Exposition of the N-glycan nomenclature system “proglycan”

Drafted by: Friedrich Altmann, Johannes Helm, Johannes Stadlmann
University of Natural Resources and Life Sciences, Vienna, Austria
Contact: friedrich.altmann@boku.ac.at

Teaser

In the beginning was the word …. but there were no words for N-glycans. Look at the entry in a relevant database for the structure that we will herein baptize A^4A^4F^6 and ask yourself, how you could label your Eppendorf vial with the therein suggested options of which good, old IUPAC code is probably still the most human-friendly one - whereby this an already simplified version:

Gal(b1-4)GlcNAc(b1-2)Man(a1-3)[Gal(b1-4)GlcNAc(b1-2)Man(a1-6)]Man(b1-4)GlcNAc(b1-4)[Fuc(a1-6)] GlcNAc

Another example using a recently published drawing tool [1], the herein used condensed IUPAC code and the respective “proglycan” cartoon and acronym.

N-glycans are just one group in the huge universe of carbohydrates, but in medical biotechnology and bio-pharmaceutics, they supremely reign our attention hierarchy. So, we assume that a comprehensive and logical naming system could be useful and therefore we apply Occam’s razor to N-glycan nomenclature.

Content

The beginning
Galactosylation (including the alfa-Gal epitope)
Gloss about structure cartoons
Core-Fucosylation
Fucose on antennae
Bisecting GlcNAc
Sialic acids
Multi-antennary glycans
LacNAc-repeats
High-mannose glycans
Plant and insect structures
Specialties
Comparison with existing systems
Prolog

The number of already existing abbreviation systems for N-glycans cannot rival that of sand grains on sea shores. However, it seems to us that none of the various modes of the currently used naming systems is capable of representing (almost) all N-glycan structures occurring in mammals, insects and plants with a decent number of characters in an unambiguous manner. The big exceptions are the machine codes such as GlycoCT [2], which are by no means apprehensible for the human eye unless somehow translated to visibility [1]. So, quite often researchers give up and just show structure cartoons. These, however, are unsuitable for oral or written transmission or for labeling of vials.

The herein introduced system has proven useful for communication with partners for already many years. The terms MMXF$^3$ and MUXF$^3$ (with or without the superscript number) enjoy widespread use in the allergy diagnosis community. Our recent experience with 40 isomeric N-glycans all composed of 5 hexoses, 4 HexNAcs and 1 fucose [3,4] reinforced our conviction that the “proglycan” system is highly useful. Therefore, we shall not surrender in our fight against the inertia exerted by beaten tracks.

For frequently occurring “default” structures, various naming systems are in use. However, in our eyes, none of these is apt to unambiguously describe a large segment of all possible structures in a systematic, easily understandable manner. A collection of such systems will be shown in the last chapter of the present treatise.

The beginning

<table>
<thead>
<tr>
<th>Proglycan Code</th>
<th>IUPAC Code</th>
<th>CFG Code (Vienna style)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GnGn</td>
<td>GlcNAcβ−2Manα, GlcNAcβ−2Manβ−4GlcNAcβ−4GlcNAc</td>
<td>![Diagram of GnGn]</td>
</tr>
<tr>
<td>MGn</td>
<td>Manβ−4GlcNAcβ−4GlcNAc, GlcNAcβ−2Manα</td>
<td>![Diagram of MGn]</td>
</tr>
<tr>
<td>GnM</td>
<td>Manβ−4GlcNAcβ−4GlcNAc, GlcNAcβ−2Manα</td>
<td>![Diagram of GnM]</td>
</tr>
<tr>
<td>MM</td>
<td>Manβ−4GlcNAcβ−4GlcNAc</td>
<td>![Diagram of MM]</td>
</tr>
</tbody>
</table>

Concealing the origin of an idea is widespread practice but we shall ignore this habit. During a brief stay in our lab in Vienna around 1990, Harry Schachter from Toronto taught us the term “GnGn” for the acceptor substrate of fucosyl transferases [5,6]. Gn stands for GlcNAc and the two Gn-s symbolize
the two terminal residues of the biantennary N-glycan. The rest is unambiguously clear. No further definitions are required.

