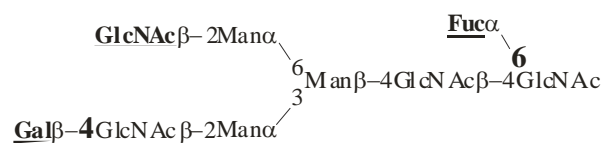


What's your name, sugar ?

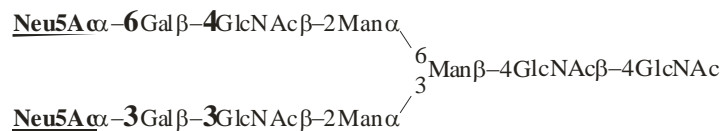
A simple abbreviation system for complex N-glycan structures

As the topic of protein glycosylation, especially N-glycosylation, is now beginning to leave the specialists' laboratories, the need for a readable and understandable text-format annotation arises. Based on the conserved basic architecture of N-glycans, the "proglycan" system indicates only the sugar residues at the non-reducing termini on the oligosaccharide. With the following rules, simple structures get very simple names and even quite complicated N-glycans can be described with terms rarely longer than the word oligosaccharide. Only the terminal residues are given beginning with that on the "upper", 6-linked antenna, then the 3-linked antenna and then substituents to the core. Examples of N-glycans which occur on human IgG and bovine fibrin, respectively, are shown below. More examples and details can be found on www.proglycan.com.

GnA⁴F⁶
(sloppy: GnAF)



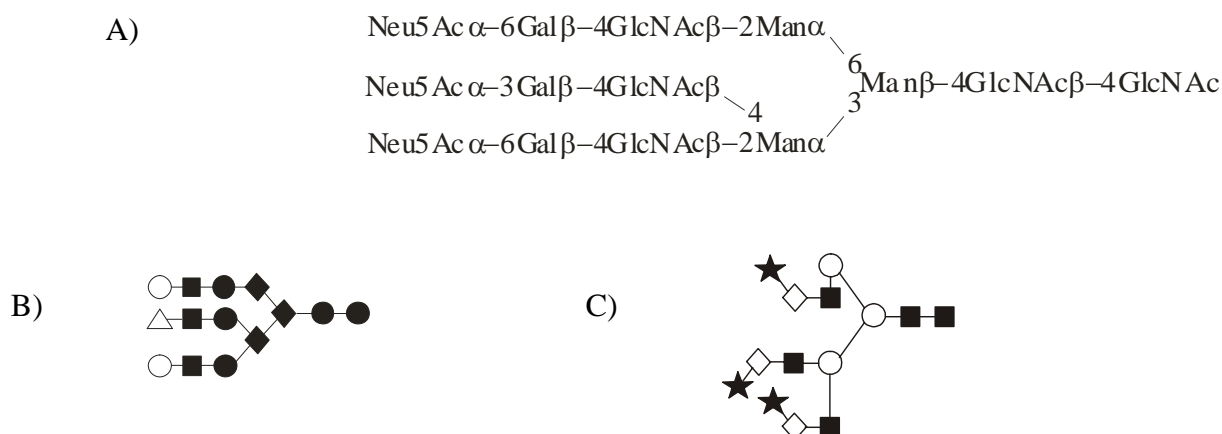
Na⁶⁻⁴Na³⁻³
(lazy: Na6-4Na3-3)



Introduction

How shall cell biologists, immunologists, hematologists, biotechnologists and other “normal” scientists ever deal with protein glycosylation if even glycobiochemists themselves cannot find practicable names for the compounds they are working with? We are either dealing with monsters of chemical formulas such as α -Neup5Ac-(2→6)- β -D-Galp-(1→4)- β -D-GlcpNAc-(1→6)-[α -Neup5Ac-(2→3)- β -D-Galp-(1→4)-D-GlcpNAc-(β 1→2)-] α -D-Manp-(1→6)-[α -Neup5Ac-(2→6)- β -D-Galp-(1→4)- β -D-GlcpNAc-(1→2)- α -D-Manp-(1→3)-] β -D-Manp-(1→4)- β -D-GlcpNAc-(1→4)- β -D-GlcpNAc-(1→4)-Asn, or we are making use of the generous offer provided by IUPAC [1] to write the same glycan this way: Neu5Ac α -6Gal β -4GlcNAc β -6(Neu5Ac α -3Gal β -4GlcNAc β -2)Man α -6(Neu5Ac α -6Gal β -4GlcNAc β -2Man α -3)Man β -4GlcNAc β -4GlcNAc; or we use a graphic program as for Fig. 1.

