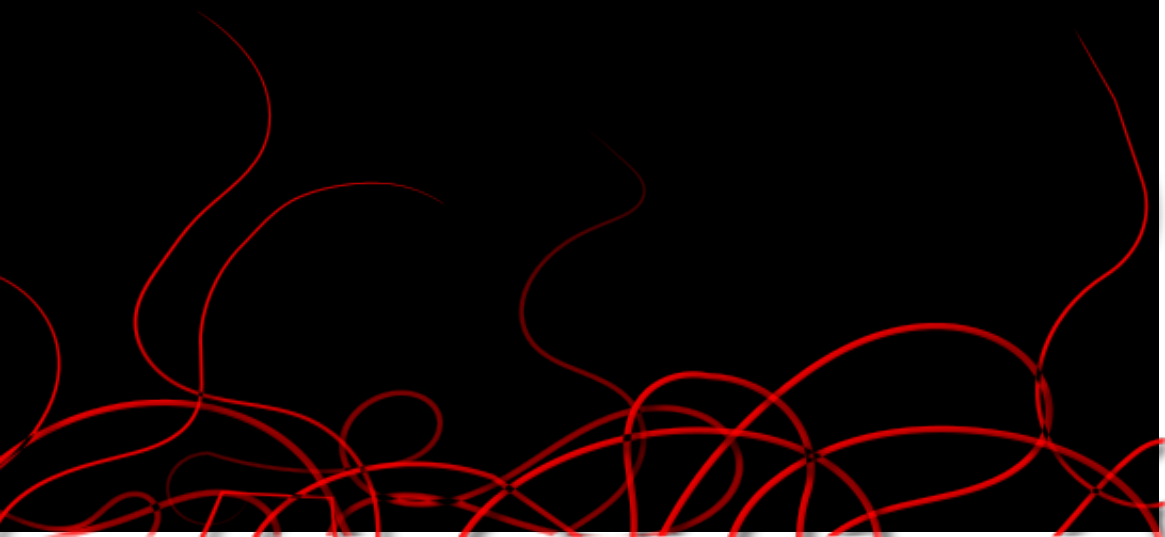


Deciphering the molecular basis of the specificity of  
**protein-carbohydrate interactions**  
by statistical analysis of 3D structural data from  
the Protein Data Bank

Miguel Ángel Rojas Macías



# Deciphering the molecular basis of the specificity of protein-carbohydrate interactions by statistical analysis of 3D structural data from the Protein Data Bank

A thesis submitted to the Faculty of Biology and Chemistry -FB08- in fulfillment of  
the requirements of the degree Doktor der Naturwissenschaften (Dr.rer.nat.) at the  
Justus Liebig University Gießen, Germany

By  
MIGUEL ÁNGEL ROJAS MACÍAS  
from México City, México

Gießen, Germany 2014



The present work was carried out at the Institute of Veterinary Physiology and Biochemistry (Faculty 10), Justus Liebig University of Gießen under the supervision of Prof. Dr. Thomas Lütteke and Prof. Dr. Thomas Wilke.

**Prof. Dr. Thomas Lütteke**

Institute of Veterinary Physiology and Biochemistry  
Justus Liebig University  
Frankfurter Str. 100  
D-35392 Gießen  
Germany

**Prof. Dr. Thomas Wilke**

Department of Animal Ecology and Systematics  
Heinrich-Buff-Ring 26-32 IFZ  
D-35392 Gießen  
Germany





*To my parents*



## Ehrenwörtliche Erklärung

Hiermit erkläre ich: Ich habe die vorgelegte Dissertation selbstständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, sind als solche kenntlichgemacht.

Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der "Satzung der Justus-Liebig-Universität zur Sicherung guter wissenschaftlicher Praxis" niedergelegt sind, eingehalten.

Gießen, Dezember 2014

.....  
Miguel Ángel Rojas Macías

## Acknowledgements

I thank Prof. Dr. Thomas Lütteke for introducing me to a very exciting field of research, for his patience, support and for the time and effort he put into reviewing my work. I am very grateful to Prof. Dr. Thomas Wilke for always finding the time to help me through the rough road to finish this thesis. I have greatly benefited from his guidance and expertise. Working with colleagues and students at the Institute of Veterinary Physiology and Biochemistry has been an unforgettable experience. The support that I received from my fellow PhD students: Mazen, Raymund and Dimitris, as well as the laughter, chats and moments that we shared were very much appreciated and necessary, I wish you all the best guys! I am also thankful to the International Giessen Graduate Centre for the Life Sciences (GGL) for all the opportunities that would not have been possible otherwise.

I am grateful to Göran Widmalm for his support and encouragement when generously hosting me in Stockholm. ECODAB has been a very rewarding project. Thank you Jonas Ståhle for being a great coworker. Good luck with your own defense and future career. I would like to extend my gratitude to all the members of the Widmalm Research Group who were so nice with me during my stay in Stockholm.

I also would like to express my sincere gratitude to Niclas Karlsson for making me part of his group. This beginning has been quite unconventional but I really appreciate your trust and patience. Varja, Barbara, Jessica, Sarah, Jin and Liaqat, my new colleagues, thank you for all the support. It is a pleasure sharing lab with all of you.

Jens, gracias por siempre alumbrar mi oscura razón. Tu presencia y apoyo han sido invaluable, así que este logro es tuyo también (unser prinzip). Gracias por este nuevo idioma que he aprendido y que ha redefinido el mundo. Jag är evigt tacksam Thomas, nu blir jag lyckligare när jag ser den gyllene säden. Gracias Stefan, por tu generosidad con los extraños, y por mantenerme con los pies en la tierra. Fawad y Jörg, les agradezco por su aliento. A mis cómplices desde esa otra vida, Varcol y Joaquín: que suerte contar con su amistad a pesar de la distancia y tiempo. Gracias Mario, por ser un contrapunto en mi historia. Espero que sigamos creciendo y acompañándonos en este exilio. Finalmente, quiero agradecer a mis padres, Ana y Aurelio, porque los privilegios que he disfrutado y que me han permitido construir mi vida se deben a su amor, trabajo, apoyo y sacrificios. Gracias por toda la libertad que siempre me han dado, y por aceptar mi ausencia. Este trabajo esta dedicado a ustedes.

The road that has led me to this point has been longer, darker and more treacherous than expected. To the rest of people that have directly or indirectly supported me during all these years: thank you.

## Publications

1. Miguel A Rojas-Macias, Jonas Stähle, Thomas Lütteke and Göran Widmalm. Development of the ECODAB into a relational database for *Escherichia coli* O-antigens and other bacterial polysaccharides. *Glycobiology*, 25(3):341-347, Mar 2015.
2. Miguel A Rojas-Macias, Alexander Loss, Andreas Bohne-Lang, Martin Frank and Thomas Lütteke. Databases and Tools of the GLYCOSCIENCES.de web server. In: Taniguchi N, Endo T, Hart GW, Seeberger P, Wong CH (Eds.), Glycoscience - Biology and Medicine, Springer, *in press*
3. Miguel A Rojas-Macias and Thomas Lutteke. Statistical Analysis of Amino Acids in the Vicinity of Carbohydrate Residues Performed by GlyVicinity. In: Lutteke T, Frank M (Eds). Methods in Molecular Biology - Glycobioinformatics, Springer, *in press*

## Summary

Glycoproteins, proteins with covalently attached sugar residues, are commonly found on the surface of almost all cells where they give rise to a vast layer called glycocalyx. The intricacy of this layer confers cells with a distinctive identity that is recognized by proteins and other receptors. Such protein-carbohydrate interactions set the basis for a broad range of essential biological events, including cell adhesion, signal transduction, the migration of leukocytes to sites of inflammation and the immunological response to carbohydrate antigens. Therefore, the need to comprehend these interactions in detail.

Carbohydrate-binding proteins show a remarkable specificity, being able to discern between very similar glycan residues. In order to better understand this precise specificity, the spatial vicinity of common monosaccharide residues is inspected in order to determine their amino acid preferences. Two datasets have been examined. Firstly, one composed of non-covalently bound carbohydrate ligands. The results of this analysis is compared to the second dataset, obtained from the study of the spatial vicinity of the monosaccharides that form the common structural core shared by different types of N-glycans. Amino acid patterns were determined based on two measures: absolute numbers and deviations from natural abundances. Superimposition of protein atoms made possible to display in more detail how interactions are established based on their stereochemical structure. This study is not limited to a given protein-carbohydrate complex, instead, all 3D structures of protein-carbohydrate complexes found in the Protein Data Bank are collected, clustered, and statistically analyzed.

The results show that polar amino acids are more frequently found in the vicinity of the analyzed monosaccharides than non-polar residues. Due to the hydrophilic nature of glycan residues, it is not surprising that the main force directing protein-carbohydrate interactions is hydrogen bonding. However, contacts between sugar residues and the non-polar surfaces of aromatic amino acids are confirmed to be also essential for the discrimination of carbohydrate residues. The role of Trp is especially remarkable since it is clearly overrepresented around all analyzed carbohydrate residues.

Monosaccharides display polar and non-polar patches on their surfaces that define the way in which interactions with proteins are established. Accordingly, changes in their conformations may also change the pattern of these patches and affect binding specificity. Linkages between proteins and glycans also influence the distribution of amino acids. For example, when comparing amino acids found in the spatial vicinity of N-glycans with those surrounding the glycosylation site, it is shown that the N-glycans interact mostly with amino acids located far from the glycosylation site. This suggests that N-glycans have a minor role at building a binding site, in contrast to non-covalently bound ligands.

This study proves that the Protein Data Bank can be effectively used as indirect source of carbohydrate 3D data. The results constitute a guide on the type of amino acid preferences that each of the analyzed monosaccharides display. The statistical analysis was performed with a bioinformatics application that was redeveloped for this purpose. This software is publicly accessible, and different parameters can be used to refine the search space. Even if the number of entries is small, in some cases the results are interesting since they indicate the beginning of a pattern which can be later corroborated once the number of entries increases.



## Zusammenfassung

Glykoproteine sind Proteine mit kovalent gebundenen Zuckerresten und werden normalerweise an der Oberfläche aller Zellen gefunden, wo sie eine große Schicht bilden, die Glykokalyx genannt wird. Die Komplexität dieser Schicht verleiht Zellen eine unverwechselbare Identität, die von Proteinen und anderen Rezeptoren erkannt wird. Die Wechselwirkungen zwischen Proteinen und Kohlenhydraten bilden die Grundlage für eine große Anzahl von wichtigen biologischen Vorgängen, wie zum Beispiel Zelladhäsion, Signalübertragung, Wanderung von Blutleukozyten zu Entzündungsherden und Immunreaktionen auf Kohlenhydrat-Antigene.

Kohlenhydrat-bindende Proteine zeigen eine bemerkenswerte Spezifität. Sie sind in der Lage, zwischen sehr ähnlichen Glykanresten zu unterscheiden. Um diese genaue Spezifität besser zu verstehen, wird die räumliche Nähe bekannter Monosaccharide untersucht, um ihre Aminosäurepräferenz zu bestimmen. Zwei Datensätze sind geprüft worden. Zuerst wurden alle nicht-kovalent gebundenen Aminosäuren in Protein-Kohlenhydrat-Komplexen ausgewählt. Die Ergebnisse wurden mit dem zweiten Datensatz verglichen, den man aus der Analyse der räumlichen Nähe der Monosaccharide erhält, welche den gemeinsamen strukturellen Kern durch unterschiedliche Arten von N-Glykanen bilden. Aminosäuremuster wurden auf zwei Wege bestimmt: Absolute Zahlen und Abweichungen von den natürlichen Häufigkeiten. Die Überlagerung von Proteinatomen erlaubt im Detail zu erkennen, wie Interaktionen auf der Grundlage ihrer stereochemischen Struktur stattfinden. Diese Studie ist nicht auf ein bestimmtes Protein-Kohlenhydrat-Komplex beschränkt, sondern es werden alle 3D-Strukturen von Protein-Kohlenhydrat-Komplexen in der Protein Data Bank gefunden, gesammelt, gruppiert und statistisch analysiert.

Die Ergebnisse zeigen, dass polare Aminosäuren häufiger in der Nähe der untersuchten Monosaccharide gefunden werden als nicht-polare Reste. Aufgrund der hydrophilen Natur der Kohlenhydrate ist es nicht verwunderlich, dass die Hauptkraft, die die Protein-Kohlenhydrat-Wechselwirkungen bestimmt, die Wasserstoffbrückenbindung ist. Allerdings konnte gezeigt werden, dass auch die Kontakte zwischen Zuckerresten und nicht-polaren Flächen der aromatischen Aminosäuren für die Unterscheidung von Kohlenhydratresten wesentlich sind. Die Rolle von Trp ist besonders bemerkenswert, da es bei allen betrachteten Kohlenhydratresten überrepräsentiert ist.

Monosaccharide haben polare und unpolare Bereiche auf ihrer Oberfläche, die die Art und Weise der Wechselwirkung mit Proteinen definieren. Daher können Änderungen in ihrer Anordnung zu einer Änderung im Muster dieser Bereiche führen und auch die Bindungsspezifität beeinflussen. Verbindungen zwischen Proteinen und Glykanen beeinflussen auch die Verteilung der Aminosäuren. Bei einem Vergleich zwischen Aminosäuren in der räumlichen Nähe von N-Glykanen und denen, die eine Glykosilierungsstelle umgeben, zeigt sich, dass N-Glykane eher mit Aminosäuren interagieren, die sich weiter von der Glykosilierungsstelle entfernt befinden. Dies legt nahe, dass N-Glykane eine unterge-

ordnete Rolle beim Aufbau einer Bindungsstelle spielen, im Gegensatz zu nicht-kovalent gebundenen Liganden.

Diese Studie beweist, dass die Protein Data Bank sehr effektiv als indirekte Quelle von 3D-Daten auf Kohlenhydrate angewendet werden kann. Die Ergebnisse bilden einen Leitfaden darüber, wie die untersuchten Monosaccharide mit unterschiedlichen Aminosäuren interagieren. Die statistische Analyse wurde mit einer Bioinformatikanwendung, die für diesen Zweck komplett überarbeitet wurde, durchgeführt. Diese Software ist öffentlich zugänglich und kann angepasst werden, um neue Parameter zu akzeptieren und verschiedene Suchaufträge durchzuführen sobald sich die Anzahl der Einträge erhöht hat. Auch wenn die Anzahl der Einträge in einigen Fällen noch gering ist, sind die Ergebnisse interessant, da sie den Anfang eines Musters zeigen, das später, wenn die Datenmenge steigt, bestätigt werden kann.



# Abbreviations

<b>Å</b>	Ångstroms. 16
<b>3D</b>	Three dimensional. 1
<b>Ala (A)</b>	Alanine. 38
<b>Arg (R)</b>	Arginine. 24
<b>Asn (N)</b>	Asparagine. 5
<b>Asp (D)</b>	Aspartic acid. 23
<b>CDG</b>	Congenital Disorders of Glycosylation. 5
<b>CRD</b>	Carbohydrate Recognition Domain. 7
<b>Cys (C)</b>	Cysteine. 39
<b>D-Galp</b>	D-Galactopyranose. 2
<b>D-GlcNAc</b>	N-acetyl-D-glucosamine. 3
<b>D-Glcp</b>	D-Glucopyranose. 2
<b>D-Manp</b>	D-Mannopyranose. 3
<b>D-Neu5Ac</b>	N-Acetyl-D-neuraminic acid. 3
<b>DNA</b>	Deoxyribonucleic acid. 3
<b>EHEC</b>	Enterohemorrhagic <i>Escherichia Coli</i> . 10
<b>ER</b>	Endoplasmic Reticulum. 5
<b>GlcNAcTs</b>	N-acetylglucosaminyltransferases. 5

<b>Gln (Q)</b>	Glutamine. 23
<b>Glu (E)</b>	Glutamic acid. 27
<b>Gly (G)</b>	Glycine. 23
<b>GPI</b>	Glycosylphosphatidylinositol. 4
<b>HAE</b>	Haemagglutinin. 10
<b>His (H)</b>	Histidine. 24
<b>Ile (I)</b>	Isoleucine. 38
<b>IUBMB</b>	International Union of Biochemistry and Molecular Biology. 2
<b>IUPAC</b>	International Union of Pure and Applied Chemistry. 2
<b>L-Fucp</b>	L-Fucopyranose. 2
<b>Leu (L)</b>	Leucine. 38
<b>LNPs</b>	Lectin nucleotide phosphohydrolases. 8
<b>Lys (K)</b>	Lysine. 27
<b>MBL</b>	Mannose Binding Lectins. 8
<b>Met (M)</b>	Methionine. 38
<b>NA</b>	Neuraminidase. 10
<b>NMR</b>	Nuclear Magnetic Resonance. 10
<b>PAMP</b>	Pathogen-associated molecular patterns. 8
<b>PCR</b>	Polymerase Chain Reaction. 3
<b>PDB</b>	Protein Data Bank. 1
<b>Phe (F)</b>	Phenylalanine. 23
<b>PHP</b>	PHP Hypertext Preprocessor. 17
<b>Pro (P)</b>	Proline. 5
<b>PRR</b>	Pattern Recognition Receptors. 8
<b>Ser (S)</b>	Serine. 29

<b>Thr (T)</b>	Threonine. 29
<b>Trp (W)</b>	Tryptophan. 20
<b>Tyr (Y)</b>	Tyrosine. 20
<b>Val (V)</b>	Valine. 38
<b>XML</b>	Extensible Markup Language. 13

# List of Figures

1.1	The disaccharides Maltose and Gentobiose . . . . .	4
1.2	The synthesis of glycan structures . . . . .	6
1.3	Carbohydrate Recognition Domains (CRD) in the major lectin families . . . . .	7
1.4	Migration of leukocytes to sites of inflammation . . . . .	9
1.5	Representation of a generic influenza virus . . . . .	11
2.1	Excerpt from a PDB entry (4M18) . . . . .	14
2.2	Pathway for data collection in GlyVicinity . . . . .	14
3.1	The GlyVicinity database . . . . .	19
3.2	Amino acid abundance in the nrdb database, the PDB and the GlyVicinity database . . . . .	20
3.3	Amino acids found in the vicinity of all glycans retrieved from GlyVicinity . . . . .	22
3.4	Amino acid frequency at decreasing distances from glycan residues . . . . .	23
3.5	Amino acids in the vicinity of $\alpha$ -D-Manp at different cluster levels . . . . .	25
3.6	Comparison between clustered and unclustered datasets . . . . .	28
3.7	Pattern of atomic interactions for each monosaccharide analyzed . . . . .	33
3.8	Interactions at atomic level between amino acids and certain glycan residues . . . . .	36
3.9	Interactions between amino acids and Trp at atomic level . . . . .	37
3.10	N-glycan core as shared by different types of N-glycans . . . . .	38
3.11	Amino acid distribution around the carbohydrate residues that conform the N-glycan core . . . . .	39
3.12	Distribution of atoms around the $\beta$ -GlcNAc residues in the first and second position in N-glycan core . . . . .	41
3.13	N-glycan sequences analyzed with GlyVicinity . . . . .	42
3.14	N-glycan sequences analyzed with GlyVicinity. Detailed . . . . .	43
4.1	Examples of CH- $\pi$ interactions from the sugar point of view . . . . .	53
4.2	The spatial position of Fucose in reference to certain glycosylation sites . . . . .	57
4.3	Examples of platforms that Trp provides to glycan residues . . . . .	58

# List of Tables

1.1	Carbohydrate names, abbreviations and chemical configurations . . . . .	3
1.2	Well established lectin families . . . . .	8
2.1	List of common errors found in PDB entries . . . . .	15
2.2	Search space for the clustered datasets in the GlyVicinity Database . . . .	16
3.1	Amino acid abundances of the nrdb database, the PDB and the GlyVicinity database . . . . .	21
3.2	Search space and distance threshold . . . . .	22
3.3	Distance thresholds around $\alpha$ -D-Manp . . . . .	26
3.4	Difference in search space between clustered and unclustered datasets for different carbohydrate residues. . . . .	45
3.5	Results obtained from the analysis of N-glycan residues . . . . .	46



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.1.1	Glycosylation . . . . .	4
1.1.2	Protein-Carbohydrate Interactions . . . . .	5
1.2	Rationale of the study . . . . .	10
1.2.1	Objectives . . . . .	12
<b>2</b>	<b>Materials and Methods</b>	<b>13</b>
2.1	Data Source . . . . .	13
2.2	Sampling method . . . . .	14
2.3	Redundancy . . . . .	15
2.4	Methods . . . . .	16
<b>3</b>	<b>Results</b>	<b>18</b>
3.1	Parameter and concept definition . . . . .	18
3.2	Redundancy . . . . .	27
3.3	Carbohydrate chemistry and interaction with proteins . . . . .	27
3.4	Aromatic amino acids in protein-carbohydrate interactions . . . . .	35
3.5	Structural patterns and glycosylation . . . . .	35
3.5.1	Atom distribution . . . . .	39
3.5.2	Amino acid sequences around glycosylation sites . . . . .	42
<b>4</b>	<b>Discussion</b>	<b>47</b>
4.1	Utility of the PDB as source of information on carbohydrates . . . . .	48
4.2	Carbohydrate chemistry and interaction with proteins . . . . .	49
4.3	Over-representation of aromatic amino acids in protein-carbohydrate interactions . . . . .	52
4.4	Structural patterns and glycosylation . . . . .	54
4.4.1	Peptide sequences . . . . .	54
4.4.2	Spatial vicinity . . . . .	55

4.5	Limitations . . . . .	58
4.6	Outlook . . . . .	60
4.7	Conclusions . . . . .	61
<b>5</b>	<b>References</b>	<b>63</b>
<b>6</b>	<b>Appendix</b>	<b>71</b>

# 1

## Introduction

CARBOHYDRATES ARE VERY ABUNDANT and diverse biomolecules. Along with amino acids, nucleotides and lipids, carbohydrates are one of the basic components of cells. In the form of glycoproteins, they constitute an important amount of the mass and structural variation in biological systems [VCE<sup>+</sup>09]. Their role as energy storage and protective elements is well known, however, it is until recent years that their function as biological information carriers is being explored.

The biological information encoded on carbohydrate structures is recognized by carbohydrate-binding proteins (e.g. lectins). Despite the high chemical similarity between carbohydrates, lectins display a very remarkable ability to recognize and bind to only determined carbohydrate structures. In this project, the presence and arrangement of amino acids in the spatial vicinity of essential monosaccharides is statistically analyzed in order to gain a better understanding of the influence that carbohydrate stereochemical conformations have on the way these residues establish interactions with proteins. This approach does not employ experimental techniques. Instead, Three dimensional (3D) structural data on protein-carbohydrate complexes are analyzed. These input data are freely available in the Protein Data Bank (PDB).

### 1.1 Background

Carbohydrates are organic compounds that follow the chemical formula  $C_x(H_2O)_n$ . The simplest units of carbohydrates, monosaccharides, can exist as open chains but they are most commonly found in the form of a ring. This ring is in most cases the product of the reaction between the hydroxyl group of C5 with the aldehyde of C1.

Monosaccharide names are abbreviated to a 3-letter code based on the nomenclature established by the International Union of Pure and Applied Chemistry (IUPAC) - International Union of Biochemistry and Molecular Biology (IUBMB). E.g. in this format Glucose is denoted as Glc, and Mannose as Man. Carbohydrate residues can exist in two forms, called isomers, which are like mirror images from each other. A prefix 'D' or 'L' is added to the carbohydrate name to indicate its isomeric form (being D the most common). Furthermore, a suffix 'p' or 'f' can be appended to the name in order to indicate its chemical structure as a pyranose or furanose. The former identifies carbohydrates whose chemical structure includes a 6-membered ring. The conformation of furanoses, on the other hand, consist of a 5-membered ring. Table 1.1 shows some common carbohydrate residues, their respective IUPAC-IUBMB nomenclature and their standard chemical representation.

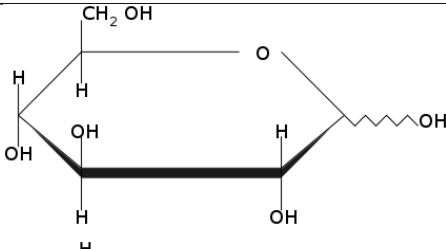
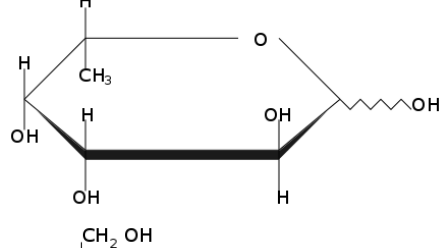
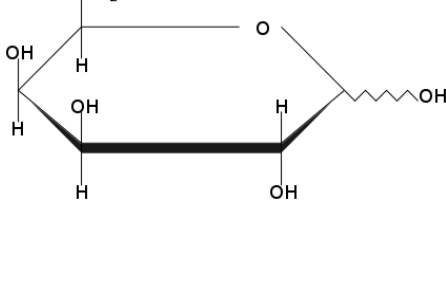
Monosaccharide and IUPAC format	Chemical representation
D-Glucopyranose (D-Glcp)	
L-Fucopyranose (L-Fucp)	
D-Galactopyranose (D-Galp)	
Continued on next page	

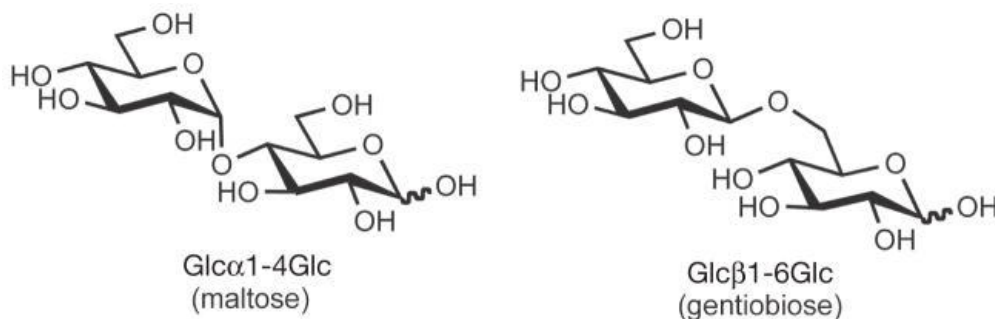
Table 1.1 – continued from previous page

Monosaccharide and IUPAC format	Chemical representation
D-Mannopyranose (D-Manp)	
N-acetyl-D-glucosamine (D-GlcNAc)	
N-Acetyl-D-neuraminic acid (D-Neu5Ac)	

**Table 1.1:** Carbohydrate names are abbreviated according to the norms established by the IUPAC-IUBMB [MW97]. Monosaccharides can exist as open chains but they are most commonly found in the form of a ring. For aldohexoses, the ring is the result of the reaction between the hydroxyl group of C5 with the aldehyde of C1, creating an hemiacetal as result. This configuration (a 5-carbon ring and an oxygen) is called pyranose and denoted with the the suffix 'p' concatenated to the carbohydrate name. A furanose, on the other hand, is represented by an 'f'

Unlike Deoxyribonucleic acid (DNA) or proteins, the biosynthesis of carbohydrates cannot be predicted from a DNA template. Carbohydrates are built by enzymes that link monosaccharides in a chain (glycosyltransferases), and enzymes that remove specific residues (Glycoside hydrolases) in a non-template-driven approach. This kind of synthesis prevents the development of techniques to amplify complex carbohydrates, such as the Polymerase Chain Reaction (PCR) for DNA. Instead, carbohydrates are studied in physiological amounts or by directed chemical synthesis.

Two monosaccharide residues can be linked in different ways into one of many possible isomers, for instance maltose and gentiobiose differ only by the linkage between D-Glcp residues (Figure 1.1) [VCE<sup>+</sup>09]. Moreover, the hydroxyl groups in single monosaccharides can serve as anchors through which they can establish linkages with other monosaccharides [VCE<sup>+</sup>09]. The complexity of carbohydrate chains is, therefore, greater than that of DNA or proteins since they are not only not sequentially assembled, but can also form branches. Accordingly, the type and order in which the monosaccharides are assembled, the anomeric configurations of the residues where glycosidic bonds are established, and the possibility of constituent monosaccharides to act as branching points contribute to a theoretically enormous number of glycan combinations. Furthermore, the complexity increases when modifications to glycan structures, such as acetylation, methylation, phosphorylation and sulfation are considered.



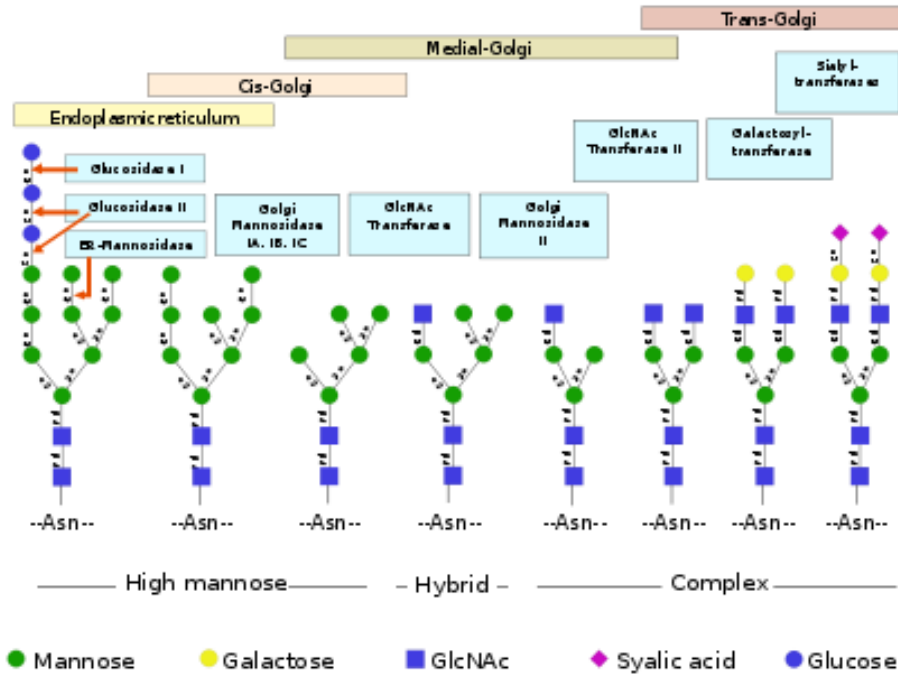
**Figure 1.1:** Maltose and gentiobiose are both disaccharides formed from the condensation of two glucose residues. They are isomer structures differentiated only by the glycosidic linkage,  $\alpha(1\rightarrow4)$  for maltose and  $\beta(1\rightarrow6)$  for gentiobiose. Taken from [VCE<sup>+</sup>09]

### 1.1.1 Glycosylation

Chains of monosaccharides are termed oligosaccharides, with the exception of chains whose structure consists of repetitive structural elements known as polysaccharides. Oligosaccharides covalently attached to proteins give rise to glycoproteins. The binding of sugar residues to proteins or lipids to produce glycoproteins or glycolipids, respectively, is known as glycosylation. Glycosylation occurs in practically all living organisms [CBRR12], [Spi02], and it is considered the most common and highly diverse of all co-translational modifications that proteins undergo in the living cell.