This substrate often fell prey to hexosaminidases leaving two isomers with just one GlcNAc, which can easily be named according to now terminal mannose residue(s) represented by an “M”. If two GlcNAc are removed, we end up with “MM”. However, if only one mannose gets exposed, we need a rule about the sequence of reading.

**RULE 1: Terminal residues are read from top to bottom = counterclockwise**
The first term describes the 6-arm antenna just annotating the non-reducing terminal sugar. The second term describes the 3-arm.

**RULE 2: Monosaccharides are depicted by one capital letter. Modifications are specified by a subsequent small letter.**

G ... glucose  
Gn ... N-acetylglucosamine  
M ... mannose  
Na ... N-acetylneuraminic acid  
A ... galactose  
Ng ... N-glycolyneuraminic acid  
F ... Fucose  
X ... Xylose

**Galactosylation (including the alfa-Gal epitope)**

Often, the antennae are elongated by galactose whose one-letter code [7] is A. Together with above mentioned three acceptors this leads to a total of 4 products if we only consider β1,4-linkages. Rarely, however, the linkage may be β1,3. To account for this ambiguity, we use superscripts.

We now know the three basic terminal elements M, Gn and A (either A³ or A⁴) of an antenna. It is clear by definition that “Gn” and “A⁴” annotate the sequences GlcNAc(β1-2)Manα1- and “A⁴” Gal(β1-4)GlcNAc(β1-2)Man(α1-, respectively. Whenever we further elongate the sequence, we need to include the terminal Gal residue. The first example for this necessity is the Galα1,3-Galβ1- element. We might resort to something like Aα1-3A⁴ or Aα1-3A³. The suggested annotation, however, is A³-⁴ or A³-³ [3], whereby the two figures point at the two residues involved. Unless – very unexpectedly – a novel structural element turns up that could be described by this combination, the terms are unambiguous.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GnA⁴</td>
<td></td>
<td>A⁴A⁴</td>
</tr>
<tr>
<td>A³Gn</td>
<td></td>
<td>A⁴A³</td>
</tr>
<tr>
<td>MA⁴</td>
<td></td>
<td>A³A⁴</td>
</tr>
<tr>
<td>GnA³-⁴</td>
<td></td>
<td>A³-⁴A⁴</td>
</tr>
</tbody>
</table>
Gloss about structure cartoons

While essentially using the cartoon format of the Consortium of Functional Glycomics (CFG), we mix in a bit of the “Oxford glycobiology” system [8], in which anomericity and linkage position are described solely by the connecting lines. In order to maintain the familiar overall appearance of structures, the 1-2 linkage of antennary GlcNAcs is shown as a bended line.

Strict adherence to the linkage angle rule allows omission of the linkage specifications by characters and this again allows to unambiguous depiction of structures even at very small size.

Core-Fucosylation

The core-fucose constitutes a third terminal residue and hence we introduce a third structure term, “F” for fucose. We could simply write e.g. A^4A^4F. In mammals, the core fucose is strictly always in α1,6. If you only work with mammalian samples you may content yourself with this simplification. However, as insect cells and plants have some relevance - certainly in biotechnology – we must consider, that here the fucose can sit or sits in the α1,3-position. Therefore, for the sake of clarity, the superscripts should be used to define the type of core-fucose.
Fucose on antennae

Lewis fucoses introduce branching of the antenna. IUPAC nomenclature uses square brackets to identify a branch. We do something similar. However, the two residues, or chains, that are linked to the root of the branch are both put in round brackets. The substitution points are defined by superscripts. So, LeX fucosylation of an A^4 antenna turns this term into (A^4F^3). This A LeA structure would be (F^4A^3) because we read the structure counter-clockwise. Nothing wrong. However, we want to be elegant and as concise as possible. Are we losing any information when omitting the superscripts? Obviously, not.