Fig. 1: Graphical depictions of a triantennary N-glycan. Most condensed format allowed by IUPAC (A); symbolic depiction used by Kole *et al.* [2] (B); and by Butler *et al.* [3] (C).



Graphical depictions with symbols certainly help perceiving the overall architecture of a glycan. In the “Oxford” system the cartoons even transport all the linkage information in a consistent, easily decipherable way [3]. Nevertheless, it requires a graphic program and one soon encounters “steric” hindrances a e.g. in the example structure shown above. In other words, cartoons and graphics are time consuming and space wasting. Many abbreviations have been used but we feel that this systems lacked the ability of describing N-glycans precisely and still in a compact form. The term GlcNAc₂Man₃GlcNAc₃Gal₃Neu5Ac₃ - application of another widely used way to fight the naming problem - fails to provide any information about linkages of the galactoses or sialic acids. Here we will learn to fully describe the above glycan with the term $\text{Na}^{6-4}[\text{Na}^{3-4}\text{Na}^{6-4}]$.

ambiguous as there is also β 1,3-linked galactose (in mammals; “type I chain”) and α 1,3-linked fucose *e.g.* in insect or plants. For an exact representation of the structure, superscripts are used and the above glycan has the correct name **GnA⁴F⁶** (or in this tutorial: **Gn·A⁴·F⁶**).

One advantage of this annotation is its inherent ability to identify branch isomers. The isomer of **GnA⁴F⁶** where the galactose is linked to the upper arm has the name **A⁴GnF⁶** (exact code) or **AGnF** (degenerate code).

Plant glycoproteins carry N-glycans with a xylose β 1,2-linked to the β -mannose of the core.

There is no other xylose linkage known and so we can write *e.g.* **GnGnXF³** (or tutorial: **Gn·Gn·XF³**).

Bisected, truncated and hybrid type N-glycans :

Rule 4: The suffix “bi” at the end of a string is used to denote the presence of a bisecting GlcNAc.

Rule 5: An **M** denotes a mannose residue on the pentasaccharide core. If even this mannose is missing, a **U** (unsubstituted) is written at this place.

Rule 6: A glycan containing more α -mannosyl residues on the 6-arm is a hybrid type glycan and is called either **Man4** or **Man5** followed by the terminal sugar on the 3-arm if not mannose.

Thus, **GnA⁴F⁶** with a bisecting GlcNAc is written as **GnA⁴F⁶bi** (or **Gn·A⁴·F⁶bi**). The number of core substituents is not limited. Anything in the third or higher position is defined as core substituent.

While the anchor points of the herein presented abbreviation system are the GlcNAc-5 and -5' residues which initiate the antennae, these may be sometimes missing and then one of the two α -mannose residues gets exposed. **MM** is the name for the core-pentasaccharide which is converted to **MGn** by the action of GlcNAc transferase I. GlcNAc transferase II will turn **MGn** into **GnGn**. In some glycans even a mannose, usually the 3-arm mannose, is absent. A “**U**” is used to denote that the 3-OH of the β -mannose is unsubstituted. Removal of the 3-arm mannose from **MM** leads to the tetrasaccharide **MU**.

Hybrid type glycans carry only one LacNAc antenna linked to the 3-arm of **Man_{3,5}GlcNAc₂**.

We suggest denoting such N-glycans as *e.g.* **Man4Gn** or **Man5A⁴**. **Man3Gn** is written as **MGn** (see above). We could even think of the case that from **Man4** the 3-arm mannose is removed which would give us **Man4U** but this might not be the most convincing argument in

favour of the herein presented system.

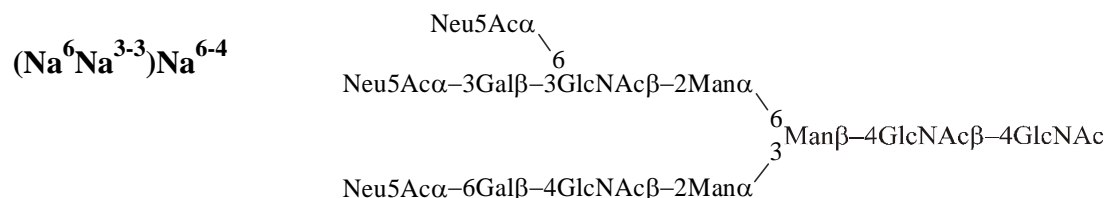
Sialic acids, straight chain B determinants and branched antennae:

Rule 7: If the terminal residue **R** is a substituent to a β -galactose, the linkage of this penultimate residue is mentioned in the form **R^{x-y}** where x gives the linkage of the terminal and y that of the galactose residue.

