

THE NEED FOR OPEN-CIP

Comparing Cahn-Ingold-Prelog Rule Implementations

John Mayfield, Daniel Lowe and Roger Sayle

NextMove Software Ltd, Cambridge, UK.

Introduction

The **Cahn-Ingold-Prelog (CIP) priority rules** rank atoms around a **stereogenic unit** to assign a stereo-descriptor that is invariant to atom order and layout, for example ***R*** (right) or ***S*** (left) for tetrahedral atoms.

A **directed acyclic graph** (digraph) is constructed for each stereogenic unit and the out edges from the root node compared and ranked according to **eight** sequence rules[1]. Each rule is applied exhaustively and tested on the entire digraph before applying the next rule[2].

Eight sequence rules (in essence)

Rule 1
a. Higher atomic number precedes lower
b. An atom node duplicated closer to the root ranks higher than one duplicated further

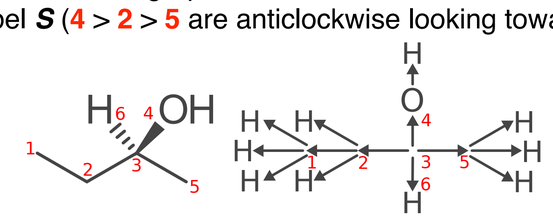
Rule 2 Higher atomic mass number precedes lower

Rule 3 *Z* precedes *E* and this precedes nonstereogenic (nst) double bonds

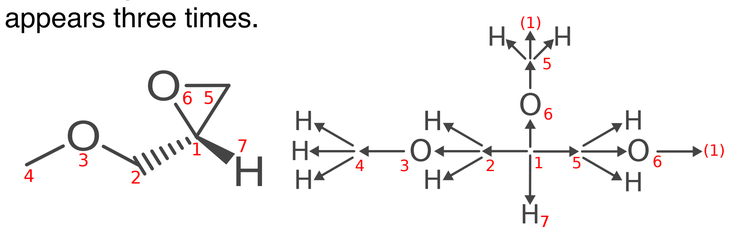
Rule 4
a. Chiral stereogenic units precede pseudoasymmetric stereogenic units and these precede nonstereogenic units (R = S > r = s > nst)
b. When two ligands have different descriptor pairs, the one with the first chosen like descriptor pairs has priority over the one with a corresponding unlike descriptor pairs
c. *r* precedes *s*

Rule 5 An atom or group with descriptor *R* has priority over its enantiomorph *S*

The application of **Rule 1a** to the digraph for **2-butanol** ranks the out edges connected to the root as giving the label ***S*** (**4** > **2** > **5** are anticlockwise looking towards **6**).



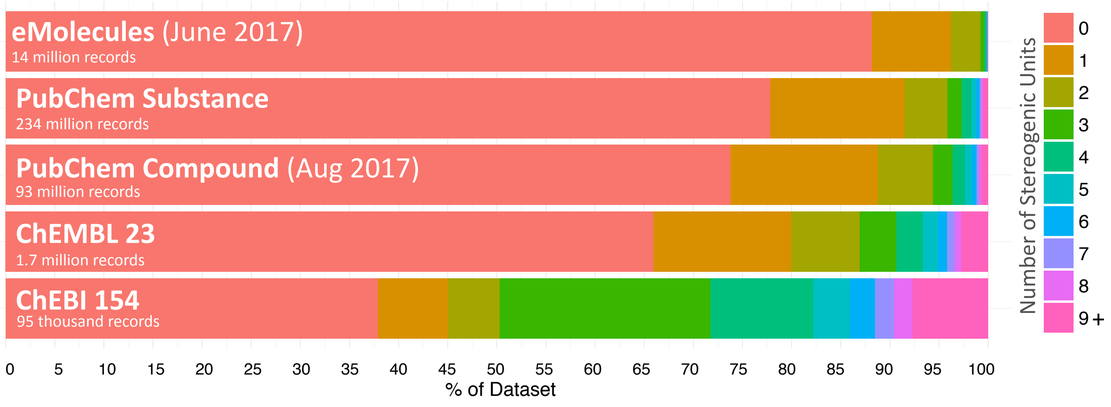
Atoms connected by **double** and **triple bonds** as well as **ring closures** result in duplicated nodes in the digraph. In the structure below atoms **5** and **6** appear twice and atom **1** (the root) appears three times.



