THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ de GRENOBLE

Discipline: Biologie Structurale et Nanobiologie

Thèse dirigée, présentée et soutenue publiquement par

M^{lle} Alessandra NURISSO

03 Mai 2010

Etudes *in silico* des interactions protéines - carbohydrates

Directeur de thèse: Dr. Anne Imberty

JURY

Prof. Jaroslav Koca, rapporteur

Dr. Alexandre G. de Brevern, rapporteur

Dr. Jesus Jimenez Barbero, membre du jury

Dr. Julie Cullimore, membre du jury

Dr. Anne Imberty, membre du jury

Dr. Serge Pérez, président du jury

Thèse préparée au Centre d'Etude et de Recherche des Macromolécules Végétales
Equipe de Glycobiologie Moléculaire (tel-00430632)

For the persons I love...

Resumé

Dans cette thèse, des méthodes classiques de modélisation moléculaire ont été utilisées pour élucider les caractéristiques structurales et dynamiques des glucides, libres ou en complexe avec les protéines. Deux sujets principaux ont été abordés: i) les lectines calcium dépendantes et les interactions avec leurs substrats glycosidiques, importants au niveau thérapeutique; ii) les interactions entre les glycolipides bactériens et les récepteurs d'origine végétale qui caractérisent la symbiose Rhizobium-légumineuses, processus clé pour comprendre le mécanisme de fixation de l'azote. La légumineuse Medicago truncatula qui vit en symbiose avec l'espèce rhizobiale Sinorhizobium meliloti sera utilisée comme modèle. i) Différentes approches d'amarrage moléculaire (docking) sur des systèmes lectines calcium dépendantes – sucres ont été comparées, afin de reproduire au mieux les données expérimentales. En utilisant l'approche la plus appropriée, les propriétés du site de liaison de la Langerine, une lectine humaine, ont été étudiées, en rationalisant le mécanisme à la base de la reconnaissance de l'épitope oligosaccharidique du virus VIH. Les caractéristiques structurales de la lectine Pseudomonas aeruginosa I ont été également étudiées par des calculs d'amarrage moléculaire, de dynamique moléculaire et d'énergie libre. ii) Une analyse conformationnelle des facteurs de nodulation (facteurs Nod), signaux bactériens nécessaires à l'induction de la symbiose, a révélé des états conformationnels spécifiques en accord avec les résultats obtenus par RMN. Les modèles par homologie du domaine LysM2 et du domaine catalytique de la kinase LYK3, récepteurs hypothétiques des facteurs Nod exprimés sur la membrane des cellules racinaires des légumineuses, ont été construits et utilisés comme support des données biologiques.

Abstract

In this thesis, classical molecular modeling methodologies have been used to elucidate structural and energetic features of carbohydrates, in a free state and in complex with proteins. Two main subjects have been considered: i) calcium dependent lectins and the therapeutically important interactions with their glycosidic substrates; ii) the interaction between bacterial glycolipids and plant protein receptors in the rhizobia-legume symbiosis, key process for understanding the nitrogen fixation mechanism. The legume Medicago truncatula and its symbiont Sinorhizobium meliloti will be used as model. i) Different flexible docking approaches on calcium dependent lectin-carbohydrate complexes have been compared, highlighting the possibility to reproduce in silico structural experimental data. Using the most appropriate docking approach, the binding site properties of langerin, a human lectin, have been elucidated, rationalizing the mechanism at the basis of HIV oligo-saccharide epitope recognition. Structural features of the Pseudomonas aeruginosa lectin 1, essential for the development of drugs against the bacterium and of diagnostics tools, have been also investigated through the application of docking, molecular dynamics and free energy calculations. ii) A conformational analysis of natural and synthetic nodulation factors, bacteria signals necessary for the symbiosis induction, revealed major conformational states in agreement with NMR measurements. Homology models of LysM domains 2 and of the catalytic domain of the kinase LYK3, hypothetic membrane-located nod factor receptors expressed in legume root cells, have been built to support biological data.

Introduction générale

Les interactions entre les protéines et les glucides sont à la base de la communication et de l'adhésion entre cellules. Plusieurs approches bioinformatiques peuvent être appliquées pour étudier ces interactions, en supportant les résultats obtenus par des approches expérimentales.

La première section de cette thèse présente une description détaillée des glucides, leurs propriétés structurales et leur importance biologique. Les lectines, protéines qui reconnaissent spécifiquement les glucides, sont également présentées, en expliquant les mécanismes généraux et les paramètres thermodynamiques qui regissent ces interactions.

Dans la deuxième section les techniques de modélisation moléculaire utilisées dans ce travail sont presentées et amplement décrites. À la fin de cette partie, le lecteur non spécialisé atteint une vision globale des techniques utilisées, de leurs limitations et des liens avec les données expérimentales.

La troisième section présente les résultats obtenus suite à l'application des approches computationnelles sur les systèmes biologiques. Tous les travaux sont présentés sous forme d'articles scientifiques en incluant une introduction complémentaire. Le premier article (publié dans *Molecular Simulation*) pose une question essentielle: comment se comportent différents logiciels d'amarrage (*docking*) dans le cadre des interactions glucides et protéines (ici des lectines dépendantes du calcium). Utilisant un ensemble représentatif de lectines, les tendances observées ainsi que la complexité de l'approche sont montrées.

Un site probable de fixation du sucre d'une lectine dépendante du calcium a été ainsi proposé alors qu'il n'avait pas été encore caractérisé. Le deuxième article (publié dans *Biochemistry*) montre un exemple de collaborations entre approches *in vitro* et *in silico*, combinants résolution de structures par crystallographie, modélisation, simulation, détermination expérimentale du rayon de gyration et microscopie électronique à transmission. À l'aide de l'ensemble de ces techniques, un modèle du trimère de l'ECD de la Langerine, en particulier de son site de fixation a été proposé, amenant à décrire un mécanisme de la déformation des membranes des granules de Birbeck.

Les troisièmes et quatrièmes articles publiés respectivement dans *Journal of Molecular Biology* & *Journal of Biochemical Chemistry*) portent sur une lectine importante d'un point de vue médical, une lectine du pathogène opportuniste *Pseudomonas Aeruginosa*. Toujours en couplant approches *in silico* et données expérimentales, l'importance de la flexibilité des disaccharides impliqués dans l'interaction a été ici plus précisément analysée. Dans le deuxième article sur cette lectine, l'importance des molécules d'eau a été mise en avant.

Les travaux plus récents (deux articles en préparation) portent sur une question importante d'un point de vue scientifique et économique: la symbiose entre les bactéries de type rhizobia et les légumineuses. Les facteurs Nod secrétés par la bactérie sont des lipochitoligo-saccharides essentiels pour la plante. Le rôle précis de la chaine lipidique est source de discussions dans la communauté scientifique. A partir de données de Résonnance Magnétique Nucléaire, la dynamique de cette chaine et surtout la stabilité de certains intermédiaires ont été analysées, amenant à émettre des hypothèses sur le mécanisme d'action de la macromolécule. Les modèles par homologie des récepteurs des facteurs Nod exprimés sur la membrane des cellules racinaires des légumineuses, ont été construits et minutieusement décrits.

Les deux dernières sections présentent les conclusions, perspectives et les annexes qui contiennent les scripts ainsi que les modes opératoires utilisés dans ce travail. Un dernier article accepté pour publication dans *Glycobiology* est encore présenté: differents outils de modélisation moléculaire ont été utilisés pour évaluer les caractéristiques conformationelles de l'acide hyaluronique.

Acknowledgements

Thanks go to: Dr. Anne Imberty, for the encouragement and guidance. Dr. Serge Perez and Dr. Jesus J. Barbero, for their enthusiasm and lively and interactive discussions; Dr. Julie Cullimore and all the European Nod Perception network for the opportunity they gave me to carry out my work; the entire CERMAV, in particular to Dr.Alain Rivet for providing me endless opportunities to improve my computer skills; Dr. Aline Thomas for proof reading; Davide Ceresetti, for his precious help and for always being there as my support and strength; Anna Szarpak, Camelia Stinga, Roberto Nervo, Michael Reynolds, Bertrand Blanchard, Teresa Ierano', Karoline Saboia Aragao, Windson Carvalho and all the Brazilian community in Grenoble, for their friendship and cheerfulness; Maria Morando for scientific and funny escapades; Eleonora Lombardo, Simona Carrisi, the Angel's dream gospel choir for keeping me sane; Dr. Giulia Caron and Prof. Giuseppe Ermondi for giving me the courage to continue my career in science; Paolo Rossi for his precious and true friendship; and my parents, Felipe and Mao for the patience and the unconditional love.

Contents

SECTION 1
INTRODUCTION

1 Introduction

1.1 Glycobiology: a brief overview

Glycobiology is the study of the structure, biosynthesis, and biology of saccharides or glycans. Together with peptides and lipids, they form one of the most abundant classes of biomolecules in nature. Glycans are necessary for structural support, energy storage and as mediators of a huge variety of biological events and thus are fundamental for the development, equilibrium and functionality of living organisms[1, 2]. They are generally located on the outer surface of cellular macromolecules, often in covalent combination with proteins or lipids, forming molecular aggregates known as glycoconjugates (Figure 1-1). The presence of glycoconjugates characterizes the interface between a cell and its outside environment and allows cell-cell recognition, discrimination and adhesion processes[2].

The importance of glycans in humans can be demonstrated by considering the enzymatic mechanism of glycosylation, the post translational modification necessary for the construction of glycoconjugates. Enzymatic deficiencies or errors during this process lead to serious medical consequences[3] including congenital muscular dystrophies[4-6], cancers[7-9], rheumatoid arthritis[10], tuberculosis[11], ulcerative colitis and Crohn's disease[12,13]. Glycosylation represents a promising target for the development of clinically useful biomarkers and therapies[14-17]. There is also interest in developing therapeutic compounds able to modulate or to inhibit carbohydrate-pathogen interactions.

Easily accessible, sugars exposed on the mammalian cells represent exciting binding opportunities for opportunistic pathogens ranging from toxins to viruses, from primitive bacteria to sophisticated eukaryotic parasites (Figure 1-1), promoting the initiation of infectious diseases[18-20]. For example, the influenza virus specifically binds through hemagglutinin onto sialic acid sugars on the surfaces of epithelial cells typically located in the nose, throat and lungs of mammals, characterizing the first phase of the infection[21]. Moving to microbes, *Escherichia coli* adheres by its lectins, proteins binding carbohydrates, to glycoproteins rich in mannose exposed on epithelial cells of the gastrointestinal or urinary tract leading to infections[22].

The ability of the immune system to recognize glycans on cell surfaces, identifying minor chemical changes of carbohydrate structures, has also profound biomedical implications,

including the suppression of immune rejection of organ transplants, blood transfusions and carbohydrate-based vaccines[1, 23, 24].

Carbohydrates show high flexibility and structural variability, populating multiple conformational states which coexist in equilibrium under physiological conditions. Each particular sugar conformation contains specific biological information. Thus, structural properties of carbohydrates need to be elucidated to fully understand the associated biological functions.

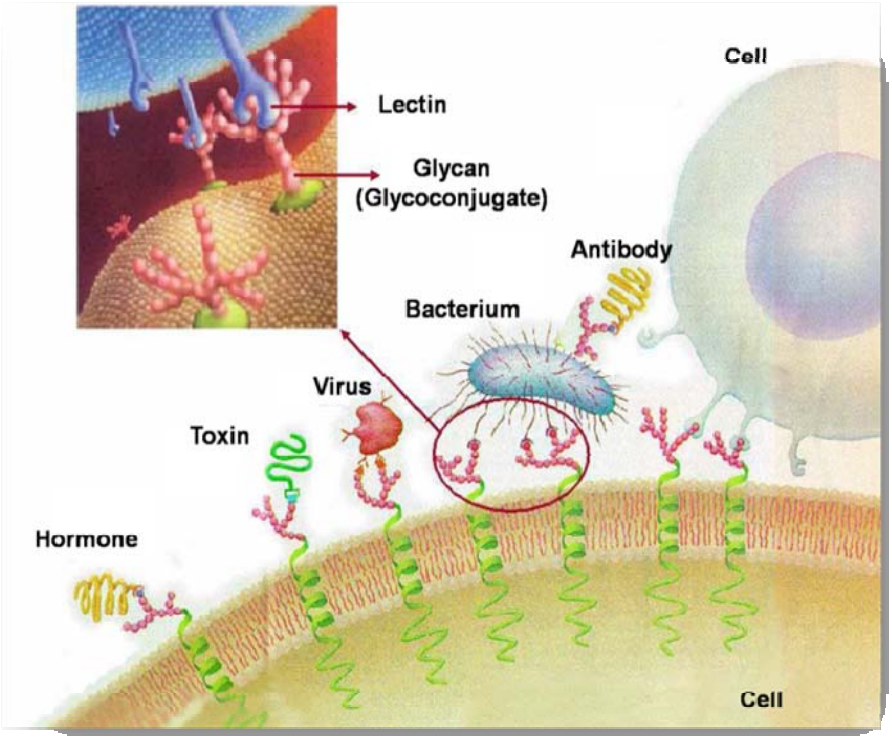


Figure 1-1 - Schematic representation of mammalian cell surfaces. At the molecular level, pathogens (toxins, virus and bacteria) exploit carbohydrate structures that decorate mammalian cell surfaces, often using those structures as a means to infect such cells. Antibodies, produced by the mammalian immune system, are sensitive to chemical changes of carbohydrate structures, recognizing external antigens and producing immune responses. Adapted from[25].

1.2 Carbohydrates, the heart of glycobiology

1.2.1 Structural features of carbohydrates

From a chemical point of view, carbohydrates can be defined as polyhydroxy aldehydes or ketones. They are commonly classified depending on the complexity of their formula as monosaccharides, disaccharides, oligosaccharides and polysaccharides[26].
A monosaccharide is the smallest sugar unit whereas a disaccharide consists of two monosaccharide units linked by a glycosidic linkage.
The term oligosaccharide is used to describe carbohydrates containing more than three monosaccharides whereas larger molecules composed of more than twenty monosaccharide units have been named polysaccharides.

1.2.2 Monosaccharides

Monosaccharides, represented by the empirical formula $C_n(H_2O)_n$, are interconvertible acyclic-cyclic hemiacetal or hemiketal units generally composed of five or more carbon atoms from which all complex carbohydrate structures are created. The equilibrium between a hydroxyaldeyde and the corresponding cyclic form is in favor of the heterocycle. Five – six membered forms can be created, corresponding to furanose or pyranose species respectively (Figure 1-2).

Configurations.

Monosaccharides exist in nature as enantiomers, non super-imposable compounds with the same molecular formula but different geometrical positioning of atoms and functional groups in space which are the exact mirror images of each other. They both differ in tendency to rotate the plane of polarized light. Depending on their geometries, enantiomers can be designated as D- or L- as prefixes of the compound names. The prefixes D- and L- refer to the configuration of the carbon atom farthest from the carbonyl group (Figure 1-3).

Anomeric forms and anomeric effect.

From the nucleophilic attack of the hydroxyl group at the carbonylic carbon, a new chiral center is generated.

The anomeric carbon is the center of the new formed hemiketal functional group. Two different configurations (anomers) may result in the new cyclic form according to the relationship between the anomeric and the most distant stereogenic center.

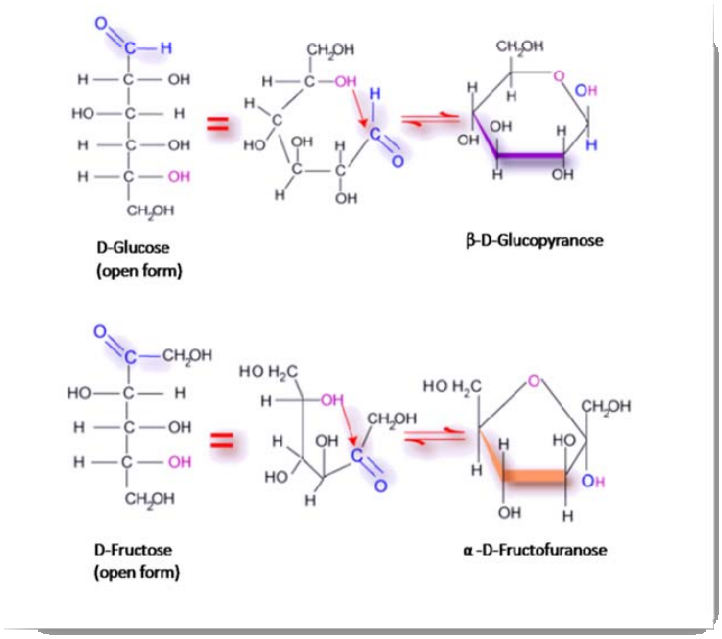


Figure 1-2 - In aldoses like D-glucose, the reaction between the aldehyde group in position 1 and the hydroxyl pendent group linked to the carbon atom C in position 5 leads to the formation of six membered cyclic rings. In the open chain structures of ketoses like D-Fructose, the keto group (carbon atom C in position 2) interacts with the hydroxyl group linked to the carbon atom C in position 5 resulting in the five membered cyclic rings.

The difference in terms of optical activity was used as the basis for classifying the anomers as α and ß[26]. In the D-series, the more dextrorotatory anomer of each sugar is designated α and the less dextrorotatory is designated as ß. In the L-series, the more levorotatory anomer is designated α whereas the lesser one is designed as ß[27].

In solution, a rapid inter conversion between α and ß via the acyclic form is observed, with changes in optical activity (mutarotation). However, α and ß configurations are not equally represented in solution. The equilibrium is commonly shifted to favor axial configurations (anomeric effect). This preference can be explained by the formation of a stabilizing hyper-conjugation between molecular orbitals (n_p) associated with the ring oxygen atom (O5) and the orbitals of the adjacent C1-O1 bond (σ*)[28]. Nevertheless, there are other interpretations of the origin of this effect, as the dipole-dipole repulsion theory[29, 30]. The anomeric effect is not only associated with O-C-O fragments but in general with systems of the type X-A-Y where X and Y are more electronegative then A[26].

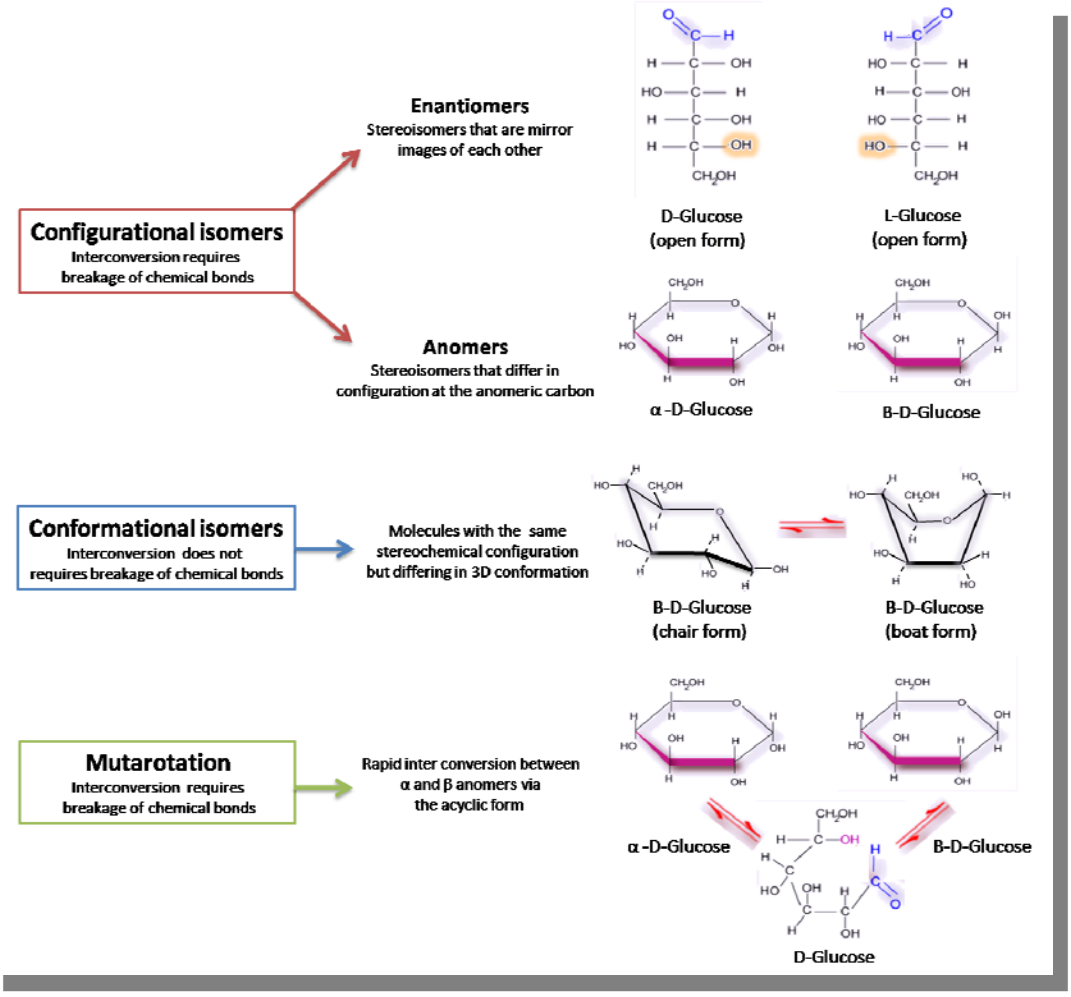


Figure 1-3 - Summary of the main structural features of monosaccharides.

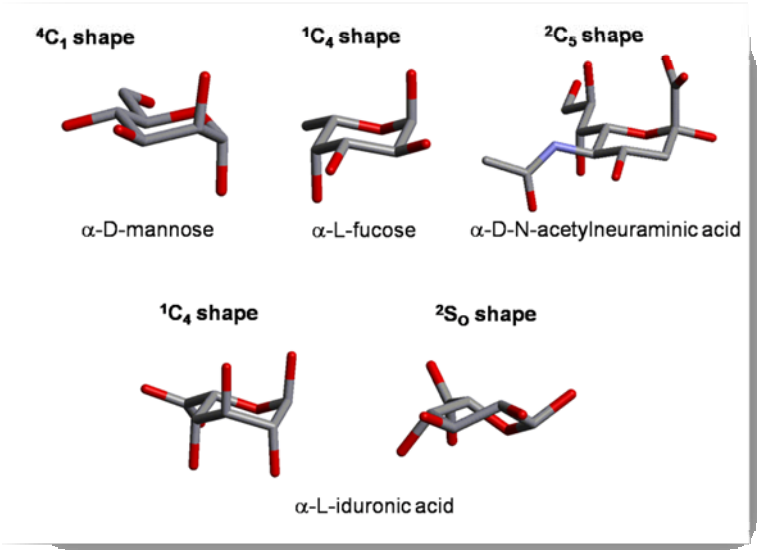


Figure 1-4 - Examples of possible favored shapes of pyranose rings.

Ring shapes.

The glycosidic rings can adopt different conformations that can be described using puckering parameters introduced by Cremer and Pople (Figure 1-5)[31]. In the case of pyranoses, chair(C) shapes are generally the most representative conformations, with C2, C3, C5 and O5 atoms all laying in the same plane. The most favored conformers are generally the $^4C_1$ (C4 is above the plane and C1 below the plane) and $^1C_4$ (C1 is above the plane and C4 below the plane) for D- and L- pyranose rings, respectively. In addition to the chair conformations, boat (B) and twist boat/skew-boat (TB or S) conformations can also be realized for pyranose rings. The iduronic acid, for example, can adopt more than one conformation in solution, with an equilibrium existing between three low energy conformers: $^1C_4$ and $^4C_1$ chair forms and an additional $^2S_0$ skew-boat conformation. When internally positioned within an oligosaccharide, the $^1C_4$ and $^2S_0$ conformations predominate[32] (Figure 1-4).

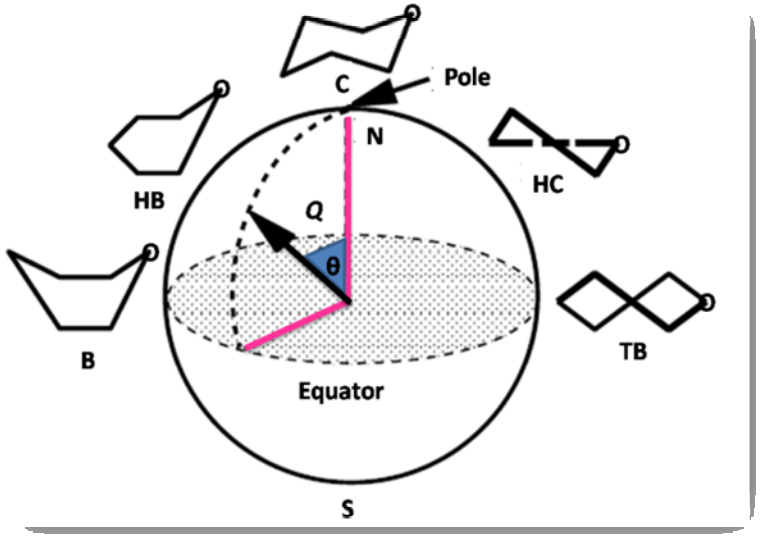


Figure 1-5 - Graphical representation of the Cremer-Pople puckering parameters for a pyranose ring. Total puckering amplitude, Q, is the radius of the sphere corresponding to the sum of the perpendicular distance of each ring atom to the ring average plane. Polar angle, θ, indicates degree of deviation from $^4C_1$ chair conformation at the North Pole (N). Other conformations as HB (half boat), HC (half chair), B (boat), and TB or S (twist boat or skew boat) are located at the equator. The $^1C_4$ chair form is at the South Pole (S). The theoretical values of Q and θ are 0.63 Å and 0° for pure $^4C_1$ chair conformation in cyclohexane[31].

Flexibility of the pendant groups

The orientation of exocyclic groups is a source of flexibility in carbohydrates. The hydroxymethyl group at C5, whose flexibility is determined by the torsion angle ω defined as Θ(O5-C5-C6-O6), can adopt a restricted number of conformations: *gauche-gauche* (*gg*), *gauche-trans* (*gt*) and *trans-gauche* (*tg*) (Figure 1-6)[33]. It was observed that, in solution, it prefers *gauche* orientations. This phenomenon depends more on solvation interactions rather than on stereo-electronic effects. In particular, *gg* conformers are the most favored conformations

found in D-Glucopyranose monomers whereas the *gt* are the most common D-galactopyranose conformers in solution[34].

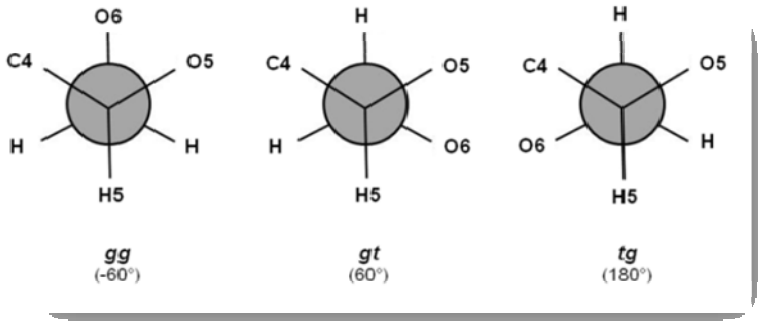


Figure 1-6 - Newman projections of the gauche-gauche (gg), gauche-trans (gt) and trans-gauche (tg) conformers of the ω dihedral angle (O5-C5-C6-O6 and C4-C5-C6-O6 ω torsion angles).

The secondary hydroxyl groups can generally rotate in space adopting staggered orientations. However, two major orientations (clockwise and anticlockwise around the ring) allow the formation of intra molecular hydrogen bonds stabilizing the structure[35].

1.2.3 Disaccharides

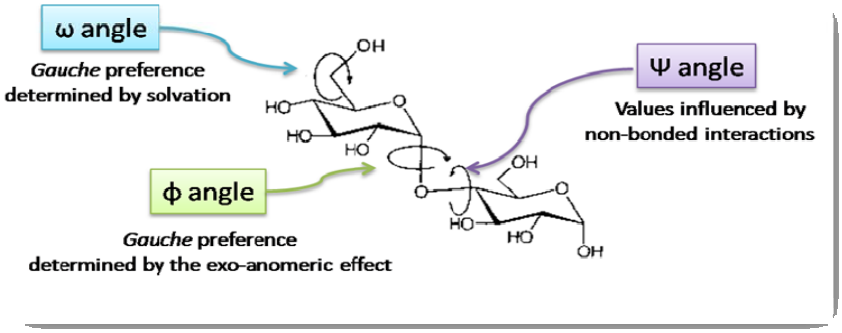
Two sugar monomers are combined together to form a disaccharide (Figure 1-7). Normally, sugar polymers are characterized by having reducing and non-reducing ends. The non-reducing end is the monomer in which the anomeric carbon is linked, by a glycosidic bond, to another monomer, preventing opening of the ring to the aldehyde or keto form. The reducing end is the monomer in which the anomeric carbon is not involved in a linkage and thus, in equilibrium with the respective open-ring form.

Glycosidic linkage

In any disaccharide, the orientation of the two pyranose rings can be easily characterized by the torsion angles φ and ψ that define the glycosidic bond. φ and ψ are defined as Θ(O5–C1–O–Cx) and Θ(C1–O1–Cx–C(x+1)), respectively. For (1-6) linkages, another torsion angle is required, usually denoted as χ Θ(O1–C6–O5–C5) torsion angle[26].
In analogy to Ramachandran's conformational plots, similar diagrams can be established for the disaccharides using Molecular Mechanics. The minima in such maps define energetically preferred conformations of the sugar dimer. Values found in theory can be further validated by comparison to experimental data (§2.3.2.1).

Effects influencing the disaccharide conformation

The *exo-anomeric* effect can be attributed to the internal stereoelectronic and hyperconjugative properties of X-C-Y fragments, where X and Y are electronegative atoms (§1.2.2). The exo-anomeric effect influences the φ-angle properties, arising from hyperconjugation within the O5-C1-O1 atomic sequence. The superposition of the orbitals n_p from O1 and σ* from O5-C1 bond shows a rotameric *gauche* preference around the exocyclic C1-O1 that defines the φ-angle [36]. The angle Ψ, associated with the Θ(C1–O1–Cx–C(x+1)) sequence, does not exhibit any particular stereoelectronic property (Figure 1-7)



.

Figure 1-7 - Summary of the effects of φ, ψ, and ω dihedral angles that influence a disaccharide conformations.

1.2.4 Oligosaccharides and glycoconjugates

Oligosaccharides are short and flexible, linear or branched chains, composed by a small number of sugar units, typically three to twenty. In nature, oligosaccharides generally occur as glycoconjugates, covalently attached to proteins or lipids, through the enzymatic glycosilation mechanism, to form glycoproteins and glycolipids, respectively[26]. Glycoproteins can be classified according to the type of linkage that connects carbohydrate-protein moieties in N- or O-glycans. N-linked glycans are oligosaccharides linked to asparagines Asn residues (Figure 1-8) whereas O-linked glycans are bound to aminoacid residues with hydroxyl groups characterizing their side chains (i.e. serine, threonine)[26].

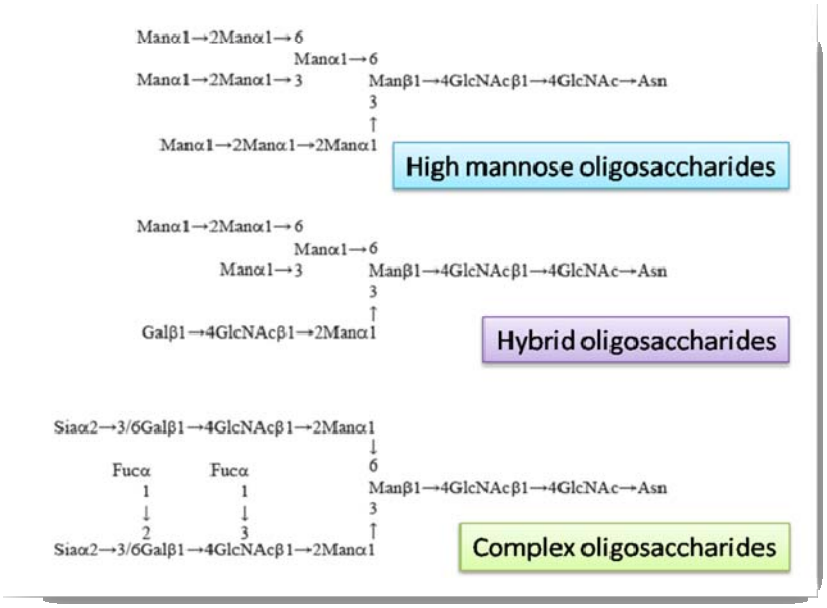


Figure 1-8 - The three major classes of N*-linked glycans identified in mammals. High mannose N-glycans contains just two* N*-acetylglucosamines with many mannose residues whereas complex oligosaccharides are formed by different combinations of mannose (Man), N-acetylglucosamine (GlcNAc), N-acetylgalactosamine (GalNAc), fucose (Fuc) and sialic acid (Sia) residues. Hybrid oligosaccharides are intermediate between high-mannose and complex oligosaccharides[26].*

Oligosaccharides formed by long unbranched anionic saccharidic chains covalently linked to proteins have been named glycosaminoglycans (Figure 1-9). They include glucosaminoglycans (heparin, heparan sulfate), galactosylaminoglycans (chondroitin sulfate and dermatan sulfate), hyaluronic acid and keratan sulfate. The hyaluronic acid is the only glycosaminoglycan synthesized without a protein core[26, 37, 38].

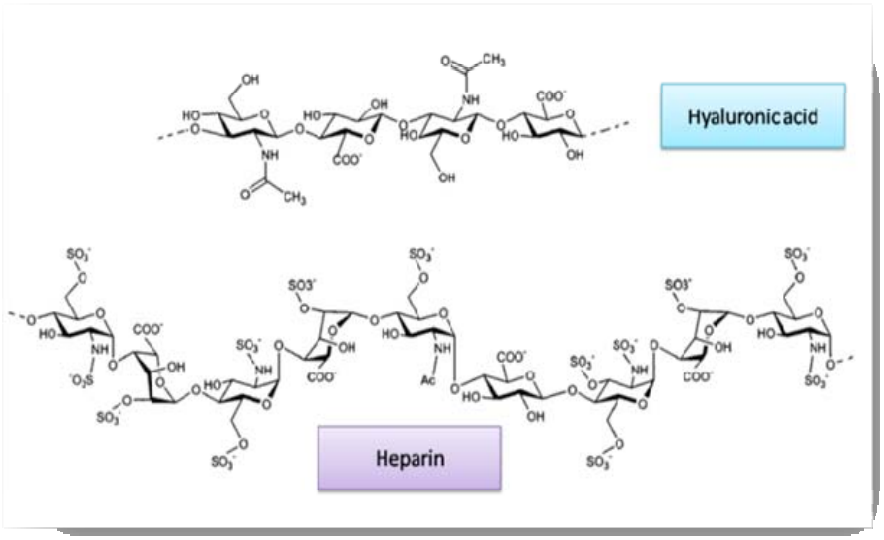


Figure 1-9 - Representation of two glycosaminoglycans: heparin, formed by [-4IdoUA(2S)α1-4GlcNS(6S)α1-] disaccharide repeating units and hyaluronic acid, composed by [-4GlcUAβ1-3GlcNAcβ1-] disaccharide units[38]. IdoUA(S2), GlcNS(6S), GlcUA and GlcNAc are the abbreviations for the monomers 2-O-Sulfonato-L-iduronate, 2-deoxy-2-sulfamido-α-D-glucopyranosyl-6-O-sulfateglucuronate, Glucuronic acid and N-acetylglucosamine respectively.

The term glycolipid designates any compound containing one or more monosaccharide residues bound by a glycosidic linkage to a hydrophobic moiety. Glycolipids can be classified

on the basis of their hydrophobic portion in glycoglycerolipids, glycophosphatidylinositols and glycosphingolipids(Figure 1-10)[39].

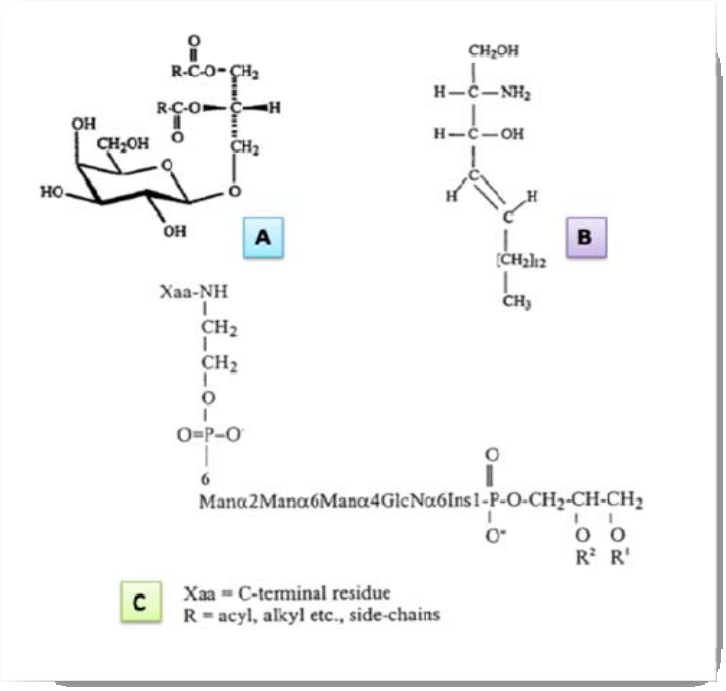


Figure 1-10 - Examples of glycolipids classified according to their lipid chains: 1,2- di-O-acyl-3-O-β-D-galactosyl-sn-glycerol (A, glycoglycerolipid); sphingosine (B, glycosphingolipid); core structure of glycophosphatidylinositols (C)[39].

The global conformation of oligosaccharides is difficult to be determined because of the high number of torsion angles that are free to rotate and because the structure is not simply influenced by the stereo-electronic properties elucidated for simple disaccharides (§1.2.3). Polar contacts and hydrophobic interactions can occur between successive and non-successive glycosidic units[38]. The solvent can also play a key role in influencing the global

shape of oligosaccharides[40, 41] and, sometimes, ions are also capable of interacting with the glycosidic hydroxyl groups, via electrostatic interactions or metal ion coordination[42].
To understand the structural variability and the complexity in studying the conformational properties of oligosaccharides, a simple trisaccharide is considered as an example (Figure 1-11). This compound can assume 18^4 different conformations, if only the φ and ψ torsion angles are assumed free to rotate by steps of 20°. If the flexibility of pendant groups is evaluated by considering only the staggered orientations, 3^{12} combinations are added to the previous ones, resulting in more than $5x10^{10}$ possible conformations. It is clear that particular investigation strategies are needed to deal with these analyses (§2.3.2).

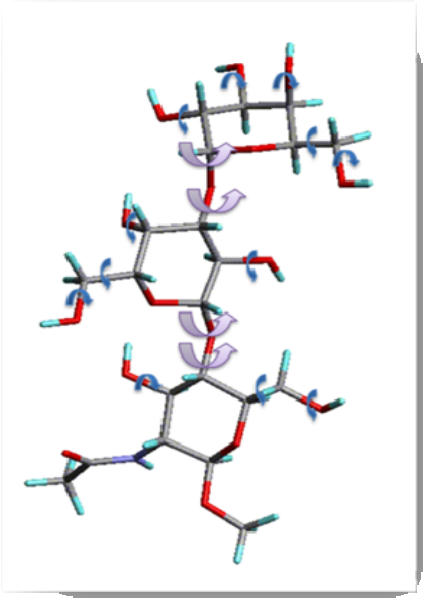


Figure 1-11 - Structure of the trisaccharide aGal(1-3)βGal(1-4)βGlcNAcOMe: the flexible torsions are indicated by colored arrows. Gal and GlcNAc are the abbreviations for the monomer Galactose and N-acetylglucosamine.

1.3 Lectins, proteins that recognize glycans

1.3.1 General features

Carbohydrates are a source of biological information written in a glycocode. In nature, a particular class of proteins, called lectins, is able to decipher this particular language and to translate it for the biological community. Like antibodies, lectins are able to recognize specific glycosidic substrates but they do not produce any immune response. Like enzymes, their structures vary in size, ion presence, subunits number and organization but they are catalytically inactive, they do not modify the carbohydrates to which they bind. Lectins possess the property of cells agglutination and polysaccharides or glycoproteins precipitation. That is because they usually exhibit architectural multivalence, consisting of structures with several, equivalent carbohydrate recognition domains (CRD) that allow cross linking between cells and consequent precipitation. However, certain lectins lack the ability to agglutinate cells because they are monovalent with respect to sugar binding[43, 44]. The occurrence in nature of proteins possessing hemagglutinating activity combined with sugar-specificity has been known since the turn of 19th century[43:] for a long time they attracted little attention, especially because it was assumed that they were confined to the plant kingdom. The attitude towards lectins began to change in the late 1960s, when it was realized that these proteins were easily available from many organisms, ranging from viruses and bacteria to plants and animals. Starting from this point, lectins were applied for the study of carbohydrates, free or in complex with proteins, in solution or anchored to cell surfaces. The deep understanding of the glycocode and lectin properties has then been successively exploited for the development of promising biomedical tools in diagnosis and therapy (Table 1). Lectin-carbohydrate binding can be quantitatively evaluated through the constant of association *Ka* that describes the equilibrium of the interaction:

$$L + C \underset{}{\overset{K_a}{\rightleftharpoons}} LC \qquad (1\text{-}1)$$

where *L* is the lectin, *C* is the carbohydrate and *LC* is the complex formed.

Table 1– Milestones in lectin research. Adapted from[43].

(CRD, Carbohydrate Recognition Domain)*

Year	Event	Scientists
1888	*Hemagglutinating activity of Ricinus communis*	*P.H.Stillmark*
1908	*Species specificity of plants hemagglutinins*	*K. Landsteiner*
1919	*Isolation of the plant lectin Concanavalin A*	*J.B. Sumner*
1949	*Human blood type specificity of plants hemagglutinins*	*W.C. Boyd; K.O. Renconen*
1952	*Use of lectins to identify cell surface sugars*	*W.M. Watkins; W.T. Morgan*
1954	*Blood antigen specific agglutinins named lectins*	*W.C. Boyd; E. Shapley*
1960	*Lectins are mitogenic*	*G. Nowell*
1963	*Agglutination of cancer cells by wheat germ agglutinin*	*J.C. Au*
1965	*Isolation of lectins through affinity chromatography*	*I.J. Goldstein*
1972	*Determination of 3D structure of Concanavalin A*	*G.M. Edelman; K.O. Hartman; C.F. Ainsworth*
1975	*Lectins in lymphocite migration*	*E.C. Butcher; I. Weissman*
1976	*Role of bacterial lectins in infectious diseases*	*I. Ofek, D. Mirelman*
1981	*Use of soybean agglutinin in transplantation*	*Y. Reisner, N. Sharon*
1988	*Identification of the CRD* in animal lectins*	*K. Drickamer*
1989	*Role of selectins in the inflammation process*	*Various*

Biochemical

Cancer

Medical

Antimicrobial

The equilibrium constant *Ka* is a function of the concentrations of the chemical species in equilibrium and can be defined as:

$$K_a = [L][C] / [LC] \qquad (1\text{-}2)$$

The equilibrium parameter *Ka* can be related to the thermodynamic parameter ΔG that corresponds to the Gibbs free energy:

$$\Delta G = -RT\ln K_a = \Delta H - T\Delta S \qquad (1\text{-}3)$$

The free energy of binding between a carbohydrate and a lectin, ΔG, is characterized by an enthalpic ΔH term and an entropic ΔS term, dependent to the temperature of the system *T*, expressed in Kelvin (*R* represents the gas constant). ΔG must be negative for interactions that occur spontaneously. When the exothermic enthalpy term is dominant, like in almost all carbohydrate-lectin interactions, the reaction is called enthalpy-driven[43]. It means that the formation of the complex is exothermic, favored by the establishment of polar/non polar contacts between ligand and protein, and, in some cases, by the formation of ion-mediated contacts (§1.3.1).

The binding of a monovalent lectin to a monovalent ligand is easily defined by the equation 1-3. However, since the affinity of single protein-glycan interaction is generally low, normally expressed in a millimolar range, many lectins bind in a multivalent fashion. Lectins are found in nature as oligomers, presenting several, equivalent carbohydrate recognition domains that increase the ability to maintain contacts with the target sugars. The term *avidity* replaces the term *affinity* whenever the strength of multivalent ligand binding is evaluated[2]. Avidity is a term commonly used to describe the combined strength of multiple bond interactions in proteins. Avidity is distinct from affinity, which is a term used to describe the strength of a single bond. As such, avidity is the combined synergistic strength

of bond affinities rather than the sum of bonds. Lectins take advantage of their multivalence state in many biological processes. For example, the lectin PA-IL is a tetrametric lectin with four carbohydrate binding sites facing different directions. This structure is ideal for exploring cell surfaces and adhering to them (§1.3.1). Langerin is a transmembrane protein with an extracellular region formed by a coiled coil of α-helices and three subunits combined together to form a trimer. Trimer formation is essential for determining selectivity towards particular oligosaccharide structures and consequently exploiting the role of endocytic receptor of Langerhans cells (§3.3). Different modes of multivalency adopted by lectins are reported in Figure 1-12.

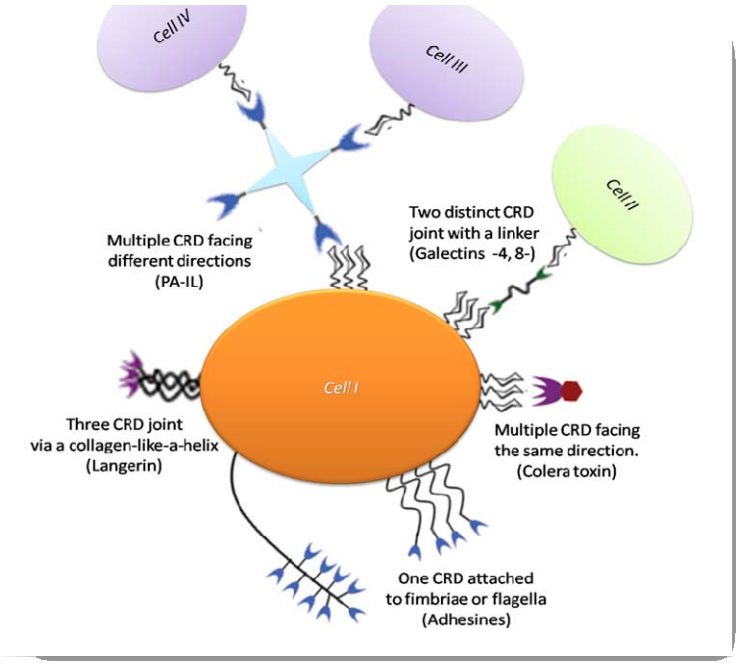


Figure 1-12 - Different modes of multivalence observed in lectins. Adapted from[45].

Lectins can be classified according to the monosaccharide for which they exhibit the highest affinity (Figure 1-13). The specificity towards particular sugars is not always well defined. Many lectins tolerate epimeric variations. For example, the plant lectin Concanavalin A can

recognize both mannose and glucose residues[46]; the bacterial lectin from *Chromobacterium violaceum* shows a specificity for both fucose and mannose monomers[47].

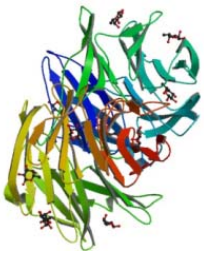
Concanavalin A lectin (5CNA)

A

Psathyrella velutina lectin (2C4D)

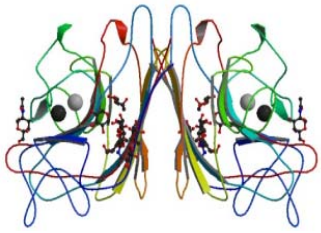
B

Aleuria aurantia lectin (1OFZ)



C

Erythina corallodendron lectin (1AX0)

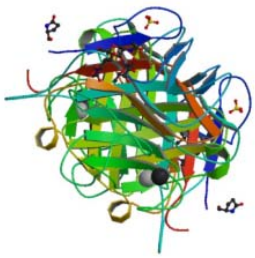


D

Psathyrella velutina lectin (2C25)

E

Griffonia simplicifolia-IVlectin(1LEC)



F

Figure 1-13 - Examples of lectins that specifically recognize mannose(A,[48]), N-*acetylglucosamine(B,[49]), fucose(C,[50]), galactose/*N-*acetylgalactosamine(D,[51]),* N-*acetylneuraminic acid(E,[49]). The lectin represented in F[52], is specific for the oligosaccharide Fucα2Galβ3(Fucα4)GlcNAc. PDB codes are reported in brackets.*

Monosaccharide specificity masks the fact that lectins often exhibit a higher affinity for di-, tri-, and tetrasaccharides that are found in nature as gycoconjugates[44]. Some lectins are only able to recognize oligosaccharides (Figure 1-13). Clearly, oligosaccharides can have more opportunities with respect to monosaccharides to create hydrophobic/polar contacts with the protein, justifying an increase in association constants.

Lectins are ubiquitous, they can be found in microorganisms, animals and plants. A classification according to their origin is reported in Table 2[53]. The role and structure of specific lectins taken from each class described in Table 2 are going to be discussed in details in the *Result* section. Two of these examples account the presence of calcium ions in their binding site that coordinate the oxygen atoms of sugar pendant groups. They can be also classified according to their binding properties as *Calcium dependent lectins*.

Table 2 – Classification and functions of lectins according to their origin.

Lectin	***Function***
MICROORGANISMS	
Bacteria	
Pili or fimbriae	Adhesion, infection
Soluble lectins	Adhesion, infection, biofilm formation
Virus	
Toxins	Adhesion, infection
Hemagglutinins	Adhesion, infection
Amoeba	
Surface lectins	Adhesion
ANIMALS	
Calnexins	Assisting protein folding of glycoproteins
M-type	E.R. protein degradation
L-type	Regulation of biosynthesis of glycoproteins
P-type (Mannose 6P-receptors)	Protein sorting post Golgi (apoptosis)
C-type (Calcium dependant lectins)	Adhesion of leucoiti to cells of bood vessels (selectins) Immunity regulation (collectins)
I-type	Adhesion (siglecs)
R-type	Target for enzymes, hormone regulation
Galectins	Glycan recognition in extracellular matrix
PLANTS	
Leguminosae lectins	Defence, symbiosis
Others	Defence

SECTION 2
COMPUTATIONAL METHODS

2 Methods: computational chemistry

Computational chemistry is a branch of chemical research that complements the information obtained from experimental data and has the aim to predict and explain chemical phenomena. With roots in the middle of the 20th century, it uses the principles of physics combined with theoretical chemistry and computer science to evaluate structures and properties of molecules through the solution of a series of equations. These are based on classical mechanical methods or quantum mechanical methods depending on the system and on the analyzed properties.

Molecular mechanics methods, also referred to as classical methods, do not take into account the movement of electrons. This approximation allows the nuclear positions to be described by classical Newtonian mechanics drastically reducing the complexity of the calculations. However, such classical approaches cannot be used to directly study chemical reactions or systems that contain complex electonic distributions. In these cases, it is necessary to use quantum mechanical or, for large systems, quantum mechanical - molecular mechanical approaches[54].

2.1 Molecular mechanics and force fields

Molecular Mechanics is based on a mathematical model of a molecule as a collection of balls corresponding to atoms with a fixed electronic distribution connected together by springs, representing the bonds.

The principle behind Molecular Mechanics is to express the total energy of that molecule, *V*, as a function of a parameterized mathematical equation (Equation 2-1). The equation constitutes a force field and can be written as:

$$V_{tot} = \Sigma V_{bond} + \Sigma V_{angle} + \Sigma V_{dihedral} + \Sigma V_{improper} + \Sigma V_{non\text{-}bonded} \qquad (2\text{-}1)$$

The total energy of the molecular system *V* is calculated as a sum of bond, angle, dihedral and non bonded (electrostatic and van der Waals) energies[55].

Specific parameters, obtained by fitting to experimental measurements and/or quantum mechanical simulations, are associated to the terms reported in equation 2.1.
The parameter set is specific to a given force field: it may be transferable, able to describe the behavior of a particular atom surrounded by different environments[54, 55]. Several force fields have been developed for carbohydrates (§2.3.1) for proteins and nucleic acids[56, 57].

2.1.1 Interatomic potentials

The different terms of Equation 2.1 are described in the next paragraphs, taking into account the AMBER force field as a model[54, 55, 58-60].

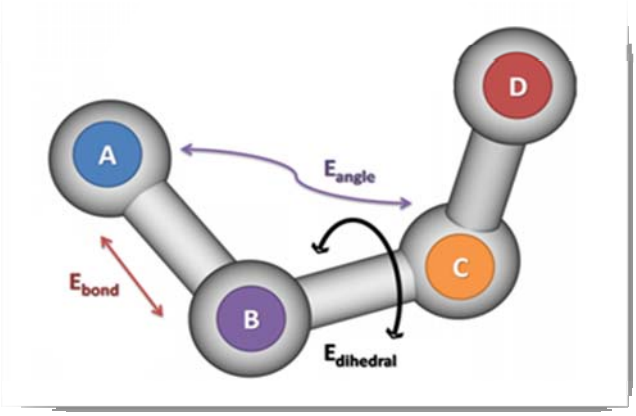


Figure 2-1– Four atoms A, B, C, D from a general biological system are represented to illustrate part of the generic terms of the Equation 2-1.

2.1.1.1 The bond stretching energy term

The first term in the Equation 2-1 describes the energy change as a bond stretches and contracts from its ideal length. The bond stretching energy term is generally treated as a simple harmonic potential (Hooke's law) where the increase in the bond energy when it is stretched is a quadratic function of the change in bond length:

$$\Sigma V_{bond} = k_{stretch}\,(r - r_{eq})^2 \qquad (2\text{-}2)$$

$k_{stretch}$ is the force constant, r and r_{eq} are the current bond length and the length at the equilibrium of a bond, respectively. The expression provides a reliable description of the energies of bond stretching only when the bond length remains close to the equilibrium value r_{eq}. This simple harmonic expression is usually sufficient to describe correctly the bonds in most chemical systems but it can be expanded and improved, including higher order terms (Figure 2-2) and increased computational cost.

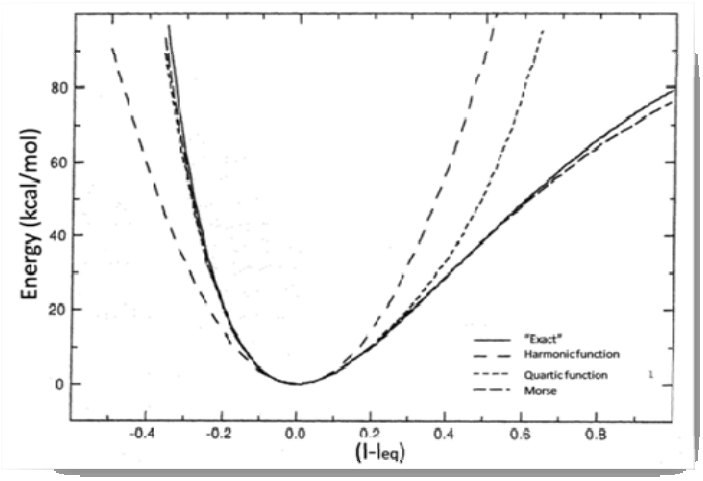


Figure 2-2 - The bond stretching energy for C-H showing the various functional forms in comparison to an "exact" quantum mechanics structure calculation[59].

For example, the Morse potential, implemented in the MM3[61] force field, takes into account bond anharmonic properties[62]. It is an exponential expression that requires the introduction of more parameters with respect to the simple harmonic function, as D, dissociation energy or well depth and a, the force constant (Equation 2-3).

$$E_{Morse}(\Delta r) = D[1 - e^{-\alpha\Delta r}]^2 \qquad (2\text{-}3)$$

2.1.1.2 The angle bending energy term

This term describes the angle bend energy between three atoms covalently bounded and in an analogous fashion to the bonded term it too is treated with a harmonic potential:

$$\Sigma V_{angle} = k_{bend}(\alpha - \alpha_{eq})^2 \qquad (2\text{-}4)$$

The angle bending term is proportional to the square of the increase in angle amplitude. K_{bend} is the force constant and α_{eq} is the equilibrium angle, where α is the specific angle. As in the bond, this may not be the ideal or perfect description for every compound, but it is usually an adequate first approximation.

2.1.1.3 The torsional energy term

As described in Figure 2-1, the torsion angle is the angle between the A-B bond and the C-D bond as viewed along the B-C bond. The energy can be evaluated by applying a cosine function that is expanded using a Fourier series:

$$\Sigma V_{dihedral} = \sum_{n=0}^{N} \tfrac{1}{2} V_n[1+\cos(n\Phi - \gamma)] \qquad (2\text{-}5)$$

V_n is the barrier height to rotation, n is the multiplicity or fold term (the number of energy minima within one full cycle), Φ is the actual torsion angle and γ is the phase shift which

determines the location of the minima. The variation of the energy of butane with the dihedral angle C-C-C-C is shown as an example in Figure 2-3: the energetic profile is obtained as a combination of multiple cosine functions of varying multiplicity.

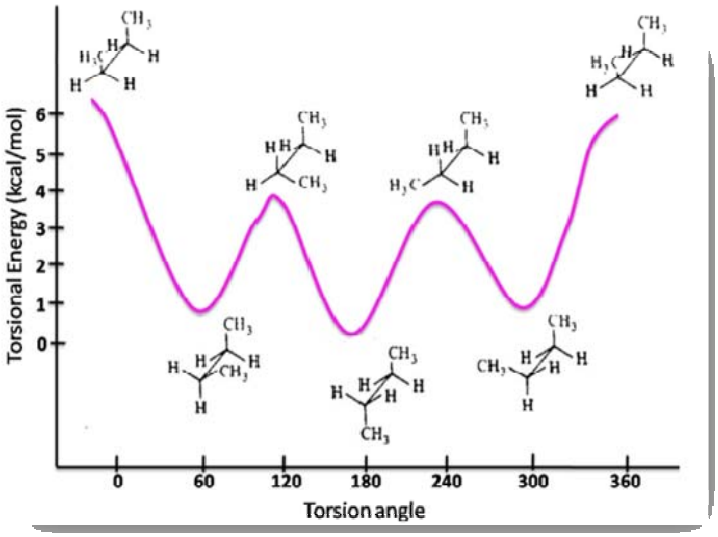


Figure 2-3 - Variation in torsional energy with C-C-C-C torsion angle for a butane fragment (P.M. Lahti, University of Massachusetts, 1995).

2.1.1.4 The improper energy term

Improper angle bending terms are generally introduced to maintain the planarity of certain atom configurations (Figure 2-4). The simplest way to assure the planarity is the application of a harmonic energy function, similar to the angle term but in this case as the angle between two planes:

$$\Sigma V_{improper} = k_\chi (\chi - \chi_{eq})^2 \qquad (2\text{-}6)$$

k_χ is the force constant whereas χ and χ_{eq} are the values of the angle at time t and at equilibrium respectively. The AMBER force field adopts a different strategy to account for improper torsions. It treats them in the same way as regular torsion angles, using a two-fold multiplicity[63].

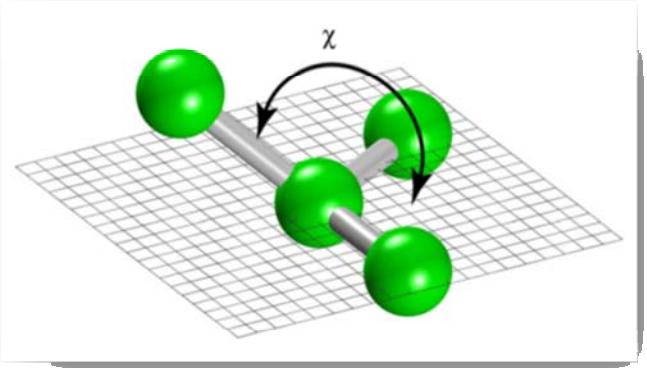


Figure 2-4 – Out of plane bending term: it preserves the correct stereo chemistry at tetrahedral centers and introduces an energy penalty associated with deviations from planarity[58].

2.1.1.5 Non bonded interactions energy terms

These terms represent the change in potential energy associated with electrostatic and van der Waals effects produced by two atoms that are not bonded (Figure 2-1).

$$\Sigma V_{non\text{-}bonded} = \Sigma V_{(elec)} + \Sigma V_{(vanDerWaals)} \qquad (2\text{-}7)$$

The electrostatic interactions are calculated by applying Coulomb's law to the system:

$$V(elec) = \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \qquad (2\text{-}8)$$

where ε_0 is the permittivity of a vacuum, ε_r is the relative permittivity of the medium, q_i and q_j are the partial charges of the two atoms and r*ij* is the distance between the atoms *i* and *j*. There is no definitive approach to the assignment of charges. They can be calculated by applying empirical methods to reproduce bulk liquid properties or they can be generated from quantum mechanical potentials to reproduce electronic distributions. In AMBER, the electrostatic potential of atoms is evaluated using the Restrained Electrostatic Potential procedure[64, 65]. This method consists of an *ab initio* calculation of the charge density for each molecule, followed by a two stage least squares fitting able to derive the charges localized at nuclear positions. In the first stage, all the charges are optimized. Methyl hydrogen atoms are then constrained to have the same charge. The second stage consists of fitting again the charges of hydrogen atoms, fixing the other charges at the values from the previous stage. The two stage protocol represents a way to mimic the correct dipole moment of a molecule. The attractive or repulsive electrostatic nature of Van der Waals interactions between atoms varies as a function of the distance (Figure 2-5). The repulsive term originates from the Pauli Exclusion Principle: at very short distances the energy varies steeply with *r* but at larger separations the decay is exponential in nature. The attractive forces, which dominate at longer distances, are due to the formation of instantaneous dipoles that can in turn induce a dipole in nearby atoms, resulting in an attractive force[59].

The Van der Waals potential $V_{(vanDerWaals)}$ can be described according to the Lennard-Jones 12-6 energy function (Equation 2-9).

$$V_{(vanDerWaals)} = \Sigma 4\varepsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}] \qquad (2\text{-}9)$$

ε is the well depth, σ is the minimum energy interaction distance, that is the distance at which the energy is zero, and r is the interatomic (j-i) distance. The first term in brackets represents the repulsion term, the second one represents the attractive term.
In the AMBER force field Van der Waals interactions are evaluated according to the Equation 2-9. However, the introduction of empirical scaling factors was necessary to smooth exaggerated stereo-electronic effects[63].

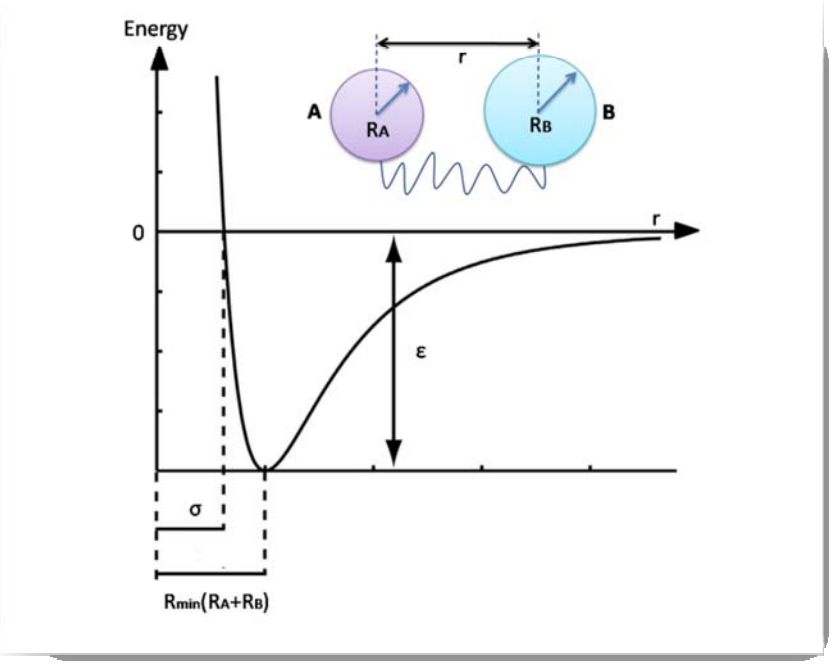


Figure 2-5 - Variation of the energy as a function of atom(A)-atom(B) distance r. σ is the minimum energy interaction distance, r_{min} is the equilibrium bond distance, ε is the depth of well.

Alternative equations can be used to achieve the contribution of Van der Waals interactions. The Buckingham-Hill potential[66], for example, replaces the repulsion term with an exponential expressions. The number of van der Waals interactions that need to be evaluated in

a biological system is high and the Lennard-Jones potential is generally the preferred form since evaluating the exponential function in the Buckingham potential is computationally expensive[59].

$$V(r) = Ae^{-Br} - \frac{C}{r^6} \qquad (2\text{-}10)$$

2.2 *In silico* strategies

Different computational strategies, used in this work for investigating properties of biological systems are going to be described. These methods rely on the energetic description of the systems as presented in the previous paragraphs.

2.2.1 Geometry optimization

Geometry optimization of molecular structures through minimization methods may be necessary in order to reach more relaxed and stable energetic states. Stable states of molecular systems correspond to global or local minima of their potential energy surface. The methods used in our work for finding minimum energy structures through the application of derivatives are successively described[54, 55, 59]. These methods gradually change the coordinates of the atoms, slowly moving the system closer and closer to the minimum.

2.2.1.1 Steepest Descent

The first derivative of potential energy and its gradient indicate the direction of the path that atoms of a system have to follow to reach the closest minimum. The energy of a system is calculated as the process starts and whenever the system orthogonally moves respect to the force, through small steps, towards the minimum. The direction is always the negative of the gradient. The steepest descent method is commonly used for poorly refined structures, whose minima are far. Close to minimum, the method does not converge well (Figure 2-6).

2.2.1.2 Conjugate gradient

In the conjugate gradient method, each minimization step is used to extract information about the gradient for computing the direction of the new vector of the next minimization step, whose path is not affected by directional restraints. The use of information from the previous step gives search lines which are "*conjugate*" to the previous search directions. The high computational cost required for this procedure is compensated by an efficient convergence towards the minimum.

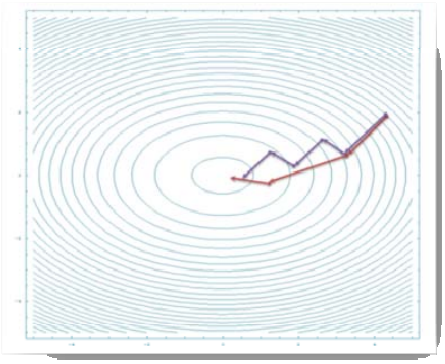


Figure 2-6- Comparison between steepest descent (violet line) and conjugate gradient (red line) minimizers when minimizing the function $x^2 + 2y^2$. The restriction to orthogonal movements limits the performance of the steepest descent minimization[59].

2.2.2 Comparative modeling

Three-dimensional structures of proteins, whose structural organization and properties are resumed in Figure 2-9, are fundamental for understanding the functional annotation of protein molecules. These structures are determined by experimental methods (X-ray, NMR) and stored in databases available on line as the Protein Data Bank (http://www.rcsb.org). However, the three dimensional investigation using experimental techniques is complex and

sometimes, comparative modeling is required to predict protein structures when no experimental structural information is available.

Comparative or homology modeling is a computational method used to generate three-dimensional structures of unknown proteins (targets) based on amino-acid sequence similarity to proteins of known structures (templates).

Success of comparative modeling is strongly dependent on the sequence similarity between the target and the template. Sequence alignments are more or less straightforward when the sequences present more than 30% of sequence identity. Between 15% and 30% of sequence identity, different homology techniques can be alternatively used to build a three dimensional model.

Threading methods, for example, are able to recognize and predict the fold of a target sequence taking that sequence and testing it on each member of a library in which known protein structures are stored. Scoring functions are then used to evaluate threading results.

Ab initio protein modeling methods try to build three-dimensional protein models only through the application of physical principles, not considering previously solved structures. This procedure tends to require a large amount of computational time and generally, only the structure of small proteins are predicted[55, 67-71].

De novo protein modeling is based on *ab initio*, comparative modeling and threading methods[72, 73].

To evaluate the different approaches for protein structure prediction, a competition called Critical Assessment of thecniques for protein Structure Prediction (CASP) was organized in 1994-1995[74]. In these CASP experiment the scientific community was invited to predict the 3D structure of novel proteins from their amino acid sequences. Until now, four CASP competitions have taken place. The CASP competitions provide a solid basis for assessing the reliability of protein models and their underlying modeling approaches[55].

2.2.3 Steps required in homology modeling

The steps involved in the process of traditional comparative modeling are described in detail in the next paragraphs and illustrated in Figure 2-7.

2.2.3.1 Template determination

Starting from the amino acid sequence of the target protein, a comparison with thousands of sequences with known 3D structures, already stored in protein databases, is necessary. The identification of the template can be achieved using different database search techniques based on sequence alignment procedures. Normally, alignments between two sequences are performed using a scoring scheme provided by matrices that dictate the rules of the alignment, according to physical, chemical features of amino acids as well as statistical properties.

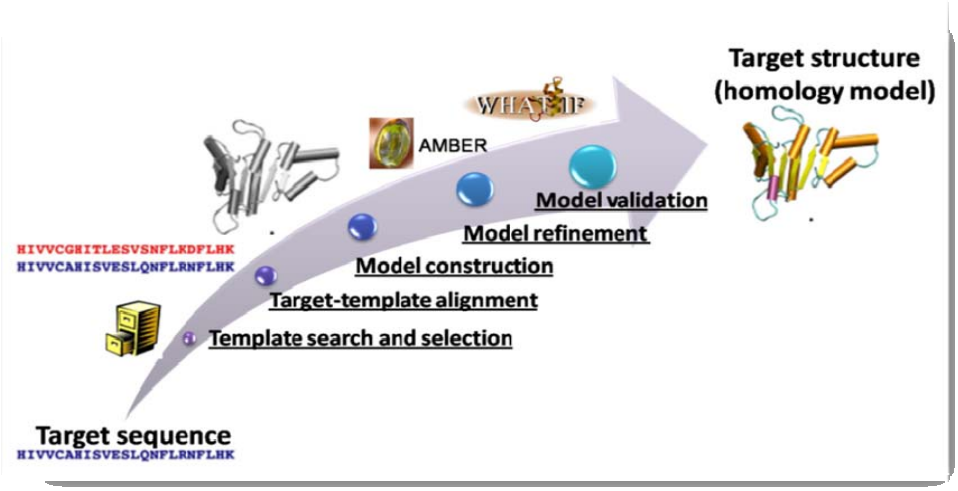


Figure 2-7 - Main steps in comparative protein structure modeling.

Differences in sequence length or variations in the location of conserved regions complicate the alignment procedure. Gaps are then introduced into the sequences to allow the simultaneous alignment of the conserved regions.

Another issue concerns the relationship between sequence identity and phylogenetic origin. It was demonstrated that sequences longer than 100 residues with more than 25% identity

are very likely related by phylogeny. Sequences presenting identities of 15%-25%, might be related, whereas they are not below[75].

In general, the global quality of an alignment is described by an alignment score that considers gap lengths by introducing penalties. The simplest alignment method is based on serial pairwise sequence alignments that compare target sequences with all the sequences of known structures, giving a set of possible homologs (FASTA[76], BLAST[77]).

The BLAST (Basic Local Alignment Search Tool) program is a tool available on line based on the Needleman and Wunsch algorithm[78]. It firstly removes the low-complexity regions composed by few type of elements from the target sequence. Then, the algorithm classifies the target sequence using a code of 3 letters and creates a list of similar triplets by using a scoring matrix. A neighborhood word score threshold T is used to reduce the number of possible triplets in the list. BLAST successively scans the database in which template sequences are stored in order to compare target-template triplets. If a match between triplets is found, this match is evaluated through a score and used to seed a possible un-gapped alignment between the target and database sequences by extending the seed in both directions. The alignment is extended until the optimal accumulated score results lower than a defined cutoff. The quality of a pairwise sequence alignment is evaluated by substitution matrices, such as PAM (Percentage of Acceptable point Mutations[79]) or BLOSUM (BLOck SUbstitution Matrices[80]), that associate high score values to pairs of identical or similar amino acids. Another method widely used to identify template sequences involves multiple sequence alignment in order to increase the sensitivity and accuracy of the search (i.e. PSI-BLAST[81]).

Once a list of possible templates is obtained, the template is selected according to the percentage of the target-template identity, also considering the template structure resolution.

2.2.3.2 Target-template alignment

The sequence-structure alignment is a crucial step in the model building process. Alignment is used to select template sequences as seen here above and to identify the most conserved regions between the target and the template. The amino acid coordinates of the template will be then used as reference for building the whole target structure. For these reasons,

once the template is identified, more sensitive and selective alignment procedures are needed to refine alignments obtained in the previous step.