N-linked glycosylation is the most extended in prokaryotes, however, there are other types: O-linked, Glycosylphosphatidylinositol (GPI) anchors, Phospho-serin and C-mannosylation. Glycosylation plays an important role in thermodynamic stability and regulation of protein folding, so that minor changes in the glycan may affect protein





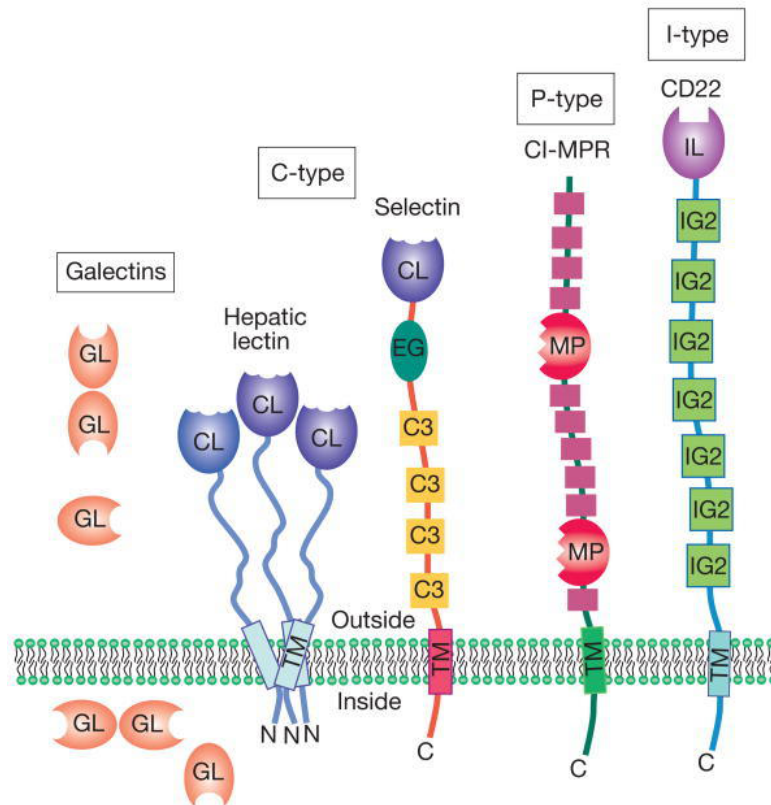
**Figure 1.2:** Adapted from [vdLLE09]. The synthesis of glycan structures is a non template driven process, exposed to multiple enzymatic pathways. The process starts on the cytosolic surface of the ER membrane with the addition, one by one, of glycan residues to form the N-glycan core. The process continues co-translationally in the ER by the transfer of the pre-assembled blocks of 14 sugars (2 N-acetylglucosamines, 9 mannoses and 3 glucoses) to a Nitrogen atom of an asparagine side chain by the oligosaccharyltransferase enzymes. Mature glycoproteins display very diverse glycan structures, however, until this step the carbohydrate moieties are homogenously assembled. The nascent glycoproteins are subsequently transferred to the medial stacks of the Golgi apparatus. Of the pre-assembled N-glycan core, only five residues remain. Here, further modifications are introduced to form more complex structures in a process called terminal glycosylation. As a final product, three types of N-linked glycoproteins are produced: complex, high-mannose and hybrid, which are then transported from the Golgi apparatus to the cell surface or other parts of the cell [TAP12].

cal information.

Coding information on the surfaces of cells implies the existence of elements able to decode this information. One of these decoder devices are the lectins, a group of carbohydrate-binding proteins of non-immune origin [dSC14]. Accordingly, protein-carbohydrate interactions are established through the precise match between complementary molecules, the information carriers (glycans) and the decoders (lectins).



Lectins can occur in all kingdoms of life. They consist of a group of very diverse proteins that belong to unrelated protein families, as reflected by their presence in different biological contexts. The domain that directly recognizes and interacts with carbohydrate residues in lectins is called Carbohydrate Recognition Domain (CRD) (See Figure 1.3). Most of CRDs have highly conserved primary and/or tertiary structures that, besides the implication of a shared ancestor, allows their classification in different families (See Table 1.2). Members of a lectin family, however, do not necessarily share affinity for the same glycan residues.



**Figure 1.3:** CRDs mediate the interactions of lectins with glycan residues. Based on conserved sequence motifs or/and tertiary structures of their CDRs, lectins have been classified in families. The figure, adapted from [VCE<sup>+</sup>09], shows schematic examples of CRDs in the major lectin families: CL stands for C-type lectin CRD, GL for galectin CRD, MP for P-type lectin CRD, and IL for I-type lectin CRD. Other domains are EG for EGF-like domain; IG2 for immunoglobulin C2-set domain; TM for transmembrane region; and C3 for complement regulatory repeat.

The glycocalyx confers cells with a distinctive identity. Lectins make use of this singularity to establish interactions with specific glycan residues. Below three examples are briefly described in order to demonstrate the importance of this binding event, as well as

the broad range of essential biological events in which protein-carbohydrates interactions are involved: the activation of the human immune complement system, the migration of leukocytes to sites of inflammation, and the attachment of pathogens and toxins to carbohydrate residues as the first stage in the cycle of infection.

<b>C-type</b>	E.g., calcium-dependent lectins such as selectins, collectins, etc.
<b>Galectins</b>	Formerly S-type lectins.
<b>I-type</b>	Immunoglobulin superfamily members, including the Siglec family.
<b>L-type</b>	Plant legume seed lectins, ERGIC-53 in ER-Golgi pathway, calnexin family.
<b>M-type</b>	$\alpha$ -mannosidase-related lectins (e.g. EDEM).
<b>N-type</b>	Lectin nucleotide phosphohydrolases (LNPs) with glycan-binding and apyrase domains.
<b>P-type</b>	E.g., mannose-6-phosphate receptors.
<b>R-type</b>	E.g., ricin, other plant lectins, GalNAc-SO <sub>4</sub> receptors.

**Table 1.2:** Non-exhaustive list of well established lectin families. This classification is based on the sequence or structural evolutionary similarity they share. Adapted from [VCE<sup>+</sup>09]

## Human immune system

Innate immunity is mediated mostly by phagocytic cells, such as macrophages, dendritic cells and neutrophils, that engulf extracellular pathogens, or virally infected host cells, and initiate adaptive immunity through the display of pathogen-associated proteins to T cells in the lymph nodes. Innate immunity serves as the first line of defense against pathogens. The adaptive immunity becomes functional afterwards, through the production of specific antibodies and cytotoxic T cells [KL12].

One of the main components of the innate immunity system are the Pattern Recognition Receptors (PRR). PRRs trigger host response by recognizing well conserved Pathogen-associated molecular patterns (PAMP) that are common to pathogens (mainly bacteria, but also fungi, viruses, and potential allergens) during the first moments after exposure [HR00], [KL12]. They are also able to identify damaged cells, for example through the change in glycosylation patterns they present on their surfaces [ITES09].

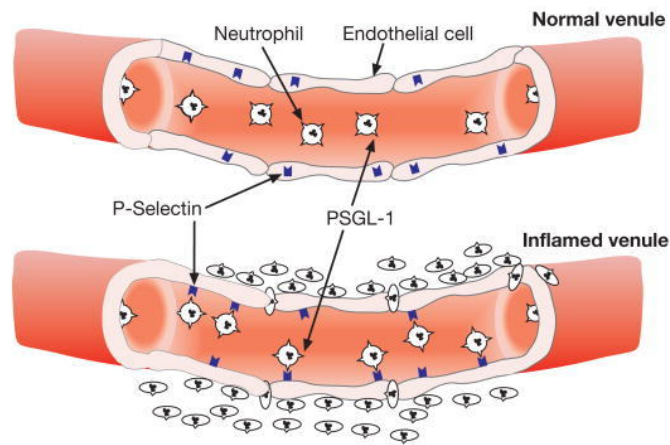
As part of the innate immune system, Mannose Binding Lectins (MBL) circulate in serum and in the liver. They carry multiple Carbohydrate Recognition Domains that, despite its possibly misleading name, are able to distinguish not only D-Mannose but also L-Fucose, and N-acetyl-D-Glucosamine residues located on the surface of a wide range of pathogens from mammalian glycans, for their identification and posterior neu-

tralization [ITES09], [HSFWH12].

MBLs are essential for the activation of the complement system via the lectin pathway [HSFWH12]. The importance of MBL becomes more noticeable when the ability to develop adaptive immune responses is compromised (e.g. AIDS patients) and the body must rely on the innate response. Moreover, recurrence of serious infections in children has been associated to mutations in the collagen domains of MBLs, especially between the 6 months and 2 years of age. It is during this age interval that the immunity inherited through the antibodies located in the maternal milk and placenta recedes [AK09].

### Migration of leukocytes to sites of inflammation

The binding of the P-members of the selectin family to their respective leukocyte receptors as the initial step in the event of inflammation is essential for the innate and adaptive immune system. In normal conditions, leukocytes flow freely in the blood stream without interacting with the endothelium. However, during the event of inflammation, selectins are expressed on endothelium surfaces. The binding of leukocytes leads to their adhesion to the vessel wall where, after a process of rolling, they trespass the endothelial cell barrier and reach the extravascular space into the site of injury (Figure 1.4) [VCE<sup>+</sup>09], [KU06], [MC06].



**Figure 1.4:** Migration of leukocytes to sites of inflammation. Adapted from [VCE<sup>+</sup>09].

### Attachment of pathogens and toxins as the first stage in the cycle of infection

Certain viruses and toxins make use of the glycans in the glycocalyx as an entrance door into the cell. The first isolated lectin of this type, and one of the most thoroughly studied, is the influenza virus hemagglutinin [VCE<sup>+</sup>09].

Influenza belongs to the orthomyxoviridae family, which includes three serotypes: A, B and C. A and B types are responsible for periodic global epidemics. These serotypes, A and B, display two carbohydrate-binding proteins on their surfaces, the Haemagglutinin (HAE) and Neuraminidase (NA) (See Figure 1.5). The function of these proteins is fused into only one, the haemagglutinin-esterase-fusion protein, in the C type.

The attachment of the virion to the cell surface is performed by the HA homotrimer. This occurs through the recognition of target cells with sialic acid moieties on their surfaces. The HA then attaches itself and after certain conformational changes fuses to the cell membrane. Afterwards, the virus RNA is released into the cytosol of the host cell. Human influenza preferably binds to sialic acid  $\alpha$ -(2->6)-linked to galactose. Avian influenza binds to sialic acid  $\alpha$ -(2->3)-linked to galactose and swine influenza viruses are able to recognize both receptors.

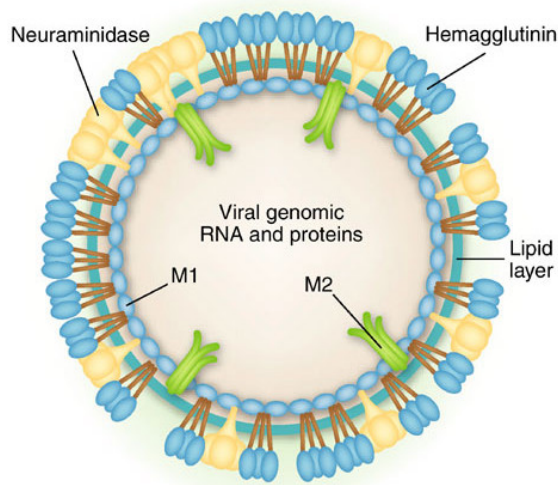
NA enzymatically cleaves the sialic acid groups from hosting cells and viral proteins. Therefore, it is responsible for the releasing of newly produced viruses from the host in order to infect other cells and for preventing the binding of other viruses [MH-BvI09]. This is a crucial point in the influenza life-cycle. Consequently, the identification of NA inhibitors is an important strategy in the fight against the influenza virus [Rob01].

Some toxins follow a similar mechanism to infect cells. The structure of the toxins of *Vibrio cholerae*, Enterohemorrhagic *Escherichia Coli* (EHEC) and its homologous, for example, are composed of an A-unit associated with a pentamer B-unit. Through CRDs receptors located on the base of its non-toxic B units, these kind of toxins bind to glycans on the surface of cells to facilitate the entrance of the toxic unit-A. This last part of the process is not yet well known [VCE<sup>+</sup>09].

## 1.2 Rationale of the study

X-ray crystallography and Nuclear Magnetic Resonance (NMR) can be considered the leading techniques used to produce detailed information about protein-carbohydrate interactions at atomic level [PN03], [Jef90], [dCFADB<sup>+</sup>12], [JP05], [Hal94]. In the PDB, the largest source of biomolecular 3D structures publicly available [BWF<sup>+</sup>00], around 7% of the total entries are glycoproteins and protein-carbohydrate complexes determined mainly through these methods. However, X-ray crystallography and NMR experiments are expensive, laborious and, unfortunately, these techniques cannot be universally applied, e.g. not all proteins can be crystallized or be purified in enough quantity for NMR experiments.

In the year of 2009 Lütteke and von der Lieth [vdLLE09b] published a bioinformatics study of protein-carbohydrate interactions where no experimental methods were employed. Instead, conclusions about the nature and traits of binding specificity in protein-



**Figure 1.5:** Influenza is a highly contagious virus, responsible for millions of deaths every year. The serotypes A and B display two carbohydrate-binding proteins on their surfaces, the haemagglutinin and the neuraminidase. The former attaches itself to sialic acids on cell surfaces whereas the latter modifies such glycan residues, preventing the attachment of new viruses and ensuring the release of newly produce influenza viruses. Adapted from [WP09]

carbohydrate interactions were drawn based on the statistic analysis of 3D structures of protein-carbohydrate complexes already available in the PDB. An important part of the study was the development of a software tool called GlyVicinity. Through the use of this application, the spatial vicinity of certain glycan residues( $\beta$ -D-Galp, D-GlcpNAc,  $\alpha$ -D-Neup5Ac, sulfated residues) was analyzed.

The main conclusions drawn by Lütteke and von der Lieth included the variability in the distribution of amino acids among the analyzed residues and the over-representation of polar amino acids. The importance of polar amino acids around glycan residues is not surprising due to their hydrophilic nature. However, a prominent presence of the aromatic amino acids, especially Trp around the carbohydrates analyzed (in lesser extent with the sulfated residues) was also found. The distribution of interactions at atomic level suggested the establishment of stacking interactions.

Studies on the affinity between sugar residues and lectins have been reported. For example, Tyagi *et al* [TYKUGO] analyzed the binding affinities between D-Glucose and the sodium-glucose cotransporters (SGLTs). Kumari *et al* [KBS11] used quantum chemical calculations to examine the binding affinities of 3-methylindole and D-glucose,  $\beta$ -D-Galactose,  $\alpha$ -D-Mannose and L-fucose. Rao *et al* [RLQ98] docked D-Galactose, D-GalpNAc and Mannose to the binding sites of 4 different lectins in order to study their interactions. The complex formed between  $\beta$ -galactosidases and galactose has been analyzed by Maksimainen *et al* [MHK<sup>+</sup>11], etc. However, the difference between the above mentioned -and similar studies- and our approach is that they examine couples of

protein-sugar complexes, whereas our analysis collects all protein-sugar complexes that have been made available to the scientific community in the PDB website. Accordingly, the observed patterns of interactions are the result of multiple observations and can be extrapolated to other cases.

The present thesis rests on the work by Lütteke and von der Lieth. Since the original study was published, the number of carbohydrate-containing PDB entries has increased considerably, which open the possibility of examining more monosaccharides. To the best of my knowledge, this is the study with the largest number of carbohydrate residues to be analyzed and whose binding affinities are compared. Furthermore, the analysis of covalently linked glycans has also been included, in contrast to the original study that was limited to only non-covalently linked ligands. The possibility of clustering the input datasets was also implemented as a means of decreasing redundancy of data and consequently, increase the reliability of the results.

### 1.2.1 Objectives

This project is supported by the following objectives:

1. Firstly, it aims at testing the utility of the PDB as an indirect source of information on carbohydrates as well as the extension in which redundancy affects the input datasets.
2. Carbohydrates share a very similar structure based on the chemical formula  $C_x(H_2O)_n$ . However, carbohydrate-binding proteins are able to recognize and bind only to certain residues. This work seeks to determine the influence that carbohydrate chemistry (linkages, position in the glycan chain, anomers and other chemical modifications e.g. acetylation) have on the way monosaccharides interact with proteins.
3. Carbohydrates are hydrophilic molecules. Therefore, interactions between polar amino acids and glycan residues are expected. However, interactions between monosaccharides and non-polar amino acids have been previously reported. This work tries to find the reason behind the large presence of aromatic amino acids in protein-carbohydrate interactions, as well as the extent in which these interactions occur among different glycan residues.
4. At least 50% of all proteins are glycosylated. The sequon Asn-xx-Threonine/Serine (where xx is any amino acid except for Pro) indicates potential glycosylation sites. However, not all sequons are glycosylated and it is not clear why. The present study seeks to find a possible structural pattern that indicates glycosylation through the analysis of 3D data on N-glycans and the comparison of interactions to non-covalently linked ligands.

# 2

## Materials and Methods

### 2.1 Data Source

The Protein Data Bank is the central repository of protein 3D structures determined mainly by NMR or crystallography. To date (November 2014), the PDB contains about 105,025 entries. The number of entries for pure carbohydrates is very small (about 20 entries). However, approximately 7% of the total entries contain carbohydrates (protein-carbohydrate complexes, glycoproteins, etc) [Lüt09].

The PDB implements a standard notation to describe 3D structures of proteins (see Figure 2.1). For example, all PDB files must contain a label `HEADER`, which reports a classification of the protein, the date of deposition, and a unique identification code. Another label is `ATOM`, that describes the spatial 3D coordinates of the atoms that constitute the protein. The label `HETATM` identifies the atomic coordinates of non standard residues, such as carbohydrates. Nevertheless, other types of residues are also designated with this label (e.g. inhibitors, solvent and ligands), which complicates the search for carbohydrate-containing entries. Additionally, in many cases carbohydrates do not receive as much attention as the protein when a structure is described. This may give rise to missing (e.g. lack of glycosidic linkages) or erroneous data (e.g. wrong stereochemistry). Furthermore, the PDB does not force to comply with the standard nomenclature to identify carbohydrates. E.g. the monosaccharide  $\beta$ -D-Mannose should be referred to as BMA. However, it is often wrongly identified as MAN in the PDB, which in turn stands for  $\alpha$ -D-Mannose.

The program `pdb2linucs` automatically locates and extracts newly released PDB entries that contain carbohydrate residues without relying only on PDB annotations. The resulting entries are then stored in a Extensible Markup Language (XML) file. This XML file is parsed and used as input source for the GlyVicinity database (See Figure

```

HEADER      SUGAR BINDING PROTEIN                                02-AUG-13   4M18
TITLE       CRYSTAL STRUCTURE OF SURFACTANT PROTEIN-D D325A/R343V MUTANT IN
TITLE       2 COMPLEX WITH TRIMANNULOSE (MAN-A1,2MAN-A1,2MAN)
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: PULMONARY SURFACTANT-ASSOCIATED PROTEIN D;
COMPND      3 CHAIN: A, B, C, D, E, F, G, H, I, J, K, L;
COMPND      4 FRAGMENT: NECK AND CARBOHYDRATE RECOGNITION DOMAIN (UNP RESIDUES 229-
COMPND      5 375);
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE      3 ORGANISM_COMMON: HUMAN;
SOURCE      4 ORGANISM_TAXID: 9606;
AUTHOR      B. C. GOH, M. J. RYNKIEWICZ, T. R. CAFARELLA, M. R. WHITE, K. L. HARTSHORN, K. ALLEN,
AUTHOR      2 E. C. CROUCH, O. CALIN, P. H. SEEBERGER, K. SCHULTEN, B. A. SEATON
REVDAT      2 11-DEC-13 4M18 1 JRNL
REVDAT      1 04-DEC-13 4M18 0
JRNL        AUTH B. C. GOH, M. J. RYNKIEWICZ, T. R. CAFARELLA, M. R. WHITE,
JRNL        AUTH 2 K. L. HARTSHORN, K. ALLEN, E. C. CROUCH, O. CALIN, P. H. SEEBERGER,
JRNL        AUTH 3 K. SCHULTEN, B. A. SEATON

ATOM 13056 CD2 PHE L 355 -71.928 -9.184 -47.655 1.00 31.97 C
ATOM 13057 CE1 PHE L 355 -69.667 -9.733 -46.167 1.00 22.50 C
ATOM 13058 CE2 PHE L 355 -72.050 -9.623 -46.344 1.00 34.73 C
ATOM 13059 CZ PHE L 355 -70.917 -9.897 -45.600 1.00 26.07 C
ATOM 13060 OXT PHE L 355 -71.001 -5.882 -51.825 1.00 31.19 O
TER 13061 PHE L 355
HETATM13062 CA CA A 401 -28.889 -6.624 -50.710 1.00 49.63 CA2+
HETATM13063 CA CA A 402 -20.729 -8.711 -51.221 1.00 49.50 CA2+
HETATM13064 CA CA A 403 -17.817 -8.468 -53.890 1.00 47.52 CA2+
HETATM13065 CA CA A 404 -22.308 21.449 -33.285 1.00102.70 CA2+
HETATM13066 C1 MAN A 405 -33.767 -5.699 -56.933 1.00 98.20 C
HETATM13067 C2 MAN A 405 -32.480 -4.887 -56.818 1.00 98.45 C
HETATM13068 C3 MAN A 405 -31.557 -5.264 -57.974 1.00 88.31 C
HETATM13069 C4 MAN A 405 -31.309 -6.767 -57.951 1.00 94.16 C

```

**Figure 2.1:** Excerpt from a PDB entry (4M18). Carbohydrates should be included in the HETATM section. Each line identifies an atom, and describes residue and atom properties: atom serial number, atom name, residue (MAN=  $\alpha$ -D-Manp), the chain identifier (A), the residue sequence number (405), the x, y and z coordinates to describe the positions of atoms in space, occupancy (1.00), temperature factor, segment identifier, element symbol and charge of the atom.

2.2). pdb2linucs runs weekly. The GlyVicinity database can then be weekly updated by an automatic scanning of the XML files generated by pdb2linucs. In this way, the data is kept up to date at the same rate as the PDB itself.



**Figure 2.2:** Pathway for data collection in GlyVicinity

## 2.2 Sampling method

In order to identify problems and ensure the quality of the data in the GlyVicinity database, an additional attribute is assigned to each entry by the program pdb2linucs [BWF<sup>+</sup>00]. By discriminating entries based on this quality attribute, GlyVicinity avoids common mistakes found in PDB entries, increasing in this way the reliability of the data. Following, an example of how this process is implemented:



A binary vector is used for each carbohydrate residue in every PDB entry. For example:

```
00000000001  residue 1
00000001001  residue 2
```

Each position in the vector identifies a specific problem, as shown in Table 2.1 [BWF<sup>+</sup>00]. The final values in the binary vector are summed up. The resulting number identifies the PDB entries that contain errors, and of what type.

Position	Description of the problem
1	Positive control.
2	Wrong assignment of anomer.
3	PDB residue name and detected monosaccharide are inconsistent.
4	Residue name given in PDB file is unknown in the list of residues (wrong positive or new residue name: no validation possible).
5	Stereochemistry of the carbohydrate could not be assigned.
6	No glycosidic O -(S -, N -) atom at C1 could be detected.
7	Bond between the anomeric C-atom and adjacent O-atom (not part of a ring) was added based on distance criteria.
8	Monosaccharide is connected to a chain, which is neither contained in the list of residues nor in the list of substituents.
9	Dihedral angles derived to calculate the stereochemistry show unusual values.

**Table 2.1:** The program pdb2linucs assigns a quality attribute to evaluate PDB entries. GlyVicinity discriminates entries based on this attribute, hence, the results obtained with the program are more reliable. This is only an extract of the list as published in [LvdL06].

## 2.3 Redundancy

During the process of transferring data from the XML file to the database, the amino acid sequence of the protein is also collected from the PDB. These sequences are later used to cluster the data in the database at different sequence identity levels (70% to 100%) in order to decrease redundancy. The program CD-HIT [LG06] version 4.6.1 was employed for this purpose. The size of the resulting clusters are summarized in table 2.2. The subset of entries clustered at 80% protein sequence identity was selected as source of data for all the analyses in this work, due to the optimal compromise between

quantity and reliability.

Dataset	Total
Unclustered data	9,069 entries
clustered at 100% identity	4,949 clusters
clustered at 90% identity	3,192 clusters
clustered at 80% identity	2,946 clusters
clustered at 70% identity	2,744 clusters

**Table 2.2:** Search space for the clustered datasets in the GlyVicinity Database

## 2.4 Methods

1. The study of carbohydrate affinity was divided in two groups based on the way they bind to proteins: covalently and non-covalently. This division allows for a better understanding of the interactions established between proteins and: a) glycan ligands and b) N-glycans.
2. Experimental methods were not used for this work. Instead, a statistical analysis with the bioinformatics tool GlyVicinity was performed. GlyVicinity is a program and a database that explores the spatial vicinity of glycan residues. The application was completely redeveloped and further improved for this work.
3. The statistical analysis used in this study is based on two measures: absolute counts and deviations from natural abundances.
  - (a) Absolute counts indicates how many times an amino acid is found in the spatial vicinity of a glycan residue. A radius of 4 Ångstroms (Å) around glycan residues was selected because it is wide enough to include many amino acids without losing specificity.
  - (b) Amino acids do not occur in nature in the same frequency. Therefore, deviations from the natural occurrences are also calculated, in order to determine if the presence of an amino acid is different in the area interacting with the monosaccharide compared to an average protein.

Deviations from natural occurrences are calculated with the formula:

$$\frac{\frac{num_{aa}}{num_{total}} - nat_{aa}}{nat_{aa}} * 100$$

Where:

$\text{num}_{\text{aa}}$  represents the absolute counts of an amino acid in a dataset,  
 $\text{num}_{\text{total}}$  the total number of amino acids analyzed in a dataset

The proteinogenic amino acids contained in all of the PDB entries were summed up to calculate their natural abundance(nat aa). The same process was repeated for the GlyVicinity database and its subset of non-redundant entries. The resulting values were compared to the published values for the nrdb database. The goal was to ensure that the data sources were not biased. Fortunately, this was not the case, and the distribution of atoms among the different sources analyzed was very similar.

4. To complement the statistical analysis, the superimposition of protein atoms around monosaccharides was implemented in GlyVicinity, and displayed in Jmol (<http://jmol.sourceforge.net/>). In this way, the protein-carbohydrate interactions at atomic level can be visually analyzed in 3D.
5. Besides GlyVicinity, other small programs were developed in order to perform diverse tasks:
  - (a) The data transference from the xml files generated by pdb2linucs to the database in GlyVicinity is performed by a program written in C++.
  - (b) The calculation of amino acid abundance in the PDB, and the GlyVicinity database is performed by a program written in PHP Hypertext Preprocessor (PHP).

# 3

## Results

THE REDEVELOPMENT OF GLYVICINITY included the design and implementation of a database as a repository of data. Figure 3.1 display graphically the resulting tables, attributes and their relationships.

### 3.1 Parameter and concept definition

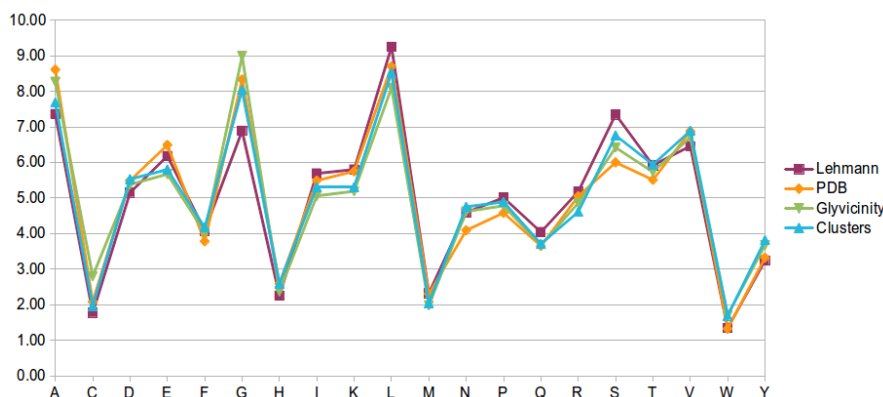
In order to statistically analyze the spatial vicinity of common monosaccharides, two important concepts must first be defined:

1. Absolute counts
2. Deviations from natural occurrences

Absolute counts are calculated simply by summing up the presence of each amino acid in the search space. However, these results should be considered carefully since they do not reflect the fact that not all proteinogenic amino acids are used in nature in the same proportion for the assembling of polypeptides.

