

2018

# Biophysical Approaches to Solve the Structures of the Complex Glycan Shield of Chloroviruses

Cristina De Castro

*University of Napoli*, [decastro@unina.it](mailto:decastro@unina.it)

Garry Duncan

*Nebraska Wesleyan University*, [gduncan@nebrwesleyan.edu](mailto:gduncan@nebrwesleyan.edu)

Domenico Garozzo

*Consiglio Nazionale delle Ricerche–Istituto di Chimica e Tecnologia dei Polimeri*

Antonio Molinaro

*University of Napoli*

Luisa Sturiale

*Consiglio Nazionale delle Ricerche–Istituto di Chimica e Tecnologia dei Polimeri*

*See next page for additional authors*

Follow this and additional works at: <https://digitalcommons.unl.edu/vanetten>



Part of the [Genetics and Genomics Commons](#), [Plant Pathology Commons](#), and the [Viruses Commons](#)

---

De Castro, Cristina; Duncan, Garry; Garozzo, Domenico; Molinaro, Antonio; Sturiale, Luisa; Tonetti, Michela; and Van Etten, James L., "Biophysical Approaches to Solve the Structures of the Complex Glycan Shield of Chloroviruses" (2018). *James Van Etten Publications*. 39.

<https://digitalcommons.unl.edu/vanetten/39>

This Article is brought to you for free and open access by the Plant Pathology Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in James Van Etten Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

---

**Authors**

Cristina De Castro, Garry Duncan, Domenico Garozzo, Antonio Molinaro, Luisa Sturiale, Michela Tonetti,  
and James L. Van Etten

# Biophysical Approaches to Solve the Structures of the Complex Glycan Shield of Chloroviruses

Cristina De Castro, Garry A. Duncan,  
Domenico Garozzo, Antonio Molinaro,  
Luisa Sturiale, Michela Tonetti, and  
James L. Van Etten

C. De Castro (*Corresponding author*), Department of Agricultural Sciences, University of Napoli, Portici, NA, Italy; email: decastro@unina.it

G. A. Duncan, Department of Biology, Nebraska Wesleyan University, Lincoln, NE, USA; email: gduncan@nebrwesleyan.edu

D. Garozzo & L. Sturiale, CNR, Institute for Polymers, Composites and Biomaterials, Catania, Italy; email: domenico.garozzo@cnr.it; luisella.sturiale@cnr.it

A. Molinaro, Department of Chemical Sciences, University of Napoli, Napoli, Italy; email: molinaro@unina.it

M. Tonetti, Department of Experimental Medicine and Center of Excellence for Biomedical Research, University of Genova, Genova, Italy; email: tonetti@unige.it

J. L. Van Etten, Department of Plant Pathology and Nebraska Center for Virology, University of Nebraska, Lincoln, NE, USA; email: jvanetten1@unl.edu

## Abstract

The capsid of *Paramecium bursaria* chlorella virus (PBCV-1) contains a heavily glycosylated major capsid protein, Vp54. The capsid protein contains four glycans, each N-linked to Asn. The glycan structures are unusual in many aspects: (1) they are attached by a  $\beta$ -glucose linkage, which is rare in nature; (2) they are highly branched and consist of 8–10 neutral monosaccharides; (3) all four glycoforms contain a dimethylated

---

Published (as Chapter 12) in Y. Yamaguchi, K. Kato (eds.), *Glycobiophysics*, Advances in Experimental Medicine and Biology 1104, pp 237-257.

doi 10.1007/978-981-13-2158-0\_12

Copyright © 2018 Springer Nature Singapore Pte Ltd. Used by permission.

rhamnose as the capping residue of the main chain, a hyper-branched fucose residue and two rhamnose residues “with opposite absolute configurations; (4) the four glycoforms differ by the nonstoichiometric presence of two monosaccharides, L-arabinose and D-mannose; (5) the N-glycans from all of the chloroviruses have a strictly conserved core structure; and (6) these glycans do not resemble any structures previously reported in the three domains of life.

The structures of these N-glycoforms remained elusive for years because initial attempts to solve their structures used tools developed for eukaryotic-like systems, which we now know are not suitable for this noncanonical glycosylation pattern. This chapter summarizes the methods used to solve the chlorovirus complex glycan structures with the hope that these methodologies can be used by scientists facing similar problems.

**Keywords:** Giant viruses, GC-MS, NMR, MALDI, N-glycosylation

## Abbreviation List

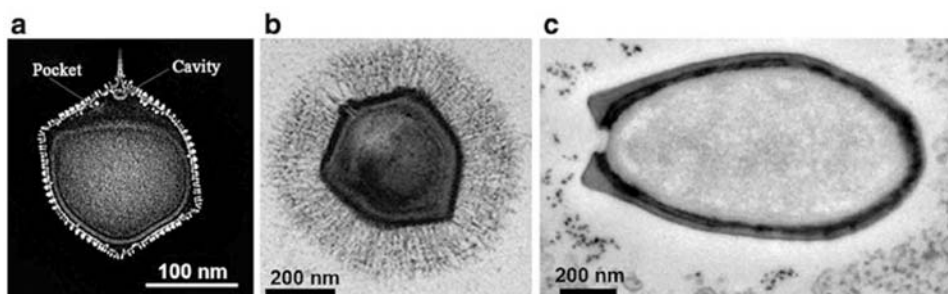
AA	acetylated alditol
AMG	acetylated methyl glycosides
AOG	acetylated octyl glycoside
COSY	correlation spectroscopy
GC-MS	gas chromatography-mass spectrometry
HMBC	heteronuclear multiple bond coherence
HSQC	heteronuclear single quantum coherence
NOESY	nuclear Overhauser effect spectroscopy
PBCV-1	<i>Paramecium bursaria</i> chlorella virus type 1
PMAA	partially methylated and acetylated alditol
ROESY	rotating frame Overhauser effect spectroscopy
TOCSY	total correlation spectroscopy

## 1 Introduction

In considering enzymes involved in manipulating carbohydrates, one usually does not think about viruses playing a role in this important activity; however, exceptions to this common belief are beginning to emerge from viruses often referred to as giant viruses. Giant viruses are a rapidly expanding group of viruses (their taxonomy is still being resolved) characterized by a large particle size and very large dsDNA genomes (typically from >300 kb to ~2.5 Mb) that encode many proteins involved in functions not normally found in typical viruses such as HIV or Ebola (Colson et al. 2013).

*Paramecium bursaria* chlorella virus 1 (PBCV-1, genus, *Chlorovirus*), the first member of the giant virus group, was discovered more than 35 years ago (Van Etten et al. 1982). PBCV-1 is an icosahedron with a diameter of 190 nm, and it has an ~331 kb genome that encodes ~416 proteins and 11 tRNAs (Dunigan et al. 2012). The virus infects the unicellular, eukaryotic, and symbiotic microalga *Chlorella variabilis* NC64A (**Fig. 1a**). Since then many new giant viruses have been discovered, with a burst of activity in the last 15 years (Abergel et al. 2015).

For example, *Mimivirus* was first described in 2004 (Raoult et al. 2004, **Fig. 1b**). It is more complex than PBCV-1 in terms of physical and genome size. The capsid has an overall size of 700 nm in diameter and a genome of 1.18 Mb, with ~1000 protein-encoding genes, and its capsid is covered with a thick layer of long fibers (~200 nm). Recently, even larger viruses have been isolated. Pandoraviruses (**Fig. 1c**) have a size reaching 1  $\mu\text{m}$  with a shape reminiscent of some types of bacteria (Philippe et al. 2013). The *Pandoravirus salinus* genome is 2.5 Mb and contains 2,556 putative protein-encoding genes, of which only 7% have recognizable relationships with genes from other known organisms. *Pithovirus sibericus* was isolated from a 30,000-year-old permafrost sample harvested from Siberia (Legendre et al. 2014). This virus is most closely related to Marseilleviridae, another group of giant viruses. Its maximum size can reach 1.5  $\mu\text{m}$ , and it has an oval shape similar to Pandoraviruses; however, its genome only has ~467 protein-encoding genes.



