UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Biologie Structurale et Nanobiologie**

Arrêté ministériel : 7 août 2006

Présentée par

« Anita SARKAR »

Thèse dirigée par **« Serge PEREZ »** et
Co-dirigée par **« Anne IMBERTY »**

préparée au sein du **Laboratoire Centre de Recherches sur les Macromolecules Vegetales** (CERMAV, CNRS)
dans **l'École Doctorale Chimie et Sciences Vivant**

Facettes de glycobioinformatique : Applications à l'étude des interactions proteines-sucres

Facets of glycobioinformatics : Applications to the study of protein-carbohydrate interactions

Thèse soutenue publiquement le **« 26 Septembre, 2012 »**,
devant le jury composé de :

Prof. Anne MILLET
Président du Jury
Dr. Isabelle ANDRE
Rapporteur
Dr. Alexandre G. de BREVERN
Rapporteur
Dr. Søren B. ENGELSEN
Membre du Jury
Dr. Michaela WIMMEROVA
Membre du Jury

… *to love and the spirit of survival*

Acknowledgements

Acknowledgements

'Not everything that can be counted counts,

and not everything that counts can be counted.'

- *Albert Einstein.*

The first person who I owe my Ph.D. to is my supervisor Dr. Serge Pérez. His belief in me to be able to complete this demanding task of a 3-year thesis, in a field I had no experience in, was the stepping stone whose fruit I shall reap in my career in research. I thank Serge for his friendship, love and care. He gave me the opportunity to explore, to be independent and learn from my mistakes, always guiding me when I strayed. I enjoyed the weekend sessions with him when we could discuss the issues we had through the week, never once making me feel that he was 'the boss'. Thank you Serge for everything!

The next person without whom the story is incomplete is Dr. Anne Imberty. Anne was my shining star throughout my stay in Grenoble. From the moment I entered the city till I was leaving Anne, her pretty smile and her support were my anchors. In hours of darkness, late into the evenings in CERMAV, when I felt that there was no way out, Anne was my shining light. Between Calcutta to Grenoble, I had already travelled two continents, missed my flight in Frankfurt, did not know anybody in France and with no telephone number, with just one conviction that everything is going to work out beautifully and I wanted to take on the adventure that had just begun. Thankfully Anne found me at the 'Gare de Grenoble' with the warmest of smiles that I have had from a stranger, and I knew from that moment that all was well.

Anne, I cannot thank you and Serge enough.

Once in Grenoble, I suddenly realized that a lot had changed. I had no internet, telephone or my English to help me in communicating. But thanks to the ever cheerful and warm French people I survived the first two months in their company seeing the onset of the colourful autumn in the French Alps. Thank you Sophie (Mathieu) for welcoming me to Grenoble with the fantastic home-made ice-cream at Place Grenette, and helping me find my way in the city. Life got so much easier with all your help. I would also like to thank Catherine (Gautier), Valerie (Chazalet), Isabelle (Caldara), Martine (Morales), Magali (Gardes), Martine (Broué) for easing out all the paper-work at all times and being such nice people. I met Michael (Reynolds) and Alessandra

(Nurisso), my first friends in our group of Glycobiologie Moléculaire, who inducted me into this strange life-of-a-Ph.D. student. Mike and Ale, shared all their experiences and knowledge to help me settle-in. Their company was and always shall be enriching and endearing. I just wish we could have spent more time together in CERMAV. Thank you Ali (Ghadban) for being such a nice friend and well-wisher. The breaks we took to share all kinds of stories, opened my eyes to how much can be missed when one is chained to one's computer. I thank Nico(las Sapay) for being so patient and friendly with all my queries and the time-outs we had, just to talk about life in general. I thank Annabelle (Varrot), Olivier (Lerouxel), Christelle (Breton), Emilie (Gillon) for the fantastic times we spent together, for all their help and support, and my friends, Raquel (Benevides), Sumaira (Kousar), Vincent (Grassot), Gaëlle (Batot), Joanna (Rochas), Soorej (Basheer), Dörte (Hundling), Julie (Arnaud), Aymeric (Audfray), Géraldine (Ganne), (Chen) Pan, Bruno (Frka-Petešić), for their support and the memorable times together. I wish you the best in life.

I would thank Sandeep (Srivastava) for the moral support during my initial days in Grenoble. Among all the mails to various people belonging to different associations in Grenoble, he was the first to respond and the most helpful. I shall always remember our trips in town during the weekends, when we tried to explain to one another where we would meet and neither could register the French pronunciation of the other ☺. In about a month's time, I met more Indian (rather Bengali) students living in Grenoble through Sandeep (mostly the crazy physicists) and then there was the beginning of lasting friendships. Inspite of living the major part of my life in the capital of Bengal, before moving out for studies, I frankly did not have so many Bengali friends! The camarederie is one never to be forgotten, that brought me warmth from India when I first saw snow fall on the French Alps. The first friend I made in the Grenobangali community was Akash (Chakraborty). I am happy and lucky to have a dear friend, confidante and support-system in him – a very humble person with many talents (some that I am still discovering). His love and care took me through a lot of rough patches and our travel through many highs and lows has made us thick friends ☺. I wish you the best in life for all your future endevours. Next I would like to thank Soumen (Mandal) for being a dear friend, for all the comic relief and *gyan* (usually) at approriate times, after hard days at work, when all I wanted to do was de-stress. Thank you for the late-night walks and your trust. Though you were introduced to me as "scary" (yes that is true!), you proved to be quite the opposite – big-hearted and mischievous. Thank you to Arpan (Krishna Deb a.k.a. Bhaaitu) for all the fun we had together and the late night cycle rides, your affection and kind words. You truly are multi-talented, multi-facetted and a strong person. Thank you to Maitreyee (Mookherjee) for the excellent desserts, your beautiful company and support. Thank you Dibyendu (Hazra) for your friendship and constant support especially during the nerve-wrecking thesis-writing phase; your cooking skills (especially the concocted chicken) too deserve a special mention ☺. I thank Kalpana (Mandal) for being a caring friend, especially while we were on the same boat (we defended on the same day!) as we could share our frustrations and jubilations together. It is a different feeling when we wish someone "*Congratulations! You are a Doctor now*" and the reply is a "Same to you ☺". A big thank you to Priyadarshini (Chatterjee), Ananda (Shankar Basu), Arpita (Chanda), Roopak (Sinha), Arijit (Roy), Satarupa (Roy Kar), Abhijit (Ghosh), Chandan (Bera), Sutirtha (Mukhopadhyay), Dipanwita (Biswas), Subhadeep (Dutta), Anupam (Kundu), Biplab (Biswas), Ayan (Bandyopadhyay) and Kuheli (Bandopadhyay) for your company, the excellent parties and the numerous moments that made Grenoble my first home away from home. Thank you Arijit da for giving me one opportunity to be superstitious on New Year's when you wish us all well and whatever you wish for each one of us comes true ☺. Thank you Shruti (Sharma) and Nadia (Aziz) for the excellent times we spent together. The one almost serendipitous friend I have made during the French soujourn is *Kaku* (Abhijit Sengupta), strangely, without him even being in France, thanks to Akash. The wish and will to help and befriend people is epitomized through him. Unique is his company and treatment of human relationships- a gem of a person. I am grateful to my childhood friends, especially, Arshi di, Sathya, Shankar da and the many big-hearted people I have known. A big thank you to my dear teachers in school, in college and at the University who nurtured the human being in me and developed me into a tuff nut ☺; Ashish Sir (Dr. Ashish Arora) at the Central Drug Research Institute, Lucknow, who showed me that science can be one's full-time career and exemplifies the grit to survive; Sundar Sir (Dr. D. Sundar) at the Indian Institute of Technology Delhi, with whom I started my journey in science and learnt many a lesson from him about hard-work, dedication, zeal and success in the face of adversity.

The tale of thank you's, running well into the third page, shall remain only a pile of words without mentioning some people whose role in my life is beyond measures: my parents and brother who are the pillars of my being, the love of my grand-parents and a little spark that they saw in me, that led to all this. There are no words in which I can express what I feel about all of you; only that I am sure that I would not have been here, finishing my thesis, being able enough to say the thank you's.

Thank you to my dearest friends, who are with me all through when it rains and when it shines. In chronological order. Seerat: without you being at the back of my mind at all times, life would be drab. No spice, no fun, and no memories. As high the tide may rise, we still are rock solid at the

bottom. Abira: you’re one of the highlights of my trips home, to the city of joy. Strong, steady and beaming, is how I have known you since we were small, and never shall it change. Shayan: sharp, determined and brave, and a friend of all-weathers. Thanks for being there! Arindam: Over the time I know you, you have travelled from one-end of my opinionated friendship metre and come to rest at the other, inhabiting it for a hefty span of time. A bundle of contradictions, and yet so dear. Be my punching bag of goodies forever. Sonu: Despite all that could have happened for us not to meet, I am glad we did. Of the many dreadful things that I associate Delhi with, I can forget everything for this one person it gave me. Without your unwavering support and selfless love, I would not have seen this day. You are the best gift I found for myself ☺. I am blessed to have you in my life and I love you with my being! And Anna: I have NEVER met someone who resembles me so much in so many ways… it is like discovering you have a twin ☺! Anna is a gem and a human being of the highest standard in my ratings. There are many many more awesome adjectives that she deserves. Without her it would be very difficult to reach the finishing line and survive the ‘scare’. Un très grande merci pour toi Anna! Grenoble and you shall remain in my heart and my memory forever like the colours of autum and the flowers of spring ☺.

Thank you Marie Curie Initial Training Network FP7 (People’s Programme) for funding me as an Early Stage Researcher in the EuroGlyoArrays (EuGAs) consortium and giving me this unique opportunity of experiencing science in different countries with different young and experienced scientists. I acknowledge the opportunity that I got through this position to be a Marie Curie almuni. During this stay I not only developed professional skills and contacts, but also strong inter-personal relationships and meet the brilliant minds of our times in science at the Nobel Laureates Meeting in Lindau, Germany. I thank the other EuGAs fellows and senior scientists for the times spent together in science and otherwise. I would like to thank Dr. Göran Widmalm and Carolina Fontana for our fruitful collaboration and would like to say that it was a great experience working with you. Hopefully, we have many collaborations to come from within the EuGAs network in the future. I am grateful to the Centre Nationale de la Recherche Scientifique (CNRS) for funding my last two months of stay in Grenoble.

It indeed is not the amount but the quality of time spent together that forges friendships. Everything else is immaterial.

I am indeed grateful to all that led to my deciding to go to Grenoble, for there I made such good friends and learnt so many lessons from life.

Merci.

Contents

Contents:

Chapters Page

Background

1. Introduction

1.1 Carbohydrates 1

1.2 The third alphabet of life 2

1.3 Glycobiology to glycomics 3

1.4 Glycosylation: a requirement of the cell 3

1.5 Bacteria: The love-hate relationship 6

References 8

2. Glycan databases & encoding

2.1 Glycan databases 9

2.2 Encoding of glycan structures 10

References 16

3. Molecular modeling: a high-throughput technique in glycobiology

3.1 High-throughput techniques in glycobiology 19

3.2 Technique and rationale of glycan 3D structure construction 22

3.3 Carbohydrate force fields 24

3.3.1 Shape 25

3.3.2 Genetic algorithm (GA) and energy evaluation 26

3.3.3 Clustering 29

3.3.4 Filtering 30

3.4 High-throughput molecular modeling 30

References 31

Contents:

Thesis aim & scope 33

Section A

4. *PolySac3DB: An annotated database of 3 dimensional structures of polysaccharides* 37

Background 40

Construction and content

Data sources – screening, conversion and information extraction 42

Data storage 45

Utility and Discussion

PolySac3DB search: Navigation and retrieval 45

Data access 45

PolySac3DB output 47

Beyond the unit cell contents 47

Conclusion 50

Availability 51

References 52

5. Introduction to *O*-antigenic polysaccharides 53

Three-dimensional structural elucidation of O-antigenic polysaccharides 59

a. Three-dimensional structural elucidation of O-antigenic polysaccharides from enteroaggregative pathogenic Escherichia coli strains O5ac and O5ab

Introduction 59

Materials & methods 60

Results & discussions 66

Conclusion 81

References 83

b. *Molecular modeling and conformational analysis of the O-antigenic polysaccharide of the pathogenic Escherichia coli strain 1303* 85

Introduction 85

Materials & methods 87

Results & discussions 87

Conclusion 91

Perspectives 92

References 93

Section B

6. *BiOligo: A 3D structural database of bioactive oligosaccharides* 95

Introduction 98

Scope of the work 99

High-throughput conformational analysis of glycan determinants 101

Construction & content 105

Database query and results 106

Conclusions 109

Computational methods 110

Contents:

References 113

7. Fucose-binding lectins 115

Soluble lectins from opportunistic bacteria binding to human fucosylated epitopes: Screening specificity by docking and glycan array 121

Introduction 121

Materials and methods 124

Results 127

Conclusions 137

References 138

GENERAL CONCLUSION & PERSPECTIVES

General conclusions 141

Review and perspectives 147

Protein-carbohydrate interactions: Computational aspects
(*In* Structural Glycobiology, *CRC Press*)

Annex 197

Annex I: Structure of carbohydrates i – ix

Annex II: PolySac3DB: i – ix

Polysaccharide families, their constituent members and the bibliographic references present in PolySac3DB

Annex III: BiOligo i – xiii

List of bioactive oligosaccharides

Annex IV: CFG glycan descriptions &
Screening specificity by docking and glycan array i – iv

Abstract

Abstract

This thesis presents an account of two important facets of glycobioinformatics, comprising database development and molecular modeling of 3D structures of carbohydrates alongside the simulation of protein-carbohydrate interactions. Classical molecular modeling techniques were used to reconstruct 3D polysaccharide structures from experimentally determined atomic coordinates, or known starting points about their structures were used as guidelines to model them. A genetic algorithm search was employed as a high-throughput technique to characterize low energy conformers of bioactive oligosaccharides. The data generated were organized into two open-access relational databases, namely, PolySac3DB and BiOligo, for use by the scientific community. The validation of the molecular techniques used were performed using solution phase NMR experiments on four enteroaggregative pathogenic *E. coli* strains, and were found to be robust and realistic. Further, the impact of the presentation of human fucosylated oligosaccharide epitopes to lectins from opportunistic gram negative bacteria was investigated in a screening study using molecular docking studies, which could help in evaluating the feasibility of using automated docking procedures in such instances as well as deciphering binding data from glycan array experiments and also correlated to isothermal calorimetry data. On comparison with high-resolution experimental crystal complexes, automated docking was found to delineate the present level of applicability, while emphasizing the need of constant monitoring and possible filtering of the results obtained. Finally, a review of the present status of the computational aspects of protein-carbohydrate interaction studies is discussed in the perspectives of using molecular modeling and simulation studies to probe this aspect of molecular and structural glycobiology.

Résumé

Le travail décrit dans ce manuscrit rassemble les résultats obtenus au cours de ma thèse de doctorat. Ils s'inscrivent dans le domaine de la glycobioinformatique. Ils ont impliqué des développements de bases de données structurales et des applications en modélisation moléculaire des interactions protéines-sucres. Les méthodes de modélisation moléculaire ont été utilisées dans la reconstruction et dans la prédiction des structures tridimensionnelles de polysaccharides et d'oligosaccharides, ces dernières étant également établies par une approche de type "haut-débit" par application d'un algorithme génétique à des fins de minimisation énergétique. Les données ainsi générées ont été organisées sous la forme de bases de données relationnelles, proprement annotées (PolySca3DB et BiOligo) qui sont en libre accès pour consultation sur internet. Ces méthodes de modélisation moléculaire ont été appliquées à la caractérisation, par RMN en solution, des conformations de basse énergie de souches pathogènes d'un polysaccharide de la bactérie *E. coli*. D'autres bactéries pathogènes de type gram négatif, interagissent avec des oligosaccharides par l'intermédiaire de protéines secrétées, telles que des lectines. Nous avons testé, au travers de l'utilisation de méthodes d'amarrage moléculaire, la possibilité d'identifier de manière automatique, la nature de ces interactions, en prenant comme cibles des épitopes oligosaccharidiques fucosylés. Les résultats de ces recherches ont été comparés, de manière critique, à ceux issus de l'application de bio-puces à sucres et de calorimétrie isotherme de titration. Les conclusions et perspectives de ces travaux sont présentées dans un article de revue consacré à l'application des méthodes de chimie computationnelle dans l'étude des interactions protéines-glucides qui viennent compléter l'arsenal des outils dédiés au champs de recherche couvert par la glycobiologie structurale et moléculaire.

Figures & tables

List of figures:

Chapter 1

Figure 1.1 Carbohydrates in the scheme of the molecular paradigm of the central dogma of life.

Figure 1.2 The gram-negative bacteria (*Escherichia coli*, *Pseudomonas aeruginosa* and *Burkholderia*) studied in this thesis. *Figure references* [8-10].

Chapter 2

Figure 2.1 The different levels of glycan encodings (not involving 3D coordinates) handled during this thesis. The example used to illustrate the variety of the notations in this figure is Blood group A Lewis B antigen.

Chapter 3

Figure 3.1 Classical force fields categorized according to their use in carbohydrate chemistry based upon their application class (see colours) indicating that biological relevance dominated trends in this area, according to a survey with data upto April 2010. Figure as published in [17].

Figure 3.2 An illustration of the basic concept of a genetic algorithm search.

Chapter 4

Figure 4.1. Schematic overview of the PolySac3DB organization and content.

Figure 4.2 Cellulose chain conformation and morphology. (A) Crystalline conformations of the cellulose chain in the 1β allomorph showing the disordered orientation of hydroxylic hydrogen atoms. (B) Relative orientation of cellulose chains of native cellulose 1β. (C) Molecular model of the microfibril of cellulose projected along the fibril axis along with the indexing of the surfaces. (D) Computer representation of the crystalline morphology and surfaces of the microfibril of cellulose made up of 36 cellulose chains.

Figure 4.3 Different levels of structural organization in starch. (A) Representation of the left-handed single chains that are parallel stranded in A-starch double helix. (B) and (C) Representations of the double helix of crystalline starch after modeling the branching point between the strands. (D) Computer representation of an ideal platelet nanocrystal showing (i) width of the platelet with the tilt angle of the double helical component, (ii) composition of the platelet and (iii) the enlarged view of the constituent repeating unit.

Chapter 5: *Introduction*

Figure 5 The interactions of *Escherichia coli* and O-antigenic polysaccharides on its surface. (A.) The *E. coli* cell (magnification: 10,000 X) showing the double-layered cell wall packing in all the soluble cellular components [1].

(B.) A magnified (1,000,000 X) portion of the *E. coli* cell illustrating the proteins, nucleic acids, polysaccharides and lipid-membranes. The internal space of the cell is filled with water, glycans, nucleotides, amino acids, metal ions and many other small molecules [1].
(C.) Schematic structure (CFG representation) of an enterobacterial lipopolysaccharide molecule [2]. The lipids are depicted by ribbons attached to GlcNAc in the lipid A part, attached to Kdo, heptoses in the inner core region, hexoses in the outer core region, and finally the O-antigenic components, most commonly hexoses.
(D.) The immune system piercing the *E. coli* cell wall (magnification: 1,000,000 X). Our blood contains proteins that recognize and destroy invading pathogens. This illustration depicts a cross-section through the bacterial cell (lower section of the figure in green, blue and purple) being attacked by the proteins in the blood serum (upper part of the figure in yellow and orange). Y-shaped antibodies recognize and attach themselves to the cell surface setting off a cascade of actions that culminate in a membrane attack complex, shown here, piercing the cell wall of *E. coli* [1].

Chapter 5.a

Figure 5.a.1 Structure of the biological repeating units of the O-antigen PS from a) E. coli O5ac and b) E. coli O5ab in CFG-notation (top), schematic chemical representation (middle) and standard nomenclature (bottom), respectively.

Figure 5.a.2 Illustrated example of the torsion angle conventions used in this study, described using the disaccharide β-D-Gal*p*-1,3-α-D-Gal*p*NAc. The *Heavy Atom Convention* is represented in the *top panel*, while the *Light Atom Convention* is illustrated in the bottom panel.

Figure 5.a.3 Relaxed adiabatic maps of the disaccharide components of the molecular model of O5ac and O5ab. The top panel illustrates the glycosidic linkages that are identical in the two *E. coli* samples, while the lower panel highlights the glycosidic linkages (Gal*p*NAc-α12-Qui*p*3NAc in O5ac and Gal*p*NAc-α14-Qui*p*3NAc in O5ab) that are the distinguishing feature between them.

Figure 5.a.4 Conformation of the ribofuranose ring (residue B) as a function of the puckering parameters Q and φ [27]. The twenty O5ac structures of lower energy obtained from the north and the south starting models are denoted in red and blue, respectively.

Figure 5.a.5 Scatter plots of r_{ij} vs Ψ^H obtained from conformational sampling on the two hexasaccharide models representing the biological repeating unit of the O-antigenic PS from *E. coli* O5ac. The families that explain the experimental data are indicated in red.

Figure 5.a.6 Selected region of the 2D ^{1}H,^{1}H-NOESY spectrum of the O-antigen PS from *E. coli* O5ac recorded at 700 MHz with a mixing time of 80 ms. Correlations from the anomeric protons are indicated with pertinent annotations.

Figure 5.a.7 Plots of the normalized volume intensities versus mixing time obtained for the intra-residue correlation between H1 and H2 of Gal*p*NAc (•), the trans-glycosidic correlation between H1 of Gal*p*NAc and H2 of Qui*p*NAc (▲) and the long-range correlation between H1 of GalpNAc and H4 of Rib*f* (♦). The data was obtained from 2D ^{1}H,^{1}H-NOESY experiments recorded at 700 MHz.

Figure 5.a.8 Scatter plots of r_{ij} vs. ΨDA obtained from the hexasaccharide models representing the biological repeating unit of the O-antigenic PS from *E. coli* O5ab. The conformational families that explain the experimental data are indicated in red.

Figure 5.a.9 (a) ^{1}H-NMR spectrum of the O-antigen PS from *E. coli* O5ab and (b) 1D ^{1}H,^{1}H-SPFGSE NOESY experiment (mixing time 80 ms) with selective excitation of the H1 resonance of Gal*p*NAc and (c) plot of the normalized intensities versus mixing time obtained for the resonance of H2 of Qui*p*3NAc by selective excitation of H1 of Gal*p*NAc in the 1D ^{1}H,^{1}H-SPFGSE NOESY experiments.

Figure 5.a.10 The schematic representation of the polysaccharide helices, describing the major populations observed in O5ac (8 repeating units) and O5ab (8 and 4 repeating units, in the upper and lower panels for O5ab, respectively) (a) as viewed perpendicularly, and (b) along the length of the helix.

Figure 5.a.11 Common epitope observed in the O-antigen PS from *E. coli* O5ac O5ab in schematic chemical (ring) representation of (*top*) and molecular models (*below*). Dashed lines denote the face exposed to the surface of the helix.

Chapter 5.b

Figure 5.b.1 Structure of the biological repeating units of the O-antigen PS from *E. coli* 1303 (with and without O-acetylation on α-L-Fuc*p*) in CFG[1]-cartoon notation (*top*), schematic chemical (ring) representation (*middle*) and linear nomenclature (*bottom*), respectively.

Figure 5.b.2 Relaxed adiabatic maps of the disaccharide components of the molecular models of O1303. The top and middle panels illustrate the glycosidic linkages that are distinct in the two varieties of the biological repeating units, due to the variable substitution on the α-L-Fuc, while the lower panel highlights the glycosidic linkages (Gal*p*-β1→3-Gal*p*NAc and Gal*p*NAc-α14-Qui*p*3ANc) that are common to both the glycans being investigated.

Figure 5.b.3 The schematic representations of the helices formed by the O-antigenic polysaccharides of O1303 and O5ab.

Figure 5.b.4 Comparison of the biological repeat units of O-antigenic polysaccharides O5ab and the O1303. The green part highlights identical stretch between the sequences, while the pink region marks the difference, which is in the substitution of one monosaccharide unit in the biological repeat.

Chapter 6

Figure 6.1 Nomenclature and structural representations commonly used for complex glycans, for example, the pentasaccharide lacto-N-fucopentaose V [Gal β1-3 GlcNAc β1-3 Gal β1-4 (Fuc) α1-3 Glc] in this figure. The relative orientation of two contiguous monosaccharide units in a disaccharide is expressed by two torsion angles Φ and Ψ around the glycosidic bond. According to the *heavy atom convention* (x+1), Φ= O5-C1-O-C_x and Ψ= C1-O-C_x-C_{x+1} for a (1→x) linkage

[1] CFG is an abbreviation for the Consortium for Functional Glycomics.

[*inset: top panel*]. Alternatively, reference to the hydrogen atoms involved in the glycosidic linkage as per the *light atom convention*, can be used Φ^H = H1-C1-O1-C_x and Ψ^H = C1-O1-C_x-H_x, for a (1→x) linkage. For a (1→6) linkage another torsion angle is required and denoted by ω, referring to O5-C5-C6-O6. The sign of the torsion angle is given in accordance with the IUPAC nomenclature [8].

Figure 6.2 The distinct conformations reported in BiOligo after a complete conformational sampling of the lacto-N-fucopentaose V structure.

Figure 6.3 A schema showing the various search modes accessible to the user and results displayed for a query made to the BiOligo database.

Figure 6.4 An illustration of the simple search (*Top*) and advanced search (*Bottom*) search options in BiOligo.

Figure 6.5 An illustration of the results in BiOligo. (*Top*) Preview (*Left*) Molecule information (*Right*) Display and download.

Chapter 7: *Fucose-binding lectins*

Figure 7.a The gram-negative bacteria *Pseudomonas aeruginosa* [1].

Figure 7.b The gram-negative bacteria *Burkholderia* found in roots of plants [4].

Chapter 7

Figure 7.1 Schematic representation of fucosylated trisaccharides and bacterial lectins used in the docking calculations (LecB, BambL and Bc2L-C-nt, from left to right).

Figure 7.2 Selected data from the glycan array v4.1 experiment performed on three bacterial lectins. Only fluorescent results for biding to terminal fucosylated epitopes presented in monovalent manner on glycans have been selected. Blue bar: average value with standard deviation, red bar: maximum response observed.

Figure 7.3 Docking of α-methyl fucoside in the binding site of BambL and LecB. The protein model is represented in red with docked ligand as sticks. The crystal structures of BambL/fucose (3ZW0) and LecB/fucose (1GZT) are represented in green with ligands as lines.

Figure 7.4 Docking of six fucosylated oligosaccharides in the binding sites of BambL. The docking pose with best "glide-score" is represented in red for all oligosaccharides. For the blood group B trisaccharide, the second best orientation is represented in yellow. Comparison with crystal structures is performed with same oligosaccharide when available (H type 1: 3ZW1, H type 2: 3ZZV, blood group B tetrasaccharide: 3ZW2) or elsewhere with fucose (3ZW0), always represented as green line.

Figure 7.5: Docking of six fucosylated oligosaccharides in the binding sites of LecB. The docking pose with best "glide-score" is represented in red for all oligosaccharides. Comparisons with Le^a trisaccharide and fucose monosaccharide are represented with green lines from corresponding crystal structures (1GZT and 1W8H).

Figure 7.6 Attempts to correlate experimental data (ΔH and ΔG) obtained for the interaction of BambL with a series of oligosaccharides and the experimental data (Glide score and Glide energy) obtained from docking.

List of tables:

Chapter 1

Table 1.1: Comparison of the possible structural isomers for nucleic acids, proteins and carbohydrates found in mammals. The numbers are calculated considering both the α and β configurations for the 10 most common mammalian monosaccharides of D-Glc [4], D-Gal [4], D-Man [4], D-Neu5Ac [4], D-GlcNAc [3], D-GalNAc [3], L-Fuc [3], D-Xyl [3], D-GlcA [3] and L-IdoA [3] and the various possible linkage positions. The number of substitutable hydroxyl groups is mentioned in square brackets. Commonly only the pyranose (and not the furanose) forms of these monosaccharides are found in mammals [2].

Table 1.2: Core structures of eukaryotic glycans.

Chapter 2

Table 2.1: Glycan databases arranged according to the information content. The ones boxed in green incorporate glycan structure information, the one in blue include both structure and 3D structure information, while the one boxed in orange provides glycan 3D structure information.

Table 2.2: Glycan databases arranged according to the focus of the database/tool and its respective glycan-encoding format.

Chapter 3

Table 3.1: High-throughput techniques for glyco-analysis: MS [5], HPLC [3], glycan array [6] and molecular modeling.

Chapter 4

Table 4.1. The classification of polysaccharide structures in PolySac3DB.

Table 4.2. Workflow of the informatics tools used in PolySac3DB.

Chapter 5

Table 5.a.1 Averaged torsion angles for each of the conformational families identified in the conformational sampling of the two hexasaccharides, representing the tetrasaccharide biological repeating unit (with the monosaccharide linked to the reducing and non-reducing ends to account for the linkage effect) of the O-antigen PS from E. coli O5ac. The maximum and minimum torsion angle values considered for each conformational family are indicated in square brackets. (a) Associated to the ribofuranose ring in ${}^{3}T_{2}$–${}^{3}E$ (N) conformations. (b) associated to the ribofuranose ring in E_{2}–${}^{3}T_{2}$ (N) or ${}^{2}E$–${}^{4}T_{3}$ (S) conformations. The term "*exo*" denotes those conformations where the *exo*-anomeric effect prevails.

Table 5.a.2 Cross relaxation rates and effective distances determined for the O-antigen polysaccharide from *E. coli* O5ac from 2D ${}^{1}H$,${}^{1}H$-NOESY experiments at 500 and 700 MHz. Calculated distances are reported for the different conformational families identified; the values used to explain the experimental data are highlighted in bold. [a distance used as reference. b Overlapping with D1/D3 and D1/D5. Cross relaxation rates were calculated subtracting the theoretical values obtained from the models. n.d. = not determined due to overlapping. n.o. = not observed or at the noise level. N = north conformation of Rib*f*. S = south conformation of Rib*f*.]

Table 5.a.3 Averaged torsion angles obtained from conformational sampling of the two hexasaccharides representing the biological repeating unit of the O-antigen PS from *E. coli* O5ab.

Table 5.a.4 Cross relaxation rates and effective distances determined for the O-antigen polysaccharide from *E. coli* O5ab from 1D ^{1}H,^{1}H-NOESY experiments at 700 MHz. Calculated distances are informed for the different conformational families identified; the values used to explain the experimental data are highlighted in bold. [a distance used as reference.]

Chapter 6

Table 6.1 The classification of 3D structures of glycan determinants in BiOligo.

Table 6.2 The list of disaccharides included as part of the sub-database GlycoLego incorporated within BiOligo.

Chapter 7

Table 7.1 Glycosidic linkage conformations for all starting model of oligosaccharides. The torsion angles about a glycosidic 1→x linkage are Φ / Ψ, with Φ = O5-C1-O1-C_x and Ψ = C1-O1-C_x-C_{x+1}.

Table 7.2 Docking results obtained for six fucosylated trisaccharides with BambL lectin.

Table 7.3 Docking results obtained for six fucosylated trisaccharides with LecB lectin.

Table 7.4 Titration microcalorimetry data for the interaction between BambL, LecB, BC2LC-nt and fucosylated ligands (all energies in kJ/mol). All data have been measured at least twice and standard deviations are below 15%. ([a] from ref [13], [b] from ref [15], [c] from ref [15])

Background

Introduction

CHAPTER 1

INTRODUCTION

1.1 Carbohydrates

Carbohydrates are the most abundantly occurring organic matter on earth, being present ubiquitously throughout the living world. Complex carbohydrates are built for high-density biocoding at par with proteins and nucleic acids, if not more. An extended paradigm of molecular biology in which biological information flows from DNA to RNA to protein and the role of carbohydrates therein is described in **Figure 1.1**.

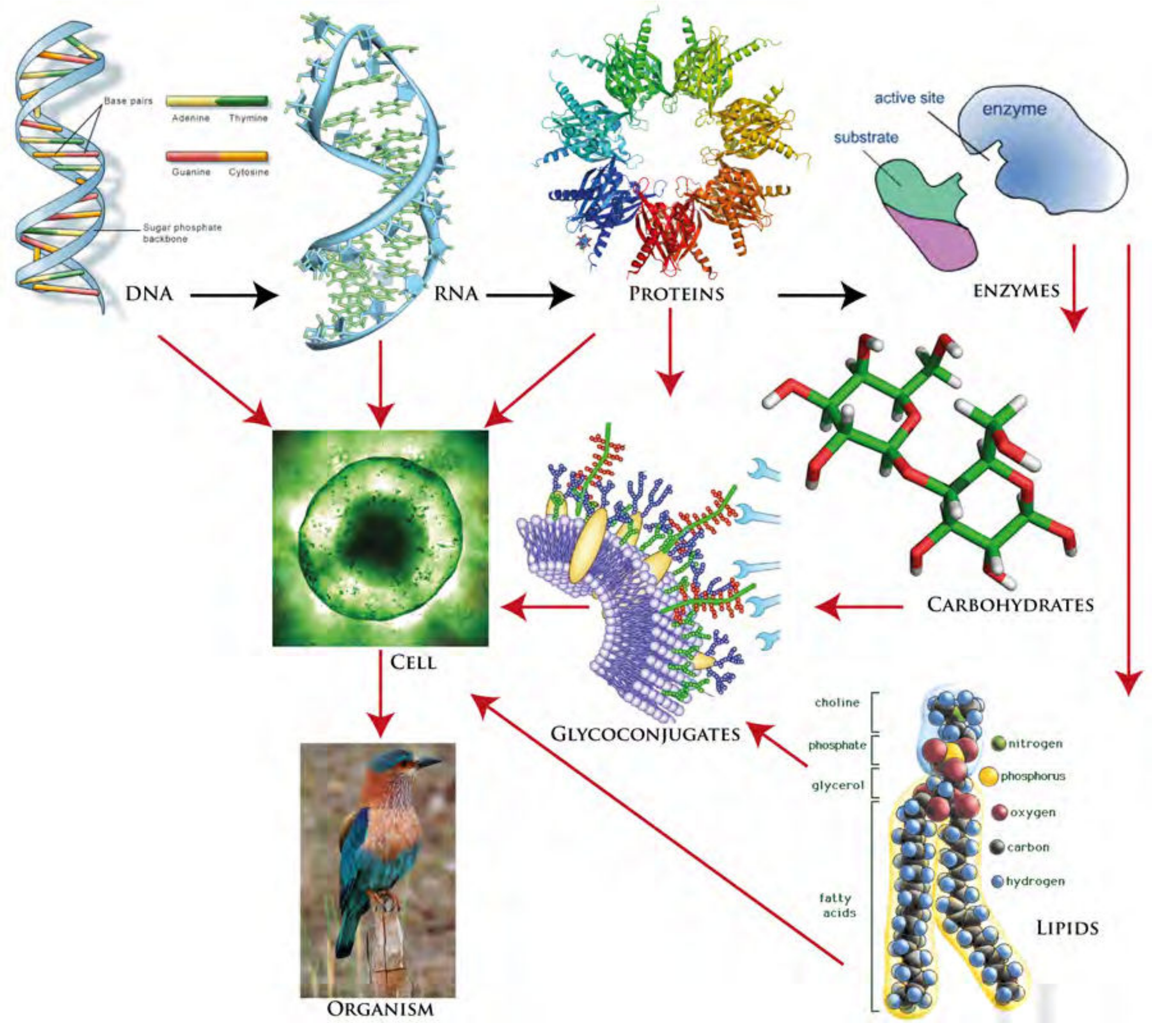


Figure 1.1: Carbohydrates in the scheme of the molecular paradigm of the central dogma of life. *(Inspired from the Essentials of Glycobiology [1])*

Out of the major classes of biological macromolecules, carbohydrates differ from nucleic acids and proteins because they are

a. Either linear or branched.
b. Their constituting monosaccharides are connected through different types of glycosidic linkages (unlike proteins that have amide bonds and nucleic acids that have 3’-5’ phosphodiester bonds). Each monosaccharide can theoretically be present in an α or a β conformation and be linked to any one of several positions on another monosaccharide in a chain or to another type of molecule.

This complexity allows carbohydrates to provide almost unlimited variations in their structures (**Table 1.1**).

Table 1.1: Comparison of the possible structural isomers for nucleic acids, proteins and carbohydrates found in mammals. The numbers are calculated considering both the α and β configurations for the 10 most common mammalian monosaccharides of D-Glc [4], D-Gal [4], D-Man [4], D-Neu5Ac [4], D-GlcNAc [3], D-GalNAc [3], L-Fuc [3], D-Xyl [3], D-GlcA [3] and L-IdoA [3] and the various possible linkage positions. The number of substitutable hydroxyl groups is mentioned in square brackets. Commonly only the pyranose (and not the furanose) forms of these monosaccharides are found in mammals [2].

Size	Nucleotides	Peptides	Carbohydrates
1	4	20	20
2	16	400	1360
3	64	8000	126,080
4	256	160,000	13,495,040
5	1024	3,200,000	1,569,745,920

1.2 The third alphabet of life

Carbohydrates form the third alphabet of life. The high-density coding capacity inherent in oligosaccharides is strongly influenced by its stereochemistry, established by variations in its [3]:

- Anomeric status
- Linkage positions

- Ring size
- Branching
- Introduction of site specific substitutions

An explosive amount of possible glycan structures can be theoretically present in biological systems. But a relatively small fraction of glycans out of this very large number of possible monosaccharide units is observed in naturally occurring biological macromolecules in a limited number of combinations [1].

1.3 Glycobiology to Glycomics

The term **glycobiology** was introduced during the 1980s [4], when it became apparent that a knowledge of sugar decorations was becoming necessary to fully describe biomolecular functions. Glycobiology was thus born at the interface of biochemistry, carbohydrate chemistry and molecular biology with the aim of studying the biosynthesis, structure and biological functions of saccharides (sugar chains/glycans). Later, the conceptual term **glycome** was described in the literature to refer to the complete set of glycan structures synthesized by an organism, at par with proteome and genome [5]. And at the beginning of the 21st century the term **glycomics** has become common, in analogy to genomics and proteomics [6] and emphasizes the holistic view of the total glycan content and its functions in a given organism, cell or tissue.

1.4 Glycosylation: a requirement of the cell

There has been an observation in the scientific community that often altered protein profiles of cells and tissues is a result of an alteration in protein expression rather than modified gene expression. This has been shifting the focus of rigorous research from the genetic code more towards post-translational modifications (PTMs). PTMs in proteins increase manifold the functional diversity of the proteome by modifying the translated protein encoded in the genetic information content. PTMs have a marked influence on all aspects of biological functions and pathogenesis. It is gradually becoming clear in biology that given the surprisingly limited number of genes in the entire mammalian genome, including humans, than what was expected before the completion of the human genome project, PTMs regulating protein function in the phenotype of cells have a much greater role than previously acknowledged. Glycosylation is the most extensive and complex form of the protein PTMs providing for the functional diversity to generate multiple phenotypes from a limited genome. The central dogma of cellular biology can be interpreted to its extended paradigm as shown in **Figure 1.1**. Thus, the study of glycans is critical to the fundamental understanding of cell biology as well as for disease treatment and prevention.

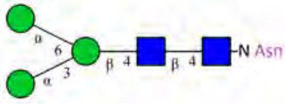
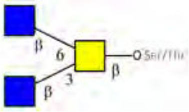
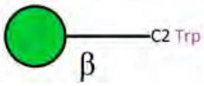
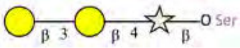
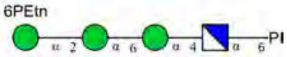
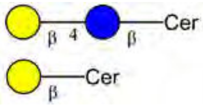
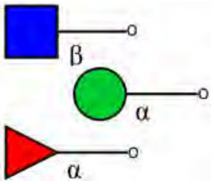
Glycans, like all other components of living cells are constantly recycled. They are assembled and attached by the action of glycosyltransferases. The degradation phase is mediated by very specific enzymes that cleave the sugars either at the non-reducing / terminal end or internally at the reducing end (called exoglycosidases and endoglycosidases, respectively) occurring at the lysosome. These monosaccharide units are then exported from the lysosome to the cytosol for re-use [1].

The variety of information transmission and reception that glycans provide to the genetic code that is translated as proteins, though nothing short of a boon for the cell in equipping it with various possibilities for adapting to changing environments, has been the bottleneck for studying sugars. The solution of this scientific mystery is necessary to fully elucidate nature’s way of biocoding in extremely efficient and compact tools and subsequently usher in a new era of biological understanding and progress.

Glycosylation is highly sensitive to alterations in cellular function, and abnormal glycosylation is indicative and used as a diagnostic technique in diseases like cancer. Through glycosylation, different cells can be labeled with the same recognition markers without having to code it into the genome. Glycosylation patterns differ among glycoconjugates of different species (driven by evolutionary selection pressures) as well as between different cell types of the same organism. Site-specific protein glycosylation suggests that the 3D structure of the protein has an important role to play in fixing the extent and type of its own glycosylation.

It has been found that the actual geometry in which the oligosaccharide presents itself to the receptor is important to trigger a biological response [7]. If the same sugar occurring on multiple proteins does not present itself in the correct geometry to the receptor, the receptor cannot distinguish between the 'self' and 'foreign' sugars.

Table 1.2: Core structures of eukaryotic glycans.

Index	Eukaryotic glycan core structures	Classification	CFG Representation of a typical example
1	N-glycans	ER[1] N-glycans are preassembled and attached *en bloc* to the amine group of asparagine residues	
2	O-glycans	Glycans bound via the -OH groups of a serine or threonine constitute the cores of the different O-glycans, α-GalNAc attached O-glycans are also called '***mucin***-type'	
3	C-glycans	Mannose is bound to the tryptophan via C2 of the indole ring in C-glycans	
4	GAGs	The core structure is attached to the -OH groups of a serine but not threonine	
5	GPI-anchors	GPI anchor- glycans are preassembled on phosphoinositol and then transferred *en bloc* involving a trans-amidation reaction that results in the cleavage of a signal peptide	
6	Glycolipids	The glycan moiety of eukaryotic glycosphingolipids is attached to ceramide (Cer)	
7	Cytoplasmic / Nuclear glycans	These are complex glycoconjugates in the nucleoplasmic and cytoplasmic compartments of the cell like cytosolic N-linked glycans, O-linked glycans and sialic-acid-containing glycans on nucleoporins. A few examples are glycogenin, O-linked GlcNAc, O-linked mannose, O-linked fucose and nuclear GAGs	

[1] ER refers to Endoplasmic Reticulum

The majority of the cell surface and secreted proteins are glycosylated, with glycans being attached via a covalent bond between their reducing end to either a nitrogen atom (from an asparagine) or an oxygen atom (belonging to a serine or threonine). The important core glycan structures found in eukaryotes have been illustrated in **Table 1.2**.

Strikingly, on measurement, it is observed that the distance between contiguous oxygen atoms (O1 and O4) of a carbohydrate residue is ~5.4 Å and that of the ends of the first three residues of an N-linked glycan (α-D-Man*p*-1,3-β-D-Man*p*-1,4-β-D-Glc*p*NAc-1,4-β-D-Glc*p*NAc) is 16 Å from head to tail. Considering the dynamic motions of the glycans, these relatively small oligosaccharides seem to shield large areas of the protein surface keeping in mind that an N-linked sugar typically would have at least two or more arms each comprising three or four monosaccharide units [7]. Moreover, combining the flexibility of the glycan-protein linkage (and its associated small motions) with the rigid glycan core would provide an amplification of the motion of the terminal arms of the glycan that would enable it to span an even larger surface, which could greatly affect the accessibility of the glycoprotein for intermolecular interactions (**Table 1.2**). The knowledge of the 3D structures of these glycoproteins shall correctly quantify such properties and the time evolution of the same, besides shedding light on properties like hydration, hydrophobicity and hydrophilicity. Also, over the same time span, the glycan shows greater dynamic fluctuations than the protein which is highlighted by nuclear magnetic resonance (NMR), thus requiring molecular dynamics (MD) to fulfill the frequent insufficiencies in the data that would add up to a complete conformational analysis.

1.5 Bacteria: The love-hate relationship

Bacteria are small, sleek and self-contained organisms, probably the most successful ones on earth, that are found everywhere, be it in hot springs or freezing water or our gut. They seem to have explored and mastered every possible way to sustain themselves on this planet.

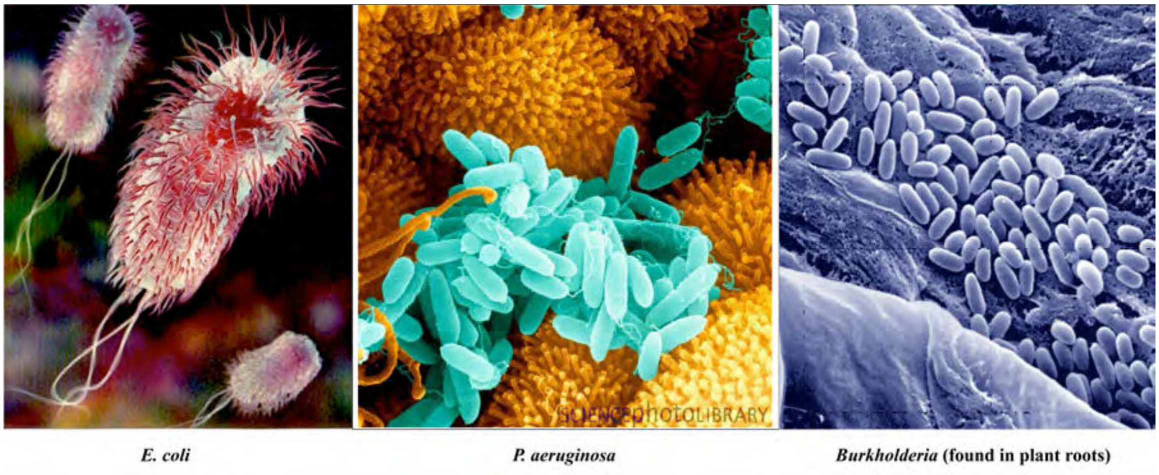


Figure 1.2: The gram-negative bacteria (*Escherichia coli*, *Pseudomonas aeruginosa* and *Burkholderia*) studied in this thesis. *Figure references* [8-10].

They can be non-pathogenic or pathogenic in nature. Bacteria can be classified as gram-positive and gram-negative[2]. The majority (almost 90-95%) of gram-negative bacteria known currently are pathogenic. This thesis includes the application of molecular modeling techniques to study complex glycan structures in three such gram-negative bacteria that cause diseases in humans, diary herds and plants, namely- *Escherichia coli* (*E. coli*), *Pseudomonas aeruginosa* and *Burkholderia ambifaria* (**Figure 1.2)**.

[2] Gram-positive and gram-negative refer to how bacteria react to gram staining. If it takes the initial stain, it will be purple and be considered gram-positive. If it does not take the initial stain, it will be pink and gram-negative. The difference is the outer casing of the bacteria. Gram-positive bacteria will have a thick layer of peptidoglycan (a sugar-protein shell) that the stain can penetrate. Gram-negative bacteria have an outer membrane covering a thin layer of peptidoglycan on the outside. The outer membrane prevents the initial stain from penetrating.

References:

1. Varki A: **Historical background and overview**. In: *Essentials of Glycobiology.* Edited by Varki A, Cummings RD, Esko JD, Freeze HH, Hart GW, Etzler ME, 2nd Edition edn. Cold Spring Harbour (NY): Cold Spring Harbor Laboratory Press; 2nd edition (October 15, 2008); 2008: 784.

2. Werz DB, Ranzinger R, Herget S, Adibekian A, von der Lieth CW, Seeberger PH: **Exploring the structural diversity of mammalian carbohydrates ("glycospace") by statistical databank analysis**. *ACS Chem Biol* 2007, **2**(10):685-691.

3. Gabius H-J (ed.): **The sugar code - Fundamentals of glycosciences**. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim; 2009.

4. Rademacher TW, Parekh RB, Dwek RA: **Glycobiology**. *Annu Rev Biochem* 1988, **57**:785-838.

5. Feizi T: **Progress in deciphering the information content of the 'glycome' – a crescendo in the closing years of the millennium**. *Glycoconjugate Journal* 2000, **17**(7):553-565.

6. Taniguchi N, Ekuni A, Ko JH, Miyoshi E, Ikeda Y, Ihara Y, Nishikawa A, Honke K, Takahashi M: **A glycomic approach to the identification and characterization of glycoprotein function in cells transfected with glycosyltransferase genes**. *Proteomics* 2001, **1**(2):239-247.

7. Dwek RA: **Glycobiology: "towards understanding the function of sugars"**. *Biochem Soc Trans* 1995, **23**(1):1-25.

8. Berger J: ***Pseudomonas aeruginosa* bacteria.** [http://www.sciencephoto.com/media/11601/enlarge]

9. Bioni_USA: **Efficacy of Bioni against *E. coli*, *Listeria* and *P. aeruginosa***. [http://bioniusa.com/2011/02/25/efficacy-of-bioni-against-e-coli-listeria/]

10. Cardiff_University: **Antibiotics from *Burkholderia***. [http://www.futurity.org/health-medicine/cystic-fibrosis-bacteria-fights-mrsa/]

Glycan databases & encoding

CHAPTER 2

GLYCAN DATABASES & ENCODING

About 70% of the protein in sequence repositories have potential N-glycosylation sites (recognized by the occurrence of the Asn-Xxx-Ser/Thr sequon[1]) [1, 2]. According to a rough estimate, more than half of all proteins in the human body are associated with glycans. A large number of carbohydrate sequences have been determined through extensive work in areas of chemical and enzymatic degradation of biomass or bacterial fermentation of glycan structures and their analysis using mass spectroscopy (MS) and nuclear magnetic resonance (NMR). The primary impetus behind the growth of glycoinformatics has been the construction of large-scale repositories to store, organize and disseminate the data that was rapidly being generated through experiments and theoretical calculations in relation to glycan sequence and structure. Additionally, various algorithms and tools have been developed that could query (search) these repositories, interlink them and be useful in further calculations and analyses of the existing data.

2.1 Glycan databases

The complex carbohydrate sequences determined till the year 1997 (~23,000 unique structures and 50,000 entries) were stored in the pioneering glycan database called Complex Carbohydrate Structure Database (CCSD) [3] developed at the Complex Carbohydrate Research Center, University of Georgia. This was created, edited and made searchable using the CarbBank search tool [4] and was the first such attempt to build such a large-scale common database, which would be available for free public use. CCSD comprised published structures of oligosaccharides and glycoconjugates that had three or more glycosyl residues, though excluding mono- and disaccharides as well as synthetic intermediates (as these were already commercially available from the Chemical Abstracts Service at the time). This project was discontinued in the mid-1990s but as the data was

[1] The N-glycosylation sequon is a sequence motif as mentioned above where Xxx can be any amino acid

made available publicly, it became the foundation for other glycan structure[2] databases that followed.

Though there have been major advances in the sequence determination of glycans, the 3D structures of complex glycans lagged behind considerably, due to their inherent complexities and variability[3]. This could become a bottleneck in identifying the functions of glycans in the many biological roles that they are involved in which are directly related to their 3D structures and surface properties.

In modern day science, databases have become an integral part of any research project design and development. Glycobiology has such repositories that store and disseminate data to the scientific community catering mostly to the area of carbohydrate chemistry. To categorize these databases and other useful tools according to the focus of the data contained in each database, we can classify them according to the **Tables 2.1 and 2.2**.

2.2 Encoding of glycan structures

One of the main requirements for databases is to store information in an organized way that facilitates its computational processing. Two approaches can be followed to encode a carbohydrate molecule:

a. *Connecting atom sets through chemical bonds*

   This approach, commonly followed in chemoinformatics and chemical file formats like InChi [18] and SMILES [19] have been developed to aid storing of molecule information in chemical databases like PubChem [20] or ChEBI [21]. IUPAC (extended), InChi and SMILES encoding are computed from the chemical drawing (ring structure) and thus, auto-generation of these encodings is possible. Yet, there are severe limitations that do not make this kind of encoding the favored choice.

[2] Structure, in glycobiology, indicates the branched representation of a glycan sequence, unless 3D is explicitly mentioned. In some places the term ‘sequence’ has been used to maintain coherence.

[3] Stereochemistry (the arrangement of the constituent atoms in space that are connected within the molecule) is the key to the structural characterization of glycans, and the source of their structural complexities and variability.

Table 2.1: Glycan databases arranged according to the information content. The ones boxed in *green* incorporate glycan structure information, the one in *blue* include both structure and 3D structure information, while the one boxed in *orange* provides glycan 3D structure information.

Database	Description	URL	Present Status
CCSD [3] **(Also called CarbBank [4])**	The Complex Carbohydrate Structure Database was an effort to link literature data to carbohydrate sequences. It was discontinued but has been succeeded (& extended) by SugaBase and SweetDB. It has also been incorporated into other databases such as KEGG.	-	Discontinued in 1997
SugaBase [5]	SugaBase is a database that combines structure data from the CCSD with NMR data. SweetDB succeeds it.	http://boc.chem.uu.nl/sugabase/databases.html	Development was stopped in 1998, but entries can still be searched
BCSDB [6]	Bacterial Carbohydrate Structure Database provides structural, bibliographic, taxonomic and other related information on bacterial carbohydrate structures. Sourced from CCSD.	http://csdb.glycoscience.ru/bacterial/	Maintained
KEGG Glycan [7]	The KEGG GLYCAN structure database is a collection of experimentally determined glycan structures. It contains all unique structures taken from CarbBank, structures entered from recent publications & structures present in KEGG[4] pathways.	http://www.genome.jp/kegg/glycan/	Maintained
GlycoSuiteDB [8, 9]	The GlycoSuite database is an annotated & curated relational database of glycan structures and is a product of Tyrian Diagnostics Ltd (formerly Proteome Systems Ltd). Currently, the database contains most published O-linked glycans & N-linked glycans in the literature from the years 1990-2005. For each structure, information is available concerning the glycan type, linkage & anomeric configuration, mass and composition. Detailed information is provided on native and recombinant sources, including tissue and/or cell type, cell line, strain and disease state. Where known, the proteins to which the glycan structures are attached are described, and cross-references to Swiss-Prot/TrEMBL are provided if applicable. The database annotations include literature references, which are linked to PubMed. Detailed information on the methods used to determine each glycan structure is noted to assess the quality of the structural assignment.	http://glycosuitedb.expasy.org/glycosuite/glycodb	Re-launched
GlycoBase (Dublin) [10]	GlycoBase is an HPLC resource that contains elution positions (expressed as glucose unit values) for more than 375 2AB-labeled N-linked glycan structures by a combination of NP-HPLC with exo-glycosidase sequencing and mass spectrometry (MALDI-MS, ESI-MS, ESI-MS/MS, LC-MS, LC-ESI-MS/MS)	http://glycobase.nibrt.ie/glycobase/show_nibrt.action	Maintained
GlycoBase (Lille) [11]	This is a compilation of glycan sequences in various animal species that have been validated using mass spectrometry and NMR. This allows searching of glycan sequences that have found to be typical for each species. Both NMR files & annotated NMR spectra can be downloaded in .jpg format.	http://glycobase.univ-lille1.fr/base/	Maintained
CFG[5]-Glycan Database [12]	Consortium for Functional Glycomics Glycan Database offers detailed structural and chemical information for thousands of synthetic glycans as well as glycans isolated from biological sources. Each glycan structure in the database is linked to relevant entries in CFG and external databases (including primary data and information about binding proteins, where available). Links are also provided to a 3D modeling feature, references, and other information. The starting data in the	http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp	Maintained

[4] KEGG → Kyoto Encyclopedia of Genes and Genomes

[5] CFG → Consortium *for* Functional *Glycomics*

	CFG portal was established using the commercial GlycoMinds to which new structures are added based on experimental evidence.		
GlycomeDB [13]	GlycomeDB through its cross-linking and inter-conversion of the carbohydrate sequences of all freely available glycan databases (CFG, KEGG, GLYCOSCIENCES.de, BCSDB & CCSD) to GlycoCT provides an overview of all carbohydrate structures in the different databases and crosslinks common structures in the different databases. One can search for a particular structure in the meta database and get information about the occurrence of this structure in the five carbohydrate structure databases.	http://www.glycome-db.org/	Maintained
JCGGDB [14]	Japan Consortium for Glycobiology and Glycotechnology DataBase is a portal that integrates all glycan-related data in Japan (glycoprotein, glycolipid, glycosaminoglycans, polysaccharides, etc.) set-up in Japan. The data is sourced from large-quantity synthesis of glycogenes and glycans, analysis and detection of glycan structure and glycoprotein, glycan-related differentiation markers, glycan functions, glycan-related diseases and transgenic and knockout animals, etc.	http://jcggdb.jp/search/search.cgi	Maintained
ECODAB [15]	*Escherichia coli* O-antigen Database contains structures of the repeating units that comprise the O-antigen.	http://www.casper.organ.su.se/ECODAB/	Maintained
EUROCarbDB [16]	EUROCarbDB is a relational database containing glycan structures, their biological context and, when available, primary and interpreted analytical data from high-performance liquid chromatography, mass spectrometry and nuclear magnetic resonance experiments. The database is complemented by a suite of glycoinformatics tools, specifically designed to assist the elucidation and submission of glycan structure and experimental data when used in conjunction with contemporary carbohydrate research workflows.	http://www.ebi.ac.uk/eurocarb/home.action, http://www.eurocarbdb.org/databases	Maintained
Glycoconjugate DB	This database comprises carbohydrate and glyco-conjugate data linked to "chemical" compounds. It consists of structural, spectroscopic (NMR, MS), synthesis pathway data etc. The database doubles up as a compounds' "library".	http://akashia.sci.hokudai.ac.jp/	Maintained
GlycoEpitope	This database provides information on polyclonal or monoclonal antibodies (that have been used as tools for analyzing expression of various carbohydrate chains and their functions), carbohydrate antigens, i.e. glyco-epitopes.	http://www.glyco.is.ritsumei.ac.jp/epitope/	Maintained
Glycosciences.de [17]	The first glycomics web-portal comprising glyco-related databases and tools in the Molecular Modeling Group of Willi von der Lieth at the German Cancer Research Center (DKFZ) in Heidelberg, Germany. The program SWEET & its successor SWEET-II were the first web-based molecular builders for carbohydrate 3D structures. In the late 1990s the SweetDB project started generating data based upon, the then discontinued CCSD, & to link 3D carbohydrate structures modeled with Sweet-II to the corresponding entries, linking them to SugaBase NMR data and others from the literature. Carbohydrates in PDB were also analyzed & incorporated. Several tools were made available to access or analyze this data.	http://www.glycosciences.de	Reinstated
Glyco3D	A site for the 3D structures of glycans and related proteins (lectins, glycosyltransferases and GAG-binding proteins, etc.).	http://glyco3d.cermav.cnrs.fr/glyco3d/	Maintained

b. *Connecting building blocks (monosaccharides) through glycosidic linkages*

Like nucleic acids and proteins, it is far more efficient to encode carbohydrates using a residue-based approach [22]. However, as compared to nucleic acids or proteins, there are a far greater number of building blocks (monosaccharides),

> arising due to the frequent modifications occurring on the parent monosaccharides. Also, since carbohydrates are frequently found to have branched structures, most of them are tree-like molecules, unlike nucleic acids and proteins. The pre-requisite for a residue-based encoding format is a controlled vocabulary of its residue names. For practical reasons, it makes sense to restrict the number of residues to as low a number as possible. Yet, the lack of clear rules to subscribe atoms of a molecule to one particular monosaccharide and not to a substituent, pose the main hurdle in encoding monosaccharide names. As explained in an excellent review [22], let us consider examples of Glc, GlcN, GlcNAc, GalNAc and GlcOAc, all of which can be called monosaccharides from a biologist's or chemist's point of view, except GlcOAc where the monosaccharide is glucose that carries an 'acetyl' substituent. However, considering these as separate monosaccharides would create a major computational complexity. On the other hand, if we think from the encoding point of view, all these examples can be related to the monosaccharides Glc and Gal, with N, NAc and OAc being their respective substituents. Even for bacteria, where the number of monosaccharides is more than 100 [23], this schema is reasonable sized and relatively much easier to maintain.

Due to the development of glycan databases approximately at about the same time in various geographical locations on the globe, but essentially independent of each other, several formats for representing glycan structures have been developed. The major formats for representing glycans that have been used to construct major glycan databases are described in **Table 2.2**.

The variety in nomenclature and structural representation of glycans makes it complex to decide the best form of illustrating the approach of the scientific investigation. The choice of notation is frequently based on whether the study is focused on the chemistry or has a more biological approach. Moreover, the information content of each representation may vary or highlight a particular aspect as compared to others. For example, while representing a complex glycan structure, chemists prefer to elucidate the structure that

Table 2.2: Glycan databases arranged according to the focus of the database/ tool and its respective glycan-encoding format.

	Database	Encoding	URL
Integrated databases	GlycomeDB	GlycoCT [24]	http://www.glycome-db.org/
	EUROCarbDB	GlycoCT [24]	http://www.ebi.ac.uk/eurocarb/, http://www.eurocarbdb.org/databases
	CFG	Glycominds Linear Code ® [25]	http://functionalglycomics.org/
	GlycoSuiteDB	IUPAC condensed [8]	http://glycosuitedb.expasy.org/glycosuite/glycodb
	GLYCOSCIENCES.de	LINUCS [26]	http://www.glycosciences.de/index.php
	JCGGDB	CabosML [27]	http://jcggdb.jp
	Glyco3D	Motif-based	http://glyco3d.cermav.cnrs.fr/glyco3d/index.php
Tools for building glycans and GPI site prediction	Sweet II [28, 29]	N/A	http://www.glycosciences.de/modeling/sweet2/doc/index.php
	Glycam [30]	N/A	http://glycam.ccrc.uga.edu/ccrc/pages/3dspt.html
	SHAPE [31]	N/A	http://sourceforge.net/projects/shapega/
	Polys [32]	N/A	-
	GPI site prediction [33]	N/A	http://mendel.imp.ac.at/gpi/gpi_server.html
Glycan biosynthetic and catabolic pathways (species specific)	KEGG-Glycan	KCF [34]	http://www.genome.jp/kegg/glycan
	CazyDB [35]	N/A	http://www.cazy.org
	LectinDB [36]	N/A	http://proline.physics.iisc.ernet.in/lectindb/
	CancerLectinDB [37]	N/A	http://proline.physics.iisc.ernet.in/cancerdb/
	Dougal [38]	LINUCS	http://www.cryst.bbk.ac.uk/DOUGAL/
	BCSDB	BCSDB linear code	http://csdb.glycoscience.ru/bacterial/
	GGDB	CabosML [27]	http://riodb.ibase.aist.go.jp/rcmg/ggdb
Structural glycan characterizat-ion	GlycoBase (Dublin/NIBRT)	Motif based	http://glycobase.nibrt.ie/glycobase.html
	Glycobase (Lille)	Linkage path	http://glycobase.univ-lille1.fr/base
	CCSD (CarbBank)	IUPAC extended [18]	http://boc.chem.uu.nl/sugabase/carbbank.html
	GMDB [39] (Glycan Mass Spectra Database)		http://riodb.ibase.aist.go.jp/rcmg/glycodb/Ms_ResultSearch

includes information about the anomeric carbon, the chirality of the glycan, the monosaccharides present and the glycosidic linkages that connect them. For others, it is more interesting to visualize the monosaccharides present and hence a symbolic/diagrammatic notation is favored. The most popular and distinct ways of encoding glycans are:

a. The symbolic/diagrammatic notation (e.g. Oxford and CFG notations)

b. Linear notations (e.g. IUPAC)

The structural encodings of glycans dealt with during the course of this thesis have been illustrated in **Figure 2.1**.

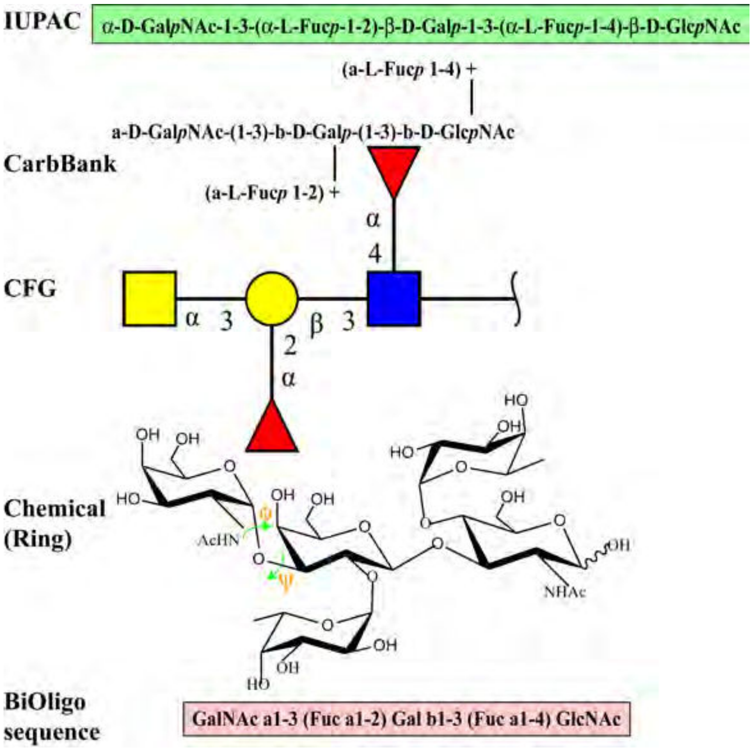

IUPAC
α-D-GalpNAc-1-3-(α-L-Fucp-1-2)-β-D-Galp-1-3-(α-L-Fucp-1-4)-β-D-GlcpNAc
CarbBank
(a-L-Fucp 1-4) +
a-D-GalpNAc-(1-3)-b-D-Galp-(1-3)-b-D-GlcpNAc
(a-L-Fucp 1-2) +
CFG
α
4
α 3
β 3
2
α
Chemical (Ring)
OH
OH
HO
O
AcHN
O
φ
ψ
OH
OH
O
O
OH
HO
HO
OH
HO
O
O
OH
O
OH
NHAc
BiOligo sequence
GalNAc a1-3 (Fuc a1-2) Gal b1-3 (Fuc a1-4) GlcNAc

Figure 2.1: The different levels of glycan encodings (not involving 3D coordinates) handled during this thesis. The example used to illustrate the variety of the notations in this figure is Blood group A Lewis B antigen.

It is evident from the tables above that the focus of glycoinformatics is still on glycan structure (i.e. the glycan composition and topology and not the 3D aspect). But the data generated by the high-throughput techniques of mass spectroscopy (MS), high-pressure liquid chromatography (HPLC) and glycan array technology needs to be translated to a structural understanding for channelizing this information towards the rational structure-based drug development and vaccine design.

References:

1. Apweiler R, Hermjakob H, Sharon N: **On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database**. *Biochimica et Biophysica Acta (BBA) - General Subjects* 1999, **1473**(1):4-8.

2. Ben-Dor S, Esterman N, Rubin E, Sharon N: **Biases and complex patterns in the residues flanking protein N-glycosylation sites**. *Glycobiology* 2004, **14**(2):95-101.

3. Doubet S, Bock K, Smith D, Darvill A, Albersheim P, Albersheim are at the University of Georgia P: **The complex carbohydrate structure database**. *Trends in biochemical sciences* 1989, **14**(12):475-477.

4. Doubet S, Albersheim P: **CarbBank**. *Glycobiology* 1992, **2**(6):505.

5. van Kuik JA, H¬ªrd K, Vliegenthart JFG: **A 1H NMR database computer program for the analysis of the primary structure of complex carbohydrates**. *Carbohydrate Research* 1992, **235**(0):53-68.

6. Toukach F, Knirel Y: **New database of bacterial carbohydrate structures**. *XVIII International Symposium on Glycoconjugates* 2005, **22**:216 - 217.

7. Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita K, Ueda N, Hamajima M, Kawasaki T, Kanehisa M: **KEGG as a glycome informatics resource**. *Glycobiology* 2006, **16**:63R - 70R.

8. Cooper C, Harrison M, Wilkins M, Packer N: **GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources**. *Nucleic Acids Research* 2001, **29**:332 - 335.

9. Cooper C, Joshi H, Harrison M, Wilkins M, Packer N: **GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update**. *Nucleic Acids Research* 2003, **31**:511 - 513.

10. Campbell MP, Royle L, Radcliffe CM, Dwek RA, Rudd PM: **GlycoBase and autoGU: tools for HPLC-based glycan analysis**. *Bioinformatics* 2008, **24**(9):1214-1216.

11. **GlycoBase (Lille)** [http://glycobase.univ-lille1.fr/base/]

12. Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R: **Advancing glycomics: implementation strategies at the consortium for functional glycomics**. *Glycobiology* 2006, **16**:82R - 90R.

13. Ranzinger R, Herget S, Wetter T, von der Lieth C-W: **GlycomeDB - integration of open-access carbohydrate structure databases**. *BMC Bioinformatics* 2008, **9**(1):384.

14. Yoshida K, Suzuki A, N. T: **[Japan consortium for glycobiology and glycotechnology; toward establishment of international network and systems glycobiology] [Article in Japanese]**. *Tanpakushitsu kakusan koso Protein, nucleic acid, enzyme* 2004, **49**(15):2313 -2318.

15. Lundborg M, Modhukur V, Widmalm Gr: **Glycosyltransferase functions of *E. coli* O-antigens**. *Glycobiology* 2010, **20**(3):366-368.

16. von der Lieth C-W, Freire AA, Blank D, Campbell MP, Ceroni A, Damerell DR, Dell A, Dwek RA, Ernst B, Fogh R *et al*: **EUROCarbDB: An open-access platform for glycoinformatics**. *Glycobiology* 2011, **21**(4):493-502.

17. Lutteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, von der Lieth C: **GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research**. *Glycobiology* 2006, **16**:71R - 81R.

18. McNaught A: **Nomenclature of carbohydrates (recommendations 1996)**. *Adv Carbohydr Chem Biochem* 1997, **52**:43 - 177.

19. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules**. *Journal of Chemical Information and Computer Sciences* 1988, **28**(1):31-36.

20. Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, Suzek TO, Wang J, Xiao J, Zhang J, Bryant SH: **An overview of the PubChem BioAssay resource**. *Nucleic Acids Research* 2010, **38**(suppl 1):D255-D266.

21. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M: **ChEBI: a database and ontology for chemical entities of biological interest**. *Nucleic Acids Research* 2008, **36**(suppl 1):D344-D350.

22. Frank M, Schloissnig S: **Bioinformatics and molecular modeling in glycobiology**. *Cell Mol Life Sci* 2010, **67**(16):2749-2772.

23. Herget S, Toukach PV, Ranzinger R, Hull WE, Knirel YA, von der Lieth CW: **Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans**. *BMC Struct Biol* 2008, **8**:35.

24. Herget S, Ranzinger R, Maass K, von der Lieth C: **GlycoCT - a unifying sequence format for carbohydrates**. *Carbohydrate Research* 2008, **343**:2162 - 2171.

25. Banin E, Neuberger Y, Altshuler Y, Halevi A, Inbar O, Dotan N, Dukler A: **A novel Linear Code nomenclature for complex carbohydrates**. *Trends in Glycoscience and Glycotechnology* 2002, **14**:127 - 137.

26. Bohne-Lang A, Lang E, Forster T, von der Lieth C: **LINUCS: linear notation for unique description of carbohydrate sequences**. *Carbohydrate research* 2001, **336**:1 - 11.

27. Kikuchi N, Kameyama A, Nakaya S, Ito H, Sato T, Shikanai T, Takahashi Y, Narimatsu H: **The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures**. *Bioinformatics* 2005, **21**(8):1717-1718.

28. Bohne A, Lang E, von der Lieth C-W: **W3-SWEET: Carbohydrate modeling by internet**. *Journal of Molecular Modeling* 1998, **4**(1):33-43.

29. Bohne A, Lang E, von der Lieth CW: **SWEET - WWW-based rapid 3D construction of oligo- and polysaccharides**. *Bioinformatics* 1999, **15**(9):767-768.

30. **Woods Group. *GLYCAM Web*. Complex Carbohydrate Research Center, University of Georgia, Athens, GA.** [http://www.glycam.com]

31. Rosen J, Miguet L, Perez S: **Shape: automatic conformation prediction of carbohydrates using a genetic algorithm**. *J Cheminform* 2009, **1**(1):16.

32. Engelsen SB, Cros S, Mackie W, Perez S: **A molecular builder for carbohydrates: application to polysaccharides and complex carbohydrates**. *Biopolymers* 1996, **39**(3):417-433.

33. Eisenhaber B, Bork P, Eisenhaber F: **Prediction of potential GPI-modification sites in proprotein sequences**. *J Mol Biol* 1999, **292**(3):741-758.

34. Hattori M, Okuno Y, Goto S, Kanehisa M: **Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways**. *Journal of the American Chemical Society* 2003, **125**(39):11853-11865.

35. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics**. *Nucleic Acids Research* 2009, **37**(suppl 1):D233-D238.

36. Chandra NR, Kumar N, Jeyakani J, Singh DD, Gowda SB, Prathima MN: **Lectindb: a plant lectin database**. *Glycobiology* 2006, **16**(10):938-946.

37. Damodaran D, Jeyakani J, Chauhan A, Kumar N, Chandra N, Surolia A: **CancerLectinDB: a database of lectins relevant to cancer**. *Glycoconjugate Journal* 2008, **25**(3):191-198.

38. **Dougal: A database of glycoprotein structures** [http://www.cryst.bbk.ac.uk/DOUGAL/]

39. Kameyama A, Kikuchi N, Nakaya S, Ito H, Sato T, Shikanai T, Takahashi Y, Takahashi K, Narimatsu H: **A strategy for identification of oligosaccharide structures using observational multistage mass spectral library**. *Analytical Chemistry* 2005, **77**(15):4719-4725.

Molecular modeling:
A high-throughput technique in glycobiology

CHAPTER 3

MOLECULAR MODELING: A high-throughput technique in glycobiology

3.1 High-throughput techniques in glycobiology

In glycobiology, the current important areas of research include establishing relationships between glycan structures and their functions (functional glycomics), monitoring glycosylation in diseased states, disease diagnosis and prognosis and elucidating molecular mechanisms underlying pathogenesis. Towards this goal there is a critical need for new and precise methodologies to be developed that are robust and sensitive at the same time. The field of glycobiology, which is at the frontiers of biomolecular research, is affected by far greater complexities as compared to proteomics or genomics as already discussed in the previous chapters.

Glycan biosynthesis is not template-driven and hence is highly sensitive and responsive to the changing cellular environment. Since glycoconjugates are rather low in abundance (frequently in the femto-molar scale) existing analytical methods need to be adapted to match their occurrence. Improvement in the sensitivity of the existing methods and development of such newer technologies for the investigation of glycans, is a challenge beyond which lies the unraveling of the glycome. Since, no universal technique is still recognized for the rapid and reliable detection, identification and characterization of glycan structures, more possibilities still need to be explored.

Glycan structural diversity and complexity stems from an elaborate and energetically expensive glycosylation process in the cell, executed by numerous sequential and competitive steps in an assembly-line-like system. The intricate work of glyco-enzymes (glycosyltransferases and glycosylhydrolases) gives rise to cell-specific glycan expression patterns, which again, are subjected to notable modifications due to changing cellular conditions like aging or disease [1, 2]. Three different routes can be taken for glycosylation analysis

a. Characterization of glycans present in intact glycoproteins
b. Characterization of glycopeptides
c. Structural analysis of chemically and enzymatically released glycans.

All the above routes provide complementary information and it is important to ascertain

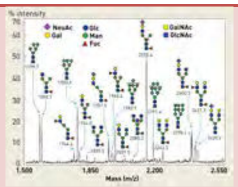
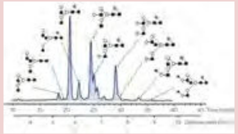
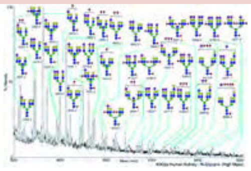
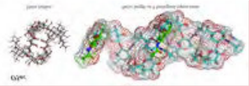
a. Which level of information is the focus of the investigation?
b. Whether the investigation is expected to yield qualitative or quantitative answers?
c. How much of the starting material is available for carrying out the experiments?

High-throughput technology development in glycomics that rationally probes glycan sequence to 3D structure requires heavy bioinformatics support. Glycoinformatics has been crucial in rationally organizing data, building retrieval systems, the analyses and automation of the experimentally generated bulk data in glycobiology.

For glycan composition and interaction studies, the already existing technologies in genomics and proteomics have been modified to suit the special requirements of glycans. A brief over-view of the status of these high-throughput techniques is provided in **Table 3.1**.

In addition to insightful information about the functional properties of glycans and their biological roles, knowledge about glycan structure and composition can provide explanations to observed discrepancies in protein molecular weight, charge or chromatographic retention for values predicted from polypeptide sequences [3]. Glycoprotein properties are strongly influenced by the overall glycan size, shape and charge. Thus, a modification in glycan structure, in say a culture medium, in a recombinant glycoprotein production system can lead to solubility problems during purification and storage, and in a glycoprotein with enzymatic activity (e.g. tissue plasminogen activator) can show modified (enhanced to decreased) activity [4].

Table 3.1: High-throughput techniques for glyco-analysis: MS [5], HPLC [3], glycan array [6] and molecular modeling.

Focus	Technique	Schematic representation	Present status
Glycan Composition	- Mass Spectrometry (MS)		Rapidly developing
	- High Pressure Liquid Chromatography (HPLC)		Rapidly developing
Glycan Interaction	- Glycan array		Rapidly developing
Glycan 3D Structure construction	Molecular Modeling		Developing

Orthogonal (chemically distinct) functions of protein glycosylation depend on the overall size, shape and charge of the glycans that are attached. The 3D structure is crucial for receptor-glycan interactions (as with all interactions). To accurately describe/map the 3D structure of glycans, molecular modeling plays a prominent part alongside experimental techniques. Due to the inherent flexibility of glycan structures, because of glycosidic linkages (whose degree varies from one linkage to another), established protein structure determination methods like X-ray crystallography and NMR need molecular modeling to aid in fully describing the structure and the conformational space sampled by the various stable conformations in solution. As glycans have two naturally active nuclei (^{13}C and ^{1}H), NMR is a key technique that is often indispensible in the determination of unusual or previously undescribed glycans, such as those present in bacterial glycoconjugates [3]. Yet, NMR requires highly purified glycans in large amounts (typically >1mg), though the introduction of the cryoprobe has increased the sensitivity in NMR analysis. Inspite of NMR databases and software now being available, data analysis still remains quite specialized. No single technology can provide the answers for all information required to solve a 3D structure of a glycan, as this needs a multi-dimensional approach.

