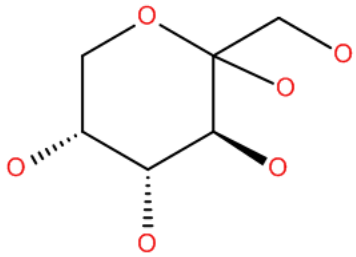
Hexopyranoses

e.g. D-Glucopyranose



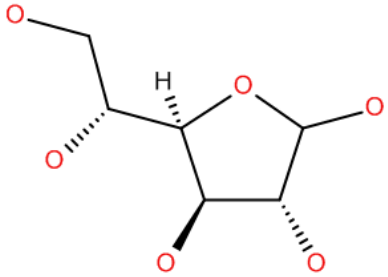
Pentopyranoses

e.g. D-Xylopyranose



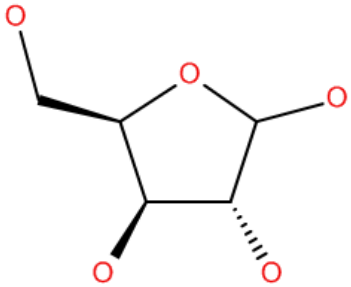
Hex-2-ulopyranoses

e.g. D-Fructopyranose



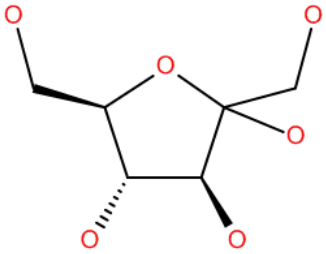
Hexofuranoses

e.g. D-Glucofuranose



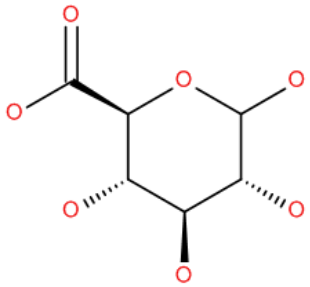
Pentofuranoses

e.g. D-Xylofuranose



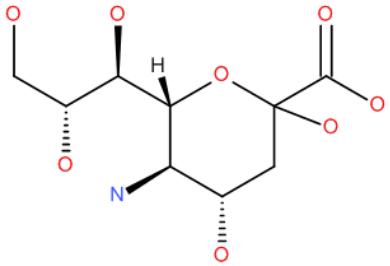
Hex-2-ulofuranoses

e.g. D-Fructofuranose



Hexopyranuronic acids

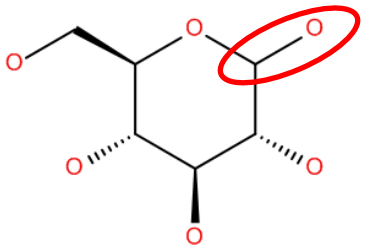
e.g. D-Glucopyranonic acid



Non-2-ulopyranosonic acids

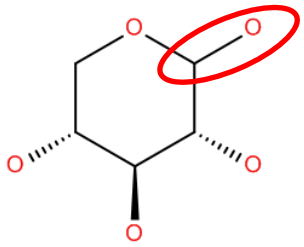
e.g. Neuraminic acid

BIOLOGICALLY-IMPORTANT MONOSACCHARIDE CLASSES



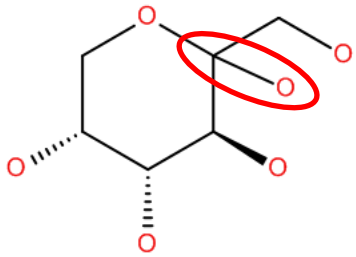
Hexopyranoses

e.g. D-Glucopyranose



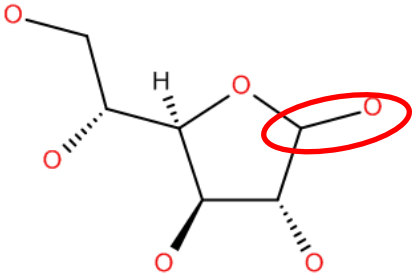
Pentopyranoses

e.g. D-Xylopyranose



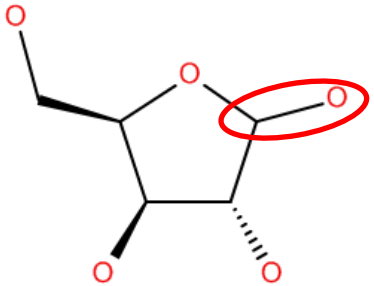
Hex-2-ulopyranoses

e.g. D-Fructopyranose



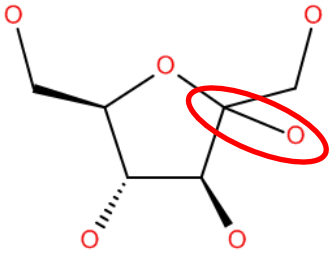
Hexofuranoses

e.g. D-Glucofuranose



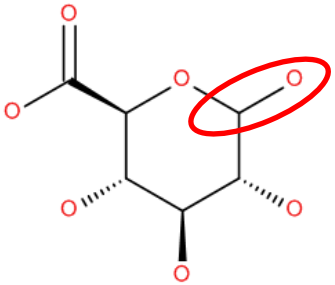
Pentofuranoses

e.g. D-Xylofuranose



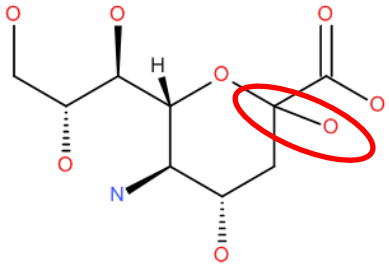
Hex-2-ulofuranoses

e.g. D-Fructofuranose



Hexopyranuronic acids

e.g. D-Glucopyranonic acid



Non-2-ulopyranosonic acids

e.g. Neuraminic acid

POLYMER MATCHING

- Matching of linear, cyclic and dendrimeric polymers and copolymers can be efficiently implemented by graph relaxation algorithms.

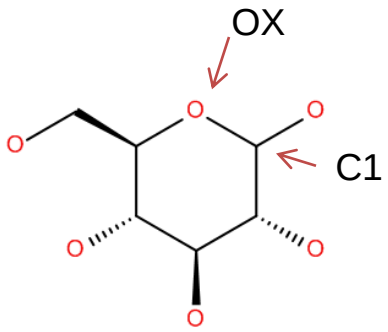