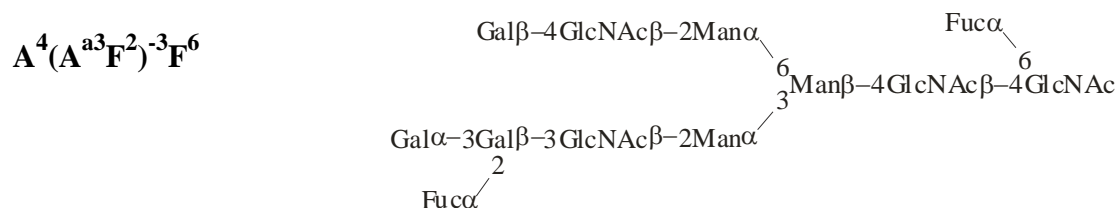
Rule 8: If an antenna branches, the two terminal residues on this branch are put in normal brackets.

Rule 9: If a residue occurs in an unusual anomeric form, this is indicated by a superscript a or b before the linkage figure. To our knowledge this only applies to galactose which is assumed to occur as the β anomer if not indicated otherwise.

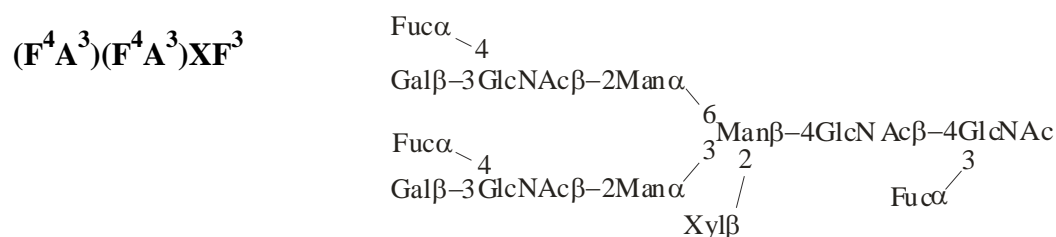
Mammalian N-glycans usually terminate their antennae with *N*-acetylneuraminic acid which can occur in α 2-3 or α 2-6 linkage and which itself is linked to galactose residues which themselves may occur in either β 1-3 or β 1-4 linkage. Thus a disaccharide is linked to the GlcNAc-5 or -5'. According to rule 7, the most common disialo N-glycan is **Na⁶⁻⁴Na⁶⁻⁴** and any of its isomers such as *e.g.* **Na³⁻³Na⁶⁻⁴** can easily be depicted by the proglycan system. One sialyl transferase transfers neuraminic to the 6-position of GlcNAc. This element forms branched antennae. The two branches to GlcNAc are put in normal brackets. The branch with the higher locant at the GlcNAc is written first. A third NeuNAc linked to the upper branch of **Na⁶⁻⁴Na⁶⁻⁴** therefore would lead to **(Na⁶Na⁶⁻⁴)Na⁶⁻⁴**. In this name the whole term **(Na⁶Na⁶⁻⁴)** stands for the 6-linked antenna, thus in tutorial style we write **(Na⁶Na⁶⁻⁴)·Na⁶⁻⁴**. With higher probability we may encounter the isomer having a type I chain:



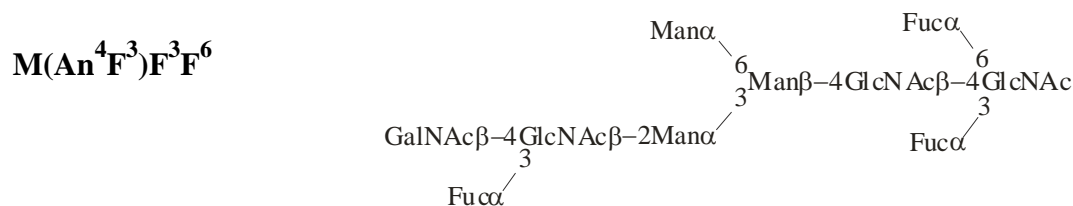
Branched antennae are part of the AB0 as well as the Lewis blood group antigens. We write **(A⁴F³)** for a Lewis X antenna and **(F⁴A³)** for a Lewis A antenna. The situation is more complex with AB0 blood groups as here the branching point is not GlcNAc but galactose and thus a glycan with a blood group B determinant on the 3 arm is written **A⁴(A^{a3}F²)⁻³F⁶** (or in tutorial style **A⁴·(A^{a3}F²)⁻³·F⁶**).



