

GlycoCT



A sequence format and namespace for
complex oligo- and polysaccharides

Version 2
„KUROI“

11-11-06

S. Herget, R. Ranzinger, W.v.d.Lieth
Central Spectroscopic Department
DKFZ Heidelberg

Content

Table of contents

Content.....	2
Introduction.....	4
History.....	4
Motivation.....	4
Glossary: Definition of terms.....	5
The sequence format – overview.....	6
A first small example.....	6
The RES section.....	7
Naming conventions and descriptors for carbohydrate portions.....	7
Basetypes.....	7
Superclasses.....	7
Stem types.....	9
Anomeric center.....	10
Configuration	10
Ringsize.....	10
Modifications of the carbohydrate stem types.....	11
Stereoloss by deoxygenation.....	11
Carbonyl function other than C1.....	11
Alditols.....	11
Double bond.....	11
Acidic function.....	11
SP2-Hybrids.....	11
Geminal	12
Substituents.....	13
Non-monosaccharide entities.....	14
Special entities.....	14
The LIN section.....	15
Compositions.....	16
Undefined and partially known linkages.....	16
Multiconnected residues.....	17
Circular subgraphs.....	17
The PRO section.....	18
The REP section.....	19
The STA section	20
The facultative ISO section.....	21
The facultative AGL section.....	22
Sorting GlycoCT – how to guarantee uniqueness.....	23
Examples.....	24
Examples: Small entities.....	24
Ketoses.....	24
Deoxy sugars.....	24
Acidic sugars.....	24
Uronic acids.....	24
Aldonic acids.....	24
Aldaric acids.....	25

Amino sugars.....	25
Thio sugars.....	25
Alditols.....	25
Intramolecular anhydrides.....	26
Unsaturated monosaccharides.....	26
Lactonized carbohydrates.....	26
Sialic acids	26
Examples: Longer sequences.....	27
XML {GlycoCT}.....	29

Introduction

History

Motivation

This format in its condensed, sorted form is designed to be a hashcode for oligosaccharides in database applications. The more verbose XML – syntax facilitates parsing of the sequences. Its information storage capacity is focused on oligo- and polysaccharides. The namespaces of all entities are controlled as much as possible, with an emphasis on a **controlled vocabulary for carbohydrates**, which is fully machine-readable.

All potential information of the established sequence formats in the glycobioinformatics can be captured with this format. We have taken the following formats into account:

- Glyde (XML) as of version 1.0, 1.1 and preliminary version 2.0
- LINUCS
- IUPAC-condensed form
- IUPAC- 2D (ASCII – graph)
- KEGG-KCF
- Glycominds LinearCode™
- BCSDB – format
- CabosML

This format can serve as an **unique identifier** for any glycan structure, even in the case of ambiguity in the structural description. None of the existing sequence formats is complete in this regard. The following structural features can be encoded:

- linear and branched structures
- compositions
- partially known topologies
- ambiguities on the linkage level
- fuzzy connectivities of subbranches
- statistical distribution of subgraphs along a carbohydrate sequence
- repeating units as completely distinct subgraphs
- circular structures
- multiconnected residues

Glossary: Definition of terms

Basetype	<p>A basetype is a description of a stereochemically defined structure without any substituent from the chemical class of polyhydroxyaldehydes or -ketones. On this level acidic functions, double bonds, deoxygenations leading to stereochemical information loss, reductions of C1 – carbonyls and additional keto functions are encoded.</p> <p><i>Synonyms: Core monosaccharide, stem type, basic molecular framework, basic carbohydrate</i></p>
Substituent	<p>A non-basetype entity with exactly one linkage to a basetype. The list of known substituents is actively maintained.</p>
Monosaccharide	<p>A basetype including its substituents.</p>
Aglycon	<p>Commonly used as a term to describe non-carbohydrate moieties at the reducing end. Within GlycoCT it is more generally a non-basetype, non-substituent entity with linkages to an monosaccharide.</p>
Oligosaccharide	<p>A metastructure of covalently connected entities, includes monosaccharides and aglyca.</p>
Sugar graph	<p>The part(s) of an oligosaccharide which are constituted of monosaccharides only.</p>
Linked unit	<p>A basetype including outgoing linkages (its connections) to other entities, excluding connections in the direction of the reducing terminus (typically C1 for aldoses).</p>
Carbohydrate component	<p>A monosaccharide with all its neighbouring monosaccharides, excluding connections in the direction of the reducing terminus.</p>
Glycosidic linkage	<p>The bond between a hemiacetalic C atom and a OH group, connecting 2 residues. The most prevalent linkage to construct oligosaccharides.</p>

The sequence format – overview

The format contains six discrete sections. It can be extended in the future by more sections to cover additional structural features and to hold more information if needed.

RES	List of connected entities (<i>Residue</i>)
LIN	Topology information. A list of all topologically unique linkages (<i>Linkage</i>)
ISO	Isotopic Labelling (<i>Isotope</i>)
PRO	Linkages between subbranches which are not resulting in unique topologies due to experimental results (<i>Probabilistic</i>)
REP	Definition of repeating units as distinct subgraphs (<i>Repeat</i>)
STA	Definition of subgraphs with statistical distribution (eg. 60% sulfated residues in glycosaminoglycans) (<i>Statistical</i>)

Uppercase letters are mandatory for the section identifiers. The main RES-section is mandatory, followed by the optional LIN section. The rest of the sections appears in the alphabetical order of their respective keywords. Each line is terminated by a „;“. Sequences start with the reducing terminus in general.

