

# WURCS: The Web3 Unique Representation of Carbohydrate Structures

Kenichi Tanaka,<sup>†,‡,▽</sup> Kiyoko F. Aoki-Kinoshita,<sup>§</sup> Masaaki Kotera,<sup>||</sup> Hiromichi Sawaki,<sup>†</sup> Shinichiro Tsuchiya,<sup>§</sup> Noriaki Fujita,<sup>†</sup> Toshihide Shikanai,<sup>†</sup> Masaki Kato,<sup>⊥</sup> Shin Kawano,<sup>#</sup> Issaku Yamada,<sup>\*,‡</sup> and Hisashi Narimatsu<sup>†</sup>

<sup>†</sup>Research Center for Medical Glycoscience, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568, Japan

<sup>‡</sup>The Noguchi Institute, Itabashi, Tokyo 173-0003, Japan

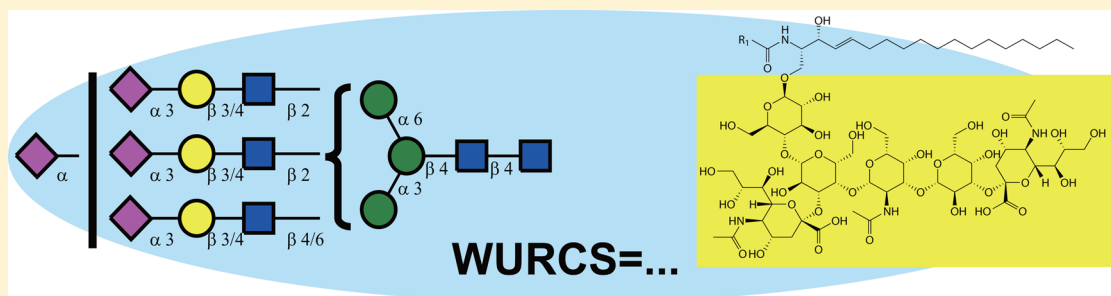
<sup>§</sup>Department of Bioinformatics, Faculty of Engineering, Soka University, Hachioji, Tokyo 192-8577, Japan

<sup>||</sup>Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo 152-8550, Japan

<sup>⊥</sup>Structural Glycobiology Team, RIKEN Global Research Cluster, Wako, Saitama 351-0198, Japan

<sup>#</sup>Database Center for Life Science, Research Organization of Information and Systems, Kashiwa, Chiba 277-0871, Japan

## Supporting Information



**ABSTRACT:** In recent years, the Semantic Web has become the focus of life science database development as a means to link life science data in an effective and efficient manner. In order for carbohydrate data to be applied to this new technology, there are two requirements for carbohydrate data representations: (1) a linear notation which can be used as a URI (Uniform Resource Identifier) if needed and (2) a unique notation such that any published glycan structure can be represented distinctively. This latter requirement includes the possible representation of nonstandard monosaccharide units as a part of the glycan structure, as well as compositions, repeating units, and ambiguous structures where linkages/linkage positions are unidentified. Therefore, we have developed the Web3 Unique Representation of Carbohydrate Structures (WURCS) as a new linear notation for representing carbohydrates for the Semantic Web.

## INTRODUCTION

The computational analysis of complex carbohydrates, or glycans, has produced a number of linear and nonlinear notations to represent these complex structures. Table 1 lists some of the major formats used today. Each representation format has advantages and disadvantages over the others to varying degrees, for different applications. For example, LinearCode<sup>1</sup> is a compact format in which users can see and understand quickly the general structure of a glycan (once they are accustomed to the monosaccharide coding scheme). However, rarely used monosaccharides cannot be represented in this format. GlycoCT<sup>2</sup> is a popular format as it uses a strict coding scheme for monosaccharides and linkages. However, as we will show in this paper, some monosaccharides cannot be represented using this format either.

In recent years, the Semantic Web has become the focus of life science database development as a means to link life science data

in an effective and efficient manner.<sup>3</sup> The Semantic Web utilizes the hyperlinks often found on the Internet as a means to link data that are related to one other. Moreover, it uses rules to apply semantics to these links such that the relationships between links are made clear. Thus, it would eventually become possible to computationally trace links across the Semantic Web to find new information or knowledge, such as functional information or related proteomics information, pertaining to a particular glycan, for example. However, in order for carbohydrate data to be applied to this new technology, there are two requirements to represent carbohydrate data that are currently not possible with existing notations: (1) a linear notation which can be used as a URI (Uniform Resource Identifier; a string that is used on the