Figure 3.2 shows a comparison of amino acid abundance calculated from different datasets (nrdb database [LBvdL00], the PDB and GlyVicinity). The sample size for each dataset is described in detail in table 3.1. It is clear that the differences among datasets are modest. On the other hand, the variation in frequency among the different amino acids is considerable. Deviations from these natural abundances indicate possible divergences in amino acid composition between the binding sites and the rest of the protein structure. Therefore, not only absolute numbers but also deviations for each case were calculated.





**Figure 3.2:** A comparison of amino acid abundance in the nrdb database, the PDB, the GlyVicinity database and the clustered database of GlyVicinity. The difference in frequency among amino acids is very notorious. On the other hand, the patterns of amino acid distribution among databases is modest.

Furthermore, two parameters are important for this type of analysis:

1. Radius around carbohydrate residues to be examined.
2. Level of redundancy to be tolerated.

The optimal values for these parameters were selected as follows:

Firstly, all protein-carbohydrate complexes determined at a resolution better or equal to 3 Ångströms were retrieved from the GlyVicinity database. Subsequently, a subset of all non-covalently bound amino acids found in a radius up to 10 Ångströms around carbohydrate residues were selected and analyzed with the GlyVicinity program. The results are depicted in Figure 3.3 ordered by polarity, ranging from polar (red) to unpolar (blue).

Visual inspection of Fig 3.3 reveals that absolute numbers vary notably (Fig 3.3 A1) but this is not the case for the respective deviations (Fig 3.3 A2). Except for Tyrosine (Tyr (Y)), Tryptophan (Trp (W)) and to certain extent Asn, whose positive deviations indicate a higher preference compared to the rest, most of amino acids occur near their expected distributions in nature. These results do not reflect the high level of specificity that carbohydrate-binding proteins possess. Thus, a threshold of 10 Ångströms around glycan residues is too large, and in consequence specificity is lost. The same analysis was performed with thresholds decreasing from 9 to 3 Ångströms in order to find the most suitable range for study (Fig 3.4 and Table 3.2). With every decrement of threshold value, the results become more precise, which can be interpreted as increased specificity due to the enforced close proximity between amino acids and glycan residues. A threshold of 3 Ångströms (Fig 3.4 G2) stops refining the results to actually change notably the distribution pattern. For instance, the deviation of Trp drops notably even though its deviations were the highest in previous cases. A radius of 3 Ångströms seems to be

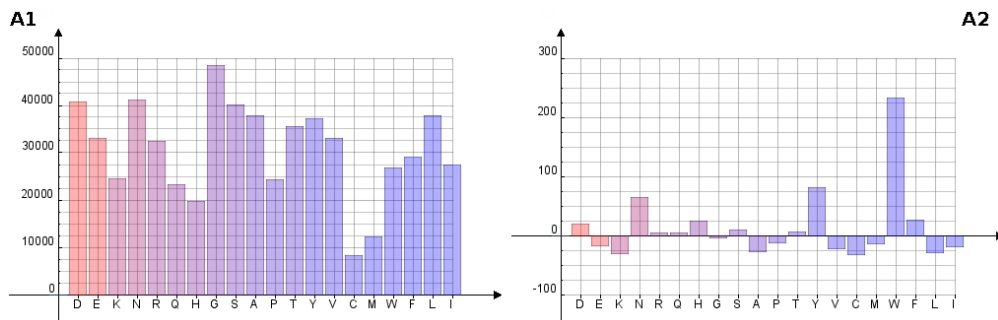
too narrow and some interactions are left out as result. A threshold of 4 Ångstroms is then considered as an optimal distance which is close enough to cover most of the interactions. This value is used for the rest of cases in this study.

Aa	Lehman	%	PDB	%	Glyvicinity	%	Cluster,	%
A	5,938,037	7.36	5,575,281	8.61	703,934	8.27	191,264	7.68
C	1,431,956	1.77	1,348,117	2.08	238,124	2.80	48,726	1.96
D	4,150,999	5.15	3,554,304	5.49	458,066	5.38	137,670	5.53
E	4,989,268	6.18	4,199,461	6.49	482,317	5.67	144,605	5.80
F	3,291,866	4.08	2,445,199	3.78	341,919	4.02	103,814	4.17
G	5,562,328	6.89	5,388,742	8.33	765,574	8.99	200,379	8.04
H	1,815,876	2.25	1,679,139	2.59	202,083	2.37	64,443	2.59
I	4,592,937	5.69	3,550,379	5.49	430,580	5.06	132,197	5.31
K	4,682,260	5.80	3,720,832	5.75	441,996	5.19	132,111	5.30
L	7,466,050	9.25	5,632,975	8.70	689,636	8.10	212,510	8.53
M	1,869,254	2.32	1,491,664	2.30	168,114	1.98	50,760	2.04
N	3,705,385	4.59	2,645,273	4.09	393,178	4.62	118,413	4.75
P	4,044,125	5.01	2,961,693	4.58	407,065	4.78	121,772	4.89
Q	3,259,036	4.04	2,359,538	3.65	312,404	3.67	92,500	3.71
R	4,181,883	5.18	3,266,148	5.05	413,235	4.86	114,868	4.61
S	5,930,098	7.35	3,884,297	6.00	546,096	6.42	168,441	6.76
T	4,777,231	5.92	3,563,672	5.51	487,719	5.73	147,996	5.94
V	5,207,912	6.46	4,451,033	6.88	571,467	6.71	171,707	6.89
W	1,081,628	1.34	851,176	1.32	145,323	1.71	42,057	1.69
Y	2,614,263	3.24	2,149,879	3.32	312,413	3.67	94,956	3.81

**Table 3.1:** Numerical comparison of Amino acid abundance in the nrdb database, the PDB, the GlyVicinity database, and GlyVicinity database clustered at 80% sequence identity. Values for nrdb were adapted from [LBvdL00]

The second parameter to consider is the level of redundancy to be tolerated in the analysis. In order to verify to what extent the data results are affected by redundancy, a new dataset was generated. In this case, all non-covalently bound amino acids present in a radius up to 4 Ångstroms around  $\alpha$ -D-Manp residues in the GlyVicinity database were retrieved. The resulting data formed an 'unclustered' dataset from which 'clus-

tered' datasets were derived through the use of CD-HIT at different identity levels (see Material and Methods section). The results of the analysis performed by GlyVicinity are shown in Figure 3.5 and Table 3.3. The scale is automatically adjusted to fit the values.

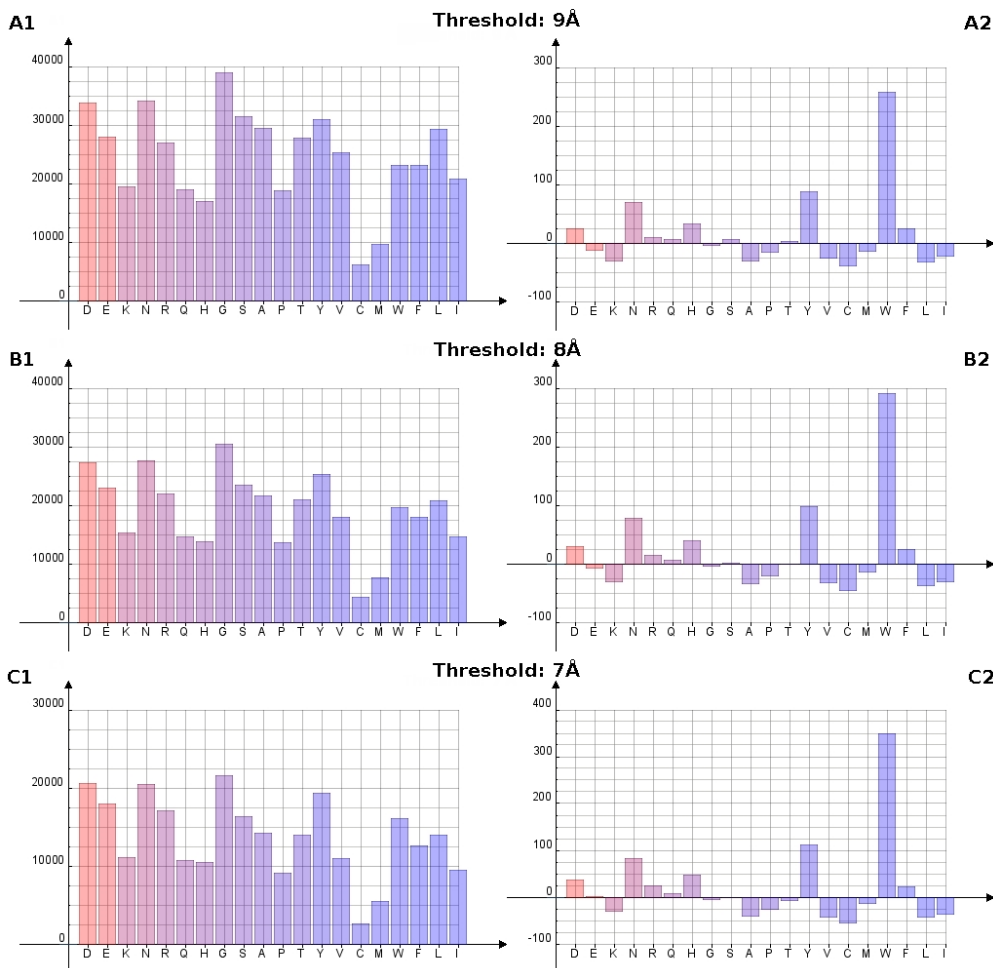


**Figure 3.3:** All proteinogenic amino acids in the database located up to 10 Ångstroms around glycan residues were selected and analyzed with GlyVicinity. The query retrieves 612,146 amino acids in the vicinity of 19,352 carbohydrate residues (11,400 chains, 4,700 PDB entries). The color indicates polarity (red= polar, blue= unpolar). A1 shows the absolute counts, or how many times each amino acid is present in the search space, and A2 indicates deviations from natural occurrences.

Threshold (Å)	Amino acids	Carboh. residues	Carboh. chains	PDB entries
9 Å	492,130	19,298	11,391	4,698
8 Å	380,877	19,225	11,368	4,695
7 Å	273,322	19,154	11,351	4,693
6 Å	192,597	18,958	11,285	4,682
5 Å	148,955	18,711	11,230	4,673
4 Å	107,763	18,197	11,118	4,652
3 Å	37,892	14,335	9,965	4,371

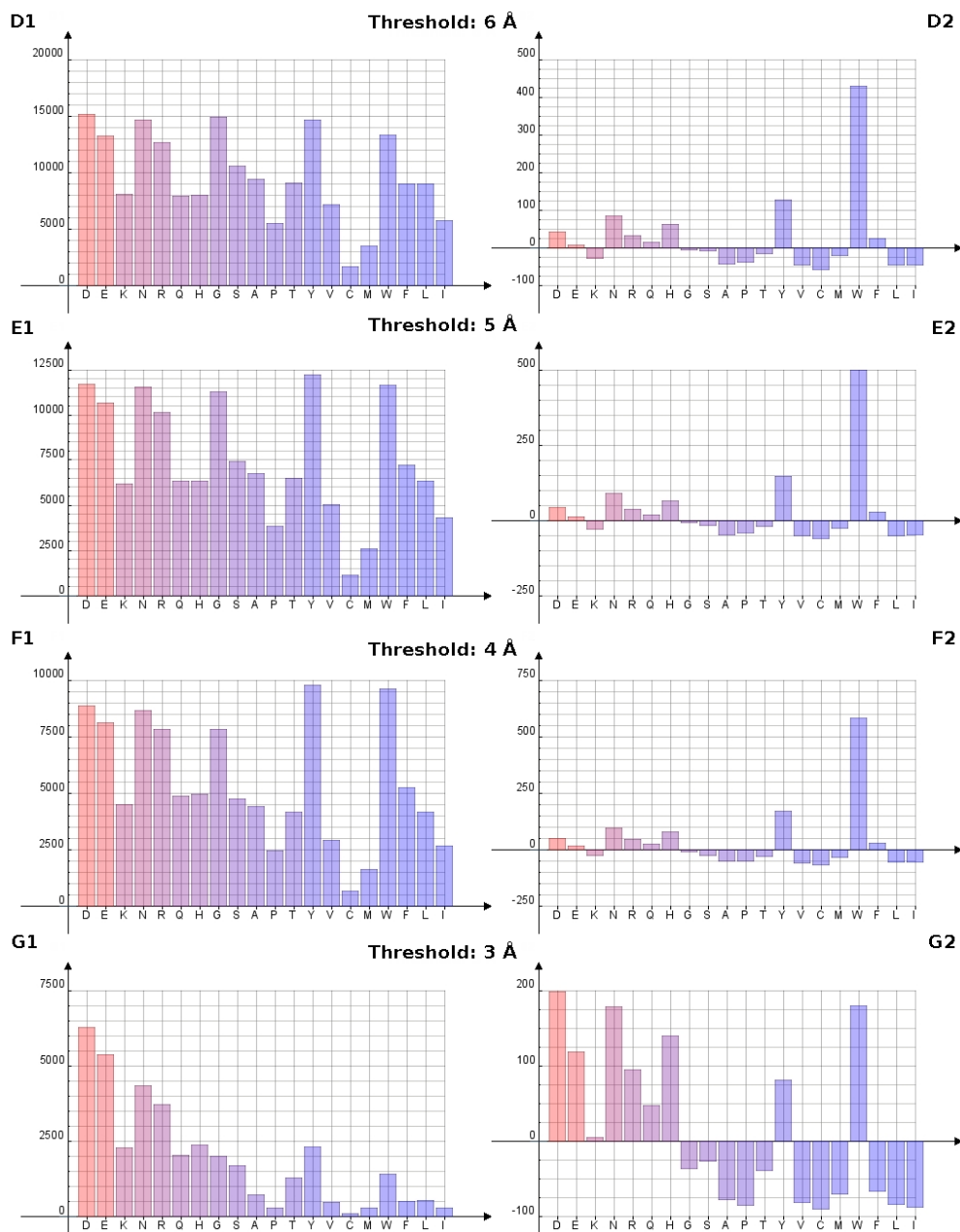
**Table 3.2:** In order to determine the radius around glycan residues most suitable to perform the analysis, a decreasing number of thresholds were examined. A radius of 4 Ångstroms is considered the best threshold since it is neither too small that many entries are discarded nor too unrestrictive that specificity gets lost. The search space becomes smaller with decreasing distance thresholds. For example, at a radius of up to 9 Ångstroms, 492,130 amino acids are found which interact with 19,298 glycan residues. These glycan residues are found in 11,391 glycan chains, and correspond to 4,698 PDB entries. These numbers decrease to 37,892 amino acids around 14,335 glycan residues found in 9,965 carbohydrate chains in 4371 PDB entries at a radius of 3 Ångstroms.





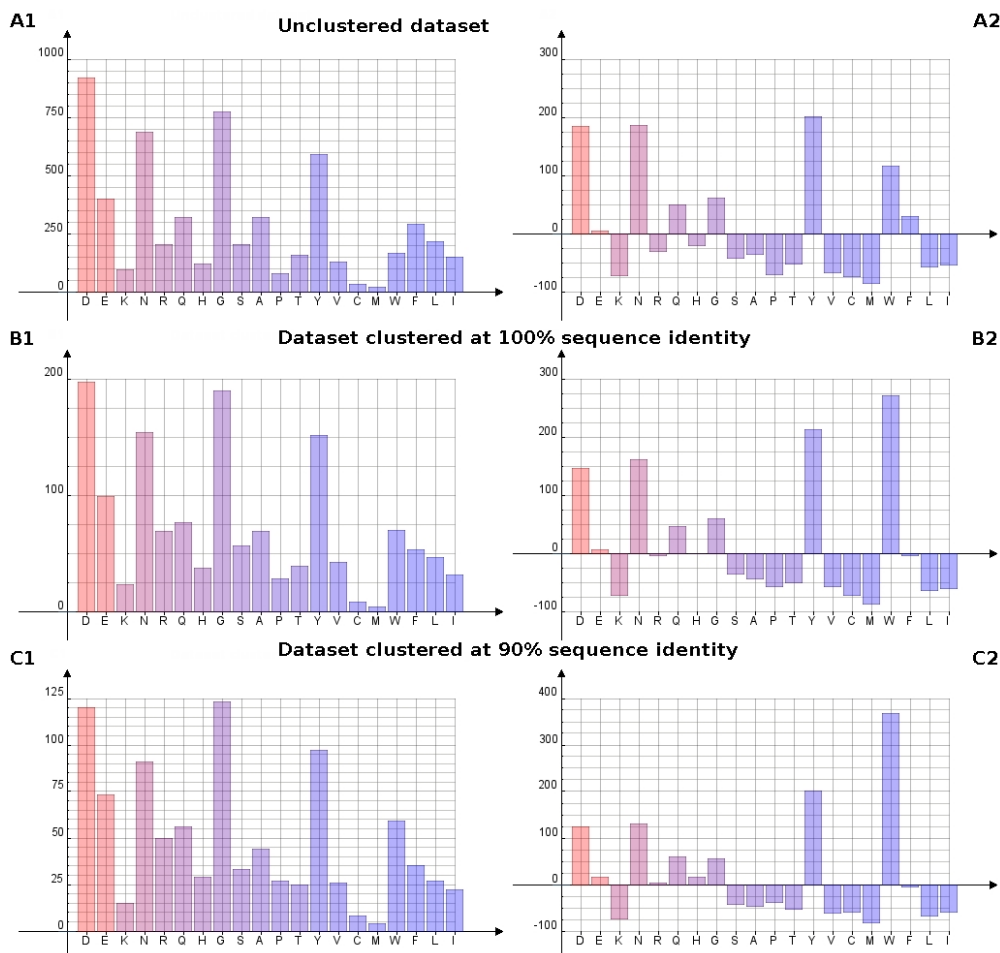
**Figure 3.4:** Comparison of the results generated by decreasing the threshold distance between amino acids and monosaccharides. The proportion of absolute counts does not show a big variation. This is not the case with the deviations, which become more refined as the values decrease. A threshold of 3 Å, however, changes notably the deviations. Apparently, this value is too small and many bonds are omitted. 4 Å is then selected as an optimal distance. Continued on next page.

The effect of redundancy on absolute counts is hardly visible for all examples with  $\alpha$ -D-Manp. This is not the case for deviations. Visual inspection showed that a) Aspartic acid (Asp (D)), Asn, Tyr and Trp have the largest positive deviations, b) Glutamine (Gln (Q)), Glycine (Gly (G)) and Phenylalanine (Phe (F)) are also over-represented and c) the rest of amino acids are under-represented (Figure 3.5). In general, the distribution is quite different from the deviations previously shown for all monosaccharides (Figure 3.3). Moreover, a difference is notable when the unclustered deviations (Figure 3.5 A2) are compared to those of the clustered dataset at 100% identity (Figure 3.5 B2). The divergence becomes increasingly accentuated through the rest of the clustered datasets.



**Figure 3.4:** Continuation

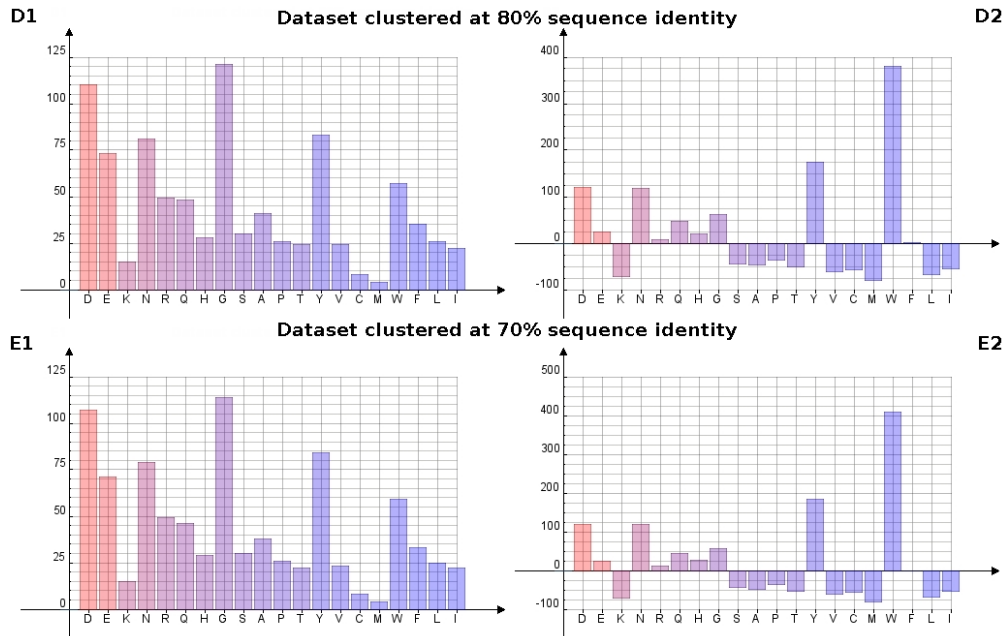
For instance, Histidine (His (H)) and Arginine (Arg (R)) become over-represented and the deviations for Asp and Asn are reduced. In general, amino acid deviations decrease along with the identity level. The exception is Trp since its value increases through clusters. The relevance of Trp was 'hidden' behind redundant data and it could not emerge until this redundancy was reduced.



**Figure 3.5:** All amino acids in the vicinity of  $\alpha$ -D-Manp were selected and clustered at different identity levels using the program CD-HIT. The deviations from natural occurrences show divergences through levels. For example, in the dataset clustered at 100% identity level, 867 amino acids around 151 carbohydrate residues in 151 carbohydrate chains are found. These numbers decrease in the clustered dataset at 80% identity to 534 amino acids, 97 carbohydrate residues and 96 glycan chains respectively. The number of PDB entries employed for the analysis also decrease from 151 to 97. A value of 80% identity has been selected as a good compromise between reduction of redundancy and statistical significance. Continued on next page.

The selection of the 'best' clustering identity threshold varies according to the precision of the analysis. It must be noticed that, as described above, the search space decreases with the increment of identity threshold. If the search space is too small, the results may not be statistically reliable but biased again by the small dataset. In the following examples, a threshold of 80% sequence identity has been chosen as a good

compromise between dataset size and reduction of redundancy, and is used in all further analyses.



**Figure 3.5:** Continuation.

Cluster	Amino acids	Carboh. residues	Carboh. chains	PDB entries
Unclustered	5,849	974	644	273
100%	867	151	151	151
90%	554	101	101	101
80%	534	97	97	97
70%	518	95	95	95

**Table 3.3:** All amino acids in the vicinity of  $\alpha$ -D-Manp were selected and clustered at different identity levels using the program CD-HIT. The deviations from natural occurrences show divergences through levels. For example, in the dataset clustered at 100% identity level, 867 amino acids around 151 carbohydrate residues in 151 carbohydrate chains are found. These numbers decrease in the clustered dataset at 80% identity to 534 amino acids, 97 carbohydrate residues and 96 glycan chains, respectively. The number of PDB entries employed for the analysis also decrease from 151 to 97.

A value of 80% identity has been selected as a good compromise between reduction of redundancy and statistical significance.

## 3.2 Redundancy

The first objective of this study was to determine up to what extent is the data from the PDB affected by redundancy. The above described example with  $\alpha$ -D-Manp shows that the results obtained vary with each cluster level. Other monosaccharide residues were analyzed (Figure 3.6, Table 3.4). In some cases ( $\beta$ -D-Manp,  $\alpha$ -D-Glcp,  $\beta$ -D-Glcp Figure 3.6 'I', 'L' and 'M' respectively) the difference between clustered and unclustered datasets is barely perceptible, whereas for others ( $\alpha$ -L-Fucp, sulfated residues,  $\beta$ -D-Galp Figure 3.6 'J', 'T' and 'S' respectively) the divergences are more notorious.

## 3.3 Carbohydrate chemistry and interaction with proteins

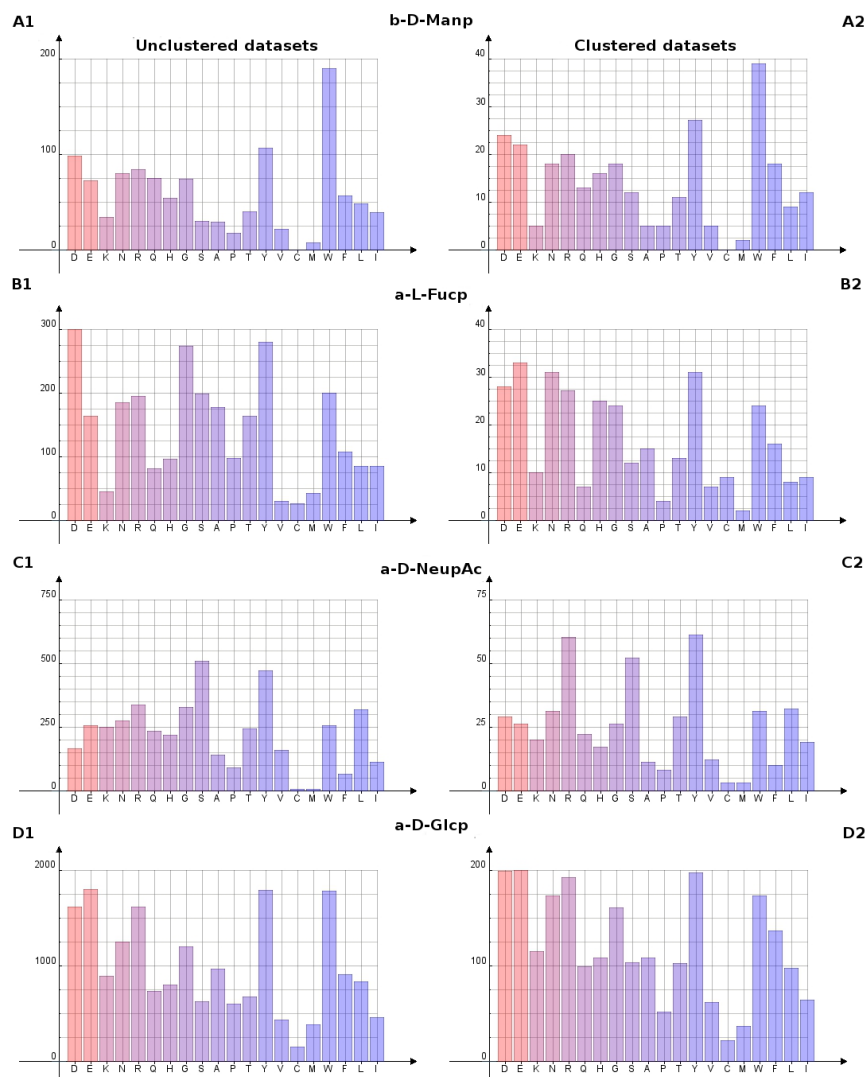
In order to reach the second objective of this thesis, the spatial vicinity of the most common monosaccharides is further inspected. As seen in Figure 3.6, most amino acids with planar groups in their side chains are overrepresented (Asn, Arg, Gln, His, Tyr, Trp) in all cases. Trp and Tyr show the largest positive deviations while those for Asn, Arg, Gln are positive, but vary greatly among glycan residues.

The N-acetylation of  $\beta$ -D-Glcp may contribute to an increment of Asn around  $\alpha$ -D-GlcpNAc. Arg and His have the third largest deviation for  $\alpha$ -D-Neup5Ac and  $\alpha$ -L-Fucp, respectively. Also variable is the deviation of Gln, which changes from under to over-represented through the cases.

Between  $\alpha$ -D-Manp (Figure 3.5 D2) and  $\beta$ -D-Manp (Figure 3.6 I2) there is a notable difference in the deviation value for Trp. A similar case occurs between  $\alpha$ -D-GlcpNAc and  $\beta$ -D-GlcpNAc (Figure 3.6 N2 and O2) though at a lesser extent. However, in most cases,  $\alpha$  or  $\beta$  anomers seem not to have a relevant influence on the results. For instance, between  $\alpha$ -D-Glcp and  $\beta$ -D-Glcp the change in deviations is minimal. The main difference between  $\alpha$ -D-Galp and  $\beta$ -D-Galp is the slight increment in the positive deviations of His, Tyr and Trp.

Figure 3.6.K2 shows deviations from natural occurrences for the sialic acid  $\alpha$ -D-Neup5Ac. The negatively charged amino acids, Asp and Glutamic acid (Glu (E)), are under-represented. The positively charged Arg and to a lesser extent His are over-represented. Lysine (Lys (K)) on the other hand, is under-represented. In fact, except for the sulfated residues, the deviation of Lys is for all cases negative.