A first small example

This figure shows a first simplistic view on the GlycoCT format:

General structure of the sequence format

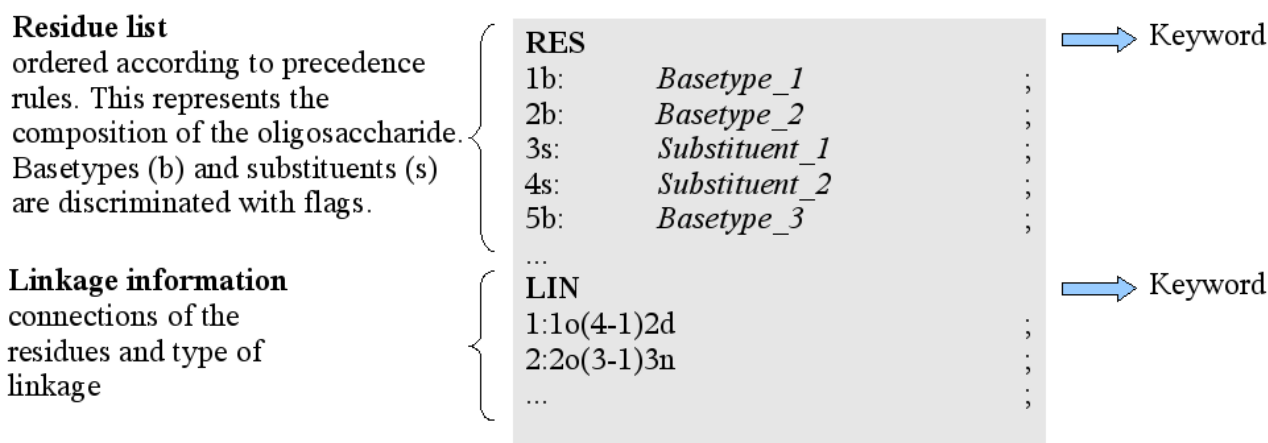


Figure 1: General Structure of the format - Adjacency List

The RES section

This section contains all occurring basetypes, substituents and other entities in the current subgraph. The list is ordered according to the general sorting scheme (see respective chapter). The entries are numbered consecutively, followed by the mandatory type identifier, which has the following possible values:

b	basetype
s	substituent
n	other chemically defined entity (freetext identifier for historical data)
i	INCHI encoded non-basetype, non-substituent chemical entity
r	repeating unit
s	statistical unit

Naming conventions and descriptors for carbohydrate portions

This namespace is designed to result in unique identifiers for carbohydrates, including common modifications of this substance class. It is designed to be easily machine-readable and to resemble traditional IUPAC definitions as much as possible. Trivial IUPAC – names for carbohydrates are not allowed.

Definition:

A carbohydrate is a polyhydroxyaldehyde or – ketone. For convenience the substance class of alditols is integrated in this framework. Inositols are not treated as carbohydrates.

Basetypes

A basetype is per definition the main chain of C-atoms as defined in IUPAC definitions for carbohydrates.

Superclasses

Each basetype descriptor is suffixed with its superclass. The following superclasses are defined:

number of linear oriented C- atoms	3-letter-code	long name
3	TRI	Triose
4	TET	Tetrose
5	PEN	Pentose
6	HEX	Hexose
7	HEP	Heptose
8	OCT	Octose
9	NON	Nonose
10	DEC	Decose
11	S11	Undecose
12	S12	Dodecose
13	S13	Tridecose
14	S14	Tetdecose

Figure 2: Superclass definitions

Basetypes with more than 14 C-atoms can be named with the S[NN]-notation, allowing for basetypes with up to 99 C-atoms to be encoded.

Stem types

A basic carbohydrate stem type is defined by its stereochemistry. The following basic stereochemically distinct entities are defined by IUPAC and will be used in the sequence format. These aldoses are the fundamentum of the namespace, each alteration from these basetypes is encoded in the carbohydrate nomenclature.

Configuration	3-letter-code	long name	superclass
D	GRO	Glyceraldehyde	TRI
D	ERY	Erythrose	TET
D	RIB	Ribose	PEN
D	ARA	Arabinose	PEN
D	ALL	Allose	HEX
D	ALT	Altrose	HEX
D	GLC	Glucose	HEX
D	MAN	Mannose	HEX
D	TRE	Threose	TET
D	XYL	Xylose	PEN
D	LYX	Lyxose	PEN
D	GUL	Gulose	HEX
D	IDO	Idose	HEX
D	GAL	Galactose	HEX
D	TAL	Talose	HEX
L	GRO	Glyceraldehyde	TRI
L	ERY	Erythrose	TET
L	RIB	Ribose	PEN
L	ARA	Arabinose	PEN
L	ALL	Allose	HEX
L	ALT	Altrose	HEX
L	GLC	Glucose	HEX
L	MAN	Mannose	HEX
L	TRE	Threose	TET
L	XYL	Xylose	PEN
L	LYX	Lyxose	PEN
L	GUL	Gulose	HEX
L	IDO	Idose	HEX
L	GAL	Galactose	HEX
L	TAL	Talose	HEX

Figure 3: Basetype names are chosen according to IUPAC

Carbohydrates with more than 4 stereogenic C-atoms

For carbohydrates with more than 4 stereogenic centers an additional rule is required. We follow established IUPAC conventions for the naming purposes (*2-Carb-2.2.2. Choice of parent name, IUPAC Nomenclature of carbohydrates*), which result in composite names. Trivial names are not supported by this format.

Example:

5-deoxy-D-glycero- α -D-galacto-non-2-ulopyranosonic acid (basetype for neuraminic acid)

5-deoxy-DGro- α -DGal-non-2-ulopyranosonic acid (abbreviated notation)

Anomeric center

This information is mandatory. Possible values:

a	alpha
b	beta
x	unknown
o	open = linear (no anomeric center exists)

Configuration

This information is mandatory. Possible values:

d	Dexter
l	Laevus
x	unknown

Ringsize

The descriptor for the ringsize follows always directly the basetype and is mandatory for all monomers. The position identifiers of the C-atoms forming a hemiacetal are written in numerical order and are separated by a „:“.

In the case of unknown ringsize and -closure the special character x encodes the ambiguity. Linear structures receive a symbolic „0:0“ ([zero]:[zero]). These descriptors are mandatory.

Modifications of the carbohydrate stem types

Modifications are noted in conjunction with the stem type in the following cases, as the stereoconfiguration is altered. This information should be stored in close proximity to the basetype.

- stereoloss by deoxygenation
- carbonyl function other than C1
- reduction of primary aldehyde function (alditols)
- double bond
- acidic function
- geminal OH
- SP²-Hybrids

Stereoloss by deoxygenation

Stereoloss is a common result of modifications of basetypes and is indicated by a „d“ (deoxy). We follow IUPAC-conventions for naming. For the definition of a basetype with stereoloss only the remaining stereogenic centers are taken into account, which results in composite names. Trivial names like Rhamnose or Paratose are not allowed.

Carbonyl function other than C1

The keyword „keto“ is used to define the position of a carbonyl function at a position different from C1. It substitutes the original carbonyl function. Keto and alditol definitions on the same C-atom are deprecated.