- Each atom records a set of possible template equivalences represented as a bit vector.

- These bit vectors are iteratively refined using the bit vectors of neighboring atoms.

OLIGOSACCHARIDE BACKBONE DETECTION

- Assign initial constraints based upon atomic number, ring membership and heavy atom degree (and optionally valence and charge)

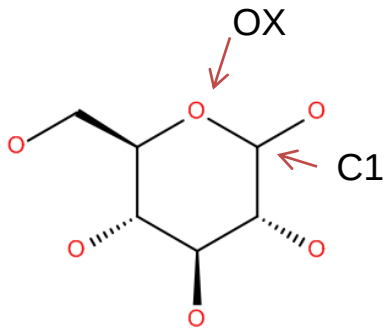
Bit C1:	C	In ring	3 neighbors
Bit OX:	O	In ring	2 neighbors



OLIGOSACCHARIDE BACKBONE DETECTION

- Perform iterative graph relaxation at each atom using its neighbor's bit masks
  - Unset bits if do not match neighbor templates

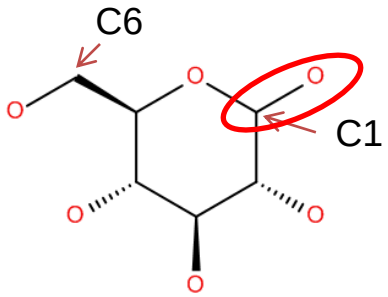
Bit C1:	OX	C2	OH
Bit OX:	C5	C1	



  - End result: either bitmasks go to zero if not a supported monosaccharide, otherwise each atom will have a bitmask with a single bit set

RECOGNIZE MONOSACCHARIDES

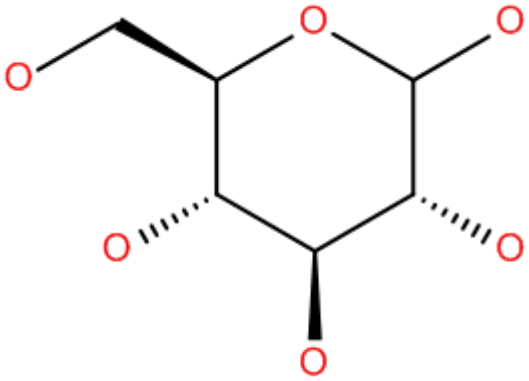
- Find all anomeric centers attached to a peptide (glycoprotein) or with a free OH
  - These are the starting points of oligosaccharide chains



- Travel around each monosaccharide ring from C1->C6
  - Note the location of any substituted OHs or deoxys
  - Note the stereo configuration of the OHs (or subsituted OHs)
  - If an OH links to another monosaccharide, recursively traverse that, building up the oligosaccharide structure

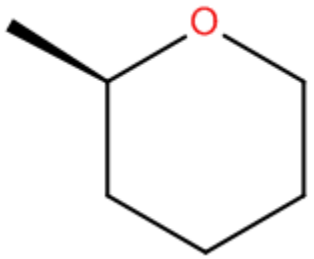
- At C6, name the monosaccharide based on the stereo configuration and record the substitution pattern

SACCHARIDE OR NOT?



