A Simplified Carbohydrate Nomenclature

HARVEY A. KLEIN*

Tufts University, School of Medicine, Boston, Massachusetts 02111

Received June 5, 1979

While the idea of using binary numbers as stereodesignators for carbohydrates is not new,[1] it has never before been developed into a working system of nomenclature with rules for naming compounds. A new system is described, including new names, with each compound being named uniquely, thus precluding the use of α and β and of R and S. This system presents a learning tool for the student, an easy method of drawing and recognizing structures for the biochemist, and a simple means of representing carbohydrate structure in a digital computer.

When in the course of studying organic chemistry or biochemistry the student encounters the carbohydrates, he is faced with two problems: (1) memorizing the names of the six-carbon sugars, and (2) memorizing their structures. The first task will usually be accomplished by the use of a clever mnemonic, such as: "*All altr*uists *gl*adly *m*ake *gum in gal*lon *t*anks."[2] The first letters in each word of the mnemonic stand for allose, altrose, glucose, mannose, gulose, idose, galactose and talose, respectively. The second task is quite difficult and consequently, when the student is asked to draw the Fischer projections of these vital organic compounds, he is at a loss.

To alleviate this difficulty, description is made here of a proposed system of nomenclature (which is also a valuable teaching aid) using simple binary, or base two arithmetic, which not only helps the student learn the structures of the whole class of sugars, but also lends insight to the notion of enantiomers.

Why choose the binary system for this task? And, furthermore, why did the binary system arise for popular usage in the first place? Well, back in the early 1940s the idea of Charles Babbage's mechanical "analytical engine" was being converted into an electronic device, first by Howard Aiken of Harvard University (the Mark I), and later by engineers at the University of Pennsylvania (the ENIAC in 1946).[3] The engineers decided to measure a voltage rather than the position of a gear to determine a number. Naturally, they first attempted to use the decimal system, i.e., to represent numerical values in a computer by a voltage between 0 and 10 V. They soon realized the inefficiency of such a system in that (1) the time for such measurements was too long, and (2) the values were too close to each other for rapid, accurate measurement. In other words, there were too many in-between voltages for efficient exploitation of such a computing system.

Then came the elegant yet simple concept of using a switch (or transistor) which would have only two states to recognize, "on" and "off", thus allowing for a rapid and reliable computing system.[4] And so the binary number system came of age.

The great advantage of the binary system lies in the fact that there are only two kinds of binary digits or "bits", namely 1 and 0. This not only defines a simplified arithmetic but also provides a means for dealing with two-valued functions or bistable systems.[4] The binary digits 1 and 0 represent "ON" & "OFF", or "YES" & "NO", etc.

To convert a decimal (base 10) number to binary (base 2) form, we divide the decimal number by 2 successively, using the remainders as our binary bits, with the first remainder as the least significant digit, d_0, etc. For example, the conversion of $11_{(10)}$ to base 2 is as follows:

$11/2 = 5$ remainder 1 (so $d_0 = 1$);

$5/2 = 2$ remainder 1 ($d_1 = 1$);

$2/2 = 1$ remainder 0 ($d_2 = 0$); and

$1/2 = 0$ remainder 1 ($d_3 = 1$).

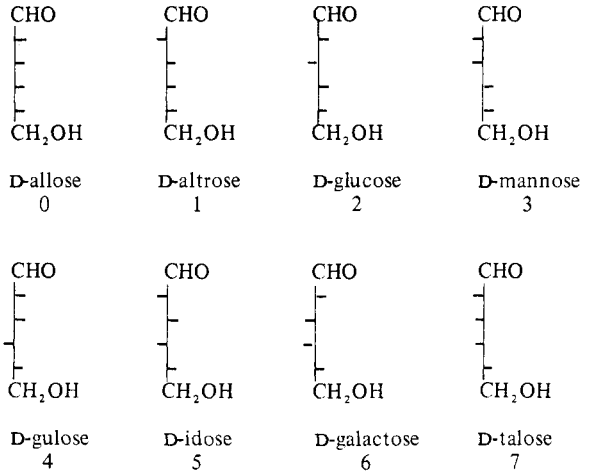
Thus $d_3d_2d_1d_0 = 1011 = 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 8 + 0 + 2 + 1 = 11_{(10)}$. We may now draw Table I.

"So," you may ask, "what does all this sweet talk about numbers have to do with sugar?" The answer is as simple as "All Altruists Gladly Make Gum In Gallon Tanks," which is the previously mentioned mnemonic for the eight D isomers of the six-carbon sugars, the aldohexoses, as shown in Chart I. These sugars may be assigned the decimal numbers shown in the chart. The convention used here is as follows: each intersection represents a carbon atom; each horizontal line represents an hydroxyl (OH) group; opposite each "hydroxyl" group, on the other side of the carbon backbone, is an understood hydrogen atom.