Since experimental carbohydrate 3D structure determination often is time-consuming, expensive and sometimes unsuccessful, the computational approach of conformation prediction is a viable option for pre-screening, or even replacing, experiments, altogether, when the experimental determination of 3D structure fails and/or is dependent upon the upgradation of existing techniques or instrumentation. The added advantage of computational approaches is that access to compounds is not required. We propose that molecular modeling can prove to be a key high-throughput technique in procuring atomic coordinates for studying 3D structure of glycans.

3.2 Technique and rationale of glycan 3D structure construction

Carbohydrates are well suited for computational conformation prediction. Of the ~20 atoms present in the pyranose unit bound within an oligosaccharide chain, 80% are rigidly-linked together: six locked in the pyranose ring and further ten which are rigidly attached to the five ring carbons. This considerably reduces the motion of the ring structure. The torsion angles, Φ and Ψ of the glycosidic linkage are the most significant for the relative orientations of the glycan units.

The basic assumption used in the modeling of oligo- and polysaccharides is based upon the observation that the structures are made up of alternating fairly rigid components (the pyranose rings) having dimensions in the order of 5 Å, which are linked by flexible glycosidic linkages. As a result there are no strong interactions between non-adjacent sugar residues. From such an assumption, the structure of an oligosaccharide can be constructed from disaccharide nuclei by the sequential addition of one residue at a time. Each added residue is fixed in the lowest energy location. The first computer softwares to be used in the area based upon these approximations were the Hard Sphere Exo-Anomeric (HSEA) program [7, 8] and the Potential Function for Oligosaccharides (PFOS) [9]. The advantage of the rigid-residue approach was that it is extremely fast. More recent tools based on these approximations are SWEET-II [10] (a web-based glycan builder) and POLYS [11]. Whereas this fast method of construction has been used successfully, it is only robust if there are no interactions between non-adjacent residues, and does not allow for a thorough conformational exploration of the conformational space. The requirement for carrying out an exhaustive search led to the development of more complex computational protocols, allowing all significant degrees of freedom (glycosidic torsion angles, primary and secondary hydroxyl groups) to be varied over the full angular range by increments. The complete sampling of the conformational space can be depicted on two-dimensional maps, called 'adiabatic (energy) maps'. While offering some guarantees to find global and local energy minima, this method suffers from its exponential complexity and has been limited to the investigation essentially of disaccharides.

The knowledge of the occurrence of the low energy conformers for a series of disaccharide residues offers the possibility to assemble in a very efficient way much more complex structures. The program POLYS [11] uses this method to generate 3D structures of large polysaccharides and glycan structures from a library of pre-optimized structures of monosaccharides and population statistics of disaccharide segments. Obviously, there are cases where long-range interactions occur and the generated structures are not valid and have to be discarded. Nevertheless, the computational process being efficient, the overall procedure can be continued and the structures presenting steric hindrances can be discarded.

More elaborate prediction methods have been developed and applied to the exploration of the hyper-dimensional conformational space of complex oligosaccharides. Among these, molecular dynamics (MD) simulations have been shown to be a suitable first to characterize the low energy regions of disaccharide molecule [12] along with the conformational pathways corresponding to some transitions from one energy minimum to another one. When applied to larger oligosaccharides, the "natural transition" between low energy conformations, while crossing high-energy barriers, becomes a real issue. Increase in computation time may become impractical, without offering any guarantee to explore all areas of the conformational space. In order to cope with such drawbacks, several MD simulations are run in parallel, for a given molecule, starting from different random selected conformations. Another way is to perform simulation at "high temperature” in order to allow the molecule to undergo conformational transitions by lowering the height of the energy barriers.

The heuristic approach aims at identifying the geographical occurrence of the low energy conformations by progressing along the low energy valleys of the energy hypersurface. The CICADA algorithm [13] has been developed and applied to the characterization of fairly complex oligosaccharides. It must be recognized that such a guided search method, which is fast, does not guarantee to find the global minimum conformation. Nevertheless, the quality of the results compared to those obtained by an exhaustive search has been established. CICADA has been shown to successfully predict the 3D structure of large oligosaccharides, while requiring an acceptable computational time [14].

The stochastic search offers a compromise between computational efficiency and quality of the exploration of the conformational hyperspace. Among them, the Monte Carlo (MC) method has been shown to be both efficient and robust for exploring the conformational space of oligosaccharides [15, 16]. The protocol requires starting randomly with a conformation with known energy, which is submitted to a random alteration and further energy evaluation. The energy difference between the old and the new conformation is evaluated. Based upon a probability function operating from the energy difference, the new conformation is rejected or accepted; in the later case, replacing the previous conformation. The efficiency of the MC method depends on parameters such as the scheme used for the random generation of the structure and the acceptance criterion. As with the molecular mechanics calculations, MC searches can be run in parallel to increase the chance of exploring satisfactorily the conformational hyperspace.

3.3 Carbohydrate force fields

For the successful modeling of glycans, the choice of the correct force field is crucial, for which the parameterization used for the involved glycans should be the deciding factor. An extensive review of carbohydrate-related molecular force-fields has been published recently [17]. Molecular mechanics potential functions that specifically cater to the special characteristics of carbohydrates, and those, which are widely used, are briefly mentioned below, along with the most recent bibliographic reference:

a. MM2, MM2CARB and MM3 [9, 18, 19]: are generic molecular force fields, developed in solid and gas phase simulation states. MM3 is the most widely used molecular force field, applicable to a wide range of molecular classes. These force fields are known to have reproduced experimental data from X-ray crystallography, gas phase structures, besides others.

b. GROMOS [20]: was first developed for MD simulations of proteins, nucleotides or sugars in aqueous or apolar solutions, or in crystalline form. It considers the exo-anomeric effect for pyranoses in its parameterization.

c. CHARMM [21]: has been designed for molecular modeling using both molecular mechanics and MD calculations.

d. AMBER [22]: initially developed for simulating proteins and nucleic acids, it was modified for carbohydrates.

e. GLYCAM [23]: was developed based on AMBER, with a parameter set specially focused towards the MD simulations of glycoproteins and oligosaccharides.

f. OPLS-AA [24]: has incorporated parameters to accommodate carbohydrates simulated in gas phase and explicit solvated simulations.

g. TRIPOS_PIM [25]: parameters were incorporated into a generalized molecular mechanics force field in solid and gas phase and validated on X-ray crystallographic data.

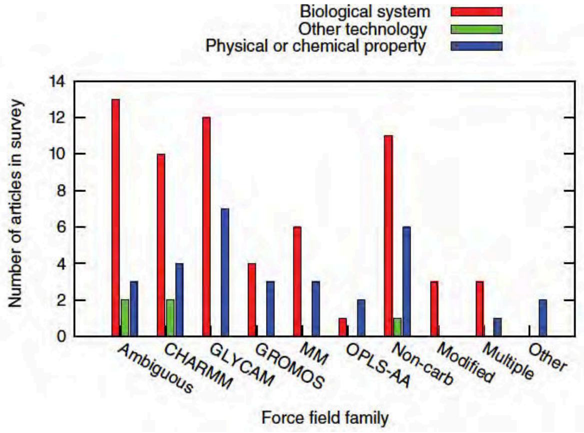


Figure 3.1: Classical force fields categorized according to their use in carbohydrate chemistry based upon their application class (see colours) indicating that biological relevance dominated trends in this area, according to a survey with data upto April 2010. Figure as published in [17].

During the course of this thesis, SYBYL X 1.3 [26] has been used extensively for the modeling of disaccharide and monosaccharide units, employing the pim_2010 parameters for atom types and partial charges incorporated in the TRIPOS force field [25].

3.3.1 Shape

The Shape software [27] has been developed for glycan conformation prediction using a genetic algorithm (GA). It is a powerful tool for automated modeling. Glycans upto a few hundred atoms in size can be investigated using the standard computer hardware on a desktop or laptop. It has been tested on ~300 oligosaccharide structures and has also been found to confirm well to the previously predicted glycan structures. Shape primarily searches the rotational torsion angles of the compound. Ring structures are detected and high-energy ring conformations are ignored as they impart high energies on the conformers, which are generally not seen in nature.

Shape uses a genetic algorithm for searching the conformational space of the glycans. The MM3 force field [28-30] is used for the energy evaluation in Shape. MM3 is considered to perform well for probing glycan structure and hence is the force field of choice working in the background of this program. Shape can be conceptually divided in separate sections

a. Input of molecular data and control parameters
b. GA search combined with the energy-evaluation back-end
c. Clustering of the results from the conformation analysis

3.3.2 Genetic algorithm (GA) and energy evaluation

Genetic algorithms are inspired from molecular evolution; especially the concept of '*survival of the fittest*' is used to reach an ensemble of conformations that have the lowest energies. Several parallel populations compete for survival based on their conformational energies. The conformations with the lowest energies are considered to be the '*fittest*' and survive to pass on their traits to the next generation of conformations. Inheriting the traits of the parents and the further application of genetic operators of mutation and crossover produces the new generation (as illustrated in **Figure 3.2**).

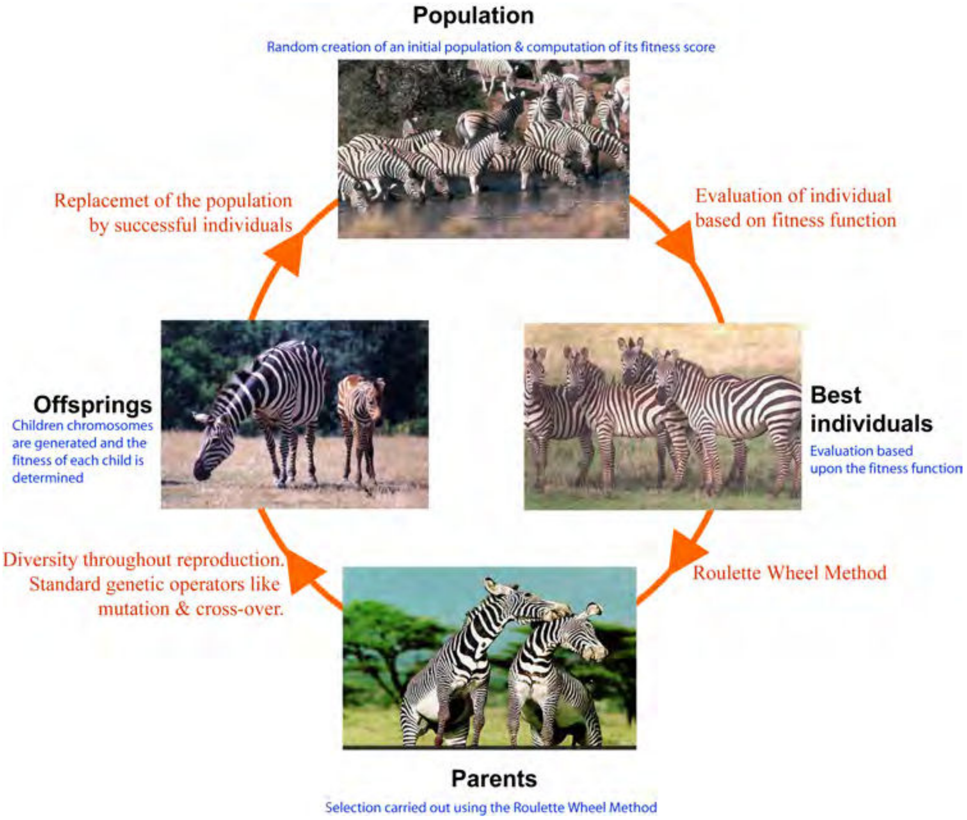

Population
Random creation of an initial population & computation of its fitness score
Evaluation of individual based on fitness function
Best individuals
Evaluation based upon the fitness function
Roulette Wheel Method
Parents
Selection carried out using the Roulette Wheel Method
Diversity throughout reproduction. Standard genetic operators like mutation & cross-over.
Offsprings
Children chromosomes are generated and the fitness of each child is determined
Replacemet of the population by successful individuals

Figure 3.2: An illustration of the basic concept of a genetic algorithm search.

The GA implementation in Shape is a generational parallel population Lamarkian GA, supporting mutation, crossover and migration. Genes are represented using real value direct encoding of torsion angles. All genes are arbitrarily arranged on a single linear chromosome. Each conformation of a molecule is called an ***individual***. Each individual has a ***genome*** describing the torsion angles of that individual. Several individuals constitute a ***population***, within which the individuals compete for survival based upon the calculated conformational energy known as ***fitness***. A low energy conformation corresponds to a high fitness score. Each step in Shape can be best described as below:

a. **Initialization**: One or several populations are initialized with a number of individuals each having a random genome composition, corresponding to random conformations.

b. **Energy evaluation**: Every individual in each population is evaluated by minimizing their structure with the MM3 force field in the back-end, and both the final geometry and energy are stored. The minimized geometries are encoded to torsion angles and saved (updated) in the genome of each individual. The fitness score of each individual is calculated based upon its conformational energy, in comparison to all other individuals of that population.

c. **Pro-creation**: All populations then propagate their next generation with the application of various genetic operators (based upon the input). These operators that act in the evolution are:

   i. ***Mutation***: randomly changing one or more torsion angles (genes), in the individual conformation. Mutation is parameterized either by frequency or probability. Two versions of mutation operators are available: one of totally random mutation, or another that sets a torsion angle to a random value, or an additive mutation, wherein the torsion angle is modified by a random amount.

   ii. ***Crossover***: creation of a new individual by combining the genomes of two or more parents. Crossover is supported through inheritance from two or more individuals with the probabilities of multiple parents. The weight of inheritance between parents is determined by the search parameters. Different scoring and selection methods are available for selecting parents for the next generation of individuals. To employ a ranking based scoring model and a roulette wheel based selection model is the most successful method.

   iii. ***Migration***: one or more individual(s) move from one population (with specified sizes) to another. In general, one large population is more efficient than several smaller ones. This observation is true especially for small compounds, which is consistent with published results [31].

d. **Termination**: The conformational search terminates in Shape when no significant improvement in the best individuals has been found for a (specified) long time.

Lamarkian evolution is supported in Shape, wherein the minimized geometry of each surviving (fit) individual is inherited by its offspring. This method significantly speeds up the convergence rate (as compared to the standard GA approach). However, for larger molecules having a more complex energy hyper-surface, it can be difficult to find a global minimum with only one population. In this case, the standard parameters provided work reasonably well.

GA is a heuristic method that does not guarantee reaching a global minimum, but when correctly parameterized it has a high probability of finding all the important low energy conformations with relatively limited computation time. For a pentasaccharide, the computation time is approximately 6 hours, on a laptop with a 2GB RAM and 20 GB hard disk, while for a decasaccharide it is close to one day. A trade-off between speed and accuracy can be optimized depending upon the user's priority. Computer clusters speed up the calculations even further and modeling of large glycan structures or a library of glycans can be performed with ease.

3.3.3 Clustering

The numerous conformations generated during the conformational sampling allow the trace-back of the evolution of the molecule being investigated. These conformations are sorted according to increasing energy. The first cluster is then seeded by the lowest energy conformation and grown by adding all such conformations that lie within the specified distance (RMSD) from the seed centroid conformation. The next cluster is similarly formed with the next lowest energy conformation that has not already been designated to a cluster and grown as mentioned before. This mode of clustering is fast, with an algorithmic complexity faster than $O(c \cdot n)$ (as mentioned in the Shape documentation), where c is the number of clusters and n is the number of conformations. This is found to be efficient in locating all low energy conformers for the test set used for Shape. The drawback that may appear for this simplified clustering technique is the appearance of small false positive cluster(s) due to the conformations lying just outside the specified cut-off distances from the previous centroids. But this is not a major problem as this can be detected in the filtering stage.

Shape supports a few different combinations of distance measurements and weighting formulas. The primary distance measurement is the RMSD of atomic positions. In this case, different weights can be applied to the atoms, for instance,

- All atoms can be assigned have equal weight.
- Atoms below a certain mass can be discarded.

- Atoms can also be weighted based on their mass, either linearly or by the square root.

The second distance measurement implemented is based on RMSD of the torsion angles, where the torsion angles can be weighted evenly, etc. The available measurement and weighting schemes allow for flexible distance measurements for the clustering result.

3.3.4 Filtering

The clustering groups the result of the conformational search into distinguishable families of low energy conformers. This is further filtered based upon ϕ/ψ maps generated by using MM3 that gives an indication of the low energy regions occupied by the conformations of the molecule being investigated. Out of the cluster centroids reported after Shape clustering, the ones that inhabit the low energy regions are selected and stored as the final results of the conformational sampling.

3.4 High-throughput molecular modeling

Thus, calculations that employ genetic algorithms (as in Shape) to describe the conformation of glycans can prove to be a valuable high-throughput mode to determine the 3D coordinates of the stable low-energy conformations that can be acquired, without the application of any constraints on the molecule. Other methods, such as high temperature MD simulations (usually performed at 700 to 1000 K) may be faster in terms of exploring the energy hypersurface, but it carries a risk of having ring inversions or getting stuck in a high-energy zone, from which it cannot exit.

Shape, for instance, when used on a single desktop machine (with a dual core processor @ 2.9 GHz, 2.00 GB of memory [RAM[1]] and 160 GB hard disk space on a 32-bit operating system), can produce atomic coordinates representing the conformational sampling of a trisaccharide in about 6 to 8 hours, for the specific genetic algorithm parameters that were employed for the job. When deployed on a large-scale cluster (as in

[1] RAM is an abbrevaiation for *Read Only Memory*.

our case, on the CECIC[2] cluster maintained by the Université de Grenoble) the calculation time evidently speeds up significantly. Since, experimental 3D structure determination of glycans is complicated due to the difficulty of obtaining enough sample amounts for biophysical characterization, problems in crystallizing sugars coupled with the inherent flexibility of the glycans, molecular modeling can serve as an active high-throughput technique in speeding up the process of determining atomic coordinates of these special molecules.

References:

1. Freeze HH, Aebi M: **Altered glycan structures: the molecular basis of congenital disorders of glycosylation**. *Curr Opin Struct Biol* 2005, **15**(5):490-498.
2. Ohtsubo K, Marth JD: **Glycosylation in cellular mechanisms of health and disease**. *Cell* 2006, **126**(5):855-867.
3. Marino K, Bones J, Kattla JJ, Rudd PM: **A systematic approach to protein glycosylation analysis: a path through the maze**. *Nat Chem Biol* 2010, **6**(10):713-723.
4. Wittwer AJ, Howard SC, Carr LS, Harakas NK, Feder J, Parekh RB, Rudd PM, Dwek RA, Rademacher TW: **Effects of N-glycosylation on in vitro activity of Bowes melanoma and human colon fibroblast derived tissue plasminogen activator**. *Biochemistry* 1989, **28**(19):7662-7669.
5. Borman S: **CARBOHYDRATE ADVANCES - Recent progress in arrays, functional analysis, and synthetic techniques is helping to make carbohydrates better understood and more useful**. *Chemical & Engineering News* 2005, **83**(32):41-50.
6. Homann A, Seibel J: **Chemo-enzymatic synthesis and functional analysis of natural and modified glycostructures**. *Natural Product Reports* 2009, **26**(12):1555-1571.
7. Lemieux RU, Bock K: **The conformational analysis of oligosaccharides by 1H-NMR and HSEA calculation**. *Archives of Biochemistry and Biophysics* 1983, **221**(1):125-134.
8. Thøgersen H, Lemieux RU, Bock K, Meyer B: **Further justification for the exo-anomeric effect. Conformational analysis based on nuclear magnetic resonance spectroscopy of oligosaccharides**. *Canadian Journal of Chemistry* 1982, **60**(1):44-57.
9. Tvaroška I, Pérez S: **Conformational-energy calculations for oligosaccharides: a comparison of methods and a strategy of calculation**. *Carbohydrate Research* 1986, **149**(2):389-410.
10. Bohne A, Lang E, von der Lieth C-W: **W3-SWEET: Carbohydrate Modeling By Internet**. *Journal of Molecular Modeling* 1998, **4**(1):33-43.
11. Engelsen SB, Cros S, Mackie W, Pérez S: **A molecular builder for carbohydrates: application to polysaccharides and complex carbohydrates**. *Biopolymers* 1996, **39**(3):417-433.
12. Ha SN, Madsen LJ, Brady JW: **Conformational analysis and molecular dynamics simulations of maltose**. *Biopolymers* 1988, **27**(12):1927-1952.
13. Koča J: **Computer program CICADA — travelling along conformational potential energy hypersurface**. *Journal of Molecular Structure: THEOCHEM* 1994, **308**(0):13-24.
14. Koča J, Pérez S, Imberty A: **Conformational analysis and flexibility of carbohydrates using the CICADA approach with MM3**. *Journal of Computational Chemistry* 1995, **16**(3):296-310.
15. Peters T, Meyer B, Stuike-Prill R, Somorjai R, Brisson J-R: **A Monte Carlo method for conformational analysis of saccharides**. *Carbohydrate Research* 1993, **238**(0):49-73.
16. Weimar T, Meyer B, Peters T: **Conformational analysis of α-D-Fuc-(1-->4)-β-D-GlcNAc-OMe. One-dimensional transient NOE experiments and Metropolis Monte Carlo simulations**. *J Biomol NMR* 1993, **3**(4):399-414.
17. Foley BL, Tessier MB, Woods RJ: **Carbohydrate force fields**. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2012, **2**(4):652-697.
18. Allinger NL, Rahman M, Lii JH: **A molecular mechanics force field (MM3) for alcohols and ethers**. *Journal of the American Chemical Society* 1990, **112**(23):8293-8307.
19. Nørskov-Lauritsen L, Allinger NL: **A molecular mechanics treatment of the anomeric effect**. *Journal of Computational Chemistry* 1984, **5**(4):326-335.
20. Hansen HS, Hunenberger PH: **A reoptimized GROMOS force field for hexopyranose-based carbohydrates accounting for the relative free energies of ring conformers, anomers, epimers, hydroxymethyl rotamers, and glycosidic linkage conformers**. *J Comput Chem* 2011, **32**(6):998-1032.

[2] CECIC is an abbreviation for *Centre d'Experimentation et de Calcul Intensif en Chimie*, which is a part of the Univeristé de Grenoble's high performance computing center called *CIMENT*.

21. Raman EP, Guvench O, MacKerell AD, Jr.: **CHARMM additive all-atom force field for glycosidic linkages in carbohydrates involving furanoses**. *J Phys Chem B* 2010, **114**(40):12981-12994.

22. Momany FA, Willett JL, Schnupf U: **Molecular dynamics simulations of a cyclic-DP-240 amylose fragment in a periodic cell: Glass transition temperature and water diffusion**. *Carbohydrate Polymers* 2009, **78**(4):978-986.

23. Kirschner KN, Yongye AB, Tschampel SM, Gonzalez-Outeirino J, Daniels CR, Foley BL, Woods RJ: **GLYCAM06: a generalizable biomolecular force field. Carbohydrates**. *J Comput Chem* 2008, **29**(4):622-655.

24. Kony D, Damm W, Stoll S, Van Gunsteren WF: **An improved OPLS-AA force field for carbohydrates**. *J Comput Chem* 2002, **23**(15):1416-1429.

25. Imberty A, Bettler E, Karababa M, Mazeau K, Petrova P, Pérez S: **Building sugars: The sweet part of structural biology.** . In: *Perspectives in Structural Biology".* Edited by M. Vijayan NYASK. Hyderabad: Indian Academy of Sciences and Universities Press, Hyderabad; 1999: 392-409.

26. Tripos: **SYBYL-X 1.3**. In., X 1.3 edn. St. Louis, Missouri, USA: Tripos International; 1991- 2011.

27. Rosen J, Miguet L, Perez S: **Shape: automatic conformation prediction of carbohydrates using a genetic algorithm**. *J Cheminform* 2009, **1**(1):16.

28. Allinger NL, Yuh YH, Lii JH: **Molecular mechanics. The MM3 force field for hydrocarbons. 1**. *Journal of the American Chemical Society* 1989, **111**(23):8551-8566.

29. Lii JH, Allinger NL: **Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics**. *Journal of the American Chemical Society* 1989, **111**(23):8566-8575.

30. Lii JH, Allinger NL: **Molecular mechanics. The MM3 force field for hydrocarbons. 3. The van der Waals' potentials and crystal data for aliphatic and aromatic hydrocarbons**. *Journal of the American Chemical Society* 1989, **111**(23):8576-8582.

31. Djurdjevic DP, Biggs MJ: **Ab initio protein fold prediction using evolutionary algorithms: Influence of design and control parameters on performance**. *Journal of Computational Chemistry* 2006, **27**(11):1177-1195.

Thesis aim & scope

Thesis aim and scope

This thesis was carried out within the framework of the Marie Curie Initial Training Network (FP7) for the Euroglycoarrays consortium. The objective of the programme was to bring together an interdisciplinary team of scientists and technologists in Europe for the development and application of 'glycan arrays'. Glycan arrays are carbohydrate microarrays displaying unnatural and natural complex carbohydrate structures, such as those found on cell surfaces and/or attached to proteins and lipids. Such glycan arrays are meant to identify the many interactions of carbohydrate binding proteins with specific carbohydrate sequences on a cell and organism-wide scale, and thus providing a fundamental understanding of these important biological recognition events.

Our role in the Euroglycoarrays consortium was to provide an understanding of the three-dimensional (3D) basis underlying the interactions of glycans, focussing on computational methods such as molecular modeling along with database development and management. A particular emphasis was given to the organization of the data into relational databases for retrieval and sustainability, as well as to make it open access, for the effective use and update by the scientific community. The challenges in the context of contemporaneous questions raised in glycosciences are presented in **Chapters 1 to 3** of this thesis.

A complete section of the thesis addresses issues pertaining to the 3D structures of polysaccharides in the solid state (as revealed by X-ray, neutron and electron diffraction studies) and in solution (as revealed by the joint use of high resolution NMR spectroscopy and molecular modeling). **Chapter 4** focuses on the organization of polysaccharide atomic coordinates in one single database called PolySac3DB [1]. The search period covered by the investigation is about 50 years and yielded structural information of 157 polysaccharide entries, which have been organized into 18 categories. Structure related information includes the glycosidic linkages present in each entry, the unit cell and expanded 3D representations of the repeat unit, information about the unit cell dimensions and space group, type of helix and most of the original diffraction diagrams, linked to the abstract of the publication, bibliographic references and the atomic coordinate files for visualisation and download. All this data has been organized in an annotated database accompanied by a user-friendly graphical user-interface, and can be accessed at http://polysac3db.cermav.fr.

Chapter 5 is an example of the determination of 3D structures of a O-antigenic bacterial polysaccharides with the joint use of high resolution NMR spectroscopy and molecular mechanics / dynamics in search of a common epitope that could be found in 4 different strains of enteroaggregative pathogenic *E. coli*. This is a collaborative endeavour with another member of the Initial Training Network (Göran Widmalm's group, Stockholm University), where the isolation of the biological material and the NMR experiments were performed. As with many other polysaccharides, the solution behavior is characterized by the occurrence of several interconverting conformations, the occupancy of which can only be explained by the results provided by the molecular modeling calculations. The structural similarities and differences shed some light on the nature of a common epitope between the different strains. These new 3D structures of bacterial polysaccharides will be included in the PolySac3DB upon acceptance of the submitted manuscript for publication.

The next section of the manuscript deals with the complex world of glycan determinants, which are recognized by glycan binding proteins. **Chapter 6** addresses the question of applying high-throughput molecular modeling to the characterization of the low energy conformers of those molecules, which are grafted on glycan arrays and used to probe the occurrence of interactions between the proteins and the glycan. To this end about 250 glycan determinants have been selected and submitted to a thorough exploration of their hyperdimensional space to determine their conformational preference. Such an endeavour is one of the first attempts to apply molecular modeling to such a large number of samples, ranging in size from trisaccharides to dodecasaccharides. This required the implementation of a widely applicable high accuracy molecular mechanics force field coupled to a genetic algorithm search (previously developed in the group) on a high-performance computing center. The 3D structures of these 250 determinants, accompanied by a total of 200 disaccharides and monosaccharides of interest, which represent more than 1300 conformers and 4000 data files, have been organized into a database called BiOligo (http://bioligo.cermav.cnrs.fr).

While BiOligo [2] aims at offering 3D information to help in deciphering binding data from glycan arrays, it provides, among other structural information, realistic starting conformations that can be used in such instances as molecular dynamics calculations, or docking oligosaccharides in glycan binding proteins. This opens the route to *in silico* screening of protein-carbohydrate interactions, provided that automated docking is capable of structurally characterizing these interactions. **Chapter 7** is an attempt to evaluate such a level of feasibility, in conjunction, at present, with data coming from glycan arrays and isothermal calorimetry. To this end a collaborative study was conducted on soluble lectins from opportunistic bacteria binding to human fucosylated epitopes on mucins. Critical evaluations of the results compared to those derived from the experimentally available high-resolution crystal structures of the complexes, delineate the present level of applicability, while suggesting some future directions for improvements.

The conclusions and perspectives for future works are encapsulated in a review chapter that covers the application of molecular modeling at large, to the field of protein-carbohydrate interactions. It is recognized that the presently available computational tools are considered as useful as the other methods of structural investigation. They can actually help in reconciling the experimental results gathered from separate experiments in different conditions and environments and in extrapolating the results. The wealth of successful applications to many different protein interactions with carbohydrates is a testimony for the maturity of the molecular modeling methods and protocols that have been developed. Nevertheless, these success cases are almost exclusively dealing with instances where proteins interact with carbohydrates, without any further catalytic actions.

A further conclusion emphasizes the urgent need of a more organized access to glycan related data and pleads for the establishment of an open-access global portal to connect glycosciences with the other branches of life sciences.

Links to the databases developed:

1. **PolySac3DB: A database of polysaccharide 3D structures**
[http://www.cermav.cnrs.fr/polysac3db/web/home]

2. **BiOligo: A 3D structural database of bioactive oligosaccharides**
[http://bioligo.cermav.cnrs.fr]

Section A

Chapter 4

CHAPTER 4

PolySac3DB: An Annotated Data Base of 3 Dimensional Structures of Polysaccharides

[1]

Abstract

Background: Polysaccharides are ubiquitously present in the living world. Their structural versatility make them important and interesting components in numerous biological and technological processes ranging from structural stabilization to a variety of immunologically important molecular recognition events. The knowledge of polysaccharide three-dimensional (3D) structure at the molecular level is important in studying carbohydrate-mediated host-pathogen interactions, interactions with other bio-macromolecules, drug design and vaccine development as well as material science applications.

Description: PolySac3DB is an annotated database that currently contains the 3D structural information of 157 polysaccharide entries that have been collected from an extensive screening of scientific literature. They have been systematically organized using standard names in the field of carbohydrate research into 18 categories representing polysaccharide families. Structure-related information includes the saccharides making up the repeat unit(s) and their glycosidic linkages, the expanded 3D representation of the repeat unit, unit cell dimensions and space group, helix type, diffraction diagram(s) (when applicable), experimental and/or simulation methods used for structure description, link to the abstract of the publication, bibliographic reference and the atomic coordinate files for visualization and download. The database is accompanied by a user-friendly graphical user interface (GUI). It features interactive displays of polysaccharide structures and customized search options for both beginners and experts. The site also serves as an information portal for polysaccharide structure determination techniques. The web-interface also references external links where other carbohydrate-related resources are available.

[1] Anita Sarkar[1]& Serge Pérez[1, 2*]

AS designed the method, developed the MySQL database, web interface and related PHP scripts and wrote the manuscript.

[1] Centre de Recherches sur les Macromolécules Végétales (CERMAV, CNRS), Grenoble, France, BP 53X, F-38041 Grenoble Cedex 9, France. (*) affiliated with Université Joseph Fourier, Grenoble
[2] European Synchrotron Research Facility (ESRF), Grenoble, France

Conclusion: PolySac3DB is established to maintain information on the detailed 3D structures of polysaccharides. All the data and features are available via the web-interface utilizing the search engine and can be accessed at http://polysac3db.cermav.cnrs.fr.

Keywords: polysaccharides, carbohydrates, three-dimensional (3D) database, graphical user interface (GUI), atomic coordinates, information portal.

Background

Carbohydrates are an essential class of biological molecules. They are ubiquitous in the living world, occurring mostly as polysaccharides and oligosaccharides, frequently in the form of conjugates with other bio-molecules like proteins (glycoproteins) or lipids (glycolipids). In comparison to the more vastly studied nucleic acids and proteins, carbohydrates have an information carrying capacity of a much higher degree by virtue of the presence of a multiplicity of chiral centers, combinations of various glycosidic linkages and a large number of functional group modifications that might include acetylation, methylation, oxidation and sulfation, creating an even greater diversity out of the already numerous possible building blocks (monosaccharides) [1]. Polysaccharides (or carbohydrate polymers) are macromolecules made up of repeating monosaccharide units linked by glycosidic bonds. They are essential cellular constituents and their roles extend far beyond being mere energy stores (e.g. starch and glycogen) and structural support agents (e.g. cellulose and chitin). They partake in regulating cell wall plasticity (e.g. pectins, alginates and carrageenans), cell signaling, governing solution properties of some physiological fluids and participating in the structural build-up of the extra-cellular matrix (e.g. glycosaminoglycans), eliciting immune responses, cancer progression and as an anti-coagulating agent for the prevention of blood clots (e.g. heparin). Polysaccharides are frequently found on the cell surface of single-celled or multicellular organisms [2] and in the extra-cellular matrix of eukaryotes [3] and are involved in host-pathogen recognition events.

Polysaccharides range in structure from linear to highly-branched. They are often quite heterogeneous, containing slight modifications of their repeating units. This high degree of complexity and inherent micro-heterogeneity of polysaccharide structures make them very difficult macromolecules to handle and explore experimentally. Their structure determines their properties and consequently their function(s). To understand the molecular basis of the native arrangements of polysaccharides and relating their properties and functions to their structures, the different levels of their structural organization must be determined. As with other macromolecules, the elucidation of the primary structure (implying the sequence of monomeric units with the respective glycosidic linkages) is a pre-requisite. Depending on their primary structures and biosynthesis, polysaccharides may have single or multiple chains in characteristic helical forms that define their secondary structure. Energetically favored interactions between chains of well-defined secondary structures result in ordered organizations, referred to as tertiary structures. A higher level of organization involving further associations between these well-structured entities results in quaternary structures.

Diffraction techniques (X-rays, neutrons and electrons) are used for structure determination of bio-macromolecules. Nuclear magnetic resonance (NMR) is used in assessing 3D structures of polysaccharides either in solution or in solid state. Molecular modeling has also become an essential component, not only as a complementary technique to be used in the elucidation of 3D crystalline structures, but also as a powerful tool in the study of the packing of polysaccharides, which can be used to build models, study chain-chain interactions [4] and calculate energies. These molecular modeling techniques can be used to construct structures starting from the content of the crystallographic unit cell to much larger macromolecular assemblies offering a unique possibility to visualize morphological features which are in many cases, the relevant level of structural organization with respect to functions or properties of carbohydrate polymers.

3D structures provide information that is indispensable in many respects of molecular interaction studies. The unification of the resources on carbohydrate polymers and their easy and free availability is necessary. Bioinformatics has played a role in unifying the resources and information available in genetics and proteomics. Similarly, glycoinformatics has a crucial role to play in the field of carbohydrates. Although a large amount of 3D information regarding the structure of polysaccharides has accumulated over time, the effort to collect, curate and disseminate this data electronically and freely to the scientific community has been feeble when compared to similar initiatives in the fields of proteomics or genomics. The Protein Data Bank (PDB [5]) contains few polysaccharide entries, though some coordinates are available, and the Cambridge Structural Database [6] is not open source. The only similar contribution, with respect to polysaccharides, has been in the form of a book chapter, wherein all the atomic coordinates of polysaccharide structures established by X-ray fiber diffraction have been reported and categorized [7]. A similar effort has been made for celluloses and cellulose derivatives in a book devoted to the structures of this important polysaccharide [8].

Here we report the construction of an annotated polysaccharide 3D structural database called PolySac3DB, which provides details of experimental and modeled structures of polysaccharides.

Construction and content

Construction

PolySac3DB is a web-based, platform-independent, manually curated database of polysaccharide 3D structures. It currently runs on an Apache web server [9] hosted at the Centre de Recherches sur les Macromolécules Végétales (CERMAV) with the application program Hypertext Preprocessor (PHP) [10]. It has been developed based on a combination of three layers. The underlying layer is the MySQL database system [11], a relational database management system [MySQL 5.1.41 (Community Server) with PBXT engine 1.0.09-rc] that stores all the structural information along with the respective publications in the back-end and provides the facility to link two or more tables in the database. The intermediate layer is an Apache-PHP application [Apache 2.x; PHP 5.3.1] that receives the query from the user and connects to the database to fetch data to the upper layer, which comprises populated HTML and PHP pages, to the web browser client. The PHP and Java scripts are embedded in the HTML web pages to this effect and are used as application programs for integrating the back-end (MySQL database) to the web pages (HTML). Apache is used as the web server for building the interface between the web browser and the application programs. HTML and PHP have been used to build the web interface.

Content

Data sources –screening, conversion and information extraction

In order to collect structural information about the constituent members of the various polysaccharide families, an extensive screening of literature was performed. This yielded 87 publications that supplied records of the atomic coordinates of polysaccharide (unit) structures established using various structure determination techniques as well as molecular modeling, predominantly containing diffraction data. Enough information could be extracted from these publications to fit the minimum information criteria set for this database and thereafter a total of 157 polysaccharide structures were incorporated into PolySac3DB. The classification of the polysaccharide structures into families is presented in **Table 1 (**The detailed table can be found in Annex II: *Supplementary Material for PolySac3DB***)**.

Table 4.1. The classification of polysaccharide structures in PolySac3DB.

Index	Polysaccharide Family
1	Agaroses
2	Alginates
3	Amyloses & Starch
4	Bacterial Polysaccharides
5	Carrageenans
6	Celluloses
7	Chitins and chitosans
8	Cudlans
9	GAGs
10	Galactoglucans
11	Galactomannans
12	Glucomannans
13	Mannans
14	Pectins
15	Scleroglucans
16	Xylans
17	Nigerans
18	Others

The information was manually extracted and curated before incorporation into the repository. The publications provided atomic coordinates within the asymmetric unit of the cell content available as fractional, Cartesian or cylindrical polar coordinates. The available data was converted to either fractional or Cartesian coordinates to generate the atomic coordinate files in standardized representations of PDB (Protein Data Bank) [5] or Mol2 (SYBYL) [12] formats. The files were generated using an in-house PHP script called PDBGenerator, developed for the construction of this database, which can convert fractional and cylindrical/polar coordinates to PDB format. Besides, SYBYL, PyMol, Mercury and Polys were also used to generate helical/expanded forms of the unit cell structures [12-15]. The aforementioned formats were chosen to provide a broad readability by various visualization programs as well as to expedite comparisons of glycan with nucleic acid and protein structure as well as computer simulation of their interactions. Application of the symmetry operators of the space groups was done to generate the atomic content of the unit cell and extend them to larger structures. Where symmetry operator information was unavailable, models were generated (wherever possible) to offer a representation of the expanded forms assumed by the polysaccharides. The 3D structures of the repeat units and the packing structures were split into two separate tables on the relational database, respectively. The workflow is described in **Table 4.2**.

Table 4.2. Workflow of the informatics tools used in PolySac3DB.

Concept	Implementation
Raw or primary data	Atomic coordinates from experiments
↓	↓
Digitization of data	Text files
↓	↓
Computational conversion to structure files	PDB/ENT, Mol2, Mol (Using molecular modeling tools, e.g. SYBYL, PDBGenerator)
↓	↓
Conversion to helical/expanded structures	PDB/ENT, Mol2, Mol (Using molecular modeling tools, e.g. SYBYL, Polys, PyMol etc.)
↓	↓
Relational database	Constructed using XAMPP comprising of the Apache web server PHP, MySQL
↓	↓
Web interface	HTML & PHP pages

The extracted data also included information about carbohydrate composition, glycosidic linkages as well as space group, unit cell dimensions (a, b, c, α, β and γ), the type of helix (that the polysaccharide chains form), which is made available via the ‘Expert Mode’. The experimental methods used in structure determination of the respective polysaccharide, the link to the abstract and the reference to the publication cited were also extracted. Particular attention was given to the recording of the available diffraction patterns, which are indeed the original experimental data from which the 3D structures were established. In the present version of PolySac3DB more than 120 diffractograms have been collected; they form a unique collection of information that have been generated over almost half a century of structural research in the area of carbohydrate polymers.

Data storage

Efficiency of data storage and management are the hallmarks of a fully functional database. At present the database comprises four tables stored within the relational database working in the back-end of PolySac3DB developed using MySQL which provides the facility of linking/relating two or more tables in a database. The important tables within the database are 'strucdata', 'images', 'polysac3dview' and 'polysac3d-dwnld' that incorporate information regarding the experimental or modeled structures and other information extracted from the publications, the diffraction data and the figure legends, and the atomic coordinates of the 3D structures for viewing and download, respectively. The tables are linked via a unique key to maintain non-redundancy in PolySac3DB. Subsequent tables can easily be added and logically connected to the existing relational database to accommodate more data about the polysaccharide structures that would be deemed relevant in the future.

Utility and Discussion

PolySac3DB search: Navigation and retrieval

The links to access various utilities and the search engine are provided on the left panel of the website via which the data content of the repository can be browsed and retrieved by the user. The 'User Guide' describes each search parameter and its output with detailed examples. A 'Discover Mode' is available that provides background information about the entry/family (mainly regarding occurrence, biosynthesis, property and function). The two-dimensional representations of the polysaccharide repeating unit have been constructed and made available through the 'Discover Mode' in PolySac3DB to aid users to find a familiar representation of the glycan. Information about the nature of the helical structure and all other information can be retrieved upon querying through the 'Expert Mode'.

Data access

Data retrieval and usability are the front-runners in terms of the goals set by the developers of an effective database. An interactive front-end was designed for PolySac3DB with HTML pages and server side scripts that extract data from the tables on the relational database for user-queries on 'Search' and display the retrieved information in a coherent manner. PolySac3DB is equipped with a user-friendly GUI for quick and easy access to the required data. The interface provides the user with options to search by ‘Name’ or ‘Family’ of the polysaccharide. This GUI was tested on different versions of four web browser clients (Google Chrome, Mozilla Firefox, Safari and Internet Explorer) with which it performed efficiently. The ‘User Guide’ gives a detailed description of the content and searchable options within the repository. The schematic overview is provided in **Figure 4.1**.

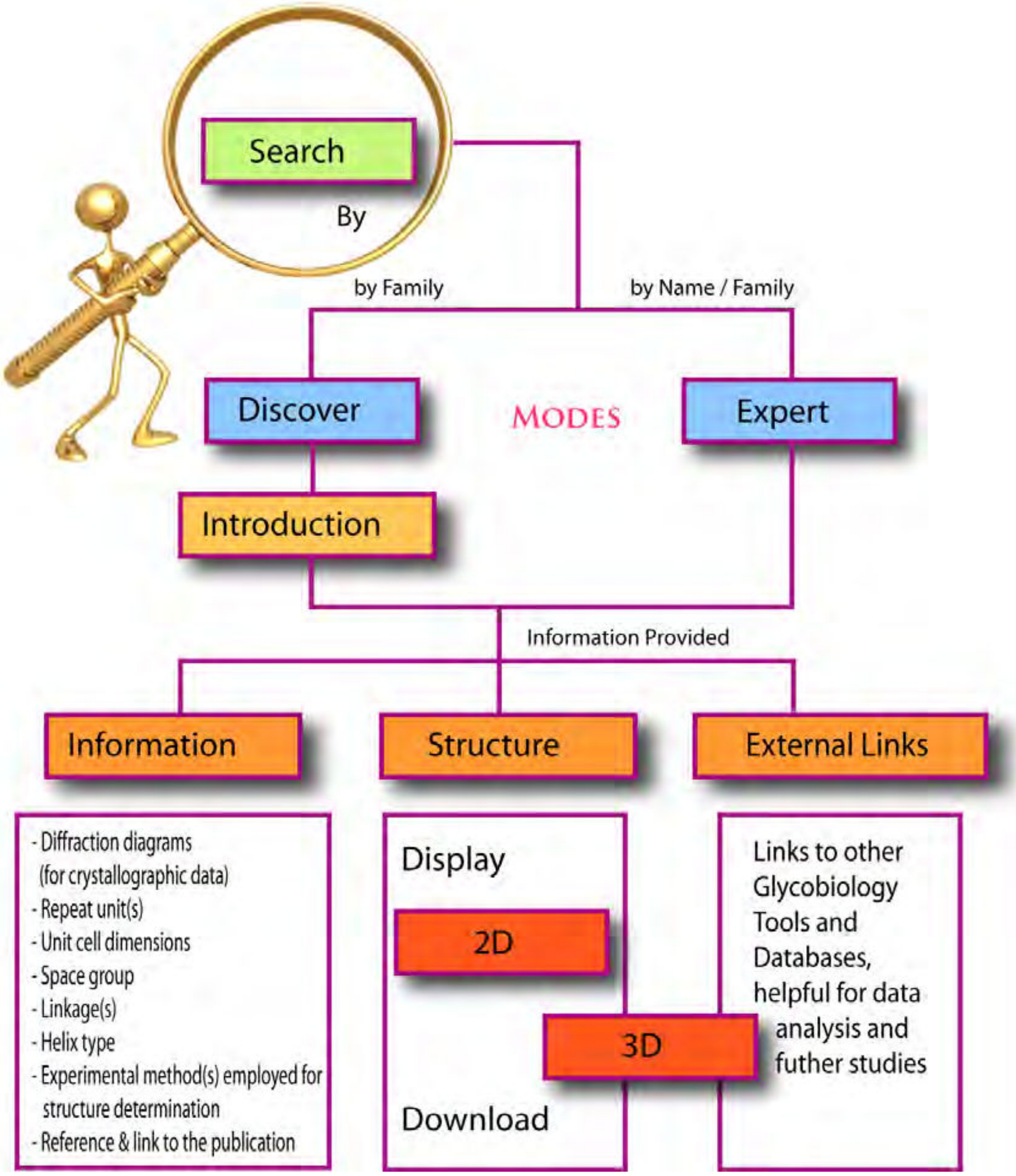

Search
By
by Family
by Name / Family
Discover
MODES
Expert
Introduction
Information Provided
Information
Structure
External Links
- Diffraction diagrams (for crystallographic data)
- Repeat unit(s)
- Unit cell dimensions
- Space group
- Linkage(s)
- Helix type
- Experimental method(s) employed for structure determination
- Reference & link to the publication
Display
2D
3D
Download
Links to other Glycobiology Tools and Databases, helpful for data analysis and futher studies

Figure 4.1. Schematic overview of the PolySac3DB organization and content.

PolySac3DB also provides an overview about the polysaccharide structure determination methods, acting as an information portal on how X-ray, neutron and electron diffraction as well as molecular modeling are applied to polysaccharides. A list of references is provided on the site on a separate web page incorporating all the publications from which the atomic coordinates of the structures in the database have been derived, besides proper referencing on the individual ‘Expert’ pages. In an effort to assimilate other relevant resources for sugars, ‘External Links’ are provided that empowers the user to explore more online glycoinformatics resources.

PolySac3DB output

The bulk of the structure information for the polysaccharide entries is made available via the ‘Expert Mode’. 3D structures can be viewed over the website via the Jmol application [16]. Jmol is an interactive web browser applet, which is an open-source, cross-platform 3D Java visualizing tool for chemical and molecular structures that provides high-performance 3D rendering with standard available hardware. Downloading the atomic coordinates for further independent use is of course another option provided via the expert mode. The GUI has been designed to retrieve, interpret and display the related information about each entry stored in the back-end on four tables of the relational database and display it interactively to the user.

Data collection was followed by data arrangement and fields were set up under which the data was categorized in the database. Since the majority of experimental structures in our dataset contained entries from crystallography, the data fields were defined upon these guidelines.

Beyond the unit cell contents

Besides providing essential structural information, the 3D crystallographic data on polysaccharides open the way to further insights into other strata of structural organization. The following describes some examples of such extensions. In the case of celluloses, the availability of an accurate description of the crystalline structures of the two allomorphs cellulose Iα [17] and cellulose Iβ [18] has provided new insights into the crystalline morphology of the native celluloses. These models were used to generate different ordered atomic surfaces, and evaluate their occurrence along with their respective features. Full atomic models of the crystalline morphology and surfaces of a micro-fibril of cellulose made up of 36 cellulose chains could be conceptualized [19]. Such a model was built as a part of the present database as shown in **Figure 4.2**.

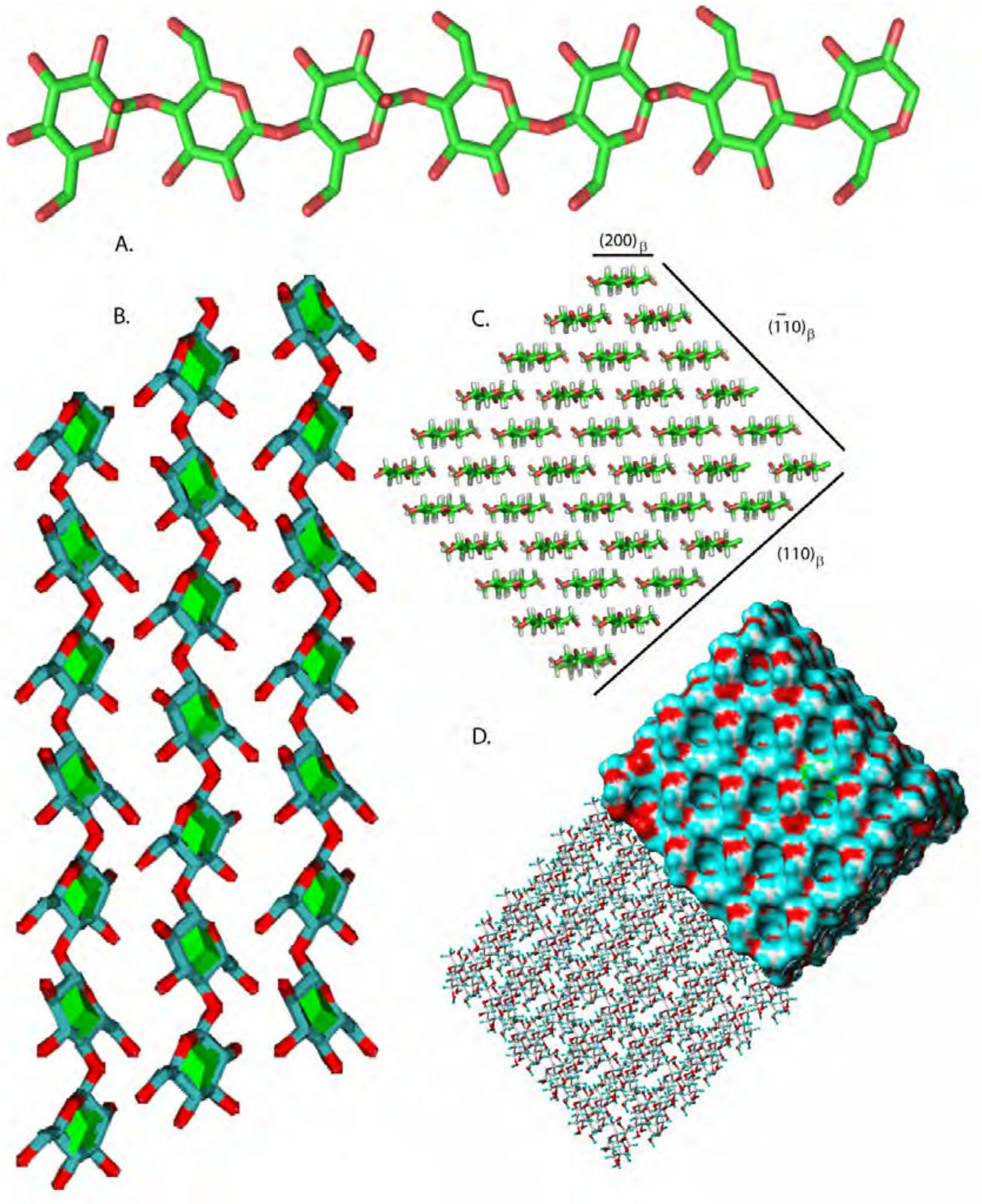


Figure 4.2. Cellulose chain conformation and morphology. (A) Crystalline conformations of the cellulose chain in the 1β allomorph showing the disordered orientation of hydroxylic hydrogen atoms. (B) Relative orientation of cellulose chains of native cellulose 1β. (C) Molecular model of the microfibril of cellulose projected along the fibril axis along with the indexing of the surfaces. (D) Computer representation of the crystalline morphology and surfaces of the microfibril of cellulose made up of 36 cellulose chains.

In other instances, the structural characterization of the branching areas of polysaccharides is difficult to assess. The reason for this may be because these branches constitute only a small fraction of the total macromolecule, or since they are located in the amorphous regions, or in less ordered regions such as between crystallites as in the case of starch. The use of advanced methods of macromolecular modeling has been essential for going from a single helix of amylose, to a unit cluster of amylopectin, by working through a series of building blocks, including single helices, double helices and branch points [20]. In the case of starch, full atomic models of a nano-crystal containing 300 double helical segments in full crystallographic register have been constructed as a part of the work on PolySac3DB. They explain the morphology of these macromolecular assemblies as revealed by transmission electron microscopy [21].

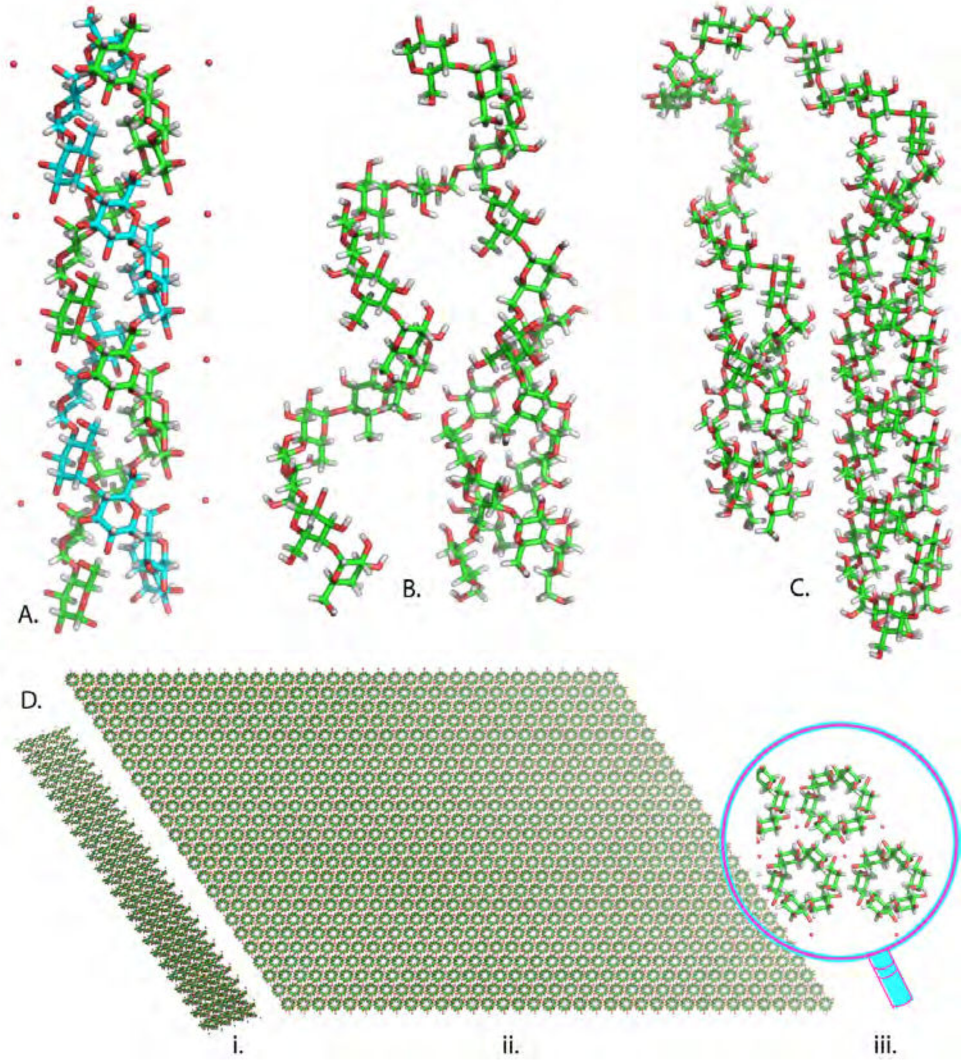


Figure 4.3. Different levels of structural organization in starch. (A) Representation of the left-handed single chains that are parallel stranded in A-starch double helix. (B) and (C) Representations of the double helix of crystalline starch after modeling the branching point between the strands. (D) Computer representation of an ideal platelet nanocrystal showing (i) width of the platelet with the tilt angle of the double helical component, (ii) composition of the platelet and (iii) the enlarged view of the constituent repeating unit.

Figure 4.3 describes the different levels of structural organizations of starch as represented in the various structures present in PolySac3DB. Cases occur where the quality of the experimental data are far from being sufficient to establish a non-ambiguous model of the 3D arrangement. For example, extensive molecular modeling has provided insights about the way chain-pairing occurs, being mediated by Ca^{2+} interactions in alginates and pectins [22].

Conclusion

The aim of the present work is to provide an organization of all polysaccharide atomic coordinates in one single database serving as a unifying repository and to categorize them in a logical fashion for the user to access the required data using pre-customized searching techniques. The search period covered by the present investigation is about 50 years, during which these structural models have been proposed in carbohydrate research. In view of the crucial role played by molecular modeling techniques, it was important to preserve, organize and distribute the macromolecular models developed. Their extensions to higher level of structures may expand our knowledge from the molecular to the microscopic level and help scrutinizing the several levels of structural organization of polysaccharides that underline their remarkable functions and properties. With the increasing number of third generation synchrotron X-ray sources, free electron lasers, new neutron spallation sources and upgrading of current large scale facilities worldwide as well as development of electron-microscopy instrumentation and techniques, one should anticipate an increase in micro, nano and single molecule diffraction data in all areas of science, most certainly including the glycosciences, databases such as PolySac3DB will become an increasingly important tool for the continued documentation, classification, access and dissemination of such data to the scientific community.

At a time when more and more carbohydrates and especially polysaccharides are being called to the fore for their increased use in a plethora of areas as diversified as tissue engineering and repair, wound healing, drug delivery systems, biofuels, bio-degradable fibers and bio-composites due to their generally non-toxic and biodegradable properties and being a renewable resource, PolySac3DB shall be an asset to the community for probing further into the behavior of this class of biological macromolecules.

Availability

The database PolySac3DB is now available at http://polysac3db.cermav.cnrs.fr

Abbreviations

Three-dimensional (3D), Graphical User Interface (GUI), Nuclear Magnetic Resonance (NMR), Protein Data Bank (PDB)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Corresponding author SP designed the framework for the project and wrote the detailed 'Discover' notes. AS designed the method, developed the MySQL database, web interface and related PHP scripts and wrote the manuscript. Both authors have read and approved the final manuscript.

Acknowledgements

The research leading to this publication has received funding from the European Commission's Seventh Framework Programme FP7/2007-2013 under grant agreement n° 215536. We are grateful to the Marie Curie Initial Training Network as part of the FP7 People Programme for training and funding. We acknowledge the help of Cyril Bras for testing the webpages for the search engine and Chen Pan for helping to obtain a few extended structures in this database.

References:

1. Gabius H-J: **Biological Information Transfer Beyond the Genetic Code: The Sugar Code**. In: *The Codes of Life.* Edited by Barbieri M, Hoffmeyer J, vol. 1: Springer Netherlands; 2008: 223-246.

2. Gabius HJ: **Cell surface glycans: The why and how of their functionality as biochemical signals in lectin-mediated information transfer.** *Crit Rev Immunol* 2006, **26**(1):43-79.

3. Iozzo RV: **Matrix Proteoglycans: From molecular design to cellular function**. *Annual Review of Biochemistry* 1998, **67**(1):609-652.

4. Pérez S, Imberty A, Scaringe Raymond P: **Modeling of Interactions of Polysaccharide Chains**. In: *Computer Modeling of Carbohydrate Molecules.* Edited by French AD, Brady JW, vol. 430: American Chemical Society; 1990: 281-299.

5. Berman H, Henrick K, Nakamura H: **Announcing the worldwide Protein Data Bank**. *Nat Struct Biol* 2003, **10**(12):980.

6. Allen F: **The Cambridge Structural Database: a quarter of a million crystal structures and rising**. *Acta Crystallographica Section B* 2002, **58**(3 Part 1):380-388.

7. Chandrasekaran R: **Molecular architecture of polysaccharide helices in oriented fibers**. *Advances in carbohydrate chemistry and biochemistry* 1997, **52**:311-439.

8. Zugenmaier P (ed.): **Crystalline Cellulose and Derivatives : Characterization and Structures**: Springer Berlin Heidelberg; 2008.

9. **Apache Web Server** [http://www.apache.org/]

10. **PHP: Hypertext Preprocessor** [http://www.php.net/]

11. Vaswani V: **MySQL: The Complete Reference**, 1 edn. Emeryville, USA: McGraw-Hill Osborne Media; 2005.

12. Tripos: **SYBYL-X 1.3**. In., X 1.3 edn. St. Louis, Missouri, USA: Tripos International; 1991- 2011.

13. Engelsen SB, Cros S, Mackie W, Pérez S: **A molecular builder for carbohydrates: application to polysaccharides and complex carbohydrates**. *Biopolymers* 1996, **39**(3):417-433.

14. Macrae CF, Bruno IJ, Chisholm JA, Edgington PR, McCabe P, Pidcock E, Rodriguez-Monge L, Taylor R, van de Streek J, Wood PA: **Mercury CSD 2.0 - new features for the visualization and investigation of crystal structures**. *Journal of Applied Crystallography* 2008, **41**(2):466-470.

15. PyMOL: **The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.** In.

16. Herraez A: **Biomolecules in the computer: Jmol to the rescue**. *Biochem Mol Biol Educ* 2006, **34**(4):255-261.

17. Nishiyama Y, Sugiyama J, Chanzy H, Langan P: **Crystal Structure and Hydrogen Bonding System in Cellulose Iα from Synchrotron X-ray and Neutron Fiber Diffraction**. *Journal of the American Chemical Society* 2003, **125**(47):14300-14306.

18. Nishiyama Y, Langan P, Chanzy H: **Crystal structure and hydrogen-bonding system in cellulose Iβ from synchrotron X-ray and neutron fiber diffraction**. *J Am Chem Soc* 2002, **124**(31):9074-9082.

19. Pérez S, Samain D: **Structure and Engineering of Celluloses**. In: *Advances in Carbohydrate Chemistry and Biochemistry*. Edited by Derek H, vol. Volume 64: Academic Press; 2010: 25-116.

20. O'Sullivan AC, Pérez S: **The relationship between internal chain length of amylopectin and crystallinity in starch**. *Biopolymers* 1999, **50**(4):381-390.

21. Pérez S, Bertoft E: **The molecular structures of starch components and their contribution to the architecture of starch granules: A comprehensive review**. *Starch - Stärke* 2010, **62**(8):389-420.

22. Braccini I, Pérez S: **Molecular basis of Ca^{2+}-induced gelation in alginates and pectins: the egg-box model revisited**. *Biomacromolecules* 2001, **2**(4):1089-1096.

Chapter 5

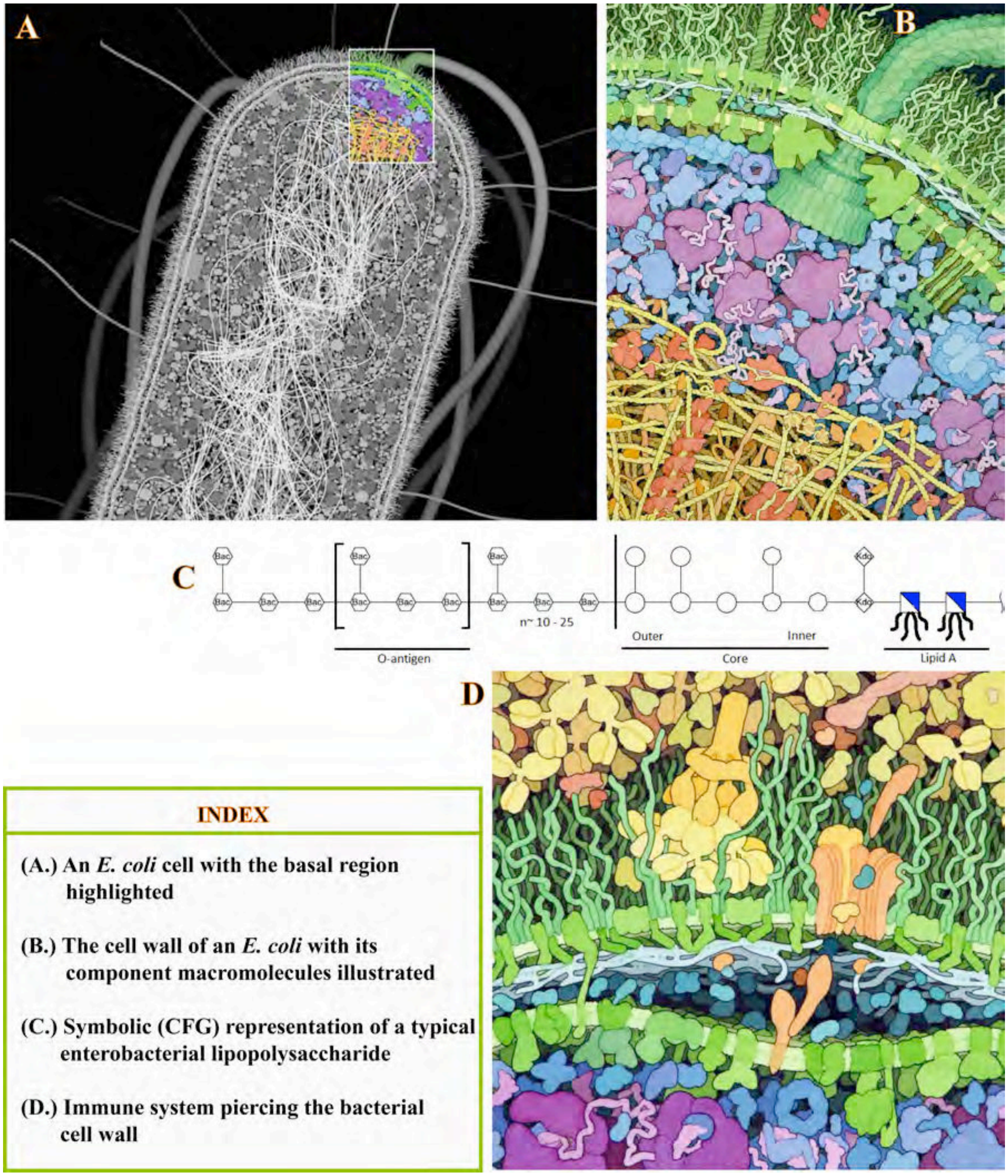
INTRODUCTION TO *O*-ANTIGENIC POLYSACCHARIDES

Escherichia coli

E. coli is the most studied cellular organism known to science and the best described system that played an important role in many of the seminal discoveries of biochemistry (since its discovery by Theodor Escherich in 1885) including the genetic code, glycolysis and the regulation of protein synthesis.

A multi-layered, multi-functional cell wall surrounds the *E. coli* cell and insulates it from environmental dangers. The outer-most membrane is the cell's first line of defense. Its outer face is primarily composed of lipopolysaccharides, which are long strings of polysaccharides with a little bundle of lipid attached at one end. The lipid anchors the molecule into the membrane while the polysaccharide chain extends into the surrounding liquid to form a sticky protective coat. The antibodies of our immune system use these lipopolysaccharides to recognize invading bacteria and mobilize our body's defense mechanism to fight the infection (**Figure 5)**

Many gram-negative bacteria contain structurally unique polysaccharides, i.e. capsular polysaccharides and lipopolysaccharide O-antigens that are often pathogen specific. These polysaccharides are the signature of the respective bacteria. In view of conducting sero-epidemiological surveillance, vaccine trials and livestock immunity studies, the generation of several pathogen polysaccharide antigen arrays is being evaluated. Conjugation strategies can be specifically applied on bacterial polysaccharides followed by immobilization on glass slides. There exist efficient and reproducible ways to prepare polysaccharide antigens either via total organic synthesis or via specific degradations of the native lipopolysaccharides to give chemically defined fragments with high antigenic activity and enable systematic evaluation antigenicity toward new diagnostic tools. Among the many open questions that remain to be evaluated to reach such a goal, the issue related to the characterization of the structural behavior of the polysaccharide in solution and the investigation of the binding conformations still



INDEX

(A.) An *E. coli* cell with the basal region highlighted

(B.) The cell wall of an *E. coli* with its component macromolecules illustrated

(C.) Symbolic (CFG) representation of a typical enterobacterial lipopolysaccharide

(D.) Immune system piercing the bacterial cell wall

Figure 5: The interactions of *Escherichia coli* and O-antigenic polysaccharides on its surface. (A.) The *E. coli* cell (magnification: 10,000 X) showing the double-layered cell wall packing in all the soluble cellular components [1].

(B.) A magnified (1,000,000 X) portion of the *E. coli* cell illustrating the proteins, nucleic acids, polysaccharides and lipid-membranes. The internal space of the cell is filled with water, glycans, nucleotides, amino acids, metal ions and many other small molecules [1].

(C.) Schematic structure (CFG representation) of an enterobacterial lipopolysaccharide molecule [2]. The lipids are depicted by ribbons attached to GlcNAc in the lipid A part, attached to Kdo, heptoses in the inner core region, hexoses in the outer core region, and finally the O-antigenic components, most commonly hexoses.

(D.) The immune system piercing the *E. coli* cell wall (magnification: 1,000,000 X). Our blood contains proteins that recognize and destroy invading pathogens. This illustration depicts a cross-section through the bacterial cell (lower section of the figure in green, blue and purple) being attacked by the proteins in the blood serum (upper part of the figure in yellow and orange). Y-shaped antibodies recognize and attach themselves to the cell surface setting off a cascade of actions that culminate in a membrane attack complex, shown here, piercing the cell wall of *E. coli* [1].

remains to be addressed. These NMR experiments have to be used in conjunction with experimental observations such as immunological data, and in the future with the glycan array screening, combined with molecular modeling. Ultimately, the identification of the nature and the size of the epitope should guide the optimum preparation of the future pathogen glycan arrays.

Chapters 5.a and **5.b** are devoted to the structural elucidation of O-antigenic polysaccharides from enteroaggregative pathogenic *Escherichia coli* strains, with the aim to fully characterize their solution behavior and to identify the nature and the size of the epitopes reacting with monoclonal antibodies.

References:

1. Goodsell D, S.: **The machinery of life**. New York: Springer New York; 2009.
2. Stenutz R, Weintraub A, Widmalm G: **The structures of *Escherichia coli* O-polysaccharide antigens**. *FEMS Microbiology Reviews* 2006, **30**(3):382-403.

CHAPTER 5.a

Three-dimensional structural elucidation of O-antigenic polysaccharides from enteroaggregative pathogenic *Escherichia coli* strains O5ac and O5ab. [1]

Introduction

The multi-layered cell wall of *Escherichia coli* forms the cell's first line of defense, protecting and insulating it from the external environment. Lipopolysaccharides (LPS) (which are long strings of polysaccharides with a little bundle of lipids attached at one end) constitute an important part of the lipid bilayer in the bacterial cell wall. The lipid anchors the molecule into the membrane while the polysaccharide chains extend out into the surrounding liquid environment to form a sticky, protective coat. Immune systems of the hosts have antibodies that recognize the LPS of the bacteria and prevent infection.

E. coli is a part of the human colonic flora. They are generally non-pathogenic although some strains are known to cause virulence causing a number of diseases in animals as well as human beings. Diarrheagenic *E. coli* strains are major pathogens associated with enteric disease in many parts of the world. Enteroaggregative pathogenic *E. coli* strains have been found to colonize intestinal mucosa, mainly of the colon, and the subsequent secretion of enterotoxins and cytotoxins [1]. Initially, this pathotype was considered to be an emerging agent of persistent infantile pediatric diarrhea, especially in developing countries [1-5]. But later it was detected to be virulent in adults and having a global distribution [6, 7].

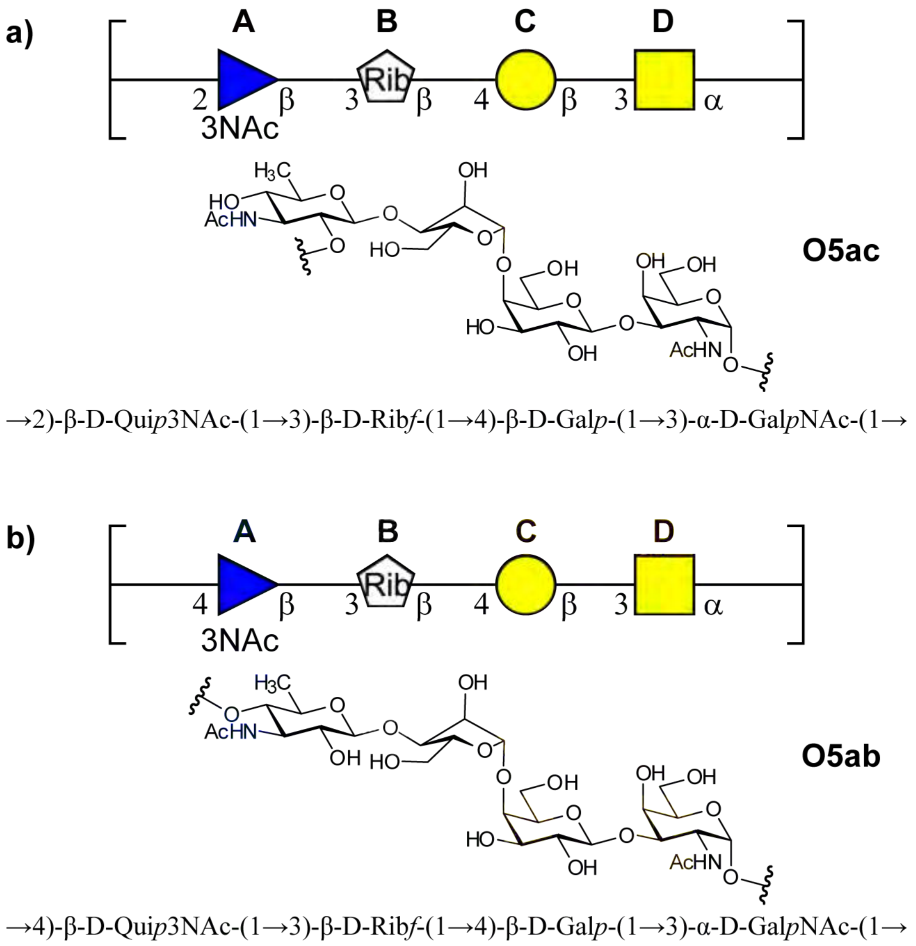
The O-antigenic polysaccharide present in the LPS of *E. coli* O5ac and O5ab are very similar to each other, with the difference being in one glycosidic linkage, i.e. the

[1] Anita Sarkar[§a], Carolina Fontana[§b], Anne Imberty[a], Serge Pérez[a] & Göran Widmalm[b]

[§] *These authors have contributed equally to the work.* AS designed the simulations, modeled the polysaccharides, calculated the low energy maps and wrote the manuscript sans the NMR section.

[a] Centre de Recherches sur les Macromolécules Végétales (CERMAV, CNRS), Grenoble, France, BP 53X, F-38041 Grenoble Cedex 9, France. [b] Department of Organic Chemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden

substitution pattern of β-D-Qui*p*3NAc that links two biological repeat units as shown in **Figure 5.a.1**. Thus, the O-antigenic polysaccharide of O5ac, in which β-D-Qui*p*3NAc is 2-substituted, and that of O5ab, where the same monosaccharide is 4-substituted, are positional isomers. The shape and biological function of polysaccharides are closely related. An understanding of the conformational behavior of the bacterial surface polysaccharide in solution can clearly elucidate this relationship. In this study we focus on the O-antigenic cell surface polysaccharides of *E. coli* O5ac and O5ab to understand their immunochemical similarities.



→2)-β-D-Qui*p*3NAc-(1→3)-β-D-Rib*f*-(1→4)-β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc-(1→

→4)-β-D-Qui*p*3NAc-(1→3)-β-D-Rib*f*-(1→4)-β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc-(1→

Figure 5.a.1. Structure of the biological repeating units of the O-antigen PS from a) *E. coli* O5ac and b) *E. coli* O5ab in CFG-notation (*top*), schematic chemical representation (*middle*) and standard nomenclature (*bottom*), respectively.

Materials & Methods

Nomenclature

The position of two contiguous monosaccharides connected by a glycosidic linkage (excluding the 1→ 6 linkage) is described by the torsion angles Φ and Ψ. The Φ/Ψ definitions used for the relaxed maps (as well their description in the results & discussion section) correspond to the heavy atom convention (as illustrated in **Figure 5.a.2**), where, for a (1→x) glycosidic linkage

Molecular modeling: Φ → O5-C1-O1-C_x and Ψ → C1-O1-C_x-C_{x+1}.

NMR: Φ^H→ H1-C1-O1-C_x and Ψ^H → C1-O1-C_x-H_x.

The different conformational families were clustered according to their characteristic Ψ^H torsion angles and denoted Ψ^+, Ψ^-, Ψ^{cis} and Ψ^{trans} for torsion angles corresponding to *gauche*+, *gauche*–, *cis* and *trans* states, respectively. Analogously, the conformational families that have a torsion angle Φ^H in the *trans* state are denoted Φ^{trans}.

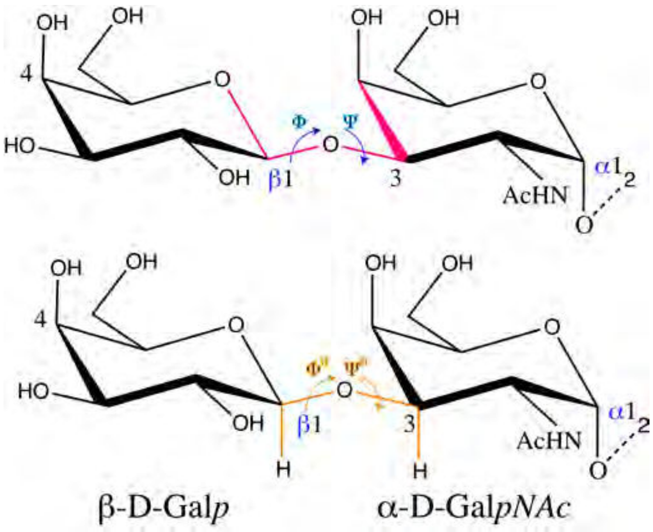


Figure 5.a.2. Illustrated example of the torsion angle conventions used in this study, described using the disaccharide β-D-Gal*p*-1,3-α-D-Gal*p*NAc. The *Heavy Atom Convention* is represented in the *top panel*, while the *Light Atom Convention* is illustrated in the bottom panel.