D-Glucopyranose

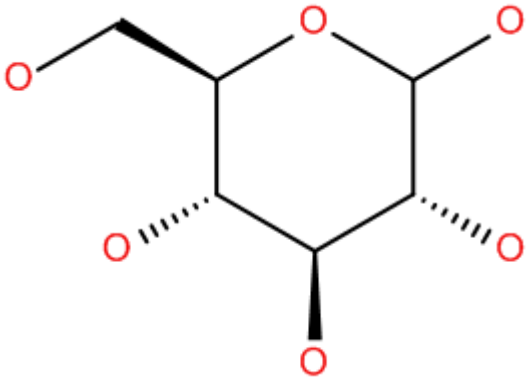
D-*gluco*-hexopyranose



(2S)-2-methyloxane

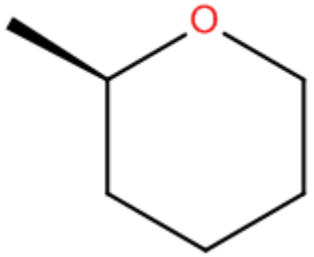
(2S)-2-methyl-tetrahydropyran

SACCHARIDE OR NOT?



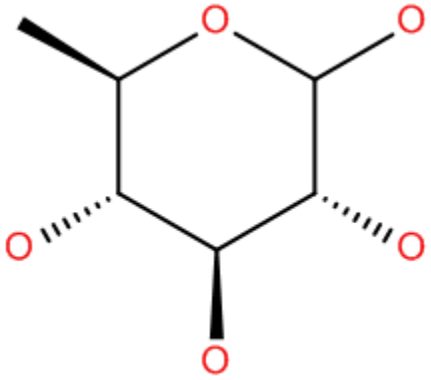
D-Glucopyranose

D-*gluco*-hexopyranose



(2S)-2-methyloxane

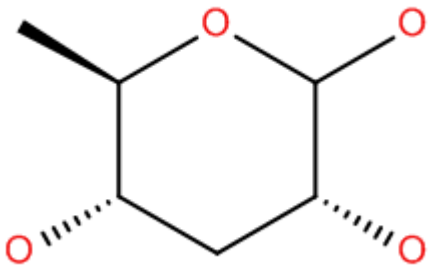
(2S)-2-methyl-tetrahydropyran



D-Quinovopyranose

6-deoxy-Glucopyranose

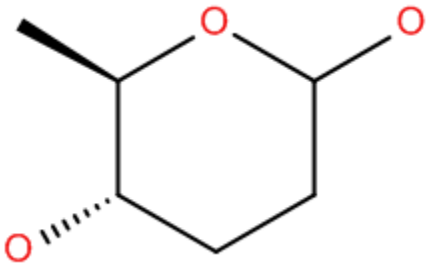
6-deoxy-D-*gluco*-hexopyranose



D-Paratopyranose

3,6-dideoxy-Glucopyranose

3,6-dideoxy-D-*ribo*-hexopyranose



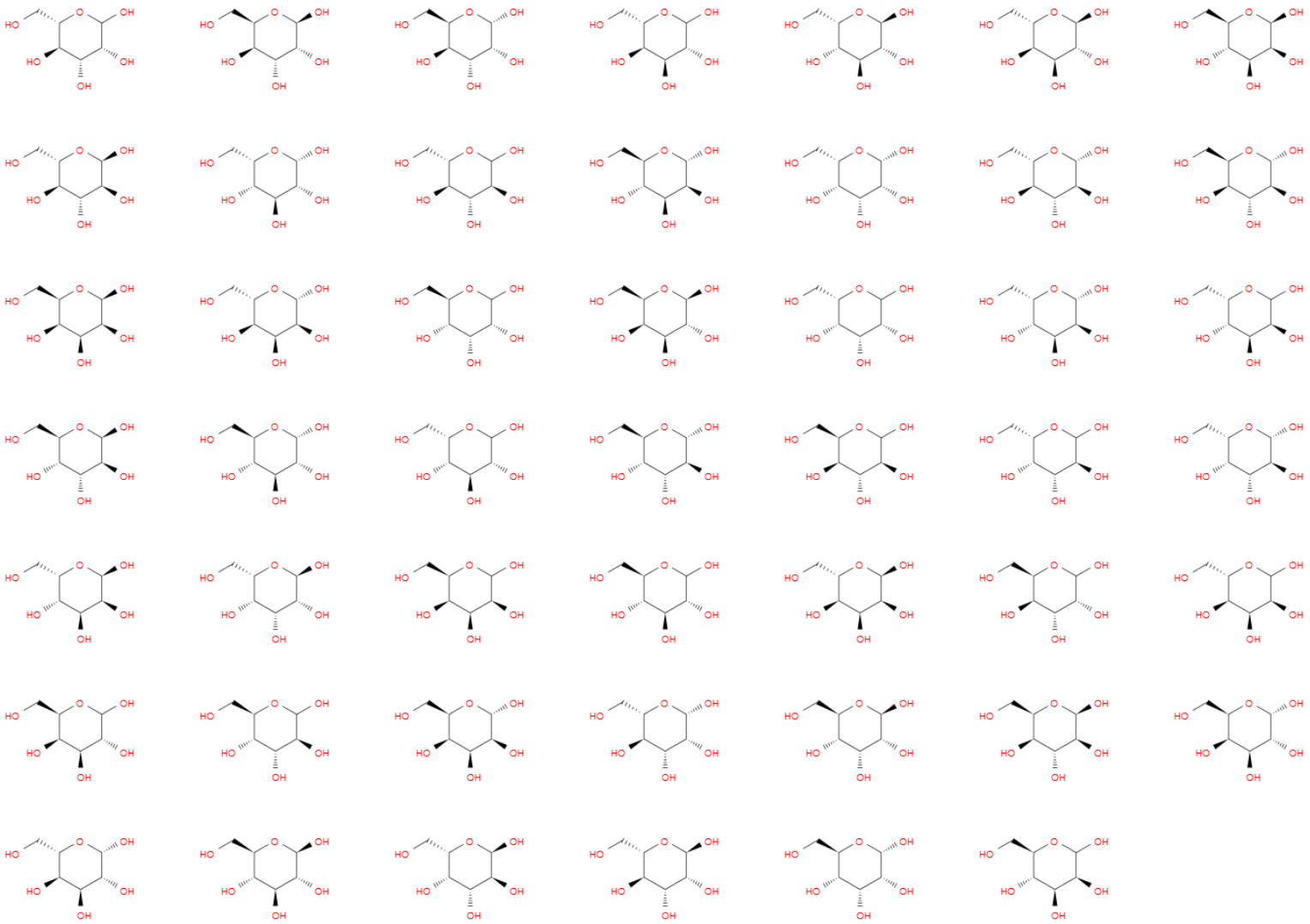
D-Amicetopyranose

2,3,6-trideoxy-Glucopyranose

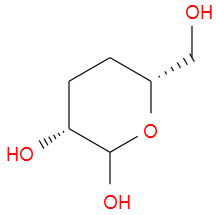
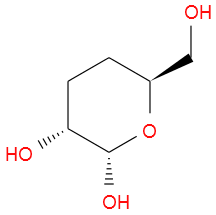
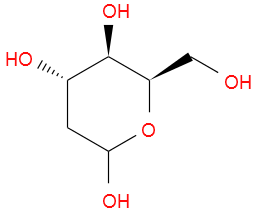
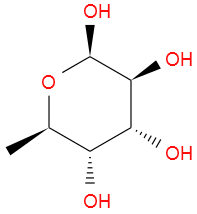
2,3,6-trideoxy-D-*erythro*-hexopyranose

SUPPORTED SACCHARIDES

