University of Naples Federico II

PhD in Chemical Sciences

Structural insight into new glycosylation patterns of microbial (bacterial and viral) origin

Immacolata Speciale

Tutor
Prof. Cristina De Castro

Supervisor
Prof. Angelina Lombardi

XXIX CYCLE

<u>Index</u>

Section I. Introduction

Section II. Results and discussion

Abbreviations

AAPM	Partially methylated alditol acetate
COSY	Correlation spectroscopy
ESI-MS	Electrospray Ionisation-Mass Spectrometry
FF	Force Field
G6P	Glucose-6-phosphate
G6PD	Glucose-6-phosphate dehydrogenase
GC-MS	Gas chromatography – Mass spectrometry
GST	Glutathione S-Transferase
HMBC	Heteronuclear multiple bond correlation
HSQC	Heteronuclear single quantum coherence
HSQC-TOCSY	Heteronuclear single quantum coherence - Total correlation spectroscopy
MALDI	Matrix-Assisted Laser Desorption/Ionization
MCMM	Metropolis Monte Carlo Minimization
MCP	Major Capsid Protein
MD	Molecular Dynamic
MM	Molecular Mechanic
NADP$^+$/NADPH	Nicotinamide-Adenine Dinucleotide Phosphate
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear overhauser enhancement spectroscopy
PBCV-1	*Paramecium bursaria* chlorella virus
PBS	Phosphate buffered saline
PDB	Protein Data Bank
RLU	Relative light unit
RMSD	Root mean square deviation
SDS-PAGE	Sodium Dodecyl Sulphate - PolyAcrylamide Gel Electrophoresis
TOCSY	Total correlation spectroscopy
T-ROESY	Total Rotating frame Overhauser effect spectroscopy
UGD	UDP-D-glucose-4,6-dehydratase

Monosaccharide symbols

Araf	β-L-arabinofuranose		
Fuc	α-L-fucose		
Gal	α-D-galactose		
Glc	β-D-glucose		
Man	α-D-mannose		
Rha	α-D-rhamnose		
Xyl	β-D-xylose		
Xulf	β-D-xylulose		

Summary

Sugars can be divided into two major subfamilies: the simple sugars (monosaccharides) and the complex sugars (oligosaccharides or polysaccharides depending on the unit of the monosaccharides). Complexity in structure arises from the elevated number of isomers that even two simple monosaccharides can generate, for instance, two glucose units can combine into 11 different disaccharides, while two identical aminoacids form only one dipeptide.

Carbohydrates, simple or complex, can be further linked to other molecules, as proteins and fats to form glycoproteins and glycolipids (known as glycoconjugates).

Sugars cover a wide range of fundamental roles, spanning from structural scaffolds to energetic intermediates and storage, they confer immunological protection, participate to cell-cell recognition processes (Fig. 1) and so on. In addition, their diversity in type and number guarantees to the organism the possibility to having a private and unique signature.
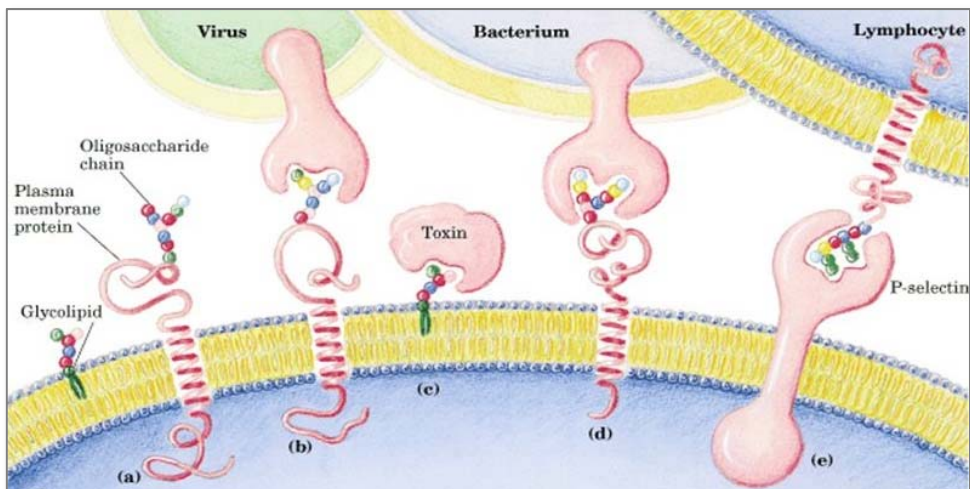
Figure 1. Schematic representation of cell-cell interaction.

In this regard, microorganisms display an extraordinary diversity that is confirmed by the continuous emergence of new glycosylation patterns: the studies of microorganisms disclose, indeed, the occurrence of atypical and rare sugars, which play distinctive roles in their interaction and adaptation with the environment, not known so far.

Therefore, the structural analysis of the glycoforms produced by the microorganisms is crucial, it is the basis to understand the biosynthetic mechanism and the possible relationship between the structure and the biological activity.

In this frame, my Ph.D. project focuses on the structural analysis of the glycosylation patterns of a special class of viruses, namely those having an autonomous glycosylation process. More in detail, I have studied the Chlorella viruses, belonging to the Giants Viruses class, deepening our understanding on the structures produced, their asset at the capsid and gaining clues on the biosynthetic process.

Chlorella viruses (family Phycodnaviridae) infect certain unicellular, eukaryotic, symbiotic chlorella-like green algae. Their importance is related to their capability to encode most, if not all, of the component required to glycosylated their major capsid protein, unlike to the other viruses which use the host biosynthetic machinery[1]. The prototype of this class of viruses is the *Paramecium bursaria* chlorella virus (PBCV-1), that infects *Chlorella variabilis* NC64A, which is a symbiont in the protozoan *Paramecium bursaria*. Previous studies established that it was constituted by a major capsid protein, called Vp54, that presents four glycosylation sites, which are unusual in several

---

[1] Van Etten JL, Gurnon JR, Yanai-Balser GM, Dunigan DD, Graves MV **2010** *Biochem Biophys Acta* 1800 (2):152-159.

aspects:[2] i) the glycans are not located in a typical Asn-X-(Thr/Ser) consensus site; ii) the glycans are attached to the protein by a β-glucose linkage; iii) the oligosaccharides are highly branched; iv) each glycan have two rhamnose residues with opposite configurations.

Therefore, the first aim of my project is to verify if the glycoform described by PBCV-1 is an isolated case or it is a structural motif shared by other Chlorella viruses.
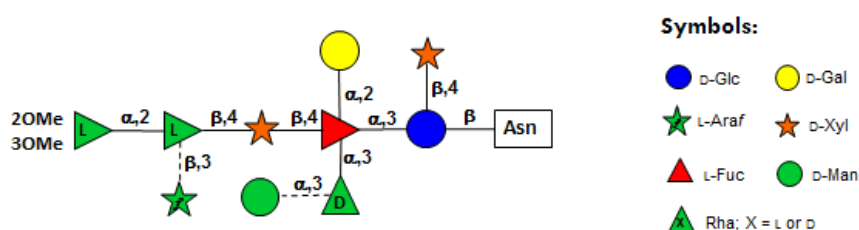
Figure 2. Vp54 structure. Dotted bonds indicate the non-stoichiometric residues[2].

For this reason, I have extended the work to the other viral isolates that belong to the same genus of PBCV-1, but also with other host specificity, such as: NY2A, which infects *Chlorella variabilis*; MT325, CVM-1 and NeJV-1, that infect *Microctinium conductrix*; ATCV-1 and TN603, which infect *Chlorella heliozoae*[3] and OSy-NE5, that infects Syngen 2-3 algae. In figure 3 the N-glycans structures from each chloroviruses, are reported.

This study discloses that the N-glycans of all *chlorella* viruses have a new structural core (Fig. 4) that do not resemble any other reported for bacteria, archea or eukarya,[4] thus it can be considered a signature for this class of

[2] De Castro C, Molinaro A, Piacente F, Gurnon JR, Sturiale L, Palmigiano A, Lanzetta R, Parrilli M, Garozzo D, Tonetti M, Van Etten JL **2013** *PNAS* 110: 13956-13960. doi: 10.1073/pnas.1313005110.
[3] Van Etten JL **2003** *Annu Rev Genet* 37: 153–195.
[4] De Castro C, Speciale I, Duncan G, Dunigan DD, Agarkova I, Lanzetta R, Sturiale L, Palmigiano A, Garozzo D, Molinaro A, Tonetti M, Van Etten JL **2016** *Angew Chem Int* Ed 55:654-65. doi: 10.1002/anie.201509150

organisms. Moreover, the oligosaccharidic core can be decorated with different monosaccharides depending on the host specificities.
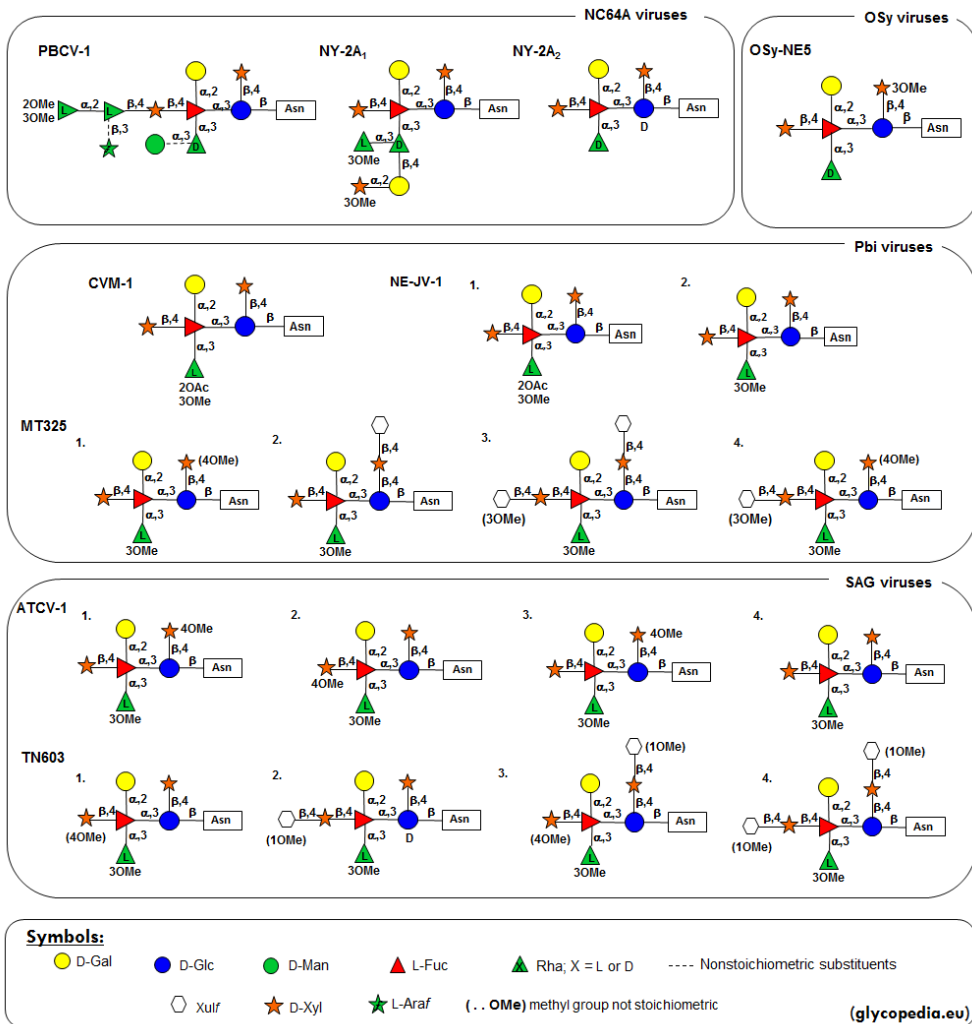
Figure 3. Representation of the *N*-glycans of all analyzed chloroviruses[4,5,6]. The *N*-glycan of the prototype, PBCV-1, is given for comparison. Different viruses are grouped according to their host specificity. All sugars are in the pyranose form except where specified.

[5] Quispe CF, Esmael A, Sonderman O, McQuinn M, Agarkova I, Battah M, Duncan GA, Dunigan DD, Smith TPL, De Castro C, Speciale I, Ma F, Van Etten JL **2016** *Virology* 500:103–113 doi: 10.1016/j.virol.2016.10.013

[6] Speciale I*, Agarkova I, Duncan GA, Van Etten JL, De Castro C. **2017** *Antonie van Leeuwenhoek* doi:10.1007/s10482-017-0861-3
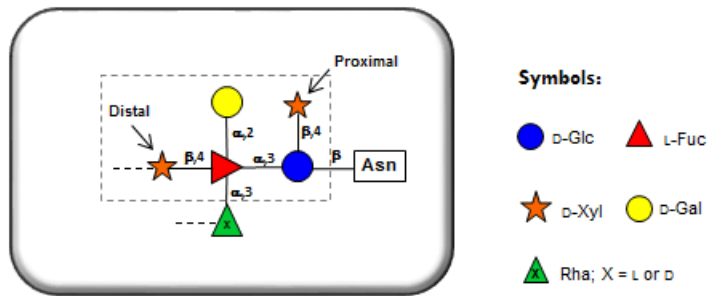
Figure 4. Oligosaccharidic core structure common to the N-glycans of major capsid proteins from eight different chloroviruses[4].

The work was extended to the spontaneous mutants, also referred to as antigenic variants, of PBCV-1 in order to carry on the structure-to-function gene analysis of the prototype virus. Recent analysis about the PBCV-1 gene, discloses that it encodes, at least, six putative glycosyltransferases. One of them, the A064R protein, has drawn our attention because 18 of 21 mutants map to the related gene (*a064r* gene). This gene consists in three domains, whose function has been disclosed and in part confirmed during this PhD.

On the basis of genetic analysis, the first domain encodes for glycosyltransferases, whereas the third domain seems to be a methyltransferase-like domain, while no clear information were deduced for the second domain.

Therefore, the second goal of my PhD project consisted in two parts. First, I have undertaken the structural studies of the antigenic variants, to understand the function of the whole A064R gene by identifying the phenotype associated to the mutants.

Antigenic variants are dived in six antigenic classes, depending on the reaction against antibodies raised against them. Variants belonging to class B (EPA and EPA2) have a domain one affected (there is a point mutation in the first domain, or it was truncated), and in both cases the N-glycans present only six monosaccharides: β-rhamnose unit is missing. On the

contrary, variant of class A and F, has this domain intact (Fig. 5) and β-rhamnose elongates the previous oligosaccharide. This lead to hypothesize that the first domain encodes for a β-(1,4)-L-rhamnosyl transferase. CME6, the only variant of class F, has both domain 1 and 2 and the glycan displays an additional α-rhamnose unit (Fig. 5). Thus, we suppose that the second domain encodes an α-(1,2)-L-rhamnosyl transferase.

Furthermore, we belief that domain 3 methylates O-2 and O-3 of α-rhamnose. In support of this hypothesis are the minor forms of the glycan from P91 and CME6. These two variants have domain 3 in its full length or truncated, respectively, and their glycans are methylated to a low extent. We believe that P91 methylation is incomplete because the methyltransferase domain does not find its ideal substrate (there is β- and not α-rhamnose). In CME6, the right substrate acceptor is present, but domain 3 of A064R is truncated and impaired in its activity.
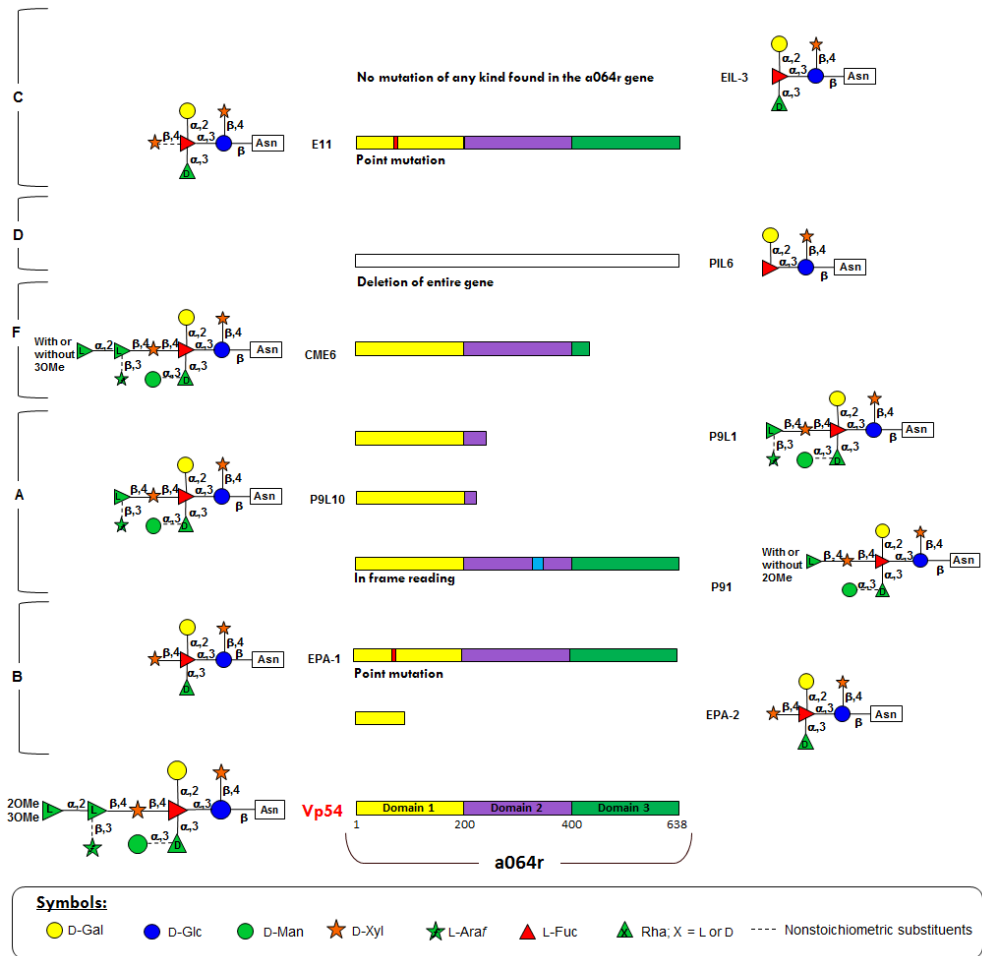
Figure 5 Schematic representation of the *N*-glycans of all antigenic variants. Different viruses are grouped according to their antigenic class (reported on the left of the figure). The *N*-glycan structure of the prototype, PBCV-1, is given for comparison. Where are reported two glycoforms, with 1 is indicated the most abundant.

Regarding the function of the first domain, its crystallographic data are available and literature reports that it encodes for a glycosyltransferase, for which the best ligand was a UDP-glucose[7]. Considering our structural data, the best ligand for this glycosyltransferase is UDP-β-L-rhamnose.

---

7 Zhang Y, Xiang Y, Van Etten JL, Rossmann MG **2007** *NIH Public Access* 15(9): 1031-1039.

Therefore, additional experiments were performed and our hypothesis proved (manuscript in preparation).

The enzyme encoded by the first domain (A064R-D1) was expressed in *E. coli* and its activity investigated by biochemical assays and by analyzing the product formed giving UDP-β-L-rhamnose as donor and a synthetic substrate resembling EPA1 glycan (hyperbranched fucose, substituted with a galactose, xylose and rhamnose residues, that presents a small lipophilic tail, at the reducing end), as acceptor (Fig. 6).
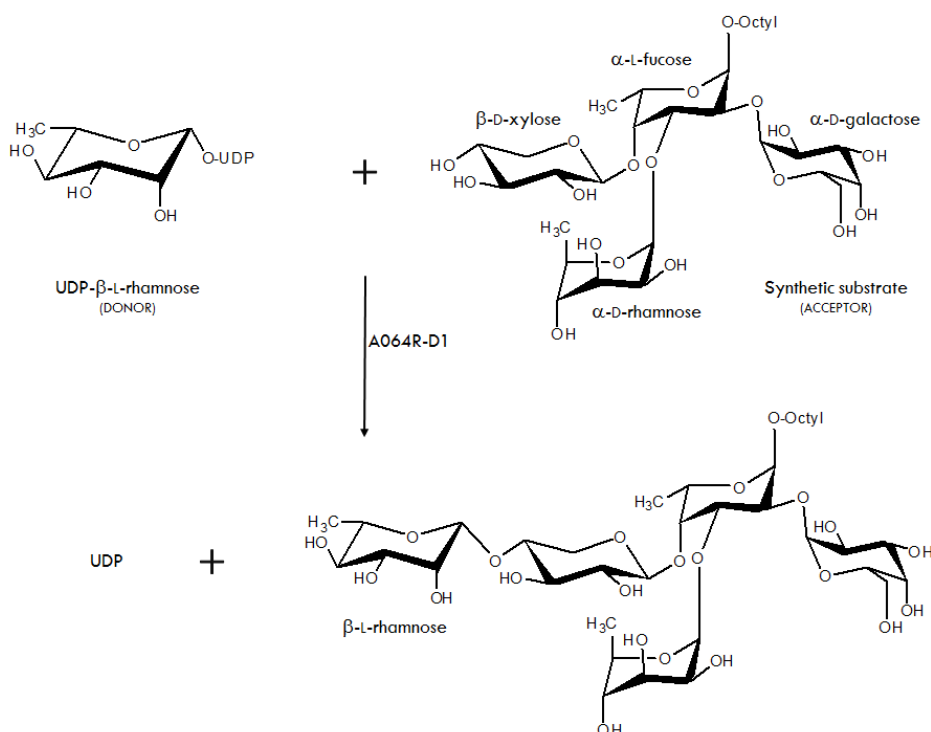
Figure 6 Schematic reaction of the β-L-rhamnosyl transferase (A064R-D1), codified by the first domain of a064r gene, that transfers the rhamnose unit form the donor to the acceptor.

The biochemical experiments were performed by using the bioluminescent assay developed from Promega (https://ita.promega.com/products/cell-signaling/glycosylation/udp_glo-glycosyltransferase-assay/) and fully confirmed our initial hypothesis about the function of the first domain of a064r, but also broadens our knowledge about it. We have proved that

this enzyme is a β-rhamnosyl transferase, manganese-dependent, and able to transfer the beta-rhamnose also onto free xylose monosaccharide (Fig. 7), indicating that its specificity for the acceptor is rather broad. In addition, we have excluded that the enzyme can work using Mg as coordinating cation and also that it does not recognizes UDP-α-Glc (Fig. 7) as instead hypothesized in the literature.
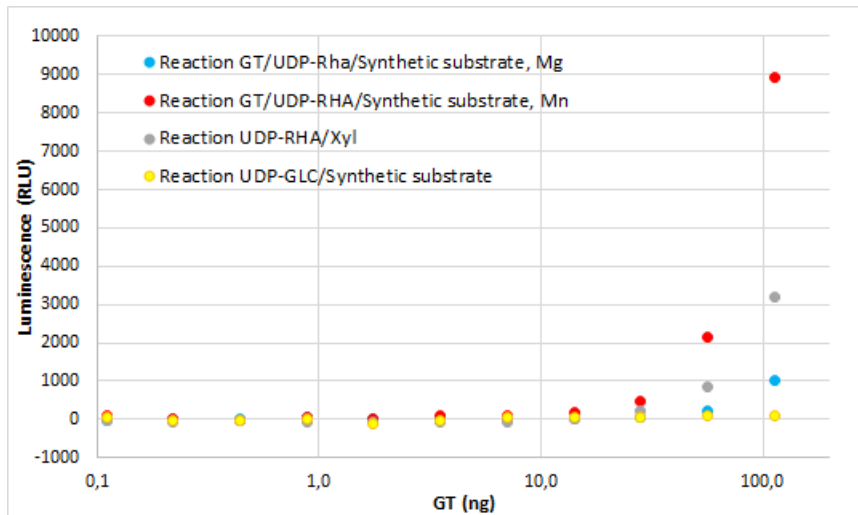
Figure 7 Detection of the glycosyltransferase activity in different conditions. Curves relative to four different reactions: in blue and in red are depicted the curves relative to the reactions between the UDP-Rha and the synthetic substrate, in presence of the enzyme and magnesium or manganese ion, respectively; in gray is indicated the reaction between the UDP-Rha and Xyl, in presence of the enzyme and both Mn^{2+} and Mg^{2+} ions; in yellow is the reaction between the UDP-Glc and the synthetic substrate, in presence of the enzyme and both ions. Of note, a high value of the luminescence indicates a large amount of the UDP produced, suggesting that the reaction occurs.

It is our idea to continue the experiments on other two domains encoded by the a064r gene (using biochemical assays and docking analysis) to address the role of the whole gene.

Another important question regards the structure of the Vp54 major capsid protein. X-ray data combined with cryo-electron microscopy disclosed that the capsid has an icosahedral shape constructed with two higher order

elements, the trysimmetron and the pentasymmetron, each in turn constituted by several capsomer unit. Each capsomer is a trimer of the capsid protein, Vp54, which consists of a 437amino acid, organized into two consecutive jelly-roll domains: D1 (residues 27-212) and D2 (residues 225-437; N-glycosylated at Asn 280, 302, 399 and 406), that are related by a 53° rotation approximately about the central threefold axis of the trimer giving to the capsomer a pseudo hexagonal symmetry.[8]

To gain insight into the function of these glycans, the original data of the Vp54 were re-examined to correct the inconsistencies reported in the first publication related to the unknown structure of the N-glycans.

Upon revising the original X-ray data by using the correct sugar templates, the overall level of information increased. Indeed, the new structure contained the first aminoacids that were originally undetected, no sugar densities related to O-linked sugar could be fitted, while, more importantly, almost all the residues of the N-glycans were placed. Information from this new structure was further implemented through a Molecular Modelling approach that fixed some faults still existing in the X-ray structure. Actually, X-ray fitting produced some residues in the wrong ring conformation, while some few, especially those located far from the polypeptide backbone, were completely absent.

The computation protocol integrated a systematic conformational search (Metropolis MonteCarlo) with Molecular Dynamic simulation that yielded in the end to determine the complete three-dimensional description of the chlorovirus PBCV-1 glycosylated major capsid protein (Fig. 8).

[8] Nandhagopal N., Simpson AA, Gurnon JR, Yan X, Baker TS, Graves MV, Van Etten JL, Rossmann MG. **2002** *PNAS* vol.99, 23, 14758-14763 DOI: 10.1073pnas.232580699.

Figure 8. Representation of correct Vp54 glycoprotein. Polypeptide backbone: in gray from 2 to 26 aa; D1 in dark green: 27 − 212; Hinge in dark blue: 213 − 224; D2 in dark red: 225 − 437. Oligosaccharidic chains: [280]Asn indicates in green; [302]Asn in gray; [399]Asn in red; [406]Asn in cyan.

This study has given a first insight into the interactions existing between the carbohydrate and the protein part (paper under revision) and are at the basis to understand which role *N*-glycans play in the capsid packaging, or at what extent stabilize the whole capsid. Preliminary studies on this last aspect were performed using again adopting the molecular modelling approach. This work is still in progress, but preliminary simulations suggest that N-glycans not are involved in the protein folding, but probably they play an important role into the capsid packaging. This idea arises by the analysis of the normal mode of the glycoprotein: when the glycans are present, the two domains are maintained close (Fig. 9), and the only permitted movement is the twisting mode; on the contrary, when the glycans are absent the two domains move away (Fig. 9). This hypothesis needs further investigation, and will be the target of future work.
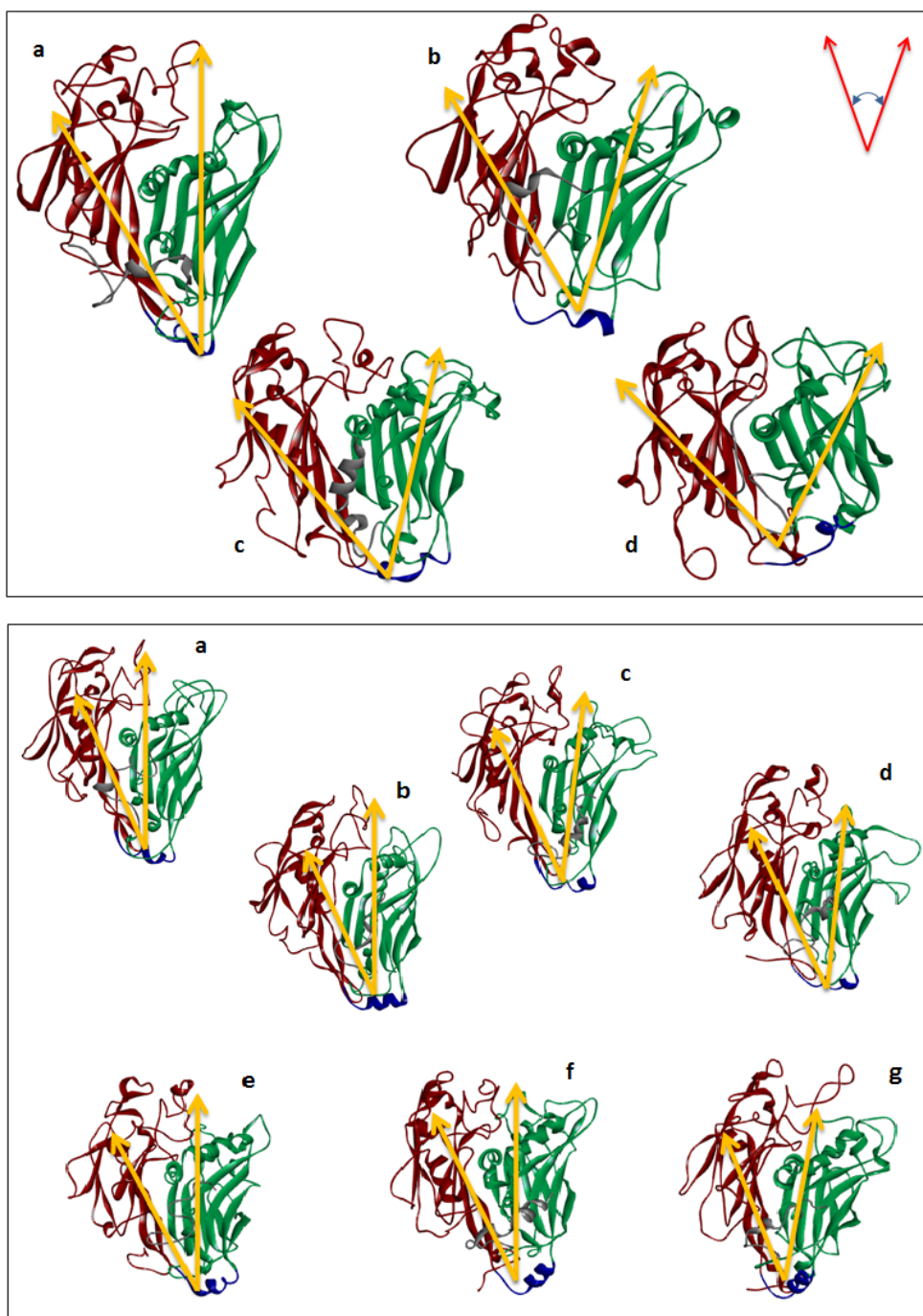
Figure 9. Structures describing the dynamic behavior of the protein with (bottom) and without (top) the glycans. The dynamic simulation was performed at 500 K, for 50 ns. The yellow arrows indicate the closure mode. Polypeptide backbone: in gray from 2 to 26 aa; D1 in dark green: 27 − 212; Hinge in dark blue: 213 − 224; D2 in dark red: 225 − 437.

Section I

Introduction

Chapter 1

Viruses

Viruses are ubiquitous members of the biosphere that are found in ecosystem on Earth (Breitbart et al. 2005; Mokili et al. 2012). They can infect all types of life forms, from animals and plants to microorganisms, including bacteria and archaea (Koonin et al. 2006).

Their functions of predation and gene transfer make viruses key drivers in the dynamics of microbial ecosystems (Suttle 2007; Mokili et al. 2012).

Viruses are supposed to be tiny even compared to cells that are just a micrometer in diameter (Fig. 1.1), indeed they cannot be seen with an optical microscope, so scanning and transmission electron microscopes are used to visualize them.
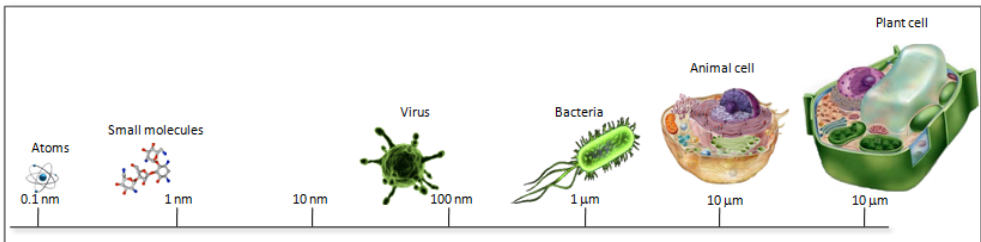
Figure 1.1 Comparative size of cells and microorganisms.

Viruses display a wide diversity in size, in the type of nucleic acid that they contain, in the structure of the capsid and in the number of protein subunits in it, in host species that infect and immunological characteristics, thus they are not usually classified into conventional taxonomic groups, but are usually grouped according to these properties (Fig. 1.2). So far, two classification systems exist: the Hierarchical virus classification system and the Baltimore Classification System.
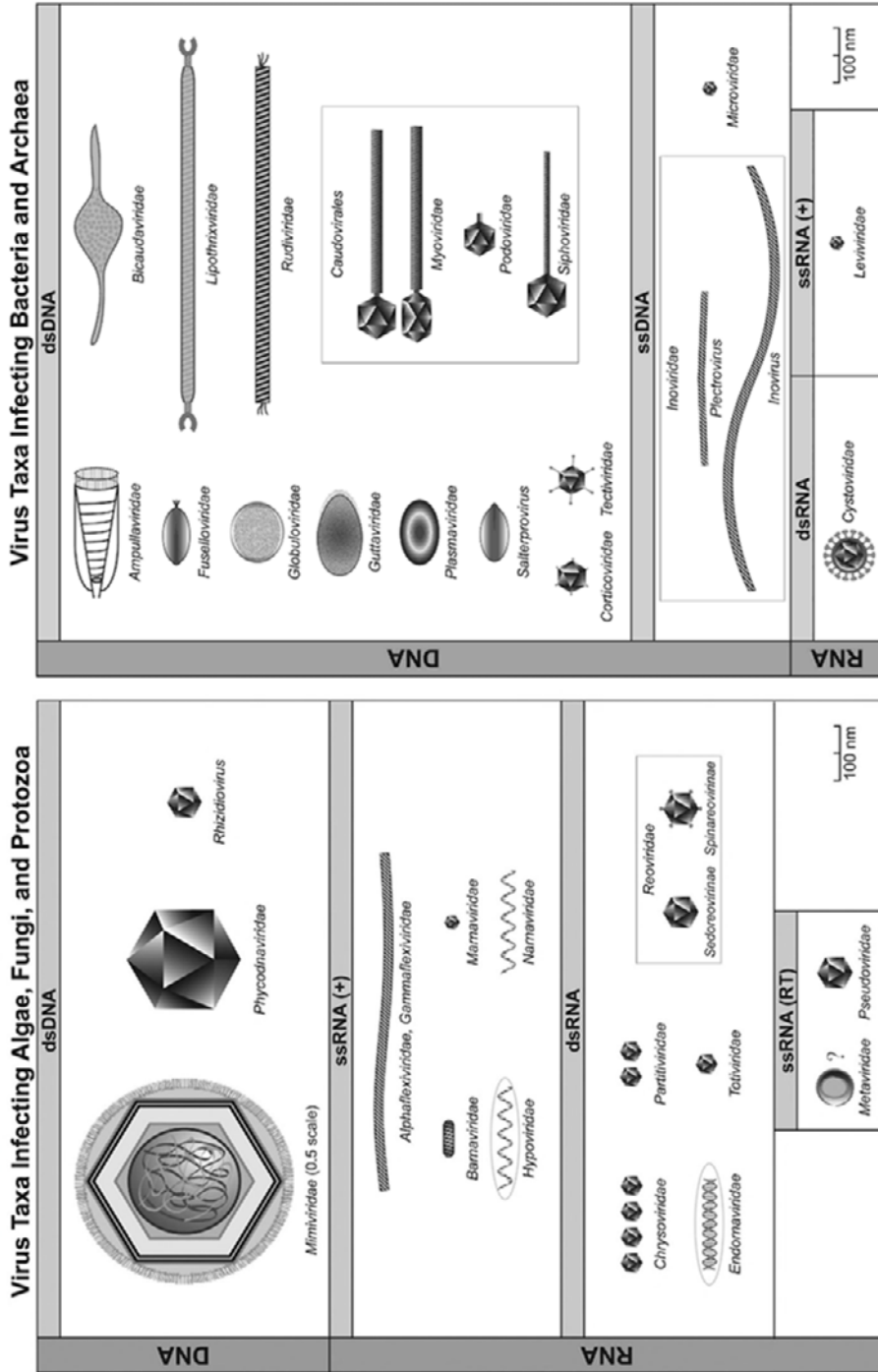
Figure 1.2 Classification of viruses.

Regarding the hierarchical classification, in 1962 Lwoff, R. W. Horne, and P. Tournier advanced a complete scheme for the classification of all viruses consisting of phylum -class -order -family -subfamily -genus -species -strain/type. This classification consider that the viruses should be grouped according to their shared properties rather than the properties of the cells or organisms that they infect. So, four main characteristics were used:

1. Nature of the nucleic acid: RNA or DNA
2. Symmetry of the capsid
3. Presence or absence of an envelope
4. Dimensions of the virion and capsid

This classification is schematized in the following figure.
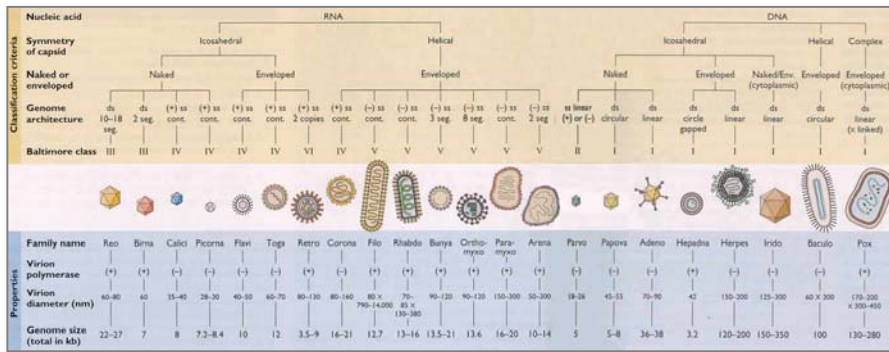
Figure 1.3 Virus classification

The Baltimore system (Baltimore 1971), instead, classified the viruses depending on the mechanism of mRNA production: they must generate mRNAs from their genomes to produce proteins and replicate themselves, using different mechanisms to achieve this, in each virus family. Viral genome may be single-stranded (ss) or double-stranded (ds), RNA or DNA, and can or cannot use reverse transcriptase (RT). Moreover, ssRNA viruses may be either sense (+) or antisense (-). Therefore, this classification organized viruses into seven groups: i) dsDNA viruses (e.g. Adenoviruses, Herpeviruses, Poxviruses); ii) ssDNA viruses (e.g. Parvoviruses); iii) dsRNA viruses (e.g. Reoviruses); iv) (+) ssRNA viruses (e.g. Togaviruses); v) (-) ssRNA viruses (e.g. Rhabdoviruses); vi) ssRNA-RT viruses (e.g. Retroviruses); vii) dsDNA-RT viruses (e.g. Hepadnaviruses).

Noteworthy, the viruses do not have their own metabolism, so they do not neither carry out their life-sustaining functions nor naturally reproduce outside a host cell (Wimmer et al. 2009). They cannot synthesize proteins, because they lack ribosomes and must use the ribosomes of their host cells to translate viral messenger RNA into viral proteins. Moreover, the viruses are not able to generate or store energy in the form of adenosine triphosphate (ATP), but have to derive their energy, and all other metabolic functions, from the host cell. For all these reasons, the typical view of viruses is of tiny burglars that sneak into cells, grab the biosynthetic controls, and compel the cell to make huge numbers of progeny that break out of the cell and keep the replication cycle going (Van Etten 2011).

1.1 Virus structure

Viruses are constituted by nucleic acid, either DNA or RNA (but not both), and a protein coat surrounded by a protective coat made of a few proteins, called **capsid** (Fig. 1.4). The capsid has three functions: i) it protects the genes from the environment, including the host cell; ii) it contains special sites on its surface that allow the virion (the infective form of the virus, as it exists outside the host cell) to attach to a host cell; iii) it serves as vehicle to enable viral entry into the host cell. The nucleic acid genome plus the protective protein coat is called the **nucleocapsid** (Fig. 1.4), which may have icosahedral, helical or complex symmetry (described in the following paragraph, 1.1.1). Proteins associates in several types of oligomeric structural subunits are known as **capsomers** (Fig. 1.4). They are closely associated and their arrangement confers the shape to the virus, that can be rod-shaped, or polygon-shaped sphere (icosahedral), or a more complicated architecture. Moreover, some viruses are also enclosed by an **envelope** of fat and protein molecules (lipoprotein bilayer) and may contain material from the membrane of a host cell in combination with their own proper. Most enveloped viruses are dependent on the envelope for their infectivity. Many viruses also develop **spikes** made of glycoprotein on their envelopes that help them to attach to specific cell surfaces.
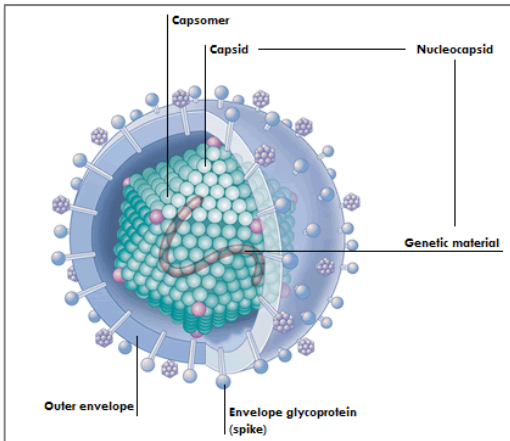
Figure 1.4 Generic representation of the virus's structure.

1.1.1 Virion Nucleocapsid structure

<u>Helical symmetry</u>

Viruses belonging into this category are composed by a single type of capsomer stacked around a central axis to form a helical structure, which may have a central cavity (Fig. 1.5a). Such nucleocapsids form short and highly rigid or elongated and flexible filaments. The genetic material, usually the single-stranded RNA (ssDNA sometimes), is bound into the protein helix by interactions between the negatively charged nucleic acid and positive charges on the protein. The length of a helical capsid is related to the length of the nucleic acid contained within it and the diameter is dependent on the size and arrangement of capsomers. The most extensively studied helical virus is tobacco mosaic virus (Fig. 1.5b).
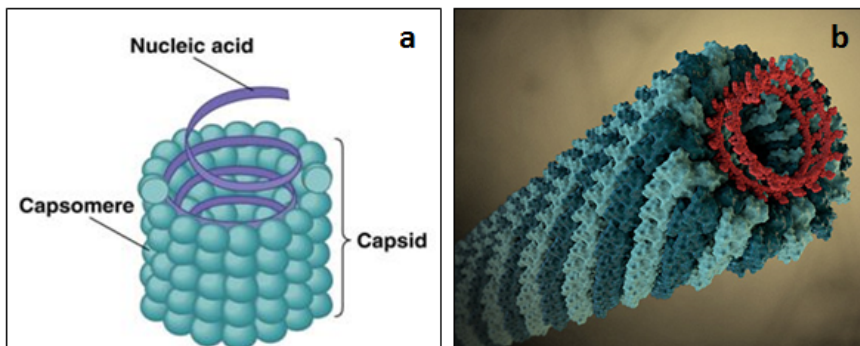
Figure 1.5 a) Helical symmetry of the nucleocapsid. b) Tobacco mosaic virus.

Icosahedral morphology (Fig. 1.6a) is characteristic of the nucleocapsids of many spherical or polyhedral viruses. An icosahedron is a polyhedron having 20 equilateral triangular faces and 12 vertices and it can have fivefold, threefold or twofold axes of rotational symmetry (Fig. 1.6b): i) lines through opposite vertices define axes of fivefold rotational symmetry; ii) the threefold symmetry axes are formed by lines through the center of opposite triangular; iii) lines through midpoints of opposite edges form the twofold axes.

A regular icosahedron is the best way of forming a closed shell from identical sub-units. The minimum number of identical capsomers required is twelve, each composed of five identical sub-units. Many viruses (such as rotavirus) have more than twelve capsomers and appear spherical, but they retain this symmetry. Capsomers at the apices are surrounded by five other capsomers and are called **pentons**, instead capsomers on the triangular faces are surrounded by six others (they are called **hexons**). Although the icosahedral structure is extremely common among viruses, size differences and slight variations exist between virions. Prolate is an example of icosahedron elongated along the fivefold axis and it is a common arrangement of the heads of bacteriophages (Fig. 1.6c). This structure is composed of a cylinder with a cap at either end.
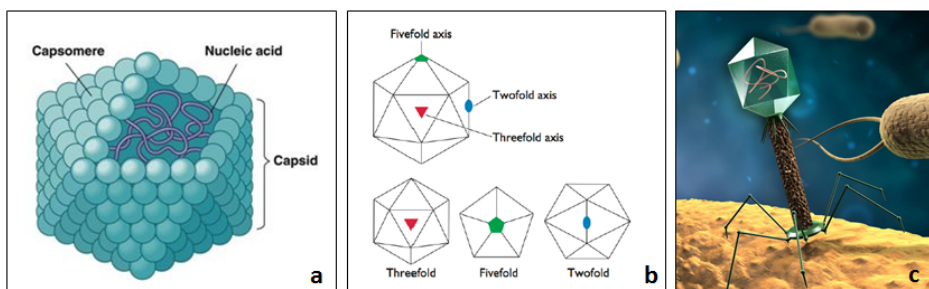
Figure 1.6 a) Icosahedral symmetry of the nucleocapsid. b) Representation of threefold, fivefold and twofold axes of rotational symmetry in icosahedral polyhedron. c) Example of icosahedral symmetry: Bacteriophage's head.

Complex symmetry

In this case, the viruses possess a capsid that is neither purely helical nor purely icosahedral, and that may possess extra structures such as protein tails or a complex outer wall. Some viruses that infect Archaea have complex structures that

are unrelated to any other form of virus, with a wide variety of unusual shapes, ranging from spindle-shaped structures, to viruses that resemble hooked rods, teardrops or even bottles. Another example of complex symmetry is poxvirus (Fig. 1.7).
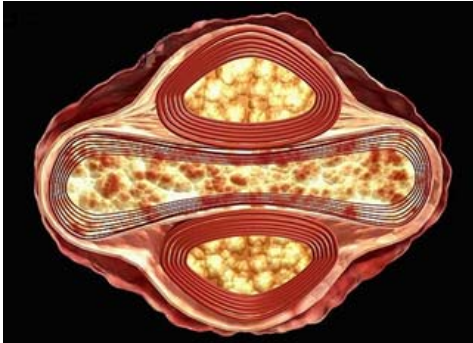
Figure 1.7 Poxvirus, example of a complex symmetry of nucleocapsid.

1.2 Virus Replication cycle

A virus must use at least part of the cell's equipment for replication of viral nucleic acids and expression of viral genes. They also use the cell's protein synthetic machinery, and the cell's metabolic energy resources (Wagner et al. 2007).

The viral replication induces dramatic biochemical and structural changes in the host cell, which may cause cell damage: it can destroy the cell or cause changes of the cell functions, preventing the normal function, even though the cells remain alive for a period time.

Virus specifically interacts with a precise protein on the host cell membrane (receptors) through proteins in the capsid or via glycoproteins embedded in the viral envelope (**attachment step**). The specificity of this interaction determines the host (and the cells within the host) that can be infected by a specific virus. Once bound to the cell membrane, the virus injects its genetic material into cell (**penetration step**) (Fig. 1.8a). The injection can occur via endocytosis (the host cell transports viral particle into himself by engulfing it like it would a food particle), membrane fusion (viruses attach the surface of the host cell and then the viral envelope fuses with the host cell membrane, emptying the now-bare virus into the cell) or penetration (attaching to the surface of the cell via receptors and injecting

only its genome into the cell, leaving the rest of the virus on the surface). Then, cell enzymes strip off the virus protein coat to release and render accessible the virus nucleic acid (**uncoating)** to the next step, the **replication** (Fig 1.8c) of the viral particle.
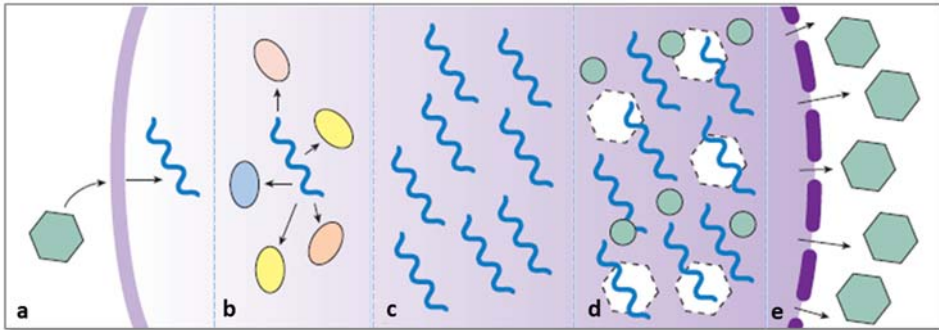
Figure 1.8 Virus replication cycle. a) Initial recognition between virus and cell and introduction of viral genetic material into the host cell; b) virus gene expression and induction of virus-induced modification of host; c) virus genome replication; d) expression of virus-associated proteins; e) assembly and release of new viruses, often resulting in cell death.

Viruses that contain DNA as genetic material and utilize the infected cell's nucleus as the site of genome replication, share many common patterns of gene expression and genome replication along with similar processes occurring in the host cell. Instead, RNA viruses have devised some way to replicate such material, since the cell does not have machinery for RNA-directed RNA replication. The replication of RNA viruses requires expression of specific enzymes that are not present in the uninfected host cell.

Viral replication proteins are enzymes involved both in nucleic acid replication and in the expression and regulation of viral genetic information. Viruses also encode enzymes and proteins involved in modifying the cell in which the virus replicates, in order to optimize the cell for virus replication. The replication occurs in the host's cell cytoplasm for single strain RNA viruses, instead double-stranded DNA viruses typically must enter the host cell's nucleus before they can replicate.

Once a virus infects its host and the viral progeny components are produced by the host's cellular machinery, the **assembly** of the virus particles happen. This is a non-enzymatic process, it is usually spontaneous. The assembly may take place in the cell's nucleus, cytoplasm, or at plasma membrane for most developed viruses.

The viruses, now are released from the host cell (**release step**) (Fig 1.8e), either by sudden rupture of the cell, waiting for the cell to die, or by gradual budding off through the cell membrane. They may invade or attack other cells, or remain dormant in the cell. A schematic replication cycle is reported in the figure 1.8.

Within this general pattern there is a wealth of variation and difference in detail. For example, a plant cell presents a unique challenge for virus entry because there is a rigid and fairly thick cell wall. Indeed, the entry into the plant cell must take advantage of some break in integrity of the cell wall. Seemingly, when the virus enters such a break and it is located in close proximity to the plant cell's plasma membrane, it does not need to interact with specific receptors. Breaks or lesions in the plant cell's wall are often produced by organisms that feed on the plant or by mechanical means. Once inside the plant cell cytoplasm, viruses are uncoated and gene expression begins following patterns similar to those described above.

1.3 N-Glycosylation

Many viruses, such as rhabdoviruses, herpesviruses, poxviruses, and paramyxoviruses, present structural proteins glycosylated. Viral surface glycoproteins are necessary for the virus because mediate its entry into host cells through receptor binding at the cell surface and may help viruses to avoid the host immune system.

Typically, viruses 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 either co-translationally or shortly after translation of the protein. Therefore, the only way to alter glycosylation of virus proteins is to either grow the virus in a different host or have a mutation in the virus protein that alters the protein glycosylation site. This happens because viruses lack genes encoding glycosyltransferases, generally.

Moreover, the majority of the virus glycoproteins, presents N-glycosylation to Asn via N-acetylglucosamine, albeit there are some viruses that present O-linked sugars attached to Ser or Thr residues via an amino sugar (acetylglucosamine or acetylgalactosamine).

In eukaryote, the N-glycan is not made on the protein, but pre-manufactured on an ER-resident lipid, and then is 'en bloc' transferred to the protein (Bieberich 2014). Hence, the glycosylation pathway begins at the cytoplasmic face of the ER membrane with the assembly of the initial sugar precursor linked to the lipid dolichol pyrophosphate (Dol-PP) (Fig. 1.9). Two GlcNAc are the first residues incorporated into dolichol-pyrophosphate. The activated monosaccharide donor is synthesized by the GlcNAc phosphotransferase, which transfers the GlcNAc-1-phosphate from UDP-GlcNAc (Elbein 1987). The oligosaccharide is further assembled, step by step, by specific glycosyltransferases that add other monosaccharides. These reactions depend on activated sugars that are provided on the cytosolic and luminal side of the ER in form of UDP-GlcNAc, GDP-mannose (GDP-Man), and UDP-glucose (UDP-Glc).

Following assembly and the flipping reaction of the precursor oligosaccharide, co-translational 'en bloc' transfer of the N-glycan from Dol-PP onto the nascent polypeptide is the next critical step in N-glycoprotein biosynthesis (Fig. 1.9). This step is catalyzed by oligosaccharyltransferase (OST), a multimeric enzyme complex (Roth et al. 2010). The precise function of these subunits is still a subject of ongoing research, however, it is clear that OST will have to manage the association with the ribosome and signal recognition particle receptor for the nascent membrane glycoprotein, the recognition of the acceptor sequence Asn-X-Ser/Thr (X cannot be proline), and the catalytic transfer of the oligosaccharide from Dol-PP onto asparagine (Bieberich 2014). Upon transferring the completed glycan onto the nascent polypeptide, the enzymes known as glycosidases remove glucose residues from the structure. This initial trimming step is thought to act as a quality control step in the ER to monitor protein folding and signals that the glycoprotein is ready for transit from the ER to the Golgi (Drickamer and Taylor 2006). However, if the protein is not folded properly, the glucose residues are not removed and thus the glycoprotein can't leave the endoplasmic reticulum. A chaperone protein (calnexin/calreticulin) binds to the unfolded or partially folded protein to assist protein folding.

The processed glycoprotein, therefore, moves to the Golgi where it is modified by hydrolytic removal of sugar residues (trimming) followed by re-glycosylation with

additional sugar residues (processing), such as galactose, fucose or sialic acid in complex N-glycoproteins (Fig. 1.9). The order of addition of sugars to the growing glycan chains is determined by the substrate specificities of the enzymes and their access to the substrate. Thus, the organization of this machinery within a cell plays an important role in determining which glycans are made. Consequently, the glycan portion of virus glycoproteins is host specific.
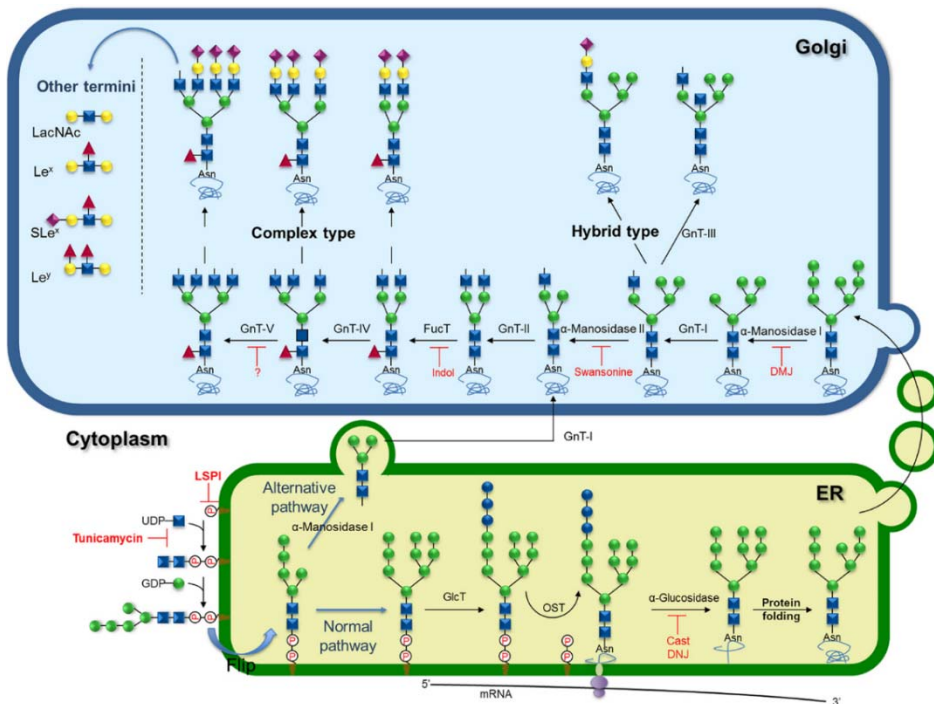
Figure 1.9 Schematic representation of biosynthesis and processing of N-linked oligosaccharides.

1.3.1 N-glycans in three domains of the life

Asparagine (N)-linked glycosylation of proteins is a fundamental and extensive post-translational modification that results in the covalent attachment of an oligosaccharide onto asparagine residues of polypeptide chains (Schwarz and Markus 2011). This protein modification occurs across the three domains of life: Bacteria, Archaea and Eukarya (Fig. 1.10). Several studies disclose that archaea exhibit the greatest diversity in glycosylation pathway compared to bacteria.

11

Indeed, Archaea use a large pool of building blocks glycans, they can synthesize more than one type of LLO (Lipid-linked oligosaccharide) in the same cell and they produce both dolichyl phosphate-linked and pyrophosphate-linked (Schwarz and Markus 2011).

By contrast, bacteria display a limited diversity and N-glycosylation is confined to a small number of species. Instead, eukaryotes share a common oligosaccharidic core sequence, a Man_5-$GlcNAc_2$ (Burda and Aebi 1999), that can be classified into three types: oligomannose, complex and hybrid depending on the mannose decoration of the conserved core.

The diversity of N-glycans found among eukaryotes is a consequence of successive elaboration taking place in the late ER and in the Golgi. Differently, in prokaryotes, the proteins are modified in the periplasm and maintained outside.
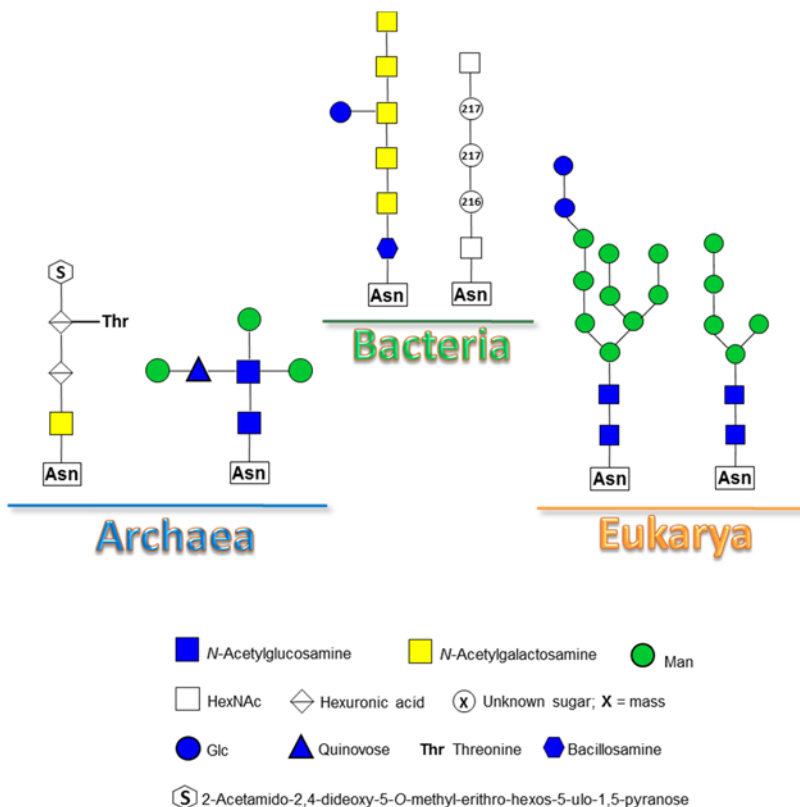
Figure 1.10 N-glycans structure produced in different domains of the life.

Noteworthy, several studies reveal that the acceptor substrate of N-glycosylation is the same for the three domains of life: an asparagine residue present within the consensus motif Asn-X-Ser/Thr (N-X-S/T), in which X is any amino acid except proline (Hart 1992). This sequence is commonly recognized by eukaryotic and prokaryotic enzymes involved in N-glycosylation.

Chapter 2

Giant viruses: Nucleo-Cytoplasmic Large DNA Viruses

Giant viruses (also called **giruses**) differ from the viruses in many of the above described features. In particular, they have genomes larger than 300 kilobase pairs and have a capsid diameter of about 200 nanometers or more (Van Etten 2011).

Giant viruses have been discovered and characterized in the last few years except to bacteriophage G, initially described about 40 years ago (Donelli et al. 1975), and the chlorella virus *Paramecium bursaria chlorella* virus 1 (PBCV-1), which was first described in 1982 (Van Etten et al. 1982).

One of the largest viral divisions, that seem to be monophyletic, includes nine recognized families, to date, collectively known as Nucleo-Cytoplasmic Large DNA Viruses (**NCLDV**; table 1) (Iyer et al. 2001): Phycodnaviridae, Poxviridae, Asfarviridae, Iridoviridae, Ascoviridae, Mimiviridae, Marseilleviridae, Pandoraviridae, Pithoviridae.

Table 1 Classification of the Nucleo-Cytoplasmic Large DNA Viruses (**NCLDVs**) and their characteristics. [a]Mimiviridae are organized in three groups based on their phylogenetic affinity (Yoosulf et al. 2012; 2014). [b]Marseilleviridae (Doutre et al. 2014). [c]Pandoraviridae (Antwerpen et al. 2015). [d]Pithoviridae (Legendre et al. 2014).