**Fig. 1.** Electron microscopy of (a) PBCV-1. (Adapted from Zhang et al. 2011). (b) *Mimivirus*. (Adapted from Ameobal Pathogen Mimivirus Infects Macrophages through Phagocytosis, Ghigo E et al. PLOS Pathogens 2008, 4(6): e1000087, doi: 10.1371/journal.ppat.1000087), (c) *Pandoravirus*. (Adapted by permission from Macmillan Publishers Ltd: Nature, Ed Young, Giant Viruses open Pandora's box, doi: 10.1038/nature.2013.13410, copyright 2013).

Thus, giant viruses differ in shape, genome size and number of encoded proteins. New members are continually being discovered, and ongoing genomic analyses indicate that some of them encode many glyco-related genes, including those for the biosynthesis of nucleotide-sugar precursors and glycosyltransferases. However, information on the role of these genes and on the structure of the glycans associated with them are still fragmentary. Furthermore, such viral glycans have been overlooked because of the lack of the appropriate tools to investigate them.

Currently, the most studied genus is Chloroviruses; accordingly, this chapter focuses on PBCV-1 as a representative chlorovirus member. We provide a brief description of its features and then describe the methods that allowed us to determine its new and unusual glycan structure.

## **2 Chloroviruses and *Paramecium bursaria* Chlorella Virus (PBCV-1)**

The plaque-forming chloroviruses (family *Phycodnaviridae*) are ubiquitous in freshwater throughout the world with titers occasionally reaching thousands of plaque-forming units (PFU) per ml of indigenous water. Chlorovirus hosts are unicellular microalgae, often referred to as zoochlorellae and mostly found in nature as endosymbionts of protists (Karakashian and Karakashian 1965).

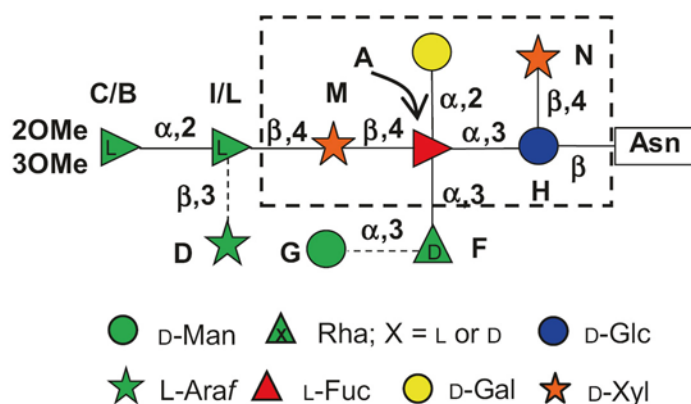
Chloroviruses are divided in four groups depending on the host selectivity, and the genomes of 43 chloroviruses have been sequenced, assembled and annotated (Jeanniard et al. 2013; Quispe et al. 2017). Collectively, the viruses encode genes from 643 predicted protein families, and at least 17 of them are predicted to be involved in manipulating carbohydrates including (1) enzymes involved in making extracellular polysaccharides such as hyaluronan and chitin; (2) enzymes that make nucleotide sugars such as GDP-L-fucose and GDP-D-rhamnose; and (3) enzymes involved in the synthesis of glycans attached to the virus major capsid proteins (Van Etten et al. 2010, 2017).

The capsid of the prototype chlorovirus PBCV-1 is an icosahedron with a spikelike structure at one vertex, which makes the first contact with the wall of the host cell (Zhang et al. 2011; Van Etten and Dunigan 2012), and a few external fibers that extend from some of the viral capsomeres (Cherrier et al. 2009). The PBCV-1 major capsid protein (named Vp54) is a glycoprotein, and the glycosylation pattern of this protein,

which is the subject of this manuscript, differs from that of all other organisms known to date.

Typically, viruses, such as HIV or Ebola, use host-encoded glycosyltransferases and glycosidases located in the endoplasmic reticulum (ER) and Golgi apparatus to add and remove N-linked sugar residues from virus glycoproteins (Vigerust and Shepherd 2007). Consequently, the glycan portion of the glycoproteins of these viruses is host-specific and resembles that of the host.

However, glycosylation of the PBCV-1 major capsid protein differs from the scenario described above because the virus encodes most, if not all, of the machinery for the process (Van Etten et al. 2010, 2017). Solving the structure of the four Vp54 N-linked glycans (De Castro et al. 2013) proved the uniqueness of PBCV-1 glycosylation: (1) Vp54 has four Asn-linked glycans, and none of the Asn are located in an Asn-X-(Thr/Ser) sequon characteristic of ER-located glycosyltransferases (Nandhagopal et al. 2002; De Castro et al. 2018); (2) the glycans are attached to Asn by a  $\beta$ -glucose linkage, which is rare in nature; (3) the glycans are highly branched and consist of eight to ten neutral monosaccharides (**Fig. 2**); (4) the four glycoforms contain a dimethylated rhamnose as the capping residue of the main chain, a hyper-branched fucose



**Fig. 2.** Structures of glycoforms of the major capsid protein Vp54 from chlorovirus PBCV-1. Residues are labeled with the letter used during the NMR assignment (Fig. 5). Glycoform 1 lacks residue **D** and the two L-rhamnose units are labeled **C** and **I**. Glycoform 2 has the arabinose residue **D**, and the two L-rhamnose residues are labeled **B** and **L**. (Adapted from Van Etten et al. 2017). Residues within the box are conserved among all chloroviruses; rhamnose **F** is a semiconserved element because its absolute configuration is virus-dependent. Additional decorations occur on this core N-glycan and represent a molecular signature for each chlorovirus.

residue, and two rhamnose residues with opposite absolute configurations; (5) the four glycoforms differ by the nonstoichiometric presence of two monosaccharides, L-arabinose and D-mannose; and (6) there is a core region (Fig. 2) strictly conserved in the N-glycans of all the chloreviruses studied to date (De Castro et al. 2016; Quispe et al. 2017; Speciale et al. 2017). The glycan structures do not resemble any structure previously reported in the three domains of life.

Solving the structure of the Vp54 glycans was a challenging issue that was accomplished by a combination of gas chromatography-mass spectrometry (GC-MS), NMR spectroscopy, and MALDI spectrometry techniques as described in the following sections.

### **3 GC-MS Approach**

GC-MS is a powerful technique. Its detection limit is very low, but it has one prerequisite: the glycan (or the glycoprotein) needs to be depolymerized and transformed into a proper volatile derivative. Depending on the type of derivative, GC-MS helps to elucidate the monosaccharides, their absolute configuration and substitution pattern. This information is of paramount importance because it supports the interpretation of NMR or MALDI spectra.