Multiple alignment methods were developed to obtain more information for the construction of the protein target structure. The three dimensional model can then use structural information derived from multiple template structures. In this case, the number of gaps decreases because filled by amino acid residues derived from different template sequences (ClustalW program[82]).

An alternate alignment method consists of including structural information, making substitution matrix scores dependent on solvent exposure, secondary structure type, and hydrogen bonding features (Fugue program[83]) Alignments can then be further improved manually before the construction of the model.

2.2.3.3 Model construction

In order to proceed with the construction of the target three-dimensional model, structurally conserved regions have to be recognized in the sequence alignment. These regions contain conserved amino acids that characterize a family of homologous proteins. Their identification is easier when several template structures are available. Algorithms like Choral, implemented in the Orcherstrar program[84], are able to recognize these regions, by firstly identifying common secondary structural elements. The target and template structures are then superimposed on their Cα atoms using least-square fitting procedures. After the determination of the conserved regions, the atomic coordinates of the template backbone corresponding to these particular regions can be transferred to the target.

The next task is to search for loops to fill the variable regions. If no similar loops exist in one of the template structures, a search can be conducted in databases containing peptide fragments derived from PDB structures (i.e. FREAD) or, for small loops, in databases containing *ab initio* fragments (i.e. PETRA)[85]. A loop is selected on the basis of similar number of residues and energetic parameters. Steric problems may be detected once the loop coordinates are transferred to the target model. Thus, loop regions have to be refined by energy minimization procedures (§2.2.1).

At this point, a pre-model formed by backbone atoms is obtained and side chain may be added. Side chain conformations are taken from the template structure whenever identical

or similar amino acid residues are superposed in the alignment. If no similar amino acids are detected, side chains are added using rotamer libraries[86].

2.2.3.4 Model refinement

In order to remove steric overlaps and to obtain a more refined and relaxed three dimensional target structure, energy minimization cycles (§2.2.1) are applied sometimes combining some molecular dynamics simulation steps (§2.2.5).

2.2.3.5 Model evaluation

The sterochemical quality of the model as well as the accuracy of parameters like bond lengths, bond angles, dihedral amplitudes and the correct chirality of amino acids have to be evaluated. Programs like PROCHECK[87] or WHATIF[88], provide tools for the model analysis. They evaluate and compare the stereochemical accuracy (φ, ψ, ω, χ angles), packing quality and folding reliability of the modeled structure to parameters derived from high resolution structures. Large deviations derived from this comparison are interpreted as strong indicators of errors in the modeled structure that needs further refinements.

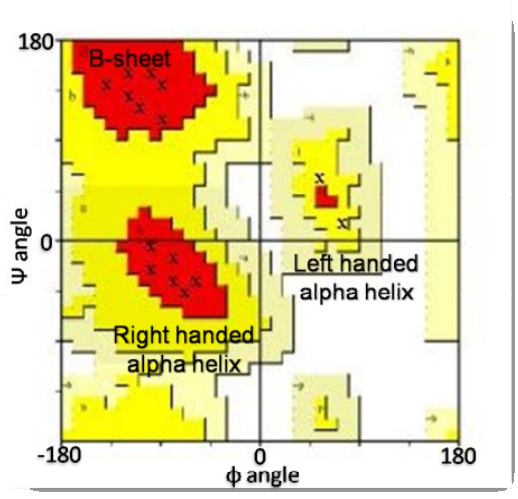


Figure 2-8 - Visualization of the Ramachandran plot (PROCHECK). The darker is the observed region, the more favored is the φ/ψ combination. Chain termini, Glycine and Proline residues are not evaluated in the plot.

One of the most important indicators of the quality of the model is the distribution of the main chain torsion angles φ and ψ that can be evaluated superposing the dihedral values on the Ramachandran plot (Figure 2-8).

To build this plot, steric favored and unfavored regions explored by φ and ψ angles of isolated dipeptides were stored: the same zones were occupied by φ/ψ torsion angles derived from NMR/X-ray structures. Thus, dihedral angles values of the model backbone must lie within the same favorable regions indicated in the map. Glycine and proline residues are not considered in the plot. Glycine residues have a hydrogen atom as side chain which is least restricted and therefore can adopt φ and ψ angles in all four quadrants of the Ramachandran plot. In contrast, proline shows only a very limited number of possible combinations of ψ and φ that arise from its 5-membered ring side chain.

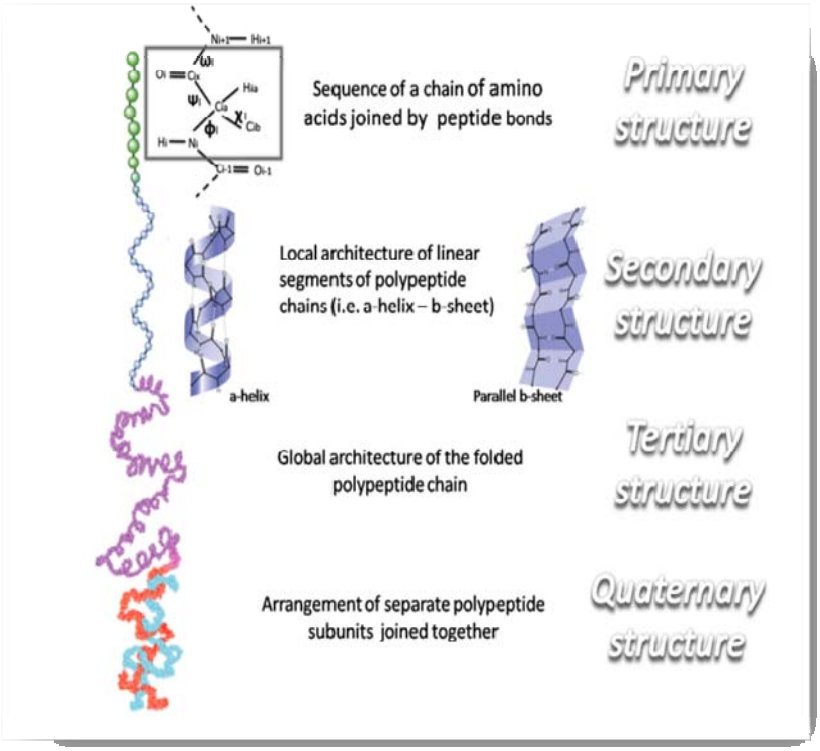


Figure 2-9 - Scheme of the structural organization of proteins.

2.2.4 Molecular docking

Molecular docking is a computational procedure that predicts the preferred orientation of a ligand to its target protein , when bound to each other to form a stable complex[89].

In order to perform computational protein–ligand docking calculations, the three dimensional structure of the target protein must be known. Difficulties in molecular docking 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 flexibility 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[55, 58, 60, 67, 71].

2.2.4.1 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 most widely used algorithms in docking programs are successively described.

2.2.4.2 Incremental construction algorithm

Incremental construction algorithms consist of the division of a ligand in 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 (Figure 2-10).

The program DOCK[90], for example, is based on this algorithm. It generates points (sphere centers) that fill the binding site and try to capture the binding site shape properties for identifying favorite 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 (§2.2.4.5). Best anchor fragments will be used for completing the construc-

tion of the ligand in the protein binding site. The best scored poses of the complete ligand will be selected.

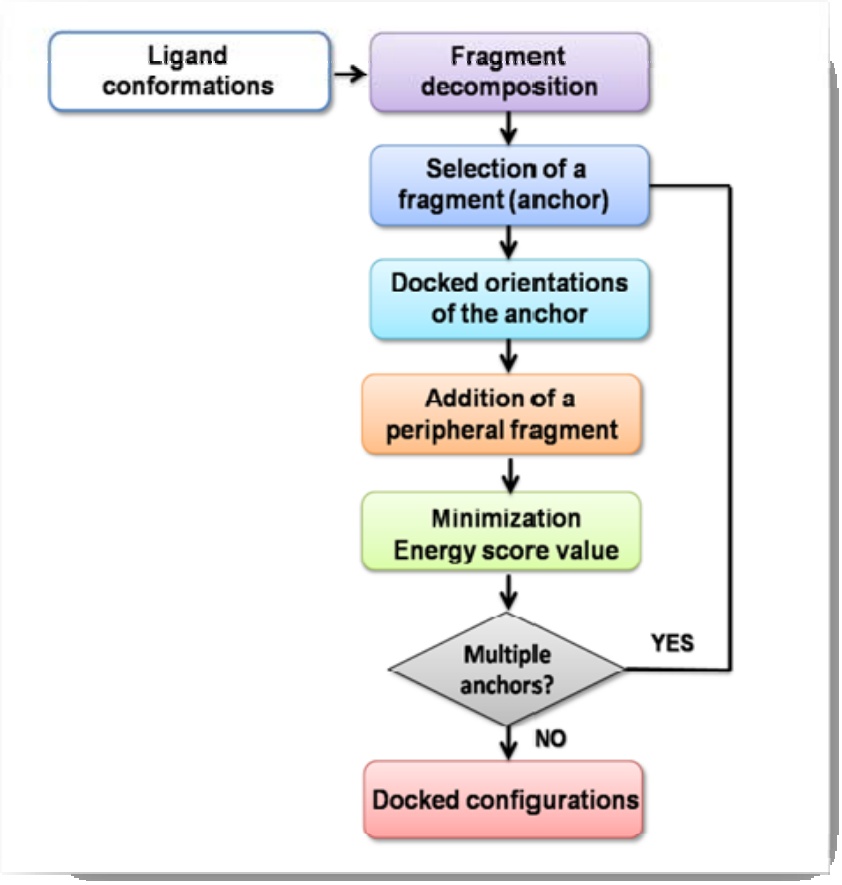


Figure 2-10 - Scheme of the incremental construction algorithm.

2.2.4.3 Genetic algorithm

Genetic algorithm is a stochastic searching approach 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.

Figure 2-11 - Main features of the genetic algorithm.

Each of these chromosomes encodes a possible conformation, characterized by genes that contain information about specific conformational variables, such as glycosidic linkages, torsion angles and pendant group conformations. To each chromosome is assigned a fitness score on the basis of the relative energetic quality of that solution in terms of protein-carbohydrate interaction. Starting from a randomly generated parent population of chromosomes, the genetic algorithm applies two major genetic operators, crossover and mutation. The crossover operator requires two parents and produces two children combining features from two different chromosomes in one. The mutation operator requires one parent and produces one child, introducing random perturbations. The emphasis of the survival of

the best individuals, evaluated in terms of energy (fitness score), ensures that the population should move toward an optimal solution, that is a correct binding mode.

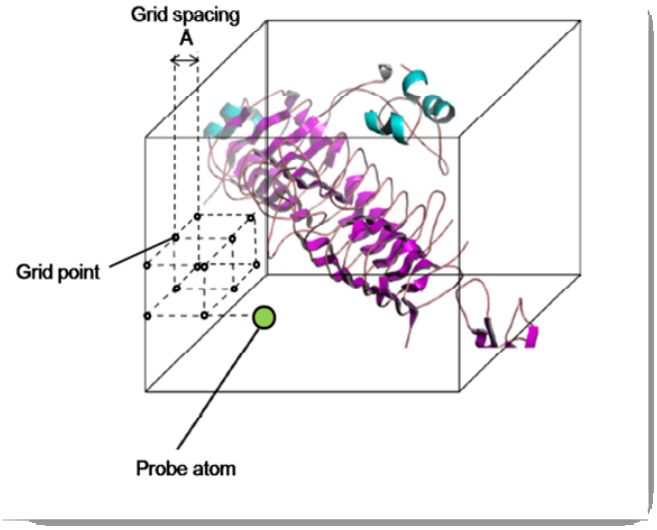


Figure 2-12 – Autodock protocol required before the genetic algorithm search. The protein is placed inside a cube shaped grid with defined spacing between the grid points. A ligand atom probe is then moved through the points. All protein-probe interaction energies are then calculated and subsequently archived in separated files.

Autodock program[91] uses this algorithm for obtaining reliable docking results. First, the protein is placed into a cube with a predefined size, characterized by a defined number of points (grid points). 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 (Figure 2-12). Afterwards, 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. It means that, once the algorithm reaches the step 3 of

the procedure (Figure 2-11), 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 (step 4) of the genetic algorithm cycle.

2.2.4.4 Miscellaneous algorithms

A variety of other sampling methods have been implemented in docking programs. Some of them include simulated annealing protocols (§2.3.2.5) and Monte Carlo simulations (§2.3.2.4).

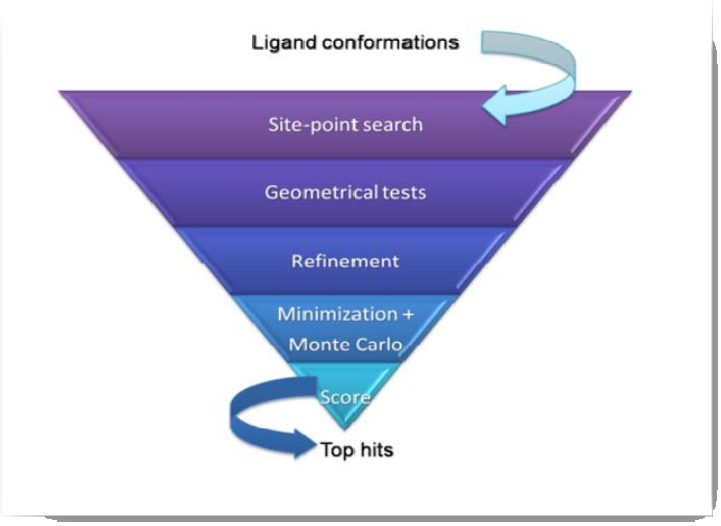


Figure 2-13- Glide docking "funnel", showing the Glide docking hierarchy.

The algorithm used in the Glide program[92], for example, 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 (Figure 2-13). 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[93]. 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.

2.2.4.5 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.

$$\Delta G = \Delta H - T\Delta S \qquad (2\text{-}11)$$

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:

$$\Delta G = -RT \ln K_a \qquad (2\text{-}12)$$

where R is the gas constant.

Some sophisticated techniques for predicting binding free energies (§2.2.6) 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. Multilinear 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 [92] is a regression-based scoring function:

$$\Delta G = C_{lipo}\sum f(r_{lr}) + C_{hbond}\sum g(\Delta r)\, h(\Delta\alpha) + C_{metal}\sum f(r_{lm}) +$$

$$+ C_{polar\text{-}phob}V_{polar\text{-}phob} + C_{coul}E_{coul} + C_{vdW}E_{vdW} + Solvation\ term \qquad (2\text{-}13)$$

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 forth 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[91], the implemented scoring function has the following form:

$$\Delta G = \Delta G_{vdw}\, \Sigma_{i,j}\, [(A_{ij}/r^{12}_{ij}) - B_{ij}/r^{6}_{ij})] + \Delta G_{hbond} +$$

$$+ \Sigma_{i,j}\, E(t)\, [(C_{ij}/r^{12}_{ij}) - D_{ij}/r^{10}_{ij})] + \Delta G_{elec}\, \Sigma_{i,j}\, q_1 q_2/\varepsilon(r_{ij})^2 \qquad (2\text{-}14)$$

$$+\Delta G_{tor} N_{tor} + \Delta G_{sol}\, \Sigma i,j\, (S_i V_j + S_j V_i)^{e(-r2ij/2\sigma 2)}$$

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 the Lennard-Jones dispersion, the directional hydrogen bonds and the Coulomb electrostatic potential taken from the AMBER force field[63]. 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[91].

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 contribute favorably to the binding affinity. In this approach, no fitting to experimental affinities is required and solvation and entropic terms are treated implicitly[94].

2.2.5 Molecular dynamics simulations

Classical molecular dynamics simulations are used in computational chemistry to simulate how a biological system evolves as a function of time[54, 58, 59, 67, 95]. These methods study structural, dynamic and thermodynamic properties of a system by solving Newton's second law of motions:

$$d^2 x_i / dt^2 = F_{xi} / m_i \qquad (2\text{-}15)$$

where F_{xi} is the force acting on the particle i of mass m_i at the position x_i and at the time t. The force is evaluated from the potential energy expression described by the force fields (§2.1). A standard method for solving the Equation 2-15 is called finite difference approach. The molecular coordinates and velocities at a time $t + \Delta t$ are obtained from the molecular coordinates and velocities at a previous time t. The iterations are repeated until sufficient time steps have been collected. The choice of a time interval Δt is important for avoiding unrealistic oscillations of the system (§2.2.5.2).

The Verlet algorithm[96] is based on this approach. It is formed by two Taylor expansions, a forward expansion, $\Delta t + 1$, and a backward expansion, $\Delta t - 1$. First, it uses the current position of a particle, x_i, to calculate the current force F_{xi}. Then, it uses the current and the previous position of the particle, x_i and x_{i-1}, to calculate the position in the next step, x_{i+1}. These two steps are repeated for every time step Δt for each atom in the molecule. This algorithm presents some disadvantages; one of them is that the velocity and temperature scaling are not explicitly included. For these reasons, modifications to the basic Verlet scheme have been proposed. The Leap-Frog integrator[97], the algorithm implemented in the package AMBER[98], was developed for explicitly including the velocity. Despite the disadvantage of having unsynchronized positions and velocities, the Leap-Frog algorithm allows the direct evaluation of velocities, useful for controlling the simulations temperature via velocity scaling.

2.2.5.1 The molecular dynamics procedure

Molecular dynamics simulations are typically carried out in four steps (Figure 2-14) under isothermal-isobaric conditions.

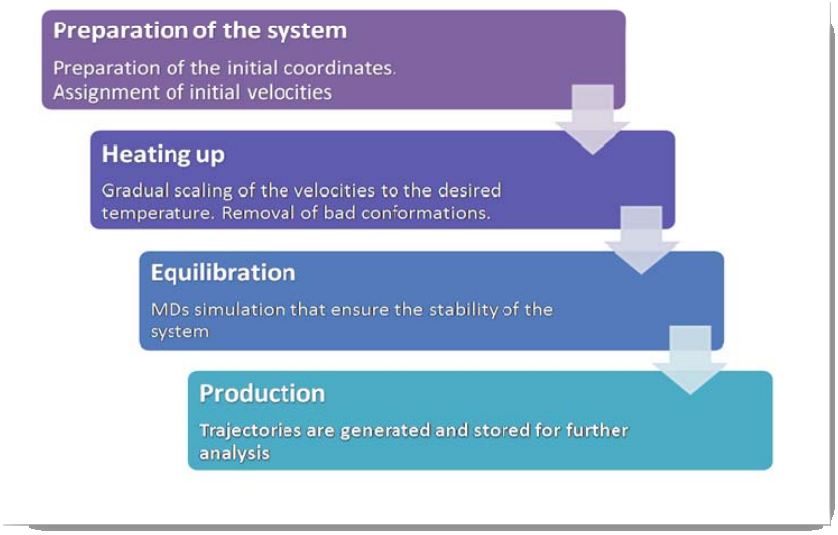


Figure 2-14 - General protocol for running MDs simulations.

In the first step, the system, derived from NMR, X-ray or homology modeling, is prepared, adding missing atoms, and submitted to minimization cycles (§2.2.1).
The second step consists of heating the system in order to remove bad contacts, increasing and assigning new atom velocities. The initial velocities are generally determined from the standard temperature-dependent Maxwell-Boltzmann distribution.
The third step is called equilibration, in which energy, temperature and RMSD of the system converge to stable values that allow the collection of trajectories (production step) that will be successively analyzed. The *RMSD* (Root Mean Square Deviation) between atoms is defined as:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{atoms}} d_i^2}{N_{atoms}}} \qquad (2\text{-}16)$$

N_{atoms} is the number of atoms taken into account for the RMSD calculation and d_i is the distance between the coordinates of the atom i when two structures of the same system are superposed. In order to set up a molecular dynamics (MDs) simulation, several choices should be made before starting the calculations.

2.2.5.2 Integration time step

The time step Δt should be small in comparison to the period of the fastest motion of the simulated system. For proteins, the fastest motions are the stretching vibrations of the bonds connecting hydrogen atoms to heavy atoms (τ = 10fs). A general rule for choosing a good time step is

$$\tau / \Delta t \sim 20 \qquad (2\text{-}17)$$

Thus, for proteins, a correct time step would be 0.5 fs, computationally expensive for long MDs simulations. The SHAKE algorithm was developed for keeping constant bond lengths notably with atoms (X-H), allowing larger time steps[99].

2.2.5.3 Hydration of the systems

In molecular dynamics simulations, the solvent can be treated in different ways. The effect of the solvent can be achieved by the use of a distance-dependent dielectric constant (*in vacuo* calculations). Alternative approaches have been developed where the solvent is considered as a continuous medium surrounding the solute. They can model the solvation effects using formulas to compute the electrostatic fields and surface area terms to evaluate hydrophobic interactions. The most widely used continuum model, PB-SA, is based on algorithms able to solve the Poisson-Boltzmann equation to take account of electrostatic interactions, including the surface area terms[100]. Another continuum model, simpler and with similar accuracy, uses the generalized Born approximation to mimic solvation effects [101]. However, the main purpose of the simulations is to reproduce the behavior of biological systems that exist, in nature, in a hydrated environment. Thus, a more reliable way to model an aqueous environment is to create a solvent box around the molecule, with discrete solvent molecules explicitly modeled in the simulations.

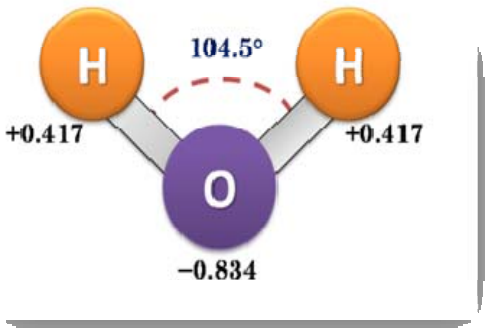


Figure 2-15 - TIP3P water model.

Several water models have been developed. They have been classified by the number of points used to define the model, whether the structure is rigid or flexible, and whether the model includes or not polarization effects.
The simplest and the most widely used model is the TIP3P[102]. It is characterized by three interaction sites, corresponding to the three atoms of the water molecule. A point charge is assigned to each atom that gets the Lennard-Jones parameters (Figure 2-15)

2.2.5.4 Periodic Boundary Conditions, PBC

In a realistic biological model, water molecules are expected to be located around the solute. If the biological model is placed in a box of water molecules, part of the solute/solvent atoms, during a simulation, will be found at the edge of the box, in contact with the surrounding vacuum. To prevent this artificial image of the liquid bulk and to assure a complete immersion of the solute during the simulations, periodic boundary conditions are employed.

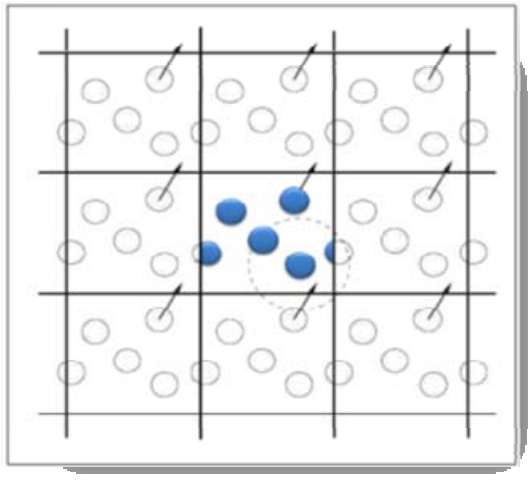


Figure 2-16 - Two-dimensional representation of the Periodic Boundaries Conditions. The cut-off for treating the non bonded interaction for a particle i *is represented with a dashed line.*

In this approach, the system, enclosed in a box of water molecules, is surrounded with replicas of itself in all the directions, to yield a periodic lattice of identical cells. When a particle moves in the central cell, its periodic image will move in the same manner in the other cells. When a particle is found at the edge, it will leave the central cell, entering from the opposite side of the same cell (Figure 2-16).
In order to reduce the computational cost of this method, an appropriate cut-off for treating the non bonded interactions may be chosen (§2.2.5.5).

2.2.5.5 Treatment of the non bonded interactions

The non-bonded interactions, composed by electrostatic and van der Waals interactions (§2.1.1.5), are considered an issue for running molecular dynamics simulations in terms of computational cost. Thus, approximations are needed.

Non bonded cut-off

As the strength of the van der Waals interactions between atoms decreases with the interatomic distance, one can truncate the Lennard Jones interactions at a specific cut-off distance. An optimal cut-off distance is determined by evaluating its effect on the variation of the total system energy. When the total energy converges quickly at a particular cut-off, the gain in going to large cut-offs is small and no longer justifies the extra computational cost. If periodic boundary conditions are applied (§2.2.5.4), the cut-off should be less or equal to half the size of the solvent box. This prevents the possibility for a particle to interact with any of its images of the same particle in the same time.

Particle Mesh Ewald summation (PME)

In periodic boundaries conditions, all charged particles of a system interact with each other in the central box and in all image boxes following the Coulomb's law modified by the appropriate translation vectors (Equation 2-8). For each charged particle, particles with equal magnitude but opposite sign are uniformly distributed around it, following a Gaussian distribution. The Particle Mesh Ewald summation is the technique commonly used to rapidly calculate the effect of long range electrostatic interactions in systems under periodic boundary conditions[103]. This method is based on the evaluation of two separated charge

distributions, summed together via fast Fourier transforms. The sum of the electrostatic interactions between charged particles and the respective neutralizing charged particles is additionally summed to a second term that contains the periodic sum of the Gaussians representing the neutralizing charge distribution[103].

2.2.6 Free energy calculations

The characterization of the thermodynamic properties of ligand – protein interactions through the evaluation of Gibbs free energy changes ΔG (Equation 2-13) is one of the most interesting applications of molecular modeling. These properties are taken into account when evaluating molecular docking results (§2.2.4.5). Free energy of binding of ligand-protein complexes can be also evaluated from molecular dynamics trajectories. Two types of theoretical approaches are commonly used to calculate the absolute and relative free energy of binding of ligand–protein complexes: Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) and thermodynamic integration (TI) methods, respectively[104-106]
MM-PBSA approach is usually used and preferred respect to the TI method when dealing with diverse sets of ligands that differ significantly in their structural and chemical composition[107].

2.2.6.1 Absolute free energy of binding: MM-PBSA method

The absolute ligand-receptor interaction energies can be obtained by performing average MM-PBSA calculations on an ensemble of uncorrelated snapshots in an implicit water environment, collected from an equilibrated molecular dynamics simulation (Figure 2-17). 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:

$$\Delta G = \Delta G_{RL} - \Delta G_{R} - \Delta G_{L} \qquad (2\text{-}18)$$

Each term of the equation 2-20 is further decomposed as follow:

$$\Delta G_{RL} = \Delta E_{MM} + \Delta G_{PBSA} - T\Delta S_{MM} \qquad (2\text{-}19A)$$

$$\Delta G_{R} = \Delta E_{MM} + \Delta G_{PBSA} - T\Delta S_{MM} \qquad (2\text{-}21B)$$

$$\Delta G_{L} = \Delta E_{MM} + \Delta G_{PBSA} - T\Delta S_{MM} \qquad (2\text{-}21C)$$

where ΔE_{MM} is the average molecular mechanical energy containing the bonds 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.

$$\Delta G_{PBSA} = \Delta G_{PB}^{el} + \Delta G_{SA}^{np} \qquad (2\text{-}20)$$

The Poisson Boltzman equation is solved for determining the solvent polar effects ΔG_{PB}^{el}[108] whereas the solvent accessible surface area is used to determine the nonpolar energetic term ΔG_{SA}^{np}[109].

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[110]. The high computational cost combined with a very slow convergence and the approximations introduce significant uncertainty in the results[111, 112]. 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[113].

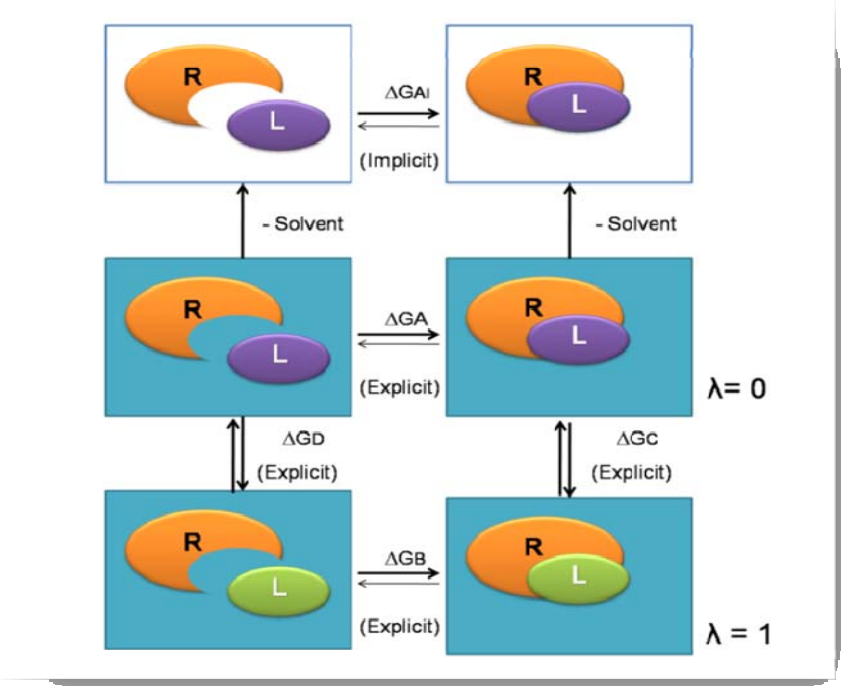


Figure 2-17 - MM-PBSA calculations determine the absolute free energy of binding of a ligand to a receptor (ΔG_{AI}) in an implicit solvent environment, whereas TI methods calculate the free-energy of binding difference between receptor–ligand complexes ($\Delta\Delta G = \Delta GC - \Delta GD = \Delta GA - \Delta GB$), where only the ligand is changed.

2.2.6.2 Relative free energy of binding: a brief introduction to thermodynamic integration TI

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 (Figure 2-17).

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:

$$\Delta G_{TI} = {}^{1}\int_{0} \langle \delta V(\lambda) / \delta(\lambda) \rangle_{\lambda} \, d\lambda \qquad (2\text{-}21)$$

In this equation, λ represents the coupling parameter that corresponds to the potential energy *V(A)* for $\lambda = 0$ and *V(B)* for $\lambda = 1$. The integration is carried out over the average of the λ derivative of the coupled potential function at given λ values. Thus, molecular dynamics 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 molecular dynamics simulations will most likely not sample enough phase space for obtaining converged results[106].

2.3 Glycomodelling

The complexity of the tridimensional determination of carbohydrate structure derives from the high mobility of these molecules and from the presence of different conformations which coexist in equilibrium[106].

Computational methods have been developed, providing complementary tools to improve and complete X-ray and NMR information in carbohydrate structural studies[114].

The variability of anomeric status, linkage position, ring size, shape and topology represent an issue for studies carbohydrate conformations *in silico*[115]. For these reasons, specific force fields have been successfully developed (§2.3.1).

2.3.1 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 necessary. Appropriate force fields for carbohydrate systems have been created, whose aim was to reproduce the particular effects that influence their global structural properties in solution[116].

The exocyclic hydroxymethyl group behavior defined by the ω-angle (O5-C5-C6-O6) and its preference for *gauche* states (§1.2.2) can be reproduced by introducing scaling factors that slightly modify the 1-4 non bonded interactions (§2.1.1.5)[34].

1-4 non bonded interactions define the influence, in terms of electrostatic and Van der Waals potentials, of two atoms located at each extremity of a torsion angle. 1-4 non bonded interactions are not treated in the same manner in all force fields and this could be a problem in simulations of 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 choice of 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[38, 117]. The widely carbohydrate force field used, GLYCAM06, is herein described in details.

2.3.1.1 Glycam06

The Glycam06 force field is highly consistent for modeling carbohydrates, glycoproteins and glycolipids[118, 119]. It can be used for describing the physico – chemical properties of complex carbohydrate derivatives and it is fully compatible with the AMBER force field[63]. Glycam06 may be used in simulation package other then AMBER through the employment of appropriate file conversion tools[120]. Parameters are 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.

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[63]. No scaling factors for treating 1–4 interactions are introduced for reproducing the *gauche* effect on ω angle rotamers[34].

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[121].

Comparison with experimental data confirmed that the force field is able to well reproduce rotational energies and carbohydrate features (§1.2.1) if combined with an appropriate charge set, except for highly polar molecules for which empirical terms have been introduced to correct energetic torsion errors[118].

In Glycam06, atomic partial charges are calculated residue by residue. For each residue, 50-100ns molecular dynamics 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[122-123]. An optimal RESP charge restraint weigh of 0.01 is applied, based on simulations of carbohydrate crystal lattices[124].

A simple computational tool has been developed to facilitate the preparation of the files necessary for running molecular dynamics simulations using Glycam interfaced to the AMBER package (*www.glycam.com*).

2.3.1.2 Other carbohydrate force fields

GROMOS-45A4, CHARMM and OPLS-AA are alternative carbohydrate force fields used, together with Glycam06, to describe conformational carbohydrate properties in computational chemistry[117] (Figure 2-18, A-B).

The GROMOS force field was early developed for molecular dynamics simulations of proteins, nucleotides, or sugars in aqueous or apolar solutions or in crystalline form but recently it has been modified to include the anomeric effects for mono and oligo pyranoses[125]. As in Glycam06, quantum mechanics methods are used for calculating bond, angle force constants whereas dihedral parameters derivation and Van der Waals terms are directly taken from previous GROMOS versions[126, 127]. An electrostatic potentials 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[125]. No distinction is done between α and ß monomers in terms of charges and anomeric atom type and electrostatic - van der Waals 1-4 scaling factors are not introduced for correctly reproduce the *gauche* effects on ω angles. 20 ns long molecular dynamics simulations in explicit waters (SCP water models[128]) were used for validating the force field, showing the capability to correctly predict the stereo-electronic effects and the stablest ring conformations but sometimes failing in reproducing their correct energies[125]. Recently, GROMOS was proposed as the more adapted force field for mimic the transit from $^4C_1$ to skew boat conformations of the iduronic acid residues in heparin molecular dynamics simulations[129].

CHARMM force field was recently developed for glucopyranose and its diastereomers[130]. Several revisions for carbohydrates have been proposed in order to extend this force field to five member sugar rings and oligosaccharides[131-133]. 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[134-136]. Preliminary parameter sets are created using small-molecules models corresponding to fragments of pyranose rings and then successively applied to complete pyranose monosaccharides. 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 type for α and ß anomers, was validated as it reproduces calculated quantum mechanical and experimental properties using molecular dynamics simulations in TIP3P water molecules.

The OPLS force field has been expanded recently to include carbohydrates[137]. In OPLS-AA-SEI (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[137]. Unique charge sets and atom types for α and ß anomers are used[137]. All non bonded parameters are imported directly from the parent force field OPLS-AA[93]. Charges are derived, as for previous force fields versions[93, 138], from standard alcohols

and acetals to simple 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[61], but now applicable to a wide range of compounds[139, 140], is widely used for the construction of adiabatic maps of disaccharides (§2.3.2.1). TRIPOS molecular mechanics force field is designed to simulate both peptides and small organic molecules[141] but parameter extension for oligosaccharides includes sulfated glycosaminoglycan fragments and glycopeptides[142, 143]. The TRIPOS force field is implemented in the molecular package Sybyl (Tripos Associates, St. Louis, MO) and commonly used for geometry optimizations (§2.2.1).

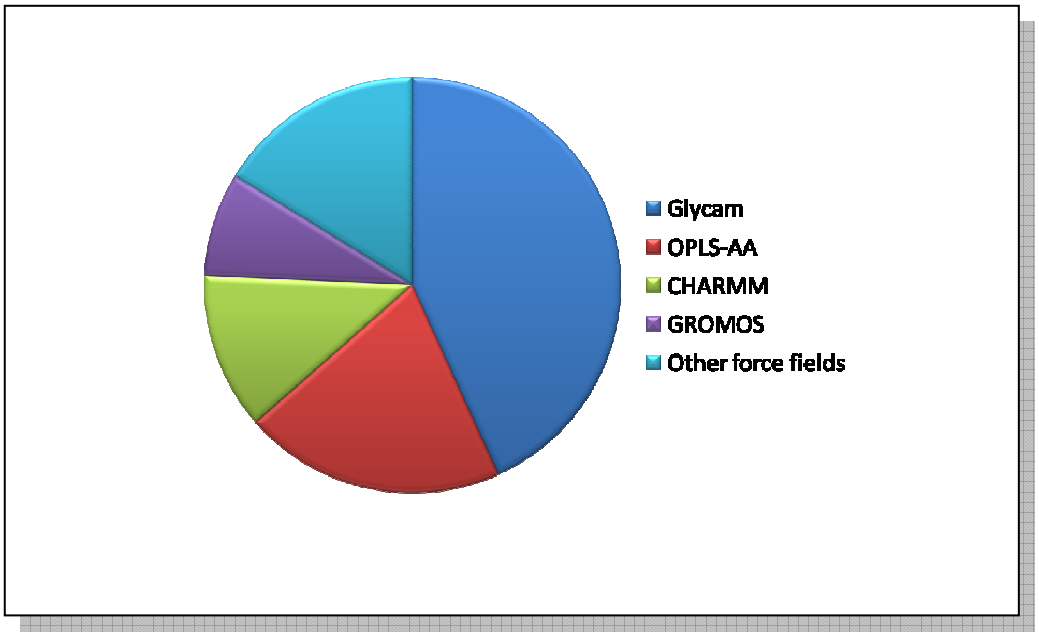


Figure 2-18A – Approximate estimate of the usage of carbohydrate force fields (GLYCAM, CHARMM, OPLS-AA, GROMOS). Each slice is proportional to the number of citations of the force field in the last 5 years, according to the ISI — Web of Science (http://scientific.thomsonreuters.com/products/wos/) [117].

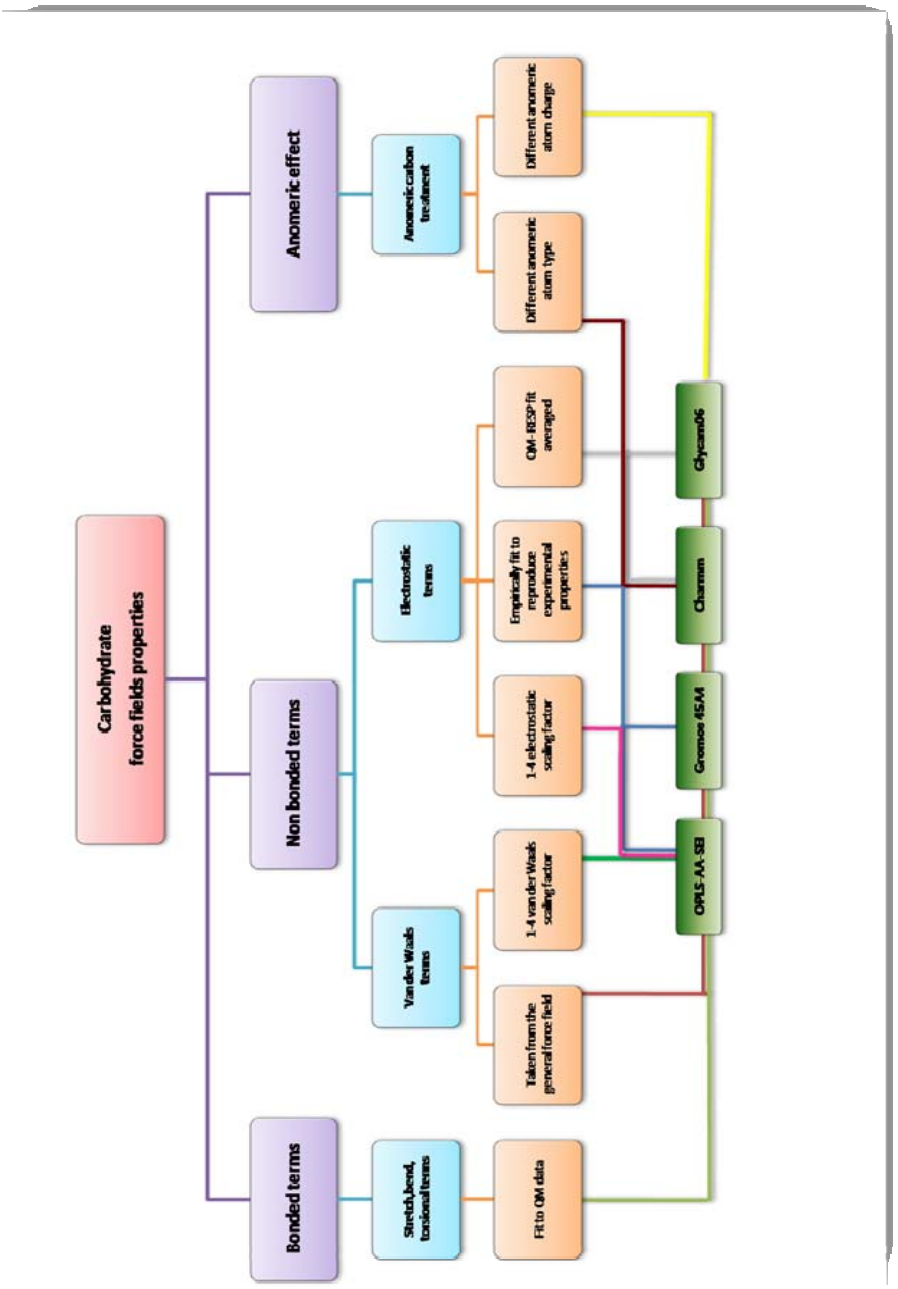


Figure 2-18B – Parameterization protocol comparison between the carbohydrate force-fields Glycam06, Gromos 45A4, Charmm, OPLS-AA-SEI.

2.3.2 *In silico* conformational studies on carbohydrate structures

The search of bioactive conformations is an important part of understanding the relationship between structure and biological activity of a molecule. A complete overview about the conformational potential of molecules can be gained by theoretical techniques. Conformational properties of carbohydrates and the respective potential energy surfaces can be evaluated through *in silico* conformational searches[55, 115].

2.3.2.1 Systematic search procedures

The systematic search is one of the most used conformational analysis methods in computational chemistry. It is performed by varying systematically each of the torsion angles of a molecule to generate all possible conformations and to evaluate the potential energy associated to each conformation.

φ and ψ are the torsion angles that define the glycosidic linkage of carbohydrates(§1.2.3). They influence the global conformation of oligosaccharides, free or in complex with receptors. Particular values for glycosidic torsion angles are associated to specific and energetically stable carbohydrate conformations. Glycosidic angles can be rotated through small increments and the corresponding potential energy can be calculated according to the applied force field. All information about energies as a function of φ/ψ changes derived from the conformational search can be plotted obtaining contour maps in which the position of minima and the heights of the transitional barriers can be visualized.

Rigid contour maps are calculated only taking into account the flexibility of φ and ψ angles. During φ /ψ rotation, the pendant groups are unable to relax, causing steric clashes and, consequently, high energetic conformations. To correctly describe the behavior of carbohydrates, more freedom may be attributed to the whole glycosidic structures, of course, not ignoring the stereo-electronic effects (§1.2).

Relaxed contour maps are built applying the method previously described for the rigid maps. A full minimization after each step, excluding the two torsion angles at the glycosidic linkage, allows the adjustments of bond lengths and angles, decreasing the conformational energies[144]. However, the minimization starts from a particular sugar conformation and the pendant groups, due to steric clashes, may not reach stable minima. At each step of the

conformational search, all possible combinations of pendant group orientations should be taken into account.

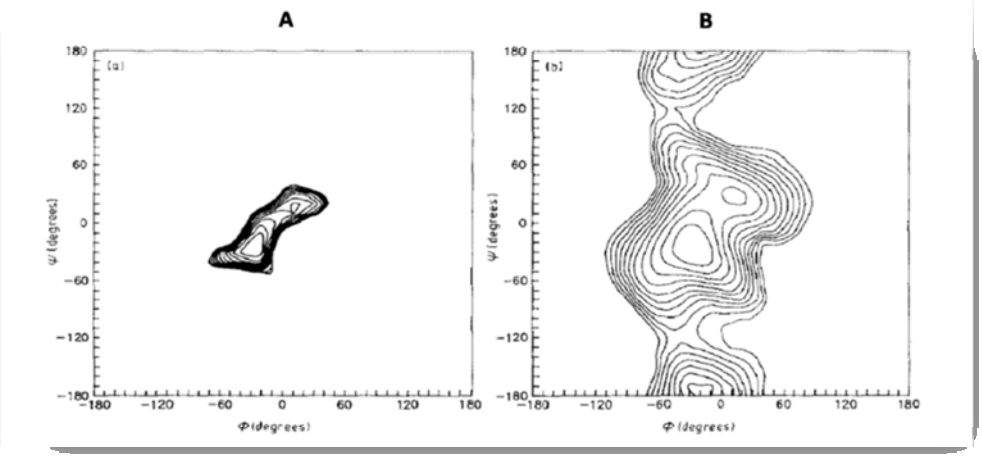


Figure 2-19 –Rigid (A) and relaxed (B) conformational maps for maltose with ω in gg orientations. Both maps use a residue optimized with MM2 for the starting geometry[145].

Thus, a population of sugar conformers can be initially created, considering all possible orientations of the pendent groups. For each conformer, relaxed maps can be calculated and consequently merged together in a unique one (adiabatic map) in which new minima are identified and barriers between minima reduced.

Adiabatic maps are widely used in computational chemistry for understanding the energetic properties of carbohydrates[146-148]. A database containing adiabatic maps of the most common disaccharides is available on line (*http://www.cermav.cnrs.fr/cgi-bin/di/di.cgi*).

2.3.2.2 Heuristic methods

In order to gain computational performance and simplicity, heuristic methods are used to search conformational space of carbohydrates. Comparison of different types of heuristic searches indicates CICADA one of the most reliable, less-time consuming, able to produce results comparable to the systematic searches described in the previous paragraph[149].

Through CICADA calculations, the potential energy surface is explored using a single-coordinate driving approach[150]: a selected torsion angle is driven with a concomitant full-geometry optimization at each increment of the conformational search, with the exception for the driven angle. The orientation of the pendent groups is also monitored for detecting energy minima. Only when CICADA detects a minimum, the conformation is fully optimized. The resulting structure is compared with the ones previously stored and consequently stored if not yet detected. The new minima are required as new starting points for further explorations.

The main advantage of the method herein described is the low computational cost: it presents a polynomial dependence on the degrees of freedom of the studied compound, in contrast to an exponential dependence found for a simple conformational search. This method also allows the exploration of all possible conformations that characterize specific families of conformers, spending the most part of the time in the most populated areas. It has been proved that CICADA is an efficient tool to explore carbohydrate conformational space, for small and complex oligosaccharides, correctly predicting the possible low-energetic conformer families[151-153].

2.3.2.3 Genetic algorithm search

The genetic algorithm (Figure 2-11) is generally used to study protein-carbohydrate/ligand complex conformations. The method, as described in the previous paragraph (§2.2.4.3), has been recently used in combination with local minimization strategies for searching the conformational space of molecules in flexible docking programs[154]. Some programs, based on this algorithm, are also completely dedicated for the description of carbohydrate conformations[155-157].

2.3.2.4 Monte Carlo methods

The name Monte Carlo, which derives from the famous Monaco casino, emphasizes the importance of randomness, or chance, in this method[158].

A particular starting conformation of a carbohydrate is submitted to constant temperature cycles of Monte Carlo, during which random changes are made in terms of orientation of pendant groups and global conformation. The new state is accepted if its energy is lower

than the energy of the preceding state. Otherwise, the configuration is accepted or rejected based on a probability expression (the Boltzmann equation). The higher the temperature of the cycle, the higher is the probability that the new state will be accepted. The process is repeated to create a large set of representative configurations of the system, covering all regions of the conformational space if the process is allowed to run for a sufficiently long time. Metropolis Monte Carlo algorithm[159] has been widely used for the conformational analysis of oligosaccharides[160-162].

2.3.2.5 MDs simulations and simulated annealing

The conformational search methodologies described previously present limitations in the applicability to flexible molecules. Flexible molecules are characterized by a high number of rotatable bonds, leading to serious problems in data handling due to the large number of generated conformers. A very common strategy to overcome this problem is the use of molecular dynamics simulations for exploring the conformational space, reproducing the time-dependent motional behavior of a molecule (§2.2.5)[163].

Molecular dynamics simulations are able to overcome energy barriers between different conformations. However, if the energy barrier is high or the number of degree of freedom is very large, then some potentially existing conformers are not reached. To enhance conformational sampling, the time of the simulation can be increased, and high temperatures can be applied to the system (Figure 2-20). At high temperatures, the molecule is able to overcome large barriers that may exist between conformations, providing enough kinetic energy to cross these barriers[55].

During simulations, the molecule can occupy distorted geometries that cannot be relaxed by a simple minimization procedure. In this case, annealed molecular dynamics simulations can be performed: in this technique, the system is cooled down at regular time intervals by decreasing the temperature of the simulation.

As the temperature reaches the lowest temperature, the molecule is trapped in the nearest minimum energy conformation. The geometry at the end of the annealing cycle is saved and used as starting point for the next simulation at high temperature.

The cycle is repeated to obtain a set of low energy conformations[164]. Using this technique, structural properties can be achieved together with thermodynamic quantities.

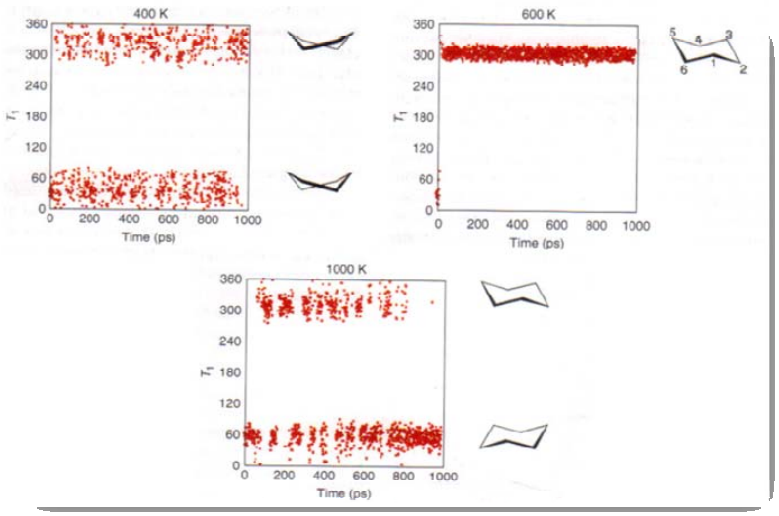


Figure 2-20 - The dependence of conformational flexibility (T1= C1-C2-C3-C4) on the simulation temperature tested on the cyclohexane, molecule whose shape mimics the pyranose ring behavior. At 400K the molecule oscillates between twist forms. At 600K, one possible minimum is reached corresponding to a chair conformation, At 1000K, both chair and twist conformations are observed but after 800ps only one of the chair conformations exists[55].

However, during simulations, the molecule can occupy distorted geometries that cannot be relaxed by a simple minimization procedure. In this case, annealed molecular dynamics simulations can be performed: in this technique, the system is cooled down at regular time intervals by decreasing the temperature of the simulation. As the temperature reaches the lowest temperature, the molecule is trapped in the nearest minimum conformation. The geometry at the end of the annealing cycle is saved and used as starting point for the next simulation at high temperature. The cycle is repeated several times to obtain a set of low energy conformations[164]. Using this technique, structural properties can be achieved together with thermodynamic quantities. Several carbohydrate conformational studies are reported in the literature, using classical molecular dynamics simulations[40, 165-167] and simulating annealing protocols[168-170].

2.3.2.6 Role of the solvent in carbohydrate conformational studies

Carbohydrates, like other systems, are strongly influenced to the solvent environment (§1.2.1). Thus, when conformational studies are performed, the solvent effect must be considered[171, 172]. The three-site water potentials models, such as TIP3P (§2.2.5), have been commonly used to describe the behavior of carbohydrate–water interactions[102]. When water molecules are explicitly modeled, water–solute interactions can be characterized through the evaluation of the radial pair distribution function, first hydration shell and averaged residence time.

The radial distribution function is a method that statistically describes the variation of water density as a function of the distance from one particular atom. The calculation of the 2D radial pair distribution function allows the identification of bridging water molecules, evaluating the water density profile between two particular atoms[173](Figure 2-21A).

The first hydration shell gives the number of water molecules at less than 3.5Å from solute oxygen atoms (Figure 2-21B).

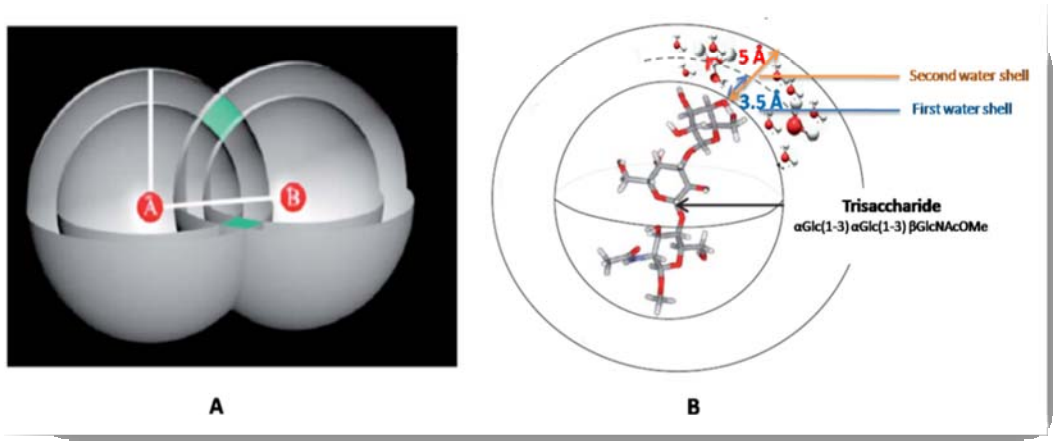


Figure 2-21 – Illustration of the intersection volume formed by two sphere shells surrounding two solute oxygen atoms, A and B, in which an oxygen atom of water molecules can be found (A)[115]. Representation of the first and second water shells for the atom O4 of the non reducing end of a trisaccharide (B).

The average residence time expresses how long water molecules occupy a specific region of space. The average residence time of water molecules in contact with hydroxyl oxygens of a carbohydrate is around 0.6–0.7 ps with some exceptions for molecules as sucrose and trehalose, where residence times of 2.7ps and 3.0 ps are calculated for some hydroxyl groups, respectively[174].

2.3.3 Glycomodelling successes: some examples

Molecular modeling studies of glycans and their receptors have been successfully used in biotechnology and medicine.

A molecular dynamics approach combined with docking calculations were necessary, for example, to explain resistance mechanism of Tamiflu® and Relenza®, anti influenza drugs. These molecules are inhibitors of the neuraminidase. This enzyme cleaves the sialic acid from the surface of glycoproteins to ensure the virus propagation[175-178].

Recently, the mechanism of binding and the impact of calcium on the H1N1 swine flu neuraminidase (Figure 2-22) was investigated through docking, dynamics and binding free energy calculations[179, 180].

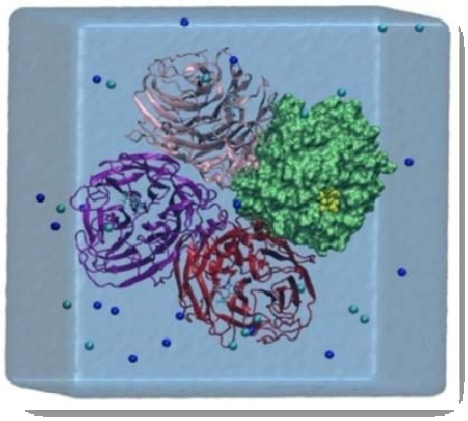


Figure 2-22 - H1N1 swine flu neuraminidase tetramer in complex with oseltamivir and ions simulated for understanding the role of calcium in the enzymatic activity and thermostability (Reproduced with permission from Dr R.C.Walker).

Molecular dynamics studies were essential for understanding the structural features of heparin, an anticoagulant and antithrombotic polysaccharide able to potentiate the inhibitory effect of antithrombin over plasma serine proteases[129, 181, 182].

Pseudomonas aeruginosa is an opportunistic gram-negative pathogen responsible for numerous nosocomial infections in immunocompromised patients (§3.4.1). The resistance to antibiotics led *Dr. Koca and al.* to search for new compounds able to inhibit part of the adhesive properties of the bacterium to pulmonary endothelial cells.

Dynamics studies were carried out on a specific protein, the lectin PA-IIL in complex with glycans that would competitively block the mechanism of adhesion[183, 184]

Dr. Mulholland et al. have applied a combination of quantum mechanics and molecular dynamics techniques to describe the reaction between hen egg white lysozyme, that hydrolyses a component of the polysaccharide cell wall in Gram-positive bacteria, and its natural oligosaccharide substrate, suggesting that the classic mechanism of action of this enzyme via an oxocarbenium ion intermediate is not correct[185].

Computational studies on glycans present numerous opportunities, not only in a medical context.

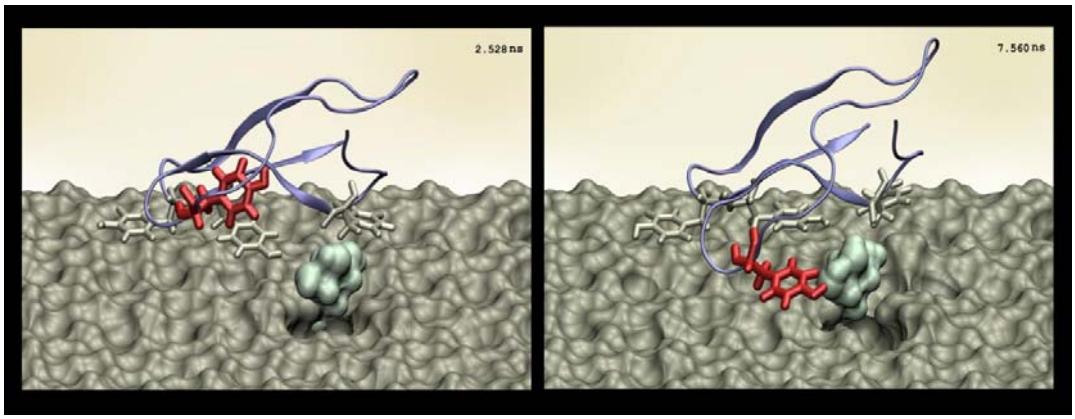


Figure 2-23 - Image showing the interaction of the binding module of cellobiohydrolase I after 2.5 and 7.5ns of molecular dynamics simulation. A tyrosine residue (in red) unfolds and interacts directly with the cellulose surface indicating a possible induced fit mechanism[186]

Crowley et al., for example, used dynamics techniques to study the enzyme cellulase, which breaks down the major plant polysaccharide, cellulose (Figure 2-23). This is the most abundant organic compound on Earth and harnessing the enzyme that digests it could be an important step towards an abundant source of biofuel; the knowledge of its mechanism of action was then elucidated for the conception of commercially viable mutants[186, 187].

SECTION 3
RESULTS

3 Results

3.1 Résumé des résultats

L'objectif de ce travail était d'utiliser des approches de mécanique moléculaire classique (§2.2) pour mieux comprendre les évènements structuraux et biologiques qui surviennent dans l'interaction entre les glucides et les récepteurs, en complément des informations obtenues à partir de méthodes expérimentales.

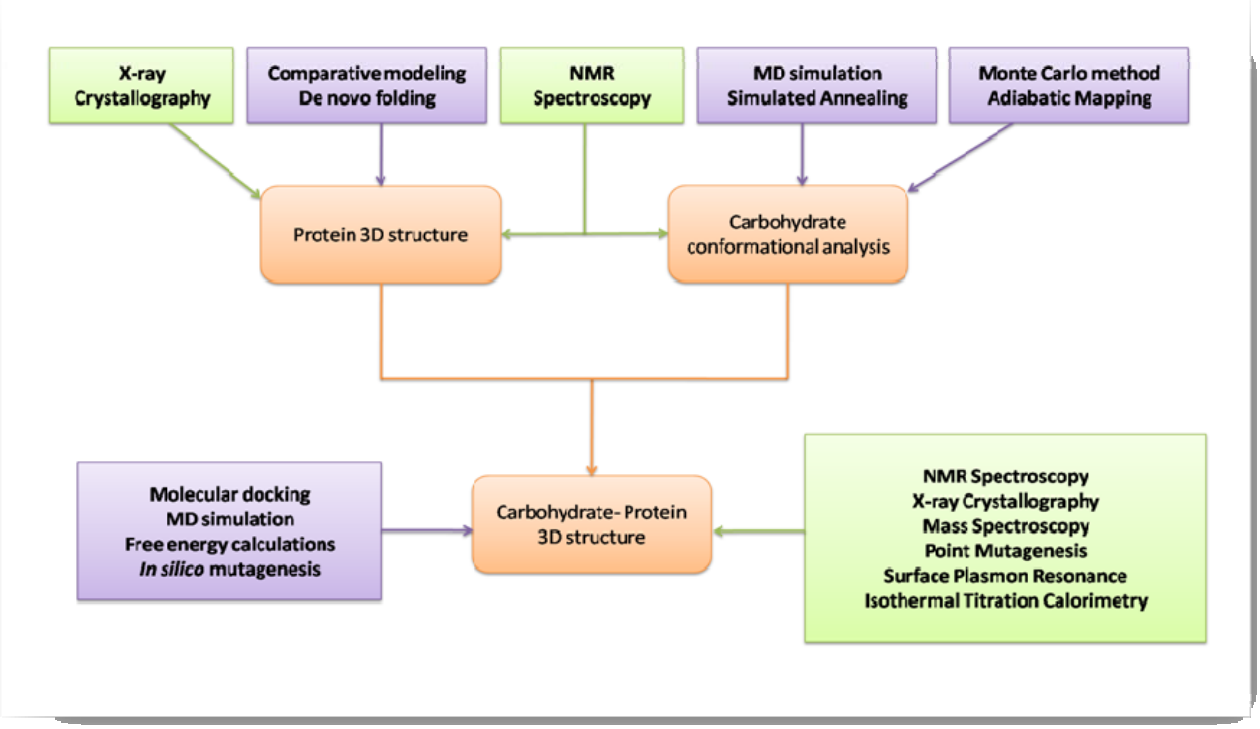


Figure 3-1 - Les rôles des méthodes in silico (violet) et expérimentales (vert) en glycobiologie structurale. The roles of computational methods (violet) alongside experimental methods (green) in structural glycobiology. Adapted from[106].

Méthodologie pour l'étude des lectines dépendantes du calcium et les interactions avec les glucides.

Les lectines dépendantes du calcium font partie de plusieurs familles de protéines qui, sont caractérisées par la présence d'un ou de deux ions de calcium dans le site de liaison, liés avec les groupes hydroxyles des sucres. La comparaison de plusieurs méthodes d'amarrage moléculaire (*docking*) a mis en évidence la capacité de ces approches à reproduire les principales interactions glucides-calcium-lectine. La méthode la plus pratique pour obtenir l'orientation expérimentale des glucides dans ce cas particulier a été choisie pour des études de *docking* sur des systèmes biologiques similaires (§3.2).

Langerine: une lectine humaine dépendante du calcium et spécifique pour le mannose.

Des fragments linéaires de mannose ont été amarrés en utilisant la structure cristallographique de la Langerine, une lectine humaine calcium-dépendante. Cette lectine est connue pour son rôle dans la reconnaissance des épitopes du virus VIH. Les résultats obtenus avec la technique de docking sont en accord avec les données structurales expérimentales qui suggèrent deux sites de liaison: un site dépendant du calcium et un site indépendant du calcium. Un modèle pour le domaine extracellulaire de la Langerine, construit en utilisant la structure trimérique de la protéine virale hemagglutinine, a été proposé. Ce modèle a montré une bonne corrélation avec les mesures hydrodynamiques et a également été utilisé pour interpréter l'organisation structurale des granules de Birbeck de la Langerine en microscopie électronique (§3.3).

PA-IL: une lectine humaine dépendante du calcium et spécifique pour le galactose.

PA-1L est une lectine dépendante du calcium dont le mécanisme d'action peut contribuer à expliquer la virulence de la bactérie *Pseudomonas aeruginosa*, principale responsable des maladies respiratoires chroniques chez les patients atteints de mucoviscidose. Les propriétés structurales de cette protéine en complexe avec des fractions oligosaccharidiques de glycosphingolipides ont été élucidées, en combinant plusieurs approches expérimentales et théoriques. L'antigène Gb3, dont la fraction glucidique est αGal1-4βGal1-4βGlc, a été proposé comme ligand naturel exprimé sur l'épithélium de l'homme (§3.4).

Les interactions entre la lectine PA-IL et trois digalactosides ont été également décrites en utilisant des approches de modélisation moléculaire et expérimentale. Nous avons montré que une molécule d'eau définie dans la structure cristalline est très importante pour la liaison digalactoside-lectine. La présence de cette molécule d'eau et l'inclusion de cette molécule d'eau dans les simulations conduisent à la prédiction de valeurs d'affinité en accord avec les experiences de microcalorimetrie (§3.4).
Les données présentées dans ces manuscrits pourraient être utilisées comme point de départ pour la conception d'inhibiteurs de la lectine PA-IL.

Roles des lectines végétales et des facteurs Nod dans la symbiose.

Le mécanisme de symbiose, très importante du point de vue agronomique et écologique, est initié par des bactéries du sol qui sécrètent des signaux, les facteurs Nod, de nature lipochitooligosaccharidique. Ces derniers vont interagir avec des récepteurs qui se trouvent au niveau de la racine des légumineuses, en changaent leur morphologie pour accueillir les bactéries fixant l'azote.
Dans le modèle *Medicago truncatula-Sinorhizobium meliloti*, les facteurs Nod sont perçus par les récepteurs de la légumineuse codés par les gènes *NFP* et *LYK3*. La connaissance des caractéristiques structurelles est nécessaire pour clarifier pleinement l'ensemble du mécanisme symbiotique. Des stratégies de modélisation moléculaire, combinées avec des analyses RMN, ont été appliquées pour étudier les conformations d'une nouvelle génération d'analogues de facteurs Nod. Ces études confortent l'idée que la partie glucidique de facteur Nod joue un rôle clé dans l'interaction avec ses récepteurs, tandis que la fraction lipidique, très flexible, agit comme régulateur de la spécificité du ligand, probablement en interaction avec un deuxième récepteur encore inconnu.
Deux modèles par homologie des récepteurs des facteurs Nod ont été construits pour supporter les données biologiques dans le cadre du projet européen NODPERCEPTION (§3.5).

Methodology for the study

of calcium-dependent protein-carbohydrate interactions

3.2 ARTICLE I: methodology for studying calcium dependent lectins *in silico*

In this article, we proposed a comparison of flexible docking methods in order to identify a reliable docking methodology able to reproduce the key carbohydrate-metal interactions that characterize the calcium dependent lectins and their carbohydrate recognition site.

Lectins display a variety of strategies for specific recognition of carbohydrates. In several lectin families from different origin, one or two calcium ions are involved in the carbohydrate binding site with direct coordination to the sugar hydroxyl groups (§1.3.1). Our work implied a molecular docking study involving a set of bacterial and animal calcium-dependent lectins. Flexible docking was performed using AutoDock, DOCK and Grid-based Ligand Docking with Energetics (GLIDE) software. All docking packages were able to predict the carbohydrate binding orientations but not in every instance the result was obvious to evaluate. DOCK showed good results according to crystallographic information but not in all tested cases the lowest energy conformation identified the experimental data. Same behavior was observed using GLIDE. However, using this program, the lowest energy pose was always a satisfactory solution, able to mimic the real carbohydrate orientation. AutoDock showed a reasonable accuracy in predicting the sugar orientation, based on docking cluster number ranking and the most accurate distances between calcium and sugar hydroxyl groups.

Autodock and GLIDE were used to predict the Galactose and N-AcetylGalactose binding mode in CEL-III, a new sea cucumber calcium-dependent lectin that displays haemolytic and cytotoxic properties.

Molecular Simulation
Vol. 34, No. 4, April 2008, 469–479

Taylor & Francis
Taylor & Francis Group

Comparison of docking methods for carbohydrate binding in calcium-dependent lectins and prediction of the carbohydrate binding mode to sea cucumber lectin CEL-III

A. Nurisso[a1], S. Kozmon[b] and A. Imberty[a1]*

[a]*Centre de Recherches sur les Macromolécules Végétales (CERMAV-CNRS), Grenoble, France;* [b]*Institute of Chemistry, Slovak Academy of Sciences, Bratislava, Slovak Republic*

(*Received 13 August 2007; final version received 23 September 2007*)

Lectins display a variety of strategies for specific recognition of carbohydrates. In several lectin families from different origin, one or two calcium ions are involved in the carbohydrate binding site with direct coordination of the sugar hydroxyl groups. Our work implied a molecular docking study involving a set of bacterial and animal calcium-dependant lectins in order to compare the ability of three docking programs to reproduce key carbohydrate-metal interactions. Flexible docking was performed using AutoDock, DOCK and Grid-based Ligand Docking with Energetics (GLIDE) softwares. All docking packages were almost able to predict the carbohydrate binding orientations but not in every instance the result was obvious to evaluate. DOCK showed good results according to crystallographic information but not in all tested cases the lowest energy conformation identified the experimental data. GLIDE presented the same difficulty in result analysis but the lowest energy pose was always a satisfactory solution, able to mimic the real carbohydrate orientation. AutoDock showed a reasonable accuracy in sugar orientation prediction based on docking cluster number ranking and most accurate distances between calcium and sugar hydroxyl groups. The latest program and GLIDE were used to predict the Gal and GalNAc binding mode in sea cucumber CEL-III, a new calcium-dependent lectin, that displays haemolytic and cytotoxic properties.

Keywords: C-type lectins; carbohydrates; autoDock; DOCK; GLIDE

1. Introduction

Lectins are ubiquitous proteins that recognise specifically carbohydrates but are devoided of catalytic activities and are not immunoglobulins [1]. This large family of proteins encompasses many biological functions that involve the deciphering of the sugar code [2], for example in intracellular glycoprotein trafficking, cell–cell signalisation or host recognition in pathogen infection. Lectins display a variety of carbohydrate binding site architectures, built in a large number of possible protein folds [3]. Due to the aliphatic character of monosaccharides, the carbohydrate binding involves a balance of hydrogen bonding and stacking of aromatic amino acids with CH of carbohydrates [4].

One particular carbohydrate binding mode involves the presence of a bridging cation ion in the binding site. Calcium-dependent lectins were first characterised in animal kingdom. The so-called C-type lectins cover a wide range of extracellular and membrane-bound proteins that contain one or several conserved carbohydrate recognition domains (CRDs) [5]. In mammals, they have multiple roles, including cell adhesion (selectins), glycoprotein clearance (asialoglycoprotein receptor), innate immunity (collectins), while in invertebrates they often play a role in non-self recognition processes involved in innate immunity and establishment of symbiosis [6,7]. Structural variations are observed in loops and disulfide bridges, but the amino acids of the sugar binding site that are involved in calcium binding are conserved. Upon carbohydrate binding, two of the sugar hydroxyl groups are involved in the coordination of calcium and also establish hydrogen bonds with the neighbouring amino acids [8]. Depending on the surrounding amino acids, C-type lectin CRD are specific for mannose and GlcNAc (and fucose) or for galactose.

More recently calcium-dependent lectins have been purified from opportunistic bacteria such as *Pseudomonas aeruginosa* [9]. These soluble lectins may play a role in host tissue recognition and/or in biofilm formation [10]. PA-IL is a tetrameric lectin specific for galactose and the involvement of O3 and O4 of galactose in the coordination of calcium is reminiscent to what is observed in animal C-type lectins [11]. The fucose/mannose specific PA-IIL exhibits a new motif of sugar binding with involvement of two close calcium ions (3.7 Å) that requires the participation of three hydroxyl groups in binding [12]; it is characterised by unusually strong (micromole) affinity for monosaccharides [13]. Whereas PA-IL has only been identified in *P. aeruginosa*, PA-IIL-like lectins were characterised from other opportunistic bacteria such as *Ralstonia solanacearum* and *Chromobacterium violaceum*

*Corresponding author. Email: imberty@cermav.cnrs.fr

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020701697709
http://www.informaworld.com

and the preference for mannose or fucose is correlated to the amino acid of a neighbouring loop [14,15].

Very recently, the calcium-dependent family of lectin has been completed by a rather surprising member. The sea cucumber lectin CEL-III adopts a double β-trefoil fold [16], also referred as R-type lectin, that has been previously characterised in many organisms and present three galactose binding sites with no calcium. However, in the Gal/GalNAc specific CEL-III lectin, five of the six sites contain a calcium ion in proper place for sugar binding and the lectin can therefore be considered as the first member of calcium-dependent β-trefoil lectin, although crystal structure of the complex has not been yet reported.