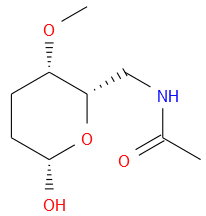
- Initial work has focussed on the most common saccharides
  - 5- and 6-membered ring forms of Hexoses, Hex-2-uloses, and Pentoses, 6-membered ring forms of Hexuronic acid and Non-2-ulosonic acid
  - Currently supported substitutions are those supported by IUPAC condensed notation (N, NAc, OAc, OMe)
- Perceived as monosaccharide if:
  - Matches one of the structures above
  - The hydroxyl group is present at the anomeric position (or substituted by N or O)
  - Another ring hydroxyl group (or substituted hydroxyl group) is present
- Covers 70.2% of the 693 monosaccharides in MonosaccharideDB

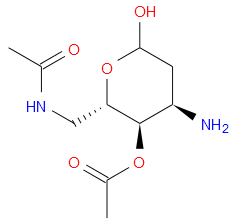


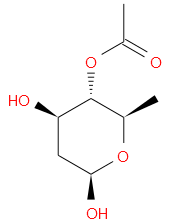
48 hexopyranoses

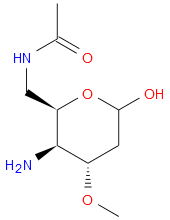





264 deoxy-hexopyranoses





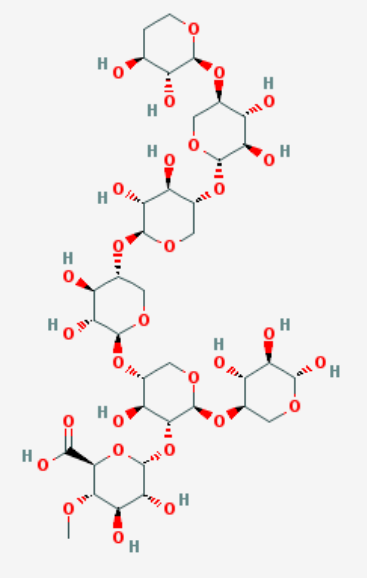




9540 substituted hexopyranoses (4 most common substituents)