Alditols

Reduction of the C1 aldehyde function is indicated by the keyword „aldi“. Only C1 reductions of basetypes are allowed.

Double bond

The keywords „en“ indicates the existence of a double bond, joining two adjacent C-atoms in the backbone. Concurrent eliminations of OH – groups have to be indicated with the deoxygenation-flag.

An unknown deoxygenation pattern can be indicated with the modification „enx“.

Acidic function

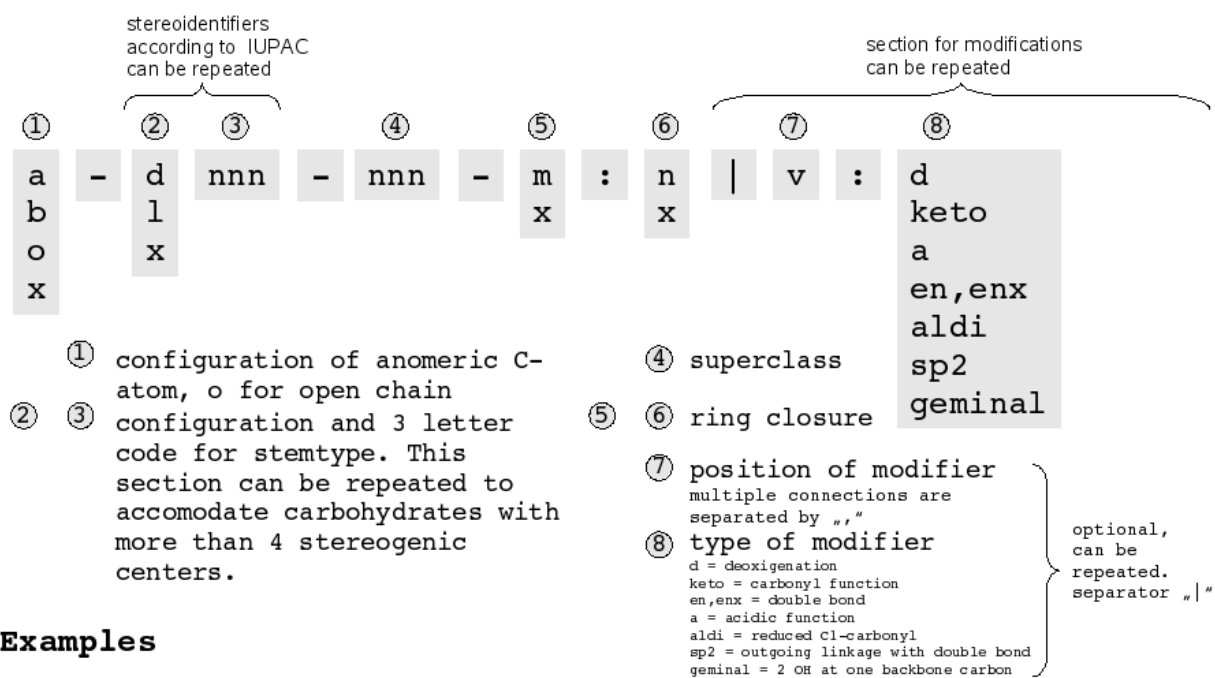
The keyword „a“ indicates the existence of an acidic function.

SP₂-Hybrids

SP² -hybridisation due to a outgoing double bond to a substituent is indicated with the keyword „sp²“.

Geminal

Should a geminal substitution occur in the backbone of the basetype, it can be indicated with the „geminal“ keyword.



Examples

b-dglc-hex-1:5	β-D-Glucose, pyranose form
x-lgro-tet-0:0 3:d	?-3-deoxy-1-grotet (CarbBank notation)
a-drib-hex-x:x 3:d 6:d	Paratose: 3,6-dideoxy-D-ribo-hexose, alpha anomer
a-lgal-hex-1:5 6:d	LFuc, alpha anomer (KEGG notation)
b-dgro-dgal-hep-1:5 7:d	7-deoxy-β-D-Gro-D-Gal-Hepp

Figure 5: General schema for monosaccharides

Substituents

An analysis of 85.000 existing entries in glycorelated databases lead to the definition of a comprehensive list of common substituents. While adding substituents to basetypes, as a general rule, basetype integrity has to be preserved as much as possible.

Other, not listed chemical entities attached to the sugar graph are consequently treated as non-monosaccharide residues.

Monovalent substituents

Symbol	Formula	Bond order	Remarks
acetyl	COCH ₃	1	
bromo	Br	1	
chloro	Cl	1	
ethyl	CH ₂ CH ₃	1	
ethanolamine	CH ₂ NHCH ₂ OH	1	
flouro	F	1	
formyl	CHO	1	
glycolyl	COCH ₂ OH	1	
hydroxymethyl	CH ₂ OH	1	
imino	NH	2	
iodo	I	1	
(r)-lactate	CH ₃ CHCOOH	1	
(s)-lactate	CH ₃ CHCOOH	1	
methyl	CH ₃	1	
n	NH ₂	1	amino
n-acetyl	NHCOCH ₃	1	aminoacetyl
n-alanine	NHCOCHNH ₂ CH ₃	1	aminoalanine
n-dimethyl	N(CH ₃) ₂	1	aminoformyl
n-formyl	NHCHO	1	
n-glycolyl	NCOCH ₂ OH	1	
n-methyl	NHCH ₃	1	
n-succinate	NCOCH ₂ CH ₂ COOH	1	
n-sulfate	NHSO ₃ H	1	
n-trifluoroacetyl	NHCOCF ₃	1	
nitrat	NO ₂	1	
phospate	PO ₃ H ₂	1	
pyruvate	COCOCH ₃	1	
sulfate	SO ₃ H	1	
thio	SH	1	

Divalent substituents

(r)-pyruvate	CH ₂ CCOOH	2 * 1	
(s)-pyruvate	CH ₂ CCOOH	2 * 1	
(r)-lactate	CH ₃ CHCO	2 * 1	
(s)-lactate	CH ₃ CHCO	2 * 1	
anhydro	-H ₂ O from basetype (intramolecular ether)	2 * 1	
lactone	-H ₂ O from basetype (intramolecular ester)	2 * 1	
epoxy	-H ₂ O from basetype (neighbouring C-atoms)	2 * 1	

Figure 6: A small, standarized substituent table is used as a seed list for GlycoCT. Future additions to this list are welcome on a „as needed“ basis for common substitutions. Attachment positions are highlighted.