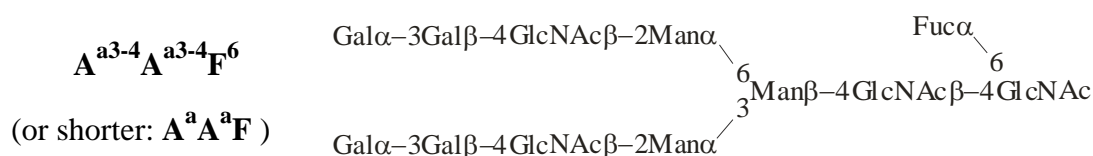
The largest N-glycan found in plants has two Lewis A determinants [4] and is written $(F^4A^3)(F^4A^3)XF^3$ (tutorial style: $(F^4A^3) \cdot (F^4A^3) \cdot XF^3$).



Honeybee glycoproteins may carry a “LacdiNAc” antenna leading to the structure $M(An^4F^3)F^3F^6$ (with separators: $M \cdot (An^4F^3) \cdot F^3F^6$).



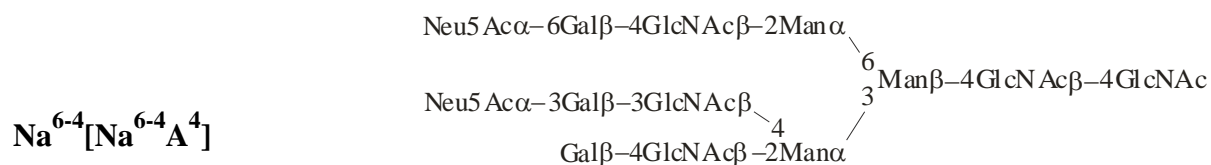
We are now also able to christen N-glycans carrying Galili (or straight-chain B) epitopes.

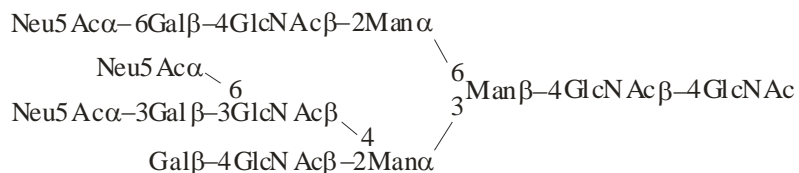
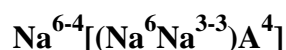
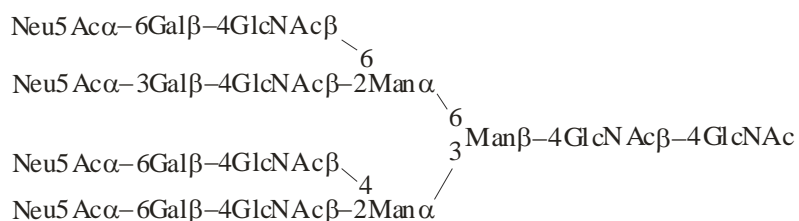


Tri- and tetraantennary structures:

Rule 10: If two antennae are linked to one core α -mannose their terminal residues are put into square brackets.

In the proglycan system it is easy to specify which type of triantennary structure is meant. The product of GlcNAc transferase IV is called **Gn[GnGn]**, that of GlcNAc transferase V **[GnGn]Gn**.





Extra long antennae, repetitive elements:

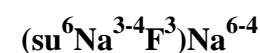
Rule 11: Chains that extend beyond a substituent of the β -galactose are written in almost “usual” style by connecting each residue and its aglycon with a hyphen.

Rule 12: Repetitions are indicated by “ xn ” in subscript font.

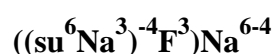
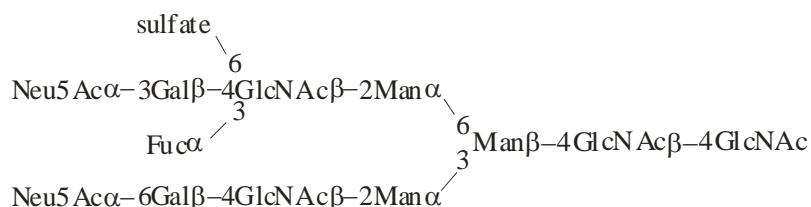
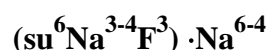
Both situations can *e.g.* be found in the case of polysialic acid chains. A glycan with five 8-linked sialic acids on the 6-arm is written as $(\text{Na}^8\text{-Na}^8\text{-Na}^8\text{-Na}^8\text{-Na}^8\text{-Na}^{6-4})\text{Na}^{6-4}$ or $(\text{Na}^8_{x5}\text{-Na}^{6-4})\text{Na}^{6-4}$.