Advanced coding:
The terms (AF) and (FA) perfectly describe the terminal structures. In the near future, we may substitute the terms by "macros", i.e. Lx for the Lewis X terminus and La for the Lewis A determinant.

Another difficulty is posed by the blood group H α1,2-fucose, which is linked to galactose, which in turn can be linked β1,3- or β1,4 to GlcNAc. So just putting "F" as the terminal sugar would leave uncertainty. Therefore – using linear code [7] – we write F^2-A^4. We – again - can save one character by omitting the A to arrive at: F^2- or F^2-3. Why not just F^4 or F^3? Because, we must not ignore the galactose. "F" would be a Fuc(α1-4)GlcNAc sub-structure, which does not exist.

RULE 3: If more than one terminal residue occurs on one antenna, these residues are put in brackets. Round brackets are used for all branching except that arising from GlcNAc-transferase IV and V (see below “Multiantennary glycans”.

RULE 4: Substituents to the β-galactose are linked to this residue by a hyphen. The terms -A^4 or -A^3 are abbreviated to the superscripts^4 or ^3.

<table>
<thead>
<tr>
<th>(FA)Gn</th>
<th>A^4(AF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LaGn</td>
<td>A^4Lx</td>
</tr>
<tr>
<td>A^3(FA)F^6</td>
<td>A^4F^2-3</td>
</tr>
<tr>
<td>A^3LxF^6</td>
<td>A^4Lh^3</td>
</tr>
</tbody>
</table>

Bisecting GlcNAc

Bisected N-glycans are a peculiar feature of IgG and especially brain glycans. To annotate this residue, we almost complete our circumnavigation of the N-glycan. Hence, the term "bi" is found at the very end of the abbreviation.

<table>
<thead>
<tr>
<th>GnGnbi</th>
<th>A^4GnF^6bi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(AGnFbi)</td>
</tr>
</tbody>
</table>
**RULE 5: Bisected GlcNAc is indicated by “bi” and is always listed as the last extension term**

**Sialic acids**

Sialic acids are (usually) not linked directly to GlcNAc and therefore the same approach as for alfa-Gal and blood-group H fucose is chosen. Thus, Neu5Ac(α2-6)Gal(β1-4)GlcNAc(β1,2)Man(α1- shrinks to Na6⁴.

Na⁴³, Na⁶⁴ and the probably not existing Na⁶⁻³ denote the other options for a sialylated antenna. *N*-glycolylneuraminic acid – extremely rare in humans, common to most animals – is abbreviated as “Ng”.

The **green code** is a suggested alleviation for cases where exact linkages are not known, or do not matter.

<table>
<thead>
<tr>
<th>A⁴Na⁶⁻⁴</th>
<th>Na⁶⁴Na⁶⁻⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>Na³⁺Ng³⁻F⁶</td>
<td>Na⁶⁻⁴Na⁶⁻⁴F⁶</td>
</tr>
<tr>
<td>&lt; NaNaF &gt;</td>
<td>Na⁶⁻⁴Na⁶⁻⁴</td>
</tr>
</tbody>
</table>

**Multi-antennary glycans**

A branch leading to three antennae can occur on either arm of an N-glycan. The two antennae ascending from the same mannose are set in square brackets. The square exclusively and immediately tells us that this branch is further branched. By that the two basic types of triantennary glycans are readily told apart.

The proglycan nomenclature reaches its limits here because the terms become lengthy and difficult to read, but then, what is the alternative? Even the strange tetra-sialylated triantennary glycan in bovine fetuin can be depicted. Note that the term [(Na³⁻³Na⁶)Na⁶⁻⁴] contains – inside the square brackets - round brackets signifying additional branching. As no other branching point is given, the root residue is GlcNAc.
**RULE 6:** Two mannose-rooted antennae are put in square brackets. The order within brackets follows the "counter-clockwise" **RULE 1**.