Molecular modelling of protein–carbohydrate interaction has proven to be a useful tool to rationalise specificity or to design glycomimetics that could be of therapeutical interest. Energy parameters that are suitable for energy minimisation and/or molecular dynamics of protein carbohydrate complexes are available for different force fields [17,18]. For predicting the carbohydrate orientation in binding sites, flexible docking methods have to be used in order to account for the possible orientations of pendent groups (i.e. hydrogen bond network directed by hydroxyl/hydroxymethyl group orientation) and also the conformational behaviour of the glycosidic linkage for oligosaccharides. Nevertheless docking methods have been used with a high rate of success for example using AutoDock program [19,20]. However, the presence of calcium ions directly involved in the binding site represents a special case. In some computer programs, the parameters for calcium have to be added to classical parameterisation. We propose here to compare three of the classically used flexible docking algorithms AutoDock [21], Dock [22] and Grid-based Ligand Docking with Energetics (GLIDE) [23] for docking monosaccharides and disaccharides in the different types of calcium dependent lectins that are displayed in Figure 1. The comparison with crystal structures will point out the strengths and weaknesses of each program. In addition, we used AutoDock 3 and GLIDE for predicting the docking of galactose in CEL-III, a new type of calcium dependent lectin for which no binding mode is yet structurally determined.

2. Flexible docking algorithms

In all of the three programs used in the study, the receptor is considered as a rigid body while the ligand is free to rotate, translate and change conformation during the docking application. Docking programs consists of two key parts: a search algorithm and a scoring function.

AutoDock is based on a hybrid search method that applied a Lamarckian genetic algorithm [21]. This exploration of binding site is based on a global searching that uses a genetic algorithm followed by an adaptive local search method derived from an optimisation of Solis and Wets algorithm [24] which has the advantage to not requiring gradient computation while it performs torsional space search. In the implementation of the genetic algorithm, the chromosome is composed of a string of real valued genes which describe the ligand translation, using the three Cartesian coordinates, the ligand orientation, involving four variables for the quaternion and the ligand conformation, defined by one real value for each torsion. The genetic algorithm begins by generating a random population of individuals which uniformly explore the grid space, followed by a specified number of generation cycles, each one consisting of a mapping and fitness evaluation, a selection, a crossover, a mutation and an elitist selection. After this step, each generation cycle is followed by the local search. At the end of every docking, AutoDock reports the docked energy, the state variables, the coordinates of docked conformation and the estimated free energy of binding [21].

DOCK [22] employs matching methods for the automated docking. In order to guide the search for ligand orientations, a negative image of the active site volume is created: spheres are only located in the receptor surface area that can interact with solvent molecules. This method allows for the limitation of possible ligand orientations to the most relevant region on the surface of the receptor. The incremental construction algorithm, called anchor-and-grow, separates the ligand flexibility into two steps: first, the largest rigid substructure of the ligand (anchor) is identified, rigidly oriented in the binding site [25]. The possible orientations are evaluated and optimised using the scoring function based on AMBER molecular mechanics force-field [26] and the energy minimiser based on the original Nelder and Mead algorithm [27]. The orientations are then collected according to their score, spatially clustered and prioritised. The remaining flexible portion of the ligand is built onto the best anchor orientations within the context of the receptor (grow). Only the interactions between ligand and protein are considered, leaving only intermolecular van der Waals and electrostatic components in the function; in addition, the receptor potential energy contribution can be pre-calculated and stored on a grid.

GLIDE program [23,28] uses a hierarchical series of filters to search for possible locations of the ligand in the active-site region of the receptor. The shape and properties of the receptor are represented on a grid by several sets of fields that provide progressively more accurate scoring of the ligand poses. Conformational flexibility is handled in GLIDE by an extensive conformational search, augmented by a heuristic screen. The scoring is carried out using Schrödinger's discretised version of the ChemScore empirical scoring function. Much as for ChemScore itself,

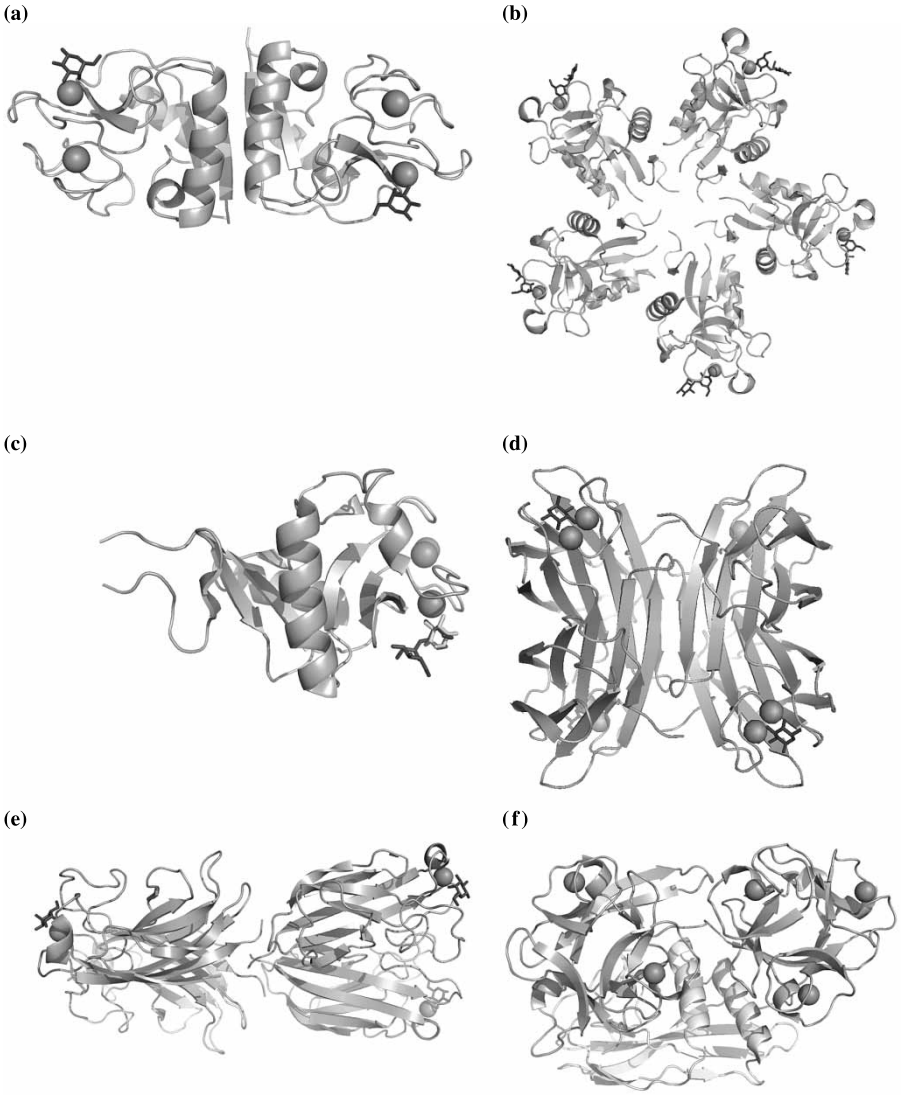


Figure 1. Graphical representation of the different calcium-dependent lectins that have been used in the present study. Peptide chains are represented by ribbon, sugar ligands by sticks and calcium atoms by spheres. (a): Tunicate lectin (C-type) complexed with galactoses (PDB code 1TLG) [38], (b): rattlesnake lectin (C-type) complexed with lactose (1JZN) [39], (c): CRD of human DC-SIGN (C-type lectin) complexed with mannose and mannobiose (1IT6) [40], (d): PA-IIL from *P. aeruginosa* complexed with fucose (1GZT) [12]. (e): PA-IL *P. aeruginosa* complexed with galactose (1OKO) [11]. (f). Sea cucumber CEL-III (1VCL) [16].

this algorithm recognises favorable hydrophobic, hydrogen-bonding, metal–ligand interactions and penalises steric clashes.

3. Materials and methods

3.1 Choice of lectin structures

All bacterial and animal calcium-dependent lectins chosen for this flexible docking study were selected from the lectin data base (http://www.cermav.cnrs.fr/lectines/). Coordinates were taken from the Protein Data Bank [29]. Lectins used in the present work are listed in Table 1 and the different architectures and quaternary arrangements are displayed in Figure 1.

3.2 AutoDock files and parameters

Receptor input files were generated using SYBYL 7.3 [30]. Monomers (or dimers when necessary) of each lectin were isolated; ligands, solvent and heteroatoms were removed except for the calcium ions located in the active sites. Hydrogen atoms were added and partial charges assigned using all-atoms charges of the AMBER force field [31]. Hydrogen atom positions were optimised

Table 1. List of calcium-dependent lectins used in docking calculations.

PDB	Organism	Protein	Ligand	Reference
C-type animal lectin				
1TLG	*Polyandrocarpa misakiensis*	Tunicate lectin	Galactose	[38]
1JZN	*Crotalus atrox*	Rattlesnake venom lectin	Lactose (βGal14βGlc)	[39]
			Dual binding:Mannobiose	
2IT6	*Homo sapiens*	DC-SIGN (CRD)	(αMan1–2αMan) Mannose	[40]
One calcium β-sandwich bacterial lectin				
1OKO	*Pseudomonas aeruginosa*	PA-IL	Galactose	[11]
Two calcium β-sandwich bacterial lectin				
1GZT	*Pseudomonas aeruginosa*	PA-IIL	Fucose	[12]
1UQX	*Ralstonia solanacearum*	RS-IIL	α-Me-mannoside	[15]
β-trefoil lectin fold				
1VCL	*Cucumaria echinata*	CEL-III	Galactose [a]	[16]
			N-acetylgalactosamine	

[a] no crystal structure of complexes available.

through energy minimisation using TRIPOS force field [32]. The programs *cnvmol2topdbq* and *addsol* included in AutoDock v.3.0.5 [21] were used to obtain the receptor atom coordinates, the partial charges, the atomic solvation parameters and fragmental atom volumes. Polar hydrogens were differentiated from non-polar hydrogens using 12-10 hydrogen bonding Lennard-Jones parameters and 12-6 hydrogen bonding Lennard-Jones parameters respectively. Additional atom type was defined for calcium ions using Lennard-Jones parameters previously proposed by Åqvist [33].

Ligands were extracted from CERMAV_3D monosaccharide and disaccharide databases (http://www.cermav.cnrs.fr/glyco3d/) and partial charges were taken according from PIM parameters for the TRIPOS force field [34]. All possible torsions in ligand molecules, including hydroxyl group rotations, were defined using the *deftors* module of AutoDock. The grid maps for van der Waals and electrostatic energies were prepared using *AutoGrid*: grid spacing was set to 0.375 Å, with 60 grid points placed on the centre of the ligand from the corresponding crystal structure complexes. Electrostatic interactions were evaluated using a distance-dependent dielectric constant to model solvent effects. Each single docking experiment consisted of 20 runs, employing a Lamarckian genetic algorithm and a rms deviation tolerance for cluster analysis of 1 Å. Applied parameters for the genetic algorithm are the default ones except for the number of energy evaluations that was set to 1×10^6. Clustering histograms created by AutoDock were analysed through the *get-docked* function to evaluate docking results.

3.3 DOCK files and parameters

The Chimera program [35] was used for preparation of ligand and receptor input files as suggested by Dock 6.1 users manual. Lectin structures were prepared using *Dock Prep* tool of Chimera with removal of ligand and solvent atoms. After addition of hydrogen atoms, partial charges were computed using the Antechamber package [36]. For calcium ions, formal charge was first assigned manually to obtain later the calculation of AM1-BCC atomic partial charge. Ligands were isolated from the original PDB lectins file and saved in MOL2 format after adding hydrogens and computing charges with Antechamber program. The active site for DOCK application was prepared using *dms* and *sphgen* programs. The maximum sphere radius was set to 4 Å while the minimum was set to 1.4 Å. In each case, spheres were selected using a root-mean-square deviation (RMSD) cut-off distance of 10.0 Å from all atom of the ligand crystal structure employing the accessory program *sphere_selector*. The interactive program *showbox* was used to visualise and define the location and size of the receptor box that defines the space for the docking conformational search taking into account the sphere set and enclosing an extra-margin of 5 Å in all the directions.

Grid maps were created using the accessory program *grid*. The *grid.in* file specified the parameters: the distance between grid points along each axis was 0.3 Å, the maximum distance between two atoms for their contribution to the energy score to be computed was set to 9999, the van der Waals energy potential parameters were taken from AMBER99 based on Lennard-Jones values using 6 and 12 as exponents of attractive and repulsive terms respectively. The van der Waals allowed overlap was set to 0.75 was and the dielectric factor coefficient to 4. An all atom model approach was chosen. All ligands were allowed to be flexible during the docking process driven by DOCK version 6.1 [10]. The input file *anchor_and_grow.in* was generated to set the default parameters for the anchor-and-grow algorithm. The number of scored conformers written was set to 20. Docking evaluations were performed with *ViewDock* utility of Chimera [35] executing the structure file *flex_scored.mol2* which contains a summary of best poses generated during the simulation.

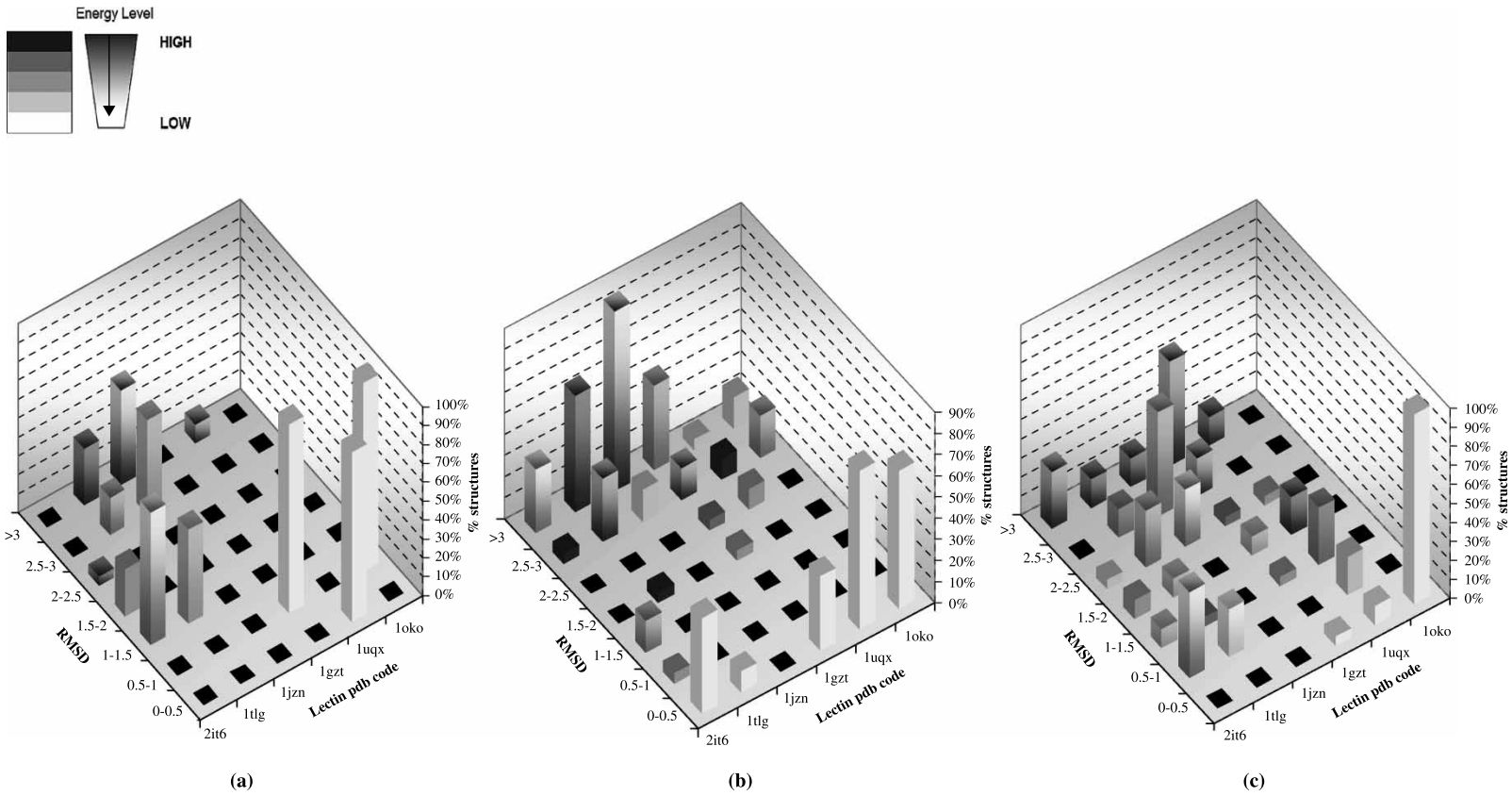


Figure 2. Graphical visualisation of docking simulation results: prediction accuracy of three different programs, AutoDock (a), DOCK (b) and GLIDE (c).

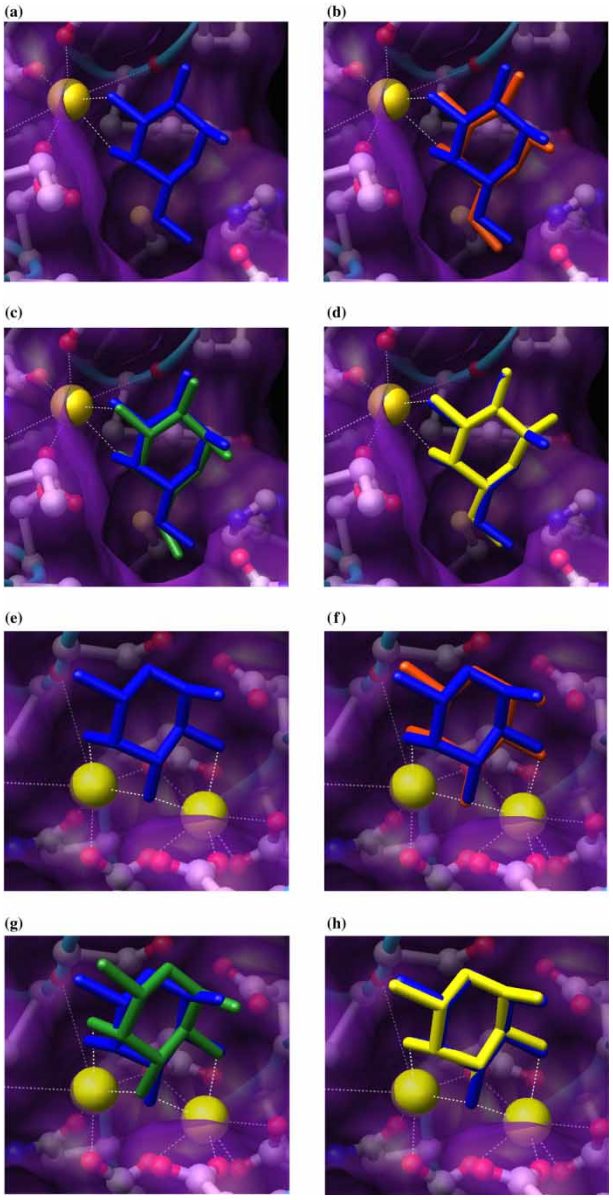


Figure 3. (a)–(d): Docking of α-D-galactose in PA-IL binding site in presence of one calcium ion; crystal structure 1OKO, ((a), blue) and comparison with lowest RMSD poses identified by AutoDock ((b), orange), DOCK ((c), green) and GLIDE ((d), yellow). (e)–(f): Docking of α-L-fucose in PA-IIL binding site in presence of two calcium ions; crystal structure 1GZT ((e), blue) comparison with lowest RMSD poses identified by AutoDock ((f), orange), DOCK ((g), green) and GLIDE ((h), yellow).

3.4 GLIDE files and parameters

The GLIDE program [23] was used in the First Discovery 2.7 package from Schrödinger, Inc. Missing hydrogens in the protein structure were added by Protein preparation routine implemented in First Discovery. Binding site was defined on the base of the available lectin structures with the bonded ligand. The box size was set to 14 Å in all three dimensions centred on the ligand. During the grid generation the parameter for Van der Waals radii scaling was scaled by 1.00 for atoms with partial atomic charge less then 0.25 and no constrains were defined. During the docking procedure the OPLS2001 partial atomic charges where assigned to the ligands. The parameters for Van der Waals radii scaling in docking was scaled by 0.80 for atoms with partial atomic charge less then 0.15. Ligand's poses were clustered within RMSD less then 0.5 Å and within maximum atomic displacement less then 1.3 Å. After the docking procedure 20 poses with the best score have been saved and used for the analyses. Docking results were analysed by GLIDE pose viewer included in MAESTRO (Schrödinger, Inc., NY).

3.5 Calculation of RMSD

We measured the RMSD between crystal and docked structures considering only the heavy atoms. All the values were determined using MAESTRO (Schrödinger, Inc., NY) and its *Superpose* command.

4. Results and discussion

4.1 Comparison of predicted binding modes and crystal structures

Since the aim of the study is to analyse how AutoDock, DOCK and GLIDE dock carbohydrates into the active sites characterised by calcium ions (i.e. the key heteroatoms implicated in sugar recognition), we first evaluated the orientation of the docked structures by comparing with crystal structures of complexes. Therefore the RMSD between crystal structure and lowest energy docking mode (and eventually most populated clusters) have been calculated and reported in Figure 2.

For AutoDock results (Figure 2(a)), histograms of cluster demonstrate that the binding mode of monosaccharides in bacterial lectins with one or two calcium (1oko, 1uqx and 1gzt) were very well predicted, within a window of 0.5 and 1 Å with respect to experimental data (Figure 3 (b),(f)). The best results were obtained for docking the monosaccharide α-methyl-mannoside (αMeMan) in *R. solanacearum* lectin (1uqx) binding site: 90% of carbohydrate conformations showed a RMSD which did not exceed 0.5 Å from the corresponding crystal structure while only 10% exceeded a rms value of 3 Å. Agreement was not so good for C-type lectins since for the rattlesnake venom lectin (1jzn), docking simulation gave 50% of structures with a RMSD between 2.5 and 3 Å and 50% with a RMSD of more than 3 Å. The tunicate lectin (1tlg) presented 50% of poses derived from the fusion of two clusters with different energy value and population number with a RMSD between 1 and 1.5 Å, followed by a percentage of 20 (RMSD 2.5–3 Å) and 30 (RMSD $>$ 3 Å) related to the other possible solutions.

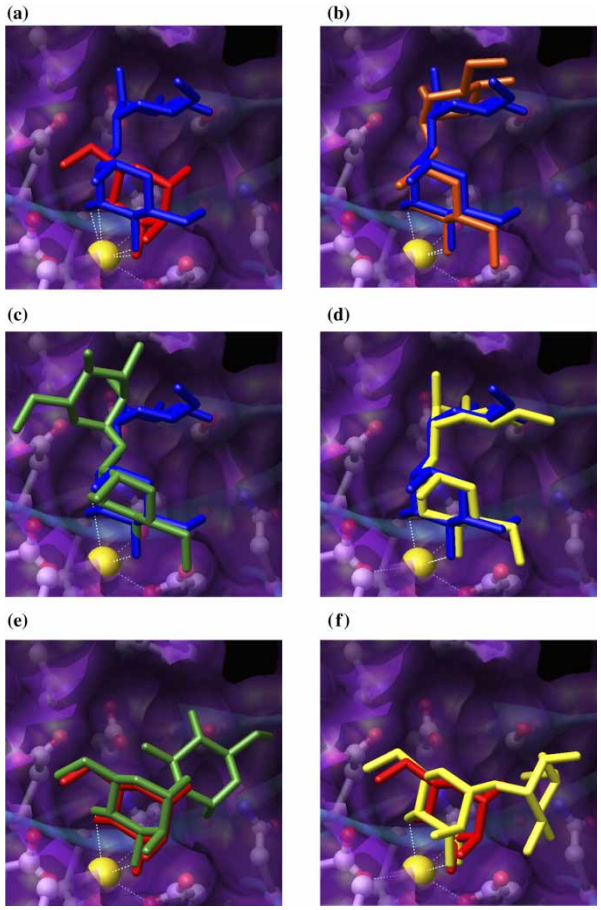


Figure 4. Docking of mannobiose in DC-SIGN binding site in presence of one calcium; crystal structure 2IT6 ((a), blue and red for the two binding modes observed in the crystal) and comparison with lowest RMSD poses identified by AutoDock ((b), orange), DOCK ((c), green), GLIDE ((d), yellow structure). Only DOCK and GLIDE (e, f) are able to reproduce the minor orientation observed for mannose in the crystal ((a), red).

However, for the human DC-SIGN (2it6), docking results always adopted the main conformation of mannobiose (Figure 4(b)) with a RMSD between 1 and 1.5 in 70% of cases. Each best pose (i.e. lower RMSD) derived from AutoDock was always the pose which had the highest number of cluster population in cluster histograms.

For DOCK results (Figure 2(b)), the best results were also obtained with bacterial lectins since for 1oko and 1uqx, the highest number of structures (65, 75%, respectively) were grouped in a single cluster with a RMSD of less then 0.5 Å. The high quality of the prediction is depicted in Figure 3(c), (d), for one-calcium and two-calcium bacterial lectins, respectively. Both major and minor orientation of the carbohydrate ligand in complex with the lectin DC-SIGN (2it6) were predicted with 35% of results fitting with the major binding mode in crystal (Figure 4(c)), 45% according with the minor binding mode (Figure 4(e)) while the rest of solutions had

Table 2. Experimental data and best docked structures: comparison in terms of root-mean-square deviation (RMSD).

Lectin PDB code	Autodock	DOCK	GLIDE
1oko	0.74	0.34	0.67
1uqx	0.43	0.41	0.29
1gzt	0.53	0.32	0.18
1jzn	2.92	2.62	2.11
1tlg	1.42	0.29	0.72
2it6 (a)	1.01	2.61	0.73
2it6 (b)[a]	–	0.48	0.55
RMSD average	1.18	1.01	0.75

[a] This row identifies the values derived from the second possible orientation of the carbohydrate in the binding site.

RMSD larger than 3 Å. Only a small cluster represented the right lactose orientation in snake lectin (1jzn) binding site with a RMSD between 2.5 and 3 Å. This group did not have the lowest energy conformation value. The same situation was noticed in docking tunicate lectin (1tlg): only 10% of poses had a RMSD which did not exceed 0.5 Å with respect of the known galactose binding mode.

In GLIDE docking analysis (Figure 2(c)), most results for the lectin PA-IL (1oko) displayed a RMSD less then 0.5 Å (Figure 3(d)). The carbohydrate orientation in RS-IIL and PA-IIL (Figure 3(h)) was also well recognised in 10% and 5% of cases, exhibiting the lowest energy conformation in each situation. The orientation of the disaccharide lactose in rattlesnake venom lectin (1jzn) binding site was reproduced in 30% of docking solutions with a RMSD between 2 and 2.5 Å. Most of results gave a RMSD between 2.5 and 3 Å (55%) followed by 15% of poses with a RMSD over 3 Å. For the tunicate lectin (1tlg), GLIDE showed 25% of results with a RMSD between 0.5 and 1 Å and favourable energy conformation. As for DC-SIGN, 43% of structures predicted the minor sugar orientation of mannobiose in 2it6 binding site, related to the major one by a 180° rotation, with a RMSD between 0.5 and 1 Å (Figure 4(d),(f)). In general, in each docking simulation performed by GLIDE, the lowest energy pose matched the conformation with the lowest RMSD, except few exceptions (Table 3) in which lowest energy poses did not differ so much with respect to crystallographic information.

Quantification of geometrical differences between the crystal structure and the closer docking mode is given in Table 2. Averaging over the six crystal structures analysed here indicated good performance of the three programs with rather limited variations among them. GLIDE displayed the lower mean RMSD (0.75 Å) with the two other ones giving values slightly higher than 1 Å. It should be noted that in the present state, the absolute values of the docking energy reported by each program are not comparable.

Table 3. Distances in Å between oxygen atoms of best docked carbohydrates and calcium ions in lectin binding site. Comparison with crystallography information.

		Crystal structure	Auto Dock	DOCK	GLIDE
1oko	Ca–O3 Gal	2.46	2.30	2.66	2.62
	Ca–O4 Gal	2.50	2.75	2.57	2.34
1uqx	Ca 1–O2 αMeman	2.51	2.32	3.07	2.92
	Ca 1–O3 αMeman	2.50	2.30	2.49	2.41
	Ca 2–O3 αMeman	2.52	2.43	2.83	2.79
	Ca 2–O4 αMeman	2.55	2.41	2.75	2.48
1gzt	Ca 1–O2 Fuc	2.54	2.24	2.52	2.77
	Ca 1–O3 Fuc	2.47	2.40	2.94	2.45
	Ca 2–O3 Fuc	2.48	2.32	2.53	2.49
	Ca 2–O4 Fuc	2.47	2.28	3.32	2.50
1jzn	Ca 1–O3 Gal	2.58	3.14	3.48	2.90
	Ca 1–O4 Gal	2.71	2.39	2.67	3.98
1tlg	Ca 1–O3 Gal	2.46	2.32	2.54	2.59
	Ca 1–O4 Gal	2.48	3.17	2.59	3.26
2it6	Ca 1–O3 Man	2.54	2.37	3.53	2.57
	Ca 1–O4 Man	2.48	2.27	2.56	2.81
2it6(b)	Ca 1–O3 Man	2.34	–	2.51	2.46
	Ca 1–O4 Man	2.62		3.13	2.65
Average relative error (%)			9.5	12.4	9.7

4.2 Comparison of calcium coordination

Since the direct involvement of calcium ion in carbohydrate binding is the characteristic feature of the lectins studied here, we analysed the ability of docking programs to predict co-ordination bonds with calcium ions. For the best docked solution, distances between the oxygen atoms from carbohydrate and the calcium ions are listed in Table 3. While the hydroxyl group oxygen atoms of carbohydrate displayed an average value of 2.5 Å to calcium ions in the crystalline state (2.49 Å ± 0.03), the docking program AutoDock had a tendency to shorten these distances (2.31 Å ± 0.07). Both DOCK (2.82 Å ± 0.38), and GLIDE (2.55 Å ± 0.15), would elongate them, and the results for both program also showed a larger distribution of values, indicating that this type of bond is maybe not sufficiently constrained in the parameterisation of these two programs. When calculating the average relative error in oxygen–calcium bond lengths, GLIDE, AutoDock and DOCK yielded values of 9.51, 12.39 and 9.70%, respectively. Thus AutoDock was the most accurate in keeping carbohydrate-calcium ion coordination distance immediately followed by GLIDE while DOCK accuracy was not so high.

It should be noted that the geometrical comparison performed above has been done using the solution with lowest RMSD compared to crystal structure. However, this "correct" solution, was not always the lowest energy one, and the programs have also been compared for their

Table 4. Heavy atom RMSDs and energies of best docked carbohydrate conformations. Italic entries identify structures in which best RMSD and lowest energy poses do not correspond.

		AutoDock RMSD/energy	DOCK RMSD/energy	GLIDE RMSD/energy
1oko	Lowest RMSD (Å)	0.74/−7.53	0.34/−35.78	0.67/−64.2
	Lowest energy (kcal/mol)			
1uqx	Lowest RMSD (Å)	0.43/−8.04	0.41/−42.82	0.29/−88.4
	Lowest energy (kcal/mol)			
1gzt	Lowest RMSD (Å)	0.53/−6.89	0.32/−36.91	0.18/−91.2
	Lowest energy (kcal/mol)			
1jzn	Lowest RMSD (Å)	*2.92/−8.30*	*2.62/−45.06*	*2.11/−62.0*
	Lowest energy (kcal/mol)	*> 4.00/−8.61*	*3.27/−46.42*	*2.35/−80.4*
1tlg	Lowest RMSD (Å)	*1.42/−6.45*	*0.29/−29.67*	*0.72/−36.5*
	Lowest energy (kcal/mol)	*> 4.00/−6.58*	*> 4.00/−31.20*	*0.79/−48.0*
2it6	Lowest RMSD (Å)	1.01/−7.75	2.61/−40.06	0.55/−55.9
	Lowest energy (kcal/mol)		0.48/−43.30[a]	0.73/−57.7[a]

[a] The asterisk corresponds to the second possible orientation of carbohydrate in the binding site.

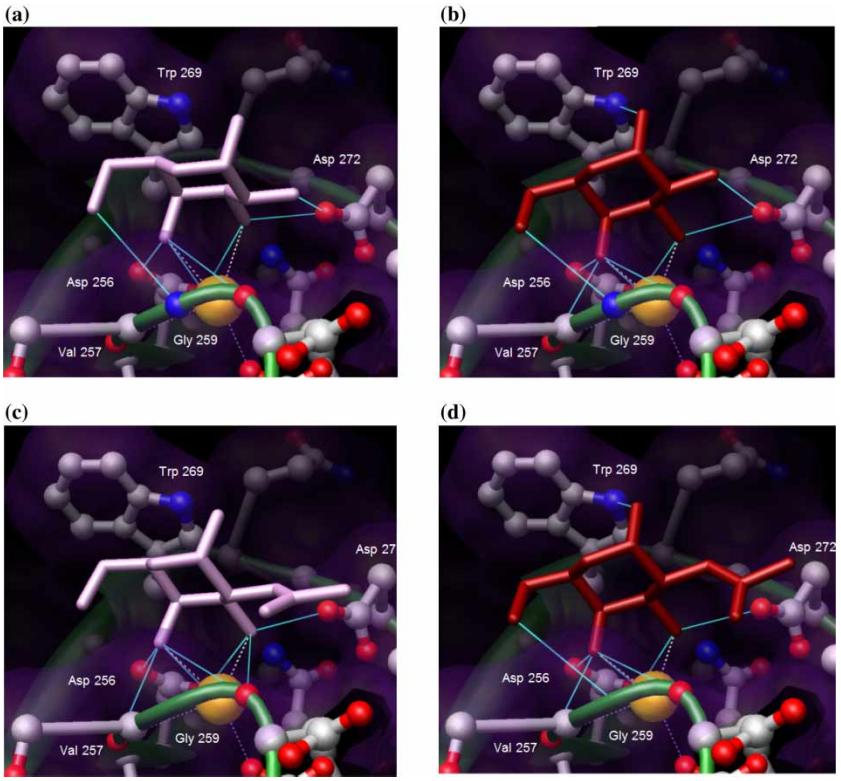


Figure 5. Modelling of the interation of galactose and *N*-acetylgalactosamine in CEL-III sub-domain 2γ according to GLIDE (plum, (a) for Gal and C for GalNAc) and AutoDock (red, (b) for Gal and (d) for GalNAc) prediction. Hydrogen bond network is coloured *cyan* while coordination bonds with calcium ion are showed in *white*.

ability to score the "correct" solution. Comparisons with best RMSD result and lowest energy one are listed in Table 4. All programs performed perfectly well for the bacterial lectins, but all of them failed in predicting the correct lowest energy conformation for C-type lectin from tunicate and snake. As seen above in Figure 2, no program could predict correctly the binding mode of lectin in tunicate lectin (1jzn). For the snake lectin (1tlg), DOCK and GLIDE could predict the correct binding with a reasonable energy cost. Alternatively, AutoDock had the advantage to predict correctly the docking mode when looking at the most populated cluster, instead of looking at the energy level.

4.3 Prediction of monosaccharide binding in sea cucumber CEL-III lectin

The sea cucumber (*Cucumaria echinata*) lectin CEL-III is a member of a new family of calcium-dependent lectin that have not yet been crystallised as complex with the ligand. The lectin adopts a double β-trefoil fold [16], also referred as R-type lectin and it is specific for galactose and *N*-acetylgalactosamine. Due to tandem repeats in the structure, several binding sites are predicted to occur and, indeed, five calcium ions have been located in putative binding sites.

According to the results described above, we used GLIDE and AutoDock for the prediction of Gal and GalNAc orientation in CEL-III. We focused our docking conformational search on the 2 γ-subdomain (named by reference to the ricin structure), concentrating on binding modes that present the lowest energy conformation for GLIDE software and the highest population number of its respective cluster in AutoDock. The resulting models are displayed in Figure 5 and the resulting hydrogen bond network and coordination geometries are listed in Tables 5 and 6. When docking galactose in CEL-III with either GLIDE or AutoDock, the resulting models indicated that O3 and O4 galactose oxygen atoms are involved directly in the coordination of calcium ion. From hydrogen bonds analysis, we found that the residue Asp 256 play a key role in the binding site forming hydrogen bonds with O3 and O4 hydroxyl groups and participating in calcium co-ordination. The residue Asp272 establishes contacts with O2 and O3 of galactose and coordinates calcium ion whereas Trp269 is involved in aromatic stacking interactions with carbohydrate ring. The large interaction area created by this interaction seems to strongly influence the sugar orientation in CEL-III binding site. The docking

Table 5. Hydrogen bonds and calcium coordination for α-D-galactose docked in sub-domain 2γ of CEL-III using GLIDE and AutoDock programs. The interatomic distances are reported in Å, weak hydrogen bonds (>3.1 Å) are indicated in italics.

Atom 1	Atom 2	GLIDE	AutoDock
Hydrogen bonds			
Gal.O1	Trp 269.NE1	–	2.74
Gal.O2	Asp 272.OD2	2.72	2.78
Gal.O3	Asp 256.OD1	2.92	2.48
Gal.O3	Asp 272.OD2	2.57	2.73
Gal.O4	Asp 256.OD1	*3.25*	3.02
Gal.O4	Asp 256.OD2	2.85	2.70
Gal.O4	Val 257.O	–	2.78
Gal.O4	Gly 259.O	*3.41*	3.07
Gal.O6	Gly 259.N	*3.18*	2.99
Coordination bonds			
O3.Gal	Ca	2.81	2.35
O4.Gal	Ca	2.55	2.17

Table 6. Hydrogen bonds and calcium coordination for α-D-GalNAc docked in sub-domain 2γ of CEL-III using GLIDE and AutoDock programs. The interatomic distances are reported in Å, weak hydrogen bonds (> 3.1 Å) are indicated in italics.

Atom 1	Atom 2	GLIDE	AutoDock
Hydrogen bonds			
Gal.O1	Trp 269.NE1	–	2.77
Gal.O3	Asp 272.OD2	2.66	2.80
Gal.O3	Asp 256.OD1	*3.15*	2.60
Gal.O4	Asp 256.OD1	*3.28*	*3.23*
Gal.O4	Asp 256.OD2	2.76	2.81
Gal.O4	Val 257.O	*3.12*	2.81
Gal.O4	Gly 259.O	*3.35*	3.08
Gal.O6	Gly 259.N	–	3.07
Coordination bonds			
O3.Gal	Ca	2.29	2.28
O4.Gal	Ca	2.56	2.35

of GalNAc did not present large variations as compared to galactose. Same contacts were maintained with the exception of the hydrogen bond between O2 of galactose and Asp272 that could not occur for GalNAc.

5. Conclusions

We evaluated and compared the reliability of three different softwares in performing flexible docking using calcium-dependent lectins as receptors and carbohydrates as ligands. This study revealed that all docking programs demonstrated the ability to accomplish docking work. The commercial package GLIDE slightly outperformed the other academic available docking engines in terms of RMSD from experimental structures but the best result did not always correspond to the lowest energy conformation. This conclusion is in agreement with recent evaluation study for docking carbohydrate derivative to Zn-containing enzyme [37]. Apart from showing a lower docking accuracy than GLIDE in terms of RMSD, DOCK program presented the same results as those noticed in GLIDE. The academic program AutoDock gave good results in tested simulations and clustering histograms created by the software always estimated the right carbohydrate position according to experimental information. Further improvements could be performed by optimising the docking parameters. Also results would be more statistically reliable by significantly increasing the number of runs. In the future, it will be of interest to also test the ability of these programs for docking large flexible oligosaccharides to calcium-dependent lectins and to evaluate the agreement between computed energies and known affinities.

Acknowledgements

The PhD thesis of A.N. and the stay of S.K. in Grenoble are financed by EEC Marie-Curie training program MEST-CT-2004-503322. This work was supported by the European Community's Human Potential Programme as a Research Training Network (Contract N° MRTN-CT-2006-035546[NODPERCEPTION]). The authors are thankful to Dr. Serge Pérez for his careful reading of the manuscript.

Note

1. Affiliated with the Joseph Fourier of Grenoble and Member of the Institut de Chimie Moléculaire de Grenoble.

References

[1] H. Lis and N. Sharon, *Lectins: carbohydrate-specific proteins that mediate cellular recognition*, Chem. Rev. 98 (1998), p. 637.
[2] H.J. Gabius, S. Andre, H. Kaltner, and H.C. Siebert, *The sugar code: functional lectinomics*, Biochim. Biophys. Acta 1572 (2002), p. 165.
[3] R. Loris, *Principles of structures of animal and plant lectins*, Biochim. Biophys. Acta 1572 (2002), p. 198.
[4] J.M. Rini, *Lectin structure*, Annu. Rev. Biophys. Biomol. Struct. 24 (1995), p. 551.
[5] K. Drickamer and M.E. Taylor, *Biology of animal lectins*, Annu. Rev. Cell Biol. 9 (1993), p. 237.
[6] A. Cambi, M. Koopman, and C.G. Figdor, *How C-type lectins detect pathogens*, Cell Microbiol. 7 (2005), p. 481.
[7] K. Drickamer and A.J. Fadden, *Genomic analysis of C-type lectins*, Biochem. Soc. Symp. (2002), p. 59.
[8] K. Drickamer, *Ca(2 +)-dependent sugar recognition by animal lectins*, Biochem. Soc. Trans. 24 (1996), p. 146.
[9] N. Gilboa-Garber, Pseudomonas aeruginosa *lectins*, Methods Enzymol. 83 (1982), p. 378.
[10] A. Imberty, M. Wimmerova, C. Sabin, and E.P. Mitchell, *Structures and roles of* Pseudomonas aeruginosa *lectins*, in *Protein–Carbohydrate Interactions in Infectious Disease*, C. Bewley ed., The Royal Society of Chemistry, Cambridge, 2006, p. 30.
[11] G. Cioci, E.P. Mitchell, C. Gautier, M. Wimmerova, D. Sudakevitz, S. Pérez, N. Gilboa-Garber, and A. Imberty, *Structural basis of calcium and galactose recognition by the lectin PA-IL of* Pseudomonas aeruginosa, FEBS Lett. 555 (2003), p. 297.
[12] E. Mitchell, C. Houles, D. Sudakevitz, M. Wimmerova, C. Gautier, S. Pérez, A.M. Wu, N. Gilboa-Garber, and A. Imberty, *Structural basis for oligosaccharide-mediated adhesion of* Pseudomonas aeruginosa *in the lungs of cystic fibrosis patients*, Nat. Struct. Biol. 9 (2002), p. 918.

[13] E.P. Mitchell, C. Sabin, L. Šnajdrová, M. Pokorná, S. Perret, C. Gautier, C. Hofr, N. Gilboa-Garber, J. Koča, M. Wimmerová, and A. Imberty, *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. Bioinform. 58 (2005), p. 735.

[14] M. Pokorná, G. Cioci, S. Perret, E. Rebuffet, N. Kostlánová, J. Adam, N. Gilboa-Garber, E.P. Mitchell, A. Imberty, and M. Wimmerová, *Unusual entropy driven affinity of* Chromobacterium violaceum *lectin CV-IIL towards fucose and mannose*, Biochemistry 45 (2006), p. 7501.

[15] D. Sudakevitz, N. Kostlanova, G. Blatman-Jan, E.P. Mitchell, B. Lerrer, M. Wimmerova, D.J. Katcoff, A. Imberty, and N. Gilboa-Garber, *A new* Ralstonia solanacearum *high affinity mannose-binding lectin RS-IIL structurally resembling the* Pseudomonas aeruginosa *fucose-specific lectin PA-IIL*, Mol. Microbiol. 52 (2004), p. 691.

[16] T. Uchida, T. Yamasaki, S. Eto, H. Sugawara, G. Kurisu, A. Nakagawa, M. Kusunoki, and T. Hatakeyama, *Crystal structure of the hemolytic lectin CEL-III isolated from the marine invertebrate* Cucumaria echinata*: implications of domain structure for its membrane pore-formation mechanism*, J. Biol. Chem. 279 (2004), p. 37133.

[17] A. Imberty and S. Pérez, *Structure, conformation and dynamics of bioactive oligosaccharides: theoretical approaches and experimental validations*, Chem. Rev. 100 (2000), p. 4567.

[18] S. Pérez, A. Imberty, S.B. Engelsen, J. Gruza, K. Mazeau, J. Jiménez-Barbero, A. Poveda, J.F. Espinosa, B.P. van Eyck, G. Johnson et al., *A comparison and chemometric analysis of several molecular mechanics force fields and parameters sets applied to carbohydrates*, Carbohydr. Res. 314 (1998), p. 141.

[19] P.M. Coutinho, M.K. Dowd, and P.J. Reilly, *Automated docking of glucosyl disaccharides in the glucoamylase active site*, Proteins 28 (1997), p. 162.

[20] A. Laederach and P.J. Reilly, *Modeling protein recognition of carbohydrates*, Proteins 60 (2005), p. 591.

[21] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson, *Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function*, J. Comp. Chem. 19 (1998), p. 1639.

[22] D.T. Moustakas, P.T. Lang, S. Pegg, E. Pettersen, I.D. Kuntz, N. Brooijmans, and R.C. Rizzo, *Development and validation of a modular, extensible docking program: DOCK5*, J. Comput. Aided Mol. Des. 20 (2006), p. 601.

[23] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry et al., *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy*, J. Med. Chem. 47 (2004), p. 1739.

[24] F.J. Solis and R.J.-B. Wets, *Minimization by random search techniques*, Math. Oper. Res. 6 (1981), p. 19.

[25] T.J.A. Ewing and I. D. Kuntz, *Critical evaluation of search algorithms for automated molecular docking and database screening*, J. Comp. Chem. 18 (1997), p. 1175.

[26] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case, *Development and testing of a general Amber force field*, J. Comput. Chem. 25 (2004), p. 1157.

[27] J.A. Nelder and R. Mead, *A Simplex-method for function minimization*, Comput. J. 7 (1964), p. 308.

[28] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, and J.L. Banks, *GLIDE: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening*, J. Med. Chem. 47 (2004), p. 1750.

[29] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *The protein data bank*, Nucleic Acids Res. 28 (2000), p. 235.

[30] SYBYL, Tripos Associates, 1699 S. Hanley Road, Suite 303, St Louis, MO 63144 USA.

[31] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M.J. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman, *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules*, J. Am. Chem. Soc. 117 (1995), p. 5179.

[32] M. Clark, R.D.I. Cramer, and N. van den Opdenbosch, *Validation of the general purpose Tripos 5.2 force field*, J. Comput. Chem. 10 (1989), p. 982.

[33] J. Aqvist, *Ion–water interaction potentials derived from free energy perturbation simulations*, J. Phys. Chem. 94 (1990), p. 8021.

[34] A. Imberty, E. Bettler, M. Karababa, K. Mazeau, P. Petrova, and S. Pérez, *Building sugars: the sweet part of structural biology*, in *Perspectives in Structural Biology*, M. Vijayan, N. Yathindra & A.S. Kolaskar, eds., Indian Academy of Sciences and Universities Press, Hyderabad, 1999, p. 392.

[35] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin, *UCSF Chimera—a visualization system for exploratory research and analysis*, J. Comput. Chem. 25 (2004), p. 1605.

[36] J. Wang, W. Wang, P.A. Kollman, and D.A. Case, *Automatic atom type and bond type perception in molecular mechanical calculations*, J. Mol. Graph. Model. 25 (2006), p. 247.

[37] P. Englebienne, H. Fiaux, D.A. Kuntz, C.R. Corbeil, S. Gerber-Lemaire, D.R. Rose, and N. Moitessier, *Evaluation of docking programs for predicting binding of Golgi alpha-mannosidase II inhibitors: a comparison with crystallography*, Proteins, 69 (2007), p. 160.

[38] S.F. Poget, G.B. Legge, M.R. Proctor, P.J. Butler, M. Bycroft, and R.L. Williams, *The structure of a tunicate C-type lectin from* Polyandrocarpa misakiensis *complexed with* D*-galactose*, J. Mol. Biol. 290 (1999), p. 867.

[39] J.R. Walker, B. Nagar, N.M. Young, T. Hirama, and J.M. Rini, *X-ray crystal structure of a galactose-specific C-type lectin possessing a novel decameric quaternary structure*, Biochemistry 43 (2004), p. 3783.

[40] H. Feinberg, R. Castelli, K. Drickamer, P.H. Seeberger, and W.I. Weis, *Multiple modes of binding enhance the affinity of DC-SIGN for high mannose* N*-linked glycans found on viral glycoproteins*, J. Biol. Chem. 282 (2007), p. 4202.

Langerin: a human C-type mannose binding lectin

3.3 ARTICLE II: *in silico* studies of human Langerin lectin

3.3.1 Introduction

All living organisms are constantly exposed to microbial pathogens present in the environment. For this reason, organisms need an efficient immune system, able to contrast pathogenic invasions and diseases. The innate immunity represents the first line of defense against pathogens and it is represented by biological structures such as skin and mucosae. Parallel to the innate immune system, the acquired or adaptive immune system uses lymphocytes, namely B and T, to recognize specific antigens. After their activation, they are immediately recruited to the site of infection to combat the pathogen. In addition, some of them are retained in the so-called memory cell *repertoire*, so that, in the event of any subsequent contact with that antigen, the acquired immune response will be faster and more efficient[188].

Dendritic cells act as messengers between the innate and adaptive immunities. They are bone marrow derived leukocytes that reside in an immature state in peripheral tissues. They have the ability to capture, internalize and process antigens as well as to migrate to lymph nodes. During the migration they undergo a maturation process ending with the presentation of processed antigens to T lymphocytes[189]

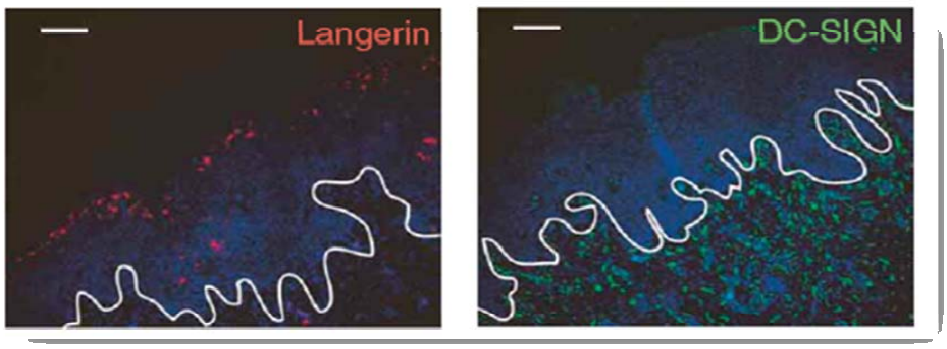


Figure 3-2 - Immunofluorescence microscopy ananlysis of human foreskin. Scale bar, 50 µm. Blue, nuclei; white, border epithelium (Langerin) and subephitelium (DC-SIGN). Adapted from[190].

Dendritic cells can be divided into several subsets that can be distinguished by the expression of specific calcium dependent lectins. Langherans cells specifically express langerin whereas other peripheral dendritic cells express DC-SIGN (*Figure 3-2*)[190]. The expression of the calcium dependent lectin langerin oriented in a type II configuration across the membrane of Langherans cells is associated with the formation of Birbeck granules, organelles formed by superimposed membranes separated by repetitive zipper-like striations, with vesicles that give the granule a typical tennis racquet shape. It has been shown that the formation of Birbeck granules is a consequence of the antigen-capture function of langerin allowing the routing of antigen into these organelles whose role is still not defined[191-193]. From a structural point of view, langerin is a type II transmembrane protein, a protein that spans the entire biological membrane, with an extracellular region consisting of a neck with a C-terminal domain containing a calcium-dependent carbohohydrate biding site with specificity towards mannose (CRD). Crystal structures of CRD has been described either in complex with mannose and maltose and a second binding site, non-dependent on calcium, has been observed[194]. The lectin associate as trimer through α coiled coil and the trimeric truncated extracellular region has been recently crystallized[195]. The single carbohydrate recognition domain shows low affinity to monosaccharides[196]. In its trimeric form (Figure 3-3), the three subunits are held in fixed positions, making the global structure more rigid and enhancing the selectivity to specific oligosaccharides[195, 197].

Figure 3-3 - Side view of the langerin trimer[195].

Recently, langerin has attracted attention for its role in immunity to human immunodeficiency virus-1 (HIV-1). Several studies have shown that DC-SIGN binds to gp120 N-glycans on HIV-1 and favor the infection of dendritic cells with subsequent transmission of HIV-1 to T-cells[190]. The entry of HIV-1 was also observed in fresh immature Langherans cells that efficiently bind to the HIV-envelope glycoprotein gp120 through langerin[190]. It was assumed that this binding leads to cell infection and consequent T-cells trans-infection-langerin mediated. However, recent data suggests that instead of promoting infection, the binding of langerin to gp120 prevents T-cell infection. The lectin captures the viral envelope glycoprotein, internalizing into Birbeck granules for a consequent degradation (Figure 3-4)[198]. The inhibition of langerin activity via monoclonal antibody or via binding-saturation with high concentrations of virus abrogates this protective effect, leading to HIV-1 infection and subsequent transmission of the virus to T cells[190]. These evidences indicate that langerin-mediated capture and internalization of HIV-1 is essential for the protective function of Langherans cells against T-cell infection. It has also been proved that the treatment of Langherans cells with lipopolysaccharides- or tumor-necrosis factors downregulates the expression of langerin, failing to prevent T-cell infection by HIV-1[199]. This suggests that in the absence of langerin, HIV-1-protective Langerans cells promote T-cell infection by the virus.

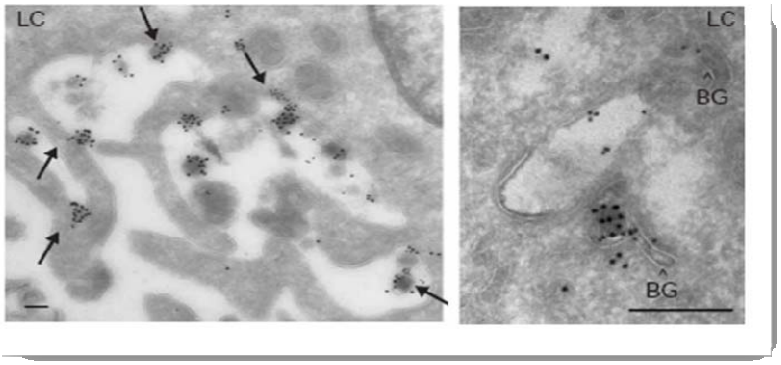


Figure 3-4 - Microscope view of HIV-I captured by langerin expressed through Langherans Cells (LC, arrows) and internalized into Borbeck granules (BG, arrowheads). Scale bar, 100nm. Adapted from[190].

Langerin also recognizes both mannose and ß-glucans present on fungal cell walls, demostating its role as fungal pathogen receptor on human Langheran cells[200].

3.3.2 Results

Linear trimannoside fragments have been docked to the calcium-dependent lectin langerin. This lectin is known for its implication in HIV glycan epitope recognition. Docking results are in good agreement with structural data on the two binding sites, suggesting the central mannose as monomer for binding to the calcium ion via 3-OH and 4-OH hydroxyl groups. An elongated model for the extracellular domain of langerin, constructed using the trimeric structure of mannose-binding protein and the helical bundle of the influenza virus hemagglutinin trimer, was proposed. This model showed good correlation with hydrodynamic measurements and was also used to interpret the organization of langerin in electron micrographs of Birbeck granules.

Structural Studies of Langerin and Birbeck Granule: A Macromolecular Organization Model[†,‡]

Michel Thépaut,[§,||,⊥,#] Jenny Valladeau,[§,▽] Alessandra Nurisso,[○] Richard Kahn,[||,⊥,#] Bertrand Arnou,[||,⊥,#,○] Corinne Vivès,[||,⊥,#] Sem Saeland,[◇] Christine Ebel,[⊥,#,△] Carine Monnier,[||,⊥,#] Colette Dezutter-Dambuyant,[▽] Anne Imberty,[○] and Franck Fieschi*[,||,⊥,#]

Laboratoire des Protéines Membranaires and Laboratoire de Biophysique Moléculaire, CEA, DSV, Institut de Biologie Structurale (IBS), 41 rue Jules Horowitz, Grenoble F-38027, France, CNRS, UMR 5075, Grenoble, France, Université Joseph Fourier, Grenoble F-38000, France, Centre Léon Bérard, INSERM U590, Lyon, France, CERMAV-CNRS, Grenoble, France, and DermImmune, 6 rue Sœur Bouvier, 69005 Lyon, France

Received November 21, 2008; Revised Manuscript Received January 28, 2009

ABSTRACT: Dendritic cells, a sentinel immunity cell lineage, include different cell subsets that express various C-type lectins. For example, epidermal Langerhans cells express langerin, and some dermal dendritic cells express DC-SIGN. Langerin is a crucial component of Birbeck granules, the Langerhans cell hallmark organelle, and may have a preventive role toward HIV, by its internalization into Birbeck granules. Since langerin carbohydrate recognition domain (CRD) is crucial for HIV interaction and Birbeck granule formation, we produced the CRD of human langerin and solved its structure at 1.5 Å resolution. On this basis gp120 high-mannose oligosaccharide binding has been evaluated by molecular modeling. Hydrodynamic studies reveal a very elongated shape of recombinant langerin extracellular domain (ECD). A molecular model of the langerin ECD, integrating the CRD structure, has been generated and validated by comparison with hydrodynamic parameters. In parallel, Langerhans cells were isolated from human skin. From their analysis by electron microscopy and the langerin ECD model, an ultrastructural organization is proposed for Birbeck granules. To delineate the role of the different langerin domains in Birbeck granule formation, we generated truncated and mutated langerin constructs. After transfection into a fibroblastic cell line, we highlighted, in accordance with our model, the role of the CRD in the membrane zipping occurring in BG formation as well as some contribution of the cytoplasmic domain. Finally, we have shown that langerin ECD triggering with a specific mAb promotes global rearrangements of LC morphology. Our results open the way to the definition of a new membrane deformation mechanism.

Dendritic cells (DCs)[1] are professional antigen-presenting cells able to specifically stimulate naive T-cells. They are in an immature state in mucosal and peripheral tissues. During migration to lymph nodes, they undergo a maturation process ending by the presentation of processed antigens to naive T lymphocytes (*1*). DCs can be divided into several subsets distinguishable by the expression of specific C-type lectins. Indeed, Langerhans cells (LCs), and a small subset of dermal DCs, specifically express langerin (CD207) (*2*) while other mucosal DCs express DC-SIGN (CD209) that has been subject to intense studies since its role in HIV transmission has been highlighted (*3*).

LCs are characterized by the presence of Birbeck granules (BG), pentalamellar and zippered membranes defining rod-shaped structures of different sizes with a central, periodically striated lamella (*4*). The correlation between langerin ac-

[†] Financial support from B. & M. Gates Foundation for M.T. postdoctoral grant and EEC Marie-Curie MEST-CT-2004-503322 training program for partial financing of A.N. "Sidaction-Ensemble contre le Sida" is also acknowledged for its financial support during the initial stages of this work.

[‡] Langerin CRD coordinates and structure factors have been deposited in the Protein Data Bank, ID code 3C22.

* Corresponding author. E-mail: fieschi@ibs.fr. Tel: +33-(0)-4-38-78-91-77. Fax: +33-(0)-4-38-78-54-94.

[§] These two authors have equally contributed to the work.

[||] Laboratoire des Protéines Membranaires, CEA, DSV.

[⊥] CNRS, UMR 5075.

[#] Université Joseph Fourier.

[▽] Centre Léon Bérard.

[○] CERMAV-CNRS.

[◇] DermImmune.

[△] Laboratoire de Biophysique Moléculaire, CEA, DSV.

[1] Abbreviations: αMan12Man, α-D-mannose-(1−2)-α-D-mannose; BG, Birbeck granules; BSA, bovine serum albumin; bp, base pairs; *c*(*s*), continuous distribution; CMS, cytomembrane sandwiching structures; CRD, carbohydrate recognition domain; DC, dendritic cells; DC-SIGN, dendritic cell-specific ICAM-3-grabbing non-integrin; DC-SIGNR, dendritic cell-specific ICAM-3-grabbing non-integrin-related protein; DLMM, double labeling minimum medium; ECD, extracellular domain; EDTA, ethylenediaminetetraacetic acid; ESRF, European Synchrotron Radiation Facility; HBSS, Hank's buffered salt solution; HIV, human immunodeficiency virus; IPTG, isopropyl β-D-1-thiogalactopyranoside; LC, Langerhans cells; Lg, langerin; M12M, α-D-mannose-(1−2)-α-methyl-D-mannoside; M12M12M, α-D-mannose-(1−2)-α-D-mannose-(1−2)-α-methyl-D-mannoside; M12M13M, α-D-mannose-(1−2)-α-D-mannose-(1−3)-α-methyl-D-mannoside; mAb, monoclonal antibody; MBP, mannose-binding protein; MIRAS, multiple isomorphous replacement with anomalous scattering; R_S, Stokes radius; $s_{20,w}$, corrected sedimentation coefficient at 20 °C in water; S-CRD, strep-tagged carbohydrate recognition domain; SDS−PAGE, sodium dodecyl sulfate−polyacrylamide gel electrophoresis; ΔCyto, deletion of the cytoplasmic domain (amino acids 1 to 28); ΔCRD, deletion of the carbohydrate recognition (amino acids 189 to 328); P231, point mutation of Pro23 into Ile.

10.1021/bi802151w CCC: $40.75
Published on Web 01/28/2009

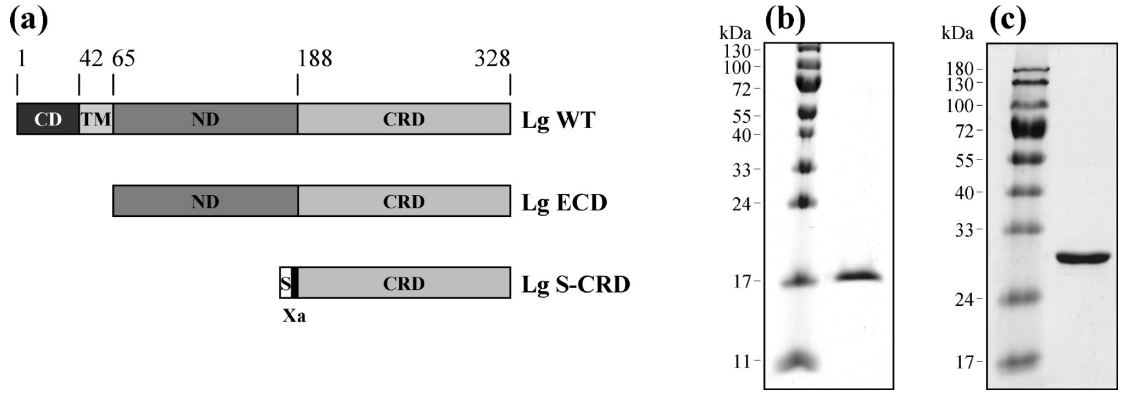


FIGURE 1: Langerin constructs used in these studies. (a) Representations of langerin domain organization and constructs used in these studies. CD, cytoplasmic domain; TM, transmembrane region; ND, neck domain required for oligomerization; CRD, carbohydrate recognition domain. For the Lg S-CRD construct a strep-tag II affinity tag was added to the N-terminus (noted S), followed by a Xa factor proteolytic site. (b, c) SDS−PAGE homogeneity analysis after purification of Lg S-CRD and Lg ECD, respectively.

cumulation and BG formation has been clearly demonstrated by the induction of their formation into fibroblasts or melanoma cell lines transfected with langerin cDNA (*2, 5*). Langerin is a type II membrane protein with an extracellular domain (ECD) containing a neck region and a C-type carbohydrate recognition domain (CRD) (*2, 6*). Its Ca^{2+}-dependent lectin properties have been confirmed (*7, 8*) with a monosaccharide specificity for mannose, fucose, and *N*-acetylglucosamine (*9*).

The roles of Birbeck granules are still a matter of debate. A disruption of the langerin gene in a mouse model did not induce any marked phenotypic alteration, apart from the disappearance of BGs. Indeed, no difference was observed compared to normal mice, in development, LC number in epidermis, antigen capture, nor LC maturation (*10*). Nevertheless, as illustrated for *Mycobacterium leprae* (*11*), a specific mobilization of langerin, in conjunction with CD1a, was demonstrated in the efficient presentation of nonpeptide antigen to T-cells. LCs, as other DCs, have been reported to bind HIV-1 and were then assumed to transmit HIV-1 to T-cells via langerin similarly to DC-SIGN (*12*). Moreover, LCs are localized in mucosal epithelia, and thus, during the initial exposition to HIV-1, they are the first DC subset in contact with the virus, emphasizing the potential importance of this DC subset in HIV infection (*13*). Recently, De Witte et al. brought new elements, criticized this point of view, and showed that langerin was able to inhibit LC infection and prevent HIV-1 transmission (*14*). They showed that HIV-1 is captured by langerin internalized into BGs and suggest its subsequent degradation. Therefore, langerin CRD may be directly involved in two phenomena: HIV-1 binding and Birbeck granule formation.

A better understanding of the interaction of langerin with HIV-1 as well as the involvement of this lectin in BG architecture requires detailed biochemical and structural characterization. Moreover, such studies are necessary to define a molecular strategy to selectively inhibit DC-SIGN without affecting langerin. The two lectins are most similar in their CRD domain while strong divergence, inducing different oligomerization states, exists in their neck domain. We have produced the CRD domain and the whole ECD of human langerin and solved the CRD structure at 1.5 Å resolution. A lower resolution structure of a slightly shorter version of langerin CRD, complexed with mannose and maltose, was very recently published (*15*). On the basis of our structure, gp120 N-glycan oligosaccharide binding was predicted by molecular modeling and compared with the published structure of the carbohydrate/CRD complexes mentioned above. Moreover, we report hydrodynamic studies on the purified langerin ECD. Resulting parameters were used to support a molecular model of the trimeric langerin ECD integrating the X-ray structure of the CRD. In parallel, electron microscopy analysis of BGs from isolated Langerhans cells is reported. From the dimension observed and the generated molecular model of the langerin ECD, a supramolecular organization of the BG is proposed. We have used different modified langerin constructs transfected in fibroblastic cells to determine the relative importance of the various domains of langerin in BG formation. Finally, we have shown that triggering of langerin ECD with specific mAb induces global rearrangements of LC morphology, which define of a novel membrane deformation mechanism.

EXPERIMENTAL PROCEDURES

Cloning and Expression of Recombinant Langerin Domains. Lg S-CRD (Figure 1) was expressed as previously described (*16*). The expression protocol of the Se-labeled Lg S-CRD used to solve the structure required a slightly different cloning. The cDNA coding for the Lg S-CRD was transferred into pET-15b vector (Novagen) since an ampicillin resistance vector was required for the production of Se-labeled Lg S-CRD. This latter plasmid was sequence checked and used to transform calcium-competent auxotrophic *Escherichia coli* BL21(DE3) selB::kan cys51E cells ($BL21_{cys}$) (*17*). Se-labeled Lg S-CRD was expressed using SeMet/SeCys double labeling (*18*). The pellet obtained from a 150 mL overnight culture was used to inoculate a 3 L culture in DLMM. The culture was grown at 37 °C for 6 h, and expression was induced by 1 mM IPTG. After 15 min incubation the cells collected by centrifugation were resuspended in 3 L of DLMM completed with 1 mM IPTG, 600 μM L-Se-cystine, and 425 μM L-Se-methionine. About 13 g of cells was collected after overnight incubation at 25 °C.

Lg-ECD construct comprises amino acids 68−328 (Figure 1). The corresponding cDNA was obtained by PCR and cloned into *Nde*I and *Bam*HI sites of pET-30b plasmid (Novagen). This construct was checked by sequencing and used to transform *E. coli* BL21(DE3) cells. Culture was

initiated from a 5% dilution of an overnight culture into Luria−Bertani medium with 50 mg/L kanamycin. Cells were grown for 3 h at 37 °C, and Lg ECD expression was induced by addition of 100 μM IPTG for an additional 3 h. Cells were harvested by centrifugation at 5000*g* for 20 min. The protein was expressed as inclusion bodies.

Protein Purification. Lg S-CRD and Se-labeled Lg S-CRD were one-step purified with the same protocol, on a streptactin superflow column (IBA GmbH), as previously described (*16*).

Since langerin ECD was expressed as inclusion bodies, a refolding step was required prior to the purification procedures. The pellet obtained from a 3 L culture was resuspended into 50 mL of buffer A (150 mM NaCl, 25 mM Tris, pH 7.8, and 4 mM $CaCl_2$). Cells were lysed by freezing at −20 °C, thawing, and sonication with addition of one complete EDTA-free tablet (Roche Diagnostics). Inclusion bodies were isolated by centrifugation at 10000*g* for 15 min at 4 °C. Refolding was performed by dilution and dialysis as previously described (*9*). Purification of functional Lg ECD proteins was achieved by affinity chromatography on a mannan−agarose column (Sigma) equilibrated in buffer A and eluted in same buffer without $CaCl_2$ but supplemented with 10 mM EDTA (buffer B). This step was followed by a Superose 6 size exclusion chromatography equilibrated in buffer A.

Retardation Assay. The Lg S-CRD functionality was tested by injecting 250 μg of purified protein on a 10 mL mannose−agarose column (Sigma). Equilibration and elution were performed in 150 mM NaCl, 25 mM Tris, pH 8, and 4 mM $CaCl_2$ buffer. As a control experiment 250 μg of bovine serum albumin (Sigma) was injected in the same conditions.

Crystallization, Data Collection, and Processing. Lg S-CRD and Se-labeled Lg S-CRD were crystallized in the same conditions as previously described (*16*). A 0.07 × 0.05 × 0.15 mm crystal of Se-labeled Lg S-CRD was cryoprotected in Paratone-N (Hampton Research) and was flash-frozen in liquid nitrogen. X-ray diffraction data were collected at FIP BM30A beamline at ESRF Grenoble. The peak (λ = 0.98009 Å) and inflection (λ = 0.98025 Å) wavelengths of the Se K adsorption edge were selected based on fluorescence from the crystal. Data sets of 360 images were collected at each wavelength with an oscillation range of 0.5° per image and an exposure time of 40 s. The crystal-to-detector distance was 256.18 mm. Data were processed using the program XDS (*19*). Se-labeled crystals are isomorphous to the native ones (*16*). Diffraction data, including previously collected native data (*16*), were merged using the program CAD from the CCP4 package (*20*).

Phasing, Model Building, and Structure Refinement. First steps of structure determination were performed using the program autoSHARP (*21*). Matthews coefficient determination (*16*) suggests that the asymmetric unit contains either three or four molecules. A MIRAS structure determination was performed using the native data set (*16*) and the two Se derivative data sets (peak and inflection). As there are two diselenium bridges per molecule, each forming one heavy-atom "supersite", and because the asymmetric unit may contain either three or four molecules, six selenium sites were initially searched and eight supersites were finally found. The electron density maps and the initial model resulting from autoSHARP were used as the starting point for model building. The structure refinement was performed on the native data set by cycling between manual building using the program COOT (*22*) and energy minimization with the program REFMAC 5 from the CCP4 package (*20*). Statistics of structure refinement for the native data set are summarized in Table 1.

Table 1: Lg S-CRD Data Collection and Structure Refinement Statistics

Data Collection Statistics for Se-Labeled Crystal		
data set	peak	inflection
wavelength (Å)	0.980089	0.980252
space group	$P4_2$	$P4_2$
unit cell parameters (Å)	a = b = 79.81, c = 90.13	a = b = 79.86, c = 90.15
resolution (Å)	50−2.15 (2.27−2.15)[a]	50−2.15 (2.27−2.15)
measured reflections	229732 (35243)	230246 (35708)
unique reflections	60525 (9451)	60639 (9585)
completeness (%)	99.3 (96.1)	99.5 (97.6)
$I/\sigma(I)$	13.7 (5.1)	14.3 (4.7)
R_{merge}[b] (%)	8.2 (24.8)	8.6 (29.1)
Structure Refinement Statistics for Native Data (*16*)		
resolution (Å)	50−1.5 (1.59−1.5)	
refinement factors		
used reflections/free (%)	84765/5.01	
R_{cryst}[c]	0.189	
R_{free}[c]	0.236	
rmsd from ideality		
bond lengths (Å)	0.014	
bond angles (deg)	1.516	
Ramachandran plot (%)		
most favored regions	86.4	
additional allowed regions	13.6	
generously allowed regions	0.0	
disallowed regions	0.0	
average *B*-factor (Å^2)		
main chains	17.7	
side chains	19.1	
all protein atoms	18.4	
waters	33.4	
all atoms	20.9	
rms B for main chain	0.8	
rms B for side chain	1.6	

[a] Values into parentheses are for the highest resolution shell. [b] $R_{merge} = \sum_h\sum_m|I_m(h) - \langle I(h)\rangle|/\sum_h\sum_m I_m(h)$. [c] $R_{cryst} = \sum||F_o| - |F_c||/\sum|F_o|$, and $R_{free} = R_{cryst}$ calculated with 5% of F_o sequestered before refinement.