**Received:** October 3, 2013

**Published:** June 4, 2014

Table 1. Some of the Major Carbohydrate Structure Formats and Their Descriptions

format name	description
CarbBank <sup>4</sup>	The carbohydrate structure format used by the Complex Carbohydrate Structure Database (CCSD) originally developed at the University of Georgia. This is a 2D structure where bars and hyphens are used to indicate linkages, and monosaccharides are specified in the following order, separated by hyphens: anomer - configuration symbol - monosaccharide abbreviation.
IUPAC <sup>5</sup>	A linear notation proposed by the International Union of Biochemistry and Molecular Biology (IUBMB), where the reducing end is specified to the right, and branches are indicated by parentheses
LINUCS <sup>6</sup>	A unique linear notation proposed and used by GLYCOSCIENCES.de, with the reducing end specified on the left. Monosaccharides, written according to Carbohydrate rules, are surrounded by square brackets and anomers, and stereochemistry is specified and separated by hyphens. Glycosidic linkages are also surrounded by hyphens, and carbon numbers of the linkages are separated by the plus sign. Branches are indicated by curly brackets.
GlycoMinds <sup>1</sup> LinearCode	A linear notation proposed by GlycoMinds, Ltd., where monosaccharides are indicated by a one- or two-letter code, linkages are indicated by "a" or "b" for anomers and a single number for the carbon number linkage on the reducing end, and branches are indicated by parentheses. This format is the most compact among existing formats, and several databases use a modified form of this format to accommodate additional codes and varying carbohydrate structures (such as ambiguous structures).
KCF <sup>7</sup>	Abbreviation for KEGG Chemical Function. This is a multiline format used by the KEGG GLYCAN database and represents glycans as graphs. Monosaccharides are represented as nodes, and glycosidic linkages are represented as edges. Each node is given a unique number, and x- and y- coordinates are specified such that the glycan structure can be drawn on a 2D plane. Edge information uses the node numbers to specify anomeric and carbon number information. Monosaccharide names are specified as text, and no particular rules are specified to represent these names.
GlycoCT <sup>condensed</sup> <sup>2</sup>	This is a multiline format used by GlycomeDB, among others, and uses a graph notation to represent glycans, similar to KCF. Monosaccharides are specified more strictly to be able to map the monosaccharides between different representations. Substituents, such as N-acetyl, are specified as separate nodes from their basetypes and are linked together in the edge section. If newlines are removed, this format can be made linear.
GLYDE-II <sup>8</sup>	This is a standard for the representation of the chemical structures of complex glycans that is based on a connection table formalism using XML syntax. This representation of monosaccharide molecules that are used as archetypes for carbohydrate residues will be dynamically generated by services provided by monosaccharideDB. In this format, an atom can be defined as a common chemical sense and includes substance like "Oxygen_atom" and "Carbon_atom" in it.

Internet to locate a particular resource, often using HTTP (hypertext transfer protocol)—most URIs are URLs (Uniform Resource Locators)—and (2) a unique notation such that any published glycan structure can be represented distinctly. This latter requirement includes the possible representation of nonstandard monosaccharide units as a part of the glycan structure, as well as compositions, repeating units, and ambiguous structures where linkages/linkage positions are unidentified. These emerge from the fact that various technologies used to determine glycan structures are often unable to identify all of the details of a glycan structure, such as anomeric configurations and linkage information.

Although the GlycoCT representation is quite close to satisfying these requirements, this format is in general nonlinear and uses a dictionary to represent monosaccharides. The nonlinear structure could be made linear by simply appending each line into a single string. However, for rare monosaccharides as could be found in bacterial structures, it would be difficult to represent these structures in GlycoCT format (see Table 6 for examples). GLYDE-II can represent a glycan containing chemical modifications. However, GLYDE-II is not easy to use as a URI, because the XML format cannot represent a unique string. Therefore, we have developed the Web3 Unique Representation of Carbohydrate Structures (WURCS) as a new linear notation for representing carbohydrates for the Semantic Web. In this new format, it is possible to represent any published carbohydrate structure as a linear string, which can be used as a unique identifier for the Semantic Web. Based on this notation, we hope that all carbohydrate databases can map their structures to this identifier such that they can be linked together appropriately.

## ■ STRUCTURE OF WURCS

Before describing the WURCS representation, we define some terms to describe the concepts used in this notation. A WURCS format is follows:

$$\text{WURCS} = \text{Version}/m,n/[\text{BMU1}]...[\text{BMU}m]\text{MLU1}...\text{MLU}n$$

where  $m$  and  $n$  are the number of monosaccharides (BMUs) and linkages (MLUs), respectively, followed by the list of sorted BMUs and the list of sorted MLUs. Next, we describe how to define and build the WURCS.

**Definitions of Structure Components.** In order to generate WURCS, first, the input glycan structure is broken down into three parts: backbone, modification, and aglycon, which are defined as follows.

**Definition 1: glycan.** A molecule consisting of backbones, their modifications, and possibly an aglycon.

**Definition 2: backbone.** The carbon backbone of a monosaccharide in a glycan containing only carbons.

**Definition 3: modification.** Components of monosaccharides other than the carbons of the backbones. Modifications include in the hydrogen and oxygen atoms attached to the carbons of the backbones in addition to heteroatoms.

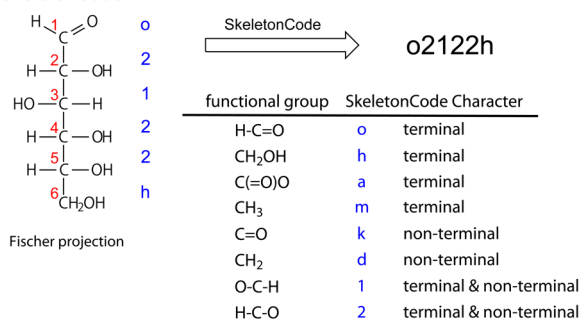
**Definition 4: aglycon.** The nonsugar component of a glycan, which is usually connected to the anomeric carbon position of a backbone.