Complicated examples:

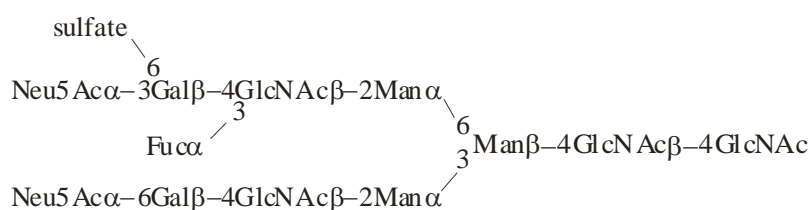
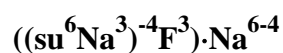
We finally will again deal with Lewis type structures and thereby we go to the extreme and try to annotate 6-sulfo-sialyl-Lewis X and 6'-sulfo-sialyl-Lewis X structures. The rules are that – as before – the substituents are sorted according to the carbon they are attached to, the higher number coming first and that the aglycon must be identified when it is not the mannose-linked GlcNAc-residue.



tutorial style.



tutorial style:



Style:

The superscript position of the linkage specifiers is not crucial for the meaning of the abbreviations but for the readability. Loss of formatting in the abbreviation $\text{Na}^{6-4}[(\text{Na}^6\text{Na}^{3-3})\text{Na}^{6-4}]$ turns it into $\text{Na}6-4[(\text{Na}6\text{Na}3-3)\text{Na}6-4]$. No information is lost thereby. Possibly, some readers may prefer this style as it is still easier to write.

Uncertainties: (Warning: This chapter may be confusing)

Supplementary rule s1: A subscript “iso” after the place for the second antenna means that we cannot specify which antenna is linked to which arm. The longer antenna or the branched one is mentioned first.

Supplementary rule s2: If the subscript “iso” follows a round bracket we cannot assign this structural element to one particular antenna.

Supplementary rule s3: If the linkage specifier is followed by “iso” then this linkage pattern cannot be confined to one particular antenna. The groups are sorted by descending linkage figures.

Supplementary rule s4: Ignorance about a particular linkage is expressed by a question mark. In human IgG we find monosialylated glycans with fucose but maybe we cannot or do not want to discriminate the two glycans $\text{Na}^{6-4}\text{A}^4\text{F}^6$ or $\text{A}^4\text{Na}^{6-4}\text{F}^6$. In such a case we indicate the existence of isomers by writing $\text{Na}^{6-4}\text{A}^4_{\text{iso}}\text{F}^6$. (or in tutorial style: $\text{Na}^{6-4}\cdot\text{A}^4_{\text{iso}}\cdot\text{F}^6$). Similarly we may speak about $[\text{Na}^{6-4}\text{Na}^{6-4}]_{\text{iso}}\text{Na}^{6-4}$ if we do not know the branch structure.

$[(\text{Na}^{6-4}\text{F}^3)\text{Na}^{6-4}]_{\text{iso}}\text{Na}^{6-4}\text{F}^6$ is any triantennary, core-fucosylated, with one sialyl Lewis x determinant on either of the three antennae, a second fucose on the non-reducing side and three sialic acids. In $[(\text{Na}^{6-4}\text{F}^3)_{\text{iso}}\text{Na}^{6-4}]_{\text{iso}}\text{Na}^{6-4}\text{F}^6$ we know the basic branch structure but not where the Lewis antigen is bound. In contrast, in $[(\text{Na}^{6-4}\text{F}^3)\text{Na}^{6-4}]_{\text{iso}}\text{Na}^{6-4}\text{F}^6$ we know that the Lewis antigen is on one of the antennae of the 6-arm mannose.

If we want to describe a mixture of $\text{Na}^{6-4}\text{Na}^{3-4}$ and $\text{Na}^{3-4}\text{Na}^{6-4}$ we can write $\text{Na}^{6-4}\text{Na}^{3-4}_{\text{iso}}$ but in a triantennary structure this would become ambiguous. In such a case we write e.g. $\text{Na}^{6-4}[\text{Na}^{6-4}\text{Na}^{3-4}]_{\text{iso}}$ if at least the branch structure is known or $[\text{Na}^{6-4}\text{Na}^{6-4}]_{\text{iso}}\text{Na}^{3-4}$ if everything is open.