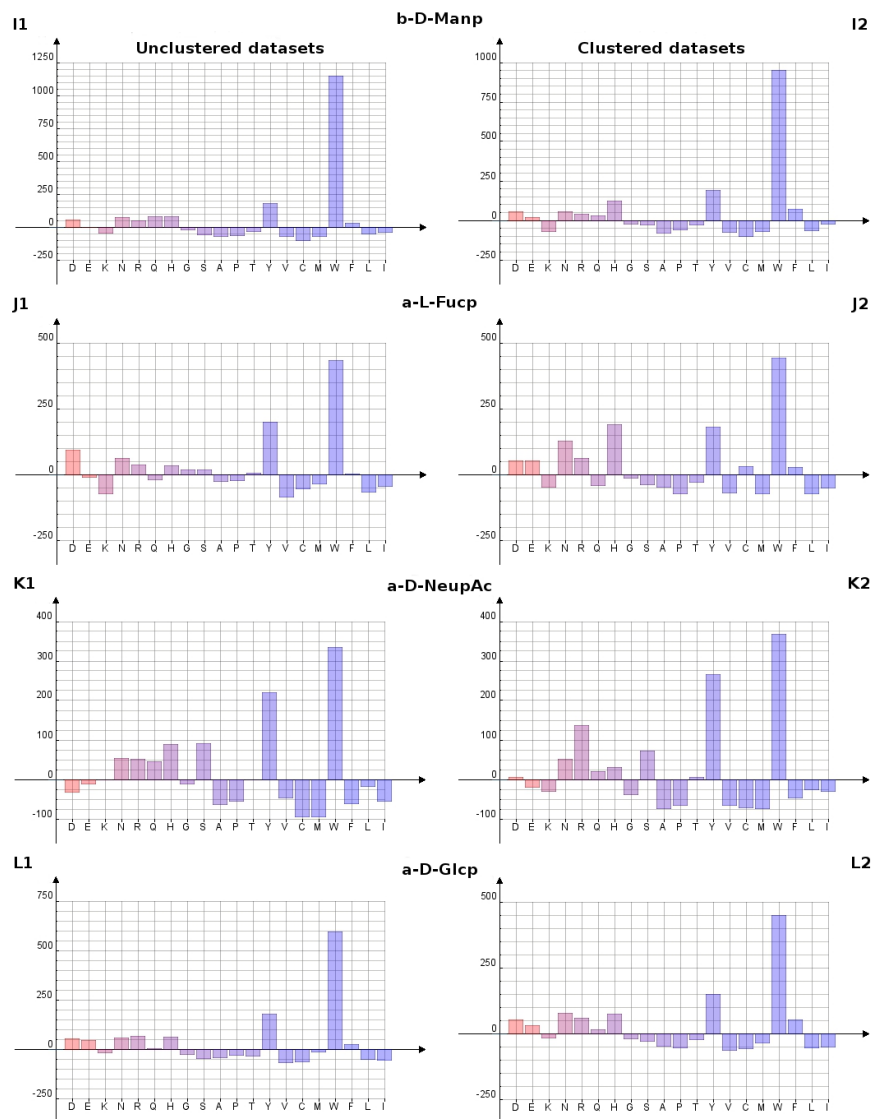
The distribution of amino acids in the vicinity of glycans shows some differences in relation to the sugar analyzed. However, a higher variability among residues would be expected to account for the remarkable specificity observed in protein-carbohydrate interactions. Since this is not the case, the analysis of the way interactions are mediated at atomic level is the key to understand this specificity in more detail. This task can



**Figure 3.6:** In order to determine at what extent redundancy affect the data, absolute counts generated from clustered (at 80% identity level) and unclustered datasets are compared. Continued on next pages.

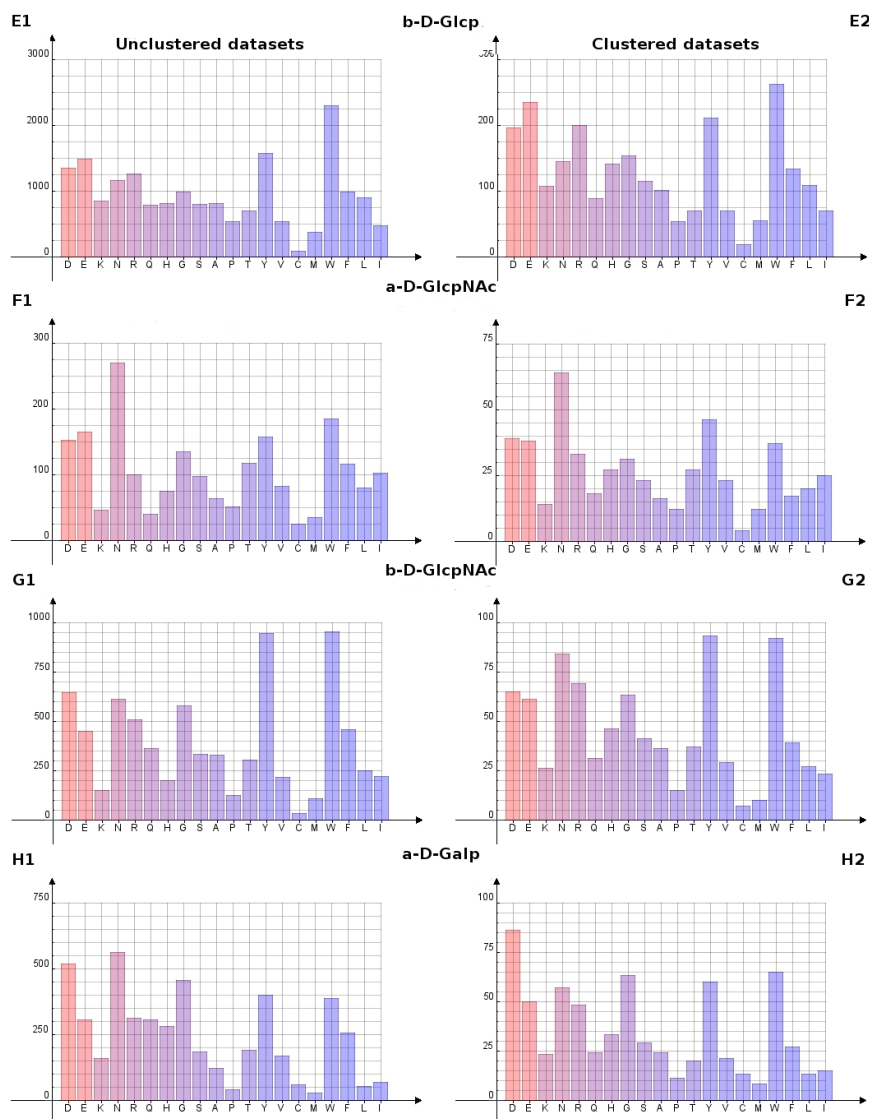
also be performed with GlyVicinity.

The spatial coordinates of all atoms in the vicinity of the selected monosaccharides are collected from the GlyVicinity database and depicted graphically with the help of Jmol. Figure 3.7 is a compendium of images that illustrate the importance of carbohydrate stereochemistry for the establishment of interactions. The first two columns show the polar (red) and apolar (blue) areas created in both sides of carbohydrate rings by means of contiguous chemical groups. In the next column, the respective monosaccha-



**Figure 3.6:** In order to determine at what extent redundancy affect the data, deviations from natural occurrences generated from clustered and unclustered datasets are compared.

rides are displayed in the center surrounded by all nitrogen and oxygen side chain atoms of polar amino acids (Arg, Lys, His, Serine (Ser (S)), Threonine (Thr (T)), Asn, Gln, Glu and Asp) which are superimposed according to their spatial coordinates. Finally, the last columns display atomic interactions of side chain atoms of Trp and Tyr, respectively. The images in the first two columns were generated with the tool YASARA [KKV02], the rest with Jmol. As described before, amino acid atoms are taken from entries determined at resolutions of at least 3 Ångstroms and clustered at 80% identity.

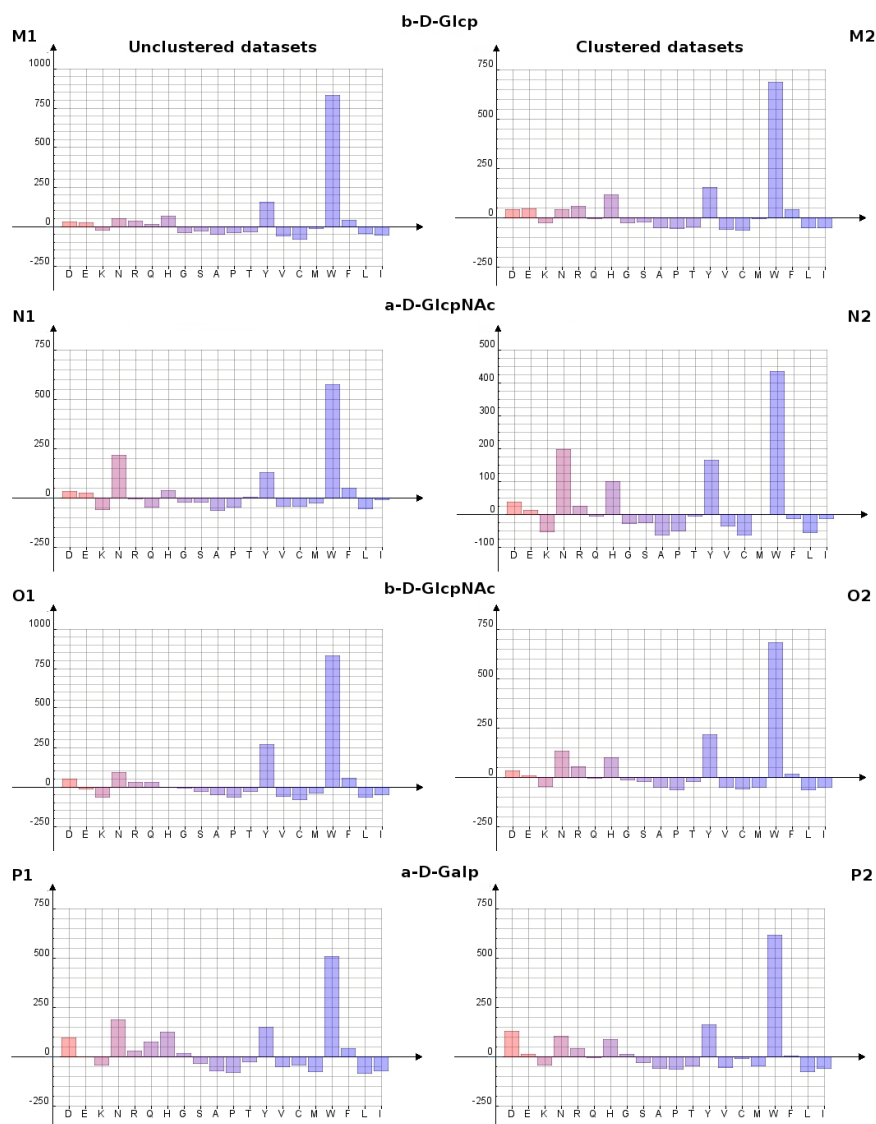


**Figure 3.6:** Continuation

The specific stereochemical features of each carbohydrate residue determine polar and apolar areas, which in turn affect the way atomic interactions are mediated. Therefore, compared to the previously shown diagrams, the difference in the distribution of atoms is more evident from residue to residue.

Visual inspection reveals that, unlike the deviations shown before,  $\alpha$  and  $\beta$  anomers show some divergences. For instance, non-polar atoms of Trp interacting with  $\beta$ -D-Manp are clearly attracted by the lower side of the ring whereas this is not exactly the case for  $\alpha$ -D-Manp (Figure 3.7 B4, A4). In the case of these two residues is difficult to observe a





**Figure 3.6:** Continuation.

difference regarding the establishment of potential hydrogen bonds (Figure 3.7 A3, B3), mainly due to the dissimilarity in the number of atoms (lesser for  $\alpha$ -D-Manp).

Polar atoms contact with both  $\alpha$ -D-Glcp and  $\beta$ -D-Glcp around the hydroxyl groups of C2, C3 and C4 (Figure 3.7 E3, F3). Non-polar atoms on the other hand, show a divergence. They prefer the upper side of the ring for  $\alpha$ -D-Glcp whereas the interactions are more equally distributed between upper and lower side of the ring for  $\beta$ -D-Glcp (Figure 3.7 E4, F4).  $\beta$ -D-GlcpNAc shows similar non-polar atom distribution like  $\beta$ -D-Glcp (Figure 3.7 G4).

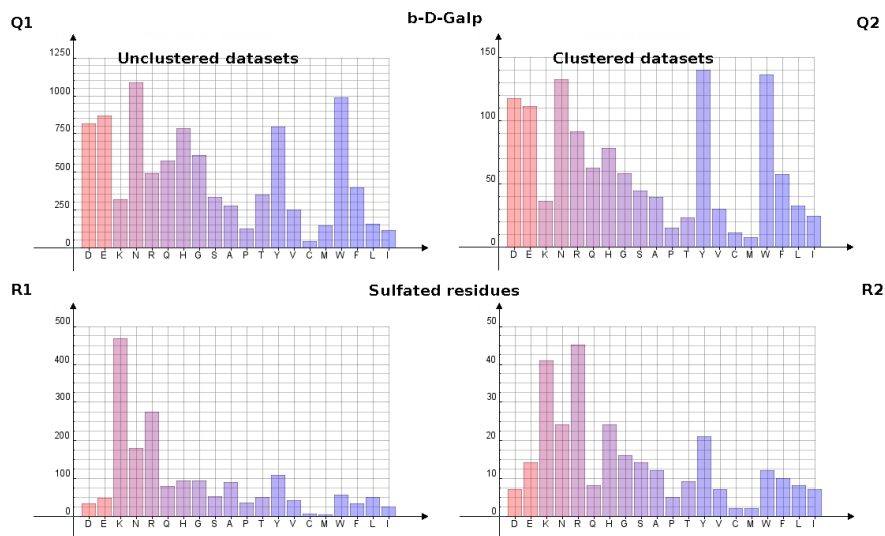


Figure 3.6: Continuation

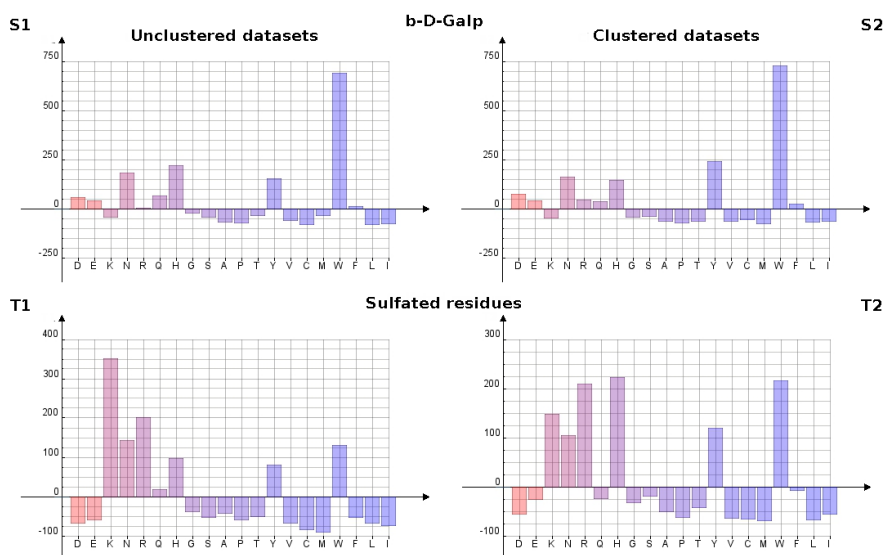
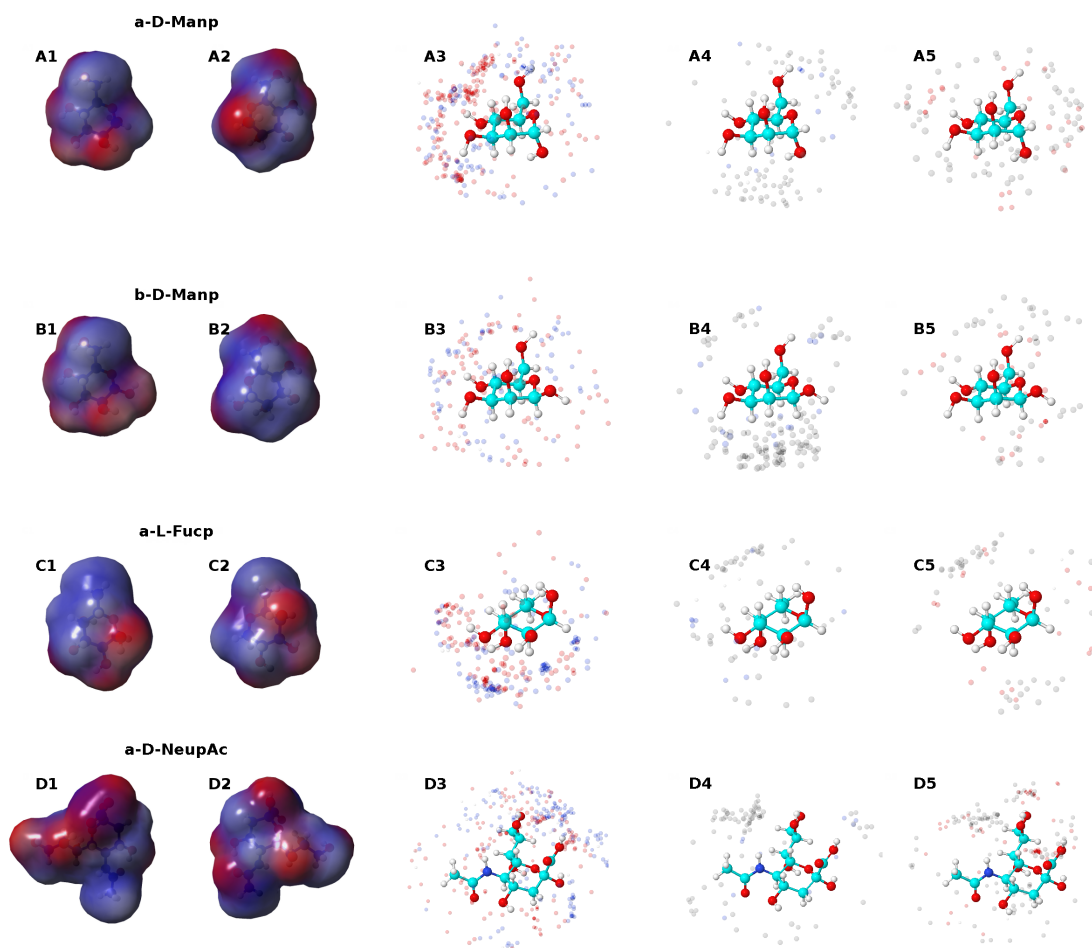


Figure 3.6: Continuation

The number of non-polar atoms for  $\alpha$ -L-Fucp (Figure 3.7 C4) and  $\alpha$ -D-Neup5Ac (Figure 3.7 D4) is small, which complicates the search for distribution patterns. Nevertheless, the observed results match well the expectations. For example,  $\alpha$ -L-Fucp shows a similar pattern as  $\alpha$ -D-Galp, with a preference for the upper side ring instead of the lower side, due to their enantiomer conformation. In the case of  $\alpha$ -D-Neup5Ac, the non polar atoms show certain predilection for the acetyl group and the hydrogens of C8 and

C9. The number of polar atoms (Figure 3.7 D3, C3) on the other hand, is sufficient to reveal the predilection for the hydroxyl groups of C6, C7, C8 and C9 of  $\alpha$ -D-Neup5Ac and C2, C3 and C4 of  $\alpha$ -L-Fucp.



**Figure 3.7:** Pattern of atomic interactions for each monosaccharide analyzed. The first two columns represent the polarity of the monosaccharide surface (red=polar, blue=unpolar), the third column shows the glycan residue with a cloud of polar atoms (O and N) superimposed from the PDB entries, the fourth and fifth columns show non polar atoms from Trp and Tyr, respectively. Continued on next page.

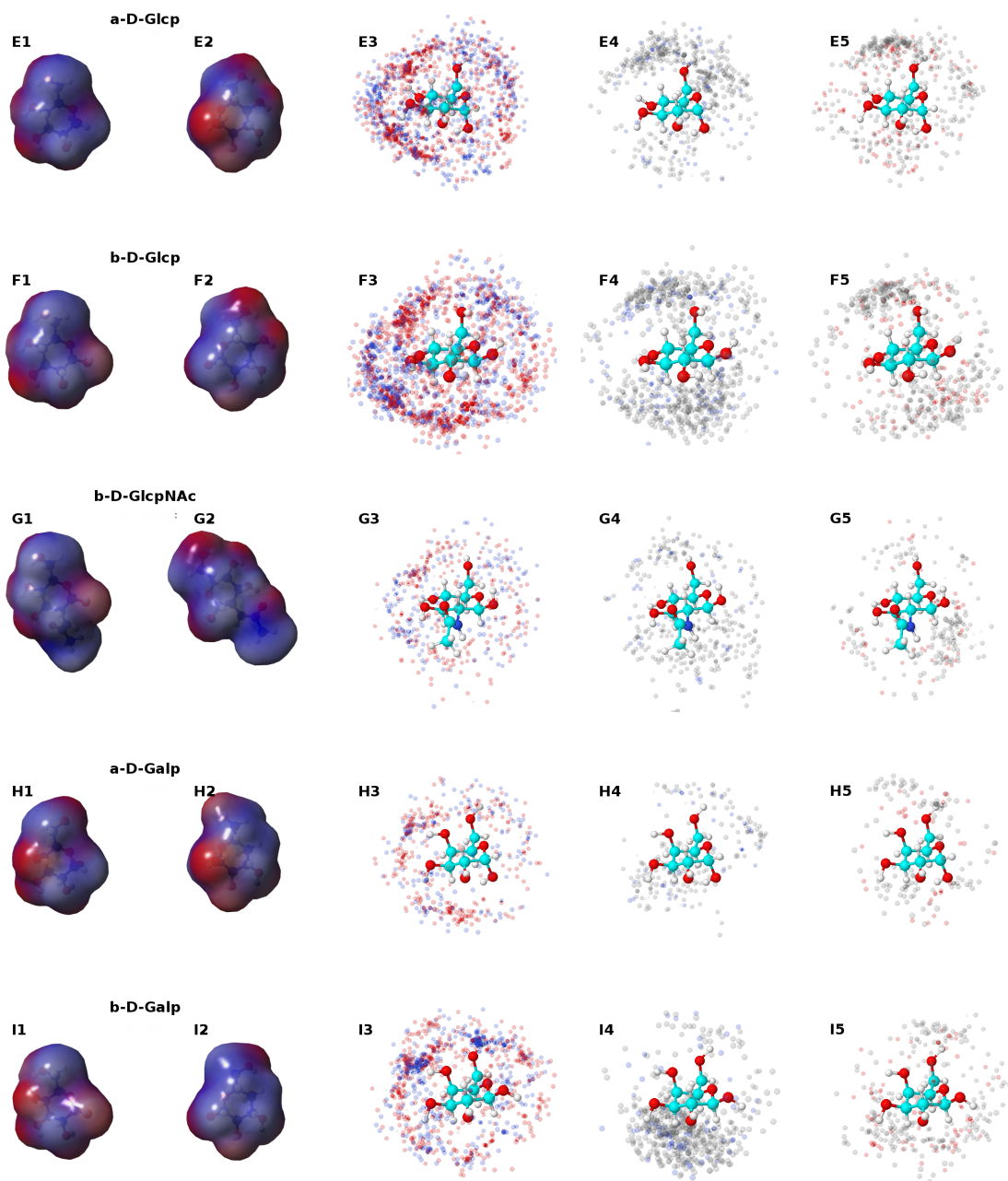


Figure 3.7: Continuation

### 3.4 Aromatic amino acids in protein-carbohydrate interactions

The calculated deviations from natural occurrences (Figure 3.6) revealed outstanding positive values for Trp and Tyr in the vicinity of all of the analyzed monosaccharides. Looking at Figure 3.7 (the analysis of residues at atomic level), it seems that the non-polar atoms follow a different interaction pattern than the polar atoms. In order to confirm this contrast, the interactions established between different carbohydrate residues and the amino acids with the largest deviations were examined. Figure 3.8 summarizes the number of interactions in which the respective atoms are involved, accentuated by gradient colors. Only  $\beta$ -D-Glcp is shown but the same pattern is observed for different carbohydrate residues.

As expected, most of interactions between glycans and polar amino acids occur at the tips of their side chains. Fewer interactions occur at the backbone atoms, since usually the backbone is buried and therefore too far away from the glycans bound in the protein surface. Most amino acids interact with carbohydrates through the atoms at the tips of their side chains. The interactions with Trp (Figure 3.8 E), however, behave differently. Unlike the other amino acids analyzed, the atoms of Trp present a distribution of interactions, which is not concentrated in some specific atoms but it is evenly dispersed along the indole ring. This characterizes the establishment of stacking interactions.

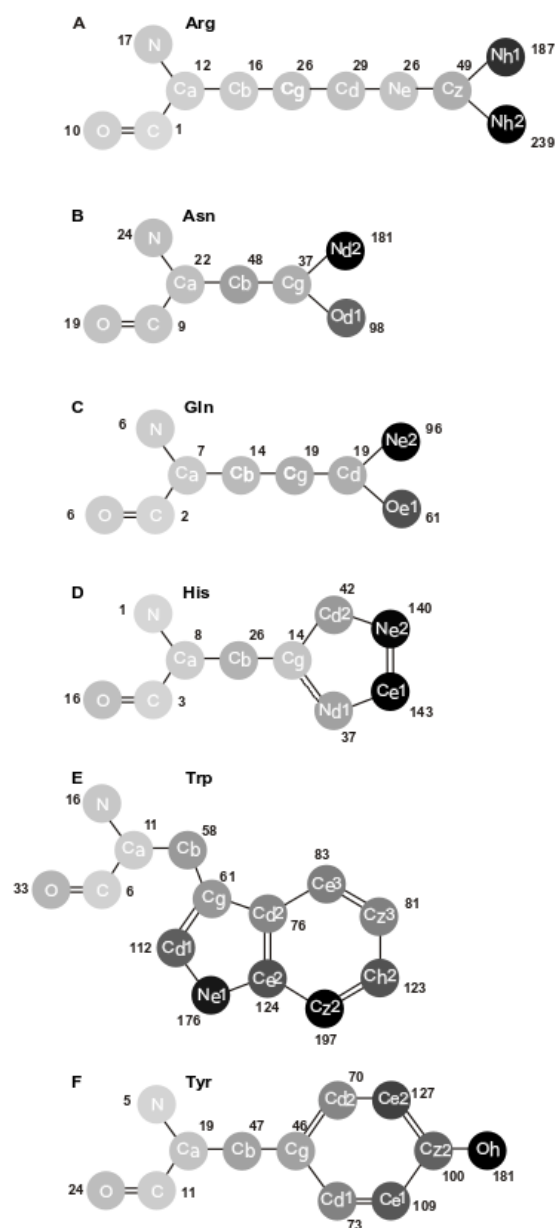
Specific monosaccharides were analyzed (Figure 3.9) to determine if the interactions with Trp were established similarly. With the exception of  $\alpha$ -D-Neup5Ac, this was the case, Trp engages in interactions differently than the rest of amino acids analyzed.

### 3.5 Structural patterns and glycosylation

Of all post-translational modifications a protein undergoes in the living cell, glycosylation is considered the most common and highly diverse. It is estimated that about 50% of proteins are glycosylated [AHS99]. However, not all sequons Asn-xx-Ser/Thr are glycosylated and it is not clear how to differentiate between occupied and unoccupied sites.

This part of the study focuses on N-glycosylation and the occupancy of glycosylation sites. Here, both the primary structure of occupied glycosylation sites, and the spatial vicinity of the individual monosaccharides that conform the N-linked glycan core are statistically analyzed. The objective is to elucidate structural interaction patterns that may affect N-glycosylation.

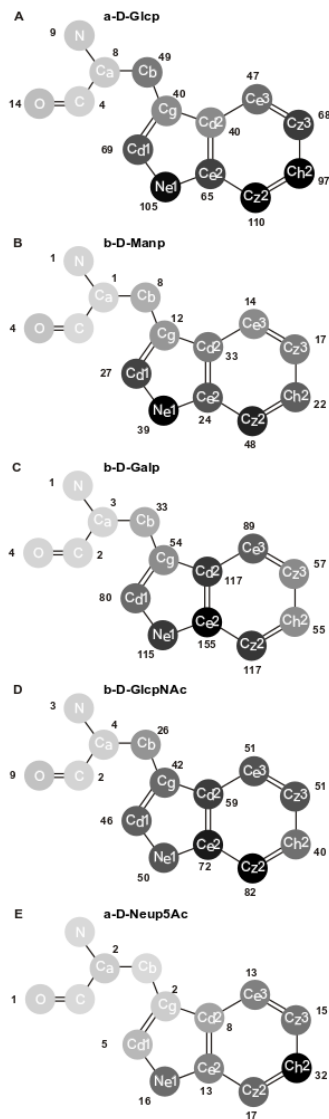
Figure 3.10 shows the common structural core shared by different types of N-glycans, only the monosaccharides in this subset of residues were analyzed. Data on glycans at



**Figure 3.8:** Gradient colors indicate the level of interactions established between glycan residues and specific amino acid atoms. The interactions with sugar residues occur mostly at the tips of the amino acid side-chains. The exception is Trp, since the interactions are more evenly distributed around the aromatic ring.

superior levels are scarce mainly due to the limitations to experimentally determine long glycan chains (see Discussion).

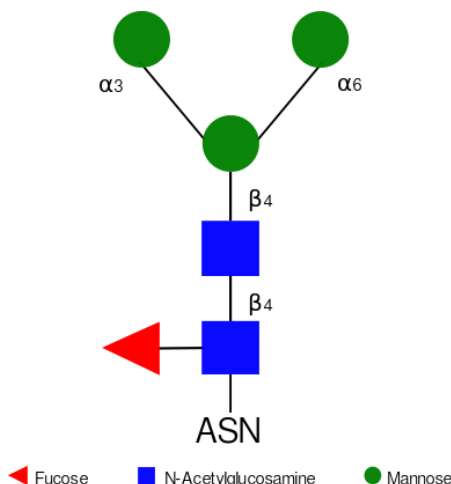
All amino acids up to 4 Ångstroms around each of the glycan residues that conform a fucosylated N-glycan core were retrieved from the GlyVicinity database. The resulting datasets were then clustered at 80% amino acid identity. Table 3.5 summarizes the datasets sizes.



**Figure 3.9:** The interactions between Trp residues and different monosaccharides are established on a similar way, distributed around the indole ring, which characterizes stacking interactions.

The absolute numbers of amino acids decrease as the analyzed glycan residues are found further away from the glycosylation site. Accordingly, for the first β-D-GlcpNAc,

a dataset of 4,153 amino acids was generated whereas for the most distant glycan,  $\alpha$ -D-Manp, the number is reduced to 313 amino acids. The amino acid distributions obtained from the analysis of the datasets by GlyVicinity are depicted in Figure 3.11 as absolute numbers and deviations from natural occurrences.



**Figure 3.10:** N-glycan core as shared by the different types of N-glycans (hybrid, high-mannose and complex). The  $\beta$ -D-GlcpNAc directly attached to the Asp is considered to be on position 1, the next  $\beta$ -D-GlcpNAc in position 2 and so on.