**Definitions for Linear Notation.** Among the three parts extracted from the glycan, the backbones, modifications, and attachment information between the backbones and modifications are represented as strings called SkeletonCodes, ALINs,

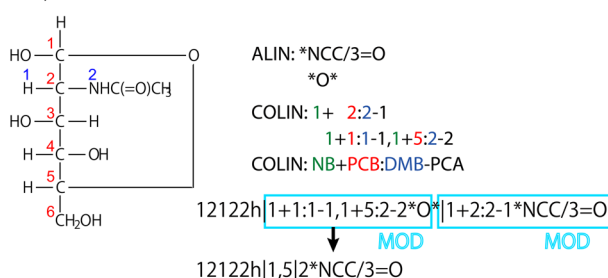
and COLINs, respectively. Each of these is formally defined as follows.

**Definition 5: SkeletonCode and SkeletonCode Character.** SkeletonCode is a string representation of backbone carbons including the adjacent atoms in modifications. Each character of a SkeletonCode is called a SkeletonCode Character. The SkeletonCode is similar to the Extended Stereocode used for representing monosaccharides in MonosaccharideDB.<sup>9</sup> The length of a SkeletonCode is dependent upon the number of carbons in the backbone (e.g., there are six SkeletonCode Characters to represent a hexose). Each SkeletonCode Character represents each carbon of the basetype backbone of the monosaccharide, which is based on the Fischer projection of the monosaccharide. In Figure 1a, the backbone carbon which is

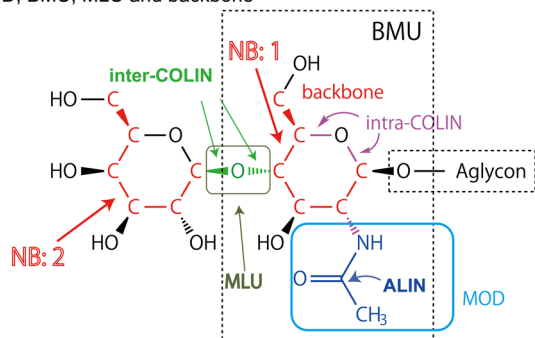
(a) SkeletonCode



(b) ALIN, COLIN and MOD



(c) MOD, BMU, MLU and backbone



**Figure 1.** Illustration of basetype string used to represent the WURCS.

numbered "1" in red, called the C-1 carbon, is connected to two atoms in modification. One is a hydrogen atom via a single bond, and the other is an oxygen atom via a double bond. In this case, "o" is assigned as the SkeletonCode Character to this C-1 carbon. As for the C-2 carbon, a hydrogen via a single bond is connected on the left side and an oxygen atom via a single bond is connected on the right side. In this case, the atomic number of the right side atom is larger than the one on the left side, hydrogen

(1) < oxygen (6), thus "2" is assigned to the C-2 carbon as the SkeletonCode Character. In the same way, the SkeletonCode Characters are assigned to all carbons in the backbone, resulting in "o2122h" as the SkeletonCode. We defined 31 SkeletonCode Characters for terminal carbons, head and tail carbons of backbones, and we defined 21 SkeletonCode Characters for non-terminal carbons (see Supporting Information Tables S3 and S4).

**Definition 6: ALIN, Atomic Linear Notation.** A string representing a modification including the adjacent carbon in a backbone. The method to generate ALIN is described in detail in the Supporting Information. Table 2 lists some examples of

**Table 2.** Examples of Modifications in ALIN Format

abbreviation	substituent group	ALIN
H	hydrogen	*H
OH	hydroxyl	*O
O	ether	*O*
Me	methyl	*C
Ac	acetyl	*OCC/3=O
NAc	N-acetyl	*NCC/3=O
P	phosphate	*OPO/3O/3=O
S	sulfate	*OSO/3=O/3=O
Pyr	pyruvate	*OCCC/4=O/3=O
PC	phosphocholine	*OPOCCNC/7C/7C/3O/3=O
PPEn	diphosphoethanolamine	*OPOCCN/5O/5=O/3O/3=O
PEtn	phosphoethanolamine	*OPOCCN/3O/3=O
N	primary amine	*N
Gc	glycolyl	*NCCO/3=O

modifications that are often found in carbohydrate structures. We simply note here that asterisks "\*" are used in the ALIN string to indicate the attached carbons in the backbones. The ALIN may be omitted in a WURCS string, as described later (see Figure 1b, Table 2).

**Definition 7: COLIN, Connection Linear Notation.** A string representing the connection between a backbone and modification (see Figure 1b). A COLIN is composed of the following four units:

NB: Number of Backbone, referring to the backbone index  
PCB: Position of the connected Carbon in the Backbone  
DMB: Direction of Modification on the Backbone carbon  
PCA: Position of connected backbone Carbon in ALIN

NB is the number of the monosaccharide being described. PCB is the number of the carbon on the BMU that is modified. DMB represents the comparison of the modifications that are connected to the same carbon of the backbone. To compare modifications, we use CIP rules which are used for *R/S* judgment in IUPAC rules.<sup>10</sup> If the modification on the left side has higher priority than the right side, DMB is "1," otherwise "2". PCA indicates the position in the ALIN string that is involved in the bond. Using these four units, a COLIN is represented as follows:

NB+PCB:DMB-PCA

A COLIN is listed just before the corresponding ALIN. The COLINs may be compressed, as described later.