### LacNAC repeats

The primary LacNAC disaccharide Galβ1-4GlcNAcβ1- that is linked to a mannose can be further elongated by the addition of GlcNAc (in β1-3 linkage to Gal), which usually is followed by the quick addition of Gal to arrive at another Galβ1-4GlcNAcβ1-, or LacNAc unit.

\[
\begin{align*}
\text{Galβ-4GlcNAcβ-2Manα-} & \quad \text{A}^4 \\
\text{GlcNAcβ-3Galβ-4GlcNAcβ-2Manα-} & \quad \text{Gn}^{3-4} \\
\text{Galβ-4GlcNAcβ-3Galβ-4GlcNAcβ-2Manα-} & \quad \text{Ln}^4 \\
\text{Galβ-4GlcNAcβ-3Galβ-4GlcNAcβ-2Manα-} & \quad \text{Ln-Ln}^4
\end{align*}
\]

The large and complex structures **N3.7.2B** in recombinant human erythropoietin [9], _aka_ EPO will thus be written as:

**High-mannose and hybrid type N-glycans**

The above defined rules can also be applied to high-mannose, _aka_ oligomannosidic type glycans, but this only makes sense if the true structure of a glycan is known, e.g., if glycans are analyzed by porous graphitic carbon chromatography [10]. Only the mannoses that occur in addition to those on the common core have to be explicitly defined and this is realized by their linkage. In the upper line in the example below the number of linkage figures always is the number of mannose residues in addition to the three of the core.
<table>
<thead>
<tr>
<th>MM</th>
<th>[\text{MM}]</th>
<th>MU</th>
<th>[\text{MU}]</th>
</tr>
</thead>
<tbody>
<tr>
<td>M^3M</td>
<td>\text{[M^6M^3]M^2} \text{aka Man6}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M^6M^3)M</td>
<td>MM^{2-2-2} \text{aka another Man5}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M^{2-6}M^3)M^2</td>
<td>(M^6M^3)M^{2-2}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M^{2-6}M^{2-3})G</td>
<td>(M^6M^3)G-G-G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M^6M^3)Gn Man5Gn</td>
<td>M^3A^4 Man4A*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* In Man4A neither the linkage of the terminal mannose nor the galactose are exactly determined. In some instances, this term may nevertheless be justified.

Plant and insect glycans
This arises from the authors’ long occupation with glycoproteins from plants and insects, where core fucose occurs in α1,3-linkage and where – in plants – a xylose is linked to the β-mannosyl residue of the core. These two peculiarities can easily be included as shown in the examples below.

<table>
<thead>
<tr>
<th>MMXF^3</th>
<th>[\text{MMXF}]</th>
<th>MUXF^3</th>
<th>[\text{MUXF}]</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMXF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGFnX</td>
<td></td>
<td>(FA)(FA)XF^3 LaLaXF^3</td>
<td></td>
</tr>
<tr>
<td>MMF^3F^6</td>
<td></td>
<td>M(AnF)F^3F^6</td>
<td></td>
</tr>
</tbody>
</table>

Specialties
The human brain contains sizable amounts of glycans with the “HNK-1” (from human natural killer cells) with sulfated glucuronic acid [11]. Annotating a structure like this requires some form of linear code and the addition of abbreviations for non-sugar substituents, in this case sulfate. Note that the hyphen binds the “su” to “Ga”, which in turn is hyphenated to the “4” (or “3”), which stands for the regular antennary galactose.
The bladder protein uromodulin aka Tamm-Horsfall protein contains glycans with sulfated GalNAc and the Sd\textsuperscript{a} determinant, which harbors a branch on the galactose residue [12]. Based on the rules for Lewis determinants, we use a round bracket and a superscript hyphen with the linkage of the galactose residue.

Another peculiar structure is that with a Lewis X determinant in the bisecting position. With the rules established so-far, even such an exotic item can be named.

To facilitate deciphering of these terms, the abbreviations are also given with colors for the 6-arm, 3-arm and the extension terms.

Mosses contain structures with methyl groups [13], non-vertebrates contain numerous “unusual” and remarkable structural features such as methylation, sulfation, and zwitterionic non-sugar substituents, again glucuronic acid and often unusual architectures such as substituted core-fucose just as an example [14,15].