#### ***3.1 Determination of the Glycan Composition***

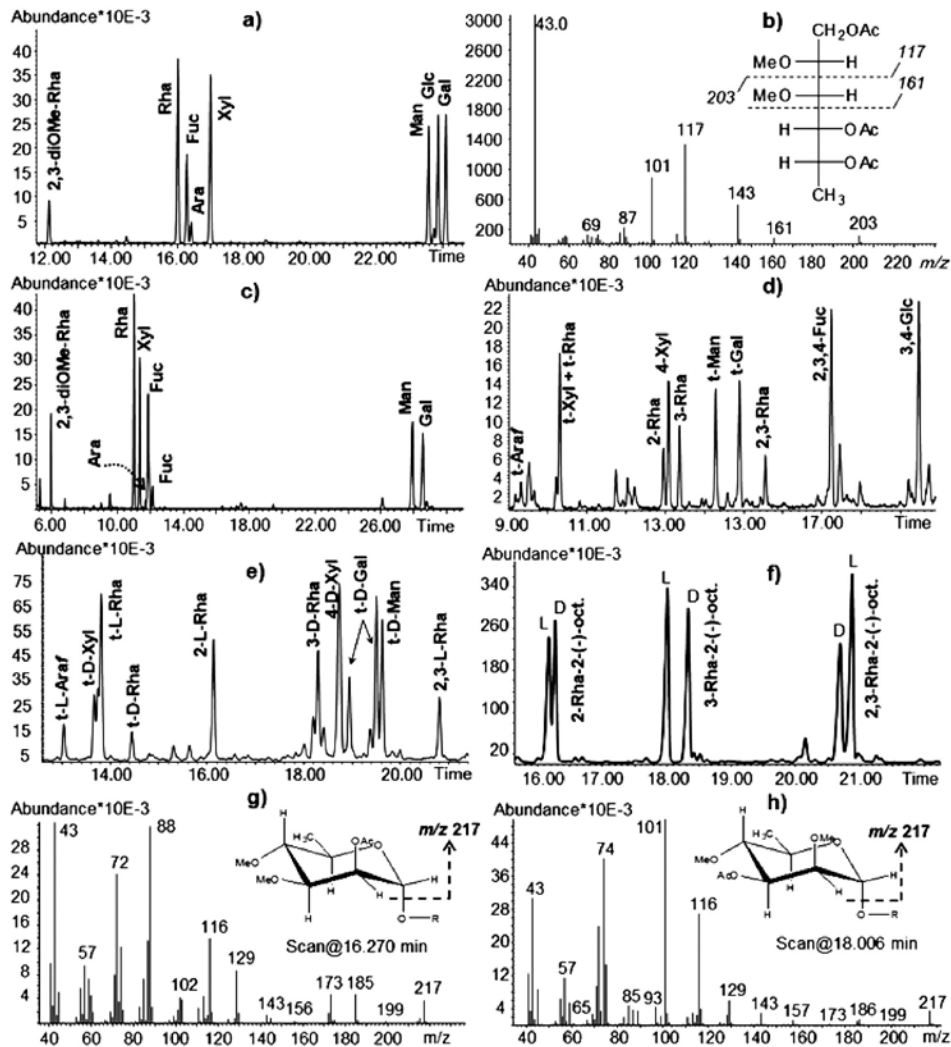
To detect the type of sugars, monosaccharides are transformed into the corresponding acetylated alditols (AA) or acetylated methyl glycosides (AMG), depending on the type of depolymerization procedure used, hydrolysis or methanolysis, respectively. Both procedures can be applied directly to the glycoprotein with no significant interference from the amino acids, and they are complementary in terms of disadvantages and advantages. The AA procedure is limited to neutral or basic monosaccharides, either aldose or ketose, but it cannot detect acidic monosaccharides unless the reduction of the carboxylic group is performed first. The advantage of this approach is that aldoses are transformed into an acyclic molecule, which gives only one peak in the chromatogram, a feature that is desirable for accurate quantification; in addition, if a monosaccharide is partially methylated, the position of the substituent is easily



inferred by applying the fragmentation rules that exist for the partially methylated and acetylated alditol derivatives. The main disadvantage is that enantiomeric monosaccharides, such as L- and D-rhamnose, cannot be distinguished, while in some cases, such as arabinose and lyxose, the two alditols are identical once the aldehydic function is reduced. In addition, ketoses produce two different alditols upon the reduction of their carbonyl function.

To solve some of the pitfalls of the AA procedure, it is good practice to combine the results with those from other types of derivatives; acetylated methylglycoside is the best alternative. This procedure (De Castro et al. 2010) is less laborious than AA and detects neutral, basic, and acidic monosaccharides but fails with neutral ketoses, which are not detected. Solvolysis of the glycoprotein or of the glycopeptides with hydrochloric methanol depolymerizes the glycan and transforms the residues in the corresponding methylglycosides, which are later acetylated. The main difference with the AA method is that the AMG approach produces cyclic sugar derivatives, either in pyranose or furanose form and with both configurations at the anomeric center; in other words, one monosaccharide can present up to four peaks in the chromatogram, with one usually more abundant than the others. Detection of the monosaccharide as cyclic derivatives removes the problem related to the production of identical derivatives, as mentioned earlier for the alditols of arabinose and lyxose.

Both AA and AMG approaches were performed on PBCV-1 glycopeptides. The AA profiles (**Fig. 3a**) contained eight monosaccharides in different proportions, of which only seven could be assigned with confidence by comparing their electron-impact (EI) MS spectrum and their retention times with that of authentic standards. The species that eluted at 12 min had an EI-MS spectrum (**Fig. 3b**) interpreted on the basis of the rules valid for partially methylated and acetylated alditols (Lönngren and Svensson 1974). In brief, the EI-MS spectrum contains cations that originate from the rupture of the carbon-carbon linkage of the molecule (**Fig. 3b**). Importantly, the carbon bearing the methoxyl, rather than an acetyl group, retains a positive charge and is revealed. Therefore, the ions displayed in the EI-MS spectrum result from the position of the methyl group(s) of the alditol, and the spectrum of species at 12 min was consistent with a derivative of a 6-deoxyhexose methylated at both 2 and 3 positions (**Fig. 3b**). The *manno*-stereochemistry of the residue was inferred later on the basis of the PMAA derivatives and NMR studies.



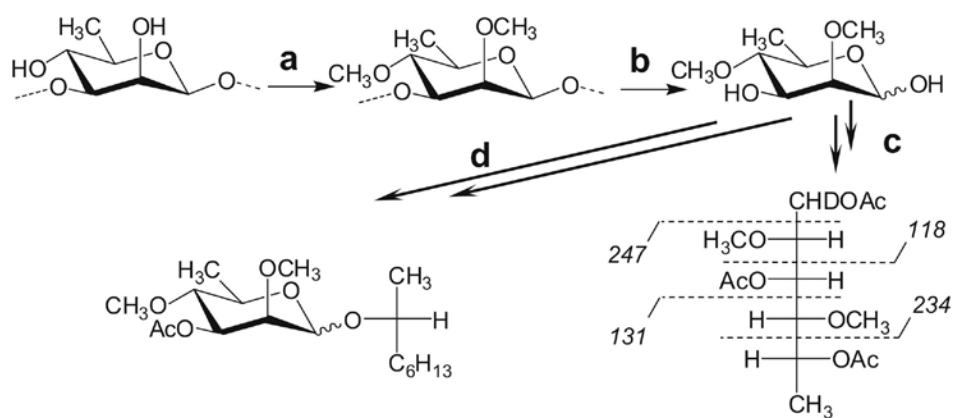
**Fig. 3.** GC-MS chromatograms from PBCV-1 glycopeptide Vp54, each analysis was performed by derivatizing 0.2–0.3 mg of sample. (a) Profile of the acetylated alditol derivatives. (b) EI-MS spectrum of the species at 12 min. of chromatogram 3A, together with the structure of the alditol and of the fragmentation pathway. (c) Profile of the acetylated methylglycosides, (d) of the PMAA, and (e) expansion of the chromatogram of the partially methylated and acetylated octyl glycosides reporting the different rhamnose species. (f) Chromatogram of the standard of the partially methylated and acetylated octyl glycosides of rhamnose 2-, 3-, and 2,3-linked. (g and h) EI-MS spectra of the peaks at 16.3 and 18.0 min. of panel (f), representing the derivatives of 2- and 3-linked octyl rhamnosides, respectively. (Fig. 3f–h is adapted from De Castro et al. (2013))

The GC-MS chromatogram of the acetylated methylglycosides (Fig. 3c) provided information similar to that from the acetylated alditols procedure; it confirmed the identity of arabinose as distinguished from

all the other residues. Notably, direct methanolysis of the glycopeptide did not cleave the linkage of the N-linked glucose, which can be accomplished by adding a hydrolytic step prior to methanolysis. Thus, by comparing results from these two derivatization procedures, it was possible to deduce which was the first monosaccharide N-linked to the protein, information that is not trivial in a system with unusual N-linked glycans.

