MACROMOLECULES OR BIG SMALL-MOLECULES?

HANDLING BIOPOLYMERS IN A CHEMICAL REGISTRY SYSTEM

Noel M. O'Boyle,[1] Evan Bolton,[2] Roger A. Sayle[1]

[1] *NextMove Software Ltd, Cambridge, UK* [2] *National Center for Biotechnology Information, Bethesda, Maryland, USA*

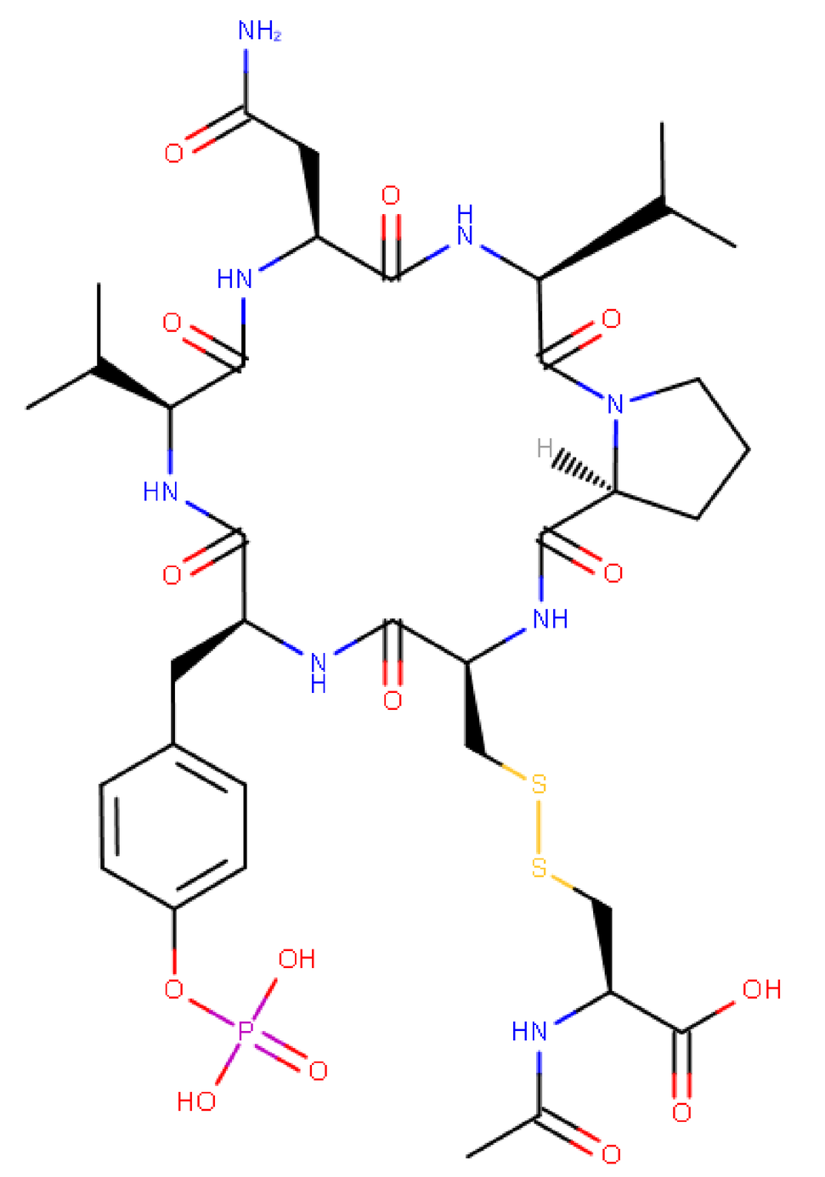
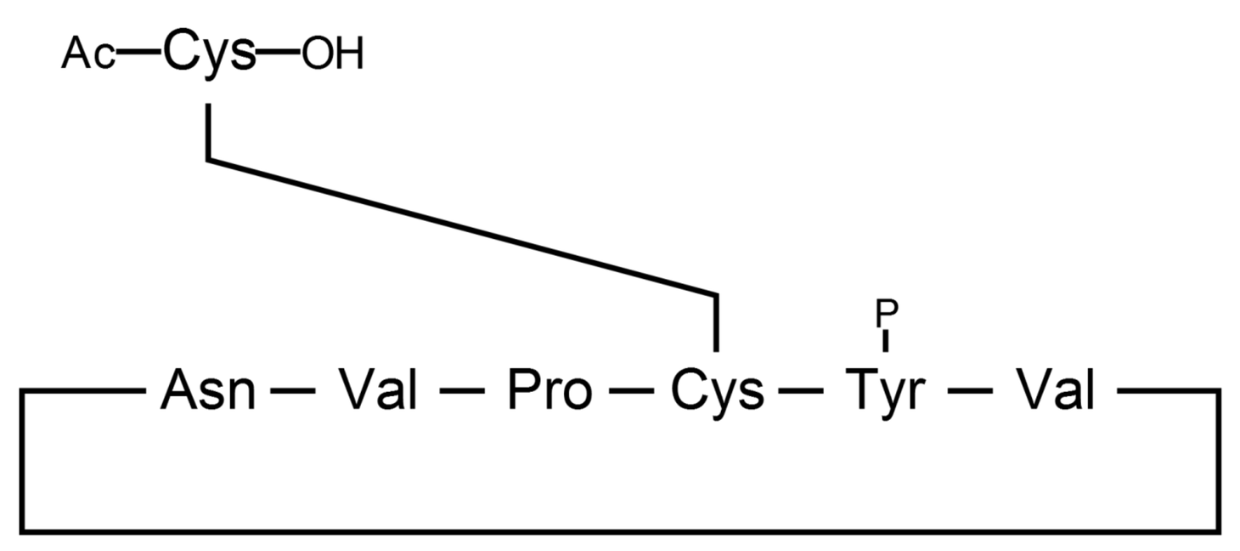
Introduction