Non-monosaccharide entities

All additional chemical entities which can not be identified as basetypes or substituents are non-monosaccharide entities. They receive a special mark-up in the residue list („n“). Two approaches to name them are possible:

1. Use a free text field to name the entity with a systematic IUPAC name. This approach is introduced for backward compatibility with old sequence format.
2. Use an INCHI code to identify the non-monosaccharide entity uniquely. These INCHI encoded non-monosaccharide entities receive the special flag „i“.

Solution #2 is preferred for future use.

Special entities

Two special entries in the residue list („REP“ and „STA“) mark connections to subgraph lists for special purposes. See the respective section descriptions for details.

The LIN section

The LIN section contains the topologically unique connectivities of the entities in the RES section. Each connectivity is numbered consecutively. The list of connectivities is ordered as described in the chapter „Sorting GlycoCT“.

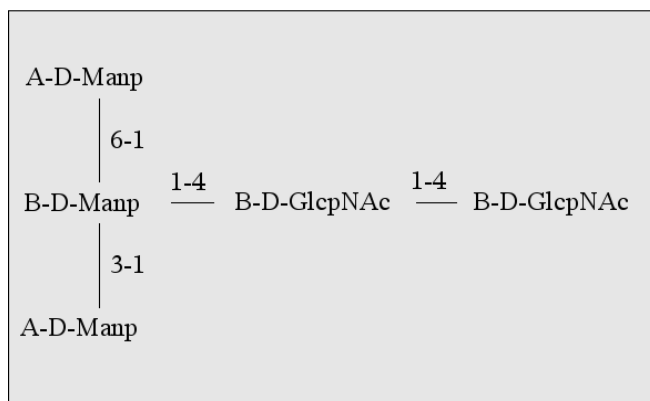
All linkages contain a **type identifier**, which indicates the substitution pattern:

o	hydrogen from OH – function removed and substituent attached at this position
h	hydrogen removed and substituent attached at this position
d	OH-function removed and substituted
n	linkage to non-monosaccharide entities, repeat or statistical units
r	prochiral H-atom removed, resulting in R-configuration
s	prochiral H-atom removed, resulting in S-configuration

The following rules should be followed when constructing the linkage information:

- Basetype structure integrity should be preserved with high priority
- Basetype-basetype linkages are always written with anomeric C receiving a deoxygenation
- If a basetype-basetype linkage is achieved by two anomeric C atoms, the C-atom which comes at second place when following the preferred direction is deoxygenated
- If no anomeric C-atom is involved in a basetype-basetype linkage, the C-atom which comes at second place when following the preferred direction is deoxygenated

N-Glycan Core



GlycoCT

RES
1b:b-dglc-hex-1-5;
2s:n-acetyl;
3b:b-dglc-hex-1-5;
4s:n-acetyl;
5:b:b-dman-hex-1-5;
6:b:a-dman-hex-1-5;
7:b:a-dman-hex-1-5;
LIN
1:1d(2-1)2n;
2:1o(4-1)3d;
3:3d(2-1)4n;
4:3o(4-1)5d;
5:5o(3-1)6d;
6:5o(6-1)7d;

Figure 7: A first example showing the decomposition of the N-Glycan core

Compositions

The format allows storing of entities with partially or completely missing linkages between its residues.

Compositions: partially defined linkages

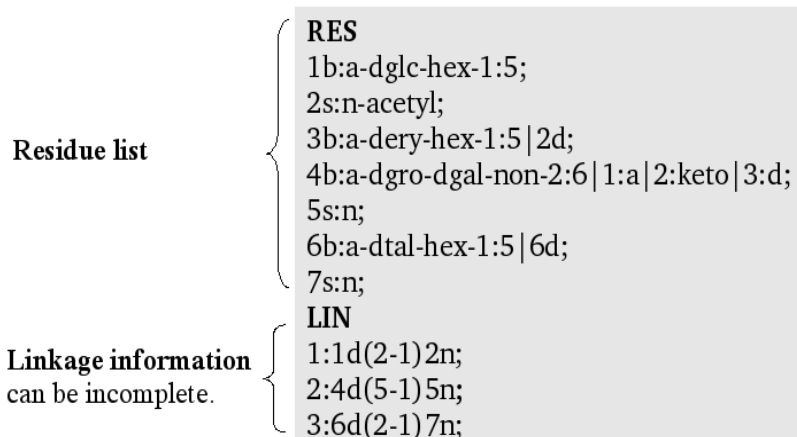


Figure 8: Compositions can be stored in GlycoCT

Undefined and partially known linkages

For different experimentally derived alternative linkages or unknown linkages all remaining linking options are listed in the format.

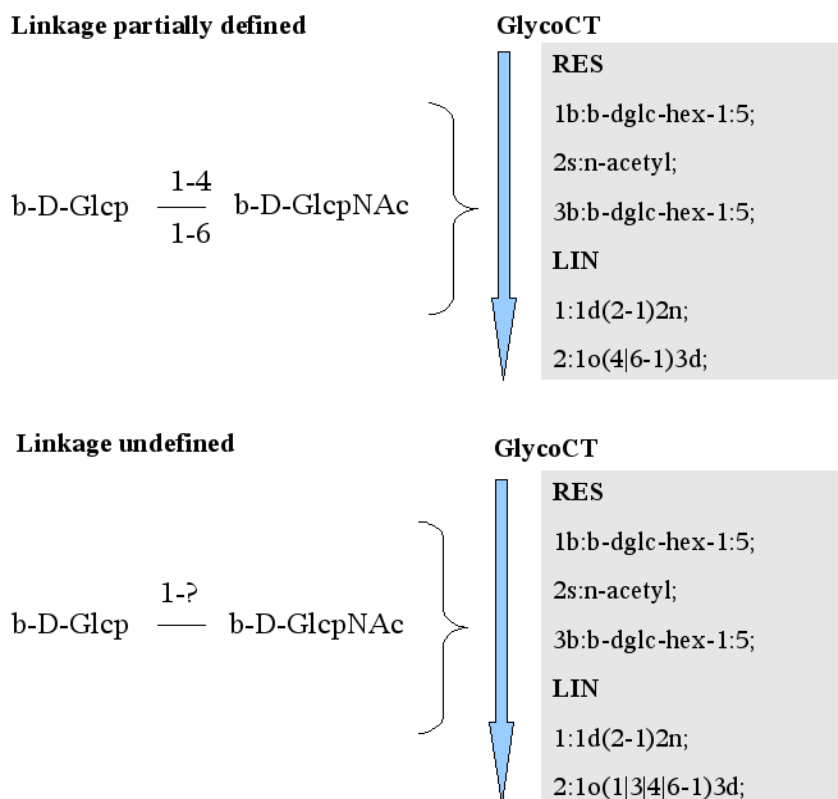
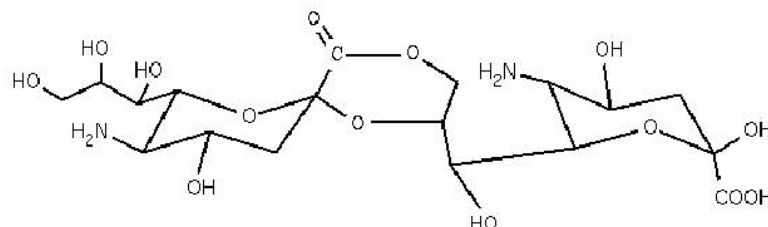