Due to this duplication, complex ring systems can generate exponentially large digraphs that are **not** computationally tractable. Further complexity in digraphs is caused by the use of fractional atomic numbers in **mancude ring-systems** and assignment of **auxiliary descriptors** for applying **Rules 3-5**.

Stereochemistry in Databases

The number of **defined** stereogenic units per **molecule** varies between databases.



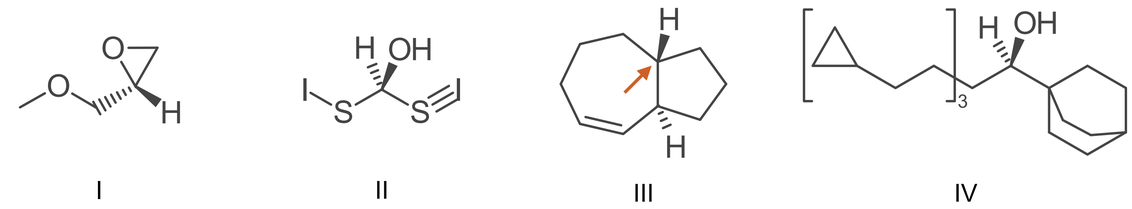
The majority of **stereogenic units** are constitutionally asymmetric and can be ranked using **Rule 1a**. However, in some datasets the number of **stereogenic units** requiring **Rule 4b** and **5** can be significant.

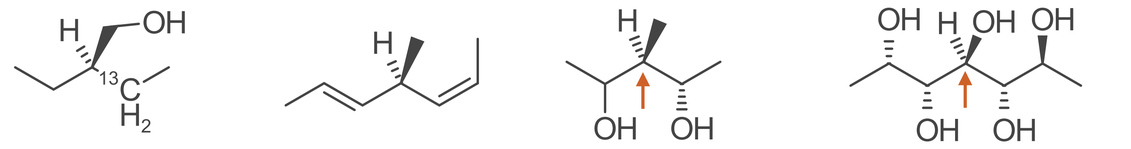
	ChEBI		ChEMBL		eMolecules		PubChem Compound[1]		PubChem Substance	
Rule 1a	281K	99.6%	1.8M	98.6%	2.4M	97.0%	53.5M	100.0%	93.1M	98.7%
Rule 1b	4		1				164		255	
Rule 2			14				3,565		6,789	
Rule 3	29		3		441		36		45	
Rule 4a	122		126		273		4		12,770	
Rule 4b	563	0.2%	4,037	0.2%	3,188	0.1%			125K	0.1%
Rule 4c			19						558	
Rule 5	285	0.1%	23.4K	1.2%	69K	2.8%	15		1.1M	1.2%
Total	**282K**		**1.9M**		**2.4M**		**53.5M**		**94.3M**	

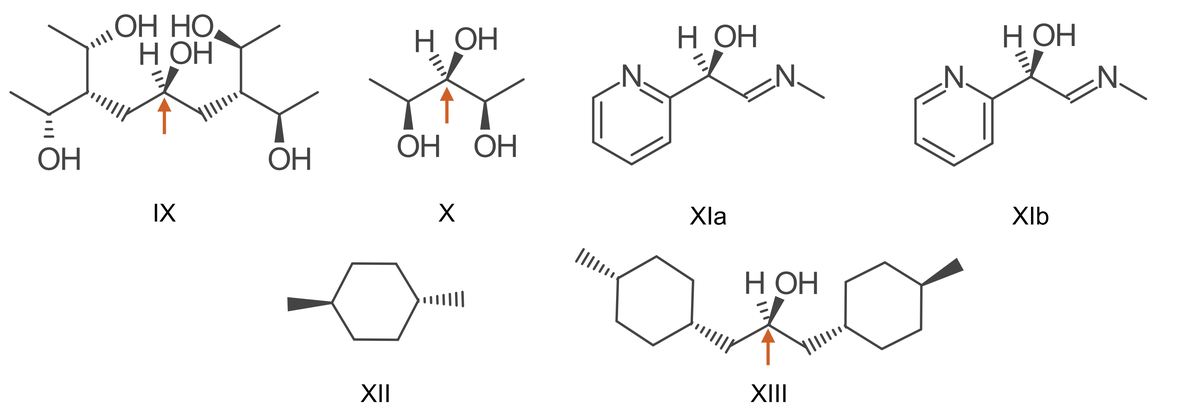
[1]As part of the PubChem Compound's processing, non-constitutional stereochemistry is removed: for example the nine stereoisomers of inositols are all represented by **CID 892**.

Results

A set of **fourteen** structures was collected to identify differences between software implementations. The structures were selected to cover all the **sequence rules** and their applications to **special cases**.







	I	II	III	IV	V	VI	VII	VIII	IX	X	XIa	XIb	XII	XIII
Centres 2.0	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*r*	*R*	*R*	*r*	*R*
JMol 14.20.3	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*r*	*R*	*R*	*r*	*R*
ACD/ChemSketch 14.05beta	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*r*	*R*	*R*	*r*	*R*
Balloon 1.6.5beta	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*r*	*R*	*R*	*r*	*R*
KnowItAll ChemWindow 2018	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*r*	*R*	*R*	*r*	*R*[5]
ChemDraw 16.0	*R*	*R*	*R*	*R*	*S*	*R*	*R*	*R*	*R*	*r*	*R*	*R*	*r*	*R*
BIOVIA Draw 2017	*R*	*R*	*R*	-	*R*	*R*	*R*	*R*	*R*	*-1*	*R*	*R*	*-1*	*R*
MarvinSketch 17.17	*R*	-	-	-	*S*	*R*	-	*R*	-	*r*	*R*	*R*	*r*	-
Indigo 1.3.0Beta.r16	*-2*	*R*	-	-	*R*	-	*R*	*R*	*R*	*r*	*S*	*R*	-	-
RDKit 2017.03.03	*S*	*R*	*S*	*R*	*S*	*R*	*R*	*S*	*R*	*R*	*R*	*R*	-	-
DataWarrior 4.6.0	*R*	*R*	*R*	-	*S*	*R*	*R*	*S*	*R*	*R*	*R*[3]	*R*	-	-
CACTVS (NCI Resolver Aug 17)	*R*	*R*	*S*	-	*S*[4]	*R*	*R*	*S*	*R*	*R*	*S*	*R*	-	-
OPSIN 2.3.1	*R*	*R*	*R*	*R*	*R*	-	-	-	-	-	*S*	*R*	-	-
LexiChem (OEChem) 20170613	*R*	*R*	-	-	*R*	-	-	-	-	-	*S*	*R*	-	-
ChemDoodle 7.0.2	*R*	*R*	-	-	*S*	-	-	*s*	-	*r*	*S*	*R*	-	-
CDK 2.0	-	*R*	*R*[5]	-	*S*	-	-	-	-	-	*S*	*R*	-	-
JUMBO 6	*R*	-	*S*	-	-	-	-	-	-	-	*S*	*S*	-	-

1. Pseudoasymmetric *r/s* labels not displayed but must be calculated due to answers given for IX and XIII
2. Runtime error occurs
3. Impossible to test as different Kekulé forms are normalised
4. *R* in CACTVS since Feb 2015, NCI resolver is old version
5. Other descriptor is assigned differently

Constitutional (Rule 1a, 1b, 2) | Geometrical + Topographical (Rule 3,4a,4b,4c,5) | Special (Mancude, Aux Descriptors)

Fix CIP Collaboration

Since submitting this work for presentation the developers: **Centres**, **JMol**, **ACD/ChemSketch**, and **Balloon** have begun a collaboration. We are in the process of submitting for publication an extended in-depth validation set and proposing sequence rule refinements and additions where they are required.

Conclusion

The CIP sequence rules provide a standard way for chemists to effectively describe the configurations of most stereogenic units. However, beyond simple cases the complexity of the rules necessitates software is used as an aid to naming configurations. The results demonstrate even then, software implementations do not all agree on the configuration.

Through the results presented here and the on-going effort of the **Fix CIP** collaboration, software should aim to converge upon consistent stereochemistry naming. An **Open CIP** software tool could provide "*blessed*" stereochemistry configuration names and provide a **standard** algorithm implementation for other vendors to integrate or adapt.

Acknowledgements

Robert Hanson, Andrey Yerin, Mikko Vainio, and Sophia Gillian Musacchio for initiating and participating in the "Fix CIP" collaboration and the many in-depth technical discussions that have lead to improvements in the tools. Karl Nedwed for providing KnowItAll results. Philip Skinner for providing ChemDraw licenses. Noel O'Boyle for feedback and suggestions.

Bibliography

1. **P-92.1.3** Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013
2. Paulina Mata. The CIP System Again: Respecting Hierarchies Is Always a Must. J. Chem. Inf. Comput. Sci., **1999,** 39 (6)

NextMove Software

www.nextmovesoftware.com

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