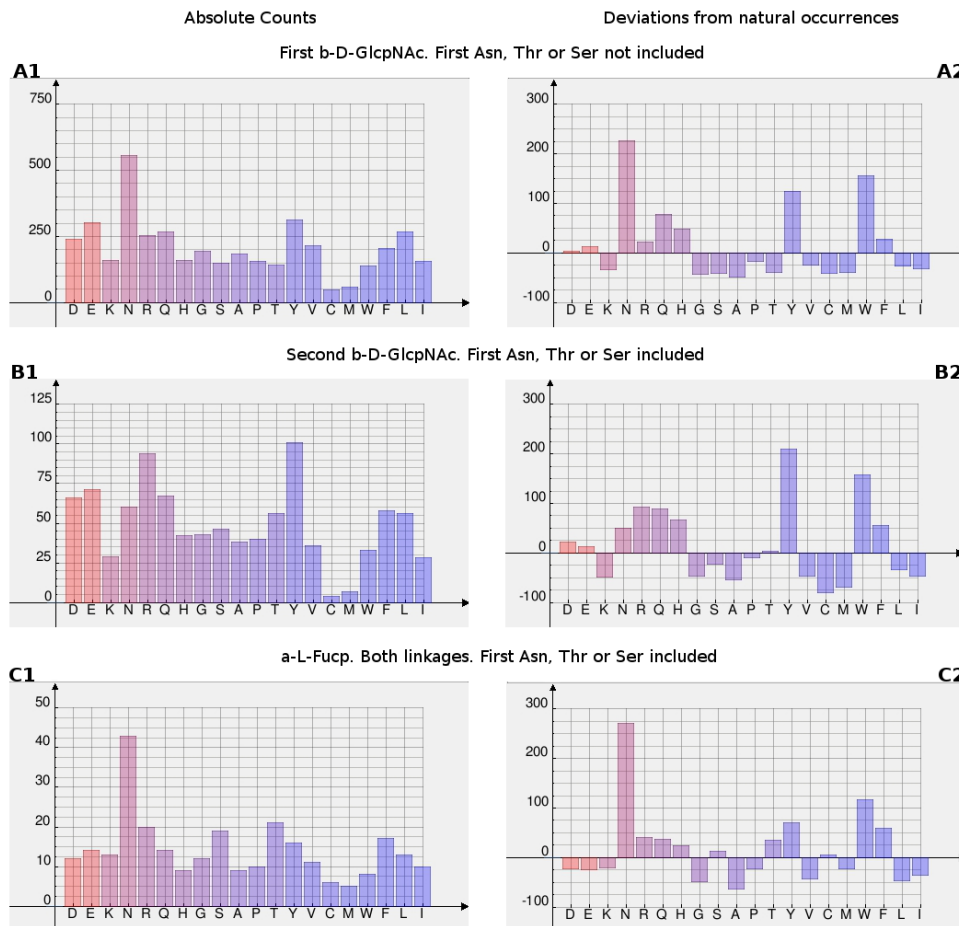
With the exception of the aromatic amino acids, the deviations in Figure 3.11 show that hydrophobic amino acids Alanine (Ala (A)), Valine (Val (V)), Leucine (Leu (L)), Isoleucine (Ile (I)), Methionine (Met (M)) are generally under-represented. On the other hand, the negatively charged amino acids (Asp, Glu) and some basic (His, Arg) and hydrophilic amino acids (Ser, Thr, Asn, Gln) are in almost all cases over-represented. The amino acid frequencies evidently vary depending on the type of glycan residue and its position in the N-glycan core. This is more visible for amino acids with large positive deviations.

The Asn and Ser/Thr that form part of the consensus sequences were excluded of the analysis for the first  $\beta$ -D-GlcpNAc. Yet, the absolute numbers for Asn are high around this and the rest of monosaccharides in the N-glycan core. Aromatic amino acids are in general over-represented. For the  $\beta$ -D-GlcpNAc in the first position, the  $\alpha$ -L-Fucp and the  $\beta$ -D-Manp, the predominant value belongs to Trp, the deviation of Tyr is the highest for the  $\beta$ -D-GlcpNAc in second position and Phe surpasses the other aromatic residues around  $\alpha$ -D-Manp. His shows mostly over-represented values, but not as high as the previously mentioned.

Despite its polar nature, Lys is under-represented in all cases. The deviations of Arg and Gln are mostly positive but also very variable. Arg around  $\alpha$ -D-Manp is remarkable



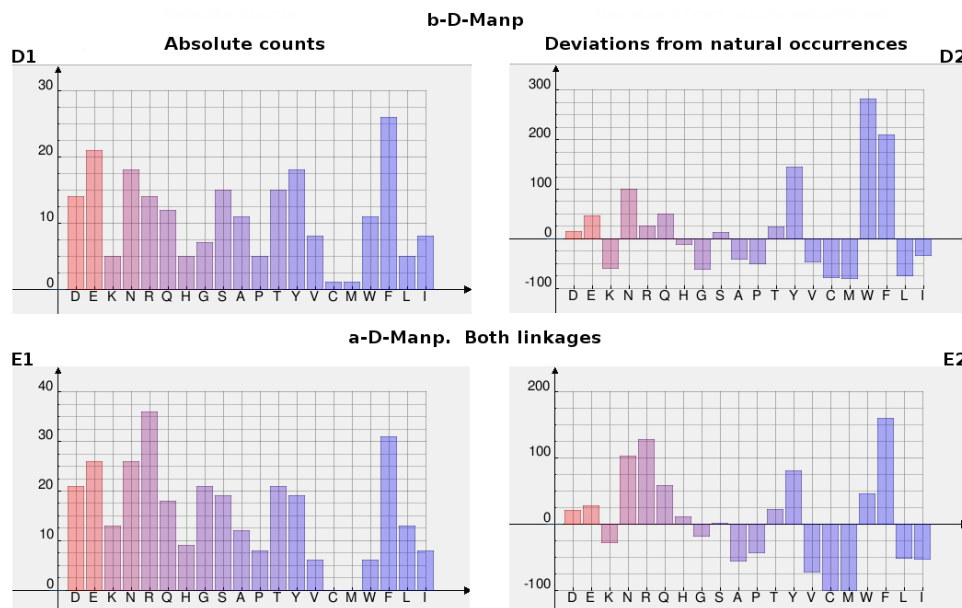
since its deviation becomes the second largest, just below Phe. Cysteine (Cys (C)) is also generally under-represented except for a small increment around  $\alpha$ -L-Fucp. It is interesting to observe that the negative deviation of Pro is less pronounced around the two  $\beta$ -D-GlcpNAc than around residues further away from the glycosylation site.



**Figure 3.11:** Amino acid distribution around the carbohydrate residues that conform the N-glycan core. The Ser and Thr closest to the glycosylation site for  $\beta$ -D-GlcpNAc were not included due to their expected high values as part of the N-glycosylation sequon. Continued on next page.

### 3.5.1 Atom distribution

The 3D position of atoms around N-glycans shows that potential hydrogen bonds for the  $\beta$ -D-GlcpNAc located in the first position in the N-glycan core are established mainly with the hydroxyl group in C1 and the NO group of the acetic acid (Figure 3.12 A1). The lowest number of interactions, on the other hand, are established with the OH groups



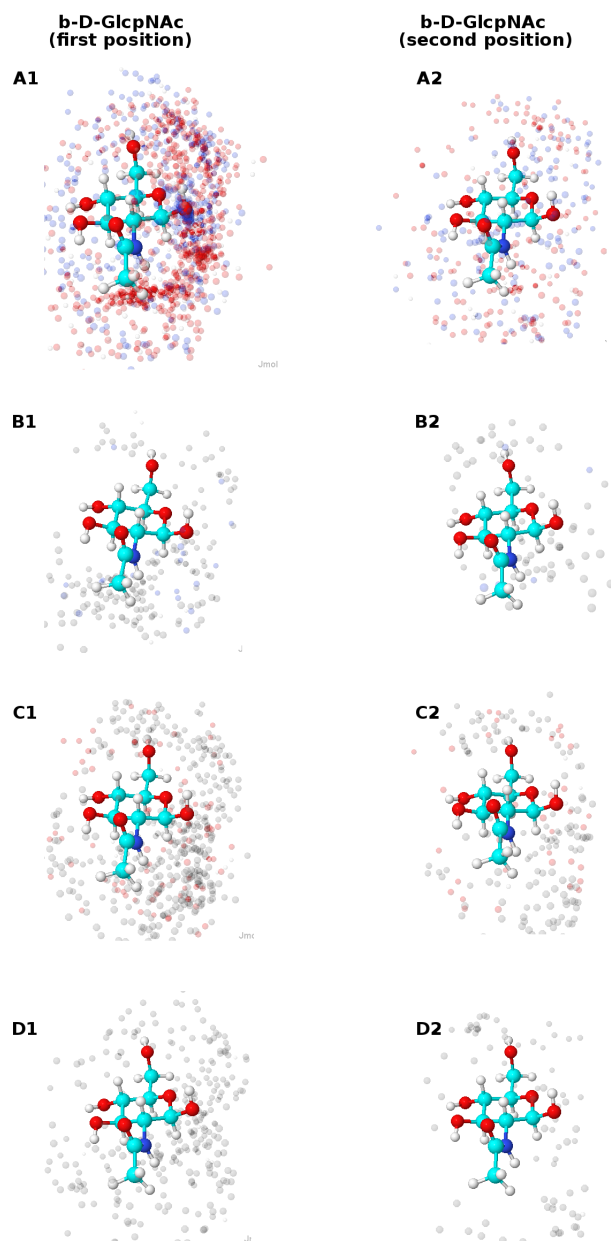
**Figure 3.11:** Continuation

of C3, C4 and C5. The  $\beta$ -D-GlcpNAc in second position (Figure 3.12 A2) does not have recognizable patterns of polar bonds, not even around the NO or C1, as it happens with the same residue in the first position.

The rest of residues present some patterns of potential hydrogen bonds. However, due to the reduced number of atoms, these patterns cannot be considered reliable. Non polar interactions with Trp are very similar between  $\beta$ -D-GlcpNAc in first and second position (Figure 3.12 B1 and B2) and also similar to those observed for non-covalently linked ligands (Figure 3.8 G4). In each case, a preference for the apolar face of the glycan residue is noticeable. The biggest difference is the slight preference for the hydroxyl groups of C5 and C6 for the  $\beta$ -D-GlcpNAc in second position in the N-glycan core.

The interactions of Tyr with both, covalently and non-covalently linked  $\beta$ -D-GlcpNAc (Figure 3.12 C1 and C2, Figure 3.7 G5) also favour the apolar face of the glycan. However, unlike the case with Trp, the patterns that arise are slightly different. Two apolar patches are especially noticeable, around C5, C6 and the acetic acid, and this pattern is very similar for Phe around the first  $\beta$ -D-GlcpNAc (Figure 3.12 D1).

Unfortunately, the number of entries for glycans at higher levels in the N-glycan core is very small. Therefore, they have not been considered for analysis.

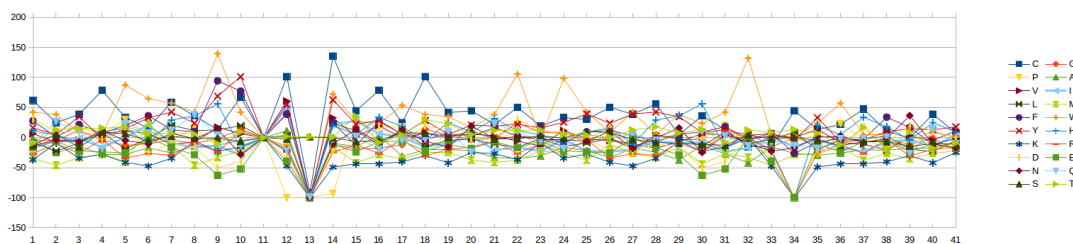


**Figure 3.12:** Distribution of atoms around the  $\beta$ -GlcNAc residues in the first and second position in N-glycan core. The first row shows polar atoms (O and N) of polar amino acids (Arg, Lys, His, Ser, Thr, Asn, Gln, Glu and Asp). The following rows show the non-polar atoms of Trp and the polar and non-polar atoms of Tyr and Phe, respectively.

### 3.5.2 Amino acid sequences around glycosylation sites

GlySeq [LFvdL05] is a tool that locates peptide sequences of glycoproteins in the PDB for their subsequent storage and analysis. In order to compare amino acid frequencies from peptide sequences to those from spatial vicinities, all sequences in the database until April 2013 were retrieved and statistically analyzed with GlySeq. The final dataset consisted of 949 non redundant primary sequences clustered at 80% identity.

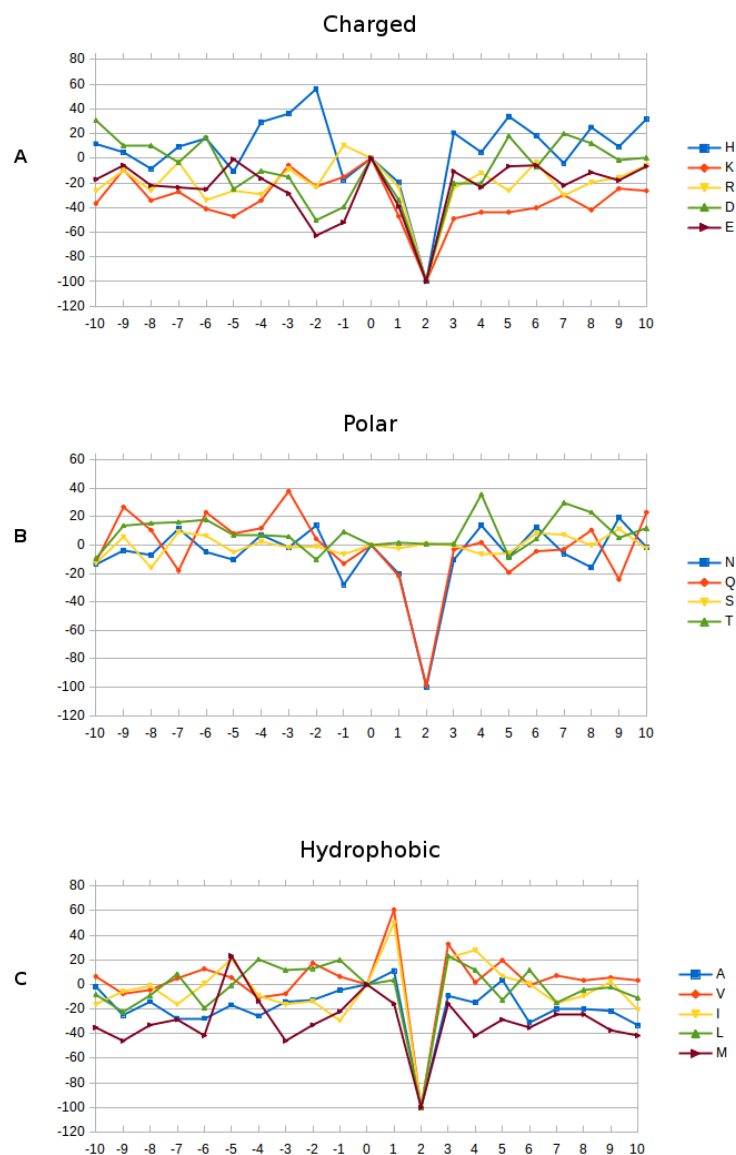
10 positions before and 30 after the glycosylated Asn (in the following referred as negative (-) and positive (+) positions, respectively) were analyzed. This range was selected to cover also those residues which may contact with the translocon complex during glycosylation [PMP<sup>+</sup>04]. However, the frequency of amino acids seems fairly constant in the areas further away from the Asn except for slight increments in certain positions, for instance, +11, +13 and especially +22 for Trp, +12 for His and +28 for Asn (See Figure 3.13). The largest variations in the distributions of amino acids can be observed in the positions flanking the glycosylation site. Therefore, only the amino acid frequencies in positions -10 to +10 are displayed in Figure 3.14 as absolute counts and deviations from natural abundances.



**Figure 3.13:** 10 positions before and 30 after the glycosylation site were statistically analyzed with GlySeq. Except for peaks in a few positions, most of amino acid variation occurs in the nearest positions to the glycosylation site.

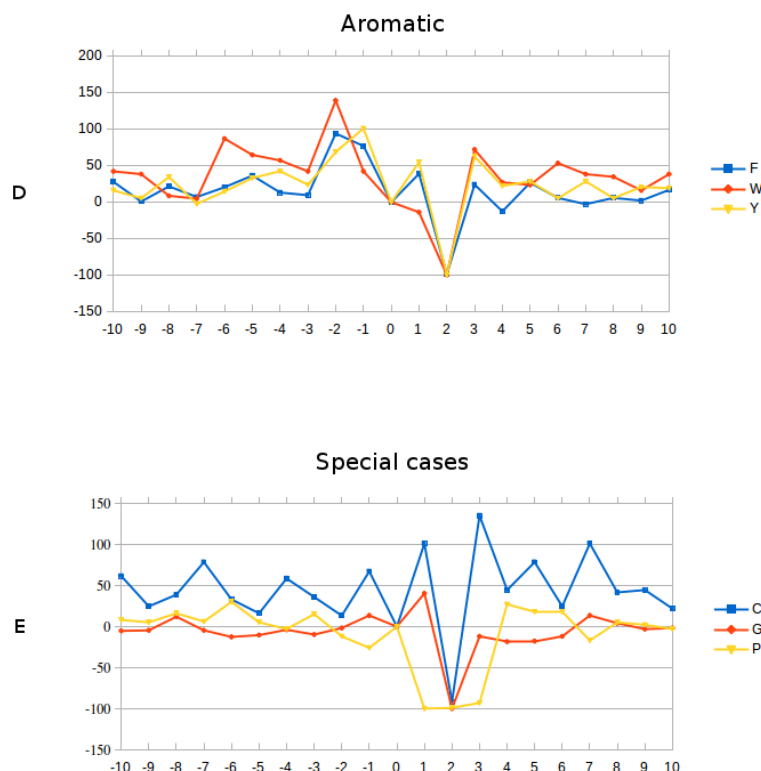
Being part of the sequon motif, the values of Asn in position 0 and Thr/Ser in position +2 were set to 0. The scale of the diagrams is reduced accordingly, which allows a better view of the rest of amino acids.

As shown above (See Figure 3.11), Asn is over-represented in the spatial vicinity of N-glycosylation sites, especially around the first  $\beta$ -D-GlcNAc and  $\alpha$ -L-Fucp. In contrast, the deviation obtained from the analysis of amino acid sequences is not high (Figure 3.14 B). Actually, Asn is under-represented in several positions (-9, -8, -5, -1, +5, +7, +8). On the other hand, similar to the spatial vicinities, aromatic amino acids are over-represented, especially in the positions preceding the glycosylation site and in +3 (Figure 3.14 D). The highest deviations in this group belong to Trp, which in position -2 reaches its highest peak. Tyr comes in a second place, surpassing Trp only in -1. Phe



**Figure 3.14:** 10 positions before and 10 positions after the glycosylation site are analyzed with GlySeq. The amino acids are organized in 5 groups, depending on their chemical properties: positively charged (His, Lys, Arg and Asp), Polar (Asn, Gln, Ser, Thr), Hydrophobic (Ala, Val, Ile, Leu), Aromatic (Phe, Trp and Tyr) and special cases (Cys, Gly and Pro). Continued on next page.

exceeds Tyr only in -2, but also becoming slightly under-represented in +4. His has the lowest deviations, with high peaks at positions -2, -3 and +5.



**Figure 3.14:** Continuation

The rest of hydrophobic amino acids have a comparatively weak presence in the positions surrounding glycosylation sites (Figure 3.14 C). Met is under-represented in both protein sequences and spatial vicinities (except for position -5). On the contrary, Val has the highest deviation within the group of hydrophobic amino acids when the primary sequence is analyzed, even though its deviation is always under-represented in the spatial vicinity of N-glycans.

Asp (Figure 3.14 A) stands out in certain positions, but in general, positively and negatively charged amino acids are under-represented around glycosylation sites, especially Lys in +4 and +5. Pro (Figure 3.14 E) is just above the average distribution as well as Gly (Figure 3.14 J) except for a peak in position +1. The most irregular distribution is shown by Cys (Figure 3.14 E) which reaches its highest peaks at +1 and +3.

Cluster	Amino acids	Carbohydrate residues	Chains	PDB entries
$\beta$ -D-Manp				
unclustered	1,154	265	211	110
80%	281	66	66	66
$\alpha$ -L-Fucp				
unclustered	2,822	456	387	151
80%	335	64	64	64
$\alpha$ -D-NeupAc				
unclustered	4,421	511	474	199
80%	502	69	69	69
$\alpha$ -D-Glcp				
unclustered	19,433	3,340	2,240	1,037
80%	2,395	438	438	438
$\beta$ -D-Glcp				
unclustered	18,692	3,637	2,919	1,148
80%	2,530	513	513	513
$\alpha$ -D-GlcNac				
unclustered	2,085	380	371	165
80%	526	106	106	106
$\beta$ -D-GlcNac				
unclustered	7,756	1,425	844	444
80%	894	200	200	200
$\alpha$ -D-Galp				
unclustered	4834	642	622	189
80%	690	104	104	104
$\beta$ -D-Galp				
unclustered	9,478	1,659	1,556	625
80%	1,243	237	237	237
Sulfated Residues				
unclustered	1,806	350	178	107
80%	288	61	61	61

**Table 3.4:** Difference in search space between clustered and unclustered datasets for different carbohydrate residues.

Residue	Depth	Linkage	Amino acids	Carb. residues	Chains	PDB entries
$\beta$ -D-GlcpNAc	1		4,153	1,781	1,781	859
$\beta$ -D-GlcpNAc	2	1,4	975	541	541	424
$\alpha$ -L-Fucp	2	1,3 1,6	282	124	123	106
$\beta$ -D-Manp	3	1,4	220	134	134	122
$\alpha$ -D-Manp	4	1,3 1,6	313	158	138	124

**Table 3.5:** Description of the results generated for the analysis of N-glycans with GlyVicinity. The second column identifies the position of the analyzed glycan in the N-glycosylation core. The third column indicates the linkages analyzed. Due to the lack of data, no distinction between 1,3 and 1,6 linkages could be made for  $\alpha$ -L-Fucp and  $\alpha$ -D-Manp. The following columns indicate the total numbers of amino acids (E.g. 975) interacting with carbohydrate residues (E.g. 541) contained in glycan chains (E.g. 541) in individual PDB entries (E.g. 424).



# 4

## Discussion

STRUCTURAL DETERMINATION OF PROTEIN-CARBOHYDRATE COMPLEXES provides with an opportunity to understand in more detail how these molecules interact. Accordingly, the analysis presented in this thesis takes advantage of the structural data on glycoproteins found in the largest source of 3D biomolecules, the PDB. These data are analyzed targeting towards the fulfillment of four key objectives: 1) To prove the utility of the PDB as an indirect source of data on carbohydrates as well as the extension in which redundancy affects the input datasets. 2) to determine the influence that carbohydrate chemistry has on the way monosaccharides interact with proteins, 3) to find an explanation for the large presence of aromatic amino acids in protein-carbohydrate interactions, as well as to determine the extent in which these interactions occur among different glycan residues, and finally 4) to find a possible structural pattern that indicates glycosylation through the analysis of 3D data on N-glycans and the comparison of interactions with non-covalently linked ligands.

The obtained results show that an analysis at atomic level must be performed to resolve specificity since calculation of deviations from typical abundances of amino acids proved not to be sufficient. Polar interactions were expected but surprisingly aromatic amino acids have the largest positive deviations in practically all cases. Furthermore, there are differences in the amino acid pattern of interactions established with covalently and non-covalently linked carbohydrates. The inherent redundancy in the database has a minor influence on the results. Except for a few cases, the removal of such redundancy is mostly useful to refine patterns already observed in the results generated from unclustered datasets. These findings support the value of the PDB as an indirect source of data on glycans.

A more extended discussion for each of the goals in this work is presented in the following sections.

## 4.1 Utility of the PDB as source of information on carbohydrates

The PDB focuses on proteins. However, some of the entries in this database consist of proteins determined by crystallography or NMR in presence of sugar residues. The PDB can then be considered as an indirect source of 3D data on carbohydrate structures. Unfortunately, using these data is not straightforward since the quality of some entries may not be acceptable. Thus, the mining of the PDB requires not only of a suitable way to find and collect carbohydrate data but also of proper curation before processing the data into useful information. In this work, the selection of data based on their quality as well as the processing are performed by the software tool GlyVicinity (see Materials and Methods).

Another challenge to consider is the redundancy of data. Redundancy is a common problem in biological databases, and the PDB is not an exception. The reasons can be several e.g. a single protein might be determined at different resolutions or in presence of different ligands, and for each result a new entry in the PDB is generated. In any case, redundant data over-represent cases and 'overwhelm' those with smaller numbers, therefore, affecting an analysis.

Protein clustering based on peptide sequence is a very helpful tool to remove redundancy [LJG01]. Once the data have been organized in groups that share different levels of protein sequence identity, only one representative per group is considered for analysis. The decision of which identity level to use depends mostly on the type of analysis pursued. In any case, the amount of data available is decisive. For example, the glycan residues selected for analysis in this study (D-Mannose, D-Glucose, D-Galactose, D-Neup5Ac, D-GlcpNAc and L-Fucose) appear in a comparably large number of PDB entries. Other monosaccharides do not have such a prominent role on biological events or their structures have not been yet determined frequently, leading to a lack of data. Furthermore, clustering reduces dataset size, which may challenge the statistical significance of the results.

The datasets in the present study have been clustered at 80% identity level, which implies that no two peptide sequences in the dataset are more than 80% identical. The unclustered datasets are affected by redundancy at different degrees. In a few cases, a different pattern of amino acid presence emerges ( $\alpha$ -D-Manp,  $\alpha$ -L-Fucp, sulfated residues, see Figure 3.6), but in general they turned out to be already reliable.

The removal of redundancy is still helpful to 'discover' hidden patterns. For instance, the positive deviation of Lys in the sulfated residues reveals the bias that redundancy can cause to an analysis. Figure 3.6 T1 shows that the deviation of Lys in the unclus-

tered dataset is predominant, nevertheless, in the clustered dataset the value of Lys is still high, but lower than His, Arg and Trp. The outstanding value of Lys in Figure 3.6 T1 is then certainly caused by the over-representation of Lys in the dataset due to the redundancy of data.

The results presented in this work depict relevant information on protein-carbohydrate interactions. In the rest of this section, these findings are shown to be in concordance with previous reports. Therefore, the utility of the PDB as indirect source of carbohydrate data is considered to have been proved.

## 4.2 Carbohydrate chemistry and interaction with proteins

The second objective set for this thesis was to determine the influence that carbohydrate chemistry exerts on protein-carbohydrate interactions. In order to reach this goal, deviations from natural occurrences were calculated for essential monosaccharides (D-Manp,  $\alpha$ -L-Fucp,  $\alpha$ -D-NeupAC, D-Glcp,  $\beta$ -D-GlcpNAc, D-Galp, sulfated residues) non-covalently linked to proteins.

Monosaccharide units are able to adopt distinct conformations; however, they are in fact rather rigid. The conformation adopted by monosaccharide residues defines the specific positioning of hydroxyl groups in the molecule: adjacent hydroxyl groups create large polar patches in the monosaccharide where hydrogen bonds are mostly established whereas non-polar patches tend to bond with aromatic amino acids. Small changes in stereochemistry can create distinct sets of interactions and therefore, delineate the remarkable specificity that characterizes the complexes between proteins and carbohydrates. C-type lectin receptors, for instance, are able to interact with a broad type of pathogens for their posterior degradation and antigen presentation through the specific recognition of mannose (expressed by viruses, fungi and mycobacteria), fucose (expressed by certain bacteria and helminths) and glucan structures (expressed by mycobacteria and fungi). This remarkable specificity is also observed in the Galectins, a family of proteins that share core sequence similarity and precise affinity for  $\beta$ -Galactoside residues [CB99].

Due to this specialized recognition ability, very variable deviations from natural occurrences for each monosaccharide were expected. However, this was not the case. Yet, these deviations can still be used to draw some conclusions. For example, polar amino acids are in general over-represented and non-polar amino acids, with exception of the aromatic residues, are in general under-represented. The main exception regarding the former claim is Lys. In fact, except for the sulfated residues, the deviation of Lys is always negative. The negative charge of  $\alpha$ -D-Neup5Ac could have influenced the over-representation of the positively charged Lys, as it occurs with His and Arg (also positively charged). However, this did not occur. Thus, the underrepresentation of Lys suggests that electrostatic charge might not be the main reason behind its values. Gln and the negatively charged amino acids also present negative deviations with some monosaccha-

ride residues. Among the aromatic residues, Trp is remarkable since its presence varies greatly (from around 900% for  $\beta$ -D-Manp to above 400% for  $\alpha$ -L-Fucp (Figure 3.6 I2,J2). Trp is clearly important in protein-carbohydrate interactions since it has the largest positive deviations in most cases.

As mentioned above, calculation of deviations from natural occurrences was not precise enough to reflect specificity. Therefore, an analysis at atomic level was performed. Figure 3.7 summarizes the results.

## Hydrogen bonds

It is not a surprise that carbohydrate recognition is to some extent achieved through the establishment of multiple classical hydrogen bonds. Hydrogen bonds are dynamic, since they are strong enough to provide stability to the complex but allow also a fast disengagement. They are highly directional, which is crucial for specificity [Qui89]. The importance of hydrogen bonding for protein-carbohydrate interactions is demonstrated by the prominent presence near binding sites of amino acids with planar polar side chains (Arg, Asn, Gln, His, Trp and Tyr). Planar residues are capable to engage in three different types of hydrogen bonding [Qui89], which explain in part their frequency in the vicinity of sugar residues.

The third column in Figure 3.7 depicts the superimposition of all side chain polar atoms (nitrogen and oxygen) in the database, which have the potential to establish hydrogen interactions with carbohydrate residues. The clouds of polar atoms (from Arg, Lys, His, Ser, Thr, Asn, Gln, Glu, Asp) adjust to the polar patches formed by the hydroxyl groups of the glycan. For example, two polar patches are formed in  $\beta$ -D-Galp, one with C1, C2 and C3 and the second with C3, C4 and C6 (Figure 3.7 I3). Accordingly, most polar atoms gather around these patches, whereas the other side of the ring (the B-face) is less crowded. The hydroxyl groups of  $\alpha$ -D-Galp are situated in such a way that despite not being oriented towards the same direction are sequentially close to each other in a sort of spiral and the polar atoms follow also this kind of arrangement.

$\alpha$ -D-Fucp is a homomorphous sugar of  $\alpha$ -D-Galp, but it lacks a hydroxyl group in C6 in comparison to the latter. Unlike most of carbohydrate residues, Fucose is normally found in the L form. In Figure 3.7 C3 it can be observed that most of the potential polar interactions of L-Fucose are found around the hydroxyl groups in C2, C3 and C4 but hardly in the vicinity of C6, where no hydrogen bond donor or acceptor is present at the sugar side. The hydroxyl in C1 seems not to attract many atoms, as it happens for  $\alpha$ -D-Galp. The reason behind seems to be that  $\alpha$ -D-Galp is found at the reducing end of carbohydrate chains more often than  $\alpha$ -L-Fucp. In this position the O1 can establish interactions with other residues easier than in other positions in the chain, where other geometrical considerations must be taken into account.

Not only changes in the position of hydroxyl groups affect the creation of polar patches but also other chemical changes. For example, N-acetyl- $\beta$ -D-glucosamine ( $\beta$ -D-GlcpNAc) is a compound derived from the linking of acetic acid to the N2 of Glucosamine. The addition of an amine group and acetic acid prevents the formation of the polar area observed for  $\beta$ -D-Glcp (Figure 3.7 F3). Therefore, the atomic interactions also change. The rotatable bonds of the C-N group that binds the acetic acid make the molecule very flexible. Moreover, the sp<sup>2</sup>-hybridization of the acetyl carbonyl carbon to build a double bond to the oxygen atom creates a flat planar configuration that is able to accommodate on different positions in order to reduce repulsion between atoms, contributing in this way to the flexibility of the glycan residue. This might explain why the polar atoms are all spread around and no clear specificity can be observed.

### CH- $\pi$ interactions

CH- $\pi$  interactions have been reported previously, and diverse approaches have been used to study the distinct factors that govern these type of interactions: NMR [VDFA<sup>+</sup>08], [TPC<sup>+</sup>07], [CAV<sup>+</sup>05], IR [SKGS<sup>+</sup>07], force fields [LR05], calorimetric [RGARQG<sup>+</sup>09], Quantum mechanical [KMSK11], [SSB05], [dCFACJBC05], ab initio molecular orbital calculations [TUM11], [TUM09], [SLS<sup>+</sup>05], [KBS11], [SSB04], etc.

CH- $\pi$  bonds are somewhat weaker and less directional than conventional hydrogen bonds. However, by distributing the contacts along the aromatic ring the attraction becomes stronger. Aromatic amino acids have been acknowledged to favor those faces in glycan residues with more axially oriented C-H bonds at one side of the sugar ring [dCFACJBC05]. Therefore, the larger the apolar patch the more attractive it becomes. For example, the amount of Trp atoms that are observed in the vicinity of the C3, C4, C5 and C6 of  $\alpha$ -L-Fucp due to the extended apolar patch they form (Figure 3.7 C4). Moreover, Figure 3.7 I4 clearly shows that the majority of Trp atoms are found packing against the B-face of  $\beta$ -D-Galp. The B-face is constituted by two apolar patches, the first formed by the hydrogens in C1, C3 and C5 and the second by the hydrogens in C4, C5 and C6. A similar case occurs with  $\beta$ -D-Manp (Figure 3.7 B4), where the spatial positions adopted by the hydroxyl groups of C1, C3 and C5 also create a B-face that is preferred by most of Trp atoms. Such a B-face cannot be formed in  $\alpha$ -D-Manp because the different stereochemistry at C1 breaks the hydrophobic patch around this atom, which explains the highly different levels of deviations of Trp for these two epimers (Figure 3.7 A4).

$\beta$ -D-Glcp is an epimer of both  $\beta$ -D-Galp and  $\beta$ -D-Manp, regarding C4 and C2, respectively. The positions of the hydroxyl groups linked to C2 and C4 of  $\beta$ -D-Glcp prevent the formation of a clear A or B face in this residue. In change, two small apolar areas are formed on the upper and lower side of the ring, which seem to attract a similar number of Trp atoms (Figure 3.7 F4). As seen before, chemical changes can affect the way polar patches are created in the molecule. The addition of an amine group and acetic acid

to N-acetyl- $\beta$ -D-glucosamine ( $\beta$ -D-GlcpNAc) contribute to modify the interactions that are seen for  $\beta$ -D-Glcp. Figure 3.7 F4 and G4 depicts the elongation of the polar patch correspondent to the lower part of the ring of  $\beta$ -D-Glcp. In  $\beta$ -D-GlcpNAc there is a slight increment in the number of Trp atoms in this part of the ring compared to the upper side.

The results are in agreement with previous analyses of CH- $\pi$  interactions. For example, in a 3D potential energy surface scan of  $\beta$ -D-Glcp,  $\beta$ -D-Manp and  $\alpha$ -L-Fucp in complex with a benzene molecule, Kozmon *et al* [KMSK11] reported strong interactions with the B-face of  $\beta$ -D-Manp, the apolar area formed by C3, C4, C5 and C6 of  $\alpha$ -L-Fucp and both faces of the  $\beta$ -D-Glcp ring. Similar results were obtained by Kumari *et al* [KBS11] for the interactions of  $\alpha$  and  $\beta$ -D-Glcp and  $\beta$ -D-Manp with an analogue of Trp.

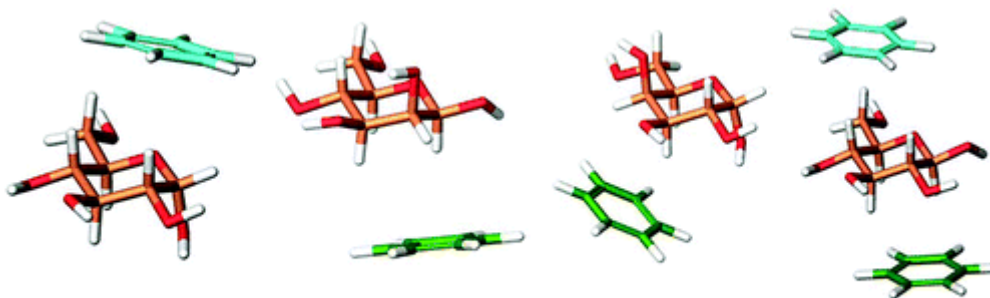
### 4.3 Over-representation of aromatic amino acids in protein-carbohydrate interactions

Hydrogen bonds in protein-carbohydrate interactions are expected to occur due to the hydrophilic nature of carbohydrates. More unexpected are the non-polar interactions established with aromatic amino acids. The third objective set for this work was to gain a better understanding on the reason behind these kind of interactions, how generalized they are as well as the role that Trp plays in this scenario.

The stacking of aromatic residues against the faces of sugar rings was already observed when the first 3D structure of an enzyme was determined (lysozyme-chitooligosaccharide) [AACJB13], Phillips DC 1967. Since then, determination of 3D structures have confirmed the remarkable frequency in which these amino acids are found in carbohydrate binding sites, to such degree that today aromatic amino acids are found in the binding site of most carbohydrate-processing enzymes and carbohydrate-binding proteins [SLS<sup>+</sup>05],[Mur02],[SB04].

During CH- $\pi$  interactions the geometrically complementary apolar surface of an aromatic ring contacts the axially oriented C-H groups of the glycan residue [MGR09] (see Figure 4.1). In this way, glycan structures basically determine the way wherein CH- $\pi$  interactions are established. Along with the hydrogen bonds, CH- $\pi$  interactions are considered essential for the recognition and discrimination of glycan residues in the binding site. For example, the families in the group of Galactose-binding proteins exhibit differences in the primary and secondary structures, in the nature and residues that form their binding sites and also in the range of functions they perform. However, besides an affinity for D-Galp all Galactose-binding proteins share potential hydrogen donor/acceptor groups around O4 and the presence of an aromatic amino acid in their binding sites [SB04]. This residue has been suggested to participate directly in the binding with sugar ligands by forming the base of the binding pocket. Moreover, the aromatic residue grants the galactose residue the freedom to move and adjust its contacts in the binding site [SB04]. Aromatic residues have also been found packing against other carbohydrate

residues [SAD<sup>+</sup>10], [NDZ<sup>+</sup>11]. Furthermore, the mutation of aromatic residues has been proved to cause a loss in affinity for the ligand, being the loss moderated when the mutated residue is also aromatic.



**Figure 4.1:** Examples of CH- $\pi$  interactions from the sugar point of view. Adapted from [AACJB13]

All aromatic amino acids are capable of engaging in CH- $\pi$  interactions. However, the deviations from natural occurrences in Figure 3.6 show that Trp is the preferred aromatic residue for all monosaccharides. The exception are the sulfated residues but its positive deviation is still large. The prevailing role of Trp fits well with previous reports [MZ01], [STARK00], [KW96], [VIGW<sup>+</sup>04]. Trp can establish hydrogen bonds through its N- $\epsilon$  atom, which can act as hydrogen donor or acceptor. However, most of interactions where Trp is involved in the binding sites are stacking. Not as outstanding as Trp, Tyr is also overrepresented in the vicinity of carbohydrate residues. Tyr has polar atoms at the tip of its side chain that can engage in hydrogen bonds, and in this sense, the corresponding pattern in Figure 3.8 is similar to that of polar amino acids. However, the atom distributions in Figure 3.7 show many similarities between Tyr and Trp and also between Tyr and polar amino acids. This suggests that the interactions with Tyr are a mixture of hydrogen bonds and stacking interactions. The role of Phe is much more discrete since this residue is found mostly in the expected amount according to its natural abundance, or slightly under this level in some cases.

Two different computational studies of Galectins, the first an analysis of the conserved Trp [MGR09] and the second an analysis of galactose- and glucose- analogue complexes [SSB05] showed that the mutation of the stacking Trp to another aromatic residue gives rise to a moderated decrease in the affinity for the glycan. The conclusion proposed in both cases was similar: the preference of Trp is not based on energetic advantage but in the size of the aromatic ring. The larger surface area of Trp would allow more freedom to the carbohydrate for fitting into the binding site and enhance its interactions with the rest of residues in the binding site. Similar conclusion was drawn in an examination of the AcAMP2 peptide and three mutants with GlcNAc residues

[CAV<sup>+</sup>05]. It was demonstrated that a larger size of the aromatic ring increased the affinity for the ligand. In accordance with these results, Figure 3.6 shows that despite of being the rarest residue in nature, the bulky Trp is definitely the preferred amino acid in the vicinity of carbohydrates.

## 4.4 Structural patterns and glycosylation

The sequon Asn-xx-Thr/Ser, where xx can be any amino acid except for Pro, identifies potential glycosylation sites. Nevertheless, not all glycosylation sites are occupied and the reason is not yet known. Therefore, in contrast to the previous analysis of non-covalently linked ligands, for N-glycans also peptide sequences were examined. The comparison of results obtained from sequences and spatial vicinities has allowed to speculate about the location of the interacting amino acids in the protein structure.

### 4.4.1 Peptide sequences

The deviations obtained from the analysis of protein sequences show that small hydrophobic amino acids (Val, Ile, Leu) as well as Cys and Gly are over-represented around occupied N-glycosylation sites (Figure 3.14). A large presence of small amino acids in the positions flanking the glycosylated site had been previously reported [PMP<sup>+</sup>04]. They are considered to give more freedom to protein conformation whereas bulky amino acids might restrict access to the glycosylated Asn [PMP<sup>+</sup>04], [KCSE97]. Surprisingly, the calculated deviations for Gly (Figure 3.14 E) and Ala (Figure 3.14 C) are not remarkable despite their size. Except for Gly in position +1, their deviations are just slightly above the expected value according to their frequency in nature.

Besides Gly, a large number of Val in position +1, the xx part of the consensus sequences, had been observed before [PMP<sup>+</sup>04]. Our results confirm these findings, however, the highest deviation in position +1 is observed for Cys (Figure 3.14 E). Its absolute numbers might be small but they are translated into one of the highest and more irregular distribution of deviations. Going up and down, it reaches its highest peaks at +1 and +3, where it surpasses all other amino acids.

Cys in the xx and +3 positions in the consensus sequence has been reported to favor core glycosylation (+3 as the immediate amino acid following a sequon) [MKSSE98], [SESK96]. It is suggested that the hydroxyl or sulfhydryl groups of Cys could be beneficial for catalytic reactions involving the amino acids in the consensus sequences [MKSSE98]. However, Cys can also hamper glycosylation when located in other positions near sequons. The potential establishment of intramolecular disulfide bonding may block the access to the oligosaccharyltransferases which would then inhibit the addition of oligosaccharides.



#### 4.4.2 Spatial vicinity

Contrary to the analysis of peptide sequences, the deviations obtained from the inspection of spatial vicinities show that small hydrophobic amino acids as well as Cys and Gly are under-represented (Figure 3.11). In general, there is a minimal interaction between glycan residues and small amino acids flanking glycosylation sites (See Figure 3.11). The values obtained for Ser and Thr also reflect this result. In order to avoid the bias towards these residues caused by the sequon definition, the Ser/Thr residues closest located to the glycosylated Asn were excluded during the examination of the first  $\beta$ -D-GlcpNAc. However, they were not excluded in the vicinity of  $\alpha$ -L-Fucp. Due to the spatial proximity between of  $\alpha$ -L-Fucp and  $\beta$ -D-GlcpNAc, large positive deviations for Ser and Thr were also anticipated, but this was not the case. The amino acids in the consensus sequence show a low overall interaction with  $\alpha$ -L-Fucp, mainly because of the distance between them being higher than 4 Å. Furthermore, there are cases (such as 1ZHN, 1FE8, 3ALA, 3QWQ, 3S2K, 4EW5 and 4EY8) where  $\alpha$ -L-Fucp is spatially oriented opposite to the consensus sequence, which prevents the establishment of interactions with these amino acids (See Figure 4.2). The Glycan residues more distant from the reducing end (Figure 3.11) show slight increments in Ser/Thr frequency but they are most probably not caused by interactions with Ser or Thr residues of the sequon but with the same type of amino acids located further away from the glycosylation site.

The position that a glycan occupies in the N-glycan core affects the amino acid distribution in its vicinity.  $\beta$ -D-GlcpNAc in Figure 3.11 clearly shows a discrepancy when this residue is located in first and second positions from the glycosylation site. The most remarkable difference is the deviation for Asn, the largest in the first position despite that the glycosylated Asn is not considered in the analysis. The  $\beta$ -D-GlcpNAc in second position shows a much smaller value for Asn but also different frequencies for Tyr, Trp and Arg. Curiously, due to their sizes, Asn, Tyr and Trp would be expected to appear in a lesser extent near the glycosylation sites, since they need more space to accommodate themselves without incurring in steric clashes. However, this was not the case.

#### Comparison between covalently and non-covalently linked carbohydrates

Not only the position in a glycan chain but also the type of linkages established between protein and carbohydrate (covalent or non-covalent) affects their interactions with amino acids. Figure 3.6 and Figure 3.11 display deviations for non-covalently bound ligands and covalently linked N-glycans, respectively. In both cases, with some exceptions, non-polar amino acids are under-represented and polar amino acids are over-represented. However, the distributions vary greatly for each case. Non-covalently bound  $\beta$ -D-GlcpNAc, for instance, shows a very large positive deviation value for Trp (Figure 3.6 N2), which does not occur for the same residue in N-glycans (Figure 3.11). The same variation is seen for  $\alpha$ -L-Fucp. Furthermore, other differences between covalently

and non-covalently linked carbohydrates can be seen in the rest of common residues in both studies. The possible source of such discrepancies is discussed in the following paragraphs.

### **Polar interactions**

As described previously, hydroxyl groups in close distance to each other create large polar patches in the monosaccharide, where potential hydrogen bonds with amino acids can be established. Polar residues can engage in multiple hydrogen bonds and contribute importantly to maintain the structural stability of the N-glycan complex. Accordingly, the deviations of positive and uncharged polar amino acids (Arg, His, Asn, Gln) in spatial vicinities are always or mostly positive but vary notably depending on the glycan residue analyzed.

Despite that amino acids surrounding glycosylation sites could influence the sequon to participate in hydrogen bonding reactions [MKSSE98], the above mentioned polar amino acids do not present remarkable deviations in the protein sequences (See Figure 3.14 A and B). Apparently, the above mentioned case of  $\alpha$ -L-Fucp not interacting with Ser/Thr in the consensus sequon can be generalized, carbohydrate residues in the N-glycan core do not establish hydrogen bonds with the amino acids surrounding the glycosylation sites, but with residues located in other positions in the peptide chain that are brought together during protein folding.

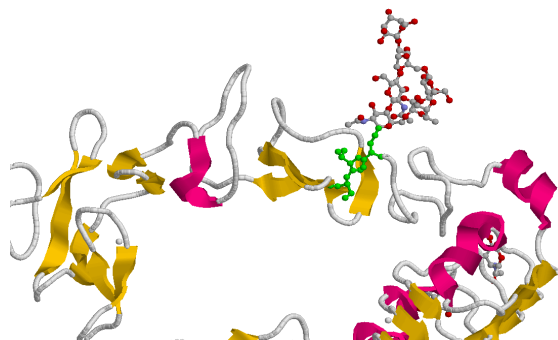
The differences in the distribution of polar amino acids between covalently and non-covalently linked ligands, may arise from the major spatial freedom that the latter have compared with the former. Covalently linked ligands form strong bonds that contribute to maintain the integrity of the N-glycan core. Non-covalently linked ligands, on the other hand, engage in hydrogen bonds mainly to bind in the binding site. These kind of bonds are weaker, which allows a fast disengagement despite that they are larger in number. Covalently linked ligands do not bind to binding sites, therefore, their polar bonds might have the primary function of providing stability to the glycan-protein complex.

### **Aromatic amino acids**

The importance of the interactions between glycan residues and aromatic amino acids has been stated previously. They play a particular function in the sense that they not only establish hydrogen bonds with glycan residues but also participate in stacking interactions. Stacking interactions have an important role stabilizing protein-carbohydrate interactions by directing the oxygen atoms away from the aromatic ring and aligning the CH bonds parallel with the glycan ring.

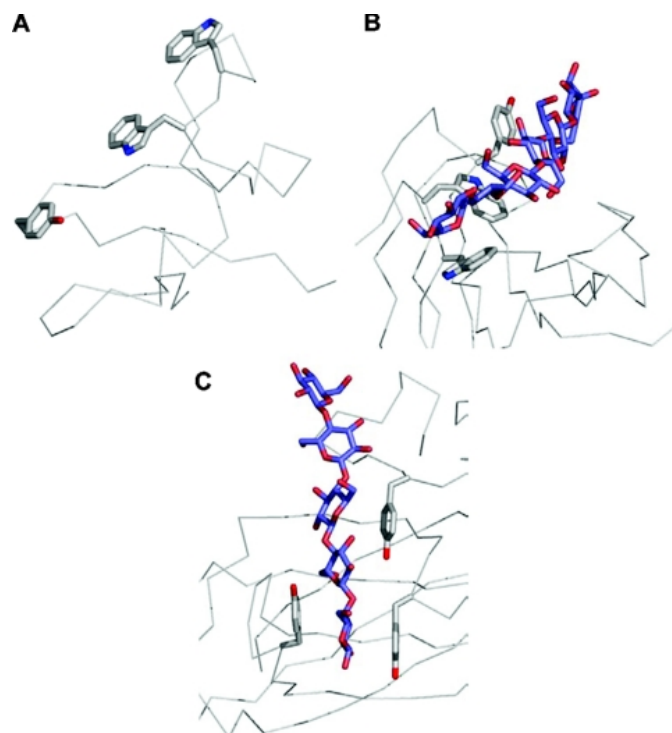
The preference for aromatic amino acids around the glycan residues in the N-glycan core is outstanding. Their deviations are positive near glycosylation sites, especially in the positions preceding the glycosylated Asn, which agrees with previous reports

[PMP<sub>+</sub>04]. As seen above, the values obtained from the analysis of spatial vicinities also show an evident preference over the rest of amino acids. However, it is interesting to notice that the deviations of Trp do not have the same predominant role when analyzing N-glycans, as it occurs with non-covalently linked counterparts (Figure 3.6). The aromatic residues in general still present the highest positive deviations for N-glycans but in this case their respective values are not as dissimilar when compared to each other.



**Figure 4.2:** In some cases the  $\alpha$ -L-Fucp in the N-glycan core does not interact with the Ser and Thr residues of the consensus sequence due to its position facing the opposite site. In the figure an excerpt of 3QWQ.

The crucial need to support the N-glycan structure might be the reason why Trp is preferred near the glycosylated Asn. In this way, the inherent flexibility of glycan chains is reduced and the complex built with protein and N-glycan becomes more stable. However, the deviations of Trp are not as prominent as it occurs for non-covalently linked ligands. Carbohydrate recognition plays an essential role for the latter since carbohydrate-binding proteins must show a great specificity to be able to differentiate even between very similar carbohydrates. Stacking interactions established with Trp contribute not only to protein stability but also to increment binding affinity. Shaping a binding site is not critical for N-glycans. In this case, the stability of the molecule may be more important. The major influence of Trp for non-covalently linked ligands might then be explained by the role that Trp plays when building a binding site. As mentioned before, the relatively large size of Trp provides a platform for the glycan to adjust and better fit into the binding site. This is better illustrated in Figure 4.3 with the a)planar, b)twisted and c)sandwich platforms that Trp provides for sugar binding. This property is not needed for N-glycans as it is stability but this stability can also be provided by the aromatic rings of Tyr and Phe. Furthermore, the smaller size of these residues would be also an advantage to accommodate themselves into the more restrictive spatial conformation of the N-glycan core. This might explain the increment in the deviations of Tyr and Phe, which surpasses those of Trp. The support for the molecule is less necessary at positions further away from the glycosylation site, which also might explain the decreasing deviations for Trp.



**Figure 4.3:** Examples of platforms that Trp provides to glycan residues - a) planar, b) twisted and c) sandwich - for better fitting into the binding site. Adapted from [PMC1133952]

## 4.5 Limitations

The Protein Data Bank contains only a few pure carbohydrate entries. It can be argued that the PDB is focused mostly on proteins but it is estimated that at least 50% of all proteins are glycoproteins [AHS99], and this number does not correspond to the sugar-containing entries found in the database. To this must be added that not all entries that are in fact included in the PDB can be used due to their poor quality, and that the removal of redundancy further decreases dataset numbers. Accordingly, despite that the utility of the PDB as source of data has been discussed above, the main limitation in this study is still the insufficient amount of data. In the presented results (see Figure 3.7) some residues have not been taken into account ( $\beta$ -D-NeupAc,  $\alpha$ -D-GlcpNAc). The analysis of O-glycosylation sites was skipped (the dataset was comprised of 369 Ser and 529 Thr unclustered O-linked glycans. These numbers are reduced when the variability of the O linkage e.g. mucin type, O-GlcpNAc, etc, is also considered). Other residues were included but the small datasets might raise the question if they are statistically significant (Phe for the  $\beta$ -D-GlcpNAc in second position of the N-glycan core, Trp for  $\alpha$ -L-Fucp).

The errors contained in some of the collected entries suggest that the same care is not given to the accompanying carbohydrate when determining a protein structure. Lack of awareness in glycochemistry can also explain this situation. On the other hand, large datasets can make it harder to analyze the results. D-Glcp is the most common carbohydrate among all of the residues considered in this work. The resulting cloud of polar atoms in its vicinity is so crowded (Figure 3.7 E3, F3) that prevents the search for a more specific pattern of interactions.

Most of carbohydrate-containing entries consist of small chains. The great flexibility that oligosaccharides show in solution, which often hampers crystal preparation, may explain in part this particularity. For N-glycans, such flexibility increases for the glycan residues located further away from the reducing end. For non-covalently linked ligands, also the reducing end can be flexible. Accordingly, the analysis of N-glycans was performed only on the N-glycan core. The glycan residues that allow the classification of N-glycans on high-mannose, complex and hybrid were not considered also due to the shortage of data. For the non-covalently linked ligands, the analysis had to be restricted to all residues without specifying their location in the glycan chain.

Aromatic residues contribute to the stability of the protein by establishing stacking interactions with sugar rings. Since the results presented here were obtained from resolved structures of glycoproteins, the question arises if the different levels in the over-representation of aromatic amino acids is a generalized feature, or there is a bias in the dataset towards more rigid structures, which yield a better electron density. If the latter is true, the results obtained in this analysis still can be used to enforce the argument that Trp plays a very important role to maintain the structural stability of the protein-carbohydrate complex.

The statistical approach used for this work reveals patterns of interactions between monosaccharides and proteins. Most of results can be explained by the polar and non-polar patches on the surfaces of glycan residues. However, there can be other factors that affect such interactions and this approach might not give much information about the cause. In such cases, it has been necessary to examine the PDB file directly in order to find a suitable explanation. For example,  $\alpha$ -L-Fucp and  $\alpha$ -D-Galp have been discussed. These residues are structurally similar except for the lack of C6 for the former. Yet C1 does not show the same attraction for  $\alpha$ -L-Fucp than for  $\alpha$ -D-Galp. The position of  $\alpha$ -D-Galp in glycan chains, which occurs more frequently at the reducing end than  $\alpha$ -L-Fucp, seems to explain this observation.

The analysis of N-glycans attempted to find a structural pattern that could differentiate between occupied and non-occupied glycosylation sites. However, this objective was not correctly set, mainly due to the difficulty to clearly define nonoccupancy. For instance, an unoccupied N-glycosylation site may seem so because it was never exposed to the oligosaccharyltransferase or due to the lack of interpretable electron density.

Moreover, deglycosylation enzymes might be employed prior to crystallization, removing possible N-glycans attached. The peptide sequences collected for this work contained potential N-Glycosylation sites which are not occupied but unoccupied glycosylation sites do not ensure definitive unoccupancy, therefore the decision to omit their analysis.

## 4.6 Outlook

This thesis is the result of an interdisciplinary work in the areas of biochemistry, glycobiology and glycobioinformatics. The generated results allow to draw some conclusions about monosaccharides and their specific interactions with amino acids. Even the cases where data are limited contribute to give a hint about atomic patterns observed in these kind of interactions. These findings can be of particular value for the design of models to experimentally study protein-carbohydrate complexes.

Glyvicinity, the software application that conducts the analysis of data is openly available. Therefore, interested users can make use of this tool for their particular work. The adjustment of parameters such as distance, resolution and redundancy thresholds help to refine the results. Future developments of this application may include the possibility to analyze neighbor glycan couples. This option would give a more complete picture of the interaction patterns. Nevertheless, I consider that the most important way in which this work could be improved is the inclusion of new criteria in the data analysis.

The number of PDB entries that contain long glycan chains may remain small due to the technical difficulties to determine these structures. Similarly, the analysis of unoccupied glycosylation sites is unlikely to be feasible with the type of data used. The available data, further reduced by clustering, limited the study of protein-carbohydrate interactions to the analysis presented in this thesis. However, the results have the potential to be much more detailed in other aspects. Advances in the variety and refinement of experimental techniques for the analysis of glycan and protein-glycan structures suggest that the amount of newly generated data can only increase. The PDB, for instance, is continuously updated. Therefore, it is reasonable to think it is a question of time before enough entries are added to the database that allow to complement the data presented in this thesis or include other monosaccharides. As mentioned in Materials and Methods, the entries that are published in the PDB must follow a standard nomenclature that describes in detail the molecule(s). These metadata could be used to take into account other variables for analysis (e.g. techniques used to characterize the protein-glycan complex, type of carbohydrate-binding proteins, or species where the glycans are expressed).

## 4.7 Conclusions

The chemistry of Carbohydrates was a prominent field of research during the first and middle part of the 20th century. However, the prevailing approach back then was to consider these molecules essentially as energy providers or structural/protective elements. Nowadays, complex carbohydrates represent the third alphabet of life, behind DNA and proteins, due to their remarkable function as biological information carriers.

Diverse studies on protein-carbohydrate interactions have been reported. Some of these studies have experimentally determined single protein-carbohydrate complexes. Others have used mutagenic models to observe affinity loss or synthesized new compounds for carbohydrate ligand screening. The approach presented in this thesis consists of the statistical analysis of all 3D structural data on glycoproteins available in the PDB. An objective of this work was to validate the use of such PDB entries as a source of data. The results obtained after curating the input data and removing redundancy produced a valuable outlook into how interactions between glycans and proteins are established. The fact that these results are in concordance with previously reported findings can be taken as proof that the complementary data stored in the PDB do not necessarily have to remain as such but can be exploited to produce new and relevant information.

Position in a glycan chain, carbohydrate modifications (e.g. acetylation) but mainly the arrangements of carbohydrate hydroxyl groups alter polar and non-polar patches on carbohydrate surfaces. This in turn alter the distribution of atoms in their spatial vicinity. Such is the principle behind carbohydrate recognition. The statistical method used in this thesis has successfully captured the concordance between polar and non-polar patches on the surface of carbohydrates and the distribution of polar and aromatic amino acids in their vicinity. Furthermore, this study has contributed to highlight the paramount importance of aromatic amino acids around all of the analyzed monosaccharides. In those cases where the linkages between proteins and glycans are non-covalent, Trp shows the most outstanding positive deviations despite of being the rarest amino acid in nature. The case of covalently linked glycans is different because also Tyr and Phe present high positive deviations. The reason behind this difference seems to be the function that Trp has. In the first case, mainly as a platform for the ligand and to help shaping the binding site. In the second case, as support of the N-glycan chain and protein folding.

The present work demonstrates the use of software tools in glycobiology to generate new knowledge. Glyvicinity, the application employed to analyze the data, is publicly available to support and guide the work of the glycoscientist interested in the topic of protein-carbohydrate interactions. In fact, Glyvicinity has already proved to be useful as predictor or starting point for experimental methods when determining protein-carbohydrate complexes: a) Siebert *et al* [SLW<sup>+</sup>09] made use of Glyvicinity in a docking experiment to determine the arrangement of amino acids that is essential for

lectin function between the *Selenocosmia Huwena* lectin and the HWTX-1 neurotoxin from *S. huwena*. The final arrangement was analyzed in GlyVicinity to find whether the glycan residues in the binding site are involved in similar interactions. b) in a molecular dynamics study Nyholm *et al* [NFK<sup>+</sup>12] analyzed  $\alpha$ -L-Fucp1,2 and  $\alpha$ -L-Fuc1,3/1,4 as the main receptor binding modes of GII.4 noroviruses. They used GlyVicinity to find out the type and frequency in which the studied fucoses engage in similar protein interactions. Other study of protein-carbohydrate interactions with clinical relevance for the peripheral nervous system was performed by Bhunia *et al* [BVE<sup>+</sup>10]. In this study, Glyvicinity was employed to collect information about the interactions between receptors and negatively charged carbohydrate moieties (uronic acids and sulfated carbohydrates). This analysis was part of a model that also included NMR and molecular modeling methods (molecular dynamics simulations as well as molecular docking) to gain more information about the way in which three ligands (the HNK-1 trisaccharide 1 and the cyclic hexapeptides c-(LSETTl) and c-(RTLPS)) and the HNK-1 receptor form a stable complex.



# 5

## References

- [AACJB13] Juan Luis Asensio, Ana Ardá, Francisco Javier Cañada, and Jesús Jiménez-Barbero. Carbohydrate-aromatic interactions. *Acc Chem Res*, 46:946-954, Apr 2013.
- [ABH<sup>+</sup>03] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Keith Roberts, Martin Raff, and Peter Walter. Essential Cell Biology. Garland *Science*, Third Avenue NY, 2003.
- [AHS99] Rolf Apweiler, Henning Hermjakob, and Nathan Sharon. On the frequency of protein glycosylation, as deduced from analysis of the swissprot database. *Biochim Biophys Acta*, 1:4-8, Dec 1999.
- [AJL<sup>+</sup>07] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular Biology of the Cell. Garland *Science*, Third Avenue, NY, 2007.
- [AK09] Kiyoko F Aoki-Kinoshita. Glycome Informatics: Methods and Applications. CRC Press, 2009.
- [AS96] Rivka Adar and Nathan Sharon. Mutational studies of the amino acid residues in the combining site of erythrina corallodendron lectin. *Eur J Biochem*, 239:668-74, Aug 1996.
- [BDERS03] Shifra Ben-Dor, Nir Esterman, Eitan Rubin, and Nathan Sharon. Biases and complex patterns in the residues flanking protein n-glycosylation sites. *Glycobiology*, 14:95-101, Sep 2003.
- [BH79] Ernst Bause and Harald Hettkamp. Primary structural requirements for N-glycosylation of peptides in rat liver. *FEBS Lett*, 108:341-344, Dec 1979.

- [BVE<sup>+</sup>10] Anirban Bhunia, Subramanian Vivekanandan, Thomas Eckert, Monika Burg-Roderfeld, Rainer Wechselberger, Julija Romanuka, Dirk Bächle, Andrei V. Kornilov, Claus-Wilhelm von der Lieth, Jesús Jiménez-Barbero, Nikolay E Nifantiev, Melitta Schachner, Norbert Sewald, Thomas Lütteke, and Hans-Christian Siebert. Why structurally different cyclic peptides can be glycomimetics of the HNK-1 carbohydrate antigen. *J Am Chem Soc*, 1:96-105, Jan 2010.
- [BWF<sup>+</sup>00] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig<sup>1</sup>, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Res*, 28:235-42, Jan 2000.
- [CAV<sup>+</sup>05] M Isabel Chavez, Cecilia Andreu, Paloma Vidal, Nuria Aboitiz, Felix Freire, Patrick Groves, Juan Luis Asensio, Gregorio Asensio, Michiro Muraki, Francisco Javier Canada, and Jesús Jiménez-Barbero. On the importance of carbohydrate-aromatic interactions for the molecular recognition of oligosaccharides by proteins: NMR studies of the structure and binding affinity of acamp2-like peptides with non-natural naphthyl and fluoroaromatic residues. *Chemistry*, 11:7060-74, Nov 2005.
- [CB99] Douglas NW Cooper and Samuel H Barondes. God must love galectins; he made so many of them. *Glycobiology*, 9:979-984, Apr 1999.
- [CBRR12] Jagat S Chauhan, Adil H Bhat, Gajendra P S Raghava, and Alka Rao. Glycopp: a webserver for prediction of N- and O-glycosites in prokaryotic protein sequences. *PloS one*, 7:e40155, Jul 2012.
- [CBV99] T Hema Thanka Christlet, M Biswas, and K Veluraja. A database analysis of potential glycosylating Asn-x-Ser/Thr consensus sequences. *Acta Crystallogr D Biol Crystallogr*, 55:1414-20, Aug 1999.
- [dCFACJBC05] Maria del Carmen Fernandez-Alonso, Francisco Javier Canada, Jesús Jiménez-Barbero, and Gabriel Cuevas. Molecular recognition of saccharides by proteins. insights on the origin of the carbohydrate-aromatic interactions. *J Am Chem Soc*, 127:7379-86, May 2005.
- [dCFADB<sup>+</sup>12] Maria del Carmen Fernandez-Alonso, Dolores Diaz, Manuel Alvaro Berbis, Filipa Marcelo, Javier Canada, and Jesús Jiménez-Barbero. Protein-carbohydrate interactions studied by NMR: from molecular recognition to drug design. *Curr Protein Pept Sci*, 13:816-30, Dec 2012.
- [dSC14] Luis CN da Silva and Maria TS Correia. Plant lectins and toll-like receptors: implications for therapy of microbial infections. *Front Microbiol*, 5:20, Feb 2014.

- [dSRL99] Michelle de Sousa, Lynne M Roberts, and J Michael Lord. Restoration of lectin activity to an inactive abrin b chain by substitution and mutation of the 2 gamma subdomain. *Eur J Biochem*, 260:355-61, Mar 1999.
- [GvH90] Ylva Gavela and Gunnar von Heijne. Sequence differences between glycosylated and non-glycosylated Asn-x-Thr/Ser acceptor sites: implications for *Protein Engineering*. *Protein Eng*, 3:433-42, Apr 1990.
- [HA01] Ari Helenius and Markus Aebi. intracellular functions of n-glycans. *Science*, 291:2364-2368, Mar 2001.
- [Hal94] Herman Van Halbeek. NMR developments in structural studies of carbohydrates and their complexes. *Curr Opin Struct Biol*, 4:697-709, Oct 1994.
- [HR00] Kjell Håkansson and Kenneth B Reid. Collectin structure: a review. *Protein Sci*, 9:1607-1617, Sep 2000.
- [HSFWH12] Sabine Heitzeneder, Markus Seidel, Elisabeth Forster-Waldl, and Andreas Heitger. Mannan-binding lectin deficiency - good news, bad news, doesn't matter? *Clin immunol*, 143:22-38, 2012 2012.
- [iP95] Anne imberty and Serge Pérez. Stereochemistry of the n-glycosylation sites in glycoproteins. *Protein Eng*, 8:699-709, Jul 1995.
- [iTES09] W K Eddie ip, Kazue Takahashi, R Alan Ezekowitz, and Lynda M Stuart. Mannose-binding lectin and innate immunity. *immunol Rev*, 230:9-21, Jul 2009.
- [Jef90] George A Jeffrey. Crystallographic studies of carbohydrates. *Acta Cryst*, B46:89-103, Apr 1990.
- [JM01] Jaak Jaeken and Gert Matthijs. Congenital disorders of glycosylation. *Annu Rev Genomics Hum Genet*, 2:129-51, Sep 2001.
- [JP05] Margaret A Johnson and B Mario Pinto. NMR spectroscopic and molecular modeling studies of protein-carbohydrate and protein-peptide interactions. *Carbohydr Res*, 339:907-928, Apr 2005.
- [KBS11] Manju Kumari, Petety V Balaji, and Raghavan B Sunoj. Quantification of binding affinities of essential sugars with a tryptophan analogue and the ubiquitous role of CH- $\pi$  interactions. *Phys Chem Chem Phys*, 13:6517-30, Apr 2011.
- [KCSE97] Lakshmi Kasturi, Hegang Chen, and Susan H Shakin-Eshleman. Regulation of n-linked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-xaa-Ser/Thr sequons that are poor oligosaccharide acceptors. *Biochem J*, 323:415-9, Apr 1997.

- [KKV02] Elmar Krieger, Günther Koraimann, and Gert Vriend. Increasing the precision of comparative models with yasara nova-a self-parameterizing force field. *Proteins*, 47:393-402, May 2002.
- [KL12] Lara M Kingeter and Xin Lin. C-type lectin receptor-induced nf-kb activation in innate immune and inflammatory responses. *Cell Mol Immunol*, 9:105-12, Mar 2012.
- [KMSK11] Stanislav Kozmon, Radek Matuska1, Vojtech Spiwok, and Jaroslav Koca. Three-dimensional potential energy surface of selected carbohydrates' ch/pi dispersion interactions calculated by high-level quantum mechanical methods. *Chemistry*, 17:5680-90, May 2011.
- [KW96] Anand R Kolatkar and William I Weis. Structural basis of galactose recognition by c-type animal lectins. *J Biol Chem*, 271:6679-85, Mar 1996.
- [KWJ87] Howard A Kaplan, Joseph K Welply, and William J. Oligosaccharyl-transferase: the central enzyme in the pathway of glycoprotein assembly. *Biochim Biophys Acta*, 906:161-173, Jun 1987.
- [LBvdL00] Wolf D Lehmann, Andreas Bohné, and Claus-W. von der Lieth. The information encrypted in accurate peptide masses-improved protein identification and assistance in glycopeptide identification and characterization. *J Mass Spectrom*, 35:1335-41, Nov 2000.
- [LFvdL05] Thomas Lütteke, Martin Frank, and Claus-W. von der Lieth. Carbohydrate structure suite (css): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res*, Jan 2005.
- [LG06] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22:1658-9, Jul 2006.
- [LJG01] Weizhong Li, Lukasz Jaroszewski, and Adam Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17:282-3, Mar 2001.
- [LR05] Alain Laederach and Peter J Reilly. Modeling protein recognition of carbohydrates. *Proteins*, 60:591-7, Sep 2005.
- [Lüt09] Thomas Lütteke. Analysis and validation of carbohydrate three-dimensional structures. *Acta Crystallogr D Biol Crystallogr*, 65:156-68, Feb 2009.
- [LvdL06] Thomas Lütteke and Claus-Wilhelm von der Lieth. The protein data bank (PDB) as a versatile resource for glycobiology and glycomics. *Bio-catal Biotransfor*, 24:147-155, Apr 2006.

- [MC06] Jian-Guo Geng Ming Chen. P-selectin mediates adhesion of leukocytes, platelets, and cancer cells in inflammation, thrombosis, and cancer growth and metastasis. *Arch Immunol Ther Exp*, 2:75-84, Apr 2006.
- [MGR09] Christophe Meyniera, Françoise Guerlesquina, and Philippe Roche. Computational studies of human galectin-1: Role of conserved tryptophan residue in stacking interaction with carbohydrate ligands. *J Biomol Struct Dyn*, 27:49-58, Aug 2009.
- [MHBvI09] Anthony Moran, Otto Holst, Patrick J Brennan, and Mark von Itzstein, editors. *Microbial Glycobiology. Structures, Relevance and Applications*. Academic Press, 2009.
- [MHK<sup>+</sup>11] Mirko Maksimainen, Nina Hakulinen, Johanna M. Kallio, Tommi Timoharju, Ossi Turunen, and Juha Rouvinen. Crystal structures of Trichoderma reesei  $\beta$ -galactosidase reveal conformational changes in the active site. *J Struct Biol*, 1:156-63, Apr 2011.
- [MK84] Ikka Mononen and Erkki Karjalainen. Structural comparisons of protein sequences around potential n-glycosylation sites. *Biochim Biophys*, 788:364-367, Aug 1984.
- [MKSSE98] JL Mellquist, L Kasturi, SL Spitalnik, and SH Shakin-Eshleman. The amino acid following an Asn-x-Ser/Thr sequon is an important determinant of n-linked core glycosylation efficiency. *Biochemistry*, 37:6833-7, May 1998.
- [Mur02] Michiro Muraki. The importance of ch/pi interactions to the function of carbohydrate binding proteins. *Protein Pept Lett*, 9:195-201, Jun 2002.
- [MW97] AD McNaught and A Wilkinson, editors. *IUPAC. Compendium of Chemical Terminology. The 'Gold Book'*. Blackwell Scientific Publications, Oxford, UK, 1997.
- [MZ01] Elizabeth Maranville and Alex Zhu. Assessment of amino-acid substitutions at tryptophan 16 in alpha-galactosidase. *Eur J Biochem*, 267:1495-501, Mar 2001.
- [NDZ<sup>+</sup>11] Anne Line Norberg, Anette I Dybvik, Henrik Zakariassen, Michael Mormann, Jasna Peter-Katalinic, Vincent GH Eijssink, and Morten Sorlie. Substrate positioning in chitinase a, a processive chitobiohydrolase from serratia marcescens. *FEBS Lett*, 585:2339-44, Jul 2011.
- [NFK<sup>+</sup>12] Waqas Nasir, Martin Frank, Chaitanya A K Koppisetty, Göran Larson, and Per-Georg Nyholm. Lewis histo-blood group  $\alpha$ 1,3/ $\alpha$ 1,4 fucose residues may both mediate binding to gii.4 noroviruses. *Glycobiology*, 9:1163-72, Sep 2012.

- [OM06] Kazuaki Ohtsubo and Jamey D Marth. Glycosylation in cellular mechanisms of health and disease. *Cell*, 126:855-67, Sep 2006.
- [PMP<sup>+</sup>04] Andrei J Petrescu, Adina L Milac, Stefana M Petrescu, Raymond A Dwek, and Mark R Wormald. Statistical analysis of the protein environment of n-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology*, 14:103-14, Feb 2004.
- [PN03] Rex A Palmer and Hideaki Niwa. X-ray crystallographic studies of protein-ligand interactions. *Biochem Soc Trans*, 31:973-9, Oct 2003.
- [Qui89] Florante A Quioco. Protein-carbohydrate interactions: basic molecular features. *Pure and Appl Chem*, 61:1293-1306, 1989.
- [REC<sup>+</sup>01] Pauline M Rudd, Tim Elliot, Peter Cresswell, Ian A Wilson, and Raymond A Dwek. Glycosylation and the immune system. *Science*, 291:2370-2376, Mar 2001.
- [RGARQG<sup>+</sup>09] Karla Ramirez-Gualito, Rosa Alonso-Rios, Beatriz Quiroz-Garcia, Aaron Rojas-Aguilar, Dolores Diaz, Jesus Jimenez-Barbero, and Gabriel Cuevas. Enthalpic nature of the ch/pi interaction involved in the recognition of carbohydrates by aromatic compounds, confirmed by a novel interplay of NMR, calorimetry, and theoretical calculations. *J Am Chem Soc*, 131:18129-38, Dec 2009.
- [RLQ98] VSR Rao, King Lam, and PK Qasba. Architecture of the sugar binding sites in carbohydrate binding proteins-a computer modeling study. *Int J Biol Macromol*, 4:295-307, Nov 1998.
- [Rob01] Noel A Roberts. Anti-influenza drugs and neuraminidase inhibitors. *Prog Drug Res*, 56:35-77, 2001.
- [RW10] R Shyama Prasad Rao and Bernd Wollenweber. Do N-glycoproteins have preference for specific sequons? *Bioinformation*, 5:208-12, Nov 2010.
- [SAD<sup>+</sup>10] Xiaoyun Su, Vinayak Agarwal, Dylan Dodd, Brian Bae, Roderick I Mackie, Satish K Nair, and Isaac KO Cann. Mutational insights into the roles of amino acid residues in ligand binding for two closely related family 16 carbohydrate binding modules. *J Biol Chem*, 285:34665-76, Nov 2010.
- [SB04] MS Sujatha and Petety V Balaji. Identification of common structural features of binding sites in galactose-specific proteins. *Proteins*, 55:44-65, Apr 2004.

- [SESK96] Susan H. Shakin-Eshleman, Steven L. Spitalnik, and Lakshmi Kasturi. The amino acid at the x position of an Asn-x-Ser sequon is an important determinant of n-linked core-glycosylation efficiency. *J Biol Chem*, 8:6363-6, Mar 1996.
- [SK14] Susan E Sparks and Donna M Krasnewich. Congenital disorders of n-linked glycosylation pathway overview. *GeneReviews*, Jan 2014.
- [SKGS<sup>+</sup>07] E Cristina Stanca-Kaposta, David P Gamblin, James Screen, Bo Liu and Lavina C Snoek, Benjamin G Davis, and John P Simons. Carbohydrate molecular recognition: a spectroscopic investigation of carbohydrate-aromatic interactions. *Phys Chem Chem Phys*, 9:4444-51, Aug 2007.
- [SLS<sup>+</sup>05] Vojtech Spiwok, Petra Lipovova, Tereza Skalova, Eva Vondrackova, Jan Dohnalek, Jindrich Hasek, and Blanka Kralova. Modelling of carbohydrate-aromatic interactions: ab initio energetics and force field performance. *J Comput Aided Mol Des*, 19:887-901, Dec 2005.
- [SLW<sup>+</sup>09] Hans-Christian Siebert, Shan-Yun Lu, Rainer Wechselberger, Karin Born, Thomas Eckert, Songping Liang, Claus-Wilhelm von der Lieth, Jesús Jiménez-Barbero, Roland Schauer, Johannes FG Vliegthart, Thomas Lütke, and Tibor Kožár. A lectin from the chinese bird-hunting spider binds sialic acids. *Carbohydr Res*, 12:1515-25, Aug 2009.
- [Spi02] Robert G Spiro. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology*, 12:43R-56R, Apr 2002.
- [SSB04] Mannargudi S Sujatha, Yellamraju U Sasidhar, and Petety V Balaji. Energetics of galactose- and glucose-aromatic amino acid interactions: implications for binding in galactose-specific proteins. *Protein Sci*, 13:2502-14, Sep 2004.
- [SSB05] Mannargudi S Sujatha, Yellamraju U Sasidhar, and Petety V Balaji. Insights into the role of the aromatic residue in galactose-binding sites: Mp2/6-311g++\*\* study on galactose- and glucose-aromatic residue analogue complexes. *Biochemistry*, 44:8554-62, Jun 2005.
- [STARK00] Miklos Sahin-Toth, Kauser M Akhoon, Joseph Runner, and H Ronald Kaback. Ligand recognition by the lactose permease of escherichia coli: specificity and affinity are defined by distinct structural elements of galactopyranosides. *Biochemistry*, 39:5097-103, May 2000.
- [TAP12] Morten Thaysen-Andersen and Nicolle H Packer. Site-specific glycoproteomics confirms that protein structure dictates formation of n-glycan type, core fucosylation and branching. *Glycobiology*, 22:1440-52, Nov 2012.

- [TPC<sup>+</sup>07] Giancarlo Terraneo, Donatella Potenza, Angeles Canales, Jesus Jimenez-Barbero, Kim K Baldrige, and Anna Bernardi. A simple model system for the study of carbohydrate-aromatic interactions. *J Am Chem Soc*, 129:2890-900, Mar 2007.
- [TUM09] Seiji Tsuzuki, Tadafumi Uchimarui, and Masuhiro Mikami. Magnitude and nature of carbohydrate-aromatic interactions: ab initio calculations of fucose-benzene complex. *J Phys Chem B*, 113:5617-21, Apr 2009.
- [TUM11] Seiji Tsuzuki, Tadafumi Uchimarui, and Masuhiro Mikami. Magnitude and nature of carbohydrate-aromatic interactions in fucose-phenol and fucose-indole complexes: Ccsd(t) level interaction energy calculations. *J Phys Chem A*, 115:11256-62, Oct 2011.
- [UR06] Kenji Uchumira and Steven D Rosen. Sulfated l-selectin ligands as a therapeutic target in chronic inflammation. *Jtrends Immunol*, 12:559-65, dec 2006.
- [VCE<sup>+</sup>09] Ajit Varki, Richard D Cummings, Jeffrey D Esko, Hudson H Freeze, Pamela Stanley, Carolyn R Bertozzi, Gerald W Hart, and Marilynn E Etzler. Essentials of Glycobiology. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2009.
- [VDFA<sup>+</sup>08] Sophie Vandenbussche, Dolores Diaz, Maria Carmen Fernandez-Alonso, Weidong Pan, Stephane P Vincent, Gabriel Cuevas, Francisco Javier Canada, Jesus Jimenez-Barbero, and Kristin Bartik. Aromatic-carbohydrate interactions: an NMR and computational study of model systems. *Chemistry*, 14:7570-8, May 2008.
- [vdLLE09] Claus-Wilhelm von der Lieth, Thomas Lütteke, and Martin Frank (Eds). Bioinformatics for Glycobiology and Glycomics: An Introduction. John Wiley and Sons, Chichester, UK, 2009.
- [VIGW<sup>+</sup>04] Jose Luis Vázquez-Ibar, Lan Guan, Adam B. Weinglass, Gill Verner, Ruth Gordillo, and H Ronald Kaback. Sugar recognition by the lactose permease of escherichia coli. *J Biol Chem*, 279:49214-21, Nov 2004.
- [TYKUGO] Navneet K Tyagi, Azad Kumar, Pankaj Goyal, Dharmendra Pandey, Wolfgang Siess and Rolf K. H. Kinne. D-Glucose-recognition and phlorizin-binding sites in human sodium/D-glucose cotransporter 1 (hSGLT1): a tryptophan scanning study. *Biochemistry*, 46(47):13616-28, nov 2007.
- [WP09] Taia T Wang and Peter Palese. Universal epitopes of influenza virus hemagglutinins? *Nat Struct Mol Biol*, 16:233-4, Mar 2009.
- [YL05] Aixin Yan and William J Lennarz. Unraveling the mechanism of protein n-glycosylation. *J Biol Chem*, 280:3121-4, Feb 2005.



# 6

## Appendix

THE DETERMINATION AND CLASSIFICATION OF O-ANTIGENS is a very complex task. The repertoire of glycan residues, the order in which they are assembled and the type of glycosidic linkages built between them account for a great diversity of O-antigen structures. ECODAB is a web-based application primarily built to support the collection of O-antigens, glycosyltransferases (GTs) and related data involved in the assembly and identification of O-antigen polysaccharides in *Escherichia Coli* (*E.coli*). Through the comparison between GT sequences of known and unknown function, ECODAB is also able to suggest GT sugar donor/acceptor specificity. Therefore, the application has evolved from being a repository to generating data on its own.

Originally, the repository of data in ECODAB was built as a file-system where every entry was stored as an independent file and represented a single serogroup. The files often included duplicated data, such as the description of the content, shared literature or the parameters used in NMR experiments, which contributed to a considerable increment of the database size with each new entry. Moreover, further development and updating resulted very difficult due to the rigidity of the system, and despite that storing and retrieval of data was ensured, the performance of the system was not optimal. ECODAB was completely redeveloped. The migration of the ECODAB repository to a relational database has facilitated the complete automation of the processes to add, remove and modify data as well as present information. Moreover, the overall performance of the application has improved, and the system has acquired a greater flexibility that allows future developments.

The re-development of ECODAB was a parallel project during my studies. The successful conclusion of this work yielded a manuscript that has been published in Glycobiology journal. Due to the lack of common points with my main research, it was decided to include the manuscript as an appendix in this thesis.

Original Article

# Development of the ECODAB into a relational database for *Escherichia coli* O-antigens and other bacterial polysaccharides<sup>†</sup>

Miguel A Rojas-Macias<sup>2,‡</sup>, Jonas Ståhle<sup>3,‡</sup>, Thomas Lütteke<sup>2</sup>, and Göran Widmalm<sup>1,3</sup>

<sup>2</sup>Institute of Veterinary Physiology and Biochemistry, Justus-Liebig-University Giessen, Frankfurter Str. 100, Giessen 35392, Germany, and <sup>3</sup>Department of Organic Chemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden

<sup>†</sup>To whom correspondence should be addressed: Tel: +46-8-16-37-42; Fax: +46-8-15-49-08; e-mail: gw@organ.su.se

<sup>†</sup>Presented in part at the International Carbohydrate Symposium, Madrid, Spain 22–27 July 2012, P236. Development of ECODAB into a relational database.

<sup>‡</sup>These authors contributed equally to this work.

Received 22 September 2014; Revised 21 October 2014; Accepted 21 October 2014

## Abstract

*Escherichia coli* O-antigen database (ECODAB) is a web-based application to support the collection of *E. coli* O-antigen structures, polymerase and flippase amino acid sequences, NMR chemical shift data of O-antigens as well as information on glycosyltransferases (GTs) involved in the assembly of O-antigen polysaccharides. The database content has been compiled from scientific literature. Furthermore, the system has evolved from being a repository to one that can be used for generating novel data on its own. GT specificity is suggested through sequence comparison with GTs whose function is known. The migration of ECODAB to a relational database has allowed the automation of all processes to update, retrieve and present information, thereby, endowing the system with greater flexibility and improved overall performance. ECODAB is freely available at <http://www.casper.organ.su.se/ECODAB/>. Currently, data on 169 *E. coli* unique O-antigen entries and 338 GTs is covered. Moreover, the scope of the database has been extended so that polysaccharide structure and related information from other bacteria subsequently can be added, for example, from *Streptococcus pneumoniae*.

**Key words:** database, ECODAB, *Escherichia coli*, glycosyltransferase specificity, O-antigen

## Introduction

The flagellated rod-shaped Gram-negative bacterium *Escherichia coli* colonizes the human gastrointestinal tract just a few hours after birth, where it becomes an important part of the lower gut flora. Normally innocuous, *E. coli* accompanies its host for life in a mutually beneficial relationship (Kaper et al. 2004). Unfortunately, this placid commensal association can be broken when *E. coli* acquires genetic elements encoding for virulence factors and turns pathogenic (Stenutz et al. 2006). Thereupon, formerly harmless *E. coli* strains may become responsible

for a variety of diarrheal and extra-intestinal infections in humans and animals, some of which have the potential to be fatal. In certain cases, when the immune system of the host turns ineffective or gastrointestinal barriers are violated, even non-pathogenic *E. coli* can provoke infection (Nataro and Kaper 1998; Stenutz et al. 2006).

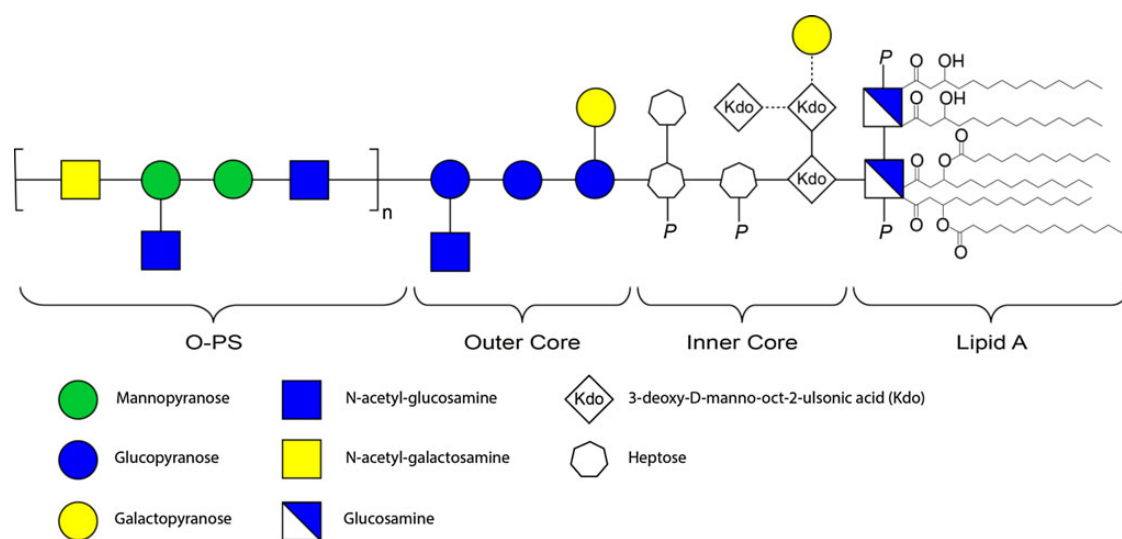
The basis for the serological identification and classification of *E. coli* strains were set by Kauffman in 1944, based on the main antigens that the bacteria express: O (somatic lipopolysaccharide), H (flagellar) and K (capsular) (Ørskov et al. 1977; Ewing 1986). O-Antigens

determine O-serogroups, which have been denoted from O1 to O187 (Stenutz et al. 2006; F Scheutz, 2014, personal communication). In this typing scheme, however, seven serogroups have been removed and certain others have been subdivided, for example, into O5ab and O5ac (Urbina et al. 2005). The combination of O- and H-antigens define a serotype. Only a limited number of laboratories are able to type K-antigens; therefore, the serotyping of O and H is considered the gold standard for subdifferentiation of *E. coli* (Ørskov and Ørskov 1992; Prager et al. 2003). To date, >60 H- and 80 K-antigens have been proposed (Stenutz et al. 2006). O-Antigens coat Gram-negative bacterial cell walls as the outermost exposed domain of polysaccharides (Figure 1), where they serve as receptors for bacteriophages, are involved in immune response, provide various defense mechanisms for the bacteria (Samuel and Reeves 2003; Li et al. 2010) and identify strain-specific surface antigens, some of which are associated with high virulence (DebRoy et al. 2011).