### 3.2 Determination of the Glycan Substitution Pattern

Evaluation of the substitution pattern of the residues in a polysaccharide is also called linkage analysis (De Castro et al. 2010). The key step is the transformation of the glycan into its methylated derivative (**Fig. 4a**); during this procedure the available free hydroxyl functions are converted into methyl ethers, *N*-acyl amino sugars are *N*-methylated, and uronic acids are esterified. At this stage, if a uronic acid is present, the methyl ester function is reduced; otherwise the residue is lost during successive steps. Next, the hydrolysis step (**Fig. 4b**) cleaves all the glycosidic linkages and releases the partially methylated monosaccharides;



**Fig. 4.** Reaction scheme for the production of the partially methylated and acetylated alditol (PMAA) derivative of a 3-linked rhamnose. **(a)** Methylation of the native glycan: the free hydroxyl groups are transformed in methyl ethers. **(b)** The methylated glycan is hydrolyzed to give a monosaccharide partially methylated. **(c)** Carbon 1 is labeled by reduction with a deuterated hydride and acetylated to produce the final partially methylated and acetylated derivative (PMAA), reported with its fragmentation scheme. Acetyl groups at carbons 1 and 5 indicate the ring closure of the monosaccharide; acetyl group at position 3 indicates a previously linked position of the residue. **(d)** Partially methylated and acetylated octyl rhamnoside obtained after methylation, hydrolysis, and octanolysis of the glycopeptide. The acetyl group is indicative of a previously linked position of the residue as in the PMAA in Fig. 4c.

the free hydroxyl groups that now appear are those originally involved in the glycosidic linkages. Successively, the monosaccharide is transformed into an alditol by reduction with a hydride, usually  $\text{NaBD}_4$ , to label the anomeric carbon; finally, acetylation yields the so-called partially methylated and acetylated alditols (PMAA, **Fig. 4c**).

This approach was applied to the glycopeptides of PBCV-1, omitting the reduction of the methyl ester function because no uronic acid was detected in the compositional analysis.

The GC-MS chromatogram presented a large array of peaks (Fig. 3d), and the interpretation of their EI-MS spectra (**Table 1**), by the rules mentioned in Sect. 3.1, identified the type of monosaccharide (pentose, hexose, or deoxyhexose) present along with their substitution pattern. Finally, determination of the stereochemistry of each species was accomplished by injecting the PMAA standards available in the laboratory collection.

Using these procedures, 11 different PMAAs were found (Fig. 3d), indicating a very complex glycan. Notably, rhamnose existed in four different forms (terminal, 2-linked, 3-linked and 2,3-linked), xylose in two forms (terminal and 4-linked), while fucose, arabinose and the three hexoses only formed one PMAA derivative.

**Table 1.** Linkage analysis of PBCV-1 glycopeptide. Unless indicated otherwise, each PMAA derivative denotes a residue in the pyranose form and the number, the alditol stereochemistry is indicated with the acronym of the corresponding monosaccharide, and the number indicates the position substituted, while “t” means that the monosaccharide is terminal

<i>Component</i>	<i>Retention time (min.)</i>	<i>Diagnostic ions<sup>a</sup></i>
t-Araf	9.30	101 <sup>a</sup> , 102 <sup>a</sup> , 118, 129 <sup>a</sup> , 161, 162
t-Xyl	10.32	101 <sup>a</sup> , 102 <sup>a</sup> , 118, 161, 162
t-Rha	10.32	89 <sup>a</sup> , 102 <sup>a</sup> , 118, 131, 162, 175
2-Rha	12.95	89 <sup>a</sup> , 130 <sup>a</sup> , 131, 190
4-Xyl	13.08	102 <sup>a</sup> , 118, 129 <sup>a</sup> , 162, 189
3-Rha	13.37	89 <sup>a</sup> , 118, 131, 202 <sup>a</sup> , 234
t-Man	14.28	102 <sup>a</sup> , 118, 129 <sup>a</sup> , 145 <sup>a</sup> , 161, 162, 205
t-Gal	14.88	102 <sup>a</sup> , 118, 129 <sup>a</sup> , 145 <sup>a</sup> , 161, 162, 205
2,3-Rha	15.57	89 <sup>a</sup> , 131, 202 <sup>a</sup> , 262
2,3,4-Fuc	17.22	87, 129 <sup>a</sup> , 146, 159, 171a, 218, 231, 290
3,4-Glc	19.48	118, 185 <sup>a</sup> , 305

Ions generated from the primary fragments by loss of a neutral molecule: acetic acid ( $m/z$  60), and/or acetic anhydride ( $m/z$  102), and/or methanol ( $m/z$  32)

These results established the nature of the 2,3-diOMe-6-deoxyhexose found in the acetylated alditol analysis (Sect. 3.1). Depending on its location in the glycopeptide, the methylation protocol could transform this residue into two different PMAAs, a 4-linked deoxyhexose if substituted at the only available position, or into a fully methylated derivative if it is located in a terminal position. Accordingly, the only PMAA derivative compatible with one of the two options was a terminal rhamnose unit.

### ***3.3 Determination of the Monosaccharide Absolute Configuration***

The acetylated methyl glycosides of enantiomeric sugars, as D- and L-rhamnose, are equivalent during GC-MS analysis, because the GC column used can separate diastereoisomers but not enantiomers. The identification of the absolute configuration is therefore possible by replacing the methanol with a chiral alcohol, usually 2-octanol, to prepare the acetylated glycoside (acetylated octyl glycoside, AOG). The advantage of this approach over the traditional one is that the isolation of mg amounts of the target monosaccharide to determine its chirality with the polarimeter is not necessary. In addition, the AOG method enables the contemporaneous determination of the stereochemistry of a mixture of monosaccharides, if the appropriate standards are available.

Preparation of the acetylated octyl glycoside standard is a procedure that consists of two steps: (1) reaction of the monosaccharide with racemic 2-(±)-octanol and (2) reaction of the monosaccharide with one optically pure enantiomer of the alcohol, for instance, 2-(-)-octanol. The first step, using L-rhamnose (L-Rha) as an example, produces two diastereoisomers, L-Rha-(+)-oct and L-Rha-(-)-oct, each equivalent to its enantiomer, D-Rha-(-)-oct and D-Rha-(+)-oct, respectively, during GC-MS analysis. The second step gives only one diastereomer, i.e., L-Rha-(-)-oct whose GC-MS retention time is also equivalent to that of its enantiomer D-Rha-(+)-oct. Thus, the two separate reactions allow the determination of the retention time of L-Rha-(-)-oct and D-Rha-(-)-oct. This information is used to assign the configuration of a monosaccharide, prepared by treating the sample with enantiomeric pure 2-(-)-octanol, followed by acetylation.

This procedure applied to PBCV-1 glycopeptides revealed that fucose and arabinose are L-configured, the three hexoses and xylose are D, while rhamnose exists in both L and D configurations. Collectively, PMAA and AOG analyses indicated that different types of rhamnose residues were

in the glycopeptide and that they differed in the type of substitution and in their absolute configuration. This finding prompted an improvement in the octyl glycoside approach that produced a derivative able to provide the absolute configuration of the residue and its substitution pattern. Accordingly, the glycopeptide was methylated and hydrolyzed as in Fig. 4b, and then the mixture of the partially methylated monosaccharide was directly converted into the corresponding partially methylated and acetylated octyl glycosides (Fig. 4d).