**Definitions for Combining Linear Notations.** WURCS is obtained by combining the generated SkeletonCodes, ALINs, and COLINs and listing them in order, separated by delimiters. We will next briefly describe MOD, BMU, and MLU, which are used in this process.



**Definition 8: MOD, linear notation of MODification.** A string representing a modification. It is composed of COLINs and ALIN. If multiple COLINs compose the modification, they are listed consecutively using “,” as a delimiter. There is no delimiter between COLIN and ALIN. A MOD is thus represented as follows:

COLIN1,...,COLIN $k$ ALIN

where  $k$  indicates the number of COLINs.

**Definition 9: Version, the Version number of WURCS.** This manuscript describes version 1.0. Version numbers are also listed as a part of a WURCS string.

**Definition 10: BMU, Basic Monosaccharide Unit.** A string representing a single monosaccharide unit, which is composed of SkeletonCodes and MODs. The target MODs are connected with only the monosaccharide. MODs are listed after the SkeletonCode, separated by a vertical bar “|”. A BMU is thus represented as follows:

SkeletonCode|MOD1...|MOD $l$

where  $l$  indicates the number of MODs. For example, GlcNAc would be represented as follows:

12122h|1+1:1-1,1+5:2-2\*O\*|1+2:2-1\*NCC/3=O

**Definition 11: MLU, Monosaccharide Linkage Unit.** A string representing a monosaccharide linkage unit, which is composed of a MOD which is not a member of a BMU. In other words, if a modification is connected with two or more backbones, a MOD is treated as an MLU. For example, glycosidic bonds are represented by MLUs.

**Definition 12: WURCS, Web3 Unique Representation of Carbohydrate Structures.** A string, representing a glycan, composed of Version, number of BMUs, number of MLUs, a set of BMUs, and a set of MLUs. In general, the WURCS string starts with “WURCS=” followed by the version number of WURCS being represented. In this manuscript, we describe version 1.0. A slash “/” follows, followed by the number of BMUs, a comma, and the number of MLU sets in the structure. Another slash follows, after which the sorted list of BMUs surrounded by square brackets follows. This is then immediately followed by MLUs (and optionally the corresponding ALINs) separated by vertical bars “|”. A WURCS string is thus represented as follows:

WURCS=Version/ $m,n$ /[BMU1]...[BMU $m$ ]

MLU1|...|MLU $n$

where  $m$  indicates the number of BMUs, and  $n$  indicates the number of MLUs.

**Compressed COLINs and ALINs.** The majority of monosaccharides in carbohydrate structures are in the ring form, and glycosidic bonds are usually formed using one oxygen atom. Thus, we propose a compressed version for MOD (COLINs and ALINs), such that many glycan structures can be compactly represented by WURCS using the compressed version of MODs. The rules for compressed MOD are as follows.

- Compressed MOD for ring form “1+1:1-1,1+5:2-2\*O\*” is “1,5”
- Compressed MOD for modification unit “1+2:2-1\*N” is “2\*N”
- ALINs “\*H,” “\*O\*,” and “\*O” should be in general omitted when it is clear whether the modification is an oxygen or hydrogen

- When the atomic position is clear (unchanged from the corresponding SkeletonCode) of the attached COLINs, it should be omitted, such as in the following: 1+1:2,3+6:0|2+1:2,3+3:1|3+1:1,4+4:2|4+1:1,5+4:2 should be compressed as 1+1,3+6|2+1,3+3|3+1,4+4|4+1,5+4.

In order to maintain the uniqueness of WURCS strings for any possible modification, we have developed a set of rules, which indicates whether a COLIN or an ALIN should be compressed. This is available at the WURCS working group Web site and will be maintained by versioning. The URL is [http://www.wurcs-wg.org/compression\\_rules.php?version=Version](http://www.wurcs-wg.org/compression_rules.php?version=Version).

**Ambiguous Linkages.** When attachment sites are unknown, a question mark (?) is used. If multiple attachment sites are possible, all the possible sites are listed and delimited by a backslash (\). Statistically present substructures are specified by adding “%.x%” or “%.y-z%” either before or after the corresponding COLIN attaching the statistically present substructure to the main glycan chain. When the position of the percentage value is placed after the COLIN, the substructure in front of the COLIN is considered the fixed structure, and the substructure behind it is the indefinite structure. The opposite holds when the “%” is before the COLIN (see Supporting Information Figure S8 for more details). The following are examples:

[12122h|1,5|2%.5%\*NCC/3=O]

[12112h|1,5][X2122h|1,5]1+1,2+4%.4-.6%

[12112h|1,5][X2122h|1,5]%.8%1+1,2+4

In such cases  $x$ ,  $y$ , and  $z$  are numbers indicating the probability or range of probabilities of the presence of the substructure. Note that this example uses compressed COLINs.