Table I

Positions of the binary system decimal values	2^4 16	2^3 8	2^2 4	2^1 2	2^0 1
decimal no.			binary equivalent		
1					1
2				1	0
3				1	1
4			1	0	0
5			1	0	1
6			1	1	0
7			1	1	1
8		1	0	0	0
9		1	0	0	1
10		1	0	1	0
11		1	0	1	1
12		1	1	0	0
13		1	1	0	1
14		1	1	1	0
15		1	1	1	1
16	1	0	0	0	0

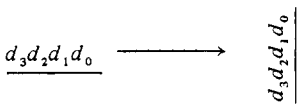
Chart I



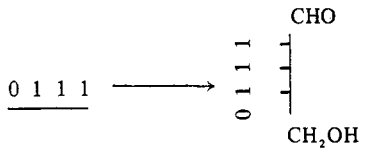
* Address correspondence to author at 1137 E. 5 St., Brooklyn, N.Y. 11230.

0095-2338/80/1620-0015$01.00/0

If we take our first four binary digits $d_3d_2d_1d_0$ and rotate the horizontal axis through an angle of +90°



and if we then write the binary representation of the decimal number of the sugar (0–7) along this vertical axis, omitting all zeros on the sugar backbone, and using talose as an example (No. 7, Chart I), we get



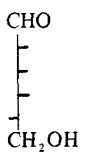
We then fill in each empty spot on the carbon backbone (where the zeros were) with a horizontal line on the right side of the backbone representing another OH group, giving

D-talose

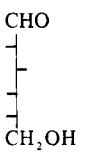
Thus, the eight isomers of D-aldohexose may now be named "(K-0)-aldohexose" through "(K-7)-aldohexose", without having to memorize any structures (or witty sayings), and their stereoisomeric formulas may be drawn directly from their new names. (Note: the *K* denotes the Klein System of Nomenclature for Carbohydrates, the *number* represents the stereoisomeric formula, and the *aldohexose* represents the class of compound.)

It is appropriate at this point to take a moment to explain one additional property of the binary system, the concept of a "one's complement". To obtain the one's complement of any binary number, we replace all zeros by ones, and all ones by zeros. If we add a number and its complement, we obtain the maximum number expressible using the amount of digits given. For example, the complement of 1010 is 0101; 1010 + 0101 = 1111 = $15_{(10)}$. The complement of 0000 is 1111; 0000 + 1111 = 1111 = $15_{(10)}$. Thus, to obtain the enantiomer of any one of the aldohexoses, we simply write its one's complement, which is equal to 15 minus the number itself.

As an illustration, to draw the enantiomer of D-talose (otherwise known as (K-7)-aldohexose) we compute 15 − 7 = $8_{(10)}$ which is represented in binary form as 1000. Then using our principle of a positive 90° rotation, we can immediately draw L-talose which is actually (K-8)-aldohexose, as follows:

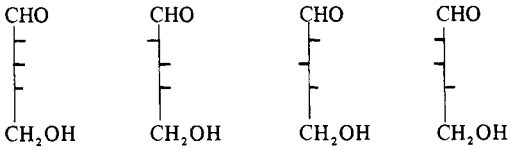


And the enantiomer of D-glucose (which is (K-2)-aldohexose) is (K-(15–2))- or (K-13)-aldohexose; $13_{(10)}$ in binary is 1101, and so (K-13)-aldohexose becomes



which is L-glucose.

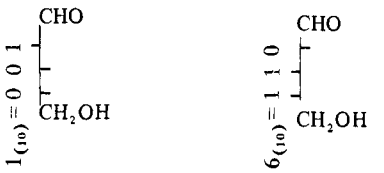
This system can be applied to the four pairs of enantiomers of aldopentoses, as follows:



(A useful mnemonic for these four sugars is "*Right Axle*".)

Now we notice that there are three asymmetric carbons. This gives us $2^3 = 8$ possible stereoisomers, or a maximum of four pairs of enantiomers. We number them 0 through 7, with the first four (numbered 0–3) being of the D configuration and the last four (numbered 4–7) being of the L configuration. The enantiomers of any of the first four can be obtained by subtracting the binary number of that sugar from 7, thus yielding the complement.