These O-antigens are assembled by a repeated sequence of oligosaccharides called O-units, each consisting of two up to seven sugar residues. In *E. coli*, O-antigens typically range from 10 to ~25 O-units, as determined by NMR spectroscopy or MALDI-TOF mass spectrometry (Linnerborg et al. 1999a,b; Lycknert and Widmalm 2004). Four separate biosynthetic pathways have been determined for the assembly of O-antigens, namely (i) the Wzy/Wzx-, (ii) the ABC-transporter-, (iii) the synthase- and (iv) the Wzk-dependent pathway (Hug et al. 2010), of which only the first two have been found in *E. coli* (Valvano et al. 2011) with the Wzy/Wzx-dependent pathway being the most abundant. The Wzy/Wzx-dependent pathway is typically used for heteropolymers while homopolymers and some disaccharide heteropolymers are formed through the ABC-transporter-dependent pathway (Valvano et al. 2011). Three classes of proteins are involved in the biosynthesis of the O-antigen: (i) proteins responsible for the biosynthesis of sugar units, (ii) glycosyltransferases (GTs) and (iii) O-antigen processing proteins such as Wzy (polymerase), Wzx (flippase) and Wzz (chain length regulating protein); the latter confers the lipopolysaccharide modal lengths of four to >100 repeating units (Kalynysh et al. 2011). The genes encoding these proteins, collectively referred to as the O-antigen gene cluster, are typically located between the *galF* and the *gnd* operon with some exceptions, most commonly

for ABC-transporter-dependent systems, where they are instead located between the *gnd* and the *his* operon. Genes encoding the biosynthesis of common sugars are found outside of the O-antigen gene cluster and usually only the presence of non-common sugars in the O-antigen structure can be predicted by analysis of the genes encoding the sugar residue biosynthesis. The process of adding O-acetyl groups and glucosylation (Wang et al. 2007) is carried out after polymerization of the O-antigen polysaccharide (O-PS) and as such the presence of O-acetyl groups does not impact the GT function specificity. The genes encoding the O-acetyl transferases are not restricted to the O-antigen gene cluster (Allison and Verma 2000; Guo et al. 2005).

The GTs constitute a very diverse family of enzymes responsible for catalyzing glycosidic bond formation by sequentially recognizing and transferring a sugar residue from an activated sugar donor to the respective acceptor (Coutinho et al. 2009). The repertoire of glycan residues, the order in which they are assembled and the type of glycosidic linkages built between them account for the great diversity of O-antigens. To accommodate for all the possible linkages, GTs show a remarkable specificity, which in turn is reflected in their peptide sequences. The precise determination of GTs' specificity is complicated by their vast variation. Moreover, biochemical determination is constrained by the limited availability of activated sugar precursors and acceptors. GT function can also be characterized by the mutation of genes to reveal the accumulated intermediate, but two problems are associated with this method: (i) the accumulated intermediate can damage the cell and (ii) the environment can influence O-antigen structure, as implied by the ability of GTs to carry out reactions with other substrates than their usual (Stevenson et al. 2008). For *E. coli*, a few cases have been reported where there is a discrepancy between the number of glycosidic linkages per repeating unit and the number of GTs. This can possibly be explained by the previously mentioned post-polymerization glucosylation (serogroups O4, O13, O73, inter alia), in some cases enzyme bifunctionality where one GT can catalyze formation of more than one linkage (serogroups O2, O139, O150, inter alia), or for some GTs a lack of an apparent function (Rocchetta et al. 1998; Samuel and Reeves 2003; Wang et al. 2005; Lundborg et al. 2010).



**Fig 1.** Schematic structure of an LPS with sugar residues in CFG-format. The O-antigen corresponds to *E. coli* O6 and the core to R2. The type of sugar residues, the order and the linkages endow O-antigens with a huge diversity.

Bioinformatics tools can be used as a means to overcome the problems with experimental methods. The *E. coli* O-antigen database (ECODAB) was created as a tool to emphasize the particularities of known *E. coli* serogroups and to facilitate the identification of common epitopes (Stenutz et al. 2006). Data on GTs were later included (Lundborg et al. 2010) and protein–protein BLAST searches (Altschul et al. 1990) were implemented with the objective of finding similarities at an amino acid sequence level between GTs with known and unknown functions. In this way, a suggestion on the probable function of GTs was possible, based on amino acid sequence data. Therefore, ECODAB has evolved from simply storing collected data to a database that can be used to produce data on its own. In the present article, a new version of ECODAB is introduced where this process has been completely automated. The sequence comparisons are also no longer restricted to those from *E. coli*, but could also include sequences from other bacteria. Moreover, the data previously stored in a file-based repository have been migrated into a relational database in order to optimize the overall system functionality and enable the possibility of future developments.

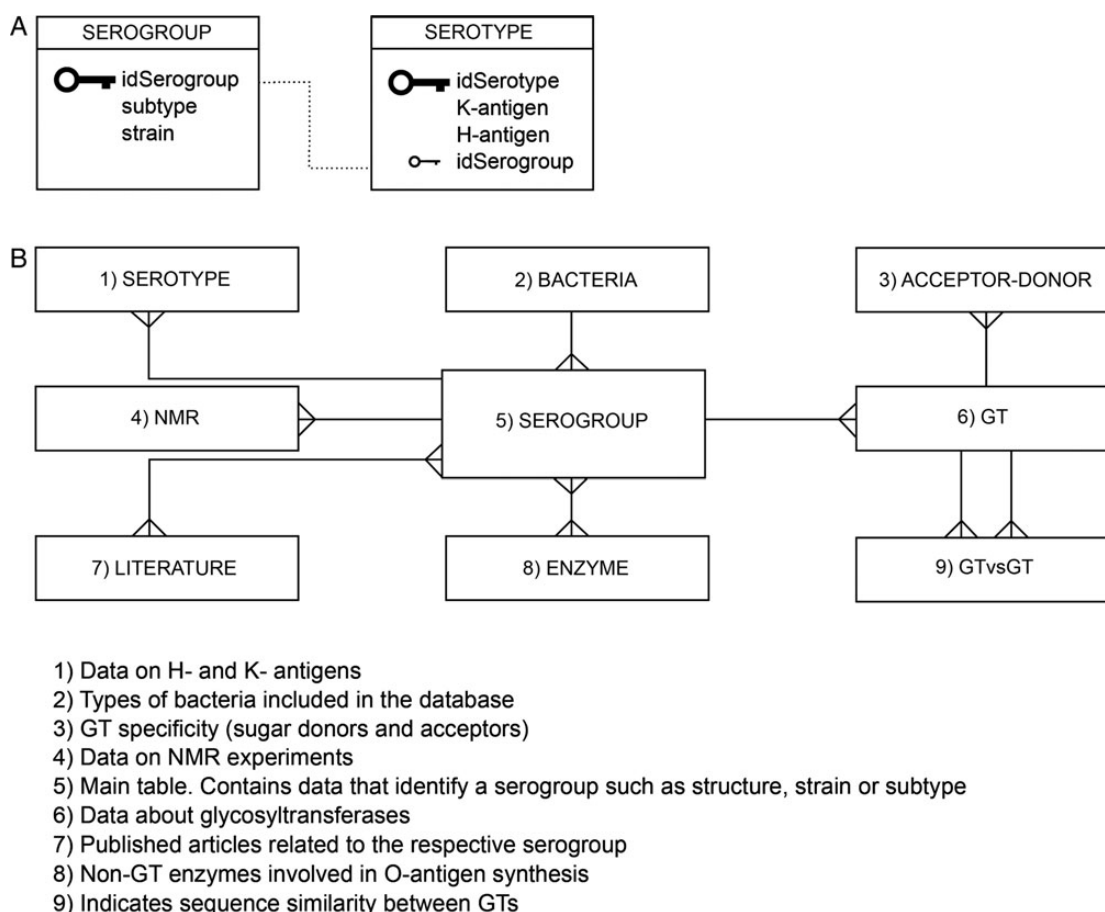
In a relational database, the major entities in a system are modelled to a set of attributes that can be visually depicted as the columns in a table. Every row in the table (a tuple or group of attribute values) is unequivocally identified by a primary key (Figure 2A), an attribute whose values are unique in the table. Relationships are established

by copying the primary key of one entry to a related entry in another table. The primary key in the second table is referred to as a foreign key. When retrieving data, the common values shared by primary and foreign keys are used to combine the columns of different tables in order to reach any value in the database (Figure 2B). However, before being displayed, the data retrieved from the database need to be ordered and processed to generate useful information.

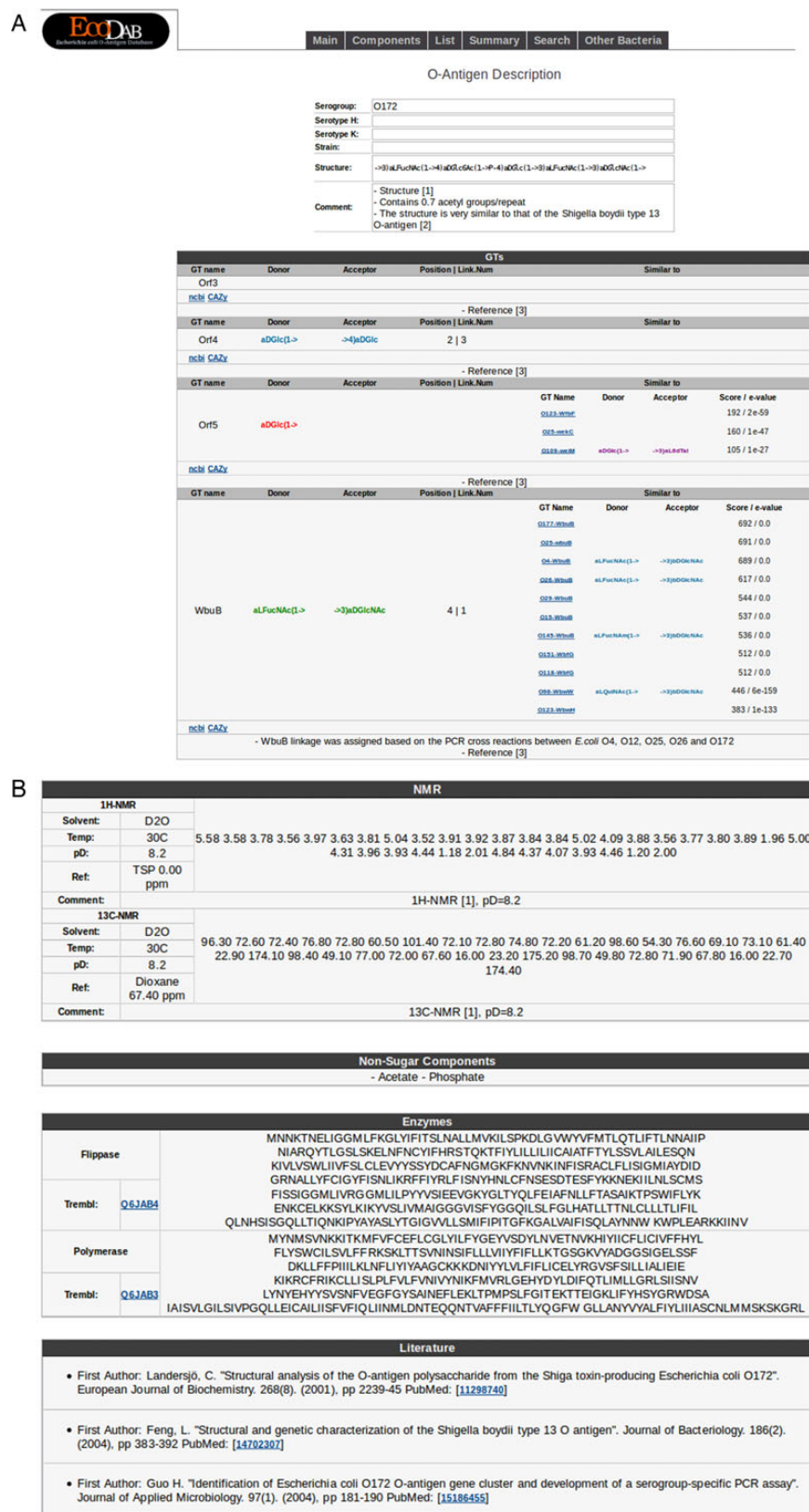
## Results and discussion

In ECODAB, all relevant data regarding serogroups are displayed in only one interface, which is organized in segments where each one describes different aspects about a serogroup: O-antigen description, GTs (Figure 3A), NMR, enzymes, non-sugar components and literature (Figure 3B).

As its name indicates, the O-antigen description section lists the attributes that identify *E. coli* strains, such as H- and K-antigens and the sequence, linkage and branching of O-antigen structures. The GT section condenses all the data on the GTs that have been determined for the corresponding serogroup and provides links to other databases such as NCBI (Benson et al. 2011) and CAZy (Lombard et al. 2014). To facilitate the interpretation of results, a color code is used as follows: (i) green when the function has been experimentally



**Fig. 2.** (A) In a relational database, the data are stored in tables. Every table must include at least one column whose values are unique, a primary key, so that every entry in the table can be unequivocally identified. Primary keys, shown as a larger key, are copied to related entries in other tables, where they are referred to as foreign keys. Relationships between tables are in this way established. The retrieval of data stored in different tables is possible through the combination of entries based on the common keys they share. (B) The blocks represent the main tables in the database and the lines indicate the relationships between them.





determined and published, (ii) blue when the function has been inferred by similarity searches and published, (iii) purple when the function has been manually assigned based on similarity searches and manual inspection but previously unpublished and (iv) the color red is used to indicate that the GT specificity (sugar donors and acceptors) was proposed by ECODAB based on amino acid and O-antigen structure comparison with GTs whose function is known (green, blue and purple).

In certain cases, these predictions are made for GTs that share similarity to a considerable number of other GTs, some of which may still have different functions. Moreover, the assignment of sugar donors and acceptors has to be consistent with the structure of the O-antigen. To overcome this situation, every GT in the system is accompanied by a list of candidate donor and acceptor sugars resulting from the most similar amino acid sequences found in the database. The abovementioned color code is also used to indicate the status of sugar donors and acceptors for each similar GT (Figure 3A). These colors may help to simplify the analysis of the results. For example, when pondering about different alternatives those with green color should be given a higher weighting.

Previous BLAST comparisons involved the manual inspection of a long list of results. The use of a stand-alone installation of BLAST+ to compare amino acid sequences has greatly improved this process since the results are generated in a fully automated way. The results from the sequence similarity search are ordered by decreasing *e*-value and the glycosidic linkages of the highest ranking GTs are compared with the target structure, when available, in order to locate mutual components. The prediction of GT function is then based on the highest ranking acceptor–donor pair with a structural similarity.

The following interface section, NMR, displays the parameters used to determine O-antigen structure by  $^1\text{H}$  and  $^{13}\text{C}$  NMR experiments, including the solvent of the compounds, the temperature of the sample, reference signal, pD values (where  $\text{pD} = \text{pH} + 0.4$ ) and the resulting chemical shift data. The next part lists the non-sugar components linked to the O-antigen structure and the amino acid sequences of genes encoding O-unit flippase and O-antigen polymerase. The latter includes a link to the respective entry in the Uniprot Knowledgebase (Magrane and Consortium 2011). Finally, the literature section lists the most relevant bibliographic information about the particular serogroup in order to support the accuracy of the data; and if available, the PubMed entry is linked.

In addition to the main interface, ECODAB implements two ways to aid the exploration of entries: a summary page and a search option. In the summary page, a table is displayed. Every cell represents a serogroup, and inside the cells a label in the format aa/bb/cc is contained: aa indicates the number of GTs with known or manually assigned function, bb is the number of GTs for each entry and cc is the number of anticipated GTs based on the structure. This table is updated automatically with each new entry in the database. For every cell, a color code is used: (i) green for serogroups whose structures are stored in ECODAB, (ii) blue when the O-antigen structure does not exist in ECODAB, but there is GT information, (iii) red for those serogroups whose structures are not stored in ECODAB and (iv) black for serogroups that have been removed.

The search option facilitates the retrieval of entries that satisfy a set of rules based on substructure, chemical shifts, GT name and/or GENE-BANK code (Benson et al. 2011). All of the parameters manually typed in the interface are used to perform the search query. Accordingly, the more parameters that were used, the more defined the search would become. Substructure search might be the most usual query parameter. In this case, ECODAB will retrieve all entries in the

database where the given matches form part of the O-antigen structure.  $^1\text{H}$  and  $^{13}\text{C}$  NMR chemical shift data can be searched individually. The query values are restricted to two decimals and they must be separated by an empty space. Moreover, the user can select whether to retrieve entries that match exactly the query values or that lie within an interval of 0.05 ppm, for example, a query value of 2.0 will retrieve chemical shifts within 1.98 and 2.02. Finally, searches can also be performed by GT name and/or GENE-BANK code.

Initially, only data on *E. coli* were collected in ECODAB. However, the new version of the application has been expanded to include data also on other bacteria. This change is particularly useful when the lack of resemblance to other GTs present in *E. coli* prevents the prediction of GT function. In these cases, other bacteria that produce analogous O-antigens and whose GT functions have been determined can support the prediction of GT functions for *E. coli*; for example, *Shigella* and *E. coli* are so closely related that some forms of *Shigella* are considered clones of *E. coli* (Lan and Reeves 2002). Hence, sequence comparison between different bacterial organisms can certainly be helpful. Data on other bacteria, for instance *S. pneumoniae*, are displayed on a similar layout as described for *E. coli*, but the interface is accessed separately in order to preserve the focus on *E. coli*.

Since its last iteration, ECODAB has grown to include data on 132 O-antigen structures and 338 GTs of an existing 180 O-antigens. A total of 188 GTs have their function predicted or experimentally determined, of which 94 were automatically predicted by ECODAB, 17 have been experimentally determined, 49 have been inferred and published and 28 were determined through similarity searches in this study (Figure 4). An example of the latter is WelM of *E. coli* O102 that shares a high-sequence similarity to WffH in O130 as well as an identical linkage, namely  $\beta\text{-D-GalNAc-(1}\rightarrow\text{3)-D-Gal}$ , present in both structures. Neither of these GTs share significant sequence similarities with any other GT in ECODAB, and it is therefore concluded that they are responsible for the formation of the aforementioned linkage.

Many GTs share common origins and consequently similar GTs tend to cluster in groups. Being able to infer a specific GT function within a group, which shares sequence similarities where GT functions were previously unknown, gives ECODAB a reference point for making automatic predictions and often increased the number of predicted GT functions considerably. This is the case for the abovementioned pair in *E. coli* O130 and O102 as well as for WfcD of the O141 serogroup. In this study, 12 of the 94 GTs that were automatically

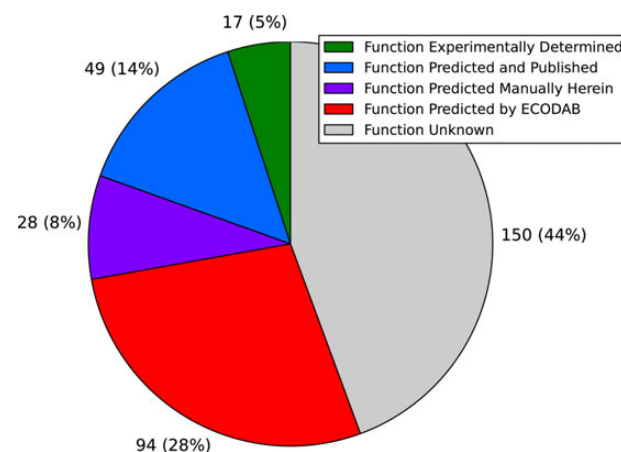


Fig. 4. The function of glycosyltransferases and relative occurrences as stored in ECODAB, based on 338 GT occurrences.

predicted by ECODAB belong to serogroups with an unknown structure and these should be able to help in future structure elucidations. The reduction in the number of GT functions determined by ECODAB compared with previous iterations is due to an increase of known published functions combined with stricter criteria for the automated predictions, rendering some previously inferred GT functions as unknown.

Currently, identification of putative GTs, based on genetic data, is becoming more common, but that is not the case for predictions of function of these proteins. The biochemical determination of GT function is challenging and only a small number of putative GTs has been studied biochemically. ECODAB takes advantage of the available data on GTs and uses this source to suggest the function of hitherto undetermined GTs by amino acid sequence similarity to GTs with known function. The results may not be conclusive, but they can be used to guide future experiments for determining O-antigen structures (Fontana et al. 2014).

The upgrade of ECODAB to implement a relational database has improved the overall performance of the application. Especially, beneficial is the concept of a Database Management System or DBMS (Codd 1990), which handles the way a relational database is created and how the stored data are organized, maintained, retrieved and manipulated. The DBMS also enforces the constraints that govern the relationships between tables and guarantee the integrity of the data by avoiding inconsistencies, duplication, absence, orphanage or obsolescence. Additionally, the DBMS along with the structure of a relational database comprise a flexible programming environment.

In conclusion, ECODAB can be freely accessed at <http://www.casper.organ.su.se/ECODAB/> and together with the Carbohydrate Structure Database (Toukach 2011; Egorova and Toukach 2014) facilitates the retrieval of structural information of O-antigen polysaccharides from *E. coli* in particular and from other bacteria in general.

## Materials and methods

ECODAB has been further developed in order to replace the previous file-system database with a relational database managed by MySQL. The programs to fully automate storage, modification, recovery and display of information were written in PHP. ECODAB can be accessed through a standard Web browser, and additional software does not need to be installed. The database is continuously updated with published data on newly characterized serogroups, mainly from *E. coli*, but the possibility exists to expand the database to include other bacteria.

In order to make predictions on GT functions, amino acid sequence similarity to GTs with determined specificity can be used to suggest a function. In ECODAB, the amino acid sequence of each newly added GT is retrieved from the NCBI portal and stored in the database. The collection of sequences is subsequently used as an input file for a locally installed version of stand-alone BLAST+ (Altschul et al. 1990) in order to perform sequence comparisons.

The *e*-value is a statistical measure of the significance of the results of a BLAST search. A value of 0.001 for example, implies the probability of 0.1% that a given similarity arose by chance alone. The lower the value, the more significant the result. In ECODAB an *e*-value threshold of 0.001 is used when selecting similar GTs and then a bit score of 90 for the display of records.

## Funding

This work was supported by GGL International funded by DAAD and a grant from the Swedish Research Council.

## Conflict of interest statement

None declared.

## Abbreviations

ECODAB, *Escherichia coli* O-antigen database; GTs, glycosyltransferases; LPS, lipopolysaccharide; O-PS, O-antigen polysaccharide.

## References

- Allison GE, Verma NK. 2000. Serotype-converting bacteriophages and O-antigen modification in *Shigella flexneri*. *Trends Microbiol.* 8:17–23.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local assignment tool. *J Mol Biol.* 215:403–410.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2011. GenBank. *Nucleic Acids Res.* 39:D32–D37.
- Codd E. 1990. *The Relational Model for Database Management*. 2nd ed. Reading: Addison-Wesley Publishing Company.
- Coutinho PM, Rancurel C, Stam M, Bernard T, Couto FM, Danchin EGJ, Henrissat B. 2009. Carbohydrate-active enzymes database: principles and classification of glycosyltransferases. In: von der Lieth C-W, Lüttke T, Frank M, editors. *Bioinformatics for Glycobiology and Glycomics: An Introduction*. West Sussex: John Wiley & Sons. p. 91–118.
- DebRoy C, Roberts E, Fraticamo PM. 2011. Detection of O antigens in *Escherichia coli*. *Anim Health Res Rev.* 12:169–185.
- Egorova KS, Toukach PV. 2014. Expansion of coverage of Carbohydrate Structure Database (CSDb). *Carbohydr Res.* 389:112–114.
- Ewing WH. 1986. The genus *Escherichia*. In: Edwards PR, Ewing WH, editors. *Edwards & Ewing's Identification of Enterobacteriaceae*. 4th ed. New York: Elsevier Science. p. 93–134.
- Fontana C, Lundborg M, Weintraub A, Widmalm G. 2014. Rapid structural elucidation of polysaccharides employing predicted functions of glycosyltransferases and NMR data: Application to the O-antigen of *Escherichia coli* O59. *Glycobiology.* 24:450–457.
- Guo H, Kong Q, Cheng J, Wang L, Feng L. 2005. Characterization of the *Escherichia coli* O59 and O155 O-antigen gene clusters: The atypical wzx genes are evolutionary related. *FEMS Microbiol Lett.* 248:153–161.
- Hug I, Couturier MR, Rooker MM, Taylor DE, Stein M, Feldman MF. 2010. *Helicobacter pylori* lipopolysaccharide is synthesized via a novel pathway with an evolutionary connection to protein N-glycosylation. *PLoS Pathog.* 6:e1000819.
- Kalynysh S, Ruan X, Valvano MA, Cygler M. 2011. Structure-guided investigation of lipopolysaccharide O-antigen chain length regulators reveals regions critical for modal length control. *J Bacteriol.* 193:3710–3721.
- Kaper JB, Nataro JP, Mobley HL. 2004. Pathogenic *Escherichia coli*. *Nat Rev Microbiol.* 2:123–140.
- Lan R, Reeves PR. 2002. *Escherichia coli* in disguise: Molecular origins of *Shigella*. *Microbes Infect.* 4:1125–1132.
- Li X, Perepelov AV, Wang Q, Senchenkova SN, Liu B, Shevelev SD, Guo X, Shashkov AS, Chen W, Wang L, et al. 2010. Structural and genetic characterization of the O-antigen of *Escherichia coli* O161 containing a derivative of a higher acidic diamino sugar, legionaminic acid. *Carbohydr Res.* 345:1581–1587.
- Linnerborg M, Weintraub A, Widmalm G. 1999a. Structural studies utilizing <sup>13</sup>C-enrichment of the O-antigen polysaccharide from the enterotoxigenic *Escherichia coli* O159 cross-reacting with *Shigella dysenteriae* type 4. *Eur J Biochem.* 266:246–251.
- Linnerborg M, Weintraub A, Widmalm G. 1999b. Structural studies of the O-antigen polysaccharide from the enteroinvasive *Escherichia coli* O164 cross-reacting with *Shigella dysenteriae* 3. *Eur J Biochem.* 266:460–466.
- Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42:D490–D495.
- Lundborg M, Modhukur V, Widmalm G. 2010. Glycosyltransferase functions of *E. coli* O-antigens. *Glycobiology.* 20:366–368.

- Lycknert K, Widmalm G. 2004. Dynamics of the *Escherichia coli* O91 O-antigen polysaccharide in solution as studied by carbon-13 NMR relaxation. *Biomacromolecules*. 5:1015–1020.
- Magrane M, Consortium U. 2011. UniProt Knowledgebase: A hub of integrated protein data. *Database (Oxford)*. 2011:bar009.
- Nataro JP, Kaper JB. 1998. Diarrheagenic *Escherichia coli*. *Clin Microbiol Rev*. 11:142–201.
- Ørskov F, Ørskov I. 1992. *Escherichia coli* serotyping and disease in man and animals. *Can J Microbiol*. 38:699–704.
- Ørskov I, Ørskov F, Jann B, Jann K. 1977. Serology, chemistry, and genetics of O and K antigens of *Escherichia coli*. *Bacteriol Rev*. 41:667–710.
- Prager R, Strutz U, Fruth A, Tschäpe H. 2003. Subtyping of pathogenic *Escherichia coli* strains using flagellar (H)-antigens: Serotyping versus *fliC* polymorphisms. *Int J Med Microbiol*. 292:477–486.
- Rocchetta HL, Burrows LL, Pacan JC, Lam JS. 1998. Three rhamnosyltransferases responsible for assembly of the A-band D-rhamnan polysaccharide in *Pseudomonas aeruginosa*: a fourth transferase, WbpL, is required for the initiation of both A-band and B-band lipopolysaccharide synthesis. *Mol Microbiol*. 28:1103–1119.
- Samuel G, Reeves P. 2003. Biosynthesis of O-antigens: Genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydr Res*. 338:2503–2519.
- Stenutz R, Weintraub A, Widmalm G. 2006. The structures of *Escherichia coli* O-polysaccharide antigens. *FEMS Microbiol Rev*. 30:382–403.
- Stevenson G, Dieckelmann M, Reeves PR. 2008. Determination of glycosyltransferase specificities for the *Escherichia coli* O111 O antigen by a generic approach. *Appl Environ Microbiol*. 74:1294–1298.
- Toukach PV. 2011. Bacterial carbohydrate structure database 3: Principles and realization. *J Chem Inf Model*. 51:159–170.
- Urbina F, Nordmark E-L, Yang Z, Weintraub A, Scheutz F, Widmalm G. 2005. Structural elucidation of the O-antigenic polysaccharide from the enteroaggregative *Escherichia coli* strain 180/C3 and its immunological relationship with *E. coli* O5 and O65. *Carbohydr Res*. 340:645–650.
- Valvano MA, Furlong SE, Patel KB. 2011. Genetics, biosynthesis and assembly of O-antigen. In: Knirel YA, Valvano MA, editors. *Bacterial Lipopolysaccharides*. Wien: Springer-Verlag. p. 275–310.
- Wang L, Liu B, Kong Q, Steinrück H, Krause G, Beutin L, Feng L. 2005. Molecular markers for detection of pathogenic *Escherichia coli* strains belonging to serogroups O138 and O139. *Vet Microbiol*. 111:181–190.
- Wang W, Perepelov AV, Feng L, Shevelev SD, Wang Q, Senchenkova SN, Han W, Li Y, Shashkov AS, Knirel YA, et al. 2007. A group of *Escherichia coli* and *Salmonella enterica* O antigens sharing a common backbone structure. *Microbiology*. 153:2159–2167.