Molecular modeling

Most of the constituent monosaccharides were built from optimized base types found in Glyco3D [8]. β-D-Qui*p*3NAc was built from the base type found in PDB (1MMY) and N-acetylated using an optimized NAc fragment from Glyco3D. The north and the south conformers of β-D-Rib*f* were extracted from two PDB structures (1QXB and 3KSM, respectively).

Energy Calculations

The geometry optimization of the starting disaccharides and oligosaccharides was performed using the Tripos force field [9], with the pim partial atomic charges [10], and an energy convergence criterion of 0.05 kcal/mol for a maximum of 1000 iterations for monosaccharides and 10,000 iterations for oligosaccharides. The dielectric constant was set at 4.0.

Relaxed Energy Maps of each Disaccharide present in the modeled oligosaccharide

The four disaccharides each of O5ac and O5ab were subjected to high temperature molecular dynamics (MD) simulations to explore the accessible conformational space of carbohydrates [11]. The conformational free energy maps were derived from the population analysis of the MD using the Boltzmann equation [12]. MM3 (implemented through the TINKER suite [13] was used to minimize the starting oligosaccharide structures of O5ac and O5ab, and to calculate the trajectories at 1000K for 10 ns. The Conformational Analysis Tools (CAT) software was used for data processing and analysis [14].

Conformational analysis of the oligosaccharides

The conformational space available to the oligosaccharides, constituting the biological repeat units of the O5ac and O5ab, have been characterized using the software called Shape [15].

Within Shape, MM3 [16, 17] and the block diagonal minimization method for geometry optimization was used with the default energy-convergence criterion ($\Delta E=0.00008*n$ kcal/mol every 5 iterations, where, n= number of atoms). MM3 allows full relaxation of the glycosidic residues taking into account the exo-anomeric effect [17, 18] and this force field also allows optimization to a nearby transition state (with the full matrix Newton-Raphson method).

Genetic algorithm

The systematic exploration of the conformational search of the O-antigenic hexasaccharides (i.e. the biological tetrasaccharide repeat and the monosacchaide required to represent the glycosidic bond connecting two biological repeats at the reducing and non-reducing ends) of O5ac and O5ab were performed using Shape. The genetic algorithm parameters for conformer generation were specified to a population size of 25 individuals to be included in every population throughout the search, while the total number of parallel populations to be used during the search was set to 20. Every generation produced by the genetic algorithm comprised

Total number of individuals = population size* total number of populations

The energy convergence criterion for the conformers generated was assigned a window size of 20 to search for improvements (i.e. the search was terminated when even after 20 generations no significant improvements in conformational energy was found), with a limit value (i.e. the highest energy difference in the entire window that is accepted as a significant improvement for the search to continue) of -0.5 kcal/mol.

Clustering

The large number of possible conformations generated during the complete conformational sampling had to be clustered in order to clearly demarcate the distinct families of low energy conformations that could be present. In this study, the conformations generated were clustered based upon the deviation of the conformation having the lowest energy as compared to the starting minimized structure (atom distances), ignoring hydrogen atoms and 1 Å tolerance for RMSD from the cluster centroid.

Filtering

The results after clustering were filtered based upon the energy. The lowest energy representative was selected from the distinct minima (as described by the adiabatic maps).

NMR spectroscopy

Preparation and purification of the O-antigen polysaccharides

The O-antigenic polysaccharides (PS) from *E. coli* O5ac and O5ab were obtained as described previously [19]. Purification by size exclusion chromatography was carried out on a HiLoad^{TM} 16/60 Superdex^{TM} 30 column (GE healthcare) using an ÅKTA^{TM} purifier system (GE healthcare).

Studies on the O-antigen PS from E. coli O5ac

The O-antigen PS from *E. coli* O5ac (3.6 mg) was deuterium-exchanged by repeated cycles of dissolution of the sample in excess of 99.9% D_2O followed by freeze-drying. The sample was dissolved in 0.5 ml 99.9% D_2O and treated with a Chelex® 100 cation exchange resin (100-200 mesh, Na^+ form, BioRad) for 1 hour in order to remove paramagnetic ions. The solution was filtered to a 5 mm NMR tube, freeze-dried, and re-dissolved in 0.6 ml of 99.99 % D_2O; sodium 3-trimethylsilyl-(2,2,3,3-$^2H_4$)-propanoate (TSP) was added as internal reference (δ_H 0.00). Oxygen was removed by three freeze-pump-thaw cycles, and finally the NMR tube was flame-sealed under vacuum. ^1H chemical shifts assignments have been reported earlier [19].

Proton cross-relaxation rates were measured using a 2D $^1H,^1H$-NOESY experiment with a zero-quantum suppression filter [20] on two different spectrometers: Bruker Avance III 700 MHz and Bruker Avance 500 MHz, both equipped with 5 mm TCI Z-Gradient CryoProbes. The experiments were recorded at 42 °C over a spectral width of 5.0 ppm, with 14k × 256 or 10k × 260 data points (at 700 MHz and 500 MHz, respectively) using 6 – 16 scans per t_1-increment and a total recycle time between scans of 16 s or 13 s (at 700 and 500 MHz, respectively) corresponding to 5 times the longer T_1. Five different cross-relaxation delays (mixing times) of 30, 40, 50, 60 and 80 ms were used. At 700 MHz a 40 kHz broad and 20 ms long adiabatic smoothed CHIRP [21] pulse was employed during the zero-quantum suppression, accompanied by a gradient pulse of strength 9% of the maximum (100 % ~ 53.0 $G \cdot cm^{-1}$). At the 500 MHz, on the other hand, a 27 kHz broad adiabatic [22, 23] smoothed CHIRP pulse was used instead, and the gradient pulse strength was set to 6% of the maximum (100 % ~ 53.0 $G \cdot cm^{-1}$). Prior to Fourier transformation forward linear prediction to 512 or 520 (at 700 and 500 MHz, respectively) in the F_1-dimension and zero-filling to 16k × 2k points were performed; 90° shifted squared sine-bell window functions were used in both dimensions. A fifth-order polynomial baseline correction was applied in both dimensions and the peaks of interest integrated by using the same integration limits at all mixing times.

The volume integral of each NOE buildup peak was divided by the volume integral of the respective autopeak to produce the normalized buildup intensities that were used to calculate the NOE buildup rates (σ) from the slope. At each magnetic field the cross relaxation rates were averaged for each proton pair, and the unknown proton-proton distances (r_{ij}) calculated using as reference the intra-residue distance between H1 and H2 of the α-D-Gal*p*NAc residue (**D**) and the following equation:

$$r_{ij} = r_{ref} \times \left(\frac{\sigma_{ref}}{\sigma_{ij}}\right)^{1/6}$$

Effective proton-proton distances from the models were calculated using the following equation:

$$\frac{1}{r_{calc}} = \langle r^{-6} \rangle^{1/6}$$

The final proton-proton distances were obtained by averaging the distances calculated at each magnetic field. The plots that were used to obtain the cross relaxation rates had a residual standard deviation of less than 16%. The experimental error in σ is estimated to be less than ± 26%, which corresponds to only ± 4% in the calculated proton-proton distances as a result of the r_{ij}^{-6} dependence, which is in the order of ± 0.1 Å.

Studies on the O-antigen PS from E. coli O5ab

The O-antigen PS from *E. coli* O5ab (6.4 mg) was deuterium-exchanged by using repeated cycles of dissolution of the sample in excess of 99.9% D_2O followed by freeze-drying. The sample was then transferred to a 5 mm NMR tube, freeze-dried and re-dissolved in 0.6 ml of 99.99% D_2O, using sodium 3-trimethylsilyl-(2,2,3,3-$^2H_4$)-propanoate (TSP) as internal reference (δ_H 0.00). ^1H chemical shifts assignments were reported earlier [24].

Proton cross-relaxation rates were measured at 700 MHz using a selective 1D single-pulse-field-gradient spin-echo (SPFGSE) NOESY experiment with a nulling 180° pulse [25] and a zero-quantum suppression filter [20]. The experiments were recorded at 27 °C over a spectral width of 7.5 ppm, with 21k data points and 256 scans per transient. A total recycle time between scans of 23 s corresponding to 8.3 times the longest T_1 was employed. Eight different mixing times of 45, 50, 55, 60, 65, 70, 75 and 80 ms were used. Selective excitation of the H1 proton of the α-D-Gal*p*NAc residue (**D**) was achieved using a 40 Hz broad RSnob [26] shaped pulse of 80 ms, flanked by pulse field gradients (sine.100) of 1 ms length, with the strength set to 15% of the maximum. The strength of the 1 ms length gradients flanking the 180° nulling pulse during the mixing time were set to 40% of the maximum. A 40 kHz broad and 20 ms long adiabatic smoothed CHIRP pulse was employed during the zero-quantum suppression, accompanied by a gradient pulse of strength 8% of the maximum. Zero-filling to 128k points and an exponential line broadening of 1 Hz were performed prior to Fourier transformation. A fifth-order polynomial baseline correction was applied and the peaks of interest integrated by using the same integration limits at all mixing times.

In addition, a 2D ^{1}H,^{1}H-NOESY experiment with a zero-quantum suppression filter [20] and a mixing time of 80 ms was recorded over a spectral width of 7.5 ppm, with 22k × 256 data points, using 6 scans per t_1-increment and a total recycle time between scans of 14 s (corresponding to 5 times the longest T_1). A 40 kHz broad and 20 ms long adiabatic smoothed CHIRP pulse was employed during the zero-quantum suppression, accompanied by a gradient pulse of strength 8% of the maximum.

The integrals of each NOE buildup peak were divided by the integral of the selective excited peak to produce the normalized buildup intensities that were used to calculate the NOE buildup rates (σ) from the slope. The intra-residue distance between H1 and H2 of the α-D-Gal*p*NAc residue (**D**) was used as reference for distance calibration. The plots that were used to obtain the cross-relaxation rates had a residual standard deviation of less than 2%.

Results & Discussions

Molecular modeling of the oligosaccharides

The repeat units of the O5ac and O5ab comprise the same constituent monosaccharides and

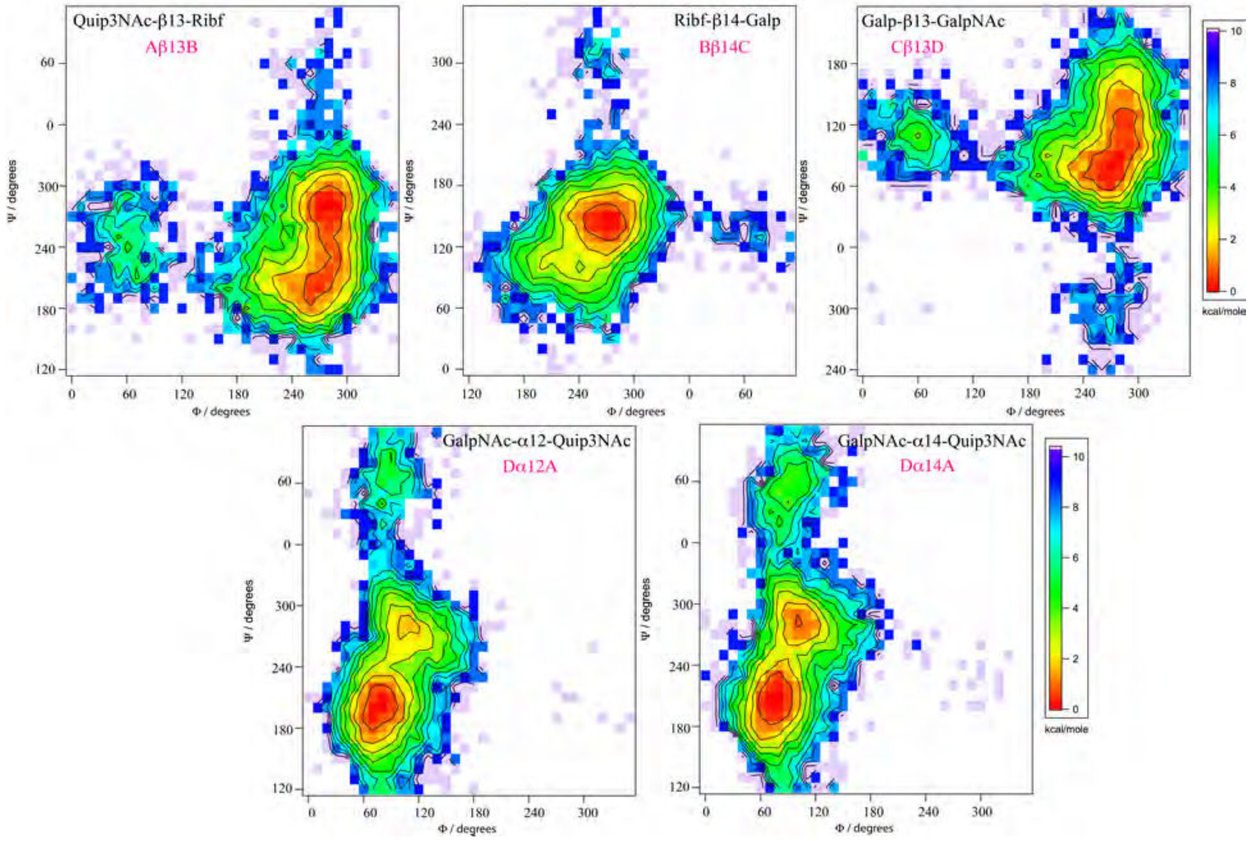


Figure 5.a.3. Relaxed adiabatic maps of the disaccharide components of the molecular model of O5ac and O5ab. The top panel illustrates the glycosidic linkages that are identical in the two *E. coli* samples, while the lower panel highlights the glycosidic linkages (Gal*p*NAc-α12-Qui*p*3NAc in O5ac and Gal*p*NAc-α14-Qui*p*3NAc in O5ab) that are the distinguishing feature between them.

glycosidic linkages with the exception being in the glycosidic linkage connecting two biological repeats in the two strains, i.e. **D**α12**A** in O5ac and **D**α14**A** in O5ab. To eliminate the linkage end effects, for each tetrasaccharide biological repeat unit of O5ac and O5ab, an α-D-Gal*p*NAc was added to the non-reducing end and a β-D-Qui*p*3NAc at the reducing end.

To build the 3D O-antigenic polysaccharide models, four glycosidic linkages were needed, namely, **A**β13**B**, **B**β14**C**, **C**β13**D** and **D**α12**A** for O5ac and **A**β13**B**, **B**β14**C**, **C**β13**D** and **D**α14**A** for O5ab, respectively. The conformational space sampled by the oligosaccharides was explored using a genetic algorithm as implemented in Shape [15]. Further, each of the constituent disaccharides were studied using high temperature MD (with TINKER using CAT scripts) by a systematic grid search approach using MM3 parameters. The conformational maps generated have been shown in **Figure 5.a.3**. Each of the disaccharides can occupy multiple conformational states, thus indicating that they are potentially flexible molecules.

O-antigen of E. coli O5ac

The Φ angle of the disaccharides Qui*p*3NAc-β13-Rib*f* (**A**β13**B**) and Rib*f*-β14-Gal*p* (**B**β14**C**) linkages of O5ac are restricted around a value of 270˚. The Ψ angle for the Qui*p*3NAc-β13-Rib*f* (**A**β13**B**) linkage is flexible and spans about 120˚ on the energy map. In the Rib*f*-β14-Gal*p* (**B**β14**C**) linkage, the Ψ angle is more restricted, centered about 145˚, and covering only ~20˚ on this axis. In the Gal*p*-β13-Gal*p*NAc (**C**β13**D**) linkage, the value for Φ is centered at about 280˚ while the Ψ angle is more flexible, spanning ~80˚ on the energy map within an energy barrier of 7 kcal/mol. The Φ angle value of the Gal*p*NAc-α12-Qui*p*3NAc (**D**α12**A**) of this oligosaccharide is much more flexible, covering approximately 40˚ on the energy hypersurface. The Ψ values are expanded between 180˚ to 230˚ for the global minimum and centered about 280˚ for a local minimum.

In the disaccharide Qui*p*3NAc-β13-Rib*f* (**A**β13**B**), we detect two populations corresponding to the energy minima, one population at Φ/Ψ ≈ 270°/180° and another at Φ/Ψ ≈ 280°/280°. The disaccharide Rib*f*-β14-Gal*p* (**B**β14**C**) presents only one global minimum at Φ/Ψ ≈ 280°/140°. For the Gal*p*-β13-Gal*p*NAc (**C**β13**D**) disaccharide, a long stretch of energy minima is observed, that can accommodate the conformational families lying between the low energy plateau, having two prominent centers at Φ/Ψ ≈ 280°/120° and Φ/Ψ ≈ 270°/70°. A higher energy island at Φ/Ψ ≈ 60˚/100˚ is also observed. Finally, for the disaccharide component Gal*p*NAc-α12-Qui*p*3NAc (**D**α12**A**), which carries the glycosidic linkage between two biological repeats of O5ac, a contiguous conformational pathway for the Φ values is followed during the conformational sampling, with two distinct minima at Φ/Ψ ≈ 80°/200°, corresponding to the global minimum, and Φ/Ψ ≈ 100°/270°, corresponding to the local minimum separated by an energy barrier of 3 kcal/mol. The comparison of these values with the NMR observations, have been recorded in **Table 5.a.1**.

O-antigen of E. coli O5ab

The disaccharide Qui*p*3NAc-β13-Rib*f* (**A**β13**B**) in O5ab has a long stretch of low energy minimum between two points, one centered at Φ/Ψ ≈ 280°/280° (corresponding to the south conformation) and Φ/Ψ ≈ 270˚/180˚ (corresponding to the north conformation), similar to that seen in O5ac. At Φ ≈ 270° there is a Ψ value coverage from 180° to 300°, extending over ~120° on the energy map. One other isolated low energy area is also detected, centered at Φ/Ψ ≈ 60°/240°. This energy space can be reached starting from the lowest energy minimum by a conformational pathway that is contiguous from 300° to 60° on the Φ axis, crossing the *trans* conformation at Φ=180°, but not through the *cis* conformation at Φ =0°. The conformation of the Φ/Ψ angles for Rib*f*-β14-Gal*p* (**B**β14**C**) is highly restricted and centered at about Φ/Ψ ≈ 270°/140°. A comparatively higher energy region separates the global minimum from the isolated (high) energy patch centered at Φ/Ψ ≈ 60°/140°. In Gal*p*-β13-Gal*p*NAc (**C**β13**D**), a contiguous pathway of conformational sampling can again be observed similar to O5ac, while moving from the Φ values of 60° to 300°, following a similar pattern of approaching the isolated energy island through the *trans* conformation. Finally, the Gal*p*NAc-α14-Qui*p*3NAc (**D**α14**A**) disaccharide, which is the sole difference between the positional isomers of O5ac and O5ab O-antigenic polysaccharides being studied, samples absolutely the same space as seen for **D**α12**A** in O5ac, though the population is more evenly distributed in the two observed minima. The values of the **D**α14**A** segment calculated using molecular modeling in comparison to NMR observations are have been mentioned in **Table 5.a.3**.

Combining the NMR and Modeling data

Based on 20 low energy models obtained from the conformational search on each of two hexasaccharides representing the biological repeating unit of the O-antigen PS from *E. coli* O5ac (one with Rib*f* in north conformation and the other in south conformation in the starting structures) were analyzed and clustered according to the conformation of the Rib*f* ring and the torsion angles Φ and Ψ across the glycosidic linkage. The conformation of the ribofuranoside ring was analyzed in the 40 models using the method described by Cremer and Pople [27], and the conformers clustered in two different populations: a major (north) population, comprising conformations between ^{3}E-$^{3}T_{2}$ and $^{3}T_{2}$-E_{2} and a minor (south) population comprising mainly conformations between $^{2}T_{3}$-E_{3}. The results of this clustering are summarized in **Figure 5.a.4**.

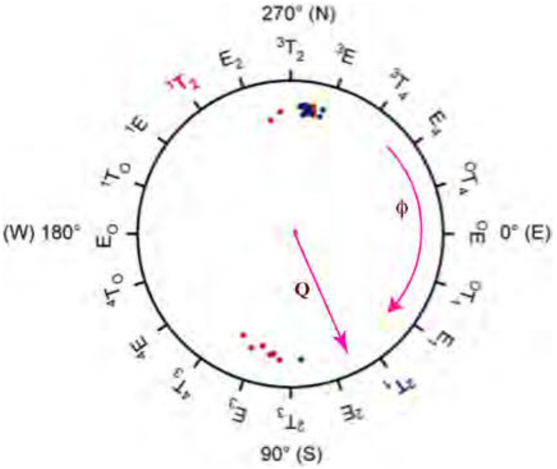


Figure 5.a.4. Conformation of the ribofuranose ring (residue B) as a function of the puckering parameters *Q* and *φ* [27]. The twenty O5ac structures of lower energy obtained from the north and the south starting models are denoted in red and blue, respectively.

Analysis of the torsion angles Φ and Ψ across the glycosidic linkages reveals several conformational families, summarized in **Table 5.a.1**. The Φ^{H}/Ψ^{H} definitions used for description of this data correspond to the light atom convention, where, Φ^{H} = H1-C1-O1-C_{x} and Ψ^{H} = C1-O1-C_{x}-H_{x}. **Table 5.a.1** shows that in most of the cases the major conformational state is the one for which the *exo*-anomeric effect prevails (Φ^{H} ~ 40° for β-D-sugars and Φ^{H} ~ –40° for α-D-sugars), with the exception of the β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc linkage (**C**β13**D**) where the Φ^{trans} conformational state (Φ^{H} ~ 180°) was also represented. In all the models where the *exo*-anomeric conformational state is present, all the Ψ^{H} glycosidic torsion angles lead to *cis*- and/or *gauche*-conformations for which the anomeric proton and the proton on the glycosyloxylated carbon are in close spatial proximity. In the *gauche*-conformations of the β-D-Qui*p*3NAc-(1→3)-β-D-Rib*f* (**A**β13**B**) and α-D-Gal*p*NAc-(1→2)-β-D-Qui*p*3NAc linkages (**D**α12**A**) both the Ψ^{+} and the Ψ^{-} states are populated (where the torsion angle Ψ^{H} is > 0° and < 0°, respectively). In the former, the Ψ^{-} state is associated with the Rib*f* (residue **B**) in $^{3}T_{2}$-^{3}E (north) conformations, whereas the Ψ^{+} state is mainly associated ^{2}E-$^{4}T_{3}$ (south) conformations. At the β-D-Rib*f*-(1→4)-β-D-Gal*p* linkage (**B**β14**C**) only the Ψ^{+} state is present. In the *gauche*-conformations of the β-D-Gal*p*-

(1→3)-α-D-Gal*p*NAc linkage (**C**β13**D**) three different states were identified and named *gauche* + (90° > Ψ^{+} > 30°), *cis* (30° > Ψ^{cis} > −30°), and *gauche* – (−30° > Ψ^{-} > −90°).

All these observations are in agreement with the maps of **Figure 5.a.3** (*top panel*), where the average torsion angles for each relevant family (described in **Table 5.a.1**) can be accommodated in the low energy regions of the corresponding maps. NMR was used to study a 14-repeat of the O5ac polysaccharide.

Table 5.a.1. Averaged torsion angles for each of the conformational families identified in the conformational sampling of the two hexasaccharides, representing the tetrasaccharide biological repeating unit (with the monosaccharide linked to the reducing and non-reducing ends to account for the linkage effect) of the O-antigen PS from *E. coli* O5ac. The maximum and minimum torsion angle values considered for each conformational family are indicated in square brackets.

Glycosidic linkage	Family	Average values (deg.)		State	Average values (deg.)		State
		Φ	Φ^{H}		Ψ	Ψ^{H}	
AB	Ψ^{-}_{AB} (a)	271 [262,290]	31 [22,51]	*exo*	185 [177,203]	−53 [−61, −33]	*gauche* –
	Ψ^{+}_{AB} (b)	283 [278,291]	43 [38,52]		278 [248,300]	45 [13,68]	*gauche* +
BC	Ψ^{+}_{BC}	279 [266,291]	42 [28,56]	*exo*	144 [134,170]	25 [14,52]	*gauche* +
CD	Ψ^{-}_{CD}	269 [255,280]	30 [16,42]	*exo*	73 [63,80]	−48 [−58, −41]	*gauche* –
	Ψ^{cis}_{CD}	288 [282,301]	50 [44,62]	*exo*	120 [96,145]	1 [−25,27]	*cis*
	Ψ^{+}_{CD}	302 [300,304]	64 [63,66]	*exo*	174 [161,187]	58 [45,71]	*gauche* +
	Φ^{trans}_{CD}	63 [50,72]	180 [168,189]	*trans*	111 [99,120]	−10 [−23,0]	
DA	Ψ^{-}_{DA}	78 [62,96]	−41 [−58, −22]	*exo*	203 [192,213]	−38 [−51, −28]	*gauche* –
	Ψ^{+}_{DA}	104 [90,120]	−15 [−30,1]	*exo*	273 [268,283]	37 [31,47]	*gauche* +

(a) Associated to the ribofuranose ring in $^{3}T_{2}$–^{3}E (N) conformations. (b) associated to the ribofuranose ring in E_{2}–$^{3}T_{2}$ (N) or ^{2}E–$^{4}T_{3}$ (S) conformations. The term "*exo*" denotes those conformations where the *exo*-anomeric effect prevails.

Relevant proton-proton distances (r_{ij}) were extracted from the models and plotted as a function of the dihedral angles Ψ^{H} in the respective glycosidic linkages. Each of the different conformational families identified in **Table 5.a.1** give rise to a very distinctive set of effective proton-proton distances (**Figure 5.a.5**), which allows for comparison with experimental observations.

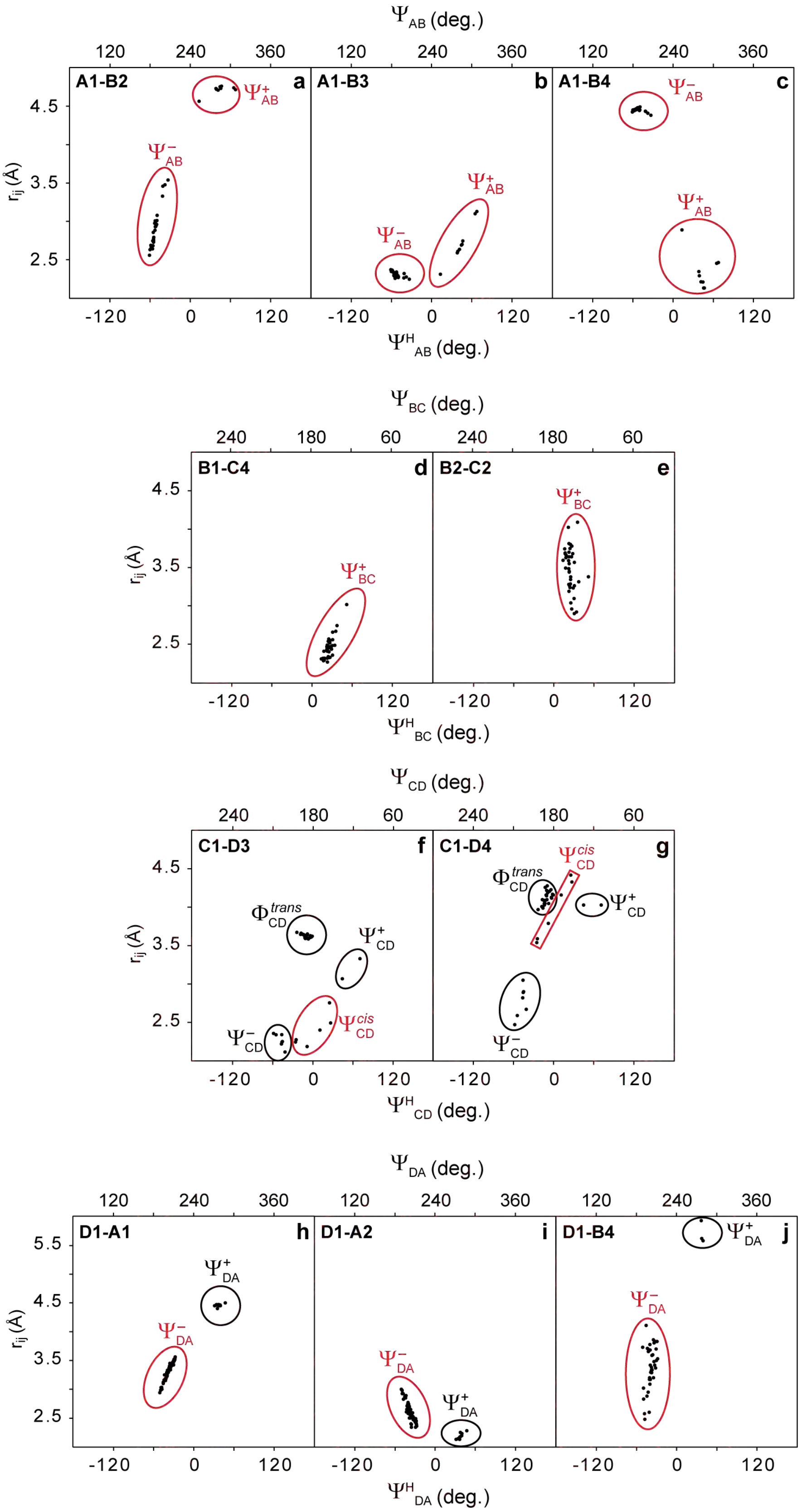


Figure 5.a.5. Scatter plots of r_{ij} vs Ψ^H obtained from conformational sampling on the two hexasaccharide models representing the biological repeating unit of the O-antigenic PS from *E. coli* O5ac. The families that explain the experimental data are indicated in red.

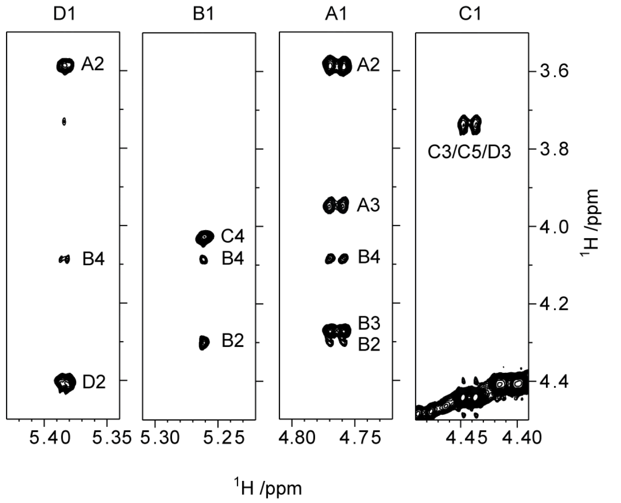


Figure 5.a.6. Selected region of the 2D ^{1}H,^{1}H-NOESY spectrum of the O-antigen PS from *E. coli* O5ac recorded at 700 MHz with a mixing time of 80 ms. Correlations from the anomeric protons are indicated with pertinent annotations.

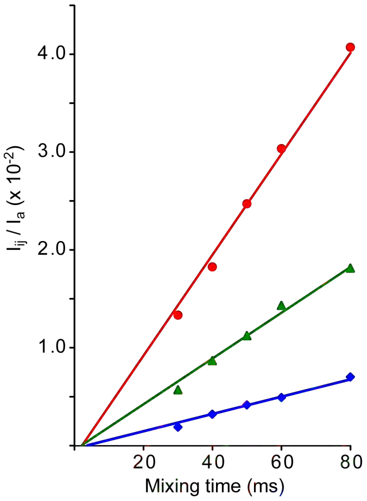


Figure 5.a.7. Plots of the normalized volume intensities versus mixing time obtained for the intra-residue correlation between H1 and H2 of Gal*p*NAc (•), the trans-glycosidic correlation between H1 of Gal*p*NAc and H2 of Qui*p*NAc (▲) and the long-range correlation between H1 of GalpNAc and H4 of Rib*f* (♦). The data was obtained from 2D ^{1}H,^{1}H-NOESY experiments recorded at 700 MHz.

The conformations of the O-antigen PS from *E. coli* O5ac were examined using proton-proton distances obtained from ^{1}H,^{1}H-NOESY experiments. The ^{1}H chemical shifts assignments have been reported earlier [19]. The ^{1}H,^{1}H-NOESY spectrum of the O-antigen PS from *E. coli* O5ac (**Figure 5.a.6**) shows a number of strong intra-residue correlations as well as inter-residue correlations across the glycosidic linkages. The cross-peaks intensities measured at different mixing times were used to generate NOE build-up curves which were analyzed as detailed by Macura *et al.* [28].

Proton-proton cross-relaxation rates were extracted from the slope of these curves (**Figure 5.a.7**) and the calculated intra-residue distance between H1 and H2 of the Gal*p*NAc residue (**D**) was used as reference to obtain the unknown distances using the isolated spin-pair approximation (ISPA) [29]. Cross-relaxation rates obtained for the different proton pairs and their corresponding effective distances are compiled in **Table 5.a.2**, and compared to those obtained by conformational sampling.

Table 5.a.2. Cross relaxation rates and effective distances determined for the O-antigen polysaccharide from *E. coli* O5ac from 2D ^{1}H,^{1}H-NOESY experiments at 500 and 700 MHz. Calculated distances are reported for the different conformational families identified; the values used to explain the experimental data are highlighted in bold.

^{1}H-^{1}H correlation	**σ_{ij} at 700 MHz (× 10^{-3} s^{-1})**	**σ_{ij} at 500 MHz (× 10^{-3} s^{-1})**	**r_{ij} (Å) NMR**	**r_{calc} (Å)**	**r_{calc} (Å) of averaged populations**
A1-B2	103	n.d.	3.08	**2.82** (Ψ_{AB}^{-}) / **4.71** (Ψ_{AB}^{+})	**2.95** [3:1]
A1-B3	446	406	2.42	**2.31** (Ψ_{AB}^{-}) / **2.63** (Ψ_{AB}^{+})	**2.37** [3:1]
A1-B4	159	144	2.87	**4.45** (Ψ_{AB}^{-}) / **2.29** (Ψ_{AB}^{+})	**2.86** [3:1]
B1-C4	354	311	2.52	**2.43** (Ψ_{BC}^{+})	
B2-C2	64	n.d.	3.34	**3.40** (Ψ_{BC}^{+})	
C1-D3[b]	788[b]	690[b]	2.21	2.25 (Ψ_{CD}^{-}) / **2.35** (Ψ_{CD}^{cis}) / 3.18 (Ψ_{CD}^{+}) / 3.63 (Φ_{CD}^{trans})	
C1-D4	n.o.	n.o.	> 3.50	2.72 (Ψ_{CD}^{-}) / **3.87** (Ψ_{CD}^{cis}) / 4.60 (Ψ_{CD}^{+}) / 4.12 (Φ_{CD}^{trans})	
D1-A1	49	n.o.	3.50	**3.28** (Ψ_{DA}^{-}) / 4.46 (Ψ_{DA}^{+})	
D1-A2	234	213	2.69	**2.58** (Ψ_{DA}^{-}) / 2.18 (Ψ_{DA}^{+})	
D1-B4	74	64	3.28	**3.19** (Ψ_{DA}^{-}) / 5.70 (Ψ_{DA}^{+})	
D1-D2[a]	463	427	2.40[a]	**2.40**	
D4-D5	446	395	2.42	**2.47**	
A1-A3	279	258	2.61	**2.59**	
B1-B2	139	119	2.95	**2.75** (N) / **2.95** (S)	**2.81** [3:1]
B1-B4	72	71	3.26	**3.49** (N) / **3.71** (S)	**3.54** [3:1]

[a] distance used as reference. [b] Overlapping with D1/D3 and D1/D5. Cross relaxation rates were calculated subtracting the theoretical values obtained from the models. n.d. = not determined due to overlapping. n.o. = not observed or at the noise level. N = north conformation of Rib*f*. S = south conformation of Rib*f*.

At the β-D-Qui*p*3NAc-(1→3)-β-D-Rib*f* linkage (**A**β13**B**) three distances were determined from the anomeric proton of residue **A** to H2, H3 and H4 of residue **B** (3.08, 2.42 and 2.87 Å, respectively). These distances were compared with the calculated distances obtained for each of the two populations identified by molecular modeling (Ψ_{CB}^{-} and Ψ_{CB}^{+} in **Figure 5.a.5a-c**). The experimental data can only be explained by considering a population distribution of 75 % and 25% of the conformational families Ψ_{CB}^{-} and Ψ_{CB}^{+}, respectively. The calculated values for the averaged populations (2.95, 2.37 and 2.86 Å, respectively) then differ only by 0.13, 0.05 and 0.01 Å, respectively, from the experimental values. At the β-D-Rib*f*-(1→4)-β-D-Gal*p* linkage (**B**β14**C**) two distances were determined: a shorter being 2.52 Å between H1 of residue **B** and H4 of residue **C**, as well as a longer distance of 3.34 Å between H2 of residue **B** and H2 of residue **C**. These results are comparable to the distances calculated from the low-energy models obtained from the conformational sampling (2.43 Å and 3.40 Å, respectively), which only comprise conformers of a Ψ_{CB}^{+} family (**Figure 5.a.5d-e**). At the β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc linkage (**C**β13**D**) the determination of the cross-relaxation rate between H1 of residue **C** and H3 of residue **D** was hampered by the overlapping of this cross-peak with two peaks arising from intra-residue correlations in residue **C** (H1-H3 and H1-H5). Therefore, the inter-residue distance was deduced from the cross-relaxation rate obtained for the three overlapping cross-peaks (1.525 s^{-1} and 1396 s^{-1} at 700 and 500 MHz, respectively) by subtracting the theoretical cross-relaxation rates predicted for each of the intra-residue cross-peaks ($r_{H1,H3}$ = 2.61 Å and $r_{H1,H5}$ = 2.41 Å). Using this approach, we found out that the distance between H1 of residue **C** and H3 of residue **D** is actually very short (~ 2.21 Å), which is a very important conformational restriction when examining the different conformational states found across this glycosidic linkage. In addition, no cross-peak of significant intensity was observed between H1 of residue **C** and H4 of residue **B**, indicating that these two atoms are separated by a distance larger than 3.5 Å (which is the shorter distance determined in this study). This data can only be explained by the proton-proton distances measured in the conformers of the Ψ_{CD}^{cis} family (2.35 Å and 3.87 Å, respectively), and indicate that the remaining sub-families (Ψ_{CD}^{-}, Ψ_{CD}^{+} and $\Phi_{\mathrm{CD}}^{trans}$in **Figure 5.a.5f-g**) are not significantly populated. At the α-D-Gal*p*NAc-(1→2)-β-D-Qui*p*3NAc linkage (**D**α12**A**) two *trans*-glycosidic effective distances could be determined, a short one between H1 in residue **D** and H2 in residue **A** (2.69 Å), and a long one between H1 of residue **D** and H1 of residue **A** (3.50 Å, determined only at the higher field). These results are consistent with the averaged proton-proton distances calculated in the conformers of the Ψ_{DA}^{-} family (2.58 Å and 3.28 Å, respectively) (**Figure 5.a.5h-i**). In addition, a long distance correlation was observed between H1 in residue **D** and H4 in residue **B** (3.28 Å), which differs only by 0.9 Å of the average value obtained for conformers of the Ψ_{DA}^{-} family ($r_{D1,B4}$ = 3.19 Å) (**Figure 5.a.5j**), providing additional evidence that the Ψ_{DA}^{+} conformational family does not represent a significant contribution to the experimental observations ($r_{D1,B4}$ = 5.70 Å).

The O-antigen PS of E. coli O5ab

In the ^{1}H,^{1}H-NOESY spectrum of the O-antigen PS from *E. coli* O5ab (data not shown) a number of intra- and inter-residue correlations similar to those observed in the ^{1}H,^{1}H-NOESY spectrum of the O-antigen PS from *E. coli* O5ac were identified across the glycosidic linkages **A**β13**B**, **B**β14**C** and **C**β13**D**. This observation indicates that the conformations across these linkages are similar to those observed for the O-antigen PS from *E. coli* O5ab. Based on this observation, we focused the analysis of the O-antigen PS from *E. coli* O5ab mainly from the point of view of the α-D-Gal*p*NAc-(1→4)-β-D-Qui*p*3NAc linkage, which is the only structural difference with respect to the O-antigen PS from *E. coli* O5ac.

The 20 low energy models obtained from the conformational search on each of two hexasaccharides representing the biological repeating unit of the O-antigen PS from *E. coli* O5ab (one with Rib*f* in the north conformation and the other in the south conformation in the starting structures) were analyzed and clustered according to the conformational states across the α-D-Gal*p*NAc-(1→4)-β-D-Qui*p*3NAc glycosidic linkage. Three conformational families were identified, all corresponding to conformational states where the *exo*-anomeric effect prevails (**Table 5.a.3**). The two major families correspond to conformers where the torsion angle Ψ^{H} leads to *gauche*-conformations, and both Ψ^{+} and Ψ^{-} states are represented. In addition, a Ψ^{trans} conformational state was observed in a few models (where the anomeric carbon of residue **D** and H4 of residue **A** are in an *trans* arrangement). The distances from H1 of the Gal*p*NAc residue (**D**) to relevant protons in the Qui*p*3NAc residue (**A**) were extracted from the models and plotted as a function of the torsion angle Ψ_{DA} (**Figure 5.a.8**).

Table 5.a.3. Averaged torsion angles obtained from conformational sampling of the two hexasaccharides representing the biological repeating unit of the O-antigen PS from *E. coli* O5ab.

Glycosidic linkage	Family	Average values (deg.)		State	Average values (deg.)		State
		Φ	Φ^{H}		Ψ	Ψ^{H}	
DA	Ψ_{DA}^{-}	77 [55,91]	−43[−65, −28]	*exo*	206 [185,220]	−35 [−58, −21]	*gauche* −
	Ψ_{DA}^{+}	103 [97,107]	−17 [−23, −12]	*exo*	279 [272,291]	42 [36,55]	*gauche* +
	Ψ_{DA}^{trans}	96 [82,104]	−24 [−38, −15]	*exo*	80 [50,107]	−164 [−192, −138]	*trans*

In order to measure effective proton-proton distances selective excitation of the anomeric proton of the Gal*p*NAc residue (**D**) was carried out using one-dimensional ^{1}H,^{1}H-NOESY experiments (**Figure 5.a.9b**) to generate NOE build-up curves (**Figure 5.a.9c**) which were analyzed in detail using the same approach as described above. Proton-proton cross-relaxation rates were extracted from the slope of these curves and a reference distance of 2.39 Å between H1 and H2 of the GalpNAc residue (**D**) was used for distance calibration. These results are compiled in **Table 5.a.4**. The effective distance between H1 of the Gal*p*NAc (residue **D**) and H4 of the Qui*p*3NAc residue (**A**) determined experimentally corresponded to 2.39 Å. This value is slightly lower than the average distance measured in conformers of the Ψ_{DA}^{+} family (2.56 Å) but slightly higher than the average distance obtained from models of the Ψ_{DA}^{-} conformational family (2.20 Å). Therefore, any of these two models, considered independently or as a mixed contribution, may explain the experimental data. The effective distance between H1 of the Gal*p*NAc residue (**D**) and H3 of the Qui*p*3NAc residue (**A**) determined experimentally (3.20 Å) is slightly higher than the average distance obtained from models of the Ψ_{DA}^{-} conformational family (2.91 Å) and significantly lower than the average distance obtained from models of the Ψ_{DA}^{+} (4.40 Å) conformational family. If a shared contribution of both conformational families is considered, where the population distribution is about 50% for each of them, experimental and calculated values only differ by less than 0.02 Å. The Ψ_{DA}^{trans} conformational family does not represent a significant contribution to the distance determined experimentally, as the average values calculated from these models differ considerably from the experimental data.

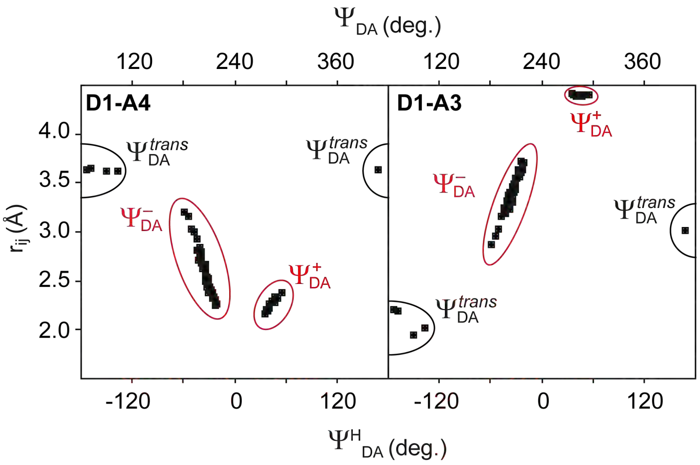


Figure 5.a.8. Scatter plots of r_{ij} vs. Ψ_{DA} obtained from the hexasaccharide models representing the biological repeating unit of the O-antigenic PS from *E. coli* O5ab. The conformational families that explain the experimental data are indicated in red.

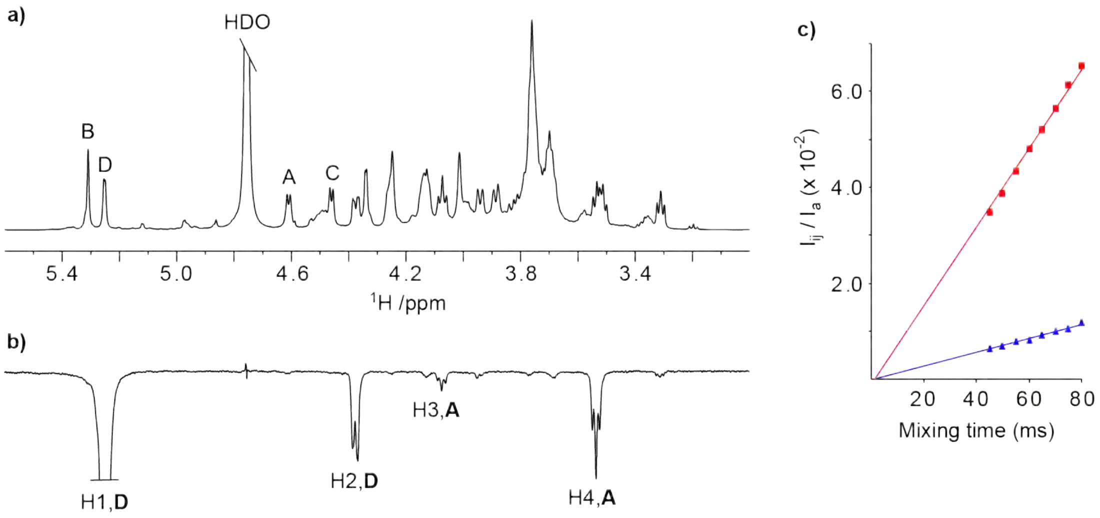


Figure 5.a.9. (a) ^{1}H-NMR spectrum of the O-antigen PS from *E. coli* O5ab and (b) 1D ^{1}H,^{1}H-SPFGSE NOESY experiment (mixing time 80 ms) with selective excitation of the H1 resonance of Gal*p*NAc and (c) plot of the normalized intensities versus mixing time obtained for the resonance of H2 of Qui*p*3NAc by selective excitation of H1 of Gal*p*NAc in the 1D ^{1}H,^{1}H-SPFGSE NOESY experiments.

Table 5.a.4. Cross relaxation rates and effective distances determined for the O-antigen polysaccharide from *E. coli* O5ab from 1D ^{1}H,^{1}H-NOESY experiments at 700 MHz. Calculated distances are informed for the different conformational families identified; the values used to explain the experimental data are highlighted in bold.

^{1}H-^{1}H correlation	σ_{ij} at 700 MHz ($\times 10^{-3}$ s^{-1})	r_{ij} (Å) NMR	r_{calc} (Å)	r_{calc} (Å) of averaged populations
D1-A3	144	3.20	**2.91** (Ψ_{DA}^{-}) / **4.40** (Ψ_{DA}^{+}) / 2.14 (Ψ_{DA}^{trans})	**3.22** [1:1]
D1-A4	819	2.39	**2.56** (Ψ_{DA}^{-}) / **2.24** (Ψ_{DA}^{+}) / 3.36 (Ψ_{DA}^{trans})	**2.37** [1:1]
D1-D2[a]	824	2.39	2.39	

[a] distance used as reference.

O-antigenic polysaccharide modeling

SYBYL X 1.3 was used to build the 3D models of the O5ac and O5ab polysaccharide chains guided by the NMR observations. Fragments comprising eight biological repeat units were constructed that corresponded to all the combinations illustrated on the disaccharide energy maps as shown in **Figure 5.a.10**. Interestingly, the only O5ac structure devoid of steric clashes corresponded to the major (75%) population observed in conformational sampling as well as in the NMR studies. The stable helical structure represented the north conformation of the β-D-Rib*f* constituent. The polysaccharide conformation of O5ac resulted in a 2-fold helix with a pitch of 15.6 Å. In the schematic representation of this helix (**Figure 5.a.10. O5ac. a**), we see that that there is no space inside the helix. The adjacent N-acetyl groups of α-D-Gal*p*NAc and β-D-Qui*p*3NAc interact. In case of O5ab, a regular helix with eight biological repeat units was constructed, that corresponded to 50% of the population observed in the NMR results. The helix pitch is measured to be 29.27 Å in a 2-fold helix and has a form resembling that observed in O5ac (**Figure 5.a.10. O5ab. (***upper panel***) b**). The β-D-Rib*f* constituent was in a conformation between ^{3}E and $^3T_2$ representing the major population as confirmed by both molecular modeling and NMR studies. The other 50% of the population had a helix pitch of 24.2 Å in a 2-fold helix (with the β-D-Rib*f* in the north conformation) and had a form distinct from that formed by the major population observed in O5ac and it had a tendency to fold upon itself, beyond a degree of polymerization equal to four (**Figure 5.a.10. O5ab. (***lower panel***) b**). Both these O5ab polysaccharide representations lacked steric hindrances. Thus, for the polysaccharide models a common pattern was observed for both O5ac and O5ab.

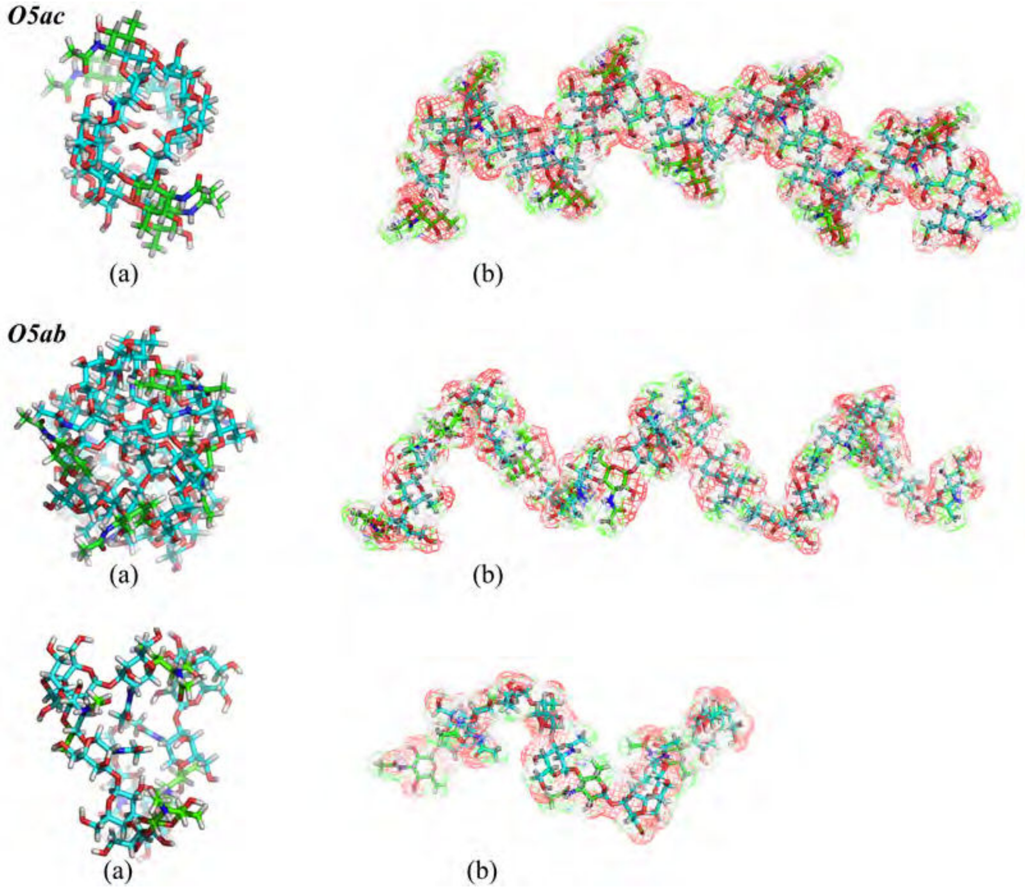


Figure 5.a.10. The schematic representation of the polysaccharide helices, describing the major populations observed in O5ac (8 repeating units) and O5ab (8 and 4 repeating units, in the upper and lower panels for O5ab, respectively) (a) as viewed perpendicularly, and (b) along the length of the helix.

Immunochemical properties of the oligosaccharide fragments

Both intra- and inter-species serological cross-reactions based on the O-antigens are observed frequently between different bacterial isolates. These cross-reactivities are based on structural similarities in the cell wall associated antigens [30].

It was reported earlier that the strains of *E. coli* O5ac and O5ab cannot be distinguished by anti-O5ac or anti-O5ab sera [19], and this strong cross-reactivity is a limitation for differentiation of these two strains using the existing serotyping methodologies. As it was shown before, the structures of the O-antigen PS of these two strains are not identical but remarkably similar, differing only at the linkage position between the α-D-Gal*p*NAc and β-D-Qui*p*3NAc residues (**Figure 5.a.1**). It seems that this structural difference is not 'visible' and hence not recognized by the mono-specific anti-sera. Thus, the conformational analysis of these two polysaccharides may help to explain the cross-reactivity observed between the two strains.

In the 3D model of the O-antigenic polysaccharide of *E. coli* O5ac, illustrated in **Figure 5.a.10** (*top panel*)**.b**, it can be observed that the α-D-Gal*p*NAc residue is located on the inside of the helix, and the external accessibility is hindered by the N-acetyl moiety of β-D-Qui*p*3NAc. Similarly, in the helical structure of the O-antigen PS from *E. coli* O5ab, as shown in **Figure 5.a.10** (*middle panel*)**.c**, the αGal*p*NAc residue is once again located in the interior of the helix and only the N-acetyl group is partially exposed to the surface. Further analysis of the models (comprising eight biological repeating units each for O5ac and O5ab) reveals that in both cases the same epitope is exposed to the surface and corresponds to the β-D-Qui*p*3NAc-(1→3)-β-D-Rib*f*-(1→4)-β-D-Gal*p* fragment as illustrated in **Figure 5.a.11**.

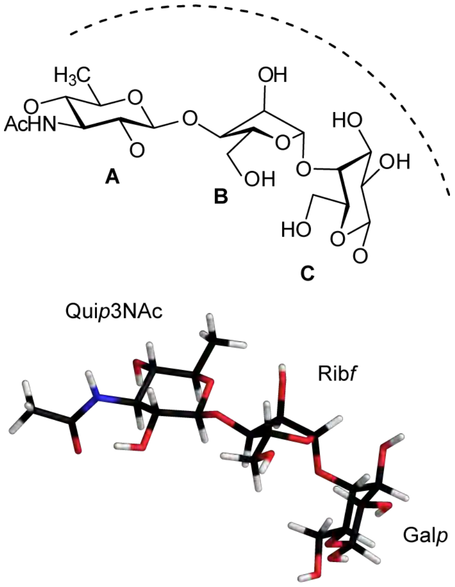


Figure 5.a.11. Common epitope observed in the O-antigen PS from *E. coli* O5ac O5ab in schematic chemical (ring) representation of (*top*) and molecular models (*below*). Dashed lines denote the face exposed to the surface of the helix.

Conclusion

It was the aim of the present work to fully characterize the conformational behavior of the two enteroaggregative pathogenic *E. coli* strains. Previous immunological investigations, accompanied by the elucidation of the sequence of the sugars present in the repeat units could not bring unequivocal conclusions about neither the eventual occurrence of a common antibody, nor its nature and size. In the present study, a combination of high resolution NMR and molecular modeling methods were used to elucidate the conformation of the two strains. The NMR study was based on the analysis of intra- and inter-residue distances using NOE build-up curves. Molecular models of the repeating units and their extension to polysaccharides were obtained, taking into account the entire conformational flexibility as assessed by the force field and genetic algorithm. The agreements between experimentally measured and calculated distances can only be obtained by considering an averaging of several low energy conformations observed in the molecular models.

Among these low energy conformations only some of them can be propagated in the form of long polymeric chain thereby allowing for investigating the eventual occurrence of any conformational epitope, which could not be found. Instead a common glycan epitope in the two strains was identified. Possibly, due to this reason the antibodies fail to differentiate between them. This aspect could be interesting to probe from the point of view of the pathogenicity of these bacteria. Indeed, some strains of *E. coli* are known to demonstrate cross-reactivity, and further studies from the immunological aspect would shed more light on this front. The extrapolation of the present conclusions to the preparation of a future pathogen glycan array would suggest that three contiguous repeat units, i.e a dodecasaccharide, are sufficient to be immobilized on glass slides. Such dimensions are compatible with a full recognition by monoclonal antibodies without generating the possible steric crowding that a long polymeric chain would cause on a glycan array.

References:

1. Nataro JP, Steiner T, Guerrant RL: **Enteroaggregative *Escherichia coli***. *Emerg Infect Dis* 1998, **4**(2):251-261.

2. Bhan MK, Khoshoo V, Sommerfelt H, Raj P, Sazawal S, Srivastava R: **Enteroaggregative *Escherichia coli* and *Salmonella* associated with nondysenteric persistent diarrhea**. *Pediatr Infect Dis J* 1989, **8**(8):499-502.

3. Bhan MK, Raj P, Levine MM, Kaper JB, Bhandari N, Srivastava R, Kumar R, Sazawal S: **Enteroaggregative *Escherichia coli* associated with persistent diarrhea in a cohort of rural children in India**. *J Infect Dis* 1989, **159**(6):1061-1064.

4. Cravioto A, Tello A, Navarro A, Ruiz J, Villafan H, Uribe F, Eslava C: **Association of *Escherichia coli* HEp-2 adherence patterns with type and duration of diarrhoea**. *Lancet* 1991, **337**(8736):262-264.

5. Wanke CA, Schorling JB, Barrett LJ, Desouza MA, Guerrant RL: **Potential role of adherence traits of *Escherichia coli* in persistent diarrhea in an urban Brazilian slum**. *Pediatr Infect Dis J* 1991, **10**(10):746-751.

6. Nataro JP, Deng Y, Cookson S, Cravioto A, Savarino SJ, Guers LD, Levine MM, Tacket CO: **Heterogeneity of enteroaggregative *Escherichia coli* virulence demonstrated in volunteers**. *J Infect Dis* 1995, **171**(2):465-468.

7. Smith HR, Cheasty T, Rowe B: **Enteroaggregative *Escherichia coli* and outbreaks of gastroenteritis in UK**. *Lancet* 1997, **350**(9080):814-815.

8. **Glyco3D: A site for glycosciences** [http://glyco3d.cermav.cnrs.fr/glyco3d]

9. Clark M, Cramer RD, Van Opdenbosch N: **Validation of the general purpose tripos 5.2 force field**. *Journal of Computational Chemistry* 1989, **10**(8):982-1012.

10. Imberty A, Bettler E, Karababa M, Mazeau K, Petrova P, Pérez S: **Building sugars: The sweet part of structural biology.** . In: *Perspectives in Structural Biology".* Edited by M. Vijayan NYASK. Hyderabad: Indian Academy of Sciences and Universities Press, Hyderabad; 1999: 392-409.

11. Frank M, Bohne-Lang A, Wetter T, Lieth CW: **Rapid generation of a representative ensemble of *N*-glycan conformations**. *In Silico Biol* 2002, **2**(3):427-439.

12. Frank M, Lutteke T, von der Lieth CW: **GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages**. *Nucleic Acids Res* 2007, **35**(Database issue):287-290.

13. **TINKER Molecular Modeling** [dasher.wustl.edu/tinker/]

14. **Conformational Analysis Tool (CAT)** [http://www.md-simulations.de/CAT/]

15. Rosen J, Miguet L, Perez S: **Shape: automatic conformation prediction of carbohydrates using a genetic algorithm**. *J Cheminform* 2009, **1**(1):16.

16. Allinger NL, Li F, Yan L, Tai JC: **Molecular mechanics (MM3) calculations on conjugated hydrocarbons**. *Journal of Computational Chemistry* 1990, **11**(7):868-895.

17. Allinger NL, Yuh YH, Lii JH: **Molecular mechanics. The MM3 force field for hydrocarbons. 1**. *Journal of the American Chemical Society* 1989, **111**(23):8551-8566.

18. Pérez S, Imberty A, Engelsen SB, Gruza J, Mazeau K, Jimenez-Barbero J, Poveda A, Espinosa J-F, van Eyck BP, Johnson G *et al*: **A comparison and chemometric analysis of several molecular mechanics force fields and parameter sets applied to carbohydrates**. *Carbohydrate Research* 1998, **314**(3-4):141-155.

19. Urbina F, Nordmark E-L, Yang Z, Weintraub A, Scheutz F, Widmalm Gr: **Structural elucidation of the O-antigenic polysaccharide from the enteroaggregative *Escherichia coli* strain 180/C3 and its immunochemical relationship with *E. coli* O5 and O65**. *Carbohydrate Research* 2005, **340**(4):645-650.

20. Thrippleton MJ, Keeler J: **Elimination of zero-quantum interference in two-dimensional NMR spectra**. *Angew Chem Int Ed* 2003, **42**(33):3938-3941.

21. Bohlen JM, Bodenhausen G: **Experimental aspects of Chirp NMR spectroscopy**. *Journal of Magnetic Resonance, Series A* 1993, **102**(3):293-301.

22. Kupče Ē: **Applications of adiabatic pulses in biomolecular nuclear magnetic resonance**. In: *Nuclear Magnetic Resonance of Biological Macromolecules.* Edited by James TL, Dötsch V, Schmitz U, vol. A. San Diego, CA: Academic Press; 2002: 82-111.

23. Tannus A, Garwood M: **Adiabatic pulses**. *NMR Biomed* 1997, **10**(8):423-434.

24. MacLean LL, Perry MB: **Structural characterization of the serotype O:5 O-polysaccharide antigen of the lipopolysaccharide of *Escherichia coli* O:5**. *Biochemistry and Cell Biology* 1997, **75**(3):199-205.

25. Stott K, Keeler J, Van QN, Shaka AJ: **One-dimensional NOE experiments using pulsed field gradients**. *J Magn Reson* 1997, **125**(2):302-324.

26. Kupce E, Boyd J, Campbell ID: **Short selective pulses for biochemical applications**. *Journal of Magnetic Resonance, Series B* 1995, **106**(3):300-303.

27. Cremer D, pople JA: **A general defintion of ring puckering coordinates**. *J Am Chem Soc* 1975, **97**(6):1354-1358.

28. Macura S, Farmer II BT, Brown LR: **An improved method for the determination of cross-relaxation rates from NOE data**. *Journal of Magnetic Resonance (1969)* 1986, **70**(3):493-499.

29. Thomas PD, Basus VJ, James TL: **Protein solution structure determination using distances from two-dimensional nuclear Overhauser effect experiments : Effect of approximations on the accuracy of derived structures**. *Proceedings of the National Academy of Sciences of the United States of America* 1991, **88**:1237-1241.

30. Nordmark E-L: **Structural and interaction studies of bacterial polysaccharides by NMR spectroscopy**. *Doctoral dissertation.* Stockholm: Stockholm University; 2004.

CHAPTER 5.b

Molecular modeling and conformational analysis of the O-antigenic polysaccharide of the pathogenic *Escherichia coli* strain 1303.[1]

Introduction

Escherichia coli is one of the major causative agents in mastitis, which is a major disease in dairy herds. Once infected, the animal is often culled and thus the dairy industry incurs considerable loss [1, 2]. Due to the economic implications of mastitis on the dairy industry and the health risk to the consumers, significant efforts have been made to identify factors that make dairy cows susceptible to infections of the mammary gland, that is most frequent at parturition. In that period, infections with *E. coli* often cause severe clinical symptoms [3, 4] accompanied by reduction in milk yield, altered milk composition and extensive damage of mammary tissue [5].

In *E. coli*, the lipopolysaccharide (LPS) is suspected to play a crucial role during infection [6]. The sequence of the bovine mastitis isolate Ec1303 was determined using chemical analyses, mass spectrometry and 1D and 2D nuclear magnetic resonance (NMR) spectroscopy methods. The O-antigenic biological repeating unit was characterized as

-[→4)-β-D-Qui*p*3NAc-(1→3)-α-L-Fuc*p*2OAc-(1→4)-β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc-(1→]-

in which the O-acetyl substitution was non-stoichiometric[2] [6]. The LPS of enterobacteria is frequently found to contain various non-stoichiometric substituents on the polysaccharide backbone [7]. These substituents can modify the biological activity of the LPS that includes, variable outer membrane stability, tolerance to cationic antibiotics, pathogenicity and sensitivity to bacteriophages that infect enterobacteria [7].

[1] Anita Sarkar & Serge Pérez. AS designed the simulations, modeled the polysaccharides, calculated the low energy maps and wrote the manuscript sans the perspectives.

[2] Non-stoichiometric compounds are chemical compounds with an elemental composition that cannot be represented by a ratio of well-defined natural numbers, and therefore violate the law of definite proportions.

The two existing serotypes of O5, designated as O5ac and O5ab, which have been described in **Chapter 5.a.**, have a close resemblance with the O-antigenic polysaccharide of Ec1303. They differ from O5ab only by the presence of α-L-Fuc*p*OAc (or α-L-Fuc*p*) as the second monosaccharide in the biological repeating units instead of a β-D-Rib*f*, and from O5ac similarly but with an additional difference of α-**1→4** instead of a α-**1→2** linking two consecutive biological repeat units.

Here, we report the conformational analysis of two molecular models of the O-antigenic polysaccharides of Ec1303, as illustrated below (**Figure 5.b.1**):

-[→4)-β-D-Qui*p*3NAc-(1→3)-α-L-Fuc*p*-(1→4)-β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc-(1→]-

-[→4)-β-D-Qui*p*3NAc-(1→3)-α-L-Fuc*p*2OAc-(1→4)-β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc-(1→]-

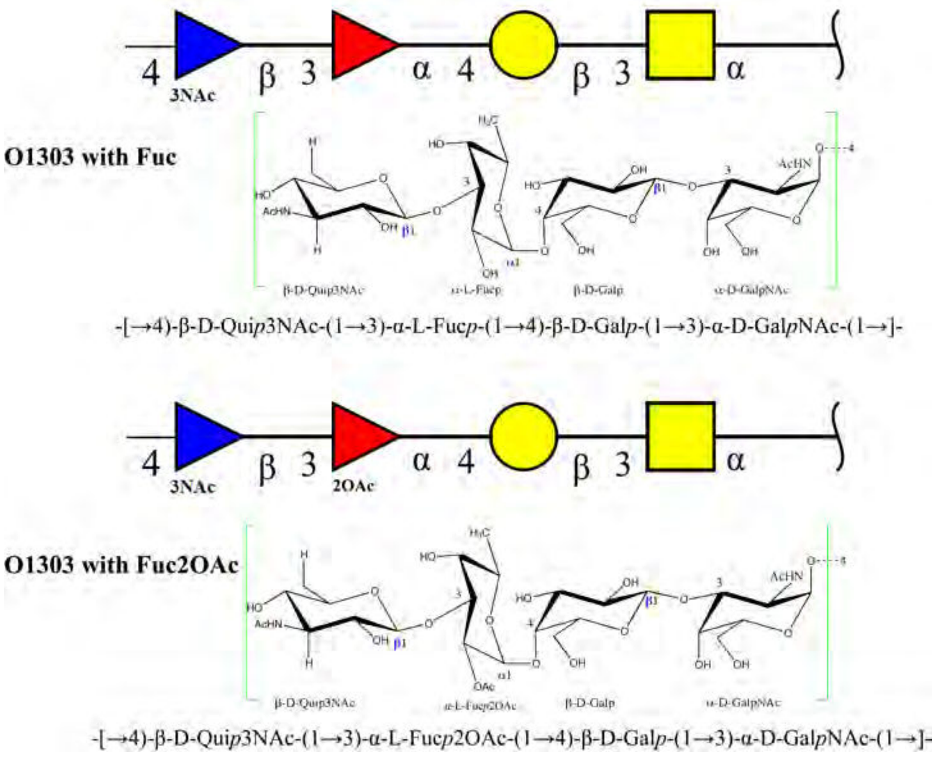

4
3NAc
β 3
α 4
β 3
α
O1303 with Fuc
β-D-Quip3NAc
α-L-Fucp
β-D-Galp
α-D-GalpNAc
-[→4)-β-D-Qui*p*3NAc-(1→3)-α-L-Fuc*p*-(1→4)-β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc-(1→]-
4
3NAc
β 3
2OAc
α 4
β 3
α
O1303 with Fuc2OAc
β-D-Quip3NAc
α-L-Fucp2OAc
β-D-Galp
α-D-GalpNAc
-[→4)-β-D-Qui*p*3NAc-(1→3)-α-L-Fuc*p*2OAc-(1→4)-β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc-(1→]-

Figure 5.b.1. Structure of the biological repeating units of the O-antigen PS from *E. coli* 1303 (with and without O-acetylation on α-L-Fuc*p*) in CFG[3]-cartoon notation (*top*), schematic chemical (ring) representation (*middle*) and linear nomenclature (*bottom*), respectively.

[3] CFG is an abbreviation for the Consortium for Functional Glycomics.

Materials and methods

The procedure followed for this study was identical to that described in **Chapter 5.a**.

Nomenclature

The Φ/Ψ definitions used for the adiabatic maps (as well their description in the *results & discussion* section) correspond to the heavy atom convention for (1→x), where, Molecular modeling: Φ → O5-C1-O1-C_x and Ψ → C1-O1-C_x-C_{x+1}. Most of the constituent monosaccharides were built from optimized base types found in Glyco3D [8]. β-D-Qui*p*3NAc was built from the base type found in PDB (1MMY) and N-acetylated using an optimized NAc fragment from Glyco3D. All the disaccharide and oligosaccharide starting structures were modeled with SYBYL X 1.3 [9] using the pim_2010 parameters for atom types and partial charges [10] and the Tripos force-field parameters [11].

Results and discussion

Molecular modeling of the oligosaccharides

The repeat units of the O1303 comprise identical glycosidic linkages and near identical constituent monosaccharides between each other, with the difference arising only with the α-L-Fuc being O-acetylated or not. Yet, to cover all possibilities of conformational behavior, for each tetrasaccharide biological repeat unit of O1303, an α-D-Gal*p*NAc was added to the non-reducing end and a β-D-Qui*p*3NAc at the reducing end, to eliminate linkage end effects.

The construction of the 3D O-antigenic models of Ec1303 required five glycosidic linkages, namely, Qui*p*3NAc-β1→3-Fuc*p*, Fuc*p*-α1→4-Gal*p*, Gal*p*-β1→3Gal*p*NAc, Gal*p*NAc-α1→4-Qui*p*3NAc, and Fuc*p*2OAc-α1→4-Gal*p*. The conformational space was sampled by a genetic algorithm-based program called Shape [12], as described in **Chapter 5.a**. Further, relaxed energy maps were calculated based upon high temperature molecular dynamics (MD) simulations (with Tinker [13] using CAT scripts [14]) with a

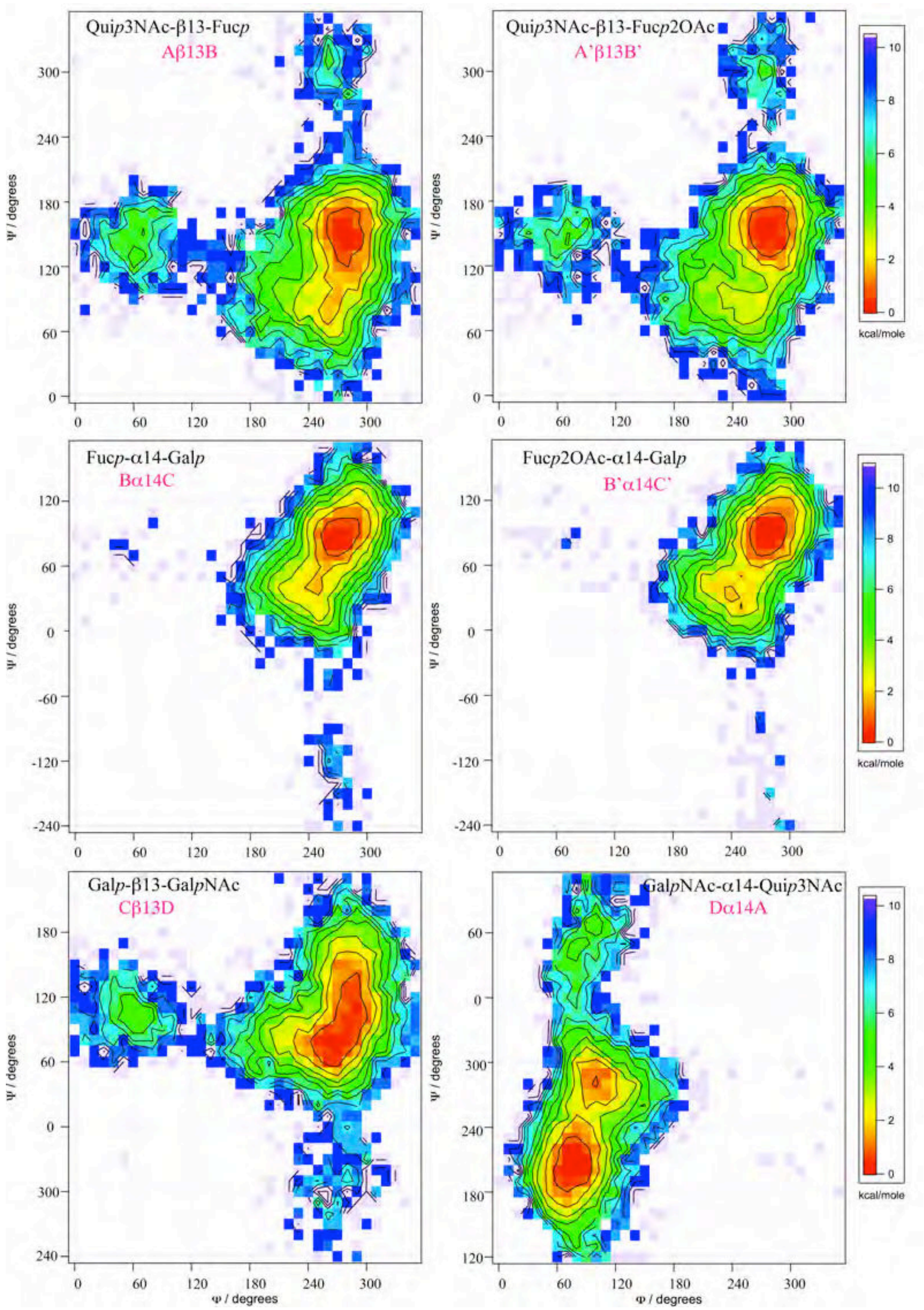


Figure 5.b.2. Relaxed adiabatic maps of the disaccharide components of the molecular models of O1303. The top and middle panels illustrate the glycosidic linkages that are distinct in the two varieties of the biological repeating units, due to the variable substitution on the α-L-Fuc, while the lower panel highlights the glycosidic linkages (Gal*p*-β1→3-Gal*p*NAc and Gal*p*NAc-α14-Qui*p*3ANc) that are common to both the glycans being investigated.

systematic grid search approach using MM3 parameters. The conformational maps generated, are illustrated in **Figure 5.b.2**. From the maps it is evident that each of the disaccharide components can inhabit distinct low energy regions that translate to distinct conformations, thus indicating the flexibility of the molecules.

O-antigens of E. coli 1303

The Φ/Ψ maps of the disaccharides Qui*p*3NAc-β1→3-Fuc*p* and Qui*p*3NAc-β1→3-Fuc*p*2OAc are identical (**Figure 5.b.2**, *top panel*). The O-acetylation at C2 of the α-L-Fuc has no effect on the conformation of the O-antigen. The global minimum is centered about a Φ/Ψ value of 280˚/150˚ in both cases. Two isolated high-energy islands are observed in both samples, one centered at Φ/Ψ ≈ 60˚/150˚ and another at Φ/Ψ ≈ 280˚/ 300˚. For the next disaccharide segment Fuc-α1→4-Gal*p* and Fuc2OAc-α1→4-Gal*p*, again containing the variably O-acetylated α–L-Fuc between the two variants of the biological repeat units of O1303, the energy maps show the sampling of a near identical conformational hypersurface, with the global minimum lying centered at Φ/Ψ ≈ 280˚/90˚ in both instances. In the Gal*p*-β13-Gal*p*NAc linkage, the value for Φ is centered at about 280˚ while the Ψ angle is more flexible, spanning ~80˚ on the energy map within an energy barrier of 7 kcal/mol. The Φ angle value of the Gal*p*NAc-α14-Qui*p*3NAc of this oligosaccharide is much more flexible, covering approximately 40˚ on the energy hypersurface. The Ψ values are expanded between 180˚ to 230˚ for the global minimum and centered about 280˚ for a local minimum. In Gal*p*-β13-Gal*p*NAc, a contiguous pathway of conformational sampling can again be observed similar to that seen in O5ac and O5ab (described in **Chapter 5.a**), while moving from the Φ values of 60° to 300°, following a similar pattern of approaching the isolated energy island through the *trans* conformation. In this disaccharide segment, the energy minima has an oblong shape that can accommodate two populations of conformations, one centered at Φ/Ψ ≈ 260˚/120˚ and the other at Φ/Ψ ≈ 260˚/80˚ Finally, the disaccharide Gal*p*NAc-α14-Qui*p*3NAc carrying the linkage that connects two biological repeat units of O1303, is seen to represent the existence of two distinct minima, one at Φ/Ψ ≈ 80°/210°, corresponding to the global minimum, and Φ/Ψ ≈ 100°/270° indicating a local minima, respectively, separated by an energy barrier of 4 kcal/mol. The shape of the hyper-energy surface of this disaccharide segment is very similar to that traced for the same disaccharide segment in O5ab.