As an example, D-arabinose in this system is (K-1)-aldopentose (shown below); 7 − 1 = 6, so the enantiomer of (K-1)-aldopentose is (K-6)-aldopentose and is drawn thusly:



(K-1)-aldopentose D-arabinose

(K-6)-aldopentose L-arabinose

We now notice an additional advantage of this system of nomenclature: in the case of the aldoses we have *n* chiral centers and, therefore, 2^n stereoisomers. If we define the quantity $S = 2^n$ as the number of stereoisomers (which are numbered from 0 through $S - 1$), then $S/2$ becomes the midpoint or the dividing line such that for $n = 3$, $S = 2^n = 2^3 = 8$, and $S/2 = 8/2 = 4$. Thus, the first four, or the first $S/2$ sugars, will always be D, and the second half of the group, or the second four (numbered from $S/2$ to $S - 1$, i.e., from 4 to 7), will always be L.

If there were a seven-carbon sugar, with five asymmetric carbons, then $n = 5$, and $S = 2^5 = 32$, and there are 32 stereoisomers of the aldoheptoses, named (K-0)- through (K-31)-aldoheptose; $S/2 = 32/2 = 16$, so the first 16 sugars (numbered from 0 through $S/2 - 1$, i.e., from 0 through 15) are all D, the next 16 (numbered from $S/2$ to $S - 1$, i.e., (K-16)- through (K-31)-aldoheptose) are all L, and to draw an enantiomer of any of these sugars we simply find its one's complement. Thus, the enantiomer of (K-15)-aldoheptose is simply (K-(31−15)) = (K-16)-aldoheptose, it is L, and its structure can be drawn directly from its name:

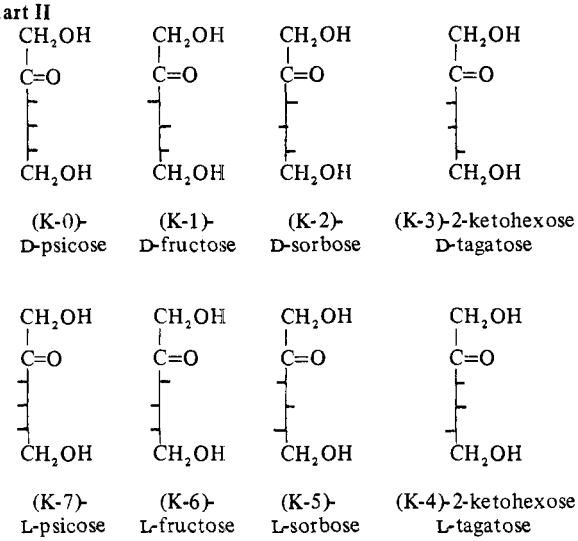
$$15_{(2)} = \begin{matrix}(16\ 8\ 4\ 2\ 1)\\ 0\ 1\ 1\ 1\ 1\end{matrix}, \text{ and } 16_{(2)} = \begin{matrix}(16\ 8\ 4\ 2\ 1)\\ 1\ 0\ 0\ 0\ 0\end{matrix}$$

and so we have,

(K-15)-aldoheptose (K-16)-aldoheptose

Applying this system to fructose, we find that fructose (which is a 2-ketohexose,) has only three asymmetric carbons,

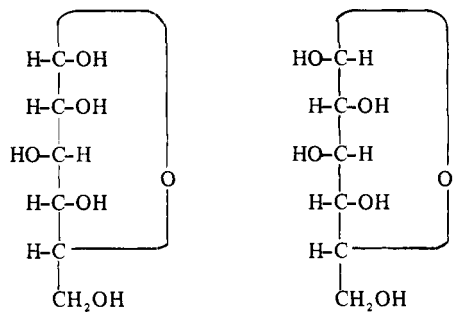
Chart II



so that $n = 3$, and $S = 2^n = 2^3 = 8$. Thus, we have eight stereoisomers of fructose, of which the first four are D and the last four are L (see Chart II). Again note that (K-0)- through (K-3)-2-ketohexose are all D and that (K-4)- through (K-7)-2-ketohexose are all L. (A useful mnemonic for these four sugars is "*P*retty *Fast*".)

Obviously, it is no longer necessary to specify *R* or *S* to describe the absolute configuration of the sugars already mentioned, since the new "name" describes the structure adequately.