Another box of plethora (sic!) is opened by the highly unusual and diverse N-glycans of microalgae [16-18]. It would be possible to somehow describe also these structures, but for the time being, it does not appear to be a pressing need.

**Comparison with existing abbreviation systems**

Finally, and hopefully not too late, we shall compare the herein portrayed naming system with existing ones. There are several more or less frequently used systems of which a few occur more often shall be presented and compared.

The most stringent and rigorous way to annotated mass spec data is to give the sum formula without even trying to interpret it in terms of types of hexoses and co

Example for \( m/z = 2369.84 \) = Hex5HexNAc4Neu5Ac2dHex1

\[ \text{Hex5HexNAc4Neu5Ac2dHex1} = H5N4S2F1 \text{ as e.g., in [19,20]} \]

or more condensed

\[ = 5_4_2_1 \text{ or 5421 as e.g., in [21] and [22], respectively} \]

or as in web.expasy.org/glycomod/

\[ = (\text{Hex})_2 (\text{HexNAc})_2 (\text{Deoxyhexose})_1 (\text{NeuAc})_2 + (\text{Man})_3 (\text{GlcNAc})_2 \]

This annotation type strictly avoids any overinterpretation of data and should be the explicit starting point of any other naming or drawing exercise. Unfortunately, it neither transports structural information nor does it spoil the human eye. MS spectra evaluation software thus often
essentially neglects the overinterpretation problem and suggests a particular structure, where actually a range of isomers is possible as recently emphasized for the H5N4F1 composition [3].

<table>
<thead>
<tr>
<th>Proglycan</th>
<th>Antibody</th>
<th>Oxford Fuc First</th>
<th>Oxford Fuc Last</th>
<th>Elaborate Oxford</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref.</td>
<td>Number of antennae</td>
<td>Type of sialic acid</td>
<td>Galactose linkage</td>
<td>Fucose linkage</td>
</tr>
<tr>
<td><img src="image1" alt="Diagram" /></td>
<td>Na(^{6-4}Na^{6-4}F) 6</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>G2FS2</td>
<td>[25], a, b</td>
<td>(√)*</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>G2FS2 α(2,6)</td>
<td>[25], c</td>
<td>(√)*</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>FA2G2S2</td>
<td>[26], a</td>
<td>√</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>F(6)A2G2S(6)2</td>
<td>[27]</td>
<td>√</td>
<td>√</td>
<td>(x)</td>
</tr>
<tr>
<td>A2S2F</td>
<td>c</td>
<td>√</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><img src="image2" alt="Diagram" /></td>
<td>A2S2G2F 6a)</td>
<td>[23], d</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>A2G2FS2 6a)</td>
<td>[28]</td>
<td>√</td>
<td>x</td>
<td>(√)</td>
</tr>
<tr>
<td><img src="image3" alt="Diagram" /></td>
<td>A(^{4}G)nF(^{6}) 6</td>
<td>√</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

8a) Examples for the absence of a stringent logic for order of symbols

Commercial sources:
- a) ludger.com/product-catalogue/standards-controls
- b) asparagi.com/product-category/glycan-standard/
- c) agilent.com/en/product/biopharma-hplc-analysis/glycan-analysis
- d) Thermo Fisher Scientific BioPharma Finder 3.0

For the expression of ideas about structure, two systems are prominent:
1) The “antibody glycan code” counting the number of galactose residues for biantennary glycans only – sufficient for antibodies.
2a) The “Oxford code” counting number of antennae and galactose residues with fucose first
2b) The “Oxford code” counting number of antennae and galactose residues with fucose later
3a/b) The elaborate “Oxford code”. Somewhat event-related additions are used to exactly specify structures, e.g., when α-Gal or N-glycolyl-neuraminic acid occur [23] or a large number of isomers are to be named [24].

The “antibody glycan code” is a perfectly fine convention in the field of – surprise - antibodies, in particular recombinant IgGs, where G2F actually stands for the isomer A₄A₄F₆ and not for any of the 40-50 possible other isobars [3].