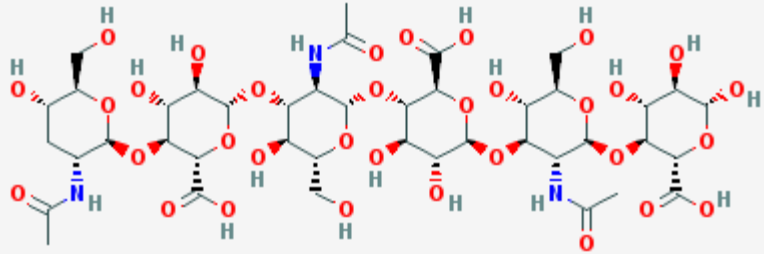
HOW MANY SACCHARIDES ARE IN PUBCHEM?

	Count
Total occurrence (of supported saccharides)	4050
…of which are monosaccharides	1628



CID 16663750

4-deoxy-L-thrPen(a1-4)Xyl(b1-4)Xyl(b1-4)Xyl(b1-4)[GlcA4Me(a1-2)]Xyl(b1-4)b-Xyl



CID 449522

3-deoxy-ribHexNAc(b1-4)GlcA(b1-3)GlcNAc(b1-4)GlcA(b1-3)GlcNAc(b1-4)b-GlcA

HOW MANY MONOSACCHARIDES ARE IN PUBCHEM?

	Count
Total occurrence (of supported monosaccharides)*	121680
…of which have defined stereochemistry	86996
……of which unique	1907
Number of possible substituted hexopyranoses†	9540
…of which observed	936

* May be present as part of a larger molecule, not necessarily an oligosaccharide

† Supported substitutions are amino, acetyl, acetylamino and methoxy

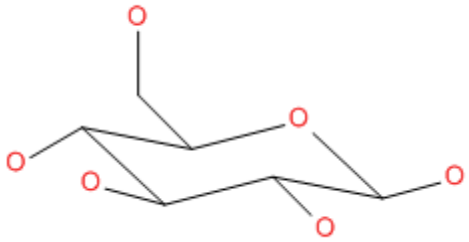
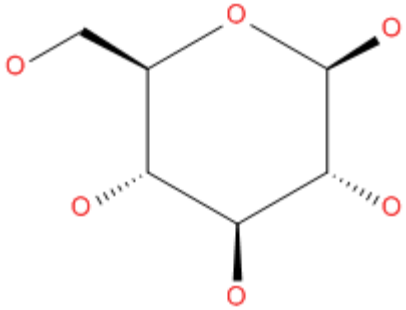
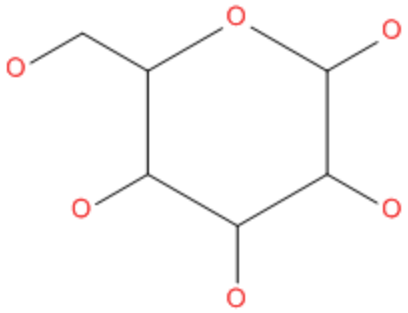
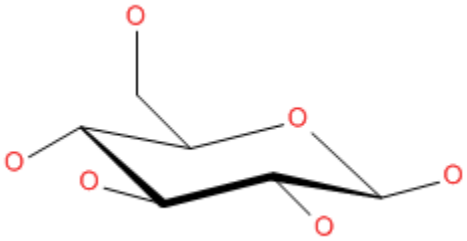
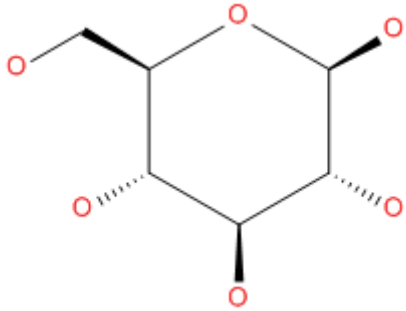
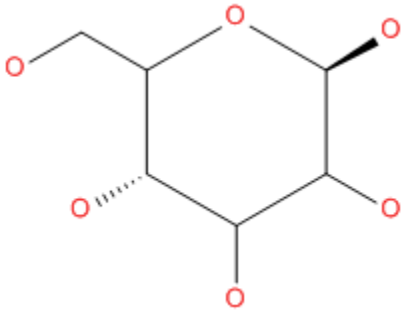
BREAKING NEWS