VIRUS FAMILY	HOSTS	GENOME MAX SIZE, KB
Phycodnaviridae	Green algae; algal symbionts of paramecia and hydras	400
Poxviridae	Animals: insects, reptiles, birds, mammals	365
Asfarviridae	Mammals	170
Ascoviridae	Invertebrates and non-mammalian vertebrates: Insects, mainly noctuids	190
Iridoviridae	Invertebrates and non-mammalian vertebrates: Insects, cold-blooded vertebrates	220
Mimiviridae[A]	Acanthamoeba; algae (probably); corals (possibly)	1280
Marseilleviridae[B]	Acanthamoeba; algae (probably)	370
Pandoraviridae[C]	Amoebae	2770
Pithoviridae[D]	Amoebae	610

The NCLDV infect animals and diverse unicellular eukaryotes, and either replicate exclusively in the cytoplasm of the host cells, or possess both cytoplasmic and nuclear stages in their life cycle.

These large viruses have a genome greater than 300 kb and up to 1300 kb (Fig. 1.11), that can encode up to 900 proteins and many tRNAs. Mimiviridae, which so far includes two closely related giant viruses isolated from Acanthamoeba – Mimivirus and Mamavirus, together with Megavirus, present the largest genome size reported (slightly larger than 1 megabase); it exceeds numerous parasitic bacteria, and approach the genome size of the simplest prokaryotes. Albeit a recent reconstruction of NCLDV evolution suggests a common ancestor, they can differ significantly in morphology, lifestyle, and genome structure.
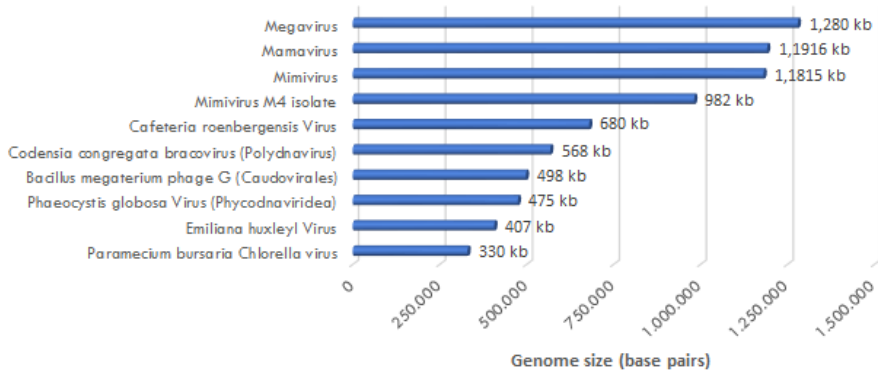
Figure 1.11 Genome size of some NCLD viruses.

Most of the NCLDV do not strongly depend on the host replication or transcription systems for completing their replication (Van Etten 2003). This relative independence of the virus reproduction from the host cell functions (apart from translation, of course) is supported by a set of conserved proteins that are encoded in the viral genomes and mediate most of the processes essential for viral reproduction. These essential, conserved proteins include DNA polymerases, helicases, and primases responsible for DNA replication, RNA polymerase subunits, chaperones involved in capsid assembly and capsid proteins themselves (Iyer et al. 2001, 2006; Yutin et al. 2010).

2.1 Phycodnaviridae family

In this thesis, the attention is focused on one family of Nucleocytoplasmic large DNA viruses: Phycodnaviridae family.

The family Phycodnaviridae consists of a morphologically similar but genetically diverse group of large (100-560 kb) double stranded DNA viruses which infect both fresh and marine water eukaryotic algae (Dunigan et al. 2006; Wilson et al. 2009). They have an internal membrane that is required for infection.

Phylogenetic analyses of the DNA polymerases of the phycodnaviruses (Fig 1.12) indicate that these viruses are more closely related to each other than to other dsDNA viruses and that they form a monophyletic group, suggesting a common ancestor (Wilson et al. 2005).
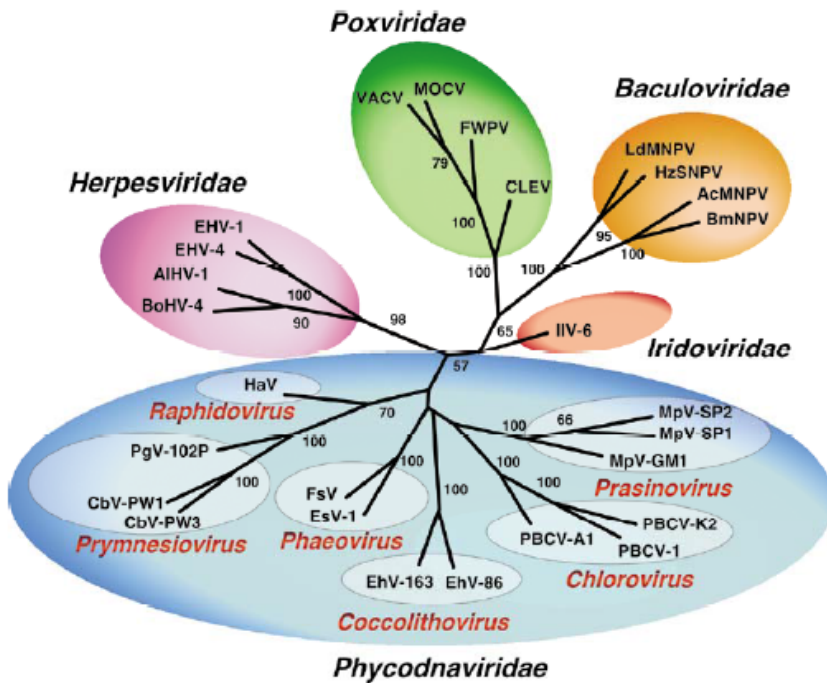
Figure 1.12 Phylogenetic tree of DNA polymerase gene fragments from members of the family *Phycodnaviridae* and other large dsDNA viruses.

Currently, there are 33 species in the Phycodnaviridae family, grouped into six genera, initially based on host range and then supported by sequence comparison of their DNA polymerases: Chlorovirus, Coccolithovirus, Phaeovirus, Prasinovirus, Prymnesiovirus, Raphidovirus. Members of the genus Chlorovirus infect chlorellalike green algae from terrestrial waters, whereas members of the other five genera infect marine green and brown algae. Phycodnaviridae play important roles in the dynamics of algal blooms, nutrient cycling, algal community structure, and possibly gene transfer between organisms.

Three viruses (*Paramecium bursaria* Chlorella virus, PBCV-1; *Emiliana huxleyi* virus, EhV; *Ectocarpus siliculosus* virus, EsV) were chosen as representative of this family. The genomes of these three viruses have only 14 genes in common (Van Etten et al. 2010), which is particularly surprising since they are all from the same virus family. This is the evidence of their long evolutionary history, indeed, over 1000 different genes exist just among these three phycodnaviruses.

Originally, all phycodnavirus virions were assumed to be large icosahedral structures (120–220 nm in diameter) with a multilaminate shell surrounding an electron dense core and lacking an external membrane. However, recent experiments indicate that not all virus structures are identical (Wilson et al. 2005). Experiments revealed that some viruses can present cylindrical spike or tail structure (as shown for PBCV-1 and EhV-86) or may have an external membrane surrounding the polyhedral capsule.

This project was focused on phycodnaviruses belonging to the genus Chlorovirus.

2.2 Chlorella Viruses (CV)

Chlorella viruses, referred to as chloroviruses (CV), are large, icosahedral, plaque forming, dsDNA-containing viruses (Yamada et al. 2006; Van Etten and Dunigan 2012, 2016), that have large genomes (ranging from 160 to 560 kb, with up to 600 protein-encoding genes; Van Etten and Dunigan 2012), when compared with viruses pathogenic for humans. They infect certain unicellular, eukaryotic, chlorella-like green algae (Van Etten and Dunigan 2016; Dunigan et al. 2006), also known as Zoochlorellae, which can establish a symbiotic relationship with the protozoan *Paramecium bursaria*, the coelenterate *Hydra viridis*, and the heliozoan *Acanthocystis turfacea* (Van Etten and Dunigan 2012, 2016). Chloroviruses infect zoochlorella isolates grown in the laboratory, but there is no evidence that these symbiotic algae grow independently of their hosts in indigenous waters. Very little is known about the role chloroviruses play in freshwater ecology (Short et al. 2011), probably they play dynamic, albeit largely undocumented, roles in regulating microbial communities in the ecosystem. Interestingly, Zoochlorellae are resistant to viruses in their symbiotic state.

Algae included in the genus *Chlorella* are small spherical or ellipsoidal, unicellular, nonmotile, asexually reproducing green algae and are among the most widely distributed and frequently encountered algae on earth (Fig. 1.13a; Oh-Hama and Miyachi 1988; Shihra and Krauss 1965). Chlorella species have a rigid cell wall and typically contain a single chloroplast, which sometimes contains a pyrenoid

body (a microcompartment associates with the carbon – concentrating mechanism) (Fig.1.13b).
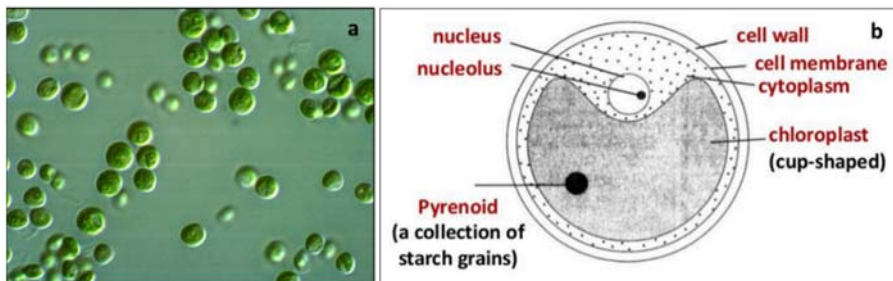
Figure 1.13 a) Chlorella green algae; b) Clorella structure.

Viruses infecting zoochlorellae are classified into four groups depending on their hosts: NC64A viruses that infect *Chlorella variabilis* strain NC64A; Pbi viruses specific for *Chlorella Pbi*; SAG viruses that infect *Chlorella heliozoae* strain 3.83; and the most recently discovered OSy viruses that infect a different strain of *Chlorella variabilis*, Syngen 2-3 strain (Quispe et al. 2016).

Forty-one chlorovirus genomes, representing NC64A, SAG and Pbi viruses, have been sequenced and phylogenetic relationships were determined using 32 concatenated proteins encoded by all viruses (Jeanniard et al. 2013). This analysis disclosed that viruses infect the same algal host cluster into monophyletic clades (Fig. 1.14). It means that viruses which infect *Chlorella* NC64A neither infect nor attach to *Chlorella* Pbi, and vice versa.
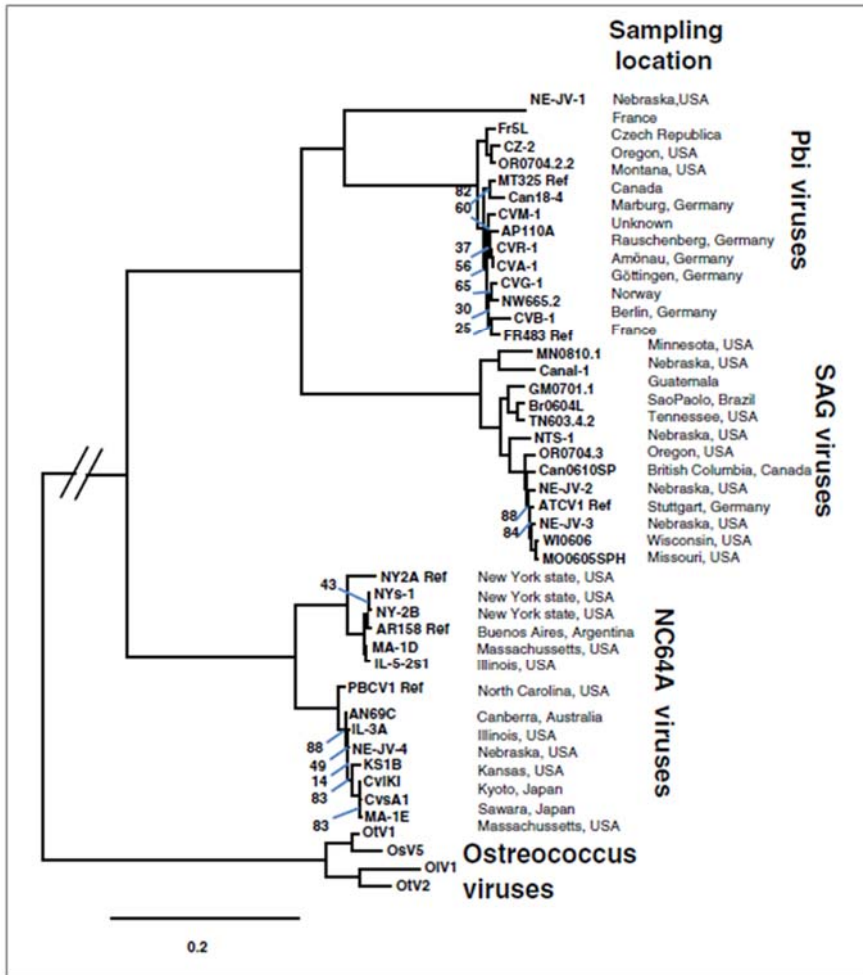
Figure 1.14 Phylogenies of chloroviruses and algal hosts. Monophyletic tree of chloroviruses based on a concatenated alignment of 32 core protein families.

Chloroviruses often encode enzymes involved in sugar metabolism (Van Etten et al. 2010). For example, two PBCV-1 encoded enzymes synthesize either GDP-L-fucose or GDP-D-rhamnose from GDP-D-mannose (Fruscione et al. 2008; Tonetti et al. 2003). In addition, many chloroviruses encode enzymes involved in extracellular polysaccharide biosynthesis, suggesting that the polysaccharides, which require a huge amount of ATP for their synthesis, are important in the viral life cycles (Van Etten and Dunigan 2012). However, this function is unknown.

They also encode functional chitinases and chitosanases (Sun et al. 1999). The discovery that members of the Chlorella genus have chitin in their walls (Takada et

al. 1991; Kapaun et al. 1992) was surprising because chitin is normally absent in green algae.

The chlorella virus genomes contain methylated bases in specific DNA sequence; this led to the discovery that chloroviruses encode multiple 5-methylcytosine (5mC) and N^6-methyladenine (6 mA) DNA methyltransferases (Van Etten et al. 1991). The purpose of the other virus-encoded DNA methyltransferases is unknown.

One further important property of these viruses is that they encode most, if not all, the components required to glycosylate their major capsid proteins (Wang et al. 1993; Van Etten et al. 2010) their viruses infecting eukaryotic organisms, that use the host biosynthetic machinery (Doms et al. 1993; Olofsson and Hansen 1998; Vigerust and Shephered 2007). Chloroviruses, not only, are able to encode glycosyltransferases, but several experimental results indicate that their glycosylation occurs independent of the endoplasmic reticulum and Golgi (Van Etten et al. 2010).

All these characteristics make the chloroviruses an unusual micro-organism, for which many characteristics of the 'normal' viruses are not verified, and this is why they are subject of several studies, including this PhD thesis for all the aspect concerning glycosylation.

Some representatives of each group of Chloroviruses were examined in this thesis: PBCV-1 and NY-2A (NC64A viruses); OSy-NE5 (OSy viruses); MT325, CVM-1 and NE-JV-1 (Pbi viruses); ATCV-I and TN603 (SAG viuses).

2.3 *Paramecium bursaria* chlorella virus (PBCV-1)

Paramecium bursaria chlorella virus (PBCV-1) is the type member of the genus Chlorovirus (*Phycodnaviridae* family) (Van Etten et al. 1991), which infects Chlorella variabilis, a symbiont of the ciliated protozoan *Paramecium bursaria* (Fig. 1.15a). PBCV-1 contains about 64% protein, 21 to 25% dsDNA, and 5 to 10% lipid (Van Etten et al. 1983).

Cryo-electron microscopy (cryoEM) and three-dimensional reconstruction (26 Å resolution; Fig. 1.15b, c) indicate that PBCV-1 has a multilayer structure composed of a double strand DNA genome and an inner protein core, a single lipid bilayer membrane, surrounded by a large (190 nm in diameter) outer icosahedral capsid shell (Yan et al. 2000; Nandhagopal et al. 2002) (schematic structure of PBCV-1 in figure 1.15d). Studies have revealed that the lipid membrane is required for the infection (Yan et al. 2000; Zhang et al. 2011).

The virion protein, named Vp54, forms an outer capsid protein, and accounts for ~40 % of the virion's protein mass (Skrdla et al. 1984; Yan et al. 2000). In addition, at least three other proteins construct the capsid, including other two glycoproteins Vp280 and Vp260, each accounting for ~1% of the mass of protein per virion (Que et al. 1994).
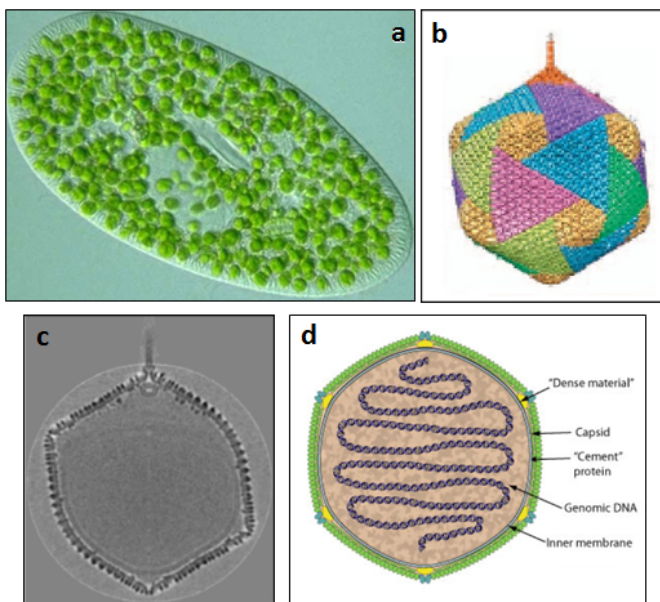
Figure 1.15 a) *Paramecium bursaria* and its symbiotic chlorella cells. b) Fivefold averaged cryo-electron micrograph (cryoEM) of PBCV-1. c) Central cross-section of the cryoEM density. d) Schematic structure of PBCV-1 virion.

Furthermore, PBCV-1 is equipped with a few external fibers (Fig. 1.16a), extending from some of the capsomers (Cherrier et al. 2009), which presumably facilitate the attachment to the host (Fig. 1.16b). In addition, one of the PBCV-1 vertices has a spike (560 Å long spike-structure; Fig. 1.16c,d), that probably aids in penetration of the wall (Van Etten and Dunigan 2012).

The spike protrudes from the surface of the virus (for 340 Å) and has an external diameter of 35 Å at the tip, and 70 Å at the base (Fig. 1.16c), and it widens up to 160 Å inside the capsid to form a closed cavity inside a large pocket between the capsid and the membrane enclosing the virus DNA (Figure 1.16d). Therefore, the internal virus membrane loses its icosahedral symmetry adjacent to the unique vertex. Consequently, the virus DNA located inside the envelope is packaged non-uniformly in the particle.
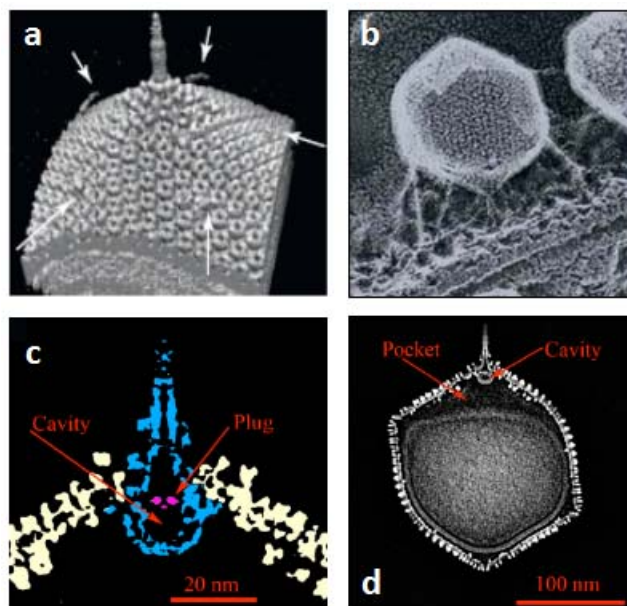
Figure 1.16 a) Surface view of the PBCV-1 spike structure and fibers (indicate with arrows). b) PBCV-1 attached to the cell wall as viewed by the quick-freeze, deep etch procedure. Note, the virions attached to the wall by fibers. c) Image of the spike structure. d) Another view of the central cross-section of cryoEM structure of PBCV-1. Note the gap between the unique vertex and the membrane enclosing the DNA.

2.3.1 Vp54 structure

The PBCV-1 major capsid protein Vp54 is a glycoprotein and represents ca. 40 % of the virion's protein mass (Skrdla et al. 1984). The cryo-EM reconstruction of PBCV-1 reveals that the external capsid is assembled from 20 triangular units (trisymmetrons; Wrigley 1969, 1970) and 12 pentagonal caps (pentasymmetrons) at the fivefold vertices. The trisymmetrons and pentasymmetrons are pseudohexagonal arrays of 66 and 30 trimeric capsomers, respectively (Fig. 1.17a), in addition, the structure of PBCV-1 capsid protein was crystallized at 2.5 Å and 2.0 Å resolution, and deposited in Protein data bank (PDB ID codes 1J5Q and 1M3Y, respectively) (Nandhagopal et al. 2002). These structures were fitted with the previous cry-EM map (Fig. 1.17b), and revealed that each capsomer is composed of three molecules of the major capsid protein Vp54 (Fig. 1.17c). This result was consistent with the previous studies, which reported that the outer capsid protein of PBCV-1 contains 5,040 copies of Vp54 organized into 1,680 trimers (Caspar and Klug 1962).
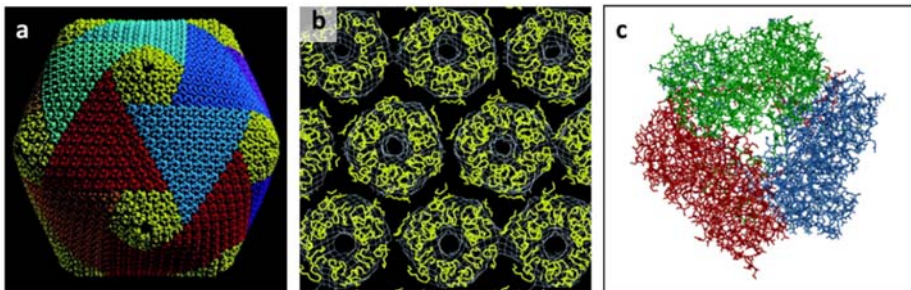
Figure 1.17 a) Quasi-atomic model of the PBCV-1 capsid based on fitting the crystal structure of the Vp54 trimer into the cryo-EM reconstruction. The pentasymmetrons are colored yellow to differentiate them from the variously colored trisymmetrons. b) PBCV Vp54 trimers (Cα backbone in yellow) near the center of a trisymmetron fitted into the cryo-EM map (white) viewed from outside the virus. c) Schematic representation of the one capsomer.

Each Vp54 monomer consists of a 437 amino acids, which were identified in Xray electron density map. The monomer fold consists of two consecutive jelly-roll domains: **D1** (residues 27-212) and **D2** (residues 225-437); this last contains four N-linked oligosaccharides sites at position Asn 280, 302, 399 and 406. The two

domains, D1 and D2, in Vp54 are related by a 53° rotation approximately about the central threefold axis of the trimer giving the capsomer a pseudo hexagonal symmetry (Nandhagopal et al. 2002). The N-linked carbohydrates point outward (Fig.1.18), on the contrary, the inner side of the capsomer is rich in basic amino acids and it is complementary to the phospholipid layer.
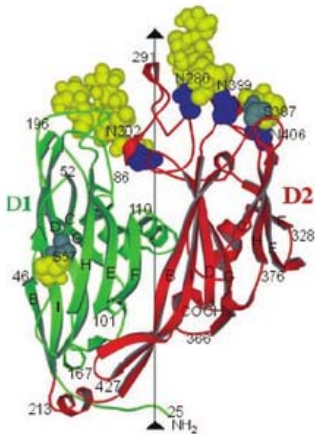
Figure 1.18 A ribbon diagram of the Vp54 monomer. The carbohydrate moieties (yellow) and glycosylated Asn residues (blue) are shown. The D1 (27 to 212 residue) is coloured in green, the D2 (225-437) in red.

Comparison of the molecular mass of Vp54 (53,790 Da) with its predicted molecular weight given from the amino acid composition (48,165 Da) indicates that the protein contains about 30 sugar moieties, of which 20 could be detected, but not identified, in the crystal density map (Nandhagopal et al. 2002).

Regarding the four N-linked glycans, they were determined using the spectroscopic approach, and consist of eight to ten neutral monosaccharide residues, for a total of four glycoforms (Fig. 1.19) (De Castro et al. 2013), and do not resemble anything in the three domains of the life, to date (paragraph 1.3.1). These N-glycans have a highly-branched architecture with a hyperbranched fucose, and a dimethylated rhamnose as a capping residue, along with two residues, arabinose and mannose, as nonstoichiometric substituents. Moreover, the rhamnose units do not possess the same absolute configuration (one is D and two are L configurated), a feature that is rare and occurring only in bacteria. Interestingly, the glycans are attached to the protein by a β-glucose linkage and not N-acetilglucosamine, typical of N-glycoproteins. Surprisingly, none of these asparagines occurs in the canonical N-X-(T/S) sequence, where X is anything but a Pro, commonly recognized

by eukaryotic and prokaryotic enzymes involved in N-glycosylation. Instead, Asn[302], Asn[399], and Asn[406] occur in the amino acid sequence (A/G)-N-T-X-T while Asn[280] is in the A-N-I-P-G sequence (De Castro et al. 2013).
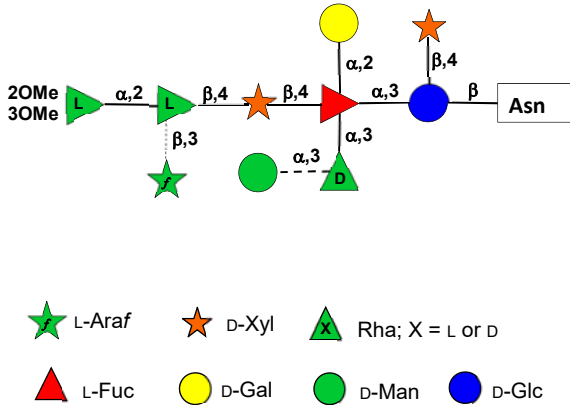
Figure 1.19 Structure of N-glycans of the major capsid protein Vp54 from chlorovirus PBCV-1. Dotted lines indicate nonstoichiometric residues.

These findings evidence that the structure of the Vp54 glycans differ from typical glycans and provide additional evidence that Vp54 glycosylation does not involve host glycosyltransferases.

2.3.2 PBCV-1 Life cycle

PBCV-1 attaches irreversibly and in specific way to the external surface of *Chlorella variabilis* cell walls (Meints et al. 1984). Three-dimensional reconstruction of PBCV-1 in the presence of its host has established that the spike first contacts the cell wall and the fibers then aid in attaching the virus to the wall (Zhang et al. 2011). The spike is too narrow to deliver DNA and serves only to puncture the wall and is then jettisoned. Unlike other phycodnaviruses, the chloroviruses encode enzymes that are involved in polysaccharide degradation, which may play a role in attachment-associated cell wall digestion. Hence, the host cell wall degrades and the viral internal membrane presumably fuses with the host membrane (Fig. 1.20b,c), promoting the entry of the viral DNA and virion-associated proteins into

the cell, leaving an empty capsid attached to the surface (Thiel et al. 2010) (Fig. 1.20d).
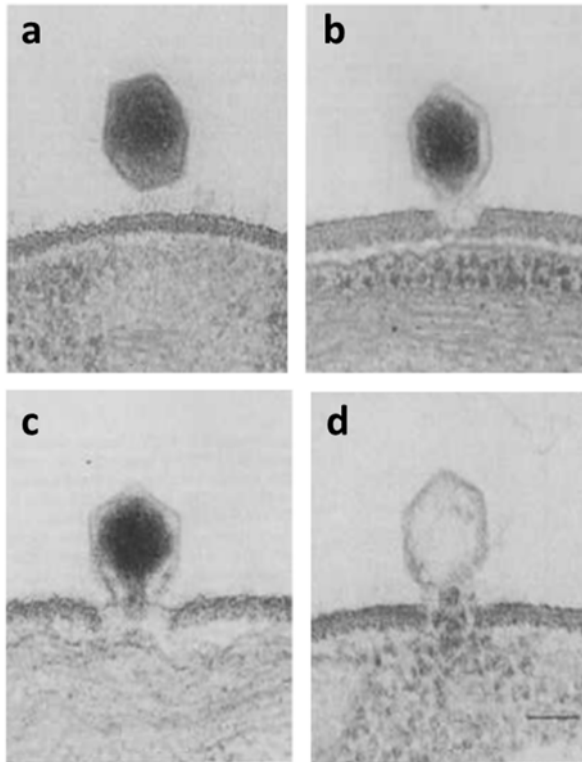
Figure 1.20 Infection of Chlorella strain NC64A by PBCV-1. a) Viral particle in close proximity to the alga. b) Attachment of PBCV-1 to the algal wall and digestion of the wall at the point of attachment. c) Viral DNA beginning to enter the host. e) An empty viral capsid remaining on the surface of the host.

This fusion process initiates rapid depolarization of the host membrane, which is hypothesized to be caused by a virus-encoded potassium ion channel (called Kcv) located in the internal virion membrane. This rapid loss of K^+ ions and associated water fluxes from the host reduce its turgor pressure, which may aid ejection of viral DNA and virion associated proteins into the host (Thiel et al. 2010). Depolarization may also prevent infection by a second virus (Greiner et al. 2009). None of the chloroviruses have a recognizable RNA polymerase gene, so it is assumed that PBCV-1 DNA and viral-associated proteins quickly move to the nucleus where early transcription begins 5 to 10 minutes post-infection (Blanc et al. 2014). In this immediate-early phase of infection, host transcription rates decrease (Rowe et al. 2014) and the host transcription machinery is reprogrammed to transcribe viral DNA. The early mRNAs are transported to the cytoplasm for translation, and at least some early proteins presumably return to the nucleus to

initiate viral DNA replication, followed by late gene transcription. Late mRNAs are transported to the cytoplasm for translation and many of these late proteins are targeted to the virus assembly centers, a localized region in the cytoplasm, where virus capsids are formed and DNA is packaged. The infectious PBCV-1 progeny viruses are released at 6/8 h post-infection (Van Etten and Dunigan 2012), and each infected cell releases ~1000 particles. A schematic diagram of PBCV-1 replication cycle is reported in figure 1.21.
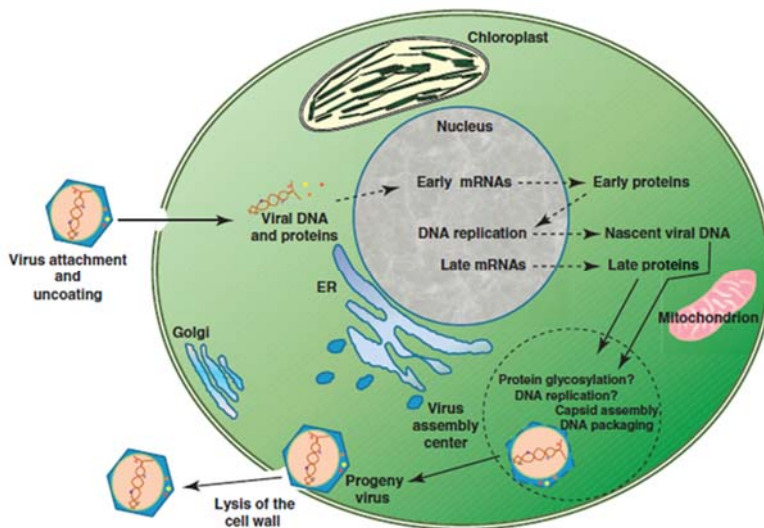
Figure 1.21 Proposed replication cycle of PBCV-1. Solid arrows indicate known events; dashed arrows are used to hypothesized events.

Some chlorella viruses have replication cycles longer than that of PBCV-1. For example, NY-2A has 18 hours replication cycle and, consequently, forms smaller plaques than PBCV-1.

2.3.3 PBCV-1 genome

The PBCV-1 double-stranded DNA genome has been recently resequenced and annotated to correct errors in the original 15-years-old sequence (Dunigan et al. 2012). The PBCV-1genome is a linear 330,61 bp, nonpermuted dsDNA molecule with covalently closed hairpin termini (Fig. 1.22) (Girton et al. 1987; Zhang et al. 1994). A 2,221 bp inverted repeat region is present adjacent to each termini (Strasser et al. 1991), but the rest of the genome is primarily a single copy of DNA sequence.
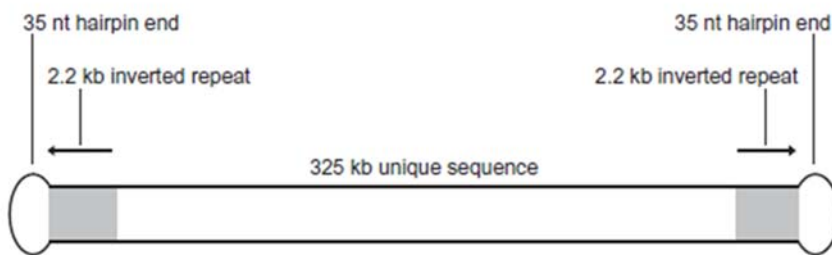
Figure 1.22 Paramecium bursaria chlorella virus genome (~330,61 bp).

Analysis of this genome revealed that it contains methylated nucleotides, similarly to the genomes of all the chlorella viruses that present methylated bases, indeed these viruses can be distinguished from one another by the site-specificity and amount of DNA methylation (Nelson et al. 1993, 1998). Moreover, the PBCV-1 genome encodes 416 CDSs (coding DNA sequences) of 40 codons or longer, and 11 transfer RNA genes (Dunigan et al. 2012). The function of a part of these CDSs has been identified: 17 of the PBCV-1 genes encode enzymes that manipulate sugars, many of which have never been found previously in a virus (Van Etten et al. 2010). Interestingly, PBCV-1 encodes, at least, six putative glycosyltransferases, presumably involved in glycosylation of the PBCV-1 major capsid protein Vp54 (Graves et al. 2001; Wang et al. 1993): *a064r* (638 aa), *a111/114r* (860 aa), *a219/222/226r* (677 aa), *a473l* (517 aa), *a546l* (396 aa), *a075l* (280 aa). The last one was identified only recently (data unpublished). These putative glycosyltransferase-encoding genes are scattered throughout the PBCV-1 genome and none of them has an identifiable signal peptide that would

target them to the ER. Furthermore, the cellular protein localization program PSORT predicts that all these proteins are soluble and located in the cytoplasm, except for A473L, which is predicted to contain one transmembrane domain (Van Etten et al. 2010). These observations suggest that the glycosylation mechanism differ from that reported for other organisms.

The genes for all glycosyltransferases are expressed early during PBCV-1 infection (Yanai-Balser et al. 2010). Thus, assuming that the proteins are stable, the enzymes should be available for either adding sugars to the Vp54 glycans or transferring the glycans.

The PBCV-1 glycosyltransferase A111/114R is the only protein with an ortholog in all the chloroviruses (Van Etten et al. 2010). This gene encodes a protein of 860 amino acids, and bioinformatic analysis suggests that the enzyme has least three domains, one of which has been annotated as a glycosyltransferase.

In addition, genetic experiments established that one of these putative glycosyltransferases (A64R) was involved in Vp54 glycosylation (Graves et al. 2001; Wang et al. 1993). Major details about this gene are discussed in the following paragraph.

2.3.4 Antigenic variants of PBCV-1

Twenty spontaneously derived, antigenic variants of PBCV-1 were identified using polyclonal antiserum prepared against intact PBCV-1 virions. Interestingly, polyclonal antisera prepared against members of each of these antigenic classes react exclusively or predominately with the Vp54 equivalents from the viruses in the class used for the immunization. This result shows that mutants are resistant to PBCV-1 polyclonal antiserum (Wang et al. 1993), suggesting a structural change in the glycan portion, because the polyclonal antiserum recognize the oligosaccharide epitopes on the major capsid protein.

Unlike other viruses that acquire their glycoprotein(s) by budding through a plasma membrane, intact infectious PBCV-1 particles are accumulated inside the host 30–40 min before virus release (Graves et al. 2001). Moreover, the nucleotide sequence of the Vp54 gene (a430l) in each of the variants is identical to the wild-

type PBCV-1, thus the peptide portion of Vp54 is not altered in the mutants (Wang et al. 1993). In addition, all the antigenic variants are grown in the same host, so the differences cannot be attributed to the host. All these observations indicate that Vp54 glycosylation is unusual. In support of this hypothesis, gene from antigenic variants were sequenced (Landstein et al. 1995), but no mutation regarding the envelope protein was detected, whereas mutations in other genes that interfere with specific steps in the viral encoded glycosylation process appeared. Surprisingly, 18 of 21 antigenic variants had mutations in gene a064r (or it was deleted). It is one of the six putative glycosyltransferases, presumably involved in glycosylation of the PBCV-1 major capsid protein Vp54, as specified in the previous paragraph.

The A64R enzyme contains 638 amino acid organized into three conserved domains (Fig. 1.23, on the top) of ca. 200 aa each. First domain presents the typical DXD motif that coordinates the phosphate of the nucleotide sugar (is demonstrated that is UDP- and not other type of nucleotides; Zhang et al. 2007) along with a divalent cation. Instead, the function of other two domains is not clear, even though genetic analysis suggests that the second domain has a glycosyltransferase function and the third domains has a weak match with a methyltransferase.

The antigenic variants are divided into six antigenic classes (Fig. 1.23, column on the left) based on their differential reaction to six different polyclonal antibodies. Regarding the antigenic classes A and E, the domain 1 of A064R is intact, while the second domain is affected. Instead, in the class B, the mutation occurs at domain 1. Class C comprises three viruses for which no mutations occur in a064r gene and one virus that present a point mutation in the first domain. All viruses that have a large deletion of the a064r gene are included in the class D. Finally, there is only one virus belonging to class F, which presents a truncation of the third domain.
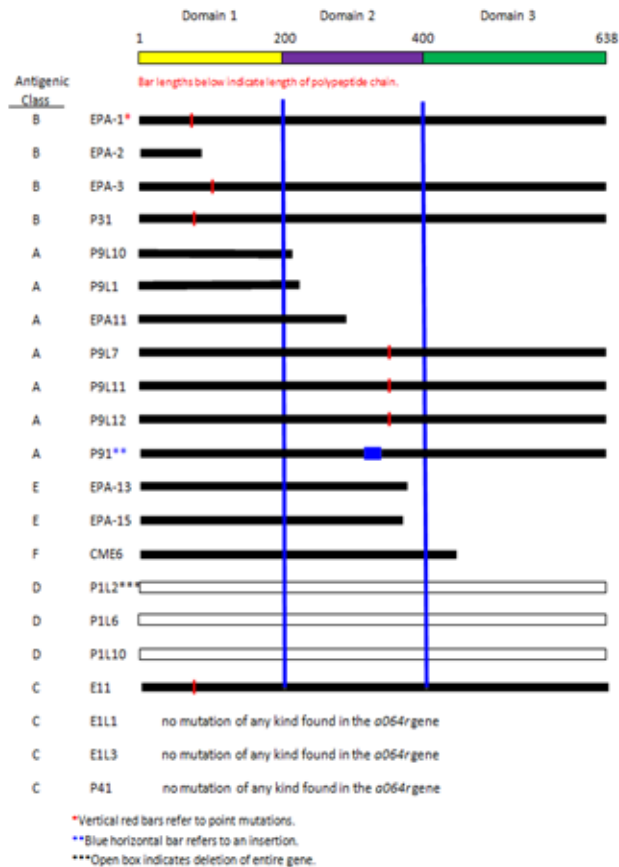
Figure 1.23 Graphic representation of the location and effect of genic mutations within the A064R protein in 21 strains of antigenic variants. The colored bar at the top represents the locations of the three protein domains within the A064R protein in 21 strains of antigenic variants. The colored bar at the top represents the locations of the three protein domains. The left-hand column (Antigenic class) represents the antigenic variant type. The lengths of the black bars represent the predicted length of the encoded protein. * Vertical red bar refers to point mutations. ** Blue horizontal bar refers to an insertion. *** Open box indicates deletion of entire gene.

The first domain (residues 1–211) of A64R was expressed in a bacterial system for crystallization studies. The crystal structure of the protein was solved at 1.6 Å resolution (Zhang et al. 2007) (Fig. 1.24). The analysis revealed that the peptide has a mixed α/β fold containing a central, six-stranded β sheet flanked by α helices. The overall fold is similar to catalytic domains of glycosyltransferases in

the GT-A group, although the amino acid similarity between them is low (less than 14% for equivalent Cα atoms). Crystal structures of A64R, complexed with UDP, CMP, or GDP, established that only UDP bound to A64R in the presence of Mn^{2+}, consistent with its high structural similarity to glycosyltransferases that use UDP as the sugar donor. The structure of the A64R, UDP-glucose and Mn^{2+} complex shows that the largest conformational change occurs when hydrogen bonds were formed with the ligands. Unlike UDP-glucose, UDP-galactose and UDP-GlcNAc did not bind to A64R, suggesting a selective binding of UDP-glucose. Thus UDP-glucose is most likely the sugar donor for A64R, consistent with glucose occurring in the virus major capsid protein glycans (Zang et al. 2007). The complex enzyme with UDP-glucose was deposited in the protein data bank (accession number 2P72).
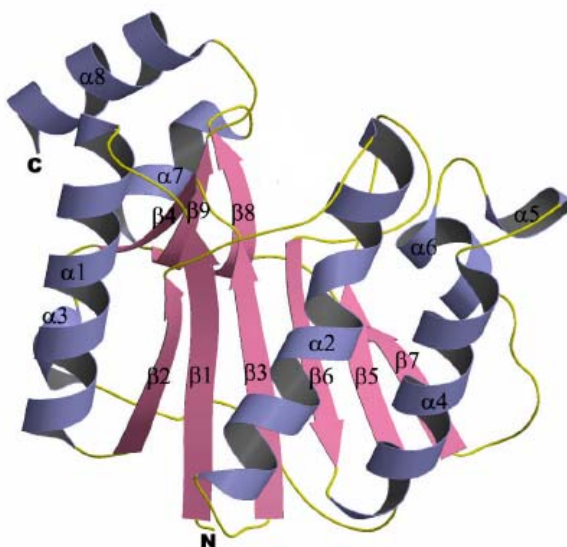
Figure 1.24 Crystal structure of the glycosyltransferase domain of A64R.

2.4 Other Chloroviruses

In the following paragraphs are summarized the important characteristics of some chloroviruses analysed in this thesis.

2.4.1 NC64A viruses (PBCV-1 and NY-2A) and OSy viruses (OSy-NE5)

NC64A viruses (as PBCV-1) infect two different *chlorella variabilis* strains, NC64A and Syngen 2–3 (referred to as Syngen) (Van Etten et al. 1983a), both endosymbionts of the protozoan *P. bursaria*. Interestingly, at the time the plaque assay was developed (1983), it was assumed that Chlorella NC64A and Syngen strains were identical and consequently the studies were focused on PBCV-1 and NC64A (Van Etten and Dunigan, 2012). However, a taxonomic study on rDNA (Kamako et al. 2005) as well as physiological analyses (Quispe et al. 2016) of zoochlorellae established that NC64A and Syngen were similar, but not identical chlorella strains.

In addition, recent studies on OSy-NE5, which was believed to take part of the same group of PBCV-1, revealed that it might only replicate in Syngen cells but not in NC64A cells. Interestingly, the OSy viruses can initiate infection in C. *variabilis* NC64A, but they are unable to complete virus replication. For this reason, it has been proposed a new group, named <u>O</u>nly <u>Sy</u>ngen (OSy) viruses.

OSy-NE5 has the same icosahedral morphology as chlorovirus PBCV-1, with a diameter of ~190 nm, which is identical to the diameter measured along the five-fold axis for PBCV-1 (Yan et al. 2000). Moreover, PBCV-1 presents a single bilayered membrane locates underneath the outer capsid shell and requires for virus infectivity because PBCV-1 is sensitive to a short exposure to chloroform (Skrdla et al. 1984). OSy-NE5 infectivity is also rapidly reduced by exposure to chloroform, suggesting that it probably has an internal membrane (results not shown).

Regarding NY-2A, it has the largest genome (370 kb) of all the characterized chloroviruses (Fitzgerald et al. 2007; Jeanniard et al. 2013), and its genome is

heavily methylated relative to that of PBCV-1 (Van Etten et al. 1985). The genome, also, presents a gene b736l, which has close homology with a546I of PBCV-1 (paragraph 2.2). Crystallization of this enzyme, B736L, discloses that it has a glycosyltransferase type B fold and a good affinity for GDP-mannose (Xiang et al. 2010).

2.4.2 Pbi viruses (MT325, CVM-1 and NE-JV-1)

Pbi viruses specific for *Chlorella Pbi*, renamed *Microactinium conductrix* (Van Etten 2010), were initially isolated from Paramecium bursaria originally collected in fresh water in Germany. They lack the genes needed to synthetize L-fucose and D-rhamnose activated precursors, unlike to the other chloroviruses, in which these genes are instead present.

As said before, these viruses infect the same algal host cluster into monophyletic clades, in other words they don't infect other chlorella strains.
Virus NE-JV-1, is one exception to this scenario, indeed it appears as a sole member of a subgroup of Pbi viruses. NE-JV-1 has 74% amino acid identity on average with the other Pbi viruses in the 32 core proteins used in the phylogenic reconstruction (Fig. 1.14). For comparison, the within-clade average protein amino acid identity is 93%, 95%, and 97% for NC64A, SAG and Pbi (excluding NE-JV-1) viruses, respectively. Between clades, the protein sequence identity ranged from 63% (NC64A vs. Pbi viruses) to 71% (Pbi vs. SAG viruses).

2.4.3 SAG viruses (ATCV-1 and TN603)

SAG viruses infect Chlorella heliozoae strain 3.83, isolated from the heliozoon Acanthocystis turfacea.
Usually, replication of the chloroviruses results in the death and lysis of the algal cell. Other evolutionarily related algal viruses are partially responsible for terminating massive algal blooms that often occur in marine environments (Suttle 2007) or they have profound effects on aquatic food webs. Recently, chlorovirus ATCV-1-like DNA sequences, representing a variety of virus genes, were detected

in 42 deep throat swabs of 92 apparently normal adults. The presence of the ATCV-1 DNA was associated with a modest but statistically significant decrease in performance on certain cognitive assessments of visual processing and visual motor speed (Yolken et al. 2014).

References

Antwerpen MH, Georgi E, Zoeller L, Woelfel R, Stoecker K, Scheidbet P **2015** Whole-genome sequencing of a pandoravirus isolated from keratitis-inducing acanthamoeba. *Genome Announc.* 3: pii: e00136-15.

Baltimore D **1971** Expression of animal virus genomes. *Bacteriological Reviews* 35,235–241.

Blanc G, Mozar M, Agarkova IV, Gurnon JR, Yanai-Balser G, Rowe JM, Xia Y, Riethoven JJ, Dunigan DD, Van Etten JL **2014** Deep RNA sequencing reveals hidden features and dynamics of early gene transcription in Paramecium bursaria chlorella virus 1. *PLoS ONE* 9:e90989. doi: 10.1371/journal.pone.0090989

Breitbart M, Rohwer F **2005** Here a virus, there a virus, everywhere the same virus. *Trends Microbiol* 13:278–284.

Burda P, Aebi M **1999** The dolichol pathway of N-linked glycosylation. *Biochem Biophys Acta* 1426:239-257.

Caspar DLD and Klug A **1962** *Cold Spring Harbor Symp. Quant. Biol.* 27,1–24.

Cherrier MV, Kostyuchenko VA, Xiao C, Bowman VD, Battisti AJ, Yan X, Chipman PR, Baker TS, Van Etten JL, Rossmann MG **2009** An icosahedral algal virus has a complex unique vertex decorated by a spike. *Proc. Natl. Acad. Sci. USA* 106:11085–11089. doi: 10.1073/pnas.0904716106

De Castro C, Molinaro A, Piacente F, Gurnon JR, Sturiale L, Palmigiano A, Lanzetta R, Parrilli M, Garozzo D, Tonetti M, Van Etten JL **2013** Structure of N-linked oligosaccharides attached to chlorovirus PBCV-1 major capsid protein reveals unusual class of complex N-glycans. *Proc Natl Acad Sci USA* 110: 13956-13960. doi: 10.1073/pnas.1313005110.

Doms RW, Lamb RA, Rose JK, Helenius A **1993** Folding and assembly of viral membrane proteins. *Virology* 193: 545-562

Donelli G, Dore E, Frontali C, Grandolfo ME **1975** Structure and physico-chemical properties of bacteriophage G. III. A homogeneous DNA of molecular weight 5 times 10(8). *J. Mol. Biol* 94:555–565.

Doutre G, Philippe N, Abergel C, Claveri J-M **2014** Genome analysis of the first Marseilleviridae representative from Australia indicates that most of its genes contribute to virus fitness. *J Virol* 88:14340–9

Drickamer K & Taylor ME **2006** Introduction to Glycobiology (2nd ed.). Oxford University Press. New York, NY, USA. ISBN 978-0-19-928278-4

Dunigan DD, Cerny RL, Bauman AT, Roach JC, Lane LC, Agarkova IV, Wulser K, Yanai-Balser GM, Gurnon J, Vitek JC, Kronschnabel BJ, Jeanniard A, Blanc G, Upton C, Duncan GA, McClung OW, Ma F, Van Etten JL **2012** *Paramecium bursaria* chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. *J. Virol.* 86:821–8834

Dunigan DD, Fitzgerald LA, Van Etten JL **2006** Phycodnaviruses: a peek at genetic diversity. *Virus Res* 117:119–132. PMID: 16516998

Elbein AD **1987** Inhibitors of the biosynthesis and processing of N-linked oligosac-charide chains. *Annu Rev Biochem 56:497–534.* doi:10.1146/annurev. bi.56.070187.002433

Fitzgerald LA, Graves MV, Li X, Feldblyum T, Nierman WC, Van Etten JL **2007** Sequence and annotation of the 369-kb NY-2A and the 345-kb AR158 viruses that infect Chlorella NC64A. *Virology* 358:472–484.

Fruscione F, Sturla L, Duncan G, Van Etten JL, Valbuzzi P, De Flora A, Di Zanni E, Tonetti M **2008** Differential role of NADP+ and NADPH in the activity and structure of GDP-D-mannose 4,6-dehydratase from two chlorella viruses. *J. Biol. Chem.* 283, 184–193.

Girton LE and Van Etten JL **1987** Restriction site map of the Chlorella virus PBCV-1 genome. *Plant Mol. Biol. 9*, 247-257.

Graves MV, Bernadt CT, Cerny R, Van Etten JL **2001** Molecular and genetic evidence for a virus-encoded glycosyltransferase involved in protein glycosylation. *Virology* 285(2):332–345. doi:10.1006/viro.2001.0937

Greiner T, Frohns F, Kang M, Van Etten JL, Käsmann A, Moroni A, Hertel B, Thiel G **2009** Chlorella viruses prevent multiple infections by depolarizing the host membrane. J. Gen. Virol. 90, 2033–2039.

Hart GW **1992** Glycosylation. *Curr Opin Cell Biol* 4:1017–23.

Jeanniard A, Dunigan DD, Gurnon JR, Agarkova IV, Kang M, Vitek J, Duncan G, McClung OW, Larsen M, Claverie J-M, Van Etten JL, Blanc G **2013** *BMC Genomics* 14:158.

Iyer LM, Aravind L, Koonin EV **2001** Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol 75*(23):11720–11734.

Iyer LM, Balaji S, Koonin EV, Aravind L **2006** Evolutionary genomics of nucleocytoplasmic large DNA viruses. *Virus Res* 117: 156–184.

Landstein D, Burbank DE, Nietfeldt JW, Van Etten **1995** JL *Virology* 214:413-420.

Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O, Bertaux L, Bruley C, Couté Y, Rivkina E, Abergel C, Claverie JM **2014** Thirty-thousand-yearold distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *P Natl Acad Sci USA*111: 4274–9.

Kamako S, Hoshina R, Ueno S, Imamura N **2005** Establishment of axenic endosymbiotic strains of Japanese Paramecium bursaria and the utilization of carbohydrate and nitrogen compounds by the isolated algae. Eur. J. Protistol. 41, 193–202.

Kapaun, E. et al. **1992** Cell wall composition of virus-sensitive symbiotic Chlorella species. *Phytochemistry* 31, 3103–3104.

Koonin EV, Senkevic TG, Dolja VV **2006** The ancient Virus World and evolution of cells *Biol. Direct.* 1:29.

Mokili JL, Rohwer F, Dutilh BE **2012** Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* 2:63–77.

Meints RH, Lee K, Burbank DE, Van Etten JL **1984** Infection of a chlorella-like alga with the virus, PBCV-1: ultrastructural studies. *Virology* 138:341–46.

Nandhagopal N, Simpson AA, Gurnon JR, Yan X, Baker TS, Graves MV, Van Etten JL, Rossmann MG **2002** *PNAS* 99,23, 14758-14763 DOI: 10.1073pnas.232580699.

Nelson M, Burbank DE, Van Etten JL **1998** Chlorella viruses encode multiple DNA methyltransferases. *Biol. Chem.* 379:423–28

Nelson M, Zhang Y, Van Etten JL. **1993** DNAmethyltransferases and DNA site-specific endonucleases encoded by chlorella viruses. In *DNA Methylation: Molecular Biology and Biological Significance*, ed. JP Jost, HP Saluz, pp. 186– 211. Basel: Birkhauser Verlag

Oh-Hama T, Miyachi S. **1988** In *Microalgal Biotechnology*, ed. LJ Borowitzka, MA Borowitzka, pp. 3–26. Cambridge: Cambridge Univ. Press

Olofsson S, Hansen JES **1998** Host cell glycosylation of viral glycoproteins – a battlefield for host defense and viral resistance. *Scand J Infect Dis* 30: 435-440

Que Q, Li Y, Wang I-N, Lane LC, Chaney WG, Van Etten JL **1994** Protein Glycosylation and Myristylation in *Chlorella* Virus Pbcv-1 and its Antigenic Variants. Virology 203:320-327.

Quispe CF, Sonderman O, Khasin M, Riekhof WR, Van Etten JL, Nickerson KW **2016** Comparative genomics, transcriptomics, and physiology distinguish symbiotic from free-living Chlorella strains. *Algal Res* 18, 332–340.

Roth J, Zuber C, Park S, Jang I, Lee Y, Kysela KG, et al. **2010** Protein N-glycosylation, protein folding, and protein quality control. *Mol Cells. Dec* 30(6):497–506

Rowe JM, Jeanniard A, Gurnon JR, Xia Y, Dunigan DD, Van Etten JL, Blanc G **2014** Global analysis of Chlorella variabilis NC64A mRNA profiles during early phase of Paramecium bursaria chlorella virus-1 infection. *PLoS ONE* 9:e90988. doi: 10.1371/journal.pone.0090988

Schwarz F, Markus A **2011** Mechanisms and principles of N-linked protein glycosylation. *Current Opinion in Structural Biology* 21:576-582.

Shihra I, Krauss RW **1965** *Chlorella Physiology and Taxonomy of Forty-One Isolates*. College Park, MD: Univ. Maryland Press

Short CM, Rusanova O, Short SM **2011** Quantification of virus genes provides evidence for seed-bank populations of phycodnaviruses in Lake Ontario, Canada. *ISME J*. 5:810–821.

Skrdla MP, Burbank DE, Xia Y, Meints RH, Van Etten JL **1984** Structural proteins and lipids in a virus, PBCV-1, which replicates in a chlorella-like alga. *Virology* 135, 308±315.

Strasser P, Zhang YP, Rohozinski J, Van Etten JL **1991** The termini of the Chlorella virus PBCV-1 genome are identical 2.2-kbp inverted repeats. *Virology 180*, 763-769.

Sun L, Adams B, Gurnon JR, Ye Y, Van Etten JL **1999** Characterization of two chitinase genes and one chitosanase gene encoded by chlorella virus PBCV-1. *Virology* 263, 376–387.

Suttle CA **2007** Marine viruses-major players in the global ecosystem. *Nat Rev* Microbiol 5:801–812.

Takeda, H. **1991** Sugar composition of the cell wall and the taxonomy of Chlorella (Chlorophyceae). *J. Phycol*. 27, 224–232.

Thiel G, Moroni A, Dunigan DD, Van Etten JL **2010** Initial events associated with virus PBCV-1 infection of *Chlorella* NC64A. *In* U Lüttge, W Beyschlag, B Büdel, eds, Progress in Botany, Vol 71. Springer-Verlag, Berlin, pp 169–183.

Tonetti M, Zanardi D, Gurnon JR et al. **2003** Chlorella virus PBCV-1 encodes two enzymes involved in the biosynthesis of GDP-L-fucose and GDP-D-rhamnose. *J. Biol. Chem.* 278, 21672–21677

Van Etten JL **2003** Unusual life style of giant chlorella viruses. *Annu Rev Genet* 37: 153–195.

Van Etten JL **2011** Giant viruses. *American Scientist* 99:304-311.

Van Etten JL, Burbank DE, Xia Y, Meints RH **1983** Growth cycle of a virus, PBCV-1, that infects Chlorellalike algae. *Virology* 126:117-125.

Van Etten JL, Dunigan DD **2012** Chloroviruses: not yet everyday plant virus. *Trends Plant Sci* 17:1-8.

Van Etten JL, Dunigan DD **2016** Giant Chloroviruses: Five Easy Questions. *PLOS Pathogens* 12(8): e1005751. doi: 10.1371/journal.ppat.1005751

Van Etten JL, Gurnon J, Yanai-Balser G, Dunigan D, Graves MV **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, Lane LC, Meints RH **1991** Viruses and virus-like particles of eukaryotic algae *Microbiological Reviews 55*, 586-620.

Van Etten JL, Meints RH, Kuczmarski D, Burbank DE, Lee K **1982** Viruses of symbiotic chlorella-like algae isolated from *Paramecium bursaria* and *Hydra viridis*. *Proc. Natl. Acad. Sci. USA* 79:3867–3871.

Van Etten JL, Schuster AM, Girton L, Burbank DE, Swinton D, Hattman S **1985** DNA methylation of viruses infecting a eukaryotic chlorella-like green alga. *Nucleic Acids Res* 13:3471–3478.

Vigerust DL, Shepherd VL **2007** Virus glycosylation: Role in virulence and immune interactions. *Trends Microbiol* 15: 211-218.

Wagner EK, Hewlett MJ, Bloom DC, Camerini D **2007** Basic Virology, 3rd Edition-Wiley-Blackwell.

Wang I-N, Li Y, Que Q, Bhattacharya M, Lane LC, Chaney WG, Van Etten **1993** Evidence for virus-encoded glycosylation specificity. *Proc. Natl. Acad. Sci. U.S.A.* 90, 3840–3844.

Wilson WH, Van Etten JL, Allen MJ **2009** The Phycodnaviridae: the story of how tiny giants rule the world. *Curr Top Microbiol Immunol* 328:1–42.

Wilson WH, Van Etten JL, Schroeder DS, Nagasaki K, Brussaard C, Delaroque N, Bratbak G, Suttle C **2005** *Phycodnaviridae*, Elsevier Academic Press, San Diego.

Wimmer E, Mueller S, Tumpey TM, Taubenberger JK **2009** Synthetic viruses: a new opportunity to understand and prevent viral disease. *Nature Biotechnology* 27(12):1163–72.

Wrigley NG **1969** An Electron Microscope Study of the Structure of *Sericesthis* Iridescent Virus. *J. Gen. Virol.* **5**, 123–134.

Wrigley NG **1970** An Electron Microscope Study of the Structure of *Tipula* Iridescent Virus. *J. Gen. Virol.* **6**, 169–173.

Yamada T, Onimatsu H, Van Etten JL **2006** Chlorella viruses. *Adv Virus Res* 66:293–336.

Yan X, Olson NH, van Etten JL, Bergoin M, Rossmann MG, Baker TS **2000** Structure and proposed assembly of large lipid-containing dsDNA viruses. *Nat Struct Biol.* 7:101–103.

Yanai-Balser GM, Duncan GA, Eudy JD, Wang D, Li X, Agarkova I., Dunigan DD, Van Etten JL **2010** Microarray analysis of Paramecium bursaria chlorella virus 1 transcription. *J Virol* 84: 532–542 doi:10.1128/JVI.01698-09.

Yolken RH, Jones-Brando L, Dunigan DD, Kannan G, Dickerson FB, Severance EG, Sabunciyan S, Talbot CC, Prandovszky E, Gurnon JR, Agarkova IV, Leister F, Gressitt KL, Chen O, Deuber B, Ma F, Pletnikov MV, Van Etten JL **2014** Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. *Proc. Natl. Acad. Sci. USA* 111:16106–16111. doi: 10.1073/pnas.1418895111

Yoosuf N, Pagnier I, Fournous G, Robert C, Raoult D, La Scola B, Colson P **2014** Draft genome sequences of Terra1 and Terra2 viruses, new members of the family Mimiviridae isolated from soil. *Virology 452–453:125–32.*

Yoosuf N, Yutin N, Colson P, Shabalina SA, Pagnier I, Robert C, Azza S, Klose T, Wong J, Rossmann MG, La Scola B, Raoult D, Koonin E **2012** Related giant viruses in distant locations and different habitats: Acanthamoeba polyphaga moumouvirus represents a third lineage of the Mimiviridae that is close to the megavirus lineage. *Genome Biol Evol* 4:1324–30.

Yutin N, Wolf YI, Raoult D, Koonin EV **2010** Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* 6:223.

Zhang Y, Strasser P, Grabherr R, Van Etten JL **1994** Hairpin loop structure at the termini of the chlorella virus PBCV-1 genome. *Virology 202*,1079-1082.

Zhang X, Xiang Y, Dunigan DD, Klose T, Chipman PR, Van Etten JL, Rossmann MG. **2011** Three-dimensional structure and function of the Paramecium bursaria chlorella virus capsid. *Proc. Natl. Acad. Sci. U.S.A.* 108:14837–14842.

Zhan Y, Xiang Y, Van Etten JL, Rossmann MG **2007** Structure and function of a chlorella virus PBCV-1 encoded glycosyltransferase. *Structure* 15:1031–1039.

Section II

Results and discussion

Chapter 3

Chlorella viruses

The most studied chlorovirus, PBCV-1, presents a major capsid protein, Vp54, that contains four *N*-linked glycans. As widely discussed in the first section, the Vp54 glycosylation is unusual in many features: i) unlike most viruses, PBCV-1 encodes most, if not all, the components required to the glycosylate its major capsid protein; ii) the glycans are not located in a typical N-X-(T/S) consensus site; iii) the glycans are not attached to the protein via *N*-acetylglucosamine, typical of Asn-linked glycoproteins; iv) the glycosylation process occurs in the cytoplasm, probably; and v) the most amazing thing is that the Vp54 structure (Fig. 3.1) is totally different from the *N*-glycans reported, so far, for the three domains of the life, Eukariots, Archea and Procariots. Indeed, the main Vp54 glycoform consists of 8-10 neutral monosaccharides arranged in highly branched fashion in which the fucose unit is completely substituted; the terminal residue at the nonreducing end is a rhamnose capped with two methyl groups; there are two rhamnose with both D and L configuration; and the β-glucose is used to attached the glycan portion to the protein moiety. Other glycoforms differ from the previous for the presence of the arabinofuranose and mannose that are in nonstoichiometric fashion.
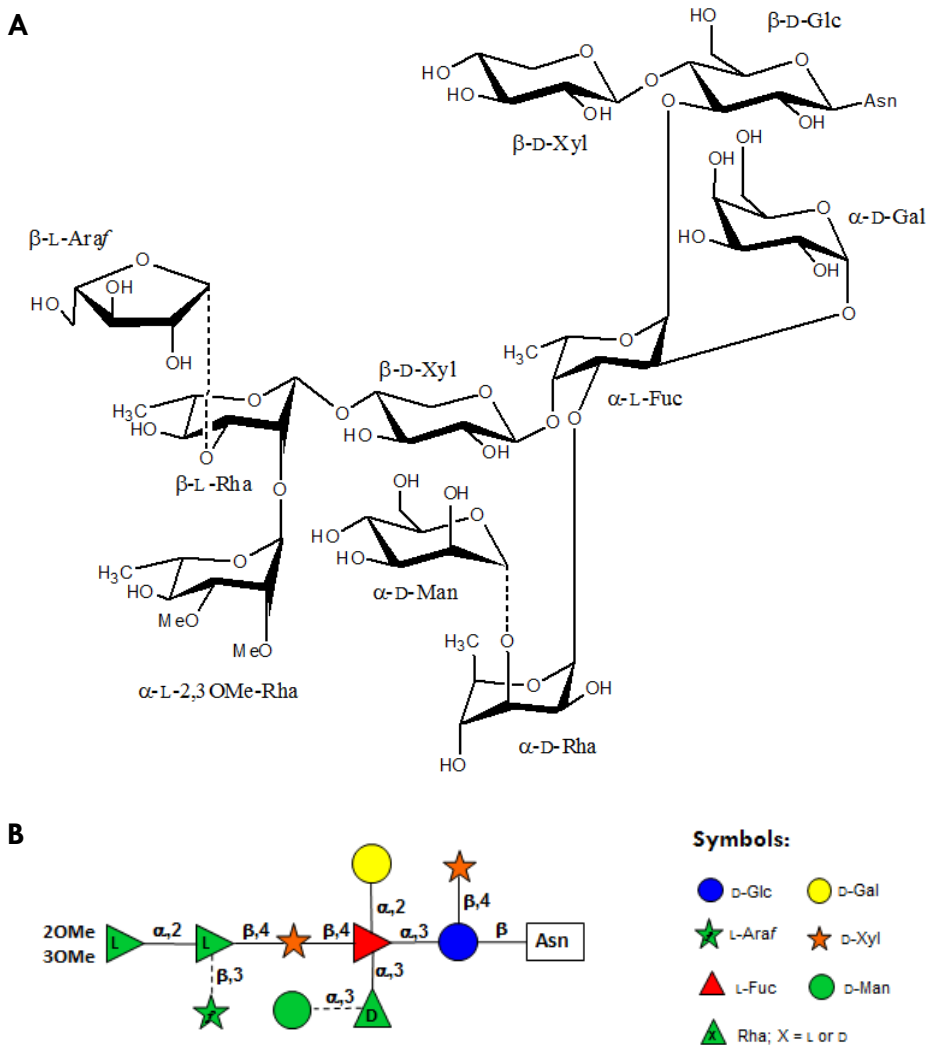
Figure 3.1 A. Structure of the PBCV-1 glycan. B. The same structure of A, depicted using the glycopedia symbols (glycopedia.eu). This representation will be used for the next structures due to its simplicity, with respect to the first one (De Castro et al. 2013). In both representations, the dotted lines indicate the monosaccharides that occur in not-stoichiometric fashion.

Prompted by the peculiarity of these motifs, the purpose of this project was the examination of the *N*-glycans structure of other chloroviruses to determine if those described for PBCV-1 were unique, or if their structural features were conserved among the chloroviruses.

Thus, I have analyzed the glycan structures of other viral isolates that belong to the same PBCV-1 genus, but that have different host specificity, such as: **NY-2A**

(NC64A virus); **MT325**, **CMV-1** and **NE-JV-1** (Pbi viruses); **ATCV-1** and **TN603** (SAG viruses); and **OSy-NE5** (OSy viruses).

3.1 Major capsid protein isolation

The major capsid proteins (MCPs) of each chlorovirus was extracted according to the procedure reported in the paragraph 8.1, and their SDS-PAGE analysis (Fig. 3.2) suggested that these MCPs were glycosylated (silver staining data were not obtained for NE-JV1 and OSy-NE5).
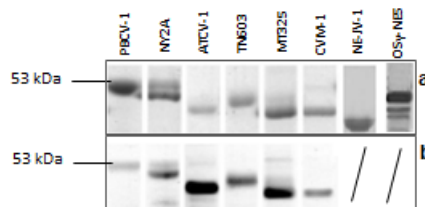
Figure 3.2 SDS-PAGE of total proteins of the MCPs from chloroviruses PBCV-1, NY-2A, ATCV-1, TN603, CMV-1, MT325, NE-JV-1 and OSy-NE5. a) Glycoprotein staining (Comassie). b) Silver staining. Note that NY-2A has two bands and that the silver staining analysis for NE-JV-1 and OSy-NE5 are not present.

Glycoprotein staining suggested that all chloroviruses encode one MCPs, except NY-2A, which presented two bands corresponding with two closely related MCPs, B585L and B617L, both glycosylated. These two MCPs could not be separated, thus further analysis was performed on the combined pair.

MCP amino acid sequence from the prototype, PBCV-1, was used to identify the MCPs from the other chloroviruses and to evaluate their molecular weight, which was always minor than the value determined experimentally using the ESI-MS (Table 3.1) or SDS-PAGE, indicating that all the proteins were glycosylated.

Table 3.1 Information about major capsid proteins of chloroviruses. For NY-2A the molecular weight of the two oligosaccharides are given, instead for TN603, MT325 and ATCV-1 only the MW of the largest oligosaccharide is indicated. Data for OSy-NE5 and NE-JV-1 were not determined, except for the OSy-NE5 gene code: OS5_219L. (n.d.: not determined)

Virus	Gene code	% AA identity	MW (without glycans)	MW	*N*-glycan MW
PBCV-1	A430L	100	48168	53790	1384
NY-2A	B585L	96	48359	52641	1348
				51250	880
NY-2A	B617L	94	48354	As above	As above
MT325	M463L	78	47652	n.d.	1186
CVM-1	548L	79	47.662	51.247	936
TN603	325L	79	48.437	54.508	1186
ATCV-1	Z280L	78	48421	54827	922

3.2 Glycopeptide(s) isolation

The major capsid protein of each chlorovirus, previously isolated, was hydrolyzed with Proteinase K, in order to minimize the length and the heterogeneity of the peptide moiety, and then purified using a size exclusion chromatography. Chromatographic peaks were checked via NMR (Fig. 3.3) to identify which fraction(s) contained the glycopeptides. Generally, one main fraction was obtained, which eluted at 50 % of the column volume and with a yield of the 15 – 20 %, with respect to the starting MCP amount. Only NY-2A produced two main fractions, named NY-2A$_1$ and NY-2A$_2$, that were analyzed separately.
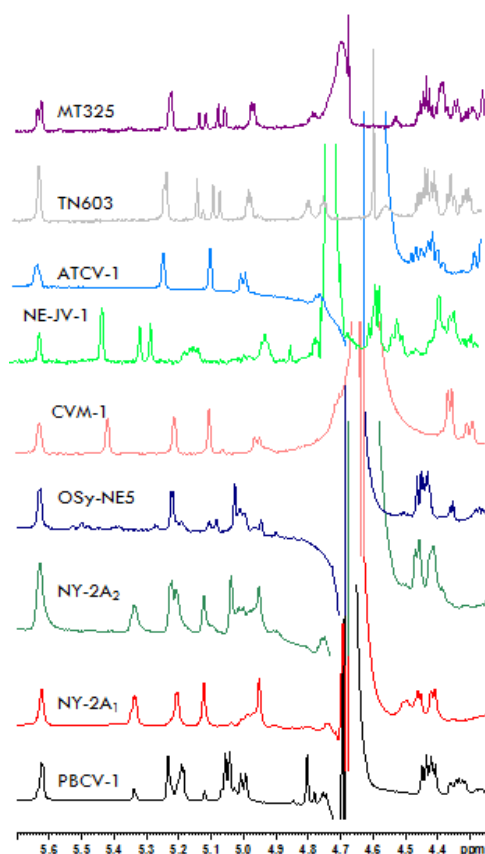
Figure 3.3 ^{1}HNMR spectra reporting the anomeric region of the glycopeptides obtained from the capsid protein of the different chloroviruses. Spectra were recorded at 600 MHz and at 310 K, except for TN603 (recorded at 950 MHz and at 318 K), MT325 (950 MHz, 310 K), ATCV1 (600 MHZ, 318 K), and NE-JV-1 (600 MHz, 323 K). The spectrum of PBCV-1 is provided as a reference. (De Castro et al. 2016; Quispe et al. 2016; Speciale et al. 2017)

3.3 Determination of the glycopeptide structure(s)

Glycopeptide structures were elucidated combining the information obtained by chemical, spectrometric and spectroscopic analysis, as elucidated below.

3.3.1 Chemical analysis

The monosaccharide composition of the glycopeptides was performed after the full set of NMR spectra. The samples, previously transforming into their corresponding acetylated methyl glycoside, were used to prepare the acetylated or methylated 2-(-)-octyl derivatives, to determine the absolute configuration of each monosaccharide. This is a very important analysis because it distinguishes between enantiomeric residues as L- and D-rhamnose, as occurs in PBCV-1.

Regarding NY-2A, two fractions were obtained from chromatographic purification, but the absolute configuration analysis was performed on fraction NY-2A$_1$. Fully acetylated 2-octylglycosydes analysis (chromatographic profile in figure 3.4) revealed the configuration of all monosaccharides except 3OMeXyl and 3OMeRha, due to the lack of a proper standard. To achieve these information, the acetyl groups were replaced with methyl groups by methylation reaction of the octyl glycosides mixture and comparing the new chromatogram with that from the appropriate standards. This analysis showed that the three xylose units were all D-configured, instead the two rhamnose did not present the same configuration: 3OMeRha is L and terminal-Rha is D (Figure 3.4).
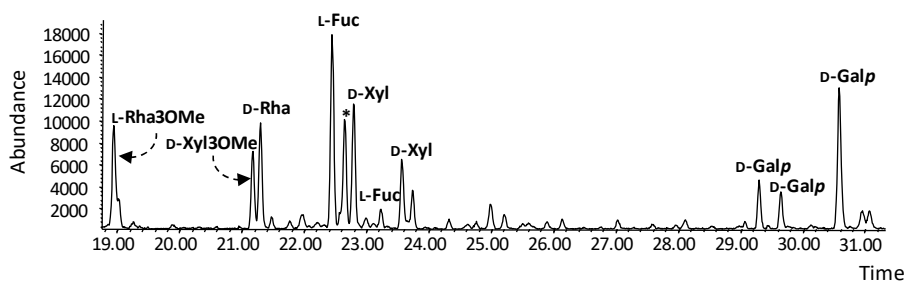
Figure 3.4 Acetylated 2-(-)-octyl glycosydes from NY-2A$_1$ oligosaccharide. * indicates impurity.

Pbi and SAG viruses, instead, contained only one rhamnose unit methylated at O-3, but the standard for this residue was not available. Thus, the samples were analyzed directly as permethylated octyl glycoside. As example, the chromatogram of CVM-1 glycan (Fig. 3.5) is reported, and inspection enabled us to determine the monosaccharide composition along with the absolute configuration.
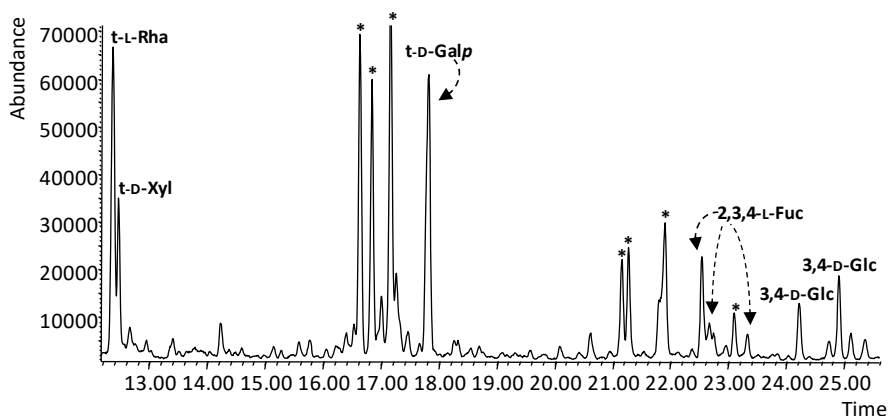
Figure 3.5 Permethylated 2-(-)-octyl glycosydes profile obtained from CVM-1. (* impurities) (De Castro et al. 2016)

In both procedures, acetylated and permethylated octyl glycosides, information regarding the xylulose, a monosaccharide that occurs in MT325 and TN603 glycopeptides, was lost because it degrades during the methanolysis reaction. Thus, a further derivatization (partially methylated and acetylated alditol derivatives) was done to prove its presence. Analysis of these derivatives did not show the xylulose unit, but established there are two types of xylose units, one terminal and one 4 linked, this last originates by the presence of xylulose at position 4 (example in figure 3.6).
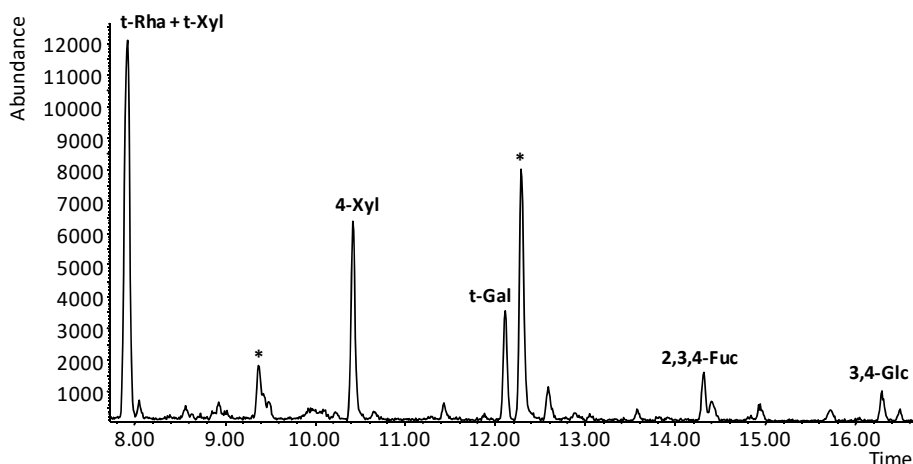
Figure 3.6 Partially methylated alditol acetates from TN603. * impurities.

<h2 style="text-align:center"><u>3.3.2 NMR analysis</u></h2>

NMR experiments were recorded at 310 K, with a field strength of 600 MHz, for the for the majority of the chloroviruses. There are some exceptions: TN603 was recorded at 950 MHz and at 318 K, to resolve the complexity of the spectra; for MT325 was used a 950 MHz field owing to the low amount of the sample; instead for ATCV-1 and NE-JV-1 only the acquisition temperature was changed (318 K and 323 K, respectively) to minimize the overlap of glycopeptide signals with that of the residual water peak.

3.3.2.1 NY-2A (NC64A viruses)

As said before, NY-2A produced two different glycopeptides fractions, that we named NY-2A$_1$ and NY-2A$_2$. Comparison of the two ^{1}H-NMR spectra, showed that NY-2A$_1$ displayed a minor number of anomeric signals, in a 1:1 ratio.

Starting from this glycan, the heteronuclear single quantum correlation (HSQC) spectrum (Fig. 3.7) showed nine anomeric signals, labeled with a letter in order of their decreasing proton chemical shift, along with two ring proton signals in the range 5.7 - 4.3 ppm, a crowded carbinolic region from 4.4 to 3.01 ppm with two

methyl groups (at 3.62 and 3.47 ppm) and three methyl groups at ca. 1.3 ppm, typical of deoxyhexoses. Analysis started from **D** residue which was N-linked to the protein because of its diagnostic carbon chemical shift (80.7 ppm), while the efficient propagation of magnetization in the TOCSY spectrum was consistent with the *gluco* stereochemistry of the residue. This information combined with those from the other experiments (COSY, TOCSY and HSQC) identified **D** as a β-glucose substituted at O-3 and O-4, owing to the displacement at low field of the respective carbons signals, similarly to what was reported for the analogue residue in PBCV-1 glycopeptide (De Castro et al. 2013). It should be stressed that the anomeric proton signal of **D** appeared rather broad, this characteristic was related to the heterogeneity of the peptide portion containing the Asn residue to which the monosaccharide was linked.

Using a similar approach, **A** was identified as a fucose unit: starting from anomeric proton (5.62 ppm) its pattern in the TOCSY spectrum displayed only three correlations (with H-2, H-3 and H-4), beyond which both COSY and TOCSY spectra failed in the detecting H-5, suggesting that this residue was a *galacto* configured. T-ROESY was used to determine the H-5, that was in turn correlated with a methyl group at 1.30 ppm. Thus, **A** was an α-fucose residue and the low field values of the carbon C2, C3 and C4, with respect to the standard value, indicated that it was fully substituted, as happen in PBCV-1. The spectroscopic pattern of **B** was comparable to that of **A**, with the difference that H-5 was connected to a hydroxymethylene group instead of a methyl group, identifying this unit as a galactose, α configured at the anomeric center.

Regarding **C**, TOCSY spectrum displayed one main correlation from the anomeric proton, whereas the connections up to the methyl group were visible from H-2 due to favorable proton-proton coupling constant. This information, combined with those of the HSQC and HMBC spectra (Fig 3.7), identified this residue as rhamnose unit, α configured at the anomeric center on the basis of its H-5/C-5 chemical shifts, and linked at O-3 of **A**. Of note, chemical analysis disclosed that the absolute configuration of this residue was **D**, similarly to what found for PBCV-1. In addition, **C** was substituted at both O-3 and O-4 and examination of HMBC spectrum identified in **I** and **L** its substituents.

The HSQCTOCSY (Fig. 3.8) propagation pattern from **L** anomeric proton contained only three densities diagnostic of the galacto ring stereochemistry of the residue and integration of this information with those from the other spectra defined this unit as a β-galactose glycosylated at O-2 due to the lowfield shifts of its C-2 (75.4 ppm) with respect to the standard value (71.7 ppm) (Bock and Pedersen 1983). HSQCTOCSY of **I** contain only the correlation with C-2, typical of *manno* configured residues. Starting from H-2 was possible to determine the position of all the other protons and carbons (Table 3.2). Thus, this residue is a terminal α-rhamnose because of the similarity of its C-5 value (69.2 ppm) with that if the reference methylglycoside (Bock and Pedersen 1983). Moreover, its C-3 value (80.4 ppm) indicated that the hydroxyl function was capped with a methyl group (^{1}H/^{13}C 3.47/57.0 ppm).

Regarding **H**, it was also methylated as suggested by the downfield shift of its C-3 (^{1}H/^{13}C 3.62/61.2 ppm), but unlike from **I**, its TOCSY pattern showed the complete magnetization propagation from H-1 up to the two protons H-5, identifying this unit as a terminal α-xylose methylated at O-3. T-ROESY spectrum indicated that this unit was located at O-2 of **L** unit.

Finally, the spectroscopic patterns of **E** and **F** residue were very similar. In both cases, there was an efficient propagation of the magnetization in the TOCSY spectrum identifying them as two beta-xylose units. Moreover, the T-ROESY experiment revealed that **E** was linked at O-4 of **A**, while **F** at O-4 of **D**. These two units was also named distal and proximal, respectively, owing to the different distance from the N-glycosylation site.

Completion of the NY-2A$_1$ glycopeptide was inferred by analyzing the T-ROESY spectrum, which disclosed that **A** was linked to O-3 of **D** and in turn had **B** at O-2. In addition, analysis of 2-octylglycosides determined the absolute configuration of all residues and in particular showed that the two rhamnose units did not possess the same absolute stereochemistry, being **C** D-configured and **I** L-configured.

The interpretation of the second chromatographic fraction (NY-2A$_2$) was facilitated because its HSQC spectrum (Fig. 3.9) contained all signals assigned previously to the nonasaccharide, for which the same letters were maintained through the attribution. Therefore, the attention was focused on the new signals and in

particular, on the anomeric proton at 5.04 ppm, labelled **C'**. TOCSY spectrum from this signal had one intense correlation with H-2 and a weak one, attributed to H-3 by integrating COSY information. This pattern indicated a *manno* configured residue and attribution of the other proton resonances from homonuclear spectra along with carbon chemical shifts from HSQC spectrum indicated that **C'** was an unsubstituted rhamnose. Inspection of the HMBC spectrum detected a long-range correlation connecting H-1 of **C'** to a carbon at 77.2 ppm, attributed to C-3 of a fucose (Table 3.3), named **A'** having the anomeric proton almost coincident with that indicated as **A**.

Thus, NY-2A produced two different glycopeptides: a nonasaccharide and its truncated form, a hexasaccharide (Fig. 3.10), which differed from the longer form because of the lack of the substituents at the rhamnose **C**, linked at the hyperbranched fucose **A**. The truncated form generated a magnetic environment different from that of the nonasaccharide, so that the chemical shifts of the other sugars were affected.
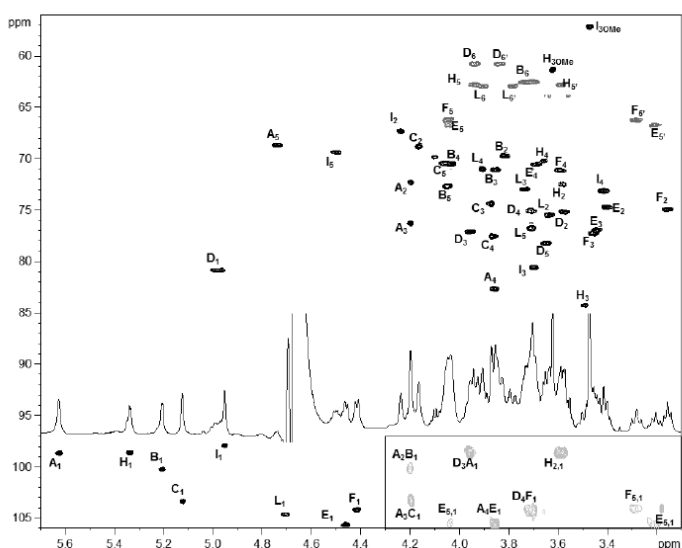
Figure 3.7 Expansion of NY-2A$_1$ glycopeptide HSQC spectrum measured at 600 MHz, 310 K, HMBC spectrum is reported in the inset at the right-bottom corner. (De Castro et al. 2016)
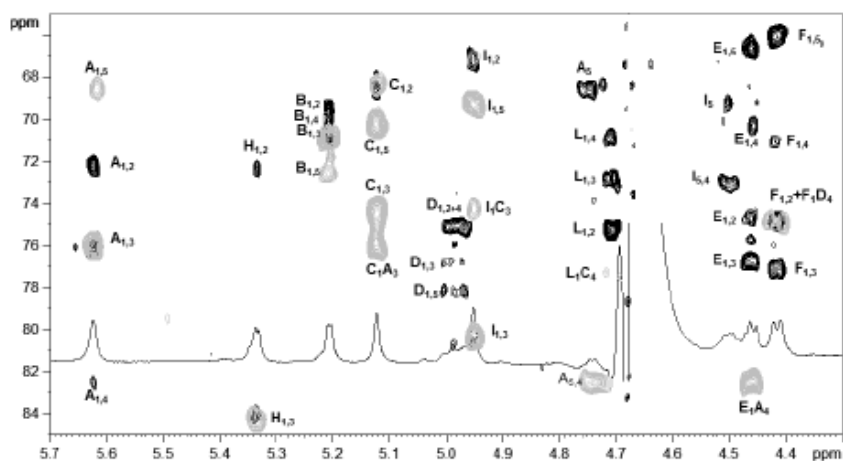
Figure 3.8 Superimposition of NY-2A$_1$ glycopeptide HSQCTOCSY (black) and HMBC (grey) spectra measured at 600 MHz, 310 K. Correlations attribution follows the letter system of Figure 3.10. (De Castro et al. 2016)
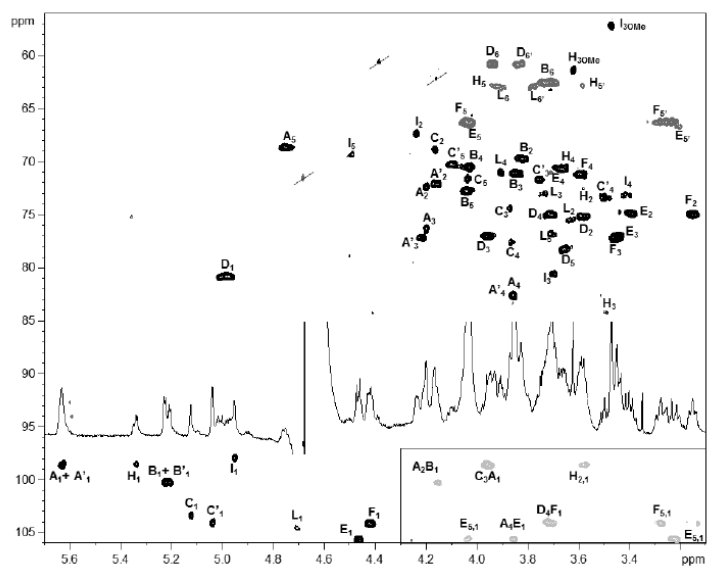
Figure 3.9 Expansion of NY-2A$_2$ glycopeptide HSQC spectrum measured at 600 MHz, 310 K, HMBC spectrum is reported in the inset at the right-bottom corner. In grey are reported carbons bearing two hydrogen atoms, that have the opposite sign with respect to the other. Signals crossed are related to the peptide moiety of the glycopeptide or to impurities. (De Castro et al. 2016)
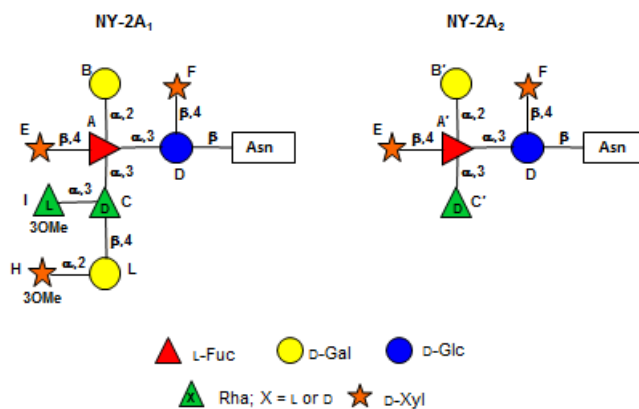
Figure 3.10 Structure of N-glycans of NY-2A. Letter labels are those used during NMR assignment. (De Castro et al. 2016)

Table 3.2 (600MHz, 310 K, D$_2$O) Proton and carbon chemical shifts deduced for NY-2A$_1$ glycopeptide. Both **H** and **I** are methylated at O-3 (^{1}H/^{13}C 3.62/6.12 ppm and 3.47/57.0 ppm, respectively). (De Castro et al. 2016)

		1	2	3	4	5 (5$_{eq}$; 5$_{ax}$)	6; 6'
A	^{1}H	5.62	4.20	4.20	3.86	4.74	1.30
2,3,4-α-L-Fuc	^{13}C	98.5	72.2	76.2	82.5	68.5	16.1
B	^{1}H	5.21	3.82	3.86	4.03	4.04	3.73; 3.69
t-α-D-Gal	^{13}C	100.1	69.5	70.9	70.4	72.6	62.4
C	^{1}H	5.12	4.16	3.87	3.85	4.06	1.43
3,4-α-D-Rha	^{13}C	103.2	68.7	74.2	77.4	70.3	18.3
D	^{1}H	4.98	3.57	3.96	3.70	3.65	3.94; 3.84
3,4-β-D-Glc	^{13}C	80.7	74.9	76.9	74.9	78.1	60.6
E	^{1}H	4.46	3.40	3.45	3.69	4.04; 3.20	-
t-β-D-Xyl	^{13}C	105.6	74.5	76.8	70.4	66.5	-
F	^{1}H	4.42	3.16	3.46	3.59	4.05; 3.28	-
t-β-D-Xyl	^{13}C	104.1	74.8	77.1	70.9	66.1	-
H	^{1}H	5.34	3.58	3.49	3.66	3.94; 3.58	-
t-α-D-3OMeXyl	^{13}C	98.5	72.3	84.1	70.1	62.7	-
I	^{1}H	4.95	4.24	3.69	3.42	4.50	1.30
t-α-L-3OMeRha	^{13}C	97.8	67.2	80.4	73.0	69.3	18.2
L	^{1}H	4.71	3.63	3.73	3.91	3.70	3.91; 3.78
2-β-D-Gal	^{13}C	104.5	75.4	72.8	70.9	76.6	62.8

Table 3.3 (600MHz, 310 K, D_2O) Proton and carbon chemical shifts deduced for NY-2A$_2$ glycopeptide. Both **B** and **F** are methylated at O-3 (^1H/^{13}C 3.62/6.12 ppm and 3.47/57.0 ppm, respectively). (De Castro et al. 2016)

		1	2	3	4	5 (5$_{eq}$; 5$_{ax}$)	6; 6'
A	^1H	5.62	4.18	4.20	3.87	4.75	1.30
2,3,4-α-L-Fuc	^{13}C	98.5	72.3	76.2	82.7	68.5	16.3
A'	^1H	5.63	4.16	4.21	3.86	4.75	1.30
2,3,4-α-L-Fuc	^{13}C	98.6	72.1	77.2	82.7	68.5	16.3
B	^1H	5.21	3.81	3.85	4.03	4.03	3.73; 3.70
t-α-D-Gal	^{13}C	100.2	69.6	71.0	70.6	72.7	62.4
B'	^1H	5.22	3.82	3.85	4.03	4.03	3.73; 3.70
t-α-D-Gal	^{13}C	100.2	69.6	71.0	70.6	72.7	62.4
C	^1H	5.12	4.16	3.87	3.86	4.06	1.42
3,4-α-D-Rha	^{13}C	103.4	68.7	74.3	77.5	70.3	18.2
C'	^1H	5.04	4.03	3.75	3.50	4.10	1.28
3,4-α-D-Rha	^{13}C	104.0	71.4	71.6	73.3	70.2	18.1
D	^1H	4.99	3.58	3.95	3.71	3.66	3.94; 3.83
3,4-β-D-Glc	^{13}C	80.7	75.1	77.0	74.9	78.3	60.6
E	^1H	4.46	3.40	3.45	3.67	4.03; 3.22	-
t-β-D-Xyl	^{13}C	105.6	74.8	77.2	70.7	66.2	-
F	^1H	4.42	3.14	3.45	3.60	4.04; 3.28	-
t-β-D-Xyl	^{13}C	104.1	74.9	77.2	71.1	66.2	-
H	^1H	5.34	3.58	3.49	3.66	3.94; 3.59	-
t-α-D-3OMeXyl	^{13}C	98.5	72.3	84.2	70.5	62.7	-
I	^1H	4.95	4.24	3.70	3.42	4.50	1.29
t-α-L-3OMeRha	^{13}C	97.9	67.3	80.5	73.1	69.2	16.2
L	^1H	4.71	3.64	3.74	3.91	3.71	3.90; 3.78
2-β-D-Gal	^{13}C	104.5	75.4	72.8	70.9	76.6	62.8

3.3.2.2 OSy-NE5 (OSy viruses)

The ^{1}H-NMR and the HSQC spectra (Fig. 3.11) of OSy-NE5 glycopeptide presented some similarities with that of NY-2A$_2$. Indeed, proton and carbon chemical shifts of **A**, **B**, **C** and **D** were very similar (Table 3.4), in agreement with their attribution as α-L-fucose, α-D-galactose, β-D-glucose and α-D-rhamnose, respectively. The main difference was in the xylose units.

The spectroscopic pattern of **E** was similar to that of **F** and the efficient propagation of the magnetization in the TOCSY spectrum identified them as two beta-xylose units, similarly to what reported for NY-2A. But, H-3 of **F** resonated at (^{1}H/[13]C 3.23/86.8 ppm) suggesting that this position was O-methylated, as confirmed by the corresponding HMBC and T-ROESY correlations.

Completion of the OSy-NE5 glycopeptide structure was inferred by analyzing the T-ROESY spectrum (Fig. 3.12), which disclosed that **A** was linked at O-3 of **D** and in turn had **B** at O-2, **C** at O-3 and **E** at O-4. In addition, **D** was substituted at O-4 with **F** (structure was reported in figure 3.13). Chemical analysis revealed that also the rhamnose unit (**C**) is D configured as occur for both PBCV-1 and NY-2A.
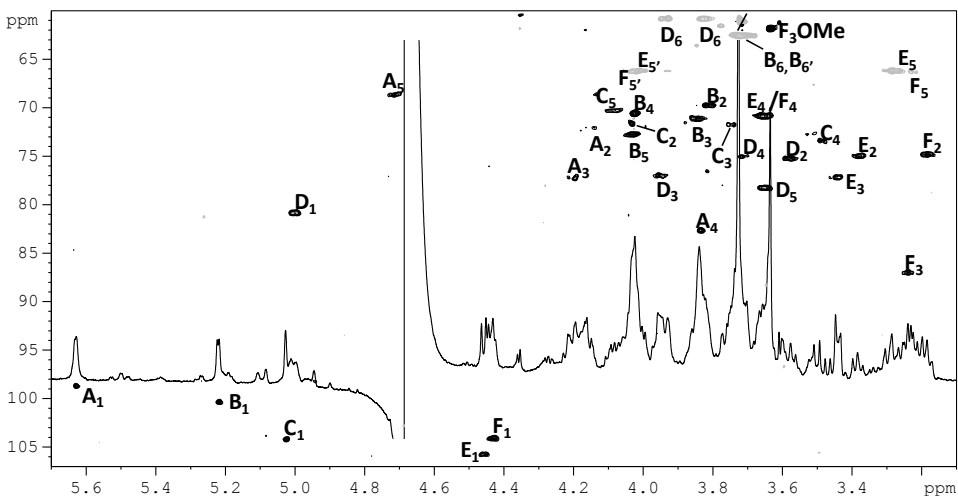
Figure 3.11 Expansion of OSy-NE5 glycopeptide HSQC spectrum measured at 600 MHz, 310 K. (Quispe et al. 2016) Carbons bearing two hydrogen atoms have densities of opposite sign compared to the others and are colored in grey. Signals crossed are related to the peptide moiety of the glycopeptide or impurities.
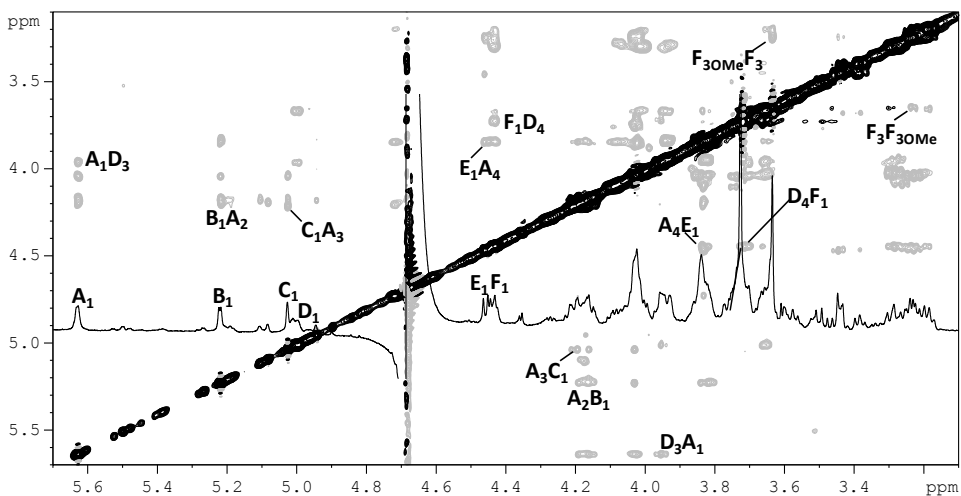
Figure 3.12 Expansion of OSy-NE5 glycopeptide T-ROESY spectrum measured at 600 MHz, 310 K.
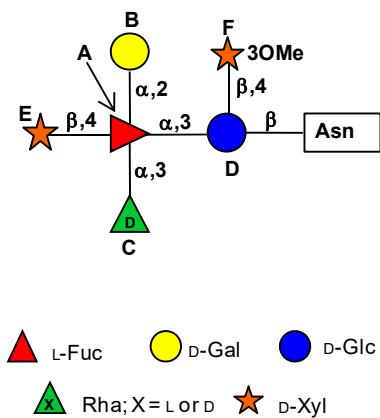
Figure 3.13 Structure of N-glycan obtained from OSy-NE5 (Quispe et al. 2016). Letter labels are those used during NMR assignment.

Table 3.4 (600MHz, 310 K, D$_2$O) Proton and carbon chemical shifts deduced for OSy-NE5 glycopeptide. **F** is methylated at O-3 (^{1}H/^{13}C 3.64/6.16 ppm).

		1	2	3	4	5 (5$_{eq}$- 5$_{ax}$)	6-6'
A	^{1}H	5.63	4.15	4.20	3.83	4.712	1.30
2,3,4-α-ʟ-Fuc	^{13}C	98.5	71.9	77.0	82.4	68.5	16.18
B	^{1}H	5.22	3.81	3.84	4.02	4.02	3.73-3.70
α-ᴅ-Gal	^{13}C	100.2	69.5	70.9	70.3	72.5	62.3
C	^{1}H	5.02	4.03	3.74	3.49	4.09	1.28
α-ʟ-Rha	^{13}C	104.0	71.5	71.5	73.2	70.0	17.9
D	^{1}H	4.99	3.57	3.95	3.71	3.65	3.94-3.82
3,4-β-ᴅ-Glc	^{13}C	80.7	75.1	76.8	74.8	78.1	60.6
E	^{1}H	4.46	3.38	3.44	3.66	4.03-3.24	-
t-β-ᴅ-Xyl	^{13}C	105.5	74.7	76.9	70.6	66.0	-
F	^{1}H	4.43	3.18	3.234	3.66	4.00-3.23	-
β-ᴅ-Xyl 3OMe	^{13}C	103.9	74.6	86.8	70.6	66.0	-

3.3.2.3 CVM-1 (Pbi viruses)

Inspection of the low field region (5.7-4.3 ppm) of the HSQC spectrum (Fig. 3.14) displayed six anomeric carbons, labeled from **A** to **F**.

D (^{1}H 5.01 ppm) was *N*-linked to the protein because of its diagnostic carbon chemical shift (80.7 ppm); thus, this residue was a glucose as inferred by the efficient propagation of magnetization in the TOCSY spectrum and by its ^{13}C chemical shift values (Table 3.5), with C-3 and C-4 (77.0 and 74.7 ppm, respectively) shifted at low field by glycosylation, similarly to what was reported for the analogue residue in previous glycopeptides. Regarding **A**, starting from the anomeric proton (5.68 ppm), its pattern in the TOCSY spectrum displayed only three correlations due to H-3/H-4 coupling constant value that hindered further propagation of the magnetization to the other protons of the residue, as typically occurs for *galacto* configured residue. H-1/C-5 long-range correlation (Fig. 3.15) determined the position of H-5 that in turn correlated with a methyl at 1.33 ppm, indicating that this unit was a fucose α configured at the anomeric center on the basis of the anomeric proton shape and its chemical shift value. Analysis of carbon chemical shift values showed that both C-3 and C-4 were shifted at low field, indeed glycosylated, while substitution at C-2 was not apparent by the chemical shift of this carbon (69.3 ppm), but it was deduced by the long-range correlation connecting H-2 to C-1 of **B**. Similarly, to the NY-2A and OSy-NE5 glycans, proton and carbon chemical shifts of **B** was very similar, in agreement with its attribution as α-D-galactose.

With regard to **C**, its anomeric proton displayed only one COSY and TOCSY correlation (H-1/H-2), while the connections up to the methyl group were visible from H-2 due to the favorable proton-proton coupling constants. These information, combined with those of the HSQC and HMBC spectra identified **C** as a terminal rhamnose unit, α configured at the anomeric center. In addition, it was methylated at O-3 due to the low field C-3 value (^{13}C 80.3 ppm), as confirmed by the long-range correlation H-3/C-3$_{OMe}$ in the HMBC spectrum (Fig. 3.14) and acetylated at O-2, as suggested by H-2 low-field chemical shift (5.447 ppm). The anomeric signals of **E** and **F** coincided at 4.41 ppm and their carbon signals, at 105.6 and

103.9 ppm, were diagnostic of residues β configured at the anomeric centre. Discrimination of the two different spin systems was achieved by analysing the correlations in the TOCSY spectrum associated with isolated signals from each of them: H-2 for **F** (3.13 ppm) and H-5eq for **E** (3.18 ppm). The efficient propagation of magnetization in the TOCSY spectrum suggested that all ring protons (other than H-5$_{eq}$) were axial oriented, as occurs for *gluco* configured residues. Indeed, H-2 (or H-5$_{eq}$) of **F** and H-5$_{eq}$ of **E** had distinct correlations each compatible with a five carbon atoms monosaccharide: a xylose. Integration of these information with those of heteronuclear spectra, established that **E** was linked at O-4 of **A**, while **F** at O-4 of **D**.

Achievement of the CVM-1 glycoprotein structure (Fig. 3.16) was inferred by analyzing the HMBC spectrum (Fig.3.14), which disclosed that **A** was linked at O-3 of **D** and in turn had **B** at O-2 and **C** at O-3.

Interestingly, the chemical shifts of **A**, α-fucose, was different from that found for the same residue in PBCV-1 and in the two previously analyzed chloroviruses, even though it was substituted in a similar fashion. This change was revealed by examining the absolute configuration of all sugars: all monosaccharides had the same absolute configuration found for other glycans analyzed so far, except for the rhamnose residue (**C**), which was L instead of D. Thus, we can conclude that the ^{13}C chemical shift displacement with respect to standard values (Bock and Pedersen 1983) in oligosaccharides depends on stereochemical factors that rule the amplitude of the α- and β-glycosylation effects as a function of the relative stereochemistry of the sugars.
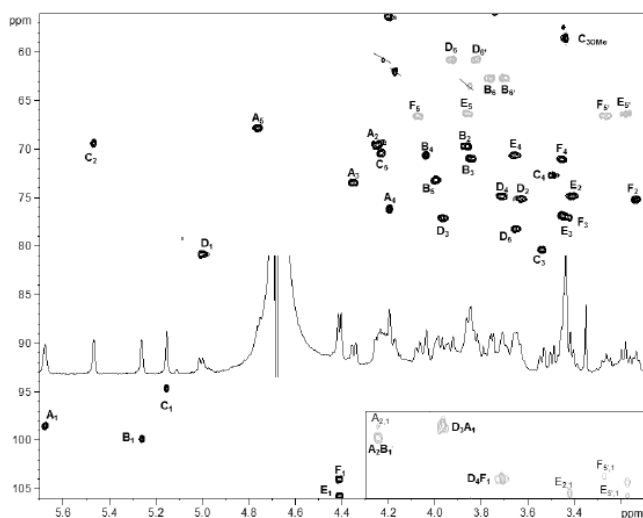
Figure 3.14 Expansion of CVM-1 glycopeptide HSQC spectrum measured at 600 MHz, 310 K; HMBC spectrum is reported at the right-bottom corner. (De Castro et al. 2016) Those appearing in gray in the HSQC spectrum represent carbons bearing two hydrogen atoms. Signals crossed indicate impurities or the peptide moiety of the glycopeptide.
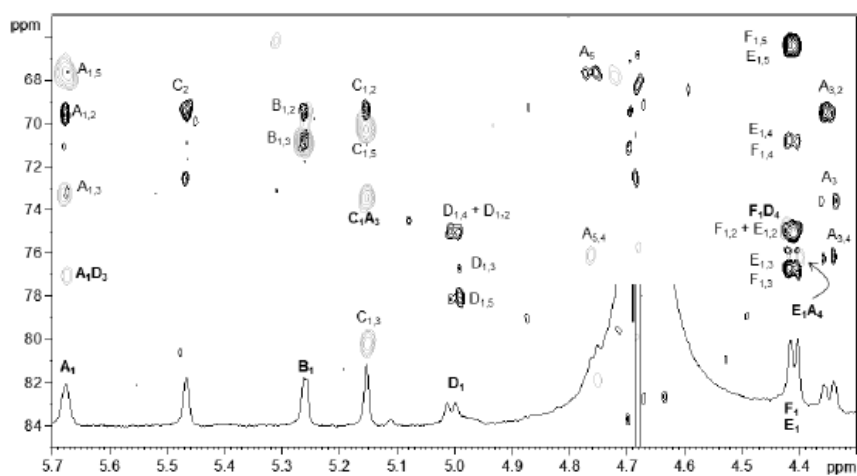
Figure 3.15 Superimposition of CVM-1 glycopeptide HSQCTOCSY (black) and HMBC (grey) spectra measured at 600 MHz, 310 K. (De Castro et al. 2016)
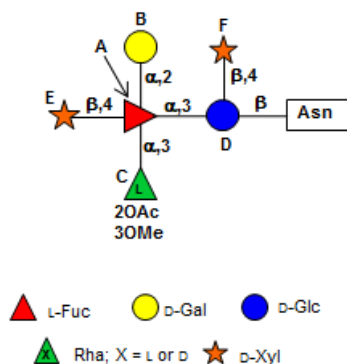
Figure 3.16 Structure of N-glycan of CVM-1 (De Castro et al. 2016). Letter labels are those used during NMR assignment.

Table 3.5 (600MHz, 310 K, D_2O) Proton and carbon chemical shifts deduced for CVM-1 glycopeptide. **C** is acetylated at O-2 ($^1H/^{13}C$ 2.19/21.6 ppm) and methylated at O-3 ($^1H/^{13}C$ 3.44/58.4 ppm).

		1	2	3	4	5 (5_{eq}- 5_{ax})	6-6'
A	^1H	5.68	4.24	4.34	4.20	4.76	1.33
2,3,4-α-L-Fuc	^{13}C	98.4	69.3	73.4	76.1	67.7	16.3
B	^1H	5.26	3.87	3.84	4.03	3.99	3.76; 3.70
t-α-D-Gal	^{13}C	99.7	69.4	70.9	70.5	73.0	62.5
C	^1H	5.15	5.47	3.54	3.50	4.23	1.35
α-L-2OAc3OMeRha	^{13}C	94.5	69.3	80.3	72.5	70.3	18.1
D	^1H	5.01	3.62	3.97	3.71	3.65	3.93; 3.83
3,4-β-D-Glc	^{13}C	80.7	75.0	77.0	74.7	78.1	60.7
E	^1H	4.41	3.41	3.42	3.65	3.86; 3.18	-
t-β-D-Xyl	^{13}C	105.6	74.6	76.9	70.4	66.3	-
F	^1H	4.41	3.13	3.45	3.45	4.07; 3.26	-
t-β-D-Xyl	^{13}C	103.6	75.1	76.8	70.9	66.5	-

3.3.2.4 NE-JV-1 (Pbi viruses)

Detailed NMR analysis supported from the NMR data of the previous glycopeptides, confirmed the presence of a terminal galactose (residue **B**) and glucose (**D**) *N*-linked to the asparagine.

Regarding **A**, close inspection of the COSY spectrum (Fig. 3.17) revealed that the signal at 4.67 ppm included two different anomeric protons, each with a distinct correlation to a H-2 signal. Accordingly, the two anomeric signals at 4.67 ppm were labelled **A** and **A'** and for each of them it was possible to trace the proton connectivity up to H-4, beyond which both COSY and TOCSY spectra failed in detecting H-5 as occurs for *galacto* configured residues. In contrast, the T-ROESY experiment (Fig. 3.18) allow to determined H-5, because connected to H-4 and H-3 via a strong and a medium NOE effect, respectively; in both cases, H-5 was connected to a methyl group. Thus, **A** (or **A'**) was a fucose, α configurated at the anomeric centre, as inferred from the proton and carbon chemical shift of the anomeric signal (Table 3.6). Analysis of the carbon chemical shift values of **A** (or **A'**) showed similar pattern found for CVM-1, indicating that C-2, C-3 and C-4 were glycosylated and suggested that the nature of the residue linked at O-3 was a rhamnose with L, and not D, absolute configuration. Moreover, T-ROESY spectrum revealed that **A** (or **A'**) was linked at O-3 of **E**.

Regarding **C** and **D**, the TOCSY spectrum (Fig. 3.17) only displayed the correlation with H-2, which instead had correlations with all the other ring protons, including a methyl group, indeed **C** and **D** were two rhamnose units. This information combined with those of the HSQC and HMBC spectra, identified these residues as terminal rhamnose units, α configurated at the anomeric center. Both residues were methylated at O-3 due to the low field C-3 value (^{13}C of **C**: 80.2 ppm; ^{13}C of **D**: 81.3 ppm), also confirmed by the long-range correlations H-3/C-3$_{OMe}$ and H-3$_{OMe}$/C-3 in the HMBC spectrum (Fig. 3.19). Differently from **D**, the O-2 position in **C** was acetylated as suggested by the low-field chemical shift of H-2 (5.61 ppm) along with the presence of an acetyl group in the spectrum ($^{1}H/^{13}C$ 2.19/21.5 ppm). In addition, the T-ROESY spectrum showed that **C** was connected at O-3 of **A**, while **D** was linked at O-3 of **A'**.

With regard the last signals, the HSQC spectrum displayed two anomeric carbon signals at ca. 105 and 103 ppm (Fig. 3.19), diagnostic of residues β configured at the anomeric centre, but the corresponding region in the proton spectrum (ca. 4.42 ppm) had a complex pattern of anomeric signals. As has occurred for the CVM-1 glycopeptide, these signals arise from the β-xylose units either proximal or distal to the peptide backbone and their full attribution is possible starting spectra interpretation from the signal of H-5 in axial position (H-5_{ax}) instead of the anomeric proton. Therefore, COSY spectrum combined with information from HSQC, identified four different H-5_{ax} protons, labelled **F-H**, each correlated with the geminal H-5_{eq} proton and to H-4; these protons did not overlap, which enabled the tracing of all correlations up to H-1. Thus, two xylose units were identified, one having H-2 at high field (ca. 3.14 ppm) containing residues **G** and **I**, and one with H-2 at a lower field (3.41 ppm) with residues **F** and **H**. The anomeric proton of **F** (4.44 ppm) did not overlap with the others and its carbon chemical shift was at 105.6 ppm; the same value was given to **H** due to the analogies between the proton chemical shifts of the two residues. Accordingly, anomeric carbon chemical shift of **G** and **I** was at 103.8 ppm. Analysis of the T-ROESY spectrum (Fig. 3.18) established that **F** was linked to O-4 of **A'**, **H** at O-4 of **A**, while **G**, or **I**, was linked at O-4 of **E** (Fig. 3.20). Thus, **F** and **H** were identified as distal xylose units, while **G** or **I** were proximal at the peptide backbone.
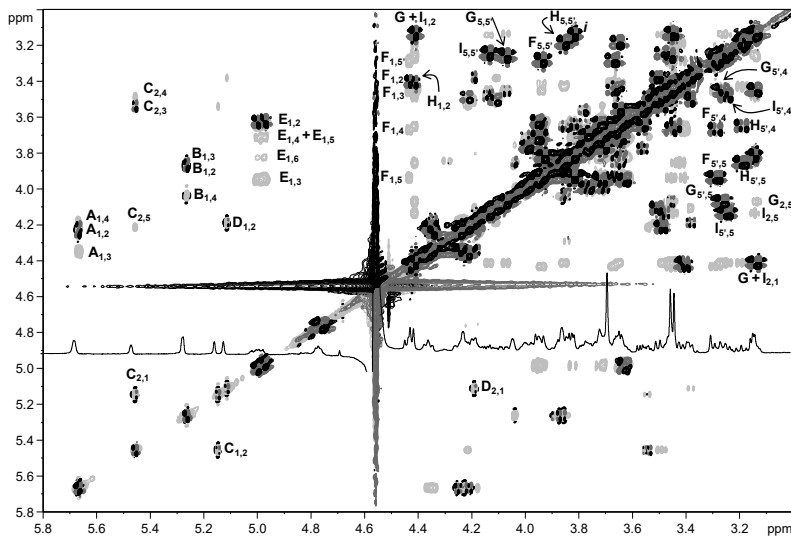
Figure 3.17 Superimposition of NE-JV-1 glycopeptide TOCSY (pale grey) and COSY (black and dark grey) spectra at 600 MHz, 323 K. (Speciale et al. 2017)
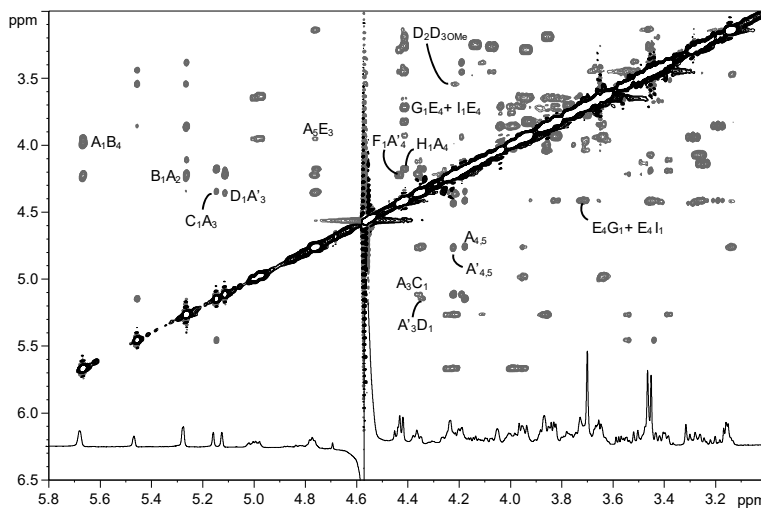
Fig. 3.18 Expansion of NE-JV-1 glycopeptide T-ROESY spectrum measured at 600 MHz, 323K. (Speciale et al. 2017)
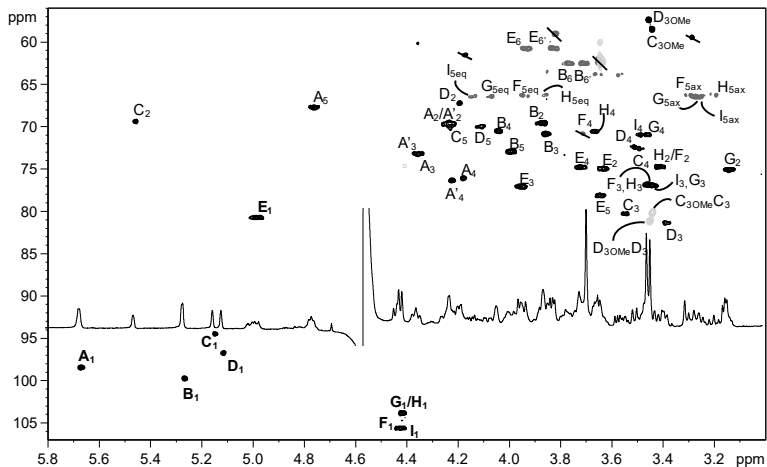
Figure 3.19 Superimposition of NE-JV-1 glycopeptide HSQC (black and dark grey) and HMBC (pale grey) spectra at 600 MHz, 323 K. (Speciale et al. 2016) Carbon bearing two hydrogen atoms are coloured in dark grey. Signals crossed are related to the peptide moiety of the glycopeptides.
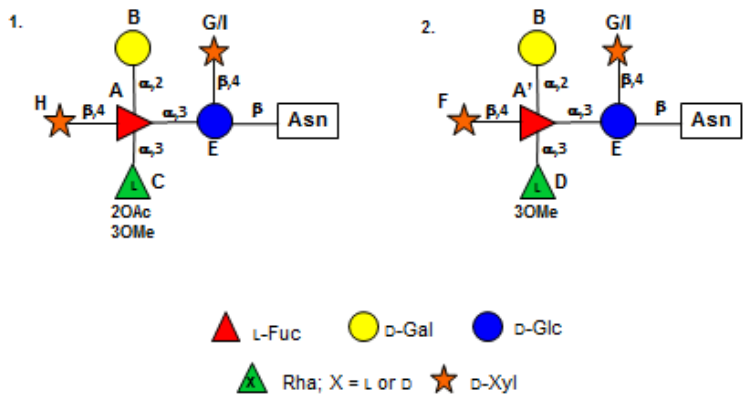
Figure 3.20 *N*-glycans structures obtained from NE-JV-1. (Speciale et al. 2017)

Table 3.6 (600MHz, 323 K, D$_2$O) Proton and carbon chemical shifts deduced for NE-JV-1 glycopeptide. **C** is acetylated at O-2 (^{1}H/^{13}C 2.19/21.6 ppm) and methylated at O-3 (^{1}H/^{13}C 3.44/58.4 ppm). **D** is substituted at position O-3 with a methyl group (^{1}H/^{13}C 3.45/57.3 ppm).

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.67	4.25	4.35	4.18	4.76	1.33
2,3,4-α-L-Fuc	^{13}C	98.3	69.6	73.1	76.0	67.7	16.3
A'	^{1}H	5.67	4.22	4.36	4.22	4.76	1.33
2,3,4-α-L-Fuc	^{13}C	98.3	69.6	73.1	76.3	67.7	16.3
B	^{1}H	5.26	3.87	3.86	4.04	3.99	3.76; 3.71
t-α-D-Gal	^{13}C	99.5	69.5	70.8	70.5	72.8	62.5
C	^{1}H	5.15	5.46	3.55	3.49	4.22	1.35
α-L-2OAc,3OMeRha	^{13}C	94.4	69.3	80.2	72.5	70.1	18.9
D	^{1}H	5.11	4.19	3.38	3.51	4.11	1.31
t-α-L-3OMeRha	^{13}C	96.7	67.1	81.3	72.4	70.0	18.9
E	^{1}H	4.99	3.63	3.95	3.72	3.64	3.93; 3.82
3,4-α-D-Glc	^{13}C	80.7	74.9	77.0	74.8	78.1	60.8
F	^{1}H	4.44	3.41	3.44	3.67	3.94; 3.29	-
t-β-D-Xyl	^{13}C	105.6	74.7	76.9	70.5	66.2	-
G	^{1}H	4.42	3.14	3.45	3.46	4.08; 3.27	-
t-β-D-Xyl	^{13}C	103.8	75.0	76.8	70.9	66.4	-
H	^{1}H	4.42	3.41	3.44	3.661	3.86; 3.19	-
t-β-D-Xyl	^{13}C	105.6	74.7	76.9	70.5	66.2	-
I	^{1}H	4.42	3.14	3.45	3.49	4.14; 3.26	-
t-β-D-Xyl	^{13}C	103.8	75.0	76.8	70.9	66.4	-

3.3.2.5 ATCV-1 (SAG viruses)

The proton spectrum of ATCV-1 glycopeptide presented several similarities with that of CVM-1: anomeric protons of **A**, **B** and **D** had the same chemical shift. Moreover, analysis of homonuclear and heteronuclear spectra suggested that proton and carbon chemical shift values (Table 3.7) of **A**, **B** and **D** were very similar to those found for CVM-1, in agreement with their attribution as α-fucose, α-galactose and β-glucose, respectively. **C** residue, similarly to that found for **D** residue of NE-JV-1 glycan, was a rhamnose α configured, which presented a methyl group at position O-3, in agreement with the low field C-3 value (81.4 ppm). All these residues were linked as in the other glycopeptides.

Of note, a more careful inspection of the TOCSY spectrum indicated that the fucose unit was heterogeneous and it was split in **A** and **A'**, differing mainly by the proton chemical shifts of H-3 and H-4.

Regarding the xylose units, the HSQC spectrum (Fig. 3.21) displayed two anomeric carbon signals at 105 and 103 ppm, diagnostic of β configured residues, but the inspection of the COSY spectrum (Fig. 3.22) disclosed a high complex pattern: it showed several H-1/H-2 cross peaks at ca. 4.42 ppm, suggesting the presence of a group of the distal xylose units (named **E**) and a group of proximal xyloses (named **F**). Attribution of the different spin systems was achieved for two of the four distal xylose units (**E** and **E'**) and two proximal residues (**F** and **F'**). Their attribution was possible starting spectra interpretation from the H-5 signals, because they were not completely overlapped.

Regarding **F'**, the COSY spectrum displayed a cross peak 4.39/3.20 relative to diastereotopic methylene protons, H-5$_{eq}$ and H-5$_{ax}$, that was shifted at highfield (64.0 ppm) with respect the standard value, because O-4 was methylated, accordingly C-4 was shifted at low field (80.4 ppm). COSY spectrum connected H-4 to H-3 and in turn with its the other ring protons. Similar approach was used to determine other xylose units. Moreover, T-ROESY confirmed that **F** group was linked at O-4 of **D**.

E' presented the same H-1 and H-5$_{eq,ax}$ chemical shifts as **F**, but diverged for the values of the other ring protons. In particular, the carbon signal C-4 resonated at

80.0 ppm suggesting that this position was methylated, accordingly with the modest displacement observed for C-3 and C-5.

Attribution of **E** was inferred combining COSY, TOCSY and HSQC information: it was a terminal xylose unit, not methylated, and linked at O-4 of the fucose unit, as found for **E'**. The other two unsigned xylose units of the group **E**, had correlations similar to those reported for **E** and **E'**, suggesting that one of this was methylated at O-4 and the other was not. Analysis of the T-ROESY spectrum established that **E'** was linked to O-4 of **A'**, whereas **E** with O-4 of **A**. This explained why the fucose was split: it conveyed the presence of methylated xylose or not. Thus, ATCV-1 produced several glycans that differ from the presence of the methyl group on the xylose units (all possible structures were represented in figure 3.23).
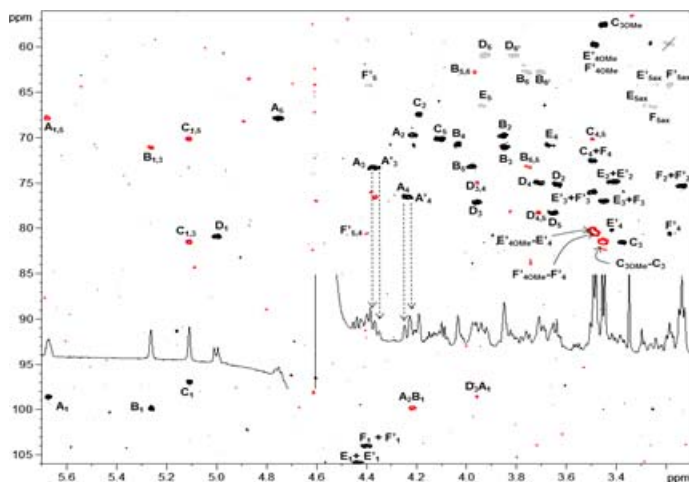
Figure 3.21 Superimposition of HSQC (black/grey) and HMBC (red) spectra of ATCV-1 glycopeptide. (De Castro et al. 2016) Spectra were recorded at 600 MHz, 318 K. Those appearing in grey in the HSQC spectrum have the opposite sign with respect to the other and represent carbons bearing two hydrogen atoms. Signals crossed are related to the peptide moiety or to impurities.
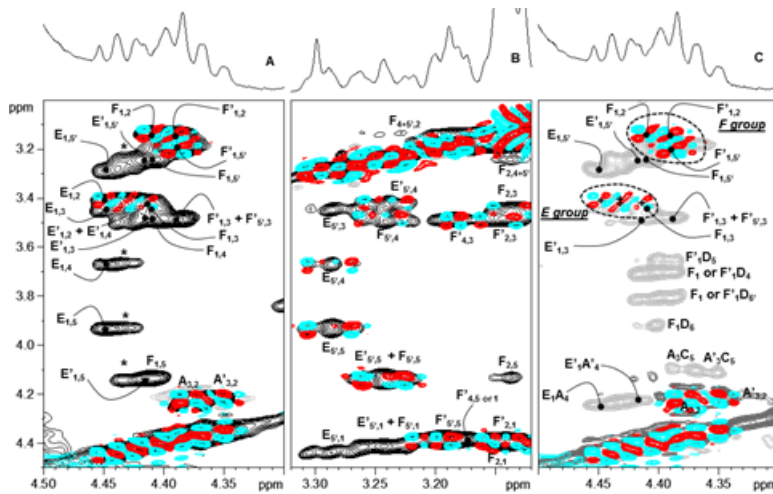
Figure 3.22 Homonuclear spectra of ATCV-1 glycopeptides recorded at 600 MHz and 318 K. (De Castro et al. 2016) A) Superimposition of TOCSY (black) and COSY (turquoise/red) spectra detailing the anomeric region of the xylose residues. B) TOCSY and COSY spectra relative to the region of H-5' signals of distal and proximal xylose units. C) Superimposition of T-ROESY (grey) and COSY (turquoise/red) spectra detailing the anomeric region of xylose residues.
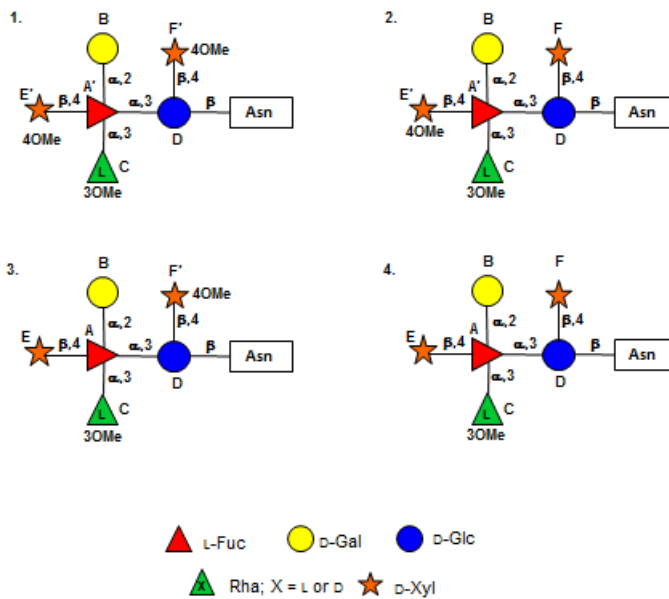
Figure 3.23 Structures of *N*-glycans from ATCV-1 virus. (De Castro et al. 2016)

Table 3.7 (^{1}H 600 MHz, ^{13}C 950 MHz, 310 K, D$_2$O) Proton and carbon chemical shifts deduced for ATCV-1 glycopeptide. **C** is methylated at O-3 (^{1}H/^{13}C 3.45/ and 3.46 /57.4 ppm) and **E'** and **F'** at O-4 (^{1}H/^{13}C 3.49/59.5 and 3.48/59.7 ppm, respectively).

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.67	4.22	4.41	4.25	4.74	1.33
2,3,4-α-L-Fuc	^{13}C	98.4	69.5	73.1	76.3	67.6	16.4
A'	^{1}H	5.66	4.22	4.36	4.23	4.74	1.33
2,3,4-α-L-Fuc	^{13}C	98.6	69.5	73.1	76.3	67.6	16.4
B	^{1}H	5.26	3.85	3.85	4.04	4.98	3.77; 3.70
t-α-D-Gal	^{13}C	99.6	69.6	70.5	70.5	72.9	62.5
C	^{1}H	5.11	4.19	3.38	3.50	4.11	1.31
t-α-L-3OMeRha	^{13}C	96.7	67.2	81.4	72.3	69.9	18.2
D	^{1}H	5.00	3.64	3.96	3.71	3.65	3.93; 3.81
3,4-β-D-Glc	^{13}C	80.7	75.0	77.0	74.7	78.1	60.7
E	^{1}H	4.45	3.41	3.46	3.67	3.94; 3.29	-
t-β-D-Xyl	^{13}C	105.6	74.6	76.8	70.6	66.3	-
E'	^{1}H	4.42	3.42	3.51	3.42	4.15; 3.26	-
t-β-D-4OMeXyl	^{13}C	105.6	74.7	75.8	80.0	64.0	-
F	^{1}H	4.41	3.14	3.45	3.49	4.14; 3.24	-
t-β-D-Xyl	^{13}C	103.8	75.1	76.8	72.4	66.3	-
F'	^{1}H	4.39	3.15	3.49	3.18	4.39; 3.20	-
t-β-D-4OMeXyl	^{13}C	103.8	75.1	75.7	80.4	64.0	-

3.3.2.6 TN603 (SAG viruses)

The proton spectrum was very similar to those found for ATCV-1 and set of NMR analysis disclosed the presence of an hyperbranched fucose (residue **A**), a terminal galactose (**B**), a terminal rhamnose with a methyl group at position O-3 (**C**) and glucose (**D**) *N*-linked to the asparagine. These residues were linked as in the other glycopeptides. Of note, **C** was not a single residue, but it indicated a group of different units of rhamnose 3OMe.

Similarly to ATCV-1, the xylose units were heterogeneous (group **E**, that contains three anomeric protons, and two for **F**) and, also in this case, the attribution of the xylose units was inferred starting from H-5. The analysis disclosed that the **E** group (**E**, **E'** and **E"**) was a terminal xylose methylated at position O-4 and linked to O-4 of fucose residue, even though the residues **E** and **E"** were not equivalent: C-3 of **E"** was upfield shifted with respect to C-3 of **E**, while the opposite was verified for C-4 chemical shifts (Table 3.8). The same pattern was found for **F** and **F"**. These variations of chemical shifts were compatible with the presence of ketose residue linked at O-4 of **E"** and **F"**. Indeed, a ketofuranose residue was individuated (**G**). Starting from H-5/C-5 signal in the HSQC spectrum, it was possible identified the H-4 and in turn the H-3. Inspection of TOCSY spectrum revealed the presence of three different H-3 protons, suggesting the presence of three different units, named **G**, **G'** and **G"**. The carbons C-3 of **G** and **G '** had a long-range correlation with their own C-1, thus C-1 of **G"** was given by exclusion. Comparing the carbon chemical shift with those reported in literature (Gorshova et al. 1995), all **G** residues were identified as β-xylulofuranose units. Of note, **G'** was methylated at O-1 as suggested by its C-1 chemical shift value (73.8 ppm) and by the long-range connection to a methyl group. HMBC spectrum (Fig. 3.24) showed a long-range correlation, indicating that the xylulose units were located at O-4 of **F"**, corroborating the assumption that **E"** was also glycosylated with a ketose.

These results revealed that TN603 was a complex mixture of glycans, that differ for the substitution of both distal and proximal xylose (Fig. 3.25). In addition, the xylulose could be methylated or was not.
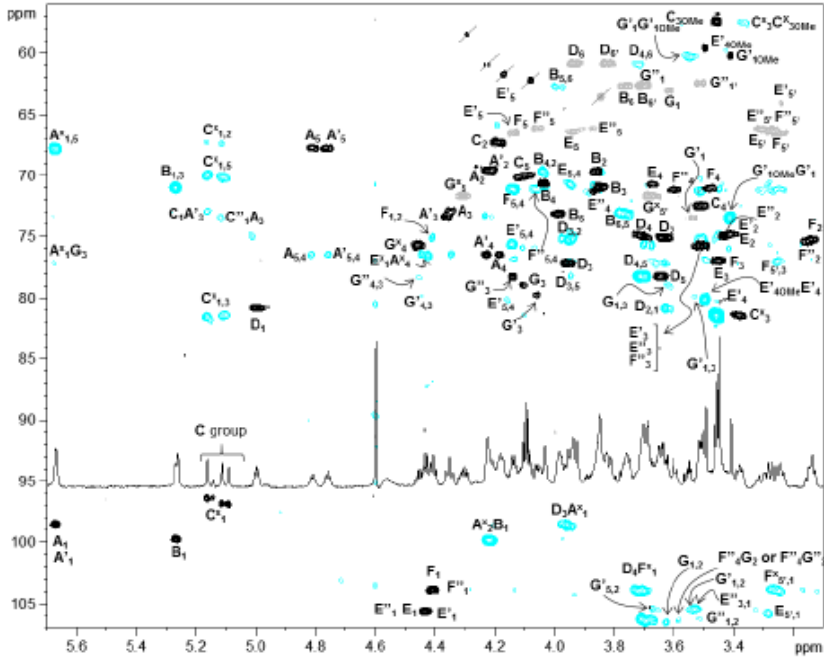
Figure 3.24 Superimposition of HSQC (black/grey) and HMBC (turquoise) spectra obtained from TN603 glycopeptide, measured at 600 MHz, 318 K. (De Castro et al. 2016) In grey were reported carbons bearing two hydrogen atoms. Signals crossed are related to the peptide moiety of the glycopeptide or to impurities.
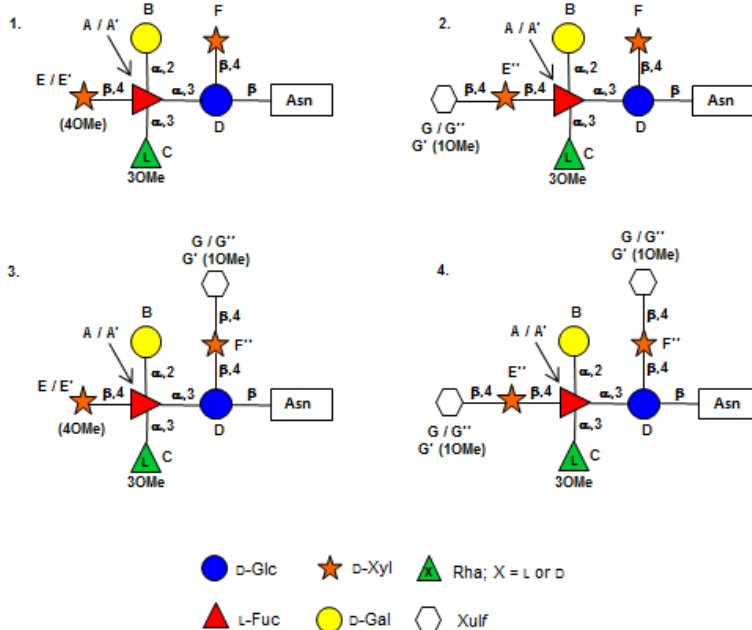
Figure 3.25 Structures of *N*-glycans obtained from TN603 glycopeptide. (De Castro et al. 2016)

Table 3.8 (^{1}H 950 MHz, ^{13}C 237.5 MHz, 318 K, D$_2$O) Proton and carbon chemical shifts deduced for TN603 glycopeptide. **C** group is methylated at O-3 (O-CH$_3$ at ^{1}H/^{13}C 3.45/57.3 ppm), **E"** at O-4 (O-CH$_3$ at 3.50/59.5), **G'** at O-1 (O-CH$_3$ at 3.41/60.1).

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.67	4.22	4.36	4.224	4.81	1.33
2,3,4-α-L-Fuc	^{13}C	98.4	69.5	73.2	76.4	67.7	16.4
A'	^{1}H	5.67	4.22	4.35	4.18	4.76	1.33
2,3,4-α-L-Fuc	^{13}C	98.4	69.5	72.9	76.4	67.7	16.4
B	^{1}H	5.26	3.85	3.85	4.04	3.98	3.76; 3.71
t-α-D-Gal	^{13}C	99.6	69.6	70.9	70.5	73.0	62.5
C	^{1}H	5.16	4.17	3.38	3.50	4.09	1.31
t-α-L-3OMeRha	^{13}C	96.3	67.2	81.3	72.3	69.8	18.2
C'	^{1}H	5.14	4.18	3.38	3.51	4.09	1.31
t-α-L-3OMeRha	^{13}C	96.3	67.2	81.3	72.3	69.8	18.2
C"	^{1}H	5.11	4.19	3.38	3.51	4.11	1.31
t-α-L-3OMeRha	^{13}C	96.8	67.1	81.3	72.3	69.7	18.2
C'''	^{1}H	5.09	4.20	3.39	3.51	4.11	1.31
t-α-L-3OMeRha	^{13}C	96.7	67.1	81.3	72.3	69.7	18.2
D	^{1}H	5.00	3.63	3.96	3.71	3.65	3.93; 3.82
3,4-α-D-Glc	^{13}C	80.7	75.0	77.0	74.8	78.2	60.8
E	^{1}H	4.43	3.40	3.45	3.67	3.94; 3.29	-
t-β-D-Xyl	^{13}C	105.5	74.7	76.8	70.6	66.3	-
E'	^{1}H	4.42	3.42	3.50	3.42	4.17; 3.24	-
β-D-4OMeXyl	^{13}C	105.6	74.7	75.8	80.0	64.0	-
E"	^{1}H	4.43	3.43	3.52	3.86	3.38; 3.32	-
4-β-D-Xyl	^{13}C	105.5	74.8	75.8	71.3	64.0	-
F	^{1}H	4.41	3.14	3.44	3.47	4.14; 3.25	-
t-β-D-Xyl	^{13}C	104.8	75.3	76.7	70.9	66.3	-
F"	^{1}H	4.41	3.15	3.50	3.63	4.05; 3.27	-
4-β-D-Xyl	^{13}C	104.8	75.2	75.7	71.0	66.0	-
G	^{1}H	3.62 (x2)	-	4.10	4.46	4.32; 3.66	-
t-β-Xul*f*	^{13}C	63.0	106.2	78.9	75.5	71.6	-
G'	^{1}H	3.54 (x2)	-	4.06	4.44	4.32; 3.66	-
t-β–1OMeXul*f*	^{13}C	73.8	105.5	79.7	75.5	71.6	-
G	^{1}H	3.70; 3.51	-	4.14	4.54	4.32; 3.66	-
t-β-Xul*f*	^{13}C	62.3	106.2	78.3	75.5	71.6	-

3.3.2.7 MT325 (Pbi viruses)

The low field region of the proton spectrum contained the same anomeric signals that were found for TN603, but group of 3OMe rhamnose (**C**) was present in minor amount. Moreover, the carbinolic region presented a new methyl group at 3.53 ppm, whereas that at 3.41 ppm, attributed at O-1 of Xul*f* in TN603, was absent. Analysis of the homonuclear spectra disclosed the presence of different type of xylulofuranose along with a complex pattern of xylose residues, which were determined starting from the axial oriented H-5 protons, similarly to the previous glycopeptides. Accordingly, different types of xyloses were found: distal xylose could be terminal (**E**) or could be binding with a Xul*f* at O-4; instead the proximal xylose could be terminal unit (**F**), or could have a methyl group at O-4 (**F'**), or could have a Xul*f* at O-4 (**F"**).

As just revealed, also MT325 presented the xylulose residues, and the inspection of the H-4/C-4 region (ca. 4.50/75.0 ppm) in the HSQC spectrum (Fig. 3.26) suggested that three different types occurred. Two of them (**G** and **G"**) were terminal, as found for TN603; instead, the third residue (**G'''**) differed for the H-3/C-3 chemical shift value (3.67/87.9 ppm; Table 3.9). This displacement of the carbon chemical shift at low field value, suggested that this position was methylated, as confirmed by the NOE effect between the methyl at 3.53 ppm and H-3 proton. This result was also supported by a weak cross peak between H-1 of distal xylose residue with H-4 of **G'''**.

Thus, similarly to was found for TN603, also MT325 produced a complex mixture of glycopeptides (Fig. 3.27), that differ in the substitution pattern of the proximal and distal xyloses.

Of note, the distal xylose in MT325 never presented the methyl group, that instead could be occur on the proximal xylose, on the contrary of TN603.
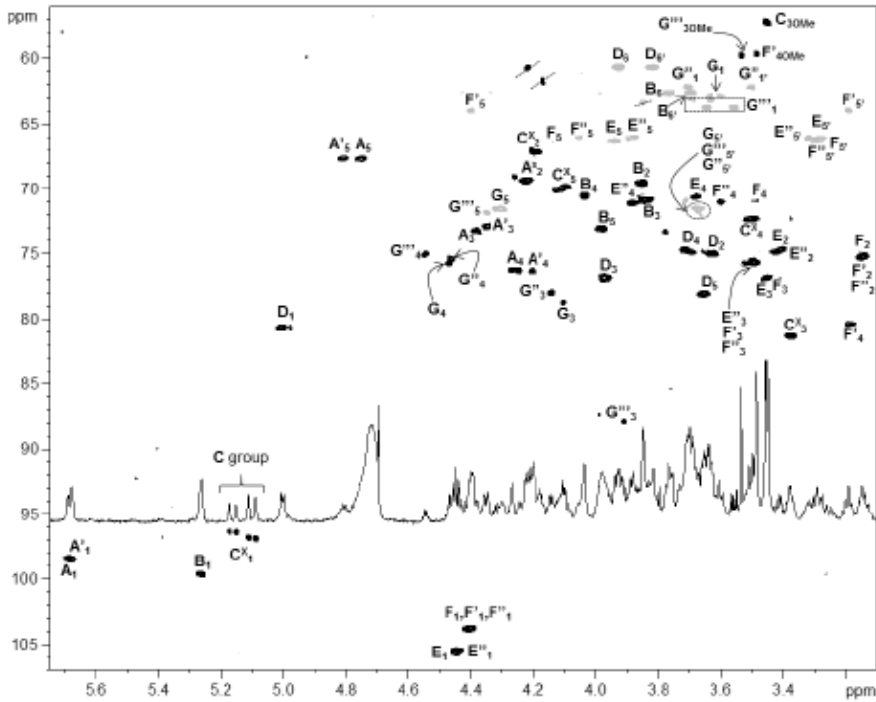
Figure 3.26 HSQC (black/grey) spectrum of MT325 glycopeptide measured at 950 MHz, 310 K. (De Castro et al. 2016) Those appearing in grey have the opposite sign with respect to the other and represent carbons bearing two hydrogen atoms. Signals crossed are related to the peptide moiety of the glycopeptide or to impurities.
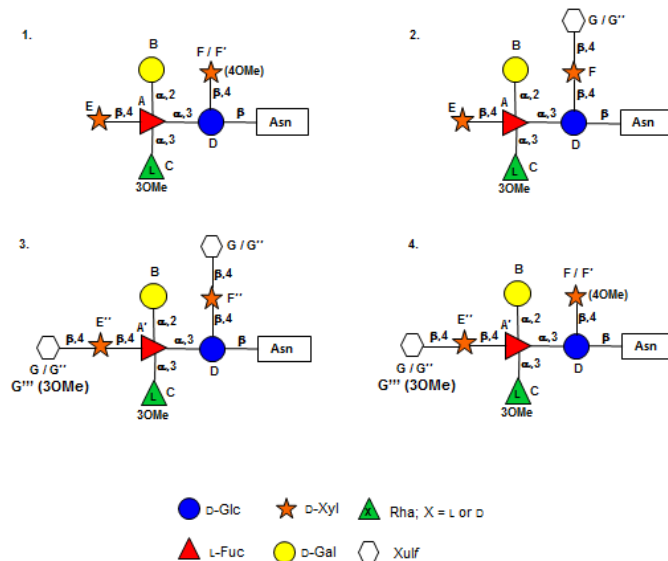
Figure 3.27 Structures of *N*-glycans from MT325 virus. (De Castro et al. 2016)

Table 3.9 (^{1}H 950 MHz, ^{13}C 237.5 MHz, 310 K, D$_2$O) Proton and carbon chemical shifts deduced for MT325 glycopeptide. **C** group is methylated at O-3 (O-CH$_3$ at ^{1}H/^{13}C 3.45/57.1 ppm), **F'** at O-4 (O-CH$_3$ at at ^{1}H/^{13}C 3.48/59.5), **G'''** at O-3 (O-CH$_3$ at 3.53/60.0). *** H-1/C-1 could not be assigned unequivocally.

|		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.69	4.22	4.39	4.27	4.75	1.33
2,3,4-α-L-Fuc	^{13}C	98.63	69.53	73.08	76.48	67.85	16.33
A'	^{1}H	5.68	4.22	4.35	4.20	4.81	1.34
2,3,4-α-L-Fuc	^{13}C	98.3	69.3	72.8	76.2	67.5	16.1
B	^{1}H	5.26	3.85	3.84	4.03	3.98	3.76; 3.70
t-α-D-Gal	^{13}C	99.6	69.4	70.7	70.4	73.0	62.6
C	^{1}H	5.17	4.18	3.37	3.51	4.10	1.31
t-α-L-3OMeRha	^{13}C	96.3	67.1	81.3	72.2	69.8	18.0
C'	^{1}H	5.15	4.18	3.38	3.54	4.10	1.31
t-α-L-3OMeRha	^{13}C	96.3	67.1	81.3	72.2	69.8	18.0
C''	^{1}H	5.11	4.20	3.38	3.49	4.12	1.31
t-α-L-3OMeRha	^{13}C	96.8	67.1	81.3	72.2	70.0	18.0
C'''	^{1}H	5.09	4.21	3.38	3.50	4.12	1.31
t-α-L-3OMeRha	^{13}C	96.8	67.1	81.3	72.2	70.0	18.0
D	^{1}H	5.00	3.62	3.97	3.70	3.65	3.92; 3.82
3,4-β-D-Glc	^{13}C	80.6	74.1	76.8	74.7	78.0	60.6
E	^{1}H	4.45	3.40	3.45	3.68	3.94; 3.29	-
t-β-D-Xyl	^{13}C	105.6	74.6	76.8	70.76	66.2	-
E''	^{1}H	4.44	3.42	3.49	3.88	3.88; 3.31	-
4-β-D-Xyl	^{13}C	105.6	74.8	75.6	71.0	65.9	-
F	^{1}H	4.41	3.14	3.45	3.47	4.14; 3.25	-
t-β-D-Xyl	^{13}C	103.8	75.1	76.8	70.9	66.1	-
F'	^{1}H	4.40	3.15	3.49	3.18	4.38; 3.19	-
β-D-4OMeXyl	^{13}C	103.8	75.41	75.6	80.3	63.9	-
F''	^{1}H	4.40	3.15	3.50	3.60	4.05; 3.27	-
4-β-D-Xyl	^{13}C	103.8	75.1	75.6	70.9	66.0	-
G	^{1}H	3.63; 3.59	-	4.10	4.47	4.32; 3.67	-
t-β-Xulf	^{13}C	62.8	ND	78.7	75.7	71.5	-
G''	^{1}H	3.70; 3.50	-	4.14	4.46	4.30; 3.68	-
t-β-Xulf	^{13}C	62.3	ND	78.0	75.3	71.5	-
G'''	^{1}H	***	-	3.67	4.54	4.35; 4.04	-
t-β-3OMeXulf	^{13}C	***	ND	87.9	75.0	71.7	-

3.3.3 MALDI mass spectrometry

In collaboration with Professor Garozzo, MALDI MS spectrometry was performed on five of seven chloroviruses. This analysis not only confirmed the NMR results of all glycopeptides, but was also used to identify the glycosylation sites of the MCPs from NY-2A and CVM-1 (Table 3.10).

Regarding NY-2A, it presents the same glycosylation sites that occurred from PBCV-1 (Asn^{280}, Asn^{302}, Asn^{399}, Asn^{406}) along with two additional, Asn^{54} and Asn^{291}(Table 3.10). This finding is interesting because Asn^{54} is located in a conserved region of all MCPs, but it is not glycosylated in PBCV-1 and CVM-1; instead Asn^{291} is not conserved, it exists only in the NY-2A B617L paralogue. Moreover, this analysis suggests that the glycosylation is site-specific, indeed the hexasaccharide is found at Asn^{54}, Asn^{280}, Asn^{291} (Fig. 3.28), while the nonasaccharide occurs at Asn^{302}, Asn^{399}, Asn^{406} (Fig. 3.29). Interestingly, the presence of five (B585L) or six (B617L) glycosylation sites did not fit with the average number of glycans calculated for these MCPs, suggesting that not all the possible sites are always occupied in each paralogue.

CVM-1, instead, presented four glycosylation sites: Asn^{47}, Asn^{279}, Asn^{285} and Asn^{293}. All positions were equivalent to those found glycosylated in PBCV-1, except Asn^{47}, which is the only one found in a typical sequon (NGS) and its glycosylation only occurs in CVM-1, even though it is conserved in all MCPs. The MS/MS spectra, showing the distinctive hexasaccharide glycoforms linked to four different glycosylation sites are reported in figures 3.30 and 3.31. These spectra showed a fragment at m/z 976, corresponding to C ions retaining the Asn amino group, revealing the molecular weight of each associated peptide moiety.

Table 3.10 Glycosylation sites that occurred from NY-2A and CVM-1 in comparison with those of PBCV-1.

Virus	Gene code	Asn sequon found
PBCV-1	A430L	^{280}NIPG; ^{302}NTGT; ^{399}NTET; ^{406}NTAT
NY-2A	B585L	^{54}NKVS; ^{280}NIPG; ^{302}NTGT; ^{399}NTET; ^{406}NTAT
NY-2A	B617L	^{54}NKVS; ^{280}NIPG; ^{291}NVAT; ^{302}NTGT; ^{399}NTET; ^{406}NTAT
CVM-1	548L	^{47}NGSV; ^{279}NLTA; ^{285}NVGY; ^{293}NTAV
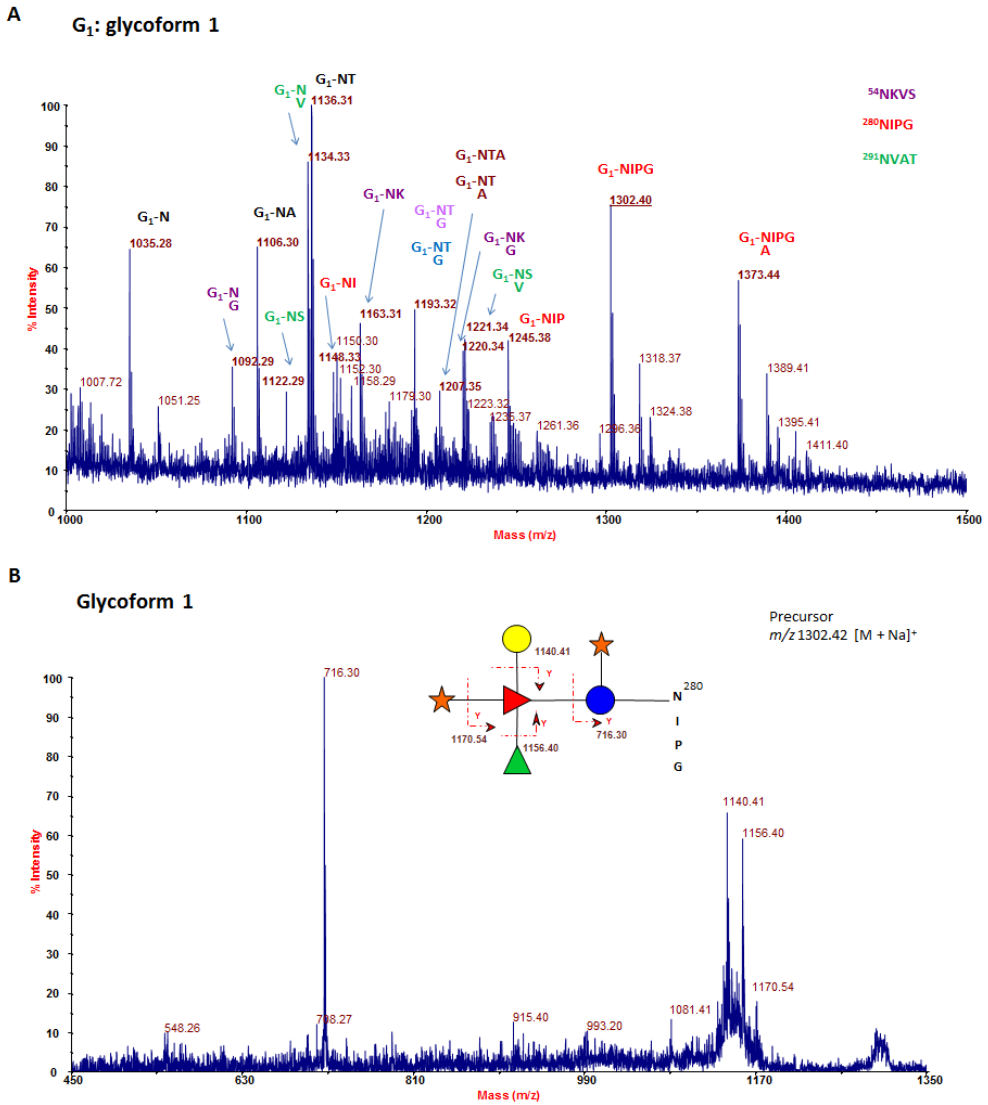
Figure 3.28 Low mass (A) MALDI-TOF MS profile of the glycopeptide mixture obtained from NY-2A MCPs. B) MS/MS analysis of selected peaks (underlined in the MALDI spectrum), relative to the first glycoform (a hexasaccharide). (De Castro et al. 2016)
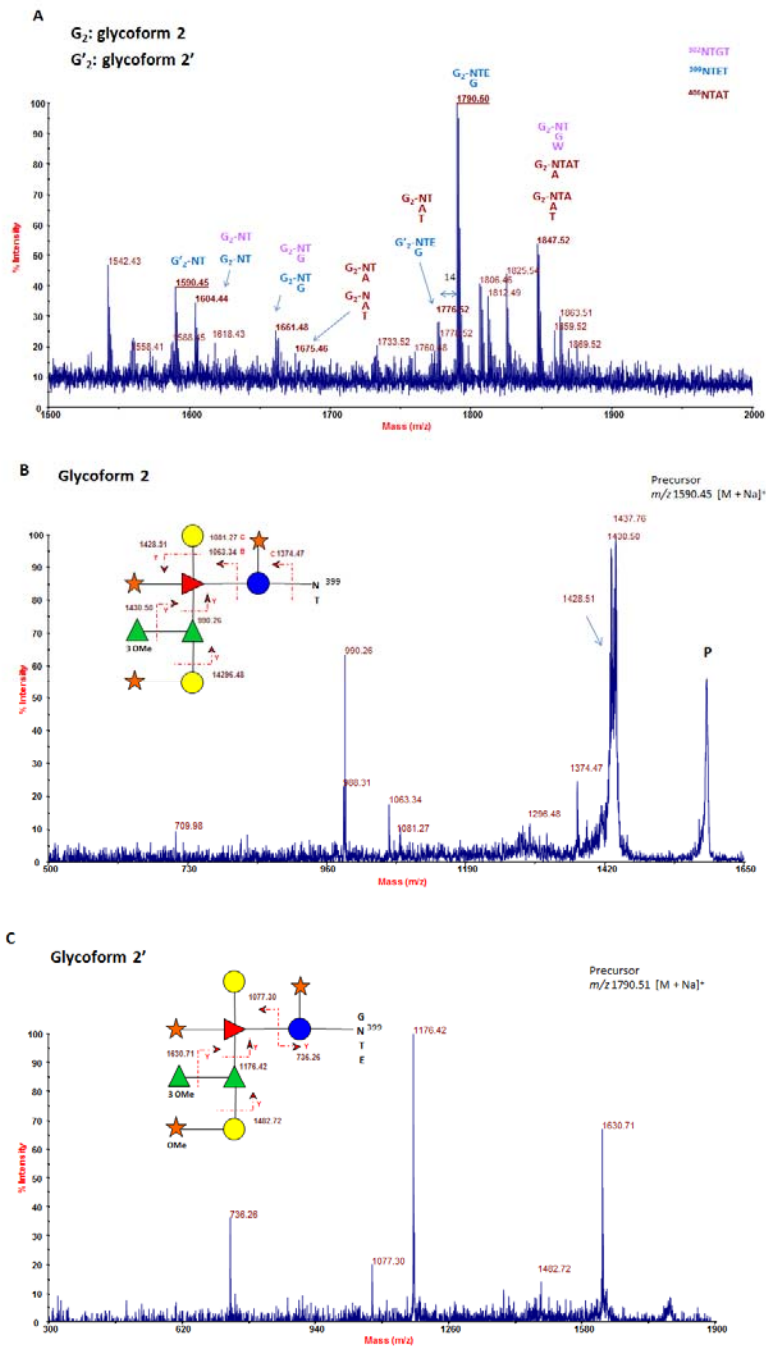
Figure 3.29 High mass (C) MALDI-TOF MS profiles of the glycopeptide mixture obtained from NY-2A MCPs. B and C) MS/MS analysis of selected peaks (m/z 1590.45 and 1790.51, respectively) of the glycoforms 2 and 2', consisting of a nonasaccharide with a not stoichiometric methyl at the galactose linked xylose unit. (De Castro et al. 2016)
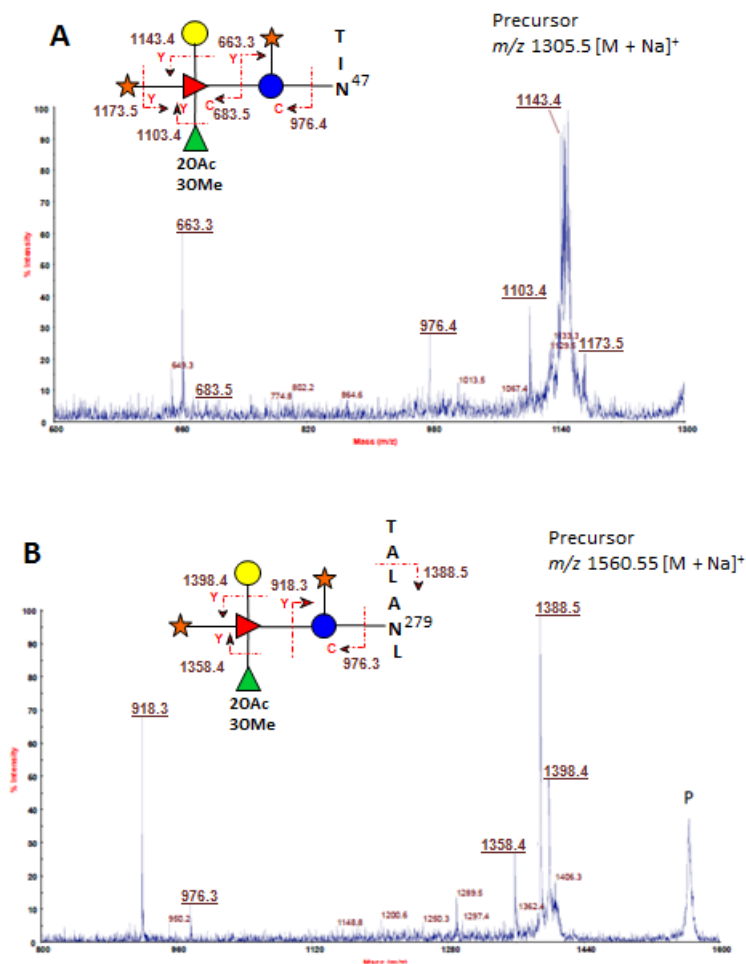
Figure 3.30 MS/MS analysis of the most significant molecular ions from CVM-1 glycopeptide fraction, showing four possible glycosylation sites: here, the [47]Asn (A) and [279]Asn (B) sites are showed, both occupied by a hexasaccharide glycoform. (De Castro et al. 2016)
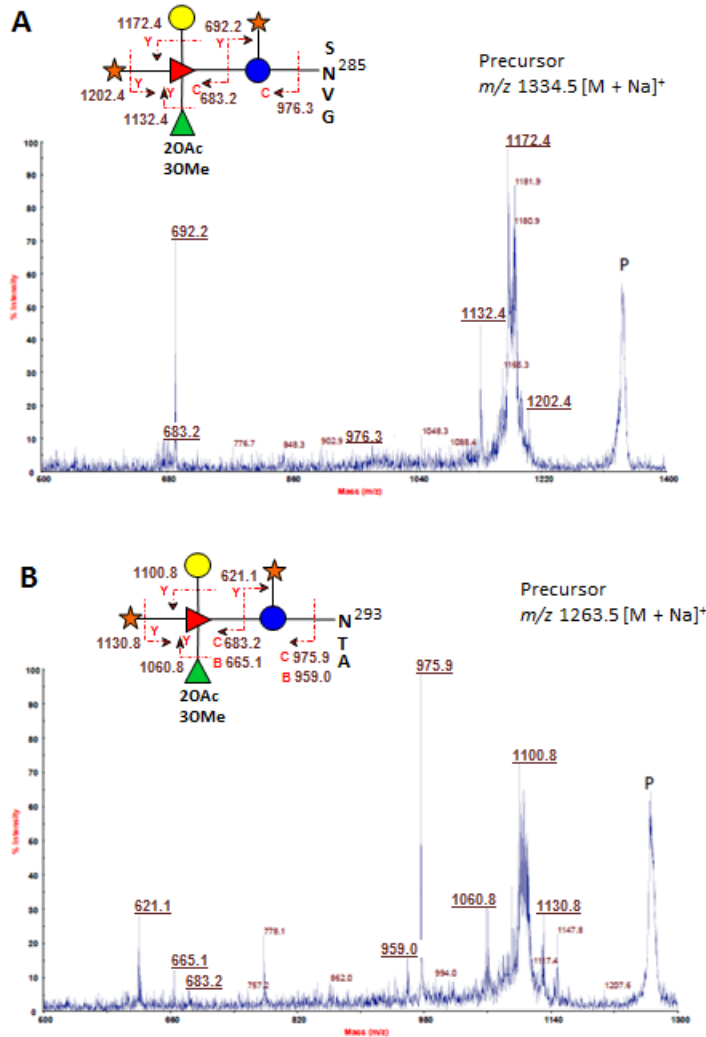
Figure 3.31 MS/MS analysis of the most significant molecular ions from CVM-1 glycopeptide fraction: [285]Asn (A) and [293]Asn (B) sites, occupied by a hexasaccharide glycoform. (De Castro et al. 2016)

Regarding ATCV-1, the MALDI spectra allowed us to understand the elevated number of xylose residues of group **E**. The reflectron positive MALDI spectrum (Fig. 3.32A) contained several ion peaks. We focused our attention on the base peak at m/z 1063.29, consistent with the sodium form of the hexasaccharide with two O-methyl groups (dhex$_2$hex$_2$Me$_2$pent$_2$Asn). Based on the NMR results, the first

methyl group was located on the deoxyhexose residue (rhamnose **C**), and the second should belong to xylose **E'** or **F'**.

Thus, MS/MS analysis was performed on the base peak to define the O-methylation. The resulting spectrum (Fig 3.32B) presented two couples of diagnostic ions: m/z 449.1, 655.2 and m/z 463.1, 641.2, that were obtained by the breaking of the bond between the fucose and glucose unit. Ions m/z 463.1 and 449.1were consistent with a hexpent Asn moiety with (m/z 463.13) and without (m/z 449.10) the methyl group, suggesting that the proximal xylose was not always methylated at O-4, as found for **F** and **F'**. The ions at m/z 641.2 and 655.2, instead, were assigned to the fragments complementary to the previous ions (m/z 463.1 and 449.1, respectively), indicated that once the proximal xylose is methylated, the distal xylose residue was not substituted or vice versa. Hence, the dominant ion of the MS glycopeptide profile (m/z 1063.29) was consistent with two species (as depicted in Fig. 3.32B), that differed for the type of xylose methylated.

A more careful analysis of the MALDI spectrum revealed two peaks, differing at higher and lower mass values for 14 u (m/z 1077.31 and m/z 1049.28), indicating that there were other two possible glycopeptide structures: one having both xylose units methylated and the other had no methyl at any xylose residue, respectively. Collectively these data confirmed that ATCV-1 had a mixture of four glycopeptides that differed from the non-stoichiometric methylation at O-4 of both proximal and distal xylose units.
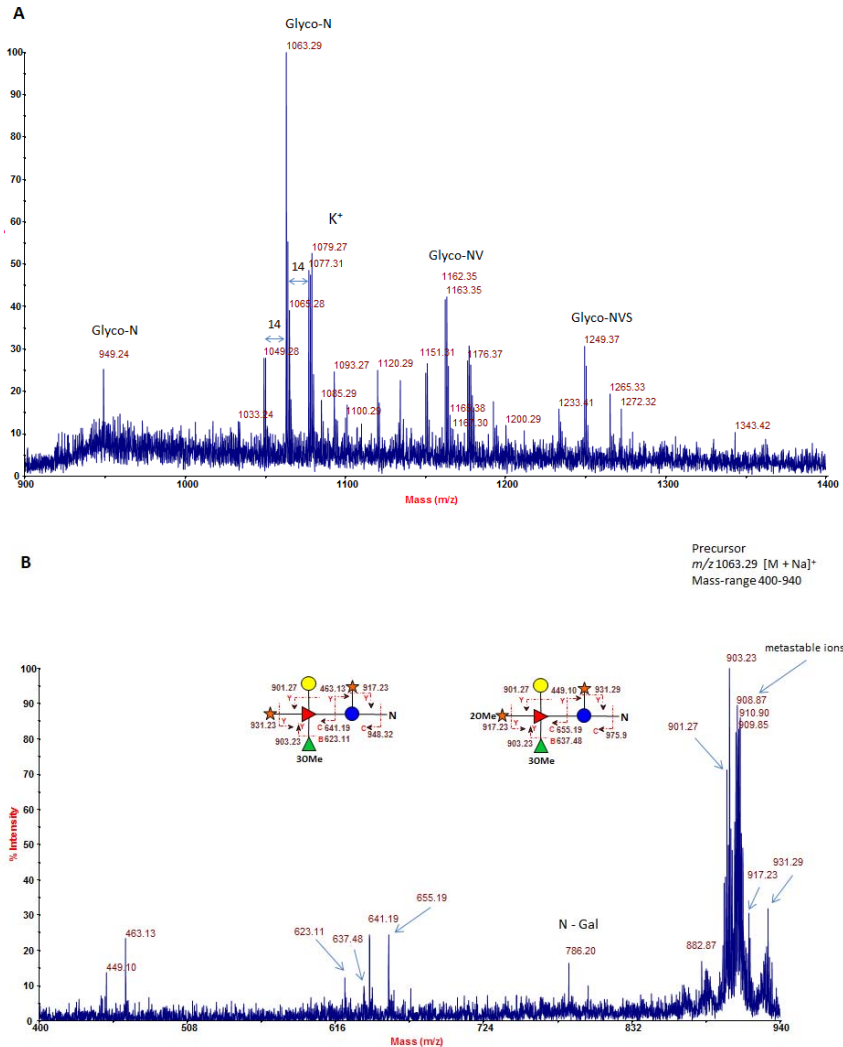
Figure 3.32 A) MALDI mass spectrum of the glycopeptide fraction from ATCV-1. B) MS/MS spectrum of the base-peak at *m/z* 1063.29. (De Castro et al. 2016)

For TN603 and MT325, MALDI mass spectra of two glycopeptide mixtures obtained by proteinase K digestion of the both capsid proteins are shown in Figures 3.33 and 3.34. MS/MS experiments were performed on many glycopeptide ions (*m/z* 1150.37, 1280.42, 1252.49 and 1266.50 for TN603; *m/z* 1209.24, 1252.25, 1063,19 and 1162.23 for MT325; data not shown) to obtain information on the structures of the associated glycan moieties. The ions at 1252 was presented in both TN603 and MT325 glycopeptide mixture and in both cases account for a couple of structures, but the MS/MS spectra, of each of them (Fig.

3.35), revealed a rather different fragmentation pattern, due to a slight variation in one of the two glycopeptide components.

Once again, the MS/MS analyses support the NMR conclusions, demonstrating that in both cases a terminal xylulose residue can be attached to either the proximal or distal xylose unit, or it could be missing. This last glycoform has low abundance and was outside to the NMR detection limits.
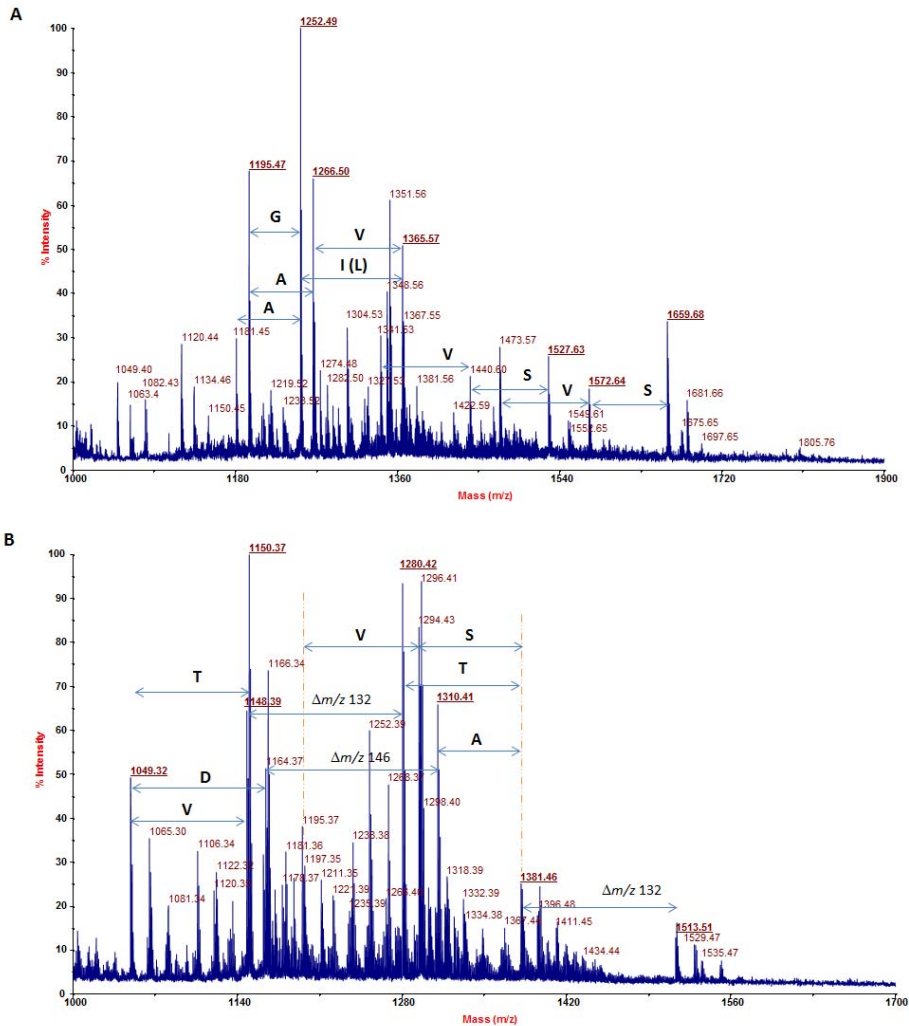
Figure 3.33 MALDI mass spectra of the glycopeptide fractions from TN603 major capsid protein. (De Castro et al. 2016)
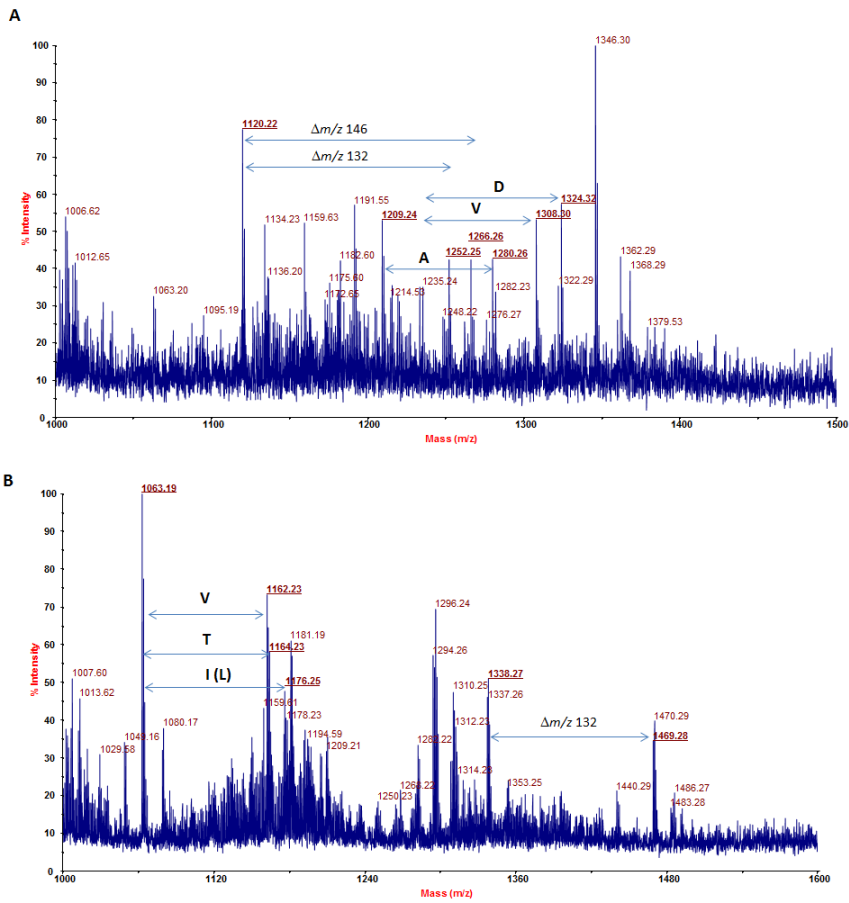
Figure 3.34 MALDI mass spectra of the glycopeptide fractions from MT325 major capsid protein. (De Castro et al. 2016)
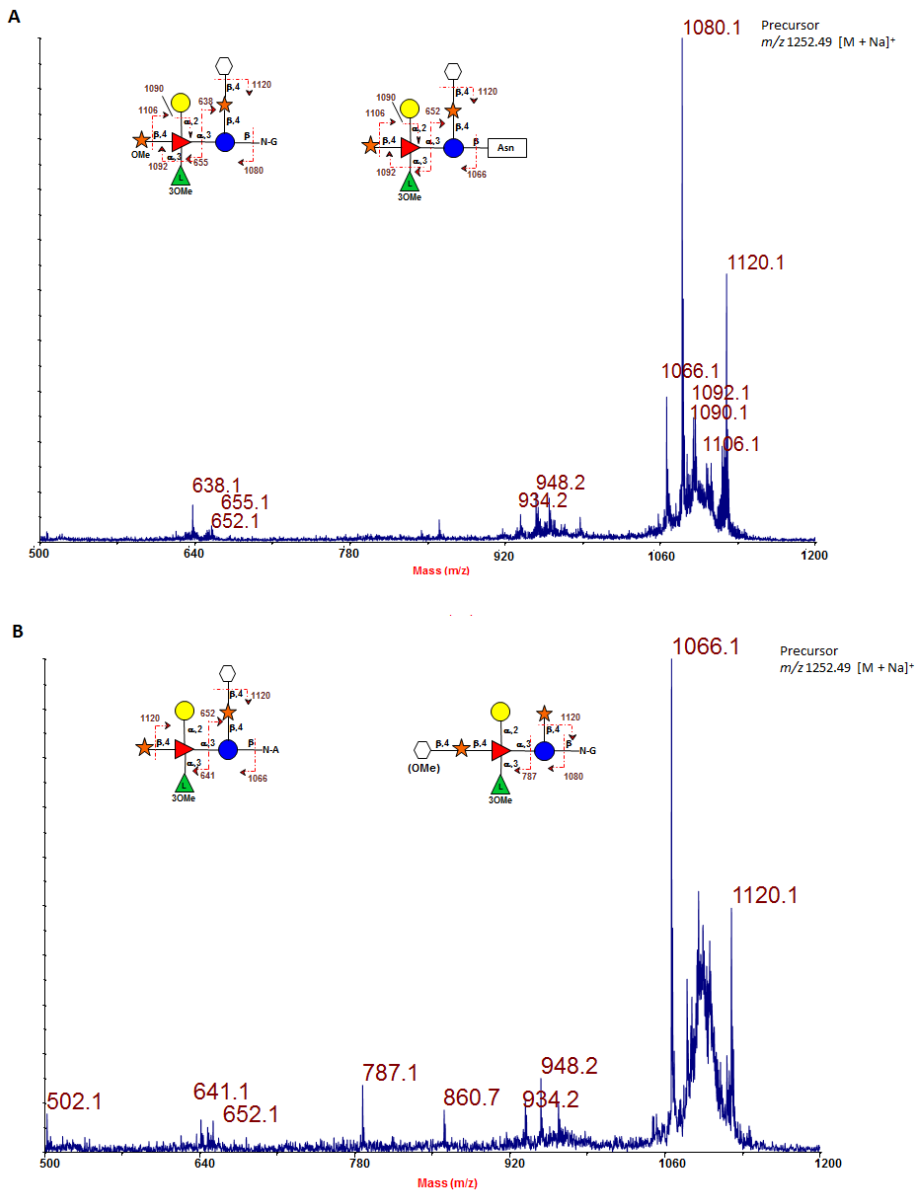
Figure 3.35 MS/MS spectra of spectrum of the base-peak at *m/z* 1252 from TN603 (A) and MT325 (B) glycoproteins. (De Castro et al. 2016)

3.4 Conclusions

The purpose of this part of the Ph.D. project is the determination of the *N*-glycans constituents of the major capsid proteins from different chloroviruses, to demonstrate that the unusual glycoform described for *Paramecium bursaria* chlorella virus 1 (PBCV-1) is not an isolated case, but it is a structural motif shared by other organisms.

In this contest, seven chloroviruses, with different host specificities, were selected and studied: NY-2A (NC64A virus); OSy-NE5 (OSy viruses); CVM-1, NE-JV-1 and MT325, (Pbi viruses); ATCV-1 and TN603 (SAG viruses).

The major capsid proteins of each virus above-mentioned was extracted (Agarkova et a. 2006) and the corresponding glycidic portion was purified (using an enzymatic digestion with proteinase K following by a chromatographic purifications) and analyzed via chemical, spectrometric and spectroscopic analysis.

These studies revealed that all MCPs of such chloroviruses are *N*-glycosylated, and more interestingly, all these glycans share an **unprecedented *N*-glycosylation pattern**, totally different to other *N*-glycans reported so far for bacteria, archea and eukarya, including unicellular algae (Mathieu-Rivet et al. 2014).

N-glycans from the chloroviruses (depicted in figure 3.36) share a unique central common core structure (Fig. 3.36) constituted by a pentasaccharide with a β-glucose linked to an asparagine residue which is not located in the typical sequon N-X-T/S. The glucose is substituted with a terminal xylose unit and a hyperbranched fucose, which is in turn connected with a terminal galactose and a second xylose residue. The third position of the fucose unit is always linked to a rhamnose, which is a semiconserved element because its absolute configuration is virus-dependent. Indeed, NC64A and OSy-NE5 viruses presented a rhamnose ᴅ configured, whereas Pbi and SAG viruses had a ʟ rhamonse. Additional decorations occur on this core *N*-glycan and represent a molecular signature for each chlorovirus.
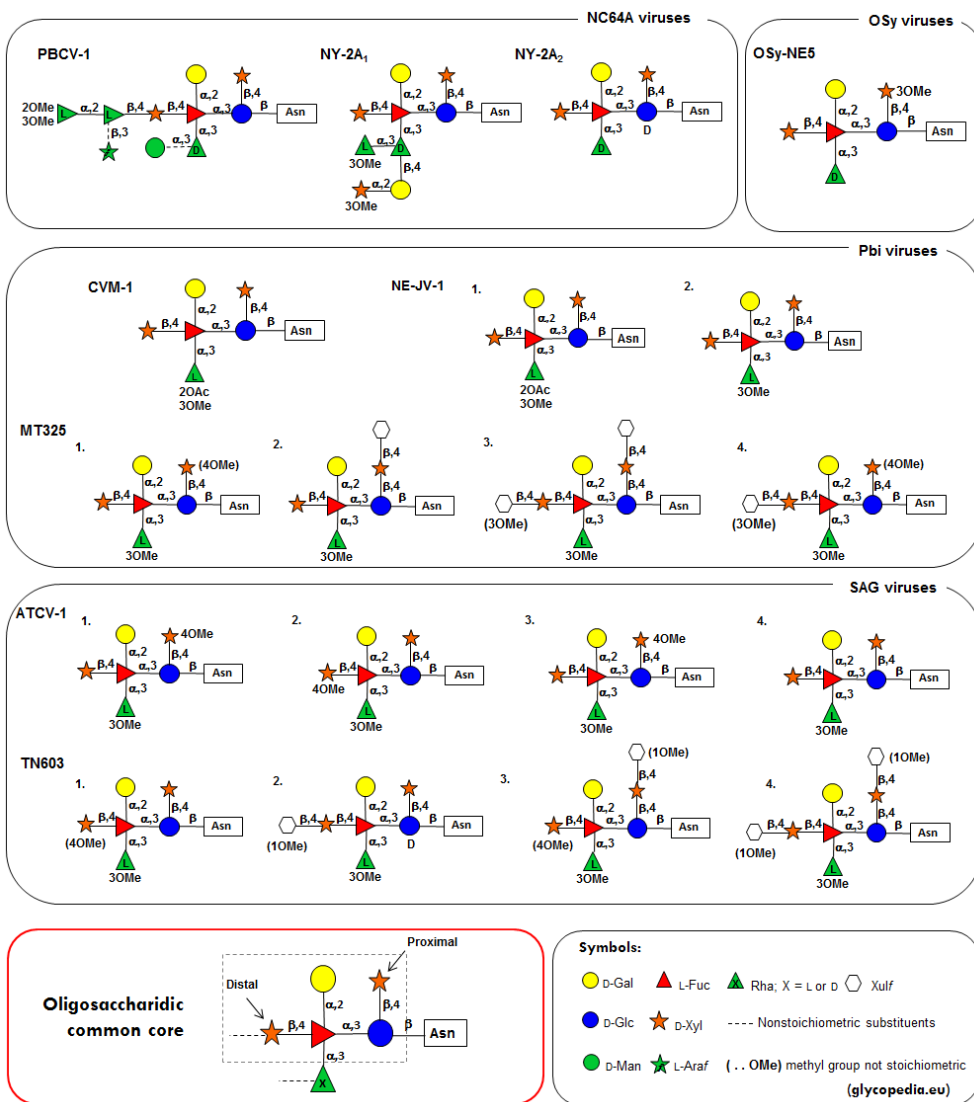
Figure 3.36 Schematic representation of the *N*-glycans of all analyzed chloroviruses. (De Castro et al. 2016; Quispe et al. 2016; Speciale et al. 2017) The *N*-glycan of the prototype, PBCV-1, is given for comparison. Different viruses are grouped according to their host specificity. All sugars are in the pyranose form except where specified. The oligosaccharide common core structure of N-glycans from chloroviruses is reported inside the red box, on the left-bottom of the figure.

Chapter 4

Antigenic variants

Twenty spontaneously derived, antigenic variants of PBCV-1, were identified and, interestingly, polyclonal antisera prepared against members of each of these antigenic classes react exclusively or predominately with the Vp54 equivalents from the viruses in the class used for the immunization. This result shows that mutants were resistant to PBCV-1 polyclonal antiserum (Wang et al. 1993), suggesting a structural change in the glycan portion, because the polyclonal antiserum recognized oligosaccharide epitopes on the major capsid protein.

In support of this hypothesis, genoma from antigenic variants was sequenced (Landstein et al. 1995) and mutations in genes that interfere with specific steps in the viral encoded glycosylation process appeared. Surprisingly, 18 of 21 antigenic variants had mutations in gene a064r (or it was deleted), which was one of the six putative glycosyltransferases, presumably involved in glycosylation of the PBCV-1 major capsid protein Vp54.

The A64R enzyme contains 638 amino acid organized into three conserved domains (Fig. 4.1). The first one presents a glycosyltransferase function, instead, the role of other two domains is not clear, even though genetic analysis suggests that also the second domain has a glycosyltransferase function and the third domains have a weak match with a methyltransferase. Thus, the structural studies of the N-glycans from different antigenic variants are needed to assign the function of the mutated genes. Therefore, the second goal of my Ph.D. project is the structural studies of the glycans of the major capsid protein from the antigenic variants in order to address the role of such gene, unknown to date.

The antigenic variants are divided into six antigenic classes (reported at the left column in the figure 4.1) based on their differential reaction to six different polyclonal antibodies: antigenic classes A and E, characterized by a mutation in the second domain; class B for which the first domain is affected; class C comprises three viruses for which no mutations occur in a064r gene and one virus that presents

a point mutation in the first domain; viruses that present a large deletion of the a064r gene belong to class D; class F comprises only one virus, and A064R presents a truncation of the third domain.

In this regard, we have analyzed one or two representative(s) mutants for each antigenic class: P9L10, P9L1 and P91 (antigenic class A); EPA-1and EPA-2 (class B); E11 and EIL-3 (class C); PIL 6 (class D); CME6 (class F).
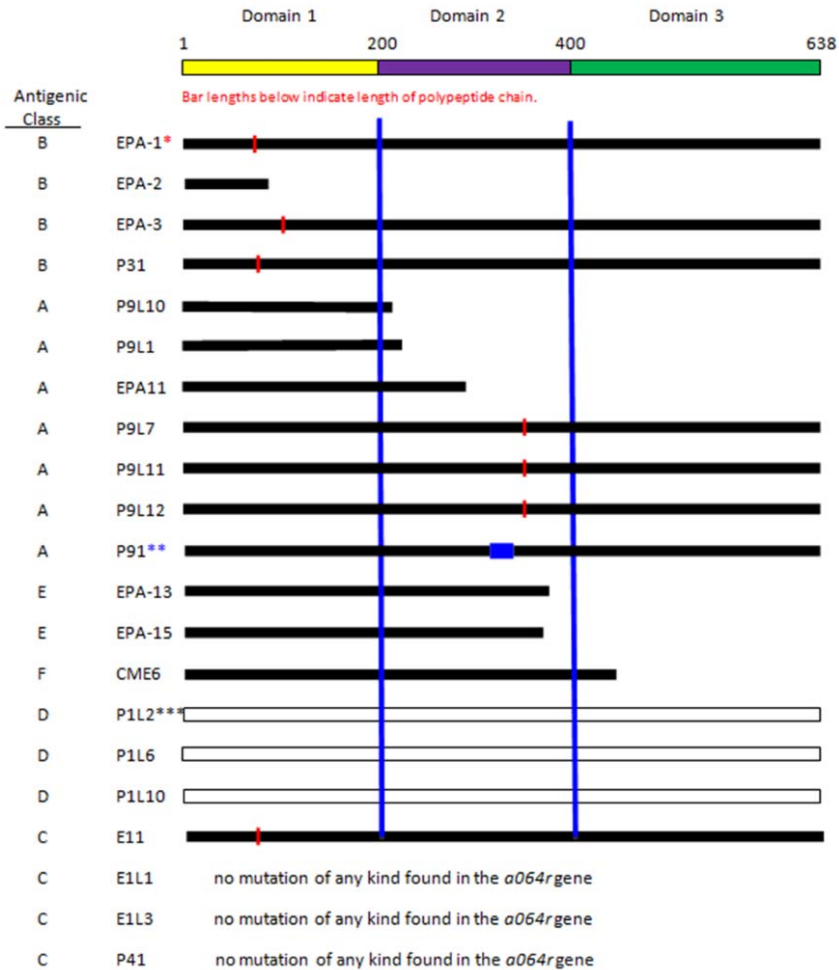
Figure 4.1 Graphic representation of the location and effect of genic mutations within the A064R protein in 21 strains of antigenic variants. The colored bar at the top represents the locations of the three protein domains. The left-hand column (Antigenic class) represents the antigenic variant type. The lengths of the black bars represent the predicted length of the encoded protein. * Vertical red bar refers to point mutations. ** Blue horizontal bar refers to an insertion. *** Open box indicates deletion of entire gene

96

4.1 MCP(s) extraction and *N*-glycans purification

The major capsid proteins of each antigenic variant were extracted according with the procedure reported in the paragraph 8.1, and their SDS-PAGE analysis (Fig. 4.2) give us an idea of their molecular weight, that decreases starting from the PBCV-1 to the E11 (data not obtained for EIL-3 and PIL6).
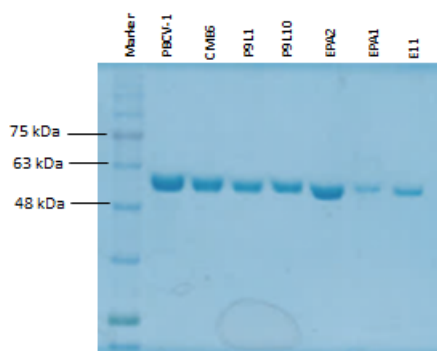
Figure 4.2 SDS-PAGE of total proteins of the MCPs from PBCV-1 and its antigenic variants CME6, P9L1, P9L19, EPA2, EPA1 and E11. Glycoproteins were revealed with Comassie staining. BLUeye Prestained Protein Ladder was used as marker.

In collaboration with Professor Tonetti, the exact molecular weight of the complete capsid of three antigenic variants (P91, EPA-1 and PIL6) was determined, using ESI-MS (Fig. 4.3) approach. This analysis allowed to predict the number of glycosylated sites (Table 4.1), knowing the molecular weight of the glycidic potion, determined combining chemical and NMR analysis.

Table 4.1 Δ Molecular weight was calculated as the difference between the calculated mass with the molecular weight of the Vp54 glycoprotein (48165 Da). Regarding EIL-3 and P91, which presented different type of glycans that differed for some non-stoichiometric substituents, all possible cases were analyzed.

Antigenic variants	Calculated MW (Da) of glycoprotein	Δ Molecular weight (Da)	Experimental Molecular weight (Da) of glycidic portion	Derived numbers of glycosilated sites
P91	52938	4773	1320 (complete) 1158 (without Man) 1290 (without OMe group) 1128 (without both)	4
EIL-3	51064,5	2899.5	1396 (complete) 748 (truncated structure)	4
PIL6	50463,2	2298.2	602	4
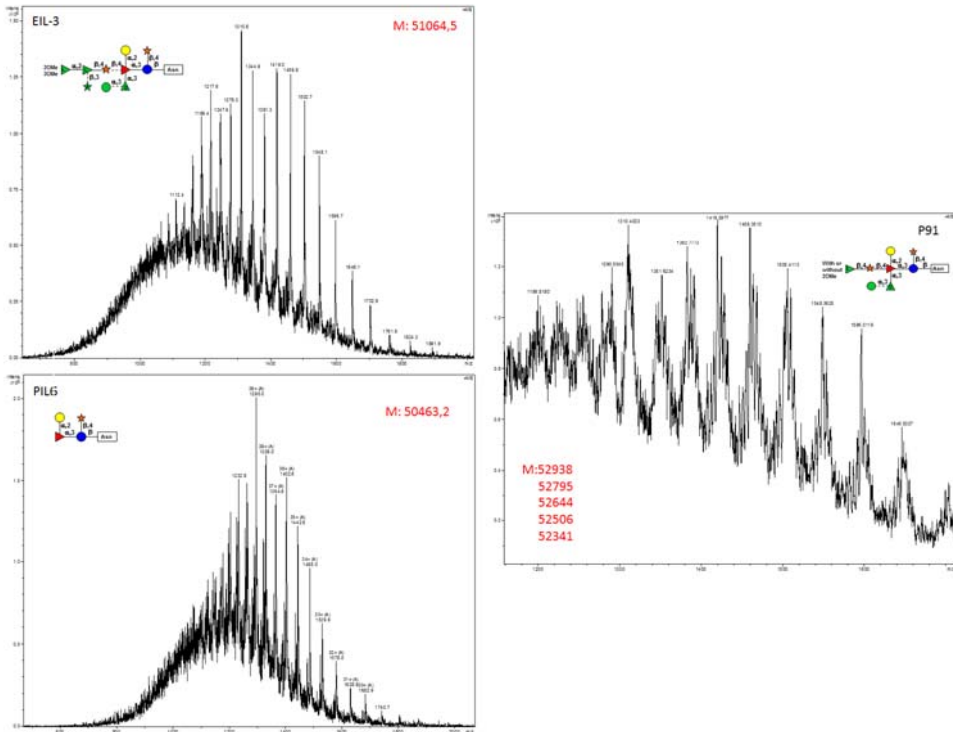
Figure 4.3 LC-MS spectra of EIL-3, P91 and PIL6.

Thereafter, the major capsid protein of each variant was hydrolyzed with Proteinase K and then purified using a size exclusion chromatography, similarly to the procedure used for the chloroviruses. Chromatographic peaks were checked via NMR and one main fraction contained the glycopeptide was obtained (^{1}H-NMR spectrum of each fraction was reported in figure 4.4), except for EIL-3, which presents two main fractions, named EIL-3$_A$ and EIL-3$_B$, that were analyzed separately. All fractions eluted at 50% of the column volume and with a yield of the 10 – 15 % with respect to the starting MCP amount.

Proton NMR spectrum and, in particular, the anomeric region is very indicative, indeed comparing this region with the corresponding of the PBCV-1, we can have an idea of the monosaccharidic composition.
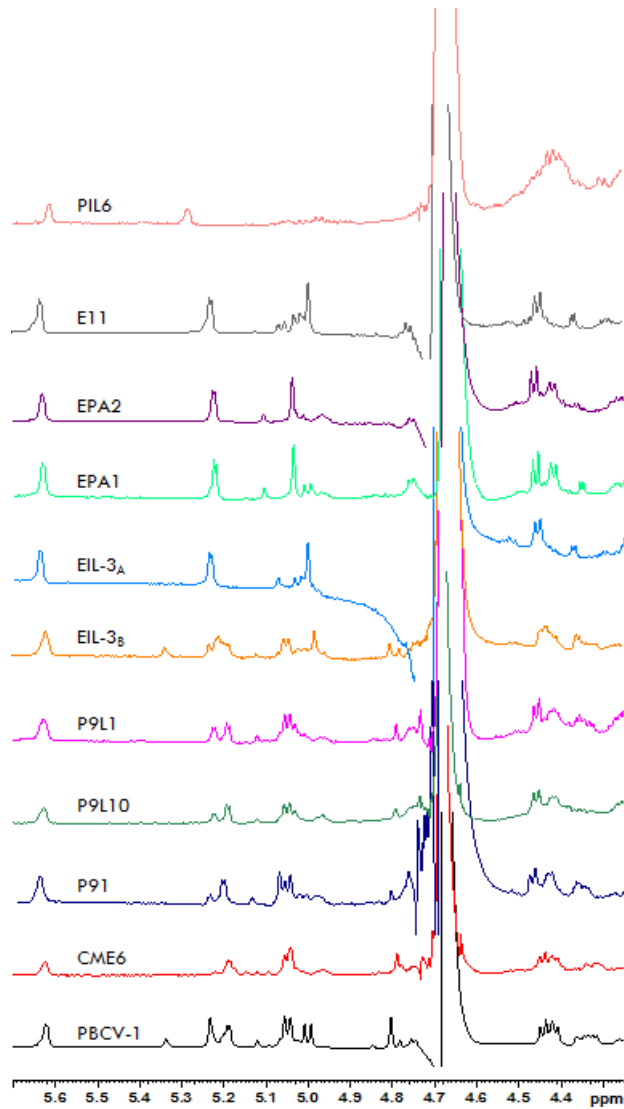
Figure 4.4 ^{1}HNMR spectra reporting the anomeric region of the glycopeptides obtained from the capsid protein of the different antigenic variants. The PBCV-1 spectrum was used as reference. Spectra were recorded at 600 MHz, at 310K, except for EIL-3$_A$, for which the temperature was changed (323 K).

4.2 Determination of glycan structures

Glycopeptide structures of each antigenic variant was investigated combining chemical, spectrometric and spectroscopic analysis, as elucidated below.

4.2.1 Chemical analysis

The monosaccharide composition of the glycopeptides was performed only for some variants and that the full set of NMR spectra was recorded.

The previous purified glycans were transformed into their corresponding acetylated methyl glycoside, that were used to prepare the methylated 2-(-)-octyl derivatives or the partially methylated alditol acetate (AAPM), to determine the absolute configuration of each monosaccharide and the monosaccharidic composition along with their branching points, respectively. The obtained results were summarized in the tables 4.2 and 4.3.

Table 4.2 Permethylated octyl glycosides data obtained from seven antigenic variants (P91, EIL-3, PIL6, E11, CME6 and EPA1). Data from PBCV-1 were reported for comparison. + indicates the sugars present; +/- means that the sugar is present, but in not stoichiometric amount; - denotes the absence of the sugar.

Permethylated octyl glycosides	PBCV-1	P91	PIL6	CME6	EPA-1
t-Araf	+	-	-	+/-	-
t-D-Xyl	+	+	+	+	+
t-L-2,3OMeRha	+	-	-	-	-
t-L-Rha	+	+	-	+	-
t-D-Rha	+	+	-	+	+
2-L-Rha	+	-	-	+	-
3-L-Rha	+	-	-	+	-
3-D-Rha	+	+	-	+/-	-
4-D-Xyl	+	+	-	+	-
t-D-Gal	+	+	+	+	+
t-D-Man	+	+/-	-	+/-	-
2,3-L-Rha	+	-	-	+	-
2,3,4-L-Fuc	+	+	+	+	+
3,4-D-Glc	+	+	+	+	+

Table 4.3 Partially methylated alditol acetate data obtained fromP91, EIL-3, PIL6 and E11. Data from PBCV-1 were reported for comparison. + indicates the sugars present; +/- means that the sugar is present, but in not stoichiometric amount; - denotes the absence of the sugar.

Partially methylated alditol acetate	PBCV-1	P91	EIL-3	PIL6	E11
t-Araf	+	-	+/-	-	-
t-Xyl	+	+	+	+	+
t-Rha	+	+	+	-	+
2-Rha	+	-	+	-	-
4-Xyl	+	+	+	-	+
3-Rha	+	+	+/-	-	-
2-Fuc	-	-	-	+	-
t-Man	+	+	+/-	-	-
t-Gal	+	+	+	+	+
2,3-Rha	+	-	+/-	-	-
2,3-Fuc	-	-	+	-	+/-
2,3,4-Fuc	+	+	+	-	+/-
3,4-Glc	+	+	+	+	+

Examples of chemical analysis are shown below.

PIL6 has the simplest glycan, as indicated from the Partially methylated alditol acetate analysis (Fig. 4.5), in which only four important peaks appeared: 2-linked fucose, terminal galactose, terminal xylose and glucose 3 and 4 substituted.
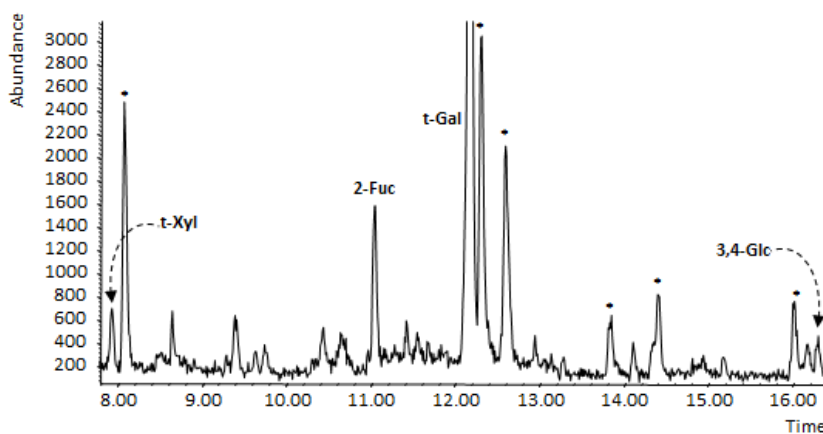
Figure 4.5 Partially methylated alditol acetate chromatographic profile obtained from PIL6. * indicates impurities.

Of interest, the AAPM chromatographic profile (Fig. 4.6) of EIL-3 displays both 2,3-fucose and 2,3,4-fucose peaks, suggesting that the sample contained a mixture of two different glycoforms: one truncated at level of the 4-fucose, with distal xylose missing with its substituent; and another with the complete glycan.
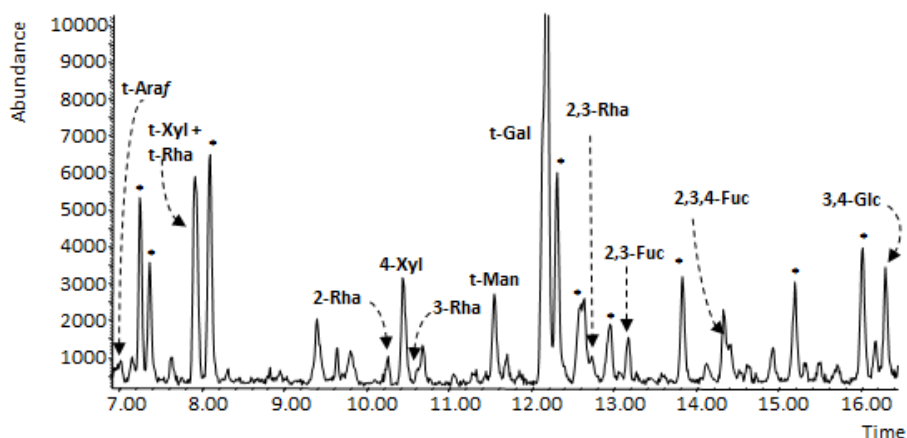
Figure 4.6 Partially Methylated Alditol Acetylated chromatographic profile of EIL-3. * indicates the impurities.

In the next example (Fig 4.7), is reported the chromatographic profile of the partially methylated and acetylated 2-(-)-Octyl Glycosides obtained for P91; a fundamental analysis to discriminate between the two rhamnose units with opposite absolute configuration.
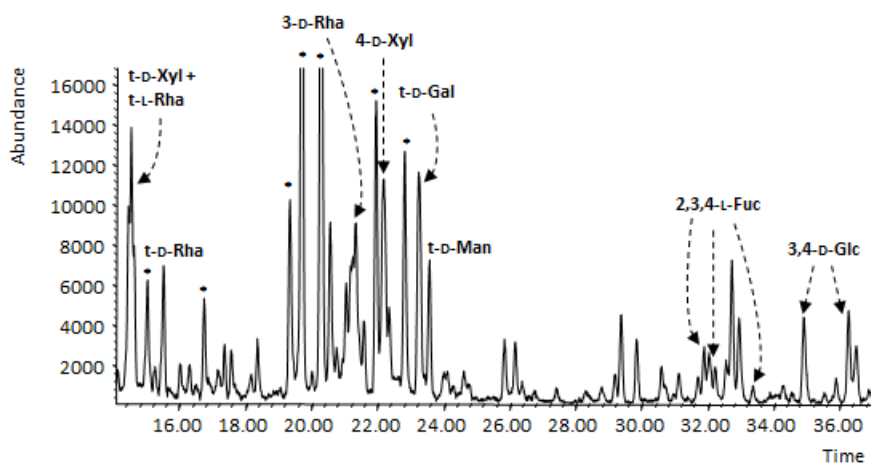
Figure 4.7 Partially Methylated and Acetylated 2-(-)-Octyl Glycosides chromatographic profile of P91. * indicates impurities.

103

All these analyses have revealed that the galactose unit was always present, instead the mannose was not in a stoichiometric amount in several viruses; thus, its amount was calculated analyzing its area in chromatographic profiles of permethylated octyl-glycosides (results are summarized in figure 4.8, in which was indicated the amount of mannose absent, thus 100% means that the mannose was not present).
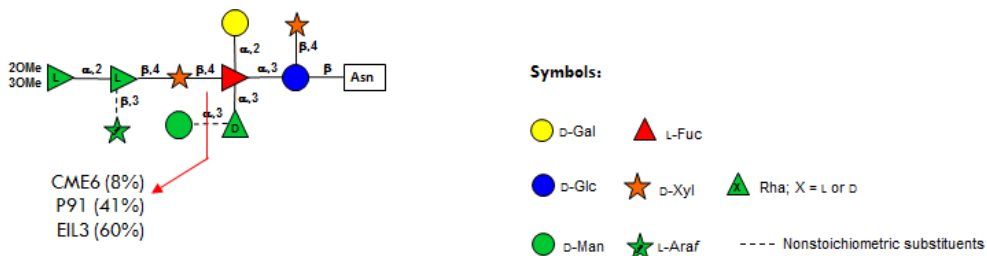
Figure 4.8 Schematic representation of *N*-glycan, in which is specified the amount of the absent mannose in the different antigenic variants. These values were calculated analyzing the octyl glycosides spectra of each mutant.

4.2.2 Spectroscopic analysis

NMR experiments were recorded at 310 K, with a field strength of 600 MHz, for each antigenic variant, except to EIL-3$_A$, for which the temperature was changed (323 K).

4.2.2.1 EPA-1 (Antigenic class B)

The heteronuclear single quantum correlation spectrum (Fig. 4.9) showed six anomeric signals, that were labelled with a letter **A**-**F** in order of their decreasing chemical shift, and one ring proton in the range of 5.7 and 4.3 ppm, a carbinolic region from 4.3 to 2.9 ppm and two methyl signals at approximately 1.3 ppm, typical of 6-deoxyresidues. NMR analysis started from the anomeric proton of **D** ([1]H 5.01 ppm) that was classified as *N*-linked residue owing to its diagnostic carbon chemical shift at 80.7 ppm. Starting from H-1 signal of this residue, TOCSY experiment exhibited a complete correlations pattern up to H-6s suggesting that this unit had the gluco stereochemistry. Furthermore, its ^{13}C chemical shift values established that **D** was a glucose with C-3 (76.9 ppm) and C-4 (74.8 ppm) shifted at low field by glycosylation, similarly to what was reported for the PBCV-1 glycopeptide used as reference (De Castro et al. 2013). Note that the COSY and TOCSY spectra indicated that the glucose unit was heterogeneous and it was split in **D** and **D'** residues, differing mainly by the peptide portion. A similar approach was utilized to determine all other residues.

Regarding **A**, TOCSY experiment displayed a correlation from the anomeric proton (5.63 ppm) with only H-2, H-3 and H-4, moreover COSY spectrum did not show any intense correlation between H-3 and H-4, typical feature of a *galacto* stereochemistry. In addition, with T-ROESY experiment (Fig. 4.10), it was possible to determine the position of H-5 proton that was in turn correlated with a methyl group at 1.31 ppm. All this information indicated that residue **A** was a fucose α-configured at the anomeric center, glycosylated at C-2, C-3 and C-4 as suggesting from the low field values of the respective carbon, with respect the standard values. **B** residue showed the same TOCSY pattern, with the difference that its H-5 had a correlation with a carbon at 62.4 ppm, as indicating by the cross-peak H-5/H-6

in the COSY spectrum, identifying this unit as a terminal galactose. The α-configuration of the anomeric centre was deduced by the comparison of carbon chemical shift values (Table 4.4).

Regarding **C**, its anomeric proton displayed one main correlation (H-1/H-2), whereas the connections up to the methyl group were visible from H-2; this information identified the residue **C** as *manno* configured unit. Inspection of other NMR spectra (COSY, TOCSY and HSQC) determined the position of all ring protons included H-6, which resonated at 1.28 ppm, and disclosed that **C** was a terminal rhamnose residue, α-configured at anomeric center.

The spectroscopic patterns of **E** and **F** residues were very similar: both presented an efficient propagation of the magnetization in the TOCSY spectrum as occured for *gluco* configured residues. However, unlike to the residue **D**, H-5$_{eq}$ of **E** and **F** presented correlations compatible with a five carbons atom monosaccharide, suggesting that they were two terminal β-xylose units. In addition, the T-ROESY (Fig 4.10) spectrum revealed that **E** was linked at O-4 of **A**, whereas **F** at O-4 of **D** (distal and proximal xyloses, respectively).

The complete glycopeptide structure (Fig. 4.11) was inferred by studying the T-ROESY experiment, which disclose that **A** was linked to O-3 of **D**, that in turn had **B** at O-2 and **C** at O-3.
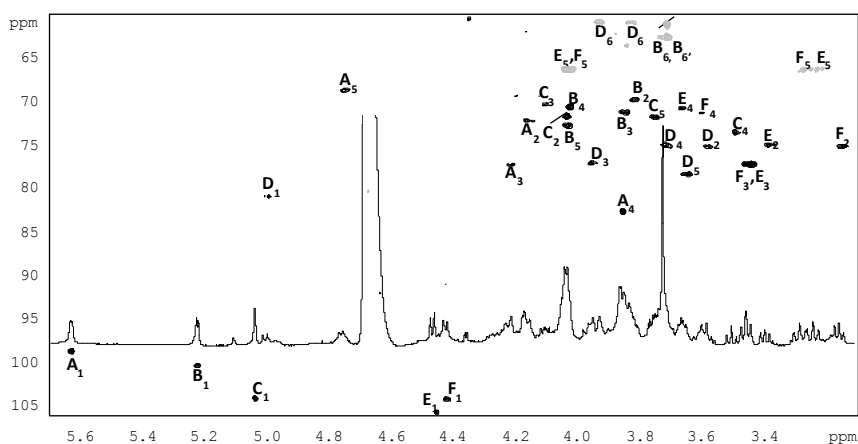
Figure 4.9 Expansion of EPA-1 glycopeptide HSQC spectrum measured at 600 MHz, 310 K. Those appearing in grey have the opposite sign with respect to the other and represent carbons bearing two hydrogen atoms. The correlations attribution follows the letter system of Table 4.4 and Figure 4.11.
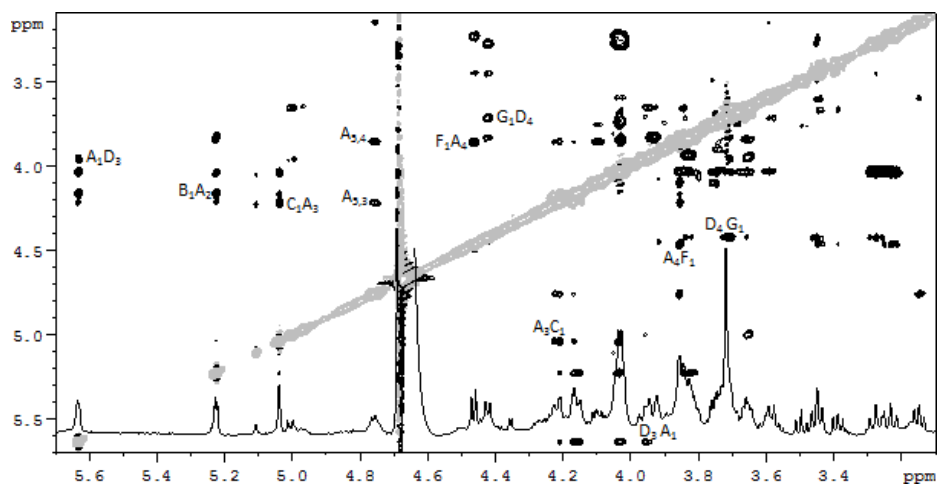
Figure 4.10 Expansion of EPA-1 glycopeptide T-ROESY spectrum measured at 600 MHz, 310 K.
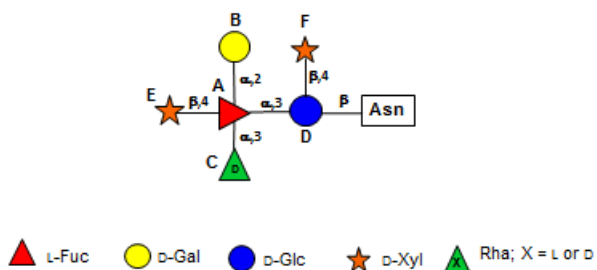
Figure 4.11 Structure of the *N*-glycan from EPA-1. Letter labels are those used during NMR assignment.

Table 4.4 Proton and carbon chemical shifts (600 MHz) obtained for EPA-1 glycopeptide, recorded in deuterated water at 310 K with acetone as internal standard.

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.63	4.16	4.22	3.85	4.76	1.31
2,3,4-α-L-Fuc	^{13}C	98.5	72.0	77.1	82.5	68.5	16.3
B	^{1}H	5.22	3.81	3.84	4.02	4.03	3.71-3.73
t-α-D-Gal	^{13}C	100.2	69.6	71.0	70.4	72.6	62.4
C	^{1}H	5.04	4.03	4.09	3.49	3.75	1.28
t-α-D-Rha	^{13}C	103.9	71.5	70.1	73.3	71.6	18.0
D	^{1}H	5.01	3.58	3.95	3.71	3.65	3.93-3.83
3,4-β-D-Glc	^{13}C	80.7	75.0	76.9	74.8	78.1	60.7
E	^{1}H	4.46	3.38	3.45	3.66	3.23-4.03	-
t-β-D-Xyl	^{13}C	105.2	74.5	76.7	70.2	65.7	-
F	^{1}H	4.20	3.14	3.45	3.59	4.03-3.27	-
t-β-D-Xyl	^{13}C	103.7	74.7	76.7	70.8	65.8	-

4.2.2.2 EPA-2 (Antigenic class B)

The proton spectrum of EPA-2 glycopeptide (at bottom of the figure 4.12) presented the same six anomeric protons of EPA-1. Moreover, the homonuclear and heteronuclear experiments (HSQC spectrum was reported in figure 412) showed the same correlation pattern for all monosaccharides, indicating that the EPA-2 glycopeptide structure was identical to that found for EPA-1. Thus, we can affirm that EPA-2 was constituted by a β-glucose linked to asparagine, as suggesting from the anomeric carbon chemical shift (80.4 ppm) in turn linked to β-xylose at O-4 and to α-fucose at O-3. The fucose was in turn glycosylated with α-galactose at O-2, α-rhamnose at O-3 and β-xylose at O-4 (Fig. 4.13).
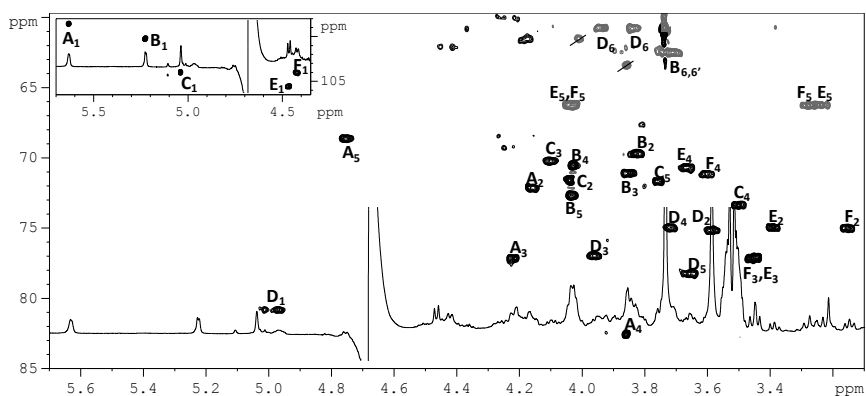
Figure 4.12 Expansion of EPA-2 glycopeptide HSQC spectrum measured at 600 MHz, 310 K. Carbons bearing two hydrogen atoms have densities of opposite sign compared to the others and are colored in grey.
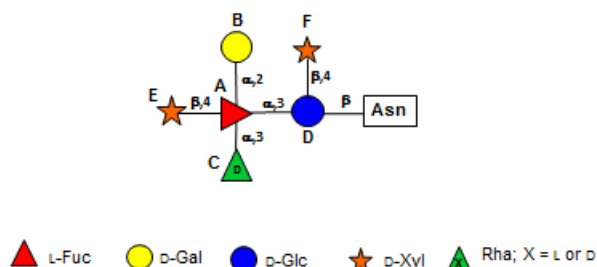
Figure 4.13 Structure of the glycan from EPA-2. Letter labels are those used during NMR assignment.

Table 4.5 (600 MHz, 310 K) Proton and carbon chemical shifts obtained for EPA-2 glycopeptide.

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.63	4.18	4.22	3.86	4.75	1.31
2,3,4-α-ʟ-Fuc	^{13}C	98.6	72.0	77.0	82.5	68.5	16.2
B	^{1}H	5.23	3.82	3.86	4.03	4.03	3.71-3.74
α-ᴅ-Gal	^{13}C	100.2	69.6	71.0	70.4	72.5	62.3
C	^{1}H	5.04	4.04	4.10	3.50	3.75	1.30
α-ᴅ-Rha	^{13}C	104.0	71.5	70.1	73.3	71.5	17.9
D	^{1}H	5.01	3.58	3.95	3.72	3.66	3.94-3.83
3,4-β-ᴅ-Glc	^{13}C	80.4	75.0	76.8	74.9	78.1	60.6
E	^{1}H	4.46	3.39	3.45	3.66	3.23-4.03	-
β-ᴅ-Xyl	^{13}C	105.6	74.9	77.02	70.6	66.1	-
F	^{1}H	4.42	3.14	3.45	3.59	4.03-3.27	-
β-ᴅ-Xyl	^{13}C	104.1	74.9	77.00	71.0	66.1	-

4.2.2.3 PIL6 (Antigenic class D)

PIL6 presents the simplest glycopeptide, indeed, the low field region of HSQC experiment (Fig. 4.14) showed four densities indicating the presence of only four monosaccharides. This glycans had several similarities with that of EPA-1: proton and carbon chemical shifts (Table 4.6) of **B**, **C** and **D** were very similar, in agreement with their attribution as α-D-galactose, β-D-glucose and α-D-rhamnose, respectively. Instead, **A** residue was a quite different.

It resulted to be a fucose α-configured at the anomeric center, glycosylated at C-2 as suggesting from the low field values of the carbon (^{13}C 75.2 ppm) with respect the standard values.

Therefore, the PIL6 glycopeptide was constituted by a β-glucose linked to the asparagine in turn glycosylated with a β-xylose at O-4 and with a α-fucose at O-3, that was in turn linked to a α-galactose at position 2 (Fig. 4.15).

Figure 4.14 Expansion of PIL6 glycopeptide HSQC spectrum measured at 600 MHz, 310 K. In grey were reported carbons bearing two hydrogen atoms that have densities of opposite sign compared to the others.

110
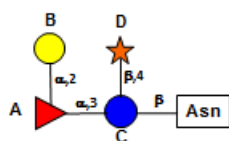
Figure 4.15 Structure of the *N*-glycan of PIL6.

▲ L-Fuc   ● D-Gal   ● D-Glc   ★ D-Xyl

Table 4.6 (600 MHz, 310 K) Proton and carbon chemical shifts obtained for PIL6 glycopeptide.

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.61	3.94	4.16	3.78	4.73	1.20
2-α-L-Fuc	^{13}C	98.7	75.2	70.5	73.7	67.6	16.6
B	^{1}H	5.28	3.84	3.90	4.02	4.05	3.85-3.83
α-D-Gal	^{13}C	101.3	70.1	70.7	70.5	72.7	63.3
C	^{1}H	4.97	3.58	3.92	3.75	3.64	3.89-3.97
3,4-α-D-Glc	^{13}C	80.8	75.0	77.3	74.7	78.2	60.1
D	^{1}H	4.42	3.15	3.44	3.49	3.26-3.93	-
β-D-Xyl	^{13}C	103.8	75.0	76.9	70.9	66.2	-

4.2.2.4 EIL-3 (Antigenic class C)

EIL-3 produced two different glycopeptide fractions: comparison of their proton spectra revealed that the most abundant glycoform (named EIL-3$_A$) had a minor number of residues than the other one (EIL-3$_B$).

Starting from EIL-3$_A$, its NMR spectra had several analogies with EPA-1, along with some differences. Indeed, **B**, **C** and **D** residues had the similar TOCSY pattern to those found for EPA1, identifying these units as α-galactose, α-rhamnose and β-glucose, respectively, even though the proton and carbon chemical shifts (Table 4.7) presented minimum differences and anomeric signal of **C** was overlapped with that of the glucose unit. These variations were justified by the different temperature at which the experiments were recorded (323 K).

Regarding **A**, TOCSY pattern revealed that it was a fucose unit, but the carbon chemical shift of C-4 (Table 4.7) was not shifted at low field value as reported for

the EPA-1, suggesting that in this case there was not an hyperbranched fucose, but that the residue was glycosylated only at O-2 and O-3. This was borne out by the presence of only one xylose unit (**E**): ^{1}H-NMR spectrum showed only one anomeric signal at 4.46 ppm and, very indicative, was the presence of only one of the two couple of H-5/C-5, in the HSQC spectrum (Fig. 4.16). Moreover, T-ROESY experiment disclosed that **E** was linked to O-4 of glucose (**D**).

The whole structure was completed analysing the T-ROESY experiment: **A** was linked to O-3 of **D**, that in turn had **B** at O-2 and **C** at O-3 (as depicted in figure 4.18).

Regarding **EIL-3$_B$** (Table 4.8), only HSQC experiment was recorded (Fig. 4.16). This spectrum was very complex due to the high number of signals and because the amount of the sample was very low, thus recording was quite close to the NMR detection limits. Nevertheless, it was possible to identify two distinct anomeric signals at ca. 4.4 ppm, indicating the presence of two types of xylose residues. This was also supported by the presence of two couples of H-5/C-5 (Fig. 4.17), in the HSQC spectrum. Significative were the presence of a density at ^{1}H/^{13}C 3.88/81.9 ppm concerning to the C-4 of the fucose unit, and three densities at ca. 3.5 ppm, relative to the methyl group, suggesting that a hyperbranched fucose and a α-rhamnose unit 2,3 methylated occurred, in according with the same monosaccharides in the PBCV-1. Finally, in HSQC experiments appeared new anomeric signals at ca. 4.8 ppm (**H** and **I**), that revealed the presence of two type of β-rhamnose, similarly to what reported for PBCV-1 (De Castro et al. 2013).

The presence of this additional glycoform was supported by chemical analysis, in particular the permethylated octyl derivatization of the sample disclosed the presence of both fucose 2,3-linked and hyperbranched fucose.

Thus, EIL-3 produced two different glycopeptides: a minor glycoform that was comparable with that of the prototype, PBCV-1, and major glycoform equivalent to that found for EPA-1 or EPA-2, a pentasaccharide, which differed from the longer form because of the lack of the substituents at the fourth position of the hyperbranched fucose (structures were reported in figure 4.18).
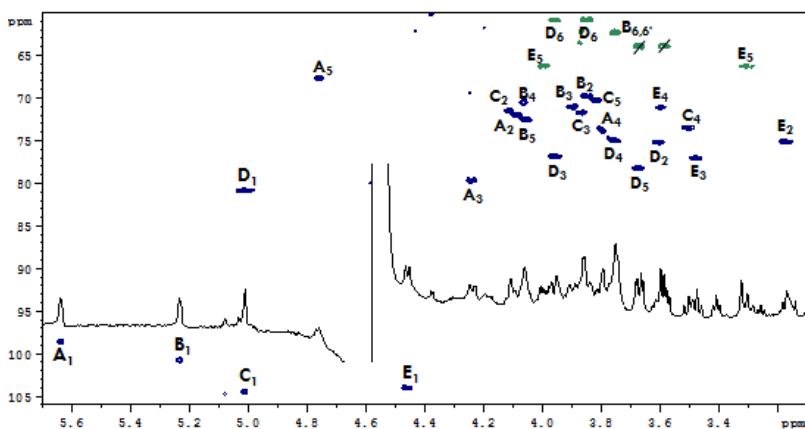
Figure 4.16 Expansion of EIL-3$_A$ glycopeptide HSQC spectrum measured at 600 MHz, 323 K. Densities appearing in green have the opposite sign with respect to the other and represent carbons bearing two hydrogen atoms. Signals crossed are related to the peptide moiety of the glycopeptide or to impurities.
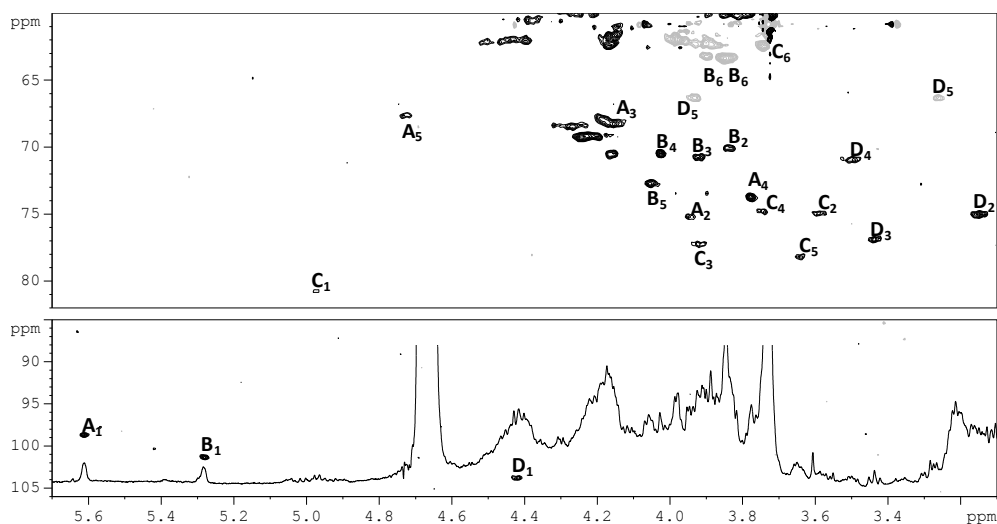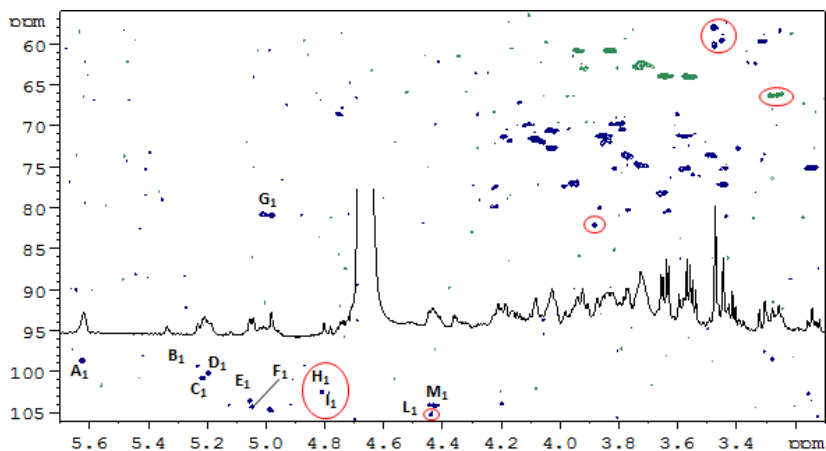

Figure 4.17 Expansion of EIL-3$_B$ glycopeptide HSQC spectrum measured at 600 MHz, 310 K. In green were reported carbons bearing two hydrogen atoms, which have the opposite sign with respect to the other. The important densities, that allow us to identify the long structure of this glycoform, are enclosed in red circles.

113
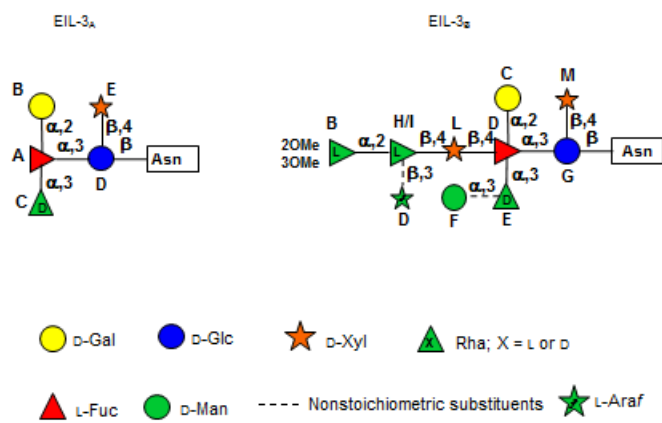
Figure 4.18 Structure of the *N*-glycans produced by EIL-3. EIL-3ₐ is the most abundant glycoform.

Table 4.7 (600 MHz, 323 K) Proton and carbon chemical shifts obtained for EIL-3ₐ glycopeptide

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	¹H	5.64	4.09	4.24	3.79	4.76	1.23
2,3,4-α-ʟ-Fuc	¹³C	98.5	71.9	79.6	73.7	67.6	16.2
B	¹H	5.23	3.85	3.90	4.06	4.06	3.76-3.75
α-ᴅ-Gal	¹³C	100.6	69.7	70.9	70.4	72.4	62.2
C	¹H	5.01	4.11	3.87	3.50	3.81	1.32
α-ᴅ-Rha	¹³C	104.4	71.4	71.6	73.3	70.1	18.0
D	¹H	5.03	3.61	3.96	3.75	3.67	3.96-3.85
3,4-β-ᴅ-Glc	¹³C	80.7	75.0	76.7	74.9	78.1	60.7
E	¹H	4.46	3.18	3.48	3.60	3.99-3.30	-
β-ᴅ-Xyl	¹³C	103.9	75.0	76.9	70.9	66.1	-

Table 4.8 (600 MHz, 310 K) Proton and carbon chemical shifts obtained for EIL-3$_B$ glycopeptide. **B** and **B'** are substituted at both O-2 and O-3 with a methyl group (chemical shifts ^{1}H/^{13}C: B-2OMe: 3.48/60.0; B-3OMe: 3.48/57.9; C-2OMe: 3.44/59.4; C-3OMe: 3.48/57.9).

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.63	4.17	4.21	3.88	4.75	1.30
2,3,4- α-L-Fuc	^{13}C	98.5	71.8	76.2	81.9	68.4	16.1
B/B'	^{1}H	5.24	3.95	3.64	3.40	4.11	1.26
α-L-2,3OMeRha	^{13}C	99.3	76.8	80.4	72.6	69.7	17.8
C	^{1}H	5.19	3.78	3.83	4.01	4.03	3.75-3.70
α-D-Gal	^{13}C	100.0	69.6	71.1	70.4	72.7	62.2
D	^{1}H	5.18	4.21	4.20	3.91	3.78-3.71	-
β-L-Ara$f$	^{13}C	100.1	77.4	74.4	83.1	62.6	-
E	^{1}H	5.05	4.19	3.86	3.58	4.12	1.29
3-α-D-Rha	^{13}C	103.4	71.2	79.8	72.4	70.5	17.9
F	^{1}H	5.04	4.08	3.86	3.60	3.81	3.72-3.91
α-D-Man	^{13}C	104.1	71.4	72.0	68.4	75.0	62.8
G	^{1}H	5.00	3.57	3.96	3.70	3.65	3.93-3.83
3,4-β-D-Glc	^{13}C	80.7	75.0	76.8	74.6	78.1	60.6
H	^{1}H	4.8	4.1	3.71	3.38	3.51	1.32
2-β-L-Rha	^{13}C	102.4	77.9	75.0	73.6	73.8	18.2
I	^{1}H	4.78	4.32	3.75	3.50	3.51	1.33
2,3-β-L-Rha	^{13}C	102.5	73.4	80.3	73.6	73.6	18.2
L	^{1}H	4.44	3.45	3.55	n.d.	4.33-3.25	-
4-β-D-Xyl	^{13}C	105.2	74.9	75.8	n.d.	65.8	-
M	^{1}H	4.43	3.14	3.44	3.60	3.28-4.04	-
t-β-D-Xyl	^{13}C	104.1	74.8	77.2	71.0	66.1	-

4.2.2.5 E11 (Antigenic class C)

Interpretation of NMR spectra (HSQC spectrum was reported in figure 4.19) of the E11 glycopeptide was facilitated because they contained all signals assigned previously to the EIL3$_A$: proton and carbon chemical shift values of residues from **A** to **E** were very similar to that found for EIL-3 $_A$, in agreement with their attribution as α-fucose, α-galactose, α-rhamnose, β-glucose and β-xylose, respectively. Even though, a more careful inspection of this spectrum revealed the presence of other three monosaccharides (labelled **A'**, **F** and **G)** in minor abundance. Therefore, the attention was focused on these new signals.

A' presented the same anomeric signal of **A**, but differed from the C-2, C-3 and C-4 carbon chemical shifts (Table 4.9): the low filed values of these carbons suggested that this residue was a fucose glycosylated at C-2, C-3 and C-4, similarly to those found for the corresponding unit in PBCV-1. The existence of a hyperbranched fucose was supported by the presence of the distal xylose (**G**), that was discriminated by the proximal xylose in the HSQC spectrum (Fig. 4.20) owing to the distinct H-1/C-1 chemical shifts; it had the most deshielded anomeric carbon value (105.7 ppm).

Regarding **F,** it had the densities of all carbons ring completely overlapped with that of rhamnose residue of EPA1 glycopeptide, indicating that this residue was another unit of rhamnose, that differed from the first one because it conveyed of the presence of a distal xylose.

Therefore, E11 was a mixture of two glycoforms (Fig. 4.21): the most abundant presents five monosaccharides, similar to that reported for EIL3$_A$ and the minor glycoform that differed from the first one for the addition of a distal xylose, obtaining a structure similar to that found for EPA1.

Figure 4.19 Expansion of E11 glycopeptide HSQC spectrum measured at 600 MHz, 310 K. In green were reported carbons bearing two hydrogen atoms, which have the opposite sign with respect to the other.
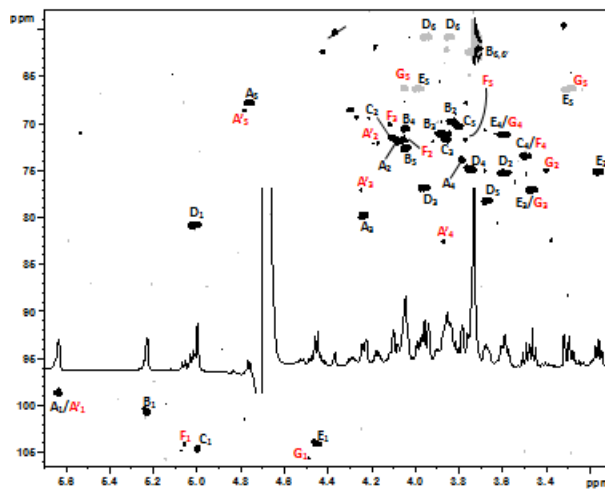
Figure 4.20 The same HSQC spectrum of the figure 3.52, but obtained increasing the intensity of the spectrum. Red letters indicated the new densities with respect to the previous figure. In grey were reported carbons bearing two hydrogen atoms, which have the opposite sign with respect to the other.
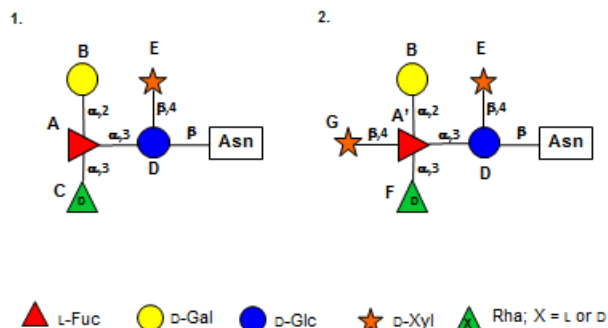
Figure 4.21 Structure of the *N*-glycans produced by E11. Letter labels are those used during NMR assignment.

Table 4.9 (600 MHz, 310 K) Proton and carbon chemical shift values obtained for E11 glycopeptides.

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.63	4.07	4.23	3.78	4.76	1.21
2,3,4-α-ʟ-Fuc	^{13}C	98.4	71.7	79.5	73.6	67.5	16.1
A'	^{1}H	5.63	4.18	4.25	3.87	4.78	1.21
2,3,4-α-ʟ-Fuc	^{13}C	98.4	71.8	76.8	82.3	68.4	16.1
B	^{1}H	5.23	3.83	3.88	4.04	4.04	3.74-3.73
α-ᴅ-Gal	^{13}C	100.5	69.6	70.8	70.3	72.3	62.2
C	^{1}H	4.99	4.10	3.85	3.49	3.80	1.30
α-ᴅ-Rha	^{13}C	104.4	71.3	71.5	73.3	70.1	17.9
D	^{1}H	5.02	3.59	3.96	3.74	3.67	3.95-3.84
3,4-β-ᴅ-Glc	^{13}C	80.6	75.0	76.6	74.6	78.0	60.5
E	^{1}H	4.45	3.16	3.46	3.59	3.98-3.29	-
β-ᴅ-Xyl	^{13}C	103.98	74.88	76.8	70.9	66.0	-
F	^{1}H	5.06	4.05	4.12	3.49	3.77	1.30
α-ᴅ-Rha	^{13}C	103.9	71.6	69.9	73.3	71.5	17.9
G	^{1}H	4.49	3.40	3.46	3.68	4.05-3.30	-
β-ᴅ-Xyl	^{13}C	105.5	74.8	76.8	70.5	66.0	-

4.2.2.6 P91 (Antigenic class A)

^{1}H-NMR spectrum of P91 glycopeptide showed a higher number of peaks with respect to the previous samples, indicating that it produced a highly complex glycan. Comparing this spectrum with that found for EPA1, it was possible to note that it contained all signals assigned previously to the EPA1 glycan, and detailed NMR analysis (HSQC in figure 4.22) confirmed the presence of a hyperbranched fucose (residue **A**), a terminal galactose (**B**), distal and proximal xylose units (**I** and **L**, respectively), and glucose (**E**) *N*-linked to the asparagine. A more careful analysis of NMR spectra revealed that the fucose and galactose units were split (named **A**/**A'** and **B/B'**), and the T-ROESY (Fig. 4.23) spectra allow us to determine that **B** was linked to O-2 of **A**, and **B'** at O-2 of **A'**.

At ca. 5.07 ppm there were three signals overlapped (**C**, **C'** and **D**), but analyzing TOCSY and COSY spectra, it was possible to distinguish three different cross-peaks H1/H2 relative to three different units, suggesting that all three residues had a *manno* configured stereochemistry. Inspection of all NMR spectra identified a 3-linked rhamnose (**C**), a terminal mannose (**D**) and a terminal rhamnose (**C'**), also supported by the proton and carbon chemical shift values (for the last two residues, the values were compared with the respective residue in PBCV-1, resulting very similar; regarding the rhamnose unit, instead, the chemical shifts were compared with that found in literature) (Table 4.10). The presence of a terminal rhamnose indicated that there were two different glycoforms in which the mannose was substituted in a not-stoichiometric fashion.

As for **F** unit, TOCSY correlation from H-1 stopped at H-2, whereas the connections with all other proton rings were visible from H-2. Thus, this residue was a terminal rhamnose and its H-5 resonated at 3.46 ppm, as occur for a β-configured rhamnose. This information was supported by the T-ROESY spectrum, in which H-1 correlated with both H-3 and H-5, a typical pattern of β-configured residue. The same pattern occurred for **G** and **H** residues, indicating that they were other two terminal β-rhamnose units. Of note, **H** unit chemical shift were similar to those of **F**; differently, **G** residue presented a methyl group at O-2, as suggested by the deshielded values of the corresponding carbon (81.74 ppm). In addition, T-ROESY

experiment (Fig. 4.23) revealed that **F** and **H** were linked to residue **I**, instead **G** at O-4 of **I'**.

Therefore, this analysis disclosed that P91 produced more than two glycoforms, which differed for the non-stoichiometric amount of mannose unit and for the presence or not of the methyl group at O-2 of β-rhamnose residue (structures were reported in figure 4.24).
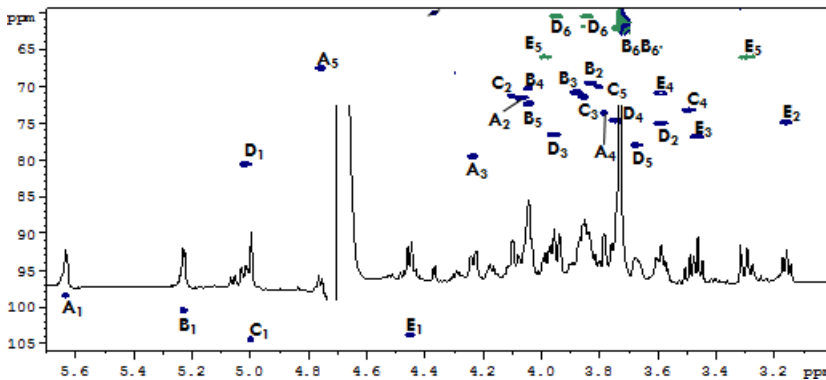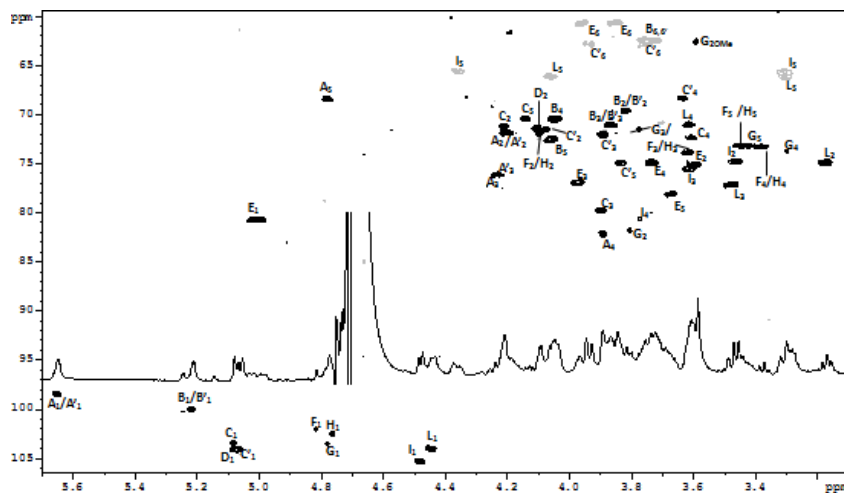

Figure 4.22 Expansion of P91 glycopeptide HSQC spectrum measured at 600 MHz, 310 K. In grey were reported carbons bearing two hydrogen atoms, which have the opposite sign with respect to the other.
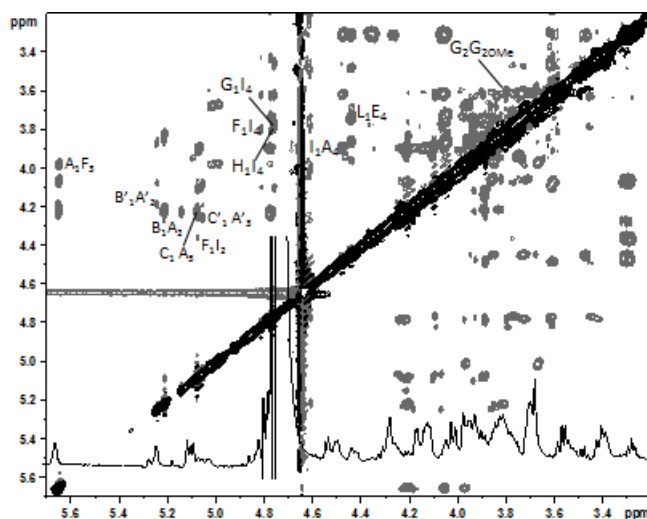


**Figure 4.23** T-ROESY spectrum of P91, recorded at 600 MHz (310 K).
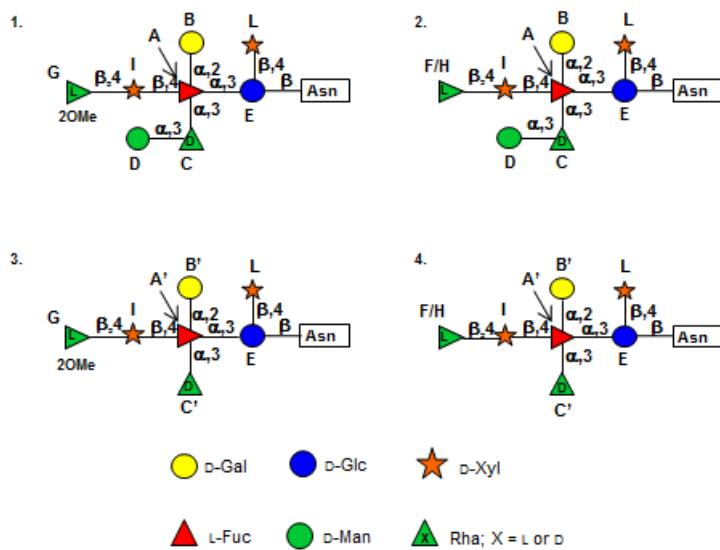
Figure 4.24 Structures of N-glycans of glycoforms of the major capsid protein from P91. Residues are labelled with the letter used during the NMR assignment.

Table 4.10 (600 MHz, 310 K) Proton and carbon chemical shift values obtained for P91 glycopeptides. **G** was substituted at O-2 with a methyl group (chemical shift values ¹H/¹³C 3.60/62.6 ppm).

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	¹H	5.65	4.21	4.24	3.89	4.78	1.33
2,3,4-α-L-Fuc	¹³C	98.4	71.8	76.1	82.1	68.4	16.1
A'	¹H	5.65	4.19	4.22	3.89	4.78	1.33
2,3,4-α-L-Fuc	¹³C	98.4	71.8	76.1	82.1	68.4	16.1
B	¹H	5.25	3.85	3.88	4.06	4.05	3.72-3.75
α-D-Gal	¹³C	100.2	71.0	71.0	70.4	72.4	62.4
B'	¹H	5.22	3.81	3.86	4.03	4.06	3.72-3.75
α-D-Gal	¹³C	100.1	69.6	71.0	70.4	72.5	62.4
D	¹H	5.08	4.21	3.62	3.75	3.64	1.32
3-α-D-Rha	¹³C	103.4	71.1	71.0	74.8	68.3	18.2
E	¹H	5.08	4.11	3.89	3.63	3.82	3.75-3.94
t-α-D-Man	¹³C	104.4	71.4	72.0	68.3	74.9	62.8
D'	¹H	5.06	4.07	3.78	3.53	4.15	1.32
t-α-D-Rha	¹³C	103.9	71.7	71.6	73.1	70.4	18.2
E	¹H	5.03	3.60	3.97	3.73	3.67	3.85-3.97
3,4-α-D-Glc	¹³C	80.7	75.1	76.9	74.8	78.2	60.6
F	¹H	4.81	4.09	3.62	3.38	3.46	1.32
β-L-Rha	¹³C	101.9	71.9	73.8	73.1	73.1	18.1
G	¹H	4.77	3.80	3.62	3.29	3.42	1.33
β-L-2OMeRha	¹³C	103.5	81.7	73.8	73.7	73.2	18.1
H	¹H	4.76	4.09	3.62	3.38	3.46	1.32
β-L-Rha	¹³C	102.4	71.9	73.8	73.1	73.1	18.1
I	¹H	4.48	3.47	3.61	3.77	3.30-4.36	-
4-β-D-Xyl	¹³C	105.3	74.7	75.5	80.6	65.6	-
L	¹H	4.44	3.18	3.48	3.62	3.31-4.06	-
β-D-Xyl	¹³C	103.9	74.8	77.1	71.0	66.1	-

4.2.2.7 P9L10 (Antigenic class A)

Homonuclear and heteronuclear experiments recorded for P9L10 presented several similarities with P91, along with some differences.

The first difference was the presence of the **C** residue at ^{1}H/^{13}C 5.19/100.0 ppm. TOCSY pattern from its anomeric signal showed five correlations and this information combined with those of the other NMR experiments led to the definition of this residue as a pentose in the furanose form, as suggested by the C-4 chemical shift (83.1 ppm) together with the C-5 value (62.8 ppm). Comparison of this unit with that found in the PBCV-1 (De Castro et al. 2013) identified it as a β-configured arabinofuranose residue.

The second was in the β-rhamnose unit: the β-2OMe rhamnose unit found in P91 was missing, even though two anomeric signal (**G** and **H**) were present. In particular, **G** accounted for a terminal β-rhamnose, instead **H** was a 3-linked β-rhamnose, as indicated by the downfield shift of its C-3 (^{1}H/^{13}C 3.71/80.0 ppm).

Finally, the HSQC spectrum (Fig. 4.25) displayed two anomeric carbon signals at 105 and 104 ppm, diagnostic of β configured residues, but a more careful analysis of the TOCSY/COSY and HSQC spectra, revealed that there were two different distal xylose residues (named **I** and **I'**). These units presented the same proton and carbon chemical shift values, except for the C-4 and C-5 signals (Table 4.11). In particular, the C-4 of **I'** resonated at slight high field as well as the C-5, with respect to the corresponding carbon of **I**. Moreover, the NOESY experiment disclosed that they were linked at O-4 of fucose unit, that was split in **A** and **A'**; in particular, **I** was linked to **A'** and **I'** at **A**. Again, the HMBC spectra (Fig. 4.25) revealed that **I** was linked at O-4 by **G** residue, instead **I'** by residue **H**.

Therefore, the P9L10 glycan presented the mannose and arabinofuranose units in non-stoichiometric amount, causing the split of the NMR signals of some residues. All possible structures are summarized in figure 4.26.
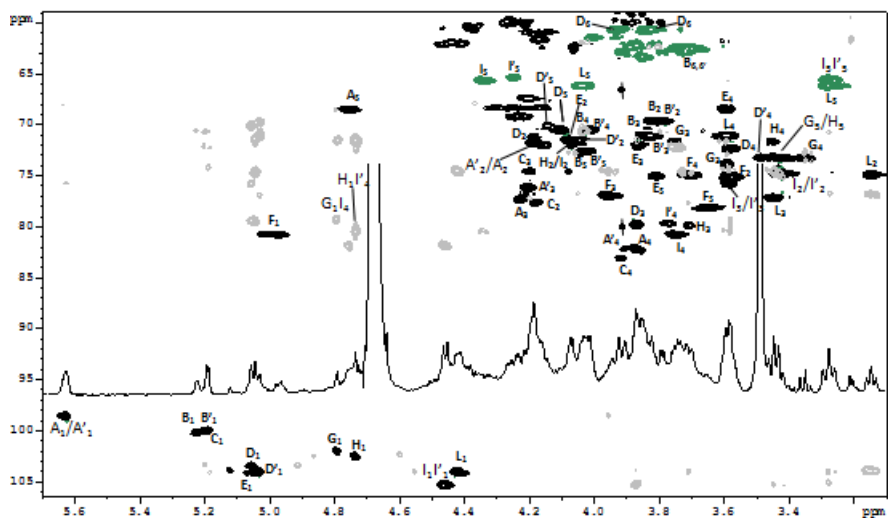
Figure 4.25 Expansion of HSQC spectrum (in black and green) overlapped with HMBC spectrum (in grey) obtained from P9L10. Experiments were recorded at 600 MHz, at 310 K, using acetone as internal standard.
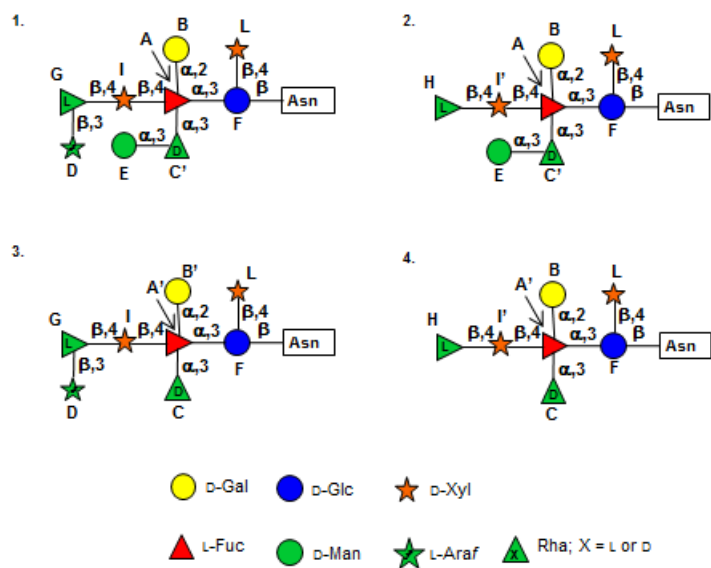
Figure 4.26 Structures of *N*-glycans of the major capsid protein from P9L10. Residues are labelled with the letter used during the NMR assignment.

124

Table 4.11 (600 MHz, 310 K) Proton and carbon chemical shift values obtained for P9L10 glycopeptides.

|		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.66	4.18	4.24	3.88	4.77	1.33
2,3,4-α-L-Fuc	^{13}C	98.5	71.8	77.3	82.0	68.5	16.1
A'	^{1}H	5.64	4.20	4.22	3.93	4.77	1.33
2,3,4-α-L-Fuc	^{13}C	98.5	71.7	76.1	82.1	68.5	16.1
B	^{1}H	5.24	3.84	3.88	4.04	4.05	3.76-3.69
α-D-Gal	^{13}C	100.1	69.6	71.0	70.4	72.5	62.7
B'	^{1}H	5.21	3.80	3.85	4.03	4.04	3.76-3.69
α-D-Gal	^{13}C	99.9	69.6	71.0	70.4	72.5	62.7
C	^{1}H	5.21	4.19	4.22	3.93	3.80-3.74	-
β-L-Ara*f*	^{13}C	99.9	77.6	74.5	83.1	62.7	-
D	^{1}H	5.08	4.20	3.88	3.60	4.13	1.32
3-α-D-Rha	^{13}C	103.4	71.2	79.9	72.2	70.4	18.1
E	^{1}H	5.07	4.09	3.88	3.61	3.82	3.73-3.93
α-D-Man	^{13}C	104.0	71.4	71.9	68.4	75.0	62.8
D'	^{1}H	5.05	4.06	3.77	3.50	4.16	1.32
α-Rha	^{13}C	104.0	71.4	71.5	73.1	70.1	18.2
F	^{1}H	4.99	3.60	3.98	3.72	3.68	3.95-3.85
3,4-β-D-Glc	^{13}C	80.7	75.1	76.9	74.8	78.0	60.6
G	^{1}H	4.81	4.09	3.60	3.36	3.47	1.32
β-Rha	^{13}C	101.9	74.5	73.8	73.2	73.1	18.3
H	^{1}H	4.75	4.08	3.72	3.46	3.45	1.32
3-β-L-Rha	^{13}C	102.4	71.9	79.9	71.6	73.1	18.3
I	^{1}H	4.48	3.44	3.61	3.76	4.36-3.29	-
4-β-D-Xyl	^{13}C	105.3	74.7	75.7	80.7	65.5	-
I'	^{1}H	4.48	3.44	3.61	3.77	4.26-3.29	-
4-β-D-Xyl	^{13}C	105.3	74.7	75.7	79.7	65.4	-
L	^{1}H	4.44	3.16	3.46	3.61	3.29-4.05	-
β-D-Xyl	^{13}C	104.0	74.8	77.1	71.0	66.1	-

4.2.2.8 P9L1 (Antigenic class A)

Comparing the proton spectra with that obtained for P9L10 (Fig. 4.27), it was evident that the two species were equal. This hypothesis was also supported by the HSQC spectrum (Fig. 4.28) that completely overlapped with that obtained from P9L10 (proton and carbon chemical shift value were similar, Table 4.12; the same letters used for P9L10 were maintained through the attribution), even though the low amount of this sample made difficult the identification of the arabino furanose unit that was in a non-stoichiometric fashion (glycoform with this residue was outside to the NMR detection limits). An evidence of the presence of this residue is the finding of a β-rhamnose 3-linked. Thus, P9L1 presents the same glycan structures that were found for P9L10 (a schematic structure of all the possible glycoforms was reported in figure 4.29), in agreement with the placing of this variant in the same antigenic class.
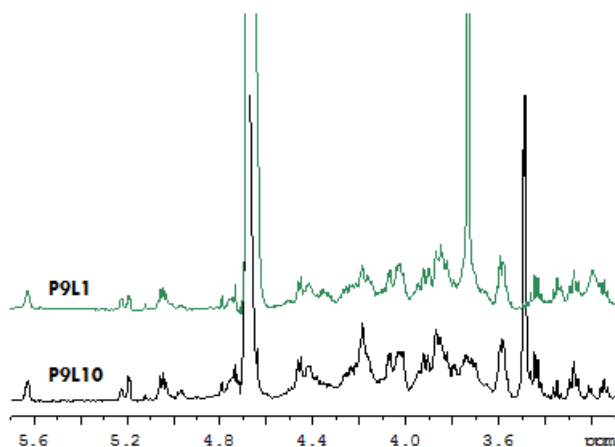
Figure 4.27 Expansion of ^{1}H-NMR spectra from P9L10 and P9L1 glycopeptides recorded at 600 MHz, and 310 K.
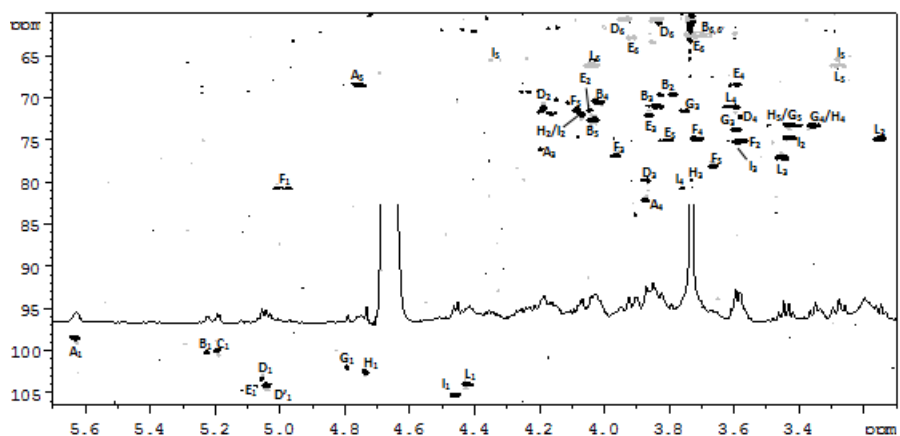
Figure 4.28 HSQC spectrum of the P9L1 glycoprotein recorded at 600 MHz, and at 310 K. In grey were reported carbons bearing two hydrogen atoms, which have the opposite sign with respect to the other.
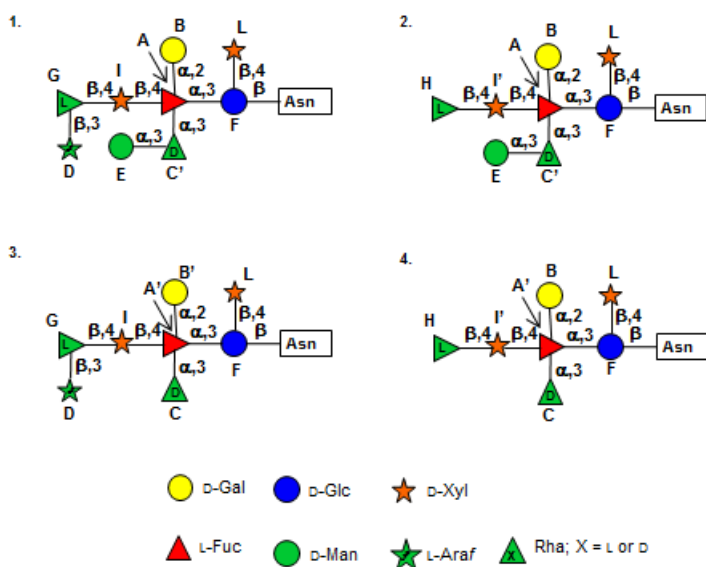
Figure 4.29 *N*-glycan structures of P9L1 glycans.

127

Table 4.12 (600 MHz, 310 K) Proton and carbon chemical shifts obtained for P9L10 glycopeptides. The Arabino furanose unit (**C**) was not determined because was outside to the NMR detection limits.

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.63	4.16	4.19	3.87	4.76	1.30
2,3,4-α-L-Fuc	^{13}C	98.5	71.8	76.1	82.2	68.5	16.2
B	^{1}H	5.23	3.83	3.84	4.03	4.04	3.71-3.68
α-D-Gal	^{13}C	100.3	69.7	71.0	70.5	72.6	62.5
D	^{1}H	5.06	4.18	3.87	3.58	4.11	1.30
3-α-D-Rha	^{13}C	103.5	71.3	79.8	72.3	70.5	18.2
E	^{1}H	5.04	4.08	3.86	3.59	3.80	3.72-3.91
t-α-D-Man	^{13}C	104.1	71.4	72.1	68.4	74.9	62.8
D'	^{1}H	5.03	4.04	3.75	3.49	4.15	1.29
t-α-Rha	^{13}C	104.1	71.4	71.6	73.3	70.1	18.2
F	^{1}H	4.98	3.57	3.96	3.71	3.66	3.94-3.84
3,4-β-D-Glc	^{13}C	80.7	75.1	76.8	74.9	78.1	60.6
G	^{1}H	4.79	4.07	3.59	3.35	3.42	1.33
β-Rha	^{13}C	102.0	71.9	73.7	73.3	73.2	18.5
H	^{1}H	4.74	4.07	3.73	3.76	3.43	1.33
3-β-L-Rha	^{13}C	102.6	71.9	79.7	79.8	73.2	18.5
I	^{1}H	4.46	3.43	3.58	3.76	4.34-3.27	-
4-β-D-Xyl	^{13}C	105.3	74.7	75.5	80.7	65.5	-
L	^{1}H	4.42	3.15	3.45	3.60	3.27-4.04	-
β-D-Xyl	^{13}C	104.0	74.8	77.1	71.0	66.1	-

4.2.2.9 CME6 (Antigenic class F)

Examination of the proton spectrum of CME6 revealed its complexity. It presented several similarities with that of PBCV-1 (De Castro et al. 2013), along with some differences. Proton and carbon chemical shift values (Table 4.13) of **A, B, C, F, I** and **L** residues were similar to the respective monosaccharides in PBCV-1, in agreement with their attribution as α-fucose, α-galactose, β-arabinofuranose, β-glucose, 4-β-xylose, and terminal-β-xylose, respectively.

The main difference was the lack of the anomeric signals relative to rhamnose substituted at O-2 and O-3 with methyl groups, also confirmed by the presence of only one methyl group (^{1}H/^{13}C 3.44/57.3 ppm).

Moreover, the HSQC spectrum (Fig. 4.30) showed three different anomeric carbons (**D, E** and **F**), at around 5.04 ppm, relative to 3-α-Rha, α-Man and t-α-Rha, respectively. The first two residues had the same proton and carbon chemical shift found for the PBCV-1 (De Castro et al. 2013). Instead, the **F** residue was a new entry. The interpretation of this unit was complicate because it overlapped with other monosaccharides, but an accurate analysis of TOCSY experiment displayed one main correlation from the anomeric proton, whereas the connections with all other proton rings were visible from H-2. A more detailed inspection of TOCSY pattern of this H-2, disclosed the presence of two cross-peaks H2/H3 relative to two different rhamnose units. Both residues were terminal rhamnose units, α configured at the anomeric center, in agreement with their C-5 chemical shift value (69.8 ppm), but **F'** presented a methyl group at position O-3, as indicating by the C-3 chemical shift value at 80.8 ppm and by a clear correlation between this carbon with a methyl group in the HMBC spectrum (highlighted with a red circle in figure 4.30). **H** and **I** residues displayed the same spectroscopic pattern described for **D** residue, identifying this unit as a rhamnose, but in both cases the C-5 chemical shift values (at ca. 73.6 ppm) specified that they were a β-configured rhamnose. Moreover, the carbon chemical shift displacement suggested that **I** residue was glycosylated at position O-2 and O-3, instead **H** residue was only glycosylated at O2. Both, **H** and **I** residues were linked to the **L** residue at position four, as revealed the T-ROESY experiment (Fig. 4.31).

Achievement of the CME6 glycopeptide structure was inferred by studying the T-ROESY, which disclosed that **G** was glycosylated at O-4 with **M** and at O-3 with **A** residue, that in turn had **B** at O-2, **L** at O-4 and **D** at O-3. Moreover, the mannose unit (**E**) was connected at O-3 of **D**, instead **H** or **I** residues had **F/F'** residues at O-4. This last information together with the presence of terminal rhamnose with or without the methyl group revealed the existence of different kinds of glycoforms (summarized in figure 4.32) that differed for the non-stoichiometric presence of terminal arabinofuranose or for the methylation at the terminal α-rhamnose residue.
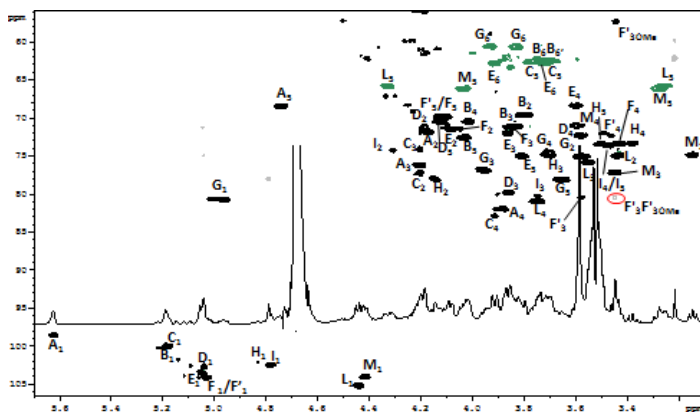
Figure 4.30 HSQC (black and green)/HMBC (gray) experiments of CME6 glycopeptide, measured at 600 MHz, 310 K. In green were reported carbons bearing two hydrogen atoms, which have the opposite sign with respect to the other. Red circle indicates the HMBC correlation of the 3OMe-rhamnose unit.
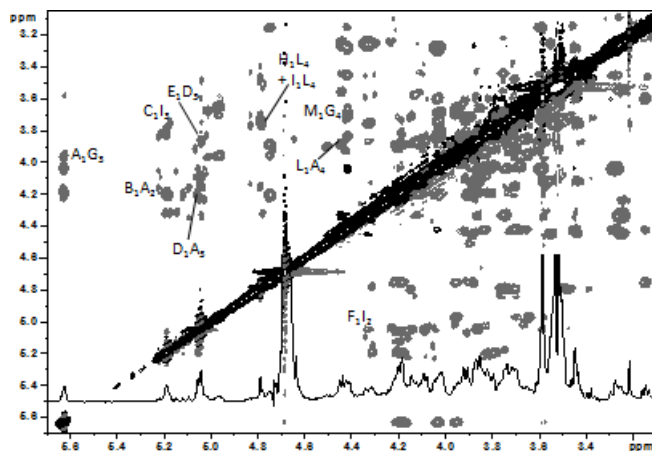
Figure 4.31 T-ROESY spectrum of CME6 glycopeptide, measured at 600 MHz, 310 K.

130
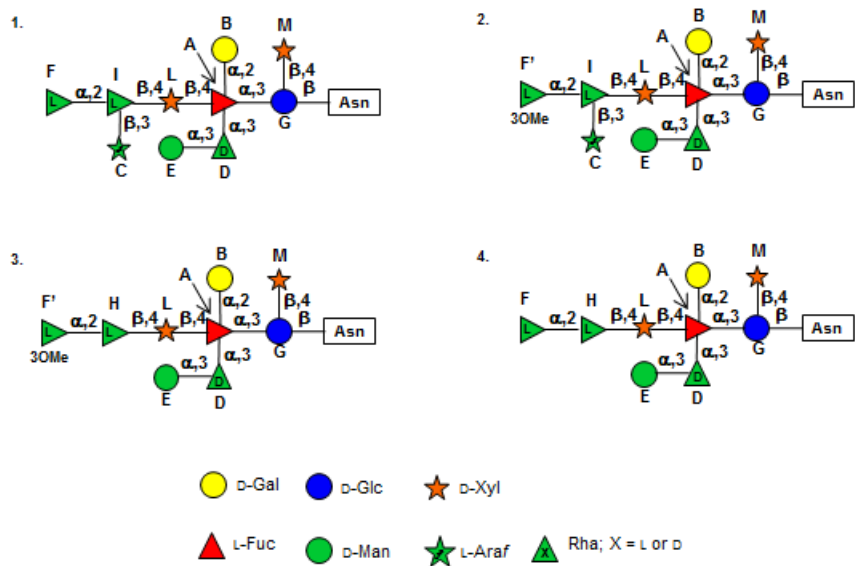
Figure 4.32 Structures of *N*-glycans of CME6 antigenic variant. Residues are labelled with the letter used during the NMR assignment.

Table 4.13 (600 MHz, 310 K) Proton and carbon chemical shifts obtained from CME6 glycopeptide. **F'** was substituted at O-3 with a methyl group (chemical shift values ^{1}H/^{13}C 3.44/57.3 ppm).

		1	2	3	4	5 (5eq; 5ax)	6; 6'
A	^{1}H	5.63	4.17	4.21	3.88	4.75	1.30
2,3,4- α-L-Fuc	^{13}C	98.5	71.8	76.2	81.9	68.4	16.1
B	^{1}H	5.19	3.78	3.83	4.01	4.03	3.75-3.70
α-D-Gal	^{13}C	99.8	69.6	71.1	70.4	72.5	62.2
C	^{1}H	5.18	4.21	4.20	3.91	3.78-3.71	-
β-L-Ara*f*	^{13}C	100.1	77.2	74.0	82.9	62.6	-
D	^{1}H	5.05	4.19	3.86	3.58	4.12	1.29
3-α-D-Rha	^{13}C	103.3	71.2	79.8	72.4	70.5	17.9
E	^{1}H	5.04	4.08	3.86	3.60	3.81	3.72-3.91
α-D-Man	^{13}C	104.1	71.4	72.0	68.4	75.0	62.8
F	^{1}H	5.02	4.05	3.85	3.43	4.11	1.28
t-α-D-Rha	^{13}C	102.7	71.3	71.4	73.3	69.8	18.2
F'	^{1}H	5.05	4.19	3.58	3.47	4.11	1.29
3-α-D-3OMeRha	^{13}C	103.9	71.2	80.5	72.4	76.8	17.9
G	^{1}H	5.12	3.57	3.96	3.70	3.65	3.93-3.83
3,4-β-D-Glc	^{13}C	79.7	75.0	76.8	74.6	78.1	60.6
H	^{1}H	4.79	4.1	3.71	3.38	3.51	1.32
2-β-L-Rha	^{13}C	102.5	77.9	75.0	73.6	73.8	18.2
I	^{1}H	4.78	4.32	3.75	3.50	3.51	1.33
2,3-β-L-Rha	^{13}C	102.5	74.3	80.3	73.6	73.6	18.2
L	^{1}H	4.44	3.45	3.55	3.75	4.33-3.25	-
4-β-D-Xyl	^{13}C	105.2	74.9	75.8	80.9	65.8	-
M	^{1}H	4.42	3.14	3.44	3.60	3.28-4.04	-
t-β-D-Xyl	^{13}C	104.1	74.8	77.2	71.0	66.1	-

4.3 Conclusions

The purpose of this part of the Ph.D. project was the determination of the *N*-glycans constituents of the major capsid proteins, from the antigenic variants of P*aramecium bursaria* chlorella virus 1, to establish the function of the gene A064R.

As previously reported, the mutants are divided into six antigenic classes and we have examined some few representatives of each class: P9L10, P9L1 and P91 (antigenic class A); EPA-1and EPA-2 (class B); E11 and EIL-3 (class C); PIL 6 (class D); CME6 (class F). Therefore, the glycidic portion of each virus (depicted in figure 4.33 along with their respective mutated gene) was purified and determined using chemical and spectroscopic analysis.
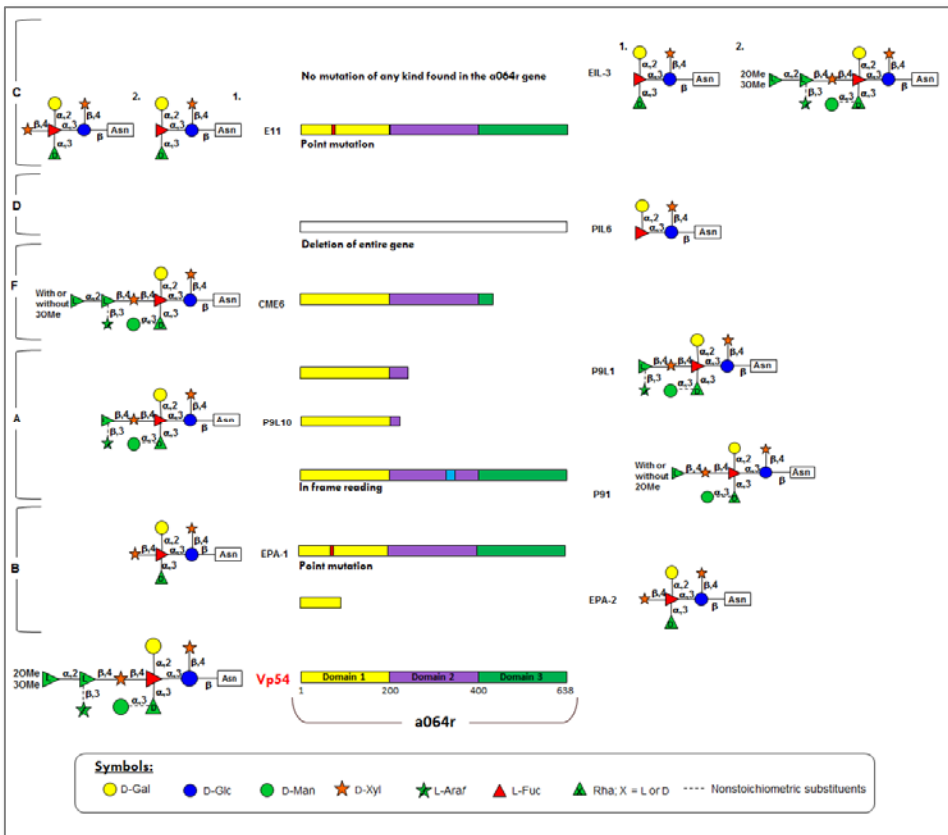
Figure 4.33 Schematic representation of the *N*-glycans of all antigenic variants. Different viruses are grouped according to their antigenic class (reported on the left of the figure). The *N*-glycan structure of the prototype, PBCV-1, is given for comparison. Where are reported two glycoforms, with 1 is indicated the most abundant.

The combination of the structural data (here descripted) with the genetic information allow us to hypothesize the function of the three domains of A064R protein.

In the antigenic class B, the domain one is affected (Fig. 4.33): EPA1 has a point mutation in the DXD motif, critical for the coordination of the sugar-nucleotide donor, instead in EPA2 the gene is truncated. As result, both variants produce the same N-glycan, in which the β-rhamnose unit is missing.

In contrast, the variants that have this domain complete, but present the second mutated or truncated (class A) present this monosaccharide. These results lead to hypothesize that the domain 1 encodes for a **β-(1,4)-L-rhamnosyl transferase**.

Regarding the variants belonging to the class A (P91, P9L10 and P9L1), they have the first domain of A064R intact, whereas the domain 2 is affected. In the P9L10 and P9L1variants, the domain 2 is truncated, while P91 presents an insertion of nine bases that keeps the gene in frame for the reading the downstream coding sequence. These antigenic variants present glycans without the terminal α-L-di-OME-rhamnose unit. Thus, we can suppose that domain 2 encodes an **α-(1,2)-L-rhamnosyl transferase**. Interestingly, in P91, the third domain is intact and it produces a minor glycoform methylated at O-2 of the β-rhamnose unit, unlike to the other two viruses belonging to the same class.

In support to the above hypothesis, is the result obtained for CME6. This variant is the only member of class F, it has a mutation that lead to the truncation of the third domain, without affecting the first two domains. Structural studies revealed that it produces the full N-glycan, namely with both α- and β-rhamnose units, while methylation is not stoichiometric and occurs only at O-3 of α-rhamnose. Therefore, this finding confimrs tt domain 2 is the glycosyltransferase involved in the formation of the α-Rha-(1,2)-β-Rha linkage, while the not stoichiometric methylation of CME6, supports that the third domain is involved in the **methylation** of the last monosaccharide added, α-rhamnose. In support of this hypothesis is the minor glycoform of P91 that presents an incomplete methylation at O-2 of the β-L-rhamnose, probably because this residue is not its ideal substrate for the

methyltransferase domain that should prefer a rhamnose α configured at the anomeric center.

Finally, analysis of the other antigenic variants has provided some clues on what A064R does not encode.

The major glycoform of E11 presents the same *a064r* mutation of EPA1 (Fig. 4.33), but its *N*-glycans miss the distal xylose. The same structure of the glycan is obtained for EIL-3 variant, for which no mutation in this gene occurs. Altogether, this information indicates that the enzyme responsible for the linkage of the distal xylose is not encoded by the *a064r* gene but elsewhere in the genome.

Finally, PIL6 produces the smallest oligosaccharide while A064R is completely deleted. The oligosaccharide matched the conserved core region common to all chloroviruses, except for the distal xylose that is not present, indeed this result proves that A064R is not involved in the biosynthesis of this oligosaccharide which instead is depends on other genes. Interestingly, this variant has a large deletion that includes not only the A064R gene, but also A075L, a strictly conserved gene in the chloroviruses of known structure, but mutated in E11 as well. This gene encodes for a protein in the cupin family and has a close homology with those involved in heparin biosynthesis. On the basis of the N-glycan structure of E11 and P1L6, a possible role for A075L is just the formation of the **β-Xyl-(1,4)-α-Fuc** linkage.

Thus, the above results enable us to deduce the function of a064r gene (Domain 1: β-(1,4)-L-rhamnosyl transferase; Domain 2: α-L-rhamnosyl transferase; Domain 3: methyl transferase) opening the way to further experiments.

In this regard, the result that the first domain of A064R encodes for a β-(1,4)-L-rhamnosyl transferase contradict previous data, which supposed that such domain encodes for a glucosyl-transferase (Zhang et al. 2007). Therefore, the experiments described in the following chapter, have been successfully performed to support our hypothesis.

Chapter 5

Determination of the function of the first domain of
a064r gene

Paramecium bursaria chlorella virus encodes, at least, six putative glycosyltransferases, presumably involved in glycosylation of the its major capsid protein Vp54 (Graves et al. 2001; Wang et al. 1993). One of these, A064R, draws our attention because 18 of 21 antigenic variants of PBCV-1 present mutations in this gene, or it was deleted. As widely described, this gene consists of three different domains, whose functions were hypothesized in the previous chapter by combining structural analyses and the genetic information: Domain 1 encodes for a β-(1,4)-L-rhamnosyl transferase onto a xylose acceptor; Domain 2 for an α-(1,2)-L-rhamnosyl transferase onto a rhamnose acceptor; Domain 3 for a dimethyl transferase specific for α-rhamnose.

Interestingly, the hypothesis that the first domain of A064R encodes for a β-(1,4)-L-rhamnosyl transferase contradicts previous data, which suppose that such domain encodes for a glycosyltransferase using as donor UDP-glucose (Zhang et al. 2007). Therefore, we tested the first domain of A064R (A064R-D1) to demonstrate our previsions.

To this end, A064R-D1 was expressed into *Escherichia coli* as fusion protein with GST to simplify the next purification steps, the donor (UDP-Rha) was generated as explained in the next paragraph, these aspects were realized through the collaboration with prof. Michela Tonetti at the University of Genoa (Italy). The oligosaccharide acceptor was a synthetic product provided from prof. Todd Lowary from the University of Alberta (Canada).

In this frame our working scheme had three targets: synthesis of UDP-β-Rhamnose, biochemical assays to determine different features of A064R-D1, structural characterization of the product obtained with A064R-D1 to confirm its activity.

5.1 Synthesis of UDP-β-L-rhamnose

On the basis of the general biosynthetic pathways, the activated form of L-rhamnose most widely used in bacteria is UDP-β-L-Rha, accordingly we have considered this and not others, as GDP-α-L-Rha. In addition, literature data clearly pointed to the affinity of A064R-D1 to UDP-nucleotides, while others could not fit in the recognition pocket of the enzyme.

To prepare UDP-β-Rha we exploited the biosynthetic pathway described for this donor and the enzymatic tools available from prof. Tonetti. UDP-Rha is commercial but its cost is prohibitive to perform the experiment planned.

The reactions pipeline needed to prepare this compound are represented in Fig. 5.1, UDP-glucose-4,6-dehydratase (UGD) converts UDP-α-D-glucose to UDP-4-keto-6-deoxyglucose, that in turn reacts with another enzyme, L780, which has 3,5-epimerization and 4-reduction activities, getting the UDP-β-L-rhamnose (Fig. 5.1). This system needs in addition NADPH therefore we added glucose-6-phosphate (G6P) and the glucose-6-phosphate-dehydrogenase (G6PD) to regenerate NADPH necessary for L780 activity.

UGD: UDP-α-D-glucose-4,6-dehydratase
L780: UDP-4-keto-6-deoxy-D-glucose-3,5-epimerase-4-reductase
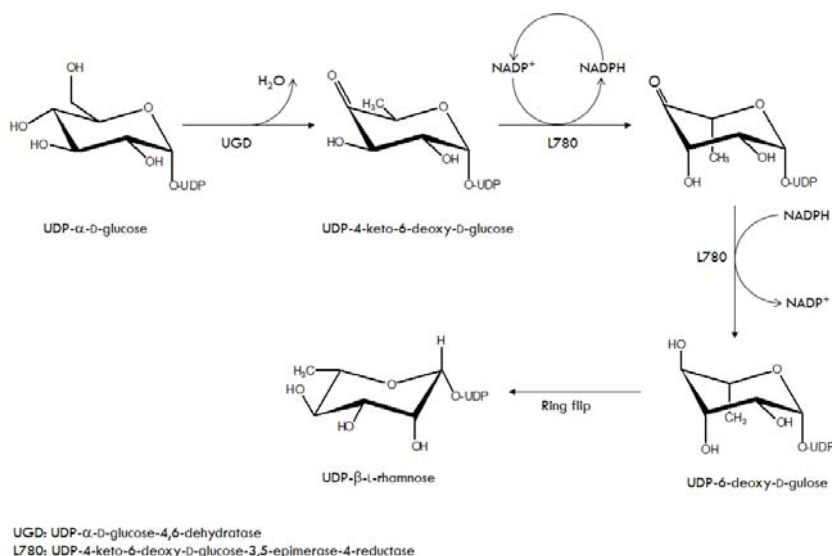
Figure 5.1 Enzymatic synthesis of the β-L-rhamnose.

This reaction was performed in not deuterated phosphate buffer supplemented with Mg^{2+} 2 mM (G6PDis Mg dependent), at room temperature for 2 hours. Then, the mixture was checked via NMR, to verify the UDP-Glc conversion in UDP-Rha. The ^{1}H-NMR spectrum (Fig. 5.2) detected the presence of UDP-rhamnose and the knowledge of the starting UDP-Glc amount, it was possible to determine the final concentration of UDP-Rha. The mixture contained UDP-Gal (10 %) which is a contaminant of the commercial UDP-Glc sold from Sigma, unreacted UDP-Glc (11%), UDP-4-keto-Glc intermediate (29%) and UDP-Rha (51%), which translated in a final UDP-Rha concentration 3.3 mM. UDP-Rha is an extremely labile activated sugar, indeed it was not purified from the other components and its solution was stored at -80°C in aliquots.
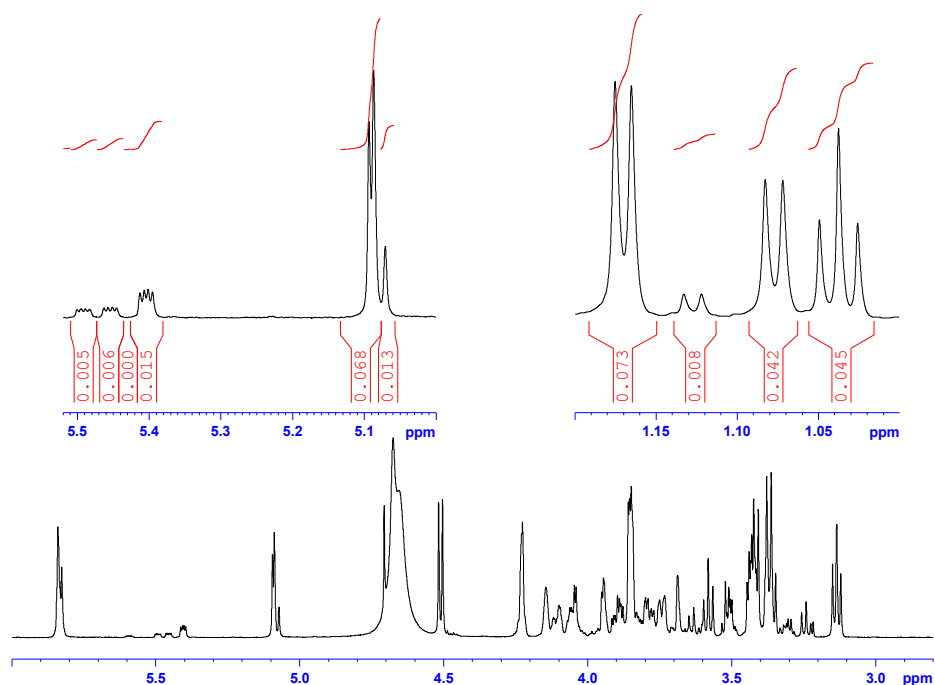
Figure 5.2 Bottom: expansion of the ^{1}H-NMR spectrum of reaction mixture, recorded at 600 MHz, at 298 K in H_2O. B. Top-left: expansion of the anomeric area and integration of the important peaks. Top-right: expansion of the methyl area and integration of the important peaks. Of note, anomeric proton of UDP-Rha (5.07 ppm) partially overlap with H-1 of G6P, and its intensity was strongly suppressed because of the high power used to reduce the water signal, indeed its ratio versus the 4-keto intermediate was appreciated by integration of the corresponding methyl groups at high field (1.17 ppm for UDP-Rha versus 1.08 ppm of the 4-keto intermediate). Intense signals at 5.1 and 4.5 ppm, are from G6P which was used in four-fold excess with respect starting amount of UDP-Glc.

5.2 Bioluminescent assay

To prove that the enzyme, encoded by the first domain of *a064r* gene, it is specific for the UDP-β-L-rhamnose, the UDP-Glo™ Glycosyltransferase assay was performed. It is a bioluminescent assay useful to monitoring the activity of glycosyltransferases (GTs).

Glycosyltransferases catalyze the transfer of a sugar moiety from an activated donor sugar (nucleotide-glycosyl donor) onto saccharide acceptors, releasing the UDP. This assay exploits the formations of UDP, converting it to ATP, in a one-pot luciferase reaction, and the light generated is detected with a luminometer. Thus, the light output is proportional to the concentration of UDP and, from this parameter, we deduce the GT activity.

The analysis starts with the construction of the standard curve of UDP, to estimate the amount of the UDP produced in the glycosyltransferase reaction and to detect the assay's trend (Table 5.1 and Fig. 5.3). Each curve was constructed with twelve points, that correspond to 12 solutions of UDP at different concentrations (Table 5.1) in the reaction buffers: PBS, Mg^{2+} 2 mM for A and PBS and Mn^{2+} 2mM for B. Luminescence values were plotted (Fig. 5.3), and the two curves obtained indicate the high accuracy of the assay, in both cases.
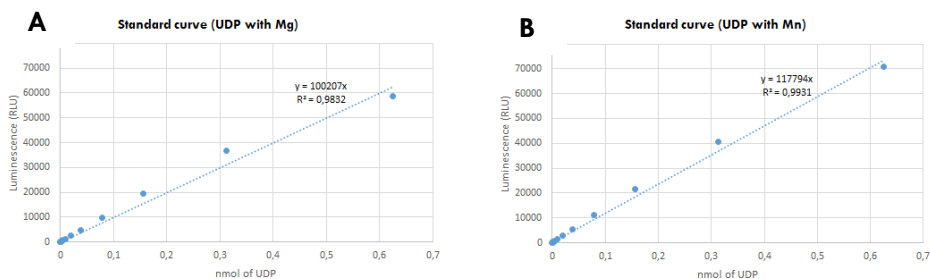
Figure 5.3 UDP standard curves of luminescence obtained for the two buffers of the reactions: A composing by Mg^{2+} ion and UDP dissolved in PBS buffer; B Mn^{2+} ion and UDP in PBS. Luminescence is expressed in Relative light unit (RLU).

Table 5.1 Luminescence values recorded for the two curves. A indicates the solutions containing UDP and Mg^{2+} ion dissolved into PBS buffer; instead, in B are present the UDP and Mn^{2+} ion into the PBS.

	A		B	
	UDP (nmol)	Luminescence (RLU)	UDP (nmol)	Luminescence (RLU)
1	0,625	58678	0,625	70688
2	0,3125	36712	0,3125	40563
3	0,15625	19333	0,15625	21399
4	0,078125	9814	0,078125	11045
5	0,0390625	4893	0,0390625	5505
6	0,019531	2544	0,019531	2809
7	0,009766	1268	0,009766	1418
8	0,004883	624	0,004883	709
9	0,002441	319	0,002441	389
10	0,001221	180	0,001221	186
11	0,00061	101	0,00061	107
12	0	28	0	30

Indeed, we answered to several queries:

- cation dependence of the transferase: literature data support the notion that A064R-D1 is Mn-dependent and not Mg-dependent.

These reactions made use of A064R-D1 with the synthetic substrate and the UDP-Rha, synthetized previously, and were performed in two buffers that differed for the cation used: Mg^{2+} or Mn^{2+}. These reactions were performed maintaining constant the ratio UDP-Rha - synthetic substrate (2:1 ratio; 2.470 nmol – 1.250 nmol, respectively) and changing the amount of the GT (Glycosyl-transferase, alias A064R-D1), that decreases proceeding from the first solutions to the last, in which was completely absent (Table 5.2). The reactions were incubated at 25 °C, for one hour, thereafter the UDP-detection reagent was added in all solutions and they were kept for another hour, in the same condition described above. The UDP-detection reagent was used to terminate the reactions and to trigger the luminescence reaction. Data were obtained as duplicates.

Table 5.2 Luminescence values obtained for the two reactions, in which the amount of the UDP-rhamnose and of the synthetic substrate are maintained equal among all solutions (2.470 nmol and 1.250 nmol, respectively), instead the amount of the glycosyltransferase increases. In each reaction is present either manganese or the magnesium ion as specified in the table. Note that the luminescence values here reported are normalized removing the background noise, constituted by the last point, in which only all components of the reactions are presented, except for the GT.

	Reaction with Mg^{2+} ion			Reaction with Mn^{2+} ion		
	GT (ng)	Replica 1	Replica 2	GT (ng)	Replica 1	Replica 2
1	112,500	1004	970	112,500	7323	8923
2	56,250	220	248	56,250	3414	2157
3	28,125	47	61	28,125	796	477
4	14,063	35	3	14,063	201	171
5	7,031	18	1	7,031	41	111
6	3,516	-38	-68	3,516	74	116
7	1,758	9	-43	1,758	36	32
8	0,879	43	-74	0,879	-51	41
9	0,439	33	-31	0,439	67	-26
10	0,220	15	-33	0,220	-35	31
11	0,110	-16	70	0,110	-22	99
12	0,000	0	0	0,000	0	0

Thus, the luminescence values were plotted (Fig. 5.4). The curves are very indicative: the reaction of the transfer of β-rhamnose occurs in both cases, but is more efficient when Mn^{2+} is used, as given by the major consumption of UDP-Rha.
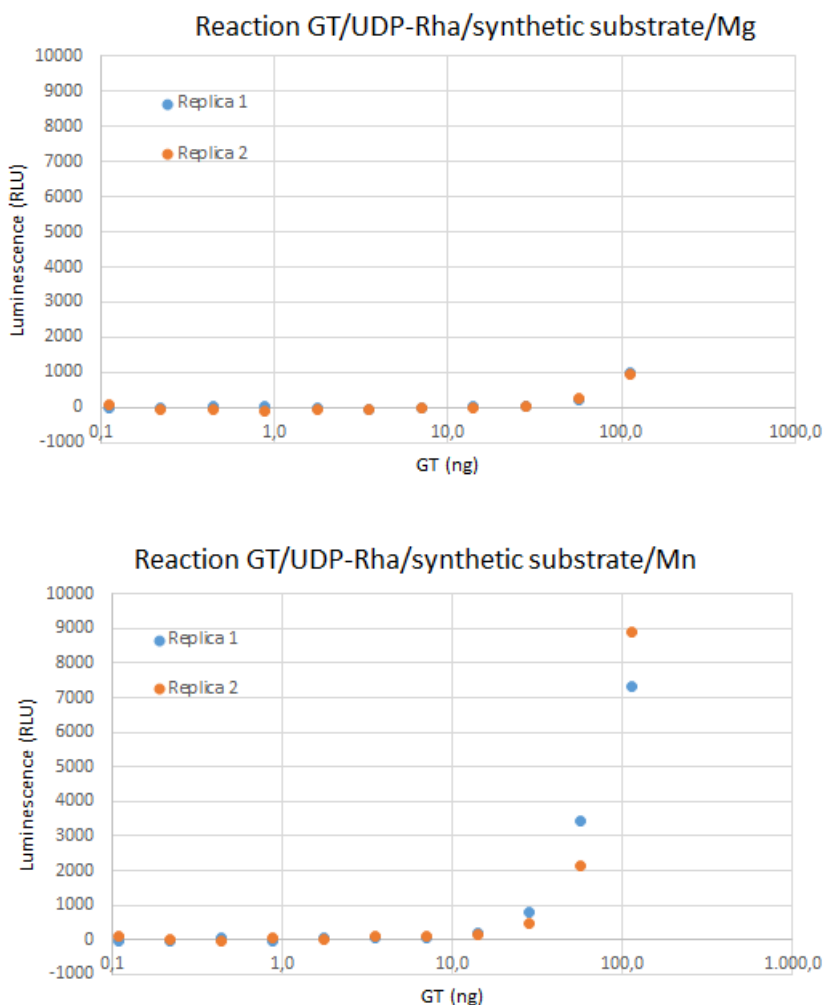
Figure 5.4 Detection of the glycosyltransferase activity in two conditions: with magnesium (top) or manganese ion (bottom). In each graph, two curves are depicted, which relate to two replicas of the same reaction, as specified in the legend.

- Specificity of A064R-D1 for acceptor or donor substrates.

In these two experiments, we changed the type of acceptor and donor used. In the first case, we used xylose monosaccharide as acceptor and kept UDP-Rha as donor. In the second, instead, the nucleotide donor was substituted with the UDP-glucose while the synthetic product was used as acceptor. This last experiment was

necessary to check the prevision reported in literature that had foreseen as nucleotide donor UDP-glucose.

In both reactions, the donor-acceptor ratio was maintained equal to the previous reactions (2:1 ratio), and constant along all solutions. GT concentration, instead, decreases from the first solutions up to the last, in which it was not present. These reactions were performed in presence of the both cations (Mg^{2+} and Mn^{2+}), for 1 h at 25 °C, as described before.

Thus, the luminescence was recorded (Table 5.3) and the values plotted (Fig. 5.5).

Table 5.3 Luminescence values obtained for two reactions: on the left are reported value of the reaction in which the acceptor substrate is the xylose residue. Its concentration is maintained constant among all solutions (1.25 nmol), as well as the concentration of the UDP-rhamnose (2,470), instead the GT amount changes. On the right, are tabled values obtained for the reaction in which the donor was changed: the UDP-Glc is present. As before, also in this case the amount of the donor and acceptor are kept constant (2.470 and 1.250 nmol, respectively), whereas the GT amount decreases. In each reaction are present both manganese and magnesium ions. Note that the luminescence values here reported are normalized as described before (Table 5.2).

| | Reaction with Xyl | | Reaction with UDP-Glc | |
	GT (ng)	Replica 1	GT (ng)	Replica 1
1	112,500	3207	112,500	95
2	56,250	843	56,250	87
3	28,125	222	28,125	69
4	14,063	17	14,063	52
5	7,031	-50	7,031	38
6	3,516	-60	3,516	-14
7	1,758	-86	1,758	-125
8	0,879	-87	0,879	-5
9	0,439	-36	0,439	-45
10	0,220	-52	0,220	-15
11	0,110	-11	0,110	62
12	0,000	0	0,000	0
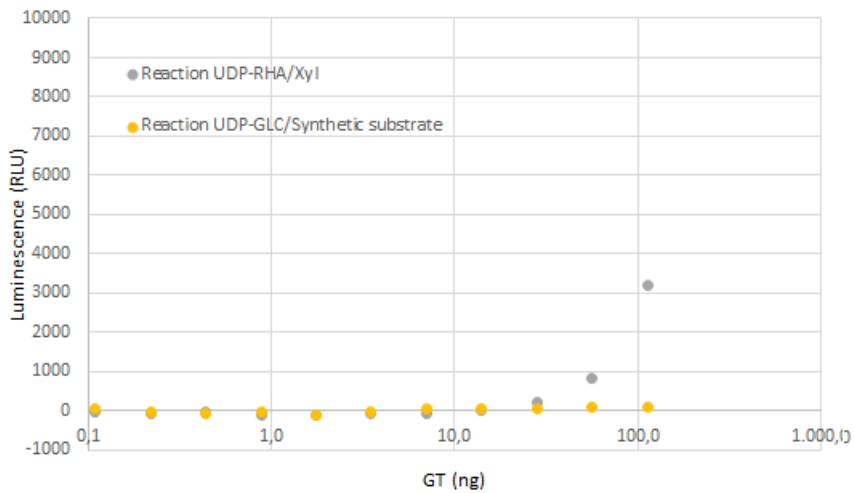
Figure 5.5 Curves relative to two different reactions: in gray is indicated the reaction between the UDP-Rha and Xyl, in presence of the enzyme and both Mn and Mg ions; in yellow is the reaction between the UDP-Glc and the synthetic substrate, in presence of the enzyme and both ions. All reactions are conducted into phosphate buffer.

The curve obtained for the reaction with xylose as acceptor, indicates that the reaction occurs, but at a lower rate if compared to that with the ideal substrate (Fig. 5.4). Therefore, the enzyme is able to transfer rhamnose residue xylose as well, but with different efficiency. Regarding the reaction in which the UDP-glucose was used as donor, instead, the result confirms that this nucleotide is not a substrate for A064R-D1.

5.3 Structural characterization of the product obtained
with A064R-D1 to confirm its activity

The hypothesis made from the work on the antigenic variants of PBCV-1 foresees that A064R-D1 is a glycosyl transferase able to catalyze the β-L-Rha-(1,4)-Xyl linkage, indeed it needs an activated sugar precursor and a suitable donor.

Indeed, we used as acceptor a synthetic substrate very similar to the structure of the EPA1 antigenic variant glycans and contained the hyperbranched fucose, substituted with a galactose, xylose and rhamnose residues, and capped with a small lipophilic tail, octyl residue, at the reducing end (Fig. 5.6). The presence of the lipophilic arm simplified the recovery of the oligosaccharide at the end of the reaction because desalting was performed by using reverse-phase cartridges.

Indeed, large scale reaction was performed in PBS buffer supplemented with Mn^{2+} ions, with a 20% excess of the sugar donor, UDP-Rha with respect to the substrate, to ensure the formation of only one oligosaccharide and not a mixture still containing the starting oligosaccharide acceptor. For this large scale reaction, I used A064R-D1 protein fused to GST and anchored to the resin, so that purification of the final oligosaccharide was limited to the soluble components of the reaction, namely: PBS, free UDP, traces of UDP-Gal, UDP-Glc, UDP-4-keto intermediate, unreacted UDP-Rha, and G6P which was used in large excess for UDP-Rha synthesis.

Thus, all components were mixed and reaction left overnight under stirring on a moving plate. No monitoring of the progress of the reaction was possible, because Mn^{2+} is a ferromagnetic nucleus and hampers the recording of NMR spectra.

Confirm of the activity of A064R-D1 was gained after cartridge purification of the sample by NMR analysis of the oligosaccharide. Indeed, Fig. 5.7 reports the spectrum of the final product along with that of the starting oligosaccharide, recorded at the same condition. Along with the appearance of a new signal at 4.81 ppm, interpretation of 2D NMR spectra confirmed that the new signal is

associated to a β-rhamnose unit linked at position 4 of the xylose unit, exactly as foreseen from the analysis of the antigenic variants.
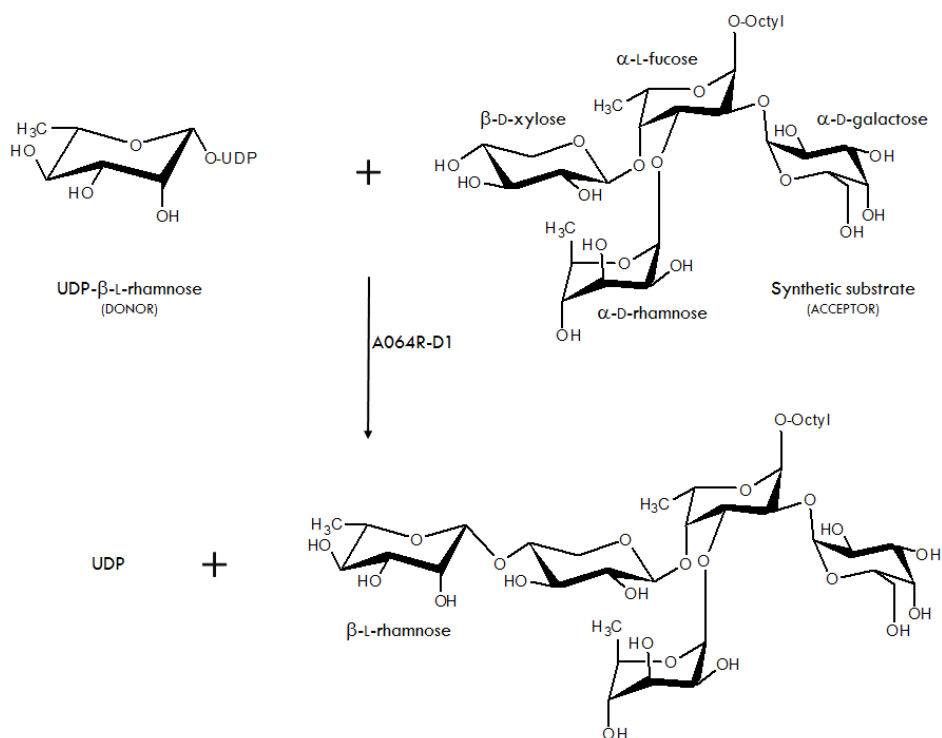
Figure 5.6 Schematic reaction of the β-L-rhamnosyl transferase, codified by the first domain of a064r gene, that transfers the rhamnose unit form the donor to the acceptor.
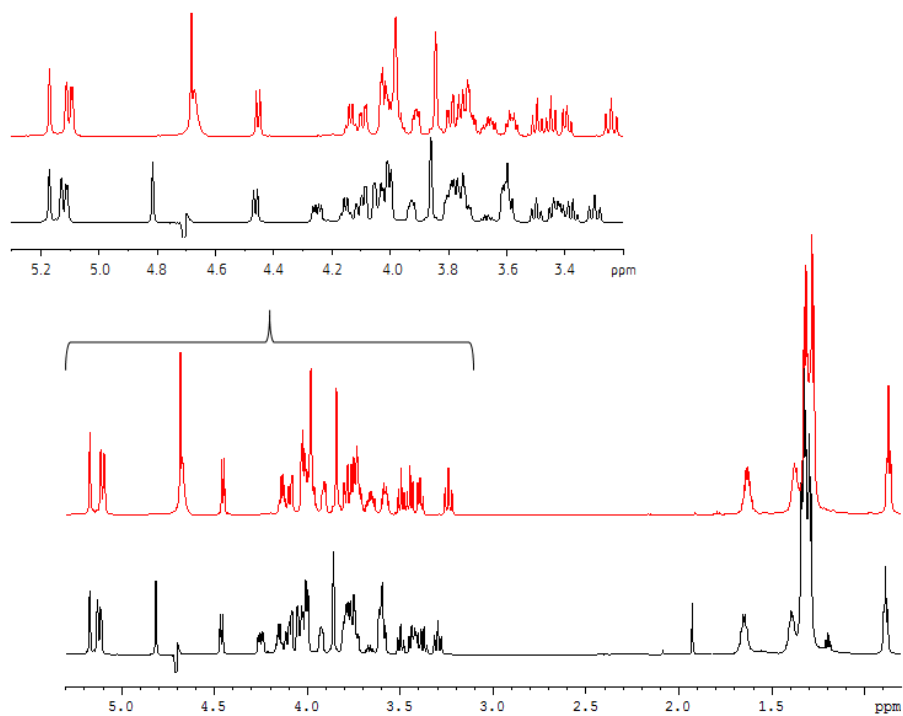
Figure 5.7 Bottom: expansion of the ¹H-NMR spectrum of the synthetic substrate before (in red) and after (in black) the reaction, recorded at 600 MHz, at 310 K in D₂0. Top-left: expansion of the anomeric and carbinolic area of both spectra. Of note, the appearance of a new signal at 4.81 ppm, relative to the anomeric proton of β-rhamnose linked at position 4 of the xylose unit.

5.4 Conclusions

This section of the Ph.D. project aims to prove our hypothesis about the function of the first domain of the enzyme encoded by the a064r gene. We supposed that it encodes for a β-L-rhamnosyl transferase, we demonstrated it, along with confirming other features that were only supposed. All this was possible by using a bioluminescence assay and performing the reaction at a large scale to analyse the biosynthetic product.

More in detail, our analysis demonstrate that A064R-D1 is a glycosyltransferase able to form the β-L-Rha-(1,4)-Xyl linkage, that it reacts only with the UDP-rhamnose and not with the UDP-glucose. In addition, we confirmed that it is a manganese-dependent transferase and that it is able to use also simple xylose as acceptor substrate.

These results lay the basis for further experiments, with the aim of to prove the activities of the other two domains of the a064r gene.

Chapter 6

Reconstruction of Major Capsid Protein Vp54 of *Paramecium bursaria* chlorella virus 1

Paramecium bursaria chlorella virus was widely studied because its unusual aspects, and in 2002 its major capsid protein (Vp54) was described using X-ray analysis (Nandhagopal et al. 2002). Its structure was crystallized at 2.5 Å and 2.0 Å resolution and deposited in Protein data bank (PDB ID codes 1J5Q and 1M3Y, respectively) (Nandhagopal et al. 2002): they contained four (1M3Y) and two (1J5Q) independent polypeptide chains in the crystallographic asymmetric unit. But, the glycan structures were unknown; consequently, the electron density data were performed fitting sugars that now we know to be inappropriate: glucosamine (not glucose) was placed near the glycosylated asparagine and mannose was used to complete the electron densities.

The knowledge of the glycans produced by PBCV-1 (De Castro et al. 2013), prompted a re-examination of the previous MCP X-ray structure, with the intent to understand which role, these glycans play in the global architecture of the MCP.

Thus, in collaboration with the group of Professor Rossmann, a detailed description of the full glycoprotein was achieved by combining two techniques, X-ray diffraction and Molecular Modeling, overcoming the limitations posed by each of them. Indeed, the crystallographic analysis was able to identify the protein backbone and the monosaccharides located close to it, but it failed to detect those further from the glycosylation sites, owing to their flexibility. On the other hand, the molecular modelling approach provided the missing information, adding the absent monosaccharides, and also examined the conformational behavior of the whole molecule.

Therefore, the aim of this part of Ph.D. project was the determination of the three-dimensional structure of the chlorovirus PBCV-1 major capsid glycoprotein Vp54 implementing the X-ray data with the molecular modelling technique.

The construction of the glycosylated MCP was performed into two steps: i) analysis of the conformation of *N*-glycans in the free form (not bound to the protein); ii) evaluation of their overall flexibility when linked to the MCP.

In both analyses, the dihedral angles of the monosaccharide junctions and of the glucose-asparagine connection were examined. More in detail, the monosaccharide junctions were defined by Φ (O_5-C_1-O-C_n) and Ψ (C_1-O-C_n-C_{n+1}) angles, in agreement with the X-ray definition (Fig. 6.1). Instead, four dihedral angles described the glucose-asparagine connection: Φ: O_5-C_1-N-C_γ; Ψ: C_1-N-C_γ-C_β; W: N-C_γ-C_β-C_α; and X: C_γ-C_β-C_α-$C_{C=O}$ (Fig. 6.2).
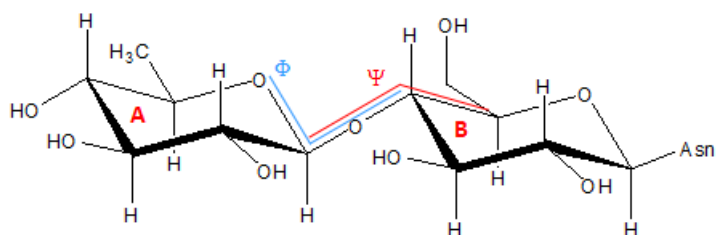
Figure 6.1 Explanation of the dihedral angles used for the glycosydic bond according with X-ray definition. Φ is defined as O_5-C_1-O-C_n; Ψ is defined as C_1-O-C_n-C_{n+1}, where n indicates the number of carbon involved in the bond of the non-reducing monosaccharide.
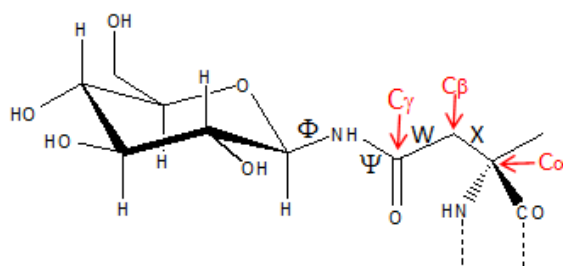
Figure 6.2 Dihedral angles annotation used for the glucose-asparagine junction. Φ is defined as O_5-C_1-N-C_γ; Ψ: C_1-N-C_γ-C_β; W: N-C_γ-C_β-C_α; and X: C_γ-C_β-C_α-$C_{C=O}$ O_5-C_1-O-C_n.

Professor Rossmann re-examined the previous X-ray diffraction data of the Vp54 crystals (Nandhagopal et al. 2002), considering the *N*-glycans of the capsid protein from PBCV-1, which were elucidated via 2D-NMR (De Castro et al. 2013). The chain A (Fig. 6.3A) of the best resolved MCP diffraction data (1M3Y) was used to evaluate which monosaccharides were present, if they were in the right ring conformation, and the dihedral angles (Table 6.1) adopted by the glycosidic junction for residues well resolved.

Table 6.1 Dihedral values measured for each chain of RX structure of Vp54. Bold numbers indicate residue not in the typical chair conformation. Dihedrals are defined as follows: $\Phi= O_5\text{-}C_1\text{-}O\text{-}C_n$; $\Psi= C_1\text{-}O\text{-}C_n\text{-}C_{n+1}$.

	Chain A				Chain B				Chain C				Chain D			
	280	302	399	406	280	30	399	406	280	302	399	406	280	302	399	406
Φ_H	-74	-81	-119	-87	-71	-85	-124	-79	-78	-77	-116	-79	-72	-77	-122	-87
Ψ_H	167	-170	-165	-174	163	-173	-159	-179	170	180	-174	152	162	-166	-165	172
W_H	-167	127	105	133	-167	130	100	137	-168	128	105	137	-167	123	105	136
X_H	64	-178	108	-61	64	179	109	-61	65	-179	108	-58	65	179	105	-60
Φ_N	**-50**	**-43**	**-32**	-	**-62**	**-56**	**-64**	-	**-49**	**-53**	**-58**	-	**-61**	**-60**	**-52**	-
Ψ_N	**-124**	**-135**	**-127**	-	**-121**	**-123**	**-112**	-	**-125**	**-114**	**-127**	-	**-136**	**-121**	**-106**	-
Φ_A	-69	-71	-64	**48**	-62	-66	-62	-32	-69	-75	-63	-	-75	-75	-64	**-82**
Ψ_A	155	151	151	**-144**	146	143	139	-92	151	149	156	-	155	158	138	**-56**
Φ_E	49	71	67	-	58	64	57	-	48	61	59	-	61	68	67	-
Ψ_E	113	94	101	-	116	95	95	-	120	96	97	-	110	94	92	-
Φ_F	**62**	63	68	-	**76**	61	71	-	**73**	70	73	-	**65**	70	54	-
Ψ_F	**-124**	-126	-120	-	**-108**	-118	-123	-	**-127**	-136	-119	-	**-118**	-128	-129	-
Φ_G	59	75	**-137**	-	-	60	**-112**	-	**-120**	81	**-112**	-	49	70	**101**	-
Ψ_G	125	113	**145**	-	-	126	**154**	-	**141**	116	**154**	-	120	113	**165**	-
Φ_M	**-48**	**-30**	**-54**	-	**-23**	**-3**	**-28**	-	**10**	**-36**	**-26**	-	**-28**	**-40**	**-44**	-
Ψ_M	**-128**	**-125**	**-101**	-	**-144**	**-142**	**-149**	-	**-170**	**-134**	**-140**	-	**-122**	**-114**	**-124**	-
Φ_I	**-26**	45	-	-	**26**	69	-	-	**3**	-134	-	-	**-63**	12	-	-
Ψ_I	**50**	-24	-	-	**96**	-45	-	-	**38**	99	-	-	**77**	9	-	-
Φ_C	**140**	-68	-	-	**-179**	-54	-	-	**-141**	-74	-	-	**-173**	-62	-	-
Ψ_C	**-134**	-96	-	-	**-157**	-113	-	-	**180**	-97	-	-	**-166**	-110	-	-

Chain A, at [302]Asn, presented seven of its nine sugar residues in the chair conformation, while the other two were distorted (Fig. 6.3C). At [280]Asn only five monosaccharides adopted the exact conformation, and the arabinose unit was not detected (Fig. 6.3B). At [399]Asn, the last two sugars of the glycan were not detected and three of those remaining were distorted (Fig. 6.3D). Finally, at [406]Asn, only glucose and fucose (that was in not a canonical chair conformation) were detected (Fig. 6.3E). Regarding the other protein chains in the 1M3Y structure, some dihedral angles of the N-glycans were slightly different compared to those of the N-glycans of chain A, some more monosaccharides were distorted and some other were undetected.
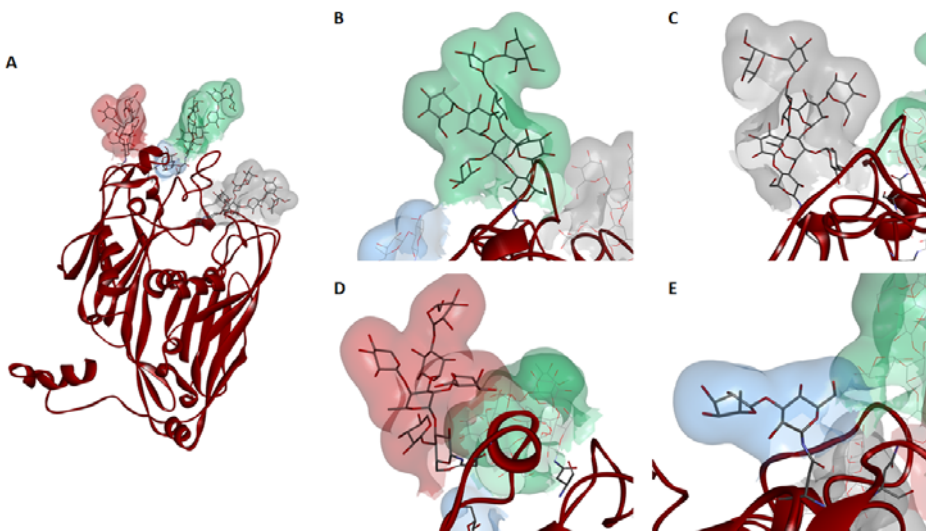
Figure 6.3 A. Chain A of the X-ray of the best resolved MCP diffraction data (1M3Y). B, C, D, E. Expansions of each chain: [208]Asn glycan, [302]Asn glycan, [399]Asn glycan, [406]Asn glycan, respectively.

Interestingly, these refined crystal structures revealed that there was no O-glycosylation at either Ser-57 or Ser-387 and resolved the density for the first 24 amino acids of Vp54 for each crystallographic independent polypeptide, differently to the original structure (Nandhagopal et al. 2002). These 24 amino acids were organized in two short α-helices connected by a linker at the internal

154

face of the MCP trimer. Comparing the crystal structures with the averaged cryoEM electron density (Zhang et al. 2011) of a Vp54 trimer resulted that these helices may help to stabilize the interactions between neighboring capsomers and minor capsid proteins located on the inner surface of Vp54.

6.2 Conformational analysis of *N*-glycans in the free form

PBCV-1 has four N-linked glycans on its major capsid protein, Vp54. These N-glycans are very similar and four glycoform exists that differ for the non-stoichiometric presence of the arabinose and mannose residues (Fig. 6.4). The most abundant glycoform (here referred to it as glyco1), has nine neutral residues and it is located at [302]Asn, [399]Asn and [406]Asn, whereas the other glycoform (glyco2), comprised ten residues (with arabinose) and is situated at [280]Asn (De Castro et al. 2013).

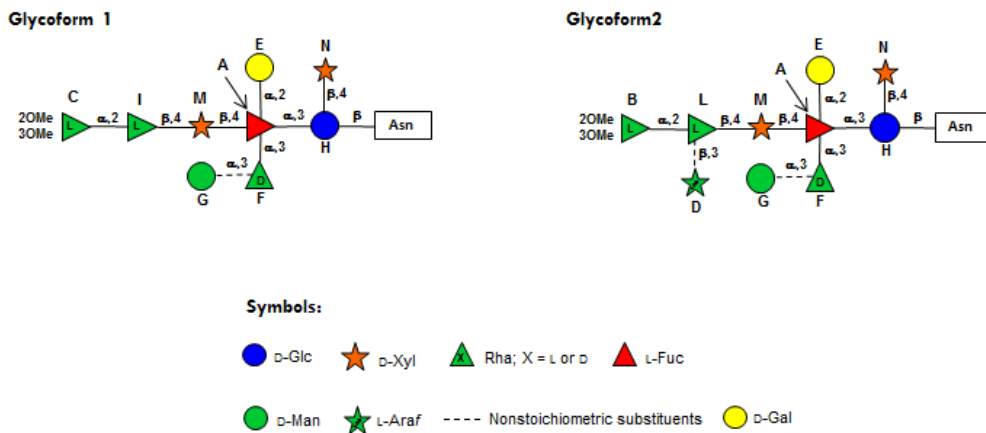**Figure 6.4** Structures of the two main glycoforms found in Vp54. Glycoform 1 is prevalent at [302]Asn, [399]Asn and [406]Asn, whereas glycoforms 2 is located at [280]Asn. Letters refer to those used during the NMR attribution.

The determination of the preferred dihedral angles (Φ and Ψ) was used to build these two glycoforms, thus the analysis of the free *N*-glycans was performed.

6.1.2 Molecular modelling approach

The two glycoforms (glyco1 and glyco2) were constructed using MacroModel 9.3.5 and, then, the Molecular Mechanic (MM) approach was used to identify a starting set of Φ/Ψ values for each glycosidic junction. However, MacroModel implemented two force fields, MM3 and Amber, both suitable for carbohydrates. Thus, to evaluate which force field was more appropriate, the dynamic behaviour of each oligosaccharide was examined running two different molecular dynamic simulations at 310 K for 20 ns, by using MM3 or Amber force field. The two ensembles of conformers were used to simulate the NOEs effects by calculating the averaged inter-proton distances. These values were compared with the experimental NMR distances, deduced from the T-ROESY spectrum (Table 6.2), and as result it was found that both force fields reproduced well the experimental data, even though the simulation performed with Amber FF had a better prediction score and matched a higher number of experimental distances. Therefore, Amber FF was used for further simulation and only the outputs obtained with this FF were further analysed.

Table 6.2 Interproton distances (Å) calculated experimentally from T-ROESY spectrum, recorded at 310 K on a 600 MHz spectrometer, or simulated for glyco1 and glyco2 oligosaccharides, for 20 ns and two different force fields. During NMR attribution, in glyco2 the β-Rha and 2,3-di-OMethyl-α-Rha residues were annotated with **B** and **L**, respectively, while in glyco1, **C** and **I** were used. These monosaccharides occupy the same position in the oligosaccharide architecture and differ for the presence of β-Ara*f* that is linked at O3 of **I** in glyco2 and absent in glyco1. Simulated distances in bold have a closer agreement with the experimental values, i.e. fall within 10% of the distance measured.

Proton pair	Experimental Dist. (T-ROESY)	Glyco1		Glyco2	
		Amber	MM3	Amber	MM3
A_1E_5	2,53	**2,71**	2,18	**2,57**	**2,46**
A_1H_3	2,65	**2,80**	3,21	**2,88**	3,06
B_1L_2	2,23	/	/	2,48	2,81
B_1B_{2OMe}	2,84	/	/	3,43	3,15
C_1I_2	2,12	2,70	2,88	/	/
C_1C_{2OMe}	2,51	**2,36**	3,43	/	/
D_1L_2	2,40	/	/	**2,56**	**2,43**
I_1M_4	2,32	**2,38**	**2,38**	/	/
L_1M_4	2,32	/	/	**2,45**	2,58
A_5N_2	2,62	**2,68**	3,65	2,94	3,04
M_1A_4	2,13	**2,32**	2,85	2,44	2,59
$N_1H_{6'}$	2,65	**2,42**	3,19	3,46	3,79
E_5H_2	2,83	**3,01**	3,24	**3,14**	3,27
C_2C_{2OMe}	2,83	**2,83**	**2,79**	/	/
B_2B_{2OMe}	2,83	/	/	**2,88**	**2,87**
H_6N_1	2,96	2,52	2,34	**2,82**	2,54
H_4N_1*	2,55	**2,32**	**2,54**	**2,47**	**2,56**
H_5N_1*	2,91	3,97	3,97	4,02	3,87
G_5F_2	3,55	3,14	2,22	2,57	2,47
$C_{3OMe}C_2$	3,24	2,52	3,80	/	/
$B_{3Ome}B_2$	3,24	/	/	**3,38**	**3,28**
# similar values		**10**	**3**	**9**	**5**

* partially overlapped NOEs

Thus, the space accessible to each residue was plotted in the scattered graph of Φ/Ψ values (Fig. 6.5). This analysis was performed for all residues, except for the glucose-Asn linkage, that in the free form had no constrained bonds, differently from what happens in the protein.

Comparison of the Φ/Ψ distribution with values obtained from crystallographic data showed that the residues near to the protein and well resolved in X-ray structure (**A**, **E** and **F**) were within the conformational space permitted to MD simulation. Whereas, the residues which appeared in distorted configuration in the X-ray structure (such as the two xyloses, **M** and **N**) along with those linked to them (**I**), or residues far from the protein, presented dihedral angle values either outside or at the edge of the allowed area (Fig. 6.5).
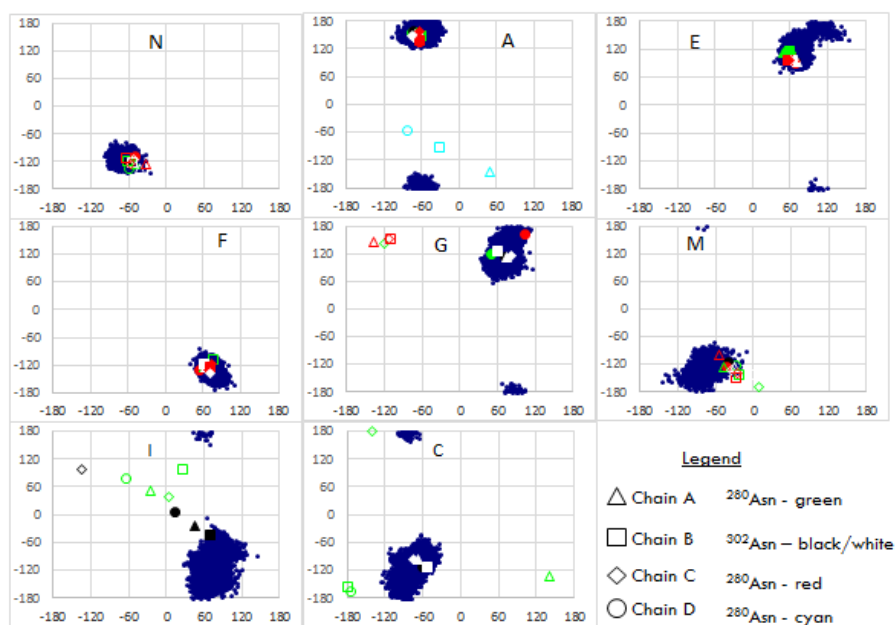

Figure 6.5 Superimposition of Φ/Ψ scattered maps (blue dots) from Amber simulation of free *N*-glycans with the dihedral values taken from the XR data. Symbols used are in the legend, open symbols mean that the residue has a non-canonical ring conformation.

Combining the X-ray data with those of MD, the two glycans were built using the dihedral X-ray values for **A, E, H** and **F** residues, calculated as the average over the four chains, and the MD values for **N, M, I, C, G** and **D** monosaccharides (values that had the highest frequency on the frequency count plot were chosen) (Table 6.3).

Table 6.3 Dihedral angles used to build the two glycans. Bold values are those taken from MD simulation.

	Φ_N Ψ_N	Φ_A Ψ_A	Φ_E Ψ_E	Φ_F Ψ_F	Φ_G Ψ_G	Φ_M Ψ_M	Φ_I Ψ_I	Φ_C Ψ_C	Φ_D Ψ_D
Glyco1	**-67 -112**	-68 149	54 114	66 -125	**63 118**	**-72 -118**	**68 -122**	**140 -134**	- -
Glyco2	**-67 -112**	-68 149	64 95	66 -125	**63 118**	**-72 -118**	**68 -122**	**-75 -105**	**-109 -161**

6.2 N-glycan linked to the MCP

The corrected glycan chains were used to build the fully glycosylated glycoprotein, connecting them at the appropriate Asn positions (glyco1 at [302]Asn, [399]Asn and [406]Asn, and the glyco2 at [298]Asn) of the polypeptide backbone, obtained from the XR structure. Of note, 1M3Y structure contained four polypeptide chains, but from the inspection of this structure seem that the N-glycans of each polypeptide chain did not interact with the other chains, thus only the chain A was considered, to simplify the system.

Before to proceed with the construction of the glycoprotein, the protein was optimized deleting the water molecules, the cations and all the glycans, and then corrected with Protein Wizard facility, as explained in the paragraph 8.4.4. Thus, the N-glycans were added to the protein using the dihedral angles reported in table 6.3. Regarding the glucose-asparagine junction of each chain, the dihedral angle values, obtained from X-ray structure, were used and they were considered as the average over the four chains, (Table 6.4).

Table 6.4 Dihedral angles of the glucose-asparagine junctions.

	Φ_H	Ψ_H	W_H	X_H
[280]**Asn**	-74	166	-168	64
[302]**Asn**	-81	-170	-166	-174
[399]**Asn**	-168	127	105	108
[406]**Asn**	64	-178	133	-61

This hybrid structure was used as starting point for further analysis: conformational search using the Metropolis Montecarlo approach and the Molecular Modelling.

6.2.1 (<u>M</u>etropolis <u>M</u>onte <u>C</u>arlo <u>M</u>inimization) - MCMM Conformational search of fully glycosylated Vp54

To determine the preferred orientations of the glycans linked at the protein, the Montecarlo Conformational search was performed, using Amber FF. The aim of this analysis was to find the preferred asset of those residues that in the X-ray structure were absent, distorted or linked to a distorted unit.

To this end, the analysis varied over all the possible values the dihedral angles Φ and Ψ of the glycosidic junctions connecting these monosaccharides, while the dihedrals of the residues well resolved in the X-ray were left free to move; the protein backbone, instead, was constrained, building two shells as detailed in the paragraph 8.4.5.

MCMM produced an ensemble of 11515 structures, with different potential energy, that were compared to the starting structure by using the superimposition criterion. Comparison among the starting structure and the MCMM ensemble was done superimposing the residues well resolved in the X-Ray: **H** (glucose), **A** (fucose) and **E** (galactose) were used for [280]Asn, [302]Asn and [399]Asn; only **H** for [406]Asn. Agreement between conformers was expressed by root mean square deviation (RMSD), and inspection of RMSD along with potential energy (plotted in the figure 6.6) disclosed the existence of several conformers with low energy (50 kJ/mol within the global minimum) and similar to the starting structure, because having low RMSD value (within 0.8 RMSD), simultaneously. In particular, two of these

conformers were of interest: the first was the best compromise between energy content and RMSD deviation (here indicated as minE_minRMSD), while the second corresponded to a structure with the minimum energy value (referred as minE).
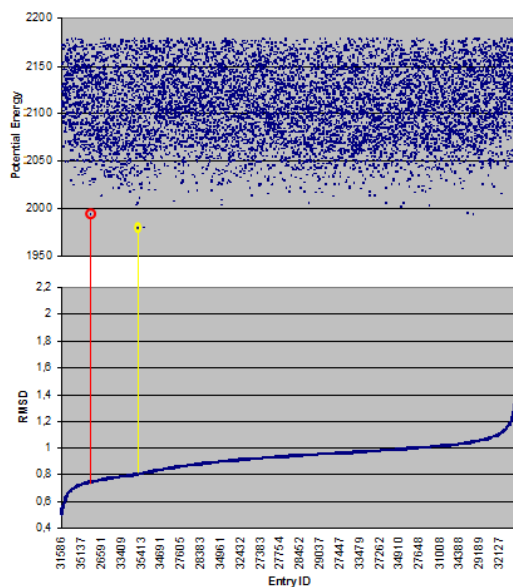

Figure 6.6 Potential energy and RMSD (root mean square deviation) plots of each structure (Entry ID). In yellow is highlighted the structure that has the lowest potential energy value ($E_{potential}$: 1997,590 kJ/mol; RMSD 2,048), whereas in red the structure with the best agreement of the two parameters ($E_{potential}$: 2012.392 kJ/mol; RMSD: 0,799).

Overlapping these structures with that used as reference, resulted that the external glycans residues occupied the same space region of the reference structure for all oligosaccharide chains, except the sequence **M-I-C-D** of ^{280}Asn glycan of minE structure (Fig. 6.7). Interestingly, the potential energy of these two conformers differed of only 15 kJ/mol, indicating that these structures could both occur in solution, interconverting one in the other. Thus, two different Molecular Dynamic simulations were performed using such conformers as starting structure, to understand the dynamic behavior of the molecule.
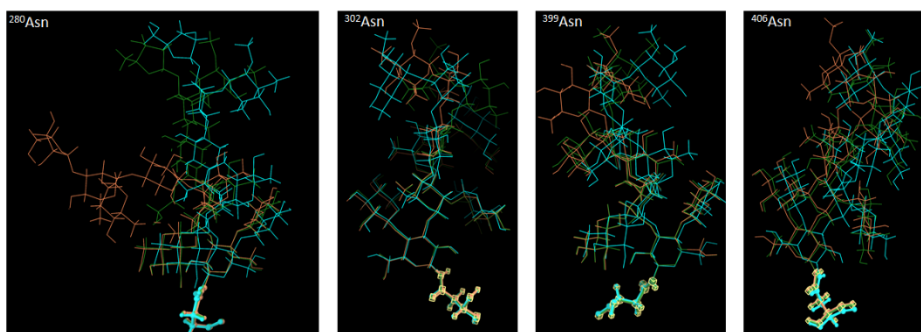
Figure 6.7 superimposition of Asn, glucose, fucose, galactose and proximal xylose for each oligosaccharidic chain. In cyan is the initial glycoprotein used as reference; in orange is the structure with low potential energy value (min E); the structure with a minimum energy and RMSD value is green (minE_minRMSD).

6.2.2 (<u>M</u>olecular <u>D</u>ynamic) – MD of fully glycosylated Vp54

The Molecular Dynamic simulation focused on the glycan part of the glycoprotein; therefore, the four oligosaccharides with the glycosylated Asn were left free to move while the protein backbone was fixed with constraints, as done during MCMM conformational search. The dihedral angles of each glycosidic junction of the whole obtained packages were extracted and compared with the corresponding distribution found for the free oligosaccharide (the scatter maps were reported from figure 6.8 to 6.12).

Inspection of the Φ/Ψ distribution of the two MD simulations were comparable and fell inside the limits defined by the MD simulation of the free oligosaccharidic chains (blu dots in figures), suggesting that the outputs converged, independently from the conformer selected for MD simulation. This result confirmed that the interconversion among different conformers was allowed.

For each glycosidic junction, the frequency of the Φ and Ψ values were plotted (Fig. from 6.8 to 6.12) and used to determine the preferred dihedral angles (values were summarized in table 6.5, along with the standard deviations σ_Φ and σ_Ψ).

Figure 6.8 Graphs obtained from minE structure as starting point. At the top: superimposition of Φ/Ψ maps of free glycans (dark blue dots) with those of same glycans linked to the ^{280}Asn (in green). At the bottom: the corresponding frequency count graphs of Φ and Ψ (in the legend) values.

Figure 6.9 Graphs obtained from minE structure as starting point. At the top: superimposition of Φ/Ψ maps of free glycans (dark blue dots) with those of same glycans linked to the [302]Asn (in gray). At the bottom: the corresponding frequency count graphs of Φ and Ψ (in the legend) values.

Figure 6.10 Graphs obtained from minE structure as starting point. At the top: superimposition of Φ/Ψ maps of free glycans (dark blue dots) with those of same glycans linked to the ^{399}Asn (in red). At the bottom: the corresponding frequency count graphs of Φ and Ψ (in the legend) values.
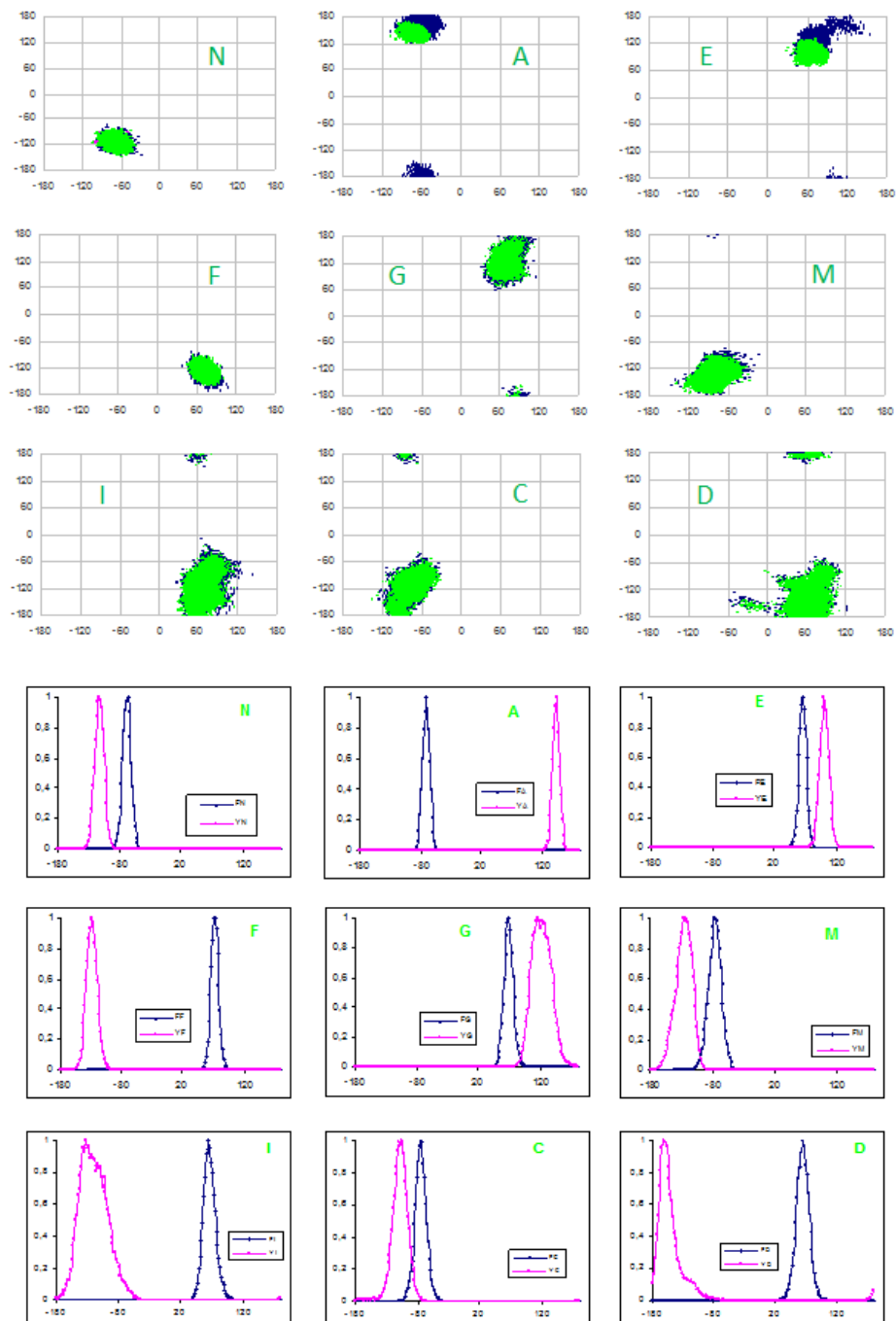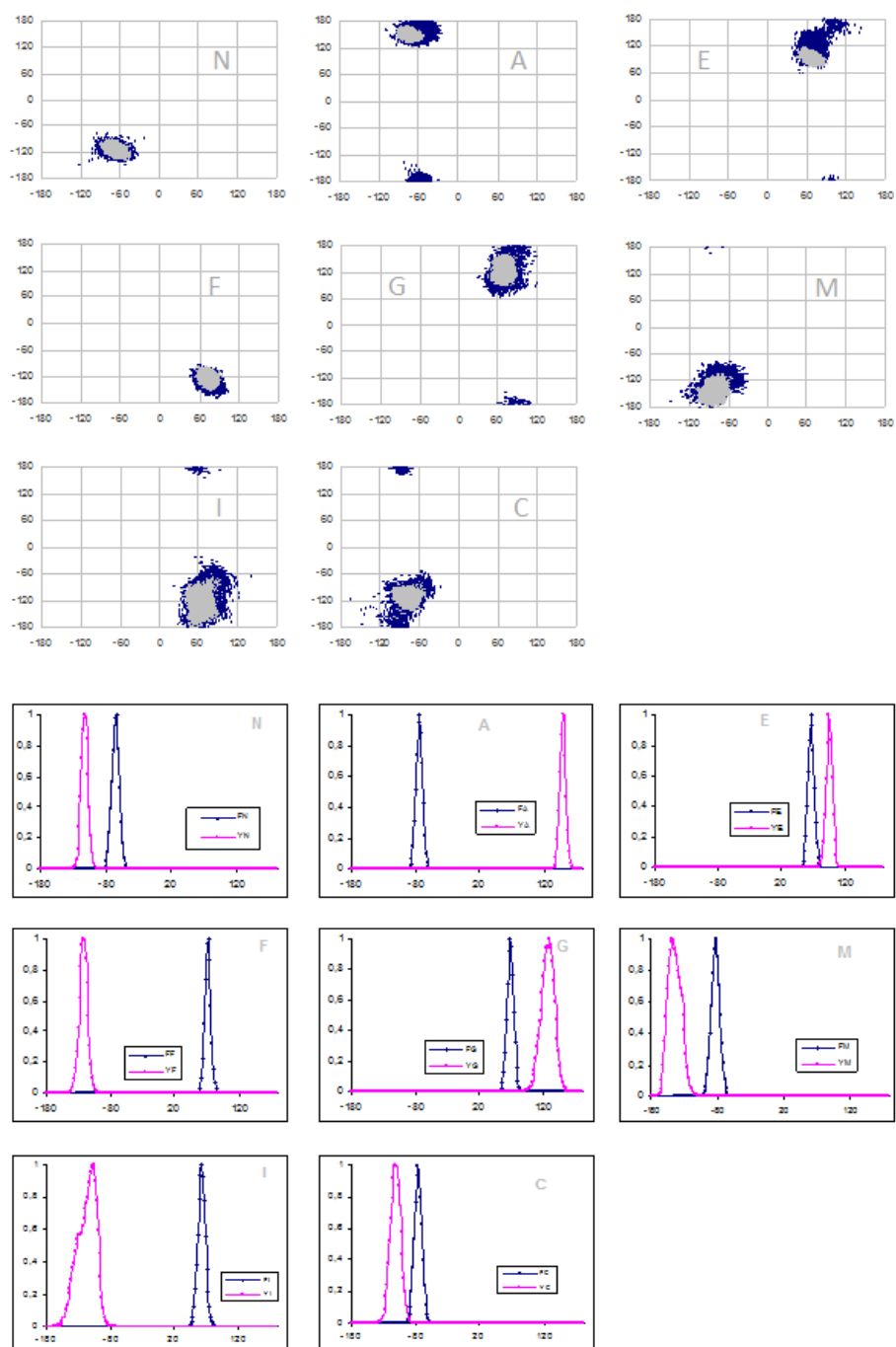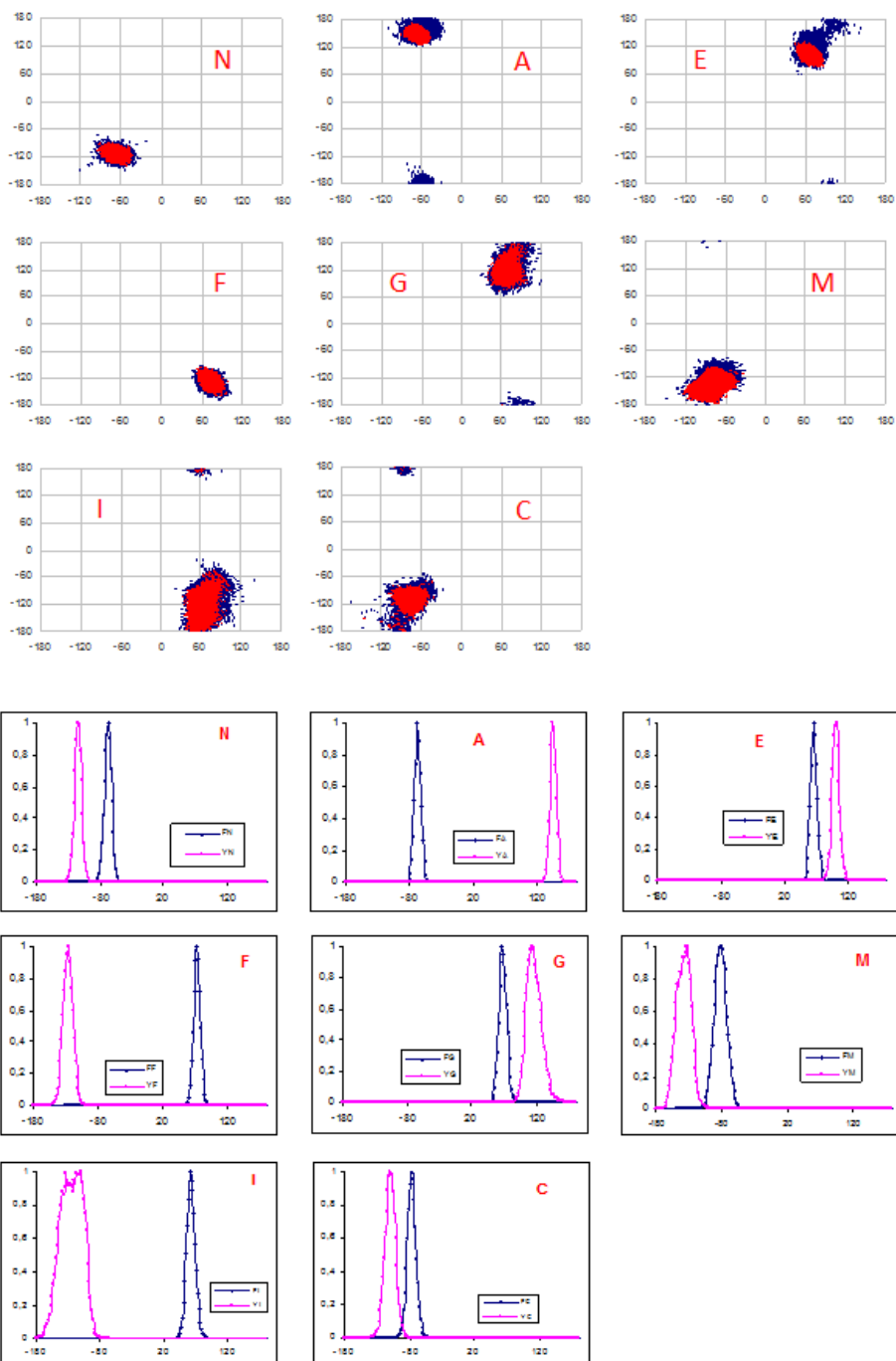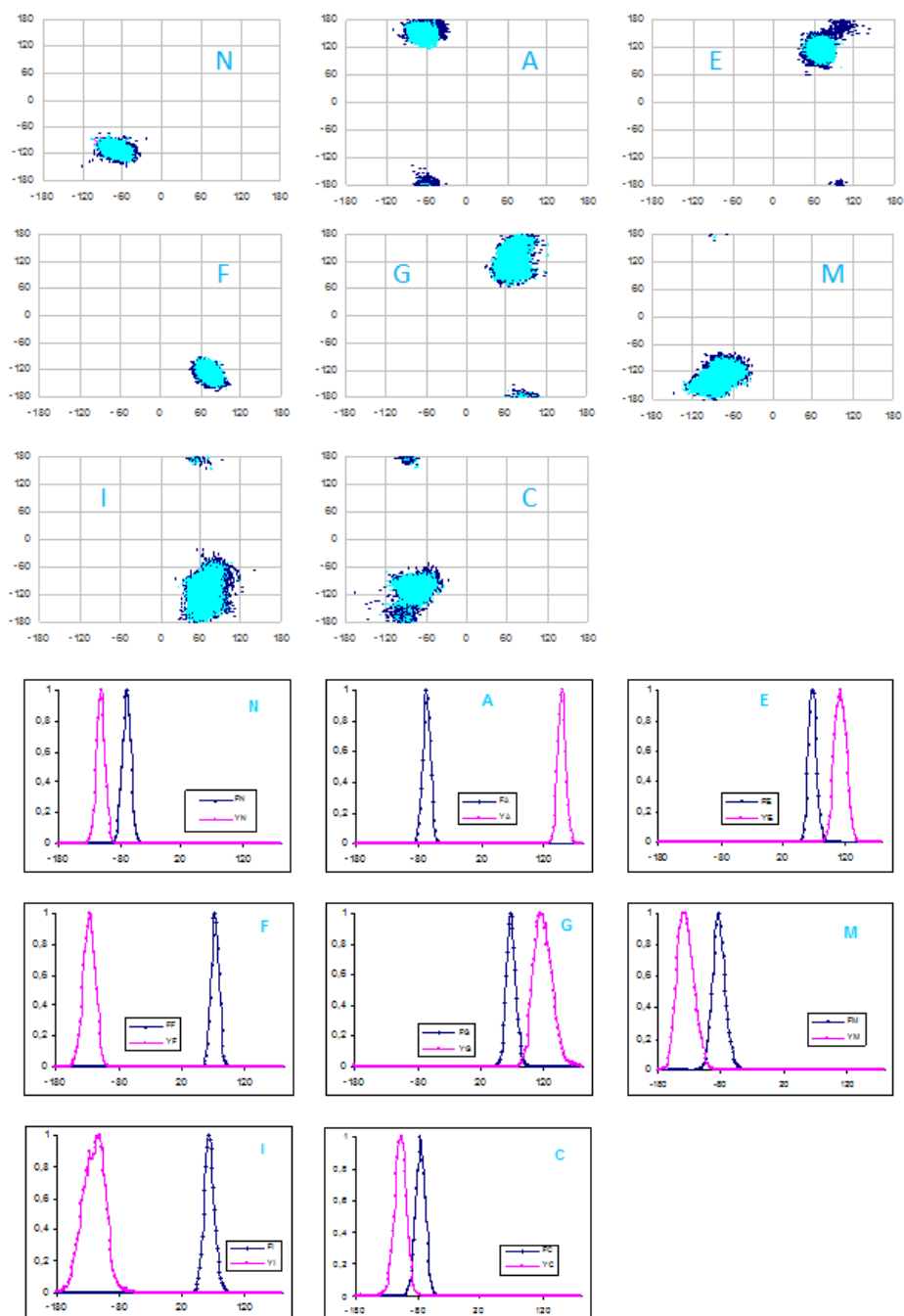
Figure 6.11 Graphs obtained from minE structure as starting point. At the top: superimposition of Φ/Ψ maps of free glycans (dark blue dots) with those of same glycans linked to the [406]Asn (in cyan). At the bottom: the corresponding frequency count graphs of Φ and Ψ (in the legend) values.
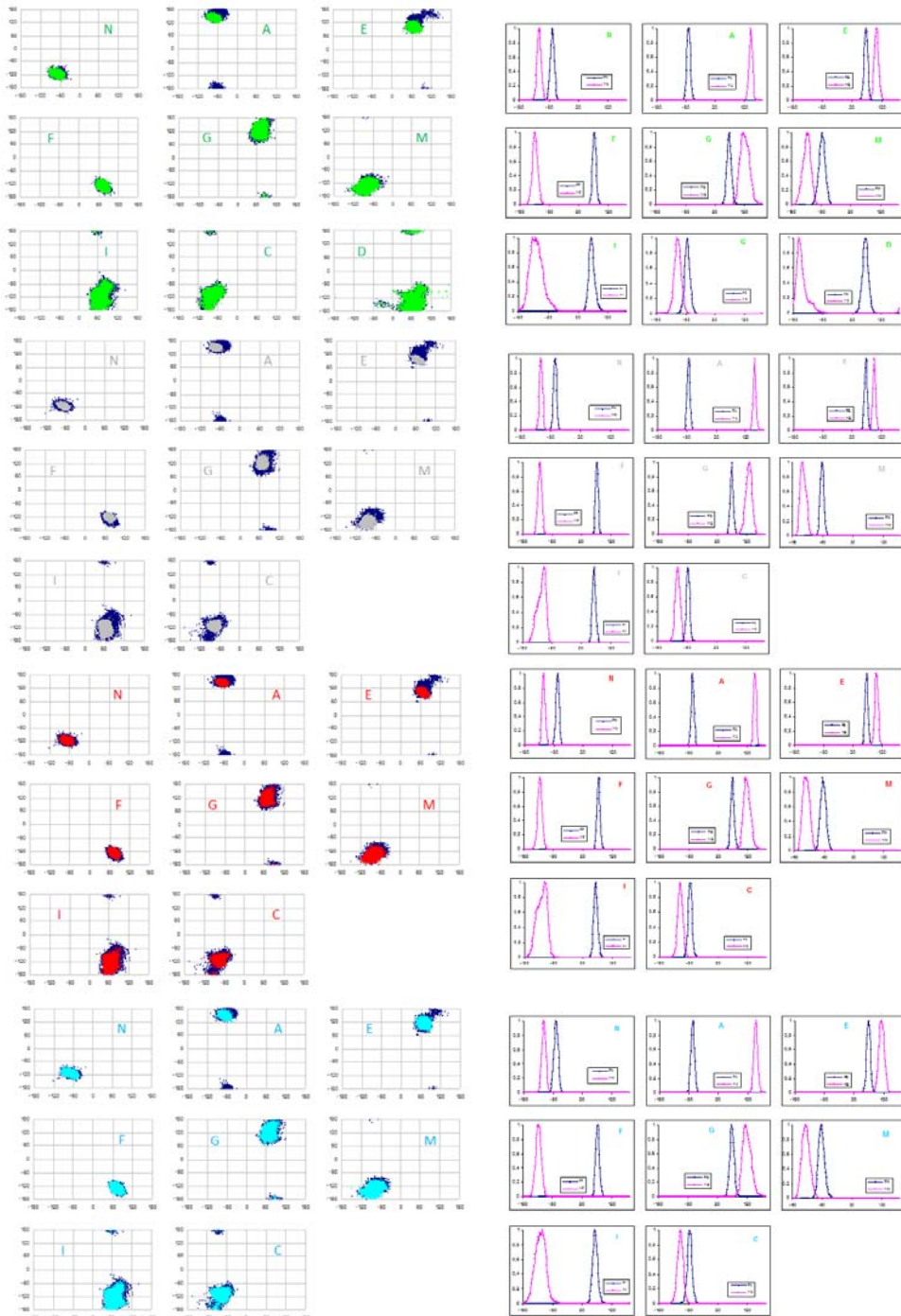
Figure 6.12 Graphs obtained from structure with minimum E and minimum RMSD values, as starting point. On the left: superimposition of Φ/Ψ maps of free glycans (dark blue dots) with those of same glycans linked to the asparagine residues (^{280}Asn in green; ^{302}Asn in gray; ^{399}Asn in red; ^{280}Asn in cyan). On the right: frequency count graphs of Φ and Ψ (in the legend) values, for each of them.

167

Table 6.5 Φ and Ψ values obtained from frequency count graphics along with their standard deviation (σΦ and σΨ, respectively).

Asn 280	Min E				Min E - Min RMSD			
	Φ	σΦ	Ψ	σΨ	Φ	σΦ	Ψ	σΨ
N (proximal Xyl)	-67	6,93	-113	7,44	-69	6,73	-113	7,01
A (Fuc)	-71	6,00	143	5,70	-69	6,06	143	5,62
E (Gal)	67	6,17	101	7,25	67	6,12	103	7,19
F (3-Rha)	73	6,22	-127	8,70	73	6,11	-127	8,83
G (Man)	69	7,76	117	16,27	69	7,60	117	16,31
M (distal Xyl)	-77	10,59	-125	13,04	-77	10,89	-127	13,06
I (2-Rha)	65	9,99	-133	23,80	63	10,14	-133	23,40
C (Rha)	-75	9,40	-105	17,75	-75	9,28	-109	16,39
D (Ara)	63	11,21	161	36,18	65	11,23	-161	35,65

Asn 302	Min E				Min E - Min RMSD			
	Φ	σΦ	Ψ	σΨ	Φ	σΦ	Ψ	σΨ
N	-65	5,52	-113	5,27	-65	5,44	-113	5,19
A	-73	4,94	151	4,59	-73	4,96	151	4,55
E	67	4,93	95	4,80	67	4,92	95	4,77
F	73	4,38	-121	6,35	73	4,35	-119	6,38
G	69	5,12	129	10,20	69	5,09	131	10,09
M	-83	6,37	-147	9,70	-83	6,30	-147	9,40
I	63	6,00	-105	14,46	63	5,86	-107	14,38
C	-77	5,87	-111	7,90	-77	5,74	-113	7,87

Asn 399	Min E				Min E - Min RMSD			
	Φ	σΦ	Ψ	σΨ	Φ	σΦ	Ψ	σΨ
N	-67	5,68	-115	5,58	-67	5,67	-115	5,61
A	-67	4,97	145	4,83	-67	4,98	145	4,82
E	67	5,07	103	5,73	67	5,13	99	5,71
F	73	5,03	-125	7,48	73	5,01	-125	7,38
G	67	6,22	113	11,67	67	6,01	113	11,89
M	-81	9,62	-131	10,74	-85	9,45	-139	11,26
I	61	6,92	-111	18,01	63	6,96	-111	17,20
C	-77	6,65	-109	9,13	-77	6,59	-111	8,59

Asn 406	Min E				Min E - Min RMSD			
	Φ	σΦ	Ψ	σΨ	Φ	σΦ	Ψ	σΨ
N	-69	7,04	-111	6,80	-71	7,24	-111	6,87
A	-67	6,09	149	8,08	-67	6,04	149	6,31
E	69	6,43	113	9,29	69	6,42	113	9,13
F	73	6,17	-127	9,34	73	6,16	-127	9,23
G	67	7,48	115	16,76	67	7,49	115	16,38
M	-83	10,44	-141	13,15	-85	10,21	-137	12,85
I	63	8,41	-111	18,27	63	8,64	-113	19,41
C	-77	8,07	-107	10,89	-77	8,17	-107	11,17

The standard deviation product ($\sigma_\Phi \cdot \sigma_\Psi$) (plotted in figure 6.13) was used to visualize the total flexibility of each connection.

Comparing the $\sigma_\Phi \cdot \sigma_\Psi$ value resulted that it increases moving from [302]Asn, to [399]Asn, [406]Asn up to [280]Asn, and a high value corresponded to the monosaccharide that appeared distorted or absent in the X-ray structure, as occurred for [406]Asn.

Thus, as general rule, high deviation values associate to large flexibility (Figs. 6.13a,b), on the contrary, low values are diagnostic of poor flexibility and are associated to residues that have an interpretable density in X-ray density map.

In addition, looking at each oligosaccharidic chains, resulted that the $\sigma_\Phi \cdot \sigma_\Psi$ value increased when moving towards the terminal part of the chain (residues **M-I-C)** (Figs. 6.13c,d,e,f), confirming that this region is more flexible and supporting the evidence that it is less resolved, if not absent at all, in the crystal structure.

Glycan at [280]Asn was an exception, it differed from this trend because the moiety built from **M-I-C-D** residues assumed two different orientations that interconverted during MD simulation, a behavior that generated large $\sigma_\Phi \cdot \sigma_\Psi$ values. This result suggests that MD simulations imitate the behavior of the molecule in solution and as such have some limitations to reproduce what happens during crystal packing where other forces come into play, selecting in this case a specific conformation of **M**-I-**C**-**D** residues.

Figure 6.13 Graphs of standard deviations of Φ (σ_Φ) and Ψ (σ_Ψ) found running MD simulations using two different glycoproteins: minE structure (A) and minE_minRMSD structure (B). C, D, E, F) standard deviation of each glycans obtained from the two glycoproteins. Note that in these last four graphics the Y scales are not the same.

Finally, the glycan dihedrals were adjusted to the averaged values obtained from the MD simulation performed using the conformer with low RMSD and low energy, as starting structure (Table 6.5), and a structure representative of Vp54 was built and used to understand the type of interaction established between each oligosaccharide and its surroundings (Fig. 6.14).
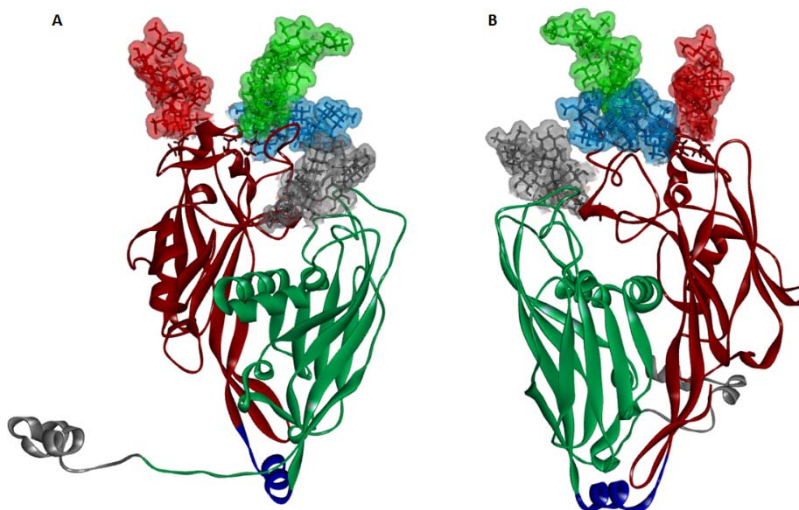
Figure 6.14 A. Final representation of Vp54 glycoprotein. B. Right rotation of 90° of A visualization. Polypeptide backbone: in gray from 2 to 26 aa; D1 in dark green: $27 - 212$; Hinge in dark blue: $213 - 224$; D2 in dark red: $225 - 437$. Oligosaccharidic chains: [280]Asn indicates in green; [302]Asn in gray; [399]Asn in red; [406]Asn in cyan.

Analysis of the Vp54 structure confirmed that all N-glycans are on the same side of the protein and point outwards. Of note, glycans at [280]Asn and [406]Asn are very close, as indicated in figure 6.15: the surface of the proximal xylose of the chain at [406]Asn makes contacts with that of the galactose unit of the glycan at [280]Asn, that in turn has a strong hydrogen bond with the Thr-391 (OH3 of Gal with γ-OH of Thr-391) (Fig. 6.15). Glycan at [302]Asn is inserted between the two domains of Vp54, D1 and D2, and establishes several interactions as visualized in figure 6.16A. Instead, glycan at [399]Asn results the chain most exposed to the environment (Fig. 6.16B).
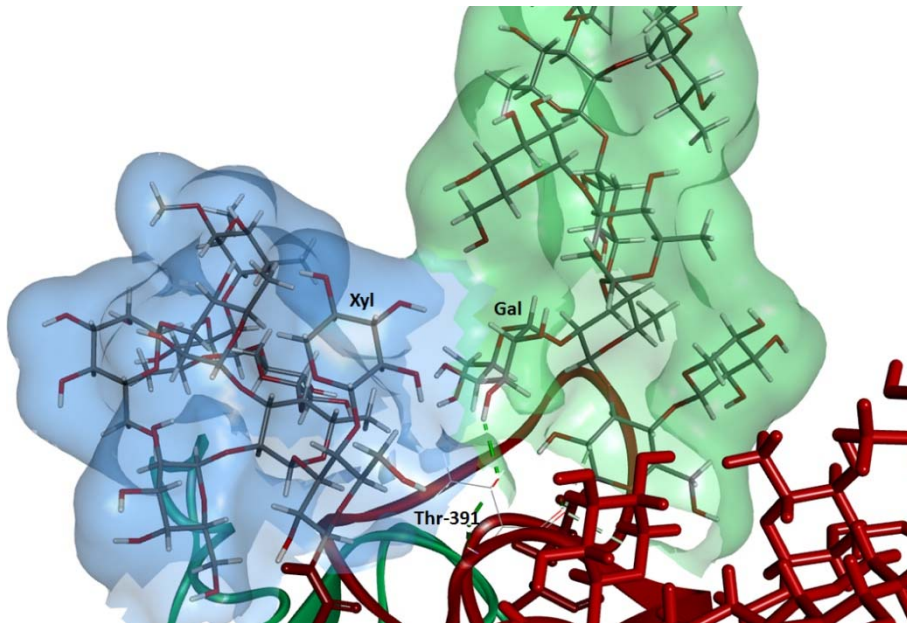
171

Figure 6.15 Representation of the contact region between glycan at [406]Asn (blue surface) and that at [280]Asn (green surface).
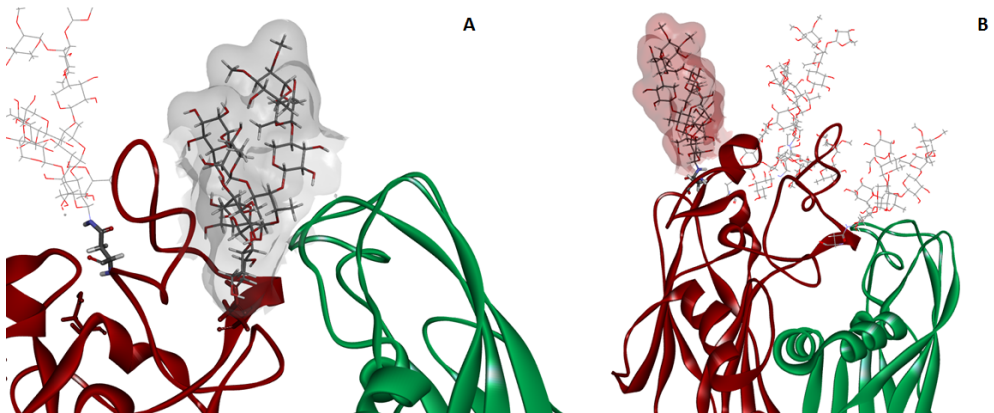

Figure 6.16 A. Representation of the position of the glycan at [302]Asn (gray surface) encloses in the domains 1 and 2 of the protein backbone. B. Position of the glycan at [399]Asn (red surface).

6.3 Conclusions

The aim of this part of PhD project is the construction of the 3-dimensional structure of major capsid protein Vp54 of the chlorovirus PBCV-1, implementing the faults of the X-ray data with a computational approach, which reconstructed the missing carbohydrate moiety and analyzed their conformational properties (De Castro et al. 2013).

The original Vp54 X-ray structure was revisited, focusing the attention on the dihedral angles of the glycosydic junction of the monosaccharides that were present in such structure and on their ring conformations. This investigation disclosed that X-ray analysis well represents the monosaccharides close to the polypeptide, on the contrary those far were in the wrong ring conformation or absent. Thus, our protocol made use of two well established approaches MCMM and MD. The MCMM conformational search produces two conformers at low energy, which were subsequently used to run two molecular dynamic. These simulations converged in the same results proving that the two conformers can interconvert one in the other, suggesting that they are not in a local minimum.

Therefore, the glycan dihedrals are adjusted and a new model of Vp54, more accurate than the one originally, published (Fig. 6.17b).
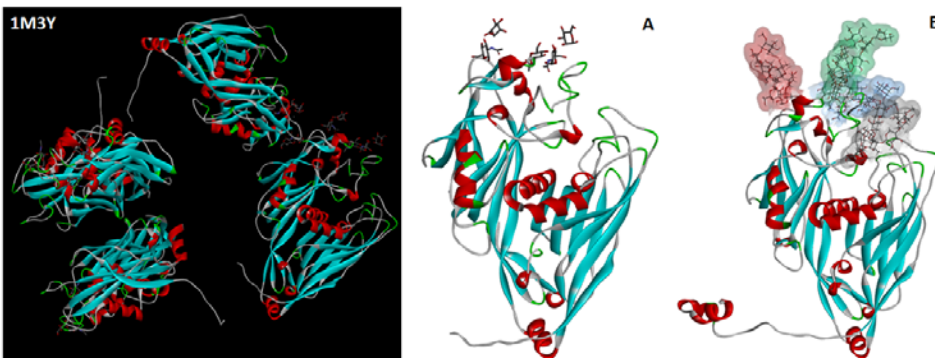

Figure 6.17. On the left: Crystallographic structure of Vp54, at 2.0 Å resolution (PDB ID code 1M3Y), containing four independent polypeptide chains in the crystallographic asymmetric unit. On the right: A. Chain A of the X-ray of the 1M3Y diffraction data. B. Structure of the revised Vp54 monomer. Oligosaccharidic chains: [280]Asn green; [302]Asn gray; [399]Asn red; [406]Asn cyan.

Knowing the correct Vp54 structure is interesting to understand which role (if any) the glycans play in the protein folding or/and in the packaging of the whole capsid. Preliminary studies will be discussed in the following chapter.

Chapter 7

Glycans function

The determination of the complete structure of Vp54 glycoprotein did not provide any information about the role that *N*-glycans play in the capsid packaging, opening many questions: are the glycans involved in the protein folding? Are all four *N*-glycans important for capsid integrity? Within the same *N*-glycan, are all the monosaccharides equally involved or only some of them are necessary for the correct folding of the capsid protein?

Thus, preliminary studies about the identification of the role of the glycans in the protein folding were performed at the laboratory of the Professor Barbero, in Bilbao.

In this contest, the Vp54 protein with and without glycans, was examined adopting a Molecular Dynamic approach, to verify the behavior of each structure, with the aim to understand if the glycans increased the stability of the protein and/or if they were involved in the protein folding. For these analyses, the Amber12 program was utilized.

7.1 Molecular dynamic

The revisited Vp54 structure, obtained from previous studies (see previous chapter), was the starting point of these experiments: the protein without glycans was obtained removing all glycans from this initial structure, instead the protein with glycans was built combining GLYCAM-Web and Discovery programs. It has been necessary to re-build the full glycosylated Vp54 structure, because the glycan structures obtained with MacroModel were not accepted as input for Amber program. The dihedral angles of the glycosydic junctions were maintained equal to those determined before.

Therefore, to determine the dynamic behavior of these two structures (Vp54 with and without glycans), different Molecular Dynamic simulations were performed in explicit water using the Amber12 package.

Two parallel analyses were started by using ff12SB FF for the protein part and GLYCAM_06j-1 FF for the glycosidic moieties. The initial structures were neutralized with two Cl⁻ ions and the whole molecules placed in a box of water, with a 12 Å cut-off. The whole systems were minimized and heated from 0 K to 300 K, using a weak restrain on the solutes. Thereafter, the systems were equilibrated at 300 K, while pressure was kept constant (1 atm) and the restraints applied to the solute were removed. Then, MD simulations of 50 ns started. Note that during the simulations no constraints were set, keeping the entire glycoprotein free to move.

The whole ensembles of conformers were analyzed, and folding (or unfolding) of the proteins checked by analysing the RMSD variation of the protein backbone with respect to the initial structure. The curves obtained were very similar (Fig. 7.1) suggesting that the glycans did not influence the protein conformation, on the contrary to what we expected.
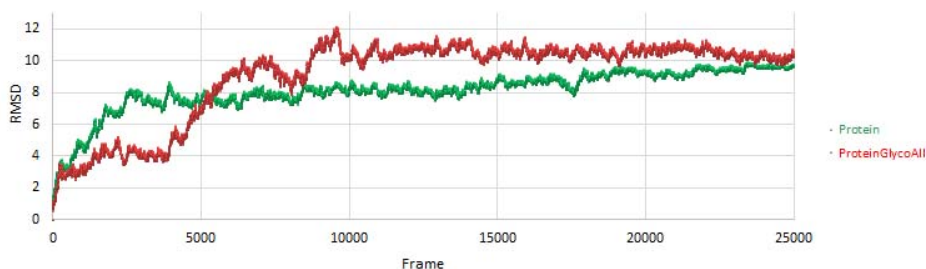

Figure 7.1 Root mean square deviation results of MD at 300 K. In red is reported the RMSD of protein equipped with glycans and in green only the protein.

To rule out the possibility that the two structures were blocked into a local minimum, a different approach was adopted. The original idea was to use the Replica Exchange Molecular Dynamic (REMD), but this protocol has a very high computational cost, thus we tried to simulate it running several independent dynamics at different temperature, and then placing one the final structure from one dynamic at another temperature, in order to explore a wide region of potential energy escaping the trapping into a local minimum. Thus, dynamics at different temperature (500 K, 600 K, 700 K, 800 K), were performed. When the simulations were conducted at high temperatures, the constraints were imposed to prevent the occurrence of the break and the flaws of the protein.

The dynamic at 500 K has given interesting results: protein with glycans presented a major value of RMSD with respect to the protein without them (Fig. 7.2), indicating that the protein fully glycosylated had a major flexibility with respect to the "naked" protein, in contrast with our initial idea that N-glycans were important for protein stabilization.
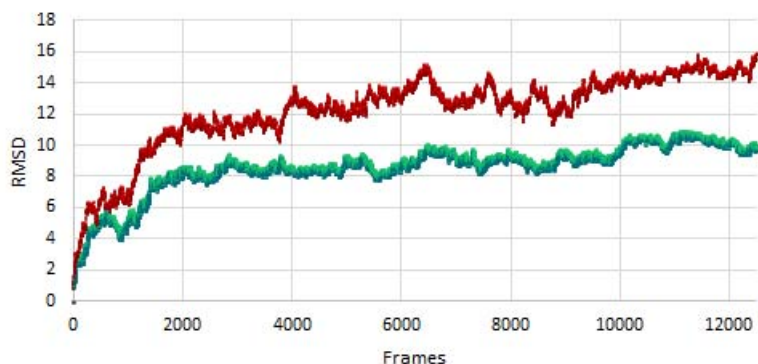
Figure 7.2 Root mean square deviation results of MD at 500 K. In red is reported the RMSD of protein equipped with glycans and in green the protein without glycans.

An accurate analysis of the ensembles of the structures revealed that the two domains of the protein, D1 and D2, changed their relative orientation. Therefore, the two domains were analyzed separately and it was found that they behave differently, with domain 2 more stable than 1 (Fig. 7.3).

Figure 7.3 RMSD curves of each domain (D1: 212 – 223 in green; D2: 224 – 436 in red) of fully glycosylated protein at 500 K (A) and at 300 K (B).

177

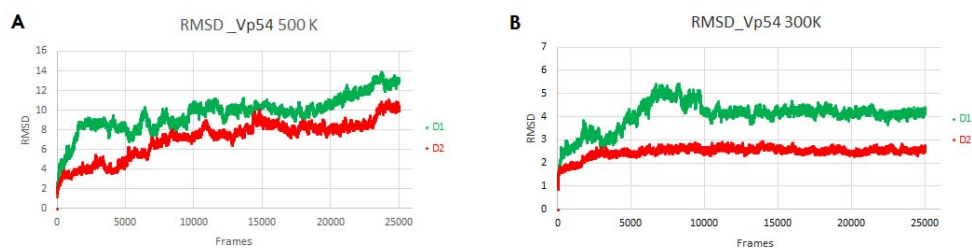Inspection of the 3D model pointed to the fact that the flexibility of the protein can be described from the relative orientation of the two domains, making necessary the analysis of the normal mode of each structure.

7.2 Normal Mode analysis

Vp54 can be schematized as two domains kept together by a flexible hinge. Generally, a protein with this structure presents three principal normal modes (Fig. 7.4): the closing, the twisting and the wobbling mode (Ma 2005).
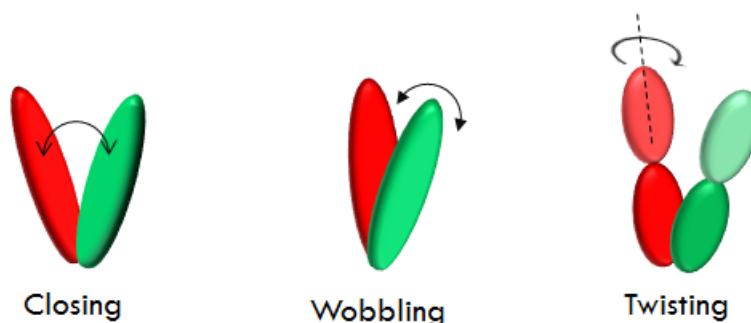
Figure 7.4 Schematic representation of the three normal modes.

Therefore, such modes were analyzed for structures obtained from both molecular dynamics (at 300 K and 500 K), using specific scripts.
For the molecular dynamic at 300 K, both closure and wobbling modes presented similar trend (Fig. 7.5), and were the same for the protein with and without the glycans, suggesting that at this temperature there was not a great difference between the two structures and confirming our previous idea that the two molecules were blocked into a local minimum.

Figure 7.5 Graph of the closure (at the top) and wobbling (at the bottom) values for protein with and without glycans (specified in the legend), at 300 K.

Regarding the dynamic at 500 K, the protein without glycans presented a closure curve that increased during the dynamic, more when compared to that of the protein glycosylated (Fig. 7.6), indicating that the two domains moved easily away when the glycans are not present (Fig. 7.7).

For the wobbling mode, the range of the distance between the two domains was considered. In this case, it was more extended with respect to that found for that at 300 K (Fig. 7.6), proving the major movement of the protein at high temperature. Once again, the protein without glycans explored a range larger (from 8 to 46 Å) compared to that of the glycosylated protein (from 11 to 35 Å).
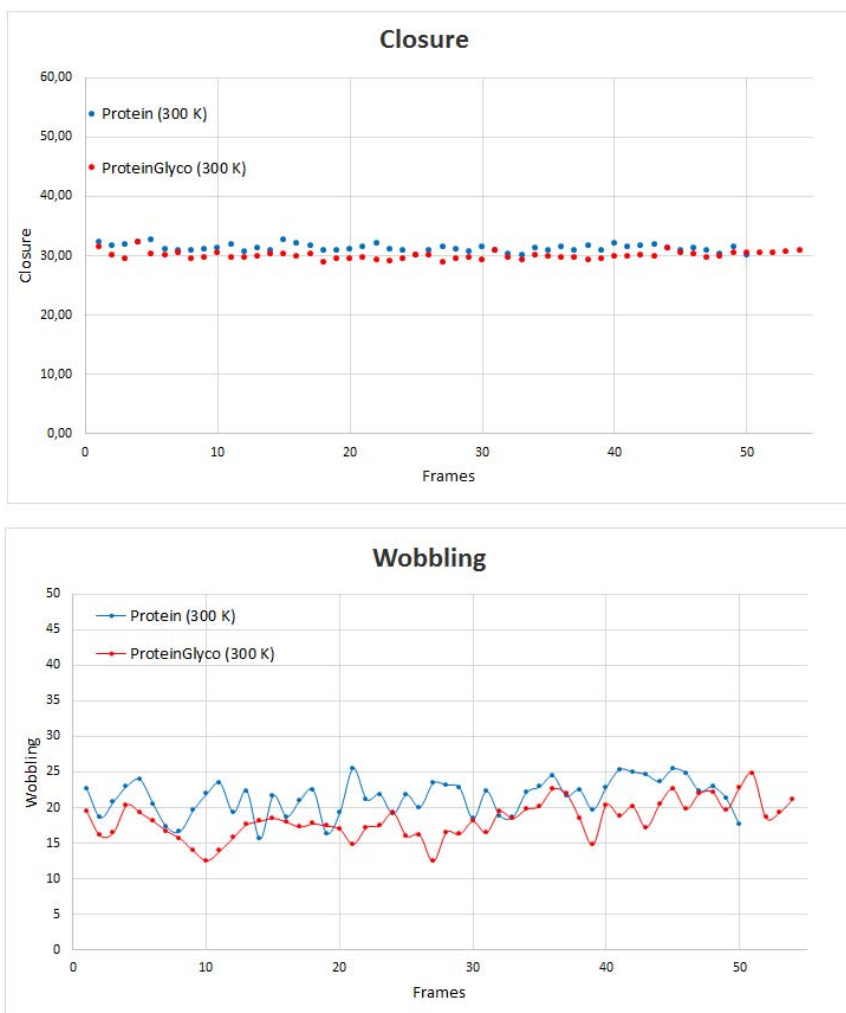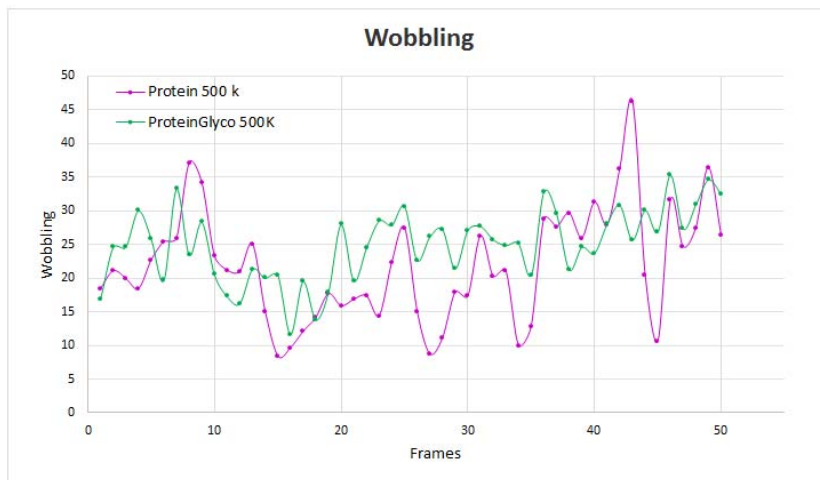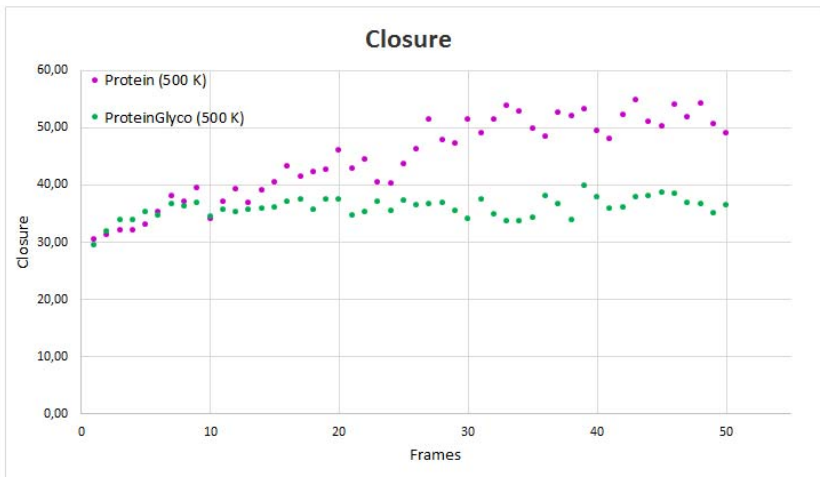
Figure 7.6 Graph of the closure (at the top) and wobbling (at the bottom) values for protein with and without glycans (specified in the legend), at 500 K.
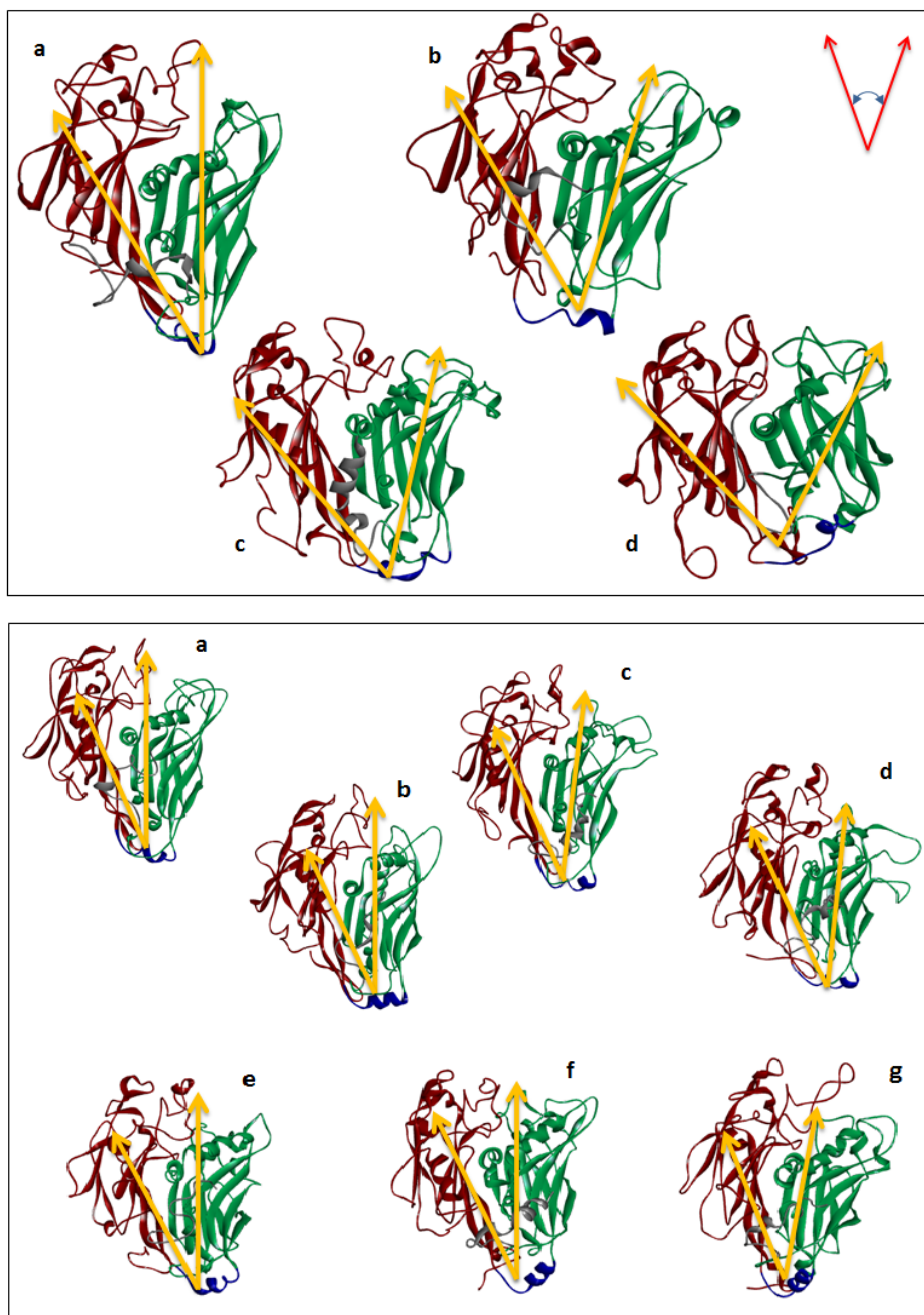
Figure 7.7 Structures describing the dynamic behavior of the protein with (bottom) and without (top) the glycans. The dynamic simulation was performed at 500 K, for 50 ns. The yellow arrows indicate the closure mode. Polypeptide backbone: in gray from 2 to 26 aa; D1 in dark green: 27 − 212; Hinge in dark blue: 213 − 224; D2 in dark red: 225 − 437.

The twisting mode was calculated as distance between two atoms belonging to amino acids of different domains: they were chosen analysing the structures obtained from dynamic at 500 K of the glycosylated protein. Three representative structures were chosen at the initial, middle and at the end of the dynamic (Fig. 7.8). The glycans that initially pointed outward of the protein, during the dynamic collapse on it, so that the glycan at [302]Asn, that was located among the two domains at the beginning of the simulation, was replaced with the glycan at [280]Asn.
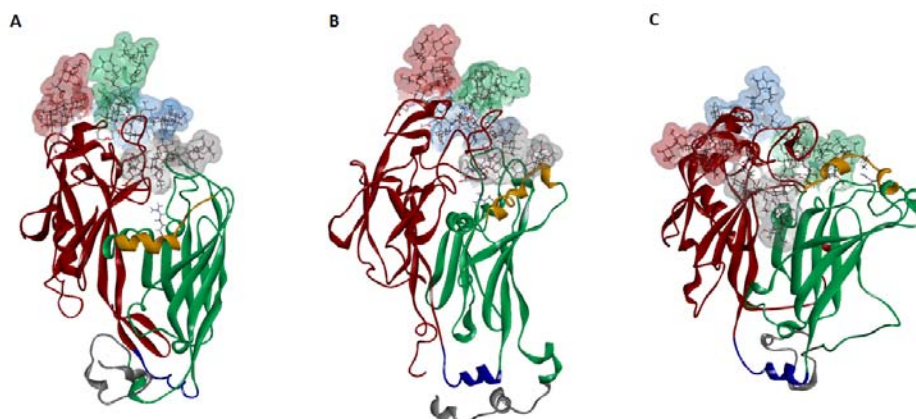
Figure 7.8 Fully glycosylated protein structures, obtained from the dynamic simulation at 500 K, taken at three times: at the beginning (A), middle (B) and end (C). Polypeptide backbone: in gray from 2 to 26 aa; D1 in dark green: 27 – 212; Hinge in dark blue: 213 – 224; D2 in dark red: 225 – 437. Oligosaccharidic chains: [280]Asn indicates in green; [302]Asn in gray; [399]Asn in red; [406]Asn in cyan.

Thus, two distances were examined during the simulations: **N**H and **C**O of Asn-280 and Arg-129; **N**H and **C**O of Asn-302 and Arg-129. Regarding the dynamic at 300 K, these distances remained constant for all the time of the simulation (Fig. 7.9).
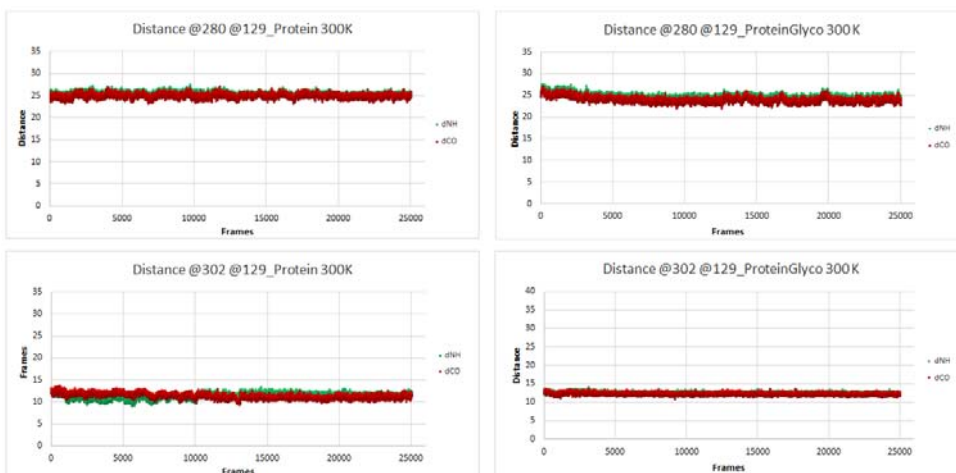
Figure 7.9 Graphs reporting the distances between Asn280 and Arg129 (at the top) and between Asn302 and Arg129 (on the bottom) during the entire dynamic simulations at 300 K, for both the polypeptide chains (with and without the glycans, indicated on the graphs).

At 500 K, the situation changed (Fig. 7.10). In the protein, the distances remained constant; on the contrary, the glycosylated protein showed a decrease of the distance between the Asn-280 and the Arg-129 and a simultaneous increase of the distance between the Asn-302 and the same arginine. These variations validate the twisting mode of the protein.

Figure 7.10 Graphs of distances between Asn280 and Arg129 (at the top) and between Asn302 and Arg129 (on the bottom) during the dynamic simulations at 350 K, for both the polypeptide chains (with and without the glycans, indicated on the graphs).

183

7.3 Conclusions

The aim of this part of the project is the determination of the role of the glycans in the protein folding, using the molecular dynamic approach.
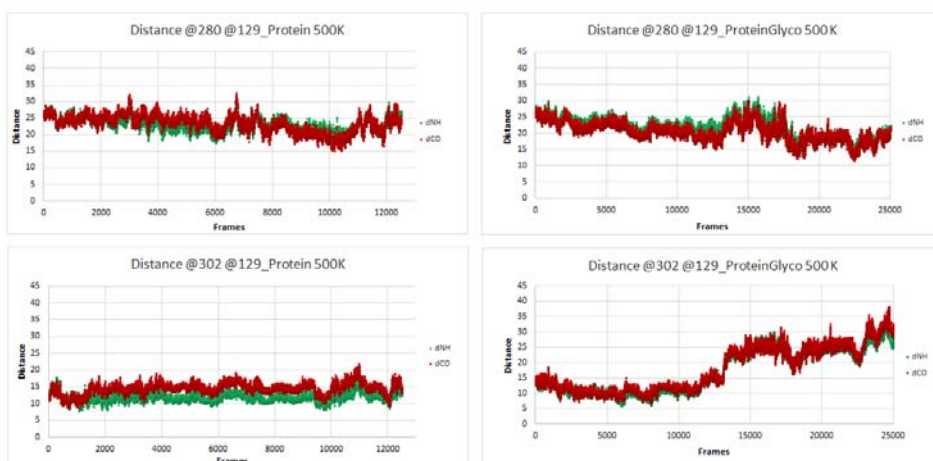
The dynamic was performed at different temperature in order to avoid the trapping in local minima and also to adapt the Replica Exchange method to the reduced computational facilities available. By this approach, we started to explore the behaviour of the protein with and without glycans, looking at the normal mode or at the way the different subunits arranged reciprocally. Preliminary results suggest that at 300 K the protein, independently form the glycosylation status, was blocked into a local minimum of the potential energy. On the contrary, at 500 k, the protein without glycans has a major value of the closure and wobbling with respect to the glycosylated protein. Accordingly, when the glycans are present, the two domains tend to remain in the same close position, and the only movement allowed is the twisting.

These results lead us to think that the glycans are not involved in the folding of the protein, but maybe are necessary to the integrity of the capsid: they probably are useful to maintain two domains at the correct distance which promotes the capsomer formation and stability.

These studies are still in progress, thus the dynamic behaviour of the glycopeptides at other temperatures will be studied. It is also interesting to prove the idea that the glycans play an important role in the capsid packaging analyzing the entire trimer of Vp54. If this idea is correct, we expect that when the glycans are not present, the trimer is not formed or it can form, but it breaks easily upon a weak solicitation. On the contrary, when the protein is glycosylated, the trimer is stable and the capsid is properly packed.

References

Agarkova IV, Dunigan DD, VanEtten JL *J Virol* **2006** Virion-associated restriction Endonucleases of Chloroviruses 80(16):8114–8123

Bock K, Pedersen S **1983** Carbon-13 Nuclear Magnetic Resonance Spectroscopy of Monosaccharides *Adv. Carbohydr. Chem Biochem* 41:27-65

De Castro C, Molinaro A, Piacente F, Gurnon JR, Sturiale L, Palmigiano A, Lanzetta R, Parrilli M, Garozzo D, Tonetti M, Van Etten JL **2013** Structure of N-linked oligosaccharides attached to chlorovirus PBCV-1 major capsid protein reveals unusual class of complex N-glycans. *Proc Natl Acad Sci USA* 110: 13956-13960. doi: 10.1073/pnas.1313005110

De Castro C, Speciale I, Duncan G, Dunigan DD, Agarkova I, Lanzetta R, Sturiale L, Palmigiano A, Garozzo D, Molinaro A, Tonetti M, Van Etten JL **2016** N-linked Glycans of Chloroviruses Sharing a Core Architecture without Precedent. *Angew. Chem. Int. Ed. 55*, 654–658; doi: 10.1002/anie.201509150

Gorshova RP, Isakov VV, Kalmykova EN, Ovodov YS **1995** *Carbohydr. Res. 268*, 249-255

Graves MV, Bernadt CT, Cerny R, Van Etten JL **2001** Molecular and genetic evidence for a virus-encoded glycosyltransferase involved in protein glycosylation. *Virology* 285(2):332–345. Doi:10.1006/viro.2001.0937

Landstein D, Burbank DE, Nietfeldt JW, Van Etten **1995** Large Deletions in Antigenic Variants of the Chlorella Virus PBCV-11 JL *Virology* 214:413-420.

Ma J **2005** Usefulness and Limitations of Normal Mode Analysis in Modelling Dynamics of Biomolecular Complexes *Structure* Vol.13, 373-380. DOI:10.1016/j.str.2005.02.002.

Mathieu-Rivet E, Kiefer-Meyer MC, Vanier G, Ovide C, Burel C, Lerouge P, Bardor M **2014** Protein N-glycosylation in eukaryotic microalgae and its impact on the production of nuclear expressed biopharmaceuticals *Front. Plant Sci.* 5:1–13

Nandhagopal N, Simpson AA, Gurnon JR, Yan X, Baker TS, Graves MV, Van Etten JL, Rossmann MG **2002** The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus *PNAS* 99,23, 14758-14763 DOI: 10.1073pnas.232580699.

Quispe CF, Esmael A, Sonderman O, McQuinn M, Agarkova I, Battah M, Duncan GA, Dunigan DD, Smith TPL, De Castro C, Speciale I, Ma F, Van Etten JL **2016** Characterization of a new chlorovirus type with permissive and non- permissive features on phylogenetically related algal strains *Virology* 500:103–113 doi: 10.1016/j.virol.2016.10.013

Speciale I, Agarkova I, Duncan GA, Van Etten JL, De Castro C. **2017** Structure of the N-glycans from the chlorovirus NE-JV-1 *Antonie van Leeuwenhoek* doi:10.1007/s10482-017-0861-3

Wang I-N, Li Y, Que Q, Bhattacharya M, Lane LC, Chaney WG, Van Etten **1993** Evidence for virus-encoded glycosylation specificity. *Proc. Natl. Acad. Sci. U.S.A.* 90, 3840–3844

Zhang Y, Xiang Y, Van Etten JL, Rossmann MG **2007** Structure and Function of a Chlorella Virus Encoded Glycosyltransferase *NIH Public Access* 15(9): 1031-1039

Zhang X, Xiang Y, Dunigan DD, Klose T, Chipman PR, Van Etten JL, Rossmann MG. **2011** Three-dimensional structure and function of the Paramecium bursaria chlorella virus capsid. *Proc. Natl. Acad. Sci. U.S.A.* 108:14837–14842

Section III

Experimental section

Chapter 8

Structural determination of the *N*-glycans from viruses

All chloroviruses and antigenic variants examined in this thesis have been grown and isolated by Professor James L. Van Etten research team in Department of Plant Pathology and Nebraska Center for Virology, University of Nebraska.

8.1 Glycopeptide(s) isolation

Isolation of the major capsid protein from chloroviruses were performed suspending the virus (1 ml, ~10^{11} plaque forming units) in phosphate buffered saline and heating the solution at 70 °C for 20 min. After cooling, the solution was centrifuged (5000 rpm, 4 °C, 30 min): the supernatant was recovered, instead the virus pellet was resuspended in a small volume of Tris buffer (50 mM Tris-HCl, pH 7.8) and the extraction was repeated twice. These two last supernatants were pooled together. The major capsid proteins were precipitated from both supernatants, adding five volumes of cold acetone and left at -20 °C for 16 hours. The proteins, checked via SDS-PAGE, were recovered by centrifugation as above and lyophilized. Major amount of MCP was always obtained in the precipitate of the first extraction.

Therefore, the glycopeptides were isolated from the major capsid protein, using an enzymatic hydrolysis with proteinase K (Sigma P6556) at 37 °C. To guarantee the complete enzymatic digestion, proteinase K was added three times (0.2 mg each time). Then, the mixture was freeze-dried and purified by size exclusion chromatography (Bio-Gel P-10, eluted with degassed water; Bio-Rad 150-4144; d = 1.5 cm, h = 118 cm, flow = 10 ml/h). The peptides were monitored with an online refractive index detector (Knauer K-2310) and fractions were pooled accordingly and monitored via ^{1}H-NMR.

Fractions containing the target glycopeptides eluted between 30-50% of the total column volume; the yield could not be determined accurately owing to the low amounts of the proteins.

8.1.1 SDS-PAGE

The extracted MCPs were checked through discontinue SDS-PAGE (Sodium Dodecyl Sulfate-PolyAcrylamide Gel Electrophoresis) analysis (Laemmli 1970), using the Mini Protean III Bio-Rad system, run at 150 constant voltages. This technique separated the biological macromolecules according to their electrophoretic mobility, which consisted of the motion of molecules under the influence of a uniform electric field.

Electrophoretic gel, consisted of acrylamide cross-linked with bis-acrylamide, made up of two parts: stacking gel (upper gel) and separating gel (lower gel), that had different composition and pH, so the samples were concentrated at the interphase among the stacking and the separating gel, before the effective separation across the sieves of the lower gel, similarly to a size exclusion frame, upon which large molecules remain on the top and small molecules migrate on the bottom. Size pores can be changed using different percentages of two components, so lower percentage gels were suitable for resolving higher molecular weight molecules, whereas much higher percentage were needed to resolve smaller molecules. In this contest, the MCPs isolated from the chloroviruses were analyzed using 8 or 12 % separating gel, instead for antigenic variants we used a 10 % separating gel. In the latter gel was added the BLUeye prestained Protein Ladder (2 μL), as protein marker, useful to determine the molecular weight of each sample. Approximately, 10 μg of protein were loaded into each well.

Gels were stained according with the silver stain procedure (Kittelberger and Hilbink 1993), for the detection of the presence of glycans on the major capsid protein, and coomassie staining (Smith 1984), to identify the proteins.

8.2 Determination of the glycopeptide structure(s)

8.2.1 NMR analysis

All NMR experiment were carried out on Bruker DRX 600 (^{1}H: 600 MHz, ^{13}C: 150 MHz) instrument equipped with a cryo-probe, except in two cases (TN603 and MT325), for which a Bruker 950 spectrometer at the facility in Grenoble was used,

to resolve the high complexity of the spectra and because of the low amount of the sample. Chemical shifts in spectra recorded in D_2O, were expressed in ppm relative to internal acetone (^1H/^{13}C 2.225/31.45 ppm) and the temperature was adjusted for each sample to minimize the overlap of glycopeptide signals with that of the residual water peak (temperatures are reported in the corresponding section).

The set of two-dimensional spectra (DQ-COSY, TOCSY, TROESY, NOESY, gHSQC, gHMBC and HSQC-TOCSY) were measured for glycopeptides from chloroviruses and antigenic variants, except for PIL6 and EIL-3B for which only the HSQC experiment was acquired. Homonuclear experiments were recorded using 512 FIDs of 2048 complex data points, whereas scan number was adjusted to find an optimal signal to noise ratio for each sample and depended on its abundance. Generally, from 24 to 40 scans per FID were used and mixing times comprises in a range of 100 to 300 ms were applied for TOCSY, TROESY and NOESY spectra acquisitions. ^1H-^{13}C heteronuclear experiments (HSQC, HMBC and HSQC-TOCSY) were acquired with 512 FIDs of 2048 complex points with 40-200 scans per FID, depending on the sample abundance.

Data processing was performed with the standard Bruker Topspin 3.1 software.

8.2.2 Chemical analysis

All GC-MS analyses were carried out on Agilent Technologies Gas Chromatograph 6850 coupled with a Mass Spectrometry (a selective mass detector 5973), equipped with a SPB-5 capillary column (Supelco, 30 m x 0.25 i.d., flow rate 1 mL/min using He as carrier gas). Electron impact mass spectra were recorded at an ionization energy of 70 eV. The temperature program used for the analyses was 150°C, for 5 min, 150 -> 280 C at 3°C/min 300°C for 5 min.

8.2.2.1 Octyl- glycosides

Absolute configuration was determined by preparing the acetylated or methylated 2-(-)-octyl derivatives. For the acetylated 2-(-)-octyl glycosides, samples were treated with 200 µl of 2-(-)-octanol and 14 µl of CH_3COCl at 60°C, for 16 h.

Excess of octanol was eliminated under air flux; then octyl glycosides were acetylated with pyridine (100 μL) and Ac$_2$O (50 μL), at 80 °C for 30 minutes (Leontein et al. 1978). Sample was dried and extracted with water and CHCl$_3$ (2000 rpm, 3 min, for 5 times) replacing the top layer with water each time. So, the organic phase was dried and it was analyzed by GC-MS.

In order to determine the absolute configuration of the OMe-residues (xylose and rhamnose), the methylated octyl-glycoside were performed (Fig. 8.1). The sample mixture was methylated according to Ciucanu and Kerek, hydrolyzed with TFA 2M (1 h at 120 °C), and then the octanolysis proceeds as described above.
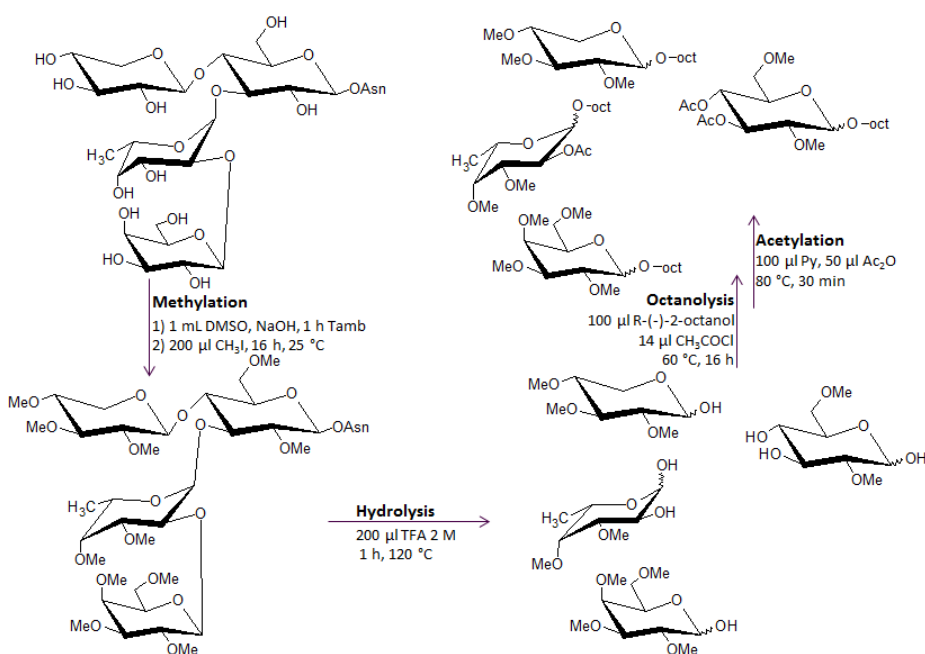
Figure 8.1 Methylated octyl-glycosides protocol.

8.2.2.3 Partially methylated alditol acetate (AAPM)

The determination of the branching point of the monosaccharide residue was obtained by analyzing partially methylated alditol acetates (AAPM), according to protocol described into De Castro et al. 2010. It consists of the following steps: methylation, hydrolysis, reduction and acetylation (Fig. 8.2).
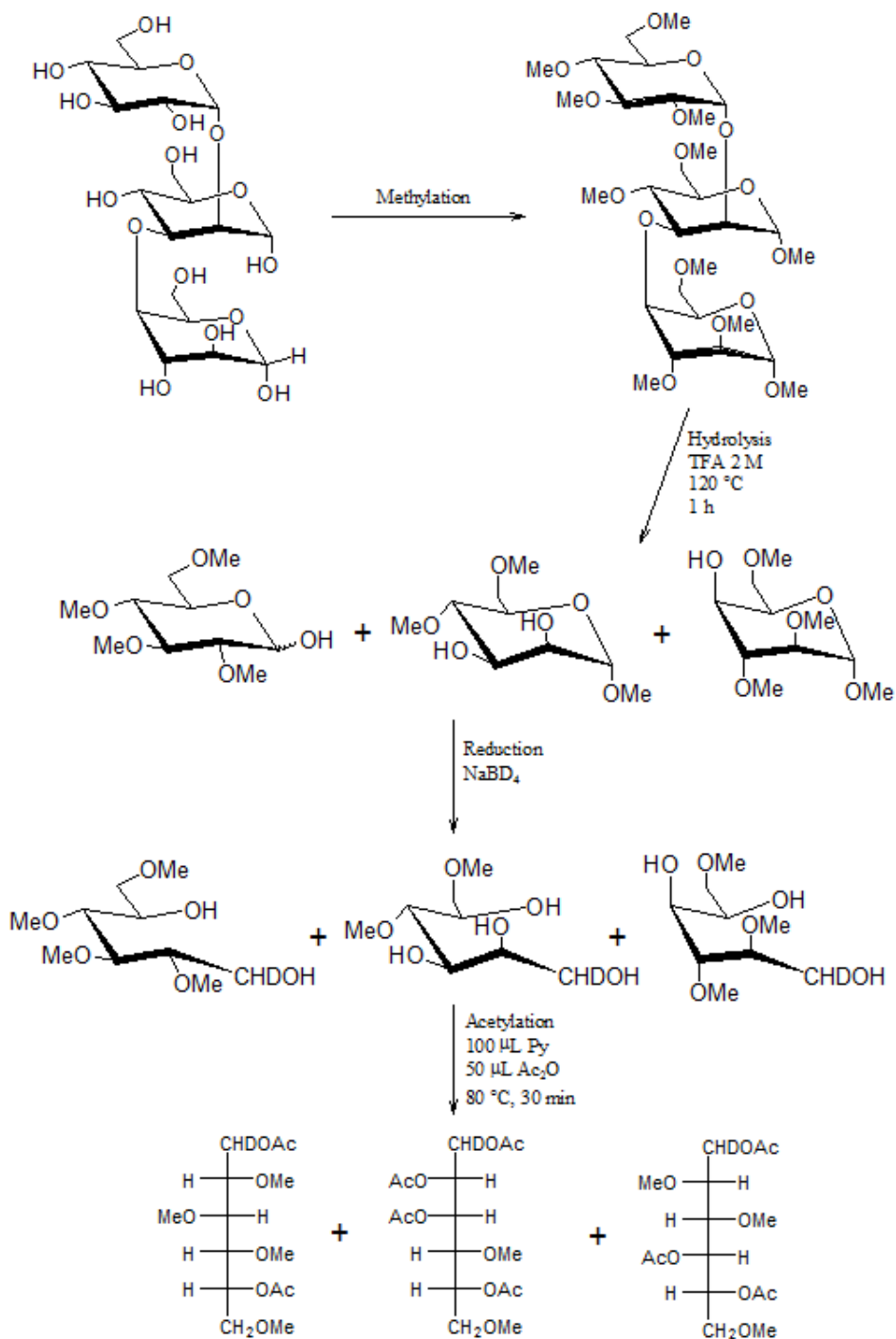
192

Figure 8.2. AAPM protocol

193

8.2.3 MALDI mass spectrometry

MALDI-TOF mass spectra were recorded in linear mode on a Perseptive (Framingham, MA, USA) Voyager STR equipped with a pulsed UV laser beam (nitrogen laser, λ 337 nm) and in reflectron mode on a 4800 Proteomic Analyzer (Applied Biosystems) provided with a Nd:YAG laser at the wavelength of 355 nm. Mass spectra were acquired twice using two different matrices, dihydroxybenzoic acid (DHB) or alphacyano-4-hydroxycinnamic acid, with the classical dried drop sample preparation. (DHB gives better results for glycopeptides) MALDI TOF/TOF MS/MS spectra were, also, recorded on the 4800 system without any collision gas. Glycopeptides were always present as sodium $[M+Na]^+$ and potassium $[M+K]^+$ adducts (pseudo molecular ions), to recognize glycopeptide from peptide pseudo-molecular ions. Since both *N*-glycan structures and glycosylation sites are extremely peculiar, classical bioinformatics tools were not useful, with the exception of FindPep tool (expasy.org), which was sometimes used for glycosylation site assignments. MS/MS spectra showed several glycosidic linkage cleavages, mainly corresponding to B, C and Y ions (Domon and Costello 1988), allowing the assignments of the branching of the oligosaccharide moieties.

8.2.4 HPLC-ESI-MS analysis

The molecular mass of proteins was detected using an Agilent 1100 HPLC system coupled to a MSD Ion Trap XCT mass spectrometer, equipped with an electrospray ion source (HPLC-ESI-MS) (Agilent Technologies, Palo Alto, CA, USA). Separations were performed on a Symmetry C18 column 1×150 mm with 3-μm particle size (Waters Corporation, Milford, MA, USA), using water and acetonitrile added with 0.1% formic acid as eluents (working gradient: 10% acetonitrile for 10 min, then linear to 95% in 50 min, with a flow rate of 30 μL/min). The column temperature was set at 25°C; the injection volume was 8 μL. Ions were detected in an ion charged control with a target ion value of 65000 and an accumulation time of 300 ms, using the following operation parameters: capillary voltage: 2600V; nebulizer pressure: 12 psi; drying gas: 8 L/min; dry temperature: 325°C; rolling averages 3, averages 8. Mass spectra were acquired in the positive ion mode in the 400-

2000 m/z mass range. Raw spectra were deconvoluted using the LC/MSD Trap Software revision 5.3.

8.3 Determination of the function of the first domain of a064r gene

8.3.1 Synthesis of UDP-β-ʟ-rhamnose

The synthesis of the UDP-β-ʟ-rhamnose was performed adding all necessary reagents together and shaking the mixture at 25 °C, for 1 hour. The mixture was composed by: 150 μl of UDP-α-ᴅ-glucose 100 mM; 600 μL of glucose-6-phosphate (G6P) 100 mM (it must be 4 times in molar excess with respect to glucose to guarantee the success of the reaction); 10 μl of glucose-6-phosphate-dehydrogenase (G6PD) 100 mM; 200 μl of $NADP^+$ 2mM; 100 μl of Mg^{2+} 40mM; 100 μL of PBS 20x; and the two enzymes, UDP-ᴅ-glucose-4,6-dehydratase (UGD, from Trichimonas) and L780 (from Mimivirus), that were anchored onto two resins. The reaction was stopped removing the resins by centrifugation and supernatant was used for the next reactions.
Note that, the G6PD, the G6P and Mg^{2+} ion, are used as regeneration system of the $NADPH/NADP^+$.

The reaction mixture was checked via NMR (600 MHz, at 298 K, in H2O, using acetone as internal standard) to verify that the synthesis was succeeded and to evaluate the amount of the UDP-rhamnose product. From the integration of the peaks relative to the rhamnose unit and knowing the initial amount of the UDP-Glc (8 mg), it was possible to established the amount of the UDP-rhamnose in the mixture (ca. 3.95 mg into 2.2 mL) and, then, its concentration (3.3 mM).

8.3.2 Bioluminescent assay

The UDP standard curve was prepared in a 96-well plate, performing a serial twofold dilution, as described below.
In the first well the UDP solution was prepared adding 0.5 μl of UDP solution 10 mM (provided in the UDP-Glo™ kit, V6961), into 200 μl of buffer (PBS 20x 10μl,

Mn^{2+} 40 mM 10μl, H2O 180 μl). In the following eleven wells, 100 μl of buffer (composed as described before) were added, and then a serial dilution was performed, transferring 100 μl from well A1 to well A2, and then from A2 to A3, and so on until the well A11; the last one was used as no-UDP control reaction. In this way, twelve solutions at different concentration of UDP were prepared. Thereafter, 25 μl, of each solution, were transferred in a second 96-well plate and 25 μl of UDP detection reagent (provided by the UDP-Glo kit) were added in each one. These solutions were kept at 25 °C, for 1 hour, and then the luminescence measure was performed. The same protocol was used for the second standard curve, that differs from the previous, for the presence of the magnesium 40 mM, instead that the manganese ion.

Reactions that involved the glycosyltransferase, the synthetic substrate and the UDP-rhamnose, in two buffers that differ for the presence of the Mg^{2+} or of the Mn^{2+}, were performed maintaining constant the ratio UDP-Rha/synthetic substrate (2:1 ratio; 2.470 nmol − 1.250 nmol, respectively) and changing the amount of the A064R-D1 glycosyl-transferase, that decreases proceeding from the first solutions to the last, in which was completely absent. The solutions at different concentrations of GT were obtained using the serial dilution twofold: 100 μl of the GT solution (5 μl of GT 150 μg/mL; 5 μl Mn^{2+}, or Mg^{2+}, 40 mM; 4.75 μL PBS 20x; 85.25 μl H2O) were added in the first well, in the other 11 wells was added 50 μl of buffer (composed by Mn^{2+}, or Mg^{2+}, 40 mM; PBS 20x and H2O) and then the dilutions were performed transferring 50 μl from the first well to the second and so on, keeping only the buffer solution in the last one. Then, 15 μl of each one were transferred into 12 wells of another 96-well plate, and 10 μl of the solution containing UDP-Rha and synthetic substrate (UDP-Rha 3.3mM, 30 μl; synthetic substrate 1mM, 50 μl; Mn^{2+}, or Mg^{2+}, 40 mM, 20 μl; PBS 20x 18.5 μl; H2O 281.5 μl) were added. The reactions were incubated at 25 °C, for one hour; thereafter, 25 μl UDP-detection were added in each one and the solutions were kept at 25 °C, for 1 hour, before to record the luminescence measure. Two replicas for each experiment (with the Mn^{2+} or Mg^{2+}) were performed.

Of note, the glycosyltransferase was expressed in *Escherichia coli* cells, as fusion protein with GST-tag, to allow easy purification, and after a proteolytic cleavage, was dissolved in buffer phosphate obtaining a concentration of 150 μg/mL. Instead, the synthetic substrate was a hyperbranched fucose, substituted with a galactose, xylose and rhamnose residues, that presents a small lipophilic tail, octyl residue, at the reducing end.

The same protocol was used for the last two reactions, in which the activated sugar donor or the acceptor substrate were changed, maintaining their ratio 2:1, as before. More in detail, in one reaction the UDP-rhamnose was substituted with the UDP-glucose 10 mM and, in the other, the synthetic substrate was changed with the xylose unit, 1mM. To keep unchanged the mole quantities of these two reagents, an appropriate volume of each one was taken. Note that, in these two reactions, both bivalent cations were added.

All luminescence measures were performed with a Synergy HT Multi-Mode Microplate Reader (BioTek instrument), using a 96-well microplate, with standard 128 x 86 mm geometry, and an integration time of 1.0 s. Luminescence was measured by the low-noise PMT detector, through an empty filter position in the Emission filter wheel, and data were analyzed using BioTek's Gen5™ or KC4™ Data Analysis Software.

8.3.3 Large scale reaction with A064R-D1 enzyme and UDP-Rha

Large scale reaction was performed in PBS buffer supplemented with Mn^{2+} ion, with a 20% excess of the sugar donor, UDP-Rha (7.86 μmol), with respect to the substrate (6.55 μmol), and the A064R-D1 protein (expressed as fusion protein with GST-tag and anchored to the resin). All components were mixed and reaction left overnight, under stirring on a moving plate.
The soluble components of the reaction (namely: PBS, free UDP, traces of UDP-Gal, UDP-Glc, UDP-4-keto intermediate, unreacted UDP-Rha, and G6P which was used in large excess for UDP-Rha synthesis) were recovered by centrifugation and then

the synthetic substrate produced was gained after Sep-Pak C18 cartridge purification, performed as described below.

The Sep-Pak cartridge was activated with 10 mL of EtOH, 4 mL of CH_3CN and 10 mL of water. After being loaded with the supernatant, the C18 cartridge was washed with 20 mL of water, 15 mL of CH_3CN/H_2O 1:4, 5 mL of CH_3CN and 20 mL of EtOH. The synthetic substrate was eluted with CH_3CN and EtOH elutions, that were collected together.

8.4 Reconstruction of Major Capsid Protein Vp54

8.4.1 X-ray diffraction

Coordinates and structure factors for the major capsid protein of PBCV-1 were recovered from the Protein Data Bank and evaluated by inspection of the structures using the program COOT. Sugars present in the structure were removed and the chemically identified oligosaccharides were modeled according to the available glycan composition. Models were refined using refmac in the CCP4 suite of programs. The final Rwork/Rfree factors were 0.1868/0.2110 and 0.1946/0.2489 for the structures with PDB accession codes 1M3Y and 1J5Q, respectively.

8.4.2 Evaluation of experimental distances of glycans

Determination of experimental distances was performed using the T-ROESY spectrum on the N-glycopeptide mixture. T-ROESY experiment was carried out in a Bruker DRX-600 instrument, described before. 512 FIDs of 2048 complex data points and 32 scans per FID were acquired with 250 ms as mixing time. Data processing and T-ROESY cross peak integrations were performed with the standard Bruker Topspin 3 program, and the elaboration of cross peak volumes into distances was performed with MS Excel program.

8.4.3 Molecular Mechanism (MM) and Molecular Dynamic (MD) of the free oligosaccharides

The optimal dihedral angles of each glycosidic junctions of the N-glycans, were evaluated with the Molecular Mechanism (MM) approach. The simulation was carried out with a dielectric constant (ε=80), as approximation of bulk water, and using the MM3* and Amber force fields, as implemented in MacroModel program, in the Schrödinger 2012. Coordinate Scan utility was used to calculate the energy maps: both Φ and Ψ were varied gradually, employing a grid step of 18°, each (Φ, Ψ) point of the map was optimised using 10000 P.R. conjugate gradients.

On the basis of the values obtained from MM approach, the two oligosaccharides, glyco1 and glyco2, were built, minimized with MM3 or Amber FFs, approximating water solvent with a dielectric constant ε=80. Molecular Dynamics simulations (with MM3 and Amber FF) were performed at 310 K; structures were initially subjected to an equilibration time of 250 ps and then a dynamic simulation was run for 20 ns, using a dynamic time step of 2.0 fs. SHAKE protocol to the hydrogen bonds was applied obtaining 10000 frames. For each frame, the interproton distances or Φ/Ψ values were extracted with the tools in Maestro program in Schrödinger suite, and then the MS Excel program was used to elaborate the Φ/Ψ scattered maps and to calculate the distance averaged values. Frequency count of Φ and Ψ was performed with Origin program to appreciate which value had the highest probability. These values were used to build the complete N-glycan used for MCMM search.

8.4.4 Construction of the complete Vp54

The revised 1M3Y.pdb was imported into MacroModel, and all (water, Hg, N-glycans and B-D chains) were delated, except Chain A that was corrected with the Protein Wizard facility. Thus, the correct bond order of each residue was assigned; charges were calculated with Propka setting at pH 7.0 (using Amber FF with ε=80); histidine protonation status along with hydroxyl, amino and thiol functions orientation was determined by the program. Flipping of Asn and formation of

disulfide bonds were not allowed. Asp-110 had steric clash and its dihedrals were set to those of Asn-110 in Chain B, were this problem did not occur.

Therefore, the revised polypeptide chain was used to build the fully glycosylated glycoprotein, connecting the oligosaccharides, glyco1 or glyco2, previously elaborated, at the appropriate Asn positions.

8.4.5 MCMM search

The fully glycosylated Vp54 was optimized applying Metropolis Montecarlo conformational search (MCMM), that was focused on the glycosidic dihedral angles, and neglected the other parameters (bond lengths, ring conformation, hydroxyl orientation...). More in detail, such search was performed varying the Φ and Ψ values of those residues that were absent, distorted or linked to a distorted unit in the X-ray: D, N, M, C and I for [280]Asn glycan; N, M, I and C for [302]Asn; N, M, G, I and C for [399]Asn; all residues except H for [406]Asn.

MCMM set up provided the minimization (Amber, $\varepsilon = 80$) for the N-glycans and the corresponding Asn, and two shells constraints were applied for the protein: atoms in the first shell (3 Å from glycans) were fixed (force constant 200.0 kJ/mol Å^2); atoms in the second shell (6 Å) were frozen; the remaining atoms were ignored.

The maximum number of steps was set to 15000, steps per rotatable bond were 100 and the energy window for saving structures was 200 kJ/mol. This analysis produced 11515 different structures, ordered by increasing potential energy; that were analyzed with the Superimposition procedure of Maestro to find those conformers that better matched the monosaccharide residues well resolved in the crystallographic data: **H**, **A** and **E** for Asn-280, -302, -399; only **H** for Asn-406. Selection focused on those conformers that had an energy 50 kJ/mol above the global minimum and with RMSD less than 0.8. Thus, two structure were chosen.

8.4.6 Molecular Dynamic of complete MCP

The two structures obtained from the MCMM analysis were used as starting point of the Molecular Dynamic simulation, that was focused on the glycan portion of the glycoprotein. Thus, the same constraints applied at the polypeptide backbone during MCMM conformational search were set: the four oligosaccharides and the glycosylated Asn were free to move; atoms within 3 Å were fixed with force constant 200.0 kJ/mol Å^2; atom in the second layer (6 Å) were kept frozen, and that left were ignored. MD simulation was performed with Amber force field approximating the bulk with the dielectric constant ε=80. The initial conformer was equilibrated for 250 ps and successively kept at 310 K for 100 ns, using a dynamic time step of 2.0 fs and collecting 30000 frames. Extraction of Φ/Ψ values was performed with Maestro and graph elaborated with MS Excel.

8.5 Glycans function

8.5.1 Construction of fully glycosylated Vp54

The glycan structures obtained with MacroModel were not accepted as input for Amber program, therefore, it has been necessary to re-build the full glycosylated Vp54 structure. The four glycans were constructed using the GLYCAM-web program, setting the correct dihedral angle values, and then they were connected to the polypeptide chain (obtained from the reconstructed Vp54) at the proper Asn, using the Discovery Studio Visualizer program.

8.5.2 Molecular dynamic using Amber12 program

The molecular dynamic simulation was performed on Vp54 protein with and without glycans and at different temperatures, using AMBER12 package.

The correct pdb file relative to each structure was loaded on Xleap module of AMBER, re-calling different force fields depending on the initially structure, in order to obtain the .tpp (topology) and .xyz (coordinates) input files. More in detail, when the only protein was considered, the ff12SB force field was loaded, instead for

the fully glycosylated protein was added another FF, GLYCAM_06j-1, that was useful to the interpretation of the glycidic portion. In all cases, Cl- ions were added to neutralize the molecules and the whole systems were enclosed in a box of water TIP3PBOX using a 12 Å cut-off.

Before starting the simulations, all systems were minimized (500 steps of deepest descendent and 500 steps of conjugate gradient algorithms) holding the solute fixed. Then, the systems were entirely minimized (maxcyc = 2500, ncyc = 1000) and, then, were warmed from 0 K up to 300 or 500 K with a 20 ps MD simulations, keeping solute fixed with weak restraint. The following 100 ps of equilibration phase were characterized by an energy minimization of the entire systems without restraints, fixing the temperature (at 300 or 500 K) and the pressure (1 atm). Finally, the MD simulation was run for 50 ns for all cases, except for the protein without glycans at 500 K (for which 25 ns were set), using a time step of 2 fs, under constant pressure (1 atm) and temperature, controlled by Langevin dynamics with a collision frequency of 1.0 ps^{-1}. During all simulations, the constant volume periodic boundaries were used and the constraints not occurred; keeping the entire molecule free to move, in order to mimic its behaviour in the solution.

These dynamics produced 25000 different frames when running for 50 ns and 12500 frames when considered 25 ns. These ensembles of structures were analyzed extrapolating the RMSD value to visualizing their behaviour in the entire simulation, using Trajin program. The RMSD was calculated considering the entire peptide backbone (NH-Cα-CO) with respect to the initial structure. Instead, the analysis of the RMSD value of the two singles domains took into account the polypeptide backbone of the corresponding domain. The obtained values were plotted using MS Excel.

8.5.3 Normal Modes

The whole packages of the structures obtained from the previous molecular dynamics, where examined with appropriate scripts to investigate the relative movements of the two domains. Regarding the closure and the wobbling modes,

the glycopeptide was considered as two subunits held together by a hinge, and the movement of these subunits were calculated taking into account the center mass. The two subunits were chosen in according with the two domains of Vp54 (D1: 27 – 212; Hinge: 213 – 224; D2: 225 – 437).

Regarding the twisting mode, the distance between **N**H and **C**O of different amino acids situated in two different domains were calculated: distance between ^{280}Asn-^{129}Arg and ^{302}Asn-^{129}Arg were investigated. Graphs were elaborated with MS Excel.

References

Ciucanu I, Kerek F **1984** A simple and rapid method for the permethylation of carbohydrates. *Carbohydr. Res.* 131: 209-217.

De Castro C, Parrilli M, Holst O, Molinaro A **2010** Microbe-associated molecular patterns in innate immunity: extraction and chemical analysis of gram-negative bacterial lipopolysaccharides *Methods in Enzymology*, Vol. 480: 89-115.

Domon B, Costello CE **1988** A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. Glycoconj. J. 5, 397–409, doi:10.1007/BF01049915.

Laemmli UK **1970** Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature.* 227 (5259): 680-685.

Leontein K, Lindberg B, Lönngren J **1978** Assignment of absolute configuration of sugars by g.l.c. of their acetylated glycosides formed from chiral alcohols. *Carbohydr. Res.* 62, 359-362

Smith BJ **1984** p. 144-146 in Methods in Molecular Biology. Volume 1: Proteins. JM Walker ed. 365 pages. Humana Press, Clifton, New Jersey.

Kittelberger R, Hilbink F **1993** Sensitive silver-staining detection of bacterial lipopolysaccharides in polyacrylamide gels. *J. Biochem. Biophys. Meth.* 26: 81-86.

Appendix

PhD Course Activity Summary

1) Attended Courses:

- Spettrometria di Massa; Pucci P.; from 30/06/2014 to 04/07/2014; 4 CFU.

- Glicoscienza; Parrilli M., Bedini E.; 17-20/06/2014; 3 CFU.

- Tecniche estrattive Solido-Liquido impiegate nella preparazione del campione per l'analisi chimica e nella produzione di estratti per usi industriali; Naviglio D.; 09-17/12/2014; 3 CFU.

- Sintesi, struttura ed applicazioni di oligonucleotidi naturali e modificati; Montesarchio D.; 04-13/ 02/2015; 3 CFU.

- Produzione Ricombinante di Proteine Naturali e Mutanti; Duilio A.; 20-24/07/2015; 3 CFU.

- Neutron scattering techniques; Zorn; 27-30/10/2015; 3 CFU.

- NMR allo stato solido; Pintacuda G.; 22-24/02/2016.

2) Attended Seminars:

Title	Speaker	Place	Date
Site Quality Control Head: Novartis Farma S.p.A	Antonio Vitiello	Dept Chemical Sciences, University of Naples Federico II	10/04/2014
Incontro con la Peroni: Azienda leader birraria	Luigi Serino	Dept Chemical Sciences, University of Naples Federico II	12/05/2014
Multifunctional bioinspired cathechol - based coatings and nanoparticles	Daniel Ruiz-Molina	Dept Chemical Sciences, University of Naples Federico II	05/11/2014
Le fitotossine: un'avventura lunga 40 anni	Antonio Evidente	Dept Chemical Sciences, University of Naples Federico II	10/12/2014
New chiral catalysts derived from Iron(II) and Bismut (III) for asymmetric synthesis	Thierre Ollevier	Dept Chemical Sciences, University of Naples Federico II	15/12/2014

Title	Speaker	Place	Date
Dall'analisi dei reflui alla depurazione degli impianti	Antonio Riccio	Dept Chemical Sciences, University of Naples Federico II	30/04/2015
Matrix metalloproteinase: inter-domain flexibility and recognition of substrates	Marco Fragai	Dept Chemical Sciences, University of Naples Federico II	08/05/2015
European large scale facilities: neutron and synchroton source	Serge Perez	Dept Chemical Sciences, University of Naples Federico II	06/05/2015
Lectins from bacteria and fungi: therapeutical targets and research tools	Anne Imberty	Dept Chemical Sciences, University of Naples Federico II	15/05/2015
Characterizing biomolecular interactions: Biacore experience	Silvia Pileri	Dept Chemical Sciences, University of Naples Federico II	19/05/2015
Pharmaceutical companies: external manufacturing and quality assurance	D. Demasi	Dept Chemical Sciences, University of Naples Federico II	25/05/2015
Biosensing and bioelectronics based on organic electrochemical devices: from monitoring drug dynamics to hybrid bio-organic devices	Salvatore Iannotta	Dept Chemical Sciences, University of Naples Federico II	11/06/2015
The bacterial cell wall by liquid state, standard and DNP solid state NMR	Jean-Pierre Simorre	Dept Chemical Sciences, University of Naples Federico II	29/06/2015
Synthetic and biofunctional studies of microbial and animal glycan toward immunoregulation	Koichi Fukase	Dept Chemical Sciences, University of Naples Federico II	29/09/2015
Innovative approaches for polysaccharide-based vaccines	Francesco Berti	Dept Chemical Sciences, University of Naples Federico II	20/11/2015
Industria e ricerca nel settore biofarmaceutico: bisogni attuali e sviluppi futuri	Sara Carillo	Dept Chemical Sciences, University of Naples Federico II	25/02/2016
Basics of detergents formulations and challenges	Giulia Bianchetti	Dept Chemical Sciences, University of Naples Federico II	16/03/2016
La Chimica: un diamante	Mario Marzullo	Dept Chemical Sciences, University of Naples Federico II	21/04/2016

3) Visiting periods in Institutions different from University of Naples "Federico II":

Host Institution	Country	Start Date	End Date
CIC BioGune (Center for Cooperative Research in Biosciences)	Bilbao, Spagna	01/10/2016	21/10/2016
CIC BioGune (Center for Cooperative Research in Biosciences)	Bilbao, Spagna	16/01/2017	30/01/2017

4) Publications:

- <u>Speciale I</u>, Tonetti M, Lowary T, Molinaro A, Duncan GA, Van Etten JL, De Castro C. Antigenic variants of the Chlorovirus PBCV-1: function of the *a064r* gene. (in preparation)

- De Castro C, <u>Speciale I</u>, Molinaro A, Van Etten JL, Klose T, Rossmann MG. **2017** Major Capsid Protein of Chlorovirus PBCV-1 Structure combining X-ray and Carbohydrate Computer Modelling Approaches. *Proc. Natl. Acad. Sci* (paper under revision)

- <u>Speciale I</u>*, Agarkova I, Duncan GA, Van Etten JL, De Castro C. **2017** Structure of the N-glycans from the chlorovirus NE-JV-1 *Antonie van Leeuwenhoek* doi:10.1007/s10482-017-0861-3

- Quispe CF, Esmael A, Sonderman O, McQuinn M, Agarkova I, Battah M, Duncan GA, Dunigan DD, Smith TPL, De Castro C, <u>Speciale I</u>, Ma F, Van Etten JL **2016** Characterization of a new chlorovirus type with permissive and non- permissive features on phylogenetically related algal strains *Virology* 500:103–113 doi: 10.1016/j.virol.2016.10.013

- Kenyon JJ†, <u>Speciale I</u>†, Hall RM, De Castro C **2016** Structure of repeating unit of the capsular polysaccharide from Acinetobacter baumannii D78 and assignment of the K4 gene cluster. *Carbohydrate Research* 434 doi: 10.1016/j.carres.2016.07.016

iii

- De Castro C, <u>Speciale I</u>, Duncan G, Dunigan DD, Agarkova I, Lanzetta R, Sturiale L, Palmigiano A, Garozzo D, Molinaro A, Tonetti M, Van Etten JL **2016** N-linked Glycans of Chloroviruses Sharing a Core Architecture without Precedent. *Angew. Chem. Int. Ed.* 55, 654–658; doi:10.1002/anie.201509150

- <u>Speciale I</u>, Paciello I, Fazio LL, Sturiale L, Palmigiano A, Lanzetta R, Parrilli M, Garozzo D, Lemaitre B, Bernardini ML, Molinaro A, De Castro C **2015** Determination of structure of the O-antigenic and the lipid a from the entomopathogenic bacterium Pseudomonas entomophila lipopolysaccharide along with its immunological properties. *Carbohydrate Research* 412: 20-27.

† Contributed equally
* Corresponding author

5) Attended congresses/workshops/summer schools/contribution:

- BMMC VII - 7th Baltic Meeting on Microbial Carbohydrates; 25 – 29/ 09/2016; Güstrow (Rostock); **Oral Comunication:** Structural Analysis of Antigenic Variants of PBCV-1 (I. Speciale, C. De Castro, M. Tonetti, J.L. Van Etten, G. Duncan, A. Molinaro).

- Ieiis 2016: 14th Biennial Meeting International Endotoxin and Innate Immunity Society; 22 – 24/09/22016; Hamburg (Germany); **Poster session:** Structural Analysis of Antigenic Variants of PBCV-1 (I. Speciale, C. De Castro, R. Lanzetta, A. Molinaro, M. Parrilli, M. Tonetti, J.L. Van Etten).

- XIV Convegno - Scuola Sulla chimica dei carboidrati; 19 - 22/06/2016; Certosa di Pontignano (Siena); **Oral Comunication:** Structural analysis of antigenic variants of *PBCV-1* (Immacolata Speciale, Cristina De Castro, Michela Tonetti, James Van Etten, Antonio Molinaro)

- Workshop Schrödinger: 15th Annual European User Group Meeting; 23 - 25/ 09/2015; The Westin Paris - Vandôme (Paris).

- ISSNP: International Summer School on Natural Products in Naples; 06 - 10 /07/2015; Stazione Zoologica Anton Dohrn (Napoli); **Poster session:** Structural Analysis of Antigenic Variants of PBCV-1 (I. Speciale, C. De Castro, R. Lanzetta, A. Molinaro, M. Parrilli, M. Tonetti, J.L. Van Etten).

- COST Action BM1003 Final Meeting: Microbial cell surface determines of virulence, antibiotic resistance and inflammation in different pathologies; 13 - 14 /10/2014; University of Naples Federico II, Complesso dei SS. Marcellino e Festo.

- XIV Convegno - Scuola Sulla chimica dei carboidrati; 22 - 25 /06/2014; Certosa di Pontignano (Siena); **Oral Comunication:** N-linked glycans from PBCV-1 and related viruses have a new type of core region (I. Speciale, D. Garozzo, J.R. Gurnon, R. Lanzetta, A. Molinaro, A. Palmigiano, M. Parrilli, F. Piacente, L. Sturiale, M. Tonetti, J.L. Van Etten, C. De Castro).

- COST Training School BM 1003: Microbial cell surface determinants of virulence as target for new therapeutics in Cystic Fibrosis; 03-05/06/ 2014; University of Naples Federico II, Complesso di Monte Sant'Angelo.