The standards needed for the identification of rhamnose in its variously acetylated and methylated forms were obtained by applying the same approach to the polysaccharide from *Kaistella flava* (Gargiulo et al. 2008), because it had a L-rhamnose linked at either 2 or 3 and at both 2 and 3 positions. The standard for the terminal rhamnose was prepared by methylation of the octyl rhamnoside.

Attribution of the GC-MS profile (Fig. 3e) was possible by extension of the fragmentation rules of the methyl glycosides to the octyl glycosides (Lönngren and Svensson 1974). In general, the most diagnostic ion is the oxonium ion that originates from the loss of the aglycon of the anomeric carbon. For the two peaks at ca. 21 min (Fig. 3f), it was  $m/z$  245 that indicated a rhamnose derivative with two acetyl groups and one methyl group, as expected from the transformation of the 2,3-linked rhamnose unit of the glycan. As for the two peaks at 16 and 18 min (Fig. 3f), they had different EI-MS spectra (Fig. 3g, h) but the same oxonium ion at  $m/z$  217, pointing to a rhamnose with two methyl groups and one acetyl, as expected for a 2- or 3-linked residue. The 2-linked rhamnose species were identified at 16 min because the EI-MS spectrum had an intense ion at  $m/z$  88 that arose from two vicinal methylated hydroxyl groups, which is only possible for a 2-linked residue. Thus, the peaks at 18 min were assigned to 3-linked rhamnose. Identification of the l and d isomer (Fig. 3f) was possible by preparing the AOG from the methylated polysaccharide with pure 2-(-)-octanol.

Finally, the PBCV-1 glycopeptide reaction with pure 2-(-)-octanol and acetylation indicated that D-rhamnose was mainly 3-linked (major form), but also terminal, while L-rhamnose was 2- or 2,3-linked or terminal (Fig. 3e). Thus, preparation of partially methylated and acetylated octyl glycosides is useful to solve glycan structures that have monosaccharides with opposite configurations; while these cases are rather rare, they are quite challenging when they exist.

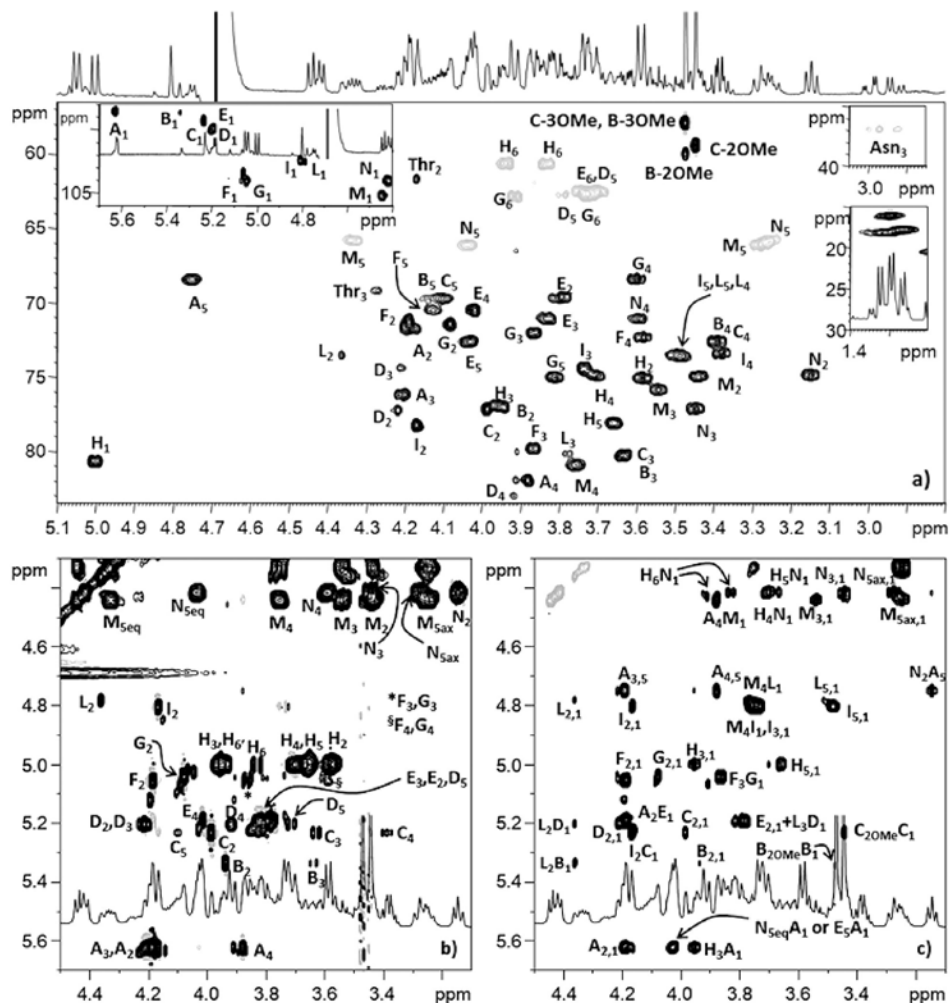
## 4 NMR Spectroscopy in Glycan Analysis, Determination of Glycan Sequence, and Anomeric Configuration

NMR spectroscopy is without doubt the best technique for providing information about several features of a glycan, including the configuration of the anomeric carbon, the stereochemistry of the residues and the way the residues are connected. This approach is well developed for samples in solution and combines the information from a certain range of experiments, viz., the homonuclear  $^1\text{H}$ - $^1\text{H}$  2D sequences COSY, TOCSY, and ROESY (or NOESY) and the heteronuclear  $^1\text{H}$ - $^{13}\text{C}$  2D sequences HSQC, HMBC, and HSQC-TOCSY. The major limitation of this technique is related to the amount of sample available, a limit that nowadays is less severe because of the availability of instruments at high field, equipped with cold probes, and gradient-selected sequences that shorten the acquisition time. Thus, the following comments describe the major shortcuts for quickly obtaining the main structural information of a glycan.

### 4.1 Configuration of the Anomeric Carbon

The regions of interest are those at 5.5–4.3 ppm and 110–90 ppm for proton and carbon spectra, respectively. Proton signals found above 5 ppm are generally considered diagnostic of a residue in the  $\alpha$ -pyranosidic form, while those below 5 ppm are diagnostic of  $\beta$ -configured pyranose sugars. In regard to the carbon signals, the anomeric carbon of an  $\alpha$ -configured pyranose residue has a chemical shift lower than a  $\beta$ -configured unit, and it is generally found at less than 102 ppm. This generalization does not consider residues in the furanose form or having a N-glycosidic linkage; for these cases, the determination of the anomeric configuration needs information on all of the proton and carbon signals.

In the case of the PBCV-1 glycan, the HSQC spectrum (**Fig. 5a**) had about 11 anomeric signals. Of these 11 signals, 6 were assigned to  $\alpha$ -configured monosaccharides (**A–C**, **E–G**) because they were found at proton signals above 5 ppm. Unit **D** represented an arabinose in the  $\beta$ -furanose form, in agreement with linkage analysis and with the carbon four (C-4) chemical shifts at 83.0 ppm (Bock and Pedersen 1983). All the other signals belonged to  $\beta$ -configured pyranose residues, with glucose **H** having an N-type glycosidic linkage as inferred from the characteristic signal at 80 ppm.