Figure 9: Partially defined linkages can be constructed with GlycoCT

Multiconnected residues

The format can handle multiply connected residues by adding additional linkages.

Example: Fragment of a sequence with a lactonized sialic acid



GlycoCT	
Residue list	RES
	1b:a-dgro-dgal-non-2:6 1:a 2:keto 3:d;
	2s:n;
	3b:a-dgro-dgal-non-2:6 1:a 2:keto 3:d;
Linkage section	4s:n;
	LIN
	1:1d(5-1)2n;
	2:1o(8-2)2d;
	3:1o(9-1)2d;
	4:3d(5-1)4n;

Figure 10: Two lactonized sialic acids represented in GlycoCT

Circular subgraphs

Circular structures can be written by adding appropriate cyclic linkages. The implications for the sorting algorithm are described in the respective chapter.

The PRO section

Some experimental techniques result in partially unresolved structures of oligosaccharides regarding the exact location of terminal residue(s). These can be encoded with the following notation:

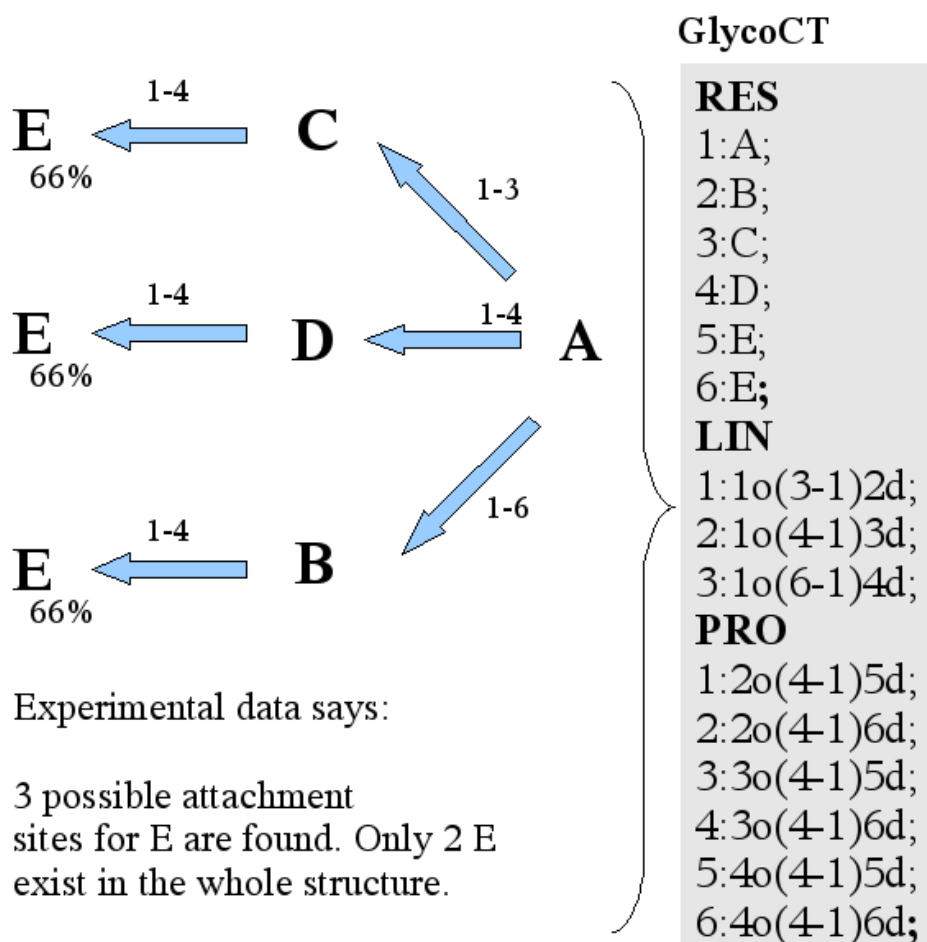


Figure 11: Ambiguous terminal subgraph connectivity can be modelled with the PRO section

Only homogenic complex carbohydrates (distinct structures) will be encoded in this notation. Mixtures of structurally different complex carbohydrates are stored as separate sequences.

The REP section

Some naturally occurring glycan structures especially from unicellular organisms or plants contain structural elements which are repeated n to m times in the resulting oligosaccharide. The section REP contains the residues and connecting linkages which form the repeat unit. The multitude of the repeat unit needs to be noted with a starting „=“ and two integer numbers divided by a „-“. If the number of repeating units is known exactly, both numbers are the same. If it is unknown, a „?-?“ is used to indicate it.

More than one repeat section may exist in a complex sugar graph. All repeating units are listed in the sequence of the general sorting scheme with consecutive numbers. The smallest possible repeat unit has to be declared.

The prerequisite for the usage of the repeating section is a number of at least **seven** repeating elements.

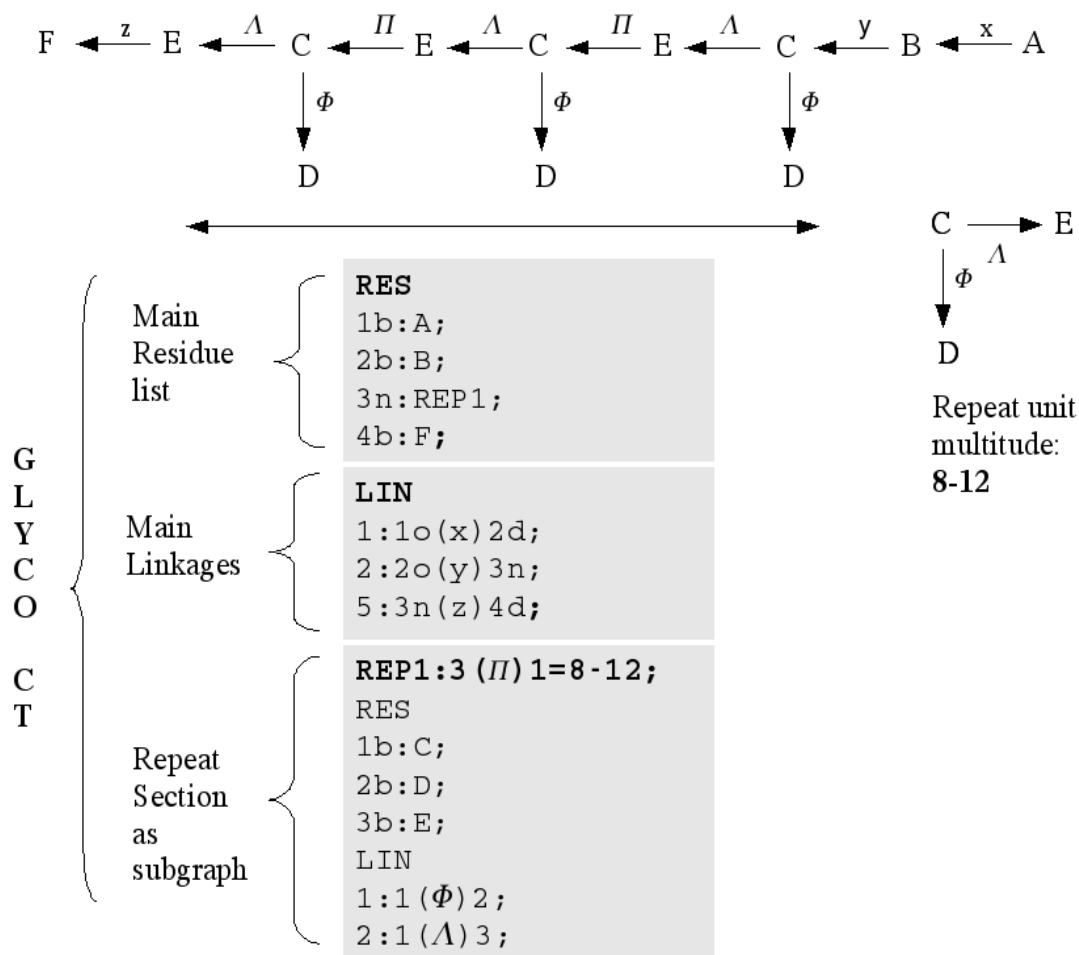


Figure 12: Repeating Units are represented as subgraphs, allowing for high flexibility

The STA section

The section STA contains information about substructures which are known to occur only on a probabilistic level. Certain residues, in most cases members of repeat units, can be marked to carry a modification on a fraction of the total sequence. This feature is encountered for instance in glycosaminoglycans, which are sulfated only to a certain degree. Example:

DERMATANSULFATE

[L-Iduronat(α 1 \rightarrow 3)D-GalNAc(4-Sulfate)(β 1 \rightarrow 4)]_n

15% of IdoA are sulfated at position 2

Repeating 100-400 units

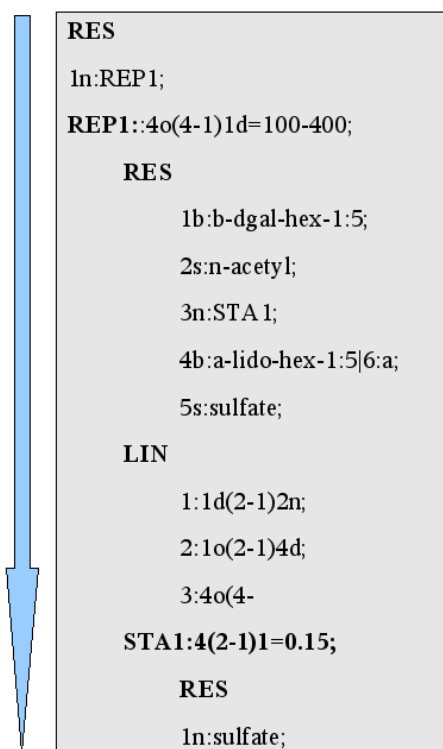


Figure 13: Subgraphs which occur randomly distributed over a oligosaccharide can be modelled with the STA section.

The facultative ISO section

Certain experimental techniques require a substitution or enrichment of atom types. These are encoded in the ISO section, using the following abbreviations:

Name	Abbreviation
Deuterium	d
Carbon-13	c13
Nitrogen-15	n15
Nitrogen-14	n14
Oxygen-17	o17
Phosphor-31	p31

These substitution types are currently defined for basetypes:

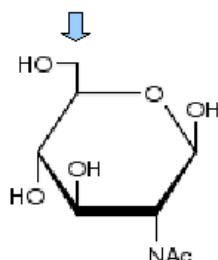
c	exchange of backbone carbon
o	exchange of O
h	exchange of H
r	exchange of R-prochiral H
s	exchange of S-prochiral H

Syntax:

[no]:[residue number][type identifier]([position-[atom_type]])

The following example shows a C13-exchange on C6 of GlcNAc:

b-D-GlcNAc, pyranose form
C13 isotopic label
on C6



RES
1b:b-dglc-hex-1:5;
2s:n-acetyl;
LIN
1:1o(2-1)2n;
ISO
1:1c(6-c13);

Figure 14: An extension of GlycoCT makes it easy to encode isotopic alterations of distinct residues

The facultative AGL section

This section is designed to contain information about special non-carbohydrate entity classes connected to the oligosaccharide.

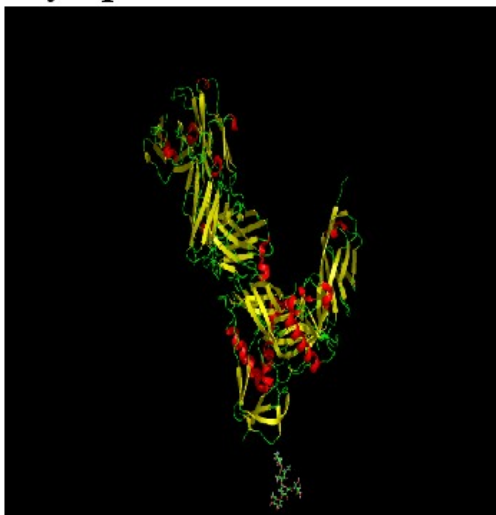
Currently the following classes are supported:

- p proteins
- l lipids
- x other molecules

The identifiers for these substance classes should be preferably chosen from existing resources such as UniProt, PubChem or LipidMaps.

The following example shows a Glycoprotein, the sugar moieties connected via the reducing terminus:

Glycoprotein



GlycoCT

RES

1b:b-dgal-hex-1:5|;

2s:n-acetyl|;

...;

LIN

1:1d(2-1)2n|;

...;

AGL

1:1o(1-p:protein_id:chain_id:residue_id:atom_id);

Namespace to be defined

Figure 15: A second example for an extension is the definition of the Aglycon (AGL) section.

Sorting GlycoCT – how to guarantee uniqueness

It is highly desirable for database applications to have an unique textual description of complex oligosaccharides, so the format can serve as a primary key in database applications to perform identity requests. The GlycoCT format can be sorted uniquely with the following rules:

Start Residue

The reducing end is generally the starting point of the sequences. If more than one connectivity group (>1 graph) is to be encoded in a single GlycoCT, the following criteria guide their sequence in the GlycoCT (rule: *size comes first*):

1. Residue count
2. Longest branch
3. Number of terminal residues
4. Number of branching points
5. Alphanumerical comparison of the GlycoCT – sequence for each putative start residue

If the sequence contains a circular structure, and cannot be resolved by the above criteria, each putative starting point is used to generate a GlycoCT code. The corresponding residue with the lowest alphanumerical sort is then used as the starting point.

Linkage Sort

Following criteria result in a unique order of all entities of the GlycoCT, when used while descending the oligosaccharides from their root residues:

1. Alphanumerical linkage sort
2. Residue count
3. Longest branch
4. Number of terminal residues
5. Number of branching points
6. Alphanumerical comparison of the GlycoCT – sequence

RES, LIN and other subgraph sections are ordered according to these rules.

Examples

Examples: Small entities

Ketoses

The default position for the carbonyl-function in carbohydrates is C1. This is included in the basetype definition. If a keto-function is at another position, the modification identifier „keto“ is used. All occurring keto functions have then to be listed explicitly in the notation. If the keyword „keto“ is found in the modifications, the default C1-carbonyl function is replaced implicitly by a OH-group. Keto functions have an impact on the resulting stereochemistry, as each non-terminal carbonyl-function replaces a stereogenic center. Trivial names for ketoses are deprecated.

IUPAC: D-Fructose, furanose ring, alpha form
CarbBank: a-D-Fruf

GlycoCT: RES
1b:a-dara-hex-2:5|2:keto;

Deoxy sugars

Deoxygenation is indicated with the keyword „d“. Resulting stereoloss is reflected in the basetype name.

IUPAC: 2,6-dideoxy-3-O-methyl- α -D-arabino-hexopyranose
CarbBank: a-2,6-deoxy-D-AraHexp3me

GlycoCT: RES
1b:a-dara-hex-1:5|2:d|6:d;

Acidic sugars

Acidic functions are generally indicated with the letter „a“ in the carbohydrate stem type.

Uronic acids

IUPAC: D-Glucopyranosyluronic acid
CarbBank: b-D-GlcpA

GlycoCT: RES
1b:b-dglc-hex-1:5|6:a;

Aldonic acids

IUPAC: D-Gluconic acid
CarbBank: D-Glc-onic

GlycoCT: RES

1b:o-dglc-hex-0:0|1:a;

Aldaric acids

IUPAC: D-Glucaric acid
CarbBank: D-Glc-aric
GlycoCT: RES
1b:o-dglc-hex-0:0|1:a|6:a;

Amino sugars

The mnemonic symbol for amino groups is the „n“.

IUPAC: 2,6-diamino-2,3,6-trideoxy- α -D-ribo-hexopyranose
CarbBank: a-D-3-deoxy-RibHexp2N6N
GlycoCT: RES
1b:a-drib-hex-1:5|3:d;
2s:n;
3s:n;
LIN
1:1d(2-1)2n;
2:1d(6-1)3n;

Thio sugars

The mnemonic symbol for thio-functions is the „sh“.

IUPAC: 3-amino-3,4-dideoxy-4-thio- α -D-galactopyranose
CarbBank: a-D-Galp3n4sh
GlycoCT: RES
1b:a-dgal-hex-1:5;
2s:n;
4s:thiol;
LIN
1:1d(3-1)2n;
2:1d(4-1)3n;

Alditols

Reduced sugars are indicated with the keyword „aldi“.

IUPAC: D-Arabinitol
CarbBank: d-Ara-ol
GlycoCT: RES

1b:o-dara-pen-0:0|1:aldi;

Intramolecular anhydrides

Intramolecular anhydrides are written with the special substituent lactone:

IUPAC: 3,6-anhydro- α -D-glucopyranose

CarbBank: 3,6-anhydro-a-D-Glc-p

GlycoCT: RES
1b:a-dglc-hex-1:5;
2s:lactone;
LIN
1:1d(3-6)2o;

Unsaturated monosaccharides

„en“ indicates double bonds. The 2 positions identifiers are separated by a delimiting comma.

IUPAC: 2,3-dideoxy- α -D-erythro-hex-2-en-pyranose

CarbBank: a-D-2-en-EryHexp

GlycoCT: RES
1b:a-dery-hex-1:5|2d|2,3:en|3d;

Lactonized carbohydrates

IUPAC: L-xylo-hex-2-ulosono-1,4-lactone (L-Ascorbat, Vitamin C, isomer)

CarbBank: L-Xyl-Hex-2ulo-1,4-lactone-onic

GlycoCT: RES
1b:o-lxyl-hex-0:0|1:a|2:keto;
LIN
1:1d(1-4)1o;

Sialic acids

IUPAC: Glycolamidoneuraminic acid (pyranose form) or
5-aminoglycolyl-3,5-dideoxy-D-glycero- α -D-galacto-non-2-ulopyranosonic acid

CarbBank: a-D-NeupGc (D - configuration makes no sense here, but CarbBank did it)

GlycoCT: RES
1b:a-dgro-dgal-non-2:6|1:a|2:keto|3:d;
1s:N-glycolyl;
LIN
1:1o(5-1)2;

IUPAC: 2-keto-3-deoxy-D-glycero- α -D-galacto-non-ulosonic acid
(KDN, pyranose form)

CarbBank: a-D-Kdnp (D - configuration makes no sense here, but CarbBank did it)

GlycoCT: RES
1b:a-dgro-dgal-non-2:6|1a|2:keto|3:d;

Examples: Longer sequences

Lewis X

b-D-GalpNAc[(3-1)a-L-Fucp](4-1)b-D-Galp



RES

1b:b-dgal-hex-1:5;
2s:N-acetyl;
3b:a-lgal-hex-1:5|6d;
4b:b-dgal-hex-1:5;

LIN

1:1d(2-1)2n;
2.1o(3-1)3d;
3:1o(4-1)4d;

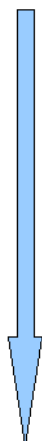
Figure 16: Lewis X

Chitin

Homopolymer b-D-GlcpNAc

[-4)b-D-GlcpNAc(1-] _n

n = 500 - 800



RES

1n:REP1;

REP1:1d(1-4)1o=250-400

RES

1b:b-dgal-hex-1:5;
2s:n-acetyl;

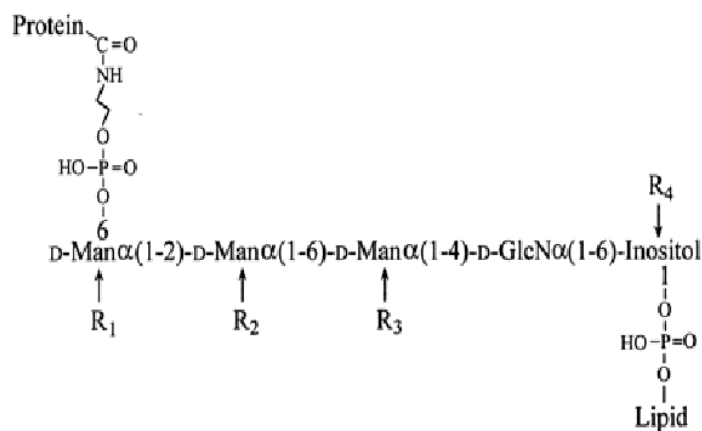
LIN

1:1d(2-1)2n;

Figure 17: One of the most abundant carbohydrates - Chitin

GPI - anchor

Glycosylphosphatidylinositol



R_n typical substitution locations, only for overview.
Not indicated in the GlycoCT.

RES
1b:a-dglc-hex-1:5;
2s:n;
3b:a-dman-hex-1:5;
4b:a-dman-hex-1:5;
5b:a-dman-hex-1:5;
LIN
1:1d(2-1)2n;
2:1o(4-1)3d;
3:3o(6-1)4d;
4:4o(2-1)5d;
AGL
1:1o(1-Phosphatidylinositol_ID);
2:5o(6-Phosphoethanolamine-Protein_ID);

Figure 18: GPI - Anchor

XML{GlycoCT}

Apart from the textual description a XML schema for GlycoCT has been defined and utilized. This facilitates parsing logic in application layers.

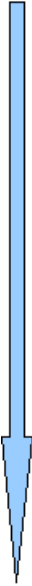
GlycoCT	XML{GlycoCT}
 <pre> RES 1n:ALPHA; 2n:BE TA; 3n:DELTA; 4n:GAMMA; /// LIN 1:1n(5-2)2n; 2:2n(2-1)3n; 3:2n(2-1)4n; </pre>	<pre> <?xml version="1.0" encoding="UTF-8"?> <BlackSugarXML Message="Simplify your life... :-)"> <RESIDUES> <R number="1" type="n" name="ALPHA" /> <R number="2" type="n" name="BE TA" /> <R number="3" type="n" name="DELTA" /> <R number="4" type="n" name="GAMMA" /> </RESIDUES> <LINKAGES> <L Number="1" Linkage="1n(5-2)2n" /> <L Number="2" Linkage="2n(2-1)3n" /> <L Number="3" Linkage="2n(2-1)4n" /> </LINKAGES> </BlackSugarXML> </pre>

Figure 19: BlackSugarXML is a simple, yet powerfull XML - syntax to map the GlycoCT information into a more structured format. The namespace for the basetypes and the linkage blocks can be extended with more attributes to simplify parsing.

GlycoCT

```

RES
1n:a-D-MAN-1:5;
2n:b-d-Tal-1:6;
3r:REP1;
4n:a-D-MAN-1:5;
///
LIN
1:1n(5-2)2n;
2:2n(2-5)3n;
3:3n(3-1)4n;
////
REPEAT SECTION///
REP1:1(3-5)1=5-8
RES
1n:a-D-MAN-1:5;
2n:Residue_in_RepeatUnit;
///
LIN
1:1n(7-7)2n;

```

XML{GlycoCT}

```

<?xml version="1.0" encoding="UTF-8"?>
<BlackSugarXML Message="Simplify your life... :-)">
  <RESIDUES>
    <R number="1" type="n" name="a-D-MAN-1:5" />
    <R number="2" type="n" name="b-d-Tal-1:6" />
    <R number="3" type="r" name="REP1" />
    <R number="4" type="n" name="a-D-MAN-1:5" />
  </RESIDUES>
  <LINKAGES>
    <L Number="1" Linkage="1n(5-2)2n" />
    <L Number="2" Linkage="2n(2-5)3n" />
    <L Number="3" Linkage="3n(3-1)4n" />
  </LINKAGES>
  <REPEAT Number="1">
    <INTERNAL InternalLinkage="1?(3-5)1=5-8">
      <SUBGRAPH>
        <RESIDUES>
          <R number="1" type="n" name="a-D-MAN-1:5" />
          <R number="2" type="n" name="Residue_in_RepeatUnit" />
        </RESIDUES>
        <LINKAGES>
          <L Number="1" Linkage="1n(7-7)2n" />
        </LINKAGES>
      </SUBGRAPH>
    </INTERNAL>
  </REPEAT>
</BlackSugarXML>

```

Figure 20: A more complex example of GlycoCT and XML{GlycoCT}, encoding a repeat unit.