## ■ WURCS METHOD

Figure 2 is an illustration of the algorithm to generate a WURCS string given a fully defined glycan structure. In general, WURCS is generated in five steps given a chemical representation of a carbohydrate structure, such as an MDL Molfile: (1) extraction of the monosaccharide backbones, (2) extraction of the aglycon structures, (3) extraction of the modifications, (4) extraction of the glycan chains, and (5) generation of WURCS. Here, we demonstrate the procedure to generate WURCS (Figure 2), using a molecule in the Protein Data Bank (PDB)<sup>31</sup> as an example. (a) This molecule is entry 4dgo, a structure where two glycan chains are attached to an oligopeptide chain. (b) In the first step, the numbers of attached heteroatoms (mainly oxygen atoms) are checked for all carbon chains, detecting the main carbon backbones for monosaccharides. The anomeric carbon atoms of all the monosaccharides are determined in the same process. (c) In the second step, the substructures other than the obtained monosaccharide backbones are regarded as aglycons if they attach to the backbones only by the anomeric carbon atoms. (d) In the third step, the substructures are regarded as modifications if they do not belong to the backbones or aglycons. An atom is also regarded as a modification if it belongs to an aglycon and attaches to a backbone. (e) In the fourth step, connected groups of backbones and modifications are regarded as glycans. (f) In the fifth step, each glycan structure is transformed into a WURCS string by sorting and combining the strings that represent backbones, modifications, and their connections using the rules described in the following section.

**Step 1: Normalization for Biologically Same Activities.** The WURCS representation is generated by first identifying the



Table 4. WURCS of Sample Glycans

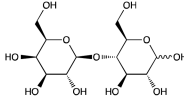
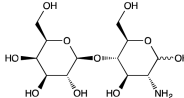
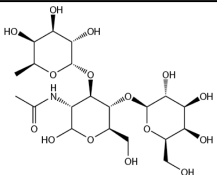
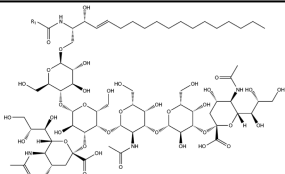
Glycan name	Chemical Structure	WURCS
Lactose		WURCS=1.0/2,1/[12112h 1,5][X2122h 1,5]1+1,2+4
Lactosamine		WURCS=1.0/2,1/[12112h 1,5][X2122h 1,5]2*N]1+1,2+4
Lewis <sup>x</sup>		WURCS=1.0/3,2/[12112h 1,5][11221m 1,5][X2122h 1,5]2*NCC/3=O]1+1,3+4 2+1,3+3
Ganglioside GD1a		WURCS=1.0/6,5/[a2d21122h 2,6]5*NCC/3=O][12112h 1,5][12112h 1,5]2*NCC/3=O][a2d21122h 2,6]5*NCC/3=O][12112h 1,5][12122h 1,5]1+2,2+3 2+1,3+3 3+1,5+4 4+2,5+3 5+1,6+4

Table 5. Number of Structures in Various Databases Converted to WURCS<sup>a</sup>

database	number of compounds	distinct number of WURCS	number of multiple WURCS
ChEBI <sup>14</sup>	4827	4648	218
PDB <sup>11</sup>	3192	722	0
KEGG compound <sup>15</sup>	2429	2322	168
GlycoNAVI <sup>16</sup>	3492	3453	93
JMSDB <sup>17</sup>	886	839	0
LfDB <sup>18</sup>	186	186	0

<sup>a</sup>Distinct number of WURCS reflects the number of unique glycan structures in the database. Number of multiple WURCS reflects the number of database entries containing more than one unique glycan structure.

have only one bond to another carbon atom. (3) Detect the carbon chains formed by linear paths between the terminal carbons (in both directions) that satisfy minNOS and minO (see Supplemental Code 1).

Next, for each detected carbon chain: (4) C-1 judgment (see Supplemental Code 2) is performed. (5) If the C-2 or later carbons seem more likely to be C-1 (see Figure 3), the carbon chain is shortened, and if the resulting carbon chain length is within the specified range of the expected carbon chain length, the carbon chain is added to the carbon chain list. While the carbon chain list is not empty: (6) The carbon chain(s) that is (are) most like the backbone (see Supplemental Code 3) in the list of carbon chains is(are) added to the backbone list, and (7) any carbon chains containing at least one carbon which is included the backbone list is excluded from the carbon chain list.

**Step 3: Identification of Aglycons and Modifications.** Consider each connected subgraph formed by removing the backbones in each molecule.

For each connected subgraph:

- Any connected subgraph attached to only anomeric atoms is considered an aglycon.

- Any connected subgraph attached to nonanomeric atoms is considered a modification.
- Among the aglycons, any single atoms directly attached to the backbone is considered a modification (e.g., the nitrogen atom of asparagine residues to which N-linked glycans are linked is considered a modification).

For each modification, generate its ALIN string (see Supporting Information S5 and Codes 4, 5, and 6).

#### Step 4: SkeletonCode and Modification Identification.

Next, each BMU is identified. For each BMU, its SkeletonCode and modification string are generated. These units are then linked together with their glycosidic bond information, followed by any ambiguous structures, listed in sorted order. The pseudo-code for this step is as follows:

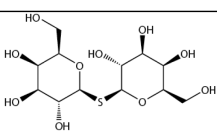
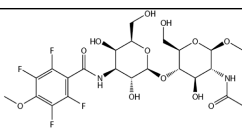
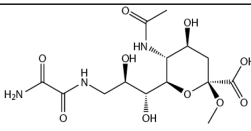
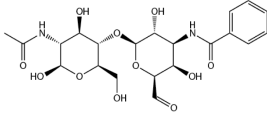
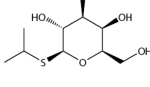
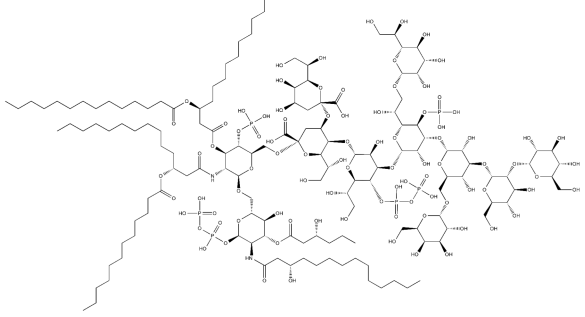
For each chain in the backbone list:

- Generate the SkeletonCode.
- Attach MOD to SkeletonCode (if MOD is attached only to this monosaccharide) to form BMU, and store extra connection information as MLU.

**Step 5: Generate WURCS.** WURCS is obtained as a character string as follows. First, the components of the glycan (SkeletonCodes, COLINs, and ALINs) are sorted (see Supplemental Codes 7, 8, and 9). Next, the order of the COLINs is determined (see Supplemental Code 10), and then the COLIN +ALINs (modifications) are sorted (see Supplemental Code 11). Thus, the main WURCS string can be determined, which results in the following format. Note that parentheses are added for clarity and are not actually used in the WURCS string.

```
WURCS=1.0/2,1/[SkeletonCode1]
(intra-COLIN1,intra-COLIN2)ALIN1]
[SkeletonCode2](intra-COLIN3)ALIN2]
(intra-COLIN4)ALIN3]
(inter-COLIN1,inter-COLIN2)ALIN4
```

Table 6. Structures in PDB That Cannot Be Uniquely Represented in Any Other Major Carbohydrate Formats

PDB ID	Chemical Structure	WURCS
1A78		WURCS=1.0/2,1/[12112h 1,5][12112h 1,5]1+1:1,2+1:1*S*
1KJR		WURCS=1.0/2,1/[12112h 1,5]3*N=^ZC(CC^ZCCCC\$4)/9 F/8F/7OC/6F/5F/3O][12122h 1,5]2*NCC/3=O]1+1,2+4
2G5R		WURCS=1.0/1,0/[a2d21122H 2,6]2*OC 5*NCC/3=O]9*N CCN/4=O/3=O]
2XG3		WURCS=1.0/2,1/[12112o 1,5]3*NC(CC^ZCCCC\$4)/3=O][ 12122h 1,5]2*NCC/3=O]1+1,2+4
3VT2		WURCS=1.0/1,0/[12112h 1,5]1:1*S]
1FI1		WURCS=1.0/11,10/[111222h 1,5][22112h 1,5][22122h 1,5] [22122h 1,5][22122h 1,5][111222h 1,5]4*OPO/3O/3=O][21 1221h 1,5]4*OPOPO/5O/5=O/3O/3=O][a2d1122h 2,6][a2d 1122h 2,6][12122h 1,5]2*NCCC^ROCCCCCCCCCCCC/7 =O/5CCCCCCCCCCCC/3=O]3*OCCC^SOCCCCCCCCCCCC CCCC/7=O/5CCCCCCCCCCCC/3=O]4*OPO/3O/3=O][221 22h 1,5]2*NCCC^SCCCCCCCCCCCCC/5O/3=O]3*OCCC^R CCC/5O/3=O]1+1,6+7 2+1,5+6 3+1,4+2 4+1,5+3 5+1,6+3  6+1,7+3 7+1,9+5 8+2,9+4 9+2,10+6 10+1,11+6

Note that here, when a monosaccharide has multiple modifications, they are separated by vertical bars "|". Some examples of WURCS strings representing major monosaccharides found in mammalian organisms are listed in Table 3.

**Ambiguous Structures.** Here, we describe how WURCS can represent ambiguity in carbohydrate structures. We list the various types of ambiguity that may occur:

- *Composition information only:* only BMUs and no MLU information needed, e.g.

WURCS=1.0/3,0/[12112h|1,5][11221m|1,5]  
[X2122h|1,5]2\*NCC/3=O]

- *Repeating units:* The smallest repeating unit must be specified.

→ The number of repeats unknown (e.g.,  $n$  times): For example, the WURCS for lactosamine will be as follows:

WURCS=1.0/2,1/<n[12112h|1,5]  
[X2122h|1,5]2\*N]1+1,2+4>

→ The number of repeats is within a range (5–10 times): For example, the WURCS for lactosamine repeating 5 to 10 times will be

WURCS=1.0/2,1/<5-10[12112h|1,5][X2122h|1,5]2\*N]  
1+1,2+4>

- *Unidentified glycosidic bond information* (unknown attachment sites). The BMUs are sorted, and possible attachment sites are indicated by backslashes.

Figure 4 is an example of an ambiguous structure (drawn using CFG representation<sup>13</sup>), where three sialyl-lactosamine structures can be attached to either of the terminal mannoses on the N-glycan core. An additional sialic acid can further modify one of these sialyl-lactosamines. In this case, the WURCS string would be as follows.



WURCS=1.0/15,14/[a1d21122h12,6l5\*NCC/3=O][a1d2  
1122h12,6l5\*NCC/3=O][12112h1,5][12122h1,5l2\*NC  
C/3=O][a1d21122h12,6l5\*NCC/3=O][12112h1,5][12  
122h1,5l2\*NCC/3=O][a1d21122h12,6l5\*NCC/3=O]  
[12112h1,5][12122h1,5l2\*NCC/3=O][21122h1,5][21  
122h1,5][11122h1,5][12122h1,5l2\*NCC/3=O][X212  
2h1,5l2\*NCC/3=O]1+2,(2+?)\ (3+?)\ (4+?)\ (5+?)\  
(6+?)\ (7+?)\ (8+?)\ (9+?)\ (10+?)l2+1,3+3l3+1,(4  
+3)\ (4+4)l4+1,(11+2)\ (12+2)l5+1,6+3l6+1,(7+3)\  
(7+4)l7+1,(11+2)\ (12+2)l8+1,9+3l9+1,(10+3)\ (10+  
4)l10+1,(11+4)\ (11+6)\ (12+4)\ (12+6)l11+1,13+6l  
12+1,13+3l13+1,14+4l14+1,15+4

The WURCS string for ambiguous glycan structure can be generated using sorting of BMU after partitioning of glycan (see Supporting Information S16, S17, and S18 and Codes 12, 13, and 14).

## RESULTS

In this section, we present examples of WURCS and some analytical results performed using WURCS. Table 4 lists examples of WURCS strings for major glycan structures including lactosamine and ganglioside GD1a, and Table 5 lists the major chemical compound databases and the number of corresponding unique WURCS strings that could be generated from their data. We selected structures from these databases that contained carbohydrate structures. Duplicates may occur for glycoconjugates with different aglycons but having the same carbohydrate structure. On the other hand, multiple carbohydrate structures may appear in a single database entry, resulting in multiple WURCS strings. In total, we could convert 12 170 carbohydrate structures in six databases (ChEBI,<sup>14</sup> PDB,<sup>11</sup> KEGG,<sup>15</sup> GlycoNAVI,<sup>16</sup> JMSDB,<sup>17</sup> and LfDB<sup>18</sup>).

Next, we took some unique structures from the PDB and generated WURCS strings for them (Table 6). In particular, we selected those structures that could not be represented by unique linear strings in any other major carbohydrate formats (listed in Table 1).

## DISCUSSION

In this work, we present a new linear text format for representing complex carbohydrate structures, and we show its advantages over previous formats, especially GlycoCT which is used heavily as a generally comprehensive glycan format. We are currently developing GlycoCT-to-WURCS converter software such that many of the existing data can be converted to WURCS (see Supporting Information S16 and Code 12).

We note that in this version of WURCS (ver. 1.0), when the reducing end is an aglycon attached to a carbon other than the anomeric position, the aglycon is currently treated as a modification (and thus included in WURCS). In general, however, reducing-end aglycons are not included in the resulting WURCS string. We also make note that the sorting functions used in WURCS may differ in some aspects from the viewpoint of glycoscience in general as we attempt to maintain uniqueness for all carbohydrate structures. Thus, in many of the sorting functions, we use the SkeletonCode as the last resort

when previous rules produce the same priority for different structures.

The advantages of WURCS allow it to be used as a unique string to encapsulate a wide range of ambiguity, which is often found in publications reporting carbohydrate structures. We described how WURCS can be used to represent ambiguous structures where glycosidic linkage positions are unknown. Cyclic structures can also be represented because of the ordering of BMUs and COLINs. WURCS was developed in order to be able to quickly identify unique structures so that it can be eventually used on the Semantic Web, possibly as a part of URIs. Thus, although this format is not human-readable, its intention is to ensure that unique identifiers can be generated for any possible carbohydrate structure published in the literature. This will then make it straightforward to build a carbohydrate structure repository by which accession numbers can be applied to any unique structure registered. In this case, both the WURCS string and WURCSKey string should be allowed when accessing carbohydrate data via a URL.

## ASSOCIATED CONTENT

### Supporting Information

Definition of terms, supplemental code, generation of ALIN string, output example of an ALIN, and SkeletonCode Characters. This material is available free of charge via the internet at <http://pubs.acs.org>. Additional information is provided separately at WURCS Working Group Web site (<http://www.wurcs-wg.org/>), where the software used to generate WURCS from MOL files is provided.

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [issaku@noguchi.or.jp](mailto:issaku@noguchi.or.jp).

### Present Address

▽Center for iPS Cell Research and Application, Kyoto University, Sakyo, Kyoto 606-8507, Japan.

### Author Contributions

K.F.A.-K., M.Ko., K.T., and I.Y. wrote the manuscript, and all authors helped finalize the text. K.F.A.-K. and I.Y. first designed WURCS, and S.K., H.S., T.S., and N.F. participated in initial discussions. K.T. developed the code to convert Molfiles to WURCS. M.Ka. developed code to convert PDB data to Molfiles. S.T. developed the pseudo code to convert glycan sequence to WURCS. H.N. oversaw the project.

### Funding

This research was supported by National Bioscience Database Center (NBDC), Japan Science and Technology 469 Agency (JST), and National Institute of Advanced Industrial Science and Technology (AIST) in Japan.

### Notes

The authors declare no competing financial interest.

## ABBREVIATIONS

WURCS, Web3 Unique Representation of Carbohydrate Structures; ALIN, Atomic Linear Notation; COLIN, Connection Linear Notation; NB, Number of Backbone; PCB, Position of the connected Carbon in the Backbone; DMB, Direction of Modification on the Backbone carbon; PCA, Position of connected backbone Carbon in ALIN; MOD, MODification; BMU, Basic Monosaccharide Unit; MLU, Basic Monosaccharide Linkage Unit; InChI, IUPAC International Chemical Identifier; URI, Uniform Resource Identifier; HTTP, hypertext transfer protocol; URLs,



Uniform Resource Locators; CCSD, Complex Carbohydrate Structure Database; IUBMB, International Union of Biochemistry and Molecular Biology; CIP rules, Cahn–Ingold–Prelog priority rules; IUPAC, International Union of Pure and Applied Chemistry; ChEBI, Chemical Entities of Biological Interest; PDB, Protein Data Bank; KEGG, Kyoto Encyclopedia of Genes and Genomes; JMSDB, Japan Monosaccharide Database; LfDB, Lectin Frontier DataBase; CFG, The Consortium for Functional Glycomics; D-Glc, D-Glucose; D-Glcp, D-Glucopyranose; D-Glcf, D-Glucofuranose; D-GlcpN, 2-Amino-2-deoxy-D-glucopyranose; D-GlcpNAc, 2-Acetamido-2-deoxy-D-glucopyranose; D-Manp, D-Mannopyranose; D-Manf, D-Mannofuranose; D-Galp, D-Galactopyranose; D-GalpNAc, 2-Acetamido-2-deoxy-D-galactopyranose; D-GalpA, D-Galactopyranuronic acid; L-Fucp, L-Fucopyranose; D-Neup5Ac, 5-Acetyl amino-3,5-dideoxy-5-Acetamido-3,5-dideoxy-D-glycero-D-galacto-non-2-ulopyranosonic acid

## REFERENCES

- (1) Banin, E.; Neuburger, Y.; Altshuler, Y.; Halevi, A.; Inbar, O.; Nir, D.; Dukler, A. A Novel Linear Code(r) Nomenclature for Complex Carbohydrates. *Trends Glycosci. Glycotechnol.* **2002**, 127–137.
- (2) Herget, S.; Ranzinger, R.; Maass, K.; von der Lieth, C. W. GlycoCT—a unifying sequence format for carbohydrates. *Carbohydr. Res.* **2008**, 343, 2162–2171.
- (3) Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Sci. Am.* **2001**, 29–37.
- (4) Doubet, S.; Albersheim, P. CarbBank. *Glycobiology* **1992**, 2, 505.
- (5) IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Abbreviated terminology of oligosaccharide chains. Recommendations 1980. *Arch. Biochem. Biophys.* **1983**, 220, 325–329. *Eur. J. Biochem.* **1982**, 126, 433–437. *J. Biol. Chem.* **1982**, 257, 3347–3351. *Pure Appl. Chem.* **1982**, 54, 1517–1522.
- (6) Böhne-Lang, A.; Lang, E.; Förster, T.; von der Lieth, C. W. LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr. Res.* **2001**, 336, 1–11.
- (7) Aoki, K.; Yamaguchi, A.; Ueda, N.; Akutsu, T.; Mamitsuka, H.; Goto, S.; Kanehisa, M. KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res.* **2004**, 32, W267–W272.
- (8) GLYDE-II. <http://glycomics.ccr.cuga.edu/core4/informatics-glyde-ii.html> (accessed December 18, 2013).
- (9) Lüttetke, T. MonosaccharideDB. <http://www.monosaccharidedb.org/> (accessed May 14, 2012).
- (10) Moss, G. P. Basic terminology of stereochemistry (IUPAC Recommendations 1996). *Pure Appl. Chem.* **1996**, 68, 2193–2222.
- (11) Berman, H. M.; Kleywegt, G. J.; Nakamura, H.; Markley, J. L. The future of the protein data bank. *Biopolymers* **2013**, 99, 218–222.
- (12) Technical Manual of the InChI. [http://www.inchi-trust.org/fileadmin/user\\_upload/software/inchi-v1.04/InChI\\_TechMan.pdf](http://www.inchi-trust.org/fileadmin/user_upload/software/inchi-v1.04/InChI_TechMan.pdf) (accessed April 4, 2014).
- (13) CFG Nomenclature for Representation of Glycan Structure. <http://www.functionalglycomics.org/static/consortium/Nomenclature.shtml> (accessed September 12, 2013).
- (14) Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; Steinbeck, C. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* **2013**, 41, D456–D463.
- (15) Goto, S.; Okuno, Y.; Hattori, M.; Nishioka, T.; Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **2002**, 30, 402–404.
- (16) Yamada, I. GlycoNAVI. <http://www.glyconavi.org/> (accessed August 18, 2013).
- (17) Shikanai, T. JMSDB. <http://jcggdb.jp/search/ChemGlycan.cgi> (accessed August 18, 2013).
- (18) Tateno, H.; Nakamura-Tsuruta, S.; Hirabayashi, J. Frontal affinity chromatography: sugar-protein interactions. *Nat. Protoc.* **2007**, 2, 2529–2537.