Stokes Radius Determination of Langerin Extracellular Domain by Gel Filtration. Superose 12 (Amersham) was calibrated with a mixture of well-characterized proteins with known R_S (gel filtration calibration kits; GE Healthcare). The column was equilibrated in buffer A for 2 column volumes at a 0.8 mL/min flow rate. Dextran blue and FAD were run to determine dead and total volume of the column. Independently, a sample of 0.2 mg of Lg-ECD was also run on the column. All proteins eluted as single peak, and their elution volumes allowed the calculation of the K_{av} by the equation:

$$K_{av} = (V_e - V_0)/(V_t - V_0)$$

V_e being the sample elution volume, V_0 the dead volume, and V_t the total volume. The known Stokes radius of each control proteins allowed to plot a calibration curve, log R_S as a function of K_{av}.

Analytical Ultracentrifugation. Sedimentation velocity experiments were performed at 42000 rpm and 20 °C, using an AN-60 rotor in a Beckman XL-I analytical ultracentrifuge, with two samples of 100 and 400 μL of Lg-ECD at 0.84 mg/mL ($A_{280} \approx 1.6$) and 0.09 mg/mL in 0.3 and 1.2 mm path length cells, equipped with quartz windows, respectively.

Solvent was buffer B. Scans were recorded overnight every 6 min at 280 nm. Data were analyzed using the programs SEDFIT and SEDPHAT (www.analyticalultracentrifugation.com) in terms of a continuous distribution $c(s)$ of sedimentation coefficients, s, and one noninteracting particle (*23*). The solvent density, $\rho = 1.007$ g/mL, viscosity, $\eta = 1.033$ mPa·s, and partial specific volume of langerin, $\bar{v} = 0.734$ mL/g, were estimated from composition with the program SEDNTERP (www.jphilo.mailway.com) and used to derive corrected sedimentation coefficients, $s_{20,w}$. The Svedberg equation relates s to protein molar mass, M, and Stokes radius, R_S: $s = M(1 - \bar{v}\rho)/(N_A 6\pi\eta R_S)$ (with N_A being Avogadro's number). Values for globular compact monomer and dimer were calculated with $R_S = f/f_{min}R_{min}$, R_{min} being the minimum theoretical value for R_S.

Molecular Modeling. The AutoDock 3.0 program (*24*) was used for docking carbohydrate ligands in langerin. Receptor input files were generated using SYBYL 7.3 for the four langerin molecules present in the asymmetric unit. Hydrogen atoms were added and partial charges assigned using AMBER all-atom charges. Hydrogen atom positions were optimized using TRIPOS force field (*25*). Polar hydrogens were differentiated from nonpolar hydrogens using 12-10 and 12-6 hydrogen-bonding Lennard-Jones parameters, respectively. Additional atom type was defined for calcium ion using Åqvist parameters as recently described (*26*). Carbohydrate ligands were built using SYBYL 7.3, and partial charges were taken to PIM parameters for the TRIPOS force field (*27*). All possible torsions in ligand molecules, including hydroxyl groups, were defined as flexible. Grid spacing was set to 0.375 Å. Two different docking tests were assayed: the first one included only the space around the calcium while the second one included the whole monomer. Each single docking experiment consisted in 100 runs employing a Lamarckian genetic algorithm with default parameters except for the number of energy evaluations, set to 1×10^6. The rms deviation tolerance for cluster analysis was set to 1 Å. The molecular model of the trimer was built by homology method, using the COMPOSER program (*28*). The longest identified trimeric coiled-coil, that of influenza virus hemagglutinin (1QU1) (*29*), was used as a template for building the trimeric neck region of langerin. The junction with the CRD domain was obtained by similarity with the MBP-A structure, another trimeric C-type lectin (1RTM) (*30*). Hydrogen atoms were added to the model and final energy minimization included geometry optimization of side chains.

LC Isolation from Human Skin. Human epidermal cell suspensions were obtained from normal skin of patients undergoing abdominal reconstructive plastic surgery after patient-informed consent and according to institutional guidelines. Skin was split-cut with a keratome set and the dermo-epidermal slices were treated for 18 h at 4 °C with 0.05% trypsin in HBSS without Ca^{2+} and Mg^{2+}. The epidermis was detached from the dermis with fine forceps. Epidermal cell suspensions were obtained by subsequent tissue dissociation and filtration through sterile gauze. Basal keratinocytes were removed by adhesion on collagen type I-coated plates (Corning-Iwaki Glass). Partial enrichment was obtained by density gradient centrifugation on Pancoll (Pan Biotech GmbH) which yielded 20−50% $CD1a^+$ LC based on 10 experiments.

Cloning of Langerin Constructs Used in Transfection Studies. Deleted forms were obtained by PCR performed on langerin cDNA (*2*). Primers were designed to amplify fragments incorporating a *Sal*I site followed by the Kozak sequence in the 5′ untranslated region and a stop codon followed by a *Not*I site to the 3′ end. Primers for "Lg ΔCRD" defined a 611 bp fragment comprising residues 1−188 of langerin. Primers for "Lg ΔCyto" defined a 929 bp fragment comprising residues 29−328 of langerin. Mutagenesis was performed using the "gene editor in vitro site-directed mutagenesis" (Promega). A primer containing the P23I mutation was designed to create a silent *Bam*HI restriction site that was subsequently used as a screen to detect mutated clones. All PCR were performed for 35 cycles (1 min denaturation at 94 °C, 1 min annealing at 60 °C, and 2 min elongation at 72 °C) with Taq polymerase. PCR products were then purified using the Wizard prep system (Qiagen), and ligations in pCRII-TOPO vector (Invitrogen) were performed. Putative positive clones were sequence checked. cDNA encoding mutated and deleted langerin forms were then transferred in pMET7 vector using the rapid DNA ligation kit (Boehringer).

Transfection and Expression of Langerin Constructs in COP5 Cells. Murine fibroblastic COP5 cells were cultured in RPMI 1640 supplemented with 10% heat-inactivated fetal bovine serum (FBS), 10 mM HEPES, 2 mM L-glutamine, 50 μM 2-mercaptoethanol, and gentamycin (80 μg/mL). The cells were then electroporated with the langerin expression constructs. Langerin expression was evaluated by FACS analysis.

Cytofluorometry Analysis. COP5 transfected cells were incubated with DCGM4 mAbs (*31*) (10 μg/mL; Beckman Coulter) or with anti-Lag mAbs (*32*) (Becton Dickinson). mAbs were revealed with goat anti-mouse Ig-FITC conjugated (Becton Dickinson). Intracellular staining was performed in the presence of permeabilization medium (0.3% saponin, 2% BSA). Negative controls were performed with an isotype Ig control. Fluorescence was analyzed with a FACScan flow cytometer (Becton Dickinson). Mortality was evaluated using PI incorporation (Sigma).

Transmission Electron Microscopy. After washing, isolated LCs or COP5 fibroblasts transfected with langerin cDNA were fixed with 2% glutaraldehyde in cacodylate buffer for 18 h. After being rinsed in cacodylate buffer with sucrose for 12 h, the cells were processed for transmission electron microscopy. Cells were postfixed with an aqueous solution of 1% osmium tetroxide in cacodylate buffer with sucrose and embedded in epoxy medium after dehydration through a graded series of ethanol. Ultrathin sections were stained with lead citrate and uranyl acetate and examined with a JEOL 1200EX electron microscope with acceleration voltage of 80 kV (Centre des Microstructures, Lyon University, France).

RESULTS

Langerin CRD and ECD Production and Purification. The truncated forms of langerin used in this study are presented in Figure 1a. Lg S-CRD was expressed as a soluble form in the *E. coli* periplasm. The strep-tag II purification tag allowed one-step purification to homogeneity (Figure 1b). The protein functionality was assessed on a mannose−agarose column

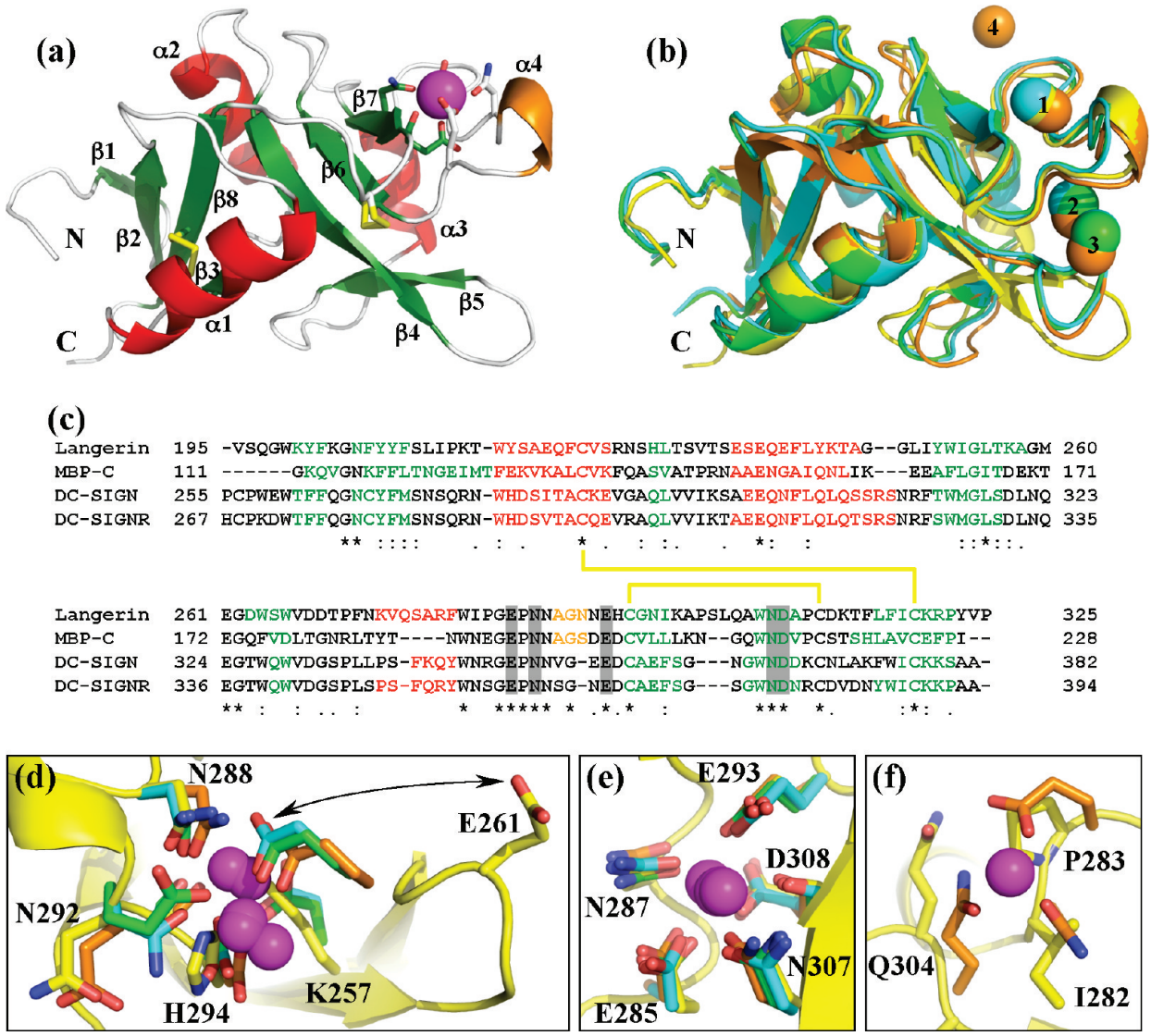


FIGURE 2: Langerin S-CRD structure comparison with other C-type lectin CRDs and structural analysis of langerin calcium-binding site. (a) Structure of Lg S-CRD. α-Helices are shown in red, 3_{10}-helix in orange, *β*-strands in green, loops in white, disulfide bridges in yellow, and Ca^{2+} atom in magenta. Side chains of residues involved in calcium binding are shown as sticks. (b) Superimposition of human langerin (3C22 in yellow), MBP (1HUP in orange), DC-SIGN (1K9I in green), and DC-SIGNR (1K9J in cyan) CRD structures. Numbering indicates calcium-binding sites. (c) Sequence alignment of CRD visible residues in structures of human langerin, MBP, DC-SIGN, and DC-SIGNR obtained from the ClustalW multiple alignment software (www.ebi.ac.uk/Tools/clustalw/index.html). DSSP predicted secondary structures (from DaliLite results) are shown with the same color code as in (a). Conserved disulfide bridges appear in yellow lines. Identical residues are indicated by a "*", conserved substitutions by a ":", and semiconserved substitutions by a ".". Conserved residues binding the carbohydrate interacting Ca^{2+} are highlighted in gray. (d) Region of calcium sites 2 (in back) and 3 (in front). Double arrow indicates Glu261 displacement due to *β*4−*β*5 loop shift. (e) Region of calcium site 1. (f) Region of calcium site 4. The structural representations were drawn with the PyMol program (www.pymol.org).

where langerin CRD elution is delayed whereas the bovine serum albumin, used as a control protein, elutes in the dead volume (Supporting Information Figure 1a).

Langerin extracellular domain (Lg ECD) was produced without an affinity tag. Its oligomeric nature allowed tight binding, through an avidity-based mechanism, on a mannan−agarose column (Supporting Information Figure 1b). Elution required EDTA, demonstrating that binding is calcium dependent and that the Lg ECD has been functionally refolded. Again, this construct was obtained to homogeneity by one-step purification (Figure 1c).

Langerin CRD Structure. The selenium derivative crystal was isomorphous to the native one (*16*). Data processing and refinement statistics are reported in Table 1.

The refined structure constructed from the native data set and the autoSHARP model, obtained from MIRAS phasing, was validated by the program PROCHECK (*33*). The resulting Ramachandran plot showed that 86.4% residues are in the most favored regions, 13.6% residues are in the additional allowed regions, and no residues are in generously allowed regions nor disallowed regions. The asymmetric unit is composed of four monomers related by noncrystallographic 222 point group symmetry. Depending on the monomer, the first 21−24 residues of Lg S-CRD, containing notably the strep-tag II, were absent from the structure. At the C-terminus only three residues are missing except in chain D where the carboxyl terminus was fixed by the two water molecules fulfilling Ca^{2+} coordinations of the chain A carbohydrate-binding site.

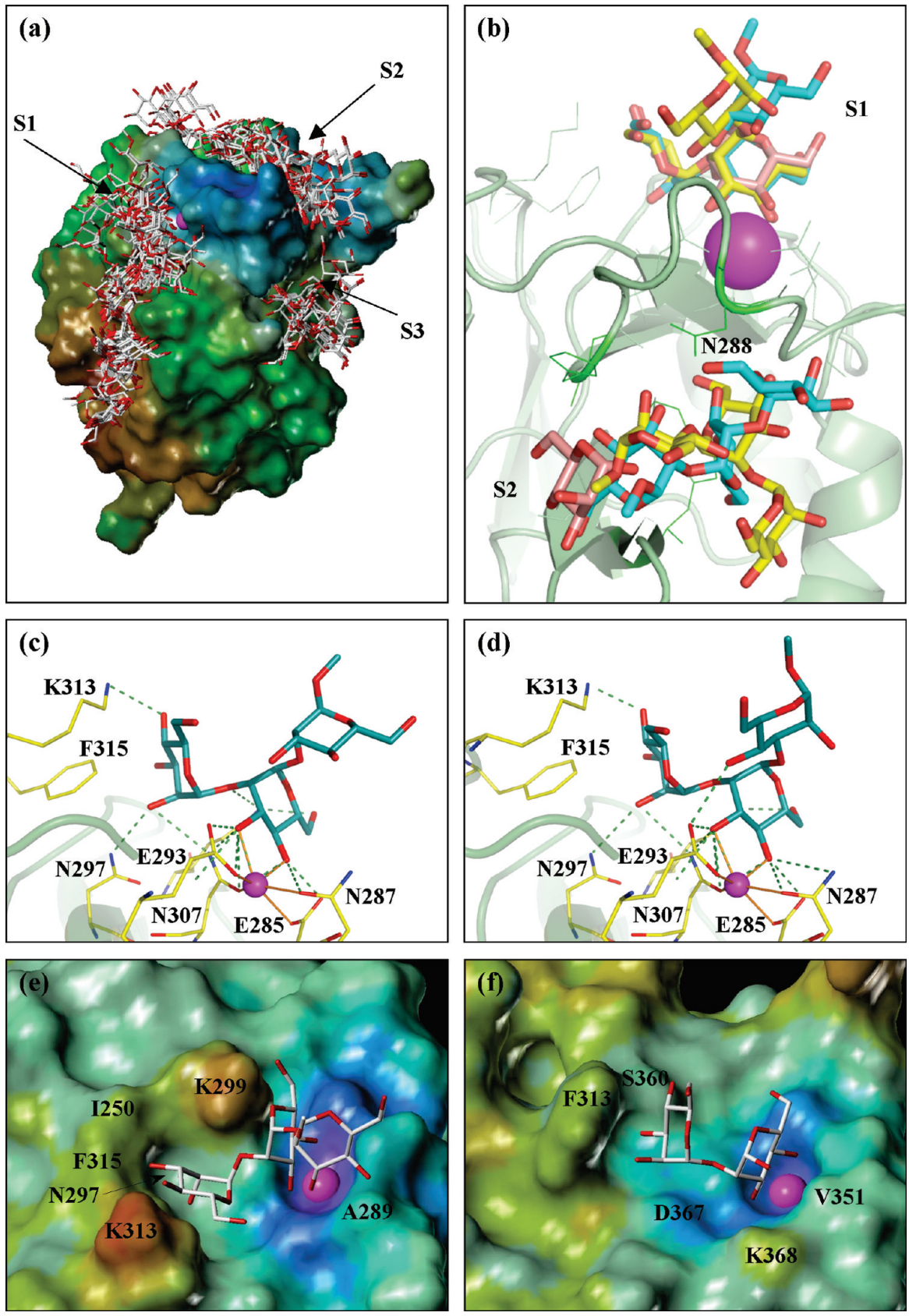


FIGURE 3: Modeling of oligosaccharide interaction with langerin CRD. (a) All predicted orientations for M12M12M represented on the accessible surface of langerin CRD color-coded according to lipophilic potential, polar (blue) to hydrophobic (brown). S1 and S2 are high probability sites; S3 is a low probability site. (b) Lower energy docking modes for M12M12M (cyan) and M12M13M (yellow) in sites S1 and S2. The mannose residue reported in the crystal structure of langerin CRD (*15*) is also represented (salmon). In the S2 site, side chains in green represent amino acids involved in oligosaccharide binding. (c, d) Lower energy docking modes in the calcium-binding site for M12M12M and M12M13M, respectively. (e) Lower energy docking modes for M12M12M together with accessible surface of langerin CRD color-coded according to electrostatic potential, acidic (blue) to basic (red). (f) Same representation for the structure of α-D-mannose-(1−2)-α-D-mannose in DC-SIGN site (2IT6).

Table 2: Hydrodynamic Properties of Lg ECD and Trimeric Model[a]

	Lg ECD		trimeric model
	analytical ultracentrifugation	gel filtration	HYDROPRO calculation
$s_{20,w}$ (S)	4.05 ± 0.1	na	3.9
R_S (nm)	5.1 ± 0.06	5.13 ± 0.3	5.4
ff_{min}	1.73 ± 0.02	na	1.8

[a] $s_{20,w}$, sedimentation coefficient; R_S, Stokes radius; ff_{min}, frictional ratio; na, not applicable.

Langerin CRD structure was compared with the structures of human mannose-binding protein (MBP) CRD (1HUP), DC-SIGN CRD (1K9I), and DC-SIGNR CRD (1K9J). At first sight, the overall structure of langerin CRD (Figure 2a) roughly observes the classical conserved fold and features of C-type lectins as illustrated by the aligned structures and sequences (Figure 2b,c). A more precise view pinpoints some specific features. Like MBP, langerin has a 3_{10}-helix near the carbohydrate-binding calcium site that is not conserved in DC-SIGN nor in DC-SIGNR structures. Additional differences concern calcium-binding sites. Only the carbohydrate-binding calcium site was conserved in langerin CRD (site 1 in Figure 2b,e) whereas site 4 was lost due to amino acid variations (Figure 2f). For sites 2 and 3 the changes are larger, implicating β4 and β5 strands, β4−β5 loop, and some residue modifications (Figure 2d). Only Asn288 was conserved, Asn292 was shifted out of the site, and at positions 294 and 257 aspartates are replaced by a histidine and a lysine, respectively. Lys257 side chain occupies the classical calcium site 2. As a consequence the β4−β5 loop, containing Glu261, is shifted far from its canonical position leading to a large groove specific to the langerin structure.

Molecular Modeling of Langerin/Oligosaccharide Complexes. A docking approach recently developed for calcium-mediated protein−carbohydrate interactions (*26*) has been used for predicting the binding mode of oligosaccharides on langerin. It has been shown that langerin can bind HIV-1 and HIV-2 gp120 proteins in a mannan-inhibitable manner (*12*), supporting gp120 glycan recognition. Moreover, data publicly available from the Consortium for Functional Glycomics [www.functionalglycomics.org/glycomics/publicdata/selectedScreens.jsp] confirm the affinity for oligomannosides. Indeed, glycan array data using the Lg-ECD or the Lg-CRD (report to data, on the consortium Web site, from G. Zurawski group and A. Skerra group, respectively) indicate that the high-mannose type N glycan, abundantly expressed on gp120 (*34*), and its derivative oligosaccharide, αMan12αMan12αMan13Man, are among the best ligands for langerin. Consequently, two linear mannosidic fragments were considered, i.e., αMan12αMan12αManOMe (M12M12M) and αMan12αMan13αManOMe (M12M13M). The docking calculation was not limited to the calcium-binding site but included the whole CRD. When looking at the complete docking results, i.e., 100 solutions, three preferred binding regions are identified at the protein surface, for both trisaccharides, as displayed in Figure 3a for M12M12M. The S1 site corresponds to the calcium-binding site but also extends for a few nanometers, indicating that longer oligosaccharides could interact. The S2 binding site corresponds to the groove created by the unusually opened β4−β5 loop (Figure 3b). This very hydrophilic region, corresponding to calcium sites in most C-type lectins, contains many residues

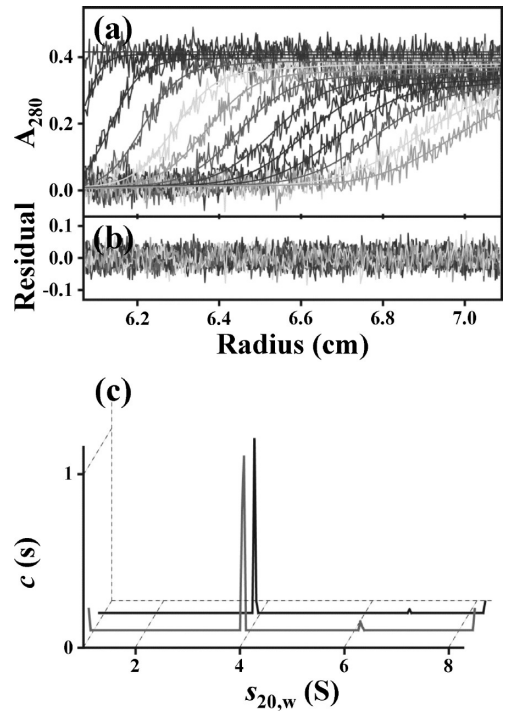


FIGURE 4: Langerin ECD sedimentation velocity analysis. (a) Superimposition of selected experimental and modeled profiles obtained at 20 °C, in 3 mm path-length cell, for up to 6 h at 42000 rpm for langerin at 0.84 mg/mL; the analysis was done without regularization and using systematic noise subtraction, considering 200 particles between 1 and 8 S with a fitted frictional ratio of 1.6. (b) Related residuals. (c) $c(s)$ for langerin at 0.84 (gray curve) and 0.09 mg/mL (black curve); $c(s)$ scale is normalized to a maximum value of 1 and shifted for the latter.

favorable to carbohydrate binding (asparagine, histidine, tryptophan) and has indeed been reported as a secondary mannose-binding site in the recent structure of langerin CRD (*15*). A third binding region, labeled S3, is close to S2, albeit with less favorable binding energies for oligosaccharides.

The most populated clusters (23% for M12M12M and 67% for M12M13M) correspond to the classical conformation with the central mannose residue coordinated to the calcium through its O3 and O4 hydroxyls (Figure 3c,d; see also Supporting Information Tables 1−3). The orientation of this residue is similar to the one observed for mannose alone in langerin (*15*) and mannose-binding proteins (*35, 36*). The nonreducing α1−2-linked mannose fits in a pocket adjacent to the main binding site and establishes hydrogen bonds with Asn297 and Lys313 (Figure 3c−e). Additional stabilization is provided by van der Waals contact between H−C4 and the Phe315 aromatic ring. On the contrary, the mannose on the reducing end does not make significant contribution to the binding, independently to its linkage with the calcium-bridged one (1−2 or 1−3 linkage).

Langerin ECD Hydrodynamic Studies. Stokes radius of langerin extracellular domain was estimated by migration over a Superose 12 size exclusion column, and an experimental Stokes radius of 5.13 nm was obtained (Table 2). This R_S value would correspond to a molecular mass of 243 kDa for a globular hydrated protein. The discrepancy with the theoretical mass of 88.2 kDa for the langerin ECD organized as a trimer, as previously described (*9*), strongly suggests a very elongated structure for langerin ECD.

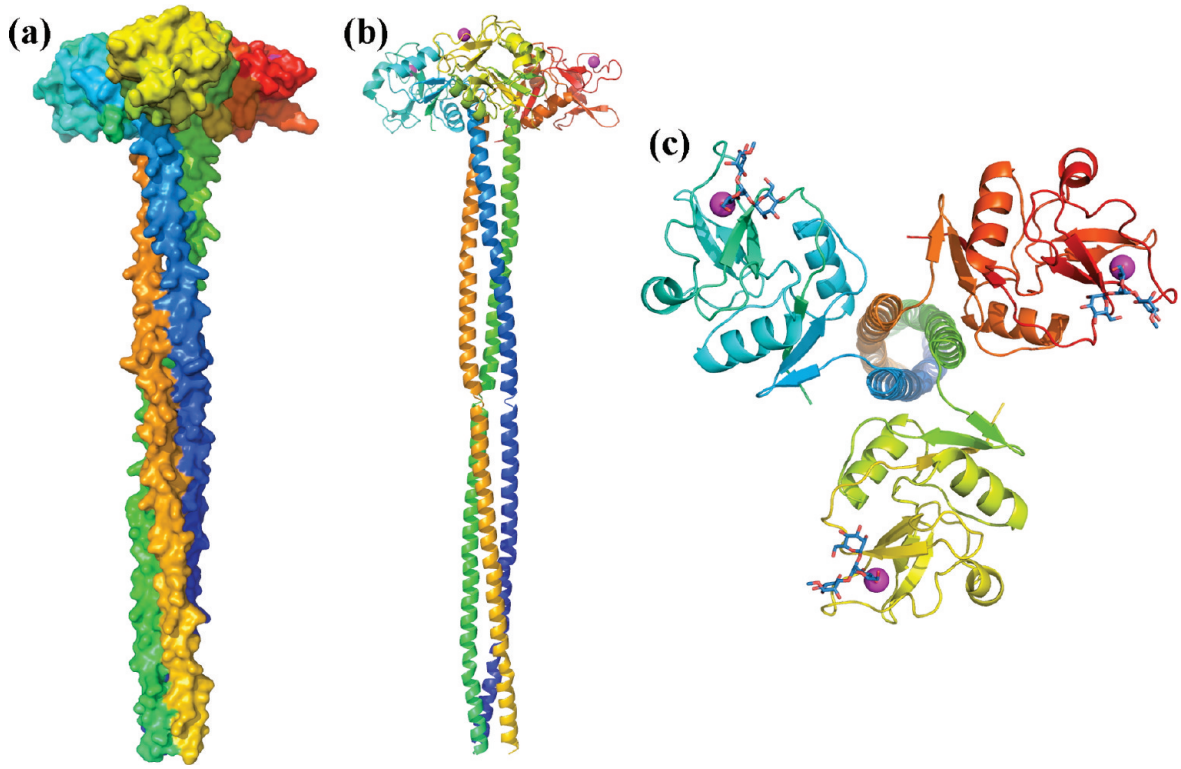


FIGURE 5: Langerin ECD trimer model. (a) Surface side view. (b) Cartoon side view. (c) Cartoon top view. Color code: rainbow colored langerin, magenta colored Ca^{2+} atom. The trimannoside M12M12M, sky blue colored, was not included in the modeling of the Lg-ECD but has been added in the figure to help visualize the interaction site of the protein.

Analytical ultracentrifugation sedimentation velocity experiments (Figure 4) were performed to confirm the association state and the Stokes radius of langerin. For samples at 0.84 and 0.09 mg/mL, the $c(s)$ analysis showed an essentially homogeneous sample with a species at $s_{20,w}$ = 4.03 S. The two data sets were thus globally analyzed with the program SEDPHAT, which provided an estimation of $s_{20,w}$ = 4.07 S and of the molecular mass = 83 kDa (the same value was obtained starting the fit with the hypothesis of a dimer of 58 kDa or a trimer of 88 kDa) in agreement with a trimer in solution. Combining the value of $s_{20,w}$ = 4.05 S with the mass of the trimer provided a Stokes radius value, R_S, of 5.1 nm as obtained with size exclusion chromatography (Table 2). The related frictional ratio of 1.7 corresponds to a very anisotropic shape, values for globular compact proteins being typically 1.25 ± 0.05. A strong correlation is thus observed between both size exclusion chromatography and velocity experiment analysis and suggests a very elongated shape for the langerin ECD.

Langerin ECD Structure Modeling. The neck region, involved in the trimerization, is a classical α-helix coiled-coil region as confirmed by sequence analysis using the COILS program (www.ch.embnet.org/software/COILS_form.html (*37*), data not shown) and by previously reported circular dichroïsm spectroscopy studies on langerin (*9*). Parts of influenza virus hemagglutinin (1QU1) and MBP-A trimeric structure (1RTM) have been used as templates for molecular modeling of trimeric ectodomain of langerin as described in Experimental Procedures. The resulting model is presented in Figure 5. The molecule is fairly elongated with the long axis extending over 20 nm. The coordinates were used as an input file in the HYDROPRO program (*38*) to calculate a theoretical Stokes radius. A value of R_S = 5.3 nm was obtained. This is in good agreement with the experimental values determined by size exclusion chromatography and sedimentation velocity (Table 2). It is also remarkable that the calculated frictional ratio is in good agreement with experiment. These strong correlations strengthen confidence in the very elongated langerin ECD structural model. The fact that the calculated frictional ratio is slightly larger than the experimental one (1.8 instead of 1.73) can be easily rationalized by the existence of flexibility of the elongated molecule, which is not taken into account in the rigid model.

Birbeck Granule Formation and Membrane Remodeling: Model of Langerin Extracellular Domain Organization. In order to propose a working model for langerin ECD contribution to BG organization, we observed numerous sections of freshly isolated Langerhans cells. As previously described, all BGs exhibit electron-dense paracrystalline structures in the center of a membrane sandwich (*39*), hereafter termed CMS for "cytomembrane sandwiching" structures. From Figure 6a, no electron-dense paracrystalline or even individual structure is visible on membranes adjacent to the BG formation site. This suggests that appearance of such structures is dependent on symmetrical elements, emerging from facing membranes. The association of these elements (involving an external ligand or not) could result in a bigger central object. Such higher order structure would then be visible, in this negative staining mode, in the center of the sandwiched membrane. Depending on the BG sections and on the membrane limits used for the estimation, the width of the BG zippered membranes can be evaluated in a range of 25−40 nm (Figure 6b).

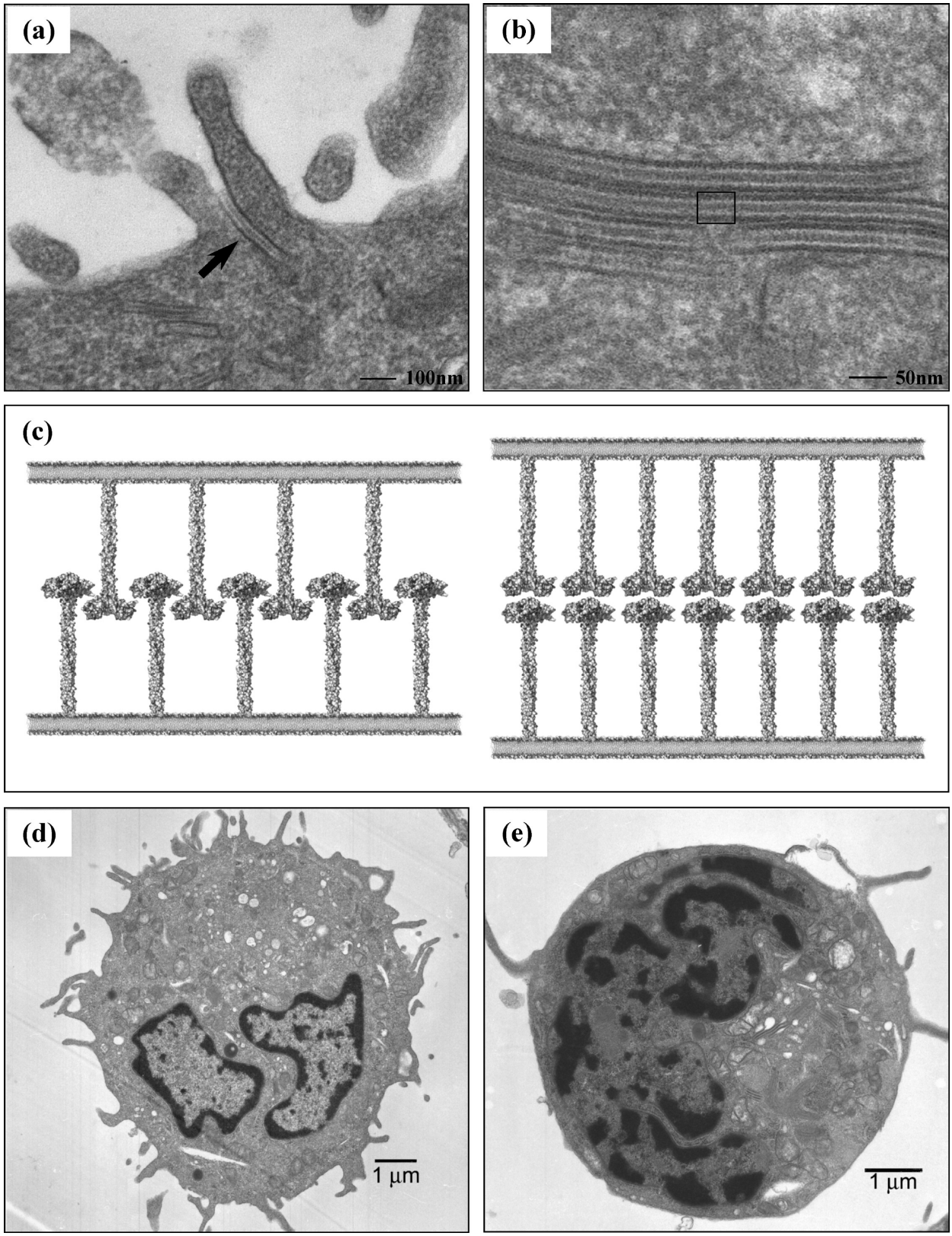


FIGURE 6: Birbeck granules in Langerhans cells and proposed macromolecular organization models. (a) Plasmic membrane invagination forming a cytomembrane sandwiching structure (CMS) (arrow) in a fresh isolated LC. (b) Stacked Birbeck granules of LC. (c) Two proposed Birbeck granule macromolecular organization models. (d, e) Langerin triggering inducing LC remodeling and Birbeck granule formation. LCs were incubated with mAb DCGM4 (e) or with anti-CD1a as control (d). A clear induction of intracellular BGs is observed in the pericentriolar region of DCGM4 treated cells, the dendritic cell shape is lost, and no dendrites can be observed.

From these various observations and from the langerin architecture deduced from sequence analysis, modeling, and hydrodynamic studies, we propose putative working models for the contribution of the langerin C-type lectin to the ultrastructural organization of the BG (Figure 6c). In such models, the CRDs displayed by facing membranes could

correspond to the central electron-dense structure. The dimensions of this model are compatible with the distances estimated from electron microscopy images.

More strikingly, we observed on fresh isolated LC that CRD engagement by DCGM4 mAb induces accumulation of BGs in the pericentriolar region (Figure 6e and Supporting Information Figure 2b). No such effect was observed with an isotype control (data not shown) or anti-CD1a mAb (Figure 6d and Supporting Information Figure 2a). Moreover, this triggering of langerin CRD also induces morphological changes: LCs become rounded and lose their dendrites (Figure 6e) compared to anti-CD1a-treated LC (Figure 6d). Besides, mannan triggers the same effect as DCGM4 mAb (data not shown). Taken together, our results demonstrate also the crucial role of the langerin ECD in the vesicle machinery for BG formation and rearrangements of LC morphology.

Langerin Carbohydrate Recognition Domain Is Required for Birbeck Granule Formation. To further explore the role of different domains of langerin in LC membrane rearrangement and BG formation, we generated mutated and deleted forms of human langerin. One form, termed Lg ΔCRD, is devoid of the CRD but retains the neck domain, whereas the second one, termed Lg ΔCyto, was deleted in the intracytoplasmic domain. Finally, langerin contains an intracellular PXXP sequence potentially recruiting SH3 domain containing proteins; therefore, a mutated form was generated with a P23I mutation in this motif (Figure 7a). After transfection in COP5 cells, we analyzed the expression of the different langerin constructs using either mAb DCGM4 for the CRD of langerin or mAb Lag recognizing the intracellular domain (*32*). As shown in Figure 7a, flow cytometry analysis showed that similar transfection efficiency and expression were observed for all constructs. As expected, mAb DCGM4 and mAb Lag reactivity was observed only when the CRD domain or the cytosolic domains are present, respectively.

After transfection in COP5 cells, we also performed electron microscopy and observed that the CRD of langerin is strictly necessary for the formation of BG structure. Indeed, no BG was observed in Lg ΔCRD transfected cells (Figure 7b,c) while BG and BG-related structures (CMS) were obtain with Lg WT as already described (data not shown) (*2*). More strikingly, we observed CMS in Lg ΔCyto transfected cells (Figure 7d). In contrast to Lg WT transfected cells (*2*), these CMS were never connected to vesicles but associated with ribosome-like condensation evoking rough endoplasmic reticulum (Figure 7e). This suggests that the cytoplasmic domain is essential to allow langerin targeting to other compartments. Interestingly, transfection of langerin mutated in the proline-rich motif (WPREPPP) of the cytosolic domain did not impair formation of superimposed membrane structures, but the latter were frequently connected to multivesicular bodies (Figure 7f,g).

Taken together, those mutants further emphasize the role of langerin CRD in the establishment of the definitive paracrystalline ultrastructure of BG.

DISCUSSION

Langerin Structure in the C-Type Lectin Family and Carbohydrate-Binding Mode. The langerin CRD presents some differences with other known C-type lectin structures for calcium binding. Langerin has only one calcium-binding site, corresponding to the canonical carbohydrate-binding site, whereas DC-SIGN/DC-SIGNR and MBP bind three to four calcium ions, respectively. As illustrated in Figure 2c, site 1 ligands (residues highlighted in gray) and the positions of their lateral chain are strictly conserved (Figure 2e), maintaining calcium site 1 integrity. For site 4, only Gln304 is conserved (Figure 2f) whereas for sites 2 and 3 many environmental changes have an obvious impact on the calcium ion loss and allow the large movement of the $\beta4-\beta5$ loop (Figure 2d). Thus a large groove is formed which is specific to langerin and not observed in DC-SIGN/DC-SIGNR or MBP.

Differences are also observed in the calcium-dependent carbohydrate-binding sites (Figure 3e,f). The DC-SIGN-binding site is characterized by the presence of Phe313 that establishes stacking with sugar B face (*40, 41*). In langerin, Phe315 is not in an appropriate position to play the same role. The presence of two lysine residues in langerin site environment, Lys299 and Lys313, is also unique to this CRD. This basic character of the binding site could be correlated with the reported affinity of langerin toward sulfated oligosaccharides (*8*). Our structure is in good agreement with the structure previously reported this year (*15*).

Oligomerization state also influences carbohydrate binding through an avidity-based mechanism. Langerin coiled-coil trimeric association compensates the CRD low affinity by the presentation of three identical carbohydrate recognition domains. Hydrodynamics studies performed on recombinant langerin ECD illustrate the very anisotropic shape of the trimer. The ECD model generated in this work is in agreement with results of the hydrodynamic studies (Table 2) but also with the known characteristics of other lectins with coiled-coil-based multimerization. Indeed, it is very similar to the well-characterized MBP trimeric structure (*30*). Uncertainties mainly reside in the relative flexibility of the different CRD domains with respect to each other.

Our modeling approach predicts that langerin binds to terminal αMan12Man moiety of gp120 in a well-defined extended pocket close to the calcium-binding site (Figure 3e). Search for possible secondary sites results in the identification of an additional region that would be favorable for mannose/oligomannose binding. The S2 binding site, which only occurs in langerin, is of high interest and closely corresponds to the secondary carbohydrate-binding site observed in recently published langerin CRD structures (*15*) (Figure 3b). Furthermore, the Asp288 variant, a result of human langerin polymorphism, has been demonstrated to bind to the mannose column with lower affinity than the major Asn288 form (*42*) (see Figure 3b for Asn288 location in S2 site). In the structure with maltose (*15*) as well as in our oligomannose binding modeling in the S2 site, this Asn288 is potentially involved in carbohydrate interaction. This strengthens the role of this langerin-specific groove in the interaction with carbohydrate and the need for future characterization of this unusual carbohydrate-binding site.

Langerin, DC-SIGN, and HIV. From the initial report on the use of DC-SIGN by HIV as a Trojan horse to invade the host organism, several groups have also shown that DC-SIGN is used by many other pathogens during their corresponding infection process (*43−47*). Hence, generation of DC-SIGN ligands, that can be used to inhibit or explore the

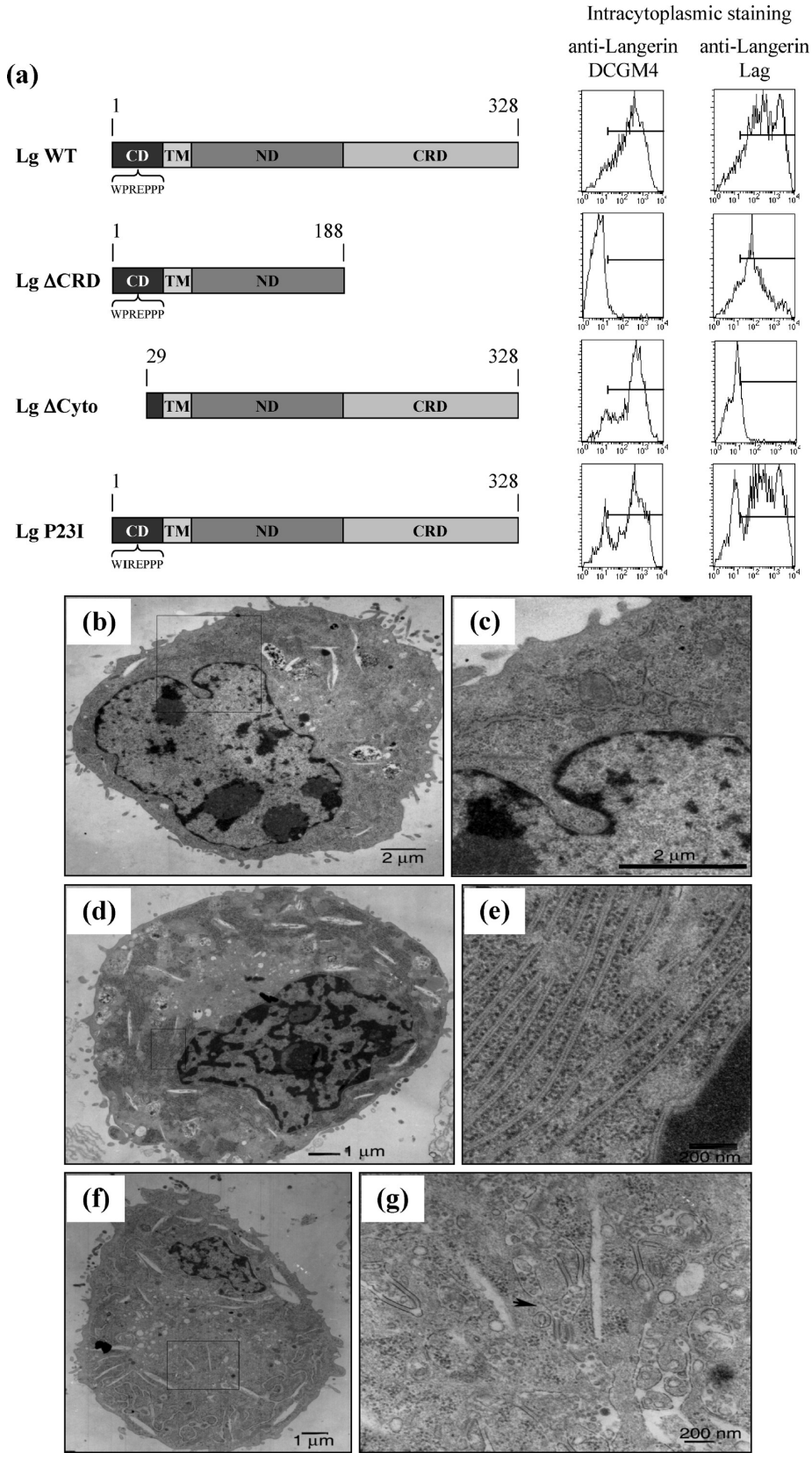


FIGURE 7: Electron microscopy (EM) of BG-related structures after transfection with different forms of human langerin. (a) Schematic representation of human langerin constructs and corresponding FACS staining on transfected COP5 cells using the anti-langerin mAbs DCGM4 and Lag following permeabilization. The Lg ΔCRD form lacks the entire carbohydrate recognition domain. The Lg ΔCyto form is deleted in the intracytoplasmic domain. The Lg P23I form contains a mutation of Pro23 into Ile in the proline-rich motif. (b−g) Electron microscopy performed on langerin COP5 transfected cells. Lg ΔCRD COP5 transfected cells shows no BG formation (b, c), whereas deletion of the intracytoplasmic domain (Lg ΔCyto) results in BG-related structures known as "cytomembrane sandwiching structures" (CMS) (d). These CMS are connected to abundant ribosomes evoking the rough endoplasmic reticulum (e). Mutation in the proline-rich motif (Lg P23I) does not allow induction of genuine BGs but rather multivesicular bodies (MVB, arrowhead) in continuity to CMS (f, g). Zoomed regions in (c), (e), and (g) correspond to squares in (b), (d), and (f), respectively.

role of DC-SIGN in pathogenesis and immunological process, is an important issue (*48*−*53*). The recent highlight on a divergent function between DC-SIGN and langerin with respect to HIV (*54*) might orient future efforts toward the design of molecules able to distinguish these two C-type lectin receptors. Thus, detailed characterization of the two binding sites is of central importance for such strategies. Panels e and f of Figure 3 present a comparison of langerin- and DC-SIGN-binding sites, respectively. In langerin, a well-defined pocket close to the calcium-binding site allows the accommodation of the terminal αMan12Man, found in high-mannose glycans (Figure 3e). DC-SIGN also binds αMan12Man with a similar orientation of the reducing mannose (*55*). However, the extended binding site is very different in the two lectins, and the bound disaccharides adopt different conformations. In DC-SIGN, the second mannose establishes a hydrogen bond with Ser360 and is partially stacked with Phe313. In langerin, Lys299 and Lys313 are directly involved in the interaction and add strong positive charges in this binding site. These two main differences, in topology and charge of the calcium-binding sites, constitute the first elements to envisage future drug design of selective compounds.

In addition to the differences in the two active sites, divergence in the oligomerization and in the binding site presentation may play a crucial role in the recognition. The occurrence of two sugar-binding sites on langerin CRD could participate in high avidity for glycosylated proteins. In both S1 and S2 sites, the oligosaccharide reducing ends are correctly oriented for binding terminal oligosaccharides from glycoconjugates. The distance between the two sites is over 20 Å, which is rather large for binding the two arms of the same biantennary high-mannose. Moreover, when considering the langerin trimeric model (Figure 5c), the estimated distance between two adjacent calcium-binding sites suggests that binding of adjacent CRDs necessarily occurs on different biantennary high-mannose. Nevertheless, gp120, which is the target of langerin on HIV (*12*), presents a shield of glycan on its exposed face (*56*) with many αMan12Man epitopes (*57*). Further characterization of DC-SIGN and langerin, in their oligomeric form, will be required to evaluate the contribution of this state with respect to their behavior toward HIV.

Langerin and Birbeck Granule Biogenesis. Although BGs have been first observed in 1961 (*4*), there is still a large part of mystery around their organization and function in the Langerhans cells. Nevertheless, BGs were described as organelles allowing a nonclassical routing for an antigen-processing pathway in the Langerhans cells (*11, 58*).

More than a simple association, a direct role of langerin in BG formation was demonstrated from observation of BG induction in murine fibroblasts (*2*) as well as in human melanoma cell line (*5*) upon langerin gene transfection. Langerin extracellular region, and more precisely its CRD, was suggested to be a key element in the BG architecture. Using Lg ΔCRD langerin, we observed that the lectin domain is strictly necessary for BG formation. This is consistent with the observation that Ca^{2+} removal can cause unzipping to various extents with inner periodical pattern disintegration (*59, 60*). Furthermore, langerin CRD point mutation W264R induces tubular-like structures different from characteristic structural features of BGs (*61, 62*). This residue is located in the middle of a large hydrophobic cluster in the CRD close to the calcium-binding site. The paracrystalline structure observed in the linear inner lamella of BG could be CRDs, possibly with associated ligands, engaged in a supramolecular organization. The observation of an induction of BG formation, upon triggering of langerin CRD with a specific mAb (DCGM4), is an additional strong argument to this scheme.

The coiled-coil neck regions allow terminal CRD presentation at rather long distances from the membrane surface. In BGs, the central electron-dense lamellar structures present spatial dimensions that are in good correlation with the size of our langerin ECD model (Figure 6b,c). Of course, these latter models are at this stage solely a "working hypothesis". Future experiments will be necessary to affine the dimension of this architecture and to investigate langerin ECD organization in such structures. The requirement of a ligand binding to initiate these high molecular weight assemblies will also need to be demonstrated.

Finally, we observed that transfection of Lg ΔCyto triggers the folding of membranes associated with ribosome-like condensation and evoking rough endoplasmic reticulum. We hypothesize that this deleted langerin mRNA is highly translated, due to the SV40 promoter, resulting in high langerin protein levels but without the correct addressing signal. Indeed, less drastic modification, the P23I mutation in the polyproline motif of the cytoplasmic domain, allows formation of CMS structures that are connected to a multivesicular body instead of genuine BG.

These latter observations demonstrate that the extracellular domain, and more particularly the CRD, is the key player in the membrane zippering as CMS structure. However, emphasis is also brought to the cytosolic domain playing a central role in membrane targeting and BG formation. The intracytoplasmic PXXP sequence is a putative SH3 binding motif, which is found in a number of molecules involved in a wide variety of biological processes (kinases regulation, subcellular localization, signaling pathways, etc.). Indeed, the altered distribution of the CMS structures, resulting from the P23I mutation, suggests the involvement of adaptor proteins. The association of Lg P23I construct with multivesicular bodies, which are protein-sorting stations of the endocytic pathway (*63*), is in accordance with previous studies. Indeed, Mc Dermott et al. have shown that BG can be formed from recycling endosomes (*64*). Both Lg ΔCyto and Lg P23I allow the formation of CMS structure but alter, to different extents, the global localization or organization of these CMS structures as genuine BG further emphasizing a role of the cytoplasmic region in the final organization of BG.

Membrane Deformation Mechanisms. Membrane deformations, for internalization as well as for budding, are crucial for many cell functions. Mechanisms leading to HIV internalization in Langerhans cell involve BG formation, simultaneously to HIV binding, through membrane deformation, invagination, and zipping. We show here that such membrane remodeling, BG formation, and even a global change of LC morphology can be induced (Figure 6) by triggering langerin with mAb DCGM4 which is specific for a CRD epitope overlapping with the calcium-dependent carbohydrate-binding site (*2*). This suggests that glycoconjugate ligands specific for the langerin CRD could be at the onset of the BG formation. Dynamic membrane remodeling

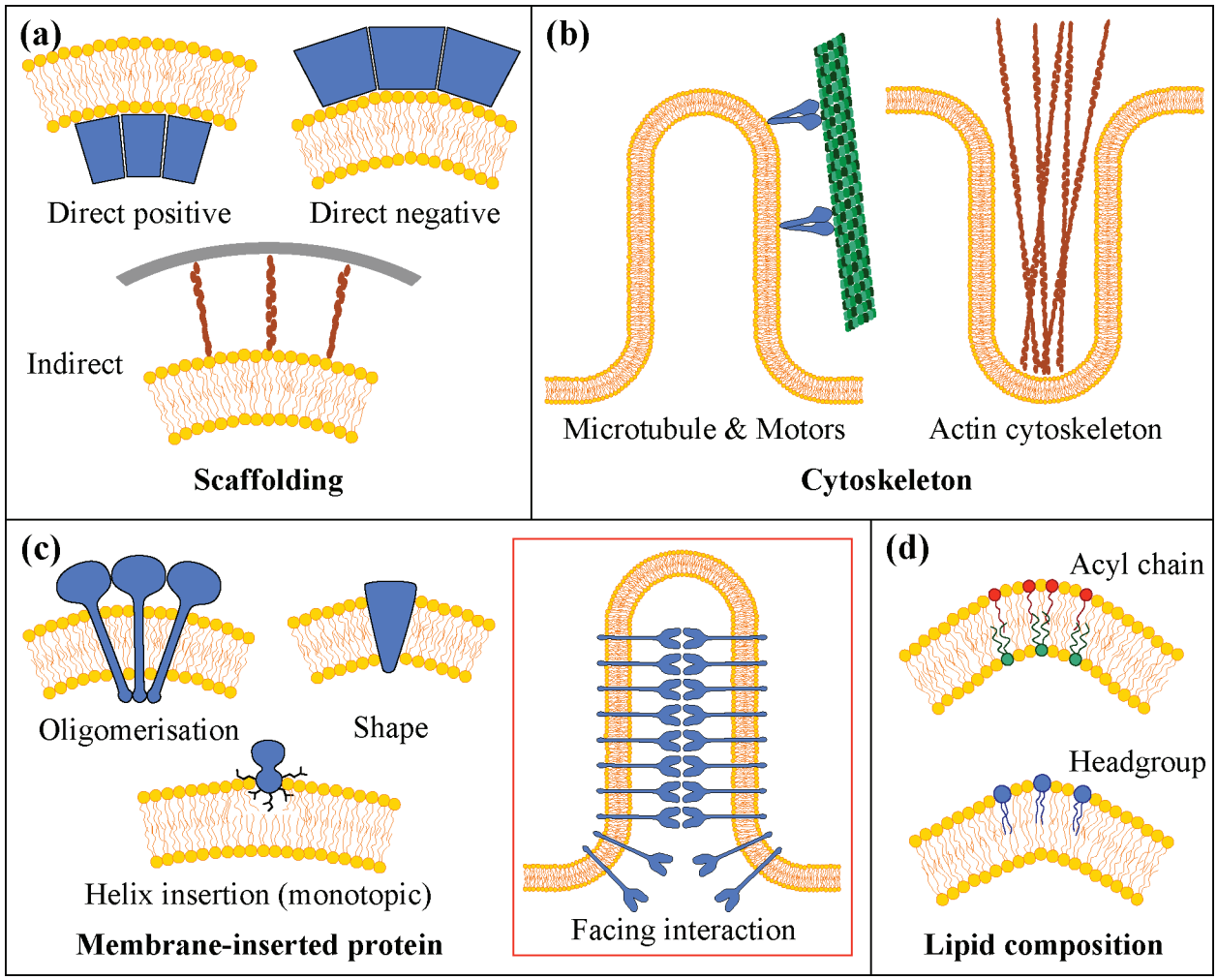


FIGURE 8: Reported membrane deformation mechanisms. (a) Scaffolding caused deformations. (b) Cytoskeleton caused deformations. (c) Membrane protein caused deformations. Red square: model of membrane zipping proposed for Birbeck granule formation. (d) Lipid composition caused deformations.

is classically performed by the interplay between lipids and protein. McMahon and Gallop have reported five general modes of membrane bending: lipid composition, cytoskeletal dependent membrane modeling, scaffolding by peripheral membrane proteins, amphipathic α-helix insertion, and finally bending resulting from the insertion of proteins in the membrane bilayer (see Figure 8, adapted from their review) (*65*). The membrane zipping arising from molecular interactions involving facing langerin receptors induces the deformation by a novel mechanism (Figure 8c, red square).

The requirement of the lectin properties of the langerin, the involvement of specific oligosaccharides and/or glycoconjugates in this membrane zipping, is still to be confirmed. However, the relative importance of the CRD, the coiled-coil region, and the cytosolic domain of langerin in membrane targeting, assembly, and stabilization of such structures has been underlined in this work for the first time.

ACKNOWLEDGMENT

We thank Dr. K. Yoneda for kindly providing anti-Lag antibody. We also thank M. P. Strub for support with selenocysteine/selenomethionine double labeling.

SUPPORTING INFORMATION AVAILABLE

Elution profiles of Langerin CRD and ECD from a mannan agarose column, zoomed sections of Figure 6d,e, and molecular modeling data. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Bell, D., Young, J. W., and Banchereau, J. (1999) Dendritic cells. *Adv. Immunol. 72*, 255–324.
2. Valladeau, J., Ravel, O., Dezutter-Dambuyant, C., Moore, K., Kleijmeer, M., Liu, Y., Duvert-Frances, V., Vincent, C., Schmitt, D., Davoust, J., Caux, C., Lebecque, S., and Saeland, S. (2000) Langerin, a novel C-type lectin specific to Langerhans cells, is an endocytic receptor that induces the formation of Birbeck granules. *Immunity 12*, 71–81.
3. Geijtenbeek, T. B., Kwon, D. S., Torensma, R., van Vliet, S. J., van Duijnhoven, G. C., Middel, J., Cornelissen, I. L., Nottet, H. S., KewalRamani, V. N., Littman, D. R., Figdor, C. G., and van Kooyk, Y. (2000) DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. *Cell 100*, 587–597.
4. Birbeck, M. S., Breathnach, A. S., and Everall, J. D. (1961) An electron microscope study of basal melanocytes and high-level clear cells (Langerhans cells) in vitiligo. *J. Invest. Dermatol. 37*, 51–63.
5. McDermott, R., Bausinger, H., Fricker, D., Spehner, D., Proamer, F., Lipsker, D., Cazenave, J. P., Goud, B., De La Salle, H.,

Salamero, J., and Hanau, D. (2004) Reproduction of Langerin/CD207 traffic and Birbeck granule formation in a human cell line model. *J. Invest. Dermatol. 123*, 72–77.
6. Valladeau, J., Clair-Moninot, V., Dezutter-Dambuyant, C., Pin, J. J., Kissenpfennig, A., Mattei, M. G., Ait-Yahia, S., Bates, E. E., Malissen, B., Koch, F., Fossiez, F., Romani, N., Lebecque, S., and Saeland, S. (2002) Identification of mouse langerin/CD207 in Langerhans cells and some dendritic cells of lymphoid tissues. *J. Immunol. 168*, 782–792.
7. Takahara, K., Omatsu, Y., Yashima, Y., Maeda, Y., Tanaka, S., Iyoda, T., Clausen, B. E., Matsubara, K., Letterio, J., Steinman, R. M., Matsuda, Y., and Inaba, K. (2002) Identification and expression of mouse Langerin (CD207) in dendritic cells. *Int. Immunol. 14*, 433–444.
8. Galustian, C., Park, C. G., Chai, W., Kiso, M., Bruening, S. A., Kang, Y. S., Steinman, R. M., and Feizi, T. (2004) High and low affinity carbohydrate ligands revealed for murine SIGN-R1 by carbohydrate array and cell binding approaches, and differing specificities for SIGN-R3 and langerin. *Int. Immunol. 16*, 853–866.
9. Stambach, N. S., and Taylor, M. E. (2003) Characterization of carbohydrate recognition by langerin, a C-type lectin of Langerhans cells. *Glycobiology 13*, 401–410.
10. Kissenpfennig, A., Ait-Yahia, S., Clair-Moninot, V., Stossel, H., Badell, E., Bordat, Y., Pooley, J. L., Lang, T., Prina, E., Coste, I., Gresser, O., Renno, T., Winter, N., Milon, G., Shortman, K., Romani, N., Lebecque, S., Malissen, B., Saeland, S., and Douillard, P. (2005) Disruption of the langerin/CD207 gene abolishes Birbeck granules without a marked loss of Langerhans cell function. *Mol. Cell. Biol. 25*, 88–99.
11. Hunger, R. E., Sieling, P. A., Ochoa, M. T., Sugaya, M., Burdick, A. E., Rea, T. H., Brennan, P. J., Belisle, J. T., Blauvelt, A., Porcelli, S. A., and Modlin, R. L. (2004) Langerhans cells utilize CD1a and langerin to efficiently present nonpeptide antigens to T cells. *J. Clin. Invest. 113*, 701–708.
12. Turville, S. G., Cameron, P. U., Handley, A., Lin, G., Pohlmann, S., Doms, R. W., and Cunningham, A. L. (2002) Diversity of receptors binding HIV on dendritic cell subsets. *Nat. Immunol. 3*, 975–983.
13. Kawamura, T., Kurtz, S. E., Blauvelt, A., and Shimada, S. (2005) The role of Langerhans cells in the sexual transmission of HIV. *J. Dermatol. Sci. 40*, 147–155.
14. de Witte, L., Nabatov, A., Pion, M., Fluitsma, D., de Jong, M. A., de Gruijl, T., Piguet, V., van Kooyk, Y., and Geijtenbeek, T. B. (2007) Langerin is a natural barrier to HIV-1 transmission by Langerhans cells. *Nat. Med. 13*, 367–371.
15. Chatwell, L., Holla, A., Kaufer, B. B., and Skerra, A. (1981) (2008) The carbohydrate recognition domain of Langerin reveals high structural similarity with the one of DC-SIGN but an additional, calcium-independent sugar-binding site. *Mol. Immunol. 45*, 1994.
16. Thépaut, M., Vivès, C., Pompidor, G., Kahn, R., and Fieschi, F. (2008) Overproduction, purification and preliminary crystallographic analysis of the carbohydrate recognition domain of human langerin. *Acta Crystallogr. F 64*, 115–118.
17. Muller, S., Senn, H., Gsell, B., Vetter, W., Baron, C., and Bock, A. (1994) The formation of diselenide bridges in proteins by incorporation of selenocysteine residues: biosynthesis and characterization of (Se)2-thioredoxin. *Biochemistry 33*, 3404–3412.
18. Strub, M. P., Hoh, F., Sanchez, J. F., Strub, J. M., Bock, A., Aumelas, A., and Dumas, C. (2003) Selenomethionine and selenocysteine double labeling strategy for crystallographic phasing. *Structure 11*, 1359–1367.
19. Kabsch, W. (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr. 26*, 795–800.
20. Collaborative Computational Project, No. (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D 50*, 760–763.
21. Vonrhein, C., Blanc, E., Roversi, P., and Bricogne, G. (2006) *Crystallographic Methods*, Humana Press, Totowa, NJ.
22. Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 60*, 2126–2132.
23. Schuck, P. (2000) Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J. 78*, 1606–1619.
24. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W., Belew, R. K., and Olson, A. J. (1998) Automated docking using a Lamarckian genetic algorithm and and empirical binding free energy function. *J. Comput. Chem. 19*, 1639–1662.
25. Clark, M., Cramer, R. D. I., and van den Opdenbosch, N. (1989) Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem. 10*, 982–1012.
26. Nurisso, A., Kozmon, S., and Imberty, A. (2008) 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. 34*, 469–479.
27. Imberty, A., Bettler, E., Karababa, M., Mazeau, K., Petrova, P., and Pérez, S. (1999) in *Perspectives in Structural Biology* (Vijayan, M., Yathindra N., and Kolaskar A. S., Eds.) pp 392−409, Indian Academy of Sciences and Universities Press, Hyderabad.
28. Blundell, T., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L., and Sutcliffe, M. (1988) 18th Sir Hans Krebs lecture. Knowledge-based protein modelling and design. *Eur. J. Biochem. 172*, 513–520.
29. Chen, J., Skehel, J. J., and Wiley, D. C. (1999) N- and C-terminal residues combine in the fusion-pH influenza hemagglutinin HA(2) subunit to form an N cap that terminates the triple-stranded coiled coil. *Proc. Natl. Acad. Sci. U.S.A. 96*, 8967–8972.
30. Weis, W. I., and Drickamer, K. (1994) Trimeric structure of a C-type mannose-binding protein. *Structure 2*, 1227–1240.
31. Valladeau, J., Duvert-Frances, V., Pin, J. J., Dezutter-Dambuyant, C., Vincent, C., Massacrier, C., Vincent, J., Yoneda, K., Banchereau, J., Caux, C., Davoust, J., and Saeland, S. (1999) The monoclonal antibody DCGM4 recognizes Langerin, a protein specific of Langerhans cells, and is rapidly internalized from the cell surface. *Eur. J. Immunol. 29*, 2695–2704.
32. Kashihara, M., Ueda, M., Horiguchi, Y., Furukawa, F., Hanaoka, M., and Imamura, S. (1986) A monoclonal antibody specifically reactive to human Langerhans cells. *J. Invest. Dermatol. 87*, 602–607.
33. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr. D 26*, 283–291.
34. Balzarini, J. (2007) Targeting the glycans of glycoproteins: a novel paradigm for antiviral therapy. *Nat. Rev. Microbiol. 5*, 583–597.
35. Ng, K. K., Drickamer, K., and Weis, W. I. (1996) Structural analysis of monosaccharide recognition by rat liver mannose-binding protein. *J. Biol. Chem. 271*, 663–674.
36. Ng, K. K., Kolatkar, A. R., Park-Snyder, S., Feinberg, H., Clark, D. A., Drickamer, K., and Weis, W. I. (2002) Orientation of bound ligands in mannose-binding proteins. Implications for multivalent ligand recognition. *J. Biol. Chem. 277*, 16088–16095.
37. Lupas, A., Van Dyke, M., and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science 252*, 1162–1164.
38. Garcia De La Torre, J., Huertas, M. L., and Carrasco, B. (2000) Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys. J. 78*, 719–730.
39. Sagebiel, R. W., and Reed, T. H. (1968) Serial reconstruction of the characteristic granule of the Langerhans cell. *J. Cell Biol. 36*, 595–602.
40. Drickamer, K. (1999) C-type lectin-like domains. *Curr. Opin. Struct. Biol. 9*, 585–590.
41. Feinberg, H., Mitchell, D. A., Drickamer, K., and Weis, W. I. (2001) Structural basis for selective recognition of oligosaccharides by DC-SIGN and DC-SIGNR. *Science 294*, 2163–2166.
42. Ward, E. M., Stambach, N. S., Drickamer, K., and Taylor, M. E. (2006) Polymorphisms in human langerin affect stability and sugar binding activity. *J. Biol. Chem. 281*, 15450–15456.
43. Lozach, P. Y., Lortat-Jacob, H., de Lacroix de Lavalette, A., Staropoli, I., Foung, S., Amara, A., Houles, C., Fieschi, F., Schwartz, O., Virelizier, J. L., Arenzana-Seisdedos, F., and Altmeyer, R. (2003) DC-SIGN and L-SIGN are high affinity binding receptors for hepatitis C virus glycoprotein E2. *J. Biol. Chem. 278*, 20358–20366.
44. Navarro-Sanchez, E., Altmeyer, R., Amara, A., Schwartz, O., Fieschi, F., Virelizier, J. L., Arenzana-Seisdedos, F., and Despres, P. (2003) Dendritic-cell-specific ICAM3-grabbing non-integrin is essential for the productive infection of human dendritic cells by mosquito-cell-derived dengue viruses. *EMBO Rep. 4*, 723–728.
45. Alvarez, C. P., Lasala, F., Carrillo, J., Muniz, O., Corbi, A. L., and Delgado, R. (2002) C-type lectins DC-SIGN and L-SIGN mediate cellular entry by Ebola virus in cis and in trans. *J. Virol. 76*, 6841–6844.
46. Geijtenbeek, T. B., Van Vliet, S. J., Koppel, E. A., Sanchez-Hernandez, M., Vandenbroucke-Grauls, C. M., Appelmelk, B., and

Van Kooyk, Y. (2003) Mycobacteria target DC-SIGN to suppress dendritic cell function. *J. Exp. Med. 197*, 7–17.
47. van Die, I., van Vliet, S. J., Nyame, A. K., Cummings, R. D., Bank, C. M., Appelmelk, B., Geijtenbeek, T. B., and van Kooyk, Y. (2003) The dendritic cell-specific C-type lectin DC-SIGN is a receptor for *Schistosoma mansoni* egg antigens and recognizes the glycan antigen Lewis x. *Glycobiology 13*, 471–478.
48. Frison, N., Taylor, M. E., Soilleux, E., Bousser, M. T., Mayer, R., Monsigny, M., Drickamer, K., and Roche, A. C. (2003) Oligolysine-based oligosaccharide clusters: selective recognition and endocytosis by the mannose receptor and dendritic cell-specific intercellular adhesion molecule 3 (ICAM-3)-grabbing nonintegrin. *J. Biol. Chem. 278*, 23922–23929.
49. Tabarani, G., Reina, J. J., Ebel, C., Vives, C., Lortat-Jacob, H., Rojo, J., and Fieschi, F. (2006) Mannose hyperbranched dendritic polymers interact with clustered organization of DC-SIGN and inhibit gp120 binding. *FEBS Lett. 580*, 2402–2408.
50. Borrok, M. J., and Kiessling, L. L. (2007) Non-carbohydrate inhibitors of the lectin DC-SIGN. *J. Am. Chem. Soc. 129*, 12780–12785.
51. Reina, J. J., Sattin, S., Invernizzi, D., Mari, S., Martinez-Prats, L., Tabarani, G., Fieschi, F., Delgado, R., Nieto, P. M., Rojo, J., and Bernardi, A. (2007) 1,2-Mannobioside mimic: Synthesis, DC-SIGN interaction by NMR and docking, and antiviral activity. *ChemMedChem 2*, 1030–1036.
52. Wang, S. K., Liang, P. H., Astronomo, R. D., Hsu, T. L., Hsieh, S. L., Burton, D. R., and Wong, C. H. (2008) Targeting the carbohydrates on HIV-1: Interaction of oligomannose dendrons with human monoclonal antibody 2G12 and DC-SIGN. *Proc. Natl. Acad. Sci. U.S.A. 105*, 3690–3695.
53. Timpano, G., Tabarani, G., Anderluh, M., Invernizzi, D., Vasile, F., Potenza, D., Nieto, P. M., Rojo, J., Fieschi, F., and Bernardi, A. (1921) (2008) Synthesis of novel DC-SIGN ligands with an alpha-fucosylamide anchor. *ChemBioChem 9*, 1930.
54. de Witte, L., Nabatov, A., and Geijtenbeek, T. B. (2008) Distinct roles for DC-SIGN+-dendritic cells and Langerhans cells in HIV-1 transmission. *Trends Mol. Med. 14*, 12–19.
55. Feinberg, H., Castelli, R., Drickamer, K., Seeberger, P. H., and Weis, W. I. (2007) Multiple modes of binding enhance the affinity of DC-SIGN for high mannose N-linked glycans found on viral glycoproteins. *J. Biol. Chem. 282*, 4202–4209.
56. McCaffrey, R. A., Saunders, C., Hensel, M., and Stamatatos, L. (2004) N-linked glycosylation of the V3 loop and the immunologically silent face of gp120 protects human immunodeficiency virus type 1 SF162 from neutralization by anti-gp120 and anti-gp41 antibodies. *J. Virol. 78*, 3279–3295.
57. Scanlan, C. N., Pantophlet, R., Wormald, M. R., Ollmann Saphire, E., Stanfield, R., Wilson, I. A., Katinger, H., Dwek, R. A., Rudd, P. M., and Burton, D. R. (2002) The broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2G12 recognizes a cluster of alpha1→2 mannose residues on the outer face of gp120. *J. Virol. 76*, 7306–7321.
58. Hanau, D., Fabre, M., Schmitt, D. A., Stampf, J. L., Garaud, J. C., Bieber, T., Grosshans, E., Benezra, C., and Cazenave, J. P. (1987) Human epidermal Langerhans cells internalize by receptor-mediated endocytosis T6 (CD1 "NA1/34") surface antigen. Birbeck granules are involved in the intracellular traffic of the T6 antigen. *J. Invest. Dermatol. 89*, 172–177.
59. Bartosik, J. (1992) Cytomembrane-derived Birbeck granules transport horseradish peroxidase to the endosomal compartment in the human Langerhans cells. *J. Invest. Dermatol. 99*, 53–58.
60. Andersson, L., Bartosik, J., Bendsoe, N., Malmström, A., Mikulowska, A., Warfvinge, K., Andersson, A., and Falk, B. (1988) in *The Langerhans Cell* (Thivolet, J., and Schmitt, D., Eds.) pp 185−191.
61. Verdijk, P., Dijkman, R., Plasmeijer, E. I., Mulder, A. A., Zoutman, W. H., Mieke Mommaas, A., and Tensen, C. P. (2005) A lack of Birbeck granules in Langerhans cells is associated with a naturally occurring point mutation in the human Langerin gene. *J. Invest. Dermatol. 124*, 714–717.
62. Mommaas, M., Mulder, A., Vermeer, B. J., and Koning, F. (1994) Functional human epidermal Langerhans cells that lack Birbeck granules. *J. Invest. Dermatol. 103*, 807–810.
63. Kleijmeer, M. J., Oorschot, V. M., and Geuze, H. J. (1994) Human resident langerhans cells display a lysosomal compartment enriched in MHC class II. *J. Invest. Dermatol. 103*, 516–523.
64. Mc Dermott, R., Ziylan, U., Spehner, D., Bausinger, H., Lipsker, D., Mommaas, M., Cazenave, J. P., Raposo, G., Goud, B., de la Salle, H., Salamero, J., and Hanau, D. (2002) Birbeck granules are subdomains of endosomal recycling compartment in human epidermal Langerhans cells, which form where Langerin accumulates. *Mol. Biol. Cell 13*, 317–335.
65. McMahon, H. T., and Gallop, J. L. (2005) Membrane curvature and mechanisms of dynamic cell membrane remodelling. *Nature 438*, 590–596.