**Fig. 5.** NMR spectra acquired for PBCV-1 glycopeptide structural studies (refer to Fig. 2 for labeling system, glycopeptide amount 0.5 mg). (a) HSQC spectrum. (b) Expansion of the TOCSY spectrum. (c) Expansion of T-ROESY spectrum. (Adapted from De Castro et al. 2013)

## 4.2 Stereochemistry of the Residues

For most of the residues, this feature can be quickly determined by analyzing the anomeric region of the TOCSY spectrum (Fig. 5b). In this experiment, the transfer of the magnetization from one proton to the next is controlled by the value of the coupling constants ( $^2J_{\text{H,H}}$  and  $^3J_{\text{H,H}}$ ), which in turn depends on the stereochemistry of the monosaccharides investigated. In the case of a pyranose ring, small  $^3J_{\text{H,H}}$  values are characteristic of axial/equatorial- or equatorial/equatorial-arranged protons and



hinder the magnetization transfer, thus decreasing the intensity of the successive correlations in the spectrum or, in some cases, preventing them completely. For example, *manno*- and *galacto*-configured residues in the pyranose form have small  $^3J_{H1,H2}$  and  $^3J_{H3,H4}$  values, respectively. Accordingly, the anomeric proton (H-1) of an  $\alpha$ -*manno*-configured monosaccharide correlates with H-2 and has very poor correlations, if at all, with the other protons in the residue; the H-1 of the  $\beta$ -isomer correlates only with H-2. A *galacto*-configured residue instead has only three correlations, from H-2 to H-4, independent of the configuration at the anomeric center. In contrast, residues with large coupling values such as *gluco*-configured forms have the complete correlation pattern, up to both H-6 protons (or up to H-5 protons in xylose residues). This approach is less straightforward for recognizing other isomers, such as those with *altro*- or *talo*-stereochemistry, or residues in the furanose form. Therefore, it is always good practice to confirm the identity of a monosaccharide by comparing its NMR data with that of a reference compound (Bock and Pedersen 1983) and with information from chemical analysis.

Regarding the PBCV-1 glycopeptides, the TOCSY spectrum in Fig. 5b identified the residues **I**, **L**, **F**, **G**, **C**, and **B** as *manno*-configured, and successive integration of TOCSY with COSY and HSQC spectra determined that all the residues were 6-deoxy-mannose or rhamnose units except **G**, which was a mannose. Similarly, **A** and **E** had a *galacto*-configuration, but **A** was a fucose, while **E** was a galactose. The residues **H**, **M**, and **N** had the complete correlation pattern (Fig. 5b), and further analysis disclosed that **H** was a glucose, while **M** and **N** were two units. The arabinose residue escaped from this scheme because it was in the furanose form; thus, its identity was determined by assignment of all the chemical shift values as reported in the previous section.

### 4.3 Sequence of the Residues

This aspect is pursued by two main approaches. Homonuclear  $^1\text{H}$ - $^1\text{H}$  NOESY and ROESY (or its variation T-ROESY) sequences exploit the nuclear Overhauser effect (NOE) that relates nuclei at less than 4 Å. Thus, NOE effects occur within the same residue (intra-residue) as well as between different sugar units (inter-residues). Both types of NOEs are useful during structural determination, but only inter-residue NOEs define the sequence of the residues in the polymer and, in most of the cases, also the substitution point.

As for the PBCV-1 glycopeptide, T-ROESY (**Fig. 5c**) displayed several correlations. For instance, the anomeric signal of **M** (a xylose unit) had three correlations: those with protons H-3 and H-5 ( $M_{3,1}$  and  $M_{5ax,1}$ , respectively) confirmed the  $\beta$  configuration of the anomeric center, while the last,  $A_4M_1$ , placed this residue unequivocally at position 4 (C-4) of **A**, the fucose residue. A similar pattern of intra- and inter-residue NOEs was detected for the other residues  $\beta$ -configured at the anomeric position, namely, **I** and **L**.

The anomeric proton of an  $\alpha$ -configured residue displays the inter-residue correlation with the proton at the position of the glycosidic linkage, while the intra-residue correlations occur with H-2. However, NOE effects report spatial proximity, meaning that care is necessary during interpretation. For instance, arabinose **D** is linked at C-3 of **L**, but it correlates with both H-2 and H-3 of this unit (correlations  $L_2D_1$  and  $L_3D_1$  in Fig. 5c). Similarly, xylose **N** is linked at C-4 of **H**, but it correlates with both H-4 and H-6 of the glucose unit (correlations  $H_4N_1$  and  $H_6N_1$  in Fig. 5c). In addition, other NOE effects connect protons of residues not even directly linked, such as H-2 of **N** with H-5 of **A** ( $N_2A_5$  in Fig. 5c), which means that the glycan adopts a conformation that brings these two parts close to each other.

The second approach consists of recording the heteronuclear  $^1H$ - $^{13}C$  HMBC spectrum that relates proton and carbon nuclei separated from two ( $^2J$ ) or three ( $^3J$ ) chemical bonds and that are scalar (or  $J$ ) coupled. Indeed, as for NOE-based techniques, HMBC detects intra- ( $^2J$  and  $^3J$ ) and inter-residue correlations ( $^3J$ ), and the intensity associated depends on the  $J$  value. The absence of a correlation does not mean that two residues are not linked but simply that the  $^3J$  value is close to null or far from the value for which the HMBC acquisition is optimized. It is good practice to record this type of spectrum to confirm the results from the NOESY (or T-ROESY) experiments as was done for PBCV-1.

## 5 Glycoproteomic Analysis of MCP from Chloroviruses by MALDI-TOF MS and MALDI-TOF-TOF MS/MS

Complete characterization of glycoproteins requires the identification of the glycosylation sites, as well as the site-specific analysis of the glycan components, which often have a broad number of glycoforms (microheterogeneity) that enhances the overall molecular complexity. Due

to the many structural components to be examined, conventional MS-based methods typically rely on a combination of two different types of analyses: (1) a basic glycomic strategy that analyzes the enzymatically or chemically released glycan mixture and (2) a glycoproteomic approach which also provides information on the glycosylation sites by analyzing the glycopeptides resulting from protein digestion. We adopted this last procedure to analyze the N-glycosylation pattern of the PBCV-1 Vp54 protein because the first approach was hampered by the lack of N-glycan releasing enzymes with the appropriate specificity. Moreover, chemical deglycosylation methods would have led to the loss of the protein counterpart and therefore loss of the structural information at the glycosylation sites. Protease digested samples often contain a complex mixture of peptides and glycopeptides, which are poorly detected because they can be heterogeneous in composition and because their ionization efficiency is less than that of the unmodified peptide. Accordingly, a number of chromatographic methods based on physical and chemical properties of the glycosylated species have been developed for the enrichment/separation of the glycopeptide fraction (Wuhrer et al. 2007; Ongay et al. 2012).

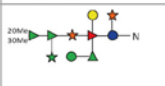
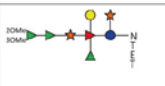
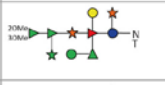
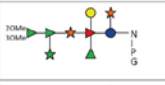
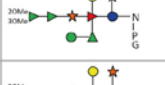

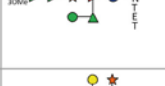

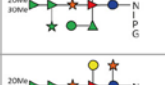
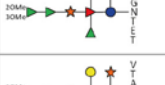
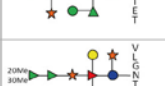





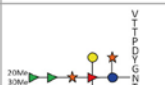
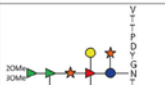
As a first step, our experimental protocol involved the enzymatic digestion of the viral Vp54 glycoprotein with two different proteases, proteinase K and thermolysin, to generate a mixture of properly sized N-linked glycopeptides (**Table 2**). With its broad specificity, proteinase K produced fractions enriched in glycopeptide fragments that were separated via size-exclusion chromatography (SEC) on a Bio-Gel P-10 column. The more selective protease, thermolysin, in association with alternative glycopeptide purification procedures, such as reverse-phase (RP) fractionation on C18 columns, hydrophilic interaction liquid chromatography (HILIC) on microcrystalline cellulose, and solid-phase extraction (SPE) on graphitized columns, produced a different set of glycopeptides. Because of its low specificity, proteinase K produced glycans with either a single amino acid or attached to short peptides, whereas thermolysin produced glycopeptides with larger peptides. Trypsin digestion, despite having more targeted cleavages of the protein backbone, was not used because it was expected to generate glycosylated peptides too large for MALDI MS and MS/MS analyses.

By using a combination of the two proteases and the various enrichment columns, we were able to isolate several fractions enriched in

glycopeptides (Table 2). In the majority of the cases, the samples were still complex mixtures, containing some peptides and mainly glycopeptides whose heterogeneity depended on both peptide and glycan components. All the fractions were analyzed off-line by MALDI-TOF MS and MALDI-TOF-TOF MS/MS in order to identify protein glycosylation sites and glycan sequences.

First-line analysis employed MALDI MS identification of glycopeptides, which always detected the positive ionization mode as adducts with sodium  $[M+Na]^+$  and potassium  $[M+K]^+$ . The spacing pattern ( $\Delta m/$

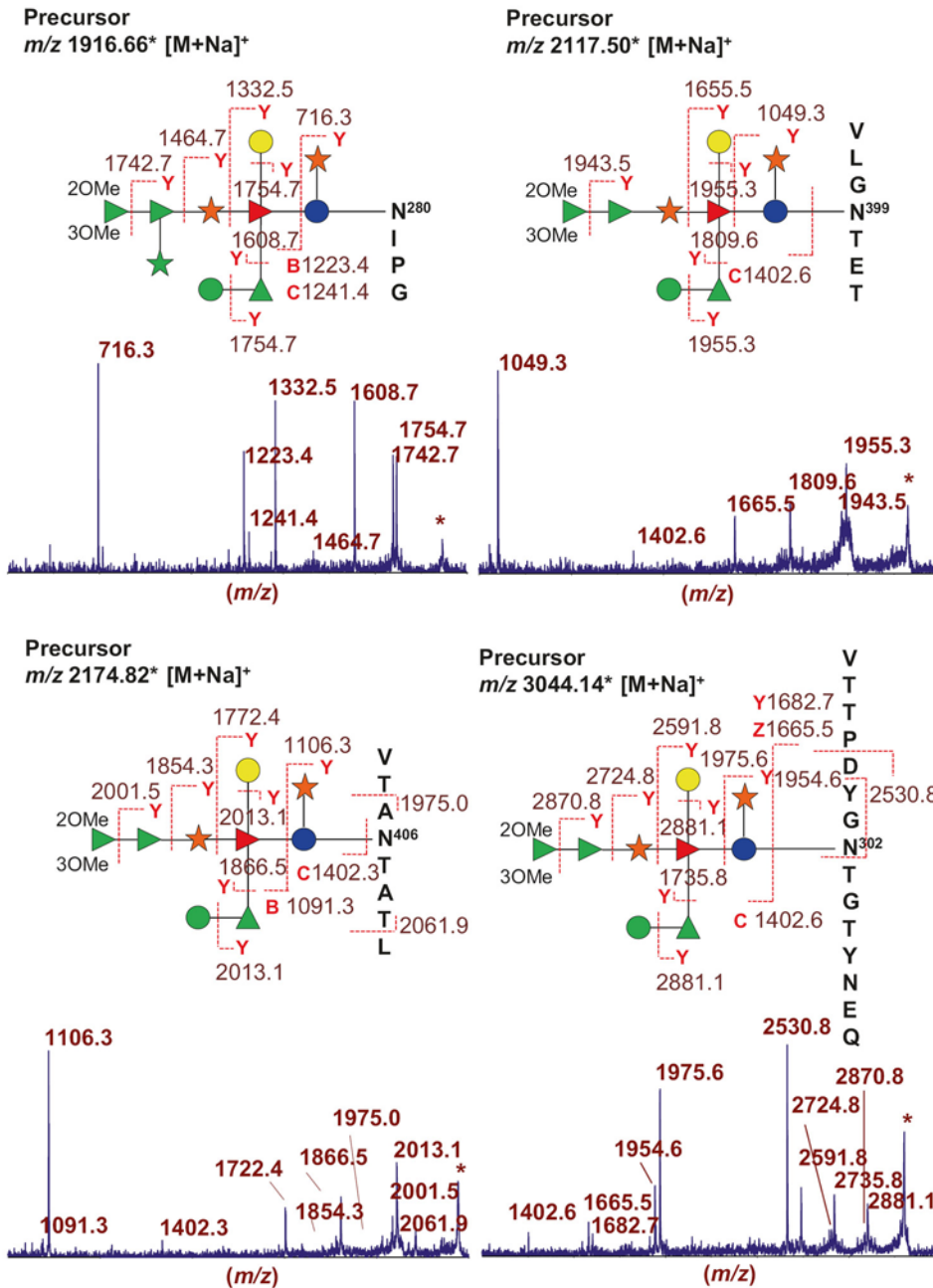
**Table 2.** Structures of Vp54 glycopeptides identified by MALDI-TOF MS and MALDI-TOF-TOF MS/MS. Individual species are reported together with the relative glycosylation site(s) assignment and the use of proteolytic enzymes and enrichment methods

$m/z$ [M+Na] <sup>+</sup> observed (theoretical)	Structure	Glycosylation site	Protease digestion glycopeptide enrichment	$m/z$ [M+Na] <sup>+</sup> observed (theoretical)	Structure	Glycosylation site	Protease digestion glycopeptide enrichment
1649.51 (1649.59)		280 Asn 302 Asn 399 Asn 406 Asn	Proteinase K Bio-Gel P10	1686.51 (1686.63)		399 Asn	Proteinase K Bio-Gel P10
1750.54 (1750.64)		302 Asn 399 Asn 406 Asn	Proteinase K Bio-Gel P10	1754.64 (1754.69)		280 Asn	Proteinase K Bio-Gel P10
1784.60 (1784.70)		280 Asn	Proteinase K Bio-Gel P10	1818.55 (1818.67)		399 Asn	Proteinase K Bio-Gel P10
1848.58 (1848.68)		399 Asn	Proteinase K Bio-Gel P10	1899.60 (1899.78)		406 Asn	Thermolysin RP-C18 Proteinase K Bio-Gel P10
1916.66 (1816.75)		280 Asn	Proteinase K Bio-Gel P10	1955.67 (1955.81)		399 Asn	Thermolysin RP-C18
1980.62 (18980.73)		399 Asn	Proteinase K Bio-Gel P10	2061.74 (2061.83)		406 Asn	Thermolysin RP-C18/SPE
2117.91 (2117.86)		399 Asn	Thermolysin RP-C18	2174.82 (2174.92)		406 Asn	Thermolysin RP-C18/SPE
2249.96 (2249.90)		399 Asn	Thermolysin RP-C18	2282.16 (2282.15)		302 Asn	Thermolysin RP-C18/HILIC
3044.14 (3044.20)		302 Asn	Thermolysin RP-C18/HILIC	3176.17 (3176.24)		302 Asn	Thermolysin RP-C18

$z=16$ ) between these two adducts proved useful to distinguish glycopeptide from peptide ions. Moreover, the mass differences between the glycopeptides disclosed that some monosaccharides were present as non-stoichiometric substituents.

Complete assignment of the more significant glycopeptide MS signals was based on the accurate measurement of the molecular mass, which inferred both peptide and glycan composition (accuracy was more than 75 ppm for Vp54 main glycopeptides), and on tandem MS analysis which provided evidence for the glycan sequence. MALDI-TOF-TOF spectra gave primarily low-energy fragments consisting of the cleavage of the glycoside bonds between adjacent sugar rings — the B-, C-, and Y-type ions (rarely Z-type ions) according to Domon and Costello nomenclature (Domon and Costello 1988). B and C ions comprise the nonreducing sugar terminal, the first arising from a glycosyl bond  $\beta$ -elimination implying the further loss of a water molecule,  $[B]^+ = [C - H_2O]^+$ , whereas, analogously, Y and Z ions include the reducing end, with  $[Z]^+ = [Y - H_2O]^+$ . Glycan Y-type fragments were found as the predominant cleavage, which generated highly informative ions retaining the peptide glycosylation sites. On the other hand, B and C ions, typically occurring at or near the nonreducing end side, provided valuable information on the molecular mass of the whole glycan moiety and on the substitution of the innermost glucose residue. Notably, species with larger peptide portions also led to y- and b-type cleavage (Biemann 1992) of the amino acid backbone linkages, allowing further characterization of the peptide side chain.