- Import of GlycosuiteDB into PubChem
  - Database of glycans taken from the literature
  - Covers multiple species
  - Mixture of partial and fully defined structures

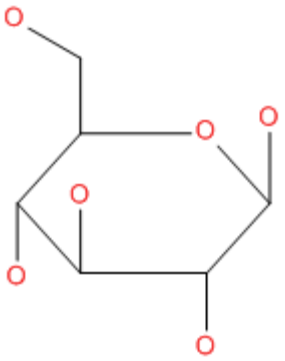
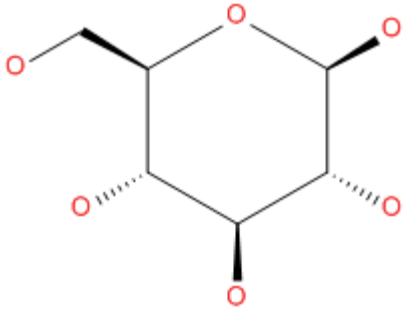
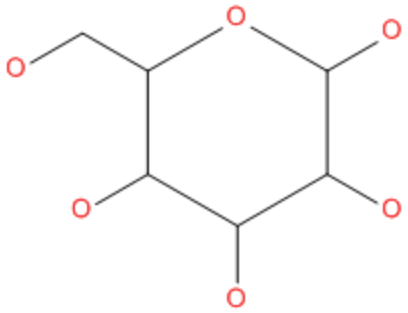
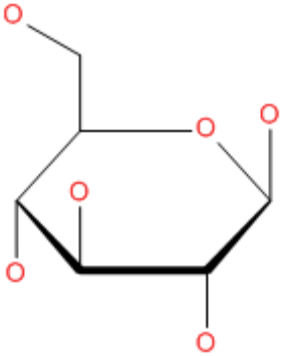
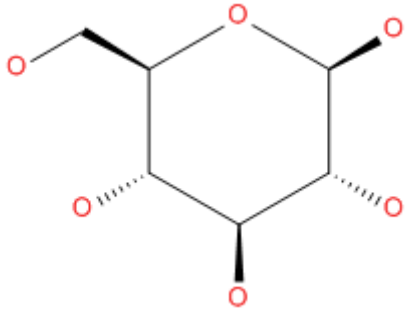
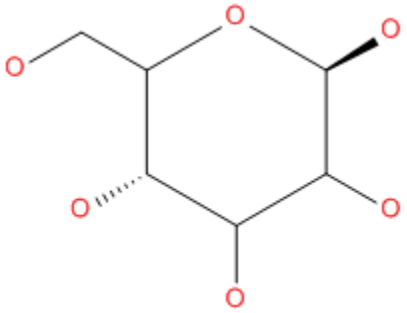
- 1487 oligosaccharide structures imported

CHALLENGES AHEAD

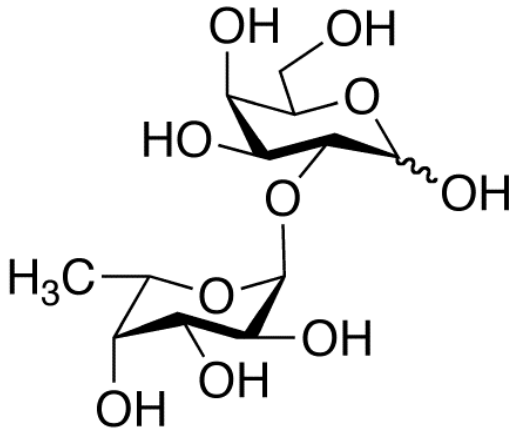
THE PERCEPTION OF SUGAR DEPICTIONS

You drew	You meant	You got
		
		

THE PERCEPTION OF SUGAR DEPICTIONS

You drew	You meant	You got
		
		

EXAMPLE



- PubChem CID 29980572
  - SMILES corresponds to **6-deoxy-L-Gul(a1-2)a-L-Alt**
- ZINC 22059715
  - SMILES corresponds to **6-deoxy-L-Gul(a1-2)a-L-Alt**
- Toronto Research Chemicals F823500
  - Diagram and name corresponds to **6-deoxy-L-Gal(a1-2)Gal**
  - But the 2D SDF file depicts chair forms and uses wedges for perspective

- Given a Mol file depiction of a sugar, the challenge is to:
  - Perceive that it is a sugar
  - Interpret the stereo correctly
  - Generate a Mol file that will be read correctly by cheminformatics toolkits that do not have advanced sugar support

ARBITRARY CHEMICAL MODIFICATIONS

- Roundtripping is straightforward for recognized super-atoms
  - Perceived from all-atom or superatom representation when reading
  - Translated to all-atom or superatom representation when writing