O-antigenic polysaccharide modeling of Ec1303

The relaxed adiabatic maps (**Figure 5.b.2**) were used to guide the construction of the polysaccharide chains using the representative oligosaccharide segments obtained from the conformational search. The polysaccharide of O1303 forms a 2-fold helix with a pitch of 29.85 Å. O5ab has a helix pitch of 29.27 Å. **Figure 5.b.3** illustrates the helix formed by O1303 in comparison to that of O5ab.

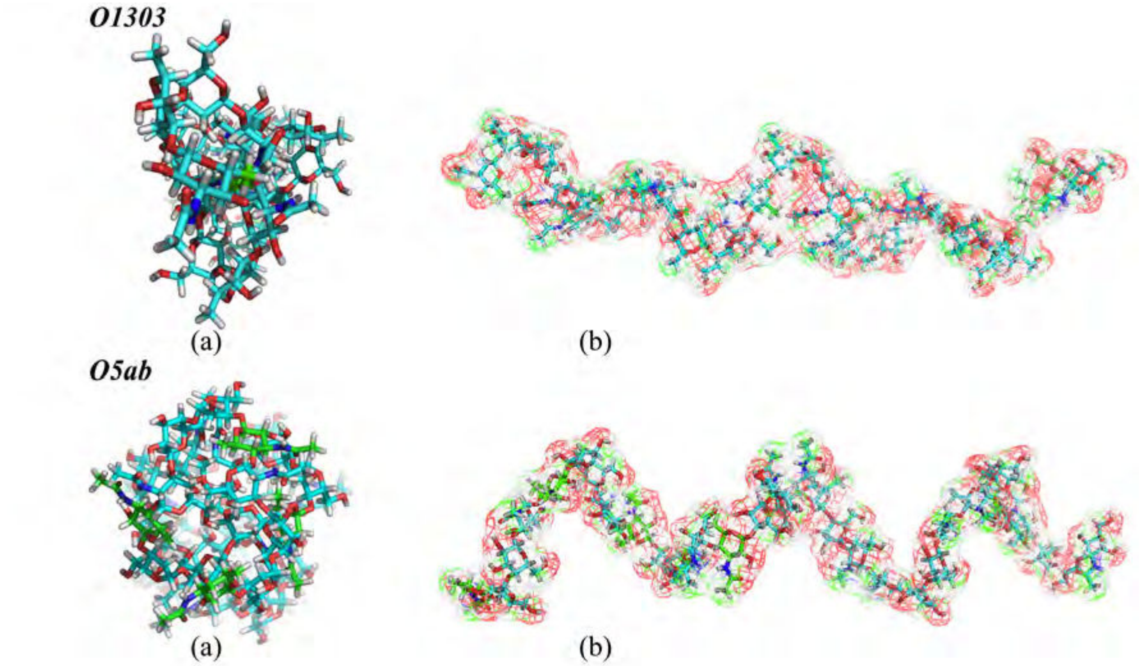


Figure 5.b.3. The schematic representations of the helices formed by the O-antigenic polysaccharides of O1303 and O5ab.

Immunochemical properties of the oligosaccharide fragments

Both intra- and inter-species serological cross-reactions based on the O-antigens are observed frequently between different bacterial isolates. These cross-reactivities are based on structural similarities in the cell wall associated antigens [15].

On the basis of structural and genetic data it has been shown previously that the mastitis isolate *E. coli* 1303 represents a new serotype and possesses the K-12 core type, which is rather uncommon among human and bovine isolates [6]. Western blot analysis of LPS from E. coli 1303 with monoclonal antibodies specific for the different *E. coli* core types showed that strain 1303 carried the K-12 core type in its LPS, which was unexpected and interesting since (i) this *E. coli* core type had been detected in previous studies in only 4% of faecal human and bovine isolates [16], (ii) the widely-used *E. coli* K-12 strains in laboratories produce an R-form LPS lacking O-antigenic polysaccharide repeating units [17], and (iii) an E. coli K-12 strain in which O-antigen assembly was restored exhibited serotype O16 [18].

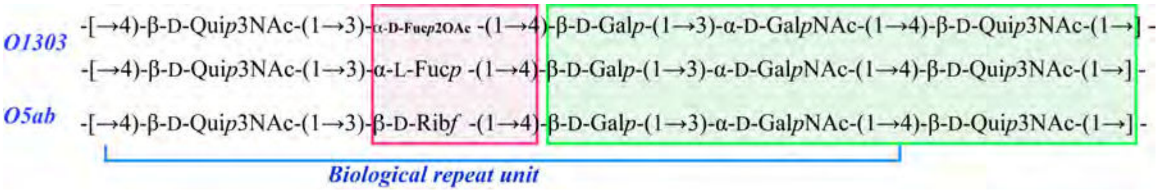

O1303
-[→4)-β-D-Quip3NAc-(1→3)-α-D-Fucp2OAc -(1→4)-β-D-Galp-(1→3)-α-D-GalpNAc-(1→4)-β-D-Quip3NAc-(1→] -
-[→4)-β-D-Quip3NAc-(1→3)-α-L-Fucp -(1→4)-β-D-Galp-(1→3)-α-D-GalpNAc-(1→4)-β-D-Quip3NAc-(1→] -
O5ab
-[→4)-β-D-Quip3NAc-(1→3)-β-D-Ribf -(1→4)-β-D-Galp-(1→3)-α-D-GalpNAc-(1→4)-β-D-Quip3NAc-(1→] -
Biological repeat unit

Figure 5.b.4. Comparison of the biological repeat units of O-antigenic polysaccharides O5ab and the O1303. The green part highlights identical stretch between the sequences, while the pink region marks the difference, which is in the substitution of one monosaccharide unit in the biological repeat.

Conclusion

It was reported earlier that the strains of *E. coli* O5ac and O5ab cannot be distinguished by anti-O5ac or anti-O5ab sera [19], and this strong cross-reactivity is a limitation for differentiation of these two strains using the existing serotyping methodologies. As it was shown before, the structures of the O-antigen PS of O5ac and O5ab are not identical but remarkably similar, differing only at the linkage position between the Gal*p*NAc and Qui*p*3NAc residues (**Chapter 5.a, Figure 5.a.1**). This structural difference seems to go unrecognized by the mono-specific anti-sera, and thus the analysis of the conformational structure of these two polysaccharides may help to explain the cross-reactivity observed between the two strains. The detection of a common glycan epitope between the two mentioned strains may provide leads to further investigations on this front. On comparing the biological repeat units of O1303 and O5ab we observe that they share 100% identity between three monosaccharides and the connecting glycosidic linkages (**Figure 5.b.4**). Two of these contiguous monosaccharides (β-D-Gal*p*-1,3-α-D-Gal*p*NAc) lie within the same biological repeating unit, while the third one (β-D-Qui*p*3NAc) is part of the next biological repeat unit. In the O-antigenic polysaccharide model of Ec1303, we observe that the α-D-Gal*p*NAc residue is located on the inside of the helix, and the access from the outside is hindered by the N-acetyl moiety of Qui*p*3NAc residue. The shapes of the helices of O1303 and O5ab are different, though the solvent accessibility of the glycan residues is similar. This may be indicative of a factor that leads to the detection of cross-reactivities between the two strains of *E. coli*, but further investigations are required to reach a definite conclusion in this regard.

The *E. coli* represents a new sub-type of serotype O5 and contains the rare K-12 core type whose expression may correlate with the ability of this strain to cause bovine mastitis.

Perspectives

The conformational features of bacterial polysaccharides are often discussed in terms of the size and shape of antigenic determinants (or *epitopes*). An epitope is the part of an antigen that is recognized by the immune system, specifically by antigen-specific membrane receptors on lymphocytes, secreted antibodies, B cells, or T cells. The part of an antibody that recognizes the epitope is called a *paratope*. The epitopes of protein antigens are divided into two categories, conformational epitopes and linear epitopes, based on their structure and interaction with the paratope [20]. A conformational epitope is composed of discontinuous sections of the antigen's sequence. These epitopes interact with the paratope based on the 3D surface features and shape or tertiary structure of the antigen. Most epitopes bind to paratopes based upon conformational characteristics.

The O-antigenic capsular polysaccharides have been found to be critical in protective immunity, as in the case of O139 Bengal strain of *Vibrio cholerae* that causes epidemic cholera. The specific interactions of the glycan epitope(s) at the antibody binding sites are responsible for the immunological behavior of these bacterial polysaccharides, which are of interest with respect to vaccine development. The glycan epitopes (usually comprising two to four sugar residues) may assume a compact and relatively rigid conformation that effectively interact with the paratopes of the antibody, or alternatively, be flexible, thus allowing antibodies to bind to select conformations, consistently being constituents of a rather flexible polysaccharide [21].

It can be interesting to probe the conformations of O-antigenic polysaccharides of *E. coli* strains (for example Ec65 that has a genetic similarity with Ec1303), to understand the basis of their cross-reactivity and subsequent pathogenesis.

References:

1. Bar D, Tauer LW, Bennett G, González RN, Hertl JA, Schukken YH, Schulte HF, Welcome FL, Gröhn YT: **The cost of generic clinical mastitis in dairy cows as estimated by using dynamic programming**. *Journal of dairy science* 2008, **91**(6):2205-2214.

2. Seegers H, Fourichon C, Beaudeau Fo: **Production effects related to mastitis and mastitis economics in dairy cattle herds**. *Vet Res* 2003, **34**(5):475-491.

3. Burvenich C, Merris VrV, Mehrzad J, Diez-Fraile A, Duchateau L: **Severity of *E. coli* mastitis is mainly determined by cow factors**. *Vet Res* 2003, **34**(5):521-564.

4. Messom GV-V, Burvenich C, Roets E, Massart-Leën A-M, Heyneman R, Kremer WDJ, Brand A: **Classification of newly calved cows into moderate and severe responders to experimentally induced *Escherichia coli* mastitis**. *Journal of Dairy Research* 1993, **60**(01):19-29.

5. Hill AW: ***Escherichia coli* mastitis**. Wallingford, UK: CAB international; 1994.

6. Duda KA, Lindner B, Brade H, Leimbach A, Brzuszkiewicz Eb, Dobrindt U, Holst O: **The lipopolysaccharide of the mastitis isolate *Escherichia coli* strain 1303 comprises a novel O-antigen and the rare K-12 core type**. *Microbiology* 2011, **157**(6):1750-1760.

7. Kojima H, Inagaki M, Tomita T, Watanabe T: **Diversity of non-stoichiometric substitutions on the lipopolysaccharide of *E. coli* C demonstrated by electrospray ionization single quadrupole mass spectrometry**. *Rapid Communications in Mass Spectrometry* 2010, **24**(1):43-48.

8. **Glyco3D: A site for glycosciences** [http://glyco3d.cermav.cnrs.fr/glyco3d]

9. TRIPOS: **SYBYL-X 1.3, Tripos, Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA**. In.; 1991 - 2011.

10. Imberty A, Bettler E, Karababa M, Mazeau K, Petrova P, Pérez S: **Building sugars: The sweet part of structural biology.** . In: *Perspectives in Structural Biology"*. Edited by M. Vijayan NYASK. Hyderabad: Indian Academy of Sciences and Universities Press, Hyderabad; 1999: 392-409.

11. Clark M, Cramer RD, Van Opdenbosch N: **Validation of the general purpose tripos 5.2 force field**. *Journal of Computational Chemistry* 1989, **10**(8):982-1012.

12. Rosen J, Miguet L, Perez S: **Shape: automatic conformation prediction of carbohydrates using a genetic algorithm**. *J Cheminform* 2009, **1**(1):16.

13. **TINKER Molecular Modeling** [dasher.wustl.edu/tinker/]

14. **Conformational Analysis Tool (CAT)** [http://www.md-simulations.de/CAT/]

15. Nordmark E-L: **Structural and interaction studies of bacterial polysaccharides by NMR spectroscopy**. *Doctoral dissertation.* Stockholm: Stockholm University; 2004.

16. Gibbs RJ, Stewart J, Poxton IR: **The distribution of, and antibody response to, the core lipopolysaccharide region of *Escherichia coli* isolated from the faeces of healthy humans and cattle**. *Journal of Medical Microbiology* 2004, **53**(10):959-964.

17. Feldman MF, Marolda CL, Monteiro MA, Perry MB, Parodi AJ, Valvano MA: **The activity of a putative polyisoprenol-linked sugar translocase (Wzx) involved in *Escherichia coli* O antigen assembly is independent of the chemical structure of the O repeat**. *Journal of Biological Chemistry* 1999, **274**(49):35129-35138.

18. Liu D, Reeves PR: ***Escherichia coli* K12 regains its O antigen**. *Microbiology* 1994, **140**(1):49-57.

19. Urbina F, Nordmark E-L, Yang Z, Weintraub A, Scheutz F, Widmalm G: **Structural elucidation of the O-antigenic polysaccharide from the enteroaggregative *Escherichia coli* strain 180/C3 and its immunochemical relationship with *E-coli* O5 and O65**. *Carbohydr Res* 2005, **340**(4):645-650.

20. Huang J, Honda W: **CED: a conformational epitope database**. *BMC Immunology* 2006, **7**(1):7.

21. Gunawardena S, Fiore CR, Johnson JA, Bush CA: **Conformation of a rigid tetrasaccharide epitope in the capsular polysaccharide of *Vibrio cholerae* O139**. *Biochemistry* 1999, **38**(37):12062-12071.

Section B

Chapter 6

CHAPTER 6

BiOligo: A 3D structural database of bioactive oligosaccharides

[1]

Abstract

The relationship between the structure and biosynthesis of glycans to their actual functions is the driving force behind the systematic exploration of their biological roles. The knowledge of the three-dimensional (3D) structures of biologically occurring oligosaccharides is needed for the understanding of biological processes involving glycoproteins and protein-glycan interactions at the molecular level. To this end, about 250 glycan determinants (bioactive oligosaccharides) have been listed and subject to systematic conformational sampling to determine their conformational preferences. They belong to widely-occurring families like the blood group antigens (A, B & O), core structures (Types 1, 2 & 4), fucosylated oligosaccharides (core & lacto-series), sialylated oligosaccharides (Types 1 & 2), Lewis antigens, GPI-anchors, N-linked oligosaccharides, globosides, among others, and have been systematically organized into an open-source database. The constituent disaccharide and monosaccharide units (currently, ~120 and 70 entries, respectively) of these bioactive oligosaccharides have also been characterized and made available through a sub-set within the BiOligo database (called GlycoLego). At present, BiOligo contains more than 400 entries of glycan determinants and their '*lego*' blocks. The BiOligo data is accessible through several search criteria like oligosaccharide name, type of constituent (monosaccharide, disaccharide or oligosaccharide), glycan category and molecular weight. All 3D structures are available for visual consultation (with basic measurement possibilities) and can be downloaded in commonly used formats. BiOligo aims at offering 3D information to help in deciphering binding data from glycan arrays, and providing realistic starting conformations to be used in molecular dynamics (MD) simulations, molecular docking of oligosaccharides with proteins or nucleic acids and improving the resolution of the structures of glycoproteins, in particular

[1] Anita Sarkar, Alain Rivet and Serge Pérez*. AS designed the method, modeled the oligosaccharides, developed the MySQL database and related PHP scripts and wrote the manuscript.
Centre de Recherches sur les Macromolécules Végétales, CNRS, BP53, 38041 Grenoble, France

with small angles X-ray scattering (SAXs) experiments. BiOligo is an open source database of bioactive glycan determinants and can be accessed via its web-interface at http://glyco3D/bioligo.cermav.cnrs.fr .

Introduction

Carbohydrates are involved in a variety of biological functions ranging from the trivial to the crucial. They play important roles in the growth, function, development and even survival of an organism. These ubiquitous and complex molecules offer a tremendous diversity which arises not only from differences in monosaccharide composition, anomeric state (α or β) and branching of monosaccharides, but also from substituted (e.g., sulphated, phosphated) components and their linkage to the aglycones (like peptides, lipids). This makes glycans 'bio-informational' molecules which are recognized by a variety of proteins, such as lectins, antibodies, receptors, toxins, microbial adhesins, enzymes, etc.), collectively known as glycan binding proteins. The number of glycans that comprise the human genome is still unknown, and it may be acknowledged that sequencing the human glycome at present may still be unrealistic given the current technology for glyco-proteomics, glycosaminoglycan-omics or glyco-lipidomics. According to R. Cummings [1] the number of glycan determinants likely to be important in their interactions with glycan binding proteins is estimated to be about 7,000. The Consortium for Functional Glycomics [2] lists 7500 entries in its glycan database.

Complex carbohydrates are in general difficult to co-crystallize, and there is still a limited number of X-ray crystallographic protein-carbohydrates complex structures that have been resolved. Hence, molecular modeling methods have been particularly helpful and widely used to characterize the conformation of complex carbohydrates [3, 4]. The flexibility and high polarity of these molecules, along with the large number of degrees of freedom are challenges for molecular modeling. Nevertheless, development of sustainable methods and tools accompanied by major improvements in computing performances, are opening the way to high-throughput molecular modeling where hundreds of complex glycan structures can be investigated during the course of time-limited investigations [5].

In contrast to genomics and proteomics, the glyosciences lack accessible, curated and annotated comprehensive data repositories that summarize and organize data such as structures, characteristics, biological origins and functions of glycans. The thrust of the glycoinformatics community thus far has mainly been carbohydrate sequence related. Web-based tools have been developed to build preliminary 3D structures starting from a sequence as implemented in SWEET-II and Glycam carbohydrate builder [6, 7].

The purpose of the present work is to provide to the scientific community a useful database called BiOligo, which stores energetically favourable and optimized 3D glycan structures that can be used as reliable starting models for interaction studies. Particular attention has been given to the database content and its open access. Further, the possible linking of the 3D information with glycan array results and other 3D databases, dealing with protein-carbohydrate interactions, is discussed in the perspectives.

Scope of the work

A comprehensive list of prominent glycan determinants involved in recognition events was established taking into account the availability of these glycans in ample quantities to fully explore the basis for glycan recognition and specificity using X-ray crystallography, nuclear magnetic resonance (NMR), and other biophysical methods as well as glycan arrays. About 250 glycan determinants that confirmed to the mentioned criteria were selected. These include representations from widely-detected families like the blood group antigens (A, B and O), core structures (Types 1, 2 & 4), fucosylated oligosaccharides (core and lacto-series), sialylated oligosaccharides (Type 1 and 2), Lewis antigens, GPI anchors, N-linked oligosaccharides, globosides, glycosaminoglycans (GAGs), etc. as listed in **Table 6.1**. In the present work, a glycan determinant is defined as a glycan structure that is required for the specific recognition by a glycan binding protein. The glycans containing three or more sugars in the current list are referred to as bioactive oligosaccharides.

A complete list of the oligosaccharides included in the present in this database version is provided in Annex III: *Supplementary Material* for BiOligo.

Specific glycan determinants result from the assembly of partial motifs i.e. disaccharide moieties. For reasons related to the construction and to the conformational analysis of complex glycan (*vide supra*), a list of partial determinants was established (**Table 6.2**). It is composed of ~120 distinct disaccharide moieties found in the glycans categorized as per **Table 6.1**.

Table 6.1. The classification of 3D structures of glycan determinants in BiOligo.

Index	BiOligo Category	No. Of entries
1.	Blood group A antigens	11
2.	Blood group B antigens	11
3.	Blood group H antigens (Blood group O)	12
4.	Blood group H antigens (Blood group O) and Globo H tetraose	1
5.	Core structures	1
6.	Core structures (Type 1 & Type 2)	4
7.	Core structures (Type 1)	4
8.	Core structures (Type 2)	16
9.	Core structures (Type 4)	1
10.	Fucosylated oligosaccharides	4
11.	Fucosylated oligosaccharides (3 Fucosyllactose core)	4
12.	Fucosylated oligosaccharides (Lacto-Series)	13
13.	GAGs	14
14.	Galα-3Gal oligosaccharides (Galili and xeno antigens)	6
15.	Galα-3Gal oligosaccharides (Isogloboseries)	3
16.	Ganglioside sugars	17
17.	Globoside sugars (P antigens) (Forssman antigens)	3
18.	Globoside sugars (P antigens) (Globo series - core structure type 4)	3
19.	Globoside sugars (P antigens) (P blood group antigens and analogues)	6
20.	Globoside sugars (P antigens) (Stage-specific Embryonic antigens : SSEA-3 & SSEA-4)	4
21.	Glucuronylated oligosaccharides	2
22.	Glycosphingolipid	2
23.	Lewis antigens	29
24.	Miscellaneous	22
25.	Miscellaneous (Blood group-related oligosaccharides)	2
26.	Miscellaneous (Chitin oligosaccharides)	4
27.	Miscellaneous (Fibriniogen related oligosaccharides)	3
28.	Miscellaneous (LDN-related oligosaccharides)	6
29.	Miscellaneous (Lewis X-related oligosaccharides)	2
30.	Miscellaneous (TF-related oligosaccharides)	4
31.	Miscellaneous (TN-related oligosaccharides)	4
32.	Miscellaneous (Trehalose-like sugars)	2
33.	N-linked oligos	18
34.	Sialylated oligosaccharide (Type 1)	11
35.	Sialylated oligosaccharide (Type 2)	12
36.	Disaccharides (***GlycoLego***)	124
37.	Monosaccharides (***GlycoLego***)	70

High-Throughput Conformational Analysis of Glycan Determinants

Nomenclature

The variety in nomenclature and structural representations of glycans makes it complex to decide the most appropriate form to deal, both with the encoding required for computational manipulation and graphical representations, which are relevant from a chemical as well a biological standpoint. The structural encoding of glycans dealt with during the development and implementation of the work are illustrated in **Figure 6.1**.

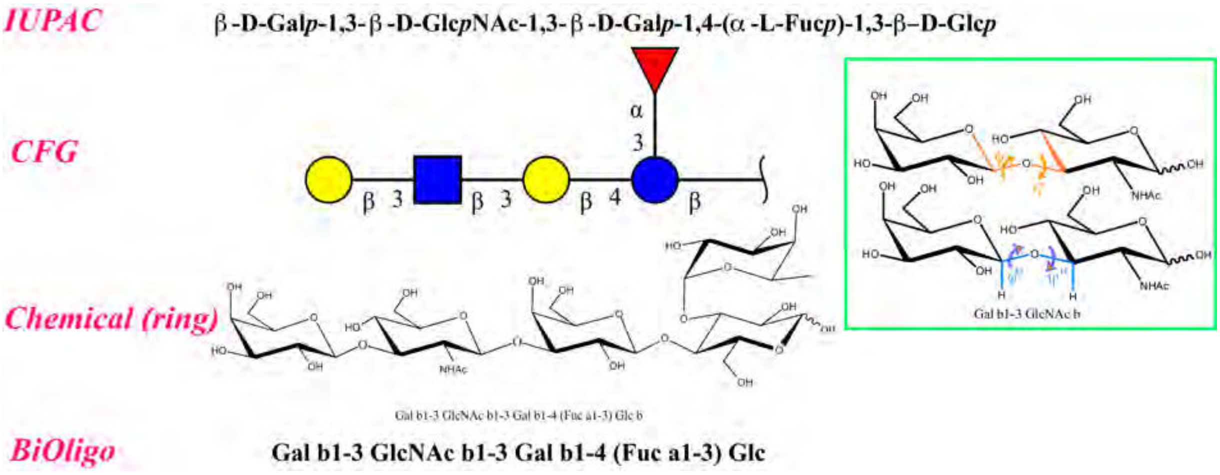

IUPAC
β-D-Galp-1,3-β-D-GlcpNAc-1,3-β-D-Galp-1,4-(α-L-Fucp)-1,3-β-D-Glcp
CFG
α
3
β 3
β 3
β 4
β
Chemical (ring)
NHAc
Gal b1-3 GlcNAc b1-3 Gal b1-4 (Fuc a1-3) Glc b
Gal b1-3 GlcNAc b
BiOligo
Gal b1-3 GlcNAc b1-3 Gal b1-4 (Fuc a1-3) Glc

Figure 6.1. Nomenclature and structural representations commonly used for complex glycans, for example, the pentasaccharide lacto-N-fucopentaose V [Gal β1-3 GlcNAc β1-3 Gal β1-4 (Fuc) α1-3 Glc] in this figure. The relative orientation of two contiguous monosaccharide units in a disaccharide is expressed by two torsion angles Φ and Ψ around the glycosidic bond. According to the *heavy atom convention* (x+1), Φ= O5-C1-O-C_x and Ψ= C1-O-C_x-C_{x+1} for a $(1\rightarrow x)$ linkage [*inset: top panel*]. Alternatively, reference to the hydrogen atoms involved in the glycosidic linkage as per the *light atom convention*, can be used Φ^H = H1-C1-O1-C_x and Ψ^H = C1-O1-C_x-H_x, for a $(1\rightarrow x)$ linkage. For a $(1\rightarrow 6)$ linkage another torsion angle is required and denoted by ω, referring to O5-C5-C6-O6. The sign of the torsion angle is given in accordance with the IUPAC nomenclature [8].

Glycan Determinants

The common computational approach to 3D structure prediction is based on searching through the conformational space of the glycan in order to find low energy regions, i.e. conformers, which the molecule is likely to populate. This can be accomplished in many different ways, and several packages have been developed for this task using a variety of different algorithms [9]. Despite their inherent structural intricacies, complex carbohydrates are particularly suited for computational conformational predictions. Of the usually 20 atoms of a hexa-pyranose unit bound within an oligosaccharide, 80% are rigidly linked together; six locked in the pyranose ring and further ten rigidly attached to the five ring carbon atoms. This reduces the motion of the ring structure. The only bonds that are of significance for the mutual orientation of two contiguous linked monosaccharide units are the torsion angles at the glycosidic linkages (Φ, Ψ and ω in the case of 1→6 linkage).

The 3D structures of complex glycans can be constructed by partitioning them into overlapping disaccharides, i.e. predicting each glycosidic linkage in the isolated disaccharide and then reassembling the complete structure. The determination of the preferred conformations of a disaccharide moiety is based on the calculation of potential energy surfaces as a function of their glycosidic torsional angles. This method provides a clear depiction of the results and an exhaustive exploration of the topology of the potential energy surface; it has a linear algorithmic complexity and is one of the fastest prediction methods available. However, it only works with oligosaccharides that have no interactions between non-adjacent residues and it requires the computation of all the potential energy surfaces of the constituting disaccharide segments.

High-throughput conformational analysis of complex glycan requires the development of methods where both speed and thoroughness are vital. These factors need to be well balanced and preferably adjustable according to user priority. It is important to perform a comprehensive general search of the conformational space to find all the important energy minima, while at the same time use as little computation time as possible. To this end a dedicated software (Shape) for automatic conformation prediction of carbohydrates using a genetic algorithm was developed [5]. Its robustness and accuracy have been tested on a series of studies on previously published conformation predictions of oligosaccharides performed using other conformation search tools. In these cases all major local minima could be found with a major improvement in computational time.

The glycan determinants (currently, 260 of these bioactive oligosaccharides have been evaluated) were submitted to the above mentioned automatic conformation prediction following the procedure described in the *Computational Methods*, which, including the mono- and disaccharide conformers, yielded a grand total of about 1200 conformers. As a typical example, the results of the exploration of the potential energy hypersurface of lacto N-fucopentaose are shown in **Figure 6.2**.

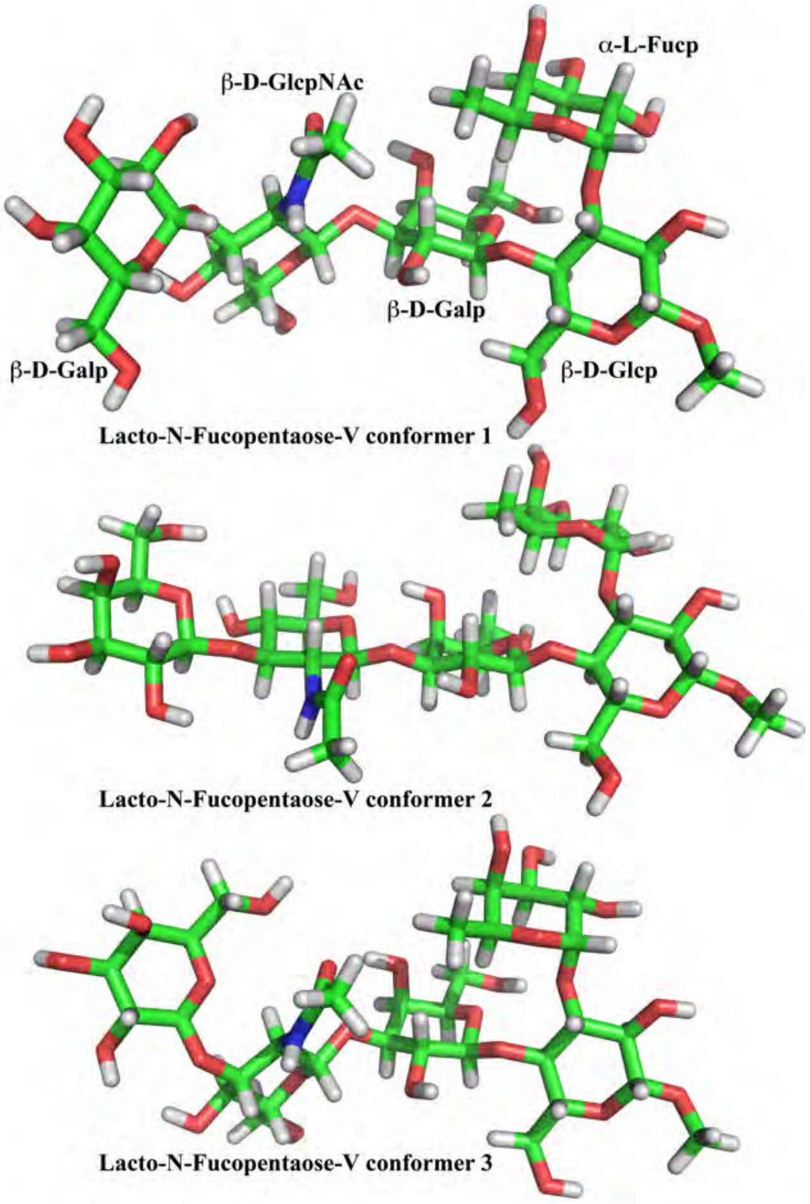

α-L-Fucp
β-D-GlcpNAc
β-D-Galp
β-D-Galp
β-D-Glcp
Lacto-N-Fucopentaose-V conformer 1
Lacto-N-Fucopentaose-V conformer 2
Lacto-N-Fucopentaose-V conformer 3

Figure 6.2. The distinct conformations reported in BiOligo after a complete conformational sampling of the lacto-N-fucopentaose V structure.

Disaccharide segments

For each disaccharide, an exhaustive search was performed using the MM3 molecular mechanics force field. It gave a complete sampling of the conformational space, yielding the construction of a relaxed adiabatic energy map, which is represented as a function of Φ and Ψ glycosidic torsion angles.

Table 6.2. The list of disaccharides included as part of the sub-database GlycoLego incorporated within BiOligo.

Fuc α1-2 Gal
Fuc α1-2 Glc
Fuc α1-3 Gal
Fuc α1-3 Glc
Fuc α1-3 GlcNAc
Fuc α1-4 Gal
Fuc α1-4 GlcNAc
Fuc α1-6 GlcNAc (*gg*)
Fuc α1-6 GlcNAc (*gt*)
Gal [3S] β1-3 GalNAc
Gal [3S] β1-4 GalNAc
Gal [3S] β1-4 Glc
Gal [3S] β1-4 Glc [6S]
Gal [3S] β1-4 GlcNAc
Gal [3S4S] β1-4 GlcNAc
Gal [3S6S] β1-4 GlcNAc [6S]
Gal [4S] β1-4 GlcNAc
Gal [4S6S] β1-4 GlcNAc
Gal [6S] β1-4 Glc
Gal [6S] β1-4 Glc [6S]
Gal [6S] β1-4 GlcNAc
Gal [6S] β1-4 GlcNAc [3S]
Gal α1-2 Gal
Gal α1-3 Gal
Gal α1-3 GalNAc
Gal α1-4 Gal
Gal α1-4 GlcNAc
Gal α1-6 Gal (*gt*)
Gal α1-6 Gal (*tg*)
Gal α1-6 Glc (*gg*)
Gal α1-6 Glc (*gt*)
Gal α1-6 Glc (*gg*)
Gal β1-4 GlcNAc [6S]
Gal β1-1 Glc
Gal β1-2 Gal
Gal β1-2 Xyl
Gal β1-3 AltNAc
Gal β1-3 Gal
Gal β1-3 GalNAc
Gal β1-3 GlcNAc
Gal β1-4 Gal
Gal β1-4 GalNAc
Gal β1-4 Glc
Gal β1-4 Glc [3S]
Gal β1-4 GlcNAc
Gal β1-4 GlcNAc [3S6S]
Gal β1-4 GlcNAc [6S]
Gal β1-6 Glc (*gg*)
Gal β1-6 Glc (*gt*)
GalNAc [4S] β1-4 GlcA
GalNAc [4S] β1-4 L-Ido
GalNAc [4S] β1-4 L-Ido [2S]
GalNAc [6S] β1-4 GlcA
GalNAc α1-3 Fuc
GalNAc α1-3 Gal
GalNAc α1-3 GalNAc
GalNAc α1-3 Man
GalNAc β1-3 Gal
GalNAc β1-4 Gal
GalNAc β1-4 GalA
GalNAc β1-4 GlcA
Glc α1-4 Glc
Glc β1-3 Glc
Glc β1-4 Glc
GlcA [4S] β1-3 GalNAc
GlcA [6S] β1-3 GalNAc
GlcA β1-3 GalNAc [4S]
GlcA β1-3 GalNAc [6S]
GlcA β1-3 GalNAc
GlcA β1-3 GlcNAc
GlcNAc [6S] β1-3 Gal
GlcNAc α1-6 GalNAc (*gg*)
GlcNAc α1-6 GalNAc (*gt*)
GlcNAc α1-6 GalNAc (*tg*)
GlcNAc β1- 3 GalNAc
GlcNAc b1-2 Gal
GlcNAc β1-2 Man
GlcNAc β1-3 Gal
GlcNAc β1-3 GalNAc
GlcNAc β1-3 GlcNAc
GlcNAc β1-4 Gal
GlcNAc β1-4 GlcA
GlcNAc β1-4 GlcNAc
GlcNAc β1-4 Man
GlcNAc β1-6 Gal
GlcNAc β1-6 Gal
GlcNAc β1-6 Gal
GlcNAc β1-6 GalNAc (*gg*)
GlcNAc β1-6 GalNAc (*gt*)
GlcNAc β1-6 GalNAc (*tg*)
GlcNAc β1-6 Man (*gg*)
GlcNAc β1-6 Man (*gt*)
GlcNS α1-4 IdoA [2S] $(^2S_0)$
GlcNS α1-4 IdoA [2S] $(^1C_4)$
GlcNS [6S] α1-4 IdoA [2S] $(^2S_0)$
GlcNS [6S] α1-4 IdoA [2S] $(^2S_0)$
IdoA [2S] $(^1C_4)$ α1-4 GlcNAc
IdoA [2S] $(^2S_0)$ α1-4 GlcNAc
IdoA [2S] $(^1C_4)$ α1-4 GlcNAc
IdoA [2S] $(^2S_0)$ α1-4 GlcNAc
IdoA [2S] $(^1C_4)$ α1-4 GlcNS [6S]
IdoA [2S] $(^2S0)$ α1-4 GlcNS [6S]
IdoA [2S] $(^1C_4)$ α1-4 GlcNS
IdoA [2S] $(^2S_0)$ α1-4 GlcNS
L-Ido [2S, 6S] α1-4 GlcNS
L-Ido a1-3 GalNAc [4S]
Man [6P] α1-3 Man
Man α1-2 Man
Man α1-3 GlcNAc
Man α1-3 Man
Man α1-4 GlcNAc
Man α1-4 Man
Man α1-6 Man (*gg*)
Man α1-6 Man (*gt*)
Man β1-4 GlcNAc
Man β1-4 Man
Neu5Ac α2-3 Gal
Neu5Ac α2-3 GalNac
Neu5Ac α2-6 Gal
Neu5Ac α2-6 GalNAc
Neu5Ac α2-6 GlcNAc
Neu5Ac α2-8 Neu5Ac
Xyl β1-2 Man
Xyl β1-3 Man

Typically the exploration of each MM3 energy map indicated the occurrence of 3 to 5 energy minima. They were ranked based on relative energies and the corresponding atomic coordinates were stored. A library of about 500 conformers was set-up corresponding to the population occurrence and statistics of all the disaccharide segments occurring in the glycan determinants under investigation.

This library is a unique collection of structural features of very diverse disaccharides that can be used to build, modify or extend 3D glycan structures. The consideration of the axial/equatorial nature at the glycosidic linkage provides a useful framework for a classification of the disaccharides moieties, independently of the surroundings of the glycan molecule.

Construction & Content

Database content

BiOligo database, in its current version, contains about 400 entries, including monosaccharides, disaccharides and oligosaccharides. The selected carbohydrates are known to be biologically active and come from in-house databases, glycan array from the '*Glycan Database'* of the Consortium for Functional Glycomics [10] as well as catalogued complex sugar libraries [11]. The source and constitution of the database include neutral and sialylated oligosaccharides from human milk and urine, cell adhesion oligosaccharides, blood groups, head groups of common glycosphingolipids, lectin-binding oligosaccharides and glycosaminoglycans.

Database construction

BiOligo is a web-based, platform-independent, manually curated database of bioactive glycan 3D structures. It currently runs on an Apache web server [12] hosted at Centre de Recherches sur les Macromolécules Végétales (CERMAV) with the application program Hypertext Preprocessor (PHP) [13]. It has been implemented using the open source MySQL database [14]. It has been developed based on a combination of three layers. The underlying layer is the MySQL database system, a relational database management system [MySQL (Community Server) with the storage-engine PBXT] that stores all the other structure-related information in the back-end and provides the facility to link two or more tables in the database. The intermediate layer is an Apache-PHP application [Apache 2.x; PHP 5.3.1] that receives the query from the user and connects to the database to fetch data to the upper layer, which comprises populated HTML pages, to the web browser client. The PHP and Java scripts are embedded in the HTML web pages for this effect and are used as application programs for integrating the back-end (MySQL database) to the web pages (HTML). Apache has been used as the web server for building the interface between the web browser and the application programs. PHP was used for writing scripts to query the database, and JavaScript (with JQuery plugin) was used to design the '*auto-complete*' function for the user-interface. The graphical user interface was developed with HTML (version 5) and CSS (version 3).

Database query and results

The schema of data organization and output is illustrated in **Figure 6.3**.

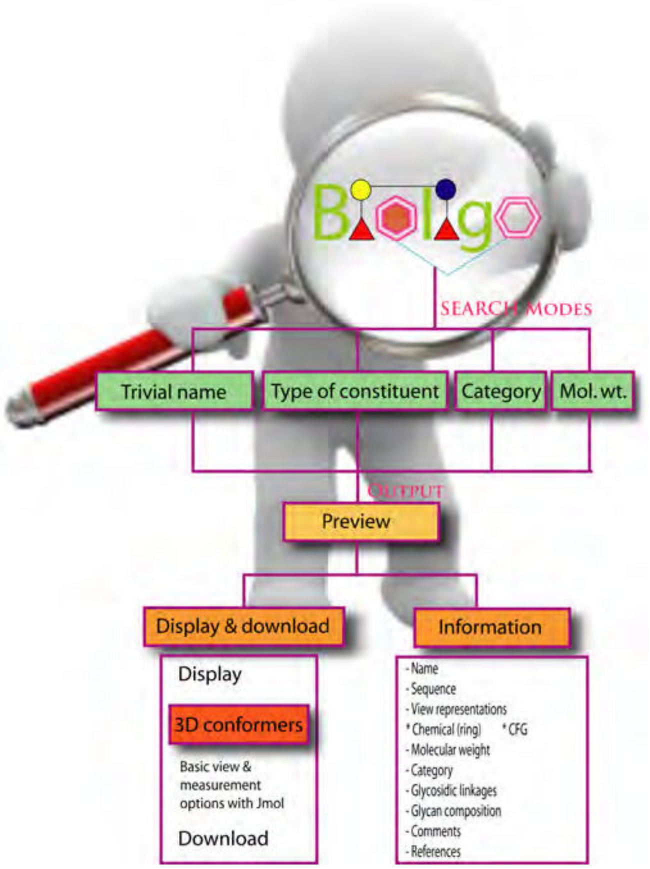
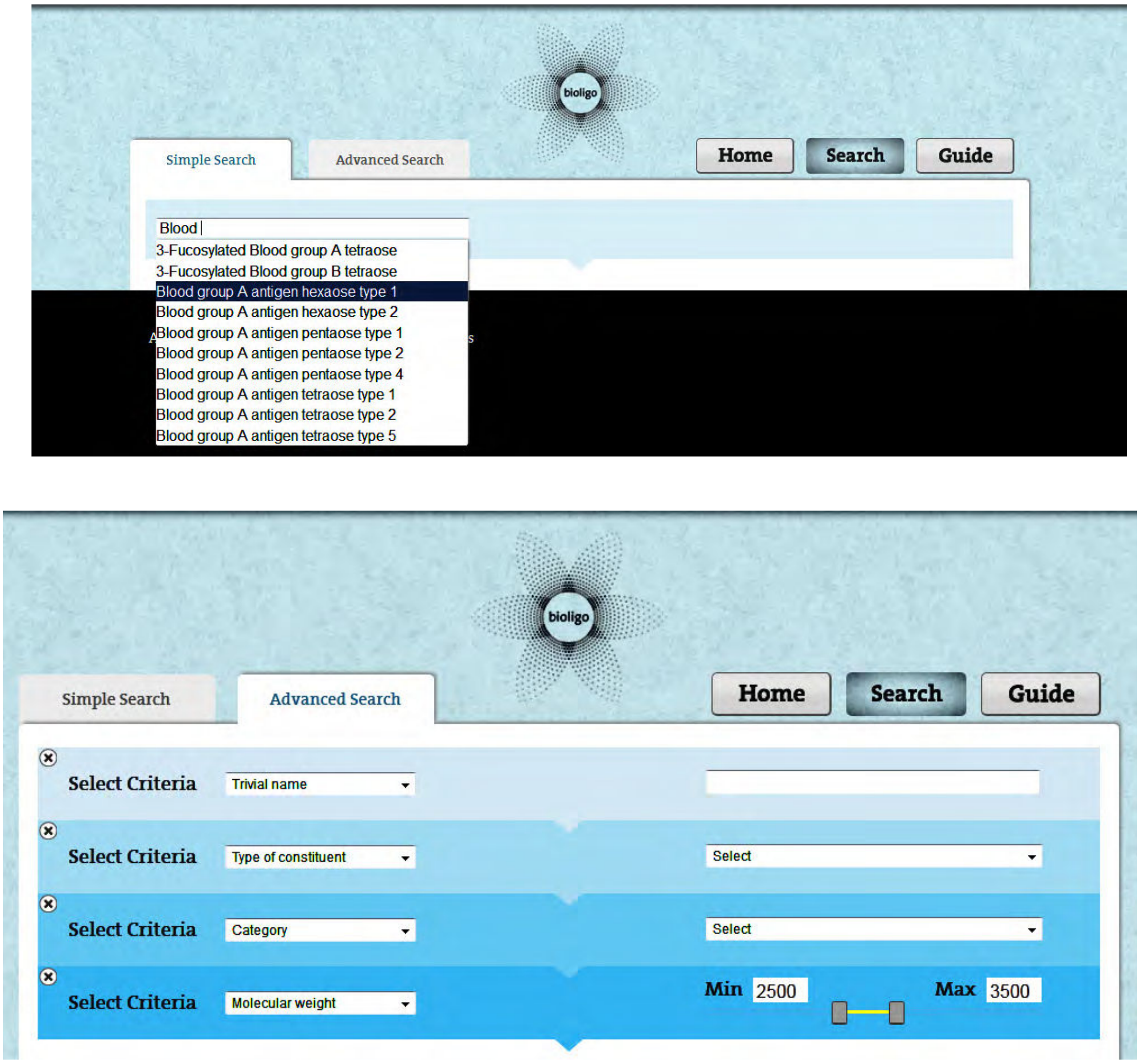


Figure 6.3. A schema showing the various search modes accessible to the user and results displayed for a query made to the BiOligo database.

Database Search and GUI features

The search page comprises primarily of two approaches to query the database, as shown in **Figure 6.4**, namely,

a. ***Simple search***: In this the user types in the text box provided, based upon which a result prompt appears to guide the user in selecting from the 'hits' found in the database. An accordion function was developed to display a preview of the results. This can be used to expand or minimize the preview of the listed results of the user query for a first glance into the entries matching the request to the database. The preview provides the glycan name, sequence, category and molecular weight to the user to make an informed choice.


bioligo
Simple Search
Advanced Search
Home
Search
Guide
Blood
3-Fucosylated Blood group A tetraose
3-Fucosylated Blood group B tetraose
Blood group A antigen hexaose type 1
Blood group A antigen hexaose type 2
Blood group A antigen pentaose type 1
Blood group A antigen pentaose type 2
Blood group A antigen pentaose type 4
Blood group A antigen tetraose type 1
Blood group A antigen tetraose type 2
Blood group A antigen tetraose type 5
bioligo
Simple Search
Advanced Search
Home
Search
Guide

Select Criteria: Trivial name

Select Criteria: Type of constituent — Select

Select Criteria: Category — Select

Select Criteria: Molecular weight — Min 2500 Max 3500

Figure 6.4: An illustration of the simple search (*Top*) and advanced search (*Bottom*) search options in BiOligo.

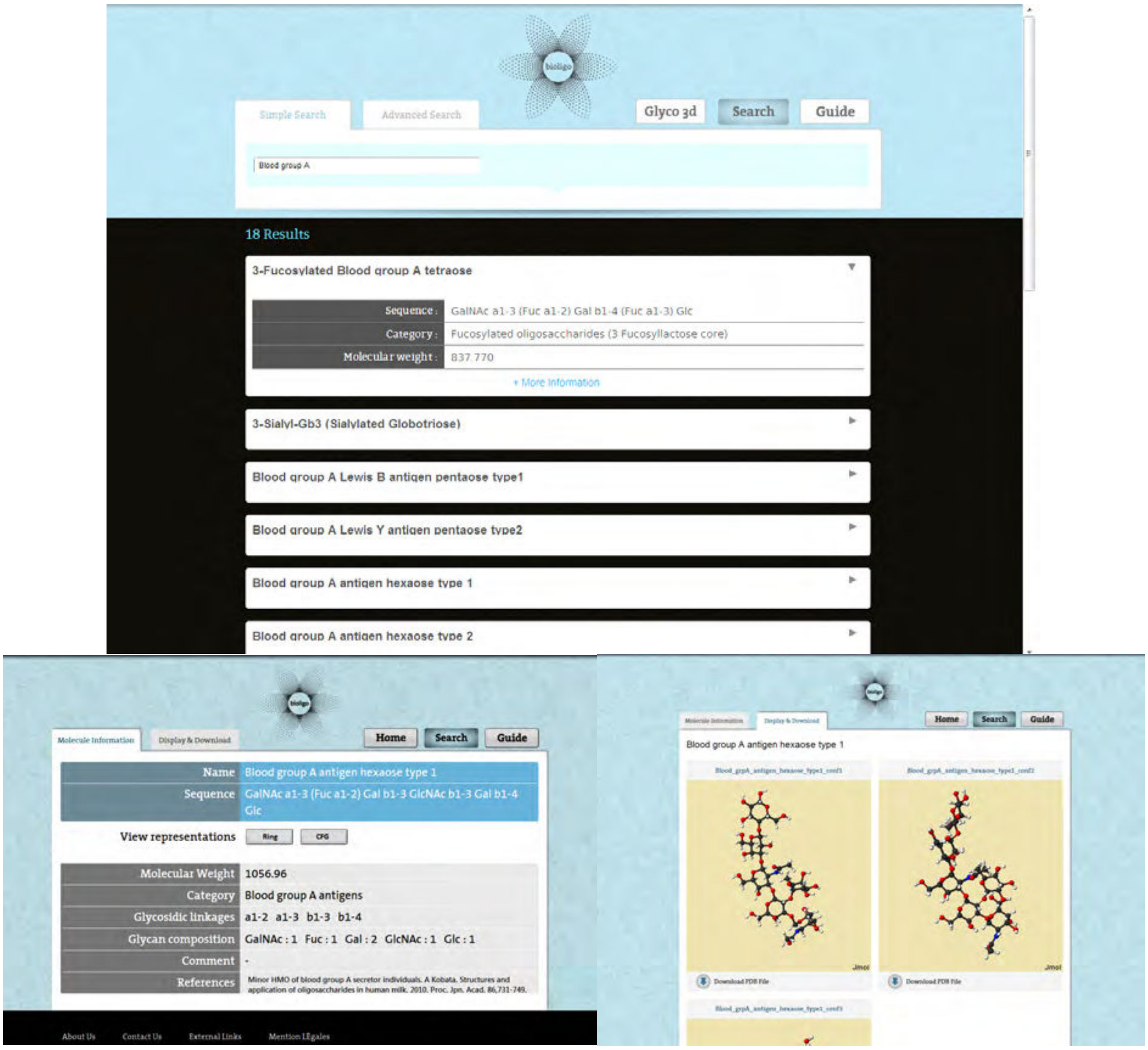
a. ***Advanced search***: This is a multi-criteria search that can be used together for querying or in various combinations as best suits the user's requirement. Four search modes are provided in BiOligo, namely, trivial name, type of constituent, category and molecular weight. A slider is provided for assigning a range of values to be queried in the molecular weight of the database entries. It consists of two cursors that can navigate on a bar for specifying the minimum and maximum limit of the search. Two text fields display the values of the cursor position on the slider bar. The slider cursors auto-adjust themselves when values are entered directly in the text boxes. This feature was developed by modifying a JQuery plug-in.

Both the simple and advanced search options are equipped with an '*auto-complete*' function. This is one of the prime features of result refinement provided in the GUI. It guides the user while querying the database. It comprises two parts (a) single field of entered text, and (b) the auto-prompt when the data is entered, through which the desired hit in the database can be selected either by scrolling down with the mouse or using the arrow keys on the keyboard.

The detailed results are organized under two tabs as shown in **Figure 6.5**, namely,

a. ***Molecule information***: This includes the trivial name of the glycan, its sequence, the chemical (ring) and CFG (Consortium of functional glycomics) cartoon representation, molecular weight, the glycan category or family into which it has been classified in BiOligo, glycan composition, i.e. the comprising glycan type and number of each such glycan in the BiOligo entry, glycosidic linkages present in it and occasionally additional comments. Each entry is associated with a reference that identifies it as a glycan determinant, and from which it has been sourced into BiOligo. The illustrative representations of the glycans are can be viewed through the '*Zoombox*' feature that was developed by modifying an existing JQuery plug-in that allows the selected image to be zoomed and highlighted.

b. ***View and download***: This tab incorporates the best representatives of the families of the most-probable low energy conformational families from the results that have passed the filtering step. The molecules are displayed using Jmol applet windows that also enable basic viewing and measurement options under the right-click options. Each of the conformations can be downloaded from this section.



Blood group A

18 Results

3-Fucosylated Blood group A tetraose

Sequence :	GalNAc a1-3 (Fuc a1-2) Gal b1-4 (Fuc a1-3) Glc
Category :	Fucosylated oligosaccharides (3 Fucosyllactose core)
Molecular weight :	837.770

+ More information

3-Sialyl-Gb3 (Sialylated Globotriose)

Blood group A Lewis B antigen pentaose type1

Blood group A Lewis Y antigen pentaose type2

Blood group A antigen hexaose type 1

Blood group A antigen hexaose type 2

Name	Blood group A antigen hexaose type 1
Sequence	GalNAc a1-3 (Fuc a1-2) Gal b1-3 GlcNAc b1-3 Gal b1-4 Glc
View representations	Ring CFG
Molecular Weight	1056.96
Category	Blood group A antigens
Glycosidic linkages	a1-2 a1-3 b1-3 b1-4
Glycan composition	GalNAc : 1 Fuc : 1 Gal : 2 GlcNAc : 1 Glc : 1
Comment	-
References	Minor HMO of blood group A secretor individuals. A Kobata. Structures and application of oligosaccharides in human milk. 2010. Proc. Jpn. Acad. 86,731-749.

Blood group A antigen hexaose type 1

Blood_grpA_antigen_hexaose_type1_conf1

Download PDB File

Blood_grpA_antigen_hexaose_type1_conf2

Download PDB File

Blood_grpA_antigen_hexaose_type1_conf3

Figure 6.5: An illustration of the results in BiOligo. (*Top*) Preview (*Left*) Molecule information (*Right*) Display and download.

Conclusions

The present work is the first report of the application of high-throughput molecular modeling to complex carbohdyrates as exemplified by the number of different molecules that have been submitted to the application of high accuracy molecular mechanics force field coupled to genetic algorithms. The present selection of more than 250 complex carbohydrate molecules is focused on glycan determinants recognized to show interactions with glycan binding proteins. All the entries incorporated in BiOligo have been sourced from experimental studies like X-ray crystallography, glycan array etc. The accompanying conformational characterization of ~120 disaccharide segments provides the foundations for further explorations as this collection of data allows the coverage of more than 80% of the constituting disaccharides found in all the glycan determinants that have been reported thus far. The same strategy can be applied to other types of carbohydrate containing molecules, for example to the repeat units found in bacterial polysaccharides.

The data generated was stored in a relational database, called BiOligo and its sub-section, GlycoLego. BiOligo is open access and can be queried by the user through the web-interfaced search engine. It categorizes the structural information into logical sections for the user to access using pre-customized searching techniques. The database will be maintained and regularly updated. In the near future, BiOligo shall be linked with existing databases dealing with crystallographic data of protein carbohydrate interactions (lectins, glycosaminoglycans intracting proteins, glycosyl transferases, and monoclonal antibodies). Efforts will be devoted to link the present database with accessible glycan information when a global consensus is reached regarding a minimum information about a glycan array experiment (similar to MIAME [15]) for which efforts are already being made in the glycobiology community.

Computational Methods

All bioactive oligosaccharides contained in the database have been sequentially built using the same protocol. First, the SWEET-II web-based tool on the *Glycosciences.de* web portal (http://www.dkfz.de/spec/sweet/doc/index.php) was used to generate a 3D model from the oligosaccharide sequence [6]. The resulting 3D model was further optimized using MM3 force field [16-18] as implemented in the TINKER package [19] and then saved in the Protein Data Bank (PDB) format. Subsequently, the carbohydrate atom and bond typing were manually checked and corrected within the SYBYL X1.3 interface [20].

The Shape software [5] has been used to perform the high-throughput computational exploration of many di- and all oligosaccharides entries, whose conformations have been reported in the present investigation. Shape uses a genetic algorithm for searching the conformational space of the glycans. The MM3 force-field [16-18] is used for the energy evaluations, which have been performed using a value of 4.0 for the dielectric constant for all calculations. The block diagonal minimization method for geometry optimization was used in MM3 with the default energy-convergence criterion ($\Delta E=0.00008*n$ kcal/mol every 5 iterations, where n= number of atoms). MM3 allows full relaxation of the glycosidic residues taking into account the *exo*-anomeric effect [16, 21] and this force field allows optimization to a nearby transition state (with the full matrix Newton-Raphson method).

The genetic algorithm implementation in Shape is a generational parallel population Lamarkian method that follows molecular evolution. The genetic operators in action are mutation, migration and crossover. A population size of 25 individuals was specified for inclusion in every population of conformations throughout the search, while the total number of parallel populations to be used during the search was set to 20. Each generation produced by the genetic algorithm comprised:

Total number of individuals = population size * total number of populations.

The energy convergence criterion for the conformers generated was assigned a window size of 20 to search for improvements (i.e. the search was terminated when even after 20 generations no significant improvements in conformational energy was found). The highest energy difference is the entire window that is accepted as a significant improvement for the search to continue (i.e. the limit) is of -0.5 kcal/mol. This is directly related to the maximum efficiency of the evolution of conformers, since this is the absolute minimum limit to the length of the conformation search. For each run, once the 'best' conformer has been found, the search still continues for a number of generations, to the specified window size, till satisfied that the results have converged.

To analyze the large amount of conformations generated by the GA, the results were clustered into distinct families of low energy conformations. The conformations were clustered using atom distances, ignoring hydrogen atoms and a 1Å tolerance for RMSD[2] from the cluster centroid. After the families of low energy conformations are clustered, a further filtering is applied based upon possible low energy regions that could be populated by the conformations of the molecules being investigated. Out of the cluster centroids reported after Shape clustering, the ones that inhabit the low energy regions are selected and stored as the final results of the conformational sampling in BiOligo.

The 3D structures deposited in BiOligo can be viewed on the interface via the Jmol application [22]. Jmol is in an interactive web browser applet, which is an open-source, cross-platform 3D Java visualizing tool for chemical and molecular structures that provides high-performance 3D rendering with standard available computer hardware. The provision to download the atomic coordinates for further independent use is provided in the PDB format.

[2] RMSD is an abbreviation for root mean square deviation.

All the calculations have been performed using the facilities of the Centre d'Expérimentation et de Calcul Intensif en Chimie (CECIC) on a cluster of computers made up of a 18-node Dell Power Edge C6100 (24GB and 48GB of central memory), 7-node Bull R424E3 (32 GB of central memory) linked by an Infiniband interconnection network, making a total of 316 processors and with access to a disk storage system offering a global capacity of 2.3 TB. This facility is part of the Grenoble University High Performance Computing Center : CIMENT.

Acknowledgements

We thank Dr. Jimmy Rosen and Sebastien Morin for their assistance in the implementation of the computer program Shape in a multiprocessor environment. The contributions of Drs. Anne Imberty, Laurence Miguet, Michael Reynolds and Petra Schwalie in setting up a prototypical version of the database are also greatly acknowledged. The implementation and the design of the database benefited from the excellent contributions of Alexandre Finet and Hervé Valentin, respectively.

The research leading to this publication has received funding from the European Commision's Seventh Framework Programme FP7/2007-2013, under the grant agreement n° 215536. : Marie Curie Programme, Initial Training Network « EuroGlycoArrays », and from the ANR program in Genomic and Plant Biotechnology : WallArrayII. Financial support from the Centre National de la Recherche Scienitifique and the European Synchrotron Radiation Facility is acknowledged, as the access to the facilities of the Centre d'Expérimentation et de Calcul Intensif en Chimie (CECIC), Grenoble.

References:

1. Cummings RD: **The repertoire of glycan determinants in the human glycome**. *Mol Biosyst* 2009, **5**(10):1087-1104.

2. **Consortium for Functional Glycomics (CFG)** [http://www.functionalglycomics.org/]

3. Imberty A, Perez S: **Structure, conformation, and dynamics of bioactive oligosaccharides: theoretical approaches and experimental validations**. *Chem Rev* 2000, **100**(12):4567-4588.

4. Pérez S: **Molecular modeling in glycoscience**. In: *Comprehensive Glycosciences: Analysis of Glycans.* Edited by Kamerling JP, vol. 2; 2007.

5. Rosen J, Miguet L, Perez S: **Shape: automatic conformation prediction of carbohydrates using a genetic algorithm**. *J Cheminform* 2009, **1**(1):16.

6. Bohne A, Lang E, von der Lieth CW: **SWEET - WWW-based rapid 3D construction of oligo- and polysaccharides**. *Bioinformatics* 1999, **15**(9):767-768.

7. **GLYCAM Web** [http://www.glycam.com]

8. McNaught AD: **International Union of Pure and Applied Chemistry and International Union of Biochemistry and Molecular Biology. Joint Commission on Biochemical Nomenclature. Nomenclature of carbohydrates**. *Carbohydr Res* 1997, **297**(1):1-92.

9. Fadda E, Woods RJ: **Molecular simulations of carbohydrates and protein–carbohydrate interactions: motivation, issues and prospects**. *Drug Discovery Today* 2010, **15**(15-16):596-609.

10. Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R: **Advancing glycomics: implementation strategies at the consortium for functional glycomics**. *Glycobiology* 2006, **16**:82R - 90R.

11. **OligoTech® - Product catalogue 2012 - Elicityl** [http://www.elicityl-oligotech.com/?fond=rubrique&id_rubrique=2]

12. **Apache Web Server** [http://www.apache.org/]

13. **PHP: Hypertext Preprocessor** [http://www.php.net/]

14. Vaswani V: **MySQL: The complete reference**, 1 edn: McGraw-Hill Osborne Media; 2005.

15. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al*: **Minimum information about a microarray experiment (MIAME) - toward standards for microarray data**. *Nat Genet* 2001, **29**(4):365-371.

16. Allinger NL, Yuh YH, Lii JH: **Molecular mechanics. The MM3 force field for hydrocarbons. 1**. *Journal of the American Chemical Society* 1989, **111**(23):8551-8566.

17. Lii JH, Allinger NL: **Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics**. *Journal of the American Chemical Society* 1989, **111**(23):8566-8575.

18. Lii JH, Allinger NL: **Molecular mechanics. The MM3 force field for hydrocarbons. 3. The van der Waals' potentials and crystal data for aliphatic and aromatic hydrocarbons**. *Journal of the American Chemical Society* 1989, **111**(23):8576-8582.

19. Pappu RV, Hart RK, Ponder JW: **Analysis and application of Potential Energy Smoothing and search methods for global optimization**. *The Journal of Physical Chemistry B* 1998, **102**(48):9725-9742.

20. TRIPOS: **SYBYL-X 1.3, Tripos, Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA**. In.; 1991 - 2011.

21. Pérez S, Imberty A, Engelsen SB, Gruza J, Mazeau K, Jimenez-Barbero J, Poveda A, Espinosa J-F, van Eyck BP, Johnson G *et al*: **A comparison and chemometric analysis of several molecular mechanics force fields and parameter sets applied to carbohydrates**. *Carbohydrate Research* 1998, **314**(3-4):141-155.

22. Herráez A: **Biomolecules in the computer: Jmol to the rescue**. *Biochemistry and Molecular Biology Education* 2006, **34**(4):255-261.

Chapter 7

FUCOSE-BINDING LECTINS

Pseudomonas aeruginosa

Pseudomonas aeruginosa is a free-living gram-negative bacteria, commonly found in soil, water, vegetation and plant surfaces and occasionally on the surfaces of animals (**Figure 7.a**).

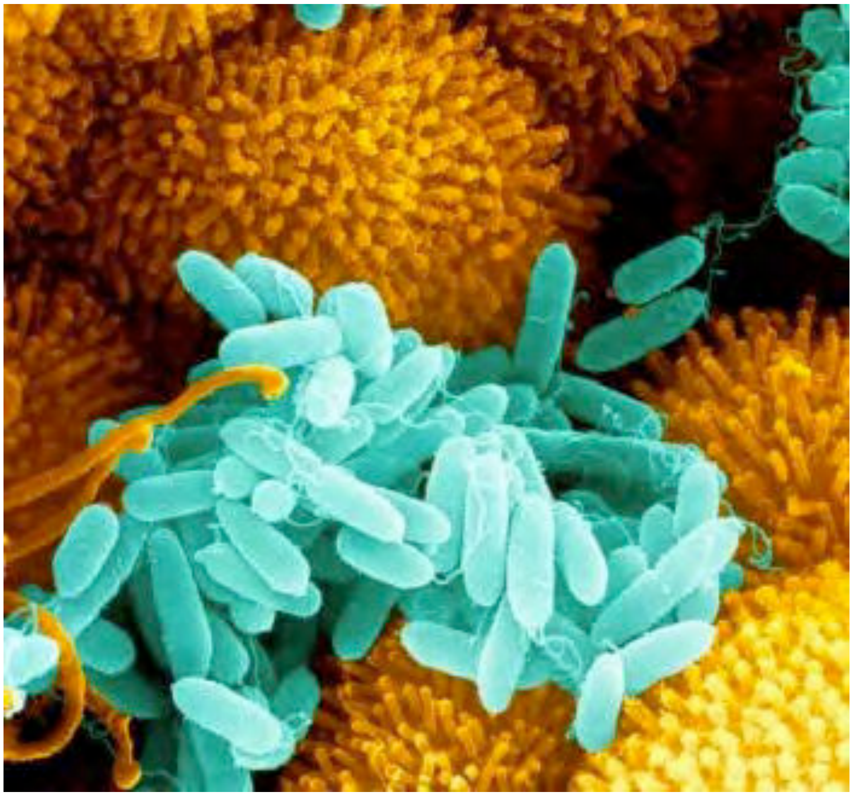


Figure 7.a: The gram-negative bacteria *Pseudomonas aeruginosa* [1].

P. aeruginosa is the epitome of an opportunistic pathogen of humans that almost never infects uncompromised tissues, yet there is hardly any tissue that it cannot infect if the tissue defenses are compromised in some manner. It causes urinary tract infections, respiratory system infections, dermatitis, soft tissue infections, bacteremia, bone and joint infections, gastrointestinal infections and a variety of systemic infections, particularly in patients with severe burns, transplants, cystic fibrosis, cancer (having undergone chemotherapy) and AIDS[1] patients who are immunosuppressed. *P. aeruginosa* is especially more dangerous to certain populations comprising people with weak immune systems, the elderly and those who have been hospitalized for long periods of time. Moreover, since this bacterium is relatively resistant to most antibacterial medications, infection can be deadly, particularly when it infects the lungs or bloodstream. The case fatality rate in these patients is near 50%.

P. aeruginosa produces high-levels of L-fucose binding, PA-IIL protein (lectin), in association with its cytotoxic virulence factors and under quorum sensing control[2]. The affinity to L-Fuc-containing oligosaccharides and its relation to the conformation has been studied as a part of this thesis.

B.2. *Burkholderia ambifaria*

Burkholderia ambifaria is a member of the *Burkholderia* cepacia complex, a group of closely related bacteria that cause lung infections in immune-compromised patients etc (**Figure 7.b**). *Burkholderia ambifaria* is usually associated with plant rhizospheres where it has bio-control effects on other microorganisms.

There is evidence that indicates that lung pathogen lectins and adhesins frequently target fucose on human epithelia [2]. The lectin BamBL from *Burkholderia ambifaria* can bind to artificial glycosphingolipid-containing vesicles, human saliva and lung tissues, which confirmed that BambL recognizes a wide spectra of fucosylated epitopes, albeit with a lower affinity for biological material from non-secretor individuals [3].

[1] Acquired immunodeficiency syndrome

[2] Quorum sensing is a system of stimulus and response correlated to population density. Many species of bacteria use quorum sensing to coordinate gene expression based upon the density of their local population. *P. aeruginosa* uses quorum sensing to coordinate the formation of biofilms, swarming motility, exo-polysaccharide production and cell aggregation. These bacteria can grow within a host without harming it, until they reach a certain concentration. Then they become aggressive, develop to the point at which their numbers become sufficient to overcome the host's immune system, and form a biofilm, leading to disease within the host.

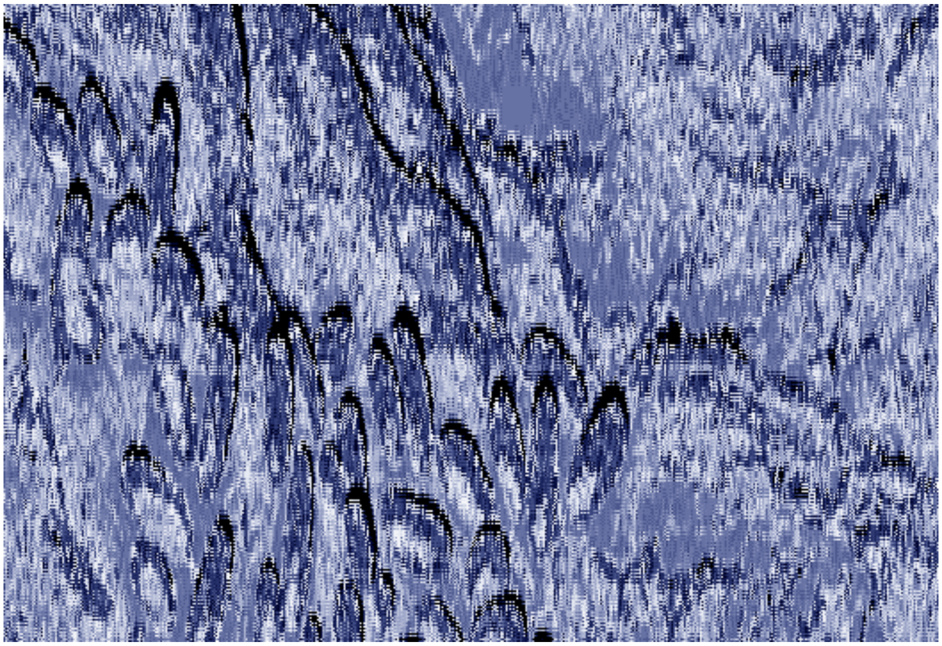


Figure 7.b: The gram-negative bacteria *Burkholderia* found in roots of plants [4].

Burkholderia cenocepacia

Burkholderia cenocepacia is a species of gram-negative bacteria that is common in the environment and may cause disease in plants. It is an opportunistic pathogen causing often-fatal infections in humans especially in patients with cystic fibrosis and chronic granulomatous disease.

Bacteria utilize different lectin topologies and fucose binding sites with different specificities towards fucosylated glycans. It is suspected that they may have potentially different responses towards humans with different histo-blood group oligosaccharides. The application of the low energy conformers generated during the process of populating BiOligo has been probed with the docking studies of fucose-containing ligand conformers with fucose-binding lectins from the bacteria introduced above. The results[3] of the docking studies were co-related to experimental (titration microcalorimetry and glycanarray) data in Chapter 7.

[3] Some studies are still on-going

References:

1. Berger J: ***Pseudomonas aeruginosa* bacteria.** In. Edited by bacteria Pa.

2. Sulak O, Cioci G, Lameignere E, Balloy V, Round A, Gutsche I, Malinovska L, Chignard M, Kosma P, Aubert DF *et al*: ***Burkholderia cenocepacia* BC2L-C is a super lectin with dual specificity and proinflammatory activity**. *PLoS Pathog* 2011, **7**(9):e1002238.

3. Audfray A, Claudinon J, Abounit S, Ruvoen-Clouet N, Larson G, Smith DF, Wimmerova M, Le Pendu J, Romer W, Varrot A *et al*: **The fucose-binding lectin from opportunistic pathogen *Burkholderia ambifaria* binds to both plant and human oligosaccharidic epitopes**. *Journal of Biological Chemistry* 2011.

4. **Antibiotics from *Burkholderia*** [http://www.futurity.org/health-medicine/cystic-fibrosis-bacteria-fights-mrsa/]

CHAPTER 7

[1]

Soluble lectins from opportunistic bacteria binding to human fucosylated epitopes: Screening specificity by docking and glycan array

Introduction

Interaction between protein and carbohydrate play important roles in many biological and pathological processes. For example, infection by bacteria is often initiated by the specific recognition of host epithelial surfaces by glycan binding proteins that are virulence factors having a major role in the first steps of adhesion and invasion. The host glycosylation is specific for species, tissues, cell types and development. The variations in animal glycans as a function of time and space is mirrored by the variety of strategies that pathogens use to exploit the host surface and escape the defence [1]. As part of the strategies adopted by microbes, a large number of their proteins, either lectins, toxins or adhesins, have the capacity to specifically recognize complex oligosaccharides present on host tissue [2, 3].

Among the human glycoconjugates that can be the targets of bacterial lectins, the A, B, and H antigens are complex fucosylated oligosaccharides present on endothelial cells and erythrocytes of all individuals of blood group A, B, or O, respectively [4]. The ABH antigens are also expressed in saliva, tears, and mucus secretions in the digestive tract of individuals who display the secretor phenotype, Se [5]. In addition, Lewis epitopes, that are also fucosylated oligosaccharides, depend on the Lewis phenotype of the individuals. The biological role of the ABO and Lewis histo-blood group systems remains to be

[1] *Publication draft*. A. Sarkar, J. Arnaud, A. Audfray, E. Gillon, D. Smith, S. Pérez, A. Imberty* & A.Thomas (two first authors participated equivalently to this work).
AS designed the method for the *in silico* docking experiments, modeled the oligosaccharides and contributed to the manuscript.

elucidated, but since 80s, several studies pointed out some correlations between the repartition of phenotype in population and the susceptibility to diseases [6, 7]. The most cited examples are the O phenotype presenting higher susceptibility towards cholera toxins [8, 9], and for gastroenteritis caused by Norwalk virus [10].

For secretor individuals, blood group related epitopes are also present in lung mucus. The nature of the oligosaccharides present in airways depends not only on ABO, Lewis and secretor phenotype but also on long-term inflammatory diseases such as chronic bronchitis and cystic fibrosis (CF). More particularly, fucosylated glycoconjugates, which are present in higher quantity in mucins [11] and N-glycans [12] of CF lungs, appear to be the target for lectins from pathogenic bacteria that are responsible for morbidity and mortality in CF patients. Soluble lectins with high affinity for human fucosylated oligosaccharides have been identified in *Pseudomonas aeruginosa* and bacteria from the *Burkholderia cepacia* complex such as *B. cenocepacia* and *B. ambifaria* [13-15]. LecB from *Pseudomonas aeruginosa* is a tetrameric protein that displays an unusually strong micromolar affinity to L-fucose in a tight binding site which requires two Ca^{2+} ions [16, 17]. BambL from *Burkholderia ambifaria* is a trimeric lectin arranged in a β-propeller fold with two similar binding sites per monomer, resulting in an hexameric arrangement of fucose binding sites [13]. Previous specificity studies indicate that both lectins bind to a large variety of fucosylated oligosaccharides, with LecB having higher affinity to Lewis A epitope [18] and BambL to H-type2 epitopes [13]. A third fucose-binding bacterial lectin of interest is the N-terminal domain of Bc2lC from *B. cenocepacia*. Bc2lC is an hexameric lectin, each of the monomer containing two domains with different specificity and oligomerisation state [19]. The N-term assembles as a TNFα-like trimer with strong specificity for H-type1 and Lewis Y oligosaccharides [15], while the C-term is a dimeric mannose-binding lectin [19].

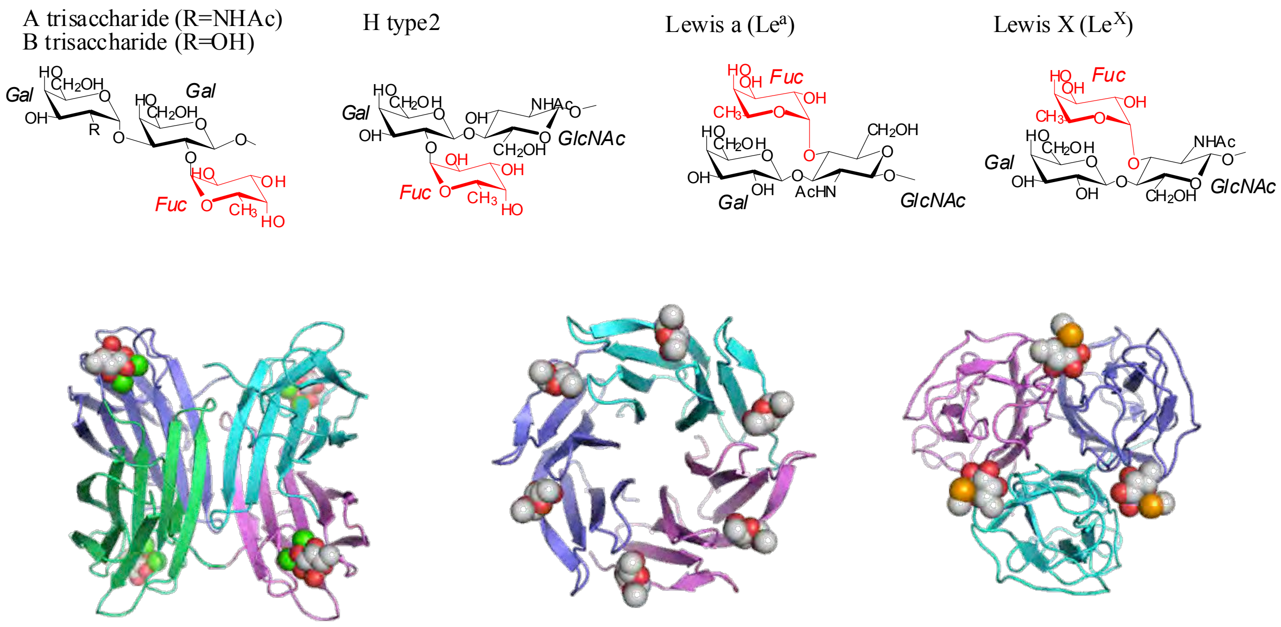


Figure 7.1. Schematic representation of fucosylated trisaccharides and bacterial lectins used in the docking calculations (LecB, BambL and Bc2L-C-nt, from left to right).

Bacteria utilize therefore various lectin topologies and fucose binding sites resulting in difference in specificity toward fucosylated oligosaccharides and therefore potentially towards human with different histo-blood group oligosaccharides. Molecular modeling can help in rationalizing these observed specificity differences. It can also help in designing glycocompounds that can compete with binding to human tissues. Several such compounds have already been synthesized for targeting LecB [20-23]. Conformational analysis of histo-blood group epitopes [24, 25] demonstrated that these oligosaccharides can adopt a limited number of well defined conformations. The results of molecular docking studies of fucose and fucose-containing oligosaccharides have been published using a variety of computational approaches, in protein targets such as histo-blood group antibodies [26, 27], virus capsid proteins [28] or bacterial lectins [29-31]. Good predictive fucose binding mode in the LecB lectin was achieved using several docking algorithms, even though the presence of two bridging calcium ions in the binding site was computationally challenging [29-32].