BI802151W

PA-IL: a bacterial galactose-binding lectin

3.4 ARTICLES III-IV: *in silico* studies of *Pseudomonas aeruginosa* lectin I

3.4.1 Introduction

Pseudomonas aeruginosa is a ubiquitous bacterium usually found in nature in a biofilm, attached to surfaces in contact with soil, or in a planktonic form in water, as unicellular organism, actively floating by means of its polar flagella (Figure 3-5). This opportunistic bacterium can cause urinary tract infections, dermatitis, pancreatitis, bone and joint infections, gastrointestinal, respiratory and a variety of systemic infections in immune compromised patients[201]. In particular, it is considered the first cause of death for persons affected by cystic fibrosis, a genetic disease in which an abnormal protein (*CFTR*) causes an altered luminal ionic milieu that lead to the accumulation of unusual thick, sticky mucus that clogs the lungs and obstructs the pancreas[202]. *P.aeruginosa*, affects the respiratory tract of cystic fibrosis patients and can produce chronic-threatening lung infections and inflammation phenomena that lead to respiratory failure and death[203].

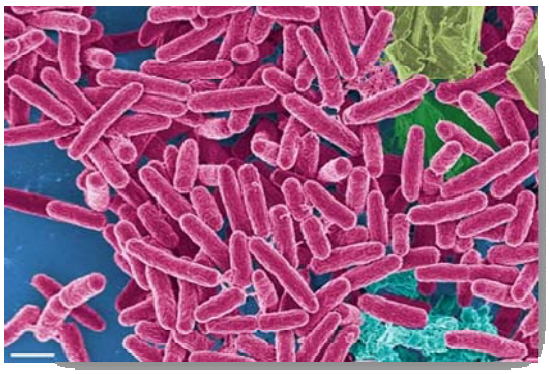


Figure 3-5 - Colonies of Pseudomonas aeruginosa under the microscope (www.waterscan.rs).

The pathogenic bacterium *P. aeruginosa* uses oligosaccharide-mediated recognition in order to adhere to the host epithelial surfaces, starting the bacterial infection and the production of a biofilm that is first cause of antibiotic resistance (Figure 3-6).

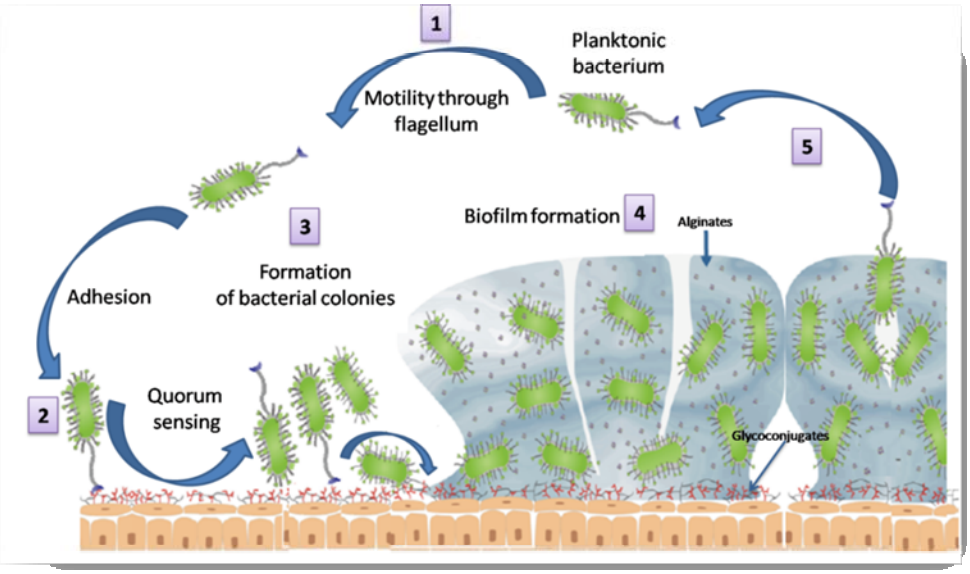


Figure 3-6 - Steps required for the biofilm formation. The bacterium P. aeruginosa*, in its planktonic form, is able to move by chemotaxis using its flagellum (1). Flagella and pili (type IV) allow the adhesion of the bacterium to cell surfaces through carbohydrate binding proteins that bind glycoconjugates (mucins) and asialo-GM1,GM2 glycolipids, respectively (2). The bacterium uses quorum sensing to coordinate the genes expression. It produces virulence factors and signals that induce the formation of bacterial colonies (3). Bacteria start to produce and accumulate exopolysaccharides (alginate), necessary for forming the biofilm (4). Successively trapped in the biofilm, bacteria communicate with each other and with the external environment via chemical signals. They can also abandon the biofilm, returning to their initial state for invading other cell surfaces (5). Adapted from*[204].

Carbohydrate-binding proteins are located on the bacterial flagella[205, 206] and pili[207]. In addition, the bacterium produces two soluble lectins, PA-IL and PA-IIL, found in its extracellular matrix or conjugated to the cell surface, that contribute to the adhesion and biofilm formation mechanisms[206]. The PA-IL and PA-IIL lectins are similar in size and both require the presence of calcium ions for binding carbohydrate. However, they do not exhibit se-

quence similarity (no sequence identity), showing different specificities towards rides[208].

PA1L was isolated in 1972[209] and its tetrameric structure solved for the first time in 2002[210]. Each PAIL monomer (12.75 kDa) contains one carbohydrate binding domain and one calcium ion in which both α and ß anomers of D-galactose are recognized ($Ka = 3.4x10^{-4}\ M^{-1}$, calculated by equilibrium dialysis[211]). It was observed that PA-1L is able to bind blood antigens from B and P groups, characterized by Galα1-3Gal, Galα1-4Gal terminal residues, respectively. Studies also revealed that the disaccharide Galα1-2Gal presents strong affinity to the lectin[212-215].

Solved crystal structures, in apo form and in complex with sugars, demonstrated that each monomer adopts a small jelly-roll type β-sandwich fold, consisting of two curved sheets, each one formed by four antiparallel β-strands with a calcium ion in the binding site[216, 217].

In the crystal structure of PA-IL co-crystallized with D-galactose, the formation of hydrogen bonds with the protein side chains involves the hydroxyl groups in the 2, 3, 4, 6 positions whereas hydroxyls occupying the 3 and 4 positions, also coordinate a calcium ion (Figure 3-7). PA-IL lectin is considered as one of the virulent factors of the bacterium *Pseudomonas aeruginosa* for its contribution to cell surfaces adhesion, biofilm formation and citotoxicity [218]. The production of this protein, encoded by the *lecA* gene, is under *quorum sensing* control. The bacterium releases specific *quorum-sensing* molecules, acylhomoserine lactones (AHL) that, in presence of the *RpoS* cofactor, regulate the *lecA* gene transcription[219]. It has been proposed that PA-IL binds these *quorum-sensing* signals in hydrophobic regions out to the classical carbohydrate binding sites, controlling their concentration but this role has yet to be confirmed[220, 221]. Previous works demonstrate that PA-IL is involved in adhesion processes, showing the capability to bind particular glycoconjugates that constitute extracellular matrix and mucus[222, 223]. Its role in stabilizing the biofilm structure was also proposed: *lecA* gene mutations or the direct inhibition of this protein lead to the formation of a biofilm with a loss of structural consistence[224]. The lectin is also known for being cytotoxic for respiratory epithelial cells, alone or in combination with other molecules like elastase and exotoxin A, disorganizing and destroying epithelial tissues[225, 226].

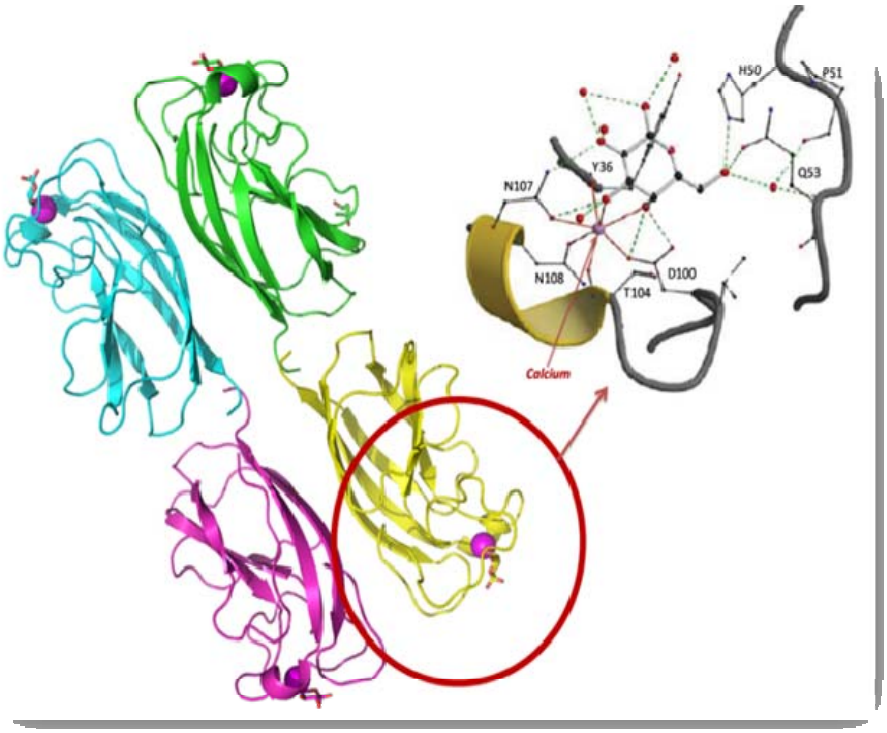


Figure 3-7 - Crystallographic structure of PA-1L in complex with D-galactose and, in the red circle, the carbohydrate binding site (PDB 1OKO).

PA-IIL, the second soluble and virulent lectin of *Pseudomonas aeruginosa*, was discovered in 1977[227]. It shows strong affinity for L-fucose (Ka = 1.6×10^6 M^{-1}, calculated by equilibrium dialysis) probably due to the presence of two calcium ions in the binding site[184, 228]. This lectin can also bind different sugars as D-fructose, D-mannose and also oligosaccharides like Lewis A (Galβ1-3(Fucα1-4)GlcNAc) in which GlcNAc residue plays a key role in the binding[228-230]. The PA-IIL crystal structure shows a tetrameric form, each monomer being of 11.73 KDa mass and consisting of nine-stranded antiparallel ß sandwich. Strong interactions between two adiacent monomers occur through head-to -tail association, resulting in the involvement of the C-terminal carboxyl group, in particular a specific glycine residue, of

each monomer in the ligand binding site of the other monomer. The two metals in the binding site interact with amino acids polar residues and sugar hydroxyl groups[228] (Figure 3-8). As for the lectin PA-1L, the production of PA-IIL is regulated by *quorum sensing*[219]. The localization of this lectin on the *Pseudomonas aeruginosa* outer membrane, combined with *in vivo* essays, demonstrates its fundamental role in glycoconjugates recognition, adhesion and biofilm construction[208, 218] PA-IIL is not cytotoxic but it can reduce mucociliary clearance of endhotelial cells[231].

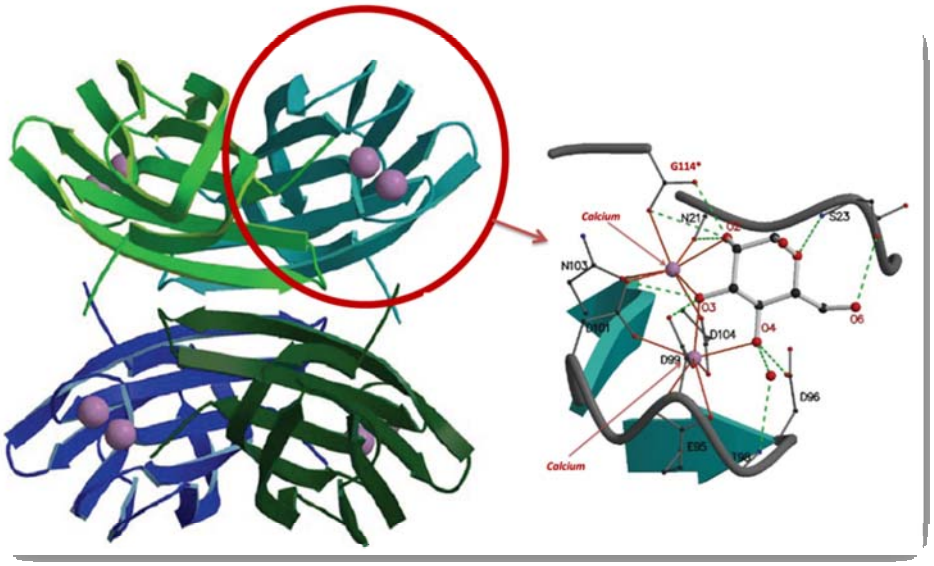


Figure 3-8 – Crystallographic structure of PA-IIL (PDB 1OUX) and representation of the carbohydrate binding site in which the lectin is found in complex with L-fucose (PDB 1GZT).

3.4.2 Results

PA-1L is a calcium-dependent lectin whose mechanism of action can contribute to explain the virulence of the bacterium *Pseudomonas aeruginosa* in developing chronic respiratory diseases in cystic fibrosis patients. In the first article, conformational searches applied on

the oligosaccharides moieties of glycosphingolipids, followed by docking calculations and MDs simulations helped in determining the structural properties of this protein in complex with its natural ligands, supporting the experimental data obtained by cell surface labeling, glycan array analysis, titration microcalorimetry and crystallography. The globotriaosylceramide antigen Gb3, whose carbohydrate moiety is αGal1-4βGal1-4βGlc, was proposed to be a likely natural ligand of the lectin on human epithelia.

The second article describes the interactions between the lectin PA-IL and three digalactoside substrates. Conformational analysis, docking calculations and MDs simulations were used to determine structural and dynamic features whereas quantitative measurements of binding parameters were performed using microcalorimetry and compared to data obtained by MDs simulations combined with MM-PBSA analysis. A bridging water molecule identified in the crystal structure was demonstrated of being important for the binding. This molecule, if included in calculations, led to the prediction of the correct affinity trend observed experimentally.

All the data presented in these manuscripts might specifically influence the possible design of PA-IL inhibitors.

doi:10.1016/j.jmb.2008.08.028 *J. Mol. Biol.* (2008) **383**, 837–853

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Structural Basis of the Preferential Binding for Globo-Series Glycosphingolipids Displayed by *Pseudomonas aeruginosa* Lectin I

Bertrand Blanchard[1]†, Alessandra Nurisso[1]†, Emilie Hollville[2], Cécile Tétaud[2], Joelle Wiels[2], Martina Pokorná[3,4], Michaela Wimmerová[3,4], Annabelle Varrot[1] and Anne Imberty[1]*

[1]*CERMAV–CNRS (affiliated with Université Joseph Fourier and belonging to ICMG), BP53, F-38041 Grenoble, France*

[2]*UMR 8126, CNRS, Université Paris-Sud 11, Institut Gustave Roussy, 94805 Villejuif Cedex, France*

[3]*NCBR, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic*

[4]*Department of Biochemistry, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic*

Received 5 July 2008; received in revised form 11 August 2008; accepted 13 August 2008
Available online 22 August 2008

The opportunistic pathogen *Pseudomonas aeruginosa* contains several carbohydrate-binding proteins, among which is the *P. aeruginosa* lectin I (PA-IL), which displays affinity for α-galactosylated glycans. Glycan arrays were screened and demonstrated stronger binding of PA-IL toward αGal1–4βGal-terminating structures and weaker binding to αGal1–3βGal ones in order to determine which human glycoconjugates could play a role in the carbohydrate-mediated adhesion of the bacteria. This was confirmed *in vivo* by testing the binding of the lectin to Burkitt lymphoma cells that present large amounts of globotriaosylceramide antigen Gb3/CD77/P^k. Trisaccharide moieties of Gb3 (αGal1–4βGal1–4Glc) and isoglobotriaosylceramide (αGal1–3βGal1–4Glc) were tested by titration microcalorimetry, and both displayed similar affinity to PA-IL in solution. The crystal structure of PA-IL complexed to αGal1–3βGal1–4Glc trisaccharide has been solved at 1.9-Å resolution and revealed how the second galactose residue makes specific contacts with the protein surface. Molecular modeling studies were performed in order to compare the binding mode of PA-IL toward αGal1–3Gal with that toward αGal1–4Gal. Docking studies demonstrated that αGal1–4Gal creates another network of contacts for achieving a very similar affinity, and 10-ns molecular dynamics in explicit water allowed for analyzing the flexibility of each disaccharide ligand in the protein binding site. The higher affinity observed for binding to Gb3 epitope, both *in vivo* and on glycan array, is likely related to the presentation effect of the oligosaccharide on a surface, since only the Gb3 glycosphingolipid geometry is fully compatible with parallel insertion of neighboring trisaccharide heads in two binding sites of the same tetramer of PA-IL.

Keywords: lectin; glycosphingolipid; *Pseudomonas aeruginosa*; adhesion; oligosaccharides

Edited by I. Wilson

Introduction

Pseudomonas aeruginosa is an opportunistic bacterium responsible for numerous nosocomial infections in immunocompromised patients for whom it may cause a wide number of diseases, such as septicaemia, urinary tract infections, pancreatitis, and dermatitis. The bacteria colonize patients with chronic lung diseases as well as those under mechanical ventilation, and these recurrent infections are often fatal for cystic fibrosis patients. *P. aeruginosa* produces a wide variety of carbohydrate-binding proteins, including the soluble lectins I (PA-

**Corresponding author.* E-mail address: imberty@cermav.cnrs.fr.

† B.B. and A.N. contributed equally to the work.

Abbreviations used: PA-IL, *Pseudomonas aeruginosa* lectin I; Gb3, globotriaosylceramide; K_a, association constant; TLC, thin-layer chromatography; MD, molecular dynamics; BL, Burkitt lymphoma; mAb, monoclonal antibody; LacCer, lactosylceramide; iGb3, isoglobotriaosylceramide; FITC, fluorescein isothiocyanate; LP, lower phase; UP, upper phase; BSA, bovine serum albumin; PBS, phosphate-buffered saline.

Table 1. Microcalorimetry data for the interaction between PA-IL and trisaccharides

Ligand	K_a (10^3 M^{-1})	K_d (μM)	$-\Delta G$ (kcal/mol)	$-\Delta H$ (kcal/mol)	$T\Delta S$ (kcal/mol)
αGal1–3βGal1–4Glc	15	68	5.7	9.1	3.4
αGal1–4βGal1–4Glc	13	77	5.6	8.4	2.8

Stoichiometry was fixed to 1. Experiments were performed twice with SD values less than 10%.

IL; gene *lecA*) and II (PA-IIL; gene *lecB*), which are specific for galactose and fucose, respectively.[1,2]

The galactophilic lectin PA-IL was the first *P. aeruginosa* lectin to be isolated by affinity chromatography using a Sepharose column.[3] It consists of 121 amino acids (12.75 kDa) associated in homotetramers.[4] The crystal structures obtained in the absence and in the presence of calcium[5,6] demonstrated that each monomer adopts a small β-sandwich fold consisting of two curved sheets, each of four antiparallel β-strands. The tetramer is assembled by 222-symmetry. The structure of the complex of PA-IL with galactose showed the presence of one calcium ion and one galactose ligand in the same binding site.[5] Oxygen atoms O3 and O4 of galactose participate in the coordination of calcium, as observed in several galactose-specific C-type lectins[7] and, more recently, in CEL-III, the β-trefoil sea cucumber lectin.[8]

PA-IL is a virulence factor, and its expression is under the control of the "quorum sensing" system.[9,10] The lectin is toxic for respiratory epithelial cells in primary culture.[11] When associated with other toxins such as exotoxin A and elastase, the presence of PA-IL induced a high rate of mortality in a mouse model of gut-derived sepsis.[12] PA-IL may be involved in pathogen adhesion since it binds to seromucinous glands and to capillaries and small blood vessels in sections of mink lungs.[13] Studies on *P. aeruginosa* mutants lacking or surproducing the lectin demonstrated its involvement in biofilm formation.[14] This finding is in agreement with the high percentage of galactose residue present in the biofilm formed by enzyme of the psl locus.[15] Nevertheless, the oligosaccharide epitopes that are involved in these different processes have not yet been characterized.

PA-IL has medium-range affinity for galactose, with an association constant (K_a) of 3.4×10^4 M^{-1} as reported from an equilibrium dialysis study.[16] When longer epitopes are considered, the lectin has a preference for α-linked terminal galactose. Competition assays with disaccharides indicate strong affinity for αGal1–6Glc (melibiose), slightly higher than that for αGal1–4Gal (galabiose) and that for αGal1–3Gal.[17] PA-IL efficiently agglutinates erythrocytes with blood group B (αGal1–3[αFuc1–2]βGal1–4βGlcNAc-R), blood group P^k equivalent to the globotriaosylceramide Gb3/CD77 antigen[18] (αGal1–4βGal1–4βGlc-Cer), and P_1 (αGal1–4βGal1–4βGlcNAc1–3βGal1–4βGlc-Cer).[19] Dual recognition of αGal1–4βGal and αGal1–3βGal capped glyco-

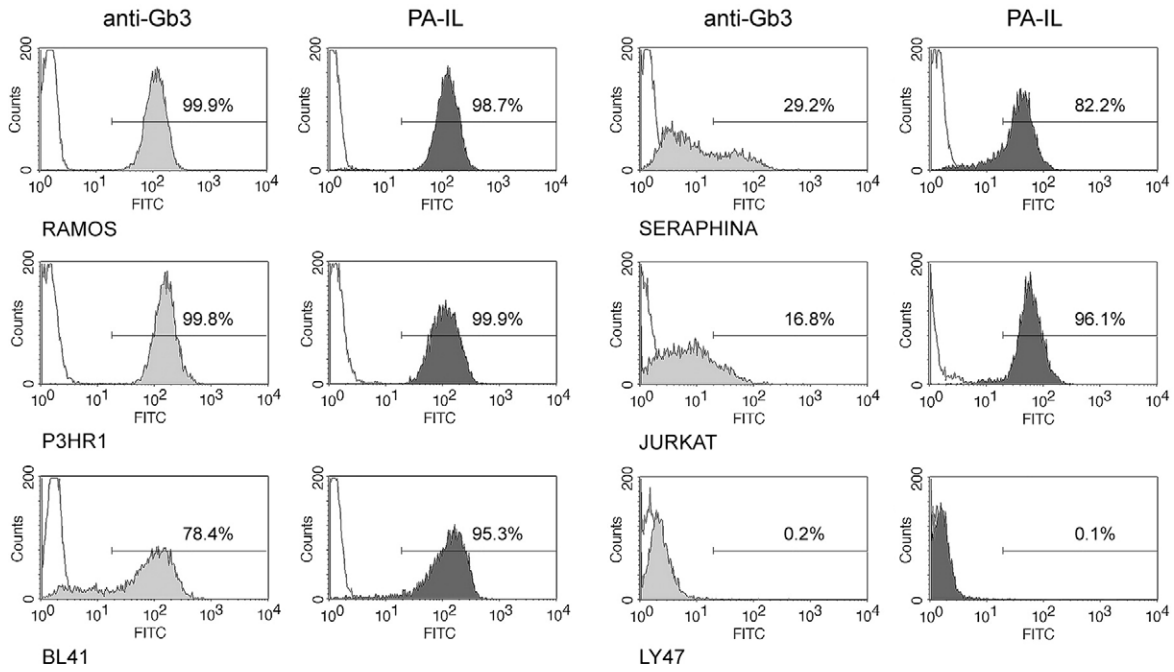


Fig. 1. Comparison of PA-IL and anti-Gb3/CD77 mAb labelings on various cell lines. Cells were labeled with biotinylated PA-IL or 1A4 anti-Gb3/CD77 mAb and an appropriate FITC-conjugated secondary antibody. Fluorescence intensity was analyzed by flow cytometry. Histograms show anti-Gb3/CD77 staining (gray) and PA-IL staining (dark gray) compared with secondary reagent staining controls (unshaded).

sphingolipids was confirmed by thin-layer chromatography (TLC).[20]

In the present work, we characterize the specificity and affinity of PA-IL for αGal1–4βGal and αGal1–3βGal epitopes by cell surface labeling combined with glycan array analysis and titration microcalorimetry. The crystal structure of PA-IL complexed with αGal1–3βGal1–4Glc trisaccharide establishes the atomic basis of the specificity and reveals how the second galactose residue makes specific contacts with the protein surface. Docking studies demonstrate that αGal1–4Gal creates another network of contacts for achieving a very similar affinity. Finally, 20-ns molecular dynamics (MD) in explicit water allows for analyzing the flexibility of each disaccharide ligand in the protein binding site.

Results

Cell surface labeling by PA-IL and anti-Gb3/CD77 monoclonal antibody

So far, most studies on the cellular specificity of PA-IL have been conducted by hemagglutination tests on erythrocytes or by immunohistochemistry on paraffin sections of various animal tissues. In both techniques, the cellular membrane is modified as compared with live cells. We therefore decided to take advantage of the high expression of the glycolipid antigen Gb3/CD77 (αGal1–4βGal1–4βGlc-Cer) on Burkitt lymphoma (BL) cells[21,22] to analyze the binding of PA-IL on these live cells and to compare it with anti-Gb3/CD77 monoclonal antibody (mAb) reactivity. Ten Burkitt cell lines were thus labeled with an anti-Gb3/CD77 mAb (1A4) or PA-IL, and their binding capacities were compared by flow cytometry. Five representative BL cell lines (Ramos, P3HR1, BL41, Seraphina, and LY47) and a T-cell line (Jurkat) are shown in Fig. 1. Gb3/CD77 was highly expressed on Ramos, P3HR1, and BL41, faintly expressed on Seraphina and Jurkat, and not detectable on LY47 cells (as demonstrated with 1A4 mAb labeling). In the case of labeling with PA-IL, all cell lines except LY47 were highly positive. Thus, four cell lines have a similar staining pattern with PA-IL and anti-Gb3/CD77, suggesting that this glycolipid may serve as a receptor for the lectin. However, PA-IL seems to recognize another epitope on Seraphina and Jurkat cell surfaces.

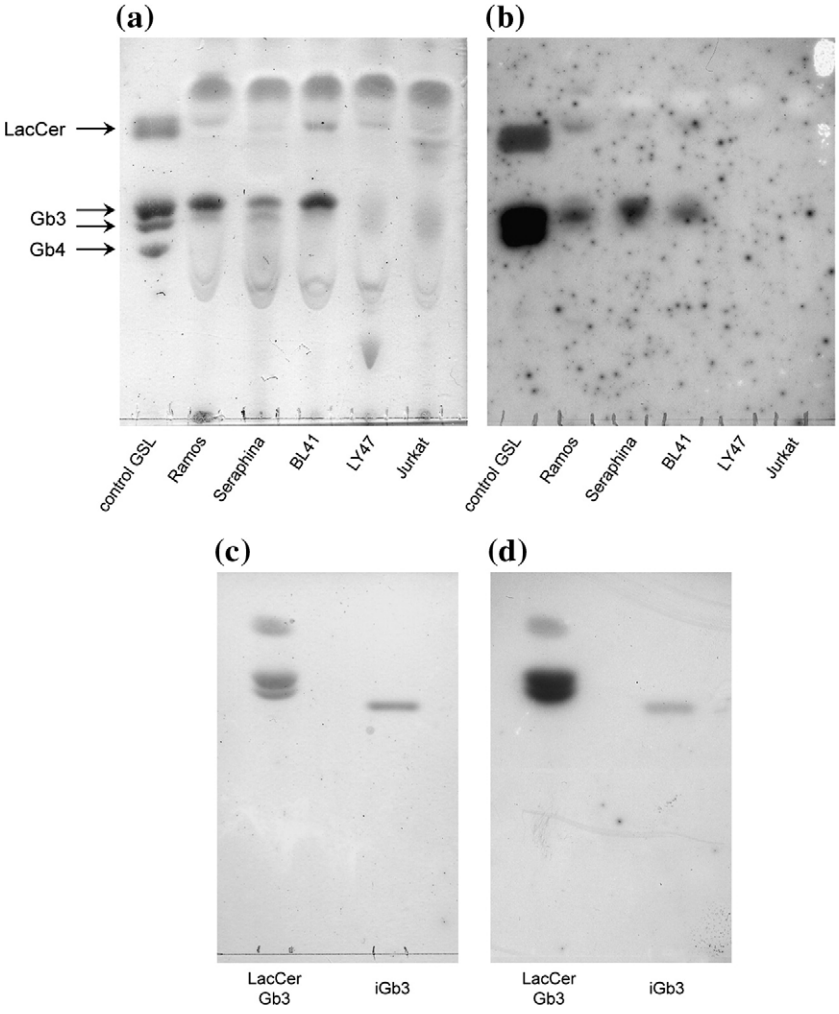


Fig. 2. Immunostaining pattern of cell line LP glycolipids and purified glycolipids. (a and c) Chemically stained with orcinol-H_2SO_4 reagent. (b and d) Immunostained with biotinylated PA-IL. Solvent system for TLC was chloroform/methanol/water (60:35:8, v/v/v).

Glycolipid recognition by PA-IL

In order to determine if PA-IL may bind αGal moiety of glycoproteins, we performed Western blot analysis of the cell line lysates with biotinylated PA-IL. We were not able to detect any specific staining especially for Seraphina and Jurkat cells (data not shown). Binding of PA-IL to lower-phase glycolipids obtained from the BL cell lines was studied by TLC immunostaining in order to confirm that the lectin recognizes Gb3. Chemical detection of glycoconjugates with orcinol (Fig. 2a) revealed that all cell lines contain low or very low amounts of a compound that had the same mobility as lactosylceramide (LacCer). Ramos and BL41 cells contain large amounts of a glycolipid that co-migrates with Gb3. Seraphina cells contain a lower amount, whereas this glycolipid is not detectable in Jurkat and LY47 cells.

TLC immunostaining was performed with biotinylated PA-IL. Figure 2b shows that PA-IL does not bind standard Gb4 (βGalNAc1–3αGal1–4βGal1–4βGlc1-1Cer) but possesses high affinity for standard Gb3 (Galα1–4Galβ1–4Glcβ1-1Cer) and surprisingly binds to standard LacCer (βGal1–4βGlc1-1Cer). Among the various LP extracts tested, glycolipids from Ramos, Seraphina, and BL41 are recognized by PA-IL, whereas no staining was detected for Jurkat and LY47. It must be noted that PA-IL staining perfectly matches the orcinol detection of Gb3 for Ramos and BL41 glycolipids, whereas PA-IL immunostaining and orcinol staining are slightly different in the case of Seraphina cells. Indeed, although orcinol stained a glycolipid that co-migrates with Gb3, PA-IL recognizes a compound with a slightly higher mobility than Gb3. Thus, these results confirm previous data obtained by Lanne *et al.*[20] showing that PA-IL is able to bind Gb3 in TLC immunostaining and that there may be another glycolipid with mobility close to Gb3 recognized by the lectin.

Isoglobotriaosylceramide (iGb3; αGal1–3βGal1–4βGlc1-1Cer) is an isomer of Gb3 with the same mobility on TLC as Gb3.[23] We therefore tested the binding capacity of the lectin for this glycolipid by TLC immunostaining. Orcinol staining revealed that chemically synthesized iGb3 has a slightly lower migration than Gb3 purified from human erythrocytes (Fig. 2c). This difference may be explained by the chain length of their respective ceramide moiety. Staining with PA-IL showed that the lectin is also able to recognize iGb3, albeit more weakly than Gb3 (Fig. 2d). Whether or not iGb3 is the compound recognized on the cell surface of the Seraphina cells remains to be determined.

Specificity of PA-IL by glycan array screening

Oligosaccharide specificity of PA-IL has been determined by glycan array experiments at the Consortium for Functional Glycomics (Fig. 3). The screening results against 241 glycan epitopes show high specificity of PA-IL toward terminal α-galactoside, with the highest preference for αGal1–4Gal, typical for Gb3/P^k and P_1 antigens. Terminal galactose in α1–6 and α1–3 linkages is also recognized, albeit with a lower observed binding. PA-IL does not display any significant binding to β-galactosides as can be seen from glycan array results.

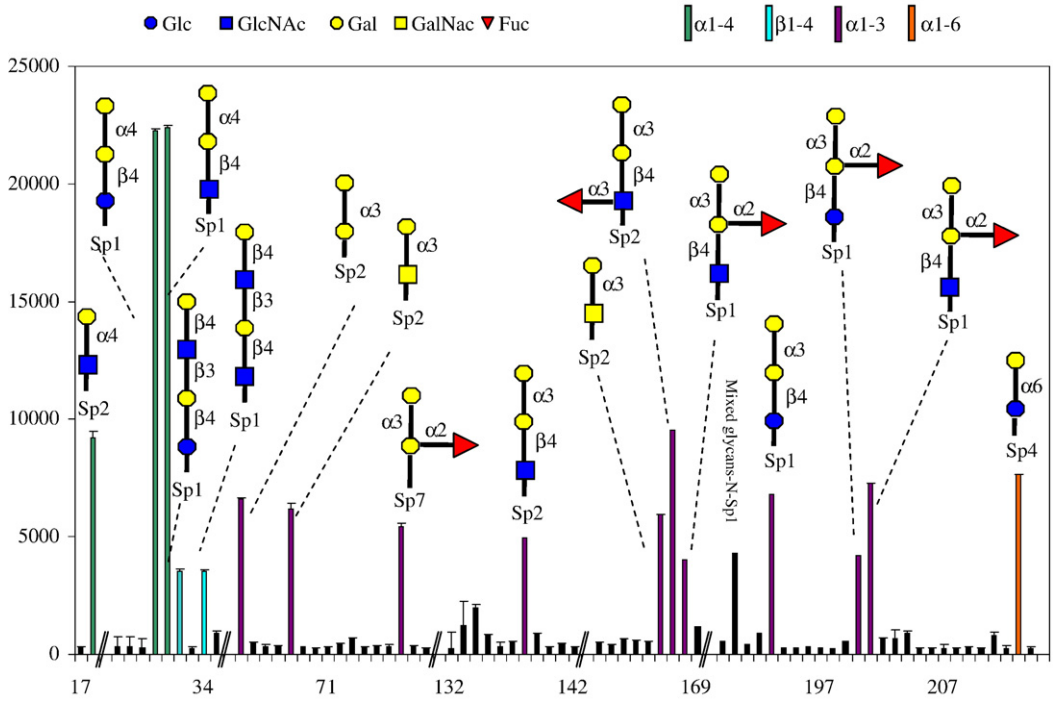


Fig. 3. Plate array screening for PA-IL specificity. PA-IL was labeled with Alexa 488 prior to arrival and screening on the glycan array version 3.8. The concentration of PA-IL used was 30 μg/ml. Color coding of the histograms is as follows: green for αGal1–4 terminal residues; purple for αGal1–3; light blue for βGal1–4; and orange for αGal1–6.

The only exception observed was with βGal1–4βGlcNAc1–3βGal1–4βGlcNAc-Sp1 (LN2) and βGal1–4βGlcNAc1–3βGal1–4βGlc-Sp1(LNnT) epitopes, for which a significant amount of bound lectin was still detectable even after extensive washing. There is no any rational explanation for the binding of βGal1–4GlcNAc epitope as the same terminal LacNAc motif is present in 15 other screened oligosaccharides. It may indicate that PA-IL can bind β-galactosides with negligible affinity but that high-density surface presentation of the βGal epitopes and lectin multivalency could lead to observable binding.

Affinity studies by isothermal titration microcalorimetry

The affinity constant and the thermodynamic binding parameters were determined using titration microcalorimetry, a method that is well suited to the characterization of protein–carbohydrate interactions, in order to characterize the interaction between PA-IL and αGal-containing compounds.[24] Titration curves for PA-IL binding to αGal1–4βGa1–4Glc and αGal1–3βGa1–4Glc are displayed in Supplementary Fig. 1S. Surprisingly, the lectin has very similar K_a values for the two trisaccharides (68–77 μM) (Table 1). This value is slightly higher than the one previously reported for the PA-IL/Gal interaction by equilibrium dialysis study.[16] For both disaccharides, the interaction is enthalpy driven, with an unfavorable entropy contribution.

Crystal structure of PA-IL/αGal1–3βGal1–4Glc trisaccharide

Co-crystals of the lectin and αGal1–3βGal1–4Glc trisaccharide were obtained in space group *P*1 with cell dimensions of $a=79.2$ Å, $b=86.5$ Å, $c=119.1$ Å, $\alpha=93.9°$, $\beta=98.2°$, and $\gamma=90.1°$ (Table 2). The asymmetric unit consists of 24 PA-IL monomers arranged in 6 tetramers, each centered on a pseudo-C222 axis (Fig. 4a). This resulted in the refinement of 2904 amino acids, 3048 water molecules, 20 ethylene glycol molecules, 24 calcium ions, and 72 carbohydrate residues with an R_{crys} of 18.5% and an R_{free} of 24.5% to 1.9-Å resolution. As previously described,[5] each monomer adopts a small β-sandwich fold consisting of two curved sheets, each consisting of four antiparallel β-strands. Tetramerization occurs by interaction between the largest sheets for one interface and by contacts between C-terminus moieties for the other interface.

In the carbohydrate binding site, clear density can be seen in each monomer corresponding to one calcium ion and one trisaccharide (Fig. 4b). The nonreducing αGal residue is buried in the binding site and participates in the coordination of the calcium ion through oxygen atoms O3 and O4. These two atoms also establish hydrogen bonds with Tyr36, Asp100, Thr104, Asn107, and Asn108 (Fig. 4c and Table 3). The position of the sugar is also stabilized by oxygen O2, which creates a hydrogen bond with the nitrogen atom of Asn107. In addition, the O6 oxygen interacts with the side chains of Glu53 and His50. Oxygen O6 also participates in a water-bridged contact through a water molecule that is conserved in the 24 monomers. This structural water molecule makes hydrogen bonds with the Pro51 main chain oxygen and Glu53 main chain nitrogen. The hydrophobic contacts are rather limited since only the CH group at C2 interacts with the side chain of Tyr61. The calcium ion is heptacoordinated with five contacts to protein side chains and two contacts to galactose oxygen atoms (Table 3). The overall orientation of galactose relative to the calcium ion and protein residues is very similar to that observed in the complex between PA-IL and galactose.[5]

Additional contacts are established by the second galactose between the oxygen O2 and the nitrogen of Gln53. The glucose made interaction with Gln53 through the O4 atom involved in the glycosidic linkage between galactose rings. This hydrogen bond is rather weak and is not observed

Table 2. Data collection and refinement statistics for the PA-IL/trisaccharide complex

	PA-IL/αGal1–3βGal1–4Glc
Data collection	
Beam line	ID14-1
Wavelength (Å)	0.934
Resolution (Å)	55.22–1.9
Highest-resolution shell (Å)	1.95–1.9
Cell dimensions	
Space group	*P*1
a, *b*, *c* (Å)	79.2, 86.5, 119.1
α, β, γ (°)	93.9, 98.2, 90.1
Measured reflections	521,421
Unique reflections	230,909
Averages multiplicity	2.3 (2.20)
Completeness (%)	93.9 (89.1)
Average I/σ (I)	6.6 (2.3)
R_{merge} (%)	11.0 (35.8)
Wilson *B*-factor (Å^2)	12.3
Refinement	
Resolution range (Å)	55.22–1.9
R_{crys}	0.185
R_{free}	0.245
Cruickshank's dispersion precision indicator based on maximum likelihood (Å)	0.158
Average B_{iso} (Å^2)	
All atoms	13.8
Protein atoms	12.5
Sugar atoms	22.3
Solvent atoms	20.2
RMSD from ideality	
Bonds (Å)	0.016
Angles (°)	1.53
Outliers on Ramachandran plots (MolProbity)	1
Protein atoms	21,634
Sugar atoms	816
Calcium atoms	24
Other hetero atoms	20
Water molecules	3048
Protein Data Bank deposition code	2VXJ

Values in parenthesis refer to the highest resolution shell.

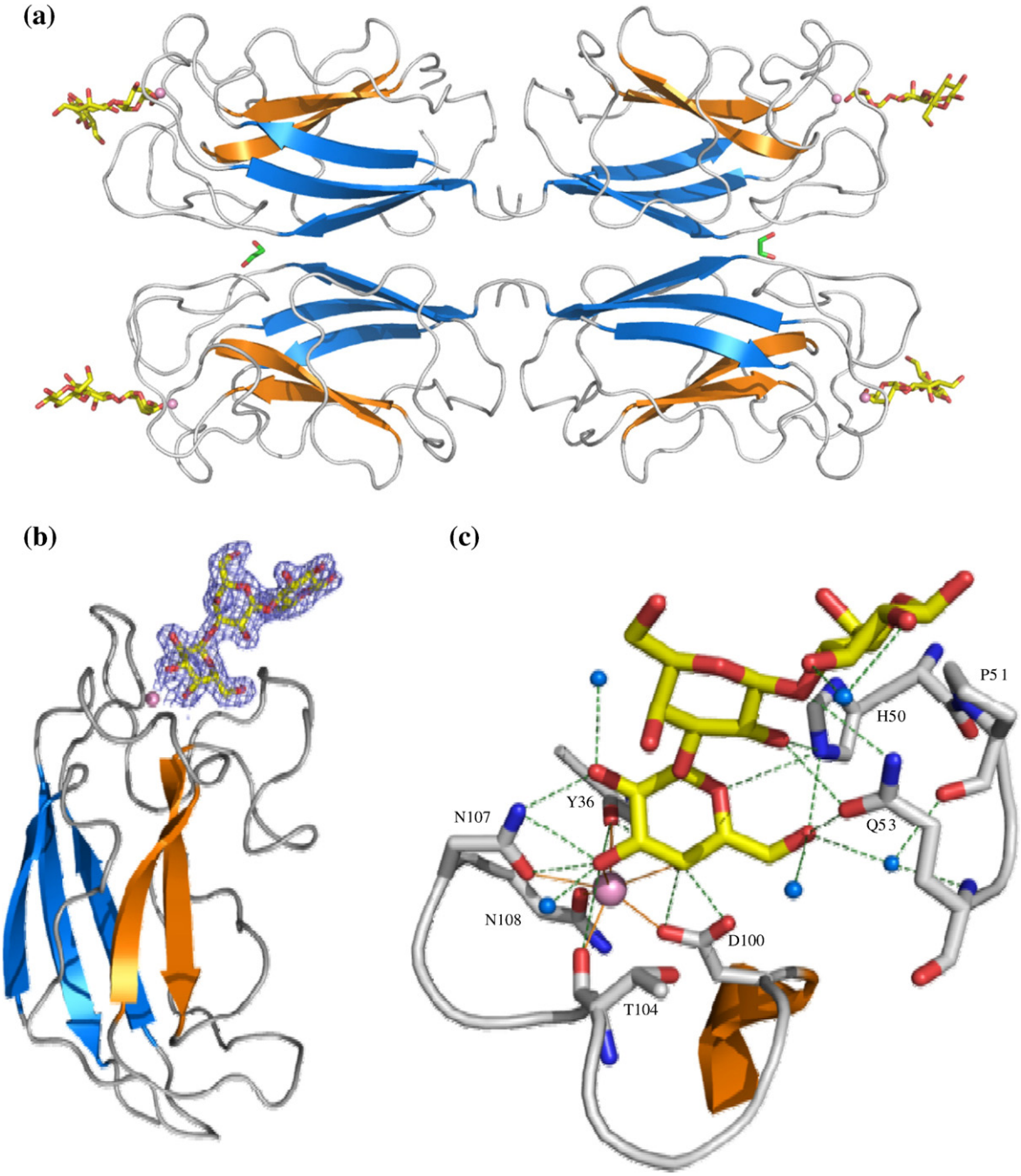


Fig. 4. Crystal structure of PA-IL/αGal1–3βGal1–4Glc complex. (a) Representation of one tetramer with the two β-sheets shown in blue and orange. Trisaccharide is represented in yellow sticks; ethylene glycol, in green sticks; and calcium ion, by a pink sphere. (b) Representation of one monomer (chain K) with the final weighted $2mF_o - DF_c$ electron density map (contoured at 1 σ, 0.34 eÅ^{-3}) around the trisaccharide. (c) View of the binding site with hydrogen bonds represented as green dashed lines and coordination contacts as continuous orange lines.

in all monomers of the asymmetric unit. The glucose orientation appears to be mainly stabilized by a hydrophobic contact between C6 and Pro51.

Analysis of the conformations adopted by the trisaccharide in PA-IL binding sites reveals no large variation among the 24 independent molecules. All carbohydrate rings are in the expected $^4C_1$ conformation with no significant distortion. As for exocyclic groups, the α-galactose that is buried in the binding site displays only one orientation of the hydroxymethyl in all the 24 monomers (O5–C5–C6–O6, ca +60°), due to its stabilization by several hydrogen bonds. In the two other monosaccharides, a variety of orientations are observed for the O6 hydroxymethyl groups. Both αGal1–3Gal and βGal1–4Glc disaccharides have been previously demonstrated to adopt several conformations at the glycosidic linkages when in solution.[25,26] The energy map of αGal1–3Gal calculated as a function of torsion angles Φ (O5–C1–O1–C3′) and Ψ (C1–O1–

Table 3. Distances of interest in calcium and carbohydrate binding site averaged from the 24 monomers in the asymmetric unit (with SD values within parentheses)

Atom 1	Atom 2	Distance (Å)[a]
Coordination of calcium ion		
Ca	Tyr36·O	2.4 (0.1)
Ca	Asp100·OD2	2.5 (0.2)
Ca	Thr104·O	2.3 (0.1)
Ca	Asn107·OD1	2.4 (0.1)
Ca	Asn108·OD1	2.4 (0.1)
Gal1·O3	Cal	2.5 (0.1)
Gal1·O4	Cal	2.5 (0.1)
Hydrogen bonds between PA-IL and αGal1–3βGal1–4Glc		
Gal1·O2	Asn107·ND2	3.0 (0.1)
Gal1·O3	Asn107·OD1	3.0 (0.1)
	Asn107·ND2	3.0 (0.2)
Gal1·O4	Tyr36·O	3.1 (0.1)
	Asp100·OD1	2.6 (0.1)
	Asp100·OD2	2.9 (0.2)
	Thr104·OG1[a]	3.1 (0.04)
Gal1·O6	His50·NE2	2.9 (0.1)
	Gln53·OE1	2.7 (0.1)
Gal1·O5	His50·NE2[a]	3.1 (0.1)
Gal2·O2	Gln53·OE1	2.7 (0.1)
	Gln53·NE2	3.1 (0.1)
	His50·NE2[a]	3.1 (0.1)
Glc3·O4	Gln53·NE2[a]	3.1 (0.1)
Hydrogen bonds with conserved water molecules		
Gal1·O6	wat	2.8 (0.1)
Bridging water molecules		
wat	Pro51·O	2.7 (0.1)
	Gln53·N	2.9 (0.1)

[a] Hydrogen bond not present in each monomer.

C3′–C4′) reveals three main energy minima.[25] As for the βGal1–4Glc linkage, which has a less flexible behavior in solution, the observed conformations of $\Phi = -73°$ (±6) and $\Psi = -110°$ (±6) do belong to the low-energy region reported for the lactose energy map.[26]

Molecular modeling of PA-IL interacting with αGal1–3Gal and αGal1–4Gal disaccharides

Possible conformations of the two disaccharides

The possible conformation of the two linkages of interest can be investigated by calculating the MM3 glycosidic linkage energy maps as a function of rotation about the two bonds while taking into account the flexibility of each ring as described previously.[27] This so-called flexible energy map of the αGal1–3Gal linkage has been previously calculated in our group.[28] The αGal1–4Gal linkage was calculated using the same approach. Both maps exhibit several low-energy regions (Fig. 5). For both disaccharides, the flexibility around the Φ torsion angle is more restricted than that around the Ψ torsion angle due to exoanomeric effect. The αGal1–3Gal disaccharide exhibits two main energy minima ($\Phi/\Psi \approx 80°/80°$ and $\Phi/\Psi \approx 100°/140°$) separated by a low-energy barrier and a remote secondary energy minimum ($\Phi/\Psi \approx 100°/-70°$) with a higher energy barrier. In the present structure of PA-IL/trisaccharide complex, the observed 24 conformations lie in a narrow range of $\Phi = 98°$ (±4) and $\Psi = 122°$ (±6), corresponding to the saddle region between the two main energy minima. The αGal1–4Gal disaccharide displays a plateau of low energy, with Φ and Ψ varying from 80° to 100° and from 90° to 180°, respectively. This behavior is independent from the hydration model used in the calculations, as recently demonstrated for this disaccharide with the use of the CHARMM program.[29] The only

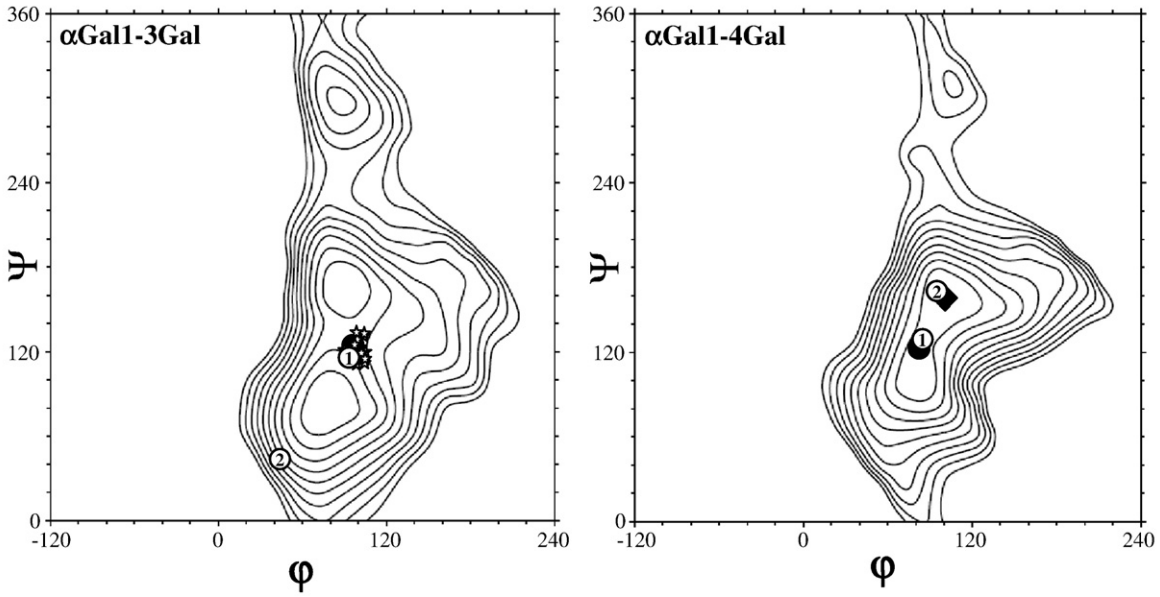


Fig. 5. Adiabatic energy maps of αGal1–3Gal (a) and αGal1–4Gal (b) disaccharides calculated as a function of Φ and Ψ dihedral angles, with isoenergy contouring at 1 kcal/mol above the absolute minimum up to 10 kcal/mol. The values observed in the crystal structure of the PA-IL/αGal1–3βGal1–4Glc complex are reported as stars. The lowest energy conformations from the docking study are reported as black dots, and the crystal structure of isolate disaccharide is shown as a black diamond. Snapshots from MD simulation are indicated by encircled numbers.

crystal structure available for αGal1–4Gal is that of the isolated disaccharide,[30] with its conformation (Φ/Ψ = 98°/158°) belonging to the low-energy region displayed in Fig. 5.

Docking of disaccharides onto PA-IL

Both αGal1–3Gal and αGal1–4Gal disaccharides were docked onto PA-IL using the AutoDock 3 program together with parameters recently validated for calcium-dependant carbohydrate binding.[31] In both cases, several possible docking modes were observed among the 100 independent runs. Nevertheless, a high prediction quality was obtained since the lowest energy docking mode always corresponds to a highly populated cluster, indicating good convergence of results (Table 4).

Results obtained with the αGal1–3Gal disaccharide validate our docking approach since the lowest energy docking mode corresponds closely to the interaction observed in the crystal structure of the PA-IL/trisaccharide complex (Fig. 6a). The nonreducing αGal coordinates the calcium ion with distances of 2.3 and 2.6 Å from oxygen O3 and oxygen O4, respectively. It establishes the same hydrogen-bond network as in the crystal except for the O6 hydroxymethyl group that adopts a different orientation and, consequently, different contacts. The reducing βGal establishes hydrogen bonds with His50 and Gln53 that stabilize this conformation. The conformation at the glycosidic linkage is also correctly predicted (Φ/Ψ = 95°/126°), corresponding to the ones observed in the crystal structure (Fig. 5). The second binding mode (10%) is not very different, reproducing correctly the nonreducing galactose in the binding site. Other runs yield completely different binding modes, but the associated energies are significantly higher.

Docking prediction for αGal1–4Gal in PA-IL also yields a low-energy conformation with a highly populated cluster (50%) and correct positioning of the nonreducing αGal on the calcium ion (Table 4). This docking mode presents the same hydrogen-bond network for the nonreducing residue as observed for the αGal1–3Gal complex (Fig. 6b). By contrast, the reducing βGal is positioned very differently from its position in the other disaccharide, which is expected due to the difference of the stereochemistry at the linkage axial–equatorial for 1–3 and axial–axial for 1–4. The reducing galactose finds a certain stability creating hydrogen bonds with His50 and Gln53. The conformation at the glycoside linkage (Φ/Ψ = 83°/122°) lies in the center of the low-energy region of the energy map (Fig. 5), indicating that no distortion of the disaccharide is needed for binding in this orientation. The other docking modes are energetically unfavorable and do not yield a correct interaction between the αGal residue and the calcium ion.

MD in the presence of water molecules

MD calculations were performed in explicit water in order to estimate the flexibility of each disaccharide into PA-IL binding site. Simulations of 10 ns were conducted on the monomer of PA-IL corresponding to 1762 atoms of proteins, one calcium ion, and 45 atoms of each disaccharide in 6187 and 5877 water molecules for αGal1–3Gal and αGal1–4Gal, respectively. Stability was checked over time, with the global RMS varying less than 1 Å from starting structures, therefore confirming that PA-IL is stable as a monomer. For both disaccharides, the nonreducing galactose and the calcium ions remain very stable in the protein binding site. Distances between calcium ion and the O3 and O4 hydroxyl groups did not show variations of more than ±0.3 Å from the mean value of 2.46 Å (Table 5). For each disaccharide, the analysis was therefore focused on the conformation at the glycosidic linkage and the contacts between the external galactose and the protein surface. A video representing the movement of the disaccharides and neighboring amino acids is provided in Supplementary Material.

Behavior of αGal1–3Gal disaccharide in PA-IL binding site

The history of the Φ and Ψ variations of the αGal1–3Gal glycosidic linkage during the 10-ns simulation is depicted in Fig. 7. The Φ torsion appears very stable and remains 85% of the time between 60° and

Table 4. Description of lowest energy binding modes as predicted by AutoDock

	Clusters (%)	Φ angle (°)	Ψ angle (°)	Lowest docked energy (kcal/mol)	Sugar into binding site[a]	RMSD αGal (Å)[b]
αGal1–3βGal						
Run A	70	94.9	125.6	−7.97	nr	0.75
Run B	10	85.5	118.2	−7.50	nr	0.90
Run C	5	117.8	122.1	−7.17	nr	1.22
Run D	10	100.1	119.9	−6.82	nr	4.15
Run E	5	112.8	121.5	−6.34	r	6.54
αGal1–4βGal						
Run A	50	82.8	121.9	−7.99	nr	0.64
Run B	12	86.8	110.3	−7.49	nr	4.16
Run C	6	91.3	127.2	−7.40	nr	0.65

[a] nr indicates nonreducing; r, reducing.
[b] RMSD of the nonreducing galactose compared with the crystal structure.

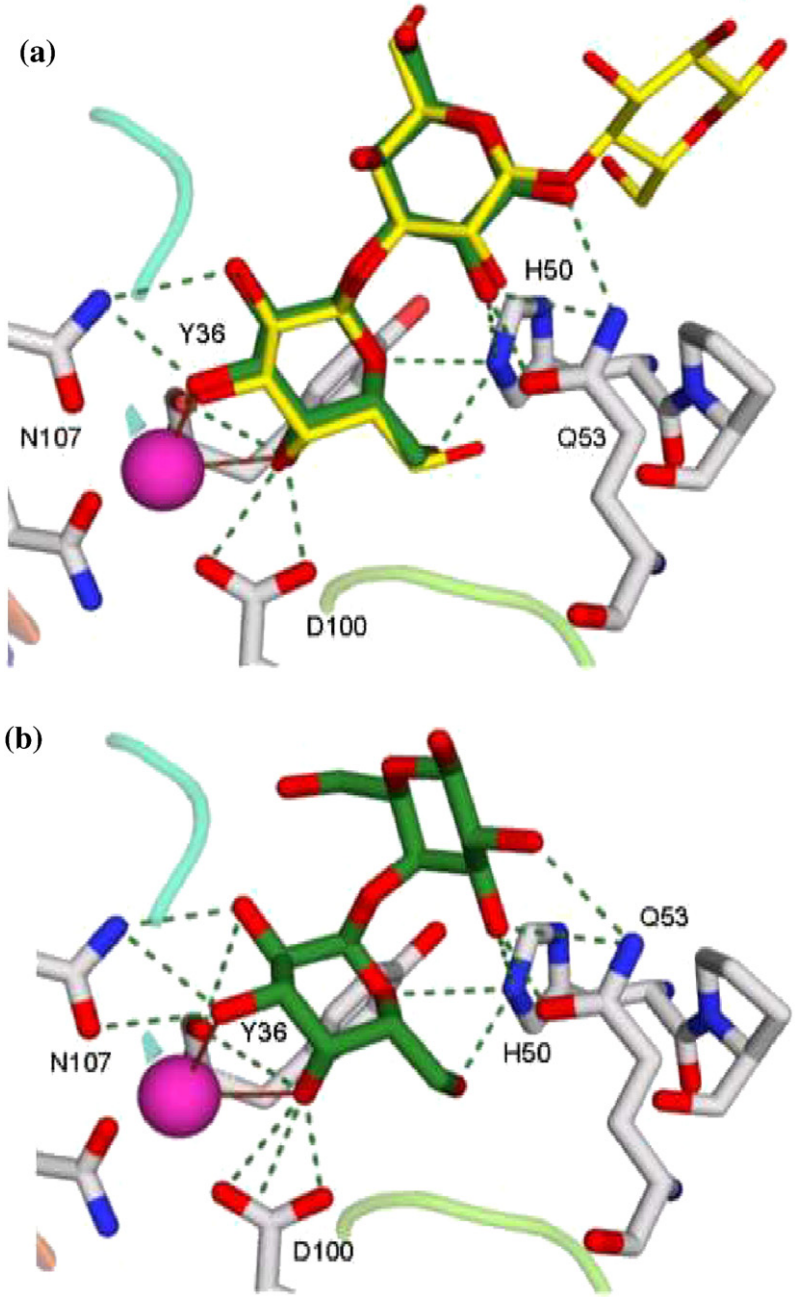


Fig. 6. Visualization of lowest energy docking results of disaccharides in PA-IL. (a) αGal1–3Gal disaccharide (green) compared with the trisaccharide observed in the crystalline complex (yellow). (b) αGal1–4Gal disaccharide (green).

100° (average value=79.1°), with very few incursions in regions with lower values except for a short time around 7 ns. The Ψ angle of this disaccharide is characterized by more freedom, varying between 60° and 180° during the simulation, corresponding to the two main energy minima of the energy map (Fig. 5). Nevertheless, 45% of recorded values are between 100° and 140° (average value=109.2°), confirming that the saddle point between the minima (i.e., the conformations observed in the crystal structure of the complex and predicted by docking) is the most stable one in the PA-IL binding site. Snapshot 1, selected in this major conformation (Fig. 7), demonstrates the major role of the interaction between the external Gal residue and amino acids His50 and Gln53. More detailed analysis (Table 5) confirms that the hydrogen bonds between the reducing βGal and these two amino acids are very stable (>30% of the trajectory). Nevertheless, the disaccharide can adopt very different conformations, such as the one displayed by snapshot 2.

Behavior of αGal1–4Gal disaccharide in PA-IL binding site

The disaccharide αGal1–4Gal is globally more static during the MD simulation, with 88% of the Φ angle value between 60° and 100° and 66% of the Ψ angle between 100° and 140° (Fig. 8). This very stable conformation, depicted in snapshot 1, corresponds closely to the one predicted by the docking procedure. Nevertheless, some variations of the value of Ψ are observed, either toward lower values (90°) or, more rarely, to higher values (150°; see

Table 5. Coordination and hydrogen-bond analysis for disaccharides in PA-IL binding site during MD simulation

Atom 1	Atom 2	Percentage occupied
αGal1–3βGal		
αGal·O3	Calcium	100.0
αGal·O4	Calcium	100.0
αGal·O2	Asn107·ND2	99.3
αGal·O3	Asn107·OD1	100.0
	Asn107·ND2	93.9
αGal·O4	Tyr36·O	48.8
	Asp100·OD1	100.0
	Asp100·OD2	99.8
	Thr104·OG1	48.8
αGal·O6	His50·NE2	93.4
	Gln53·OE1	25.6
	Asp100·OD1	7.8
αGal·O5	His50·NE2	30.3
βGal·O2	Gln53·OE1	35.5
	Gln53·NE2	37.5
	His50·NE2	39.3
βGal·O1	Gln53·OE1	15.9
	Gln53·NE2	31.2
αGal1–4βGal		
αGal·O3	Calcium	100.0
αGal·O4	Calcium	100.0
αGal·O2	Asn107·ND2	99.4
αGal·O3	Asn107·OD1	100.0
	Asn107·ND2	94.4
αGal·O4	Tyr36·O	51.0
	Asp100·OD1	100.0
	Asp100·OD2	99.8
	Thr104·OG1	57.5
αGal·O6	His50·NE2	94.4
	Gln53·OE1	18.0
αGal·O5	His50·NE2	43.7
βGal·O3	Gln53·OE1	38.4
	Gln53·NE2	13.9
	His50·NE2	64.3
βGal·O2	Gln53·OE1	41.3
	Gln53·NE2	15.5

snapshot 2), corresponding to the conformation of the disaccharide in solid state.[30] The whole range of the low-energy plateau calculated with MM3 (Fig. 5) is therefore explored during the simulation, but with strong preference for values ($\Phi/\Psi \approx 90°/130°$), that corresponds to the stronger hydrogen-bond network. Indeed, the high occupancy of the hydrogen bond between the atom NE2 of His50 and βGal O3 (64.3%) together with the interaction between the side chain of Gln53 and βGal O3 and βGal O2 (41.3%) explained the reduced flexibility of the reducing monomer of αGal1–4Gal (Fig. 8).

Discussion

The present study provides the atomic basis for the previously reported affinity of PA-IL for αGal-bearing oligosaccharides. Indeed, PA-IL agglutinates human erythrocytes that bear the B epitope (terminated by αGal1–3Gal) more strongly than the A- and O(H)-type ones.[19] The lectin can also be used for differentiation between P-positive (P_1 and P^k both bearing αGal1–4Gal) and P-negative (p) red blood cells.[32] In solution, we demonstrated that both αGal1–4βGal1–4Glc and αGal1–3βGal1–4Glc trisaccharides bind to the lectin with very similar medium-range affinity [K_d (dissociation constant) = 65–85 μM], and structural work confirms the occurrence of hydrogen bonds with the second galactose in both cases. However, in the glycan array experiments, PA-IL binds much more efficiently to αGal1–4Gal- than to αGal1–3Gal-containing oligosaccharides; this is confirmed by the labeling of Burkett cells. It is therefore very likely that the presentation of glycolipids on the cell surface (or glycoconjugates on the microarray slide) strongly influences the binding. PA-IL is a tetramer with a rectangular shape; on the small face, two binding sites are 30 Å apart (Fig. 9a). Higher affinity could occur if the glycolipids have the appropriate presentation for multivalency. Indeed, when building the sphingolipids from the complexes determined by crystallography and modeling, only the αGal1–4βGal1–4βGlc binding mode is compatible with a correct parallel orientation of the lipid tails in the close binding sites of the PA-IL tetramer (Fig. 9b).

In mammals other than humans and apes, the αGal1–3Gal antigen is abundantly expressed on erythrocytes and endothelial cells.[33] PA-IL was successfully used for labeling epithelia and endothelia in mice and mink models,[13,34] and it may have application in xenotransplantation research. In healthy humans, the αGal1–3Gal disaccharide is mostly associated to blood group B antigen, whereas there is no direct biochemical evidence for the expression of iGb3. However, iGb3 was recently proposed to be one of the candidates recognized by human natural killer T cells under pathophysiological conditions, such as cancer and autoimmune disease,[35] and its direct lysosomal precursor (iGb4) was detected in human thymus by ion-trap mass spectrometry.[36] It will therefore be of great interest to determine if iGb3 is expressed in the Seraphina cells. On the contrary, the activity of Gb3 synthase has been clearly identified in tissues of several human and mice organs.[37,38] Gb3 is not only highly expressed on a narrow range of lymphocytes and associated B-cell lymphomas[22,39] but also a more general cancer-related marker. Shiga toxin-1 (also called verotoxin-1) displays a very fine specificity for Gb3 and its B-subunits are presently used not only to label cancer cells but also to induce apoptosis of cancer or to elicit antitumor immunity.[40–42] PA-IL specificity has a somewhat broader specificity than verotoxin since it also weakly recognizes αGal1–3Gal epitope, but the lectin has potential applications for cell typing and/or tumor targeting. In addition, the fine characterization of molecular basis for disaccharide specificity will be an aid for the development of high-affinity ligands that may be used as antiadhesive compounds against infection by *P. aeruginosa*.[43,44]

Materials and Methods

Materials

The 1A4 (mouse monoclonal immunoglobulin M anti-Gb3/CD77) ascite was provided by Dr. S. Hakomori

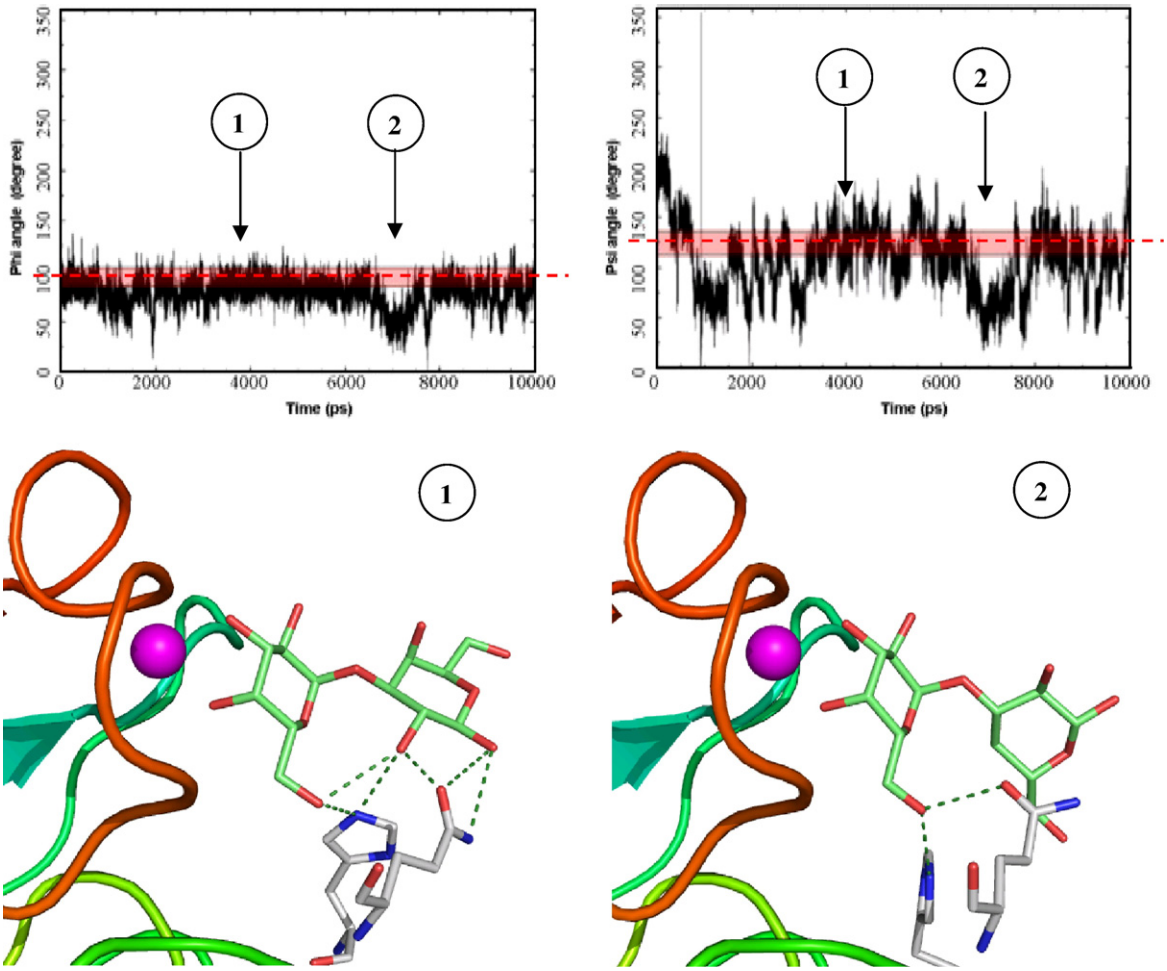


Fig. 7. Analysis of MD trajectory of the PA-IL/αGal1–3Gal complex. Plots of Φ and Ψ dihedral angles of the disaccharide as a function of time. The red area represents the range of values observed in the crystal structure of the complex, and the dotted red line represents the value predicted in the lowest energy docking mode. Two snapshots corresponding to conformations at 3900 ps (1) and 7000 ps (2) are also displayed.

(Seattle, WA). Biotinylated PA-IL was generated using a FluoReporter® Mini-Biotin-XX Protein Labeling Kit (Molecular Probes). A fluorescein isothiocyanate (FITC)-conjugated goat F(ab′)2 antimouse (Caltag Laboratories) was used for detection of 1A4 mAb, and biotinylated PA-IL was revealed with FITC-conjugated streptavidin (DakoCytomation). Purified glycolipids used as controls (LacCer, Gb3, Gb4) were obtained from Sigma-Aldrich, and isoGb3 was kindly provided by Dr A. Bendelac (Chicago, IL). PPMP was obtained from Matreya. αGal1–3βGal1–4Glc trisaccharide was purchased from Carbohydrate Synthesis.