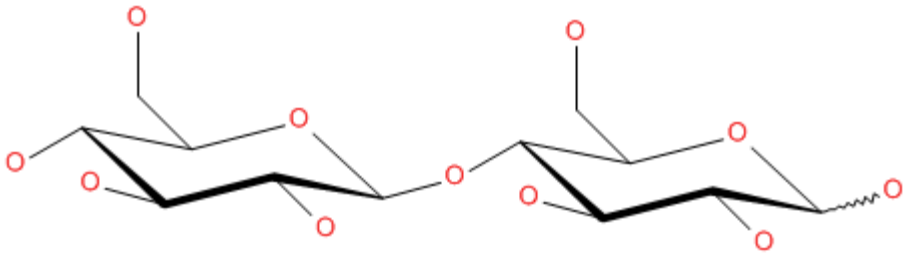
- For “unknown” monomers or chemical modifications, the all-atom representation must explicitly be stored along with the connection point(s)
  - E.g. as SMILES, “*CCC” for 1-propyl
  - A lookup table could be used for common modifications

- Note: arbitrary chemical modifications are not representable in many output formats
  - Be careful of relying on a particular polymer file format

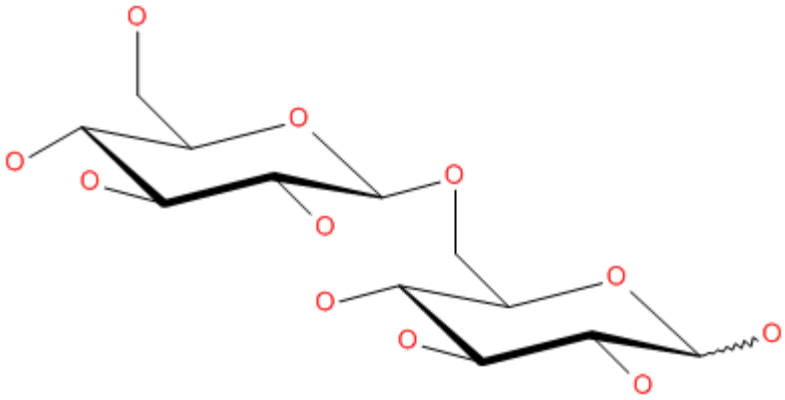
VARIABLE ATTACHMENT POINTS

Reported oligosaccharide structures often contain linkages whose exact attachment point is unknown

Glc(β1→?)Glc



4 possibilities



Challenge: Is it possible to store this information in a Mol file or SMILES string?

VARIABLE ATTACHMENT POINTS

- SMILES does not support positional variation
  - However ChemAxon have used dangling bonds to implement an extension
    - C*.C1CCCCC1 |c:1,3,5,m:1:2.7.3|
- V3000 Mol file
  - Both MarvinSketch and Accelrys Draw store variable attachment points in the same way using the bond block
    - "M V30 7 1 8 7 ENDPTS=(3 3 2 1) ATTACH=ANY"
  - ChemSketch uses its own extension, while ChemDraw cannot store the information
- V2000 Mol file
  - Accelrys Draw will not save as V2000, ChemDraw cannot store the information
  - MarvinSketch uses its own extension, as does ChemSketch