In this paper, the structural recognition between a series of fucosylated oligosaccharides representative of histo-blood group oligosaccharides (**Figure 7.1A**) and three fucose-binding lectins namely BambL, Bc2L-C-Nter and LecB (**Figure 7.1B and 7.1C**) is investigated using computational approaches so as to gain a greater insight into the atomic interactions between the glycan and the lectins. The binding of the fucosides was estimated through molecular docking calculations. The different low energy conformations of the histo-blood group oligosaccharides have been taken from a recently developed structural database BiOligo. The theoretical calculations presented in this paper are compared to the semi-quantitative binding data derived from glycan array screening and with thermodynamic data determined from titration microcalorimetry experiments.

Materials and Methods

Molecular docking

Preparation of ligands and proteins

The three-dimensional (3D) structures of the low energy conformations of the oligosaccharides were taken from the BiOLigo database (http://glyco3d.cermav.cnrs.fr/bioligo). BiOLigo is a database for 3D structures of glycan determinants that contains about 250 entries of bioactive oligosaccharides, accompanied by a total of 200 disaccharides and monosaccharides of interest, which are represented by ~1200 conformers (A. Sarkar, A. Rivet & S. Pérez, manunscript in preparation). Coordinates of low-energy conformers (2 to 4) of each of these oligosaccharides were taken from the database. They have been determined using high accuracy molecular mechanics force-field coupled to genetic algorithm implemented on a high performance computing meso-center [33]. Details of the starting conformation are listed in **Table 7.1**.

Table 7.1: Glycosidic linkage conformations for all starting model of oligosaccharides. The torsion angles about a glycosidic 1→x linkage are Φ / Ψ, with Φ = O5-C1-O1-C_x and Ψ = C1-O1-C_x-C_{x+1}.

	Conf.	H-type 1	H-type 2	A-tri	B-tri	Le^a	Le^X
Fucα1-2Gal	1	-80 / -96	-79 / -94	-95 / -167	-95 / -167		
	2	-97 / -168	-95 / -169	-69 / -87	-70 / -89		
	3	-104 / -166	-76 / -89	-89 / -176	-88 / -176		
Fucα1-3GlcNAc	1						-78 / 150
	2						-151 / 92
Fucα1-4GlcNAc	1					-79 / -98	
	2					-104 / -167	
Galβ1-3GlcNAc	1	-74 / 137				-76 / 142	
	2	-89 / 67				-64 / 172	
	3	-64 / 169					
Galβ1-4GlcNAc	1		-76 / -113				-75 / -104
	2		-88 / -178				-77 / -102
	3		-80 / 64				
Galα1-3Gal	1				73 / 77		
	2				74 / 67		
	3				91 / 177		
GalNAcα1-3Gal	1			72 / 78			
	2			71 / 68			
	3			90 / 175			

Crystal structures of three lectin/carbohydrate complexes were taken from the Protein Data Bank [34] and used as starting point, namely LecB/fucose (code 1GZT), BambL/H type 2 (code 3ZZV) and BC2L-C-nt / fucoside (code 2WQ4). In all protein-carbohydrate complexes, the hydrogen atoms were built assessing a pH of 7 with the Protein Preparation wizard within the Schrödinger Suite 2012 (Schrödinger Inc., L.L.C. New York, NY). Histidine residues were treated as neutral. The two calcium ions present in the binding site of LecB were kept in the procedure. No crystallization water molecules were considered. A standard energy minimization was performed using the Impref algorithm using the OPLS2005 force field [35] with a convergence of heavy atoms to a 0.30Å RMSD.

Docking Calculations

The docking calculations were performed using the program Glide version 5.8 in Simple Precision [36] from the Schrödinger Suite 2012 (Schrödinger Inc., L.L.C. New York, NY). For each lectin, the binding site was defined on the basis of the crystallographic structures with the bonded ligands. For BambL, that has two fucose-binding sites per chain, with high similarities, the intramolecular site was selected. The box side was set to 14 Å in all directions centered on the ligand. During the grid generation, the parameters for van der Waals radii scaling was scaled by 1.00 for atoms with partial charges less than 0.25. The ring conformational sampling was not allowed and no other constraints were defined. During the docking procedure, the OPLS2005 partial atomic charges were assigned to the ligands. The parameters for van der Waals radii scaling in docking was set at 0.80 for atoms with partial charges less than 0.15. Ligand poses were clustered within RMSD less than 0.30 Å and within maximum atomic displacements less than 1.3 Å. Up to 10,000 poses were set to be retained for the initial phase of the docking upon the Glide runs, and submitted to energy minimization. A distance dielectric constant of 4 was used. After the docking procedure, up to 1,000 poses with the best score were saved and used for further analysis. Docking results were analyzed using the Glide pose viewer included in MAESTRO (Schrodinger, Inc., N.Y.). The RMSD between crystal and docked structures were measured considering all the heavy atoms of fucose. All the values were determined using MAESTRO and the 'Superpose in place' command.

Specificity and affinity experiments

Material

Human histo-blood group oligosaccharides were purchased from Elicityl (Crolles, France). The three bacterial lectins were produced recombinantly in *E. coli* as previously described for LecB [17], BambL [13] and BC2LC-nt [15].

Glycan array

Purified LecB and BC2L-C-nt lectin samples were labeled with Alexa Fluor 488-TFP (Invitrogen, CA) according to manufacturer's instructions and re-purified on a D-Salt polyacrylamide-desalting column (Pierce, Rockford IL). Alexa-labeled proteins were used for glycan array screening with standard procedure of the Core H of the Consortium for Functional Glycomics (http://www.functionalglycomics.org, Emory University, Atlanta, GA, USA). The labeled lectins were assayed on version 4.1 of the CFG glycan array comprising 465 natural and synthetic glycans and the data were analyzed in a dose-dependent manner as previously described [37] at 10, 1 and 0.1 µg ml^{-1} of LecB and 1, 0.2 and 0.05 µg ml^{-1} of Bc2L-C-nt and BambL dissolved in 20 mM HEPES, 140 mM NaCl, 5 mM $CaCl_2$, pH 7.5.

Isothermal Titration Calorimetry (ITC)

Recombinant lyophilized LecB was dissolved in buffer (20 mM Tris/HCl pH 7.5, NaCl 150 mM) and degassed. Protein concentration was checked by measurement of A_{280} by using a theoretical molar extinction coefficient of 6990 M^{-1} cm^{-1}. Carbohydrate ligands were dissolved in the same buffer, degassed and loaded in the injection syringe. ITC was performed with a ITC200 microcalorimeter (MicroCal Inc.). Lectin solution was placed in the 200 µl sample cell, at 25°C. Titration was performed with 20 of 2 µl injections of carbohydrate ligands every 300 s. Data were fitted with MicroCal Origin 7 software, according to standard procedures. Fitted data yielded the stoichiometry (n), the association contant (K_a) and the enthalpy of binding (ΔH). Other thermodynamic parameters (i.e. changes in free energy, ΔG, and entropy, ΔS) were calculated from the equation $\Delta G=\Delta H-T\Delta S= -RTlnK_a$ in which T is the absolute temperature and R=8.314 J mol^{-1} K^{-1}. Two independent titrations were performed for each ligand tested.

Results

Specificity of the three bacterial lectins

The fine specificity of the three bacterial lectins of interest was checked using glycan array v4.1 from the Consortium for Functional Glycomics. The overall glycan array data for BambL were described previously [13] and the ones for LecB and BC2L-C-nt are available from the CFG web site (http://www.functionalglycomics.org/). Briefly, BambL and BC2L-C-nt only attached to the oligosaccharides presenting at least one fucose residue, whereas LecB displays also some affinity towards mannose-containing glycans. In order to set up a comparison with the docking results, binding data were analyzed by selecting only glycans having one fucosylated epitope at the non-reducing position. This resulted in data for 53 glycans with three lectins at three concentrations (see **Table S1**, **S2** and **S3** in Annex IV: *Supplementary Material* for Screening specificity by docking and glycan array). The different concentrations gave very consistent data, and only the results corresponding to one lectin concentration are given in **Figure 7.2** with a selection of the histo-blood groups of interest for the present investigation.

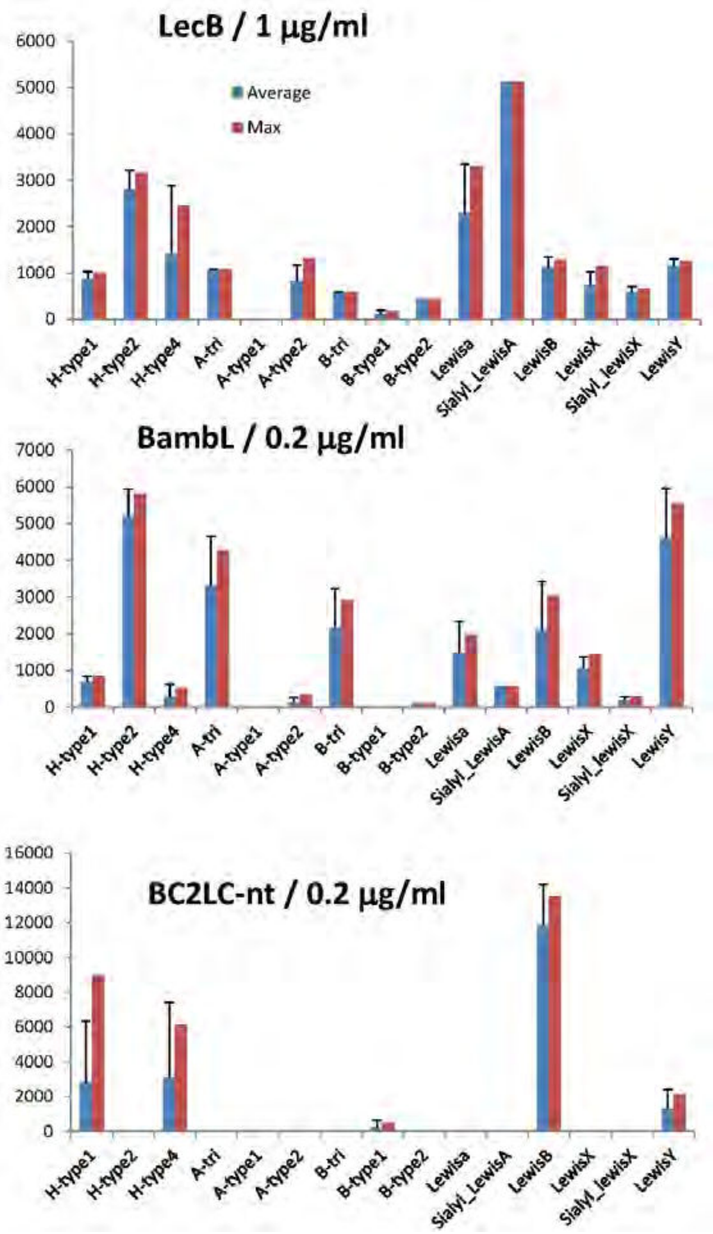


Figure 7.2: Selected data from the glycan array v4.1 experiment performed on three bacterial lectins. Only fluorescent results for biding to terminal fucosylated epitopes presented in monovalent manner on glycans have been selected. Blue bar: average value with standard deviation, red bar: maximum response observed.

Very different binding patterns are observed for the three bacterial lectins. The glycan array for LecB confirms its preference for Lewis A epitopes with higher affinity for the sialylated form (only one glycan on the array). H-type 2 is also a strong ligand, while other fucosylated glycans give moderate but significant labeling, except for A-type 1 and B-type 1 that are negative. BambL was described previously as a fucose-binding lectin with broad specificity [13] and the present analysis confirms its preference for H-type 2 and Lewis Y oligosaccharides. BC2L-C-nt has the narrowest specificity among the three lectins since, among the selected epitopes, it binds only to the ones that contain the Fucα1-3Galβ1-3GlcNAc motif, i.e. H type 1 and Lewis B. It should be noted that it also binds, albeit less strongly to H type 3 (Fucα1-3Galβ1-3GalNAcα) that is not present in **Figure 7.2** but listed in Annex IV: *Supplementary Material* for Screening specificity by docking and glycan array.

Docking calculations

The strategy for docking was first validated by docking the α-methyl fucoside into the binding site of BambL and LecB. In both cases, the "glide-score" which approximates the ligand binding free energy, indicated a very favorable interaction and indeeed the resulting pose was in good agreement with the location of the monosaccharide in the corresponding crystal structures of lectin complexes. It should be noted that the fit between the predicted and observed location of fucose is better for BamBL than for LecB (**Figure 7.3**). Indeed the presence of the two calcium ions in the binding site of LecB presents a difficult challenge for docking calculations [31].

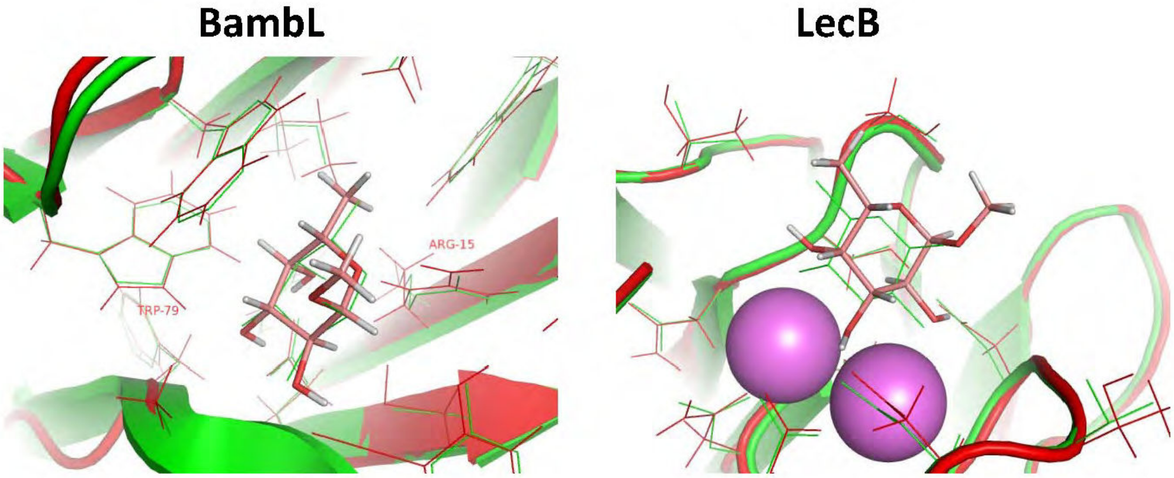


Figure 7.3: Docking of α-methyl fucoside in the binding site of BambL and LecB. The protein

model is represented in red with docked ligand as sticks. The crystal structures of BambL/fucose (3ZW0) and LecB/fucose (1GZT) are represented in green with ligands as lines.

Six histo-blood group fucosylated trisaccharides were selected for the docking procedures to two different lectins (calculations with Bc2LC-nt are in progress and will be included in the submitted version of the manuscript). In each case, between 2 to 3 starting conformations were used for the oligosaccharides, as selected from BiOligo database. By default, the Glide program would frequently invert carbohydrate rings so this option was disabled. For all trisaccharides, a good sampling was observed. The conformation with the best "glide-score" was selected for each run and the data are reported below.

Table 7.2: Docking results obtained for six fucosylated trisaccharides with BambL lectin.

	Glide	Glide	Φ	Ψ	Φ	Ψ	O3…W79.NE1	O5…ARG15.NH2
	Score	Energy						
H type 1			Fucα1-2Gal		Galβ1-3GlcNAc			
Conf_3	-4.9	-40.2	-110	-123	-48	149	2.7	2.6
Conf_2	-4.5	-40.0	-104	-121	-24	113	3.0	2.6
Conf_1	-4.3	-38.0	-97	-114	-56	144	3.0	2.6
H type 2			Fucα1-2Gal		Galβ1-4GlcNAc			
Conf_1	-4.7	-41.5	-115	-140	-64	-108	2.9	2.8
Conf_2	-4.7	41.1	-117	-153	-68	-101	3.2	2.8
Conf_3	-4.7	-41.9	-115	-142	-65	-108	2.9	2.8
A tri			Fucα1-2Gal		GalNAcα1-3Gal			
Conf_1	-3.8	-33.7	-85	-148	62	71	3.0	2.6
Conf_3	-3.7	-33.8	-81	-129	65	66	6.4	3.1
Conf_2	-2.9	-28.1	-83	-136	67	74	2.9	2.9
B tri			Fucα1-2Gal		Galα1-3Gal			
Conf_1	-3.9	-37.9	-114	61	58	54	3.1	2.8
Conf_2	-3.5	-35.7	-120	52	128	135	3.0	2.8
Conf_3	-3.5	-32.3	-83	-132	59	70	2.9	2.9
Lewis a			Fucα1-4GlcNAc		Galβ1-3GlcNAc			
Conf_2	-4.3	-40.6	-99	-162	-54	179	3.1	2.8
Conf_1	-2.6	-26.4	-76	-98	-67	134	10.0	8.0
Lewis X			Fucα1-3GlcNAc		Galβ1-4GlcNAc			
Conf_2	-3.2	-33.5	-103	106	-67	-88	2.9	2.9
Conf_1	-2.5	-32.1	-85	142	-68	100	12.6	6.6

Glide Score: an empirical scoring function that approximates the ligand binding free energy (kcal/mol)
Glide Energy: glide evdw + glide ecoulomb scores (kcal/mol)

The results for docking six fucosylated oligosaccharides in BambL binding sites are listed in **Table 7.2**. As expected, the use of several different conformations for the oligosaccharides has a favorable effect on the docking procedure, since in many cases, the starting lowest energy conformation (conf1) is not always the one that produces the best docking "glide-score". In some cases all different starting conformations converged to the same docking pose (see H type 2) whereas in others, variable poses were obtained. In all cases, but two, the fucose is correctly bound in the main binding sites, as checked by the occurrence of two hydrogen bonds (**Table 7.2**). Close inspections of the results indicate that the conformation of the monosaccharide ring is preserved, and that all the conformations at the glycosidic linkages belong to the low energy regions of the corresponding disaccharide segments. On rare occasions, the *cis* conformation of the N-acetyl substituent is found. At the present time, the reasons underlying the occurrence of such an unusual conformation have not been identified; to which extent this finding indicates a flaw in the energy parameterization, or a transient, but still valid conformational state remains to be scrutinized.

When comparing the docking results with available crystal structures of BambL complexed with oligosaccharides, the conformation and positions of best "glide-score" for H-type 1 and H-type 2 trisaccharide display good agreement between theoretical and experimental binding modes (**Figure 7.4**). The agreement is not as good for blood group B oligosaccharide, but it has to be noted that in the crystal structure, the tetrasaccharide has to adopt a constrained shape because of the branching point, that is not required for the trisaccharide used in the docking procedure. Interestingly, Lewis A and Lewis X trisaccharides, that display only moderate affinity for BambL do not bind in their lowest energy minimum (conf 1 in **Table 7.1**). Indeed this rather rigid low energy conformation correspond to the crystal structure of Lewis X [38] and to the solution conformation with stacking of galactose and fucose ring. Such arrangement is not possible in the binding site due to the shape of the fucose-binding pocket and a distorted shape is predicted.

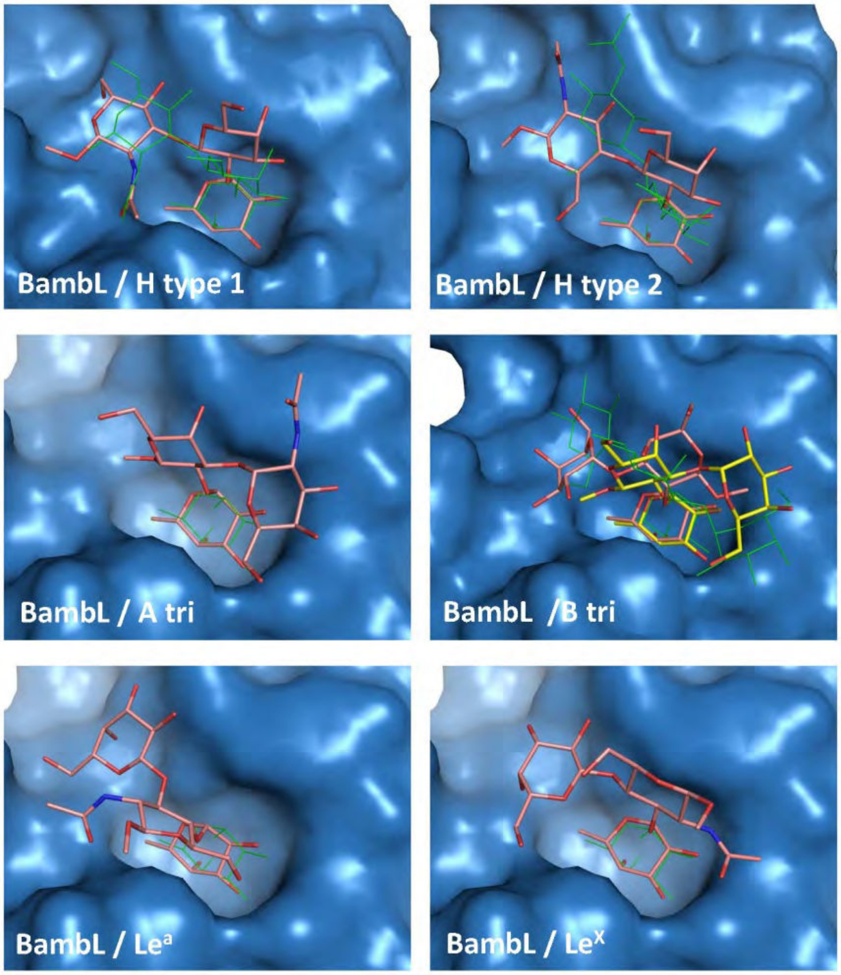


Figure 7.4: Docking of six fucosylated oligosaccharides in the binding sites of BambL. The docking pose with best "glide-score" is represented in red for all oligosaccharides. For the blood group B trisaccharide, the second best orientation is represented in yellow. Comparison with crystal structures is performed with same oligosaccharide when available (H type 1: 3ZW1, H type 2: 3ZZV, blood group B tetrasaccharide: 3ZW2) or elsewhere with fucose (3ZW0), always represented as green line.

When compared to glycan array, the general agreement is good with higher score for H-type 2 than for blood group oligosaccharides, and weaker one for Lewis X. However, H-type 1 and Lewis A have higher score than expected from the glycan array.

The results for docking of six fucosylated oligosaccharides in LecB binding sites are listed in **Table 7.3**. The position of the fucose in the binding site is evaluated by the distance between its oxygen atoms and the two calcium ions, since in the crystal structure, these atoms are directly coordinated with distances between 2.5 and 2.6 Å. It can be seen that in most cases, the fucose is well located, although not as deeply buried in the binding site as it is in the crystal structure.

Table 7.3: Docking results obtained for six fucosylated trisaccharides with LecB lectin.

	Glide	Glide	Φ	Ψ	Φ	Ψ	O2…Ca1	O4…Ca2
	score	Energy						
H type 1			Fucα1-2Gal		Galβ1-3GlcNAc			
Conf_2	-5.1	-47.2	-156	-168	-48	129	2.8	3.0
Conf_3	-3.7	-38.9	-73	70	-48	142	2.8	2.9
Conf_1	-3.2	-36.6	-90	-173	40	120	2.7	4.5
H type 2			Fucα1-2Gal		Galβ1-4GlcNAc			
Conf_2	-4.0	-42.4	-104	-173	22	-153	2.8	3.8
Conf_3	-3.8	-32.8	-72	-90	-77	60	2.6	3.2
Conf_1	-3.6	-33.8	-92	-75	-68	-107	2.7	3.1
A tri			Fucα1-2Gal		GalNAcα1-3Gal			
Conf_3	-3.9	-41.0	-76	-132	96	160	3.5	2.6
Conf_2	-3.1	-36.7	-52	-77	82	68	2.9	3.8
Conf_1	-2.9	-30.3	-84	117	91	76	2.6	3.7
B tri			Fucα1-2Gal		Galα1-3Gal			
Conf_1	-3.2	-35.2	-71	-93	47	53	3.7	2.6
Conf_2	-3.2	-36.0	-72	105	78	153	10.3	11.8
Conf_3	-3.1	-31.4	-79	-134	75	-139	3.9	2.7
Lewis A			Fucα1-4GlcNAc		Galβ1-3GlcNAc			
Conf_1	-4.5	-44.4	-152	-153	-54	169	2.7	3.1
Conf_2	-3.0	-35.5	-80	-97	23	-77	2.7	3.1
Lewis X			Fucα1-3GlcNAc		Galβ1-4GlcNAc			
Conf_2	-3.8	-40.8	-78	17	-51	86	2.7	3.1
Conf_1	-3.2	-41.2	-69	-12	-84	-97	2.7	4.9

Glide Score: an empirical scoring function that approximates the ligand binding free energy (kcal/mol)
Glide Energy: glide evdw + glide ecoulomb scores (kcal/mol)

In this case, no much convergence from different starting conformations is observed. Each starting low energy conformation leads to different poses; some of them exhibit large variations between the initial shape and the final one. All the conformations at the glycosidic linkages belong to the low energy regions of the corresponding disaccharide segments (energy maps are given in Annex IV: *Supplementary Material*). LecB has been cocrystallized with Lewis A [18] and in the crystal structure, the trisaccharide adopts the solution structure with the lowest energy conformation, characterized by stacking of galactose and fucose. The docking procedure used in the present study does not reproduce this conformation. The best pose exhibits a different orientation at the Fucα1-2Gal linkage, which results in a different conformation for the trisaccharide. However, in the case of the Lewis and blood group oligosaccharides the docked conformations in the combining site retain lowest energy conformation of the oligosaccharides. While being tempted to conclude at a successful level of prediction, it seems wiser, in the absence of reproduction of the binding mode of Lewis A, to reinvestigate further the computational protocol. This investigation is on-going.

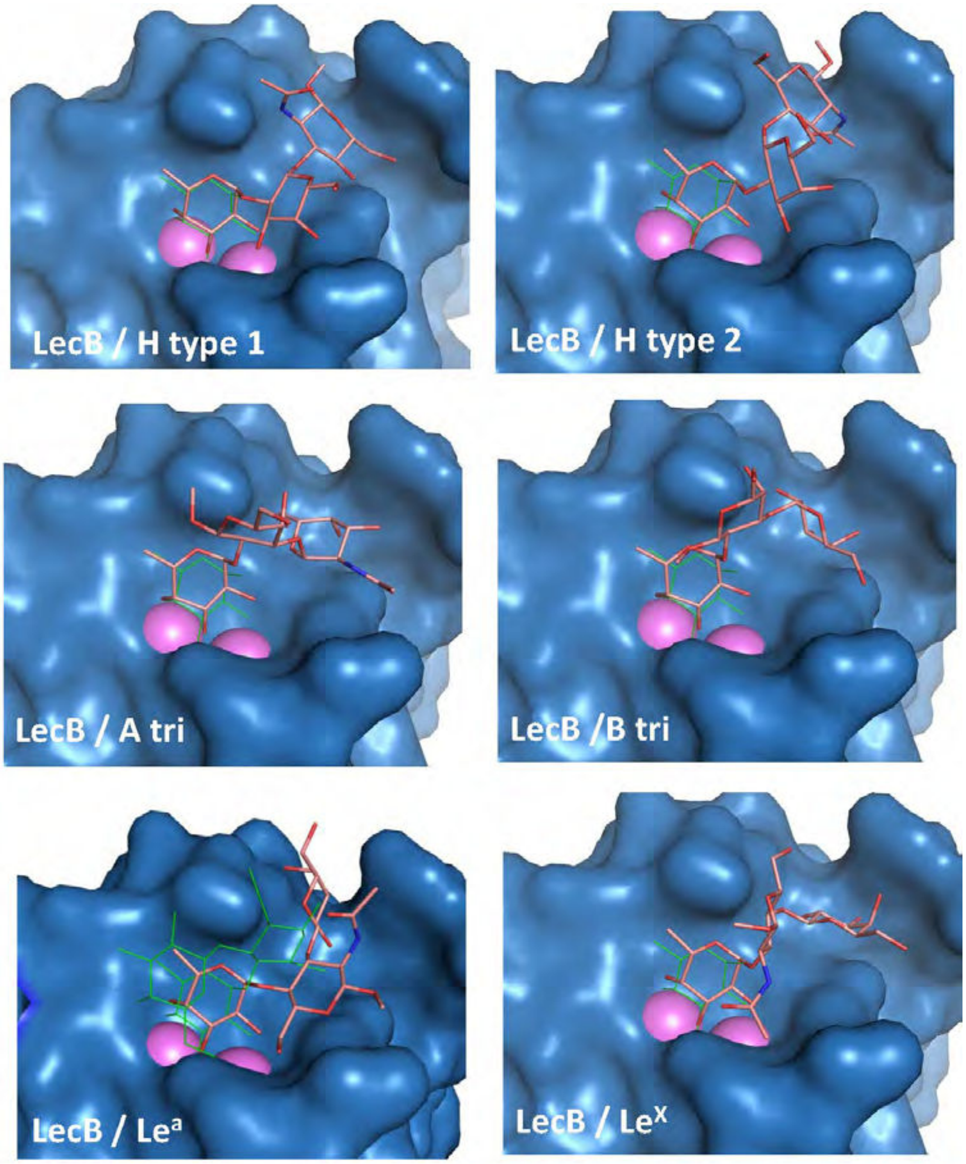


Figure 7.5: Docking of six fucosylated oligosaccharides in the binding sites of LecB. The docking pose with best "glide-score" is represented in red for all oligosaccharides. Comparisons with Le^a trisaccharide and fucose monosaccharide are represented with green lines from corresponding crystal structures (1GZT and 1W8H).

Titration microcalorimetry

For BambL, the docking modes and “glide-scores” obtained are in general agreement with the binding data from the glycan array: some of the divergences observed may be attributed to the presentation and/or to the multivalency on the chips. Thermodynamic data are available in the literature for BambL [13] and BC2LC-nt [15]; the ones for LecB have been measured.

Table 7.4: Titration microcalorimetry data for the interaction between BambL, LecB, BC2LC-nt and fucosylated ligands (all energies in kJ/mol). All data have been measured at least twice and standard deviations are below 15%

	BambL [a]			**LecB**			**BC2LC-nt** [b]		
Ligand	K_D (μM)	-ΔG	−ΔH	K_D (μM)	-ΔG	−ΔH	K_D (μM)	-ΔG	−ΔH
H-type1-tetra	26.1	26.2	17.6				77.2	23.5	23.0
H-type1 tri				1.39	33.4	32.0	77.2	23.5	23.0
H-type 2-tetra	7.5	29.3	44.4	0.48	36.1	39.3	213	16.6	24.9
A-tri	0.46	36.0	53.1	10.3	28.5	25.2			
A- type 2 (penta)	120	22	13.7						
B- type 2 (penta)	95.3	23	25.9	15.5	27.4	40.0			
Le^a-tri[c]				0.21	38.1	35.0	132	22.1	48.1
Le^a-tetra	18.2	27.1	28.7	0.20	38.2	49.1			
Le^X-tri[c]				3.44	31.1	22.3	196	21.2	38.7
Le^X-tetra	34.8	25.4	39.1	1.15	33.9	30.8			
Sialyl Le^a tetra				0.29	37.3	38.9			
Sialyl Le^x penta				1.47	33.3	52.6			

[a] from ref [13]
[b] from ref [15]
[c] from ref [15]

In the present state of calculations, the docking results obtained with LecB are not of a sufficient quality to be tested against thermodynamic data. The energy validation has therefore been performed only using BambL data obtained previously with a variety of oligosaccharides [13].

Both glide score and glide energy of binding can be compared with microcalorimetry data. For the experimental part, both enthalpy and free energy of binding are of interest. The free energy is the most interesting one since it is the direct evaluation of affinity, while the enthalpy of binding is expected to correlate better with calculated energy of binding. All correlations are displayed in **Figure 7.6**.

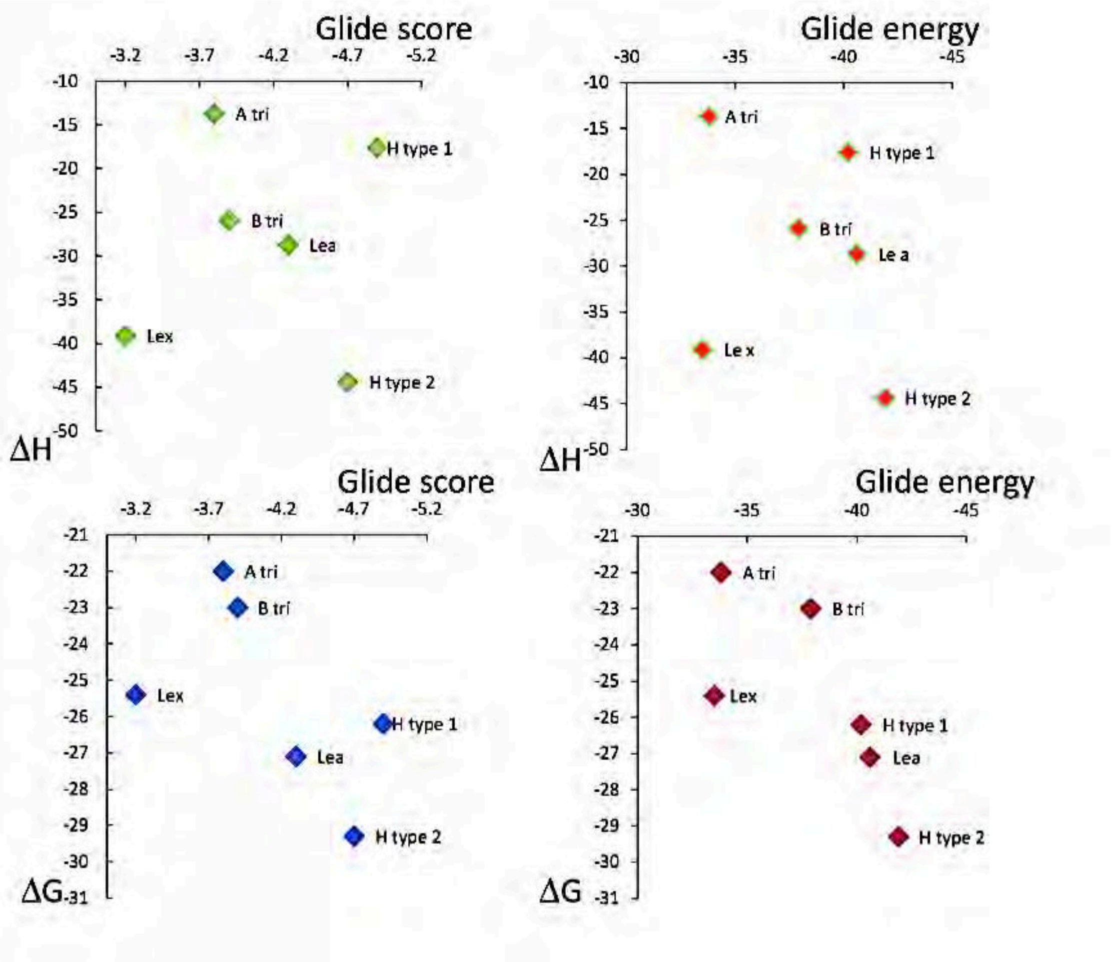


Figure 7.6: Attempts to correlate experimental data (ΔH and ΔG) obtained for the interaction of BambL with a series of oligosaccharides and the experimental data (Glide score and Glide energy) obtained from docking.

Interestingly, the glide data do not correlate well with the enthalpy of binding, resulting in scattered plots. The correlation with the free energy of binding is clearly better. The best correlation is obtained when comparing glide score and free energy of binding from ITC experiment (R^2= 0.55) indicating therefore that the glide score has predictive value for affinity between lectin and oligosaccharides. Only Lewis x trisaccharide is clearly out of the range. Its ommission from the data sets yields a R^2 value of 0.86. Thes data should be however considered as very preliminary since experiments were in general conducted on oligosaccharides longer than the ones that are used for the docking (i.e. H type1 tetrasaccharide instead of H-type 1 trisaccharide). Since BambL has a narrow binding site with preference for short glycan, this can influence the ITC data.

Conclusions

The development of a fast and easy docking protocol would be very useful for analysing the large amount of binding data generated by glycan array. Once the high affinity oligosaccharides are identified, it is of high interest to determine in which orientation and conformation they are bound to the lectin, in order to develop glycocompounds that can block protein/glycan interaction in pathological process. This approach would also be useful for designing lectin mutants with highest specificity for targeted glycan of biological interest. Molecular dynamics simulation with explicit water environment is of course well suited for calculations of free energy of binding. However, the combined use of database of oligosaccharides conformation associated with fast docking procedure appears as a medium-throughput screening approach for the analysis of glycan array data.

Acknowledgments

The work was supported by CNRS and ANR grant NeoLect 11 BSV5 002 01 (A.I. and A.A.), The authors acknowledge the COST action BM1003 "Microbial cell surface determinants of virulence as targets for new therapeutics in Cystic Fibrosis" and the Initial Training Network “EuroGlycoArrays” (A.S). Support from the Association Vaincre la Mucoviscidose and GDR Pseudomonas is appreciated. The glycan array resources were provided by the Consortium for Functional Glycomics (GM62116).

References:

1. Bishop JR, Gagneux P: **Evolution of carbohydrate antigens--microbial forces shaping host glycomes?** *Glycobiology* 2007, **17**(5):23R-34R.

2. Imberty A, Varrot A: **Microbial recognition of human cell surface glycoconjugates**. *Curr Opin Struct Biol* 2008, **18**:567-576.

3. Sharon N: **Carbohydrate-lectin interactions in infectious disease**. *Adv Exp Med Biol* 1996, **408**:1-8.

4. Watkins WM, Morgan WT: **Neutralization of the anti-H agglutinin in eel serum by simple sugars**. *Nature* 1952, **196**:825-826.

5. Henry S, Oriol R, Samuelsson B: **Lewis histo-blood group system and associated secretory phenotypes**. *Vox Sang* 1995, **69**:166–182.

6. Greenwell P: **Blood group antigens: molecules seeking a function?** *Glycoconj J* 1997, **14**(2):159-173.

7. Marionneau S, Cailleau-Thomas A, Rocher J, Le Moullac-Vaidye B, Ruvoen N, Clement M, Le Pendu J: **ABH and Lewis histo-blood group antigens, a model for the meaning of oligosaccharide diversity in the face of a changing world**. *Biochimie* 2001, **83**(7):565-573.

8. Berger SA, Young NA, Edberg SC: **Relationship between infectious diseases and human blood type**. *Eur J Clin Microbiol Infect Dis* 1989, **8**:681-689.

9. Heggelund JE, Haugen E, Lygren B, Mackenzie A, Holmner S, Vasile F, Reina JJ, Bernardi A, Krengel U: **Both El Tor and classical cholera toxin bind blood group determinants**. *Biochemical and biophysical research communications* 2012, **418**(4):731-735.

10. Lindesmith L, Moe C, Marionneau S, Ruvoen N, Jiang X, Lindblad L, Stewart P, LePendu J, Baric R: **Human susceptibility and resistance to Norwalk virus infection**. *Nature medicine* 2003, **9**(5):548-553.

11. Lamblin G, Degroote S, Perini JM, Delmotte P, Scharfman A, Davril M, Lo-Guidice JM, Houdret N, Dumur V, Klein A *et al*: **Human airway mucin glycosylation: A combinatory of carbohydrate determinants which vary in cystic fibrosis**. *Glycoconj J* 2001, **18**(9):661-684.

12. Glick MC, Kothari VA, Liu A, Stoykova LI, Scanlin TF: **Activity of fucosyltransferases and altered glycosylation in cystic fibrosis airway epithelial cells**. *Biochimie* 2001, **83**(8):743-747.

13. Audfray A, Claudinon J, Abounit S, Ruvoën-Clouet N, Larson G, Smith DF, Wimmerová M, Le Pendu J, Römer W, Varrot A *et al*: **The fucose-binding lectin from opportunistic pathogen *Burkholderia ambifaria* binds to both plant and human oligosaccharidic epitopes**. *Journal of Biological Chemistry* 2012, **287**:4335-4347.

14. Garber N, Guempel U, Gilboa-Garber N, Doyle RJ: **Specificity of the fucose-binding lectin of *Pseudomonas aeruginosa***. *FEMS Microbiol Lett* 1987, **48**:331-334.

15. Šulák O, Cioci G, Delia M, Lahmann M, Varrot A, Imberty A, Wimmerová M: **A TNF-like trimeric lectin domain from *Burkholderia cenocepacia* with specificity for fucosylated human histo-blood group antigens**. *Structure* 2010, **18**:59-72.

16. Mitchell E, Houles C, Sudakevitz D, Wimmerova M, Gautier C, Pérez S, Wu AM, Gilboa-Garber N, Imberty A: **Structural basis for oligosaccharide-mediated adhesion of *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients**. *Nature Struct Biol* 2002, **9**:918-921.

17. Mitchell EP, Sabin C, Šnajdrová L, Pokorná M, Perret S, Gautier C, Hofr C, Gilboa-Garber N, Koča J, Wimmerová M *et al*: **High affinity fucose binding of *Pseudomonas aeruginosa* lectin PA-IIL: 1.0 Å resolution crystal structure of the complex combined with thermodynamics and computational chemistry approaches**. *Proteins: Struct Funct Bioinfo* 2005, **58**:735-748.

18. Perret S, Sabin C, Dumon C, Pokorná M, Gautier C, Galanina O, Ilia S, Bovin N, Nicaise M, Desmadril M *et al*: **Structural basis for the interaction between human milk oligosaccharides and the bacterial lectin PA-IIL of *Pseudomonas aeruginosa***. *Biochem J* 2005, **389**:325-332.

19. Šulák O, Cioci G, Lameignère E, Balloy V, Round A, Gutsche I, Malinovská L, Chignard M, Kosma P, Aubert F *et al*: ***Burkholderia cenocepacia* BC2L-C is a super lectin with dual specificity and proinflammatory activity** *PLoS Pathogens* 2011, **7**:e1002238.

20. Andreini M, Anderluh M, Audfray A, Bernardi A, Imberty A: **Monovalent and bivalent N-fucosyl amides as high affinity ligands for *Pseudomonas aeruginosa* PA-IIL lectin**. *Carbohydr Res* 2010, **345**:1400-1407.

21. Johansson EM, Crusz SA, Kolomiets E, Buts L, Kadam RU, Cacciarini M, Bartels KM, Diggle SP, Camara M, Williams P *et al*: **Inhibition and dispersion of Pseudomonas aeruginosa biofilms by glycopeptide dendrimers targeting the fucose-specific lectin LecB**. *Chem Biol* 2008, **15**(12):1249-1257.

22. Marotte K, Préville C, Sabin C, Moumé-Pymbock M, Imberty A, Roy R: **Synthesis and binding properties of divalent and trivalent clusters of the Lewis a disaccharide moiety to *Pseudomonas aeruginosa* lectin PA-IIL**. *Org Biomol Chem* 2007, **5**:2953-2961.

23. Marotte K, Sabin C, Préville C, Moumé-Pymbock M, Wimmerova M, Mitchell EP, Imberty A, Roy R: **X-ray structures and thermodynamic of interaction of PA-IIL from *Pseudomonas aeruginosa* with disaccharide derivatives**. *ChemMedChem* 2007, **2**:1328-1338.

24. Imberty A, Mikros E, Koca J, Mollicone R, Oriol R, Pérez S: **Computer simulation of histo-blood group oligosaccharides. Energy maps of all constituting disaccharides and potential energy surfaces of 14 ABH and Lewis carbohydrate antigens**. *Glycoconj J* 1995, **12**:331-349.

25. Lemieux RU, Bock K, Delbaere LTJ, Koto S, Rao VSR: **The conformations of oligosaccharides related to the ABH and Lewis human blood group determinants**. *Can J Chem* 1980, **58**:631-653.

26. Agostino M, Jene C, Boyle T, Ramsland PA, Yuriev E: **Molecular docking of carbohydrate ligands to antibodies: structural validation against crystal structures**. *J Chem Inf Model* 2009, **49**(12):2749-2760.

27. Imberty A, Mollicone R, Mikros E, Carrupt PA, Pérez S, Oriol R: **How do antibodies and lectins recognize histo-blood group antigens ? A 3D-QSAR study by comparative molecular field analysis (CoMFA)**. *Bioorg Med Chem* 1996, **4**:1979-1988.

28. Koppisetty CA, Nasir W, Strino F, Rydell GE, Larson G, Nyholm PG: **Computational studies on the interaction of ABO-active saccharides with the norovirus VA387 capsid protein can explain experimental binding data**. *Journal of computer-aided molecular design* 2010, **24**(5):423-431.

29. Mishra NK, Kulhánek P, Šnajdrová L, Petřek M, Imberty A, Koča J: **Molecular dynamics study of *Pseudomonas aeruginosa* lectin-II complexed with monosaccharides**. *Proteins* 2008, **72**:382-392.

30. Mishra SK, Adam J, Wimmerová M, Koča J: **In silico mutagenesis and docking qtudy of *Ralstonia solanacearum* RSL Lectin: Performance of docking software to predict saccharide binding**. *Journal of chemical information and modeling* 2012, **52**:1250-1261.

31. Nurisso A, Kozmon S, Imberty A: **Comparison of docking methods for carbohydrate binding in calcium-dependent lectins and prediction of the carbohydrate binding mode to sea cucumber lectin CEL-III**. *Mol Simul* 2008, **34**:469-479.

32. Mishra NK, Kriz Z, Wimmerova M, Koca J: **Recognition of selected monosaccharides by Pseudomonas aeruginosa Lectin II analyzed by molecular dynamics and free energy calculations**. *Carbohydrate research* 2010, **345**(10):1432-1441.

33. Rosen J, Miguet L, Perez S: **Shape: automatic conformation prediction of carbohydrates using a genetic algorithm**. *J Cheminform* 2009, **1**(1):16.

34. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**: 235-242.

35. Jorgensen WL, Tirado-Rives J: **Potential energy functions for atomic-level simulations of water and organic and biomolecular systems**. *Proc Natl Acad Sci U S A* 2005, **102**(19):6665-6670.

36. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK *et al*: **Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy**. *Journal of medicinal chemistry* 2004, **47**(7):1739-1749.

37. Smith DF, Song X, Cummings RD: **Use of glycan microarrays to explore specificity of glycan-binding proteins**. *Methods Enzymol* 2010, **480**:417-444.

38. Pérez S, Mouhous-Riou N, Nifant'ev NE, Tsvetkov YE, Bachet B, Imberty A: **Crystal and molecular structure of a histo-blood group antigen involved in cell adhesion: the Lewis x trisaccharide**. *Glycobiology* 1996, **6**(5):537-542.

General Conclusions

&

Perspectives

GENERAL CONCLUSIONS

Glycans are vital to all cells in every living species. These special biomolecules decorate the cell surface and modulate several cellular functions. Glycans can be analogized to identity badges containing access codes to enter cells or cellular organelles. Depending upon which side of the gate they are on they can have beneficial or harmful effects for the host. Considering that the human genome comprises a much smaller number of genes than previously anticipated, glycans (and lipids) provide the additional empowerment to the cell to adapt to changing environments and pathogen attacks. Approximately 2% of human genes are associated with glycosylation, known from large-scale sequencing and gene function predictions.

A lot of effort has been invested in deciphering the 3D coordinates of glycans, and especially polysaccharides; yet, these valuable data are not readily accessible to the community and lie hidden in literature. As in proteomics and genomics, the speed of advancement in glycomics is dependent upon the development of specific bioinformatics data repositories to start with, to make all assimilated data easily accessible, to have a global vision of what has been achieved and what remains to be done to complete the puzzle of molecular machines. Towards this end, for polysaccharide structures, atomic coordinates were sourced from the literature, from techniques such as X-ray crystallography, neutron diffraction, electron diffraction, nuclear magnetic resonance (NMR) and molecular modeling, to extract data about the asymmetric unit of the cell content. The extracted data was converted to either fractional or Cartesian coordinates to generate the atomic coordinate files in standardized representations of Protein Data Bank (PDB) or Mol2 (SYBYL) formats. The files were generated using an in-house PHP script called PDBGenerator, developed for the construction of PolySac3DB, which was designed to convert fractional and cylindrical/polar coordinates to PDB format. Further, this data was used to generate helical/expanded forms of the unit cell structures. The 3D atomic coordinates thus collected or generated were classified into 18 families of well-recognized polysaccharide classes. They were then organized into a relational database, PolySac3DB, that provides, besides atomic coordinates, other related information about the polysaccharide structure such as the diffraction diagrams, linkages present, space group, the unit cell parameters, type of helix (since the majority of the data assimilated were from diffraction experiments), link to the abstract, and bibliographic references. The 3D coordinates were made available for viewing (and basic manipulation via Jmol) and download. PolySac3DB is a compilation of scientific research, covering almost 50 years of work in carbohydrate science. PolySac3DB is hosted at CERMAV_CNRS (Centre de Recherches sur les Macromolecules Vegetales) and shall be regularly updated with new structures that match the scope of this database.

Perception and protection are the primary functions of the cell wall. In *E. coli*, the multi-layered cell wall forms the first line of defense against environmental dangers. It is composed of a lipid bilayer decorated with special lipids called lipopolysaccharides (LPS). Our immune system actually uses these LPS to identify bacteria when they try to invade our bodies. Antibodies recognize these LPS and mobilize our defenses to fight infection. The O-antigenic polysaccharides of *E. coli* O5ac and O5ab have emerged as new pathotypes of persistent infantile pediatric diarrhea and now have also been detected to cause infection in adults. The O-antigenic polysaccharide present in the LPS of *E. coli* O5ac and O5ab are positional isomers with the difference lying in the substitution pattern of one monosaccharide of the tetrasaccharide biological repeating unit. The O-antigenic cell surface polysaccharides of *E. coli* O5ac and O5ab were studied to understand their immunochemical similarities and also the shape of the polysaccharide that generally has a direct influence on its biological function. Using molecular modeling, the oligosaccharide biological repeats were built and their conformational energy hypersurface was sampled. Viable models were selected based upon their energy. Further, validation was performed using high temperature molecular dynamics (MD) simulations, and torsion angle functions of the constituting disaccharide segments were plotted and matched with the models generated. Finally, NMR experiments were carried out for both the samples and the models were found to be in close agreement to the experimental results. The models, which were the best represented in the NMR analysis of the polysaccharide samples were selected for the construction of the polysaccharide models. Both O5ac and O5ab (models with four biological repeats) were found to have 2-fold helices that had a common pattern of arrangement of the monosaccharides on it. In both cases, the α-D-Gal*p*NAc residue was found to be located on the inside of the helix, and external accessibility was curtailed by the N-acetyl moiety of β-D-Qui*p*3NAc. Further, analysis on molecular models of eight repeating biological repeats revealed that in both cases the same epitope is exposed to the surface and corresponds to the β-D-Qui*p*3NAc-(1→3)-β-D-Rib*f*-(1→4)-β-D-Gal*p* fragment. This may be an indication to a common glycan epitope due to which antibodies cannot differentiate between O5ac and O5ab. Further, immunological validation is required to confirm this point.

O1303 is another *E. coli* strain that is one of the prime causative agents in mastitis, a major disease in dairy herds, which causes considerable loss to the dairy industry every year. Here again, the LPS of O1303 is suspected to play a crucial role during infection. Comparative sequence analysis of the O1303 with the O5 serotypes shows a close resemblance between the two. O5ab differs from O1303 in just one monosaccharide (β-D-Rib*f* instead of α-L-Fuc*p*, respectively), while O5ac has an additional difference in the glycosidic linkage connecting two biological repeat units. Conformational analysis of the oligosaccharide fragments were carried out using a genetic algorithm as employed for the investigations on O5ac and O5ab. In addition, Φ/Ψ energy maps were generated and used for validation of the models. The latter were then used to guide the construction of polysaccharide chains. The polysaccharides of O1303 and O5ab, both formed 2-fold helices had a comparable helix pitch, though the shape of the helix was observed to differ. Yet, a similar trend of the α-D-Gal*p*NAc residue lying in the interior of the helix was observed, with a guarded accessibility due to the N-acetyl group of the β-D-Qui*p*3NAc. Unfortunately, sufficient quantities of the O1303 sample were not available for NMR experiments. Further investigations with NMR or other biophysical methods accompanied by immunological tests shall be useful in elucidating the cross-reactivity between the strains.

The flexibility of glycan 3D structures is a blessing for the cell, but a bottleneck for glycobiology to correctly characterize them. Moreover, experimental procedures are also constrained by the availability of sufficient amounts of glycan samples to conduct tests on them. Molecular modeling in glycoinformatics is a way to overcome this hindrance. Using this as a high-throughput technique would lead to the rapid generation of dependable 3D coordinates that can be used to aid experimental techniques (for validation and further exploration) as well as other simulations to understand the multi-faceted roles of glycans. Towards this goal, the 3D structures of mono-, di- and polysaccharides have been modeled during the course of this thesis. This employs a genetic algorithm based upon the concepts of Lamarkian evolution. The theory centers on the idea of acquired characteristics and the inheritance thereof. It assumes that complexity of organisms increases over time and there is a direct transmission of phenotypic traits from parents to offspring. The method for conformational sampling used during the course of this thesis, considers one conformation to be one individual within a population of conformations that is generated from the starting structure. The torsion angles are the inheritable traits depending upon whether the conformation energy is feasible to adopt a stable state. On this basis, more than 400 glycan determinants including bioactive oligosaccharides and several of their constituting *lego* blocks have been generated. The results have been organized into a relational database, BiOligo (and its sub-set of constituting di- and monosaccharide fragments, called GlycoLego), containing information about the most stable representatives of the families of low energy conformers. Other searchable fields in this database include the trivial name, constituent type (mono-, di- or oligosaccharide), category into which the entry is grouped in BiOligo and the molecular weight. The 3D structures are available for viewing and basic manipulation via Jmol and can also be downloaded. The entries included in this database are known glycan determinants characterized already by X-ray crystallography, nuclear magnetic resonance (NMR), other biophysical methods as well as glycan arrays. The structures provided can be used in deciphering binding data from glycan arrays, and providing realistic starting conformations to be used in molecular dynamics (MD) simulations, molecular docking of oligosaccharides with proteins or nucleic acids and improving the resolution of the structures of glycoproteins, in particular with small angle X-ray scattering (SAXS) experiments. BiOligo is an open source database provided to the scientific community that shall be updated with new glycan determinants being identified.

A static picture of glycosylation is, however, not sufficient to reflect dynamic developmental and disease-related fluctuations that are critical factors in the final shaping of oligosaccharide conformations. Bacterial infection, for example, often occurs by the specific recognition of the host epithelial surface by glycan binding proteins that are virulence factors with major roles to play in invasion and adhesion initiation. As host glycosylation varies between tissues, species, cell types, developmental stage, physiological condition of the cell etc., pathogenic bacteria mirror this variability in the host glycan as a function of time and space to escape the host's defense mechanism and penetrate the host cell surface. The structural recognition between various low energy fucosylated oligosaccharide conformations of histo-blood group oligosaccharides (derived from BiOligo) and the fucose binding lectins of *Pseudomonas aeruginosa* (LecB) and *Burkholderia ambifaria* (BambL) [calculations for *Burkholderia cenocepacia* (Bc2lC-n-ter) are still in progress] were studied using molecular docking. The theoretical calculations presented in this work compared well with glycan array data for BambL, in general, except for two oligosaccharides (H type 1 and Lewis A). In case of LecB, the fucose is well-placed (evaluated by the distance between its oxygen atoms and the Ca^{2+} ions) in the binding site of the docked complexes though not as deeply buried as observed in the crystal structure. However, in case of LecB, the docking calculations do not reproduce the stacked galactose and fucose conformation as recorded in the crystal structure with this ligand. The best pose exhibits a different orientation at the Fuc-α12-Gal linkage. For the Lewis x and blood group oligosaccharides, however, the docked conformations retain their low energy conformations. The docking with LecB is being re-calculated to reproduce the binding mode observed in the crystal structure of Lewis a and only then the docking scores could be considered for comparison with thermodynamic data. For BambL, the docking scores are in general agreement with the binding data from glycan arrays. These results are still at a preliminary stage and need more refinement for correlation. The experimental results are also preliminary as the oligosaccharides are longer than the ones used for docking.

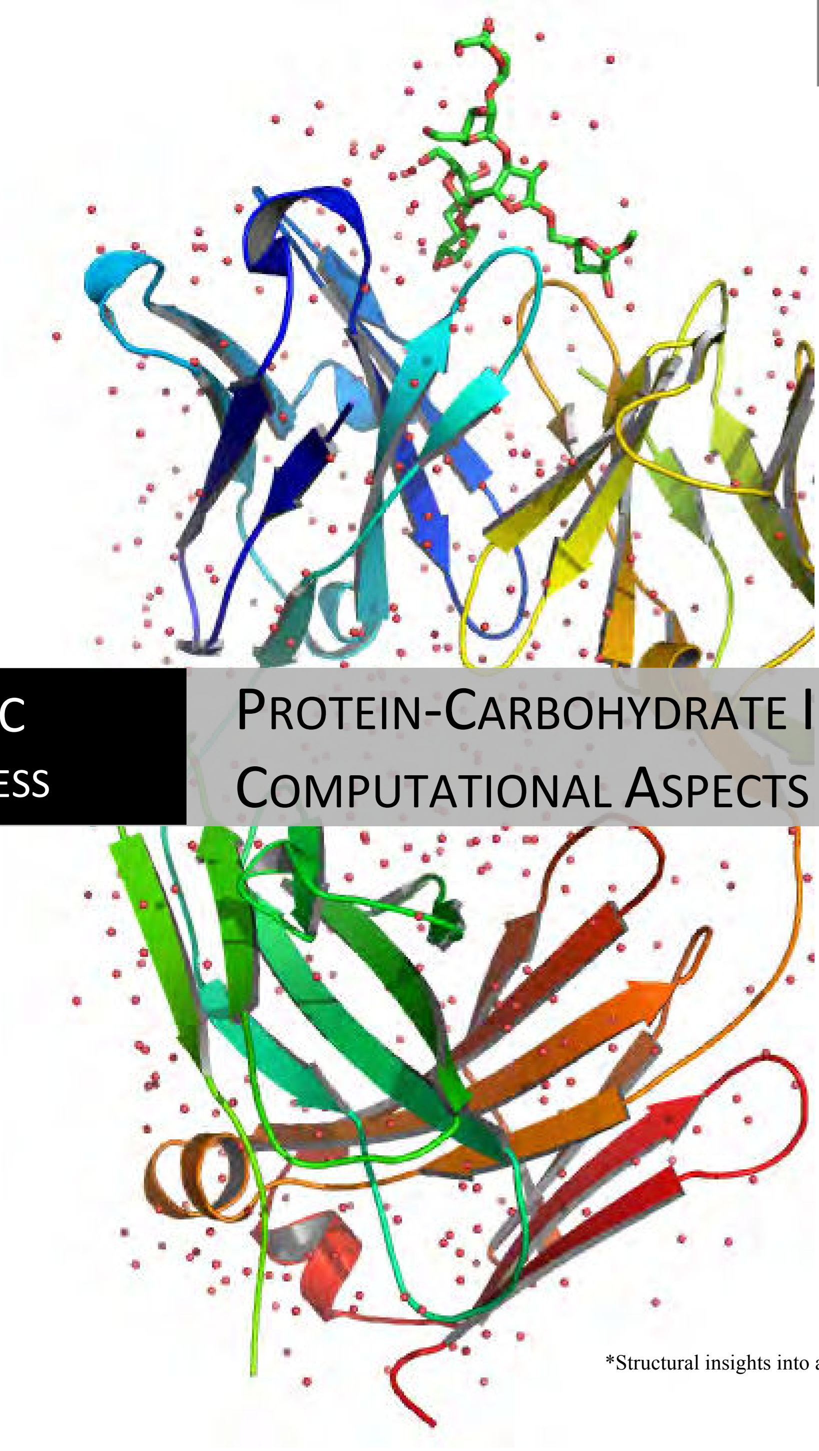
Glycan specificity is a crucial requirement for the plethora of cellular and sub-cellular interactions occurring all at the same time in an orchestrated fashion. In such a crowded setting it is very important for the substrate to find the right ligand conformation, ignoring all the other distractions that appear. This is achieved through a multi-point interaction filter for the correct recognition to occur through a perfect match of shape and chemistry. Thus, it is imperative to characterize glycan structure and functions to understand how they manipulate proteins, which are the work-horses of the cell, to help keep the cell in a robust healthy state, as well as to probe the aspects of when and why they fail to carry out their work efficiently - the sweet aspect of protein-carbohydrate interactions.

Structural Glycobiology

[In press]

CRC
PRESS

PROTEIN-CARBOHYDRATE INTERACTIONS: COMPUTATIONAL ASPECTS



*Structural insights into antibody recognition of mycobacterial polysaccharides - 3HNS

| Anita Sarkar & Serge Pérez |

CHAPTER 8

Protein-Carbohydrate Interactions: Computational Aspects

[1]

8.1 Introduction
8.2 Specific features of carbohydrate modeling
8.3 Protein-carbohydrate interactions
8.3.1 van der Waals and electrostatic interactions
8.3.2 CH/π Interactions
8.3.3 Solvation-desolvation
8.4. Force fields designed for carbohydrates
8.5 Computational tools for docking carbohydrates on proteins
8.5.1 Molecular docking
8.5.1.1 Difficulties in molecular docking
8.5.1.2 Docking algorithms
8.5.1.2.1 Incremental construction algorithms
8.5.1.2.2 Genetic algorithms
8.5.1.2.3 Hierarchical algorithms
8.5.1.3 Scoring functions
8.5.2 Molecular dynamics simulations
8.5.3 Molecular robotics
8.5.4 Free energy calculations
8.5.4.1 Relative free energy of binding
8.6 Case studies
8.6.1 Recognition
8.6.1.1 Lectins
8.6.1.2 Antibodies
8.6.1.3 Chemokine-glycosaminoglycan interactions
8.6.1.4 Transport
8.6.2 Synthesis: glycosyltransferases
8.6.3 Glycosyl hydrolases / glycosidases
8.6.3.1 Glycosyl hydrolases on single chain
8.6.3.2 Glycosyl hydrolases on a solid substrate
8.7 Conclusion

[1] Anita Sarkar[1] and Serge Pérez[1,2]
[1] Centre de Recherches sur les Macromolécules Végétales (CERMAV, CNRS), Grenoble, France
[2] European Synchrotron Research Facility (ESRF), Grenoble, France
anita.sarkar@cermav.cnrs.fr, serge.perez@cermav.cnrs.fr, serge.perez@esrf.eu

8.1 Introduction

In Nature, carbohydrates form an important family of biomolecules, as simple or complex carbohydrates, either alone or covalently linked to proteins or lipids. Most of the earlier studies on carbohydrates focused on plant polysaccharides, such as cellulose, starch, pectins, etc., largely because of their wide range of applications. More recently, the role of carbohydrates in biological events has been recognized and glycobiology has emerged as a new and challenging research area at the interface of biology and chemistry. Of special interest are the carbohydrate-mediated recognition events that are important in biological phenomena, which gives a pivotal role to the study of protein-carbohydrate interactions. Actually, the binding protein partners of carbohydrates encompass a wide variety of macromolecules involved in functions such as recognition, biosynthesis, modification, hydrolysis, etc. (**Figure 8.1**).

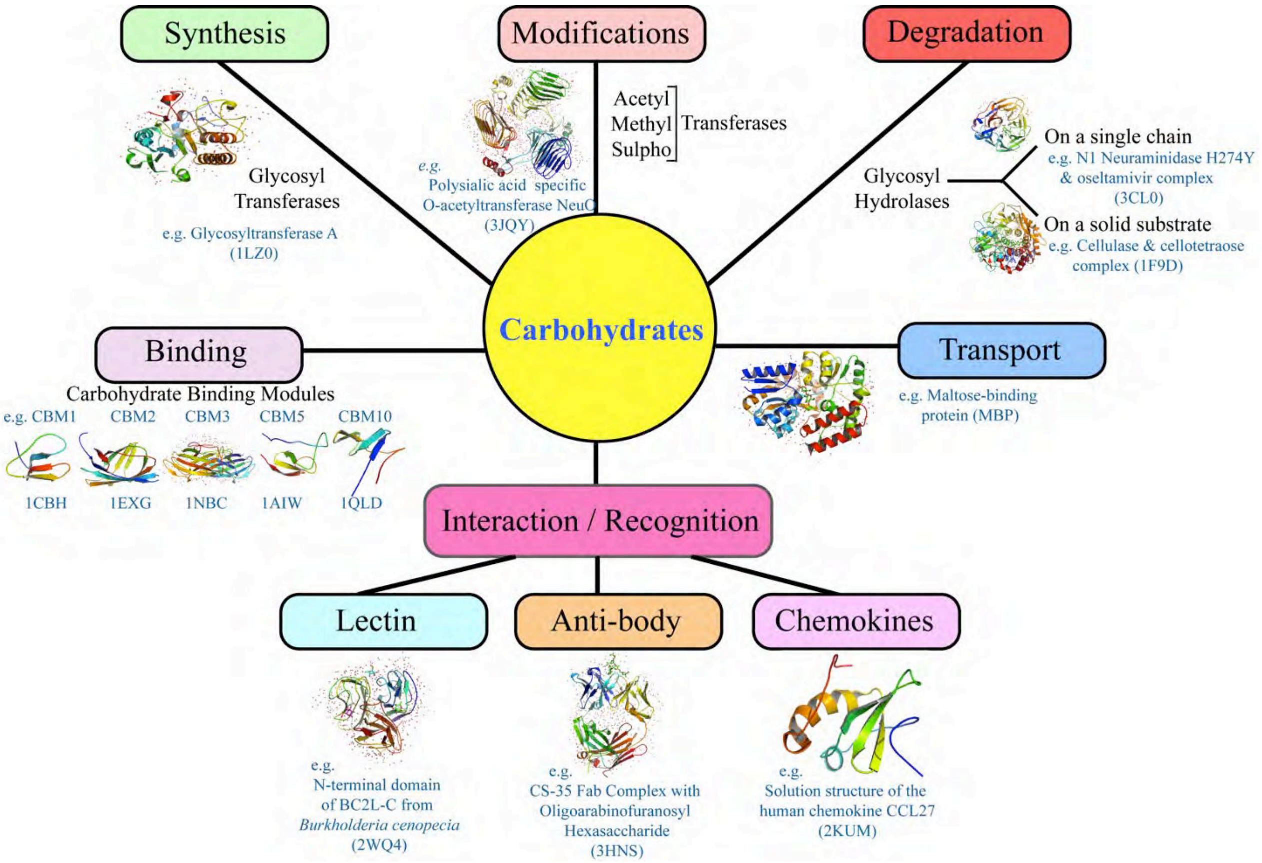

Synthesis
Glycosyl Transferases
e.g. Glycosyltransferase A (1LZ0)
Modifications
Acetyl
Methyl
Sulpho
Transferases
e.g. Polysialic acid specific O-acetyltransferase NeuO (3JQY)
Degradation
Glycosyl Hydrolases
On a single chain
e.g. N1 Neuraminidase H274Y & oseltamivir complex (3CL0)
On a solid substrate
e.g. Cellulase & cellotetraose complex (1F9D)
Carbohydrates
Binding
Carbohydrate Binding Modules
e.g. CBM1
CBM2
CBM3
CBM5
CBM10
1CBH
1EXG
1NBC
1AIW
1QLD
Transport
e.g. Maltose-binding protein (MBP)
Interaction / Recognition
Lectin
Anti-body
Chemokines
e.g. N-terminal domain of BC2L-C from *Burkholderia cenopecia* (2WQ4)
e.g. CS-35 Fab Complex with Oligoarabinofuranosyl Hexasaccharide (3HNS)
e.g. Solution structure of the human chemokine CCL27 (2KUM)

Figure 8.1. Synopsis of the families of proteins interacting with carbohydrates, illustrated with examples from the Protein Data Bank (PDB) for Synthesis [1], Modifications (acetyltransferase [2]), Degradation by Glycosyl Hydrolases (a) On a single chain [3], (b) On a solid substrate [4], Carbohydrate Binding Modules (CBMs) [5-9], Transport [10], Interaction/Recognition (a) Lectin [11], (b) Anti-body [12], (c) Chemokines [13].

Determination of the three-dimensional (3D) structural and dynamical features of complex carbohydrates, carbohydrate polymers, and glycoconjugates, along with the understanding of the molecular basis of their associations and interactions represent the main challenges in structural glycoscience [14].

Elucidation of the 3D structures and the dynamical properties of oligosaccharides is a prerequisite for a better understanding of the relationships between structures and functions, involving the biochemistry of recognition processes and the subsequent rational design of carbohydrate-derived drugs. Seemingly, the elucidation and the understanding of the different structural levels of polysaccharides are required to relate structure to properties. Ultimately, some polysaccharides are also carriers of biological information that can only be deciphered if their interactions with other biological macromolecules are understood. Unfortunately, oligosaccharides, either in their free form or as part of glycoconjugates, are inherently difficult to crystallize and structural data from X-ray studies are sparse [15]. In solution, the flexibility of certain glycosidic linkages produces multiple conformations which coexist in equilibrium. The use of several spectroscopic methods, with appropriate time resolution, is necessary for analysis of the conformational behavior of such molecules [16, 17]. As for polysaccharides, they differ from other biological macromolecules because the diffraction data that can be obtained are not sufficient to permit crystal structure determination based on the data alone. Hence, procedures for molecular modeling of carbohydrates and carbohydrate polymers have been devised as an important tool for structural studies of these compounds. Various molecular modeling methods have been developed [18] and have been widely used for the determination of oligosaccharide and polysaccharide conformations [19]. The progress made in algorithms and computational power allows for the simulation of carbohydrates in their natural environment, that is, solvated in water or in organic solvent, in concentrated solution. These developments along with their applications have been thoroughly reviewed in a previously published chapter [20].

Carbohydrates, along with proteins and nucleic acids, constitute one of the central building blocks of life. The interactions between proteins and carbohydrates play a role in numerous biological processes such as protein specificity in antibody-antigen recognition, cell-cell adhesion, enzyme-substrate specificity, molecular transport, etc. They are critical to the onset, detection, and, potentially, also the prevention of human diseases such as cancer. The interactions between proteins and complex carbohydrates such as polysaccharides are also involved in the biosynthesis and biodegradation of the major raw materials on Earth. Experimental assessment of the carbohydrate-recognition by X-ray crystallography is impeded by difficulties of co-crystallizing proteins and carbohydrates. Nevertheless, highly resolved protein-carbohydrate complexes gathered from X-ray synchrotron investigations have accumulated to the point where it has been possible to compare the experimentally derived structures with those predicted from computational methods. Some general features governing the protein-carbohydrate interactions have been derived, and computational tools have evolved and improved accordingly. These tools provide efficient ways to increase our understanding of the different contributions to the binding energy. These developments allow searching the conformational space efficiently and yield reliable estimates of the binding free energy. They allow to explore *in silico* cases where the experimental data are lacking, and provide sound structural information for a rational design of bio-active carbohydrates or carbohydrate mimetics.

In this chapter we aim to review the significant contributions and the present status of the application of computational methods to the characterization and prediction of protein-carbohydrate interactions.

8.2 Specific features of carbohydrate modeling

Carbohydrates have a potential information content that is several orders of magnitude higher than any other biological macromolecule. The diversity of carbohydrate structures results from the broad range of monomers (>100) of which they are composed and the different ways in which these monomers are joined (glycosidic bonds). Thus, even a small number of monosaccharide units can provide a large number of different oligosaccharides (also referred to as glycans), including branched structures, a unique feature among biomolecules. For example, the number of all possible linear and branched isomers of a hexasaccharide exceeds 10^{12} [21].

The carbohydrate recognition mechanism depends on (i) the sequence of the mono-saccharides in the glycan (i.e. glucose vs. mannose), (ii) the anomeric centers (i.e. α or β), (iii) the linkage positions (i.e. 1-3 vs. 1-4), and (iv) the chemical modifications to the core glycan (i.e. sulfation, phosphorylation, methylation, acetylation, etc.). The strength of this interaction is also determined by the carbohydrate conformation and orientation with respect to the binding site.

Carbohydrates and their derivatives possess many hydroxyl groups and thus a large number of rotatable bonds. Due to the many hydroxyl groups, these compounds are usually highly water soluble and their logP is often negative. The surface of carbohydrates and their derivatives is composed of hydrophobic and hydrophilic patches formed by nonpolar aliphatic protons and polar hydroxyl groups. This leads to anisotropic solvent densities around carbohydrate molecules. In aqueous environments, favorable interactions of water molecules with the hydrophilic patches result from electrostatic interactions and hydrogen bonding. Conversely, the interaction of water with hydrophobic surface patches is unfavorable. Such equilibrium between hydrophobic and hydrophilic patches forms the basis for such properties as carbohydrate solubility in water, or such functions as molecular recognition.

Another essential feature of carbohydrates is their conformational flexibility [22]. Compared to drug-like molecules, carbohydrates are typically much more flexible. The relative orientations of two consecutive monosaccharide units in a disaccharide moiety are expressed in terms of the glycosidic linkage torsional angles Φ and Ψ around the glycosidic bonds which are defined as $\Phi = \mathrm{O5–C1–O–C_x}$ and $\Psi = \mathrm{C1–O–C_x–C_{(x+1)}}$ for a (1→x) linkage (**Figure 8.2**). The energetically favorable conformations of a carbohydrate dimer may be easily shown on energy plots called (Φ, Ψ) maps which are somewhat similar to the Ramachandran plots used to visualize the backbone dihedral angles of the constituent amino acids in proteins. These plots feature multiple minima with the separating energy barriers being over 10-15 kcal/mol.

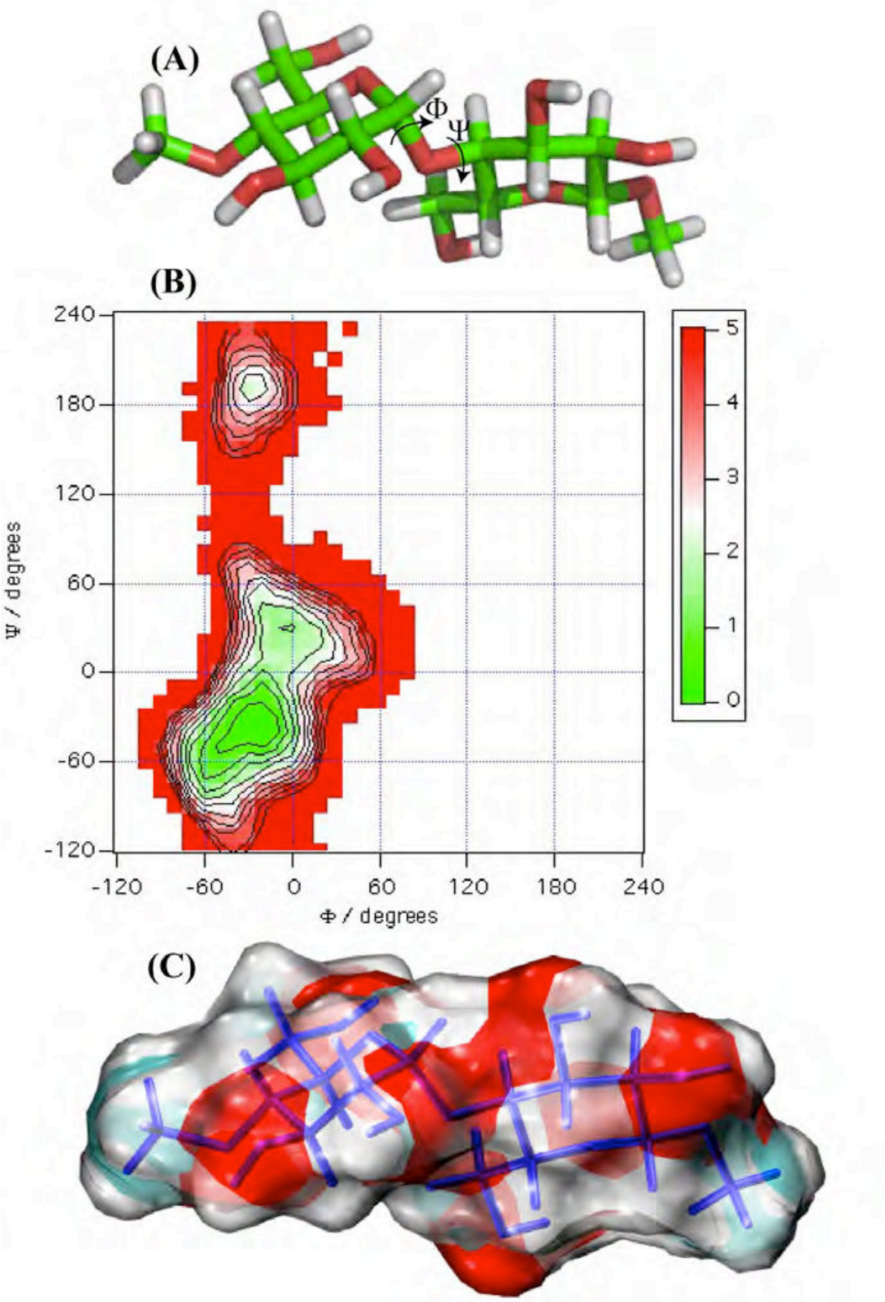


Figure 8.2. Molecular representation of the disaccharide (α-D-Glcp-(1-4)-β-D-Glcp) with the Φ and Ψ torsion angles shown on the glycosidic linkage. The potential energy surface shows conformational energy with respect to the Φ and Ψ torsion angles. The favored low-energy Φ/Ψ combinations are shown in light color, while the high energy regions are shown in red and the inaccessible regions are shown in white. The surface of the disaccharide is composed of hydrophobic (green) and hydrophilic (red) patches, formed by nonpolar aliphatic protons and polar hydroxyl groups.

However, carbohydrates in complex were found to adopt conformations belonging to different minima. These observations underline the necessity for thoroughly sampling the conformational space of carbohydrate oligomers during docking. While this may be feasible for glycosidic bonds, the number of degrees of freedom increases rapidly when, in addition to this, we take into account the orientation of the hydroxyl groups.

8.3 Protein-carbohydrate interactions

As with other types of macromolecular interactions, the formation of the complex is driven by favorable changes in enthalpy (ΔH) and entropy (ΔS). Thermodynamic measurements have indicated that the binding free energy of monosaccharide to proteins is quite small. ΔG increases in a significant manner whenever disaccharides or higher oligosaccharides are interacting with proteins. Whenever such proteins are interacting with carbohydrates, a high "avidity" is observed as a result of a multivalent effect. The binding free energy between a carbohydrate molecule and a protein partner (ΔG) is indeed the variable of interest to be assessed. It is assumed to be composed of independent contributions in terms of van der Waals forces, electrostatic interactions with or without encompassing hydrogen bonding, the hydrophobic effect etc.

8.3.1 van der Waals and electrostatic interactions

From the large number of hydrogen bond donors and acceptors present in carbohydrates, complex and dense networks of hydrogen networks with proteins arise. The complexity of such networks is enhanced by the competition occurring with the water molecules for hydrogen bonds. The overall enthalpic gain from hydrogen bonding may be counter-balanced by some entropic cost.

8.3.2 CH/π Interactions

These characterize the enthalpy of binding of carbohydrates to protein. It is defined as a type of hydrogen bond occurring between a hydrogen atom attached to a carbon and the π systems of arenes. Typically, this is a weak effect. Despite the full recognition of this effect, its computation requires a high level of theory and is not fully taken into account in the computational procedures [23].

As observed in many crystal structures of protein-carbohydrate complexes, aromatic residues of the proteins are often stacked against some faces of the carbohydrates. Such an arrangement results from the hydrophobic effect wherein small hydrophobic moieties of the solute induce an ordering of the water molecules at the solvent interface. The resulting decrease of the hydrophobic surface area induces a decrease in solvent ordering and a consequent favorable change in entropy. Alternatively, a non-classical hydrophobic effect has also be documented to occur in lectin-carbohydrate complexes, where the complex formation is driven by enthalpy due to favorable interactions between the solute forming the complex as well as favorable interactions between the solvent molecules [24, 25].

8.3.3 Solvation-desolvation

As a result of docking carbohydrates into proteins, the number of atomic contacts between ligand and protein is maximized and the subsequent structure is such that the carbohydrate lies more or less flat on the protein surface [26]. However, X-ray crystal structures show contradictory features, with carbrohydrate residues extending into the surrounding solvent. These structures might be correctly computed if the impact of solvation and desolvation on the binding free-energy were properly taken into account.

8.4 Force fields designed for carbohydrates

To study carbohydrate structures and properties using molecular modeling techniques, molecular mechanics potential energy functions and parameters specific for this class of molecules are required. Appropriate force fields for carbohydrate systems have been created, with the aim of reproducing the particular effects that influence their global structural properties in solution [27]. The exocyclic hydroxymethyl group behavior is defined by the ω-angle (O5-C5-C6-O6) and its preference for *gauche* states can be reproduced by introducing scaling factors that slightly modify the 1-4 non-bonded interactions [28]. 1-4 non-bonded interactions define the influence, in terms of electrostatic and van der Waals potentials. 1-4 non-bonded interactions are not treated in the same manner in all force fields (**Figure 8.3**) and this could be a problem in simulating complex systems in which two different force fields have to be used. In these cases, the separate treatment of 1-4 non bonded interactions can assure a full compatibility among the force fields. The potential impact of choosing the 1-4 scaling factors often becomes irrelevant when glycans bind to proteins because generally their freedom in the binding site is reduced. In literature, several reviews describe and compare the performance of carbohydrate force fields used in glycomodeling [29, 30].

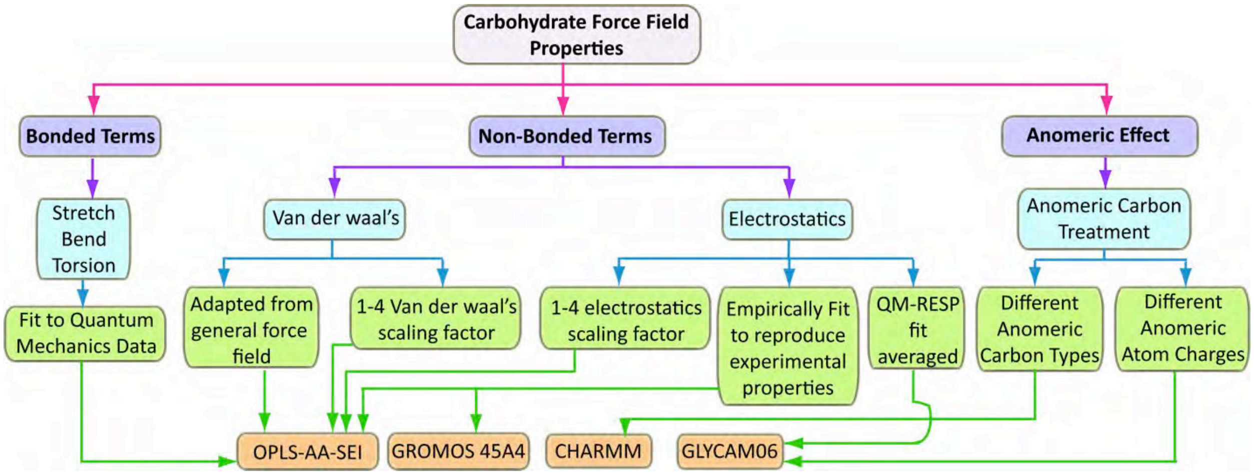

Carbohydrate Force Field Properties
Bonded Terms
Non-Bonded Terms
Anomeric Effect
Stretch Bend Torsion
Van der waal's
Electrostatics
Anomeric Carbon Treatment
Fit to Quantum Mechanics Data
Adapted from general force field
1-4 Van der waal's scaling factor
1-4 electrostatics scaling factor
Empirically Fit to reproduce experimental properties
QM-RESP fit averaged
Different Anomeric Carbon Types
Different Anomeric Atom Charges
OPLS-AA-SEI
GROMOS 45A4
CHARMM
GLYCAM06

Figure 8.3. Parameterization protocol comparison between the carbohydrate force-fields: GLYCAM06; GROMOS 45A4, CHARMM, OPLS-AA-SEI.

To simulate the behaviour of carbohydrates *in vacuo* or in solution (e.g., to study ring puckering [31] or rotational barriers of oligosaccharides), either established force fields or special parameterizations may be used [32-36]. Such force fields allow investigation and prediction of the deformation of carbohydrate rings. These special force fields (as well as previously established ones) have been employed repeatedly for molecular dynamics simulations (MD) of protein-carbohydrate complexes [37, 38]. In some cases, the simulations were successfully used for estimating binding free energies [39-42].

Despite the many possible advantages of established force fields, they were not designed to predict binding free energies or enthalpies in protein-ligand docking. Since solvent molecules are usually modeled explicitly, force fields do not need to include extra terms for hydrophobic effects. The special CH/π interactions are not taken into account [43, 44].

Some force fields do model hydrogen bonds explicitly, while others regard it as part of the electrostatic interaction. Irrespective of the approach, displacement of water molecules competing for hydrogen bonds is not accounted for.

Some force fields correlate well with *ab initio* calculations for *ab initio* optimized geometries [45]. A recent comparison of the results of *ab initio* and force field calculations underlines the difficulties in predicting binding enthalpies in protein-carbohydrate complexes using existing force fields; for example the stabilizing interaction energy for the interaction between fucose and tryptophan is heavily overestimated by the AMBER*[1] force field [46].

GLYCAM06 is a widely used force field for modeling carbohydrates, glycoproteins, glycolipids, as well as for protein-carbohydrate complexes [34, 47]. It can be used for describing the physico-chemical properties of complex carbohydrate derivatives and it is fully compatible with the AMBER force field. Parameters have been developed taking into account a test set of 100 molecules from the chemical families of hydrocarbons, alcohols, ethers, amides, esters, carboxylates, molecules of mixed functional groups as well as simple ring systems related to cyclic carbohydrates and fit to quantum mechanical data. GLYCAM06 may be used in simulation package other than AMBER through the employment of appropriate file conversion tools.

To facilitate the parameter transferability, all atomic sequences have an explicitly defined set of torsion terms, with no generic terms, and PARM94 parameters, the same used in AMBER, are used for modeling the carbohydrate van der Waals terms [48]. No scaling factors for treating 1-4 interactions are introduced for reproducing the *gauche* effect on ω angle rotamers [28].

In GLYCAM06, the stereoelectronic effects that influence bond and angle variations at the anomeric carbon atom are included in a unique anomeric atom type. This feature permits to mimic the ring flipping observed in glycosidic monomers that occur, for example, during catalytic events [49]. Comparison with experimental data confirmed that the force field is able to reproduce rotational energies and carbohydrate features quite well if combined with an appropriate charge set, except for highly polar molecules for which empirical terms have been introduced to correct energetic torsion errors [34]. The atomic partial charges are calculated

[*1] As implemented in the Maestro program (1995 version)

residue by residue. For each residue, 50 to 100 ns MD simulation is performed, 100-200 snapshots are extracted and charges are calculated by fitting to the averaging quantum mechanics molecular electrostatic potential (ESP). This strategy is adopted for incorporating the dependence of molecular conformations on partial charges. Restraints are employed in the ESP fitting procedure (RESP) to ensure that the charges on all aliphatic hydrogen atoms are zero since C-H aliphatic hydrogen atoms are not significant for reproducing dipole moments [50, 51]. An optimal RESP charge restraint weight of 0.01 is applied, based on simulations of carbohydrate crystal lattices [52].

GROMOS-53A6 (CARBO), CHARMM and OPLS-AA are alternative carbohydrate force fields used, together with GLYCAM06, to describe conformational carbohydrate properties in computational chemistry. The GROMOS force field was earlier developed for MD simulations of proteins, nucleotides, or sugars in aqueous or apolar solutions or in crystalline form but it has been modified to include the anomeric effects for mono- and oligo-pyranoses [53, 54]. As in GLYCAM06, quantum mechanics methods are used for calculating bond and angle force constants whereas dihedral parameters derivation and van der Waals terms are directly taken from previous GROMOS versions [55, 56]. An ESP fitting procedure, with restraints on aliphatic hydrogen atoms and averaging over atom types, is chosen for reproducing the electrostatic potential, using a trisaccharide as a model for charge development [54]. No distinction is done between α and β monomers in terms of charges and anomeric atom type and electrostatic - van der Waals 1-4 scaling factors are not introduced so as to correctly reproduce the *gauche* effects on ω angles. Twenty nanosecond long MD simulation in explicit water [57] was used for validating the force field, showing the capability to correctly predict the stereo-electronic effects and the most stable ring conformations but sometimes failing to reproduce their correct energies. GROMOS was proposed as the more adapted force field to mimic the transition from $^{4}C_{1}$ to *skew boat conformations* of the iduronic acid residues in heparin MD simulations [58].

The CHARMM force field was extended to glucopyranose and its diastereomers [59]. Several revisions for carbohydrates have been proposed in order to extend this force field to five member sugar rings and oligosaccharides [60, 61]. The same hierarchical parameterization procedure and treatment of 1-4 non-bonded interactions are used to ensure a full compatibility with other

CHARMM biomolecular force fields [62-64]. Preliminary parameter sets are created using small-molecule models corresponding to fragments of pyranose rings and then successively applied to complete pyranose monosaccharide structures. Missing dihedral parameters are developed by fitting over 1800 quantum mechanical hexopyranose conformational energies. Both partial atomic charges and Lennard-Jones parameter values, taken from previous CHARMM versions, are adjusted to reproduce scaled quantum mechanical carbohydrate-water interaction energies and distances, and further refined to reproduce experimental heats of vaporization and molecular volumes for liquids. The force field, with different atom types for α and β anomers, was validated as it reproduces calculated quantum mechanical and experimental properties using MD simulations with TIP3P water models.

The OPLS force field has been expanded to include carbohydrates [65]. In OPLS-AASEI (Scaling Electrostatic Interactions) force field, 1-4, 1-5 and 1-6 scaling factors are introduced to improve the prediction of Φ/Ψ conformations properties, as well as anomeric effects and relative energies [65]. Unique charge sets and atom types for α and β anomers are used. All non-bonded parameters are imported directly from the parent force field OPLS-AA [66]. Charges are derived, as done for previous force field versions [66, 67], from standard alcohols and acetals to simply reproduce consistent energetic properties, and then transferred to carbohydrates.

Other force fields are employed to understand carbohydrate properties *in silico*. In particular, MM3, a force field initially meant for hydrocarbons, is applicable to a wide range of compounds. The MM3 force field for amides, polypeptides and proteins [68, 69] is widely used for the construction of adiabatic maps of disaccharides. TRIPOS molecular mechanics force field is designed to simulate both peptides and small organic molecules [70] but parameter extension for oligosaccharides includes sulfated glycosaminoglycan fragments and glycopeptides carbohydrate interactions [71, 72]. The TRIPOS force field is implemented in the molecular package SYBYL [73] and commonly used for geometry optimizations.

8.5 Computational tools for docking carbohydrates on proteins

8.5.1 Molecular docking

Molecular docking is a computational procedure that aims at predicting the preferred orientation of a ligand to its target protein, when bound to each other to form a stable complex [74]. In order to perform computational protein–ligand docking calculations, the 3D structure of the target protein must be known. Each docking program operates slightly differently, but they share common features that enable them to: (i) search for locations on the protein surface that lead to favorable interactions with the ligand, (ii) sample the conformational space of the ligand, (iii) compute the interaction energy (or score the binding) between the protein and ligand. The interaction with the ligand relies on both the protein backbone fold in the region of the binding site, as well as on the orientation of the side chains in the binding site. One of the most significant limitations in docking is that it is generally performed while keeping the protein surface rigid, which prevents the consideration of the effects of induced fit within the binding site.

8.5.1.1 *Difficulties in molecular docking*

These difficulties are mostly due to the high number of degrees of freedom characterizing a protein-ligand system that increase the computational cost of the calculations. Thus, several approximations about the flexible states may be introduced in molecular docking experiments. The simplest approximation (rigid docking) considers only the three translational and three rotational degrees of freedom of the protein and of the ligand, treating them as two distinct rigid bodies. The most widely used algorithms at present enable the ligand to fully explore its conformational degree of freedom in a rigid-body receptor [75-79].

8.5.1.2 Docking algorithms

The docking algorithms can be grouped into deterministic and stochastic approaches. Deterministic algorithms are reproducible whereas stochastic algorithms include random factors that do not allow the full reproducibility. The following describes the most widely used algorithms in docking simulations.

8.5.1.2.1 Incremental construction algorithms

These algorithms consist of the division of a ligand into rigid fragments. One of the fragments is selected and placed in the protein binding site. The reconstruction of the ligand is then carried out *in situ*, adding the remaining ligand fragments. For example, DOCK [80] uses incremental construction algorithm to treat ligand flexibility. It generates points (sphere centers) that fill the binding site and try to capture the binding site shape properties for identifying favorable regions in which the ligand atoms may be located. The ligand is divided along each flexible bond to generate rigid segments. An anchor fragment is then selected from all the rigid pieces and oriented in the active site by matching ligand atoms with sphere centers. Fragments are then added and all possible placements are scored on the basis of their interactions with the protein using the energetic scoring function. Best anchor fragments are used for completing the construction of the ligand in the protein binding site. The best scored poses of the complete ligand are selected.

8.5.1.2.2 Genetic algorithms

Genetic algorithms are stochastic searching approaches that use techniques inspired by evolutionary biology to find reliable results. It mimics the process of evolution by manipulating a collection of data structures called chromosomes.

AutoDock [81] uses this algorithm for obtaining reliable docking results. First, the protein is placed inside a cube with a predefined size, characterized by a defined number of points (grid points). In the second step, probes corresponding to the different atom types of the ligand are then moved through the cube and, in particular, at each point, protein-probe interaction energies are calculated and stored in affinity maps. Thirdly, a conformational search of the ligand is performed applying the Lamarkian genetic algorithm. Its characteristic is that environmental adaptations of an individual’s phenotype can become heritable traits, transferred to its genotype. At this stage, a minimization or local search is performed and the results are taken into account modifying the initial conformation that will enter in a new iteration of crossover and mutation of the genetic algorithm cycle.

8.5.1.2.3 Hierarchical algorithms

The algorithm used in GLIDE [82] can be defined as a hierarchical algorithm. It uses an exhaustive systematic search for discovering the most favored ligand conformations in the protein active site, with a screening based on progressively restricted energetic cut-offs. Fields containing information of the protein receptor properties are calculated before the algorithm search. Then a set of initial ligand conformations is produced. Initial screens are performed over the whole phase space available to the ligand to locate promising ligand poses in the respective receptor fields. Ligands are minimized in the field of the receptor using a standard molecular mechanics energy function [66]. Finally, the lowest-energy poses are subjected to a Monte Carlo procedure that examines torsional minima. A composite scoring function is then used to select the correct docked poses.

A variety of other sampling methods like simulated annealing have been implemented in docking programs.

8.5.1.3 Scoring functions

Energy scoring functions are necessary to evaluate the free energy of binding between proteins and ligands. The equation below is the Gibbs-Helmholtz equation that describes the ligand-receptor affinity:

$$\Delta G = \Delta H - T\Delta S$$

ΔG gives the free energy of binding that is the measure of energetic changes between two states represented by the bound and unbound state of the receptor and the ligand. ΔH is the enthalpy, T the temperature expressed in Kelvin and ΔS is the entropy of the system. ΔG is related to the binding constant K_a by the equation:

$$\Delta G = -RT \ln K_a$$

(where R is the gas constant.)

Some sophisticated techniques for predicting binding free energies are currently too slow to be used in molecular docking of large sets of compounds. Thus, fast scoring functions have been developed. Empirical scoring functions use a set of parameterized terms describing properties known to be important in protein-ligand binding to construct an equation for predicting affinities. Multi-linear regression is used to optimize these terms using a set of known protein–ligand complexes. These terms usually describe polar-apolar interactions, loss of ligand flexibility (entropy) and desolvation effects. The Glide Score 2.5 [82] is a regression-based scoring function:

$$\Delta G = C_{lipo} \sum f(r_{lr}) + C_{Hbond} \sum g(\Delta r)\, h(\Delta\alpha) + C_{metal} \sum f\,(r_{lm}) + C_{polar\text{-}phob} V_{polar\text{-}phob} + C_{coul} E_{coul} +$$

$$C_{vdw}E_{vdw} + Solvation\ term$$

The first term describes the lipophilic and aromatic interactions whereas the polar terms are included in the second (hydrogen bonds, separated into differently weighted components that depend on the electrostatic properties of donor and acceptor atoms) and third (ionic interactions) terms. The fourth term rewards instances in which a polar but non-hydrogen bonding atom is found in a hydrophobic region. Coulomb and van der Waals interaction energies between the ligand and the receptor are evaluated as well as the solvation effect.

Force field based scoring functions (AutoDock, DOCK) are based on the non-bonded terms of the classical molecular mechanics force fields. A Lennard-Jones potential describes van der Waals interactions whereas the Coulomb energies describe the electrostatic interactions. In AutoDock (Morris *et al.* 1998), the implemented scoring function has the following form:

$$\Delta G = \Delta G_{vdw}\, \Sigma_{i,j}\, [(A_{ij}/r^{12}{}_{ij}) - B_{ij}/r^{6}{}_{ij})] + \Delta G_{Hbond} + \Sigma_{i,j}\, E(t)\, [(C_{ij}/r^{12}{}_{ij}) - D_{ij}/r^{10}{}_{ij})] + \Delta G_{elec}\, \Sigma_{i,j}$$

$$q_1 q_2/\varepsilon(r_{ij})^2 + \Delta G_{tor} N_{tor} + \Delta G_{sol}\, \Sigma_{i,j}\, (S_i V_j + S_j V_i)^{e(-r2ij/2\sigma 2)}$$

where the five ΔG terms are coefficients empirically determined using linear regression analysis from a set of protein ligand complexes with known binding constants. The summations are performed over all pairs of ligand atoms, *i*, and protein atoms, *j*. The first three terms describe the Lennard-Jones dispersion, the directional hydrogen bonds and the Coulomb electrostatic potential taken from the AMBER force field [48]. ΔG_{tor} is an empirical measure of the unfavorable entropy of ligand binding due to the restriction of conformational degrees of freedom whereas N_{tor} is the number of ligand rotatable bonds. In the fifth term, for each atom type within the ligand, fragmental volumes of the surrounding protein atoms V are weighted by an exponential function and summed, evaluating the percentage of volume around the ligand atom that is occupied by protein atoms. This percentage is then weighted by the atomic solvation parameter S of the ligand atom to give the desolvation energy [81]. Several developed docking approaches use knowledge-based scoring functions based on statistical observations of intermolecular close contacts in protein-ligand X-ray databases, which are used to derive potentials of mean force. This method assumes that the frequency of close intermolecular interactions between certain ligand and protein atoms contributes favorably to the binding affinity. In this approach, no fitting to experimental affinities is required and solvation and entropic terms are treated implicitly [83].

8.5.2 Molecular Dynamics Simulations

In MD simulations (**Figure 8.4**), an ensemble of configurations is generated by applying the laws of motion to the atoms of the molecule. The concept behind MD simulation involves calculating the displacement coordinates in time (trajectory) of a molecular system at a given temperature. Finding positions and velocities of a set of particles as a function of time is done classically by integrating Newton's equation of motion in time. MD simulations are usually carried out as a micro-canonical (constant-NVE) or canonical (constant-NVT) ensemble. As a consequence, all other thermodynamic quantities must be determined by ensemble averaging. In a classical system, Newton's equations of motion conserve energy and thus provide a suitable scheme for calculating a micro-canonical ensemble. However, canonical ensemble can readily be performed by coupling the molecular system to a constant-temperature bath, which rescales the atomic velocities according to the desired temperature. In a similar manner, constant-pressure simulations can be performed by scaling through coupling to a constant-temperature position, as the pressure can be calculated from the virial theorem.

Several algorithms have been developed for MD simulations. Such simulations follow a system for a limited time. Physically observed properties are computed as the appropriate time averages through the collective behavior of individual molecules. For the results to be meaningful, the simulations must be sufficiently long so that the important motions are statistically well sampled. Experimentally accessible spectroscopic and thermodynamic quantities can be computed, compared, and related to microscopic interactions. It should be noted that MD is severely limited by the available computer power. With presently available computers, it is feasible to perform a simulation with several thousand explicit atoms for a total time of up to a few microseconds. To explore the conformational space adequately, it is necessary to perform many such simulations. In addition, it may be possible that carbohydrate molecules undergo dynamical events on longer timescales. These motions cannot be investigated with standard MD techniques. Another way is to use high-temperature dynamics to allow the molecule to assume high-energy conformations. This approach has to be used with caution since it can make the molecules acquire 'non-physiological' conformations.

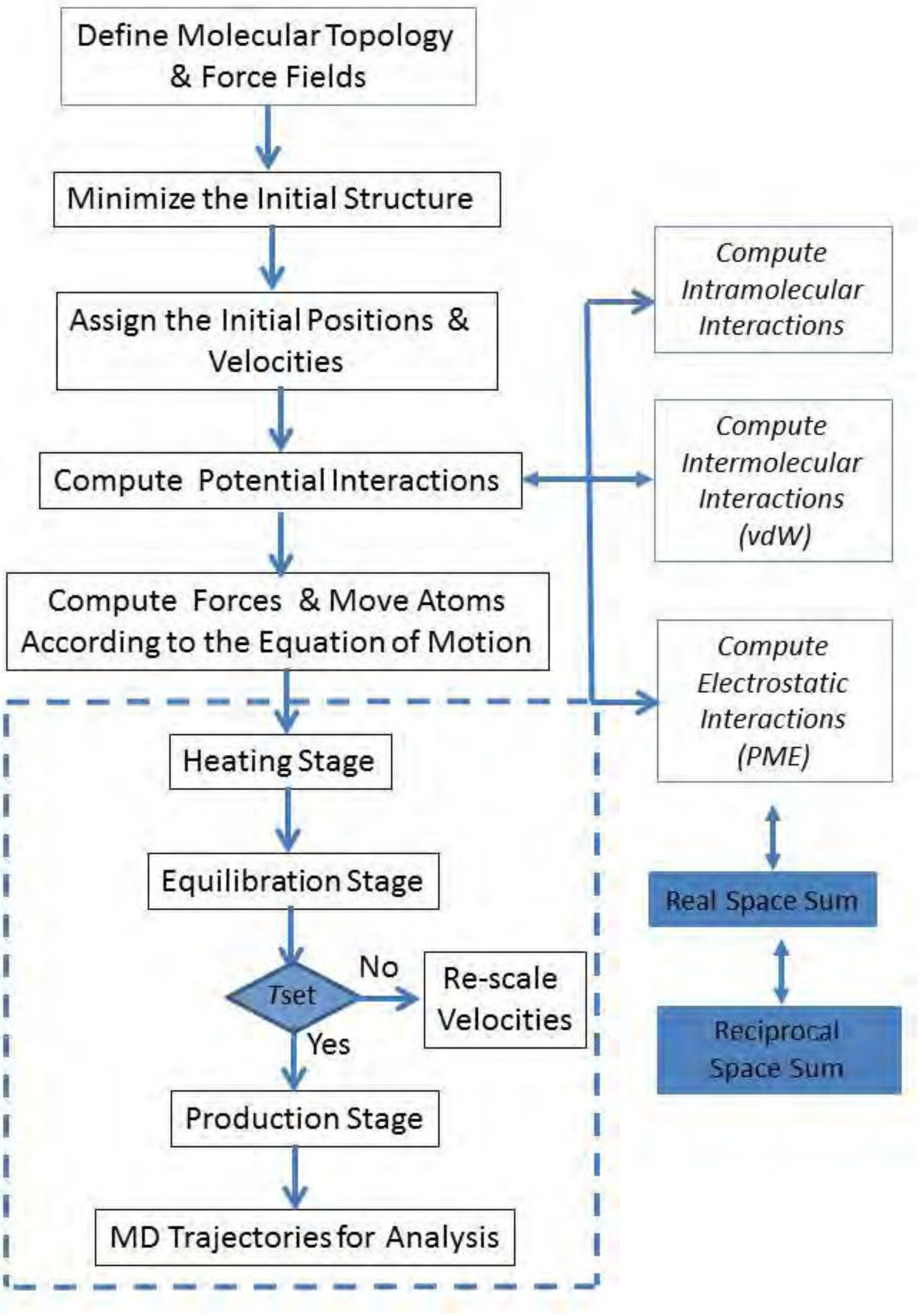

Define Molecular Topology & Force Fields
Minimize the Initial Structure
Assign the Initial Positions & Velocities
Compute Potential Interactions
Compute Forces & Move Atoms According to the Equation of Motion
Heating Stage
Equilibration Stage
Tset
No
Yes
Re-scale Velocities
Production Stage
MD Trajectories for Analysis
Compute Intramolecular Interactions
Compute Intermolecular Interactions (vdW)
Compute Electrostatic Interactions (PME)
Real Space Sum
Reciprocal Space Sum

Figure 8.4. General scheme and the practical sequential approach of the molecular dynamics simulations.

8.5.3 Molecular Robotics

Whereas a number of simulations are performed using MD calculations it must be recognized that they are usually performed on short time scales and have therefore allowed modeling of the dynamic properties of equilibrium states. Indeed, simulations that are needed to capture an entire conformational event, particularly with explicit simulation of solvents are usually too short, relative to the characteristic time of conformational changes occurring upon binding. Consequently, alternative methods such as essential dynamics or normal mode analysis have been successfully applied to selectively enhance conformational sampling along specific directions of motions, and identify large collective motions that may occur in the protein upon binding to a carbohydrate. Novel methods are being developed with the aim of simulating molecular motions that can occur on large spatial and temporal scales.

In addition, to simplify models, methods alternative to MD simulation can be applied to perform an effective exploration of the conformational space. Algorithms originally developed to compute robot motions, have been extended and proposed as alternative methods to compute molecular motions (**Figure 8.5**). Robotics-based algorithms have been applied to the study of several problems such as ligand docking and accessible pathways in flexible receptors, or conformational changes of proteins, due to loop motions, domains motions etc. A methodology named "molecular robotics" has been developed that separates the search for conformational pathways into two stages. The first stage consists of the exploration of geometrically feasible motions, using the robotics-based approach, whereas the second stage uses molecular mechanics for an evaluation of solutions found in the previous stage, while taking into account explicit simulation of solvents. Such a conformational search method handles large molecular motions in a continuous way and within very short computing times. The key advantage of the robotics-based approach is that it enables fast exploration of high-dimensional conformational spaces thanks to the combination of a geometrical treatment of the main molecular constraints, with the performance of path-planning algorithms.

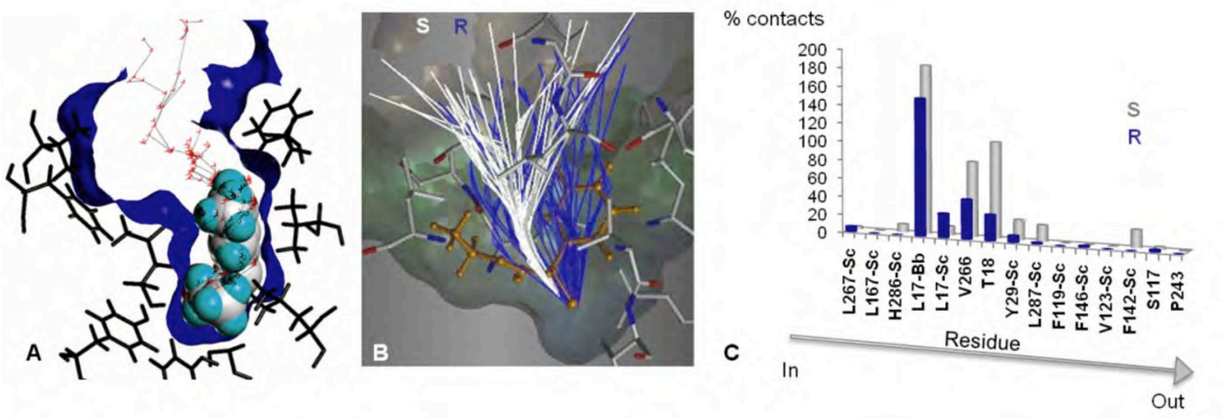


Figure 8.5. Illustration of the molecular robotics approach to investigate the role of substrate accessibility to the active site on Burkholderia cepacia lipase enantioselectivity. Conformational exploration of the active-site pocket using Path-Planing algorithms in order to search exit paths of the ligand from its catalytic position (A). Exit paths computed for the R- and S-enantiomers (50 paths for each enantiomer). The distribution obtained for the R-enantiomer (blue) appears clearly larger and less constrained, than for the S-enantiomer (white) (B). Histogram representing for each enantiomer the relative frequency of interatomic contacts (averaged among the 50 paths) with amino acid residues (C). This automated analysis of ligand-protein contacts enables to highlight amino acid hindering the displacement of enantiomers and thus provide target residues for engineering enantioselectivity [84].

Based on robotics background, computationally efficient methods have been developed in recent years for sampling and exploring conformational space of biological macromolecules. Combined with methods in computational physics such as normal mode analysis [85], or using appropriate multi-scale molecular models [86], robot path-planning algorithms relying on a mechanistic modeling of (macro)molecules are able to compute large-amplitude conformational transitions in proteins with several orders of magnitude faster than standard simulation methods such as MD [84, 87]. These robotics-inspired methods have also been developed to simulate ligand displacement inside an active-site pocket of a protein considering both partners as flexible molecular models with very low computational cost [87-89] and provide information about the interactions between the ligand and the protein and about the required conformational changes that are important for understanding the complex biochemical processes. Such methods have already been successfully applied for rational enzyme engineering [90, 91], showing the efficiency and the potential of molecular robotics methods to guide the engineering of enzyme mutants with improved activity, selectivity and specificity.

8.5.4 Free Energy Calculations

The absolute ligand-receptor interaction energies can be obtained by performing average Molecular Mechanics / Poisson-Boltzmann Surface Area (MM-PBSA) calculations on an ensemble of uncorrelated snapshots in an implicit water environment, collected from an equilibrated MD simulation (**Figure 8.6**). MM-PBSA is a method that approximates the average free energy of binding ΔG between the ligand *L* and the receptor *R* in an implicit aqueous environment as:

$$\Delta G = \Delta G_{RL} - \Delta G_R - \Delta G_L$$

Each term of the above equation is further decomposed as follows:

$$\Delta G_{RL} = \Delta E_{MM} + \Delta G_{PBSA} - T\Delta S_{MM}$$

$$\Delta G_R = \Delta E_{MM} + \Delta G_{PBSA} - T\Delta S_{MM}$$

$$\Delta G_L = \Delta E_{MM} + \Delta G_{PBSA} - T\Delta S_{MM}$$

where, ΔE_{MM} is the average molecular mechanical energy containing the bond angles, torsion angles, van der Waals and electrostatic energetic terms described in the force field. The solvation free energy term ΔG_{PBSA} term contains the electrostatic and non-polar solvent contributions.

$$\Delta\ G_{PBSA} = \Delta G_{PB}{}^{el} + \Delta G_{SA}{}^{np}$$

The Poisson-Boltzmann equation is solved for determining the solvent polar effects $\Delta G_{PB}{}^{el}$ [92] whereas the solvent accessible surface area is used to determine the non-polar energetic term $\Delta G_{SA}{}^{np}$ [93]. Finally, $T\Delta S_{MM}$ represents the entropic term, due to the loss of degrees of freedom upon association. The evaluation of this term represents an issue in computational chemistry, commonly performed by using a quasi-harmonic method or by normal-mode analysis [94]. The high computational cost combined with a very slow convergence and the approximations introduce significant uncertainty in the result [95, 96]. Thus, the entropy contribution can be neglected in case of a comparison of states of similar entropy is desired such as a series of similar ligands binding to the same protein receptor [97].

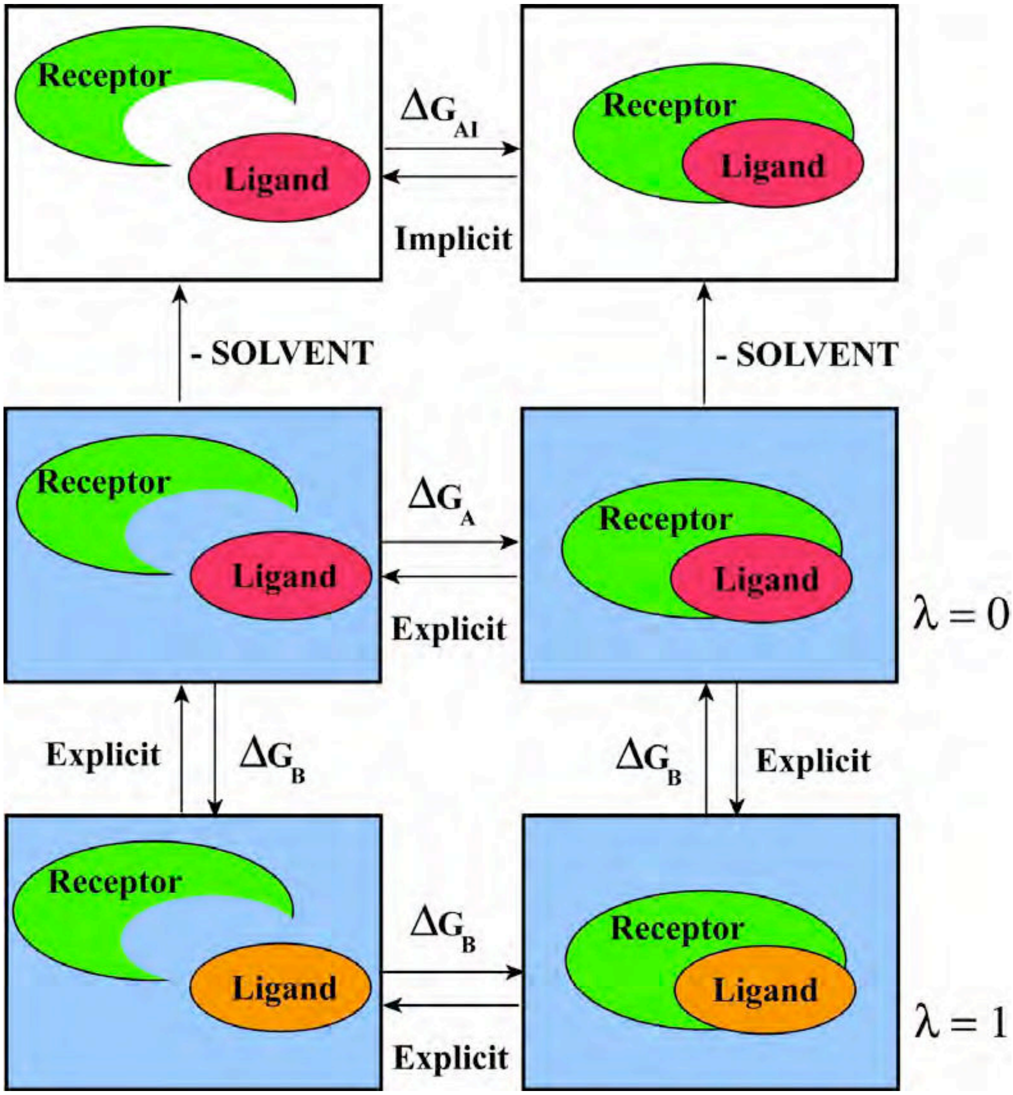


Figure 8.6. MM-PBSA calculations determine the absolute free energy of binding of a ligand to a receptor (ΔG_{AI}) in an implicit solvent environment, whereas Thermodynamics Integration methods calculate the free-energy of binding difference between receptor-ligand complexes ($\Delta\Delta G = \Delta G_C - \Delta G_D = \Delta G_A - \Delta G_B$), where only the ligand is changed.

8.5.4.1 Relative free energy of binding

Thermodynamic Integration (TI) calculations compute the free energy difference between two closely related systems A and B by slowly transforming the initial state A to the final state B. The two states are coupled via a parameter λ that serves as an additional, non-spatial coordinate. This parameter describes the transformation from the reference system A to the target system B and allows the free energy difference between the states to be computed as:

$$\Delta G_{TI} = 1\int 0 <\delta V(\lambda) / \delta(\lambda)>\lambda\ d\lambda$$

In this equation, λ represents the coupling parameter that corresponds to the potential energy V(A) for λ = 0 and V(B) for λ = 1. The integration is carried out over the average of the λ derivative of the coupled potential function at given λ values. Thus, MD simulations in explicit water at different discrete λ points are performed and the value of the integral is calculated numerically. For TI calculations, the system should not undergo significant conformational changes during the transformation, otherwise MD simulations will most likely not sample enough phase space for obtaining converged results [98].

8.6 CASE STUDIES

Avoiding the risk of transforming this section into a catalog, only select examples are provided that deal with the relevant classes of the macromolecules for which a range of the conformational features of protein-carbohydrate interactions have been reported throughout application of computational methods.

8.6.1 Recognition

8.6.1.1 Lectins

Lectins are oligomeric proteins that can specifically recognize carbohydrates, which as per present knowledge act like molecular tools to decipher sugar-encoded messages. They play biologically important roles in recognition processes involved in fertilization, embryogenesis, inflammation, metastasis and parasite-symbiote recognition, from microbes and invertebrates to plants and vertebrates. In the plant kingdom, lectins have been demonstrated to play a role in defense against pathogens or predators and hypothesized to be involved in establishing symbiosis with mushrooms and with bacteria of the *Rhizobia* species. Among the proteins that interact non-covalently with carbohydrates, lectins bind mono- and oligosaccharides reversibly and specifically while displaying no catalytic or immunological activity.

More than 700 crystal structures of lectins have been solved, most of them as complexes with carbohydrate ligands [99]. At present the 3D Lectin Database [100] makes 922 lectin structures available. The wealth of experimental data obtained from the crystallographic studies of oligosaccharides with lectins provided an essential driving force to the development of molecular modeling methods of complex oligosaccharides in their interactions with proteins. These confirmed the flexible conformational behavior of oligosaccharides that was anticipated from earlier calculations. Studies of molecular recognition of the histo-blood group oligosaccharides by lectins paved the way for the conformational analysis of complex carbohydrate-protein interactions, an area that has been thoroughly reviewed [29]. Several docking procedures have been developed, and tested against the experimental data available.

Increasing crystallographic explorations of oligosaccharide–lectin complexes have made significant progress in the characterization of the binding sites of lectins, which are usually rather shallow, located near the surface and thus accessible to solvent. This allows predicting the binding mode of complex carbohydrates to proteins [101]. In several lectin families of different origins, one or two calcium ions are involved in the carbohydrate binding site with direct coordination to the sugar hydroxyl groups. Thanks to the availability of well documented 3D structures of lectins in their native and complexed form, they have been considered as rich playground to develop and test the robustness of docking methods in predicting the binding mode of complex carbohydrates to proteins.

Flexible docking methods of AutoDock, DOCK and Grid-based Ligand Docking with Energetics (GLIDE) were compared for a set of bacterial and animal calcium-dependent lectins and their calcium-dependent sites [102]. DOCK represented crystallographic information well but its lowest energy conformations did not confirm to experimental data for all tested cases. GLIDE results were similar to that of DOCK but the lowest energy poses were always satisfactory that could mimic the real carbohydrate orientation. AutoDock showed reasonable accuracy in sugar orientation and reported the most accurate distances between calcium ions and the sugar hydroxyl groups.

8.6.1.2 Antibodies

The major role of carbohydrates in blood group transfusion and in organ transplants dramatically highlights the importance of carbohydrate-protein interactions as key to major biological processes. The two major histo-blood group carbohydrate determinants [103] are the antigen families, so-called ABH(O) and the Lewis determinants. The majority of the ABO antigens are expressed on human erythrocytes, at the ends of long polylactosaminic chains while a minority of the epitope is expressed on neutral glycosphingolipids. Despite the key role played by these determinants, the description at the molecular level of the interactions occurring between the antigens and the antibodies is only beginning to be resolved and characterized, for instance, crystal structures of Fab against Lewis determinants [104-107]. The exhaustive investigation of the cross-reaction patterns on nine antibodies against 12 carbohydrate antigens has been conducted through computational methods [108, 109]. Three-dimensional descriptors of the molecular properties of the carbohydrate antigens were used in Comparative Molecular Field Analysis (CoMFA). Processing of the QSAR data gave indications on the carbohydrate epitopes essential for antibody recognition while yielding insights into the nature of the molecular recognition.

The successful transplantation of pig organs to human (xenotransplantation) is prevented by the occurrence of carbohydrate antigens on the surface of pig organs which are recognized by xeno-reactive antibodies in the human bloodstream. *In silico* protocol aimed at analyzing the interaction between these xeno-antigens and antibodies interactions has been developed [110] and applied [111] to the determination of the structures of these terminating carbohydrate antigens in complex with a panel of xenoreactive antibodies.

Cell surface complex carbohydrates and polysaccharides are potent targets for recognizing pathogen infections or cancerous cells. As such they offer promising or already successful vaccine components against various pathologies. Consequently, their interactions with antibodies are of a significant interest. The elucidation of the molecular basis of the formation of the complexes but also the balance between the enthalpic and entropic contribution involved in the binding are both required. For the time being, only an appropriate combination of computational and experimental methods may help in establishing these features, in view of developing broad-serotype coverage vaccines.

A majority of life-threatening cases of septicemia, meningitis, and pneumonia occur from the deleterious action of surface capsular polysaccharides of bacteria. Whereas these polysaccharides may have similar carbohydrate sequences, they may markedly differ in immunogenicity, antigenicity, virulence and geographical dispersion, for example the case of Group B

Streptococcus agalactiae and *Streptococcus pneumomia*.

The generation of the antibody complexed with carbohydrate antigens was performed through a combination of comparative antibody modeling and automated ligand docking. Subsequently, several 10 ns molecular dynamic simulations were performed using the Molecular Mechanics-Generalized Born Surface Area (MM-GBSA) method with explicit hydration, augmented by conformational entropy estimates. While providing detailed insight into the molecular details and the energy components involved in the formation of the complexes, the analysis offered a comprehensive interpretation of a large body of biochemical and immunological data related to antibody recognition of bacterial polysaccharides [112].

Shigella flexneri is the main causal agent of the endemic form of bacillary dysentery. The O-antigen is the polysaccharide moiety of the lipopolysaccharide; it is the major target of the serotype-specific protective humoral response elicited upon host infection by *Shigella flexeneri*. The repeating unit of the O-antigen is a pentasaccharide. The availability of the X-ray structure of the Fab/$[AB(E)CD]_2$ complex, at a resolution of 1.80 Å [113], along with a sufficient amount of well characterized pentasaccharides, and IgG monoclonal antibody allowed a thorough analysis of the complexes by Saturation Transfer Difference (STD) NMR experiments and extensive MD simulations (**Figure 8.7**). The study brought into light information on the dynamics of the corresponding antibody-carbohydrate complexes that is available neither from the X-ray structure nor from the NMR analysis independently [114]. The proposed protocol making use of MD simulations and STD-NMR is likely to facilitate the design of either ligands or carbohydrate recognition domains, according to needed improvements of the natural carbohydrate-receptor properties.

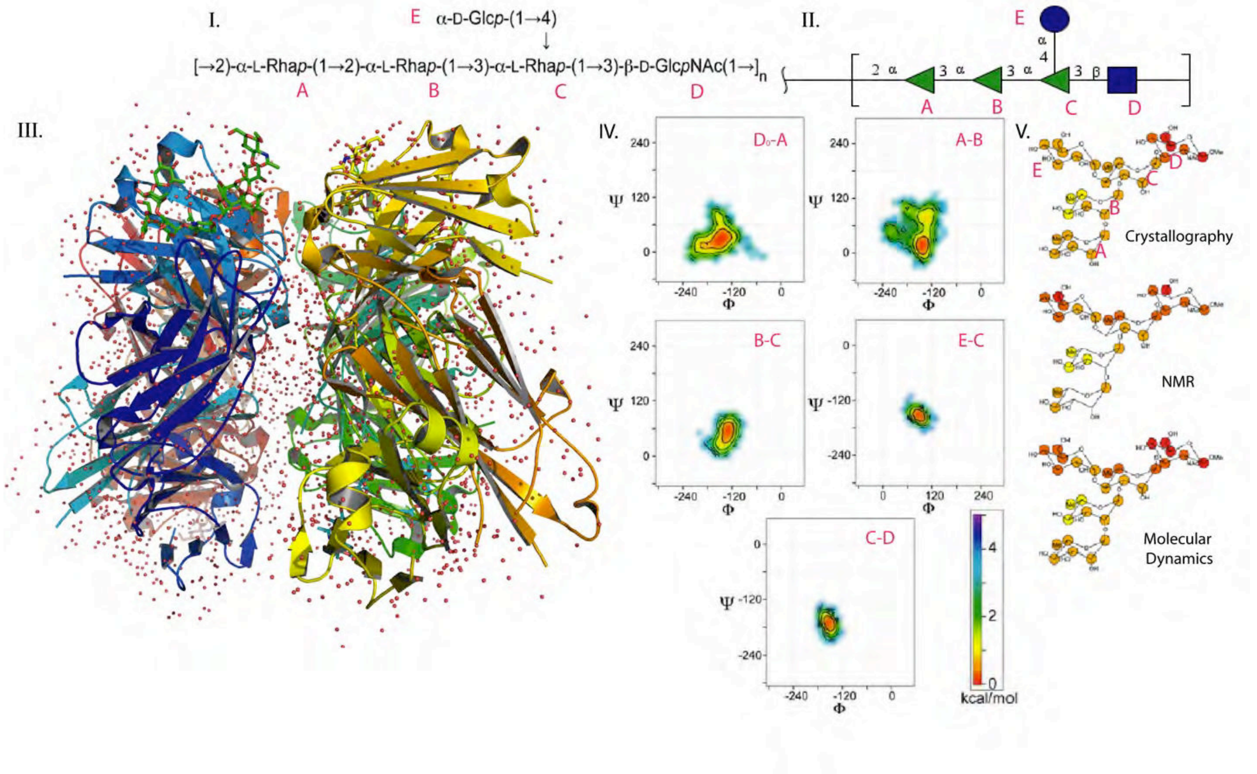


Figure 8.7. Features of *Shigella flexeneri* O-antigen interacting with monoclonal antibody.

(I.) Primary structure of the *Shigella flexneri* SF2a O-Ag [115] common AB(E)CD linear backbone repeat unit.

(II.) CFG representation of Shigella flexneri common AB(E)CD linear backbone repeat unit, where the green triangles represent Rhamnose, the circle denotes Glucose and the blue square denotes *N*-acetylgalatosamine.

(III.) Crystal structure of synthetic O-antigen decasaccharide from serotype 2a *Shigella flexneri* (PDB i.d. 3BZ4) in complex with a protective monoclonal antibody Fab F22-4.

(IV.) ϕ,ψ maps of MD simulations for the glycosidic linkages of 2 repeat units of the bound conformation of the *Shigella flexneri* O-Antigen D_0 AB(E)CD pentasaccharide.

(V.) Comparison between the predicted Saturation Transfer Difference (STD) values of the 2 repeat units of the truncated crystal structure of F22-4 and the measured STD NMR intensities and the predicted values of the 50 MD simulation snapshots of AB(E)CD.

8.6.1.3 Chemokine-Glycosaminoglycan Interactions

The glycosaminoglycans (GAG) comprise a family of complex anionic polysaccharides including: (i) glucosaminoglycans (heparin, heparan sulfate), (ii) galactosylaminoglycans (chondroitin sulfate and dermatan sulfate), (iii) hyaluronic acid and keratan sulfate. In addition to their participation in the physicochemical properties of the extracellular matrix, GAG fragments are specifically recognized by protein receptors and they play a role in the regulation of many processes, such as hemostasis, growth factor control, anticoagulation and cell adhesion [116].

Given the importance of protein-GAG interactions, oligosaccharide fragments are important targets for drug design.

Docking of GAG oligosaccharide in protein receptor binding sites presents two main difficulties: (a) the binding site does not generally adopt a pocket or crevasse shape that would allow for easy identification and (b) both the ligand and the protein presents a high flexibility of side chains. In addition to a simple molecular visualization program, the analysis of the projection of the electrostatic potential on the Connolly surface of the protein, for example with the MOLCAD program (included in SYBYL [73], has been proven to be useful. The GRID program [117], that allows prediction of the most energetically favorable region for binding of small probes on protein surface, is very successful in identifying sulfate-binding regions. For predicting the orientation of the oligosaccharide on the protein surface, the AutoDock program [81] that considers flexibility at glycosidic linkages and pendant groups (hydroxyl groups, hydroxymethyl, etc.) can be used for charged oligosaccharide fragments. It should be kept in mind that such an approach generally yields several families of conformations and that further simulations, including MD in the presence of explicit water and counter ions have to be envisaged for a thorough investigation.

The conformational behavior of the heparin pentasaccharide responsible for high affinity to antithrombin III has been the subject of several investigations. This study is complicated by the fact that a conformational change occurs in the protein upon binding [118, 119]. The first model obtained using homology modeling for the protein and hand-docking of the pentasaccharide, allowed the determination of the basic amino acids involved in the recognition of the sulfate and carboxylate groups [120]. A study making use of several newly developed docking programs, arrived at the same prediction for the binding site [121]. In the crystal structures of the complex between antithrombin III and pentasaccharide [118, 122] a cluster of basic amino acids has been demonstrated to interact with the oligosaccharide's sulfate and carboxylate groups. The conformation of the bound pentasaccharide is also subjected to induced fit upon binding (**Figure 8.8**). At the present time, both X-ray crystallography studies and NMR data coupled with molecular modeling [123] agree that the binding is accompanied by dihedral angle variations of two glycosidic linkages and conformational shift of the 2-O-sulfated iduronic residue.

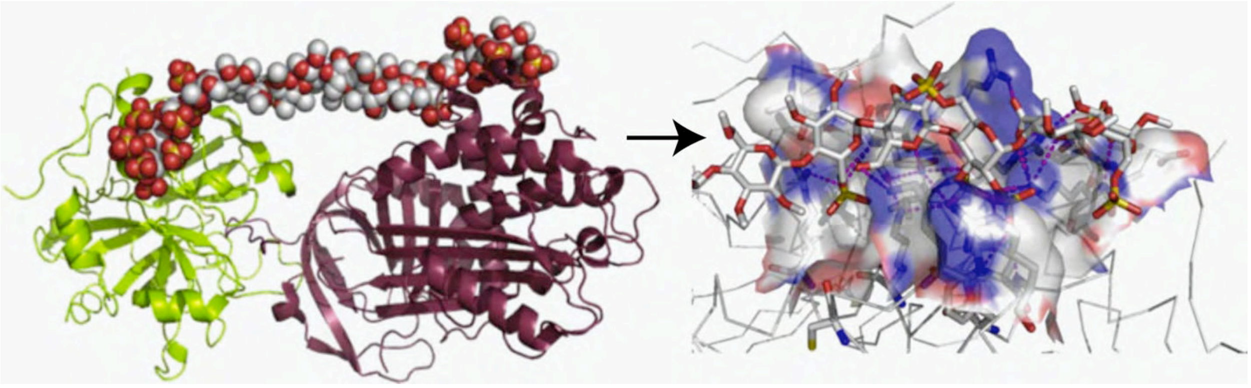


Figure 8.8. General view (*left hand side panel*) of the crystal structure of ternary complexes between antithromin (*reddish-brown ribbon*), thrombin (*green ribbon*), and heparin analog [124, 125]; (*right-hand side panel*) blow-up of the binding site of antithrombin interacting with the specific heparin fragment.

Amongst the numerous proteins that bind heparin, the fibroblast growth factors (FGFs) have received special attention because they are involved in the control of cell proliferation, migration and differentiation. Heparin fragments have been co-crystallized as ternary complexes with two FGF and their receptors (FGFRs) and the minimal binding sequences could be determined [126, 127]. Analysis of the crystal structures together with molecular modeling demonstrated that upon binding, the regular helical shape of heparin is kinked at one point by both modification of one glycosidic linkage conformation and one iduronate ring shape [128]. Such "induced-fit" of the ligand in GAG/protein interactions is very likely to happen since it is classically observed in lectin-oligosaccharide interactions.

Chemokines, derived from chemo-attractant cytokines, represent a large family of small proteins which, based on their physiological features, have been classified as "inflammatory" (or inducible) or "homeostatic" (or constitutive) [129]. Their roles include events as diverse as development, angiogenesis, neuronal patterning, hematopoiesis, viral infection, wound healing and metastasis. Given the importance of protein-GAG interactions, oligosaccharide fragments are important targets for drug design. Chemokines interact with GAGs in general and heparan sulfate in particular. This binding is thought to create a local concentration, or a gradient, of chemokines on tissues where some GAGs are specifically expressed. Modeling studies have therefore been used for describing the interaction between chemokines and heparan sulfate. One interesting structural feature is that chemokines may exist in solution as monomer or dimer (sometimes tetramer) but they bind GAGs in the dimeric or tetrameric state. Depending on the dimerization mode and positions of basic amino acids in the peptide sequences, chemokines will present positively charged clusters on their accessible surfaces that define several possibilities for binding heparan sulfate [130, 131].

8.6.1.4 Transport

Carbohydrates such as the malto-oligosaccharides of lactose, sucrose, raffinose, fructo-oligosaccharides, L-fucose, trehalose, oligo-alginate, oligo galacturonate etc. constitute a source of carbon for many organisms. These molecules have to be transported across channel and pores and their motion is critically important for understanding mechanism of many cellular processes. At the protein level, this is achieved by a family of proteins, collectively referred to as transporters. These trans-membrane proteins allow permeation of sugars: their structures along with the mechanistic transport model are the subject of intense research. The recent high-resolution structural elucidation of transporters is enabling investigation into the MD of fundamental transport processes.

Transport across the membrane is mediated by channel-forming proteins, of which maltoporin has been most extensively studied. The elucidation of the first high-resolution structure of maltoporin [132] revealed the general model of specific channel-forming membrane proteins: a beta-barrel with 18 anti-parallel strands. Like the general diffusion porins, the functional unit of maltoporin is a trimer with long loops exposed to the cell exterior and short turns exposed to the periplasm. A striking feature is a consecutive stretch of aromatic residues in the channel arranged in a left-handed helical path, which has been described as the "greasy slide".

The translocation mechanism of malto-oligosaccharides across the maltoporin membrane channel (**Figure 8.9**) has been investigated by MD calculations [133] (see the movie at http://spider.iwr.uni-heildelberg.de/fischer/research/maltoporin.mpeg). The first event is the binding of sugar to the first residue of the "greasy slide" which occur via van der Waals interactions to the hydrophobic face of the glucosyl ring. Deeper penetration into the channel occurs throughout guided diffusion of the oligosaccharide along the "greasy slide". A gradual dehydration of the malto-oliogosaccharide favours the establishment of transitory hydrogen bonds between the sugars' hydroxyl groups and the surrounding amino-acids. This is made possible by the conformational flexibility occurring at the glycosidic linkages and at the primary hydroxyl groups. The presence of the charged side chains (referred to as "polar tracks") mimics the lost hydration shell to the sugar by providing hydrogen bonds to the hydroxyl groups of the carbohydrate. The polar tracks are divided into donor and acceptor lanes all along the greasy slide. The movement of the carbohydrate residues to the next binding site of the greasy slide in combination with a rearrangement of hydrogen bonds is referred to as the "register shift". The continuous making and breaking of hydrogen bonds results in the oligosaccharide moving through the porin in a capillary-like fashion.

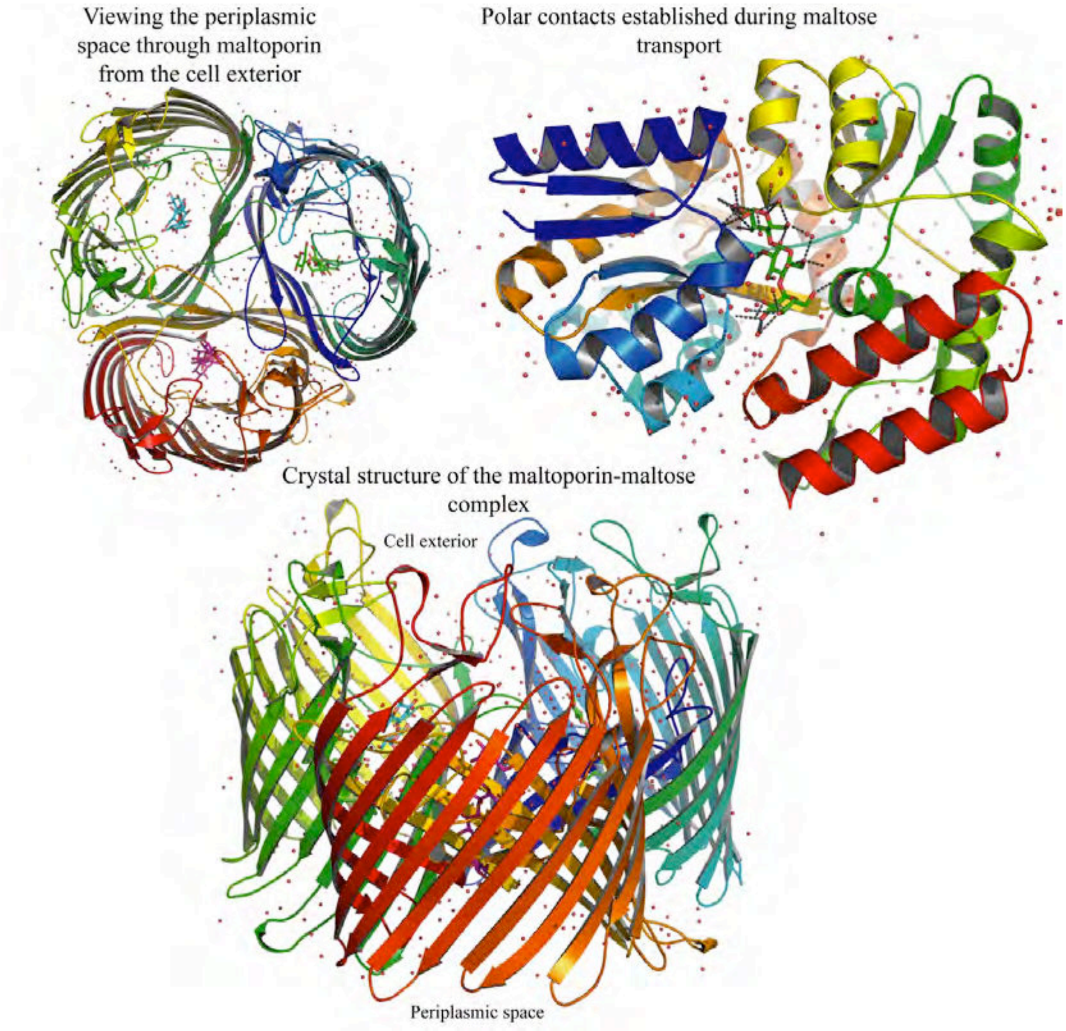

Viewing the periplasmic space through maltoporin from the cell exterior
Polar contacts established during maltose transport
Crystal structure of the maltoporin-maltose complex
Cell exterior
Periplasmic space

Figure 8.9. Three-dimensional structure of maltoporin [134] along with snapshots of the interaction of malto-oligosaccharides within the channel.

Within the super family of carbohydrate transporter exists the MFS family, which exploits the electrochemical potential to shuttle substrates across cell membranes. These transporters are thought to use an alternating-access mechanism to upload and download substrates. The elucidation of the 3D structure of a fucose transporter [135] opened the way to MD simulation of the L-fucose residue complexed to the trans-membrane protein inserted into a POPE bilayer to mimic the bacterial membrane. Structural, biochemical and computational analysis provided insights into the function of the transporter, with the identification of key amino-acids that play an essential role in the active transport path.

As the structures of other unique transport systems are revealed, the power of computational methods in transporter analysis and prediction will grow exponentially.

8.6.2 Synthesis: Glycosyltransferases

The central process of oligosaccharides, polysaccharides and glycoconjugate biosynthesis is performed by the action of glycosyltransferases (GT). These enzymes constitute a large family of proteins which are present in prokaryotes, eukaryotes, and viruses and mediate a wide range of functions from structures and storage to signaling. GTs are responsible for the formation of the glycosidic bond by attaching a sugar moiety of an appropriate donor substrate, mainly a nucleotide sugar, to a specific acceptor substrate. These proteins are highly stereo- and regio-selective and they are usually classified by their preferred sugar substrates, acceptor molecules and the types of glycosidic linkage they generate [136, 137].

Molecular modeling of GTs along with their interactions with the nucleotide sugar and the specific acceptor substrate is difficult. The number of available crystal structures is still limited [136]; only a limited number of folds has been observed. In these crystal structures, the ratio of loops to secondary elements is high, and many of them do not describe the entire catalytic domain as the electron densities are not clear due to the flexible polypeptide extremities and/or several loops. Flexible loops appear to play an important role in substrate binding. For some of these enzymes, structural and calorimetric binding studies indicate that an obligatory ordered binding of donor and acceptor substrates, linked to a donor substrate-induced conformational change, and the direct participation of UDP in acceptor binding, induces a large conformational change. It has been shown that the open state (free enzyme) has no or little affinity for the oligosaccharide acceptor. Alternatively, the closed active conformation creates a pocket that serves as the binding site for the acceptor. Starting from an available crystal structure of a GT, or using such theoretical approaches as fold recognition [138], a 3D model of the GT of interest is first constructed. Further, a combination of methods like fold recognition and molecular modeling might aid the prediction of the acceptor specificities of the putative glycosyltransferase. Docking of substrates also appears to be a difficult task because of the conformational flexibility of the nucleotide sugar and the presence of phosphate and divalent cation. Appropriate energy parameters have been developed based on the AMBER force field interfaced with CICADA for conformational search [139]. Interacting with the protein, the nucleotide sugar(s), the metal ions, and the oligosaccharide acceptor(s) are then submitted to a docking procedure which is followed by energy optimization of the amino acid side chains surrounding the substrates. At present such molecular modeling procedures are aimed at revealing the key catalytic amino acids and the nucleotide-sugar donor specificity and are performed in conjunction with site-specific mutagenesis and biochemical analysis [140]. The availability of a well-resolved crystal structure of glycosyltransferase GT(51), a penicillin-binding protein, has opened the way to investigate how computational methods can be used to explore drug targets for antibiotic resistance. Docking and scoring methodology (Surflex-Dock and FlexX-Pharm) have been applied resulting in the discovery of nine novel potential leads for GT(51) inhibition [141]. Detailed characterization of the mechanisms involved in either the inversion or retention of stereochemistry can only be interpreted with the use of *ab initio* molecular orbital study [142-144].

8.6.3 Glycosyl Hydrolases / Glycosidases

The hydrolysis of glycosidic bonds in carbohydrates, polysaccharides, glycoproteins, glycolipids etc. is performed by glycosidases. These enzymes are classified into endo- and exo-types. Exo-type glycosidases attack and hydrolyze monoglycosides into free sugar and aglycon. When acting on oligo- or polysaccharides, they liberate a monosaccharide unit from the non-reducing end. Endo-type glycosidases act on oligo- and polysaccharides and catalyzes the hydrolysis of an internal glycosidic linkage thereby liberating two carbohydrate moieties, or in releasing an oligosaccharide (or polysaccharide) and monoglycoside of the reducing end. Some glycosidases are capable of acting as both exo- and endo-types. The reactions resulting from the catalytic action of glycosidases can also be characterized by the anomeric configuration of the glycosidic bond of the substrate that the enzyme attacks, i.e. with retention or inversion of the anomeric configuration.

8.6.3.1 Glycosyl hydrolases on single chain

Computational methods are essentially used to dock the oligosaccharide, polysaccharides etc. into the active state (which is usually identified throughout systematic mutations). The glycosidases and their carbohydrate ligands are considered in their energetically stable conformations and their interaction energies are compared. The resulting docked structures are used to propose a model for substrate and conformer selectivity based on the dimensions of the active site. The docking of substrates and inhibitors indicate the dimensions of the binding site which are usually large, extending over several monosaccharide units, beyond and towards the cleaving site. The key amino acids which may be involved in the catalytic mechanism can be identified from these results. Such computational protocols have been applied to the study of several classes of glycosidases. Most recent examples incorporating state-of-the-art modeling tools have been used to investigate the features of heparanase interacting with heparin [145, 146]. These docking ligand-protein complex models can interpret the substrate specificity of heparanase, providing a rationale for the design of polysaccharides that may act as inhibitors of the enzymatic activity of heparanase. Predicted heparin/complexes show that the interactions of the heparin binding domains in combination with the catalytic domain can be targeted for the design of inhibitors.

Enzyme inhibitors can be classified into substrate analogs and transition state analogs. Both types of analogs inhibit the enzyme via generally competing with the substrate for binding to the active site of the enzymes but are not affected by the enzyme. Substrate analogs mimic the structural features of the substrates, whereas transition state analogs have some structural characteristics that are unique to the transition state.

Structural analysis of influenza virus neuraminidase [147] and neuraminidase in complex with sialic acid [148] led to the design of a potent inhibitor of neuraminidase activity: zanamivir [149]. Based on the efficacy of zanamivir (Relenza^{TM}), another neuraminidase inhibitor was also developed: oseltamivir phosphate (Tamiflu^{TM}) [150]. Both, Relenza, a carbohydrate-based drug, and Tamiflu, a carbocyclic mimetic, are potent and clinically effective anti-influenza drugs [151]. Despite the efficacy of these drugs, major concerns remain regarding the development of resistance to these drugs, which is already occurring. Point mutations in the influenza virus enzyme neuraminidase have been reported that lead to dramatic loss of activity for known neuraminidase inhibitors cited above. A more sound understanding of the molecular basis of such resistance is needed toward developing improved next-generation drugs. Modeling the binding of ligands with neuraminidase has been undertaken using explicit solvent all-atom MD simulations, free energy calculations and residue-based decomposition. The simulations predicted the effects of a known mutation at one amino acid (R292K) and provided clues as to the origins of resistance to the mutant. The results significantly enhance experimental observations [152].

The likelihood of future influenza pandemics (including the possibility of highly pathogenic H5N1 strains), highlighted the need for additional computational methods. The binding properties of the H5N1 influenza virus neuraminidase have been inferred from molecular modeling [153]. They concerned the binding properties between sialic acid, methyl 3’-sialyl lactoside, methyl 6’-sialyllactoside and H5N1 influenza virus neuraminidase using molecular docking and MD simulations. The obtained results indicate that, in the complex, sialic acid undergoes a conformational transition of the ring. Meanwhile, methyl 3’-sialyl lactoside establishes only weak interactions with a key loop of the neuraminidase, in contrast to what is observed for the complex with methyl 6’-sialyl lactoside. The differences could be attributed to the occurrence of distinct conformations about the glycosidic linkages. As these molecular modeling results are consistent with available experimental data on the specificity of neuraminidase, they provide sound structural information for a rational design of novel and specific inhibitos of H5N1 neuraminidase as potential therapeutics for the treatment of avian flu.

8.6.3.2 Glycosyl hydrolases on a solid substrate

Many polysaccharides occur in the form of highly packed 3D arrangements as a result of extensive inter- and intra-molecular hydrogen bonding networks and van der Waals interactions.

These features render the structures completely insoluble in water (e.g. cellulose, chitin) and provide them with substantial resistance from attack by most enzymes. The hydrolysis of cellulose in Nature is the result of plant cell wall degrading complexes, referred to as cellulosomes [154] including cellulases. Cellulases consist of a core of glycoside hydrolases and cellulose-binding modules (also referred to as Carbohydrate-Binding Modules or CBMs) and a linker that binds the two enzymatic components. By playing the dual role of recognizing and adhering to the solid state surface of the polysaccharide, and maintaining the proximity effect, the presence of CBMs is a key factor in the ability of the enzyme to efficiently breakdown insoluble polysaccharides. Once bound to the crystalline substrate, the active center of the core domains of cellulases can attack the cellulose chains. The cellulases are classified into two types: the exo- and endo-cellulases, depending on whether or not the cellulose can recognize the reducing end of the cellulose chains. The morphology of the native crystals is of course an essential feature in the enzymatic digestion of crystalline cellulose, and this is a major scientific and industrial question [155]. It is established that the enzymatic breakdown and degradation of cellulose requires a complex of enzymes working together (**Figure 8.10**). The general picture that has emerged from earlier investigations in this area indicates that the cooperation of at least three types of enzymes is required for efficient digestion of crystalline cellulose into glucose. These are (i) endoglucanases (EC 3.2.1.4), which cleave the chains randomly, (ii) cellobiohydrolases (EC 3.2.1.91), which recurrently cleave cellobiose from the chain-end of cellulose, and (iii) β-glucosidases (EC 3.2.2.21), which hydrolyze cellobiose. As for the cellulose-binding modules, numerous studies have established that three aromatic residues are needed for binding onto cellulose crystals, and that tryptophan residues contribute to higher binding affinity than tyrosines. However, evidence has accumulated showing that different binding sites for the same cellulose-binding domains could occur.

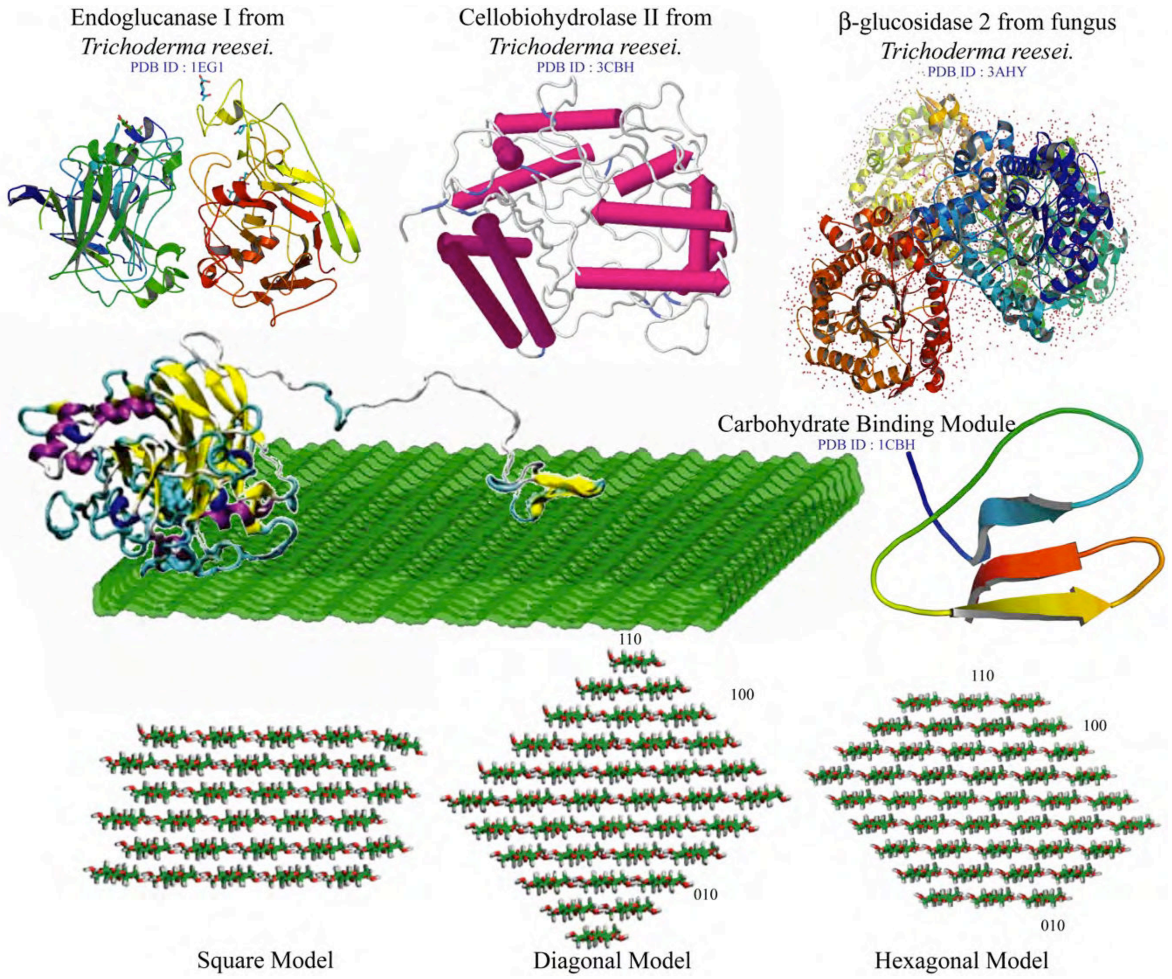


Figure 8.10. The enzymatic digestion of cellulose. The top panel depicts the three dimensional structures of the three main categories of enzymes (from *Trichoderma reesei*) that digest crystalline cellulose. The central panel provides a visual as to how cellulose digesting enzyme (shown here as cellobiohydrolase I or Cel7A [156]) interacts with the cellulose crystalline arrangement and breaks down cellulose into glucose and also illustrates a carbohydrate binding module (CBM; PDB i.d. 1CBH). The bottom panel shows the three faces of the cellulose Iα crystal models in projection with the Miller indices of their constituent crystal planes.

A systematic study of the carbohydrate binding module (CBM) protein of Cel7A of *Trichoderma reesei*, with the cellulose Iα crystal model has been performed using a combined Grid docking search and MD calculations [157]. Three types of cellulose Iα crystal models with infinite dimensions were constructed, each consisting of different crystallographic faces, i.e. (1 1 0), (1 0 0) and (0 1 0). The (1 1 0) complex models exhibited larger affinities at the interface than the other ones. It was found that the CBM was more stably bound to the (1 1 0) surface when it was placed in an anti-parallel orientation with respect to the cellulose fiber axis. The predicted directional specificity of the CBM at the optimum positions was consistent with the observed processing direction of the Cel7A [158]. In the solvated dynamic state, the curved (1 1 0) surface resulting from the fiber twist somewhat assisted a complementary fit with the CBM at the interface.

Much can be learned about the processivity by conducting carefully designed MD simulation of the binding of the catalytic domains of cellulases with various substrate configurations, solvation models and thermodynamics protocols [159].

Computational model of Cellobiohydrolase I (Cel7A) from *Trichoderma reesei* on a cellulose (1 0 0) surface displaying the large catalytic domain (left), linker (middle single strand), and cellulose binding module (right small domain). A cellodextrin strand is shown peeled out of the surface of the cellulose and threaded into the catalytic tunnel of Cel7A [156]. The investigation requires the consideration of approximately 800,000 atoms. In order to face such a computational challenge, most of the numerical simulations shall require major modifications of existing code and algorithms.

8.7 CONCLUSION

In the past few years, there has been an increase in the development and application of computational methods aimed at establishing the molecular features characterizing the protein-carbohydrate interactions. Quite naturally, these computational methods are becoming reliant on experimental studies for the elucidation of structural and dynamics feature in the field of glycoscience. Significant steps have been made among which the developments and implementations of force fields capable of taking into account the specificity of carbohydrates (stereo-electronic effect, gauche effect etc.) and their compatibility with the computational tools that have been developed for proteins. Recently, methods for handling many rotatable bonds in flexible docking of conformationally flexible carbohydrates have been established. It has been recognized that the surface of carbohydrates and their derivatives that are composed of hydrophobic and hydrophilic patches remain a source of complexity in modeling. Nevertheless, the balance between hydrophobic and hydrophilic patches is essential for carbohydrate solubility and for molecular recognition. The occurrence of such a feature combined with the enhanced conformational flexibility is a unique characteristic that explains how complex oligosaccharides can be transported throughout trans-membrane proteins in a capillary-like and yet selective fashion.

Calculations of binding free energies and enthalpies with the required accuracy remain to be improved and tested against well-characterized experimental data. Certainly, calculation of free energy perturbations is a promising approach for the prediction of carbohydrate-receptor binding affinity. Such calculations cannot be performed without a full understanding of solvation. Progresses in this area imply a better handling of hydration and the major role played by solvation and desolvation of both carbohydrates and proteins in their isolated state and during the course of their interactions.

At present these computational tools are considered as useful as the other methods of structural investigation. They can actually help in reconciling the experimental results gathered from separate experiments in different conditions and environments and in extrapolating the results. The wealth of successful applications for many different protein interactions with carbohydrates is a testimony to the maturity of the modeling methods and protocols that have been developed. Nevertheless, these success cases are almost exclusively dealing with cases where proteins interact with carbohydrates, without any further catalytic actions.

Complementary computational methods need to be developed and/or integrated to allow the study of enzymatic reaction and the subsequent optimization of bio-catalyzers. These methods, based on molecular robotics algorithms, would be used for an efficient virtual screening of configurational and conformational spaces of high dimensions. The on-going developments of robotics algorithms are likely to provide efficient tools to explore the dynamic functionality of enzymes. These are based on efficient path-planning algorithms and fast geometric operators designed for complex articulated chains. The aim is to reduce, in a significant but relevant way, the exploration of the combinatorial space of the enzyme sequences, based on the geometric feasibility for a ligand either to access or to leave the catalytic site in a “productive” way. The enhancement of the predictive performances of such algorithms will require the use of simplified energy functions to pre-filter the conformations that are non-viable to construct the network of concerted motions while the enzyme is interacting with the ligand.

The investigation of the catalytic mechanism of inverting and retaining carbohydrate active enzymes requires high level density functional theory (DFT), hybrid quantum mechanical and molecular mechanical (QM/MM) calculations. The studies of the catalytic reaction and the dynamic motions undergone by the enzymes are being investigated independently at present. Consequently, developments are required to set up MD 'hybrid methods' based on the principles of quantum mechanics with the aim of studying the dynamics of electronic effects and charge transfer within the catalytic site. Such hybrid methods would incorporate *ab initio* dynamics as developed by Carr and Parinello (CPMD) and a 'classical' MD force field. Applications of these computational methods will allow exploiting further the protein-carbohydrate interactions, especially for therapeutic purposes. Design of transition state analog inhibitors of glycosyl hydrolases and glycosyl transferases requires knowledge of the mechanism of the enzymatic reaction along with the geometry and charge distribution of transition state.

Extremely challenging cases are being identified. Many of the carbohydrates with biological functions are found at the surfaces of proteins and cells. Some physico-chemical principles that underline their associations are being considered to model such systems, for example patches of glycolipids and glyco-surfaces. As the concept of "glyco-landscape" is being recognized, new modeling protocols need to be developed. They require novel computational tools capable of constructing the landscape resulting from the side-by-side arrangements of glycoconjugates such as glycolipids. A new paradigm will emerge, and the attention will not only be given to the interaction of a protein with a single carbohydrate unit (the so-called "tree vision") but instead the interaction with glyco-surfaces (the so-called "glyco-canopy", as an analogy to the crown canopy, i.e. the uppermost layer in a forest formed by the crown of the trees).

This concept is likely to become more prevalent as the field of research dealing with the solid state degradation of plant cell walls polysaccharides by enzymes offers formidable challenges. Plant biomass is an alternative natural source for chemical and feed stocks with a replacement cycle short enough to meet the demand of the world fuel market. The enzymatic hydrolysis of cellulose is still considered as a main limiting step of the biological production of biofuels from lignocellulosic biomass. This step involves the action of three types of cellulose degrading enzymes acting in a synergistic way. In view of designing a functional kinetic model integrating the respective properties of each enzyme along with their synergies, much can be learned by conducting carefully designed computer simulations of the binding of the cellulose-binding-domains and the catalytic domains of cellulases with various substrates, solvation models and thermodynamics protocols. Such an extraordinary computational challenge is delineating the new frontiers of the area of protein-carbohydrate interactions.

ACKNOWLEDGEMENT

The research leading to this Chapter has received funding from the European Commission's Seventh Framework Programme FP7/2007-2013 under grant agreement n° 215536. We are grateful to the Marie Curie Initial Training Network as part of the FP7 People Programme for training and funding.

References:

1. Parthasarathy S, Ravindra G, Balaram H, Balaram P, Murthy MR: **Structure of the *Plasmodium falciparum* triosephosphate isomerase-phosphoglycolate complex in two crystal forms: characterization of catalytic loop open and closed conformations in the ligand-bound state**. *Biochemistry* 2002, **41**(44):13178-13188.

2. Jansma AL, Kirkpatrick JP, Hsu AR, Handel TM, Nietlispach D: **NMR analysis of the structure, dynamics, and unique oligomerization properties of the chemokine CCL27**. *J Biol Chem* 2010, **285**(19):14424-14437.

3. Collins PJ, Haire LF, Lin YP, Liu J, Russell RJ, Walker PA, Skehel JJ, Martin SR, Hay AJ, Gamblin SJ: **Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants**. *Nature* 2008, **453**(7199):1258-1261.

4. Parsiegla G, Reverbel-Leroy C, Tardif C, Belaich JP, Driguez H, Haser R: **Crystal structures of the cellulase Cel48F in complex with inhibitors and substrates give insights into its processive action**. *Biochemistry* 2000, **39**(37):11238-11246.

5. Brun E, Moriaud F, Gans P, Blackledge MJ, Barras F, Marion D: **Solution structure of the cellulose-binding domain of the endoglucanase Z secreted by *Erwinia chrysanthemi***. *Biochemistry* 1997, **36**(51):16074-16086.

6. Kraulis J, Clore GM, Nilges M, Jones TA, Pettersson G, Knowles J, Gronenborn AM: **Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing**. *Biochemistry* 1989, **28**(18):7241-7257.

7. Raghothama S, Simpson PJ, Szabo L, Nagy T, Gilbert HJ, Williamson MP: **Solution structure of the CBM10 cellulose binding module from *Pseudomonas* xylanase A**. *Biochemistry* 2000, **39**(5):978-984.

8. Tormo J, Lamed R, Chirino AJ, Morag E, Bayer EA, Shoham Y, Steitz TA: **Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose**. *EMBO J* 1996, **15**(21):5739-5751.

9. Xu GY, Ong E, Gilkes NR, Kilburn DG, Muhandiram DR, Harris-Brandts M, Carver JP, Kay LE, Harvey TS: **Solution structure of a cellulose-binding domain from *Cellulomonas fimi* by nuclear magnetic resonance spectroscopy**. *Biochemistry* 1995, **34**(21):6993-7009.

10. Cuneo MJ, Changela A, Beese LS, Hellinga HW: **Structural adaptations that modulate monosaccharide, disaccharide, and trisaccharide specificities in periplasmic maltose-binding proteins**. *J Mol Biol* 2009, **389**(1):157-166.

11. Sulak O, Cioci G, Delia M, Lahmann M, Varrot A, Imberty A, Wimmerova M: **A TNF-like trimeric lectin domain from *Burkholderia cenocepacia* with specificity for fucosylated human histo-blood group antigens**. *Structure* 2010, **18**(1):59-72.

12. Murase T, Zheng RB, Joe M, Bai Y, Marcus SL, Lowary TL, Ng KK: **Structural insights into antibody recognition of mycobacterial polysaccharides**. *J Mol Biol* 2009, **392**(2):381-392.

13. Schulz EC, Bergfeld AK, Ficner R, Muhlenhoff M: **Crystal structure analysis of the polysialic acid specific O-acetyltransferase NeuO**. *PLoS One* 2011, **6**(3):e17403.

14. Woods RJ, Tessier MB: **Computational glycoscience: characterizing the spatial and temporal properties of glycans and glycan-protein complexes**. *Curr Opin Struct Biol* 2010, **20**(5):575-583.

15. Pérez S, Gautier C, Imberty A, Ernst B, Hart GW, Sinaý P: **Oligosaccharide Conformations by Diffraction Methods**. In: *Carbohydrates in Chemistry and Biology.* Wiley-VCH Verlag GmbH; 2008: 969-1001.

16. Peters T, Pinto BM: **Structure and dynamics of oligosaccharides: NMR and modeling studies**. *Current Opinion in Structural Biology* 1996, **6**(5):710-720.

17. Rice KG, Pengguang W, Brand L, Lee YC: **Experimental determination of oligosaccharide three-dimensional structure**. *Current Opinion in Structural Biology* 1993, **3**(5):669-674.

18. Woods RJ: **The application of molecular modeling techniques to the determination of oligosaccharide solution conformations**. In: *Reviews in Computational Chemistry.* John Wiley & Sons, Inc.; 2007: 129-165.

19. Pérez S, Kouwijzer M: **Shapes and interactions of polysaccharide chains**. *ChemInform* 2000, **31**(46):258-293.

20. Pérez S: **Molecular modelling in glycoscience** In: *Comprehensive Glycoscience, 211 : Analysis of Glycans.* Edited by Kamerling JP, vol. 2: Elsevier; 2007: 347-388.

21. Laine RA: **Invited Commentary: A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 1012 structures for a reducing hexasaccharide: the isomer barrier to development of single-method saccharide sequencing or synthesis systems**. *Glycobiology* 1994, **4**(6):759-767.

22. Tvaroska I, Pérez S: **Conformational-energy calculations for oligosaccharides: a comparison of methods and a strategy of calculation**. *Carbohydrate Research* 1986, **149**(2):389-410.

23. Spiwok V, Lipovova P, Skalova T, Buchtelova E, Hasek J, Kralova B: **Role of CH/π interactions in substrate binding by *Escherichia coli* β-galactosidase**. *Carbohydr Res* 2004, **339**(13):2275-2280.

24. Chervenak MC, Toone EJ: **A direct measure of the contribution of solvent reorganization to the enthalpy of binding**. *Journal of the American Chemical Society* 1994, **116**(23):10533-10539.

25. Poveda A, Asensio JL, Espinosa JF, Martin-Pastor M, Canada J, Jimenez-Barbero J: **Applications of nuclear magnetic resonance spectroscopy and molecular modeling to the study of protein-carbohydrate interactions**. *J Mol Graph Model* 1997, **15**(1):9-17, 53.

26. Tschampel SM, Woods RJ: **Quantifying the role of water in protein-carbohydrate interactions**. *J Phys Chem A* 2003, **107**(43):9175-9181.

27. Sorin EJ, Pande VS: **Empirical force-field assessment: The interplay between backbone torsions and noncovalent term scaling**. *J Comput Chem* 2005, **26**(7):682-690.

28. Kirschner KN, Woods RJ: **Solvent interactions determine carbohydrate conformation**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10541-10545.

29. Imberty A, Perez S: **Structure, conformation, and dynamics of bioactive oligosaccharides: theoretical approaches and experimental validations**. *Chemical reviews* 2000, **100**(12):4567-4588.

30. Pérez S, Imberty A, Engelsen SB, Gruza J, Mazeau K, Jimenez-Barbero J, Poveda A, Espinosa J-F, van Eyck BP, Johnson G *et al*: **A comparison and chemometric analysis of several molecular mechanics force fields and parameter sets applied to carbohydrates**. *Carbohydrate Research* 1998, **314**(3-4):141-155.

31. Dowd MK, French AD, Reilly PJ: **Modeling of aldopyranosyl ring puckering with MM3 (92)**. *Carbohydrate Research* 1994, **264**(1):1-19.

32. Glennon TM, Merz Jr KM: **A carbohydrate force field for amber and its application to the study of saccharide to surface adsorption**. *Journal of Molecular Structure: THEOCHEM* 1997, **395,Äì396**(0):157-171.

33. Homans SW: **A molecular mechanical force field for the conformational analysis of oligosaccharides: comparison of theoretical and crystal structures of Man-α1-3Man-β-1-4GlcNAc**. *Biochemistry* 1990, **29**(39):9110-9118.

34. Kirschner KN, Yongye AB, Tschampel SM, Gonzalez-Outeirino J, Daniels CR, Foley BL, Woods RJ: **GLYCAM06: A generalizable Biomolecular force field. Carbohydrates**. *Journal of Computational Chemistry* 2008, **29**(4):622-655.

35. Momany FA, Willett JL: **Computationalstudies on carbohydrates: in vacuostudies using a revised AMBER force field, AMB99C, designed for α-(1→4)** *Carbohydrate Research* 2000, **326**(3):194-209.

36. Woods RJ, Dwek RA, Edge CJ, Fraserreid B: **Molecular mechanical and molecular dynamical simulations of glycoproteins and oligosaccharides. 1. Glycam-93 parameter development**. *Journal of Physical Chemistry* 1995, **99**(11):3832-3846.

37. Bradbrook GM, Forshaw JR, Pérez S: **Structure/thermodynamics relationships of lectin–saccharide complexes**. *European Journal of Biochemistry* 2000, **267**(14):4545-4555.

38. Tempel W, Tschampel S, Woods RJ: **The xenograft antigen bound to *Griffonia simplicifolia* Lectin 1-B4**. *Journal of Biological Chemistry* 2002, **277**(8):6615-6621.

39. Bryce RA, Hillier IH, Naismith JH: **Carbohydrate-protein recognition: Molecular dynamics simulations and free energy analysis of oligosaccharide binding to Concanavalin A**. *Biophysical Journal* 2001, **81**(3):1373-1388.

40. Clarke C, Woods RJ, Gluska J, Cooper A, Nutley MA, Boons G-J: **Involvement of water in carbohydrate-protein binding**. *Journal of the American Chemical Society* 2001, **123**(49):12238-12247.

41. Liang G, Schmidt RK, Yu HA, Cumming DA, Brady JW: **Free energy simulation studies of the binding specificity of mannose-binding protein**. *The journal of physical chemistry* 1996, **100**(7):2528-2534.

42. Pathiaseril A, Woods RJ: **Relative energies of binding for antibody-carbohydrate-antigen complexes computed from free-energy simulations**. *Journal of the American Chemical Society* 2000, **122**(2):331-338.

43. Laughrey ZR, Kiehna SE, Riemen AJ, Waters ML: **Carbohydrate-π interactions: what are they worth?** *J Am Chem Soc* 2008, **130**(44):14625-14633.

44. Ramirez-Gualito K, Alonso-Rios R, Quiroz-Garcia B, Rojas-Aguilar A, Diaz D, Jimenez-Barbero J, Cuevas G: **Enthalpic nature of the CH/π 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* 2009, **131**(50):18129-18138.

45. Spiwok V, Lipovová P, Skálová T, Vondráčková E, Dohnálek J, Hašek J, Králová B: **Modelling of carbohydrate–aromatic interactions: *ab initio* energetics and force field performance**. *Journal of Computer-Aided Molecular Design* 2005, **19**(12):887-901.

46. Vandenbussche S, Díaz D, Fernández-Alonso MC, Pan W, Vincent SP, Cuevas G, Cañada FJ, Jiménez-Barbero J, Bartik K: **Aromatic–carbohydrate interactions: An NMR and computational study of model systems**. *Chemistry – A European Journal* 2008, **14**(25):7570-7578.

47. Tessier MB, DeMarco, M. L., Yongye, A. B., Woods, R. J.: **Extension of the GLYCAM06 biomolecular force field to lipids, lipid**

bilayers and glycolipids. *Mol Simul* 2008, **34**:349-364.

48. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA: **A second generation force field for the simulation of proteins, nucleic acids, and organic molecules**. *Journal of the American Chemical Society* 1995, **117**(19):5179-5197.

49. Biarnes X, Nieto J, Planas A, Rovira C: **Substrate distortion in the Michaelis complex of *Bacillus* 1,3-1,4-β-glucanase. Insight from first principles molecular dynamics simulations**. *The Journal of biological chemistry* 2006, **281**(3):1432-1441.

50. Basma M, Sundara S, Calgan D, Vernali T, Woods RJ: **Solvated ensemble averaging in the calculation of partial atomic charges**. *J Comput Chem* 2001, **22**(11):1125-1137.

51. Woods RJ, Khalil M, Pell W, Moffat SH, Smith VH: **Derivation of net atomic charges from molecular electrostatic potentials**. *Journal of Computational Chemistry* 1990, **11**(3):297-310.

52. Woods RJ, Chappelle R: **Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates**. *Journal of Molecular Structure: THEOCHEM* 2000, **527**(1-3):149-156.

53. Hansen HS, Hünenberger PH: **A reoptimized GROMOS force field for hexopyranose-based carbohydrates accounting for the relative free energies of ring conformers, anomers, epimers, hydroxymethyl rotamers, and glycosidic linkage conformers**. *Journal of Computational Chemistry* 2011, **32**(6):998-1032.

54. Lins RD, Hunenberger PH: **A new GROMOS force field for hexopyranose-based carbohydrates**. *J Comput Chem* 2005, **26**(13):1400-1412.

55. Schuler LD, Daura X, van Gunsteren WF: **An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase**. *Journal of Computational Chemistry* 2001, **22**(11):1205-1218.

56. Schuler LD, Van Gunsteren WF: **On the Choice of Dihedral Angle Potential Energy Functions for n-alkanes**. *Mol Simul* 2000, **25**(5):301 - 319.

57. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J: **Interaction models for water in relation to protein hydration**. *Intermolecular Forces* 1981:331-342.

58. Gandhi NS, Mancera RL: **Free energy calculations of glycosaminoglycan-protein interactions**. *Glycobiology* 2009, **19**(10):1103-1115.

59. Guvench O, Greene SN, Kamath G, Brady JW, Venable RM, Pastor RW, Mackerell AD, Jr.: **Additive empirical force field for hexopyranose monosaccharides**. *J Comput Chem* 2008, **29**(15):2543-2564.

60. Guvench O, Hatcher ER, Venable RM, Pastor RW, Mackerell AD: **CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses**. *Journal of chemical theory and computation* 2009, **5**(9):2353-2370.

61. Hatcher E, Guvench O, Mackerell AD: **CHARMM additive all-atom force field for acyclic polyalcohols, acyclic carbohydrates and inositol**. *Journal of chemical theory and computation* 2009, **5**(5):1315-1327.

62. MacKerell AD, Banavali N, Foloppe N: **Development and current status of the CHARMM force field for nucleic acids**. *Biopolymers* 2000, **56**(4):257-265.

63. MacKerell AD, Bashford D, Bellott, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S *et al*: **All-atom empirical potential for molecular modeling and dynamics studies of proteins**. *The Journal of Physical Chemistry B* 1998, **102**(18):3586-3616.

64. Mackerell AD, Feig M, Brooks CL: **Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations**. *Journal of Computational Chemistry* 2004, **25**(11):1400-1415.

65. Kony D, Damm W, Stoll S, Van Gunsteren WF: **An improved OPLS-AA force field for carbohydrates**. *J Comput Chem* 2002, **23**(15):1416-1429.

66. Damm W, Frontera A, Tirado–Rives J, Jorgensen WL: **OPLS all-atom force field for carbohydrates**. *Journal of Computational Chemistry* 1997, **18**(16):1955-1970.

67. Jorgensen WL, Maxwell DS, Tirado-Rives J: **Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids**. *Journal of the American Chemical Society* 1996, **118**(45):11225-11236.

68. Allinger NL, Li F, Yan L, Tai JC: **Molecular mechanics (MM3) calculations on conjugated hydrocarbons**. *Journal of Computational Chemistry* 1990, **11**(7):868-895.

69. Lii J-H, Allinger NL: **The MM3 force field for amides, polypeptides and proteins**. *Journal of Computational Chemistry* 1991, **12**(2):186-199.

70. Clark M, Cramer RD, Van Opdenbosch N: **Validation of the general purpose TRIPOS 5.2 force field**. *Journal of Computational Chemistry* 1989, **10**(8):982-1012.

71. Imberty A, Hardman KD, Carver JP, Perez S: **Molecular modelling of protein-carbohydrate interactions. Docking of**

monosaccharides in the binding site of concanavalin A. *Glycobiology* 1991, **1**(6):631-642.

72. Pérez S, Meyer C, Imberty A: **Practical tools for molecular modeling of complex carbohydrates and their interactions with proteins**. *Molecular Engineering* 1995, **5**(1):271-300.

73. TRIPOS: **SYBYL-X 1.3, Tripos, Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA**. In.; 1991 - 2011.

74. Lengauer T, Rarey M: **Computational methods for biomolecular docking**. *Current Opinion in Structural Biology* 1996, **6**(3):402-406.

75. Bultinck PW, H. D.; Langenaecker, W; Tollenare, J. P.: **Computational medicinal chemistry for drug discovery**: Marcel Dekker, Inc.; 2004.

76. Eklund R: **Computational analysis of carbohydrates : dynamical properties and interactions**. In.: Stockholm University, Sweden; 2005.

77. Höltje H-D, Sippl W, Rognan D, Folkers G: **Molecular modeling: basic principles and applications**, 3rd Edition edn; 2008.

78. Kranjc A: **Predicting structural determinants and ligand poses in proteins involved in neurological diseases: bioinformatics and molecular simulation studies**. *PhD Thesis*. 2009.

79. Walker RC: **The development of a QM/MM based linear response method and its application to proteins**. London: Imperial College; 2003.

80. Ewing TJA, Makino S, Skillman AG, Kuntz ID: **DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases**. *Journal of Computer-Aided Molecular Design* 2001, **15**(5):411-428.

81. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ: **Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function**. *Journal of Computational Chemistry* 1998, **19**(14):1639-1662.

82. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK *et al*: **Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy**. *Journal of Medicinal Chemistry* 2004, **47**(7):1739-1749.

83. Muegge I, Martin YC: **A general and fast scoring function for protein-ligand interactions: A simplified potential approach**. *Journal of Medicinal Chemistry* 1999, **42**(5):791-804.

84. Barbe S, Cortés, J., Siméon, T., Monsan, P., Renaud-Siméon, M., André, I.: **Solvent-dependent lid motions of *Burkholderia cepacia* lipase investigated by a mixed molecular modelling and robotics approach.** In: *Proteins: Structure, Function and Bioinformatics.* 2011.

85. Kirillova S, Cortés J, Stefaniu A, Siméon T: **An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins**. *Proteins: Structure, Function, and Bioinformatics* 2008, **70**(1):131-143.

86. Haspel N, Moll M, Baker M, Chiu W, Kavraki L: **Tracing conformational changes in proteins**. *BMC Struct Biol* 2010, **10**(Suppl 1):S1.

87. Cortés J, Barbe S, Erard M, Siméon T: **Encoding molecular motions in voxel maps**. *IEEE/ACM Trans Comput Biol Bioinformatics* 2011, **8**(2):557-563.

88. Cortés J, Jaillet, L., Siméon, T.: **Disassembly path planning for complex articulated objects**. *IEEE Transactions on Robotics and Automation* 2008, **24 (2)**:475 - 481.

89. Cortés J, Thanh Le D, Iehl R, Simeon T: **Simulating ligand-induced conformational changes in proteins using a mechanical disassembly method**. *Physical Chemistry Chemical Physics* 2010, **12**(29):8268-8276.

90. Guieysse D, Cortés J, Puech-Guenot S, Barbe S, Lafaquière V, Monsan P, Siméon T, André I, Remaud-Siméon M: **A structure-controlled investigation of lipase enantioselectivity by a path-planning approach**. *Chembiochem* 2008, **9**(8):1308-1317.

91. Lafaquière V, Barbe S, Puech-Guenot S, Guieysse D, Cortés J, Monsan P, Siméon T, André I, Remaud-Siméon M: **Cover Picture: Control of lipase enantioselectivity by engineering the substrate binding site and access channel (ChemBioChem 17/2009)**. *Chembiochem* 2009, **10**(17):2677-2677.

92. Gilson MK, Honig B: **Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis**. *Proteins: Structure, Function, and Bioinformatics* 1988, **4**(1):7-18.

93. Sitkoff D, Sharp KA, Honig B: **Accurate calculation of hydration free energies using macroscopic solvent models**. *The journal of physical chemistry* 1994, **98**(7):1978-1988.

94. Srinivasan J, Cheatham TE, Cieplak P, Kollman PA, Case DA: **Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices**. *Journal of the American Chemical Society* 1998, **120**(37):9401-9409.

95. Basdevant N, Weinstein H, Ceruso M: **Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ Domains, a Case Study**. *Journal of the American Chemical Society* 2006, **128**(39):12766-12777.

96. Gohlke H, Case DA: **Converging free energy estimates: MM-PB(GB)SA studies on the protein–protein complex Ras–Raf**. *Journal of Computational Chemistry* 2004, **25**(2):238-250.

97. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W *et al*: **Calculating structures and free energies of complex molecules: Combining Molecular Mechanics and Continuum Models**. *Accounts of Chemical Research* 2000, **33**(12):889-897.

98. DeMarco ML, Woods RJ: **Structural glycobiology: A game of snakes and ladders**. *Glycobiology* 2008, **18**(6):426-440.

99. Krengel U, Imberty A: **Crystallography and lectin structure database**. In: *Lectins.* Edited by Carol LN. Amsterdam: Elsevier Science B.V.; 2007: 15-50.

100. **Online database of 3 dimensional structures of lectins** [http://www.cermav.cnrs.fr/lectines]

101. Sulak O, Lameignere E, Wimmerova M, Imberty A: **Specificity and affinity studies in lectin/carbohydrate interactions**. In: *Carbohydrate Chemistry.* vol. 35: The Royal Society of Chemistry; 2009: 357-372.

102. Nurisso A, Kozmon, S., Imberty, A.: **Comparison of docking methods for carbohydrate binding in calcium-dependent lectins and prediction of the carbohydrate binding mode to sea cucumber lectin CEL-III**. *Mol Simul* 2008, **34**:469-479.

103. Clausen H, Hakomori S-i: **ABH and related histo-blood group antigens; immunochemical differences in carrier isotypes and their distribution**. *Vox Sanguinis* 1989, **56**(1):1-20.

104. de Geus DC, van Roon A-MM, Thomassen EAJ, Hokke CH, Deelder AM, Abrahams JP: **Characterization of a diagnostic Fab fragment binding trimeric Lewis X**. *Proteins: Structure, Function, and Bioinformatics* 2009, **76**(2):439-447.

105. Ramsland PA, Farrugia W, Bradford TM, Mark Hogarth P, Scott AM: **Structural convergence of antibody binding of carbohydrate determinants in Lewis Y tumor antigens**. *Journal of Molecular Biology* 2004, **340**(4):809-818.

106. van Roon A-MM, Pannu NS, de Vrind JPM, van der Marel GA, van Boom JH, Hokke CH, Deelder AM, Abrahams JP: **Structure of an Anti-Lewis X Fab fragment in complex with its Lewis X antigen**. *Structure* 2004, **12**(7):1227-1236.

107. van Roon A-MM, Pannu NS, Hokke CH, Deelder AM, Abrahams JP: **Crystallization and preliminary X-ray analysis of an anti-LewisX Fab fragment with and without its LewisX antigen**. *Acta Crystallographica Section D* 2003, **59**(7):1306-1309.

108. Imberty A, Mikros E, Koca J, Mollicone R, Oriol R, Pérez S: **Computer simulation of histo-blood group oligosaccharides: energy maps of all constituting disaccharides and potential energy surfaces of 14 ABH and Lewis carbohydrate antigens**. *Glycoconjugate Journal* 1995, **12**(3):331-349.

109. Imberty A, Mollicone R, Mikros E, Carrupt P-A, P√©rez S, Oriol R: **How do antibodies and lectins recognize histo-blood group antigens? A 3D-QSAR study by comparative molecular field analysis (CoMFA)**. *Bioorganic &amp; Medicinal Chemistry* 1996, **4**(11):1979-1988.

110. Agostino M, Sandrin MS, Thompson PE, Yuriev E, Ramsland PA: ***In silico* analysis of antibody-carbohydrate interactions and its application to xenoreactive antibodies**. *Molecular Immunology* 2009, **47**(2-3):233-246.

111. Agostino M, Sandrin MS, Thompson PE, Yuriev E, Ramsland PA: **Identification of preferred carbohydrate binding modes in xenoreactive antibodies by combining conformational filters and binding site maps**. *Glycobiology* 2010, **20**(6):724-735.

112. Kadirvelraj R, Gonzalez-Outeiri√±o J, Foley BL, Beckham ML, Jennings HJ, Foote S, Ford MG, Woods RJ: **Understanding the bacterial polysaccharide antigenicity of *Streptococcus agalactiae* versus *Streptococcus pneumoniae***. *Proceedings of the National Academy of Sciences* 2006, **103**(21):8149-8154.

113. Vulliez-Le Normand B, Saul FA, Phalipon A, Bélot F, Guerreiro C, Mulard LA, Bentley GA: **Structures of synthetic O-antigen fragments from serotype 2a *Shigella flexneri* in complex with a protective monoclonal antibody**. *Proceedings of the National Academy of Sciences* 2008, **105**(29):9976-9981.

114. Theillet Fo-X, Frank M, Vulliez-Le Normand B, Simenel C, Hoos S, Chaffotte A, Bélot F, Guerreiro C, Nato F, Phalipon A *et al*: **Dynamic aspects of antibody:oligosaccharide complexes characterized by molecular dynamics simulations and saturation transfer difference nuclear magnetic resonance**. *Glycobiology* 2011, **21**(12):1570-1579.

115. Simmons DA, Romanowska E: **Structure and biology of *Shigella flexneri* O antigens**. *J Med Microbiol* 1987, **23**(4):289-302.

116. Kjellen L, Lindahl U: **Proteoglycans: structures and interactions**. *Annual Review of Biochemistry* 1991, **60**(1):443-475.

117. Goodford PJ: **A computational procedure for determining energetically favorable binding sites on biologically important macromolecules**. *Journal of Medicinal Chemistry* 1985, **28**(7):849-857.

118. Johnson DJD, Li W, Adams TE, Huntington JA: **Antithrombin-S195A factor Xa-heparin structure reveals the allosteric mechanism of antithrombin activation**. *EMBO J* 2006, **25**(9):2029-2037.

119. Skinner R, Abrahams J-P, Whisstock JC, Lesk AM, Carrell RW, Wardell MR: **The 2.6 Å structure of antithrombin indicates a conformational change at the heparin binding site**. *Journal of Molecular Biology* 1997, **266**:601-609.

120. Grootenhuis PDJ, Vanboeckel CAA: **Constructing a molecular model of the interaction between antithrombin-III and a potent heparin analog**. *Journal of the American Chemical Society* 1991, **113**(7):2743-2747.

121. Bitomsky W, Wade RC: **Docking of glycosaminoglycans to heparin-binding proteins: Validation for aFGF, bFGF, and antithrombin and application to IL-8**. *Journal of the American Chemical Society* 1999, **121**(13):3004-3013.

122. Jin L, Abrahams JP, Skinner R, Petitou M, Pike RN, Carrell RW: **The anticoagulant activation of antithrombin by heparin**. *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**(26):14683-14688.

123. Hricovini M, Guerrini M, Bisio A, Torri G, Petitou M, Casu B: **Conformation of heparin pentasaccharide bound to antithrombin III**. *Biochem J* 2001, **359**(Pt 2):265-272.

124. Dementiev A, Petitou M, Herbert JM, Gettins PGW: **The ternary complex of antithrombin-anhydrothrombin heparin reveals the basis of inhibitor specificity**. *Nat Struct Mol Biol* 2004, **11**(9):863-867.

125. Li W, Johnson DJD, Esmon CT, Huntington JA: **Structure of the antithrombin-thrombin-heparin ternary complex reveals the antithrombotic mechanism of heparin**. *Nat Struct Mol Biol* 2004, **11**(9):857-862.

126. Pellegrini L, Burke DF, von Delft F, Mulloy B, Blundell TL: **Crystal structure of fibroblast growth factor receptor ectodomain bound to ligand and heparin**. *Nature* 2000, **407**(6807):1029-1034.

127. Schlessinger J, Plotnikov AN, Ibrahimi OA, Eliseenkova AV, Yeh BK, Yayon A, Linhardt RJ, Mohammadi M: **Crystal structure of a ternary FGF-FGFR-heparin complex reveals a dual role for heparin in FGFR binding and dimerization**. *Mol Cell* 2000, **6**(3):743-750.

128. Raman R, Venkataraman G, Ernst S, Sasisekharan V, Sasisekharan R: **Structural specificity of heparin binding in the fibroblast growth factor family of proteins**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(5):2357-2362.

129. Rossi D, Zlotnik A: **The biology of chemokines and their receptors**. *Annu Rev Immunol* 2000, **18**:217-243.

130. Handel TM, Johnson Z, Crown SE, Lau EK, Sweeney M, Proudfoot AE: **Regulation of protein function by glycosaminoglycans - as exemplified by chemokines**. *Annual Review of Biochemistry* 2005, **74**:385-410.

131. Lortat-Jacob H, Grosdidier A, Imberty A: **Structural diversity of heparan sulfate binding domains in chemokines**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(3):1229-1234.

132. Schirmer T, Keller T, Wang Y, Rosenbusch J: **Structural basis for sugar translocation through maltoporin channels at 3.1 A resolution**. *Science* 1995, **267**(5197):512-514.

133. Dutzler R, Schirmer T, Karplus M, Fischer S: **Translocation mechanism of long sugar chains across the maltoporin membrane channel**. *Structure* 2002, **10**(9):1273-1284.

134. Dutzler R, Wang YF, Rizkallah PJ, Rosenbusch JP, Schirmer T: **Crystal structures of various malto-oligosaccharides bound to maltoporin reveal a specific sugar translocation pathway**. *Structure* 1996, **4**(2):127-134.

135. Dang S, Sun L, Huang Y, Lu F, Liu Y, Gong H, Wang J, Yan N: **Structure of a fucose transporter in an outward-open conformation**. *Nature* 2010, **467**(7316):734-738.

136. **3D Glycosyltransferase database** [http://glyco3d.cermav.cnrs.fr/cgi-bin/rxgt/rxgt.cgi]

137. Bréton C, Šnajdrová L, Jeanneau C, Koča J, Imberty A: **Structures and mechanisms of glycosyltransferases**. *Glycobiology* 2006, **16**(2):29R-37R.

138. Heissigerova H, Bréton C, Moravcova J, Imberty A: **Molecular modeling of glycosyltransferases involved in the biosynthesis of blood group A, blood group B, Forssman, and iGb(3) antigens and their interaction with substrates**. *Glycobiology* 2003, **13**(5):377-386.

139. Petrova P, Koca J, Imberty A: **Potential energy hypersurfaces of nucleotide sugars: Ab initio calculations, force-field parametrization, and exploration of the flexibility**. *Journal of the American Chemical Society* 1999, **121**(23):5535-5547.

140. Botte C, Jeanneau C, Snajdrova L, Bastien O, Imberty A, Bréton C, Marechal E: **Molecular modeling and site-directed mutagenesis of plant chloroplast monogalactosyldiacylglycerol synthase reveal critical residues for activity**. *J Biol Chem* 2005, **280**(41):34691-34701.

141. Yang M, Zhou, Lu, Zuo, Zhili, Tang, Xiangyang, Liu, Jian, Ma, Xiang: **Structure-based virtual screening for glycosyltransferase51**. *Mol Simul* 2008, **34**:849-856.

142. André I, Tvaroška I, Carver JP: **On the reaction pathways and determination of transition-state structures for retaining α-galactosyltransferases**. *Carbohydrate Research* 2003, **338**(9):865-877.

143. Kozmon S, Tvaroška I: **Catalytic mechanism of glycosyltransferases: hybrid quantum mechanical/molecular mechanical study of the inverting *N*-acetylglucosaminyltransferase I**. *Journal of the American Chemical Society* 2006, **128**(51):16921-16927.

144. Krupička M, Tvaroška I: **Hybrid quantum mechanical/molecular mechanical investigation of the β-1,4-Galactosyltransferase-I mechanism**. *The Journal of Physical Chemistry B* 2009, **113**(32):11314-11319.

145. Gandhi NS, Freeman C, Parish CR, Mancera RL: **Computational analyses of the catalytic and heparin-binding sites and their

interactions with glycosaminoglycans in glycoside hydrolase family 79 endo-β-D-glucuronidase (heparanase) *Glycobiology* 2012, **22**(1):35-55.

146. Sapay N, Cabannes É, Petitou M, Imberty A: **Molecular model of human heparanase with proposed binding mode of a heparan sulfate oligosaccharide and catalytic amino acids**. *Biopolymers* 2012, **97**(1):21-34.

147. Varghese JN, Laver WG, Colman PM: **Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution**. *Nature* 1983, **303**(5912):35-40.

148. Varghese JN, McKimm-Breschkin JL, Caldwell JB, Kortt AA, Colman PM: **The structure of the complex between influenza virus neuraminidase and sialic acid, the viral receptor**. *Proteins: Structure, Function, and Bioinformatics* 1992, **14**(3):327-332.

149. von Itzstein M, Wu W-Y, Kok GB, Pegg MS, Dyason JC, Jin B, Phan TV, Smythe ML, White HF, Oliver SW *et al*: **Rational design of potent sialidase-based inhibitors of influenza virus replication**. *Nature* 1993, **363**(6428):418-423.

150. Kim CU, Lew W, Williams MA, Liu H, Zhang L, Swaminathan S, Bischofberger N, Chen MS, Mendel DB, Tai CY *et al*: **Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site: Design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity**. *Journal of the American Chemical Society* 1997, **119**(4):681-690.

151. McCullers JA, Hoffmann E, Huber VC, Nickerson AD: **A single amino acid change in the C-terminal domain of the matrix protein M1 of influenza B virus confers mouse adaptation and virulence**. *Virology* 2005, **336**(2):318-326.

152. Chachra R, Rizzo RC: **Origins of resistance conferred by the R292K neuraminidase mutation via molecular dynamics and free energy calculations**. *Journal of chemical theory and computation* 2008, **4**(9):1526-1540.

153. Raab M, Tvaroška I: **The binding properties of the H5N1 influenza virus neuraminidase as inferred from molecular modeling**. *Journal of Molecular Modeling* 2011, **17**(6):1445-1456.

154. Doi RH, Kosugi A: **Cellulosomes: Plant-cell-wall-degrading enzyme complexes**. *Nat Rev Microbiol* 2004, **2**(7):541-551.

155. Demain AL, Newcomb M, Wu JHD: **Cellulase, clostridia, and ethanol**. *Microbiol Mol Biol Rev* 2005, **69**(1):124-+.

156. Crowley MFU, E. C.; Iii, C. L. B.; Walker, R. C.; Nimlos, M. R.; Himmel, M. E.: **Developing improved MD codes for understanding processive cellulases**. In: *Journal of Physics: Conference Series.* Institute of Physics Publishing: 2008; 12049-12012056; 2008.

157. Yui T, Shiiba H, Tsutsumi Y, Hayashi S, Miyata T, Hirata F: **Systematic docking study of the carbohydrate binding module protein of Cel7A with the cellulose Iα crystal model**. *Journal of Physical Chemistry B* 2010, **114**(1):49-58.

158. Imai T, Boisset C, Samejima M, Igarashi K, Sugiyama J: **Unidirectional processive action of cellobiohydrolase Cel7A on *Valonia* cellulose microcrystals**. *Febs Letters* 1998, **432**(3):113-116.

159. Hashimoto H: **Recent structural studies of carbohydrate-binding modules**. *Cellular and Molecular Life Sciences* 2006, **63**(24):2954-2967.

Annex

Annex I

Annex I

STRUCTURE OF CARBOHYDRATES

I. Important parameters for the 3D description of glycans

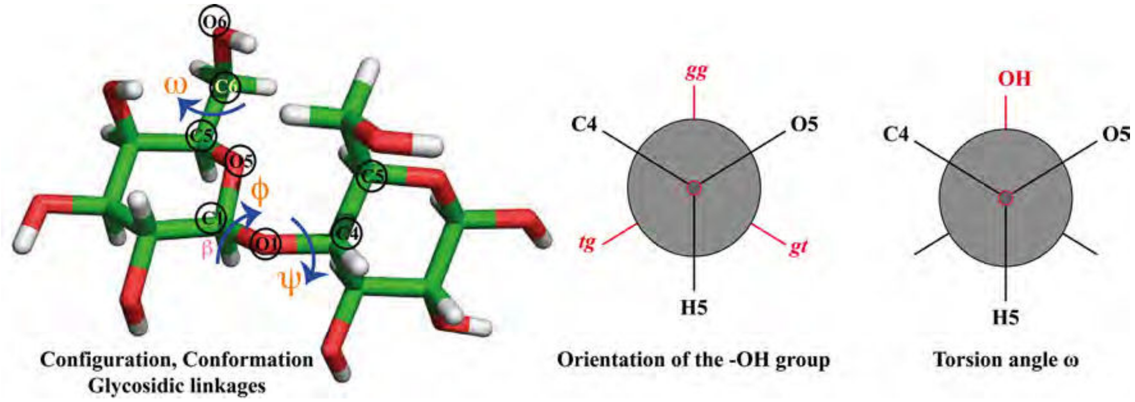


Figure I.1 Parameters for glycan structure description.
[This figure describes the 'heavy atom convention' for the torsion angle description for (1→x) linkages, where, Φ→O5-C1-O-C_x, Ψ→ C1-O1-C_x-C_{x+1} and ω→O5-C5-C6-O6].

I.1. Constitution, Configuration & conformation

Constitution by definition is the reversible formation of cyclic hemiacetals from the corresponding poly-hydroxy-aldehydes or –ketones (denoted as ring tautomerism).

Configuration of the sugar refers to its spatial arrangement of bonds that can only be altered by breaking of bonds. For instance, the D/L and the R/S configurations of organic molecules can only be changed by breaking one or more bonds connecting the chiral atom.

Conformation is the spatial arrangement of atoms in a molecule (usually of substituent groups) that are free to assume different positions in space through the free rotation of atoms about a single chemical bond. It can be changed without breaking bonds. For example, rotation about single bonds produces the *cis*/*trans* conformations.

Constitution and configuration (**Figure I.1**) describe permanent geometry descriptors for the glycan. In order to change them a chemical change would be required, implying an addition of (a relatively) high energy or catalyst, which would alter the identity of the glycan itself. On the other hand, *conformations* can change in solution at room temperature without altering the nature of the glycan. The activation energy lies, in general, below 63 kJ/mol for monosaccharides.

For furanoses and pyranoses, one can speak about definite conformations and energetic preferences of some forms (though mostly a single form abounds) as these are separated but sufficiently high-energy barriers for conformational inter-conversion. Energy differences of a few kJ/mol suffice to allow one stable conformation to dominate to almost 100% over less stable ones in equilibrium *in solution* (the effect being especially pronounced for pyranoses). In the *crystalline form*, however, monosaccharides and their derivatives are conformationally homogeneous.

In solution, the stable conformations of a monosaccharide and their relative proportions to one another are influenced by the following:

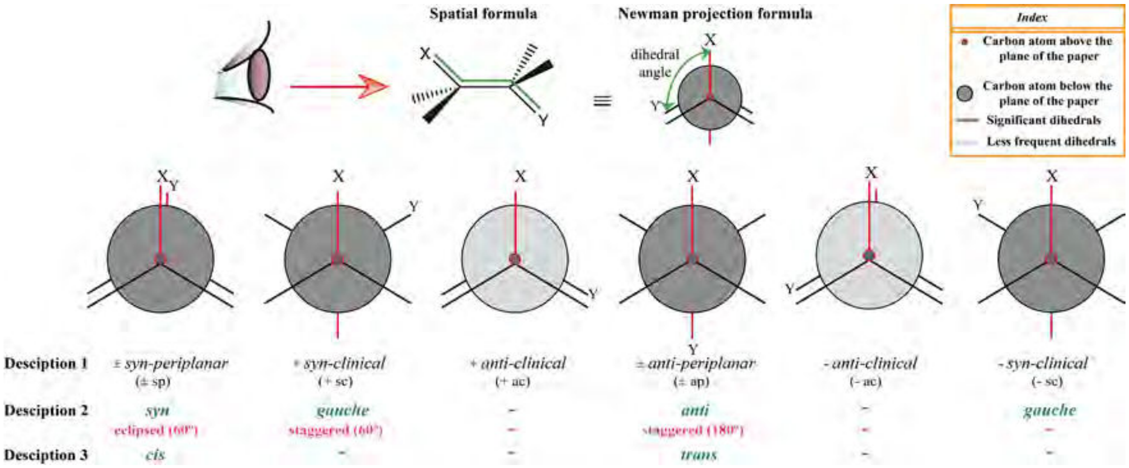
a. influence of the environment
b. underlying constitution and configuration

It can thus be inferred that conformations indicate the fine structure of the compound and come closest to representing the true shape of the molecule.

Thus, extrapolating from the above, physical properties and behavior of the glycans can be traced down to the predominant conformation(s). These conformations can be assigned using physical methods, especially NMR spectroscopy. There are 3 'non-bonded interactions' that stabilize a certain conformation in comparison to others:

a. steric (*van der Waals*) interactions
b. polar (*electrostatic*) interactions
c. hydrogen bonding

The Newman projection formulae are convenient for the pictorial representation of conformations. In such a projection, the angle formed between a bond protruding out (in front) of the plane of the paper and a bond behind the plane of the paper, is known as the *dihedral or torsion angle* (**Figure I.2**).


Spatial formula
Newman projection formula
X
Y
dihedral angle
X
Y
Index
Carbon atom above the plane of the paper
Carbon atom below the plane of the paper
Significant dihedrals
Less frequent dihedrals
X
Y
X
Y
X
Y
X
Y
X
Y
X
Y

Desciption 1	± *syn-periplanar* (± sp)	+ *syn-clinical* (+ sc)	+ *anti-clinical* (+ ac)	± *anti-periplanar* (± ap)	- *anti-clinical* (- ac)	- *syn-clinical* (- sc)
Desciption 2	*syn* eclipsed (60°)	*gauche* staggered (60°)	- -	*anti* staggered (180°)	- -	*gauche* -
Desciption 3	*cis*	-	-	*trans*	-	-

Figure I.2 Torsion angle descriptions.

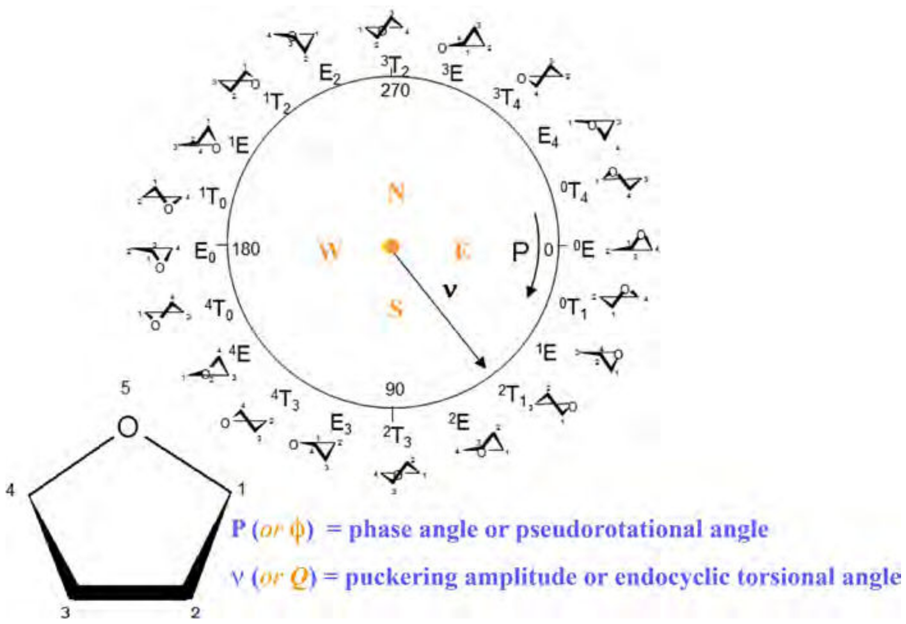
In general, steric interactions are the deciding factor in the relative stabilization or de-stabilization of certain conformations of a molecule.

I.2. Ring puckering

The puckered conformations of furanose (five-membered glycan ring) and pyranose (six-membered glycan ring) are central to analyzing the action of enzymes on carbohydrates [1]. Sugar rings can adopt different conformations that can be described using the puckering parameters described by Cremer and Pople [2] as illustrated in **Figure I.3** and **I.4** for furanoses and pyranoses, respectively. In every case, there are four or more atoms that define the plane. In order to visualize which atoms are above or below the plane, the molecule has to be oriented such that the atoms are numbered clockwise when viewed from the top. The atom above the plane is prefixed as a superscript, while the one below is suffixed as a subscript (e.g., $^4C_1$). Pyranose and furanose forms of glycans can exist as different conformers and can inter-convert if an energy penalty is met.

I.2.a. Furanoses

Pentoses like ribose form five-membered furanose rings. Upon cyclization, the C1 carbon becomes chiral giving rise to either an α or a β anomer. For furanoses, the two possible conformers are the Twist (T) and the Envelope (E). The conformational flexibility of the furanose ring is represented by the puckering parameters as described in **Figure I.3**. The figure describes the amplitude **ν** (or the endocyclic torsion angles, ν_0, ν_1, ν_2, ν_3 and ν_4) of the furanose ring. Each point on the circle represents a specific phase (or pseudorotation[1]) angle **P**. The molecule can move from one conformation to another on the circle, crossing an intermediate low energy barrier.



E2
3T2
270
3E
3T4
1T2
1E
E4
1T0
0T4
E0
180
N
W
E
S
P
0
0E
4T0
ν
0T1
4E
1E
90
2T1
5
4T3
E3
2T3
2E
4
1
3
2
O
P (or ϕ) = phase angle or pseudorotational angle
ν (or Q) = puckering amplitude or endocyclic torsional angle

Figure I.3 The schematic representation of the puckering parameters in a furanose ring [3]. The pseudo-rotational wheel of furanoses encompasses the 20 twist and envelope shapes.

I.2.b. Pyranoses

Hexoses such as glucose, on the other hand, are mostly found in the six-membered pyranose ring forms, as illustrated in **Figure I.4**. The chair [**C**] is the representative conformation for most pyranoses with $^4C_1$ and $^1C_4$ being the most-favored. In this conformation the ring oxygen O5 and the ring carbons C2, C3 and C5 lie on the same plane. Depending on whether the C4 is above the plane (with C1 below the plane) or vice-versa the sugar is said to be in either $^4C_1$ or $^1C_4$ conformation, respectively. Generally, the D- and the L- pyranose rings have a preference for the $^4C_1$ and $^1C_4$

[1] Pseudorotation is defined as the ready change of the flexible conformations from one form to another. This term is used frequently for furanoses.

conformations, respectively. Beside the chair, boat (or twist boat) and skew conformations can also be observed for pyranose rings. Usually a ring conformation is stable with one predominant form, but it does happen that the ring flips to another alternative conformation. This requires considerable energy since high-energy barriers separate the low-energy conformational states. The energy barrier for the inter-conversion of one chair conformation into another amounts to about 42kJ/mol [4]. This rare situation may arise if there are bulky substituents at the axial position. The ring flip involves a large conformational change due to all the axial groups becoming equatorial and vice-versa. It is assumed that the ring inversion via the high-energy half-chair conformation, along with the skew and boat conformations [4]. This factor is thus important to consider when molecular modeling involves a sugar with such a behavioral trait of flexible rings.

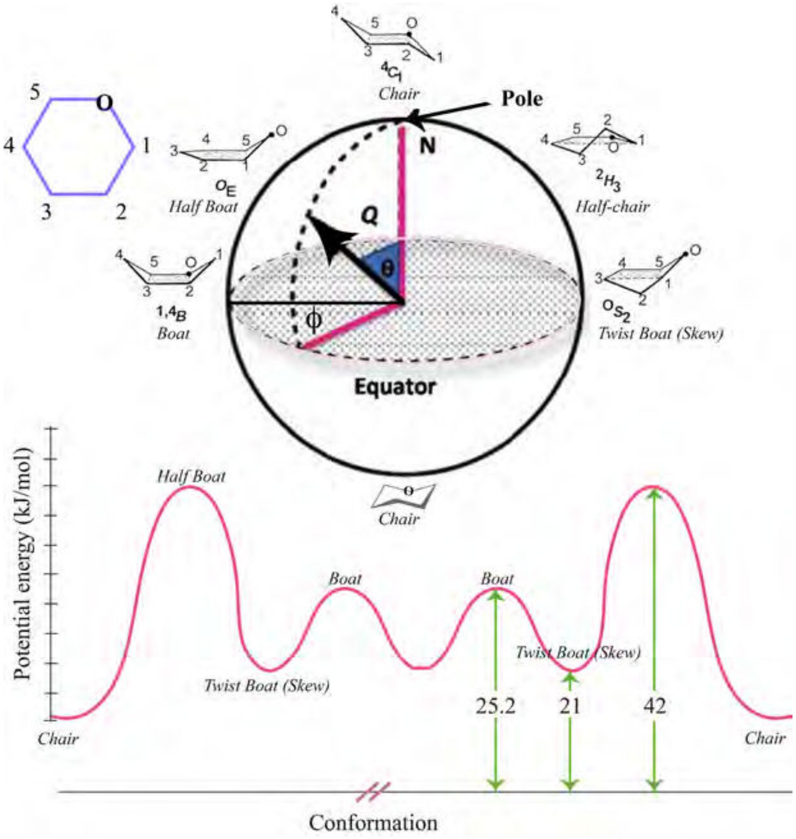


Figure I.4 (a) The puckering parameters for pyranoses describing the polar angle (θ), the azimuthal or phase angle (φ) and the puckering amplitude (Q) [3]. [The phase angle is also denoted sometimes as *P* and the puckering amplitude as ν] (b) Energy barriers for the interconversion of one chair conformation into another chair conformation in a pyranose ring system [4].

I.3. Exocyclic hydroxymethyl groups

The exocyclic primary alcohol groups can adopt a number of low-energy conformations. The conformation of the exocyclic hydroxymethyl groups is best described as an equilibrium that exists between three staggered rotamers that correspond to their local minima. In furanoses and pyranoses, they are called ***gt*** (*gauche-trans*), ***gg*** (*gauche-gauche*), and ***tg*** (*trans-gauche*). The corresponding torsion angle (**ω**) between the terminal oxygen and the ring oxygen in α-D-Glc*p*, for example, is shown in **Figure I.5**. In pyranoses, the two most frequent positions occupied by the primary hydroxyl groups correspond to those that avoid interactions between O4 and O6 [3]. However, each of the secondary hydroxyl groups can rotate almost freely.

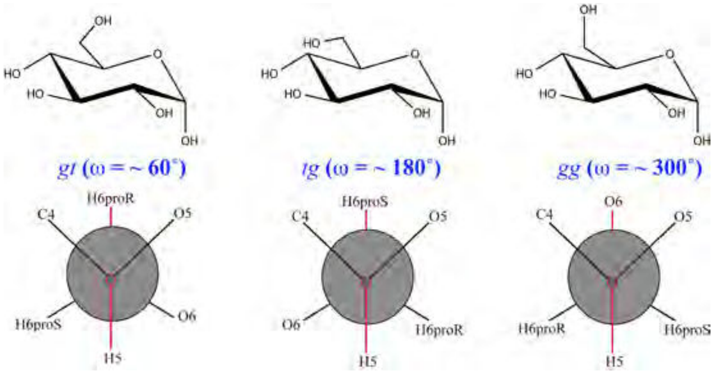


Figure I.5: The ring and Newman projections of the ***gt*** (*gauche-trans*), ***tg*** (*trans-gauche*) and ***gg*** (*gauche-gauche*) rotameric conformers of the **ω (→**O5–C5–C6–O6) torsion angle in α-D-Glc*p*.

I.4. Glycosidic linkages

The monosaccharides are connected to build longer chains of di-, oligo and polysaccharides via a condensation reaction. This occurs at the –OH function at the anomeric carbon with the hydroxyl of another monosaccharide, with the elimination of a water molecule to form an acetal. Thus, a glycosidic linkage is formed (**Figure I.1**).

I.4. Torsion Angles

The relative orientation of two consecutive monosaccharides linked by a glycosidic bond in a disaccharide can be characterized by the Φ and Ψ torsion angles. In this thesis, the 'Heavy Atom Definition' (commonly followed by crystallographers) has been used where Φ→O5-C1-O-C_x and Ψ→C1-O1-C_x-C_{x+1} (where x is the number of the carbon atom of the second monosaccharide with which the glycosidic bond is formed), except in **Chapter 5** and **Chapter 6**, where the 'Light Atom Definition' has also been used for describing the NMR results (**Figure 1.6**). For 1→6 glycosidic linkages, the ω (O5–C5–C6–O6) torsion angle is another parameter that is important. In analogy to the Ramachandran plots generated for proteins, in glycobiology Φ / Ψ maps can be generated using molecular mechanics or molecular dynamics calculations. The minima on these φ/ψ maps describe the energetically preferred disaccharide conformations. For oligosaccharides containing 1→ 6 linkages, variation in rotamer population have a direct effect on the glycan structure and function. Water seems to induce the disruption of hydrogen bonds within the glycan thus allowing rotamer populations to be determined by internal electronic and steric repulsions between the oxygen atoms [5].

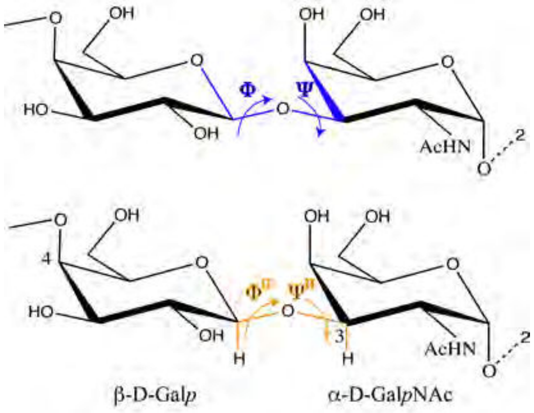


Figure I.6: The conventions used in this thesis illustrated using the disaccharide β-D-Gal*p*-1,3- α-D-Gal*p*NAc (a) Heavy atom convention: Φ→O5-C1-O1-C4, Ψ→ C1-C2-O1-C4, (b) Light atom convention: Φ^H→ H1-C1-O1-Cx and Ψ^H → C1-O1-Cx-Hx.

I.5. Reducing and non-reducing ends

The hemiacetal or hemiketal functions in the disaccharides and higher carbohydrate species retain the aldehyde or the ketone functional group and hence the ability to reduce inorganic ions. Thus, this part of the sugar is called the reducing end [6]. The non-reducing end is at the monomer whose anomeric carbon is engaged in a glycosidic bond, thus preventing the opening of the ring to the aldehyde or keto form.

Oligosaccharides are flexible molecules containing several freely rotatable bonds. Glycans are also difficult to model due to their highly polar functionality and differences in electronic arrangements, like the anomeric, exo-anomeric and the gauche[2] effects that occur during conformational and configurational changes [3].

I.7. Anomeric and exo-anomeric effect

Anomers are special epimers[3], which in the cyclic forms differ in chirality at the anomeric (hemi-acetal or hemiketal) carbon only. In the straight-chain format, anomers have an identical configuration. When the stereochemistry at C1 matches that of the last stereogenic center, the sugar is an alpha (α) anomer and when they are oppositely oriented then the sugar is a beta (β) anomer. The anomers can inter-convert through **mutarotation**[4]. The content of the α-anomer in the equilibrium mixture following mutarotation is observed to be greater than what could be explained by the conformational energy of the –OH group, for example, in α-D-methyl-glucopyranoside the α-anomer predominates with a ratio of 2:1 (the α-anomer has an axial –OH group while the β-anomer has an equatorially oriented –OH group at C1). This tendency of the α-anomer to predominate is called the anomeric effect.

Anomeric effect, also known as the endo-anomeric effect, is the propensity of heteroatoms at C1 to be oriented axially. The lone pair being donated comes from the ring oxygen.

Exo-anomeric effect is similar to the endo-anomeric effect, with the source of the lone pair of electron that is donated being different. The lone pair in the exo-anomeric effect is donated by the substituent at C1. This results from the interaction of the lone pairs on the exocyclic oxygen with the endocyclic C-O bond. This interaction is favorable only in the *gauche* conformer about the exocyclic C-O bond. Thus, the anomeric effect not only

[2] The term "gauche" refers to conformational isomers (conformers) where two vicinal groups are separated by a 60° torsion angle. IUPAC defines groups as gauche if they have a "synclinal alignment of groups attached to adjacent atoms".

[3] Epimers are monosaccharides differing in chirality at only one carbon. In the straight-chain format, epimers have –H and –OH switched at one backbone carbon, but not at any other.

[4] Mutarotation is the change in the optical rotation that occurs by epimerization (i.e. change in the equilibrium between two epimers, when the corresponding stereocenters interconvert). Cyclic sugars show mutarotation as α and β anomers interconvert. The optical rotation of the solution depends on the optical rotation of each anomer and their ratio in the solution.

influences the axial/equatorial isomerism in sugars, but also has a strong influence on the conformation of oligosaccharides.

Gauche effect The **Gauche effect** characterizes any gauche rotamer, which is actually more stable than the anti rotamer. Though it is recognized that the *gauche* effect in glycans is a solvent-dependent phenomenon, the mechanism through which this effect is induced as well as the physical origin of such conformational preferences remain unknown [7].

References:

1. Barnett CB, Naidoo KJ: **Ring Puckering: A metric for evaluating the accuracy of AM1, PM3, PM3CARB-1, and SCC-DFTB carbohydrate QM/MM simulations**. *The Journal of Physical Chemistry B* 2010, **114**(51):17142-17154.

2. Cremer D, Pople JA: **General definition of ring puckering coordinates**. *Journal of the American Chemical Society* 1975, **97**(6):1354-1358.

3. Pérez S: **Molecular modeling in glycoscience**. In: *Comprehensive Glycosciences: Analysis of Glycans.* Edited by Kamerling JP, vol. 2; 2007.

4. Mackie W: **Carbohydrates structure and biology; Edited by Jochen Lehmann translated by Alan Haines. pp 274. Georg Thieme, Stuttgart. 1998. DM 85 ISBN 3-13-110771-5**. *Biochemical Education* 1998, **26**(3):257-257.

5. Kirschner KN, Woods RJ: **Solvent interactions determine carbohydrate conformation**. *Proceedings of the National Academy of Sciences* 2001, **98**(19):10541-10545.

6. Taylor ME, Drickamer K: **Introduction to glycobiology / Maureen E. Taylor, Kurt Drickamer**, 2nd ed edn: Oxford University Press; 2006.

7. Kirschner KN, Woods RJ: **Solvent interactions determine carbohydrate conformation**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10541-10545.

Annex II

Table 1

Polysaccharide Families, their constituent members and the bibliographic references present in PolySac3DB

No.	Family Name	Polysaccharide Member	Reference
1	Agaroses	Agarose (single)	[1]
		Agarose (double)	[2]
		Agarose Molecular Models	[3]
2	Alginates	Poly-α-L-Guluronic Acid	[4]
		Poly-β-D-Mannuronic Acid	[5]
		Alginate Molecular Models	[6, 7]
3	Amyloses & Starch	A Starch	[8, 9]
		Starch Nanocrystals	[10]
		Amylopectins	[11]
		B Starch	[12]
		Amylose DMSO	[13]
		Amylose KOH	[14]
		Amylose Triacetate	[15]
		Amylose tri-O-ethyl (TEA3)	[16]
		Amylose V	[17]
		Amylose V propanol complex	[18]
4	Bacterial Polysaccharides	Dextran (high T polymorph)	[19]
		Dextran (low T polymorph)	[20]
		Exo-polysaccharide (*Burkholderia cepacia*)	[21]
		α (2-8)-linked Sialic Acid Polysaccharide	[22]
		M41 Capsular Polysaccharide (*E. coli*)	[23]
		O-antigenic polysaccharide (*E. coli* 1303)	Manuscript in preparation
		O-antigenic polysaccharide (*E. coli* O5ab)	Manuscript in preparation

		O-antigenic polysaccharide (*E. coli* O5ac)	Manuscript in preparation
		O-antigenic polysaccharide (*E. coli* O65)	Manuscript in preparation
		Capsular Polysaccharide (*Rhizobium trifolii*)	[24]
		Gellan Native K	[25]
		Gellan K	[26]
		Gellan Li	[27]
		RMDP17	[28]
		Welan (Ca)	[29]
		Xanthan	[30]
5	Carrageenans	Iota Carrageenan	[31]
		Iota Carrageenan (Na salt)	[32]
		Kappa Carrageenan	[33]
6	Celluloses	Cellulose I α	[34]
		Cellulose I β	[35]
		Cellulose I triacetate	[36]
		Cellulose II	[37]
		Cellulose II hydrate	[38]
		Cellulose II hydrazine	[39]
		Cellulose II triacetate	[40]
		Cellulose III$_I$	[41]
		Cellulose IV$_I$	[42]
		Cellulose microfibrils	[43]
7	Chitins & Chitosans	Chitin I (Chitin β)	[44]
		Chitin II (Chitin α)	[45]
		Chitosan (anhydrous)	[46]
		Chitosan (high T Polymorph)	[47]
8	Curdlans	Curdlan I (Native)	[48]
		Curdlan II	[49]
		Curdlan III	[50]
9	GAGs	Chondroitin (unsulphated)	[51]

		Chondroitin 4-sulphate Ca	[52]
		Chondroitin 4-sulphate K	[53]
		Chondroitin 4-sulphate Na	[54]
		Dermatan 4-sulphate Na (allomorphs I, II, III)	[55]
		Hyaluronate I & II Sodium	[56]
		Hyaluronate III Sodium	[57]
		Hyaluronate I Potassium	[58]
		Hyaluronate II Potassium	[59]
		Hyaluronate III Potassium	[60]
		Hyaluronate Calcium	[61]
		Hyaluronic acid	[62]
		Heparin (dp 12) Heparin (dp 18, 24, 30, 36)	[63, 64]
		Keratan-6-sulphate	[65]
10	Galactoglucans	Galactoglucan	[66]
11	Galactomannans	Galactomannan	[67]
12	Glucomannans	Konjac glucomannan	[68]
13	Mannans	Mannan I	[69]
		Mannan II	[70]
		α-D-1,3-Mannan	[71]
		Mannan dihydrate	[72]
14	Pectins	Pectic Acid	[73]
		Calcium Pectate	[74]
		Sodium Pectate	[73]
		Polyuronides Molecular Models	[6, 7]
		Arabinan	[75]
		Arabino-Galactan Type I	[75]
		Arabino-Galactan Type II	[75]
		RG-I	[75]
		RG-II	[76]

15	Scleroglucans	Scleroglucan	[77]
16	Xylans	Xylan (β-1,3)	[78]
		Xylan (β-1,4)	[79, 80]
17	Nigeran	Nigeran	[81]
18	Others	Inulin hemihydrate	[82]
		Inulin monohydrate	[82]
		α-D-glucan	[83]
		α-1,3-glucan triacetate	[84]

References:

1. Foord SA, Atkins EDY: **New X-ray diffraction results from agarose: Extended single helix structures and implications for gelation mechanism**. *Biopolymers* 1989, **28**(8):1345-1365.

2. Arnott S, Fulmer A, Scott WE, Dea IC, Moorhouse R, Rees DA: **The agarose double helix and its function in agarose gel structure**. *J Mol Biol* 1974, **90**(2):269-284.

3. Kouwijzer M, Pérez S: **Molecular modeling of agarose helices, leading to the prediction of crystalline allomorphs**. *Biopolymers* 1998, **46**(1):11-29.

4. Atkins EDT, Nieduszynski IA, Mackie W, Parker KD, Smolko EE: **Structural components of alginic acid. II. The crystalline structure of poly-α-L-guluronic acid. Results of X-ray diffraction and polarized infrared studies**. *Biopolymers* 1973, **12**(8):1879-1887.

5. Atkins EDT, Nieduszynski IA, Mackie W, Parker KD, Smolko EE: **Structural components of alginic acid. I. The crystalline structure of poly-β-D-mannuronic acid. Results of X-ray diffraction and polarized infrared studies**. *Biopolymers* 1973, **12**(8):1865-1878.

6. Braccini I, Grasso RP, Pérez S: **Conformational and configurational features of acidic polysaccharides and their interactions with calcium ions: a molecular modeling investigation**. *Carbohydrate Research* 1999, **317**(1-4):119-130.

7. Braccini I, Pérez S: **Molecular basis of Ca^{2+}-induced gelation in alginates and pectins: the egg-box model revisited**. *Biomacromolecules* 2001, **2**(4):1089-1096.

8. Imberty A, Chanzy H, Pérez S, Buleon A, Tran V: **The double-helical nature of the crystalline part of A-starch**. *J Mol Biol* 1988, **201**(2):365-378.

9. Popov D, Buléon A, Burghammer M, Chanzy H, Montesanti N, Putaux JL, Potocki-Véronèse G, Riekel C: **Crystal structure of A-amylose: A revisit from synchrotron microdiffraction analysis of single crystals**. *Macromolecules* 2009, **42**(4):1167-1174.

10. Pérez S, Bertoft E: **The molecular structures of starch components and their contribution to the architecture of starch granules: A comprehensive review**. *Starch - Stärke* 2010, **62**(8):389-420.

11. O'Sullivan AC, Pérez S: **The relationship between internal chain length of amylopectin and crystallinity in starch**. *Biopolymers* 1999, **50**(4):381-390.

12. Imberty A, Pérez S: **A revisit to the three-dimensional structure of B-type starch**. *Biopolymers* 1988, **27**(8):1205-1221.

13. Winter WT, Sarko A: **Crystal and molecular structure of the amylose-DMSO complex**. *Biopolymers* 1974, **13**(7):1461-1482.

14. Sarko A, Biloski A: **Crystal structure of the KOH-amylose complex**. *Carbohydrate Research* 1980, **79**(1):11-21.

15. Sarko A, Marchessault RH: **Crystalline structure of amylose triacetate I. Stereochemical approach**. *Journal of the American Chemical Society* 1967, **89**(25):6454-6462.

16. Bluhm TL, Zugenmaier P: **The crystal and molecular structure of tri-o-ethylamylose (TEA 3)**. *Carbohydrate Research* 1979, **68**(1):15-21.

17. Winter WT, Sarko A: **Crystal and molecular structure of V-anhydrous amylose**. *Biopolymers* 1974, **13**(7):1447-1460.

18. Nishiyama Y, Mazeau K, Morin M, Cardoso MB, Chanzy H, Putaux J-L: **Molecular and crystal structure of 7-fold V-amylose complexed with 2-propanol**. *Macromolecules* 2010, **43**(20):8628-8636.

19. Guizard C, Chanzy H, Sarko A: **Molecular and crystal structure of dextrans: a combined electron and X-ray diffraction study. 1. The anhydrous, high-temperature polymorph**. *Macromolecules* 1984, **17**(1):100-107.

20. Guizard C, Chanzy H, Sarko A: **The molecular and crystal structure of dextrans: a combined electron and X-ray diffraction study. II. A low temperature, hydrated polymorph**. *J Mol Biol* 1985, **183**(3):397-408.

21. Strino F, Nahmany A, Rosen J, Kemp GJL, Sá-correia I, Nyholm P-G: **Conformation of the exopolysaccharide of *Burkholderia cepacia* predicted with molecular mechanics (MM3) using genetic algorithm search**. *Carbohydrate Research* 2005, **340**(5):1019-1024.

22. Brisson JR, Baumann H, Imberty A, Pérez S, Jennings HJ: **Helical epitope of the group B meningococcal alpha(2-8)-linked sialic acid polysaccharide**. *Biochemistry* 1992, **31**(21):4996-5004.

23. Moorhouse R, Winter WT, Arnott S, Bayer ME: **Conformation and molecular organization in fibers of the capsular polysaccharide from *Escherichia coli* M41 mutant**. *J Mol Biol* 1977, **109**(3):373-391.

24. Lee EJ, Chandrasekaran R: **The "pseudo double-helical" structure of the gel-forming capsular polysaccharide from *Rhizobium trifolii***. *Carbohydrate Research* 1992, **231**:171-183.

25. Chandrasekaran R, Radha A, Thailambal VG: **Roles of potassium ions, acetyl and L-glyceryl groups in native gellan double helix: an X-ray study**. *Carbohydr Res* 1992, **224**:1-17.

26. Chandrasekaran R, Puigjaner LC, Joyce KL, Arnott S: **Cation interactions in gellan: An X-ray study of the potassium salt**. *Carbohydrate Research* 1988, **181**:23-40.

27. Chandrasekaran R, Millane RP, Arnott S, Atkins EDT: **The crystal structure of gellan**. *Carbohydrate Research* 1988, **175**(1):1-15.

28. Bian W, Chandrasekaran R, Rinaudo M: **Molecular structure of the rhamsan-like exocellular polysaccharide RMDP17 from *Sphingomonas paucimobilis***. *Carbohydrate Research* 2002, **337**(1):45-56.

29. Chandrasekaran R, Radha A, Lee EJ: **Structural roles of calcium ions and side chains in welan: an X-ray study**. *Carbohydr Res* 1994, **252**:183-207.

30. Moorhouse R, Walkinshaw M D, Arnott S: **Xanthan gum - Molecular conformation and interactions**. In: *Extracellular Microbial Polysaccharides.* WASHINGTON, D. C.: American Chemical Society; 1977: 90-102.

31. Arnott S, Scott WE, Rees DA, McNab CG: **Iota-carrageenan: molecular structure and packing of polysaccharide double helices in oriented fibres of divalent cation salts**. *J Mol Biol* 1974, **90**(2):253-267.

32. Janaswamy S, Chandrasekaran R: **Three-dimensional structure of the sodium salt of iota-carrageenan**. *Carbohydrate Research* 2001, **335**(3):181-194.

33. Millane RP, Chandrasekaran R, Arnott S, Dea ICM: **The molecular structure of kappa-carrageenan and comparison with iota-carrageenan**. *Carbohydrate Research* 1988, **182**(1):1-17.

34. Nishiyama Y, Sugiyama J, Chanzy H, Langan P: **Crystal structure and hydrogen bonding system in cellulose Iα from synchrotron X-ray and neutron fiber diffraction**. *Journal of the American Chemical Society* 2003, **125**(47):14300-14306.

35. Nishiyama Y, Langan P, Chanzy H: **Crystal structure and hydrogen-bonding system in cellulose Iβ from synchrotron X-ray and neutron fiber diffraction**. *Journal of the American Chemical Society* 2002, **124**(31):9074-9082.

36. Stipanovic AJ, Sarko A: **Molecular and crystal structure of cellulose triacetate I: A parallel chain structure**. *Polymer* 1978, **19**(1):3-8.

37. Langan P, Nishiyama Y, Chanzy H: **X-ray structure of mercerized cellulose II at 1 Å resolution**. *Biomacromolecules* 2001, **2**(2):410-416.

38. David ML, John B: **Structure of cellulose II hydrate**. *Biopolymers* 1981, **20**(10):2165-2179.

39. David ML, John B, Litt MH: **Structure of a cellulose II-hydrazine complex**. *Biopolymers* 1983, **22**(5):1383-1399.

40. Roche E, Chanzy H, Boudeulle M, Marchessault RH, Sundararajan P: **Three-dimensional crystalline structure of cellulose triacetate II**. *Macromolecules* 1978, **11**(1):86-94.

41. Wada M, Chanzy H, Nishiyama Y, Langan P: **Cellulose III$_1$ crystal structure and hydrogen bonding by synchrotron X-ray and neutron fiber diffraction**. *Macromolecules* 2004, **37**(23):8548-8555.

42. Gardiner ES, Sarko A: **Packing analysis of carbohydrates and polysaccharides. 16. The crystal structures of cellulose-IV$_I$ and cellulose-IV$_{II}$**. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* 1985, **63**(1):173-180.

43. Pérez S, Samain D: **Structure and engineering of celluloses**. In: *Advances in carbohydrate chemistry and biochemistry.* Edited by Derek H, vol. Volume 64: Academic Press; 2010: 25-116.

44. Gardner KH, Blackwell J: **Refinement of the structure of β-chitin**. *Biopolymers* 1975, **14**(8):1581-1595.

45. Carlstrom D: **The crystal structure of α-chitin (Poly-N-Acetyl-D-Glucosamine)**. *J Cell Biol* 1957, **3**(5):669-683.

46. Yui T, Imada K, Okuyama K, Obata Y, Suzuki K, Ogawa K: **Molecular and crystal-structure of the anhydrous form of chitosan**. *Macromolecules* 1994, **27**(26):7601-7605.

47. Mazeau K, Winter WT, Chanzy H: **Molecular and crystal structure of a high-temperature polymorph of chitosan from electron diffraction data**. *Macromolecules* 1994, **27**(26):7606-7612.

48. Okuyama K, Otsubo A, Fukuzawa Y, Ozawa M, Harada T, Kasai N: **Single-helical structure of native curdlan and its aggregation state**. *Journal of Carbohydrate Chemistry* 1991, **10**(4):645 - 656.

49. Chuah CT, Sarko A, Deslandes Y, Marchessault RH: **Packing analysis of carbohydrates and polysaccharides. Part 14. Triple-helical crystalline structure of curdlan and paramylon hydrates**. *Macromolecules* 1983, **16**(8):1375-1382.

50. Deslandes Y, Marchessault RH, Sarko A: **Triple-helical Structure of (1,3)-β-D-glucan**. *Macromolecules* 1980, **13**(6):1466-1471.

51. Sattelle BM, Shakeri J, Roberts IS, Almond A: **A 3D-structural model of unsulfated chondroitin from high-field NMR: 4-sulfation has little effect on backbone conformation**. *Carbohydrate Research* 2010, **345**(2):291-302.

52. Cael JJ, Winter WT, Arnott S: **Calcium chondroitin 4-sulfate: molecular conformation and organization of polysaccharide chains in a proteoglycan**. *J Mol Biol* 1978, **125**(1):21-42.

53. Millane RP, Mitra AK, Arnott S: **Chondroitin 4-sulfate: Comparison of the structures of the potassium and sodium salts**. *Journal of Molecular Biology* 1983, **169**(4):903-920.

54. Winter WT, Arnott S, Isaac DH, Atkins ED: **Chondroitin 4-sulfate: the structure of a sulfated glycosaminoglycan**. *J Mol Biol* 1978, **125**(1):1-19.

55. Mitra AK, Arnott S, Atkins ED, Isaac DH: **Dermatan sulfate: molecular conformations and interactions in the condensed state**. *J Mol Biol* 1983, **169**(4):873-901.

56. Guss JM, Hukins DW, Smith PJ, Winter WT, Arnott S: **Hyaluronic acid: molecular conformations and interactions in two sodium salts**. *J Mol Biol* 1975, **95**(3):359-384.

57. Winter WT, Smith PJ, Arnott S: **Hyaluronic acid: structure of a fully extended 3-fold helical sodium salt and comparison with the less extended 4-fold helical forms**. *J Mol Biol* 1975, **99**(2):219-235.

58. Mitra AK, Raghunathan S, Sheehan JK, Arnott S: **Hyaluronic acid: molecular conformations and interactions in the orthorhombic and tetragonal forms containing sinuous chains**. *J Mol Biol* 1983, **169**(4):829-859.

59. Mitra AK, Arnott S, Sheehan JK: **Hyaluronic acid: molecular conformation and interactions in the tetragonal form of the potassium salt containing extended chains**. *J Mol Biol* 1983, **169**(4):813-827.

60. Arnott S, Mitra AK, Raghunathan S: **Hyaluronic acid double helix**. *J Mol Biol* 1983, **169**(4):861-872.

61. Winter WT, Arnott S: **Hyaluronic acid: the role of divalent cations in conformation and packing**. *J Mol Biol* 1977, **117**(3):761-784.

62. Haxaire K, Braccini I, Milas M, Rinaudo M, Pérez S: **Conformational behavior of hyaluronan in relation to its physical properties as probed by molecular modeling**. *Glycobiology* 2000, **10**(6):587-594.

63. Khan S, Gor J, Mulloy B, Perkins SJ: **Semi-rigid solution structures of heparin by constrained X-ray scattering modelling: new insight into heparin-protein complexes**. *Journal of Molecular Biology* 2010, **395**(3):504-521.

64. Mulloy B, Forster MJ, Jones C, Davies DB: **N.M.R. and molecular-modelling studies of the solution conformation of heparin**. *The Biochemical journal* 1993, **293 (Pt 3)**:849-858.

65. Arnott S, Gus JM, Hukins DW, Dea IC, Rees DA: **Conformation of keratan sulphate**. *J Mol Biol* 1974, **88**(1):175-184.

66. Chandrasekaran R, Lee EJ, Thailambal VG, Zevenhuizen LPTM: **Molecular architecture of a galactoglucan from *Rhizobium meliloti***. *Carbohydrate Research* 1994, **261**(2):279-295.

67. Chandrasekaran R, Radha A, Okuyama K: **Morphology of galactomannans: an X-ray structure analysis of guaran**. *Carbohydrate Research* 1998, **306**(1-2):243-255.

68. Yui T, Ogawa K, Sarko A: **Molecular and crystal structure of konjac glucomannan in the mannan II polymorphic form**. *Carbohydr Res* 1992, **229**(1):41-55.

69. Chanzy H, Pérez S, Miller DP, Paradossi G, Winter WT: **An electron diffraction study of the mannan I crystal and molecular structure**. *Macromolecules* 1987, **20**(10):2407-2413.

70. Millane RP, Hendrixson TL: **Crystal structures of mannan and glucomannans**. *Carbohydrate Polymers* 1994, **25**(4):245-251.

71. Yui T, Ogawa K, Sarko A: **Molecular and crystal structure of the regenerated form of (1 → 3)-α-D-mannan**. *Carbohydrate Research* 1992, **229**(1):57-74.

72. Ogawa K, Miyanishi T, Yui T, Hara C, Kiho T, Ukai S, Sarko A: **X-Ray diffraction study on (1 →3)-α-D-mannan dihydrate** *Carbohydrate Research* 1986, **148**(1):115-120.

73. Walkinshaw MD, Arnott S: **Conformation and interactions of pectins. I. X-ray diffraction analyses of sodium pectate in neutral and acidified forms**. *J Mol Biol* 1981, **153**(4):1055-1073.

74. Walkinshaw MD, Arnott S: **Conformations and interactions of pectins. II. Models for junction zones in pectinic acid and calcium pectate gels**. *J Mol Biol* 1981, **153**(4):1075-1085.

75. Engelsen SB, Cros S, Mackie W, Pérez S: **A molecular builder for carbohydrates: Application to polysaccharides and complex carbohydrates**. *Biopolymers* 1996, **39**(3):417-433.

76. Pérez S, Rodríguez-Carvajal MA, Doco T: **A complex plant cell wall polysaccharide: rhamnogalacturonan II. A structure in quest of a function**. *Biochimie* 2003, **85**(1-2):109-121.

77. Bluhm TL, Deslandes Y, Marchessault RH, Pérez S, Rinaudo M: **Solid-state and solution conformation of scleroglucan**. *Carbohydrate Research* 1982, **100**(1):117-130.

78. Atkins EDT, Parker KD: **The helical structure of a β-D-1,3-xylan**. *Journal of Polymer Science Part C: Polymer Symposia* 1969, **28**(1):69-81.

79. Nieduszynski I, Marchessault RH: **Structure of β-D-(1,4')-xylan hydrate**. *Nature* 1971, **232**(5305):46-47.

80. Nieduszynski IA, Marchessault RH: **Structure of β-D-(1,4')-xylan hydrate**. *Biopolymers* 1972, **11**(7):1335-1344.

81. Pérez S, Roux M, Revol JF, Marchessault RH: **Dehydration of nigeran crystals: crystal structure and morphological aspects**. *J Mol Biol* 1979, **129**(1):113-133.

82. André I, Mazeau K, Tvaroska I, Putaux JL, Winter WT, Taravel FR, Chanzy H: **Molecular and crystal structures of inulin from electron diffraction data**. *Macromolecules* 1996, **29**(13):4626-4635.

83. Ogawa K, Okamura K, Sarko A: **Molecular and crystal structure of the regenerated form of (1 →3)-α-D-glucan**. *International journal of biological macromolecules* 1981, **3**(1):31-36.

84. Yui T, Sarko A: **Molecular and crystal structure of (1,3)-α-D-glucan triacetate**. *International journal of biological macromolecules* 1992, **14**(2):87-96.

Annex III

List of Bioactive Oligosaccharides in BiOligo

Index	Oligo	Formula	Type
1	Blood group A antigen tetraose type 1	GalNAc a1-3 (Fuc a1-2) Gal b1-3 GlcNAc b	Blood group A antigens
2	Blood group A antigen hexaose type 1	GalNAc a1-3 (Fuc a1-2) Gal b1-3 GlcNAc b1-3 Gal b1-4 Glc	Blood group A antigens
3	Blood group A antigen hexaose type 2	GalNAc a1-3 (Fuc a1-2) Gal b1-4 GlcNAc b1-3 Gal b1-4 Glc	Blood group A antigens
4	Blood group A antigen pentaose type 1	GalNAc a1-3 (Fuc a1-2) Gal b1-3 GlcNAc b1-3 Gal	Blood group A antigens
5	Blood group A antigen pentaose type 2	GalNAc a1-3 (Fuc a1-2) Gal b1-4 GlcNAc b1-3 Gal	Blood group A antigens
6	Blood group A antigen pentaose type 4	GalNAc a1-3 (Fuc a1-2) Gal b1-3 GalNAc b1-3 Gal	Blood group A antigens
7	Blood group A antigen tetraose type 2	GalNAc a1-3 (Fuc a1-2) Gal b1-4 GlcNAc b	Blood group A antigens
8	Blood group A antigen tetraose type 5	GalNAc a1-3 (Fuc a1-2) Gal b1-4 Glc	Blood group A antigens
9	Blood group A antigen triose	GalNAc a1-3 (Fuc a1-2) Gal b	Blood group A antigens
10	Blood group A Lewis B antigen pentaose type1	GalNAc a1-3 (Fuc a1-2) Gal b1-3 (Fuc a1-4) GlcNAc	Blood group A antigens
11	Blood group A Lewis Y antigen pentaose type2	GalNAc a1-3 (Fuc a1-2) Gal b1-4 (Fuc a1-3) GlcNAc	Blood group A antigens
12	B antigen hexaose type 1	Gal a1-3 (Fuc a1-2) Gal b1-3 GlcNAc b1-3 Gal b1-4 Glc	Blood group B antigens
13	B antigen hexaose type 2	Gal a1-3 (Fuc a1-2) Gal b1-4 GlcNAc b1-3 Gal b1-4 Glc	Blood group B antigens
14	Blood group B antigen pentaose type 1	Gal a1-3 (Fuc a1-2) Gal b1-3 GlcNAc b1-3 Gal	Blood group B antigens
15	Blood group B antigen pentaose type 2	Gal a1-3 (Fuc a1-2) Gal b1-4 GlcNAc b1-3 Gal	Blood group B antigens
16	Blood group B antigen pentaose type 4	Gal a1-3 (Fuc a1-2) Gal b1-3 GalNAc b1-3 Gal	Blood group B antigens
17	Blood group B antigen tetraose type 1	Gal a1-3 (Fuc a1-2) Gal b1-3 GlcNAc b	Blood group B antigens
18	Blood group B antigen tetraose type 2	Gal a1-3 (Fuc a1-2) Gal b1-4 GlcNAc	Blood group B antigens
19	Blood group B antigen tetraose type 5	Gal a1-3 (Fuc a1-2) Gal b1-4 Glc b	Blood group B antigens
20	Blood group B antigen triose	Gal a1-3 (Fuc a1-2) Gal b	Blood group B antigens
21	Blood group B Lewis Y antigen pentaose type1	Gal a1-3 (Fuc a1-2) Gal b1-3 (Fuc a1-4) GlcNAc	Blood group B antigens
22	Blood group B Lewis Y antigen pentaose type2	Gal a1-3 (Fuc a1-2) Gal b1-4 (Fuc a1-3) GlcNAc	Blood group B antigens
23	Blood group H antigen pentaose type 2	Fuc a1-2 Gal b1-4 GlcNAc b1-3 Gal b1-4 Glc	Blood group H antigens (Blood group O)
24	Blood group H antigen tetraose type 1	Fuc a1-2 Gal b1-3 GlcNAc b1-3 Gal	Blood group H antigens (Blood group O)
25	Blood group H antigen tetraose type 2	Fuc a1-2 Gal b1-4 GlcNAc b1-3 Gal	Blood group H antigens (Blood

			group O)
26	Blood group H antigen triaose type 2	Fuc a1-2 Gal b1-4 GlcNAc b	Blood group H antigens (Blood group O)
27	Blood group H antigen triose type 1	Fuc a1-2 Gal b1-3 GlcNAc b	Blood group H antigens (Blood group O)
28	Blood group H antigen triose type 3	Fuc a1-2 Gal b1-3 GalNAc a	Blood group H antigens (Blood group O)
29	Blood group H antigen triose type 4	Fuc a1-2 Gal b1-3 GalNAc b	Blood group H antigens (Blood group O)
30	Blood group H antigen triose type 5	Fuc a1-2 Gal b1-4 Glc	Blood group H antigens (Blood group O)
31	Blood group H antigen triose type 5 (2'-Fucosyllactose)	Fuc a1-2 Gal b1-4 Glc b	Blood group H antigens (Blood group O)
32	Blood group H antigen triose type 6	Fuc a1-2 Gal b1-4 Glc b	Blood group H antigens (Blood group O)
33	Lacto-N-fucopentaose I	Fuc a1-2 Gal b1-3 GlcNAc b1-3 Gal b1-4 Glc b	Blood group H antigens (Blood group O)
34	LNT-2	GlcNAc b1-3 Gal b1-4 Glc b	Blood group H antigens (Blood group O)
35	Blood group H antigen tetraose type 4 & Globo H tetraose	Fuc a1-2 Gal b1-3 GalNAc b1-3 Gal	Blood group H antigens (Blood group O)
36	Core 4	GlcNAc b1-3 (GlcNAc b1-6) GalNAc a	Core structures
37	a2-3 Neu5Ac on Core 1 of Core 2	Neu5Ac a2-3 Gal b1-3 (GlcNAc b1-6) GalNAc a	Core structures (Type 1 & Type 2)
38	a2-3 Neu5Ac on Core 1 with a-galactose on Core 2	Neu5Ac a2-3 Gal b1-3 (Gal a1-3 Gal b1-4 GlcNAc b1-6) GalNAc a	Core structures (Type 1 & Type 2)
39	a2-3 Neu5Ac on Core 1 with a-galactosylated Core 2	Neu5Ac a2-3 Gal b1-3 (Gal b1-4 GlcNAc b1-6) GalNAc a	Core structures (Type 1 & Type 2)
40	a2-3 Neu5Ac on Core 1 with Lex on Core 2	Neu5Ac a2-3 Gal a1-3 (Gal a1-4 (Fuc a1-3) GlcNAc a1-6) GalNAc a	Core structures (Type 1 & Type 2)
41	a2-3 Neu5Ac on Core 1	Neu5Ac a2-3 Gal b1-3 GalNAc a	Core structures (Type 1)
42	Core type 1 triose	Gal b1-3 GlcNAc b1-4 Gal	Core structures (Type 1)

43	Lacto-N-tetraose	Gal b1-3 GlcNAc b1-3 Gal b1-4 Glc b	Core structures (Type 1)
44	Lacto-N-triose	GlcNAc b1-3 Gal b1-4 Glc	Core structures (Type 1)
45	a2-3 Neu5Ac on Core 2	Gal a1-3 (Neu5Ac a2-3 Gal a1-4 GlcNAc a1-6) GalNAc a	Core structures (Type 2)
46	a-Galactose on Core 2	Gal b1-3 (Gal a1-3 Gal b1-4 GlcNAc b1-6) GalNAc a	Core structures (Type 2)
47	b-galactosylated Core 2	Gal b1-3 (Gal b1-4 GlcNAc b1-6) GalNAc a	Core structures (Type 2)
48	Core 2	Gal b1-3 (GlcNAc b1-6) GalNAc a	Core structures (Type 2)
49	Core type 2 triose	Gal b1-4 GlcNAc b1-4 Gal	Core structures (Type 2)
50	disialyl Core 2	Neu5Ac a2-3 Gal b1-3 (Neu5Ac a2-3 Gal b1-4 GlcNAc b1-6) GalNAc a	Core structures (Type 2)
51	disialyl Core 2 with sLex on Core 2	Neu5Ac a2-3 Gal b1-3 (Neu5Ac a2-3 Gal b1-4 (Fuc a1-3) GlcNAc b1-6) GalNAc a	Core structures (Type 2)
52	Lacto-N-hexaose	Gal b1-4 GlcNAc b1-6 (Gal b1-3 GlcNAc b1-3) Gal b1-4 Glc b	Core structures (Type 2)
53	Lacto-N-neohexaose	Gal b1-4 GlcNAc b1-6 (Gal b1-4 GlcNAc b1-3) Gal b1-4 Glc b	Core structures (Type 2)
54	Lacto-N-neooctaose	Gal b1-4 GlcNAc b1-3 Gal b1-4 GlcNAc b1-3 Gal b1-4 GlcNAc b1-3 Gal b1-4 Glc	Core structures (Type 2)
55	Lacto-N-neotetraose	Gal b1-4 GlcNAc b1-3 Gal b1-4 Glc b	Core structures (Type 2)
56	LN2 Type 2 (Di-N-Acetyl-D-Lactosamine) [I antigen]	Gal b1-4 GlcNAc b1-3 Gal b1-4 GlcNAc b	Core structures (Type 2)
57	LN3 Type 2 [I antigen]	Gal b1-4 GlcNAc b1-3 Gal b1-4 GlcNAc b1-3 Gal b1-4 GlcNAc b	Core structures (Type 2)
58	N-Acetyl-D-Lactosamine (LacNAc)	Gal b1-4 GlcNAc b	Core structures (Type 2)
59	Para-Lacto-N-neohexaose	Gal b1-4 GlcNAc b1-3 Gal b1-4 GlcNAc b1-3 Gal b1-4 Glc	Core structures (Type 2)
60	SLex on Core 2	Neu5Ac a2-3 Gal b1-3 (Gal b1-4 (Fuc a1-3) GlcNAc b1-6) GalNAc a	Core structures (Type 2)
61	Core type 4 (Elicityl)	Gal b1-3 GalNAc b1-3 Gal	Core structures (Type 4)
62	3'-sulpho Lewis A [3S-Gal-3(Fuc)-GlcNAc]	Gal [3S] b1-3 (Fuc a1-4) GlcNAc b	Fucosylated oligosaccharides
63	Fuc-(Glc-Man-(Gal-(Fuc)-GlcNAc-Man)-Man-GlcNAc)-GlcNAc	Fuc a1-6 (Glc b1-2 Man a1-6 (Gal b1-4 (Fuc a1-3) GlcNAc b1-2 Man a1-3) Man b1-4 GlcNAc b1-4) GlcNAc b	Fucosylated oligosaccharides
64	Fuc-(GlcNAc-Man-(Gal-GlcNAc-Man)-Man-GlcNAc)-GlcNAc	Fuc a1-6 (GlcNAc b1-2 Man a1-6 (Gal b1-4 GlcNAc b1-2 Man a1-3) Man b1-4 GlcNAc b1-4)GlcNAc b	Fucosylated oligosaccharides

65	3S-Gal-4(Fuc)-GlcNAc	Gal [3S] b1-4 (Fuc a1-3) GlcNAc b	Fucosylated oligosaccharides
66	a-3Galactosyl-3Fucosyllactose	Gal a1-3 Gal b1-4 (Fuc a1-3) Glc	Fucosylated oligosaccharides (3 Fucosyllactose core)
67	a-4Galactosyl-3Fucosyllactose	Gal a1-4 Gal b1-4 (Fuc a1-3) Glc	Fucosylated oligosaccharides (3 Fucosyllactose core)
68	3-Fucosylated Blood group A tetraose	GalNAc a1-3 (Fuc a1-2) Gal b1-4 (Fuc a1-3) Glc	Fucosylated oligosaccharides (3 Fucosyllactose core)
69	3-Fucosylated Blood group B tetraose	Gal a1-3 (Fuc a1-2) Gal b1-4 (Fuc a1-3) Glc	Fucosylated oligosaccharides (3 Fucosyllactose core)
70	3-Fucosyllactose	Gal b1-4 (Fuc a1-3) Glc b	Fucosylated oligosaccharides (Lacto-Series)
71	3'-Sialyl-3-fucosyllactose	Neu5Ac a2-3 Gal b1-4 (Fuc a1-3) Glc b	Fucosylated oligosaccharides (Lacto-Series)
72	6'-Sialyllactose	Neu5Ac a2-6 Gal b1-4 Glc b	Fucosylated oligosaccharides (Lacto-Series)
73	A-tetra Lactose	GalNAc a1-3 (Fuc a1-2) Gal b1-4 Glc b	Fucosylated oligosaccharides (Lacto-Series)
74	A-tetra type 2	GalNAc a1-3 (Fuc a1-2) Gal b1-4 GlcNAc b	Fucosylated oligosaccharides (Lacto-Series)
75	B-tetra Lactose	Gal a1-3 (Fuc a1-2) Gal b1-4 Glc b	Fucosylated oligosaccharides (Lacto-Series)
76	B-tetra type 2	Gal a1-3 (Fuc a1-2) Gal b1-4 GlcNAc b	Fucosylated oligosaccharides (Lacto-Series)
77	Difucosyllactose	Fuc a1-2 Gal b1-4 (Fuc a1-3) Glc b	Fucosylated oligosaccharides (Lacto-Series)
78	Lacto-N-difucohexaose I	Fuc a1-2 Gal b1-3 (Fuc a1-4) GlcNAc b1-3 Gal b1-4 Glc b	Fucosylated oligosaccharides (Lacto-Series)
79	Lacto-N-fucopentaose II	Gal b1-3 (Fuc a1-4) GlcNAc b1-3 Gal b1-4 Glc b	Fucosylated oligosaccharides (Lacto-Series)
80	Lacto-N-fucopentaose III	Gal b1-4 (Fuc a1-3) GlcNAc b1-3 Gal b1-4 Glc b	Fucosylated oligosaccharides (Lacto-Series)
81	Lacto-N-fucopentaose V	Gal b1-3 GlcNAc b1-3 Gal b1-4 (Fuc a1-3) Glc b	Fucosylated oligosaccharides

			(Lacto-Series)
82	Lacto-N-neofucopentaose	Gal b1-4 GlcNAc b1-3 Gal b1-4 (Fuc a1-3) Glc	Fucosylated oligosaccharides (Lacto-Series)
83	Chondroitin tetrasaccharide	deoxy-GlcA b1-3 GalNAc [4S] b1-4 GlcA b1-3 GalNAc [4S]	GAGs
84	Heparin decasaccharide	deoxy-IdoA [2S] a1-4 GlcN [2S6S] (a1-4 IdoA [2S] a1-4 GlcN [2S6S])4	GAGs
85	Heparin heptasaccharide	deoxy-GlcN [2S6S] (a1-4 IdoA [2S] (a1-4 GlcN [2S6S])3	GAGs
86	Heparin hexasaccharide	deoxy-GlcN [2S6S] a1-4 IdoA [2S] (a1-4 GlcN [2S6S] a1-4 IdoA [2S])2	GAGs
87	Heparin hexasaccharide	IdoA [2S] a1-4 GlcN [2S6S] (a1-4 IdoA [2S] a1-4 GlcN [2S6S])2	GAGs
88	Heparin hexasaccharide	deoxy-IdoA [2S] a1-4 GlcN [2S6S] (a1-4 IdoA [2S] a1-4 GlcN [2S6S])2	GAGs
89	Heparin nonasaccharide	deoxy-GlcN [2S6S] (a1-4 IdoA [2S] a1-4 GlcN [2S6S])4	GAGs
90	Heparin octasaccharide	deoxy-GlcN [2S6S] a1-4 IdoA [2S] (a1-4 GlcN [2S6S] a1-4 IdoA [2S])3	GAGs
91	Heparin pentasaccharide	GlcN [2S6S] a1-4 IdoA [2S] a1-4 GlcN [2S6S] a1-4 IdoA [2S] a1-4 GlcN [2S6S]	GAGs
92	Heparin pentasaccharide	deoxy-IdoA [2S] (a1-4 GlcN [2S6S] a1-4 IdoA [2S])2	GAGs
93	Heparin tetrasaccharide	deoxy-IdoA [2S] a1-4 GlcN [2S6S] a1-4 IdoA [2S] a1-4 GlcN [2S6S]	GAGs
94	Hyalonuric acid hexasaccharide	GlcA b1-3 GlcNAc b1-4 GlcA b1-3 GlcNAc b1-4 GlcA b1-3 GlcNAc b	GAGs
95	Hyalonuric acid tetrasaccharide	GlcA b1-3 GlcNAc b1-4 GlcA b1-3 GlcNAc b	GAGs
96	SR123781A (Heparin 16)	Glc [2S3S4S6S] a1-4 Glc [2S3S6S] a1-4 Glc [2S3S6S] b1-4 Glc [6S] a1-4 (Glc b1-4 Glc a1-4)3 Glc b1-4 Glc [6S] Glc b1-4 Glc a1-4 [2S3S6S] IdoA a1-4 Glc	GAGs
97	Xeno lewis X (Gal Lewis x)	Gal a1-3 Gal b1-4 (Fuc a1-3) GlcNAc b	Galα-3Gal oligosaccharides (Galili and xeno antigens)
98	Xeno lewis a	Gal a1-3 Gal b1-3 (Fuc a1-4) GlcNAc	Galα-3Gal oligosaccharides (Galili and xeno antigens)
99	Galili antigen pentaose	Gal a1-3 Gal b1-4 GlcNAc b1-3 Gal b1-4 Glc	Galα-3Gal oligosaccharides

			(Galili and xeno antigens)
100	Galili-tri	Gal a1-3 Gal b1-4 Glc b	Galα-3Gal oligosaccharides (Galili and xeno antigens)
101	Xeno antigen type 1	Gal a1-3 Gal b1-3 GlcNAc	Galα-3Gal oligosaccharides (Galili and xeno antigens)
102	Xeno antigen type 2 (Galili antigen triose Gal a3 epitope)	Gal a1-3 Gal b1-4 GlcNAc	Galα-3Gal oligosaccharides (Galili and xeno antigens)
103	Isoglobopentaose (iGB5)	Gal b1-3 GalNAc b1-3 Gal a1-3 Gal b1-4 Glc	Galα-3Gal oligosaccharides (Isogloboseries)
104	Isoglobotetraose (iGB4 & Cytolipin R)	GalNAc b1-3 Gal a1-3 Gal b1-4 Glc	Galα-3Gal oligosaccharides (Isogloboseries)
105	Isoglobotriose (iGB3)	Gal a1-3 Gal b1-4 Glc	Galα-3Gal oligosaccharides (Isogloboseries)
106	3'-Sialyllactose	Neu5Ac a2-3 Gal b1-4 Glc b	Ganglioside sugars
107	Fucosyl GM1	Fuc a1-2 Gal b1-3 GalNAc b1-4 (Neu5Ac a2-3) Gal b1-4 Glc	Ganglioside sugars
108	GA1 (aGM1)	Gal b1-3 GalNAc b1-4 Gal b1-4 Glc b	Ganglioside sugars
109	GA2 (aGM2)	GalNAc b1-4 Gal b1-4 Glc b	Ganglioside sugars
110	GD1a	Neu5Ac a2-3 Gal b1-3 GalNAc b1-4 (Neu5Ac a2-3) Gal b1-4 Glc b	Ganglioside sugars
111	GD1b	Gal b1-3 GalNAc b1-4 (Neu5Ac a2-8 Neu5Ac a2-3) Gal b1-4 Glc b	Ganglioside sugars
112	GD2	GalNAc b1-4 (Neu5Ac a2-8 Neu5Ac a2-3) Gal b1-4 Glc b	Ganglioside sugars
113	GD3	Neu5Ac a2-8 Neu5Ac a2-3 Gal b1-4 Glc b	Ganglioside sugars
114	GM1a	Gal b1-3 GalNAc b1-4 (Neu5Ac a2-3) Gal b1-4 Glc b	Ganglioside sugars
115	GM1b	Neu5Ac a2-3 Gal b1-3 GalNAc b1-4 Gal b1-4 Glc b	Ganglioside sugars
116	GM2	GalNAc b1-4 (Neu5Ac a2-3) Gal b1-4 Glc b	Ganglioside sugars
117	GM3	Neu5Ac a2-3 Gal b1-4 Glc b	Ganglioside sugars
118	GT1a	Neu5Ac a2-8 Neu5Ac a2-3 Gal b1-3 GalNAc b1-4 (Neu5Ac a2-3) Gal b1-4 Glc	Ganglioside sugars
119	GT1b	Neu5Ac a2-3 Gal b1-3 GalNAc b1-4 (Neu5Ac a2-8 Neu5Ac a2-3) Gal b1-4 Glc b	Ganglioside sugars
120	GT1c	Gal b1-3 GalNAc b1-4 (Neu5Ac a2-8 Neu5Ac a2-8 Neu5Ac a2-	Ganglioside sugars

		3) Gal b1-4 Glc b	
121	GT2	GalNAc b1-4 (Neu5Ac a2-8 Neu5Ac a2-8 Neu5Ac a2-3) Gal b1-4 Glc b : Na	Ganglioside sugars
122	GT3	Neu5Ac a2-8 Neu5Ac a2-8 Neu5Ac a2-3 Gal b1-4 Glc	Ganglioside sugars
123	Forssman antigen pentaose	GalNAc a1-3 GalNAc b1-3 Gal a1-4 Gal b1-4 Glc	Globoside sugars (P antigens) (Forssman antigens)
124	Forssman antigen triose	GalNAc a1-3 GalNAc b1-3 Gal	Globoside sugars (P antigens) (Forssman antigens)
125	Isoforssman antigen pentaose	GalNAc a1-3 GalNAc b1-3 Gal a1-3 Gal b1-4 Glc	Globoside sugars (P antigens) (Forssman antigens)
126	Globo-A	GalNAc a1-3 (Fuc a1-2) Gal b1-3 GalNAc b1-3 Gal a1-4 Gal b1-4 Glc	Globoside sugars (P antigens) (Globo series - core structure type 4
127	Globo-B	Gal a1-3 (Fuc a1-2) Gal b1-3 GalNAc b1-3 Gal a1-4 Gal b1-4 Glc	Globoside sugars (P antigens) (Globo series - core structure type 4
128	Blood group H antigen tetraose type 4 & Globo H tetraose	Fuc a1-2 Gal b1-3 GalNAc b1-3 Gal	Globoside sugars (P antigens) (Globo series - core structure type 4
129	Globoside (NAc) (P-antigen)	GalNAc b1-3 Gal a1-4 Gal b1-4 GlcNAc b	Globoside sugars (P antigens) (P blood group antigens and analogues)
130	Globotetraose (Gb4) (P antigen)	GalNAc b1-3 Gal a1-4 Gal b1-4 Glc	Globoside sugars (P antigens) (P blood group antigens and analogues)
131	P1 antigen (Globotriose analogue type 2)	Gal a1-4 Gal b1-4 GlcNAc b	
132	Globotriose analogue type 1	Gal a1-4 Gal b1-3 GlcNAc	
133	3-Sialyl-Gb3 (Sialylated Globotriose)	Neu5Ac a2-3 Gal a1-4 Gal b1-4 Glc	
134	Pk antigen (Globotriose or Gb3)	Gal a1-4 Gal b1-4 Glc b	
135	Globopentaose (Gb5) [Stage specific embryonic antigen 3a (SSEA-3a)]	Gal b1-3 GalNAc b1-3 Gal a1-4 Gal b1-4 Glc	Globoside sugars (P antigens) (Stage-specific Embryonic

			antigens : SSEA-3 & SSEA-4)
136	Globo-H hexaose [Stage specific embryonic antigen 3b (SSEA-3b)]	Fuc a1-2 Gal b1-3 GalNAc b1-3 Gal a1-4 Gal b1-4 Glc	Globoside sugars (P antigens) (Stage-specific Embryonic antigens : SSEA-3 & SSEA-4)
137	SSEA-4 tetraose (Stage-specific embryonic antigen 4)	Neu5Ac a2-3 Gal b1-3 GalNAc b1-3 Gal	Globoside sugars (P antigens) (Stage-specific Embryonic antigens : SSEA-3 & SSEA-4)
138	SSEA-4 hexaose (Stage-specific embryonic antigen 4)	Neu5Ac a2-3 Gal b1-3 GalNAc b1-3 Gal a1-4 Gal b1-4 Glc	Globoside sugars (P antigens) (Stage-specific Embryonic antigens : SSEA-3 & SSEA-4)
139	Glucuronyl_Lactose	GlcA b1-3 Gal b1-4 Glc	Glucuronylated oligosaccharides
140	Glucuronyl-Lacto-N-tetraose	GlcA b1-3 Gal b1-3 GlcNAc b1-3Gal b1-4 Glc	Glucuronylated oligosaccharides
141	Lacteneo sphingolip core	Gal b1-4 GlcNAc b1-3 Gal b1-4 Glc b : Cer	Glycosphingolipid
142	GPI anchor	Man a1-2 Man a1-6 Man a1-4 GlcNH2 a1-6 myo-inositol	Glycosphingolipid
143	3'-SiaDi-LN	Neu5Ac a2-3 Gal b1-4 GlcNAc b1-4 Gal b1-4 GlcNAc b	Lewis antigens
144	3'-Sialyl-Lewis c	Neu5Ac a2-3 Gal b1-3 GlcNAc b	Lewis antigens
145	3'-su-Lewis a	Gal [3S] b1-3 (Fuc a1-4) GlcNAc b	Lewis antigens
146	3'-su-Lewis x	Gal [3S] b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
147	6-su-GalNAc-SiaLewis x	Neu5Ac a2-3 Gal b1-4 (Fuc a1-3) GlcNAc [6S] b	Lewis antigens
148	6-su-Gal-SiaLewis x	Neu5Ac a2-3 Gal [6S] b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
149	di-Lewis x b1-4 Lewis x [di-Lewis x]	Gal b1-4 (Fuc a1-3) GlcNAc b1-4 Gal b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
150	Lacto-N-difucohexaose II	Gal b1-3 (Fuc a1-4) GlcNAc b1-3 Gal b1-4 (Fuc a1-3) Glc b	Lewis antigens
151	Lewis a hexaose	Gal b1-3 (Fuc a1-4) GlcNAc b1-3 Gal b1-4 (Fuc a1-3) Glc	Lewis antigens
152	Lewis a Lewis x	Gal b1-3 (Fuc a1-4) GlcNAc b1-3 Gal b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
153	Lewis a LN	Gal a1-3 (Fuc a1-4) GlcNAc b1-3 Gal b1-4 GlcNAc b	Lewis antigens
154	Lewis a tetraose	Gal b1-3 (Fuc a1-4) GlcNAc b1-3 Gal b	Lewis antigens
155	Lewis a triose	Gal b1-3 (Fuc a1-4) GlcNAc b	Lewis antigens
156	Lewis b pentaose	Fuc a1-2 Gal b1-3 (Fuc a1-4) GlcNAc b1-3 Gal	Lewis antigens
157	Lewis b tetraose	Fuc a1-2 Gal b1-3 (Fuc a1-4) GlcNAc b	Lewis antigens
158	Lewis c	Gal b1-3 GlcNAc b	Lewis antigens
159	Lewis X hexaose (also Lacto-N-	Gal b1-4 (Fuc a1-3) GlcNAc b1-3 Gal b1-4 (Fuc a1-3) Glc	Lewis antigens

	neodifucohexaose, LeX-LeX)		
160	Lewis x on Core 2	Gal b1-3 (Gal b1-4 (Fuc a1-3) GlcNAc b1-6) GalNAc a	Lewis antigens
161	Lewis X tetraose [Gal-(Fuc)-GlcNAc-Gal]	Gal b1-4 (Fuc a1-3) GlcNAc b1-3 Gal b	Lewis antigens
162	Lewis x triaose [SSEA-1]	Gal b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
163	Lewis y Lewis x	Fuc a1-2 Gal b1-4 (Fuc a1-3) GlcNAc b1-4 Gal b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
164	Lewis y pentaose	Fuc a1-2 Gal b1-4 (Fuc a1-3) GlcNAc b1-3 Gal	Lewis antigens
165	Lewis y tetraose [Blood group H type 2]	Fuc a1-2 Gal b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
166	Sialyl Lewis a (sLeA) tetraose [CA19-9 antigen]	Neu5Ac a2-3 Gal b1-3 (Fuc a1-4) GlcNAc b	Lewis antigens
167	Sialyl Lewis X (sLeX) pentaose	Neu5Ac a2-3 Gal b1-4 (Fuc a1-3) GlcNAc b1-3 Gal	Lewis antigens
168	Sialyl Lewis X (sLeX) tetraose	Neu5Ac a2-3 Gal b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
169	SLex-Lex (SDLeX)	Neu5Ac a2-3 Gal b1-4 (Fuc a1-3) GlcNAc b1-3 Gal b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
170	SLex-Lex-Lex	Neu5Ac a2-3 Gal b1-4 (Fuc a1-3)GlcNAc b1-3 Gal b1-4 (Fuc a1-3) GlcNAc b1-3 Gal b1-4 (Fuc a1-3) GlcNAc b	Lewis antigens
171	tri-Lewis x	Gal b1-4 (Fuc a1-3) GlcNAc b1-4 Gal b1-4 (Fuc a1-3) GlcNAc b1-4 Gal b4 (Fuc a1-3) GlcNAc b	Lewis antigens
172	Gal b1-4 GalNAc b1-4 Gal b1-4 Glc b	Gal b1-4 GalNAc b1-4 Gal b1-4 Glc b	Miscellaneous
173	(GlcNAc)3-GalNAc	GlcNAc b1-3 (GlcNAc b1-4) (GlcNAc b1-6) GalNAc a	Miscellaneous
174	3'-KDN-Lewis c	KDN a2-3 Gal b1-4 GlcNAc b	Miscellaneous
175	3'-KDNLN	KDN a2-3 Gal b1-3 GlcNAc b	Miscellaneous
176	6P-Man3	Man [6P] a1-3 Man a1-3 Man a	Miscellaneous
177	deoxy Chitotetraose	GlcNAc b1-4 GlcNAc b1-4 GlcNAc b1-4 deoxy GlcNAc	Miscellaneous
178	deoxy Chitotriose	GlcNAc b1-4 GlcNAc b1-4 deoxy GlcNAc	Miscellaneous
179	Fuc-Gal-Xyl	Fuc a1-2 Gal b1-2 Xyl a	Miscellaneous
180	Gal a (1-3') LacNAc	Gal a1-3 Gal b1-4 GlcNAc b	Miscellaneous
181	Gal-GalNAc-(Neu5Ac)-Gal-Glc	Gal b1-3 GalNAc b1-4 (Neu5Ac a2-3) Gal b1-4 Glc b	Miscellaneous
182	Gal-GlcNAc-(Gal-GlcNAc)-Gal	Gal b1-4 GlcNAc b1-2 (Gal b1-4 GlcNAc b1-3) Gal b	Miscellaneous
183	GalNAc-(Fuc)-Gal-Glc	GalNAc a1-3 (Fuc a1-2) Gal b1-4 Glc b	Miscellaneous
184	GlcNAc b(1-3') LacNAc	GlcNAc b1-3 Gal b1-4 GlcNAc b	Miscellaneous

185	GlcNAc b (1-4,6) GalNAc	GlcNAc b1-4 (GlcNAc b1-6) GalNAc a	Miscellaneous
186	GlcNAc-(Fuc)-deoxyGlcNAc	GlcNAc b1-4 (Fuc a1-3) deoxy GlcNAc	Miscellaneous
187	GlcNAc-(Fuc)-GlcNAc	GlcNAc b1-4 (Fuc a1-3) GlcNAc b	Miscellaneous
188	GlcNAc-Gal-Glc	GlcNAc b1-6 Gal b1-4 Glc b	Miscellaneous
189	GlcNAc-Man-(GlcNAc-Man)-Man	GlcNAc b1-2 Man a1-3 (GlcNAc b1-2 Man a1-6) Man a	Miscellaneous
190	Neu5Ac-Gal-(Fuc)-GlcNAc-OMe	Neu5Ac a2-3 Gal b1-4 (Fuc a1-3) GlcNAc b	Miscellaneous
191	Neu5Ac-Gal-(Neu5Ac)-GlcNAc	Neu5Ac a2-3 Gal b1-3 (Neu5Ac a2-6) GlcNAc b	Miscellaneous
192	tetrafluoro-4-methoxy-benzamide-GalN-GlcNAc	C8O2F4 GalpN b1-4 GlcpNAc b	Miscellaneous
193	Tk	GlcNAc b1-3 (GlcNAc b1-6) Gal b1-4 GlcNAc b	Miscellaneous
194	Fucosylated A antigen type 5	GalNAc a1-3 (Fuc a1-2) Gal b1-4 (Fuc a1-2) Glc	Miscellaneous (Blood group-related oligosaccharides)
195	Fucosylated B antigen type 5	Gal a1-3 (Fuc a1-2) Gal b1-4 (Fuc a1-2) Glc	Miscellaneous (Blood group-related oligosaccharides)
196	Chitohexaose	GlcNAc b1-4 GlcNAc b1-4 GlcNAc b1-4 GlcNAc b1-4 GlcNAc b1-4 GlcNAc b	Miscellaneous (Chitin oligosaccharides)
197	Chitopentaose	GlcNAc b1-4 GlcNAc b1-4 GlcNAc b1-4 GlcNAc b1-4 GlcNAc b	Miscellaneous (Chitin oligosaccharides)
198	Chitotetraose	GlcNAc b1-4 GlcNAc b1-4 GlcNAc b1-4 GlcNAc b	Miscellaneous (Chitin oligosaccharides)
199	Chitotriose	GlcNAc b1-4 GlcNAc b1-4 GlcNAc b	Miscellaneous (Chitin oligosaccharides)
200	As-Fibrinogen	Gal b1-4 GlcNAc b1-4 Man a1-3 (Gal b1-4 GlcNAc b1-4 Man a1-6) Man a1-4 GlcNAc b1-4 GlcNAc b	Miscellaneous (Fibrinogen related)
201	asialo-, agalacto-N-glycan from porcine fibrinogen	GlcNAc b1-2 Man a1-3 (GlcNAc b1-2 Man a1-6) Man b1-4 GlcNAc b1-4 (Fuc a1-6) GlcNAc b	Miscellaneous (Fibrinogen related)
202	di-sialylated N-glycan from porcine fibrinogen	Neu5Ac a2-6 Gal b1-4 GlcNAc b1-2 Man a1-3 (Neu5Ac a2-6 Gal b1-4 GlcNAc b1-2 Man a1-6) Man b1-4 GlcNAc b1-4 (Fuc a1-6) GlcNAc b	Miscellaneous (Fibrinogen related)
203	2'-F-Di-LN	Fuc a1-2 Gal b1-4 GlcNAc b1-3 Gal b1-4 GlcNAc b	Miscellaneous (LDN-related)
204	3-SLDN	Neu5Ac a2-3 GalNAc b1-4 GlcNAc b	Miscellaneous (LDN-related)
205	6'-SiaDi-LN	Neu5Ac a2-6 Gal b1-4 GlcNAc b1-4 Gal b1-4 GlcNAc b	Miscellaneous (LDN-related)
206	Sialylated LDN [6-SLDN]	Neu5Ac a2-6 GalNAc b1-4 GlcNAc b	Miscellaneous (LDN-related)

207	Bi-LDN	GalNAc b1-4 GlcNAc b1-4 Man a1-6 (GalNAc b1-4 GlcNAc b1-4 Man a1-3) Man b1-4 GlcNAc b1-4 GlcNAc b	Miscellaneous (LDN-related)
208	Fuc a (1-3) Lac-di-Nac	GalNAc b1-4 (Fuc a1-3) GlcNAc b	Miscellaneous (LDN-related)
209	3' sulfo Lewis X [3S-Gal-(Fuc)-GlcNAc]	Gal [3S] b1-4 (Fuc a1-3) GlcNAc b	Miscellaneous (Lewis X-related)
210	4' sulfo Lewis X [4S-Gal-(Fuc)-GlcNAc]	Gal [4S] b1-4 (Fuc a1-3) GlcNAc b	Miscellaneous (Lewis X-related)
211	6-LacNAc-TF	Gal b1-4 GlcNAc b1-6 (Gal b1-3) GalNAc a	Miscellaneous (TF-related)
212	6-Siab-TF	Neu5Ac b2-6 (Gal b1-3) GalNAc a	Miscellaneous (TF-related)
213	6'-SiaaTF	Neu5Ac a2-6 (Gal b1-3) GalNAc a	Miscellaneous (TF-related)
214	GlcNAc b (1-2') TF	GlcNAc b1-2 Gal b1-3 GalNAc a	Miscellaneous (TF-related)
215	3,6-(LacNAc)2Tn	Gal b1-4 GlcNAc b1-3 (Gal b1-4 GlcNAc b1-6) GalNAc a	Miscellaneous (TN-related)
216	3-6-STn	Neu5Ac a2-3 (Neu5Ac a2-6) GalNAc a	Miscellaneous (TN-related)
217	3-LacNAc-Tn	Gal b1-4 GlcNAc b1-3 GalNAc a	Miscellaneous (TN-related)
218	6-LacNAc-Tn	Gal b1-4 GlcNAc b1-6 GalNAc a	Miscellaneous (TN-related)
219	Disialosylpentaose	Neu5Ac a2-3 Gal b1-4 Glc b1-1 (Neu5Ac a2-3) Gal	Miscellaneous (Trehalose-like sugars)
220	Galactosyl-lactose	Gal b1-4 Glc b1-1 b Gal	Miscellaneous (Trehalose-like sugars)
221	(Man)2-GlcNAc	Man a1-3 (Man a1-4) GlcNAc b	N-linked oligos
222	(Man)3 chitobiose	Man a1-3 (Man a1-6) Man b1-4 GlcNAc b1-4 GlcNAc b	N-linked oligos
223	(Man)3a	Man a1-3 (Man a1-6) Man a	N-linked oligos
224	(Man)3b	Man a1-3 (Man a1-6) Man b	N-linked oligos
225	(Man)4	Man b1-2 Man a1-2 Man b1-3 Man b	N-linked oligos
226	(Man)4-(Fuc)-GlcNAc	Man a1-2 Man a1-3 (Man a1-6) Man b1-4 (Fuc a1-3) GlcNAc b	N-linked oligos
227	(Man)5	Man a1-6 (Man a1-3) Man a1-6 (Man a1-3) Man b1-4 GlcNAc b1-4 GlcNAc b	N-linked oligos
228	(Man)5-OMe	Man a1-2 Man a1-3 (Man a1-3 Man a1-6) Man b	N-linked oligos
229	(Man)6	Man a1-6 (Man b1-3) Man b1-6 (Man a1-2 Man b1-3) Man a	N-linked oligos
230	(Man)6 (GlcNAc)2	Man a1-6 (Man a1-3) Man a1-6 (Man a1-2 Man a1-3) Man b1-4 GlcNAc b1-4 GlcNAc b	N-linked oligos
231	(Man)7	Man a1-2 Man a1-6 (Man a1-3) Man a1-6 (Man a1-2 Man a1-	N-linked oligos

		3) Man b1-4 GlcNAc b1-4 GlcNAc b	
232	(Man)9-GlcNAc-GlcNAc	Man b1-2 Man a1-6 (Man b1-2 Man b1-3) Man b1-6 (Man a1-2 Man a1-2 Man b1-3) Man b1-4 GlcNAc b1-4 GlcNAc b	N-linked oligos
233	Gal-GlcNAc-Man-(Gal-GlcNAc-Man)-Man-GlcNAc	Gal b1-4 GlcNAc b1-2 Man a1-3 (Gal b1-4 GlcNAc b1-2 Man a1-6) Man b1-4 GlcNAc b	N-linked oligos
234	Man-3 chitobiose core Fuc	Man a1-6 (Man a1-3) Man b1-4 GlcNAc b1-4 (Fuc a1-2) GlcNAc b	N-linked oligos
235	Mono-sialylated N-glycan from porcine fibrinogen	Neu5Ac a2-6 Gal b1-4 GlcNAc b1-2 Man a1-3 (Gal b1-4 GlcNAc b1-2 Man a1-6) Man b1-4 GlcNAc b1-4 (Fuc a1-6) GlcNAc b	N-linked oligos
236	N-glycan mixture from horseradish peroxidase	Man a1-3 (Xyl b1-2) (Man a1-6) Man b1-4 GlcNAc b1-4 GlcNAc b	N-linked oligos
237	N-Man5 GlcNAc 1-4	Man a1-6 (Man a1-3) Man a1-6 (GlcNAc b1-4) (GlcNAc b1-2 Man a1-3) Man b1-4 GlcNAc b1-4 GlcNAc b	N-linked oligos
238	tri-sialylated N-glycan from porcine fibrinogen	Neu5Ac a2-8 Neu5Ac a2-(3,6) Gal b1-4 GlcNAc b1-2 Man a1-3 (Neu5Ac a2-6 Gal b1-4 GlcNAc b1-2 Man a1-6) Man b1-4 GlcNAc b1-4 (Fuc a1-6) GlcNAc b	N-linked oligos
239	Sialylated triose type 1	Neu5Ac a2-3 Gal b1-3 GlcNAc b	Sialylated oligosaccharide (Type 1)
240	Di-sialyl-lacto-N-tetraose	Neu5Ac a2-3 Gal b1-3 (Neu5Ac a2-6) GlcNAc b1-3 Gal b1-4 Glc b	Sialylated oligosaccharide (Type 1)
241	LS-tetrasaccharide a	Neu5Ac a2-3 Gal b1-3 GlcNAc b1-3 Gal b1-4 Glc b	Sialylated oligosaccharide (Type 1)
242	LS-tetrasaccharide b	Gal b1-3 (Neu5Ac a2-6) GlcNAc b1-3 Gal b1-4 Glc b	Sialylated oligosaccharide (Type 1)
243	LS-tetrasaccharide c	Neu5Ac a2-6 Gal b1-3 GlcNAc b1-3 Gal b1-4 Glc b	Sialylated oligosaccharide (Type 1)
244	Neu5Ac-Gal-(Neu5Ac)-deoxy-GalNAc	Neu5Ac a2-3 Gal b1-3 (Neu5Ac a2-6) deoxy GalNAc	Sialylated oligosaccharide (Type 1)
245	Sia2TF	Neu5Ac a2-3 Gal b1-3 (Neu5Ac a2-6) GalNAc a	Sialylated oligosaccharide (Type 1)
246	Sialylated tetraose type 1	Neu5Ac a2-3 Gal b1-3 GlcNAc b1-3 Gal	Sialylated oligosaccharide (Type

			1)
247	Disialylated tetraose type 1	Neu5Ac a2-8 Neu5Ac a2-3 Gal b1-3 GlcNAc	Sialylated oligosaccharide (Type 1)
248	Disialylated pentaose type 1	Neu5Ac a2-8 Neu5Ac a2-3 Gal b1-3 GlcNAc b1-3 Gal	Sialylated oligosaccharide (Type 1)
249	Trisialylated pentaose type 1	Neu5Ac a2-8 Neu5Ac a2-8 Neu5Ac a2-3 Gal b1-3 GlcNAc	Sialylated oligosaccharide (Type 1)
250	LS-hexasaccharide d (3'-SiaLN-LN-LN)	Neu5Ac a2-3 Gal b1-4 GlcNAc b1-3 Gal b1-4 GlcNAc b1-3 Gal b1-4 GlcNAc b	Sialylated oligosaccharide (Type 2)
251	LS-tetrasaccharide d (3'-SiaLN-LN-LN)	Neu5Ac a2-3 Gal b1-4 GlcNAc b1-3 Gal b1-4 Glc	Sialylated oligosaccharide (Type 2)
252	3'-Sialyl-N-acetyllactosamine	Neu5Ac a2-3 Gal b1-4 GlcNAc b	Sialylated oligosaccharide (Type 2)
253	6'-Sia-6-su-LacNAc	Neu5Ac a2-6 Gal [6S] b1-4 GlcNAc	Sialylated oligosaccharide (Type 2)
254	6'-Sialyl-N-acetyllactosamine	Neu5Ac a2-6 Gal b1-4 GlcNAc b	Sialylated oligosaccharide (Type 2)
255	deoxy-3'SLN	Neu5Ac a2-3 Gal b1-4 deoxy GlcNAc	Sialylated oligosaccharide (Type 2)
256	deoxy-6'SLN	Neu5Ac a2-6 Gal b1-4 deoxy GlcNAc	Sialylated oligosaccharide (Type 2)
257	Neu5Ac-Gal-(Fuc)-GlcNAc	Neu5Ac a2-3 Gal b1-4 (Fuc a1-3) GlcNAc b	Sialylated oligosaccharide (Type 2)
258	3-Sialyl-LN type 2 [Sialylated triose type 2, 3'-SLN]	Neu5Ac a2-3 Gal b1-4 GlcNAc b	Sialylated oligosaccharide (Type 2)
259	Sialylated tetraose type 2	Neu5Ac a2-3 Gal b1-4 GlcNAc b1-3 Gal	Sialylated oligosaccharide (Type 2)
260	Disialylated tetraose type 2	Neu5Ac a2-8 Neu5Ac a2-3 Gal b1-4 GlcNAc	Sialylated oligosaccharide (Type 2)
261	Trisialylated pentaose type 2	Neu5Ac a2-8 Neu5Ac a2-8 Neu5Ac a2-3 Gal b1-4 GlcNAc	Sialylated oligosaccharide (Type 2)

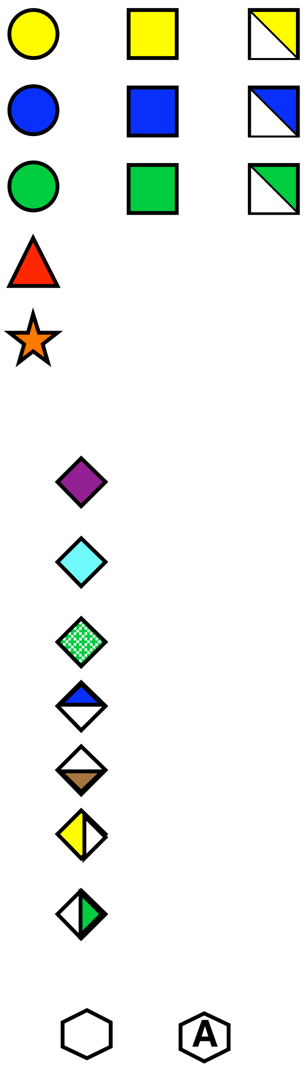
Annex IV

Essentials Second Edition Symbols : RGB colors

Hexoses: Circles   **N-Acetylhexosamines: Squares**

Hexosamines: Squares divided diagonally

Print in color



- **Galactose stereochemistry: Yellow (255,255,0) with Black outline**
- **Glucose stereochemistry: BLUE (0,0,250) with Black outline**
- **Mannose stereochemistry: GREEN (0,200,50) with Black outline**
- **Fucose: RED (250,0,0) with Black outline**
- **Xylose: (5-pointed star) ORANGE (250,100,0) with Black outline**

Acidic Sugars (Diamonds)

- **Neu5Ac: PURPLE (125,0,125) with Black outline**
- **Neu5Gc: LIGHT BLUE (200,250,250) with Black outline**
- **KDN: GREEN (0,200,50) with Pattern & Black outline**
- **GlcA: BLUE (0,0,250)/Upper segment with Black outline**
- **IdoA: TAN (150,100,50)/Lower segment with Black outline**
- **GalA: RED (250,0,0)/Left segment with Black outline**
- **ManA: GREEN (0,200,50)/Right segment with Black outline**

Other Monosaccharide

(use letter designation inside symbol to specify if needed)

Essentials Second Edition Symbols : Black & White

Hexoses: Circles **N-Acetylhexosamines: Squares**

Hexosamines: Squares divided diagonally

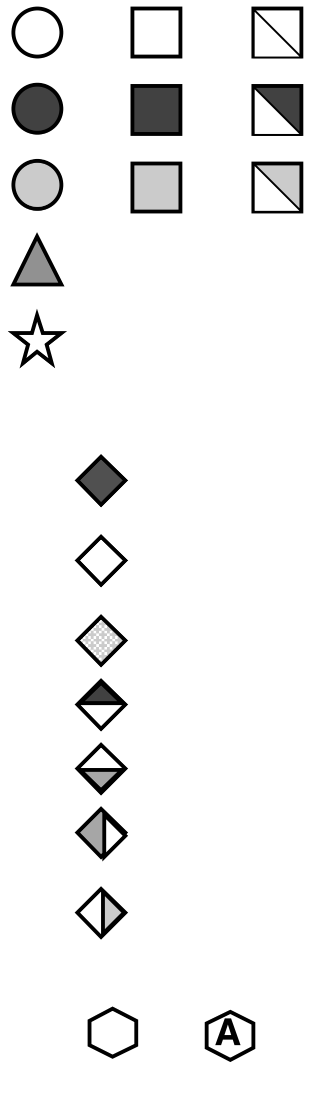
Print in black & white

- **Galactose stereochemistry: white with Black outline**
- **Glucose stereochemistry: Black with Black outline**
- **Mannose stereochemistry: Grey with Black outline**
- **Fucose: Dark Grey with Black outline**
- **Xylose: (5-pointed star) with Black outline**

Acidic Sugars (Diamonds)

- **Neu5Ac: Dark Grey with Black outline**
- **Neu5Gc: White with Black outline**
- **KDN: Light Grey Pattern & Black outline**
- **GlcA: Grey upper segment with Black outline**
- **IdoA: Grey Lower segment with Black outline**
- **GalA: Grey Left segment with Black outline**
- **ManA: Grey Right segment with Black outline**

Other Monosaccharide
(use letter designation inside symbol to specify if needed)



Supplementary material

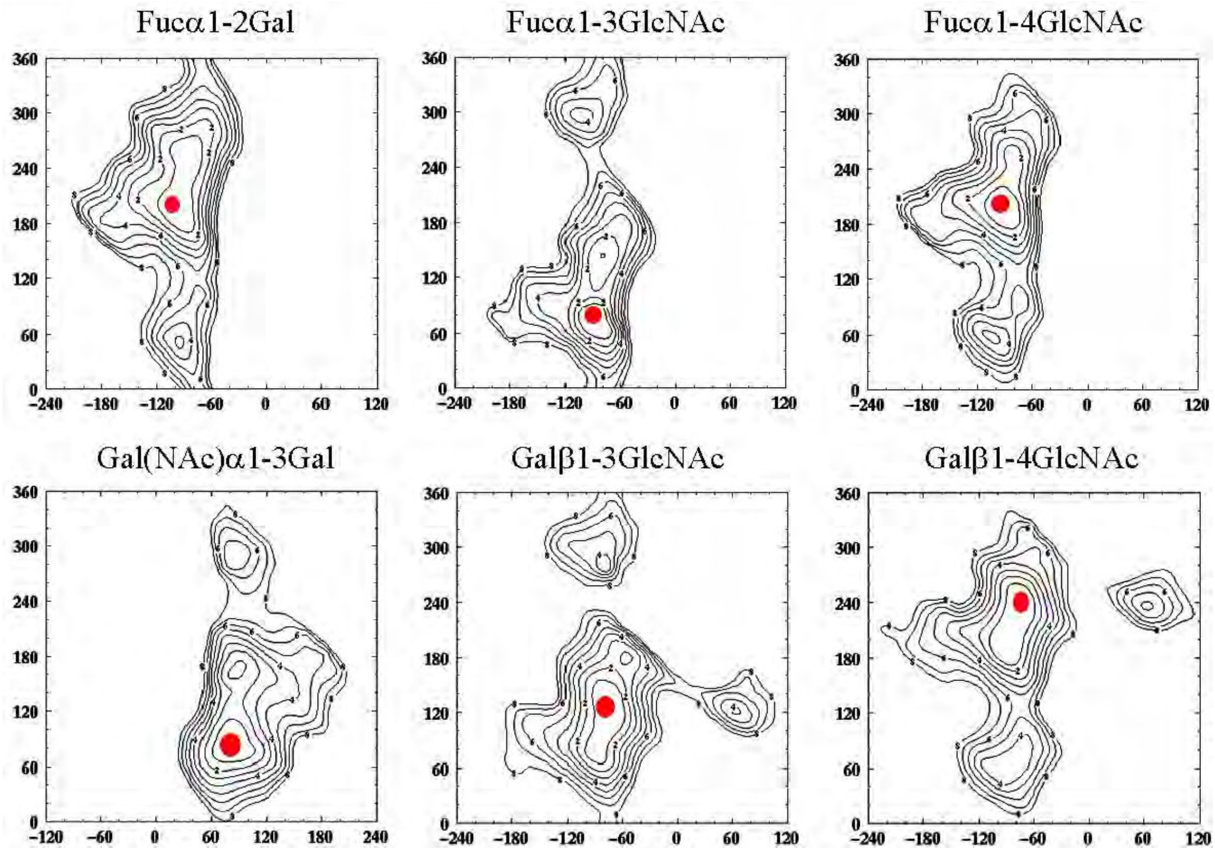
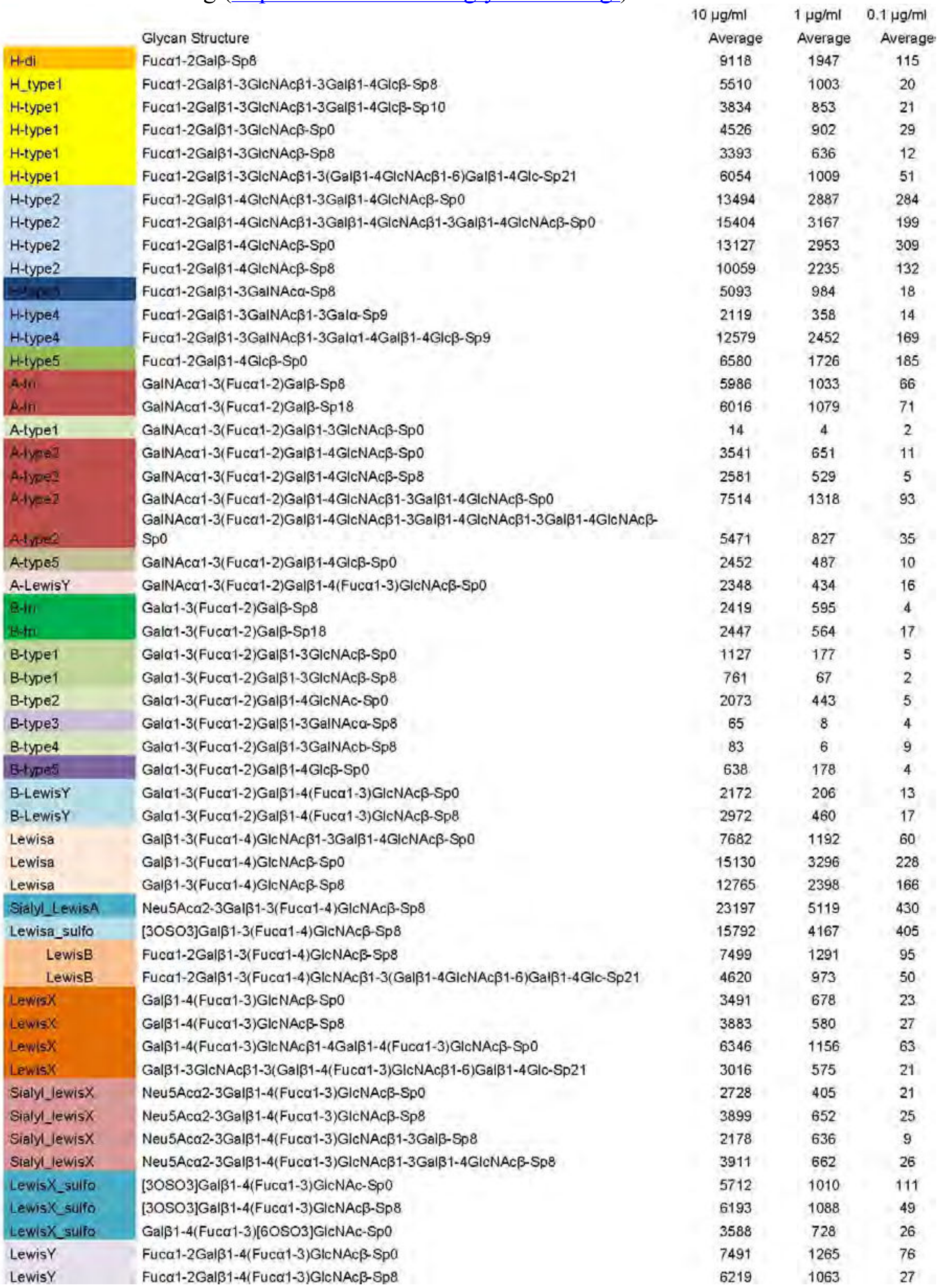


Figure S1. Energy maps for all glycosydic linkages of interest taken from Glyco3D (http://glyco3d.cermav.cnrs.fr/) as a function of Φ andΨ torsion angles with Φ = O5-C1-O1-C$_x$ (X-axis) and Ψ= C1-O1-C$_x$-C$_{x+1}$ (Y-axis). Red dots represent the lowest energy conformation.

Table S1: Binding intensities (FU → fluorescence units) for labeled LecB protein with glycan array chips v4.1 from the consortium for functional glycomics. Full data available on the web site of cfg (http://www.functionalglycomics.org/)

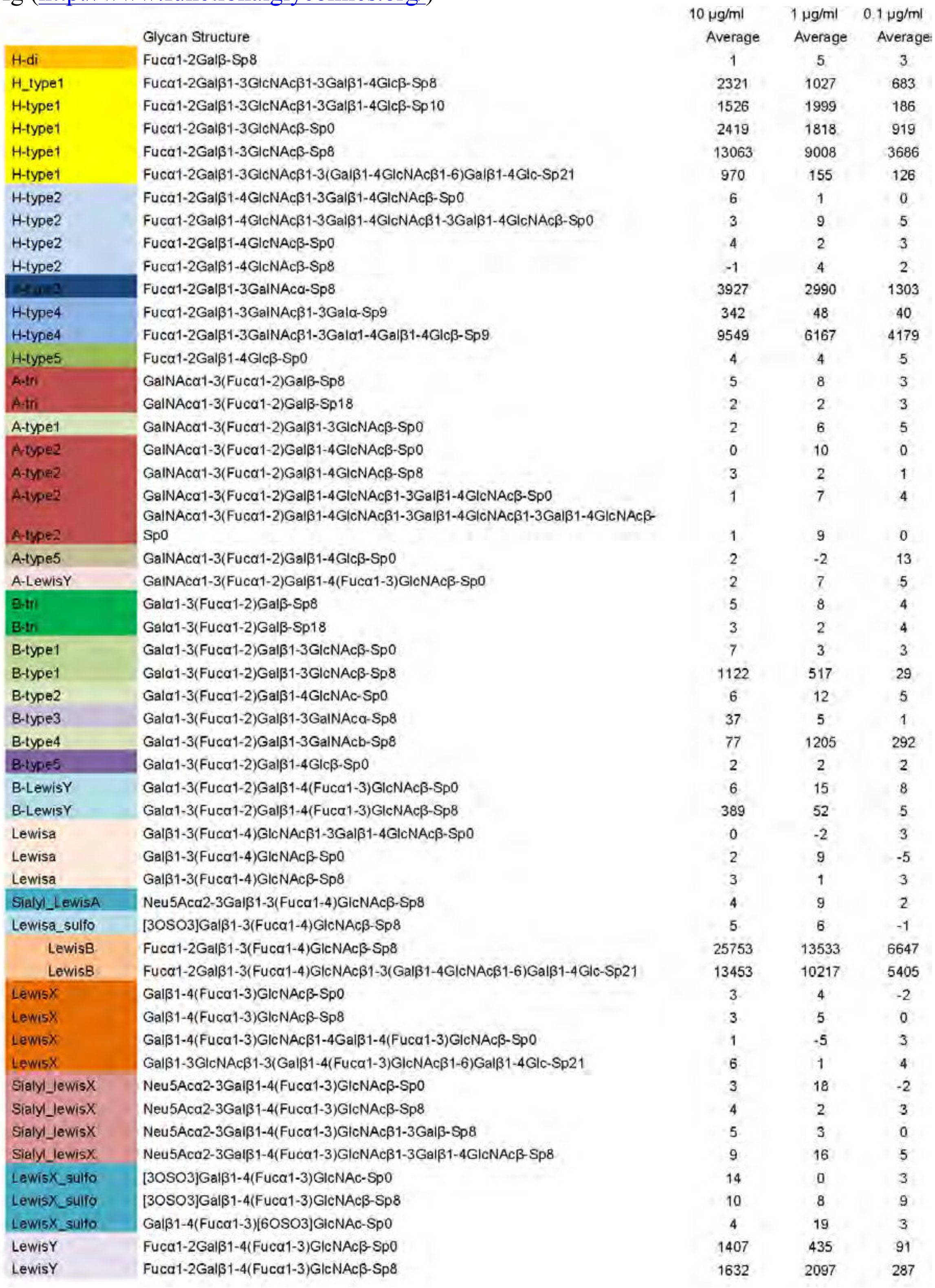


		10 µg/ml	1 µg/ml	0.1 µg/ml
	Glycan Structure	Average	Average	Average
H-di	Fucα1-2Galβ-Sp8	9118	1947	115
H_type1	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-Sp8	5510	1003	20
H-type1	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-Sp10	3834	853	21
H-type1	Fucα1-2Galβ1-3GlcNAcβ-Sp0	4526	902	29
H-type1	Fucα1-2Galβ1-3GlcNAcβ-Sp8	3393	636	12
H-type1	Fucα1-2Galβ1-3GlcNAcβ1-3(Galβ1-4GlcNAcβ1-6)Galβ1-4Glc-Sp21	6054	1009	51
H-type2	Fucα1-2Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	13494	2887	284
H-type2	Fucα1-2Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	15404	3167	199
H-type2	Fucα1-2Galβ1-4GlcNAcβ-Sp0	13127	2953	309
H-type2	Fucα1-2Galβ1-4GlcNAcβ-Sp8	10059	2235	132
H-type3	Fucα1-2Galβ1-3GalNAcα-Sp8	5093	984	18
H-type4	Fucα1-2Galβ1-3GalNAcβ1-3Galα-Sp9	2119	358	14
H-type4	Fucα1-2Galβ1-3GalNAcβ1-3Galα1-4Galβ1-4Glcβ-Sp9	12579	2452	169
H-type5	Fucα1-2Galβ1-4Glcβ-Sp0	6580	1726	185
A-tri	GalNAcα1-3(Fucα1-2)Galβ-Sp8	5986	1033	66
A-tri	GalNAcα1-3(Fucα1-2)Galβ-Sp18	6016	1079	71
A-type1	GalNAcα1-3(Fucα1-2)Galβ1-3GlcNAcβ-Sp0	14	4	2
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ-Sp0	3541	651	11
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ-Sp8	2581	529	5
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	7514	1318	93
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	5471	827	35
A-type5	GalNAcα1-3(Fucα1-2)Galβ1-4Glcβ-Sp0	2452	487	10
A-LewisY	GalNAcα1-3(Fucα1-2)Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	2348	434	16
B-tri	Galα1-3(Fucα1-2)Galβ-Sp8	2419	595	4
B-tri	Galα1-3(Fucα1-2)Galβ-Sp18	2447	564	17
B-type1	Galα1-3(Fucα1-2)Galβ1-3GlcNAcβ-Sp0	1127	177	5
B-type1	Galα1-3(Fucα1-2)Galβ1-3GlcNAcβ-Sp8	761	67	2
B-type2	Galα1-3(Fucα1-2)Galβ1-4GlcNAc-Sp0	2073	443	5
B-type3	Galα1-3(Fucα1-2)Galβ1-3GalNAcα-Sp8	65	8	4
B-type4	Galα1-3(Fucα1-2)Galβ1-3GalNAcb-Sp8	83	6	9
B-type5	Galα1-3(Fucα1-2)Galβ1-4Glcβ-Sp0	638	178	4
B-LewisY	Galα1-3(Fucα1-2)Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	2172	206	13
B-LewisY	Galα1-3(Fucα1-2)Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	2972	460	17
Lewisa	Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	7682	1192	60
Lewisa	Galβ1-3(Fucα1-4)GlcNAcβ-Sp0	15130	3296	228
Lewisa	Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	12765	2398	166
Sialyl_LewisA	Neu5Acα2-3Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	23197	5119	430
Lewisa_sulfo	[3OSO3]Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	15792	4167	405
LewisB	Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	7499	1291	95
LewisB	Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ1-3(Galβ1-4GlcNAcβ1-6)Galβ1-4Glc-Sp21	4620	973	50
LewisX	Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	3491	678	23
LewisX	Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	3883	580	27
LewisX	Galβ1-4(Fucα1-3)GlcNAcβ1-4Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	6346	1156	63
LewisX	Galβ1-3GlcNAcβ1-3(Galβ1-4(Fucα1-3)GlcNAcβ1-6)Galβ1-4Glc-Sp21	3016	575	21
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	2728	405	21
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	3899	652	25
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ-Sp8	2178	636	9
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp8	3911	662	26
LewisX_sulfo	[3OSO3]Galβ1-4(Fucα1-3)GlcNAc-Sp0	5712	1010	111
LewisX_sulfo	[3OSO3]Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	6193	1088	49
LewisX_sulfo	Galβ1-4(Fucα1-3)[6OSO3]GlcNAc-Sp0	3588	728	26
LewisY	Fucα1-2Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	7491	1265	76
LewisY	Fucα1-2Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	6219	1063	27

Table S2: Binding intensities (FU) for labeled BambL protein with glycan array chips v4.1 from the consortium for functional glycomics. Full data available on the web site of cfg (http://www.functionalglycomics.org/)

		10 µg/ml	1 µg/ml	0.1 µg/ml
	Glycan Structure	Average	Average	Average
H-di	Fucα1-2Galβ-Sp8	26289	7796	2255
H_type1	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-Sp8	5409	780	153
H-type1	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-Sp10	3345	481	97
H-type1	Fucα1-2Galβ1-3GlcNAcβ-Sp0	5730	864	154
H-type1	Fucα1-2Galβ1-3GlcNAcβ-Sp8	3942	749	134
H-type1	Fucα1-2Galβ1-3GlcNAcβ1-3(Galβ1-4GlcNAcβ1-6)Galβ1-4Glc-Sp21	5050	638	171
H-type2	Fucα1-2Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	23214	4190	1251
H-type2	Fucα1-2Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	25593	5653	1443
H-type2	Fucα1-2Galβ1-4GlcNAcβ-Sp0	24469	5203	1672
H-type2	Fucα1-2Galβ1-4GlcNAcβ-Sp8	22830	5809	1626
H-type3	Fucα1-2Galβ1-3GalNAcα-Sp8	6101	1054	309
H-type4	Fucα1-2Galβ1-3GalNAcβ1-3Galα-Sp9	1518	58	15
H-type4	Fucα1-2Galβ1-3GalNAcβ1-3Galα1-4Galβ1-4Glcβ-Sp9	4065	527	151
H-type5	Fucα1-2Galβ1-4Glcβ-Sp0	13229	3390	1401
A-tri	GalNAcα1-3(Fucα1-2)Galβ-Sp8	12180	2348	786
A-tri	GalNAcα1-3(Fucα1-2)Galβ-Sp18	21223	4272	1161
A-type1	GalNAcα1-3(Fucα1-2)Galβ1-3GlcNAcβ-Sp0	12	5	8
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ-Sp0	1269	43	9
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ-Sp8	935	58	14
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	3428	335	69
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	1284	88	15
A-type5	GalNAcα1-3(Fucα1-2)Galβ1-4Glcβ-Sp0	1158	60	19
A-LewisY	GalNAcα1-3(Fucα1-2)Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	1280	139	28
B-tri	Galα1-3(Fucα1-2)Galβ-Sp8	11862	2914	836
B-tri	Galα1-3(Fucα1-2)Galβ-Sp18	8217	1401	382
B-type1	Galα1-3(Fucα1-2)Galβ1-3GlcNAcβ-Sp0	274	5	10
B-type1	Galα1-3(Fucα1-2)Galβ1-3GlcNAcβ-Sp8	96	6	12
B-type2	Galα1-3(Fucα1-2)Galβ1-4GlcNAc-Sp0	2000	119	32
B-type3	Galα1-3(Fucα1-2)Galβ1-3GalNAcα-Sp8	6	1	6
B-type4	Galα1-3(Fucα1-2)Galβ1-3GalNAcb-Sp8	12	14	4
B-type5	Galα1-3(Fucα1-2)Galβ1-4Glcβ-Sp0	383	29	10
B-LewisY	Galα1-3(Fucα1-2)Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	813	38	8
B-LewisY	Galα1-3(Fucα1-2)Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	1808	241	22
Lewisa	Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	3885	503	99
Lewisa	Galβ1-3(Fucα1-4)GlcNAcβ-Sp0	10858	1976	342
Lewisa	Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	11845	1975	296
Sialyl_LewisA	Neu5Acα2-3Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	4115	568	177
Lewisa_sulfo	[3OSO3]Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	7890	1064	355
LewisB	Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	17009	3043	992
LewisB	Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ1-3(Galβ1-4GlcNAcβ1-6)Galβ1-4Glc-Sp21	7712	1182	481
LewisX	Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	5379	831	99
LewisX	Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	6660	1164	216
LewisX	Galβ1-4(Fucα1-3)GlcNAcβ1-4Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	8701	1448	292
LewisX	Galβ1-3GlcNAcβ1-3(Galβ1-4(Fucα1-3)GlcNAcβ1-6)Galβ1-4Glc-Sp21	4805	783	226
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	1176	110	5
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	2116	299	65
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ-Sp8	864	150	11
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp8	1744	227	32
LewisX_sulfo	[3OSO3]Galβ1-4(Fucα1-3)GlcNAc-Sp0	4119	596	189
LewisX_sulfo	[3OSO3]Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	4777	672	160
LewisX_sulfo	Galβ1-4(Fucα1-3)[6OSO3]GlcNAc-Sp0	4355	546	124
LewisY	Fucα1-2Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	22439	5562	1021
LewisY	Fucα1-2Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	19573	3642	824

Table S3: Binding intensities (FU) for labeled Bc2L-C-nt protein with glycan array chips v4.1 from the consortium for functional glycomics. Full data available on the web site of cfg (http://www.functionalglycomics.org/)



		10 µg/ml	1 µg/ml	0.1 µg/ml
	Glycan Structure	Average	Average	Average
H-di	Fucα1-2Galβ-Sp8	1	5	3
H_type1	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-Sp8	2321	1027	683
H-type1	Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-Sp10	1526	1999	186
H-type1	Fucα1-2Galβ1-3GlcNAcβ-Sp0	2419	1818	919
H-type1	Fucα1-2Galβ1-3GlcNAcβ-Sp8	13063	9008	3686
H-type1	Fucα1-2Galβ1-3GlcNAcβ1-3(Galβ1-4GlcNAcβ1-6)Galβ1-4Glc-Sp21	970	155	126
H-type2	Fucα1-2Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	6	1	0
H-type2	Fucα1-2Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	3	9	5
H-type2	Fucα1-2Galβ1-4GlcNAcβ-Sp0	4	2	3
H-type2	Fucα1-2Galβ1-4GlcNAcβ-Sp8	-1	4	2
H-type3	Fucα1-2Galβ1-3GalNAcα-Sp8	3927	2990	1303
H-type4	Fucα1-2Galβ1-3GalNAcβ1-3Galα-Sp9	342	48	40
H-type4	Fucα1-2Galβ1-3GalNAcβ1-3Galα1-4Galβ1-4Glcβ-Sp9	9549	6167	4179
H-type5	Fucα1-2Galβ1-4Glcβ-Sp0	4	4	5
A-tri	GalNAcα1-3(Fucα1-2)Galβ-Sp8	5	8	3
A-tri	GalNAcα1-3(Fucα1-2)Galβ-Sp18	2	2	3
A-type1	GalNAcα1-3(Fucα1-2)Galβ1-3GlcNAcβ-Sp0	2	6	5
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ-Sp0	0	10	0
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ-Sp8	3	2	1
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	1	7	4
A-type2	GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	1	9	0
A-type5	GalNAcα1-3(Fucα1-2)Galβ1-4Glcβ-Sp0	2	-2	13
A-LewisY	GalNAcα1-3(Fucα1-2)Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	2	7	5
B-tri	Galα1-3(Fucα1-2)Galβ-Sp8	5	8	4
B-tri	Galα1-3(Fucα1-2)Galβ-Sp18	3	2	4
B-type1	Galα1-3(Fucα1-2)Galβ1-3GlcNAcβ-Sp0	7	3	3
B-type1	Galα1-3(Fucα1-2)Galβ1-3GlcNAcβ-Sp8	1122	517	29
B-type2	Galα1-3(Fucα1-2)Galβ1-4GlcNAc-Sp0	6	12	5
B-type3	Galα1-3(Fucα1-2)Galβ1-3GalNAcα-Sp8	37	5	1
B-type4	Galα1-3(Fucα1-2)Galβ1-3GalNAcb-Sp8	77	1205	292
B-type5	Galα1-3(Fucα1-2)Galβ1-4Glcβ-Sp0	2	2	2
B-LewisY	Galα1-3(Fucα1-2)Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	6	15	8
B-LewisY	Galα1-3(Fucα1-2)Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	389	52	5
Lewisa	Galβ1-3(Fucα1-4)GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp0	0	-2	3
Lewisa	Galβ1-3(Fucα1-4)GlcNAcβ-Sp0	2	9	-5
Lewisa	Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	3	1	3
Sialyl_LewisA	Neu5Acα2-3Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	4	9	2
Lewisa_sulfo	[3OSO3]Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	5	6	-1
LewisB	Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ-Sp8	25753	13533	6647
LewisB	Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ1-3(Galβ1-4GlcNAcβ1-6)Galβ1-4Glc-Sp21	13453	10217	5405
LewisX	Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	3	4	-2
LewisX	Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	3	5	0
LewisX	Galβ1-4(Fucα1-3)GlcNAcβ1-4Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	1	-5	3
LewisX	Galβ1-3GlcNAcβ1-3(Galβ1-4(Fucα1-3)GlcNAcβ1-6)Galβ1-4Glc-Sp21	6	1	4
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	3	18	-2
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	4	2	3
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ-Sp8	5	3	0
Sialyl_lewisX	Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ1-3Galβ1-4GlcNAcβ-Sp8	9	16	5
LewisX_sulfo	[3OSO3]Galβ1-4(Fucα1-3)GlcNAc-Sp0	14	0	3
LewisX_sulfo	[3OSO3]Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	10	8	9
LewisX_sulfo	Galβ1-4(Fucα1-3)[6OSO3]GlcNAc-Sp0	4	19	3
LewisY	Fucα1-2Galβ1-4(Fucα1-3)GlcNAcβ-Sp0	1407	435	91
LewisY	Fucα1-2Galβ1-4(Fucα1-3)GlcNAcβ-Sp8	1632	2097	287