PA-IL cloning and production

The recombinant protein PA-IL was cloned using following procedure: The *lecA* gene was amplified by polymerase chain reaction using genomic DNA from *P. aeruginosa* ATCC 33347 as a template with the primers 5′-CGG AGA TCA CAT ATG GCT TGG AAA GG-3′ and 5′-CCG AGA CAA GCT TTC AGG ACT CAT CC-3′ (NdeI and HindIII restriction sites are underlined). After digestion with NdeI and HindIII, the amplified fragment was introduced into pET25(b+) vector (Novagen, Madison, WI), resulting in plasmid pET25pa1l.

Escherichia coli BL21(DE3) cells harboring the pET25-pa1l plasmid were grown in 1 l of Luria broth at 37 °C. When the culture reached an optical density of 0.5–0.6 at 600 nm, isopropyl-β-D-thiogalactopyranoside was added to a final concentration of 1 mM. Cells were harvested after a 3-h incubation at 30 °C, washed, and resuspended in 10 ml of the loading buffer (20 mM Tris–HCl and 100 μM $CaCl_2$, pH 7.5). The cells were broken by cell disruption (Constant Cell Disruption System, UK). After centrifugation at 10,000*g* for 1 h, the supernatant was further purified by affinity chromatography on Sepharose 4B (GE Healthcare). PA-IL was eluted with 1 M NaCl in loading buffer. The purified protein was intensively dialyzed against distilled water for 7 days, lyophilized, and kept at −20 °C.

Cell lines

All cell lines were originally established from endemic or sporadic cases of human BL, except Jurkat, which was derived from a human T lymphoma. These cell lines were cultured in RPMI 1640 medium (Invitrogen) containing 2 mM L-glutamine, 1 mM sodium pyruvate, 20 mM glucose, and 20 μg/ml of gentamicin and supplemented with 10% heat-inactivated fetal calf serum. Ramos PPMP and

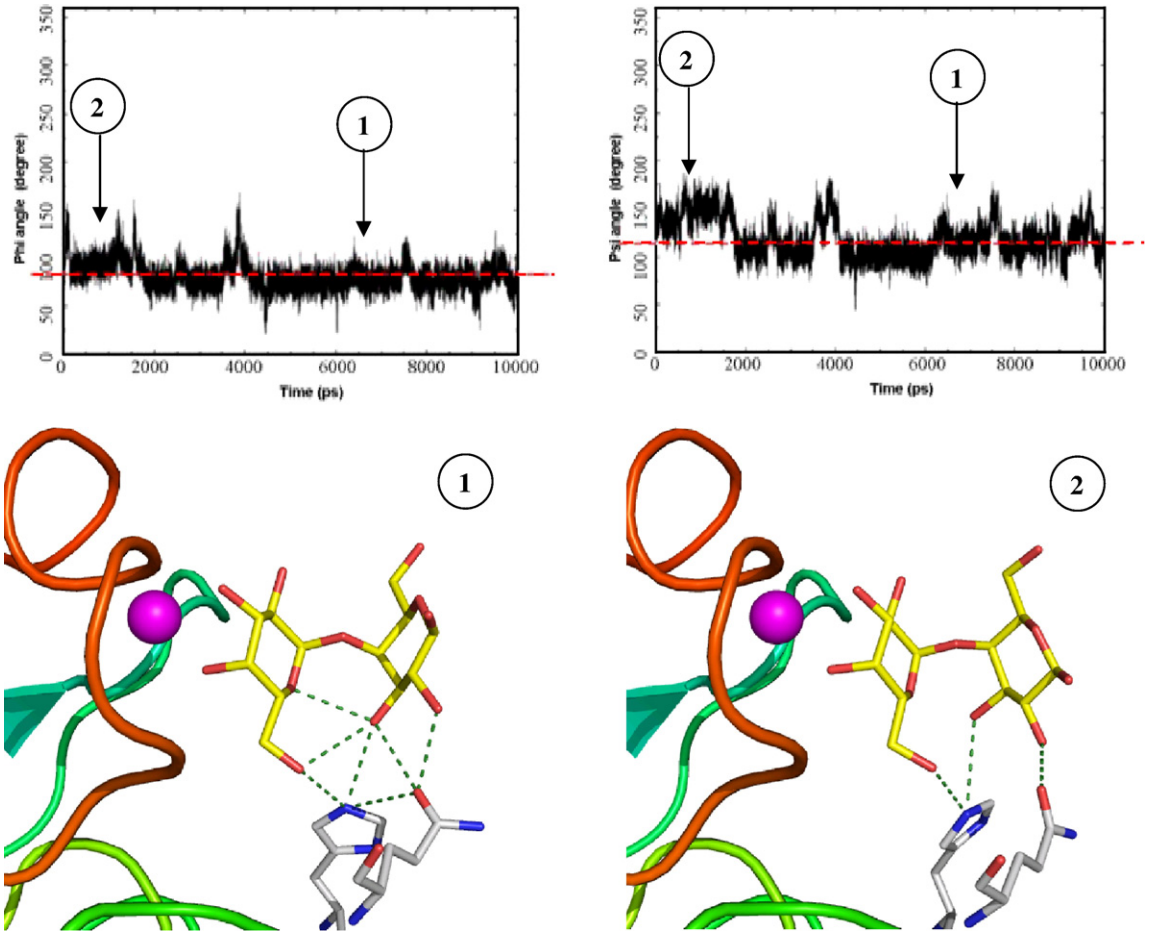


Fig. 8. Analysis of MD trajectory of the PA-IL/αGal1–4Gal complex. Plots of Φ and Ψ dihedral angles of the disaccharide as a function of time. The dotted red line represents the value predicted in the lowest energy docking mode. Two snapshots corresponding to conformations at 6500 ps (1) and 1000 ps (2) are also displayed.

Seraphina PPMP were obtained after 10 days of culture in media containing 2 μM D,L-threo-PPMP (D,L-threo-1-phenyl-2-hexadecanoylamino-3-morpholino-1-propanol HCl), a glucosylceramide synthase inhibitor causing reversible glycolipid depletion.

Surface immunofluorescence labeling

Cells (3×10^5) were incubated with 50 μl of primary reagent (1A4 mAb or biotinylated PA-IL) for 30 min at 4 °C. After washing, cells were incubated with 50 μl of secondary reagent for 30 min at 4 °C. Cells were then washed and analyzed by flow cytometry (FACSCalibur, Becton Dickinson). Data were analyzed using Cell Quest software (Becton Dickinson).

Glycolipid purification and TLC

Glycolipids were extracted from 150×10^6 cells sonicated twice in isopropanol/hexane/water (55:25:20, v/v/v). After centrifugation at 500*g*, supernatants were dried under N_2 and then partitioned according to Folch's procedure. Folch's lower phase (LP) and upper phase (UP) were dried under N_2. UPs were purified on C18 Bond Elut cartridge (Varian). After elution with methanol and chloroform/methanol (2:1, v/v), glycolipids were dried under N_2. LP and UP glycolipids were then taken in chloroform/methanol (2:1, v/v) in a quantitative manner: 50 μl for 1×10^8 cells. Glycolipids were separated on High-Performance TLC plates (EM Science Merck) using a solvent system of chloroform/methanol/water containing 0.05% $CaCl_2$ (60:35:8, v/v/v). For chemical detection of glycolipids, TLC plates were sprayed with 0.5% orcinol in 10% sulfuric acid and then heated at 120 °C for 10 min. For TLC immunostaining, dried plates were blocked for 2 h with 5% bovine serum albumin (BSA) in phosphate-buffered saline (PBS) at room temperature and reacted overnight at 4 °C with biotinylated PA-IL (2.5 μg/ml in PBS 0.5% BSA). After washing, plates were incubated for 1 h with ^{125}I-labeled streptavidin (GE Healthcare) (0.2 μCi/ml in PBS 0.5% BSA). Plates were washed, dried, and submitted to autoradiography.

Protein crystallization and data collection

Crystals were obtained by the hanging-drop vapor diffusion method using 2-μl drops containing a 50:50 (v/v) mix of protein and reservoir solution at 20 °C. Lyophilized protein was dissolved in water (10 mg ml^{-1}) and incubated for 1 h with αGal1–β3Gal1–4Glc (0.297 mM) at room temperature prior to co-crystallization. Crystals of the complex were obtained after optimization of condition 23 of the Clear Strategy Screen II (Molecular Dimension

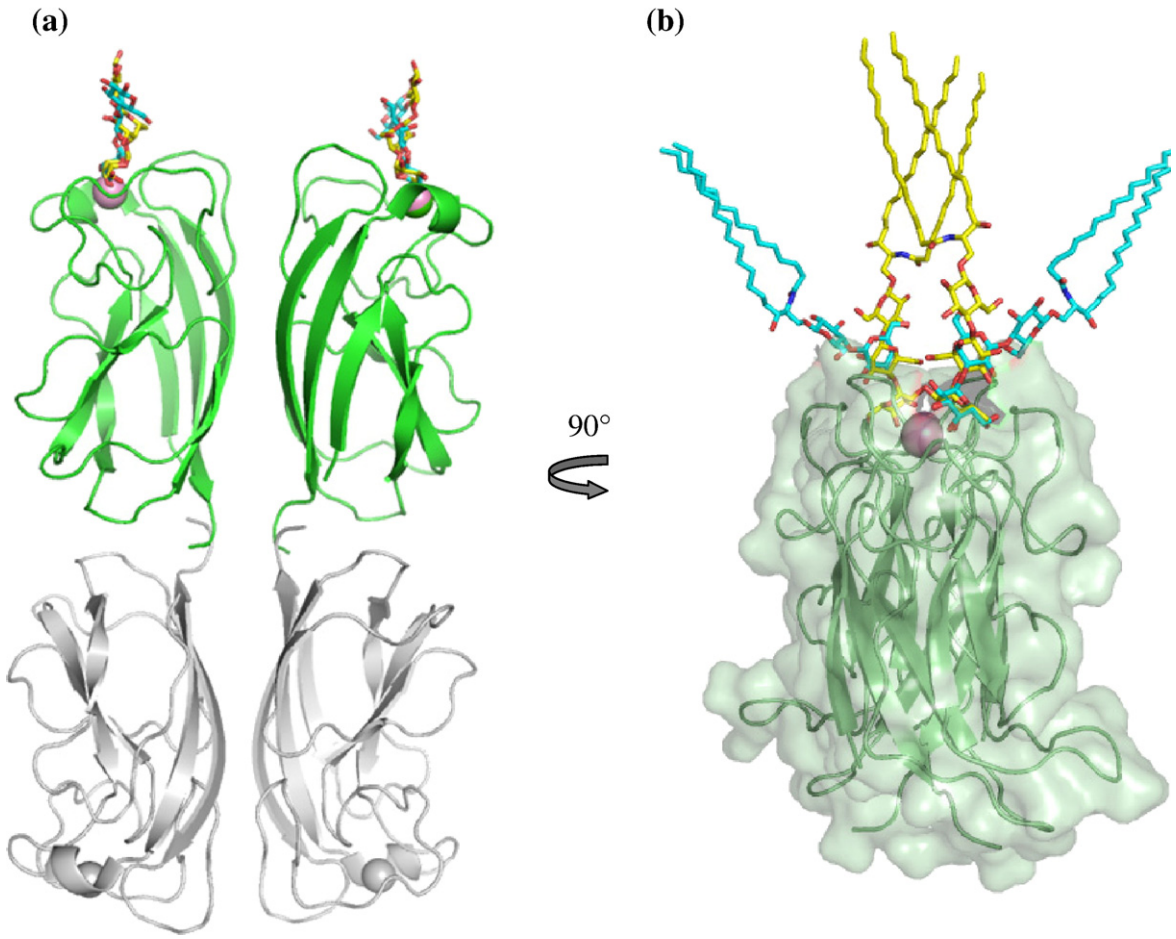


Fig. 9. (a) Superimposition of αGal1–3βGal1–4Glc as observed in the crystal structure (blue sticks) and αGal1–4βGal1–4Glc as predicted by modeling (yellow sticks). The tetramer of PA-IL is represented by a ribbon; calcium, by a pink sphere. (b) Representation of one PA-IL dimer with superimposition of docked αGal1–3βGal1–4βGlc-Cer (blue sticks) and αGal1–4βGal1–4βGlc-Cer (yellow sticks) with modeled lipid moiety.

Limited) using 10% polyethylene glycol 5KMME, 25 mM KSCN, and 100 mM NaAc, pH 4.6. Diffraction data were collected at 100 K, and 30% ethylene glycol was used as cryoprotectant. Crystals belong to space group *P*1 with 24 subunits per asymmetric unit. Data were collected at the European Synchrotron Radiation Facility (Grenoble, France) at station ID14-1 and on an ADSC Quantum 210 CCD detector. The data were processed using MOSFLM[45] and scaled and converted to structure factors using SCALA. All further computing was performed using the CCP4 suite unless otherwise stated.[46]

Structure solution and refinement

Molecular replacement technique was used to solve the structure with PHASER,[47] using the tetrameric coordinates of PA-IL structure with calcium (Protein Data Bank code 1L7L) as the search model. Six tetramers were found, but since their 222-fold symmetry has been broken, the position of the 24 monomers was optimized by rigid-body refinement. Five percent of the observations were set aside for cross-validation analysis,[48] and hydrogen atoms were added in their riding positions and used for geometry and structure-factor calculations. The structure was refined by restrained maximum-likelihood refinement using REFMAC[49] iterated with manual rebuilding in Coot.[50] The incorporation of the ligand was performed after inspection of the $mF_o - DF_c$ weighted maps, and the initial maps revealed clear density for at least one or two galactose residues of the αGal1–3βGal1–4Glc moiety. Water molecules were introduced automatically using Coot and inspected manually. The stereochemical quality of the model was assessed with the program Procheck,[51] and coordinates have been deposited in the Protein Data Bank under code 2VXJ. Molecular drawings were prepared using PyMOL Molecular Graphics System (DeLano Scientific, Palo Alto, CA).

Glycan microarray analysis

PA-IL was labeled with Alexa Fluor 488-TFP (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions and purified on a D-salt polyacrylamide desalting column (Pierce, Rockford, IL). Alexa-labeled PA-IL was used to probe on the glycan array version 3.8. The concentration of PA-IL used was 30 μg/ml and followed the standard procedure of Core H of the Consortium for Functional Glycomics‡.

‡ http://www.functionalglycomics.org/

Microcalorimetry

Purified and lyophilized PA-IL was dissolved in buffer (0.1 M Tris–HCl buffer containing 3 μM $CaCl_2$, pH 7.5) at a concentration of 0.05 mM and degassed. Protein concentration was checked by measurement of optical density using a theoretical molarity extinction coefficient of 28,000 (1 cm). Carbohydrate ligands were dissolved directly into the same buffer at a concentration of 1.7 mM, degassed, and placed in the injection syringe. Isothermal titration calorimetry was performed with a VP-ITC MicroCalorimeter from MicroCal Incorporated. PA-IL was placed into the 1.4478-ml sample cell, at 25 °C, using 10-μl injections of carbohydrate every 300 s. Carbohydrate ligand was also titrated into buffer alone. Data were fitted with MicroCal Origin 7 software, according to standard procedures. Fitted data yielded the 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:

$$\Delta G = \Delta H - T\Delta S = RT\ln K_a$$

where T is the absolute temperature and $R = 8.314\ \mathrm{J \cdot mol^{-1} \cdot K^{-1}}$. Two to three independent titrations were performed for each ligand tested.

Molecular mechanics calculations of disaccharides

Adiabatic maps were calculated for αGal1–3βGal and αGal1–4βGal disaccharides taking into account the glycosidic linkages, defined by the torsion angles φ = O5–C1–O1–CX, ψ = C1–O1–CX–CX + 1, and ω = O5–C5–C6–O6 through a rotation in 20° increments over the whole angular range. The 16 individual relaxed maps were computed for each disaccharide with different starting geometries of the pendent groups: two staggered positions of the hydroxymethyl groups (*gt* and *tg*) and the *clockwise* and *counterclockwise* orientations of the secondary hydroxyl groups.

At each step of the conformational search, geometry optimization of the disaccharide was performed applying the MM3 force field[52] that has been demonstrated to be well adapted to carbohydrate specificity.[53] The structure relaxation was performed using a block diagonal method with the convergence termination criterion of $n \times 0.00008$ kcal/mol per five iterations, where n is the number of atoms. A dielectric constant of 78.5 was used in all the calculations in order to reproduce an aqueous environment. The isoenergy contour maps were then visualized using the program XFarbe.[54]

Docking of disaccharides in PA-IL

Automated docking simulations were conducted with the AutoDock 3.05 suite of programs.[55] The crystallographic structure of one monomer of PA-IL was used for the preparation of the receptor input file in Sybyl 7.3 (Tripos Associates, St. Louis, MO). Hydrogen atoms were added in the structure, and positions of atoms were optimized through an energy minimization with TRIPOS force field.[56] Disaccharide ligands were constructed with Sybyl, and charges were assigned according to the PIM parameters for the TRIPOS force field.[57]

Receptor atomic solvation parameters and fragmental volumes were assigned using the program *addsol* included in AutoDock. Hydrogen bonds and van der Waals interactions were modeled using 12-10 and 12-6 Lennard–Jones parameters, respectively. Calcium atom was defined as a new atom type as recently described.[31] Electrostatic grid maps with a grid spacing of 0.375 Å and 60 grid points were centered on the ligand. The Lamarckian genetic algorithm and the pseudo-Solis and Wets methods were applied using default parameters. The number of energy evaluations was set to 1 million in all docking jobs. The 20 best docking poses for the ligand αGal1–3βGal and the 100 best docking conformations for the ligand αGal1–4βGal were taken into account for calculating the clustering histograms generated. Building of sphingolipids based on the best docking solutions was performed as recently described.[58]

MD simulations

Ten-nanosecond MD simulations of the disaccharides αGal1–3βGal and αGal1–4βGal were performed using AMBER 8 package (University of California). The starting structures of these simulations were a PA-IL monomer complexed with a calcium ion and a disaccharide as derived from AutoDock results, using the lowest energy conformations that always belong to the most populated cluster from the histograms. For the simulation, the AMBER force field parm99[59] was used for the lectin, while for carbohydrates, parameters were taken from the GLYCAM06 force field,[60] with partial charges calculated at the HF/6-31G* level followed by an RESP fitting. The calcium ion was considered as an individual atom with two positive charges, a van der Waals radius of 1.79 Å, and a well depth of 0.014 kcal/mol.[61]

The Xleap module of AMBER was used for preparing the input files, including the addition of hydrogen atoms, the electrostatic neutralization, and the solvation of the systems. The PA-IL/αGal1–3βGal and PA-IL/αGal1–4βGal systems were immersed in a bath of TIP3P water molecules to a depth of 10 Å. The equilibration of the system was carried out through energy minimization of water molecules (500 steps of steepest descendent and 500 steps of conjugate gradient) with restraints on the solute atoms followed by 80-ps-long MD simulations, warming the system to 298 K. The equilibration phase continued with an energy minimization of the total systems without restraints. The systems were warmed from 10 to 298 K during 70 ps of MD followed by 130 ps of dynamics at constant temperature and constant pressure of 1 atm.

The MD production phase was developed during 10 ns under a constant pressure of 1 atm and a constant temperature of 298.15 K controlled by the Langevin thermostat with a collision frequency of 1.0 $\mathrm{ps^{-1}}$. During the simulations, SHAKE algorithm[62] was turned on and applied to all hydrogen atoms and the particle-mesh Ewald method was used for treating the electrostatic interactions, with a cutoff of 10 Å. An integration time step of 2 fs was employed.

Minimization, equilibration, and production phases were carried out by SANDER module, while the analyses of the simulations were performed using the Ptraj module of AMBER 8. The visualization of the trajectories was performed using VMD software.[63] Data processing and two-dimensional plots were created using Scilab and Xmrg software.

Accession number

Coordinates and structure factors have been deposited in the Protein Data Bank with accession number 2VXJ.

Acknowledgements

This work was supported by CNRS, the French Ministry of Research, the Ministry of Education of the Czech Republic (MSM0021622413), and the EEC European Community through programs MEST-CT-2004-503322 (CermavTrain) and MRTN-CT-2006-035546 (NODPERCEPTION). The glycan array resources were provided by the Consortium for Functional Glycomics through grant GM62116. Funds in support of this work from the Association pour la Recherche sur le Cancer (ARC 3454), the Association Vaincre la Mucoviscidose, and the GDR Pseudomonas are gratefully acknowledged. We also acknowledge the European Synchrotron Radiation Facility for access to synchrotron data collection facilities.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2008.08.028

References

1. Gilboa-Garber, N. (1982). *Pseudomonas aeruginosa* lectins. *Methods Enzymol.* **83**, 378–385.
2. Imberty, A., Wimmerova, M., Mitchell, E. P. & Gilboa-Garber, N. (2004). Structures of the lectins from *Pseudomonas aeruginosa*: insights into molecular basis for host glycan recognition. *Microb. Infect.* **6**, 222–229.
3. Gilboa-Garber, N., Mizrahi, L. & Garber, N. (1972). Purification of the galactose-binding hemagglutinin of *Pseudomonas aeruginosa* by affinity column chromatography using Sepharose. *FEBS Lett.* **28**, 93–95.
4. Avichezer, D., Katcoff, D. J., Garber, N. C. & Gilboa-Garber, N. (1992). Analysis of the amino acid sequence of the *Pseudomonas aeruginosa* galactophilic PA-I lectin. *J. Biol. Chem.* **267**, 23023–23027.
5. Cioci, G., Mitchell, E. P., Gautier, C., Wimmerova, M., Sudakevitz, D., Pérez, S. *et al.* (2003). Structural basis of calcium and galactose recognition by the lectin PA-IL of *Pseudomonas aeruginosa*. *FEBS Lett.* **555**, 297–301.
6. Karaveg, K., Liu, Z. J., Tempel, W., Doyle, R. J., Rose, J. P. & Wang, B. C. (2003). Crystallization and preliminary X-ray diffraction analysis of lectin-1 from *Pseudomonas aeruginosa*. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **59**, 1241–1242.
7. Drickamer, K. (1996). Ca(2+)-dependent sugar recognition by animal lectins. *Biochem. Soc. Trans.* **24**, 146–150.
8. Hatakeyama, T., Unno, H., Kouzuma, Y., Uchida, T., Eto, S., Hidemura, H. *et al.* (2007). C-type lectin-like carbohydrate recognition of the hemolytic lectin CEL-III containing ricin-type β-trefoil folds. *J. Biol. Chem.* **282**, 37826–37835.
9. Schuster, M., Lostroh, C. P., Ogi, T. & Greenberg, E. P. (2003). Identification, timing, and signal specificity of *Pseudomonas aeruginosa* quorum-controlled genes: a transcriptome analysis. *J. Bacteriol.* **185**, 2066–2079.
10. Winzer, K., Falconer, C., Garber, N. C., Diggle, S. P., Camara, M. & Williams, P. (2000). The *Pseudomonas aeruginosa lectins* PA-IL and PA-IIL are controlled by quorum sensing and by RpoS. *J. Bacteriol.* **182**, 6401–6411.
11. Bajolet-Laudinat, O., Girod-de Bentzmann, S., Tournier, J. M., Madoulet, C., Plotkowski, M. C., Chippaux, C. & Puchelle, E. (1994). Cytotoxicity of *Pseudomonas aeruginosa* internal lectin PA-I to respiratory epithelial cells in primary culture. *Infect. Immun.* **62**, 4481–4487.
12. Laughlin, R. S., Musch, M. W., Hollbrook, C. J., Rocha, F. M., Chang, E. B. & Alverdy, J. C. (2000). The key role of *Pseudomonas aeruginosa* PA-I lectin on experimental gut-derived sepsis. *Ann. Surg.* **232**, 133–142.
13. Kirkeby, S., Wimmerova, M., Moe, D. & Hansen, A. K. (2007). The mink as an animal model for *Pseudomonas aeruginosa* adhesion: binding of the bacterial lectins (PA-IL and PA-IIL) to neoglycoproteins and to sections of pancreas and lung tissues from healthy mink. *Microbes Infect.* **9**, 566–573.
14. Diggle, S. P., Stacey, R. E., Dodd, C., Camara, M., Williams, P. & Winzer, K. (2006). The galactophilic lectin, LecA, contributes to biofilm development in *Pseudomonas aeruginosa*. *Environ. Microbiol.* **8**, 1095–1104.
15. Ma, L., Lu, H., Sprinkle, A., Parsek, M. R. & Wozniak, D. J. (2007). *Pseudomonas aeruginosa* Psl is a galactose- and mannose-rich exopolysaccharide. *J. Bacteriol.* **189**, 8353–8356.
16. Garber, N., Guempel, U., Belz, A., Gilboa-Garber, N. & Doyle, R. J. (1992). On the specificity of the D-galactose-binding lectin (PA-I) of *Pseudomonas aeruginosa* and its strong binding to hydrophobic derivatives of D-galactose and thiogalactose. *Biochim. Biophys. Acta*, **1116**, 331–333.
17. Chen, C. P., Song, S. C., Gilboa-Garber, N., Chang, K. S. & Wu, A. M. (1998). Studies on the binding site of the galactose-specific agglutinin PA-IL from *Pseudomonas aeruginosa*. *Glycobiology*, **8**, 7–16.
18. Wiels, J. & Tursz, T. (1995). CD77 Workshop Panel report. In *Leukocyte Typing V* (Schlossman, S. F., Boumsell, L. & Gilks, W., eds), pp. 597–559. Oxford University Press, Oxford, UK.
19. Gilboa-Garber, N., Sudakevitz, D., Sheffi, M., Sela, R. & Levene, C. (1994). PA-I and PA-II lectin interactions with the ABO(H) and P blood group glycosphingolipid antigens may contribute to the broad spectrum adherence of *Pseudomonas aeruginosa* to human tissues in secondary infections. *Glycoconjugate J.* **11**, 414–417.
20. Lanne, B., Ciopraga, J., Bergstrom, J., Motas, C. & Karlsson, K. A. (1994). Binding of the galactose-specific *Pseudomonas aeruginosa* lectin, PA-I, to glycosphingolipids and other glycoconjugates. *Glycoconjugate J.* **11**, 292–298.
21. Nudelman, E., Kannagi, R., Hakomori, S., Parsons, M., Lipinski, M., Wiels, J. *et al.* (1983). A glycolipid antigen associated with Burkitt lymphoma defined by a monoclonal antibody. *Science*, **220**, 509–511.
22. Wiels, J., Holmes, E. H., Cochran, N., Tursz, T. & Hakomori, S. (1984). Enzymatic and organizational difference in expression of a Burkitt lymphoma-associated antigen (globotriaosylceramide) in Burkitt lymphoma and lymphoblastoid cell lines. *J. Biol. Chem.* **259**, 14783–14787.
23. Zhou, D., Mattner, J., Cantu, C., 3rd, Schrantz, N., Yin, N., Gao, Y. *et al.* (2004). Lysosomal glycosphingolipid recognition by NKT cells. *Science*, **306**, 1786–1789.
24. Dam, T. K. & Brewer, C. F. (2002). Thermodynamic studies of lectin–carbohydrate interactions by isothermal titration calorimetry. *Chem. Rev.* **102**, 387–429.
25. Imberty, A., Mikros, E., Koca, J., Mollicone, R., Oriol, R. & Pérez, S. (1995). 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 J.* **12**, 331–349.
26. Martin-Pastor, M., Espinosa, J. F., Asensio, J. L. & Jiménez-Barbero, J. (1997). A comparison of the geometry and of the energy results obtained by application of different molecular mechanics force fields to methyl alpha-lactoside and the C-analogue of lactose. *Carbohydr. Res.* **298**, 15–49.
27. Imberty, A., Tran, V. & Pérez, S. (1989). Relaxed potential energy surfaces of N-linked oligosaccharides: the mannose-alpha-(1–3)-mannose case. *J. Comp. Chem.* **11**, 205–216.
28. Corzana, F., Bettler, E., Hervé du Penhoat, C., Tyrtysh, T. V., Bovin, N. V. & Imberty, A. (2002). Solution structure of two xeno-antigens: αGal-LacNAc and αGal-Lewis X. *Glycobiology*, **12**, 241–250.
29. Kuttel, M. M. (2008). Conformational free energy maps for globobiose (α-D-Galp-(1→4)-β-D-Galp) in implicit and explicit aqueous solution. *Carbohydr. Res.* **343**, 1091–1098.
30. Svensson, G., Albertsson, J., Svensson, C., Magnusson, C. & Dahmen, J. (1986). X-ray crystal structure of galabiose, *O*-α-D-galactopyranosyl-(1→4)-D-galactopyranose. *Carbohydr. Res.* **146**, 29–38.
31. Nurisso, A., Kozmon, S. & Imberty, A. (2008). 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.* **34**, 469–479.
32. Sudakevitz, D., Levene, C., Sela, R. & Gilboa-Garber, N. (1996). Differentiation between human red cells of P^k and p blood types using *Pseudomonas aeruginosa* PA-I lectin. *Transfusion*, **36**, 113–116.
33. Galili, U., Shohet, S. B., Kobrin, E., Stults, C. L. & Macher, B. A. (1988). Man, apes, and Old World monkeys differ from other mammals in the expression of alpha-galactosyl epitopes on nucleated cells. *J. Biol. Chem.* **263**, 17755–17762.
34. Kirkeby, S., Hansen, A. K., d'Apice, A. & Moe, D. (2006). The galactophilic lectin (PA-IL, gene LecA) from *Pseudomonas aeruginosa*. Its binding requirements and the localization of lectin receptors in various mouse tissues. *Microb. Pathog.* **40**, 191–197.
35. Zhou, D. (2006). The immunological function of iGb3. *Curr. Protein Pept. Sci.* **7**, 325–333.
36. Li, Y., Teneberg, S., Thapa, P., Bendelac, A., Levery, S. B. & Zhou, D. (2008). Sensitive detection of isoglobo and globo series tetraglycosylceramides in human thymus by ion trap mass spectrometry. *Glycobiology*, **18**, 158–165.
37. Fujii, Y., Numata, S., Nakamura, Y., Honda, T., Furukawa, K., Urano, T. *et al.* (2005). Murine glycosyltransferases responsible for the expression of globo-series glycolipids: cDNA structures, mRNA expression, and distribution of their products. *Glycobiology*, **15**, 1257–1267.
38. Kojima, Y., Fukumoto, S., Furukawa, K., Okajima, T., Wiels, J., Yokoyama, K. *et al.* (2000). Molecular cloning of globotriaosylceramide/CD77 synthase, a glycosyltransferase that initiates the synthesis of globo series glycosphingolipids. *J. Biol. Chem.* **275**, 15152–15156.
39. Mangeney, M., Richard, Y., Coulaud, D., Tursz, T. & Wiels, J. (1991). CD77: an antigen of germinal center B cells entering apoptosis. *Eur. J. Immunol.* **21**, 1131–1140.
40. Lingwood, C. A. (1999). Verotoxin/globotriaosyl ceramide recognition: angiopathy, angiogenesis and antineoplasia. *Biosci. Rep.* **19**, 345–354.
41. Tetaud, C., Falguieres, T., Carlier, K., Lecluse, Y., Garibal, J., Coulaud, D. *et al.* (2003). Two distinct Gb3/CD77 signaling pathways leading to apoptosis are triggered by anti-Gb3/CD77 mAb and verotoxin-1. *J. Biol. Chem.* **278**, 45200–45208.
42. Vingert, B., Adotevi, O., Patin, D., Jung, S., Shrikant, P., Freyburger, L. *et al.* (2006). The Shiga toxin B-subunit targets antigen *in vivo* to dendritic cells and elicits anti-tumor immunity. *Eur. J. Immunol.* **36**, 1124–1135.
43. Imberty, A., Chabre, Y. M. & Roy, R. (2008). Glycomimetics and glycodendrimers as high affinity microbial antiadhesins. *Chem. Eur. J.* **14**, 7490–7499.
44. Sharon, N. (2006). Carbohydrates as future anti-adhesion drugs for infectious diseases. *Biochim. Biophys. Acta*, **1760**, 527–537.
45. Leslie, A. G. W. (1992). Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4/ESF-EACMB Newsletter on Protein Crystallography*, **26**.
46. Collaborative Computational Project Number 4. (1994). The CCP4 Suite: programs for protein crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **50**, 760–763.
47. McCoy, A. J. (2007). Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **63**, 32–41.
48. Brünger, A. T. (1992). Free *R*-value—a novel statistical quantity for assessing the accuracy of crystal-structures. *Nature*, **355**, 472–475.
49. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **53**, 240–255.
50. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **60**, 2126–2132.
51. Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. (1993). Procheck—a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291.
52. Allinger, N. L., Yuh, Y. H. & Lii, J.-H. (1989). Molecular mechanics. The MM3 force field for hydrocarbons. *J. Am. Chem. Soc.* **111**, 8551–8566.
53. Pérez, S., Imberty, A., Engelsen, S. B., Gruza, J., Mazeau, K., Jiménez-Barbero, J. *et al.* (1998). A comparison and chemometric analysis of several molecular mechanics force fields and parameter sets applied to carbohydrates. *Carbohydr. Res.* **314**, 141–155.
54. Preusser, A. (1989). Algorithm 671: FARB-E-2D: fill area with bicubics on rectangles—a contour plot program. *ACM Trans. Math. Software*, **15**, 79–89.
55. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.* **19**, 1639–1662.
56. Clark, M., Cramer, R. D. I. & van den Opdenbosch, N. (1989). Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **10**, 982–1012.
57. Imberty, A., Bettler, E., Karababa, M., Mazeau, K., Petrova, P. & Pérez, S. (1999). Building sugars: the sweet part of structural biology. In *Perspectives in Structural Biology* (Vijayan, M., Yathindra, N. & Kolaskar, A. S., eds), pp. 392–409. Indian Academy of Sciences and Universities Press, Hyderabad, Andhra Pradesh.
58. Campanero-Rhodes, M. A., Smith, A., Chai, W., Sonnino, S., Mauri, L., Childs, R. A. *et al.* (2007). *N*-glycolyl GM1 ganglioside as a receptor for Simian virus 40 (SV40). *J. Virol.* **81**, 12846–12858.

59. Wang, J., Cieplak, P. & Kollman, P. A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comp. Chem.* **21**, 1049–1074.
60. Kirschner, K. N., Yongye, A. B., Tschampel, S. M., Gonzalez-Outeirino, J., Daniels, C. R., Foley, B. L. & Woods, R. J. (2008). GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **29**, 622–655.
61. Bradbrook, G. M., Gleichmann, T., Harrop, S. J., Habash, J., Raftery, J., Kalb, J. *et al.* (1998). X-ray and molecular dynamics studies of concanavalin-A glucoside and mannoside complexes: relating structure to thermodynamics of binding. *J. Chem. Soc., Faraday Trans.* **94**, 1603–1611.
62. Ryckaert, J. P., Cicotti, G. & Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comp. Chem.* **23**.
63. Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD—Visual Molecular Dynamics. *J. Mol. Graphics*, **14**, 33–38.

JBC Papers in Press. Published on April 21, 2010 as Manuscript M110.108340
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.M110.108340

ROLE OF WATER MOLECULES IN STRUCTURE AND ENERGETICS OF *PSEUDOMONAS AERUGINOSA* PA-IL LECTIN INTERACTING WITH DISACCHARIDES

Alessandra Nurisso[†], Bertrand Blanchard[†], Aymeric Audfray[†], Lina Rydner[‡], Stefan Oscarson[‡], Annabelle Varrot[†], Anne Imberty[†]*

From [†]CERMAV-CNRS (affiliated to Université Joseph Fourier and ICMG), BP 53, 38041 Grenoble cedex 9, France, [‡]Centre for Synthesis and Chemical Biology, UCD School of Chemistry and Chemical Biology, University College Dublin, Belfield, Dublin 4, Ireland

Running head: Water in lectin/sugar complexes

Address correspondence to : Dr. Anne. Imberty, CERMAV-CNRS, BP53, 38041 Grenoble cedex 09, France; Tel: +33-476037636; Fax: +33-476547203; email: Imberty@cermav.cnrs.fr

The calcium-dependent lectin I from *Pseudomonas aeruginosa* (PA-IL) binds specifically to oligosaccharides presenting an α-galactose residue at their non-reducing end, such as the disaccharides αGal1-2βGalOMe, αGal1-3βGalOMe and αGal1-4βGalOMe. This provides a unique model for studying the effect of the glycosidic linkage of the ligands on structure and thermodynamics of the complexes by means of experimental and theoretical tools. The structural features of PA-IL in complex with the three disaccharides were established by docking and molecular dynamics (MD) simulations and compared to those observed in available crystal structures, including PA-IL/ αGal1-2βGalOMe complex, that was solved at 2.4 Å resolution and reported herein. The role of a structural bridge water molecule in the binding site of -PAIL was also elucidated through MD simulations and molecular mechanics Poisson-Bolzmann surface area MM-PBSA approach. This water molecule establishes three very stable hydrogen bonds with O6 of non-reducing galactose, oxygen from Pro51 main chain and nitrogen from Gln53 main chain of the lectin binding site. Binding free energies for PA-IL in complex with the three disaccharides were investigated and the results were compared with the experimental data determined by titration microcalorimetry. When the bridge water molecule was included in the MM-PBSA calculations, the simulations predicted the correct binding affinity trends with the 1-2 linked disaccharide presenting three times stronger affinity ligand than the two other ones. These results highlight the role of the water molecule in the binding site of PA-IL and indicate that it should be taken into account when designing glycoderivatives active against *P. aeruginosa* adhesion.

Lectins are carbohydrate-binding proteins of non-immune origin with no enzymatic activity, responsible for selective glycan recognition in bacteria, animals and plants (1,2) Due to their specificities toward sugars, lectins are involved in many biological processes such as cell-cell communication, embryogenesis, cell maturation and can also play a role in pathology including tumor growth and host-pathogen interactions (2).

The opportunistic Gram negative bacterium *Pseudomonas aeruginosa* is the major mortality factor for cystic fibrosis patient, causing endobronchial infections and related neutrophilic inflammatory responses (3). The bacterium produces two soluble lectins, Lectin I and Lectin II from *Pseudomonas aeruginosa* (PA-IL and PA-IIL), located in its cytoplasm and on its outer membrane (4). These lectins are expressed under the control of the quorum sensing system and are considered as virulence factors that have been proposed to be involved in adhesion to glycoconjugates on respiratory epithelia and biofilm formation (5,6).

PA-IIL is a tetrameric fucose-binding lectin, each monomer containing two calcium ions in the binding site with micro molar affinity toward carbohydrate ligands that participate in the ions coordination. The structures and thermodynamic properties of this lectin were elucidated through experimental and *in silico* techniques (7,8). PA-IL has not sequence similarity with PA-IIL and associates has a different tetramer with specificity to galactose (9). Solved crystal structures, in apo form and in complex with sugars, demonstrated that each

monomer adopts a small jelly-roll type β-sandwich fold, consisting of two curved sheets, each one formed by four antiparallel β-strands with a calcium ion in the binding site (10). We previously described the structural basis and thermodynamic properties of this lectin in complex with oligosaccharide moieties of glycosphingolipids combining several approaches such as cell surface labeling, glycan array analysis, titration microcalorimetry, crystallography and molecular modeling (11). Together with previous binding studies (12,13), these data established that the globotriaosylceramide antigen Gb3, which carbohydrate moiety is αGal1-3βGal1-4βGlc, is a likely natural ligand of the lectin on human epithelia.

The aim of the present work is to unravel the structural and thermodynamic features of the PA-IL lectin in complex with three isomeric disaccharides, methyl 2-*O*-α-D-galactopyranosyl-β-D-galactopyranoside (αGal1-2βGalOMe), methyl 3-*O*-α-D-galactopyranosyl-β-D-galactopyranoside (αGal1-3βGalOMe) and methyl 4-*O*-α-D-galactopyranosyl-β-D-galactopyranoside (αGal1-4βGalOMe) that all bind to the lectin and only differ by the stereochemistry of their glycosidic linkage. We used flexible molecular docking on one lectin monomer coupled with explicitly solvated molecular dynamics (MD) simulations to evaluate the structure and flexibility of the binding sites together with the role of the solvent. Molecular Mechanics Poisson-Bolzmann Surface Area (MM-PBSA) analyses were used on the calculated trajectories to evaluate the free energy of binding of the complexes. The *in silico* structural and energetic observations were compared to experimental data. In particular, the crystal structure of PA-IL in complex with αGal1-2βGalOMe was solved at 2.4 Ă resolution and thermodynamic of PA-IL binding to the three oligosaccharides was measured by titration microcalorimetry. The modeling data, in agreement with the experimental one, showed the importance of one structural bridge water molecule always present in the PA-IL binding site.

MATERIAL AND METHODS

Material

The recombinant protein PA-IL cloned in plasmid pET25pa1l was expressed and purified as described previously (11). The purified protein was lyophilized and stored at -20°C. Oligosaccharide derivatives αGal1-2βGalOMe and αGal1-3βGalOMe were purchased from Carbohydrate Synthesis (Carbohydrate Synthesis, Oxford, UK). αGal1-4βGalOMe was synthesized following published procedures (14,15), the structure was checked by NMR and found to be identical to published data.

Microcalorimetry

Lyophilized PA-IL was dissolved in buffer (0.1MTris–HCl buffer containing 3 μM $CaCl_2$, pH 7.5) at a concentration of 0.05 mM and degassed. Protein concentration was checked by measurement of optical density at 280 nm using a theoretical molarity extinction coefficient of 27,600 (1 cm). Carbohydrate ligands were dissolved directly into the same buffer at concentration range from 0.07 to 1.7 mM and degassed. Isothermal titration calorimetry was performed with a VP-ITC MicroCalorimeter (MicroCal Inc.). Titration was performed on PA-IL in the 1.4478 ml sample cell using 10 μl injections of carbohydrate every 300 s at 25 °C. Carbohydrate ligand was also titrated into buffer for control. Data were fitted with MicroCal Origin 7 software, according to standard procedures using a single-site model with a stochiometry of 1. Fitted data yielded the Ka and the enthalpy of binding (ΔH). Other thermodynamic parameters (i.e., changes in free energy, ΔG, and entropy, ΔS) were calculated from the equation:

$$\Delta G = \Delta H - T\Delta S = RT\ln K_a$$

Where T is the absolute temperature and $R=8.314 J \cdot mol^{-1} \cdot K^{-1}$. Two independent titrations were performed for each ligand tested.

Protein crystallography

Crystals of PA-IL complexed with αGal1-2βGalOMe were obtained by the hanging-drop vapour diffusion method using 2-μl drops containing a 50:50 (v/v) mix of protein and reservoir solution at 20 °C. Lyophilized protein was dissolved in water (10 mg ml^{-1}) and incubated for 1 h with αGal12βGalOMe at room temperature prior to co-crystallization. Crystals of the complex were obtained after optimization of condition 23 of the Clear Strategy Screen II (Molecular Dimension Limited) using 20% polyethylene glycol 6K, 1M Lithium chloride, and 100 mM sodium acetate, pH 4. Crystals were frozen in liquid nitrogen in the presence of 25% ethylene glycol as cryoprotectant.

Diffraction data were collected at 100 K, at the European Synchrotron Radiation Facility (Grenoble, France) at station ID14-eh2 using an ADSC Q4 CCD detector. Crystals belong to space group $P2_1$ and diffracted to 2.4 Å resolution. The data were processed using MOSFLM (16) and scaled and converted to structure factors using SCALA. All further computing was performed using the CCP4 suite (17) unless otherwise stated.

Molecular replacement technique was used to solve the structure with PHASER (18), using the A chain of PA-IL/galactose complex (10), (pdb code 1OKO) (19) after removing of water and ligand molecules. Eight monomers were found, corresponding to two tetramers of PA-IL. Five percent of the observations were set aside for cross-validation analysis (20), and hydrogen atoms were added in their riding positions and used for geometry and structure-factor calculations. The structure was refined by restrained maximum-likelihood refinement using REFMAC (21), iterated with manual rebuilding in Coot (22). The incorporation of the ligand was performed after inspection of the mFo −DFc weighted maps, and the initial maps revealed clear density for one or two galactose residues of the αGal12βGalOMe ligand, depending on the sites. Water molecules were introduced automatically using Coot and inspected manually. The stereochemical quality of the model was assessed with the program Procheck (23), and coordinates have been deposited in the Protein Data Bank under code 2WYF. Molecular drawings were prepared using PyMOL Molecular Graphics System (DeLano Scientific, Palo Alto, CA).

Flexible molecular docking:

To perform docking calculations, the x-ray structure of the monomer A of the lectin PA-IL was taken into account (2VXJ). The structure was edited using the molecular modeling package Sybyl (Tripos Associates, St. Louis, MO), removing crystallographic water molecules and heteroatoms, with the exception of the calcium ion in the binding site, and adding hydrogen atoms which position was optimized through energy minimization with the TRIPOS force field (24). The available X-ray coordinates of the trisaccharide αGal-4βGal1-4Glc were considered in this study and used as positional reference for the non-reducing galactose residue for all studied disaccharides.

Partial charges for the protein and carbohydrates atoms were calculated using the Kollman All parameters (25) and PIM parameters for the TRIPOS force field (26), respectively. The program *deftors* implemented in AutoDock3 (27) was used to define the rotational degrees of freedom for the ligands during the docking calculations. The PA-IL monomer was considered as a rigid body with receptor atomic solvation and fragmental volumes parameters assigned by the *addsol* utility. Calcium ion was treated using the method previously described (28). The electrostatic grid maps were centered on ligands and calculated using the *autogrid* tool with a grid spacing of 0.375 Å and 70 grid points that include an exhaustive space of the PA-IL binding site. Polar hydrogens were differentiated from non-polar hydrogens using 12-10 and 12-6 hydrogen bonding Lennard-Jones parameters, respectively. Flexible docking runs (100) were carried out with AutoDock 3.0, employing the Lamarckian GA combined with the Solis and Wets local search. Default parameters were used, except for the number of energy evaluations, set to 4,000,000 per GA run. The best docked structures for each run were collected, the ligand poses with the most favorable van der Waals and electrostatic interactions were clustered in a histogram and their coordinates extracted as pdb files.

MM3 energy maps of disaccharides

MM3 was used for the conformational analysis of each disaccharide by calculating Φ/Ψ relaxed potential energy maps. MM3 is considered one of the best force fields for carbohydrates since it allows full relaxation of the glycosidic residues taking into account the exo-anomeric effect (29,30). The notations used for the torsion angles of the glycosidic linkage are as follows: Φ=O5–C1–O1–CX, Ψ=C1–O1–CX–CX+1, and ω=O5–C5–C6–O6. Several starting conformations (16 altogether) were considered and generated using Sybyl software (Tripos Associates, St. Louis, MO) by considering possible orientations of the hydroxymethyl and secondary hydroxyl groups. Systematic searches for Φ/Ψ rotations were performed with a step of 20° at a dielectric constant of 80 to mimic an aqueous environment. The block-diagonal minimization method, with the default energy convergence criterion of 0.00008 * n kcal mol_1 per five iterations, n being the number of atoms, was

used for optimizations. Individual relaxed maps were created and then combined together to obtain adiabatic energy maps in which only the lowest energy conformer at each Φ/Ψpoint is considered. Energy contour plots were visualized with the program Xfarbe (31).

Molecular dynamics simulations

PA-IL monomer in complex with disaccharides derived from docking results were simulated using the AMBER program version 8 (University of California) with AMBER force field parm99 for the lectin (32) and Glycam06 force field (33) for carbohydrates. The calcium ion was treated with a charge of +2, a radius of 1.79 Å, and a well depth 0.014 kcal/mol (34). All systems were neutralized by Na^+ ions using the Xleap module of AMBER. A truncated octahedron of TIP3P water molecules was added to a distance of 10 Å on each side of the complexes. During the construction of the three systems with Xleap, the structural water molecule observed in the binding site of PA-IL crystal structures (10,11) was retained. The simulations were carried out using the particle mesh Ewald technique (35) with 10Å non bonded cutoff and 2 fs integration time step. SHAKE algorithm was applied to all hydrogen atoms to eliminate the fastest X-H vibrations for a longer simulation time step(36).

The equilibration protocol started with 1000 steps of minimization of water molecules and ions in order to allow water molecules to assume a lower energetic geometry whereas the solute was constrained. The resulting systems were then subjected to 5000 steps of minimization, 2000 in steepest descent, with no restraints, reaching a rms-gradient of 0.1 to assure the relaxation of the structures, followed by 100 ps of heating from 10K to 300K in constant volume ensemble with weak restraints on the solute (10kcal/(mol $Å^2$). Density equilibration (50 ps) followed by 500ps of constant pressure without restraints at 300K completed the equilibration step with convergences of energies, temperature, pressure, density of the systems. The production phase was carried out with constant pressure boundary conditions and constant temperature controlled by the Langevin thermostat. A total of 10 ns of MDs production were run and trajectories were collected recording the coordinates every 0.2 ps.

Analogue methodology was applied for 10ns-MDs simulations of the free disaccharides and the PA-IL monomer in a non bounded state. Trajectories were analyzed using the Ptraj module of AMBER and R program and visualized using the VMD molecular visualization program (37). Pictures were generated using Pymol (DeLano Scientific, Palo Alto, CA) and Chimera (University of California, San Francisco, CA).

Calculation of free energy of binding by MM-PBSA

The free energy of binding was evaluated using the MM-PBSA method (38) implemented in AMBER10 (University of California). This approach is based on the generation of multiple structures generally from a single MDs trajectory, representing the ligand, the receptor and the protein-ligand complex, respectively. Usually, water molecules and counterions are removed from every snapshot. Snapshots (1000) equally spaced at 10 ps intervals were generated from the three trajectories calculated for PA-IL monomer in complex with αGal1-2βGalOMe, αGal1-3βGalOMe and αGal1-4βGalOMe. The same procedure was repeated while including the crystallographic bridge water molecule, in order to investigate its role in the binding energetic. Snapshots were also generated from the trajectories of the disaccharides in water and PA-IL in water. Free energy of binding was then calculated as described (38). The solvation free energy takes into account the non polar contribution calculated from the solvent accessible surface area A (39) and the electrostatic contribution estimated solving the Poisson–Boltzmann equation via the pbsa module implemented in AMBER. The PB equation was solved using a grid spacing of 0.375 Å for the cubic lattice; the solvent dielectric constant was set to 80.0, the internal dielectric constant was set to 1.0, the solvent radius was set to 1.4 Å, no counterions were included. In addition, the contribution of each monosaccharides to the binding energy was evaluated by performing the same calculations while omitting one monosaccharide or the other.

$T\Delta S$ represents the energy cost of vibrational, translational, rotational and conformational changes in solute during the complex formation. Due to the high computational cost of this calculation and its approximate nature, it is often omitted from the free energy of binding estimation (40). 100 snapshots from each trajectory were used to calculate the entropy through a normal mode analysis, using the Nmode module implemented

in AMBER. Before the entropic estimation the structure of each snapshot was subjected to 1000 cycle of minimization, setting the distance-dependent dielectric function to 4 to reproduce the impact of the water environment.

RESULTS AND DISCUSSION

Crystal structure of PA-IL/αGal12βGalOMe disaccharide

Co-crystals of the lectin and αGal1-2βGal-O-Met disaccharide were obtained in space group $P2_1$ with cell dimensions of *a*=49.9 Å, *b*=99.8 Å, *c*=91.3 Å, and γ=100.8° (Table 1). The asymmetric unit consists of eight PA-IL monomers arranged in two tetramers, each tetramer is assembled by pseudo 222-symmetry (Figure 1A). Clear density for a calcium ion and a disaccharide was observed in all binding sites (Figure 1B), except in chain B and E where only the calcium ion and the contacting galactose residue can be located. This resulted in the refinement of 968 amino acids, 648 water molecules, 8 calcium ions, and 14 carbohydrate residues with an Rcrys of 17% and an Rfree of 26% at 2.4 Å resolution (Table 1).

As previously described (10,11), each monomer adopts a small β-sandwich fold consisting of two curved sheets, each consisting of four antiparallel β-strands. The non-reducing αGal residue is buried in the binding site and participates in the coordination of the calcium ion through oxygen atoms O3 and O4. These two atoms also establish hydrogen bonds with Tyr36, Asp100, Thr104 and Asn107 (Figure 1C and Table S1 of supplemental material). The position of the sugar is also stabilized by oxygen O2, which creates a hydrogen bond with the nitrogen atom of Asn107 and by oxygen O6 oxygen that interacts with the side chains of Glu53 and His50. The hydroxymethyl group that carries this oxygen always adopts a *gauche-trans* orientation (O5-C5-C6-O6 = 72° (+/-15°). Oxygen O6 also participates in a water-bridged contact through a water molecule that is conserved in all monomers. This structural water molecule makes hydrogen bonds with the Pro51 main chain oxygen and Glu53 main chain nitrogen. The hydrophobic contacts are rather limited since only the CH group at C2 interacts with the side chain of Tyr61. The interaction of non-reducing galactose including calcium ion and water molecule appears to be very similar to observations in the structures of PA-IL complexed with galactose (10) or αGal14αGal14Glc.

The reducing galactose is located in the solvent exposed area at the surface of the site. In all binding sites, it establishes two additional hydrogen bonds with the protein through contacts between oxygen O4 and the nitrogen of Gln53 and between the O3 galactose atom and oxygen of Gln53. The conformation at the glycosidic linkage between the galactose residues does not exhibit much variations with Φ angle adopting average value of 59° (+/-9°) in agreement with the exo-anomeric effect and the Ψ angle a value of -154° (+/-6°).

Microcalorimetry data

The interaction between PA-IL and the three disaccharide derivatives were quantified by titration microcalorimetry, a method that is well suited for protein–carbohydrate interactions (41). Since the dissociation constant of PA-IL towards αGal-containing trisaccharides were previously reported to be in the order of 100 μM, the titration was performed with an excess of ligand as recommended for low affinity systems (42) and the stochiometry was fixed to 1, taking into account data from crystallography. Titration curves for PA-IL binding to αGal1–2 βGalOMe, reported in Figure 2A, displays large exothermic peaks characteristics of exothermic, i.e. enthalpy driven interaction. After integration of data (Figure 2B), association and dissociation constants as well as thermodynamics contribution have been calculated (Table2). K_d Values are very similar for αGal1-3 βGalOMe and αGal1-4 βGalOMe (132 and 115 μM respectively) and are close to data obtained with trisaccharides αGal1-3βGal1-4Glc and αGal1-4βGal1-4Glc (11). Interestingly, the affinity is three times stronger for αGal1-2βGalOMe (K_d=27μM) than for the other disaccharides. For all disaccharides, the interaction is enthalpy driven with unfavorable entropy contribution that varies from 29 to 62 % of the free energy of binding (Figure 2C and Table 2).

Docking of disaccharides in the binding site of PA-IL

The Autodock 3 software was used for docking the three disaccharides into the PA-IL pocket since this approach was previously validated for sugar-protein interactions (43,44), including the case of bridging calcium ion (28). One hundred runs were calculated for each disaccharide with all rotatable bonds considered

as flexible. For each disaccharide, the cluster with higher population is listed in Table 3. In all cases, these clusters are highly populated (between 72% and 97%) and always contain the lowest energy solution (Figure 3). The lowest energy docking modes all present the non-reducing galactose residue with expected coordination of calcium ion by oxygen atoms O3 and O4 and hydrogen bonds to Tyr36, Asp 100, Thr104 and Asn107 evaluated using the utility FindHBond implemented in Chimera. Only the O6 oxygen atom does not establish the expected hydrogen bond to His50 and Gln53. This appears to be due to the omega angle (O5-C6-C6-O6) that adopts a value close to 0° in the docking procedure that differs significantly from the gauche-trans conformation observed in the crystals.

In all docked disaccharides, the galactose residue at the reducing end establishes hydrogen bonds with the protein that involve His50 and Thr53 (Figure 3). In the system PA-1L/αGal12GalOMe, the reducing galactose is stabilized by a hydrogen bond network involving the oxygen O3 and O4 of the second monomer and the atoms OE1 and NE2 from the residue Gln53 (distances of 2.5 Å, 3.1 Å and 3.3 Å, 2.5 Å respectively). Concerning the reducing end of αGal13GalOMe, the oxygen O2 is involved in a hydrogen bond network with the atoms OE1, NE2 from the residue Gln53 (2.5 Å, 3.1 Å) and the atom N2 from His 50 (2.8 Å). Same conditions were found in the case of PA-1L/ αGal14GalOMe in which O2 and O3 from the second galactose monomer were involved in polar contacts with the atoms from the amino acids previously mentioned. In particular, the oxygen O2 establishes hydrogen bonds with the atoms OE1 and NE2 from the residue Gln53 (2.7 Å, 3.2 Å) while the oxygen O3 creates bonds with the atoms OE1, NE2 from the residue Gln53 (2.5 Å, 3.0 Å) and the atom N2 from His 50 (2.9 Å).

Docking results can be structurally compared with data derived from crystallography for the 1-2 and the 1-3 linked disaccharides. The conformations at the glycosidic linkages do not show large variations between docked and experimental structures: The αGal1-2Gal docked complex displays αGal1-2Gal docked conformation with Φ,Ψ values of (62.3°, 153.6°) that compare well to the average values of (60°, 154°) from the crystal structure described above while the αGal1-3Gal docked conformation varies from (95.9°, 135.1°) in the docking approach to average observed values of (98°, 121°) in the crystal structure of the complex between PA-IL and the trisaccharide (11). The low RMSD calculated for disaccharides between experimental and theoretical data for the disaccharides αGal1-2GalOMe and αGal1-3GalOMe as well as the coherent oxygen-metal ion distances and glycosidic angles values validate our docking approach and support the prediction of the αGal1–4GalOMe binding mode. However, the absence of the conserved water molecule clearly induced a change in conformation for O6 hydroxymethyl group and could therefore affect the behavior of His50 and Gln53 that are of strong importance for disaccharide binding.

Molecular dynamics

Seven 10 ns MDs simulations in explicit water have been conducted on different systems as detailed in Table 4. The systems contain disaccharides, PA-IL monomer or the complex of both. Simulations were carried out under periodic boundaries conditions with constant pressure and temperature as described in the Material and Methods section.

Conformational analysis and MD trajectories of disaccharides

The potential conformational space of each disaccharide can be represented as a function of Φ and Ψ potential energy contour maps with the MM3 software. Only the αGal12Gal map was calculated herein since the two others ones were obtained previously (11,45). The maps are typical of α-linkages, with two main energy minima separated by about 100° in Ψ and a remote one with higher energy. The molecular simulation of each disaccharide in water displays rapid interconversion between the two main minima (Figure 4). The population in each conformation has been calculated (Table S2 in supplemental material) and reported on the energy maps with the trajectory of the simulations (Figure 4).

The distribution of conformations for the hydroxymethyl group at C6 was also analyzed and the gauche-trans one (ω approx +60°) is the most populated one (75 to 80%) for all galactose residues (Table S2). This simulation is in agreement with previous studies on galacto monosaccharides in solid state (46) and in solution (47).

Stability of the monomeric PA-IL in MD trajectories and analysis of the binding site

The PA-IL monomer was subjected to several 10 ns simulations either alone in solution, or in complex with the different disaccharides (Table 4). In all cases, the calcium ion was present in the binding site. The stability of the global tertiary structure of the monomer was evaluated by monitoring the evolution of the backbone RMSD as a function of time. Except for some fluctuations, the plots show that the RMSD increases at the beginning of the simulations, reaching a relatively stable value that does not exceed 2 Å (Figure S1 in Supplemental Material). More precise analysis was performed on the stability of the carbohydrate recognition site, defined by the calcium ion and the amino acids His50, Gln53, Asp100, Thr104 and Asn107. In the presence of the disaccharides, the binding site RMSD in all the systems does not present any excursions, with an average of 0.8 (+/- 0.2) Å (Figure S2 in Supplemental Material). In the absence of bound carbohydrate, the RMSD of the binding site increases and displays a mean value of 1.2 (+/- 0.5) Å.

In order to check the stability of the protein in regions close to the binding site, the theoretical B-factors were calculated by multiplying the average atomic positional fluctuation of the backbone atoms by $8/3\pi 2$ (Figure 5). PA-IL, both in a bounded and non-bounded state, shows regions with higher mobility. In particular, fluctuations are observed in proximity of the C and N-terminal residues and in for amino acids 20 to 25, 70 to 75 and 85 to 95 which correspond to solvent exposed loops opposite to the carbohydrate binding site area. The simulation of the unliganded PA-IL monomer also shows slight mobility in loops corresponding to amino acids 50 to 55 and 100 to 107, which belong to the carbohydrate binding site.

Dynamic features of the carbohydrates in the PA-IL binding site

The molecular dynamics of the PA-IL/disaccharide complexes were analyzed in terms of flexibility of the glycosidic linkage for the bound carbohydrates (Figure 4). When the αGal1-2βGalOMe is bound to the lectin, its flexibility is reduced to one low energy conformation centered on Φ, Ψ values of 60°/ -150° that corresponds closely to the conformations observed in the crystal structure of PA-IL/αGal1–2GalOMe complex described above. The second energy minimum, corresponding to Φ value around -100°, is never occupied in the simulations in which disaccharides are present in PA-IL binding site.

On the contrary, the dihedral history of the Φ and Ψ angles of αGal1-3βGalOMe bound to PA-IL is quite similar to the one of the free disaccharide. Both main energetic minima (Φ/Ψ ≈ 80°/80° and Φ/Ψ ≈ 100°/150°) are visited during the simulation (Figure 4 and Table S2 and S3). The disaccharide in complex with PA-IL shows an equal population of the two minima while the minimum centered on Ψ=80° is more populated for the disaccharide in solution. For comparison, in the structure of PA-IL/trisaccharide complex (11), the αGal1-3Gal linkage adopts a Φ/Ψ conformation of 98°/121° that corresponds to the plateau between the two main minima with a relative energy of 2 kcal/mol.

The αGal1-4Gal energy map displays an extended energetic allowed region with almost no energy barrier between the main energy minima. In MDs simulations of PA-IL/αGal1–4βGalOMe complex, the Φ and Ψ region centered on 80°/110° was the most visited in the case of complex with PA-IL (approx. 80%) whereas the free disaccharide fluctuates rapidly in all the conformational space of the energy plateau and does not display significant conformational preferences.

The comparison between the conformational behaviors of the three disaccharides demonstrates a clear difference between the 1-2 linked disaccharides and the two others. The αGal1-2βGalOMe is the only one that displays a very limited conformational freedom when bound to PA-IL, whereas the flexibility of the two other ones is not significantly affected. These observations are in agreement with the calorimetry data that demonstrates a higher entropy penalty for binding of the αGal1-2βGalOMe disaccharide, however compensated by strong enthalpy of binding.

Calcium coordination and hydrogen bond network

The PA-IL binding site contains one calcium ion which coordinates 5 atoms from amino acids and two oxygen atoms from the bound galactose, namely O3 and O4. The ion – galactose oxygen coordination is well conserved and stable in all trajectories (Figure 3S).

Oxygens O3 and O4 are constantly involved in coordination of calcium with mean distances of 2.5 Å (+/- 0.1), similar to the ones observed in crystal structure (10).

The hydrogen bond network has been followed in the simulations of PA-IL/disaccharide systems by tracking the pair interactions in terms of distance and occupancy between the possible atoms acting as hydrogen bond donors or acceptors in each lectin/carbohydrate binding site (Table 5, Table S3). In the primary galactose binding site, the most conserved hydrogen bond network involves the Asn107 side chain that can receive hydrogen bonds from O2 and O3 and donate to O3 (occupancy > 90%). The side chain of Asp100 can receive, along all simulations, hydrogen bonds from O4 (≈ 100%). The amino acid His50 contributes to the stability of the complexes by giving hydrogen bonds to the oxygen atoms O6 (occupancy ≈ 90%) and O5 (occupancy ≈ 30%). Less permanent hydrogen bonds are also observed between the non-reducing galactose and other amino acids: the Thr104 main chain together with the Gln53 sidechain can be involved in polar contacts, accepting hydrogen bonds from HO3 and (≈ 60%) and HO6 (≈ 70%) respectively.

The reducing galactose establishes different polar contacts with the protein, depending on the stereochemistry of the glydosidic linkage involving the side chains of His50 and Gln53. The side chain of the amino acid Gln53 gives the most relevant contribution to the stability of all the reducing sugar moieties: it establishes a hydrogen bond network with the oxygens O3 and O4, the oxygen O2 and the oxygens O2 and O3 in the complexes PA1L/αGal1-2βGalOMe, PA1L/αGal1-3βGalOMe and PA1L/αGal1-4βGalOMe respectively. The side chain of the His50 can give hydrogen bonds to the oxygens O2, O3 in the complex PA1L/αGal1-3βGalOMe and to the oxygen O2 in the system PA1L/αGal1-4βGalOMe. The higher occupancy of stable hydrogen bonds observed for PA1L/αGal1-2βGalOMe can be correlated with the strong enthalpy of interaction measured by titration microcalorimetry.

Role of water molecules in the PA1L binding site

The crystallographic water molecule located in the cavity between the main chain of the lectin (Pro51 and Gln53) and the oxygen O6 of the α-galactose remains stable along all the simulations due to the occurrence of a dense hydrogen bond network. This water molecule can accept one hydrogen bond from nitrogen atom N of the Gln53 main chain and give one to Pro51 main chain oxygen (Table 5). A third hydrogen bond formed through polar contacts with oxygen O6. This scheme is found for all αGal1-2βGalOMe, αGal1-3βGalOMe and αGal1-4βGalOMe ligands (occupancy ≈ 90%). The water molecule is stable for the location of the oxygen atom, but it tumbles freely with the hydrogen atoms jumping between the several positions of the hydrogen bond network as displayed in the snapshots of Figure 6.

In order to evaluate the possible involvement of other water molecules in PA-IL/disaccharide interactions, the water density maps were calculated around each carbohydrate in complex with the lectin. The density maps were computed separately for the oxygen and the hydrogen atoms. Three significant bridging water molecules were identified in the complex PA1L/αGal1-2βGalOMe, two in the complex PA1L/αGal1-3βGalOMe and only one in the complex PA1L/αGal1-4βGalOMe (Figure 6). In all cases, the stronger density corresponds to the stable crystallographic water molecule that is present in the three simulations and that is described above. For simulations in presence of αGal1-2βGalOMe and αGal1-3βGalOMe, additional density peaks indicate the occurrence of a water molecule close to oxygen O2 and O3 of the buried galactose residue and to Asn107 side chain. In the PAIL/αGal1-2βGalOMe, a third region of hydration is identified between the O2 of the buried galactose and the O1 of the other one.

The 2D-RDF plots (Figure 7) were calculated according to the methodology described by S.B. Engelsen (48) in order to evaluate the variations in location and residence time of the water molecules identified by the density map calculations.

Only the crystallographic water molecule persists for the entire simulations in the binding pocket of all the systems with the maximum residence time (10 ns) while maintaining strong polar contacts to the sugar non-reducing end, and the protein backbone. In the PA1L/αGal1-2βGalOMe system, the bridging water molecule that mediates the interactions between the oxygen atom O2 and O3 of the α-galactose and Asn107 is present in 68 % of the MDs simulation with an average residence time of 36.7 ps. Concerning the PA1L/αGal1–

3βGalOMe complex, this water is present only in the first 1ns of the simulation (average residence time of 4.3 ps). The third water molecule identified in the PA-IL/ αGal1-2βGalOMe between the oxygen O2 of α-galactose and O1 of β-galactose is present during 8.8 ns (average residence time of 16.4ps). An equivalent water molecule bridging the two residues was identified in the simulation of this disaccharide free in solution (≈ 75% occupancy). The presence of this solvent residue could be considered an exclusive feature of this disaccharide.

Free energy analysis from MD trajectories

MM/PBSA analysis was performed using 1000 snapshots from each 10 ns MD simulation. An attempt to calculate the TΔS contribution using a harmonic approximation yielded essentially the same results for 100 snapshots of the three complexes with large standard deviations (≈ 19 +/-13kcal/mol), due mostly to the vibrational contribution of the systems. A quasi-harmonics approach approach could be used to calculate entropy, but the 10ns trajectories are normally not sufficient to reach a convergence of entropic values (49). The TΔS contribution was therefore not included in the analysis, an approximation classically used when comparing binding of different ligands and/or different mutants (50,51).

In a first round, the MM-PBSA energies were calculated without explicit water molecules (Table 6). The resulting energies were very similar for the different disaccharides (-10.6 to -10.1 kcal/mol) and did not correspond to the variations observed for the experimental values (Table 2). Previous studies have demonstrated that explicit consideration of structural water molecules could be important for ΔG calculations (51-53). The crystallography water molecule, present and stable during all simulations with carbohydrate ligands, has therefore been included in the calculations (Table 5). The resulting MM/PBSA energy is significantly lower for the αGal1-2βGalOMe ligand compared to the two other ones (2.3 and 3.3 kcal difference towards αGal1-3βGalOMe and αGal1-4βGalOMe). This trend compares very well with the experimental data. MM-PBSA approach also provides detailed information concerning the forces that are involved during the formation of carbohydrate-lectin complexes (Table 6) in which the electrostatic interactions are the main contributors to the energy of binding. The proposed energetic values can be considered qualitative results but in agreement with the experimental data.

When evaluating the contribution of each monosaccharides to the binding energy (Table S5), the non-reducing residue, i.e. buried aGal, is the major contributor (60 to 80% of the binding energy). Nevertheless, a clear effect of the water molecule is observed since the binding contribution of the buried galactose increases by about 10% when the water molecule is present (Table S5). This is in agreement with the direct hydrogen bond contacts established between this residue and the bridging water.

The applied method underscores and confirms the importance of including specific water molecules in computational studies for the prediction of free energies of binding.

CONCLUSIONS

The interaction between PA-IL and three digalactosides different only by the linkage position has been evaluated by experimental and theoretical tools. An excellent agreement is obtained that rationalizes the preference for the α2 linked disaccharide. This molecule undergoes a strong reduction of flexibility upon binding, which corresponds to high entropy penalty. However, it also establishes a high number of hydrogen bonds with higher occupancy between the external galactose and the protein surface as demonstrated by crystal structure and molecular dynamics. The resulting interaction enthalpy term overpasses the entropy cost and results in the stronger affinity constant. Molecular docking and molecular dynamics simulations data are in agreement with structural data derived from crystallography and they can be considered a valid method for the prediction of the behavior of different ligands in PA1L binding site. Although αGal1-2Gal disaccharide is not a human epitope, it is present on the surface of some parasites such as *Trypanosoma cruzi* (54) and its high affinity binding by PA-IL lectin may be of interest for specific labeling.

Molecular dynamics simulation was a useful tool for the analysis of the solvent in PA-IL binding site. Structural waters were identified using different *in silico* approaches such as density map calculations and 2D radial distribution function analysis. This study confirmed the stability of a structural bridge water molecule always present along the

simulations and also revealed the presence of other structural water molecules, with lower residence time and occupancy, in disaccharide /PA1L complexes. MM-PBSA was used for calculating the free energy of binding. This study could be considered as a qualitative approach in order to understand and also to predict the trend of the energetic features of ligands in complex with PA-IL. Our work shows that the inclusion of the stable bridge water molecules in MM-PBSA calculation was fundamental for understanding the trend of the energies of binding. This is in agreement with previous works which demonstrated that explicit consideration of structural water molecules could be important for ΔG calculations.(51-53) This work provides information that is key to strengthening the fundamental understanding of how specific carbohydrates bind to the PA 1L binding site.

The first assays for occupying a bridging water site by a synthetic side chain were recently performed on ConA (55) and PA-IIL (56) but with no gain yet in affinity. Considering the multivalency of lectins, the synthesis of multivalent ligands is a classical way for gaining avidity. Indeed, the first tetravalent galactose-containing ligand for PA-IL demonstrated a 850 fold in affinity compared to the monomeric one (57). Combining the detailed knowledge about structure and energetics of the binding site together with multivalency approach will be the route for the design of drugs active against *P. aeruginosa* infections.