This is not yet a comprehensive treatise of all possible uncertainties and their annotation. The brave reader may by now feel some dizziness and hence we propose to postpone any further elaborations of these subtleties to future times when hopefully the basic concept of this abbreviation system will be widely used.

All too often we will lack information about some glycosidic linkages and then we write *e.g.* $\text{Na}^{6-?}\text{Na}^{3-?}$ or $\text{Na}^{?-?}\text{Na}^{?-?}$. Such cases are, however, better addressed by a simplified version of the proglycan system, the degenerate code.

Degenerate code:

Supplementary rule s5: The degenerate code does not use superscripts to describe linkages.

Supplementary rule s6: The degenerate code only depicts the terminal residue of each branch.

Supplementary rule s7: The degenerate code does not fix each and every structural detail and the abbreviations shall be understood as comprising all possible interpretations.

In many instances we will not know every structural detail about an N-glycan or we will deal with mixtures of isomers. Consider *e.g.* a peak of mass 2246 in a MALDI mass spectrum telling us that we have a diantennary N-glycan with 2 sialic acids but we do not know in which linkage they occur. We have already learned to call this group of possible structures $\text{Na}^{?-?}\text{Na}^{?-?}$. We suggest to simplify this term to **NaNa**. **NaNaF** then carries a core α 1,6-fucose. **AA** is the desialylated form and **GnGn** has even lost the galactose residues. Besides, the term **GnGn** is normal code anyway. In the case of triantennary N-glycans we have the choice of writing **AAA** (“triple A”) – if we do not know the branch structure – or **[AA]A** or **A[AA]** (see rule 10). A tetraantennary structure is always **[AA][AA]** or **[NaNa][NaNa]Fbi** but never **AAAA**.

Use of the degenerate code greatly simplifies the abbreviations and thereby yields terms more suitable for oral communication, lab protocols or vial labelling. In papers, however, it must be very clear what you are talking about. But when it is sufficiently defined that you deal with *e.g.* plant N-glycans, you can forget the $(\text{F}^4\text{A}^3)(\text{F}^4\text{A}^3)\text{XF}^3$ and ease yourself to a **(FA)(FA)XF**.

Conclusions:

The “proglycan” system turns the annotation of N-glycan structures into a simple task which no longer requires rectangles, squares, legends and graphic programs. Even rather complicated structures can be given unambiguously whereas simple structures likewise get simple names. By not mentioning the conserved interior of an N-glycan we arrive at terms simpler than those provided by the recently introduced linear code [4] which, however, is applicable to all kinds of oligosaccharides. Maybe this suggestion contributes a little bit to relieving at least one (large) area of glycobiology from its wall flower fate.

References:

- [1] IUPAC –IUBMB, Nomenclature of Carbohydrates (Recommendations 1996) 2-Carb-38
- [2] Koles, K., Patrick H. C. van Berkel, P.H.C., Frank R. Pieper, F.R., Jan H. Nuijens, J.H., Maurice L.M., Mannesse, M.L.M., Johannes F.G. Vliegthart, J.F.G. and Johannes P. Kamerling, J.P. (2004) *Glycobiology* **14**, 51-64
- [3] Butler, M., Quelhas, D., Critchley, A.J., Carchon, H., Hebestreit, H.F., Hibbert, R.G., Vilarinho, L., Teles, E., Matthijs, G., Schollen, E., Argibay, P., Harvey, D.J., Dwek, R.A., Jaeken, J. and Rudd, P.M. (2003) *Glycobiology* **13**, 601-622
- [4] Wilson, I.B.H., Zeleny, R., Kolarich, D., Staudacher, E., Stroop, C.J.M., Kamerling, J.P. and Altmann, F. (2001) *Glycobiology* **11**, 261-274
- [2] Ehud, B., Neuberger, Y., Altshuler, Y., Halevi, A., Inbar, O., Dotan, N. and Dukler, A. (2002) *Trends Glycosci. Glycotechnol.* **14**, 127-137

Acknowledgment: This manuscript is dedicated to Harry Schachter, the pioneer of N-glycan branching, who coined the term “GnGn” for the essential intermediate of complex-type N-glycan biosynthesis. I hope he likes what we made of it.