Yet another advantage of this system becomes apparent when we notice that in cyclic structures, such as in the two hemiacetal forms of D-(+)-glucose,



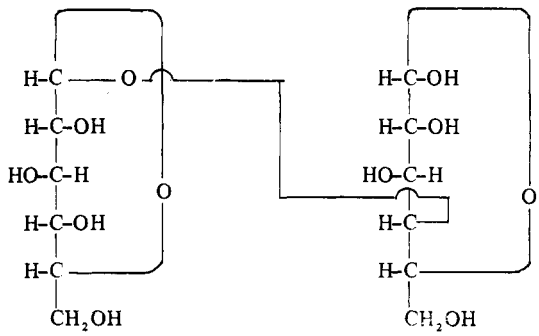
α-D-(+)-glucopyranose β-D-(+)-glucopyranose

the α form will always have the anomeric OH on the right side, whereas the β form will always have the anomeric OH on the left side. The implication is that the α form will always be even numbered (i.e., K-0, 2, 4, 6, etc.) and the β form will always be odd numbered (i.e., K-1, 3, 5, 7, etc.). Thus, with the new system of nomenclature it is no longer necessary to specify α or β, since the "name" (or number) of the compound automatically tells us this fact.

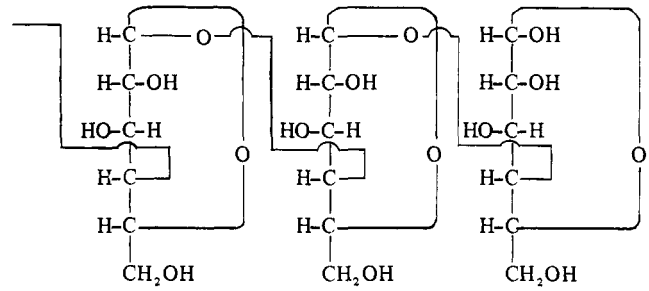
Now in considering the cyclic forms of the aldohexoses, we encounter five instead of four asymmetric carbons, yielding $2^5 = 32$ stereoisomers instead of 16 stereoisomers. Still the system retains its simplicity because α-D-(+)-glucopyranose and β-D-(+)-glucopyranose may now be renamed (+)-(K-4)-pyranose and (+)-(K-5)-pyranose, respectively.

Of course, with a compound such as maltose, the common name (maltose) should be retained, but the structural name of α-maltose should be changed from 4-*O*-(α-D-glucopyranosyl)-D-glucopyranose to 4-*O*-[(K-4)-pyranosyl]-(K-4)-pyranose (see structure below).

Turning to some of the more complex molecules, e.g., starches, we see that amylose, the unbranched chain of homopolymers of glucose with α (1→4) linkages, is simply an



(+)-maltose (α anomer)



amylose

unbranched chain of homopolymers of (K-4)-pyranose with α (1→4) linkages, and may be named [4-*O*-{(K-4)-pyranosyl}]$_n$-(K-4)-pyranose, where $n = 1, 2, 3, \ldots$.

Glycogen is also a homopolymer of (K-4)-pyranose (α-D-glucopyranose), with α (1→4) linkages, and branching every 8 to 12 glucose residues via α (1→6) linkages.[5]

Let us now compare the IUPAC and Klein systems in relation to the aldohexoses. In order to draw the stereoisomeric formula of one of the aldohexoses (assuming the use of a mnemonic device to remember the names of these compounds in the correct sequence, although brute force memorization would suffice), we must go through the following steps.

A. Conventional (IUPAC) System:

(1) Write down a mnemonic representing the eight names of the aldohexoses (or write the names themselves).

(2) Draw the "grid" or "cross formulas" for *all eight* sugars.

(3) Choose the desired sugar and recopy its Fischer projection into the text we happen to be writing (e.g., paper, exam, etc.) at the time.

B. Klein System:

(1) Same as 1A above.

(2) Place a number under each name (or each mnemonic), starting with 0, as follows:

All Altruists Gladly Make Gum In Gallon Tanks
0 1 2 3 4 5 6 7

(3) Immediately and directly, draw the *one* formula needed by converting the number previously written (in step 2B) into binary form, and writing its equivalent structure sideways:

sideways

Although it may seem complicated at first, with very little practice steps 2B and 3B (in the Klein system) can be performed in one-eighth (1/8) the time it takes to perform steps 2A and 3A in the conventional system.

Thus, the proposed system is much more than just a simpler system of nomenclature; it is a tool which can be used to eliminate the tedium of the time-consuming steps 2A and 3A—the drawing of the "grid". It is also a slick method of

getting from the conventional name to the Fischer projection and is, therefore, a most valuable and much needed teaching aid for the student and time saver for the chemist and biochemist.

Finally, this system allows for an easy and direct method of representing sugar structure in a digital computer, suggesting possibilities for more sophisticated and elegant analyses of carbohydrates than have been done to date.