REFERENCES

1. Goldstein, I. J., and Hayes, C. E. (1978) *Adv. Carbohydr. Chem. Biochem.* **35**, 127-340
2. Lis, H., and Sharon, N. (1998) *Chem. Rev.* **98**, 637-674
3. Doggett, R. G., and Harrison, G. M. (1972) *Infect. Immun.* **6**, 628-635
4. Gilboa-Garber, N. (1982) *Methods Enzymol.* **83**, 378-385
5. Imberty, A., Wimmerova, M., Mitchell, E. P., and Gilboa-Garber, N. (2004) *Microb. Infect.* **6**, 222-229
6. Tielker, D., Hacker, S., Loris, R., Strathmann, M., Wingender, J., Wilhelm, S., Rosenau, F., and Jaeger, K.-E. (2005) *Microbiology* **151**, 1313-1323
7. Mishra, N. K., Kulhánek, P., Šnajdrová, L., Petřek, M., Imberty, A., and Koča, J. (2008) *Proteins* **72**, 382-392
8. Mitchell, E., Houles, C., Sudakevitz, D., Wimmerova, M., Gautier, C., Pérez, S., Wu, A. M., Gilboa-Garber, N., and Imberty, A. (2002) *Nature Struct. Biol.* **9**, 918-921
9. Garber, N., Guempel, U., Belz, A., Gilboa-Garber, N., and Doyle, R. J. (1992) *Biochim. Biophys. Acta* **1116**, 331-333
10. Cioci, G., Mitchell, E. P., Gautier, C., Wimmerova, M., Sudakevitz, D., Pérez, S., Gilboa-Garber, N., and Imberty, A. (2003) *FEBS Lett.* **555**, 297-301
11. Blanchard, B., Nurisso, A., Hollville, E., Tétaud, C., Wiels, J., Pokorná, M., Wimmerová, M., Varrot, A., and Imberty, A. (2008) *J. Mol. Biol.* **383**, 837-853
12. Gilboa-Garber, N., Sudakevitz, D., Sheffi, M., Sela, R., and Levene, C. (1994) *Glycoconj. J.* **11**, 414-417
13. Lanne, B., Ciopraga, J., Bergstrom, J., Motas, C., and Karlsson, K. A. (1994) *Glycoconj. J.* **11**, 292-298
14. Garegg, P. J., and Hultberg, H. (1982) *Carbohydr. Res.* **110**, 261-266
15. Garegg, P. J., and Oscarson, S. (1985) *Carbohydr. Res.* **137**, 270-275
16. Leslie, A. G. W. (1992) *Jnt CCP4/ESF-EACMB. Newsletter on Protein Crystallography,* **26**
17. Collaborative Computational Project Number 4. (1994) *Acta Crystallogr. D. Biol. Crystallogr.* **D50**, 760-763
18. McCoy, A. J. (2007) *Acta Crystallogr. D. Biol. Crystallogr.* **63**, 32-41
19. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res.* **28**, 235-242
20. Brünger, A. T. (1992) *Nature* **355**, 472-475
21. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) *Acta Crystallogr. D. Biol. Crystallogr.* **53**, 240-255
22. Emsley, P., and Cowtan, K. (2004) *Acta Crystallogr. D. Biol. Crystallogr.* **60**, 2126-2132
23. Laskowski, R. A., MacArthur, M. W., and Thornton, J. M. (1998) *Curr. Opin. Struct. Biol.* **8**, 631-639
24. Clark, M., Cramer, R. D. I., and van den Opdenbosch, N. (1989) *J. Comput. Chem.* **10**, 982-1012
25. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. J., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) *J. Am. Chem. Soc.* **117**, 5179-5197
26. Imberty, A., Bettler, E., Karababa, M., Mazeau, K., Petrova, P., and Pérez, S. (1999) Building sugars : The sweet part of structural biology. In: Vijayan, M., Yathindra, N., and Kolaskar, A. S. (eds). *Perspectives in Structural Biology*, Indian Academy of Sciences and Universities Press, Hyderabad
27. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998) *J. Comp. Chem.* **19**, 1639-1662
28. Nurisso, A., Kozmon, S., and Imberty, A. (2008) *Mol. Simul.* **34**, 469-479
29. Allinger, N. L., Yuh, Y. H., and Lii, J.-H. (1989) *J. Amer. Chem. Soc.* **111**, 8551-8566
30. Pérez, S., Imberty, A., Engelsen, S. B., Gruza, J., Mazeau, K., Jiménez-Barbero, J., Poveda, A., Espinosa, J. F., van Eyck, B. P., Jonhson, G., French, A. D., Kouwijzer, M. L. C. E., Grootenhuis,

D. J., Bernardi, A., Raimondi, L., Senderowitz, H., Durier, V., Vergoten, G., and K., R. (1998) *Carbohydr. Res.* **314**, 141-155
31. Preusser, A. (1989) *ACM Trans. Math. Software* **15**, 79-89
32. Cheatham, T. E., 3rd, Cieplak, P., and Kollman, P. A. (1999) *J. Biomol. Struct. Dyn.* **16**, 845-862
33. Kirschner, K. N., Yongye, A. B., Tschampel, S. M., Gonzalez-Outeirino, J., Daniels, C. R., Foley, B. L., and Woods, R. J. (2008) *J. Comput. Chem.* **29**, 622-655
34. Bradbrook, G. M., Gleichmann, T., Harrop, S. J., Habash, J., Raftery, J., Kalb, J., Yariv, J., Hillier, I. H., and Helliwell, J. R. (1998) *J. Chem. Soc. Faraday T* **94**, 1603–1611
35. Darden, T., York, D., and Pedersen, L. (1993) *J. Chem. Phys.* **98**, 10089-10092
36. Ryckaert, J. P., Cicotti, G., and Berendsen, H. J. C. (1977) *J. Comp. Chem.* **23**
37. Humphrey, W., Dalke, A., and Schulten, K. (1996) *J. Molec. Graphics* **14**, 33-38
38. Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E., 3rd. (2000) *Acc. Chem. Res.* **33**, 889-897
39. Sitkoff, D., Sharp, K. A., and Honig, B. (1994) *J. Phys. Chem.* **98**, 1978–1988
40. Gilson, M. K., and Zhou, H. X. (2007) *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21-42
41. Dam, T. K., and Brewer, C. F. (2002) *Chem. Rev.* **102**, 387-429.
42. Turnbull, W. B., and Daranas, A. H. (2003) *J. Am. Chem. Soc.* **125**, 14859-14866
43. Rockey, W. M., Laederach, A., and Reilly, P. J. (2000) *Proteins* **40**, 299-309
44. Wen, X., Yuan, Y., Kuntz, D. A., Rose, D. R., and Pinto, B. M. (2005) *Biochemistry (Mosc).* **44**, 6729-6737
45. Corzana, F., Bettler, E., Hervé du Penhoat, C., Tyrtysh, T. V., Bovin, N. V., and Imberty, A. (2002) *Glycobiology* **12**, 241-250
46. Marchessault, R. H., and Pérez, S. (1979) *Biopolymers* **18**, 2369-2374
47. Nishida, Y., Ohrui, H., and Meguro, H. (1984) *Tetrahedron Lett.* **25**, 1575-1578
48. Andersson, C., and Engelsen, S. B. (1999) *J. Mol. Graph. Model.* **17**, 101-105
49. Fogolari, F., Brigo, A., and Molinari, H. (2003) *Biophys. J.* **85**, 159-166
50. Gohlke, H., Kiel, C., and Case, D. A. (2003) *J. Mol. Biol.* **330**, 891-913
51. Wong, S., Amaro, R. E., and McCammon, J. A. (2009) *J. Chem. Theory Comput.* **5**, 422-429
52. Lepsik, M., Kriz, Z., and Havlas, Z. (2004) *Proteins* **57**, 279-293
53. Treesuwan, W., and Hannongbua, S. (2009) *J. Mol. Graph. Model.* **27**, 921-929
54. Avila, J. L., M., R., and Velazquez-Avila, G. (1992) *Am. J. Trop. Med. Hyg.* **4**, 413-421
55. Kadirvelraj, R., Foley, B. L., Dyekjaer, J. D., and Woods, R. J. (2008) *J. Am. Chem. Soc.* **130**, 16933-16942
56. Andreini, M., Anderluh, M., Audfray, A., Bernardi, A., and Imberty, A. (in press) *Carbohydr. Res.*
57. Cecioni, S., Lalor, R., Blanchard, B., Praly, J. P., Imberty, A., Matthews, S. E., and Vidal, S. (2009) *Chem. Eur. J.* **15**, 13232-13240

FOOTNOTES

[1] The abbreviations used are: PA-IL : Lectin I from *Pseudomonas aeruginosa;* MM-PBSA : Molecular Mechanics Poisson- Bolzmann Surface Area; ; αGal1-2βGalOMe : methyl 2-*O*-α-D-galactopyranosyl-β-D-galactopyranoside; αGal1-3βGalOMe : methyl 3-*O*-α-D-galactopyranosyl-β-D-galactopyranoside; αGal1-4βGalOMe : methyl 4-*O*-α-D-galactopyranosyl-β-D-galactopyranoside

[2]Coordinates and structure factors have been deposited in the Protein Data Bank with accession code 2wyf.

[3] The on-line version of this article (available at http://www.jbc.org) contains supplemental data

ACKNOWLEDGMENTS

The work was supported by CNRS. We acknowledge EEC European Community programs MEST-CT-2004-503322 (CermavTrain) and MRTN-CT-2006-035546 (NODPERCEPTION) for A.N. salary, French Ministry of Research for B.B. salary and Science Foundation Ireland for L.R. salary We thank the ESRF, Grenoble, for access to synchrotron data collection facilities. Financial support is gratefully acknowledged from the Association Vaincre la Mucoviscidose and the GDR Pseudomonas

FIGURE LEGEND

Figure 1. **Crystal structure of PA-IL/αGal1-2βGalOMe complex**. A. Representation of one tetramer with the two β-sheets coloured in blue and orange. The disaccharide is represented as sticks and the calcium ion as pink sphere. B. Representation of one monomer (chain A) with the final weighted 2mFo –DFc electron density map (contoured at 1 σ, 0.34 e Å-3) around the disaccharide. C. View of the binding site with hydrogen bonds represented as green dashed lines and co-ordination contacts as solid orange lines

Figure 2. **Titration microcalorimetry of PA-IL by disaccharide derivatives**. A Titration curve at 25°C of PA-IL (0.49 mM) by αGal1-2βGalOMe (1.7 mM) from 29 automatic injections of 10 µL sugar added every 300 sec to PA-IL containing cell (B) Total heat released as a function of total ligand concentration. The solid line represents the best least-squares fit. (C) Free energy, enthalpy contribution and entropy contribution for the binding PA-IL with the three disaccharides.

Figure 3. **Graphical representation of the lowest energy docking poses of disaccharides on PA-IL**. The docked disaccharides are represented by white sticks :(A) αGal1-2βGalOMe, (B) αGal1-3βGalOMe and (C) αGal1-4βGalOMe. Available crystallographic structures are represented as dark lines for the complexes of PA-IL with (A) αGal1-2βGalOMe (this work) (B) αGal1-3βGal14Glc (11) and (C) galactose (10).

Figure 4 **- Plots of Φ and Ψ dihedral angles of the disaccharides as a function of time for the three disaccharides in solution or complexed with a monomer of PA-IL.** The MM3 adiabatic maps of each disaccharides are represented with and without superimposition of the 10 ns dynamics simulations trajectories.

Figure 5. **Theoretical B-factor of PA-IL residues (backbone atoms only) calculated for all the simulations: red line represents the B-factor calculated for the residues of the monomer PA-IL**. The B-factors calculated for the residues of PA-IL in complex with disaccharides are color-coded as follow: red for the free monomer, green for αGal1- 2βGalOMe/PA-1L, blue for αGal1-3βGalOMe/PA-1L (B) and yellow for PA1L/αGal1- 4βGalOMe/PA-1L (C).

Figure 6. **Representation of the density/solvent distribution of the first water shell around the ligands.** The density was calculated separately for the oxygen atoms (violet surfaces) and hydrogen atoms (gray surfaces) of water molecules**.** The water density map reveals the presence of water density peaks (red circles) around the non-reducing end of the carbohydrates, represented in red circles, (A) PA1L/αGal1-2βGalOMe, (B) PA1L/αGal1-3βGalOMe and (C) PA1L/αGal1-4βGalOMe

Figure 7. **2D-RDF analysis of the presence of water molecules at given place during the 10 ns simulation of PA1L/αGal1–2βGalOMe system** (for the other disaccharides, see Figure 4S in supplemental material). (A). Analysis of the crystallographic bridge water molecule. (B) Analysis of two additional water molecules present between the oxygen O3 of the non-reducing galactose and the nitrogen N of the residue Asn107 and between the oxygen O2 of the non-reducing end and the glycosidic oxygen of the disaccharide (GalO).

Table 1. Data collection and refinement statistics for the PA-IL/ αGal1-2βGal-O-Met complex

	PA-IL/Galα1-2Galβ-O-Met
Structural data	
Beam line	ID14-eh1
Wavelength (Å)	0.934
Resolution (Å)	37.11-2.40 (2.53-2.40)*
Cell dimension	
space group	$P2_1$
Unit cell (Å)	a=9 b=99.8 c=91.3, β =100.8°
Measured reflexion/ unique	34140 / 4992
Average multiplicity	3.1 (3.1)
Completeness (%)	99.2 (99.8)
Average I/σ(I)	8.5 (3.1)
R_{merge} (%)	11.4 (35.2)
Wilson *B*-factor	20.8
Refinement	
Resolution range (Å)	37.11-2.40
R_{work} (%)	0.17
R_{free} (%)	0.26
Average Biso ($Å^2$)	
All atoms	13.3
Protein atoms	12.6
Sugar atoms	30.43
Solvent atoms	16.57
RMSD from ideality	
angles (°)	1.6
Bonds (Å)	0.015
Water molecules	648
Number of disaccharide	6
Number of galactose	2
Calcium atoms	8
Protein Data Bank deposition code	2WYF

* Values in parenthesis refer to the highest resolution shell

Table 2. Titration microcalorimetry data for the interaction between PA-IL and disaccharides. *

Ligand	Ka ($10^3 M^{-1}$)	Kd (μM)	ΔG (kcal/mol)	ΔH (kcal/mol)	- TΔS (kcal/mol)
αGal1-2βGalOMe	27 (+/- 2)	37	-6.03	-11.5 (+/- 0.2)	5.4
αGal1-3βGalOMe	7.6 (+/- 0.2)	132	-5.29	-8.6 (+/- 0.0)	3.3
αGal1-4βGalOMe	8.8 (+/- 0.9)	115	-5.37	-7.6 (+/- 0.5)	2.2

*Experiments were performed twice and standard deviations are expressed between parenthesis for Ka and ΔH. The stochiometry was fixed to 1 during the fitting procedure

Table 3. Description of lowest energy binding modes between PA-IL and the disaccharides as predicted by Autodock3. Data from crystallography are reported between parenthesis for comparison..

	Clusters (%)	Lowest energy (kcal/mol)	RMSD (Å) *	Φ angle (°)	Ψ angle (°)	Ca-O3 distance (Å)	Ca-O4 distance (Å)
αGal1-2βGalOMe	72	-5.3	1.3	62 (64)	-153 (-152)	2.8	2.5
αGal1-3βGalOMe	97	-5.0	0.8	96 (98)	135 (127)	2.5	2.5
αGal1-4βGalOMe	83	-5.6		94	129	2.4	2.5

*RMSD was calculated between the best docking poses and available crystal structures from this work and (11).

Table 4. Details of the different MDs simulations

	A	B	C	D	E	F	G
Protein				PA-IL	PA-IL	PA-IL	PA-IL
Ligand *	G12G	G13G	G14G		Gl2G	G13G	G14G
N. atoms system	2490	2514	2460	22452	22656	22647	22587
N. atoms protein				1763	1763	1763	1763
N. atoms ligand	48	48	48		48	48	48
N. waters	814	822	804	6896	6948	6945	6925
N. of Ca++				1	1	1	1
N. of Na+				1	1	1	1

* G12G: αGal1-2βGalOMe, G13G: αGal1-3βGalOMe, G14G: αGal1-4βGalOMe

Table 5. Hydrogen bonds data with an occupancy > 15% and angle cutoff of 100 degrees collected along the MDs simulations.*

Acceptor	Donor H-X	PA-1L/ αGal1-2βGalOMe		PA-1L/ αGal1-3βGalOMe		PA-1L/ αGal1-4βGalOMe	
		Occupancy %	Distance Å	Occupancy %	Distance Å	Occupancy %	Distance Å
αGal.O2	Asn107.HD21	99.5	3.0 (0.2)	99.5	3.0(0.2)	99.6	3.0(0.1)
αGal.O3	Asn107.HD21-	94.3	3.2 (0.2)	91.2	3.2(0.2)	93.0	3.2(0.2)
Asn107.OD1	*αGal.H3O*	100.0	2.9 (0.1)	100.0	2.9(0.1)	100.0	2.9(0.1)
Thr104.O	*αGal.H3O*	55.9	3.3 (0.1)	60.0	3.3(0.1)	60.9	3.3(0.1)
Asp100.OD1	αGal.H4O	100.0	2.6 (0.1)	100.0	2.6(0.1)	100.0	2.6(0.1)
Asp100.OD2	αGal.H4O	99.9	3.0 (0.1)	99.9	3.0(0.1)	99.9	3.0(0.1)
αGal.O5	His50.HE2	16.2	3.3 (0.1)	31.3	3.2(0.2)	43.7	3.2(0.2)
αGal.O6	His50.HE2	99.1	3.0 (0.1)	92.7	3.0(0.2)	96.5	3.0(0.2)
Gln53.OE1	αGal.HO6	71.8	2.8 (0.2)	50.1	2.9(0.3)	37.1	2.9(0.3)
αGal.O6	Gln53.HE22	15.5	3.0 (0.2)	16.3	3.0(0.2)	24.9	3.1(0.2)
Gln53.OE1	βGal.H2O			58.0	3.0(0.3)	60.0	3.0(0.3)
βGal.O2	His50.HE2			46.1	3.1(0.2)		
βGal.O2	Gln53.HE21			54.0	3.1(0.2)	59.6	3.0(0.3)
Gln53.OE1	βGal.H3O	75.0	3.0 (0.3)			65.9	2.9 (0.3)
βGal.O3	Gln53.HE21	36.6	3.1(0.2)			38.2	3.1(0.2)
βGal.O3	His50.HE2					49.7	3.2(0.2)
βGal.O4	Gln53.HE21	61.1	3.1(0.2)				
Gln53.OE1	βGal.HO4	53.7	3.1(0.3)				
αGal.O6	WAT.H2	99.2	2.8(0.2)	96.7	2.9(0.2)	97.8	2.9(0.2)
Pro51.O	WAT.H1	99.6	2.8(0.2)	98.0	2.8(0.2)	99.1	2.8(0.2)
WAT.O	Gln53.HN	93.4	3.1(0.2)	84.2	3.1(0.2)	90.0	3.0(0.2)

*The hydrogen bonds reported in italic are present if a cutoff angle of 60° is considered. The occupancy is defined as the percent over the whole trajectory in which both the distance and the angle criteria are satisfied

Table 6 - Thermodynamics of binding for PA-IL calculated using the MM-PBSA approach.*

	ΔG_{vdw}	$\Delta G_{ele\text{-}int}$	ΔG_{nonpol}	$\Delta G_{ele\text{-}sol}$	$\Delta G_{mm\text{-}pbsa}$	ΔG_{exp}
PA1L/G12G	-11.7 (4.1)	-89.9 (8.5)	-3.7 (0.2)	94.7 (6.7)	-10.5 (4.0)	-6.03 (0.01)
PA1L/ G12G +WAT	-12.5 (4.3)	-100.9 (9.6)	-4.0 (0.2)	99.8 (7.1)	-17.7 (4.7)	
PA1L/ G13G	-10.6 (4.0)	-85.1 (6.2)	-3.5 (0.2)	88.5 (6.0)	-10.6 (4.5)	-5.29 (0.02)
PA1L/ G13G +WAT	-11.8 (4.1)	-94.6 (8.9)	-3.8 (0.2)	95.7 (6.6)	-15.3 (4.6)	
PA1L/ G14G	-10.3 (3.9)	-86.6 (7.2)	-3.3 (0.2)	90.0 (7.2)	-10.1 (3.8)	-5.37 (0.06)
PA1L/ G14G +WAT	-11.1 (4.2)	-94.4 (10.6)	-3.6 (0.2)	95.6 (8.0)	-14.3 (4.8)	

*All values are in Kcal/mol. Data in parentheses are standard deviations. G12G: αGal1-2βGalOMe, G13G: αGal1-3βGalOMe, G14G: αGal1-4βGalOMe

Figure 1

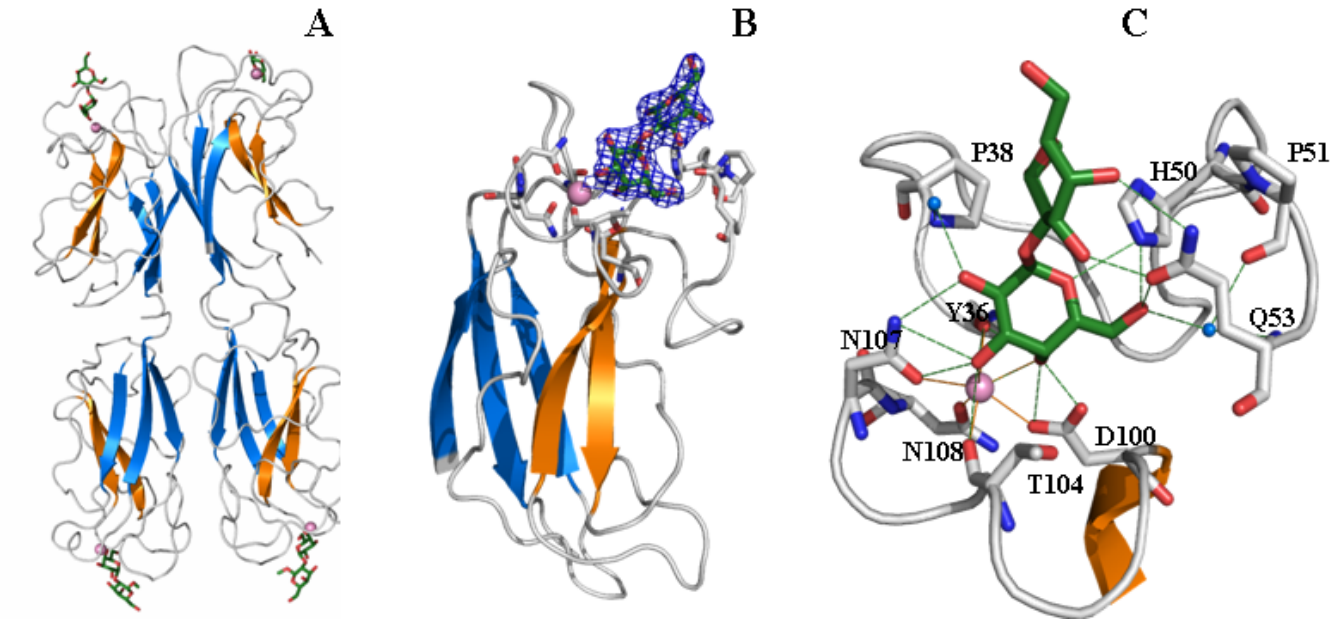


Figure 2

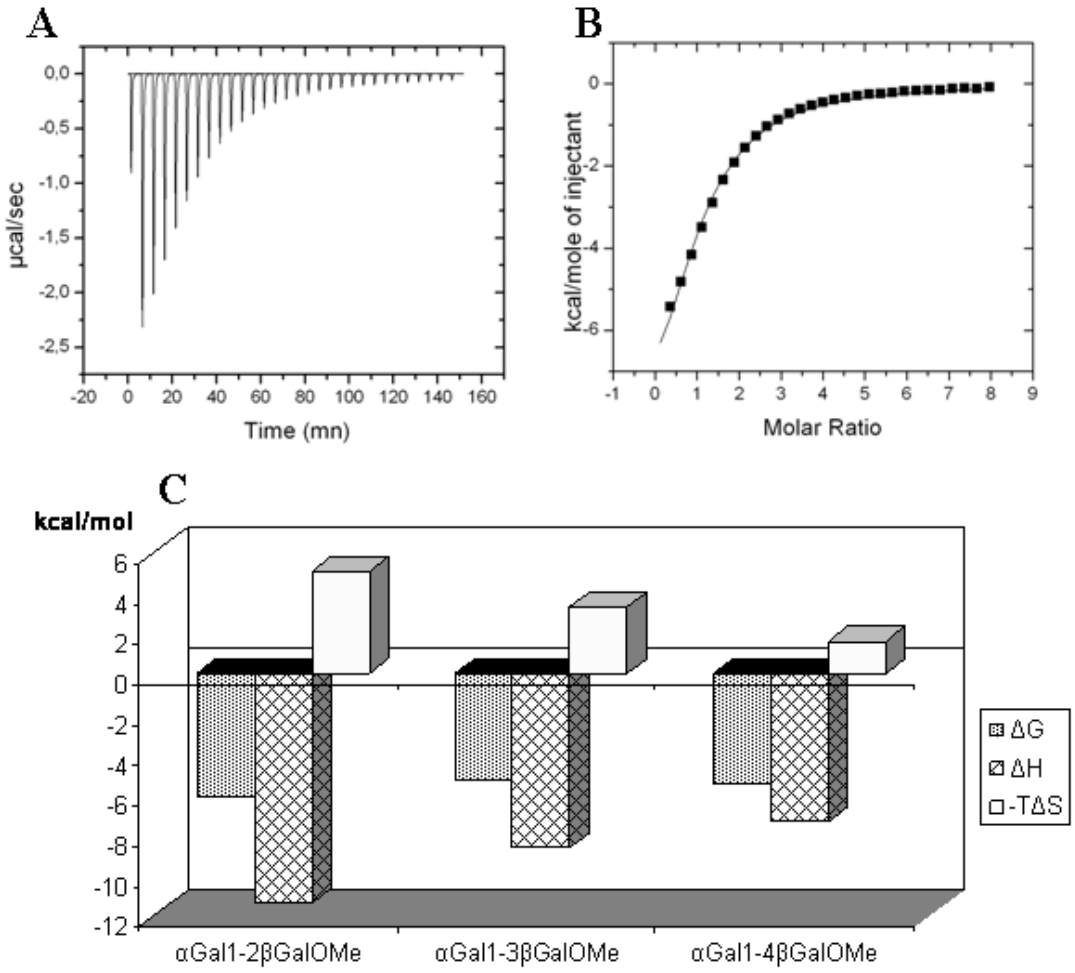

A
µcal/sec
0,0
-0,5
-1,0
-1,5
-2,0
-2,5
-20 0 20 40 60 80 100 120 140 160
Time (mn)
B
kcal/mole of injectant
0
-2
-4
-6
-1 0 1 2 3 4 5 6 7 8 9
Molar Ratio
C
kcal/mol
6
4
2
0
-2
-4
-6
-8
-10
-12
αGal1-2βGalOMe
αGal1-3βGalOMe
αGal1-4βGalOMe
ΔG
ΔH
-TΔS

Figure 3

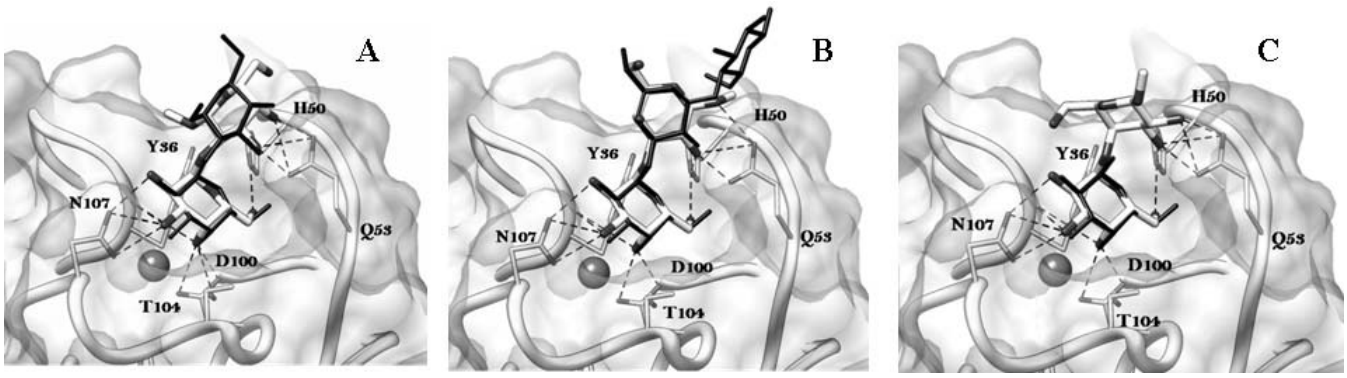

A
H50
Y36
N107
Q53
D100
T104
B
H50
Y36
N107
Q53
D100
T104
C
H50
Y36
N107
Q53
D100
T104

Figure 4

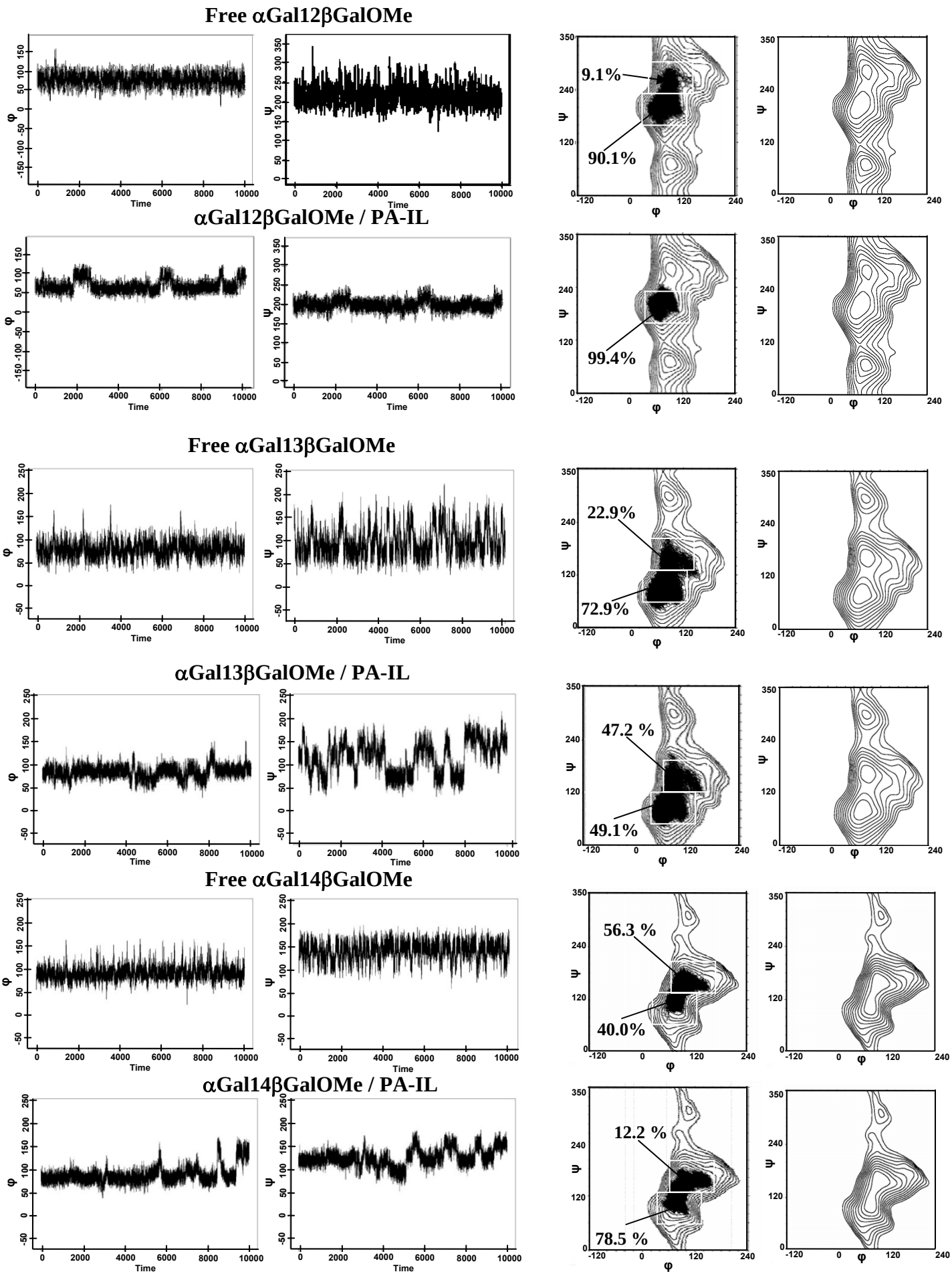

Free αGal12βGalOMe
9.1%
90.1%
αGal12βGalOMe / PA-IL
99.4%
Free αGal13βGalOMe
22.9%
72.9%
αGal13βGalOMe / PA-IL
47.2 %
49.1%
Free αGal14βGalOMe
56.3 %
40.0%
αGal14βGalOMe / PA-IL
12.2 %
78.5 %
Time
φ
ψ

Figure 5

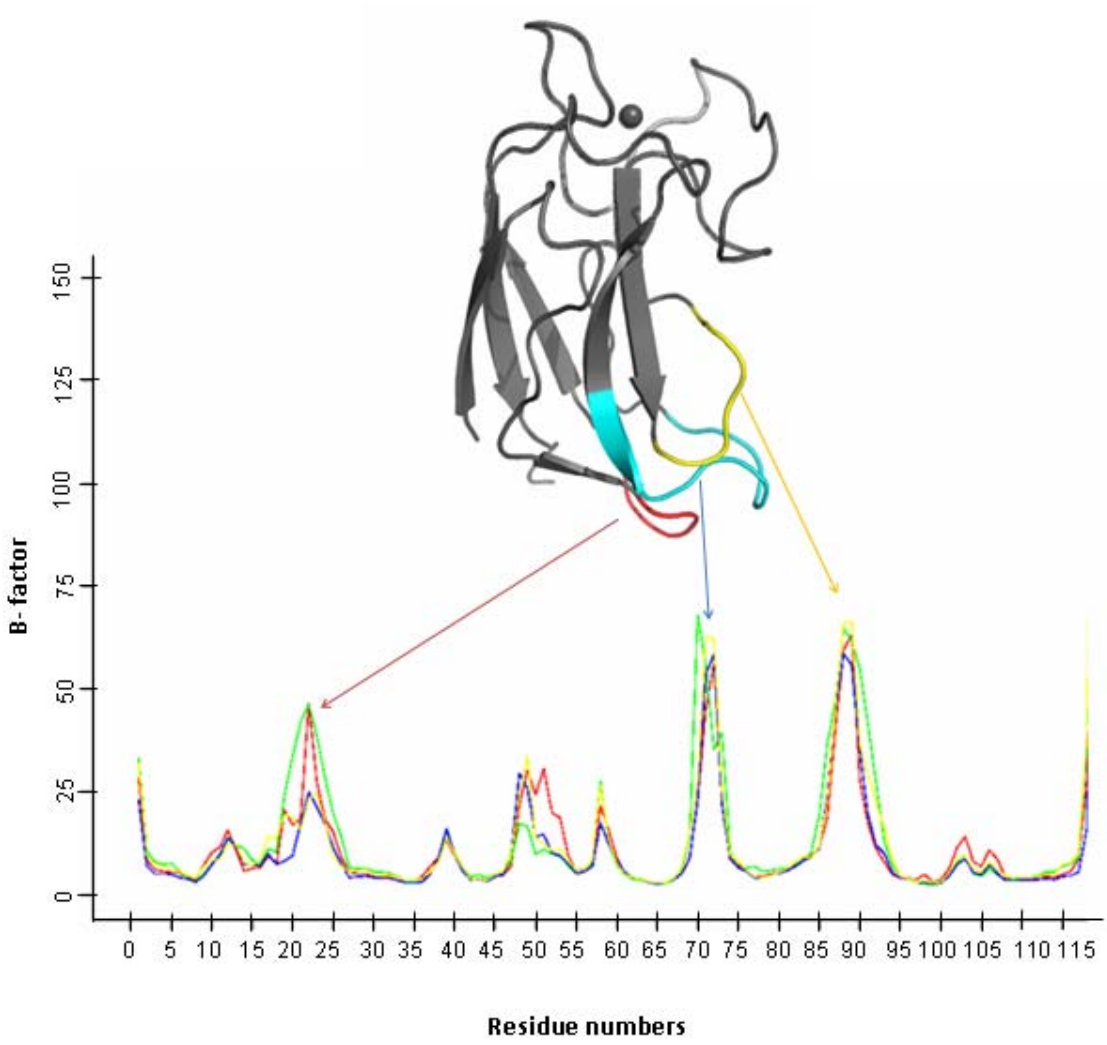

B-factor
0
25
50
75
100
125
150
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115
Residue numbers

Figure 6

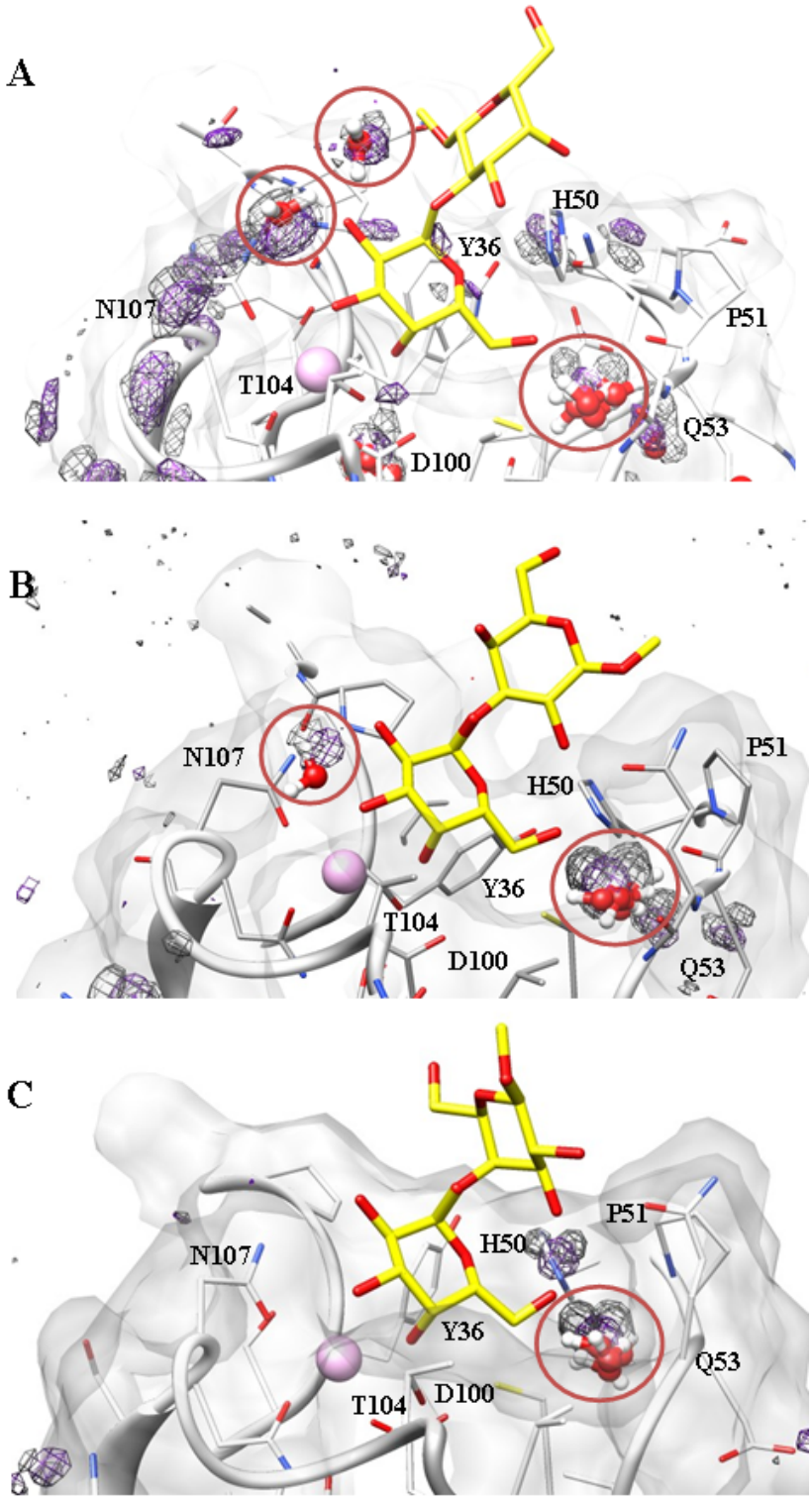
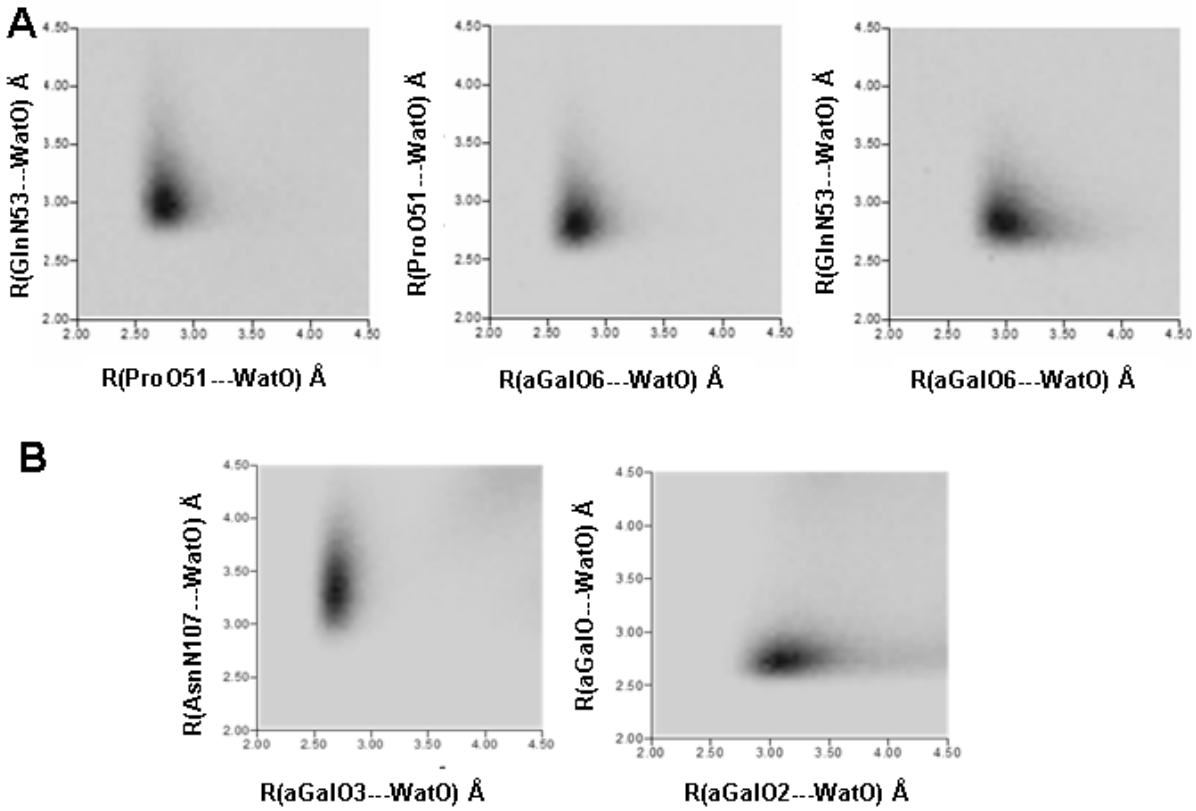

A
H50
Y36
N107
P51
T104
Q53
D100
B
N107
P51
H50
Y36
T104
D100
Q53
C
P51
N107
H50
Y36
Q53
T104
D100

Figure 7


A
R(GlnN53---WatO) Å
R(ProO51---WatO) Å
R(ProO51---WatO) Å
R(aGalO6---WatO) Å
R(GlnN53---WatO) Å
R(aGalO6---WatO) Å
B
R(AsnN107---WatO) Å
R(aGalO3---WatO) Å
R(aGalO---WatO) Å
R(aGalO2---WatO) Å

Plant lectin domains involved in bacterial symbiosis and their ligands

3.5 ARTICLES V-VI: *in silico* studies of Nod Factors and its receptors

3.5.1 Introduction

Plants require macronutrients like nitrogen and phosphorus for healthy growth. However, these elements are often present in nature in limited quantities.

To facilitate the uptake of nutrients, plants establish mutualistic symbiosis with other organisms. Most plants are able to interact with arbuscular mycorrhiza fungi: this symbiotic relationship implies the formation of branched feeding structures, the arbuscules, that greatly improve the bioavailability of phosphate and other elements[232].

In contrast, the symbiosis with nitrogen-fixing bacteria is limited to only a few plant families.

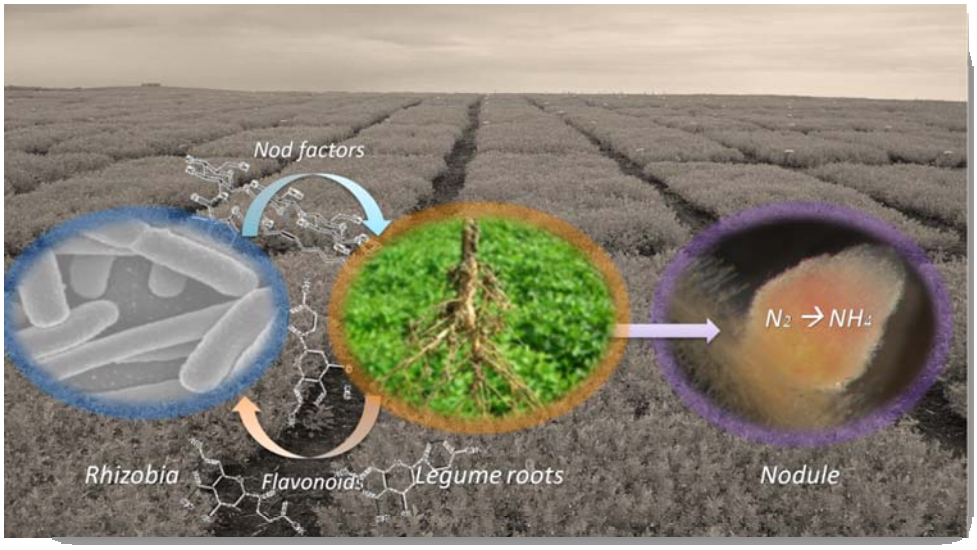


Figure 3-9 - The molecular dialogue between Rhizobia and legume plant allows the nitrogen fixation. The main actors of the symbiosis are herein represented.

The best studied nitrogen-fixing symbiosis is characterized between legumes (*Fabaceae*) and gram-negative soil bacteria called *Rhizobia.* This process is widely studied because nitrogen fixation can act as a renewable and environmentally sustainable source of nitrogen that can ideally replace the use of fertilizer nitrogen, source of environmental pollution[233].

This interaction leads to the formation of a completely new organ, the root nodule, in which bacteria can fix atmospheric nitrogen, then used by the plant[234, 235]. Bacterial chemotaxis towards plant roots is a crucial event in legume-*Rhizobia* interactions. Legume plants exude nutrients as amino acids, sugars, dicarboxylic acids which are chemoattractants for the bacteria that successively interact with root tips, probably by means of plant lectins[236]. During this process, flavonoids, metabolic aromatic products of legume plants, are released near the emerging root hair zone in micromolar concentration. They activate bacterial genes responsible for the production of specific signal molecules called Nod factors.

Nod factors have been described as the key to legume door[237]: their detection by legume hosts induces the expression of genes that lead to morphological changes of root hairs, accompanied by electrophysiological changes and nodule organogenesis[235]. Once Nod factors are released, the tip of a root hair where *Rhizobia* are bound curls back on itself, trapping the bacteria in a pocket. Successively, host cell wall gets gradually disrupted and rhizobial cells come into direct contact with the plant cell plasma membrane.

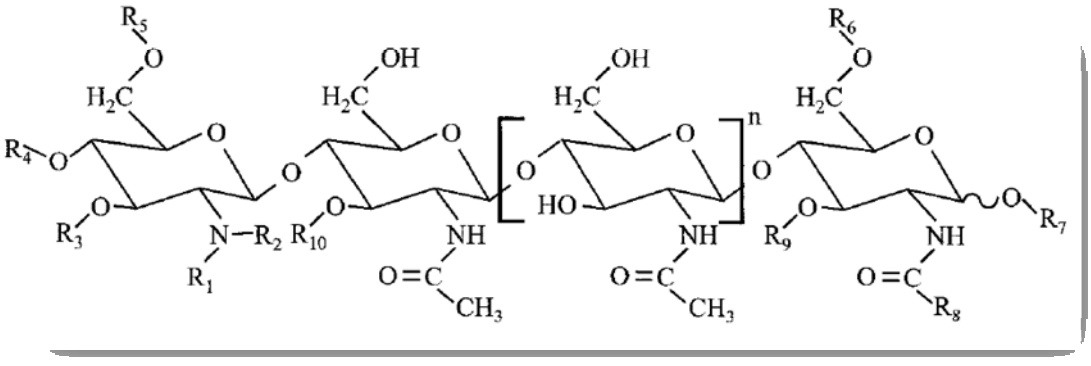


Figure 3-10 - General structure of Nod factors produced by Rhizobia. The identity of the substitutions (R1-R10) and of the degree of oligomerization (n) varies as a function of their origin.

In the same time, pericycle and cortical root plant cells are activated for division to form a nodule primordium that will be then reached by *Rhizobia* in an endocytotic way (root hair infection). In the newly formed structure, bacteria can differentiate in bacterioids that show high nitrogenase activity, catalyzing the reduction of atmospheric nitrogen N_2 to ammonia (Figure 3-9)[238]. The key event in nodule development and bacterial invasion is the synthesis and release of Nod factors that are active at very low concentrations (10^{-12}M to10^{-19}M)[239]. The structures of Nod factors have been widely studied and characterized[240]. However, it has not yet been determined how these molecules are perceived. The fact that specific structures lead to responses on host legumes at low concentrations suggests that they are perceived by specific plant receptors[241]. Nod factors are formed by an oligomeric backbone composed by three, four or five ß-1,4-linked N-acetyl-D-glucosamine residues. The non-reducing sugar moiety is substituted with a fatty acid whose structure varies between different rhizobial species. Further substitutions can be found both on the reducing and on the non-reducing end of the saccharidic backbone, determining the host specificity of a Nod factor[240]. For example, the Nod factor produced by the soil bacterium *Sinorhizobium meliloti*, symbiont of the legume plant *Medicago truncatula*, is characterized by O-sulphate (R6), O-acetyl (R5) and methyl (R8) groups. It is a tetraoligosaccharide (n=1) with a C16:2, Δ2, 9 fatty acid chain (R1) (Figure 3-10). Studies with rhizobial mutants have revealed that the sulphate group at the reducing end is essential for the nodulation[242] whereas the acetyl group and the fatty acyl chain C16:2 are essential for the initial symbiotic responses[243]. This evidence leads to the hypothesis that two receptors may be required for the Nod factor perception: a signaling receptor, necessary for the initial response, and an entry receptor, necessary for the infection[243]. Legume genes essential for symbiotic nodulation have been isolated[244-247]: they encode for proteins, receptor-like kinases characterized by two or three small extracellular motifs, (LysM motifs) separated by short amino acid linkers containing *CXC* cysteine motifs. LysM domains are short domains (~40 amino acids) that have been previously identified in other organisms, in particular in bacteria, and often displayed the capability to bind N-Acetilglucosamine GlcNAc residues[248]. They are likely candidate to bind Nod factors, while the intracellular kinase domain can act in signal transduction. In the legume plant *Medicago truncatula,* a two step model of Nod factor perception has been proposed (Figure 3-11).

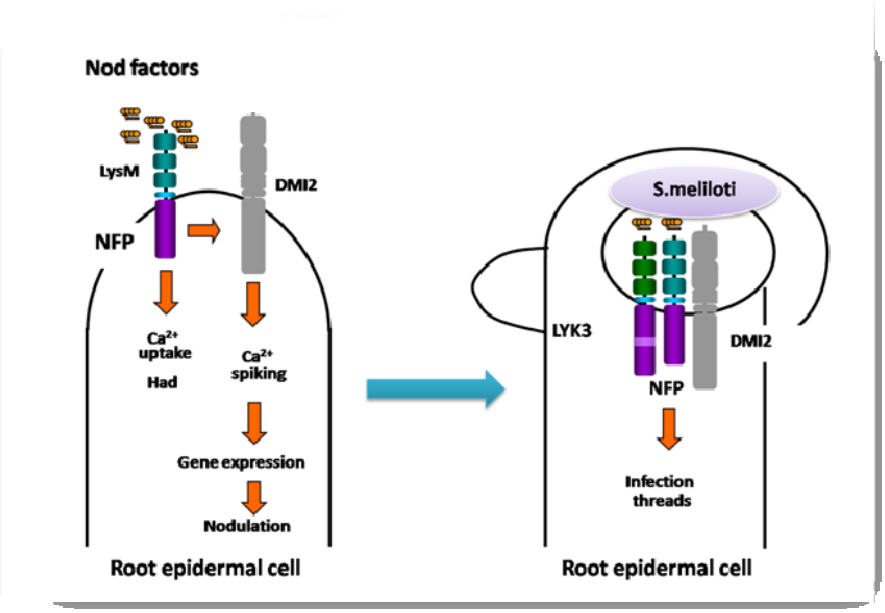


Figure 3-11 - The two step model for Nod factor perception (S. Meliloti Nod factors) in Medicago truncatula. DM2 is not reported because included in the nuclear membrane. Adapted from[241].

The first step involves the gene *NFP* encoding for three LysM and a receptor like kinase which is not enzymatically active. This protein binds to Nod factors from *Sinorhizobium meliloti* and leads to calcium uptake and root hair deformation. Due to the aberrant kinase domain of *NFP*, another kinase protein, encoded by the gene *LYR*, can also contribute to transmit signals to the plant. *DMI2/DMI1* genes encode for a leucine rich repeat receptor like kinase and for a channel like protein respectively. These proteins are required for the induction of calcium spiking and for the expression of genes that allow the formation of nodules.

The second step, the infection, influenced by Nod factors, is controlled by the genes *LYK3* that encodes a LysM-RLK protein with autophosphorylation activity and *DMI2*[241]. Among all the available amino acidic sequences, only three tridimensional LysM domains struc-

tures (two isolated from bacteria, *1e0g* and *1y7m*, one isolated from humans with the PDB code *2djp)* have been reported in the literature[249, 250]. The three structures present the same *βaaβ* secondary structure with two α helices packing onto the same side of an antiparallel ß sheet but a low sequence identity (~ 20%). This information has been used for modeling structures of LysM domains from legumes (Figure 3-12). The structures of the three LysM domains encoded by the *NFP* gene from the plant *Medicago truncatula* have been characterized by homology modeling. Molecular docking calculations predict the most favored binding modes of Nod factors[247, 251]. Studies on *Lotus japonicus* Nod receptor encoded by the gene *NFR5* have demonstrated *in vivo* the importance of the second LysM domain for Nod factor binding, together with the importance of a hydrophobic amino acid residue (Figure 3-12). A homology model of this domain has been proposed to identify a possible binding site[252]. *Ohnuma et al.* characterized the carbohydrate binding site of two LysM domains of the chitinase from the fern *Pteris ryukyuensis* using NMR and calorimetric methods[253]. They identified a residue critical for binding, proposing a homology model and possible binding modes for chitooligosaccharides (Figure 3-12).

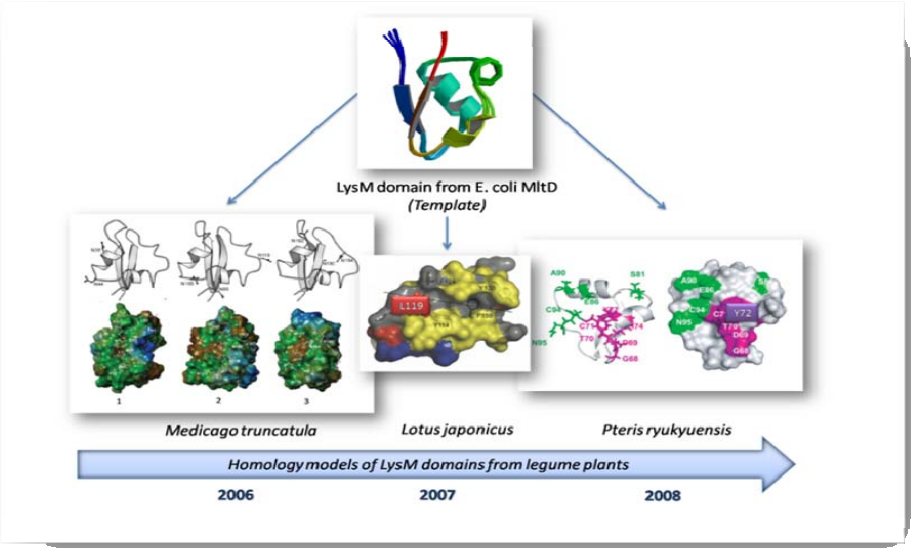


Figure 3-12 - Homology models available in the literature of LysM domains from M. truncatula, L. japonicus and P. ryukyuensis Chitinase A.

3.5.2 Results

The establishment of the agronomically and ecologically important legume-rhizobia symbiosis is initiated by beneficial soil bacteria which secrete lipochitooligosaccharidic Nod factor signals. The latter are perceived by legume plants that will change their morphology to accommodate bacteria, able to start the nitrogen fixation. In the *Medicago truncatula* - *Synorizobium meliloti* model, Nod factors are perceived by recently-cloned legume receptors, encoded by the *NFP* and *LYK3* genes). It is clear that the knowledge of structural features is required to fully clarify the entire symbiotic mechanism. Using molecular modeling strategies, combined with NMR analyses performed by Maria Morando and Dr. J.J.Barbero in Madrid, conformational studies of a new generation of Nod factor analogues have been performed. These studies provide support for the idea that the carbohydrate moiety of Nod factor plays a key role in the interaction with its receptors, while the lipid moiety, highly flexible, acts as regulator of the ligand specificity, probably interacting with a second and still unknown partner.

Homology modeling techniques were used to predict the three dimensional structures of receptors involved in the Rhizobia – legume symbiosis (*Medicago truncatula – Sinorhizobium meliloti* model). These models have been used and still used as reference structures to support biological data by the NODPERCEPTION European network whose activities are described in the next pages.

Unlocking the key to the legume–rhizobia symbiosis

Effect of rhizobia inoculation on soybean growth (centre) and nodules on inoculated roots (inset)

The symbiotic relationship between legumes and rhizobia is of great ecological and agronomic importance. **Julie Cullimore** of the NODPERCEPTION network expands on an EC project, providing interdisciplinary training to young researchers across Europe, to advance knowledge in this area

Growing crops commercially is a complex process, and in many cases about half of the economic cost is incurred through the production, supply and application of nitrogen fertilisers. Legumes – plants from the Fabaceae family, such as peas, beans, soybeans, lucernes, lupins and clovers, are a major source of protein-rich foods and animal feed, but do not require added nitrogen fertilisers, as they are able to make their own. This property is attributable to the ability of legumes to form a symbiosis with nitrogen-fixing soil bacteria called rhizobia. Each legume recognises its own group of rhizobia, and allows only these specific bacteria to infect their roots, producing root outgrowths called 'nodules'. Within these nodules the bacteria use plant-made energy sources to convert atmospheric nitrogen gas to a form of fixed nitrogen that the legumes can then use as their own nitrogen fertiliser. Some of this fixed nitrogen is later released into the soil, and because of this property legumes are also able to increase soil fertility, in turn bringing significant benefits to subsequent plants in crop rotations. Growing legumes is thus extremely important to sustainable agriculture because it reduces the need to use chemically-produced nitrogen fertilisers, which are energy-demanding to produce and use, and furthermore can also lead to environmental pollution.

The question of how the rhizobia are specifically recognised by their host legume plants is the key subject being addressed by the NODPERCEPTION project, an EC-funded Marie Curie Research Training Network bringing together eight European partners, who currently employ a total of 10 young researchers in biology, biophysics and chemistry. The key to recognition is the fact that the rhizobia all make different types of recognition signals, known as Nod factors: these signals (termed lipochitooligosaccharides, because they have a special oligosaccharide backbone and a lipid tail) instruct the plant to make a nodule and then act a bit like entry tickets, allowing the 'partner' rhizobia to enter into the nodule to fix the nitrogen.

The NODPERCEPTION project aims to:

- Improve our understanding of the molecular mechanisms of interaction of the Nod-factor signal with its plant receptor protein
- Address the question of how this molecular interaction leads to recognition of the partner rhizobia, both from other rhizobia and also from the millions of different bacteria in the soil
- Advance our understanding of how activation of the Nod-factor receptor activates the plant root cells, leading to nodule formation and infection by the partner rhizobia

Receptors and signals

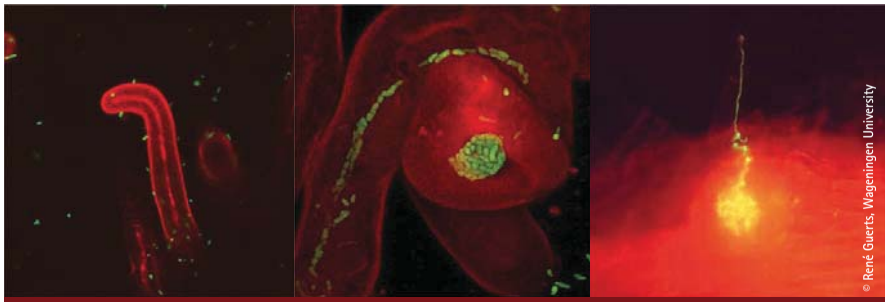
Tackling these aims requires all the partners to work on the same experimental system. We have chosen to work on the receptors and signals for the model legume, *Medicago truncatula*, for which a number of International initiatives (including the EC Grain Legumes Integrated project or 'GLIP') have developed genetic and genomic tools, including those based on the sequencing of the genome. These tools have accelerated the progress of our research while furthermore, because of the conservation between legume genomes, information obtained on *Medicago truncatula* is readily applicable to Nod factor recognition in important legume crops.

The study of molecular signals, receptors, and plant responses requires the adoption of a multidisciplinary approach, integrating the work of chemists, biophysicists and biologists, who play complementary roles in pursuit of the project's overall objectives. For example, within the NODPERCEPTION network the chemists synthesise modified Nod factors and the biochemists produce receptor proteins; the two types of molecule are then used for structural and interaction studies by the biophysicists, and for cellular and plant studies by the biologists. A range of techniques are used, including Nuclear Magnetic Resonance (NMR), Surface Plasmon Resonance (SPR) and microscopy measurements of Fluorescent Resonance Energy Transfer (FRET). These powerful techniques allow us to relate molecular changes to changes at the level of individual plant cells, which in turn informs us of how the plants respond to the bacterium itself. Thus, by combining studies at different levels of resolution a complete picture can be built up, ranging from the molecular structures involved in Nod factor perception right through to the production of rhizobia-containing nodules on the legume roots.

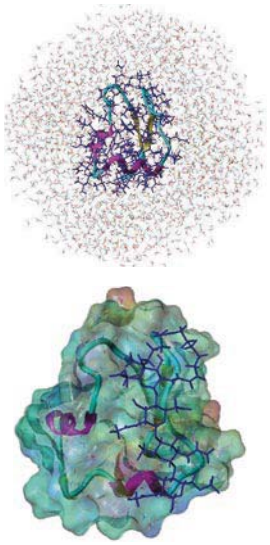
Nod factors instruct the plant to make a nodule and then act a bit like entry tickets, allowing the 'partner' rhizobia to enter into the nodule to fix the nitrogen

Such an integrative, broad-based approach would have been unimaginable to previous generations of researchers, and it is the development of new, improved technologies in biology, biophysics and chemistry over recent years which have made ambitious projects like NODPERCEPTION technically feasible. However, the interdisciplinary nature of NODPERCEPTION requires not only the work of the different specialists across a range of fields, but also a scientific environment conducive to working across National boundaries towards a common objective. In this aspect it must be acknowledged that the EC is instrumental to the project by providing the funding necessary to both link the partner laboratories, and also to employ enthusiastic young researchers to carry out the work.

For many of the young researchers this is the first time they have participated in an international, multidisciplinary project, which can be a very challenging environment. The fact that NODPERCEPTION is funded by the Marie Curie Actions means that many of the young researchers are based outside their home country, while the nature of the subject matter demands that they work closely with other partners in the network. Understanding the different disciplines and how they can be used for the project is an important part of their training, and is based on workshops organised by the young researchers themselves. Bi-annual meetings between the eight partners are an important means of coordinating the work and allowing the young researchers to present their results and to discuss how to proceed. Most of the young researchers are carrying out this work as part of a PhD project, and the network allows them to undertake part of their project at another partner's laboratory, where they use complementary techniques and are exposed to different schools of thought. Presenting their work at network meetings and at International conferences trains the young researchers in scientific communication.



Fluorescent microscopy shows rhizobia attaching and entering a root hair cell and then entering the developing nodule



© Alessandra Nurisso / CERMAV-CNRS

A combination of NMR spectrometry and molecular modelling is used to understand the Nod factor – receptor interaction

The network also organises training in complementary skills such as scientific writing.

This multifaceted approach to the training brings significant benefit to the young researchers themselves. They develop critical and conceptual thinking which enables them to formulate scientific hypotheses and then to find the best possible approach to test them so as to answer complex scientific questions. The ability to see the bigger picture, to put individual research projects in context and, above all, to communicate with scientists in other disciplines, are qualities which the young researchers believe will be enormously important to their future careers. Providing effective training to young researchers boosts European research in turn, because it helps develop skilled researchers who are able to work more effectively in the emerging modern, interdisciplinary, collaborative research environment.

As rhizobial and Nod factor inoculants are used commercially to improve legume growth in low fertiliser regimes, further advances in our knowledge of Nod factor recognition could potentially help improve the design and use of inoculants to favour the establishment of the legume-rhizobia symbiosis using environmentally-friendly technologies. However, our work on Nod factor perception, which is largely fundamental, may also lead to advances in knowledge which could be of real benefit to other systems. For example, it could lead to further studies on the establishment of a related plant symbiosis with mycorrhizal fungi, which is an important factor in improving the nutrient uptake of many plant species, including cereals. Also, knowledge of the structural basis of the recognition of the unusual lipo-chitooligosaccharide Nod factor molecule could help advance the studies on the recognition of other lipids and oligosaccharides, which play very important signalling roles in plants and animals. ★

Providing effective training to young researchers boosts European research because it helps develop skilled researchers who are able to work more effectively in the emerging modern, interdisciplinary, collaborative research environment

At a glance

Full Project Title

NODPERCEPTION : an integrated, interdisciplinary approach to Nod factor perception

Project Information

NODPERCEPTION is a Marie Curie Research Training Network currently training eight PhD students and two post-doctoral researchers using multidisciplinary techniques

Project Partners

- Dr Julie Cullimore, Dr Clare Gough and Dr Jean-Jacques Bono, INRA-CNRS, Toulouse, France,
- Dr René Geurts, Wageningen University, The Netherlands,
- Dr Giles Oldroyd and Prof Allan Downie, The John Innes Centre, Norwich, UK,
- Dr Gabriella Endre, Biological Research Center, Szeged, Hungary,
- Prof Dorus Gadella, Amsterdam University, The Netherlands,
- Dr Anne Imberty and Dr Hugues Driguez, CERMAV-CNRS, Grenoble, France,
- Prof Jesús Jiménez-Barbero and Dr Javier Cañada, CSIC, Madrid, Spain,
- Prof Slawomir Pikula, Nencki Institute of Experimental Biology, Warsaw, Poland.

Contact Details

Dr Julie Cullimore
Coordinator
T: +33 5 61 28 55 13 / 53 22
F: +33 5 61 28 50 61
E: Julie.Cullimore@toulouse.inra.fr
W: http://medicago.toulouse.inra.fr/NODPERCEPTION

Julie Cullimore

Project coordinator

Julie Cullimore is an INRA Director of Research, working in the Laboratory of Plant-Microbe Interactions near Toulouse. She initially worked in the UK before moving to France in 1991.

NMR AND MODELING REVEAL THE STRUCTURAL FEATURES OF SYNTHETIC NODULATION FACTORS

M.A.Morando[1*], A. Nurisso[2*], A. Imberty[2], F.J. Cañada[1], J.J. Barbero[1]

[1]Centro de Investigaciones Biológicas, Consejo Superior de Investigaciones Científicas, Madrid, Spain.

[2]CERMAV-CNRS (affiliated with Université Joseph Fourier and belonging to ICMG), BP 53, F-38041 Grenoble, France.

Key words: Nodulation factors, NMR, molecular dynamics, conformational study

* *These authors contributed equally to this work*

Abstract

Nodulation (Nod) factors are lipochitoligosaccharides produced by Rhizobia bacteria involved in the symbiotic process of leguminous plants. Analogs of nod factors have been synthesized. Their conformational behaviour is of interest because it may be directly correlated to the biological activity.

Conformational studies have been performed combining molecular dynamics simulations in explicit water and NMR: data revealed that the glycosidic head group can assume restricted conformations whereas chemical modifications of the lipid chains, highly flexible in a water environment, influence the global shape of the molecules. The synthetic compounds, for which structural properties have been characterized in this work, compared to the native one, may be promising probes for understanding the role of Nodulation factors in the molecular recognition process. Collected structural data could be used in future to rationalize and understand their biological activity and affinity towards a putative receptor.

Introduction

Nodulation factor signals (Nod factors) are lipochitoligosaccharides synthesized and secreted by Rhizobia bacteria that play a fundamental role in the chemical dialogue between bacteria and legume plants in the rhizosphere[1]. Their production is under the control of bacterial genes activated by flavonoids derived from host plants. Active at pico-nano molar concentrations, nodulation factors are able to interact with legume plant roots and trigger formation of nitrogen fixing nodules necessary to start the nitrogen fixation[2-4]. The plant's perception of these signals depends on specific receptor like kinases containing LysM domains but the structural details of the interaction are still to be elucidated[5-7]

As the symbiosis induces soil fertility, the deep understanding of this process could be clearly useful for developing products able to promote this process.

Nodulation factors have a common structure characterized by a backbone of three to five N-acetylglucosamine (GlcNAc) residues, bearing an amide-bond fatty acyl chain at the non-reducing end and a variety of additional substituents. The substituents depend on the bacterial strain and the structural variations of the basic carbohydrate skeleton as well as the variation on the lipid moiety determine the host specificity in the symbiotic process[8].

The Nod factors from *Sinorhizobium meliloti*, the symbiont of *Medicago* plant, consist of a chitotetraose with O-sulfate on the reducing sugar, and O-acetyl and C16:2 Δ2,9 fatty acid chain on the terminal non-reducing one. The sulphate is required for all biological activity by *Medicago* plant cells, while the acetyl and acyl chains are important for infection and nodulation[9-11].

The structural role of the lipid chain is still subject of discussions: it has been proposed to be required for for receptor specificity, to play a role in membrane-anchoring or to help in the formation of micro - aggregates. The use of synthetic Nod factors and analogs demonstrated that the length and the structure of the acyl chain play a role on both the morphogenic activity on legume roots[12] and the affinity towards receptor-enriched cell suspensions[13, 14].

In the absence of crystal structures of natural nodulation factors or their derivates, NMR and molecular modeling have to be used to understand the effect of lipid chain modifications on conformation and shape of Nod factors. Natural nodulation factors have been studied by a combination of dynamics calculations in implicit water combined with NMR experiments[15]. The glycosidic backbone of the nodulation factors from *S. fredii* displays a major and stable

conformation, with a solvent dependent orientation of the fatty acyl chain that can mostly adopt a quasi parallel orientation to the oligosaccharide chain. Simulated annealing and NMR studies were also carried out on natural Nodulation factors from *S. meliloti* to obtain information about their conformational behavior[16]. In this work, the importance of the lipid moieties has been highlighted. Distinct lipid orientations were found, supporting the idea that they can contribute to a high degree of ligand specificity to Nod factor receptors.

We present here the conformational study of Nod factor analogs presenting insertion of a benzamide group in the acyl chain at the sugar non reducing end (Figure 1). The synthesis of **1** and **2** was previously described and the reported affinities to culture cell are in the same range as compound **N**, a deacetylated analog of the natural Nod factor of *S. meliloti*[13]. Additional analogs with different saturation schemes (compounds **3** and **4**) have also been included in the present study. Molecular dynamic (MD) simulations and measurement of nuclear Overhauser enhancements (NOEs) have been performed in order to assess the relative orientation of the lipid and tetrasaccharide moieties in a uniformly hydrated environment and to explore how chemical modifications can influence the shape and behavior of these compounds in solution.

Materials and Methods

Material

Compounds **1** and **2** have been synthesized as described previously[13]. Synthesis of compound **3** and **4** has been described by J.M. Beau et al.[17].

Nomenclature

A schematic representation of the nodulation factors taking into account in this study, along with labeling of the heavy atoms, is shown in Figure 1.

In this study we consider nine main conformational degrees of freedom.

The flexibility of the carbohydrate scaffold and the hydroxyl groups is described by the following torsion angles:

φ = O5-Cl-O1-C4 $\quad$ ψ = Cl-O1-C4-C5 $\quad$ ω = O5-C5-C6-O6

The torsion angle between the carbohydrate scaffold and lipid chain in the nodulation factor analogs is defined by

η = Oa-Ca-C1e-C2e

The orientation of the fatty acid group linked to the carbohydrate scaffold is defined by the five torsion angles:

α1 = C1e-C2e-O1f-C1f (ortho compound) / C2e-C3e-O1f-C1f (meta compounds)
α2 = C2e-O1f-C1f-C2f (ortho compound) / C3e-O1f-C1f-C2f (meta compounds)
α3 = O1f-C1f-C2f-C3f
α4 = C1f-C2f-C3f-C4f
α5 = C2f-C3f-C4f-C5f

The behaviour of the acyl chain of the native compound was described taking into account the following torsion angles:

α1 = C1f-C2f-C3f-C4f
α2 = C2f-C3f-C4f-C5f
α3 = C3f-C4f-C5f-C6f
α4 = C4f-C5f-C6f-C7f
α5 = C5f-C6f-C7f-C8f

NMR spectroscopy

For NMR measurements, ~1 mg of each compound was dissolved in 1 ml of D_2O. Compounds were recovered and dissolved again in 1 ml of H_2O/D_2O 15% for the spectra in water solution. In both cases, the concentrations were between 1-2 mM.

