Part One
Chemical Basis

The Sugar Code. Edited by Hans-Joachim Gabius

ISBN: 978-3-527-32089-9

1
The Biochemical Basis and Coding Capacity of the Sugar Code

Harold Rüdiger and Hans-Joachim Gabius

Teaching the biochemistry of carbohydrates is not simply an exercise in terminology. It has much more to offer than commonly touched upon in basic courses, if we deliberately pay attention to the far-reaching potential of sugars beyond energy metabolism and cell wall stability. In fact, then there is no reason why complex carbohydrates should shy at competition with nucleic acids and proteins for the top spot in high-density biocoding. On the contrary, sugars have ideal properties for this purpose, as will be concluded at the end of this chapter. In this sense, an obvious explanation why research in glycosciences (structural and functional glycomics and lectinomics) has lagged behind the fields of genomics and proteomics, also in the public eye, is 'that glycoconjugates are much more complex, variegated, and difficult to study than proteins and nucleic acids' [1]. What is a boon for decorating cell surfaces with a maximum number of molecular messages at the same time has been and still is a demanding challenge for analytical and synthetic chemistry (please see Chapters 3–5 for details on how to address it properly). That the sugar code can give up its secrets rather easily depends on a solid understanding of the unique aspects of its biochemical basis. With hindsight, E. Chargaff's rule on the regularities of nucleotide composition, derived from biochemical analysis of DNA in 1949/1950, was conspicuously more than a descriptive parameter. Its ratio of unity signified base complementarity. It is thus worthwhile to carefully examine basic carbohydrate biochemistry to define what makes sugars genuinely special in chemical terms. In doing so, we guide readers from the etymological roots of frequently used terms to raising awareness for what common structural depictions tell us about information coding by sugars.

1.1
Etymological Roots

Elementary analysis of hexoses revealed presence of carbon (Latin '*carbo*' = 'coal') and water (Greek '*hydor*' = 'ὕδωρ') in a stoichiometric proportion, that is, $C_n(H_2O)_m$, with $n \geq m$. This result explains the origin of the term 'carbohydrates'. Its synonym

The Sugar Code. Edited by Hans-Joachim Gabius

ISBN: 978-3-527-32089-9

'sugar' goes back to the Sanskrit word '*sarkar*', which means sugar cane and its product. This word has entered many languages (for example Persian, Greek, Latin and Arabic). The Greek variant '*saccharon*' ('σάκχαρον') and its Latin form '*saccharum*' have led us to designate sugars as 'saccharides'. When using 'glycoside', the Greek term '*glykys*' ('γλυκύς'), translated into 'sweet', is the etymological root.

The most abundant hexose is glucose (Glc) or grape sugar. Its name is derived from the Latin loanword 'glucus'. This sugar and its 2-deoxy-2-acetamido derivative *N*-acetylglucosamine (GlcNAc) are the building blocks for rigid cell walls, that is, the polymers cellulose and chitin (please see Chapter 12 for details). GlcNAc is ubiquitously present in glycosaminoglycan chains and peptidoglycans as well as in glycoconjugates such as glycoproteins with *N*- and *O*-glycans (please see Chapters 6–11 and 29). De-*N*-acetylation without proper sulfation results in the occurrence of small amounts of glucosamine (GlcN) in glycosaminoglycan chains (please see Chapter 11). Historically, the scientific detection of the natural presence of GlcN and its derivative GlcNAc dates back to a fortunate menu selection in 1875 (see Info Box 1 and also Info Box 1 in Chapter 12). An answer why glucose and its derivatives are so abundant in Nature is given in the next paragraph by closely looking at the different projection formulas.

Info Box 1

'In 1875 a young physician named Georg Ledderhose was working during the summer semester in the laboratory of Friedrich Wöhler in Göttingen when Ledderhose's uncle, Felix Hoppe-Seyler, a noted physiological chemist, invited him to dinner. At his uncle's suggestion he took the remains of the lobster they had eaten back to the laboratory, where he found that the claws and the shell dissolved in hot concentrated hydrochloric acid and that on evaporation the solution yielded characteristic crystals. He soon identified the crystalline compound as a new nitrogen-containing sugar, which he named *glycosamin*' [N. Sharon. Carbohydrates. *Sci Am* 1980; **243**, 80–97].

Note: F. Hoppe-Seyler was the founder of the *Zeitschrift für physiologische Chemie* in June 1877, which today is known as *Biological Chemistry*.

1.2 What Projection Formulas Tell Us

The classical chain structure introduced by E. Fischer figures how the $(H_2O)_m$ molecules are distributed over the carbon backbone. A series of hydroxy groups is present together with an aldehyde function at C1 (Figure 1.1a). Of note, their high density establishes a platform with exceptional properties (please see below). When looking at experimental numbers in Figure 1.2, it becomes obvious that the chain structure will not at all be the predominant form of a carbohydrate.

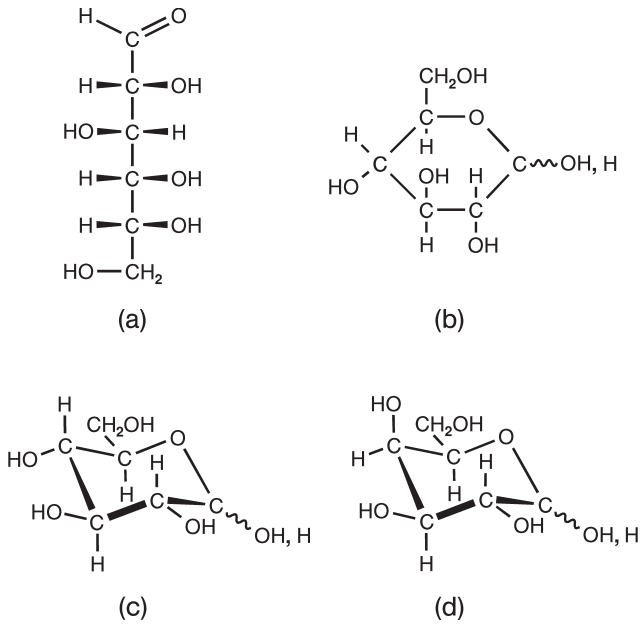

Figure 1.1 Illustration of the two types of projection formulas and the chair-like conformation of D-glucose. The open-chain (Fischer) (a) and hexopyranose (Haworth) projection formulas (b) as well as the $^4C_1$ low-energy chair-like pyranose conformation (c) are presented. Structural variability at the anomeric center (α or β) is symbolized by a wavy line. For further information on assignment of anomeric positions and contributions of pyranose and open-chain forms to the equilibrium, please see Figure 1.2 and its legend. Epimer formation from D-glucose (c) to D-galactose (d) leads to the axial positioning of the 4-hydroxy group in D-galactose and changes in the topological signature of hydroxy and polarized C—H groups.

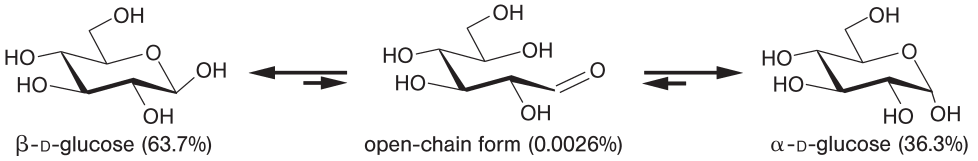

Figure 1.2 Illustration of the equilibrium including the two anomeric forms of D-glucose. The percentages of presence of the two anomeric hexopyranose and the open-chain forms in equilibrium are given in the bottom line.

As early as 1883, B. Tollens demonstrated that monosaccharides fail to react in common tests for free aldehydes. This observation intimates an intramolecular reaction. Its product is depicted in the Haworth projection in Figure 1.1b. This structure, in chemical terms a semialdehyde, is a derivative of the heterocyclus pyran, which explains the term 'pyranose'. Therefore, a monosaccharide symbol can be accompanied by an italic '*p*'. However, it is often omitted, since most of the physiologically important hexoses occur as pyranoses. Like cyclohexane, the

pyranose ring is not planar, but adopts a low-energy chair-like conformation (Figure 1.1c). Each substituent can assume either an axial or an equatorial position. Thus, the chemically equivalent groups are subdivided into geometrically different constellations and these are energetically not identical in the chair conformation. Owing to the free rotation around a single bond, an axial substituent can 'collide' with other axial groups in its vicinity (1,3-diaxial 'clashes'). Evidently, the lowest energy level is attained for a pyranose if all substituents larger than a hydrogen atom reside in an equatorial position and this formula represents glucose, the most abundant sugar in Nature (Figure 1.1c). Placing a plane through carbon atoms C2, C3, C5 and the ring oxygen atom readily explains why this conformation is referred to as $^4C_1$ (C = chair) (for further information on conformational flexibility of the pyranose ring, please see Figure 2.1). It harbors a characteristic topological signature of hydroxy groups, ready for directional hydrogen bonds in protein/carbohydrate–carbohydrate recognition or coordination bonds with Ca^{2+} (please see Chapters 13, 16 and 21). The remaining ambiguity at the C1 position, the anomeric center, is clarified with the reaction mechanism given in Figure 1.2.

All D-sugars can present this hydroxy group in either axial (α) or equatorial (β) positions. Mutarotation (Lat.: '*mutare*' = 'to alter'; alteration of optical rotation using polarized light) is the consequence of the equilibrium between α- and β-anomers, when starting measurements with a pure anomer. This phenomenon was first observed in 1846 by A.P. Dubrunfaut, who discovered D-fructose one year later. The geometry at the C1-atom of a hexopyranose has not only a bearing on optical properties. It also determines the shape of disaccharides and thus enhances total coding capacity (please see below). What happens if a hydroxy group other than at the anomeric center changes position? Is this only a subtle change?

This process yields epimers. Epimerization will entail emergence of the mentioned 1,3-diaxial clashes. Being a source of structural destabilization, their number will most likely be restricted to a minimum in natural glycans, that is to one in the 4-epimer galactose (and also the 2-epimer mannose) as shown in Figure 1.1d. By the way, 3-epimerization of D-glucose to D-allose leads to two clashes, whereas the 5-epimer L-idose can avoid them by adopting the $^1C_4$ conformation [please see below: case study of D-glucuronic acid (GlcA) versus its natural 5-epimer L-iduronic acid (IdoA)]. Beyond 1,3-diaxial contacts and, of course, the alteration of the signature of hydroxy group presentation, a further parameter is automatically affected in epimers: the spatial distribution of the positively polarized C—H bonds. When comparing the presentation of C—H bonds in D-glucose/D-galactose one should pay attention to the constellation between the C3 and C5 atoms (Figure 1.1c and d). The 4-epimer has a contiguous stretch of these C—H bonds, inviting polar contact with π-electrons. This patch is ideally suited for C—H/π interaction. On these grounds the presence of an aromatic residue in proteins, preferably of tryptophan, is predicted in binding sites of proteins specific for D-galactose (please see Figure 13.1 for the answer to this question). In sum, epimerization – together with derivative formation – accounts for the origin of many constituents of natural glycans.

A final point arising from inspecting the structures in Figure 1.1 is versatility for oligomer formation. In comparison to the phosphodiester and peptide bonds,

strictly constrained in this respect, a large panel of glycosidic linkages can be formed. This unique property is visualized by arrows in Figure 1.3. That this statement is not of solely academic value is underscored by the natural occurrence in honey and other sources of all theoretically possible diglucosides (Table 1.1). In forming a glycosidic bond, the anomeric center of one partner is always involved. Even if the linkage points (for example 1–4) are identical, the decision on the anomeric position will markedly bear upon the biochemical properties of the

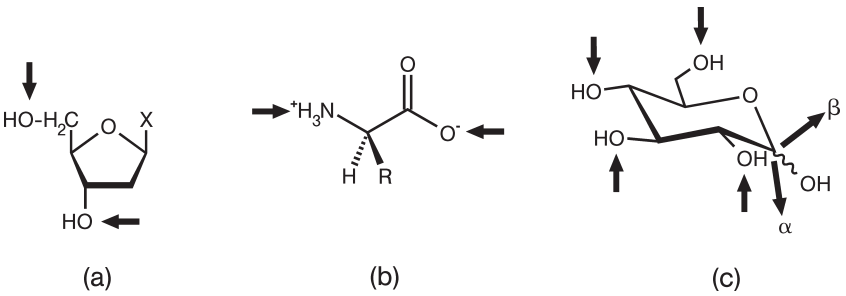

Figure 1.3 Illustration of the linkage points for oligomer formation in biomolecules by arrows. The phosphodiester bond in nucleic acid biosynthesis (a) and the peptide bond in protein biosynthesis (b) yield linear oligomers. In contrast, the glycosidic linkage in oligosaccharides can involve any hydroxy group, opening the way to linear and also branched structures (c) (for an example of branching, please see Figure 1.5). Using D-glucose (please see Figure 1.1c) as an example, its active form UDP-Glc allows conjugation of this sugar to carbohydrate acceptors to any hydroxy group, as symbolized by arrows directed towards the hydroxy groups (for a list of resulting diglucosides, please see Table 1.1). The anomeric position in chain elongation can vary, as symbolized by two bold arrows pointing away from the molecule (for structures of the anomers and the two 1–4-linked diglucosides with different anomeric positions, please see Figures 1.2 and 1.4).

Table 1.1 Naturally occurring disaccharides formed from two glucose units.

Type of linkage	Common name
α1–2	Kojibiose
β1–2	Sophorose
α1–3	Nigerose
β1–3	Laminaribiose
α1–4	Maltose
β1–4	Cellobiose
α1–6	Isomaltose
β1–6	Gentiobiose
α1–1'α	Trehalose

All disaccharides are conversion or degradation products of natural polysaccharides and glycosides, except for trehalose, which is present in bacteria, fungi and insects.

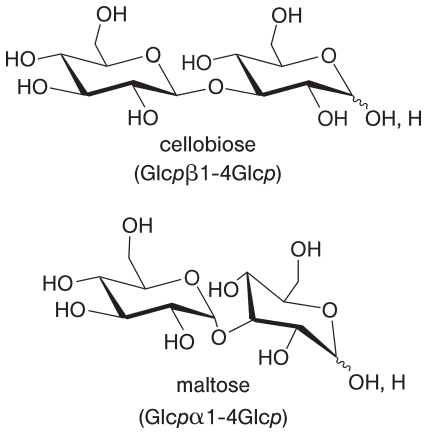

Figure 1.4 Illustration of the two 1–4-linked diglucosides cellobiose and maltose. The α- and β-anomers of D-glucose (please see Figure 1.2 for structures) produce diglucosides with different shapes, underscoring spatial consequences of anomer selection.

product. Cellulose and glycogen/starch are telling examples, and Figure 1.4 illustrates the structures of the respective α/β1–4-linked diglucopyranosides. Taken together, the projection formulas tell us that (i) hydroxy groups in a hexapyranose are presented in high density with implications for hydrogen and C—H/π bonding as well as oligomer formation, and (ii) distinct hexoses will be energetically favored (glucose and its epimers mannose and galactose). What are the actual consequences of these structural aspects for the coding capacity of oligosaccharides?

1.3 The Coding Capacity of the Sugar Code

A quantitative measure of this characteristic is the total number of 'words' (isomers) built from a set of 'letters' (monomers). As deduced from Figure 1.3, calculations to define the range of structural permutations are simple in the cases of nucleotides and amino acids. It is only the linear sequence that counts to master this task. In order to completely define a disaccharide structurally, however, it is not sufficient to determine just the sequence, as already explained above. Beyond this parameter, the linkage points, the anomeric position and also the ring size (pyranose/furanose) must be known. These attributes distinguish carbohydrates from nucleotides and amino acids, and there is more.

Compactness of structural units is achieved by branching. Toward this end, a monosaccharide will then engage more than two hydroxy groups for glycosidic linkages. Figure 1.5 shows a classical example for branched oligosaccharides, that is, the ABH(0) histo-blood group epitopes. Delineation of their biochemical nature was aided by application of an eel lectin (see Info Box 2; its crystal structure is shown in Figure 16.1g), their structures are given in Figure 1.5. Branching is also

Info Box 2

'An early observation of Landsteiner had established that for artificial antigens a simple substance with a structure closely related to, or identical with, the immunologically determinant (haptenic) group of the antigen can combine with the antibody and thereby competitively inhibit the reactions between antigen and antibody. Although this principle is employed today in many forms of inhibition tests, in the 1950s it had not been used to find the determinants in naturally occurring antigens. By 1952 we had accumulated a large selection of anti-H reagents of human and animal origin and we decided, with no great expectation of the outcome, to screen them for inhibition of the agglutination of 0 cells with the component sugars present in the blood-group active substances. Somewhat to our surprise one of the many reagents, that from the eel, *Anguilla anguilla,* was quite strongly inhibited by L-fucose and to a greater extent by α-methyl L-fucoside and not by the other monosaccharides. Our conclusions were somewhat tentative at first because this was an isolated result with a rather exotic reagent, but the inference that L-fucose in α-linkage is more important than the other sugars for H specificity was reinforced when we were given some plant agglutinins (later called lectins)...' [W.M. Watkins. A half century of blood-group antigen research: some personal recollections. *Trends Glycosci Glycotechnol* 1999; **11**, 391–411; for illustration of the folding of the eel agglutinins, please see Figure 16.1g].

a common feature of *N*- and *O*-glycan structures (please see Chapters 6–8), and has a bearing on bioaffinity in receptor (lectin) binding [2]. Consideration of the factors of sequence, of linkage-point and ring-size permutations as well as branching sets glycans far apart from nucleic acids and proteins in terms of coding capacity. In actual numbers, only 4096 (4^6) hexanucleotides are possible with the four letters in the DNA language and still 6.4×10^7 (20^6) hexapeptides from 20 proteinogenic amino acids, but the staggering number of 1.44×10^{15} hexasaccharides from 20 monosaccharides [3]. Even though not every combination is realized, since oligomer synthesis is confined to using exclusively the anomeric center of the activated donors (please see also Table 1.1), the case for an enormous potential

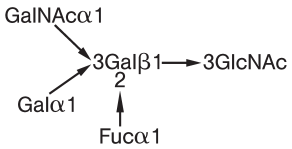

Figure 1.5 Illustration of the linkage pattern in ABH(0) histo-blood group tri- and tetrasaccharides. The core H(0)-trisaccharide (type I: α1–2-fucosylated Gal*p*β1–3GlcNAc*p*), whose L-fucose part is freely accessible to the eel lectin (please see Info Box 2), can be extended in α1,3-linkage by either *N*-acetylgalactosamine (A epitope) or galactose (B epitope). A branched structure is generated, as intimated by arrows in Figure 1.3. For structures of the individual 'letters' of the ABH(0) 'words', please see Figure 1.6.

of glycans as bioinformatic toolbox is nonetheless convincing. The basic alphabet of the sugar language is comprised by a set of 'letters'. The structures, symbols and acceptor characteristics of common monosaccharides are compiled in Figure 1.6. In addition to building up glycans of cellular glycoconjugates (glycoproteins, glycolipids and proteoglycans) and also cell walls (please see Chapters 6–12 for details), saccharides can also enter low-molecular-weight molecules. An instructive example from basic biochemistry is the conjugation of two GlcA molecules to bilirubin. This glycosylation enhances solubility and in consequence promotes excretion of this otherwise hardly soluble compound. Glycosylation processes of this type are also fairly common in plants, for example, concerning alkaloids, cyanohydrins, phenolics or terpenoids. By doing so, a wide variety of plant glycosides is generated such as the famous *Digitalis* compounds, opening a fertile field for synthetic glyco-randomization with the therapeutic/biotechnological aims of glyco-optimization [4]. All in all, the monosaccharides in Figure 1.6 establish the third alphabet of life.

The modification of nucleotides and amino acids is a powerful means to enlarge the size of a basic alphabet. Akin to posttranslational protein phosphorylation or sulfation glycan epitopes, too, are subjected to such reactions in order to convey particular properties to specific sites [5, 6]. Examples of how the cores for routing signals to lysosomes and to endothelial cells in the liver or for a determinant of cell communication in the nervous system look like are given in Figure 1.7 (top panel) (for further information on lectins binding these epitopes, please see Chapters 19.3 and 30.7 as well as Figure 16.1j). Beyond mammalian biochemistry, sulfation of egg jelly glycans at fucose residues is relevant for fertilization in invertebrates, this topic explained in Chapter 24.

Illustration of the anticoagulant pentasaccharide of heparin serves a second purpose besides documenting presence of sulfations, especially the rare 3-*O*-sulfation in the structure's center (Figure 1.7, bottom panel). It gives an impression of the ingenious combinatorial way to turn a seemingly dull repetition of the basic disaccharide [-4GlcNAcα1–4GlcAβ1-]$_n$ into an amazing structural (heparanomic) complexity. In fact, 48 different dimers are possible. As alluded to above, the epimerization from D-GlcA to L-IdoA (for structures, please see Figure 1.6) has significant consequences, because IdoA is an ideal hinge for shape rearrangements [7]. The conformational flexibility of the $^1C_4$ form to adopt a ^2S_O skew-boat structure (Figure 1.6) underlies this property (for further details on designation of ring geometry, please see next chapter, and on proteoglycans and drug design with heparin, please see Chapters 11 and 28.5). As current research is revealing, these substitutions are by no means a rare or random event. Thus, the enzymatic machinery for glycan assembly, processing and remodeling as well as site-specific substitution is elaborate and well developed, as expected for professional tasks in information handling. Whereas only two sulfotransferases are assigned to protein substitution in the human Golgi region, a total of 35 enzymes adds sulfate groups to carbohydrates [8]. The overall investment in genomic coding pays off by producing the glycomic complexity. It is not template derived and can be adjusted dynamically by the availability of enzymes, substrates and acceptors in space and

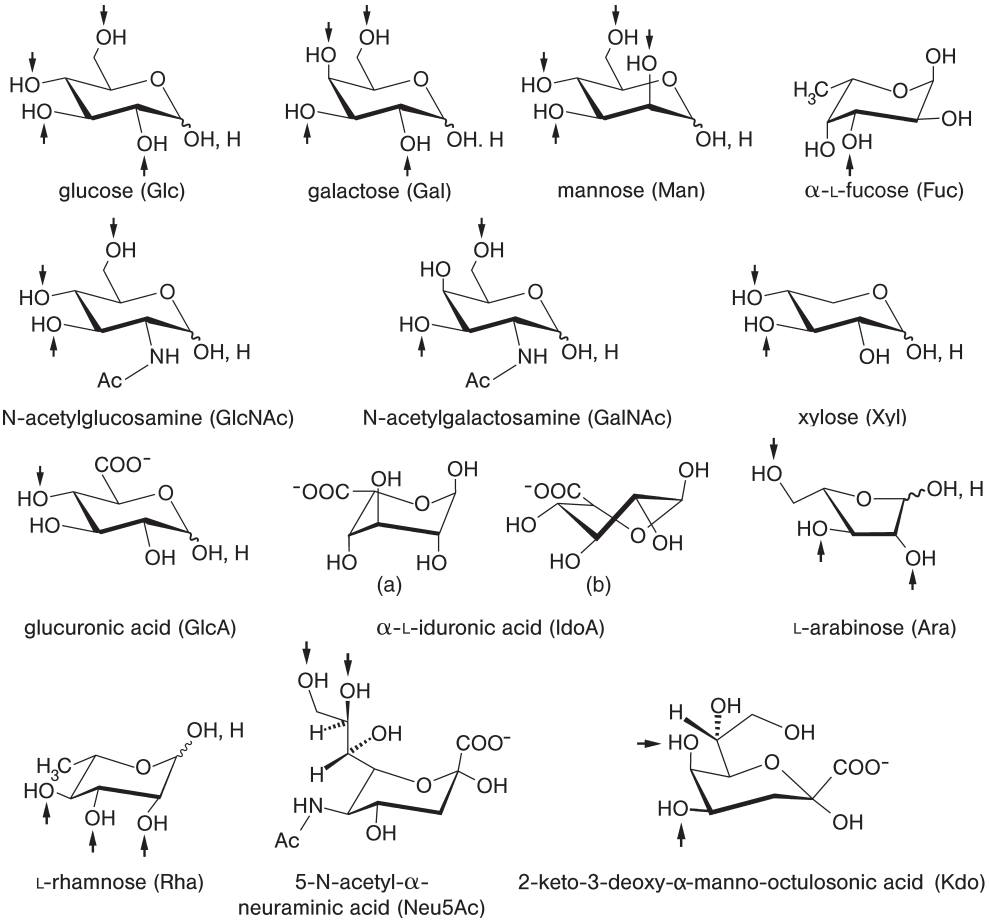

Figure 1.6 Illustration of the alphabet of the sugar language. Structural representation, name and symbol as well as the set of known acceptor positions (arrows) in glycoconjugates are given for each letter. Four sugars have L-configuration: fucose (6-deoxy-L-galactose), rhamnose (6-deoxy-L-mannose) and arabinose are introduced during chain elongation, whereas L-iduronic acid (IdoA) results from postsynthetic epimerization of GlcA at C5. The $^{1}C_{4}$ conformation of IdoA (a) is in equilibrium with the $^{2}S_{O}$ form (b) in glycosaminoglycan chains where this uronic acid can be 2-sulfated (please see Figure 1.7d). All other 'letters' are D-sugars. Neu5Ac, one of the more than 50 sialic acids, often terminates sugar chains in animal glycoconjugates. Kdo is a constituent of lipopolysaccharides in the cell walls of Gram-negative bacteria, and is also found in cell wall polysaccharides of green algae and higher plants. Foreign to mammalian glycobiochemistry, microbial polysaccharides contain the furanose ring form of D-galactose and also D/L-arabinose indicated by an italic '*f*' derived from the heterocyclus furan. The α-anomer is prevalent for the pentose arabinose, for example, in mycobacterial cell wall arabinogalactan and lipoarabinomannan. β1–5/6-Linked galactofuranoside is present in the arabinogalactan and the β1–3/6 linkage in lipopolysaccharides.

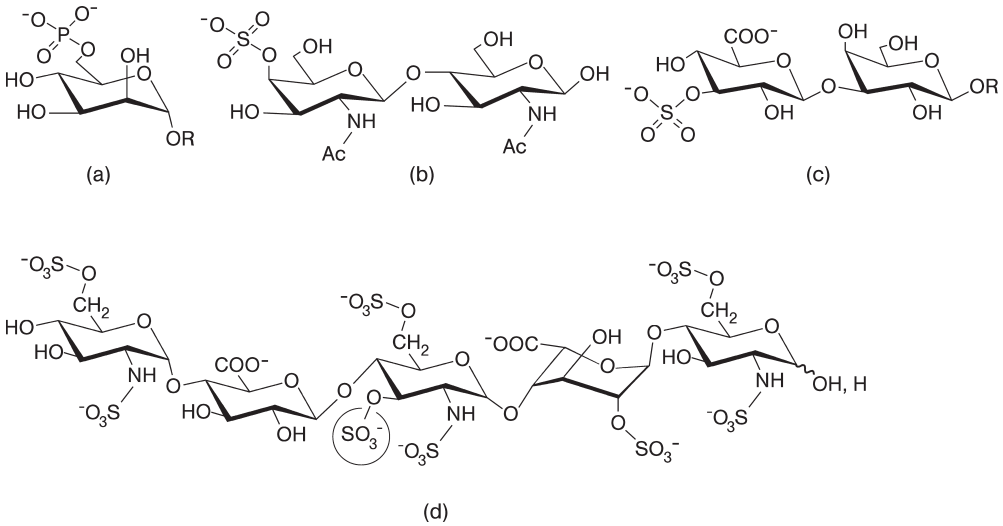

Figure 1.7 Illustration of phosphorylated (phosphated) and sulfated (sulfurylated) glycan 'words'. 6-Phosphorylation of a mannose moiety (in the context of a mannose-rich pentasaccharide) is the key section of a routing signal in lysosomal enzymes (a), 4-sulfation of the GalNAcβ1–4GlcNAc (LacdiNAc) epitope forms the 'postal code' for clearance from circulation by hepatic endothelial cells of pituitary glycoprotein hormones labeled in such a manner (b), the HNK (human natural killer)-1 epitope (3-sulfated GlcAβ1–3Galβ1–4GlcNAc) is involved in cell adhesion/migration in the nervous system (c) and the encircled 3-*O*-sulfation in the pentasaccharide's center is essential for heparin's anticoagulant activity (d). All sugars are in their pyranose form. Please note that the central GlcN unit has *N*,*O*-trisulfation and that the 2-sulfated IdoA, given in the $^{1}C_{4}$ conformation, can also adopt the hinge-like $^{2}S_{O}$ skew-boat structure (please see Figure 1.6; about 60% or more for the $^{2}S_{O}$ form in equilibrium depending on the structural context) when present within glycosaminoglycan chains of the proteoglycan heparin. 2-Sulfation of IdoA serves two purposes: favoring the hinge-like $^{2}S_{O}$ conformation and precluding reconversion to GlcA.

time [5, 9], and its structural and functional aspects are the topic of the following chapters. That said, we have built the evidence for the following conclusions.

1.4 Conclusions

'Carbohydrates are ideal for generating compact units with explicit informational properties, since the permutations on linkages are larger than can be achieved by amino acids, and, uniquely in biological polymers, branching is possible. Moreover, the oligosaccharide units are not flexible but exhibit highly specific structures with only limited degrees of freedom' [10]. This statement highlights that the sugar code has a third dimension. What this means is explained in the next chapter and its relevance for protein–carbohydrate interactions is outlined in Chapter 13.

Summary Box

Carbohydrates form the third alphabet of life. Compared to amino acids and nucleotides their versatility for isomer formation (code words) is unsurpassed. The resulting high-density coding capacity of oligosaccharides is established by variability in (i) anomeric status, (ii) linkage positions, (iii) ring size, (iv) by branching and (v) introduction of site-specific substitutions.

Note: Nomenclature rules for carbohydrates are presented in detail in *Carbohydr Res* 1997; **297**, 1–92.

References

1 Roseman S. Reflections on glycobiology. *J Biol Chem* 2001;276:41527–42.

2 André S *et al*. Substitutions in the *N*-glycan core as regulators of biorecognition: the case of core-fucose and bisecting GlcNAc moieties. *Biochemistry* 2007;46: 6984–95.

3 Laine RA. The information-storing potential of the sugar code. In: *Glycosciences: Status and Perspectives* (Eds.: Gabius H-J, Gabius S), pp. 1–14. Chapman & Hall, London, 1997.

4 Griffith BR *et al*. 'Sweetening' natural products via glycorandomization. *Curr Opin Biotechnol* 2005;16:622–30.

5 Reuter G, Gabius H-J. Eukaryotic glycosylation – whim of nature or multipurpose tool? *Cell Mol Life Sci* 1999;55:368–422.

6 Gabius H-J. Cell surface glycans: the why and how of their functionality as biochemical signals in lectin-mediated information transfer. *Crit Rev Immunol* 2006;26:43–80.

7 Casu B *et al*. Conformational flexibility: a new concept for explaining binding and biological properties of iduronic acid-containing glycosaminoglycans. *Trends Biochem Sci* 1988;13:221–5.

8 Hemmerich S *et al*. Strategies for drug discovery by targeting sulfation pathways. *Drug Discov Today* 2004;9:967–75.

9 Gabius H-J *et al*. The sugar code: functional lectinomics. *Biochim Biophys Acta* 2002; 1572:165–77.

10 Winterburn PJ, Phelps CF. The significance of glycosylated proteins. *Nature* 1972;236: 147–51.