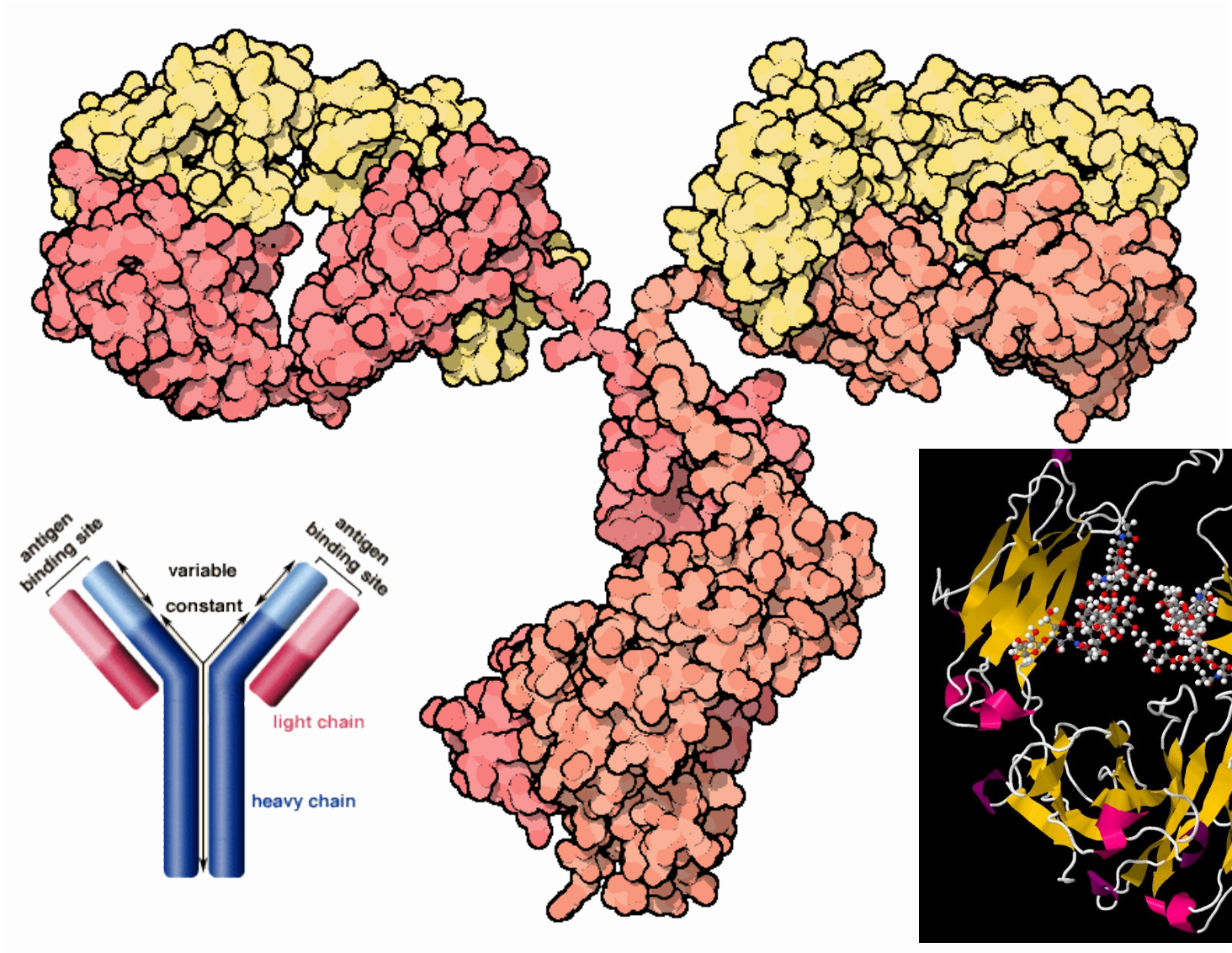
ACKNOWLEDGEMENTS

- Evan Bolton PubChem
- Daniel Lowe NextMove Software

http://nextmovesoftware.com

http://nextmovesoftware.com/blog

noel@nextmovesoftware.com



Left: The Biology Project, Immunology
http://www.biology.arizona.edu/immunology/tutorials/antibody/structure.html
Center: David Goodsell, Antibodies: September 2001 Molecule of the Month, DOI: 10.2210/rcsb_pdb/mom_2001_9

WOULD A SUGAR BY ANY OTHER NAME TASTE AS SWEET?

- Carbohydrates
  - Originally compounds of formula $(CH_2O)_n$, but now used as generic term for monosaccharides, oligosaccharides, derivatives thereof
- Saccharides
  - Considered by some to be synonymous with carbohydrates, otherwise mono-, oligo- and polysaccharides
- Monosaccharides
  - Aldoses, ketoses, uronic acids, and many more
- Sugars
  - Loose term applied to monosaccharides and lower oligosaccharides
- Glycans
  - "Biochemical" term for oligo- and polysaccharides especially as components of glycoproteins and proteoglycans

http://www.chem.qmul.ac.uk/iupac/class/carbo.html

External database	Number of sequences in external database	URL
BCSDB (4)	8119	http://www.glyco.ac.ru/bcsdb3/
CCSD (2)	23 402	http://www.genome.jp/dbget-bin/www_bfind?carbbank
CFG (5)	8873	http://www.functionalglycomics.org/
EUROCarbDB	13 467	http://www.ebi.ac.uk/eurocarb/
Glycobase(Lille) (11)	247	http://glycobase.univ-lille1.fr/base/
GLYCOSCIENCES.de (12)	23 285	http://www.glycosciences.de/
KEGG (6)	10 969	http://www.genome.jp/kegg/glycan/
PDB (13)	905	http://www.rcsb.org/pdb/

Taken from Table 1, Ranzinger et al., Nucleic Acids Research, 2001, 39, D373.

HOW MANY MONOSACCHARIDES ARE IN PUBCHEM?

	Count
Total occurrence (of supported monosaccharides)*	121680
…of which have defined stereochemistry	86996
……of which unique	1907
Number of possible substituted hexopyranoses†	9540
…of which observed	936
Number of possible hexopyranose substitution patterns†	5^4=625
…of which observed	75

* May be present as part of a larger molecule, not necessarily an oligosaccharide

† Supported substitutions are amino, acetyl, acetylamino and methoxy