The increasing importance of biotherapeutics ("biologics") to the pharmaceutical industry presents a challenge for traditional cheminformatics systems. The formats and depictions appropriate for small molecules may not be appropriate for biopolymers such as polypeptides. Indeed, recently it has been asserted that it is impractical to represent biopolymers at the atomic level [1].

Using efficient perception and conversion routines as implemented in Sugar & Splice, biopolymer structures can be interconverted between macromolecule and small-molecule file formats, allowing appropriate representations and depictions to be generated depending on the context. In this way, existing registry systems designed for small-molecules may be extended to handle macromolecules.

Biopolymers should be handled differently to small molecules

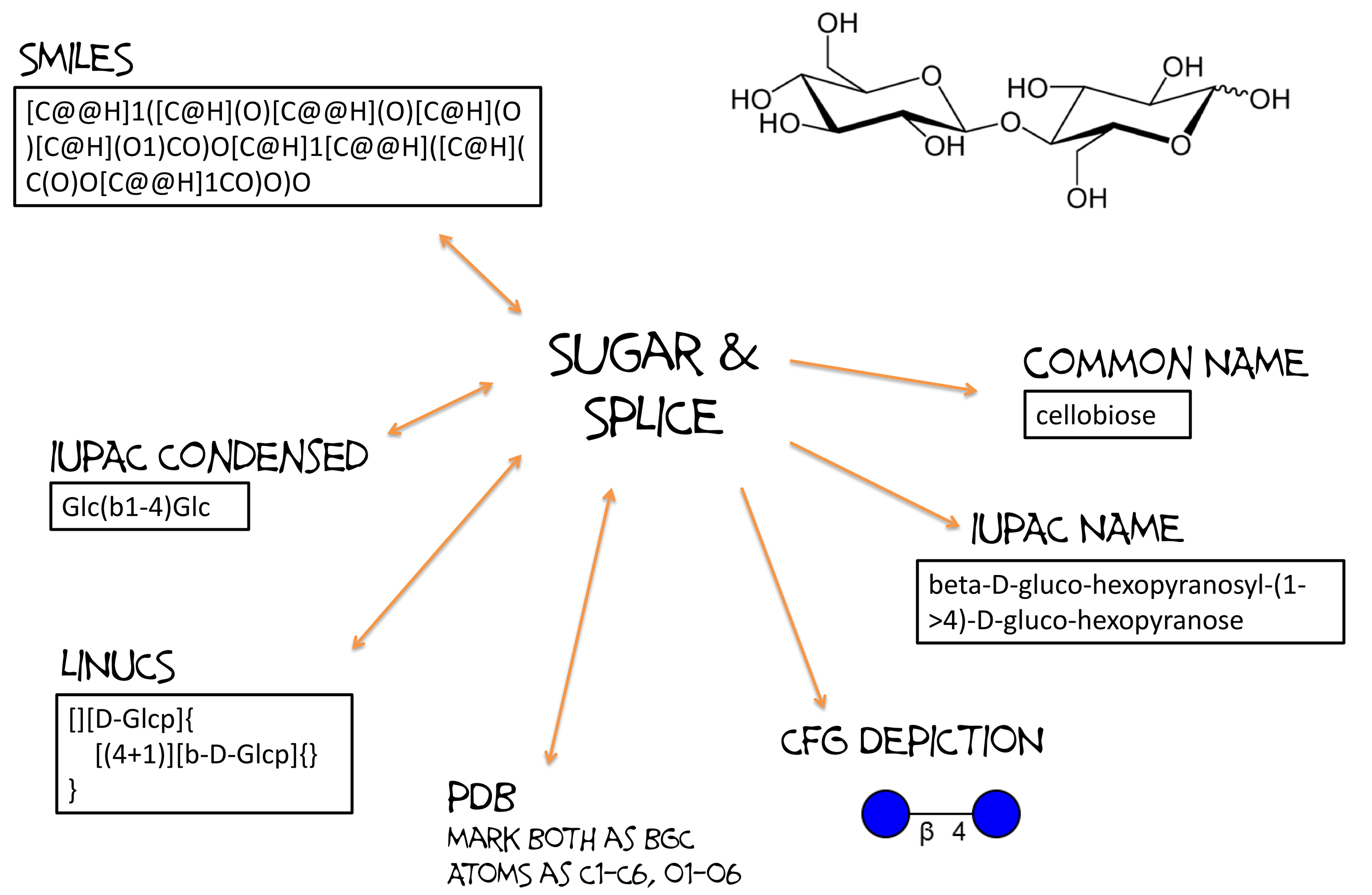
In contrast to the all-atom representations used for small-molecules, biopolymer file formats and depictions are superatom representations of the molecular structure that highlight the identity of monomer units and the nature of the connections between them. The following table illustrates the differences between all-atom and superatom representations of PubChem CID10033747.