The “Oxford code” and its variants likewise have a definite raison d’être especially when rather large and only partially defined structures shall be named. E.g., the term 3A2SF comprises a number of related structures that are most often not told apart by the analytical results [30,31].

When particular, exactly known N-glycan structures are to be described, the elaborate “Oxford code” is an option, but the proglycan system appears to be a big step ahead. The following table tries to give an overview of abbreviation systems. By no means does it claim to be comprehensive. The readers contribution to update and complete this table is highly encouraged.

Finally, we dare to devise a decision triangle that opposes the Oxford and proglycan systems with plain composition as the starting point and save haven.
List of possible “structure terms”:

<table>
<thead>
<tr>
<th>6-arm term</th>
<th>3 arm term</th>
<th>core extension</th>
</tr>
</thead>
</table>

**in both first (6-arm) and second (3-arm) position:**

- **Gn** a GlcNAc linked to the invariant pentasaccharide core
- **U** indicates no substituent of the β-mannose.
- **A** Galactose linked to GlcNAc, linkage not specified
- **A⁴ / A⁴** Galactose β1,4-linked to GlcNAc
- **A³ / A³** Galactose β1,3-linked to GlcNAc
- **An⁴ / An⁴** GalNAc β1,4-linked to GlcNAc
- **(AF) / (A⁴F³) / (A⁴F³) / Lx** Lewis X determinant
- **(FA) / (F⁴A³) / (F⁴A³) / La** Lewis A determinant
- **F² / Lh³** Blood group H determinant on type I chain
- **F² / Lh⁴** Blood group H determinant on type II chain
- **La** Lewis A determinant
- **Lh** Blood group H determinant, probably Lh³, maybe Lh⁴
- **Ln** LacNAc extension (Ln⁻¹, Ln³, Ln⁴)
- **Lx** Lewis X determinant
- **Na** Sialic acid linked to Gal-GlcNAc, linkages not specified
- **Na⁶-⁴ / Na⁶-⁴** Neu5Ac-α2,6-Gal-1,4-GlcNAc-
- **Na³-⁴ / Na³-⁴** Neu5Ac-α2,3-Gal-1,3-GlcNAc-
- **(Na⁶Na³)⁴** Neu5Ac-α2,6(Neu5Ac-α2,3-Gal-1,3)-GlcNAc- (as in fetuin)

**Possible in first (6-arm) position:**

- **M** 6-arm mannose of the invariant pentasaccharide core
- **M³ / M³** a mannose in 3-linkage to 6-arm mannose of the core
- **M⁶ / M⁶** a mannose in 6-linkage to 6-arm mannose of the core; rare
- **[GnGn]** GlcNAc-β1,6-(GlcNAc-β1,2-)Man
- **[A⁴A⁴]** branched 6-arm with two Gal residues (in analogy to [GnGn])

**Possible in second (3-arm) position:**

- **M** 3-arm mannose of the invariant pentasaccharide core
- **M² / M²** a mannose in 2-linkage to the invariant pentasaccharide core
- **M²-² / M²-²** two mannoses in series
- **[GnGn]** GlcNAc-β1,4-(GlcNAc-β1,2-)Man (in analogy to [GnGn])
- **[A⁴A⁴]** branched 3-arm with two Gal residues (in analogy to [GnGn])

**Possible in the extension positions:** – if present; fucosylation before bisection

- **F⁶ / F⁶** α1,6-fucosylation of reducing end GlcNAc
- **F³ / F³** α1,3-fucosylation of reducing end GlcNAc
- **F³F⁶ / F³F⁶** difucosylation of reducing end GlcNAc
- **X** β1,2-xylsosylation of the β-mannose (written before fucosylation)
- **bi** indicates presence of a bisecting GlcNAc
A-bi  bisecting LacNAc
(AF)-bi bisecting Lewis X

Order of extension terms:


Non-sugar substituents

su  sulfate
ac  acetyl
me  methyl
po  phosphate
pc  phosphocholine
pe  phosphoethanolamine

References:


