The Symbolic Representation of Monosaccharides in the Age of Glycobiology

Abstract.

This chapter offers a general background for students embarking in glycoscience to grasp an essential component of the field, i.e. the alphabet of the building blocks that constitute the many naturally occurring glycans and complex carbohydrates. There is a need to conform to the recommendations of nomenclatures of carbohydrates whilst the constraints required by the developing field of glycobiology in terms of visualization and encoding. The present chapter offers a unified presentation of the nomenclature and symbols that form part of the "language" used to communicate more effectively. It covers about 120 monosaccharides which have been identified as the building blocks of the vast majority of glycans. As a picture speaks a thousand words, it is hoped that this language could be understood as quickly and effectively as possible and used by those working in bioinformatics/database compilation and any one working with oligo, polysaccharides and complex carbohydrates.

1.0 Developments of Carbohydrate Nomenclature and Representation

Monosaccharides are the chemical units from which all members of the major family of natural products, the carbohydrates, are built. They are the individual carbohydrate building blocks, i.e. the monomeric constituents of more complex architectures that will be referred to as glycans, an assembly of sugars either in free forms or attached to another molecule or macromolecule. Glycans occur as: (i) oligosaccharides (comprising 2 to 10 monosaccharides linked together either linearly or branched); (ii) polysaccharides (for glycan chains built up from more than 10 monosaccharides but the distinction with oligosaccharides is not strictly drawn); (iii) glycoconjugates (when the glycan chains are covalently linked to proteins (glycoproteins), lipids (glycolipids) or naturally occurring aglycones (e.g as in antibiotics, saponins, alkaloids). Glycobiology is the study of structure, chemistry, biosynthesis and biological functions of glycans and their derivatives.

1.1. From Fischer Projections to IUPAC/IUBMB Recommendations

Emil Fischer elucidated the structure of glucose and its isomers using ingenious chemical and polarimetric methods (Fischer, 1890), the work being recognized as one of the outstanding achievements of early structural work (Lichtenthaler, 2002). Monosaccharides with an aldehydic carbonyl (or potential aldehydic) group are called aldoses; with a ketonic carbonyl (or potential ketonic carbonyl) group are called ketoses. Glyceraldehyde ('glycerose' in carbohydrate terms) is the simplest aldose (a triose containing an aldehydic

group) having one asymmetric center, and therefore two stereoisomers (enantiomers); there are four aldotetroses, eight aldopentoses and 16 aldohexoses.

Fischer projection formulas of the D-enantiomers of the common aldotriose, aldotetroses, aldopentoses and aldohexaoses are presented in Figure 1, including their trivial names, and their abbreviations when defined. Fischer assigned to the dextrorotatory glucose (via glucaric acid) the projection with the OH group at C-5 pointing to the right. Much later (Bijvoet, 1951) it was proved correct in the absolute sense.

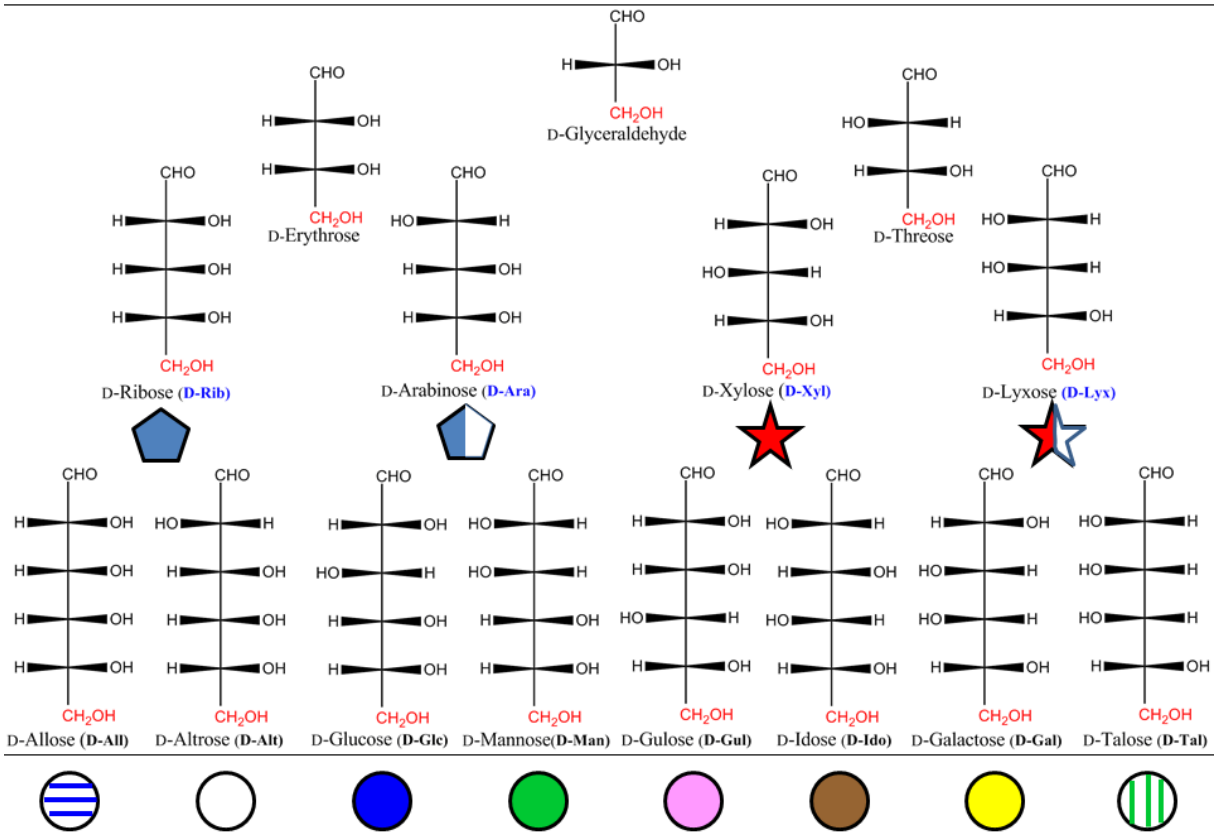


Figure 1. Fischer projection formulas of the D-enantiomers of the common aldotriose, aldotetroses, aldopentoses and aldohexoses, along with their trivial names and abbreviations. Only the D-forms are shown; the L-forms are the mirror images. (*A saying for remembering the structures of the eight aldohexoses is : "ALL ALTruists GLadly MAke GUms In GAllon TAnks"*

Note: The word "carbohydrate" was derived via the German "Kohlenhydrat" and the French "Hydrate de Carbone" to describe the simplest polyhydroxycarbonyl compounds having formulae $C_n(H_2O)_n$. Many monosaccharides having somewhat modified formulae (having amino groups, devoid of specific hydroxyl groups) have been found. Therefore, the word "carbohydrate" no longer has an exact definition, nor has the word "sugar" which is often used as a synonym for "monosaccharides", but is also applied to more complex monosaccharide-containing molecules. Indeed, in everyday usage "sugar" refers to table sugar which is sucrose or saccharose, a disaccharide molecule made up of two monosaccharides, D-glucose and D-fructose.

In aqueous medium, monosaccharides with a suitable carbon-chain length, having both hydroxyl and carbonyl functions, undergo intramolecular (cyclic) hemiacetal formations. The equilibrium, of which the formation is accelerated under weak acidic or alkaline conditions, favors cyclic forms, and "open chains" forms occur only in trace amounts. Stable five- (4 C and 1 O atom) and six (5 C and 1 O atom) membered ring forms are the result. The drawing of the cyclic forms of the Fischer projection formulas does not provide a realistic

representation. More realistic drawings of the cyclic forms were introduced by Haworth in the 1920s, and are referred to as Haworth representations. A perspective drawing of the ring offers a simplified model. The ring is oriented almost perpendicular to the plane of the paper, but viewed from slightly above so that the edge closer to the viewer is drawn below the more distant edge, with the intracyclic oxygen behind and the anomeric carbon at the right-hand end. To define the perspective, the ring bonds closer to the viewer are often thickened. Figure 2 is a schematic representation of a pyranose ring closure in D-glucose that shows the reorientation at C5 necessary to allow ring formation.

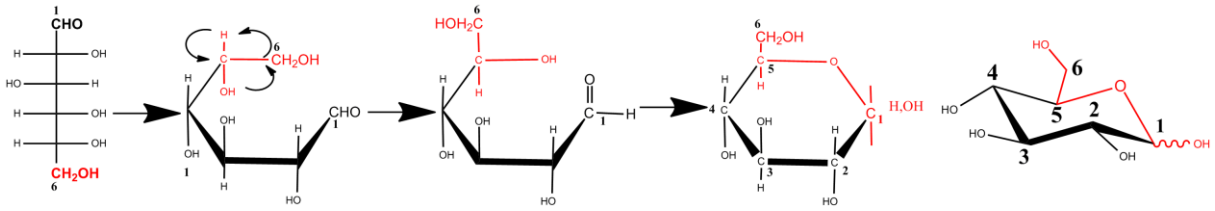


Figure 2: From linear (Fischer) representation to cyclic structure (Haworth) and chair conformation

In the case of D-glucose, the hydroxyl group at C5 reacts intra-molecularly with the aldehyde group at C1. As the carbonyl carbon atom C1 of the open-chain form becomes an additional asymmetric carbon in the hemiacetal formation, two pyranose rings are formed. To describe the stereochemistry around C1 (denoted by the anomeric carbon atom), the terms α and β have been chosen. In the α-anomer, the exocyclic oxygen atom at the anomeric center is formally *cis* in the Fisher projection, to the oxygen attached to the anomeric reference atom; in the β anomer these oxygen atoms are formally *trans* (Figure 3).

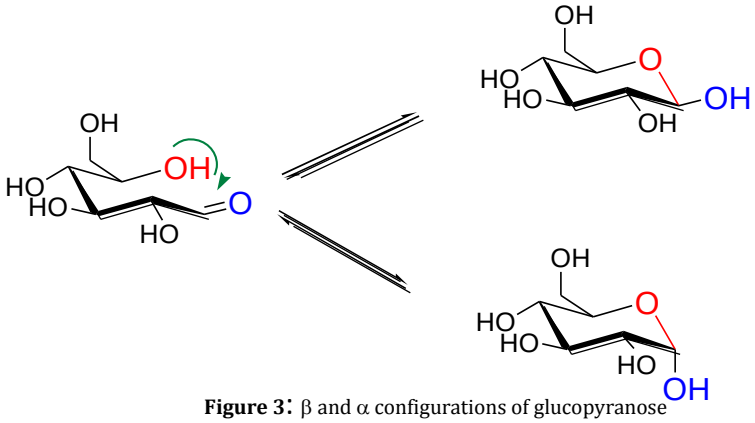


Figure 3: β and α configurations of glucopyranose

Mixtures of anomers.

In solution, most simple sugars and many of their derivatives occur in a monosaccharide-specific equilibrium of α-pyranose, β-pyranose, α-furanose, β-furanose and acyclic (open chain) forms. This process is called mutarotation as it refers to changes in optical rotation to an equilibrium value when pure anomeric forms of monosaccharides are dissolved in water. In general the acyclic forms are only present in trace amounts (e.g. <0.03% in the case of D-glucose) (Figure 4). The presence of a mixture of two anomers of the same ring size may be

indicated in the name by the notation α,β, e.g. α,β-D-glucopyranose. In formulae, the same situation may be expressed by use of a wavy line.

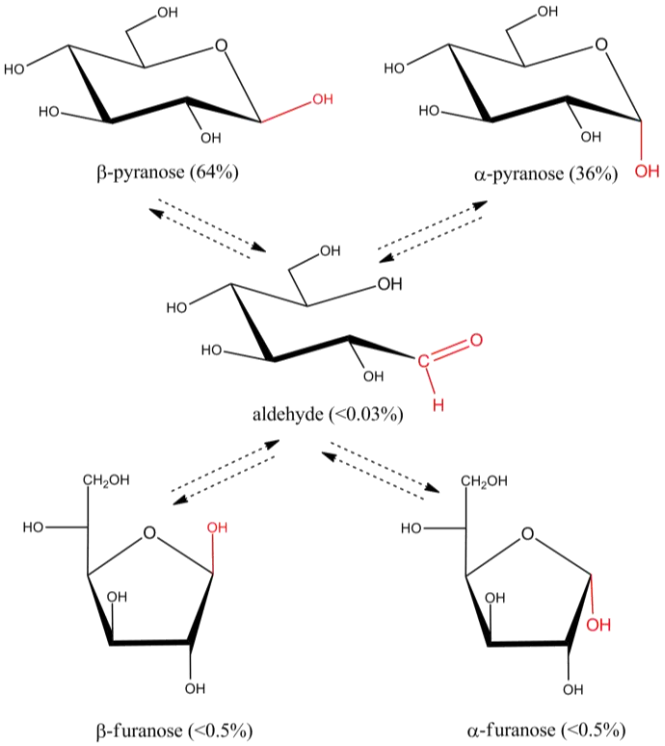


Figure 4: Mutarotation scheme in D-glucose.

1.2. The Conformational Descriptors.

Furanose ring structures occur in envelope (*E*) and twist (*T*) conformations which can be represented on a pseudo-rotational wheel. As the difference in energy between the different conformations on the wheel is generally low, two regions having low energy conformations occurring in the Northern and Southern can be identified (Figure 5).

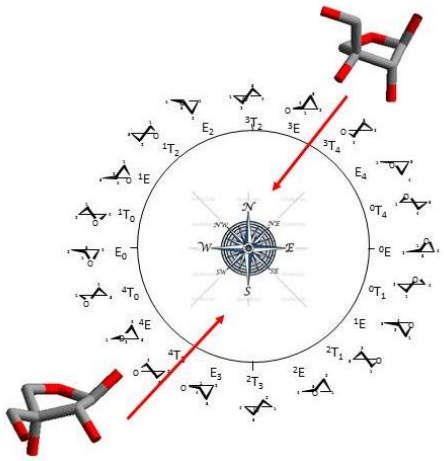


Figure 5. Schematic representation of the five-member ring structures, along with the map for ring-puckering parameters: The pseudo-rotational wheel of the five-member ring encompasses the 20 twist and envelope shapes. The molecular drawings of the low-energy conformers of the monosaccharide apiose North and South, are presented.

For the sake of clarity they are referred to as *N* (North) and *S* (South) forms. Because furanoses can adopt several low energy conformations, the Haworth projection still appears to be the simplest means to avoid the complexity of structural representation.

Six-membered ring structures can occur in two chair (*C*), six boat (*B)*, six skew (*S*), and twelve half-chair (*H*) conformations. These variants are defined by the locants of the ring atoms that lie outside a reference plane. In practice, the two chair conformations have the lowest energy, and strongly dominate. The preference for these low energy conformations is dictated by the relative orientations of the hydroxyl groups. In the case of D-glucopyranoses, only the $^{4}C_{1}$ conformation is of importance, whereas the $^{1}C_{4}$ conformation dominates in α-D-idopyranose. Cases occur as in β-D-arabinopyranose where both chair conformations are in equilibrium.

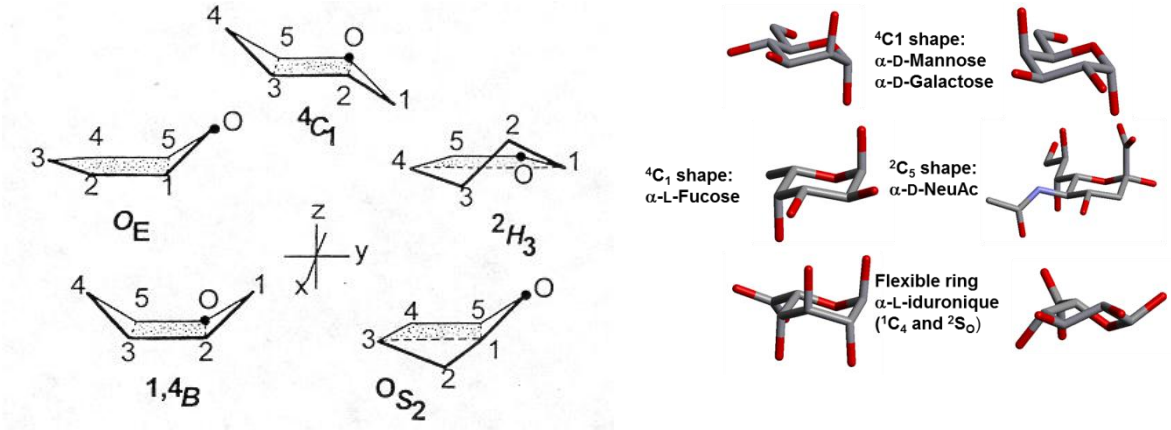


Figure 6. Schematic representation of the puckering parameters describing the conformations of six-member rings along with five low-energy conformations $^{O}_{E}$, $^{1,4}B$, $^{0}S_{2}$, $^{2}H_{3}$, $^{4}C_{1}$. The low-energy shapes taken by common monosaccharides are shown.

1.3. Naming Monosaccharide Derivatives.

The hydroxyl groups of monosaccharides can undergo a series of chemical modifications; methylation, esterification (phosphate, acyl esters, sulfate esters, …), deoxygenation to form deoxysugars. Many monosaccharides have *N*-acetamido groups, such as GlcNAc, GalNAc, and Neu5Ac. In rare cases, the *N*-acetamido group is de-N-acetylated to form amino groups. These are found in heparan sulfate, glycosylphosphatidylinositol (GPI) anchors, and many bacterial glycan structures. Amino groups can be modified with sulfates, similar to hydroxyl groups, as found in heparan sulfate.

The IUPAC/IUBMB document entitled "Nomenclature of Carbohydrates (ref: IUPAC, IUBMB) provides recommendations in giving systematic names to monosaccharide derivatives. (Kamerling, 2007). Nevertheless, the use of trivial names still persists.

Of special biochemical importance are the esters of monosaccharides with phosphoric and sulfuric acid. They are generally termed as phosphates and sulfates (regardless of the state

of ionization or the counter ions). In abbreviations, an italic capital P (*P*) is used to indicate a –PO3H2 group or 'PO2H- group. An italic capital S (S) is used in the case of sulfuric acid preceded by the appropriate locant after the carbohydrate name.

2.0 At the Instigation of Glycobiology

The multiple monosaccharide building blocks can be linked to various regio-chemistries and stereo-chemistries, and the resulting glycan be assembled on protein or lipid scaffolds. The diversity of glycoconjugate structures comprises an "information-rich" system capable of participating in a wide range of biological functions. There comes the question of representing and encoding the many monosaccharides in ways that meet the practices of several scientific communities and are compatible with the requirement of bioinformatics.

Like nucleic acids and proteins, it is far more efficient to encode glycans using a residue-based approach. However, as compared to nucleic acids or proteins, there are a far greater number of residues arising, due to the frequent modifications occurring on the parent monosaccharides. Also, since glycans are frequently found to have branched structures, most of them are tree-like molecules, unlike nucleic acids and proteins. The pre-requisite for a residue-based encoding format is a controlled vocabulary of residue names. For practical reasons, it makes sense to restrict the number of residues to as low a number as possible. Yet the lack of clear rules to subscribe atoms of a molecule to one particular monosaccharide and not of a substituent, pose the main hurdle in encoding monosaccharide names. The variety in nomenclature and structural representation of glycans makes it difficult to decide the most appropriate form of illustrating the approach of the scientific investigation. The choice of notation may be based on whether the study is focused on the chemistry or on the biology side of glycoscience. Moreover, the information content of each representation may vary or highlight a particular aspect as compared to others. While representing a complex glycan structure, chemists could favor a representation that includes information about the anomeric carbon, the chirality, the monosaccharides present along with the glycosidic linkages connecting them. For others, it could be more appropriate to visualize the monosaccharides present, and hence a symbolic/diagrammatic representation would be favored.

2.1. The Symbol and Text Nomenclature for Representation of Glycan Structures.

The representation of monosaccharides as symbols is a result of a concerted effort under the umbrella of the Consortium for Functional Glycomics (ref: CFG). After evaluating widely used symbol nomenclatures, an ad hoc nomenclature committee selected a version originally put forth by Stuart Kornfeld and later adapted by the editors of the textbook 'Essentials of Glycobiology' (Varki *et al.*, 2008 - Cold Spring Harbor Laboratory Press). Further considerations were concerned with the adequacy for the annotation of mass spectra, as well as achieving

common ground with as many other interested groups as possible. It resulted in the following set of rules:

- Avoid same shape/color but different orientation to represent different sugars since annotations of mass spectra are hard to show as horizontal cartoons.
- Choice of symbols should be logical and simple to remember. Each monosaccharide type (e.g. Hexose) should have the same shape, and isomers are to be differentiated by color /black/white/shading.
- Use same shading/ color for different monosaccharides of same stereochemical designation e.g., Gal, GalNAc, GalA should all be the same shading/color.
- To minimise variations, sialic acids and uronic acids are same shape. Only major uronic and sialic acid types are represented.
- While color is useful, the system should also function with black and white, and colored representations should survive black-and white printing or photocopying
- Positions of linkage origin are assumed to be common ones unless indicated. Anomeric notation and destination linkage indicated without spacing/dashes.
- Modifications of monosaccharides (e.g., sulfation, O-acetylation) indicated by attached small letters, with numbers indicating linkage positions, if known.

Following these principles, a partial set of symbols could be established (Figure 7).**Hexoses:** Circles, **N-Acetylhexosamines:** Squares, **Hexosamines:** Squares divided diagonally

- **Glucose stereochemistry: BLUE (0,0,250)**
- **Galactose stereochemistry: Yellow (255,255,0)**
- **Mannose stereochemistry: GREEN (0,200,50)**
- **Fucose: RED (250,0,0)**
- **Xylose: ORANGE (250,100,0)**

Acidic Sugars (Diamonds)

- **Neu5Ac: PURPLE (125,0,125)**
- **Neu5Gc: LIGHT BLUE (200,250,250)**
- **Kdn: GREEN (0,200,50)**
- **GlcA: BLUE (0,0,250)**
- **IdoA: TAN (150,100,50)**
- **GalA: Yellow (255,255,0)**
- **ManA: GREEN (0,200,50)**

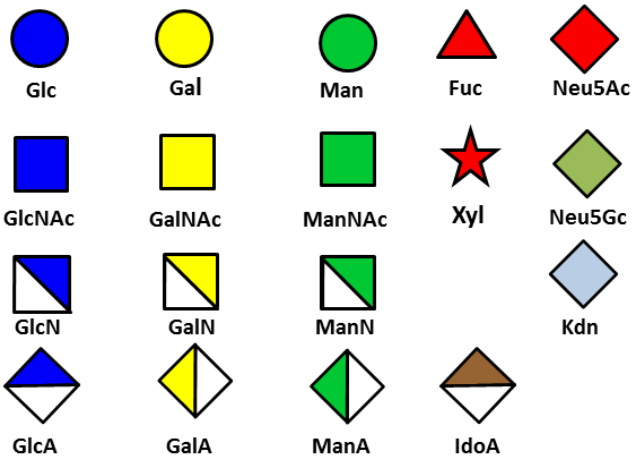
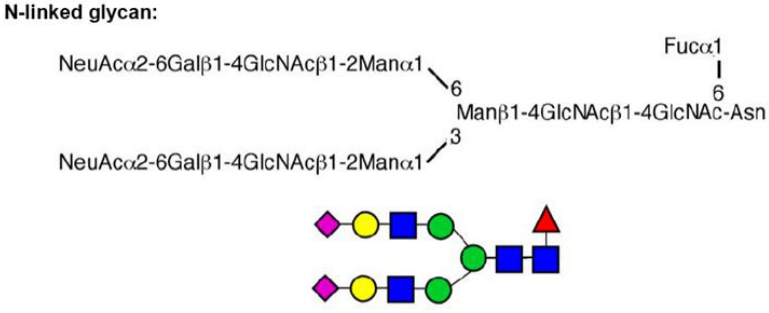


Figure 7. First generation of symbols proposed by the Consortium for Functional Glycomics (CFG).

The symbolic representation dealing essentially with mammalian glycans has gained wide acceptance and is being used to describe glycan composition and biosynthesis as shown in Figure 8.



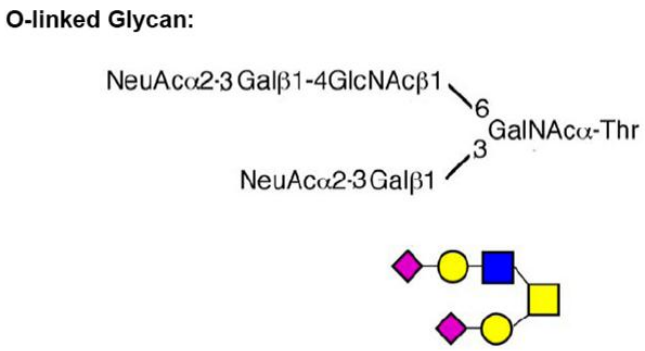


Figure 8. Typical structural and symbolic depictions of representative of N- and O-linked glycans.

The preliminary set of symbol notation for carbohydrates dealt with a limited number of monosaccharide units, which indeed provides a satisfactory coverage to describe the mammalian oligosaccharides which have been characterized so far.

Monosaccharides	Abundance (%)
D-GlcNAc	31.8
D-Gal	24.8
D-Man	18.9
Neu5Ac	8.3
L-Fuc	7.2
D-GalNAc	4.8
D-Glc	2.5
D-GlcA	0.3
D-Xyl	0.1
L-IdoA	0.1
Others	1.2

Table 1: Abundance of monosaccharide in mammalian carbohydrates, taken from Werz et al., 2007.

2.2. The Extension of the Symbol Notation to More Complex Glycans

The symbol notation has been extended to describe the structure of pathogen polysaccharides (Berger *et al.*, 2008) with the objective of providing a quick and easy way to visually distinguish between polysaccharides, for understanding cross-reactivity of sera on pathogen glycan arrays. Those glycans that are encountered in pathogens are more diverse, and contain greater structural variability than glycans found in mammals. There is a need to identify monosaccharides (by shape and color) and with the connectivity and substituent differences between polysaccharides. An extra level of complexity is found, occurring from the fact a diversity of monosaccharides with multiple modifications, which can occur both in the D- and L-configurations, and may be found in both pyranose and furanose forms.

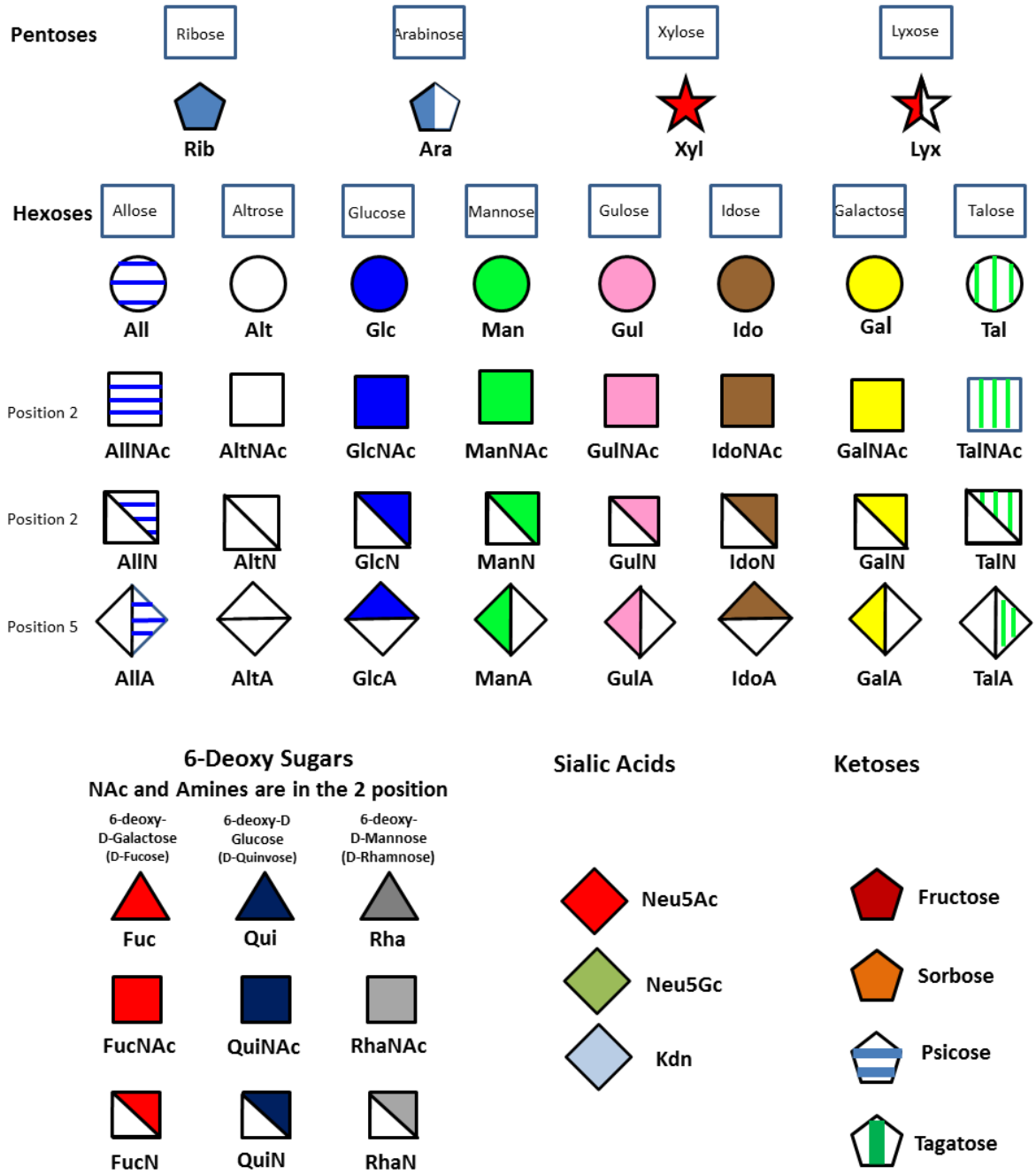


Figure 9: Extended Carbohydrate Symbol Notations of Pentoses and Hexoses, including, N-Acetyl and amines (in the two positions) and acids in the five position.

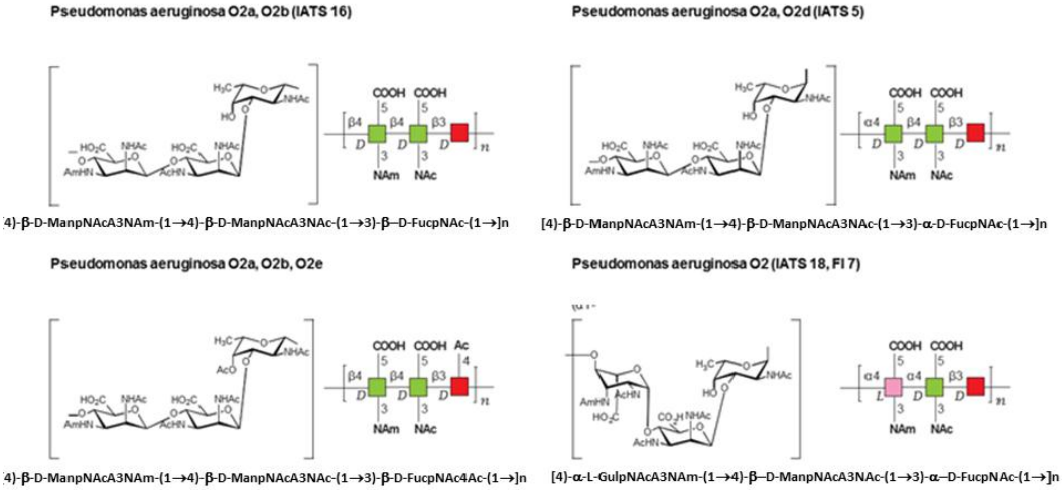
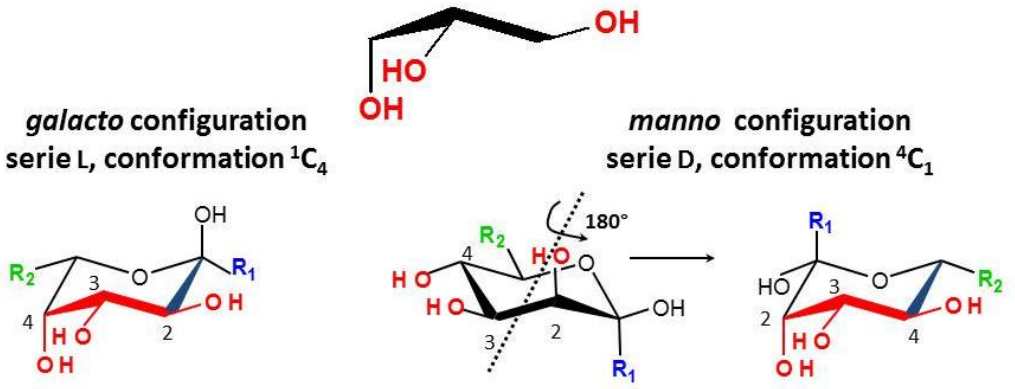


Figure 10. Because of their structural complexity of bacterial polysaccharides, the use of symbol notation provides a most obvious way to identify differences and similarities between the constituting monosaccharide units in the repeating sequences.

3.0 Beyond the Symbol Notation and Three-Dimensional Representation

It is fully recognized that no symbol system will ever convey a full appreciation of the three-dimensional structures of glycans, essential information for the understanding of how glycans and proteins interact (Varki *et al.*, 2008). This is clearly illustrated by the example given in Fig. 11 which shows the many monosaccharide units that can satisfy the stereochemical requirements resulting from the spatial arrangements of 3 contiguous hydroxyl groups, for recognition by the binding site of the PA-IIL lectin from *Pseudomonas aeruginosa* (Michell *et al.*, 2005).



Monosaccharide	R_1	R_2	Monosaccharide	R_1	R_2
α-L-Galactopyranose	H	CH_2OH	β-D-Mannopyranose	H	CH_2OH
α-L-Galacturonic acid	H	COOH	β-D-Mannuronic acid	H	COOH
α-L-Fucopyranose	H	CH_3	β-D-Rhamnopyranose	H	CH_3
β-D-Arabinopyranose	H	H	β-D-Lyxopyranose	H	H
β-D-Fructopyranose	CH_2OH	H	β-D-Tagatopyranose	CH_2OH	H

Figure 11 : Stereochemical requirements for binding Lectin PA II-L from *Pseudomonas aeruginosa.*

There is indeed a need to link the symbol notation to a graphical representation that maintains the stereochemical feature of each monosaccharide. The results of such an exercise have been reported (Berger *et al.,* 2008). Typical representations of symbol notations and graphical representations are given for cyclic furanose and pyranose forms (Figure 12), cyclic forms of hexopyranuronic acids (Figure 13), cyclic forms of 2-ketoses (Figure 14); three-representatives of 6-deoxy sugars, with NAc and N at the position 2 (Figure 15) and Neu5Ac, Neu5Gc, Kdo, Kdn, Heptose (Figure 16).

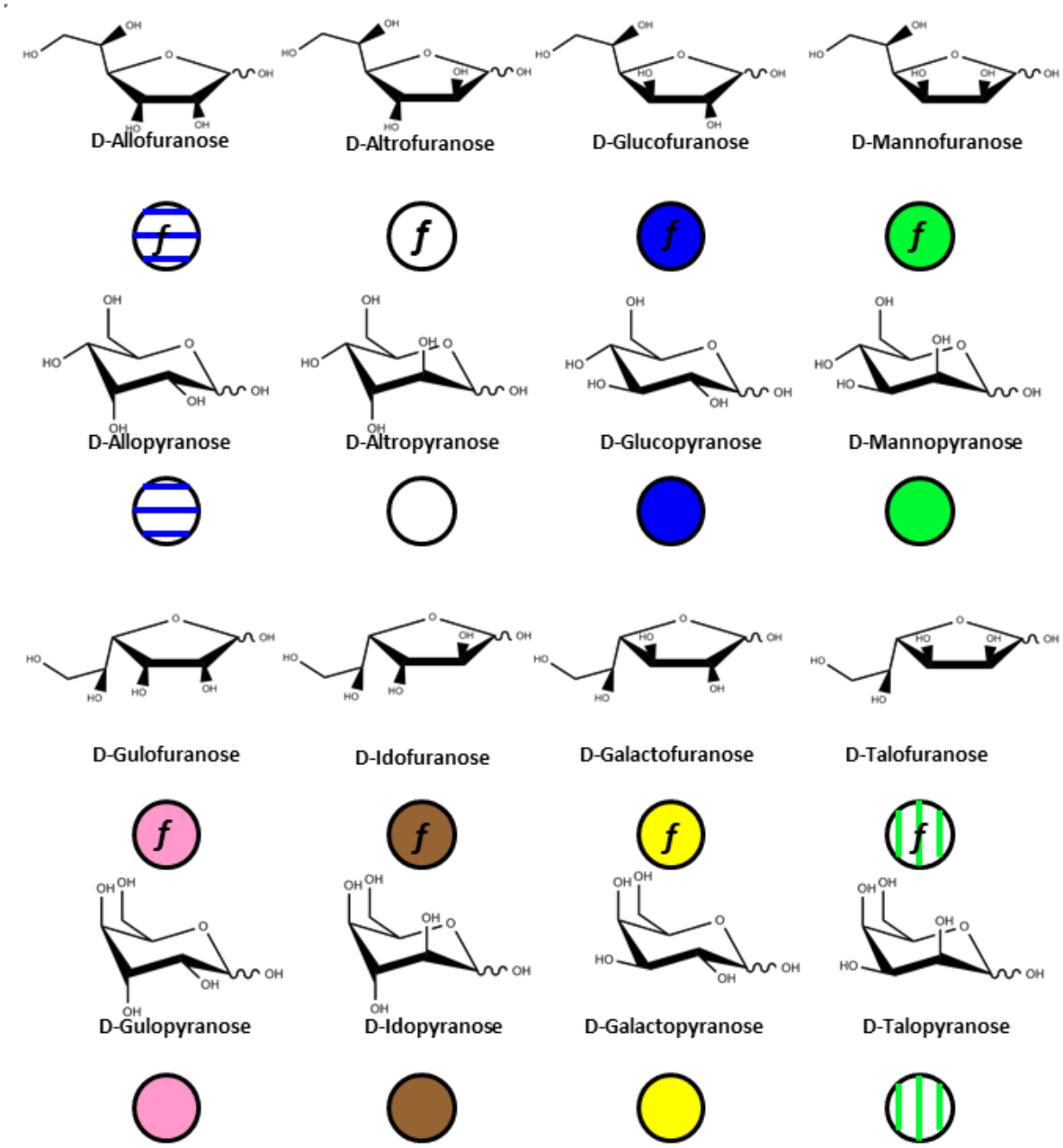


Figure 12. Three-dimensional and symbol representations of cyclic furanose and pyranose forms.

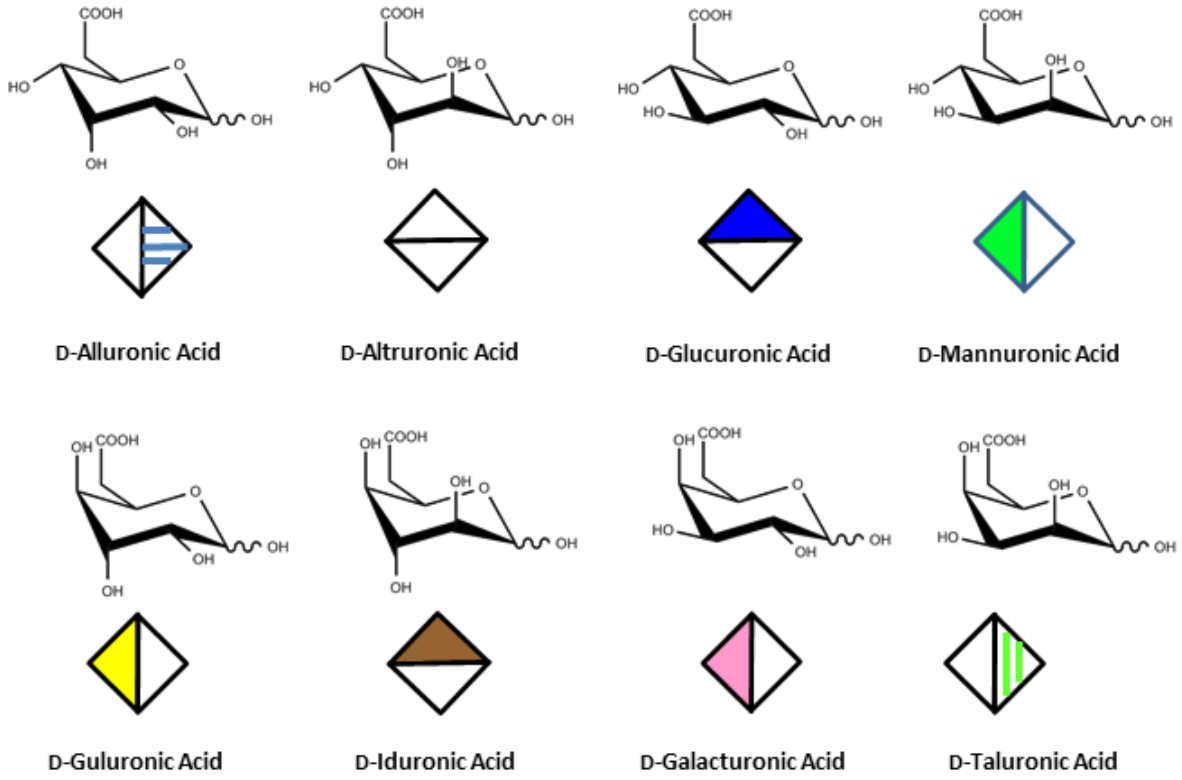


Figure 13: Three-dimensional and symbol representations of cyclic forms of hexopyranuronic acids.

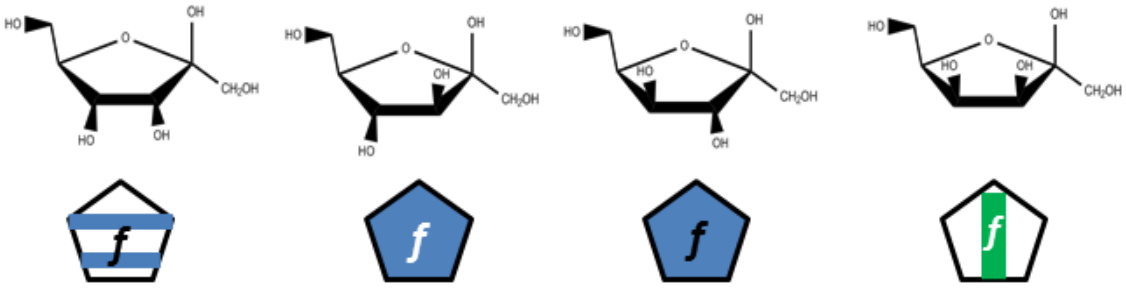


Figure 14: Three-dimensional and symbol representations of cyclic forms of 2-ketoses

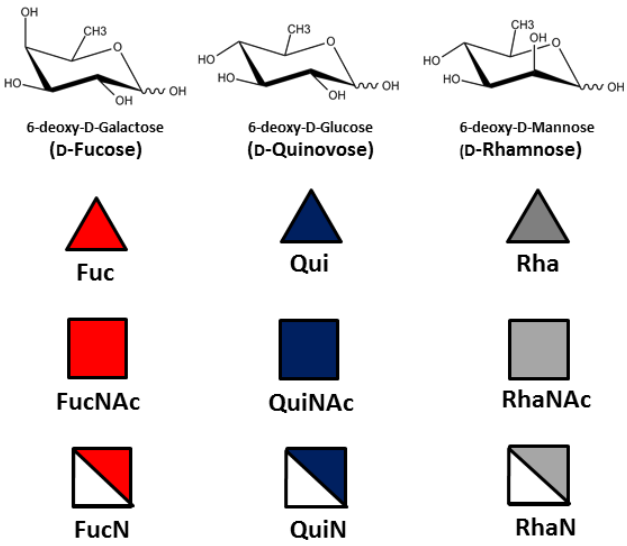


Figure 15: Three-dimensional and symbol representations of three-representatives of 6-deoxy sugars, with NAc and N at the position 2.

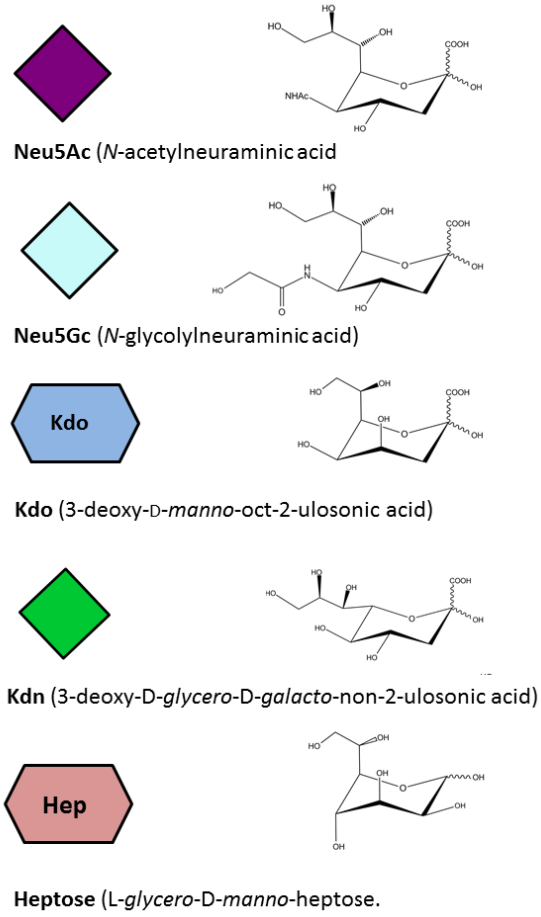


Figure 16: Three-dimensional and symbol representations of Neu5Ac, Neu5Gc, Kdo, Kdn and Heptose

3.2. Encoding and Representation

The most popular ways of encoding and representing glycans are: (i) Linear notation (e.g. IUPAC); (ii) Symbolic / diagrammatic notation e.g. Oxford and the notation proposed by the Consortium for Functional Glycomics (CFG).

It is nevertheless possible to concatenate the symbolic representation of monosaccharides with a limited number of structural descriptors to achieve a fairly exhaustive nomenclature which can be further used in constructing three dimensional structures of glycans. This can be achieved using the following set

Residue Letter Name: Rib, Ara, Xyl, Lyx, All, Alt, Glc, Man, Gul, Ido, Gal, Tal, ….and abbreviated trivial name.

[O-ester and ethers]: (when present) are shown attached to the symbol with a number, e.g.

6Ac for 6-*O*-acetyl group

3S for 3-*O*-sulfate group

6P for 6-*O*-phosphate group

6Me for 6-*O*-methyl group

36Anh for 3,6-anhydro

Pyr for pyruvate group

Absolute Configuration: D or L
The D-configuration for monosaccharide and the L configuration for Fucose and Idose are implicit and do not appear in the symbol. Otherwise the L configuration is indicated in the symbol, as in the case of Arabinose or L-Galactose.
For those occurring in the furanose form, a letter N or S is inserted in the symbol, indicating the northern (*N*) or Southern (*S*) conformation of the five membered ring.

Anomeric configuration: The nature of the glycosidic configuration (α or β) is explicitly set within the symbol.

Ring conformation
All pyranoses in the D-configuration are assumed to have $^{4}C_{1}$ chair conformation; those in the L configuration are assumed to have $^{1}C_{4}$ chair conformation. Otherwise, the ring conformation is indicated in the symbol, as $^{2}S_{0}$ in the case of α-L-Idopyranose.
N or *S* indicates the conformation of the five membered rings on the conformational wheel.

While maintaining the spirit of using the symbolic representation for monosaccharides (and towards glycans) this set of rules provides the necessary extension to the construction of three-dimensional structures, allowing encoding for computational manipulation whilst maintaining IUPAC nomenclature.

As a result of an extensive search, 120 monosaccharides have been identified as the building blocks of the vast majority of oligosaccharides, complex carbohydrates such as “glycan determinants” (blood group antigens, core structures, fucosylated oligosaccharides, sialylated oligosaccharides, Lewis antigens, GPI-anchors, N-linked oligosaccharides, globosides, ….), glycosaminoglycans, plant and algal polysaccharides as well as some bacterial polysaccharides. Each monosaccharide can be assigned a symbolic representation in accordance with the rules set forward, and its cyclic structure represented. (Figure 17).

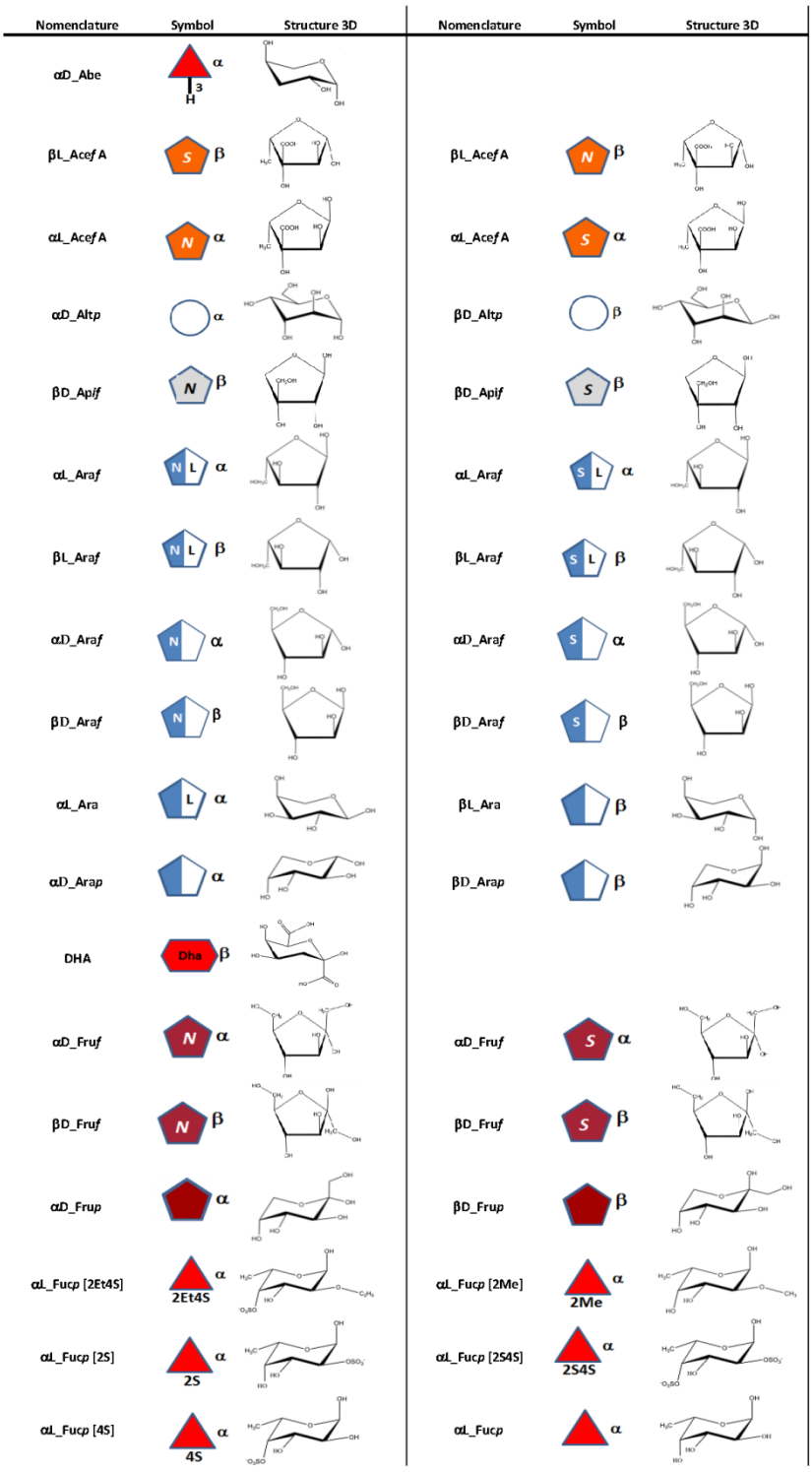


Figure 17 : The symbolic and structural representation of 120 monosaccharides, constituents of the vast majority of glycans.

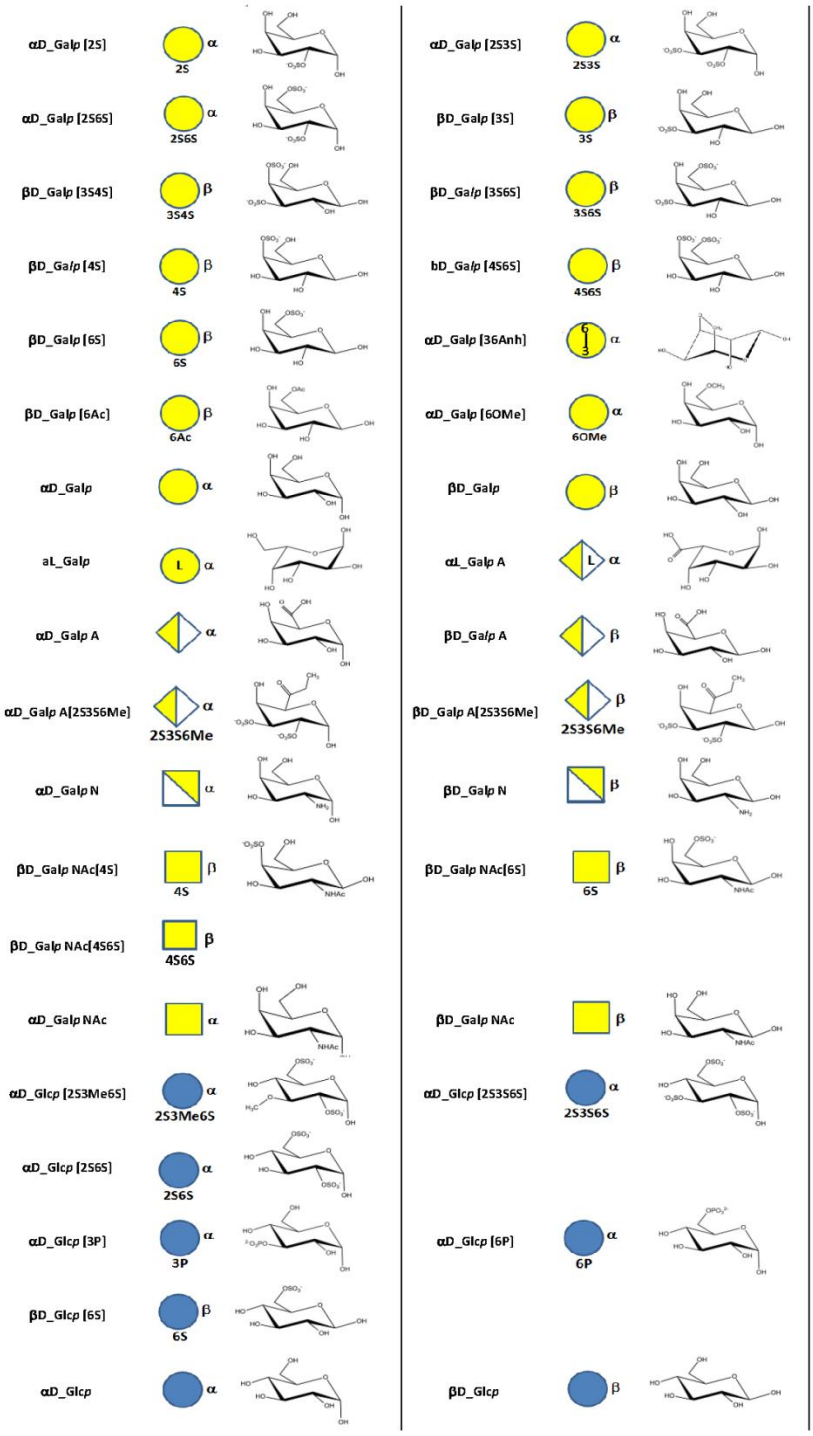


Figure 17 : The symbolic and structural representation of 120 monosaccharides, cosntituents of the vast majority of glycans.

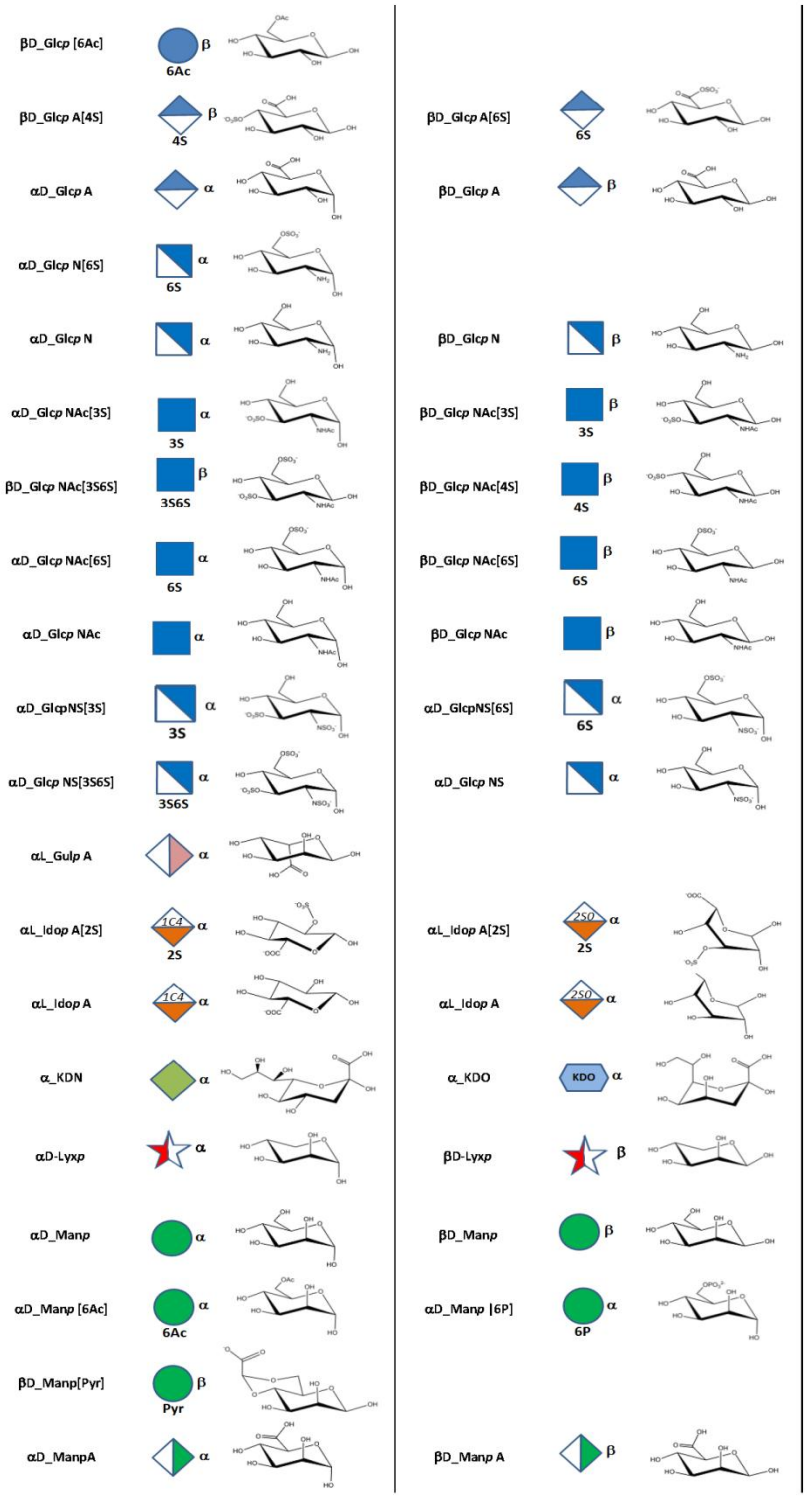


Figure 17 : The symbolic and structural representation of 120 monosaccharides, constituents of the vast majority of glycans.

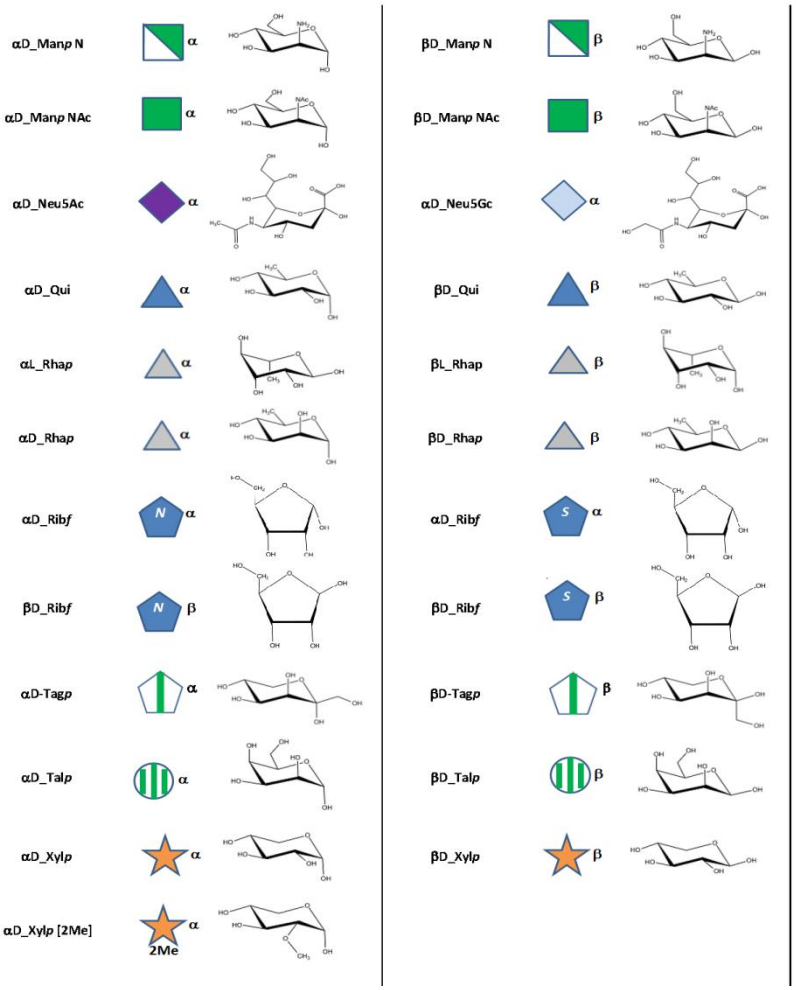


Figure 17 : The symbolic and structural representation of 120 monosaccharides, constituents of the vast majority of glycans.

This information is part of the body of an annotated data base that contains the three-dimensional structures of the low energy conformations. They have been subjected to systematic conformational sampling to determine their conformational preferences, using molecular mechanics optimization. Whereas most of the monosaccharides exhibit a fairly rigid ring conformation some cases exist where several ring shapes can occur such as in iduronic acid, idose, and all furanosides. In these cases, the low energy conformations are available for each entry. This set of structural information is required to construct three-dimensional structures of complex glycans (ref: Glyco3d) and provides the required input to the Polys builder (Engelsen *et al.,* 2013).

References

Berger, O., McBride, R., Razi, N. & Paulson, J. (2008) Symbol Notation Extension for Pathogen Polysaccharides, The Scripps Research Institute, Consortium for Functional Glycomics.

Bijvoet, J.M., 1951, Peerdeman, A.F. van Bommel, A.J., Determination of the Absolute Configuration of Optically Active Compounds by Means of X-rays, Nature, 168, 271-272.

Consortium for Functional Glycomics. : http://www.functionalglycomics.org

Fischer, E., (1890) Ber., 23, 2114

Glyco3D, A Site for Glycosciences: http://www.glyco3d.cermav.cnrs.fr

Engelsen, S.B, Hansen, P.I., Perez, S. (2013) POLYS: An Open Source Software Package for Building Three-Dimensional Structures of Polysaccharides, Biopolymers.

International Union of Pure and Applied Chemistry (IUAPC) Nomenclature Home Page; http://www.chem.qmw.ac.uk/iupac

International Union of Biochemistry and Molecular Biology (IUBMB) Nomenclature Home Page; http://www.chem.qmw.ac.uk/iubmb

Kamerling, J.P. (2007) Basic Concepts and Nomenclature Recommendations in Carbohydrate Chemistry, in *Comprehensive Glycosciences, From Chemistry to System Biology*, Vol. 1, pp. 1-37, Kamerling, J.P., Ed., Elsevier.

Lichtenthaler, F.W. (2002) Emil Fischer, His Personality, His Achievements, and his Scientific Progeny, Eur. J. Org. Chem., 24, 4095-4122.

McNaught, A. (1997) Nomenclature of Carbohydrates (Recommendations 1996). Adv. Carbohydr. Chem. Biochem., 52, 43-177; International Union of Pure and Applied Chemistry and International Union of Biochemistry and Molecular Biology. Joint Commission on Biochemical Nomenclature. Nomenclature of Carbohydrates. Carbohydr. Res., 297, 1-92.

Michell, E.P., Sabin, C., Snajdrova, L., Budova, M., Perret, S., Gautier, C., Hofr, C., Gilboa-Garber, N., Koca, J., Wimmerova, M., Imberty, A. (2005) High Affinity Fucose Binding of PA-IIL 1.0 Ang. Resolution Crystal Structure of the Complex Combined with Thermodynamics and Computational Chemistry Approaches, Proteins. Struct. Funct. Bioinfo, 58, 735-746.

Varki, A. (2008) Historical Background and Overview, in *Essentials of Glycobiology*,(2nd edition, pp. 784. Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Hart, G.W., Etzler, M.E., eds., Cold Spring Harbor Laboratory Press.

Werz, D.B., Ranzinger, R., Herget, S., Adibekian, A., von der Lieth, C.-W., Seeberger, P.H. (2007) Exploring the Structural Diversity of Mammalian Carbohydrates ("'Glycospace") by Statistical DataBank Analysis, ACS Chemical Biology, 2, 685-691.