All-atom representations	Superatom representations
	**IUPAC depiction [2]** 
	IUPAC Condensed [3] Ac-L-Cys(1)-OH.cyclo[L-Asn-L-Val-L-Pro-L-Cys(1)-L-Tyr(P)-L-Val]
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	**HELM [1]** 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:R3-4:R3|PEPTIDE2,CHEM1,4:R2-1:R1|CHEM1,PEPTIDE3,1:R2-1:R1|PEPTIDE3,PEPTIDE2,1:R2-1:R1$$$

Perception of biopolymer structure from all-atom representations

One way to perceive biopolymer structure would be to use a series of SMARTS patterns for each monomer; however, this is inefficient. Instead, Sugar & Splice uses an iterative graph relaxation algorithm that matches all patterns simultaneously. Each atom records a set of which pattern atoms it may match; this set of matches is iteratively refined based on the set of patterns matched by neighbouring atoms until each atom matches either a single pattern atom or none.

Finally, the structure of the biopolymer is built up by connecting the identified monomers.



Convert between macromolecule and small-molecule representations

The ability to perceive biopolymer structure (as described above) allows one to convert from an all-atom representation to a superatom representation. This can be performed losslessly if the particular representation supports the perceived structure, or provides a method to incorporate an all-atom representation for unknown monomers.

Converting from a superatom representation to an all-atom representation is straightforward and can be done by dictionary lookup (e.g. for amino acids) or algorithmically (e.g. for saccharides).

Handling unknown connection points

Sometimes the identity of the connecting locant is unknown. For example, the IUPAC condensed format for oligosaccharides uses a "?" to indicate this as in the disaccharide "Man(a1-?)Glc". This can be interconverted to/from SMILES using ChemAxon's extended SMILES notation:

[C@H]1([C@H]([C@H]([C@@H]([C@H](O1)CO)O)O)O)*.C1([C@@H]([C@H]([C@@H]([C@H](O1)CO)O)O)O)O |m:11:22.21.20.19|

↕

Man(a1-?)Glc

Similarly there are extensions for MOL files to support unknown connection points, for both version V2000 and version V3000.

Performance

When run over 47.5 million PubChem SMILES strings, the Sugar & Splice perception algorithm takes just under 40 mins (Intel Core i7-3770@3.4Gz) and perceives 7,507 oligonucleotides, 267,294 oligopeptides and 2,101 oligosaccharides.

Conclusions

The conceptual gap between small-molecules and macromolecules can be bridged using efficient perception and interconversion routines. This allows the appropriate tool to be used for a particular task; e.g. generate IUPAC depiction for a peptide, but convert to SMILES for substructure searching.

Bibliography

1. Zhang T, Li H, Xi H, Stanton RV, Rotstein SH. **HELM: A Hierarchical Notation Language for complex biomolecule structure representation.** *J. Chem. Inf. Model.* **2012**, *52*, 2796.
2. IUPAC-IUB Joint Commission on Biochemical Nomenclature. **Nomenclature of Cyclic Peptides. 2004**. *Provisional Recommendations.*
3. IUPAC-IUB Joint Commission on Biochemical Nomenclature. **Nomenclature and Symbolism for Amino Acids and Peptides.** *Pure Appl. Chem.* **1984**, *56*, 595.

NEXTMOVE SOFTWARE

www.nextmovesoftware.co.uk
www.nextmovesoftware.com

NextMove Software Limited
Innovation Centre (Unit 23)
Cambridge Science Park
Milton Road, Cambridge
UK CB4 0EY