By combining all the MS data with peptide sequence information, as well as those provided by the FindPept bioinformatic tool ([www.expasy.org](http://www.expasy.org)) and the compositional analysis, we were able to define the PBCV-1 N-glycosylation pattern. MS and MS/MS analyses helped to identify individual amino acid sequences associated with Vp54 glycosylation (at <sup>280</sup>NIPG, <sup>302</sup>NTGT, <sup>399</sup>NTET, <sup>406</sup>NTAT), as well as the characterization and relative evaluation of their attached glycoforms. Of note, none of the asparagine is in a typical consensus sequence; **Fig. 6** shows MS/MS analysis of the four glycopeptides that were more representative and reveal the complex pattern of Vp54 N-glycosylation.



**Fig. 6.** MALDI-TOF-TOF MS/MS spectra of selected glycopeptides from Vp54. Each precursor ion, as detected by MALDI-TOF MS, is associated with one of the four distinct Vp54 glycosylation sites. (Adapted from De Castro et al. 2013). The horizontal axis indicates the mass/charge ratio ( $m/z$ ) of the ions, and numbers are omitted to avoid crowding. Most relevant peaks are labeled with the mass value

## 6 Conclusions

The structure of the glycans of the PBCV-1 major capsid protein Vp54 had remained elusive because the tools commonly used to study eukaryotic-like glycosylation patterns failed. This glycosylation pattern has been solved only recently (De Castro et al. 2013) by combining three techniques, GC-MS, NMR and MALDI, and this same approach was effective when applied to other chlorovirus encoded glycoproteins (De Castro et al. 2016; Quispe et al. 2017; Speciale et al. 2017). As a result, the structures of other chlorovirus N-glycans are now available, confirming that glycosylation is a trait common to these viruses. The results also revealed that chlorovirus glycans share the same core oligosaccharide structure that is unique (Fig. 2) and not shared by any other form of life.

Chloroviruses are members of a rapidly expanding group of viruses characterized by large particle size with huge genomes and are often referred to as giant viruses. Many giant viruses have putative genes involved in carbohydrate metabolism, such as *Mimivirus* and *Megavirus* (both in the *Mimiviridae* family) for which we have direct evidence that a thick layer of polysaccharides coats their surface (De Castro and Tonetti, unpublished results). The take-home message is that viral autonomous glycosylation probably occurs in a wide range of viruses. The insightful use of the available methods has now added a new topic to the flourishing field of glycobiology.

## References

- Abergel C, Legendre M, Claverie JM (2015) The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol Rev* 39:779–796
- Biemann K (1992) Mass spectrometry of peptides and proteins. *Annu Rev Biochem* 61:977–1010
- Bock K, Pedersen C (1983) Carbon-13 nuclear magnetic resonance spectroscopy of monosaccharides. *Adv Carbohydr Chem Biochem* 41:27–66
- Cherrier MV, Kostyuchenko VA, Xiao C et al (2009) An icosahedral algal virus has a complex unique vertex decorated by a spike. *PNAS* 106:11085–11089
- Colson P, De Lamballerie X, Yutin N et al (2013) “Megavirales”, a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Arch Virol* 158:2617–2521
- De Castro C, Parrilli M, Holst O et al (2010) Microbe-associated molecular patterns in innate immunity: extraction and chemical analysis of gram-negative bacterial lipopolysaccharides. *Methods Enzymol* 480:89–115

- De Castro C, Molinaro A, Piacente F et al (2013) Structure of N-linked oligosaccharides attached to chlorovirus PBCV-1 major capsid protein reveals unusual class of complex N-glycans. *PNAS* 110:13956–13960
- De Castro C, Speciale I, Duncan G et al (2016) N-linked glycans of chloroviruses sharing a core architecture without precedent. *Angew Chem Int Ed* 55:654–658
- De Castro C, Klose T, Speciale I et al (2018) Structure of the chlorovirus PBCV-1 major capsid glycoprotein determined by combining crystallographic and carbohydrate molecular modeling approaches. *PNAS* 115:E44–E52
- Domon B, Costello CE (1988) A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconj J* 5:397–409
- Dunigan DD, Cerny RL, Bauman AT et al (2012) *Paramecium bursaria* chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. *J Virol* 86:8821–8834
- Gargiulo V, De Castro C, Lanzetta R et al (2008) Structural elucidation of the capsular polysaccharide isolated from *Kaistella flava*. *Carbohydr Res* 343:2401–2405
- Jeanniard A, Dunigan DD, Gurnon JR et al (2013) Towards defining the chloroviruses: a genomic journey through a genus of large DNA viruses. *BMC Genomics* 14:158
- Karakashian SJ, Karakashian MW (1965) Evolution and symbiosis in the genus *Chlorella* and related algae. *Evolution* 19:368–377
- Legendre M, Bartolia J, Shmakov L et al (2014) Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *PNAS* 111:4274–4279
- Lönngren J, Svensson S (1974) Mass spectrometry in structural analysis of natural carbohydrates. *Adv Carbohydr Chem Biochem* 29:41–106
- Nandhagopal N, Simpson AA, Gurnon JR et al (2002) The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. *PNAS* 99:14758–14763
- Ongay S, Boichenko A, Govorukhina N et al (2012) Glycopeptide enrichment and separation for protein glycosylation analysis. *J Sep Sci* 35:2341–2372
- Philippe N, Legendre M, Doutre G et al (2013) Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* 341:281–286
- Quispe CF, Esmael A, Sonderman O et al (2017) Characterization of a new chlorovirus type with permissive and non-permissive features on phylogenetically related algal strains. *Virology* 500:103–113
- Raoult D, Audic S, Robert C et al (2004) A huge virus that infects amoebae contains genes that are not usually part of the viral repertoire and defines a family of ancient nucleocytoplasmic DNA viruses. *Science* 306:1344–1350
- Speciale I, Agarkova I, Duncan GA et al (2017) Structure of the N-glycans from the chlorovirus NE-JV-1. *Anton van Leeuw* 110:1391–1399
- Van Etten JL, Dunigan DD (2012) Chloroviruses: not your everyday plant virus. *Trends Plant Sci* 17:1–8
- Van Etten JL, Meints RH, Kuczmarski D et al (1982) Viruses of symbiotic *Chlorella*-like algae isolated from *Paramecium bursaria* and *Hydra viridis*. *PNAS* 79:3867–3871



- Van Etten JL, Gurnon JR, Yanai-Balser GM et al (2010) Chlorella viruses encode most, if not all, of the machinery to glycosylate their glycoproteins independent of the endoplasmic reticulum and Golgi. *Biochim Biophys Acta* 1800:152–159
- Van Etten JL, Agarkova I, Dunigan DD et al (2017) Chloroviruses have a sweet tooth. *Viruses* 9:E88
- Vigerust DJ, Shepherd VL (2007) Virus glycosylation: role in virulence and immune interactions. *Trends Microbiol* 15:211–218
- Wuhrer M, Catalina MI, Deelder AM et al (2007) Glycoproteomics based on tandem mass spectrometry of glycopeptides. *J Chromatogr B* 849:115–128
- Zhang X, Xiang Y, Dunigan DD, Klose T et al (2011) Three-dimensional structure and function of the Paramecium bursaria chlorella virus capsid. *PNAS* 108:14837–14842