And so, since sugar (in one form or another) is by far one of the most abundant compounds in nature, it is one of nature's biggest blessings to mankind. Ergo, we may now literally "count our blessings" and draw the one we need, rather than drawing them all each time.

ACKNOWLEDGMENT

I wish to express my gratitude to Professor Henry Yuska, Chairman of the Department of Chemistry at Brooklyn College, Brooklyn, New York, for his help, understanding, guidance, and encouragement; and to Professor Henry G. Mautner, Chairman of the Departments of Biochemistry and Pharmacology at Tufts University School of Medicine, Boston, Massachusetts, for his valuable advice on the preparation of this paper.

REFERENCES AND NOTES

(1) S. Neelakantan, *Curr. Sci.*, **38**, 353 (1969).
(2) L. F. Fieser and M. Fieser, "Organic Chemistry", 3rd ed, Reinhold, New York, 1959, p 359.
(3) M. L. James, G. M. Smith, and J. C. Wolford, "Applied Numerical Methods for Digital Computation with FORTRAN", International Textbook Co., Scranton, Pa., Oct 1968, p 4.
(4) "McGraw-Hill: Encyclopedia of Science and Technology", Vol. 9, McGraw-Hill, New York, 1971, p 258.
(5) A. L. Lehninger, "Biochemistry", 2nd ed, Worth Publishers, New York, 1975, p 266.

ADDITIONAL REFERENCES

Feldman, A., *J. Org. Chem.*, **24**, 1556 (1959).
Morrison, R. T., and Boyd, R. N., "Organic Chemistry", 3rd ed, Allyn and Bacon, Boston, 1973.
Neelakantan, S., *Curr. Sci.*, **39**, 85 (1970).
Neelakantan, S., *Indian J. Chem. Educ.*, **2**, 15 (1971).
Ore, O., "Number Theory and Its History", McGraw-Hill, New York, 1948.
Pigman, W., Horton, D., and Herp, A., "The Carbohydrates", Vol. II-B, 2nd ed, Academic Press, New York and London, 1970.
Rosenblatt, D. H., *J. Chem. Educ.*, **42**, 271 (1965).

The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds

J. MOCKUS* and R. E. STOBAUGH

Chemical Abstracts Service, Columbus, Ohio 43210

Received December 3, 1979

The Chemical Abstracts Service (CAS) Chemical Registry System is a computer-based information system that uniquely identifies chemical substances on the basis of their molecular structure. Substances that have several possible chemically equivalent representations are difficult to portray precisely by a single structure diagram or connection table. Among the major causes of this problem are aromatic rings, whose alternating single and double bonds can be represented in more than one way, and tautomerism, an equilibrium involving single/double bond shifts coupled with hydrogen migration. The CAS Chemical Registry System handles the problem by algorithmically recognizing tautomeric and alternating bond structures, replacing the explicit single and double bonds with special normalized bonds, and associating the migrating tautomeric hydrogen with groups of atoms rather than just single atoms. This article describes the normalization techniques used in handling alternating bonds and tautomeric bonds, as well as substructure search aspects involving these bond types, and denormalization procedures required for algorithmic structure display and name generation.

INTRODUCTION

The Chemical Abstracts Service (CAS) Chemical Registry System is a computer-based system for the unique identification of chemical substances on the basis of structure.[1] The initial, experimental system, Registry I, began operation in 1964 and established the viability and validity of the registration concept for fully defined organic substances. In 1968, the scope of the system was increased as it began to handle additional classes of substances. The system, now known as Registry II, began to be integrated into the CAS indexing operation. In 1974, the most recent version, Registry III, made major adjustments in the Registry structure records to provide increased support to the process of generating names for the *Chemical Abstracts* (CA) Chemical Substance Index, and also to computer-based structure output operations through explicit identification of the ring systems present in a substance. As its use has expanded, the CAS Chemical Registry System has proven to be reliable and consistent as a structure identification method and has become an essential CAS production tool supporting CA index input and compilation. It has also found widespread interest and support in the scientific and technical community.

The foundation of the CAS Chemical Registry System is an algorithm that generates a unique and unambiguous machine-readable description of the molecular structure of a substance. The principal component of the machine record is a connection table, a detailed description of the atoms and bonds that comprise the basic structure of the substance. Other components describe stereochemical characteristics, isotopic labeling, and derivatives (salts, hydrates, etc.).

The representation of a chemical substance by a unique structure diagram or connection table poses problems to both chemists and chemical information systems when the substance has several possible representations, chemically equivalent but structurally distinct. Resonant or aromatic bonds which have characteristics of both single and double bonds are one major

0095-2338/80/1620-0018$01.00/0