^1H NMR NOESY and TOCSY spectra were obtained using the standard pulse sequences with watergate for experiments in H_2O/D_2O and presaturation for experiments in D_2O provided by manufacturer at two different spectrometers: BRUKER NMR spectrometer operating at

frequency of 800 MHz (spectra in D_2O of molecules **1**, **2**, **4**) and VARIAN NMR spectrometer operating at the frequency of 900 MHz (spectra in H_2O/D_2O 15% of compounds **1**, **2**, **3**). Spectra were collected with mixing times between 60 and 300 ms at the temperatures ranging from 278-288 K. For DOSY experiments, all samples were prepared in D_2O. The standard BRUKER DOSY protocol was used at 298 K on an AVANCE 500 MHz equipped with a broad-band z-gradient probe[16].Thirty-two 1D ^1H spectra were collected with a gradient duration of $\delta = 2$ ms and an echo delay of $\Delta = 100$ ms. Acquisition times of 8-15 mins (8-16 scans) were required for the samples. The ledbpg2s pulse sequence, with stimulated echo, longitudinal eddy current compensation, bipolar gradient pulses and two spoil gradients, was run with a linear gradient (53.5 G cm^{-1}) stepped between 2% and 95%. The 1D ^1H spectra were processed and automatically baseline corrected. The diffusion dimension, zero-filled to 1k, was exponentially fitted according to preset windows for the diffusion dimension ($-8.5 < \log D < -10.0$).

Molecular dynamics protocol

Initial models building

Nodulation factors starting structures were built usyng Sybyl 7.3 (Tripos Associates, St. Louis, MO). As starting point, the lowest energy conformations of each glycosidic linkage were selected from available adiabatic maps of β-D-GlcpNAc(1,4)β-D-GlcpNAc (http://www.cermav.cnrs.fr/glyco3d/). The acyl chains were built in extended conformation. For the linkage between the carbonyl and benzene group (η) of coumpounds 1 to 4, two different starting conformations were considered, *syn* and *anti*. Compound **1** was built only in *anti* conformation according to data derived from NMR experiments (Results session).

To the glycosidic moiety, which is identical in all compounds, the parameters from GLYCAM06 force field[18] have been assigned while acyl chains were described using parameters adopted from Wang et al[19]. For charge calculations, the carbohydrate moiety and the lipid one were separated in two compounds. A methyl group was used as cap for the sugar scaffold whereas the ACE (CH_3CO) group was used as cap for the lipid chains in order to mimic the system present at the linkage with the sugar scaffold. Charges were calculated using the RED server[20]. The tool allows the automatic calculation of the electrostatics properties consistent with the different force fields applied. The charge for each glycosidic unit was set to -1 for the

presence of the sulphate group whereas the lipid portion was set as neutral. The carbohydrate and lipid portions were then merged together. A new topology input file was created for each system in which electrostatic and van der Waals scaling factors (*SCEE* and *SCNB*) were set to the unity for the sugar scaffold (Glycam force field) and 1.2 /2 for the lipid portion (Amber force field).

Computer experiment details

Each compound was placed in a 10 Å depth truncated octahedral box of explicit TIP3P waters and sodium ions were added in order to neutralize the systems. Equilibration and production phases were carried out by PMED module implemented in Amber 10 (*University of California*). The equilibration phase consisted on energy minimization of the solvent to remove the initial bad geometries followed by an energy minimization of the entire system without restraints. The system was then heated up from 10K to 288 K during 100 ps MDs with weak restraints on the solute followed by 100 ps dynamics at constant temperature and constant pressure of 1 atm. The MDs production phase lasted 10 ns under constant pressure of 1 atm and constant temperature of 288 K, according to NMR experiments, controlled by the Langevin thermostat with a collision frequency of 1.0 ps-1. During the simulations, the SHAKE algorithm was turned on and applied to all hydrogen atoms[21]. A cut-off of 10 Å for all non bonded interactions was adopted. An integration time step of 2 fs was employed and periodic boundaries conditions were applied throughout. During all simulations the particle mesh Ewald (PME) method was used to compute long range electrostatic interactions[22].

The analyses of the simulations were performed using Ptraj module of AMBER10 (*University of California*). The visualization of the trajectories was performed using VMD software[23]. Data were processed and plotted usyng R software. Figures were prepared using PyMOL Molecular Graphics System (*Scientific, Palo Alto, CA*).

Results and Discussion

Molecular dynamics simulations

10 ns MDs simulations in an explicit water environment were performed for each compound described in Fig.1, taking into account the different conformers in solution around the angle η. Details about the simulations are reported in Table 3S.

Analysis of the carbohydrate moiety

The conformational behavior of the tetrasaccharide moiety has been analyzed during the trajectories of all compounds. All the pyranose rings maintain their $^4C_1$ conformation along the simulations. The amido groups at position 2 of all monosaccharides adopt stable conformations with H-C2-N2-H torsion close to 180° and a *trans* orientation of the NH-CO bond. Hydroxymethyl groups at position 6 show a more flexible behavior with frequent transition for the ω torsion. The *gauche-gauche* (ω =-60°) and *gauche-trans* (ω =60°) rotamers are observed during most of the simulation time, in agreement with stable conformation for *gluco* configuration[24]. Only the GlcNAc reducing unit presents mainly a *gg* orientation. The *gt/tg/gg* ratio for all the compounds is reported in Table 1S.

The scaffold shape is analyzed by monitoring the values of the Φ and Ψ conformations at each glycosidic linkage. The trajectories are superimposed on the adiabatic map of the parent disaccharide β-D-GlcpNAc(1,4)β-D-GlcpNAc (Fig.2 and Fig1S). The conformations of all compounds mainly cluster in the lower energy conformation region (≈ 95% occupancy) corresponding to the minimum of the adiabatic map, with the exception of minor torsion angle populations that occupy close areas. Even in the absence of transition to remote low energy conformations, local flexibility is observed with variations of up to 100° in Ψ. The average glycosidic torsion angles for each compound can be found in Table 1. The average angle values were similar in each compound and the Φ and Ψ torsions of the glycosidic scaffold appear to be very stable around the main minima ≈ -80°/-130°; significant differences arose only in minor populations. For instance, the glycosidic linkage B of the compound **1** populates a region of

space characterized by Φ and Ψ torsion angles of ≈ 100/-130 for 1% of the MD simulation, but the same conformation is not found in the compounds meta substituted.

The hydrogen bonds between adjacent residues are also analyzed. The occupancy of the potential hydrogen bonds along the simulations is reported and listed in Table 2. Strong polar contacts between the acceptor oxygen O5 of each GlcNAc residue and the donor hydroxyl group in position 3 of the successive residues can be established along the simulations (occupancy ≈ 90%). These strong interactions contribute to the stability of the sugar moiety in solution and explain the restricted variation of Φ and Ψ angles along the simulations. Exclusive polar contacts are found for the compound 1, in which the nitrogen atom of the non reducing end is involved in contacts with the oxygen atom from the lipid moiety. The sulfate group at the reducing end of the carbohydrate moiety does not establish contact with the rest of the oligosaccharide.

Conformational analysis: molecular dynamics studies on the lipid chain

Analysis of torsion angles. The flexibility of the different lipid chains was analyzed by monitoring the torsion angles in the portion adjacent to the carbohydrate and looking for contact between the carbohydrate and the acyl chain. The torsion angles analyzed are the five ones adjacent to the benzyl group (or equivalent ones for compound **N**) since the ones between the carbohydrate and the benzyl ring do not display flexibility. Possible values for each torsion angle were plotted in histograms (Figure 3-2S) where density peaks reveal the most frequent dihedral values in all the simulations.

For the lipid chain of the native nod factor **N**, the planar torsion angle adjacent to the carbohydrate is followed by the dihedral α1, whose histogram contains a spread torsion distribution mainly around -97◦; α2 angles values are mostly distributed around two peaks centered on 68◦ and 178◦ whereas α3, α4 and a5 present three different peaks around 65◦, 180◦ and -70◦. However, in the five most populated families, they mostly assume orthogonal features promoting a parallel behavior respect to the carbohydrate scaffold.

The compound **1** that is ortho substituted appears to be more rigid than the other Nod factor analogs. The torsion angle α1 is rather flexible and can assume values between 120◦ to -120◦. The torsion angle α 2 is much more rigid with values close to 180◦. The torsion angles α1 and α2

strongly influence the three dimensional shape of compound **1** This behavior is mostly due to the stable hydrogen bond that the oxygen of the acyl chain establishes with the nitrogen of the benzamide group. The torsions α3, α4 and α5 result more flexible: main density peaks were found though the histograms analysis, around 65◦/178◦/-65◦, 73◦/178◦ and 93◦/-108◦ respectively.

In the compounds with a meta substitution, differing in terms of insaturation at the position 4 of the acyl chain (Fig. 1), the rotamers at α1 are different. The *anti* conformer of the compound **2** (angle η = 180◦) presents distinct distribution of the torsion angle α1 around 180◦ with less populated geometries around 5◦. The *syn* form of the same compound (angle η = 0◦) presents distribution of the torsion angle α1 around 125◦ and less values around 45◦. The molecule **3**, with a single bond on the acyl chain, presents more degree of freedom around α1 respect to the structures previously described. The angles α2, α3 and α4 present mostly flexible behavior with general preference for the staggered conformation. The conformational behavior of torsion angle α5 depends on the nature (single, double or triple bond) of the adjacent linkage. The variation of flexibility of this torsion angle affects only the possible shape of the extremity of the lipid.

Family clustering. A clustering in families has been applied to 10000 snapshots from each MDs simulation in order to visualize general tendencies. A particular snapshot from the MDs simulation is associated to a family only if the five dihedral angles differ to the corresponding density peak by less than 2* σ degrees. The characteristics of the five most populated families are listed in Table 3 and 2S. The low population of each family, reported in percentage, reflects the high flexibility of the acyl chains: several snapshots have not been classified because of the narrow range that characterizes each class. The most representative structures from each family were superimposed and displayed in Figure 4.

The five representative families of the native Nod factor **N** have a common α1 angle value according to the main peak of the histogram previously described (-98◦). Whereas the angle α2 has values around 68◦ in the three most populated families and 178◦ in the other ones, the angle α3 shows common orthogonal features. Only the most populated family (12.8%) presents α4 values around 63◦; in the other cases this angle is around 180◦. α5 presents angle values around 63◦ for the families 1 and 3, 178◦ for the families 2 and 4, and -73◦ for the family 5.

The ortho compound **1** shows two distinct values for the angle α1, around 158◦, in the most populated families, and -128◦. The orthogonal features of the α2 reflect the relative rigidity of this compound, due to the hydrogen bonds between the lateral chain and the carbohydrate scaffold. The angle α3 oscillates between -65◦ (families 1, 4, 5) and 65◦ (families 2 and 3). The α4 dihedral results planar with an exception for the family 1, in which this angle takes values around 73◦ while the dihedral α5 mainly shows values around -108◦ (families 1,2,4), 93◦ (family 3) and 158◦ (family 5).

The most populated family of the molecule **2** syn (7.7%) presents values α1 and α2 around -60◦ and -68 respectively. Structures of this family also have α3 angles around 178◦ (same behavior is reported for the families 2, 4, and 5) and α4 angles around -63◦**.** The dihedral angle α5 reports values around -98◦**,** like the structures that belong to the families 4 and 5. The anti conformer of the synthetic Nod factor **2** shows a main populated family (14.9%) with α1 angle around -165◦. The angle α2 is planar, as reported for most part of the families, while the angle α3 assumes values around -68◦, like for the families 3 and 4. The dihedral angle value α4 oscillates around -68◦ whereas the angle α5 assumes values around 108◦. In the less populated families, α4 and α5 dihedral values are around 178◦/68◦ and -103◦ respectively.

The molecule **3** syn presents two main families populated for the 11.9 % and 10.1%. These families share common geometrical features. In particular, angles α1, α4 and α5 assume values around 145◦, 178◦ and -178◦ respectively. α2 assumes values around -68◦ (family 1) and 178◦ (family 2), whereas α3 assumes values around -65◦ (family 1) and 65◦ (family 2). The compound **3** anti shows 17.5% structures clustered in the family 1, with angles α1 distributed around 135◦, α2 and α3 around 63◦, α4 around -115◦ and α5 around -145◦.

The families for the meta substituted compounds **4** were classified taking into account the four dihedrals α1, α2, α3, α4 because the angle α5 does not show peaks corresponding to a particular dihedral value in the histogram .The syn conformation presents a main family (27.7%) with α1 around 125◦. Snapshots belonging to this family have dihedrals α2 around -103◦, α3 around -63◦ and α4 around -175◦. This family mainly differs from the others around the angle α2 (173◦/-63◦) and α3 (178◦). The angle α4 oscillates between -65◦ and -175◦. The anti conformation of the compound **4** presents more flexibility in water, reflected to the poor % of conformers that characterize each family. The dihedral α1 takes values around -175◦, except for the family 3

(5◦), and α2 around -178◦, except for the family 3 (65◦). The angles α3 and α4 assume different values around -68◦ (families 1, 2, 3), 63◦ (family 4) and -178◦ (family 5).

General shape. In order to rationalize the different sizes that these molecules can adopt in solution, the average radius of gyration was calculated over 10 ns for each structure. Structural information about the *syn* and *anti* conformations of the meta substituted compounds were joined together, in order to have a general idea about the shape that they can adopt in solution (Table 4S).

The smallest radius of gyration was obtained for the compound **1**, 7.3 +/- 0.4 Å. This reflects the restricted shape that this molecule can assume along the simulation, compared to the other studied molecules. The Nod factor analogs meta substituted have radius of gyration between 7.5 and 7.8. Å. The larger values are obtained for the compounds **2** and **4**, where the presence of an unsaturated bond in the acyl chain results in an elongated form (7.8+/- 0.8 and 7.7 +/- 0.8 Å). Compound **3** and native Nod factor **N** have radius of gyration of 7.5 +/- 0.8 and 7.6 +/- 0.8 Å, respectively.

The distance between the sulfate group at the carbohydrate reducing end and the extremity of the acyl chains was also evaluated. As expected, the smallest C-S distance were obtained for the ortho substituted compound **1** (12.0 +/-5.1 Å), followed by the meta substitute compound **3** and the native molecule **N**, with distances of 12.9 +/-5.4 Å and 14.6 +/-4.8 Å, respectively. According to the information obtained though the evaluation of the radius of gyration (Table 4S), these compounds occupy a more limited space in a water environment in which lipid moieties show preferred conformations close to the saccharidic portion along the simulations. The compounds **2** and **4**, in which the lipid chains assume an axial position respect to the sugar scaffold for a significant amount of time, show an increased C-S distance of 14.6 +/- 5.6 Å and 15.3 +/- 5.2 Å respectively. This behavior is probably due to the orthogonal features that the dihedral angles α1-3 of these compounds mostly assumes along the simulations.

Inter-residue contacts. In order to correlate MD simulations and NMR experiments, inter-residue distances were measured for the carbon-linked proton and short distances were analyzed. For the chitotetraose moiety, only distances between the anomeric proton and the H3, H4, H5a, H6a and H6b across the glycosidic linkages display values smaller than 4 Å. No significant

differences could be observed between compounds and only the data corresponding to compound 1 are listed in Table 4.

Contacts between the proton of the benzyl group and the non reducing end of the sugar scaffold are observed for compounds meta substituted 2 to 4.

In particular, for the syn conformers, the hydrogen atom H2-E makes contacts with the hydrogens H3-A (4.1 Å), H1-A (4.1 Å), HO6-B (4.4 Å), H6a-B (3.7 Å) and H6b-B (4.0 Å). For the compounds anti, the hydrogen atom H6-E is involved in contacts with the same atoms listed in the previous case, with analogs average distances.

As for contacts between the acyl chain and the carbohydrate moiety, measurement of inter-proton distances along the simulations reveals possible geometries in which the lipid chains can assume relatively close positions respect to the sugar scaffold for a brief period of time. However, the average inter-proton distances from the different MD simulations indicate that the native compound and the ortho substituted molecule are the only two molecules that establish relatively stable van der Waals contacts between sugar and lipid scaffolds along the simulations. Concerning the native compound **N**, the hydrogen atoms H1-F and H2-F from the lipid chain can establish rather stable contacts (< 4.5 Å) with the hydrogen atoms H6a-B and H6b-B of the carbohydrate moiety. The ortho substituted molecule **1** presents contacts between the hydrogen atoms HC2-f /HC3-f from the lipid chain and the hydrogen atoms HC3-a from the sugar scaffold (Table 4).

Comparison between NMR and theoretical data

To complete the structural characterization of the Nodulation factor analogs, comparisons with NMR experimental data were performed when possible.

First, the aggregation state of the Nodulation factor analogues was studied. The amphypatic character of these molecules, due to the presence of the hydrophobic chains, could confer non-monomer states to these compounds.

By means of Diffusional Order Spectroscopy (DOSY) it was possible to confirm that at concentration between 0.1 and 1 mM, the molecules did not show any observable aggregation. Thus, the experimental data were unambiguously correlated with monomeric species, which are those expected at the very low physiologically active concentrations.

For all these molecules (**1-4**), the NMR assignments were accomplished through NOESY and TOCSY experiments, either in D_2O and in H_2O/D_2O. The concentration of monomeric species (~1mM) was sufficient to obtain a significant number of assignments. Nevertheless, the severe signal overlapping characteristic of oligosaccharidic structures precluded a complete assignment of all the pyranose rings hydrogen resonances. A similar situation occurred in the lipid proton region. However, there was a clear distinction between signals arising from the anomeric protons (H-1), for the H2-H6 sequence (3.2-5.1 ppm), for the lipid moiety (0.5-1.5 and 2.15-2.3 ppm) and the aromatic region (6.7-7.6 ppm) (Tables 5-8). The amide region (8-8.7ppm) was also analyzed using experiments in H_2O. The NOESY and TOCSY spectra in H_2O/D_2O contained information on the contacts between the anomeric proton H1 of each residue and the H4 (although with overlapping) and H6 protons of the following residues. When these contacts do exist at the same time, they can be safely correlated with allowed exo-anomeric conformations in which the glycosidic torsion angles assume values of $\Phi/\ \psi$ around -80/-130. Indeed, in all compounds, these cross peaks were found in the NOESY spectra, corresponding to the major orientations found in the MD simulations (Fig.6- 7).

During the 10 ns simulations, the $^4C_1$ chair conformation of each glycosidic residue remained stable. This evidence was confirmed by the proton-proton averaged distances listed in Table 4. The calculated distances can be also directly correlated to the intra residual NOE cross peaks (Fig. 7).

The anti-conformation of the amide torsion angles of the sugar residues predicted in the MDs simulations was confirmed by the measurement of the coupling constants J1-3 between each amide proton and the proton H2 of the sugar residues, displaying values > 8 Hz in NMR spectra, typical for anti-type torsional angles.

Concerning the torsion angle η that defines the conformation and planarity of the linkage with the aromatic moiety, the NMR data revealed the presence of two different geometries, anti and eclipsed, inter-convertible in solution, corresponding to two orientations around this dihedral angle. In fact, the NOESY experiments for the meta-substituted structures showed two mutually exclusive NOE contacts between the amide proton and the aromatic protons (Fig. 5A). Thus, this experimental evidence confirmed the existence of two main stable conformations.

In contrast, for the ortho-substituted compound **1**, one unique NOE contact between the amide proton and the aromatic one was found (Fig. 5B). This evidence indicates the presence of only

one possible value for the planar torsion angle and, therefore, of only one stable conformation in solution. This information was taken into account in the MDs simulations: only one conformer for compound **1** was considered.

The 10 ns MDs simulations revealed possible, although short timing, interactions between sugar scaffold and the lipid moieties. Unfortunately the strong overlapping observed in the NOE spectra did not allow the non ambiguous NOE assignments of possible sugar-lipid interactions that, in any case, cannot be excluded.

Nevertheless, for the ortho-substituted compound, it was possible to assign two NOE contacts between protons belonging to the lipid chain and the non-reducing end of the tetrasaccharide (Fig. 6), confirming the short average inter proton contacts derived from the MD simulation of **1** (Table 4).

Conclusions

Natural and synthetic nodulation factors are flexible molecules, which may adopt a variety of shapes in water solution. Still, their average three dimensional structures of synthetic nodulation factors can be characterized using a combination of experimental NMR and MD simulation data. The results of these studies indicate that chemical modifications influence the spatial disposition of the lipid chain and the shape that they can assume, while the carbohydrate portion displays a relatively stable dynamic behaviour. The lipid moieties have a considerable degree of freedom and can assume rather different orientations in equilibrium. The classification of the acyl chain conformations in solution gives information about the preferential disposition of these molecules, influenced by an explicit water environment. Still, different shapes and dynamics are accessible to the natural and synthetic molecules, depending on the chemical nature of the substitution, the aromatic ring, and the character of the insaturation of the lipid chain. Thus, different biological activities have been reported for the different analogues.

The data herein reported support the idea that the carbohydrate portion plays a key role in the nod perception mechanism, while the lipid moiety modulates ligand specificity to nodulation factor receptors. From the results of the MD simulation of nod factors analogs and the respective native compound, an incoming protein receptor would initially be expected to recognize the carbohydrate scaffold. The high flexibility of the acyl moieties could play an important role in

modulating the recognition process and the biological response, decreasing accessible portions of the oligosaccharide to the binding site. The different behavior of the lipid moieties elucidated in this study could explain and support the different biological activity of these compounds.

The receptor binding site, probably located on extracellular domains of plant kinases, can clearly influence the preferred conformation of these molecules. However, in absence of structural data concerning the receptor, these structural studies might be useful for explaining the different biological activity of these molecules.

Acknowledgments

The PhD thesis of M.A. Morando and A. Nurisso are financed by European Community's Marie-Curie fellowship MRTN-CT-2006-035546 (NODPERCEPTION). Madrid group thanks MICINN (Spain) for financial support (Grant CTQ2009-08536). The authors thank Dr. Ross C. Walker (SDSC) for help and precious scientific discussions.

References

1. Dénarié, J.; Debellé, F.; Promé, J. C., Rhizobium lipo-chitooligosaccharide nodulation factors: signaling molecules mediating recognition and morphogenesis. *Annual Review of Biochemistry* **1996,** 65, (1), 503-535.
2. Cullimore, J. V.; Ranjeva, R.; Bono, J. J., Perception of lipo-chitooligosaccharidic Nod factors in legumes. *Trends in plant science* **2001,** 6, (1), 24-30.
3. Long, S. R., Rhizobium symbiosis: Nod factors in perspective. *The Plant Cell* **1996,** 8, (10), 1885.
4. Spaink, H. P., Root nodulation and infection factors produced by Rhizobial Bacteria. *Annual Reviews in Microbiology* **2000,** 54, (1), 257-288.
5. Radutoiu, S.; Madsen, L. H.; Madsen, E. B.; Jurkiewicz, A.; Fukai, E.; Quistgaard, E. M. H.; Albrektsen, A. S.; James, E. K.; Thirup, S.; Stougaard, J., LysM domains mediate lipochitin–oligosaccharide recognition and Nfr genes extend the symbiotic host range. *The EMBO Journal* **2007,** 26, (17), 3923.
6. Limpens, E.; Franken, C.; Smit, P.; Willemse, J.; Bisseling, T.; Geurts, R., LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. *Science* **2003,** 302, (5645), 630.

7. Madsen, E. B.; Madsen, L. H.; Radutoiu, S.; Olbryt, M.; Rakwalska, M.; Szczyglowski, K.; Sato, S.; Kaneko, T.; Tabata, S.; Sandal, N., A receptor kinase gene of the LysM type is involved in legumeperception of rhizobial signals. *Nature* **2003,** 425, (6958), 637-640.
8. D'Haeze, W.; Holsters, M., Nod factor structures, responses, and perception during initiation of nodule development. *Glycobiology* **2002,** 12, (6), 79R.
9. Ardourel, M.; Demont, N.; Debellé, F.; Maillet, F.; De Billy, F.; Promé, J. C.; Dénarié, J.; Truchet, G., Rhizobium meliloti lipooligosaccharide nodulation factors: different structural requirements for bacterial entry into target root hair cells and induction of plant symbiotic developmental responses. *The Plant Cell Online* **1994,** 6, (10), 1357.
10. Lerouge, P.; Roche, P.; Faucher, C.; Maillet, F.; Truchet, G.; Promé, J. C.; Dénarié, J., Symbiotic host-specificity of Rhizobium meliloti is determined by a sulphated and acylated glucosamine oligosaccharide signal. **1990**.
11. Roche, P.; Debellé, F.; Maillet, F.; Lerouge, P.; Faucher, C.; Truchet, G.; Dénarié, J.; Promé, J. C., Molecular basis of symbiotic host specificity in Rhizobium meliloti: nodH and nodPQ genes encode the sulfation of lipo-oligosaccharide signals. *Cell* **1991,** 67, (6), 1131-1143.
12. Demont-Caulet, N.; Maillet, F.; Tailler, D.; Jacquinet, J. C.; Prome, J. C.; Nicolaou, K. C.; Truchet, G.; Beau, J. M.; Denarie, J., Nodule-inducing activity of synthetic Sinorhizobium meliloti nodulation factors and related lipo-chitooligosaccharides on alfalfa. Importance of the acyl chain structure. *Plant Physiology* **1999,** 120, (1), 83.
13. Grenouillat, N.; Vauzeilles, B.; Bono, J. J.; Samain, E.; Beau, J. M., Simple Synthesis of Nodulation-Factor Analogues Exhibiting High Affinity towards a Specific Binding Protein. *Angewandte Chemie International Edition* **2004,** 43, (35).
14. Gressent, F.; Drouillard, S.; Mantegazza, N.; Samain, E.; Geremia, R. A.; Canut, H.; Niebel, A.; Driguez, H.; Ranjeva, R.; Cullimore, J., Ligand specificity of a high-affinity binding site for lipo-chitooligosaccharidic Nod factors in Medicago cell suspension cultures. *Proceedings of the National Academy of Sciences* **1999,** 96, (8), 4704.
15. Gonzalez, L.; Bernabe, M.; Felix Espinosa, J.; Tejero-Mateo, P.; Gil-Serrano, A.; Mantegazza, N.; Imberty, A.; Driguez, H.; Jimenez-Barbero, J., Solvent-dependent conformational behaviour of lipochitoligosaccharides related to Nod factors. *Carbohydrate research* **1999,** 318, (1-4), 10-19.
16. Groves, P.; Offermann, S.; Rasmussen, M. O.; Cañada, F. J.; Bono, J. J.; Driguez, H.; Imberty, A.; Jiménez-Barbero, J., The relative orientation of the lipid and carbohydrate moieties of lipochitooligosaccharides related to nodulation factors depends on lipid chain saturation. *Organic & Biomolecular Chemistry* **2005,** 3, (8), 1381-1386.
17. Beau, J.; Denarie, J.; Greiner, A.; Grenouillat, N.; Maillet, F.; Vauzeilles, B. Synthetic compounds useful as nodulation agents of leguminous plants and preparation processes thereof. Patent number: US 2007/0067921 A1, 2004.
18. Woods, R. J.; Dwek, R. A.; Edge, C. J.; Fraser-Reid, 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.
19. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** 25, (9), 1157-1174.
20. E.Vanquelef, G. M., J.C. Delepine, P. Cieplak & F.-Y. Dupradeau. ; . R.E.D. Server: a web service designed to automatically derive RESP and ESP charges and to generate force field

libraries for new molecules and molecular fragments. Université de Picardie Jules Verne - Burnham Institute for Medical Research 2009.
21. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977,** 23, (3), 327-341.
22. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993,** 98, 10089.
23. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *Journal of molecular graphics* **1996,** 14, (1), 33-38.
24. Kirschner, K. N.; Woods, R. J., Solvent interactions determine carbohydrate conformation. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** 98, (19), 10541.

Figure 1 – Structures of the native (N) and synthetic Nod Factors (1-4) investigated in this study including labelling of the heavy atoms and torsion angle definitions on the acyl chain.

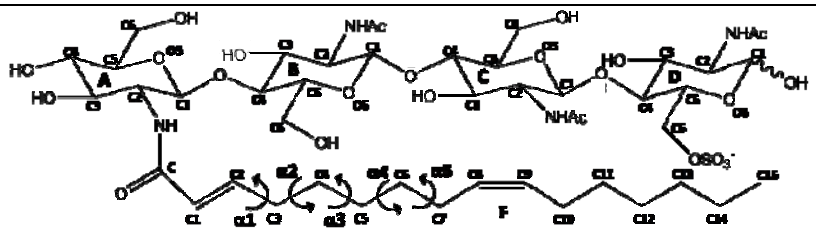
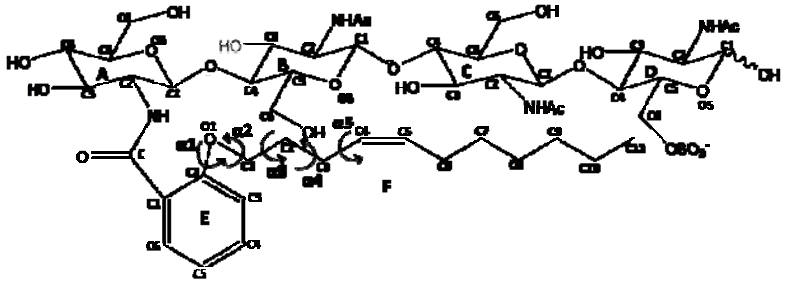
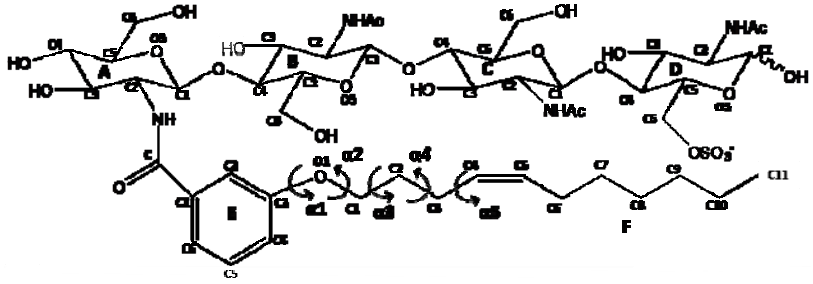
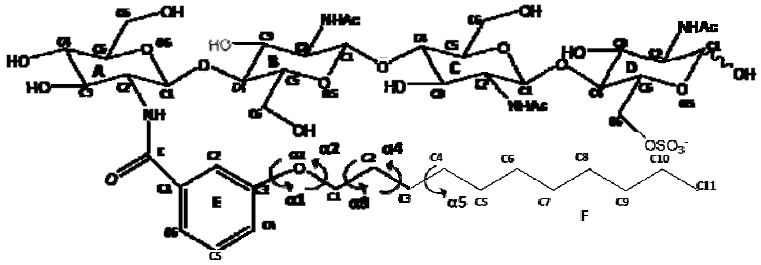
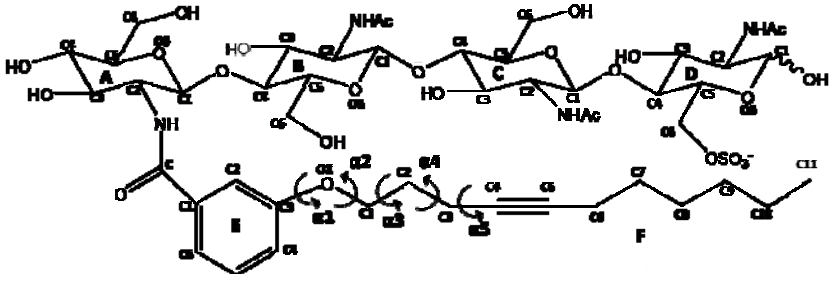
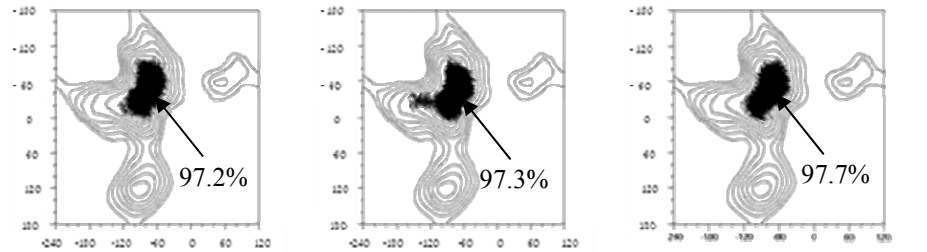
N	
1	
2	
3	
4	

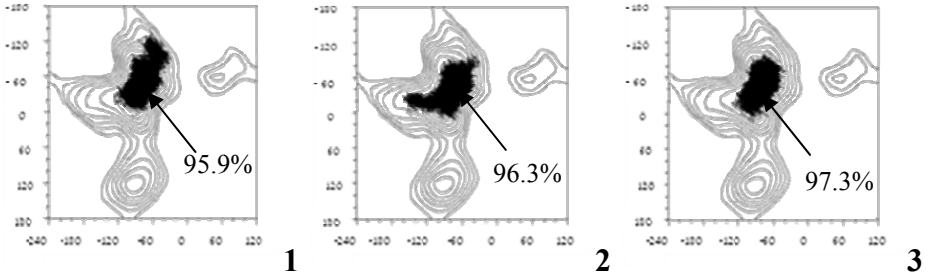
Figure 2 – Trajectories of the glycosidic torsion angles φ (x axis) and ψ (y axis) for the Nod Factors' tetrasaccharide scaffold. Results for the compounds N, 1, 4(syn/anti) during 10 ns dynamics simulations were superimposed on the MM3 adiabatic maps of β-D-GlcpNAc(1,4)β-D-GlcpNAc. The glycosidic torsions angles are labelled according to Figure 1.

N



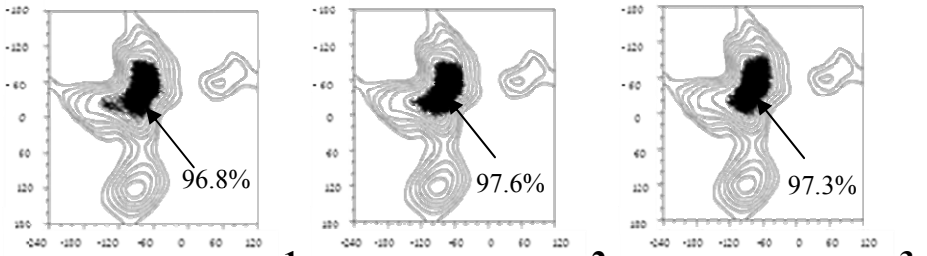
1 **2** **3**

1



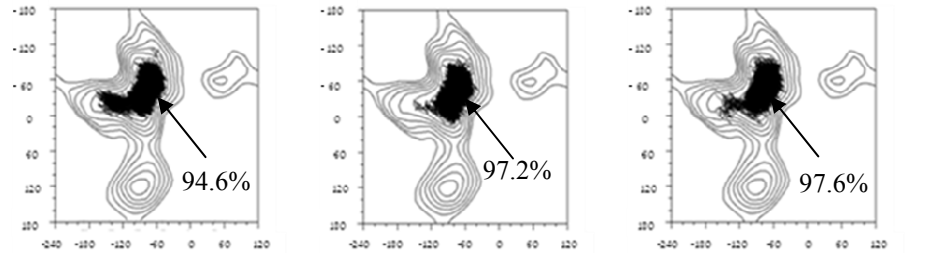
1 **2** **3**

4syn



1 **2** **3**

4anti



1 **2** **3**

Figure 3 – Histograms of the dihedral values distribution that characterize the Nod factors, lipid moieties of the molecules **N**, **1**, **4** (syn/anti).

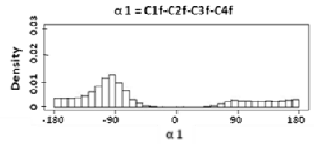
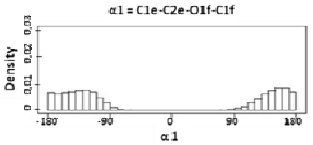
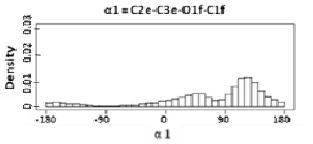
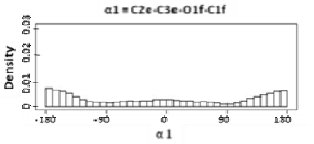
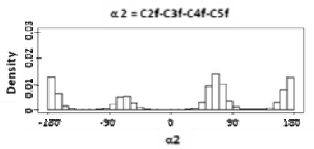
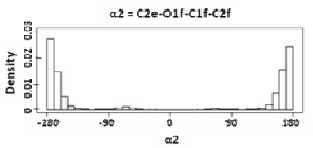
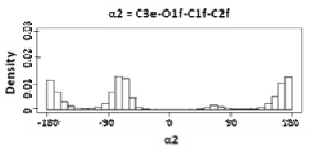
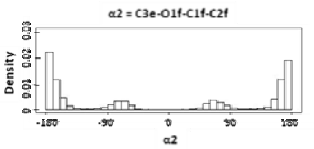
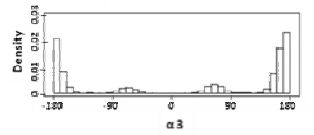
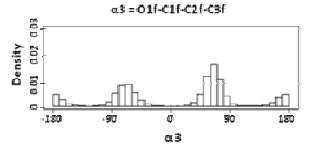
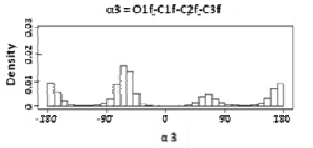
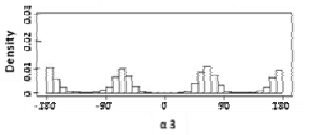
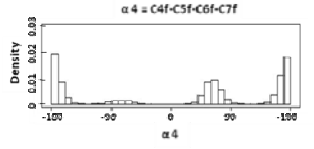
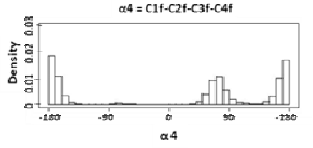
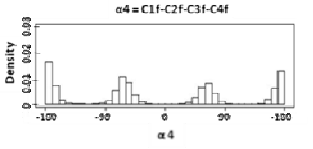
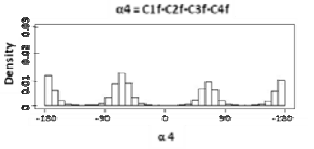
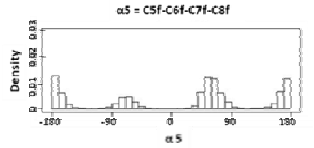
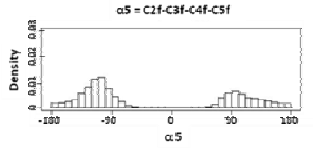
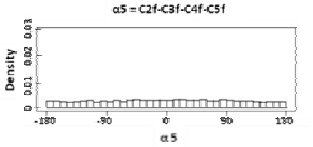
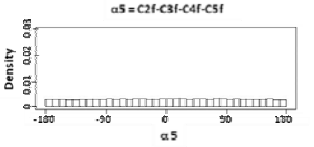
	N	**1**	**4syn**	**4anti**
α1				
α2				
α3				
α4				
α5				

Figure 4 – Superimposed representative snapshots from the five most populated conformational families of each compound (Fam1= red, Fam2= green, Fam3= blue, Fam4= yellow, Fam5=purple)

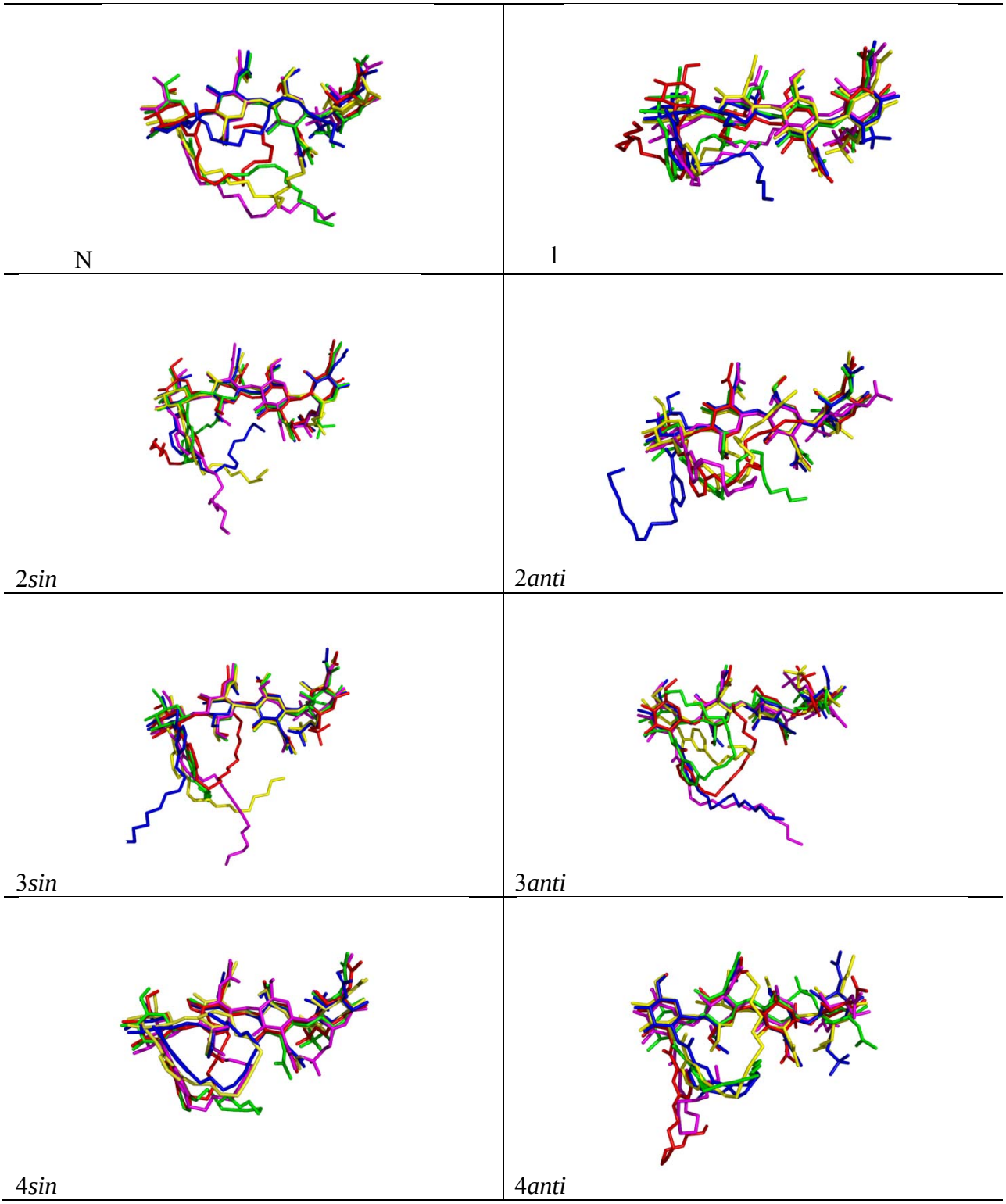
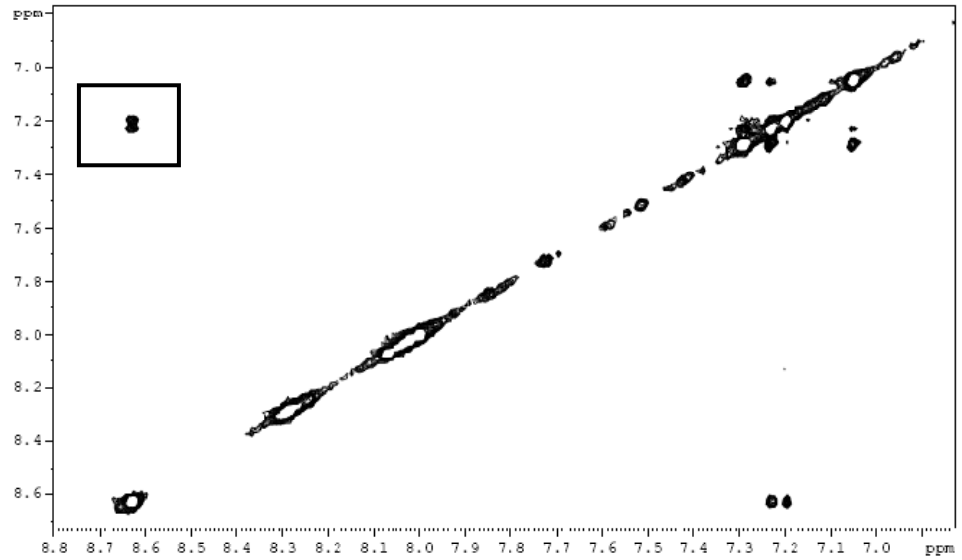
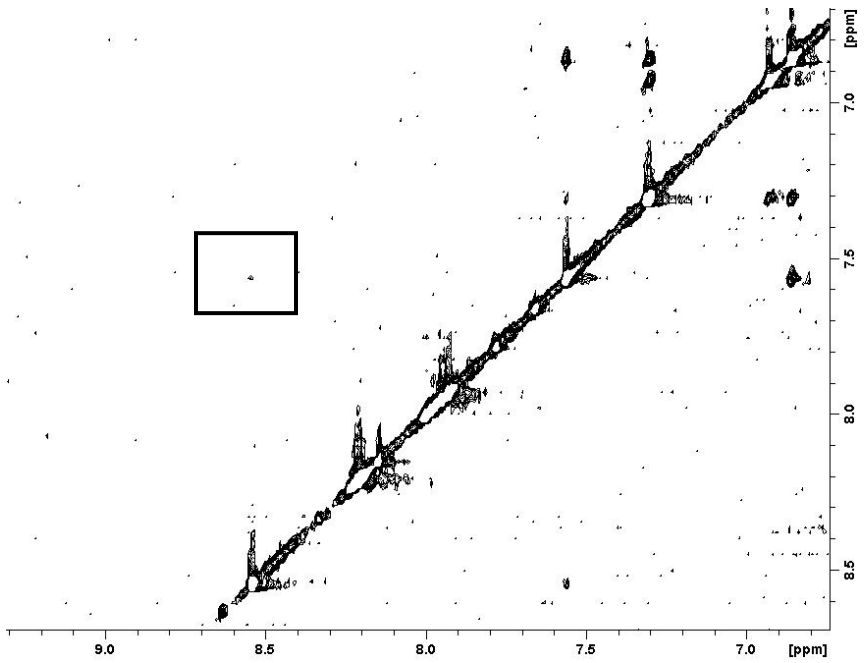


Figure 5- Two mutually exclusive NOE contacts between the NH and the aromatic protons show two distinct stable conformations for the compound 2 (same NOE contacts were found for compounds 3, 4). In figure 3B, one exclusive NOE contact between the NH and one aromatic proton show one stable conformation for the molecule 1.



A



B

Figure 6 – 2 D NOESY, D_2O, 288K, 800 MHz, (1-4 ppm). Lipid-tetrasaccharide contacts at a concentration of 1~mM for the compound 1

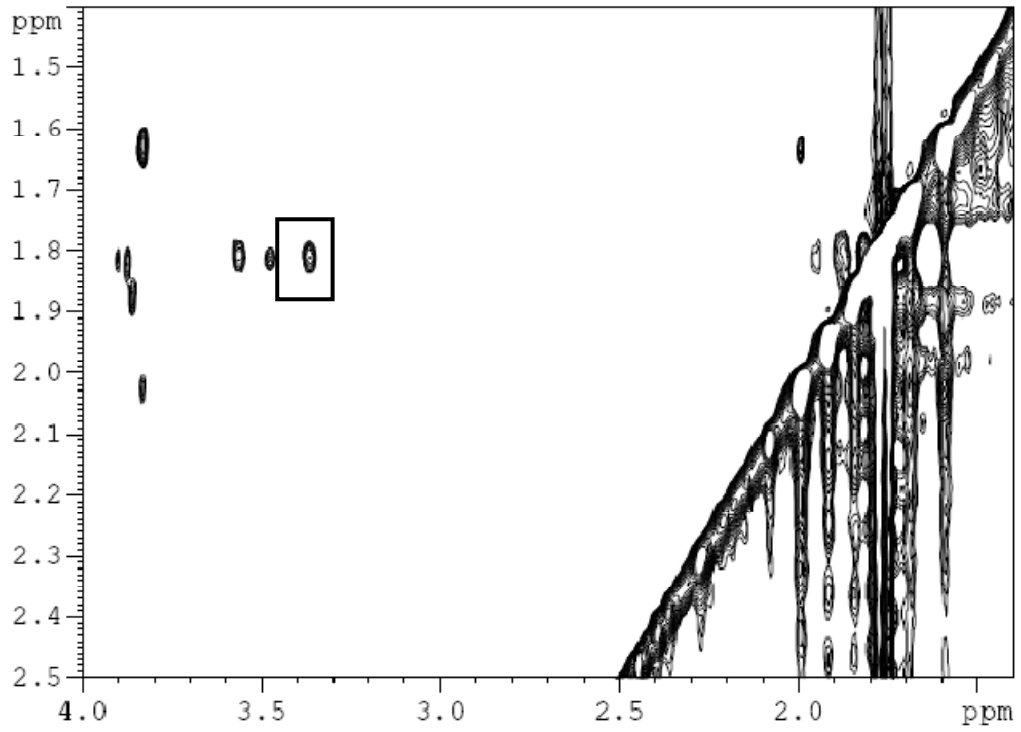
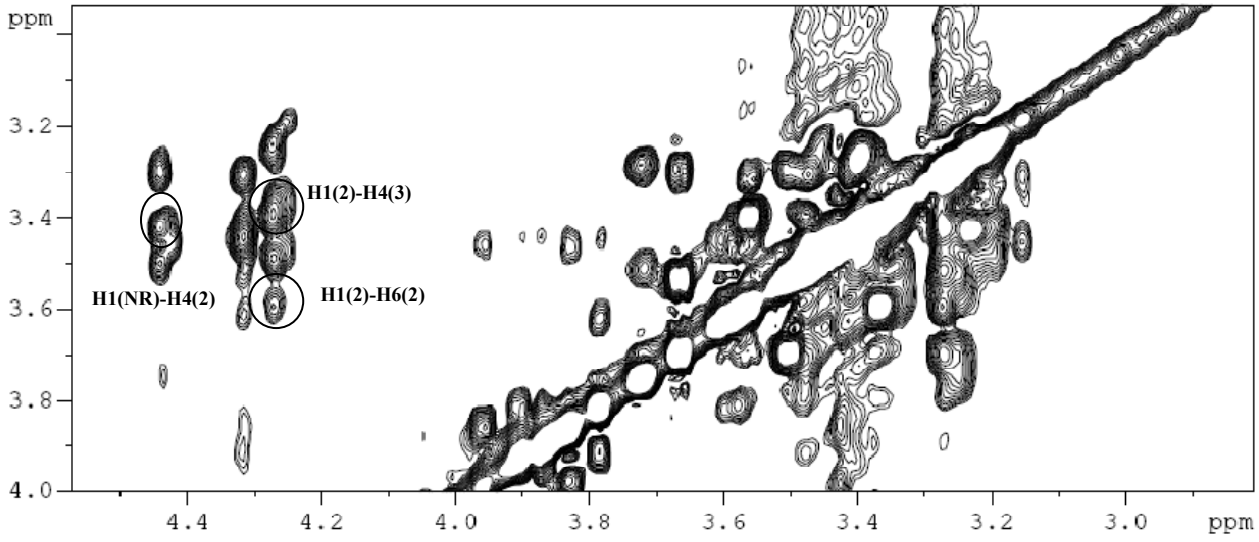


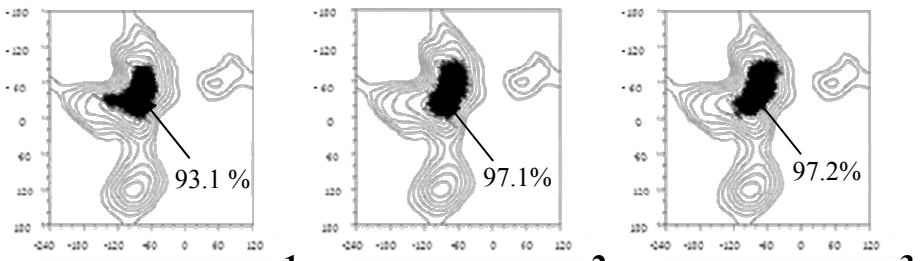
Figure 7 – 2D NOESY, D_2O, 288K, 800 MHz, (2.9-4.5 ppm). NOE contacts between the anomeric proton H1 of each residue and the H4, H6 protons of the following residues of the NF tetrasaccharide scaffold for the compound 1.



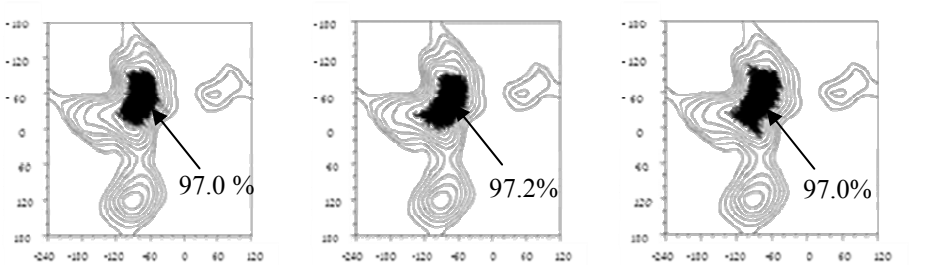
Figures (Supplemental material)

Figure 1S – Trajectories of the glycosidic torsion angles φ (x axis) and ψ (y axis) for the Nod Factors' tetrasaccharide scaffold. Results for compounds 3 and 2 (syn/anti) during 10 ns dynamics simulations were superimposed on the MM3 adiabatic maps of β-D-GlcpNAc(1,4)β-D-GlcpNAc. The glycosidic torsions angles are labelled according to Figure 1.

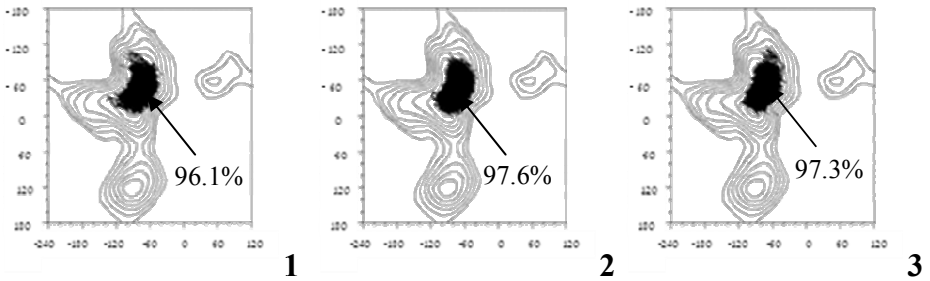
2syn



2anti



3syn



3anti

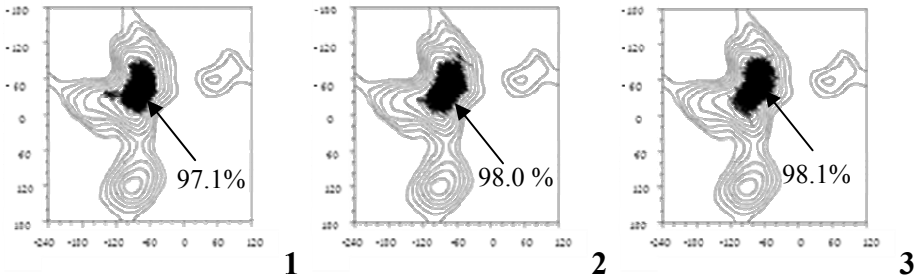
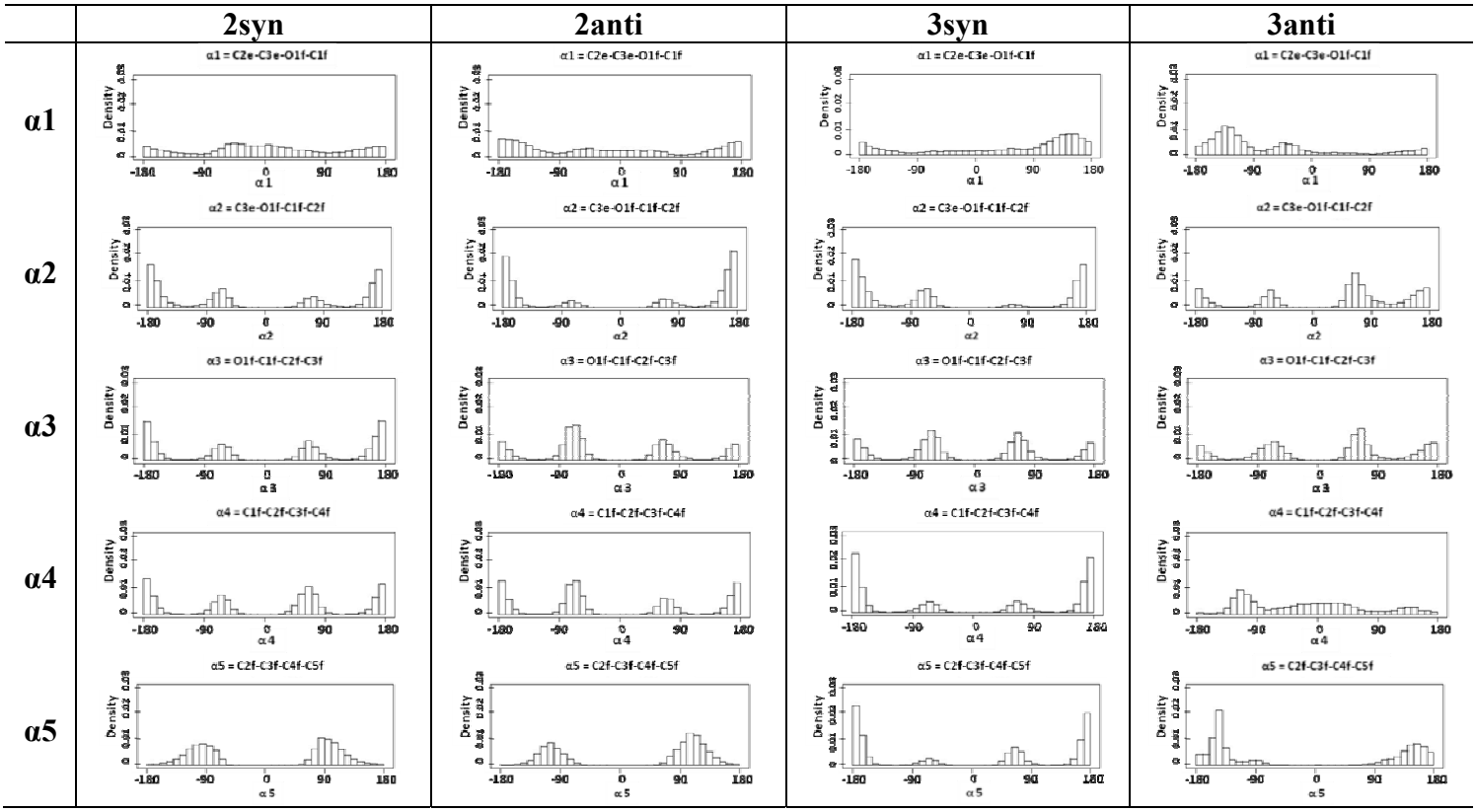


Figure 2S – Histograms of the dihedral values distribution that characterize the Nod factors, lipid moieties of the molecules **2** (syn/anti), **3** (syn/anti).



Tables

Table 1 – Average of glycosidic torsion angle values calculated from 50000 snapshots of each trajectory. Standard deviations are reported in brackets.

Compound	φ1	ψ1	φ2	ψ2	φ3	ψ3
N	-79.1 (10.0)	-128.1 (15.6)	-78.6 (10.8)	-130.5 (15.6)	-79.1 (9.6)	-129.1 (12.9)
1	-79.4 (10.9)	-133.1 (16.2)	-78.6 (12.3)	-131.4 (14.9)	-79.1 (10.0)	-129.7 (13.8)
2syn	-84.0 (11.2)	-138.8 (14.5)	-79.7 (10.1)	-132.4 (16.0)	-78.5 (9.8)	-128.5 (13.6)
2anti	-79.5 (11.2)	-129.3 (15.4)	-78.3 (10.4)	-130.8 (15.4)	-79.3 (9.6)	-129.6 (14.0)
3syn	-80.7(10.3)	-132.2 (14.6)	-79.5 (10.0)	-133.6 (14.4)	-78.6 (9.4)	-129.0 (13.0)
3anti	-81.3 (9.7)	-135.5(13.4)	-78.4 (9.9)	132.0 (13.5)	-77.7 (9.5)	-128.6 (13.2)
4syn	-80.4 (10.2)	-131.3 (16.1)	-79.0 (10.1)	-132.1 (15.4)	-79.2 (9.7)	-129.3 (13.6)
4anti	-81.6 (12.3)	-134.2 (16.0)	-79.1 (10.0)	-131.9 (15.1)	-78.8 (10.4)	-129.6 (13.4)

Table 2 - Inter hydrogen bonds data with an occupancy > 20%, distance <3.5 Å and angle cutoff of 120° collected along the MDs simulations. The occupancy is defined as the percent over the whole trajectory in which both the distance and the angle criteria are satisfied.

Inter hydrogen bonds			Occupancy (%)							
Acceptor	**Donor**		**N**	**1**	**2sin**	**2anti**	**3sin**	**3anti**	**4sin**	**4anti**
O5.A	HO3.B	O3.B	80.7	86.6	91.4	82.5	85.7	91.5	83.1	86.7
O5.B	HO3.C	O3.C	84.4	86.6	83.5	83.2	87.6	88.8	89.0	85.6
O5.C	HO2.D	O2.D	79.7	91.4	78.9	79.4	80.9	82.7	78.3	81.7
O1.E	HN.A	N2.A		98.9						

Table 3 - Geometrical features of the five most visited conformations (families) of the native (**N**) and synthetic Nod factors **1** and **4** along 10 ns of MDs simulation. The families were classified according to the lipid flexibility. Letter code for the dihedral angles is written according to the Nomenclature section. The population of conformers is reported in per cent (P%).

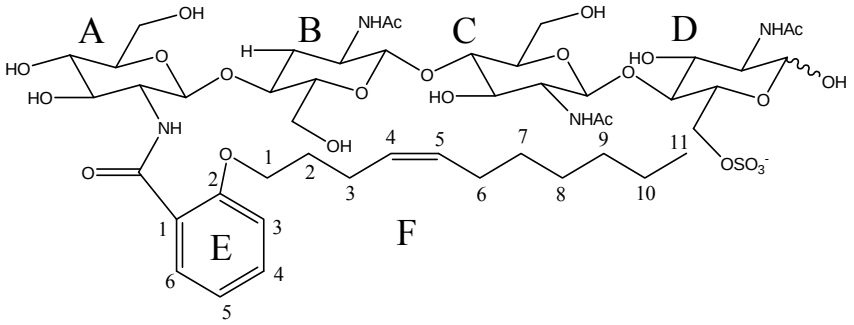
	α1	**α2**	**α3**	**α4**	**α5**	**P%**
N						
Fam1	-98	68	178	63	63	12.8
Fam2	-98	68	178	178	-178	8.4
Fam3	-98	68	178	178	63	4.6
Fam4	-98	178	178	178	-178	4.2
Fam5	-98	178	178	178	-73	3.4
1						
Fam1	158	178	-65	73	-108	13.3
Fam2	158	178	65	178	-108	13.2
Fam3	-128	178	65	178	93	11.6
Fam4	-128	178	-65	178	-108	8.9
Fam5	-128	178	65	178	158	6.4
4syn						
Fam1	125	-103	-63	-175		27.7
Fam2	45	173	-63	-65		11.9
Fam3	45	173	178	65		9.3
Fam4	125	173	178	-65		6.6
Fam5	125	-63	178	-175		5.4
4anti						
Fam1	-175	-178	-178	-68		6.0
Fam2	-175	-178	-68	-68		5.7
Fam3	5	-178	-178	-68		4.9
Fam4	-175	-178	-178	63		4.5
Fam5	-175	65.0	63	-178		4.2

Table 4 – Relevant average proton-proton inter/intra residue distances from the MDs simulations Distances derived from the simulation of the compound 1, here reported, are representative for all the simulations. Standard deviations are reported in brackets.

Intra proton-proton distances		**Inter proton-proton distances**	
Distances (Å)		**Distances (Å)**	
H1.A-H3.A	2.7 (0.2)	H1.A-H4.B	2.3 (0.2)
H1.A-H5.A	2.6 (0.2)	H1.A-H6.Ba	3.5 (0.5)
H1.B-H3.B	2.7 (0.2)	H1.A-H6.Bb	3.6 (0.9)
H1.B-H5.B	2.6 (0.2)	H1.B-H4.C	2.3 (0.2)
H1.C-H3.C	2.7 (0.2)	H1.B-H6.Ca	3.4 (0.6)
H1.C-H5.C	2.6 (0.2)	H1.B-H6.Cb	3.8 (0.9)
H1.D-H3.D	2.7 (0.2)	H1.C-H4.D	2.3 (0.2)
H1.D-H5.D	2.6 (0.2)	H1.C-H6.Da	4.4 (0.7)
H1.D-H3.D	2.7 (0.2)	H1.C-H6.Db	3.5 (0.5)
H1.D-H5.D	2.6 (0.2)	H3.A-H2.F*	4.0 (1.1)
		H3.A-H3.F*	4.0 (1.1)

*Specific for the compound **1**

Table 5 – H NMR data (**δ** ppm) for compound 1.

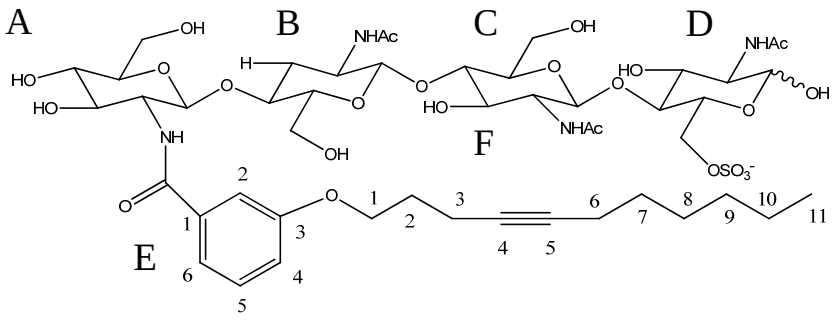


	δH1	**δH2**	**δH3**	**δH4**	**δH5**	**δH6a-b**	**δNH***	**CH3**
D	4,92	3,63	3,80	3,43	3,92	3,89	7,95	1,75
NOESY	3,63 H2							
C	4,33	3,50	3,41	3,33	3.28	3,57-3,38	7,93	1,78
NOESY	3,41 H3 3.28 H5 3.89 H6A							3,89 H6A 7,93 NH
B	4,28	3,48	3,48	3,38	3.21	3,68-3,52	8,21	1,78
NOESY	3,47 H3 3.33 H4B 3,58 H6B							3,48 H2 8,21 NH
A	4,45	3,72	3,45	3,26	3,45	3,4-3,25	8,50	
NOESY	3.72 H2 3,38 H4C 3,68 H6C 3,45 H3/H5						4.44 H1 3,72 H2 3,45 H3 3,24 H4 7,5H5E	
E	/	6,94	7,31	6,86	7,56	/		
NOESY					8,5 NHD			

	H1	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**
F	3.96/ 3.93	1.73/ 1.63	2.09/ 1.99	5,22	5.30	1.74	1,002	0,92	0,88	0,84	0,55
NOESY		3,45 H3D	3,45 H3D								

*chemical shift measured at 900MHz 288K

Table 6 – H NMR data (**δ** ppm) for compound **4**.



	δH1	**δH2**	**δH3**	**δH4**	**δH5**	**δH6**	**δNH***	**CH3**			
D	4.92	3.62	3.80	3.43	3.91	3.89	7.81	1.81			
NOESY	3.6 H2						3.63 (H2) 1.81 (Me)				
C	4.33	3.50	3.40	3.35	3.29	3.57	7.93	1.75			
NOESY	3.4 H3 3.29 H5						3.46 (H3) 1.83 (Me)				
	3.89 H6A										
B	4.28	3.47	3.50	3.41	3.20		8.26	1.79			
NOESY	3.5 H3 3.35 H4B 3.57 H6B						3.47 (H3) 1.79 (Me)				
A	4.47	3.74	3.50	3.28	3.48	3.68	8.63				
NOESY	3.74 H2 3.41 H4C 3.68 H6C 3.48 H3/H5						3.5 (H3) 3.74 (H2) 7.19 H2E 7.26 H6E				
E	/	7.19	/	7.04	7.28	7.26					
NOESY		8.63NHD 4.04 H1F 2.23 H2F		4.04H1F 3.88 H1F 2.23 H2F		8.63NHD					
	H1	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**
F	4.04/3.88	2.23	1.77	/		1.89	1.14	1.03	0.92	0.92	0.55

*chemical shift measured at 900MHz 288K

Table 7 – H NMR data (**δ** ppm) for compound **3**.

	δH1	**δH2**	**δH3**	**δH4**	**δH5**	**δH6**	**δNH***	**CH3**			
D	4.95	3.63	3.84	3.43	/	/	8.01	1.81			
NOESY							3.63 (H2) 1.81 (Me)				
C	4.38	3.50	3.40	/	/	/	8.26	1.83			
NOESY							3.40 (H3) 1.83 (Me)				
B	4.37	3.53	3.48	3.41	/	/	8.28	1.83			
NOESY							3.47 (H3) 1.83 (Me)				
A	/	3.76	/	/	/	/	8.64				
NOESY							7.20 H2E 7.16 H6E				
E	/	7,16	/	7,02	7,27	7,2					
NOESY		8.63NHD 4.04 H1F 2.23 H2F		4.04H1F 3.88 H1F 2.23 H2F		8.63NHD					
	H1	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**
F	3,93/3,96	1,58/1,2	1,14	1,05	1,05	1,05	1,05	1,05	1,0	0,6	0,6

Table 8 – H NMR data (**δ** ppm) for compound **2**.

	δH1	**δH2**	**δH3**	**δH4**	**δH5**	**δH6**	**δNH**	**CH3**			
D	4,9	3,60	3,81	3,44		3,88					
NOESY	3.6 H2										
C	4.27	3.48	3,27	3.37	3.27	3.59	/	1.80			
NOESY	3.27 H3						3.47 (H3) 1.79 (Me)				
B	4.23	3.47	3.47	3.41	3.24	/					
NOESY	3.37 H4C 3.59 H6C										
A	4.43	3.73	3.51	3.29	3.51	4.43					
NOESY	3.73 H2 3.41 H4C 3.69 H6C 3.51 H3/H5										
E	/	7.08	/	7.13	6.95	7.18					
	H1	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**
F	4,04/4,02	2,02	1,60	5,34	5,43	1,72	0,90	0,86	0,81	0,77	0,54

Tables (Supplemental material)

Table 1S – gt/gg/tg ratio of the ω angle values for the Nod factors' tetrasaccharide scaffold along 10 ns dynamics simulations.

N			
A	ω1 ≈ 60 64.6%	ω1 ≈ 180 1.9%	ω1 ≈ -60 33.5%
B	ω2 ≈ 60 24.8%	ω2 ≈ 180 0.7%	ω2 ≈ -60 74.5%
C	ω3 ≈ 60 30.1%	ω3 ≈ 180 1.6%	ω3 ≈ -60 68.4 %
D	ω4 ≈ 60 --------	ω4 ≈ 180 1.5%	ω4 ≈ -60 98.5%
1			
A	ω1 ≈ 60 72.8%	ω1 ≈ 180 2.6%	ω1 ≈ -60 24.6%
B	ω2 ≈ 60 64.8%	ω2 ≈ 180 2.6%	ω2 ≈ -60 32.6%
C	ω3 ≈ 60 48.2%	ω3 ≈ 180 1.0%	ω3 ≈ -60 50.8 %
D	ω4 ≈ 60 6.6%	ω4 ≈ 180 5.2%	ω4 ≈ -60 88.2%
2syn			
A	ω1 ≈ 60 85.8%	ω1 ≈ 180 2.4%	ω1 ≈ -60 11.8%
B	ω2 ≈ 60 46.2%	ω2 ≈ 180 1.2 %	ω2 ≈ -60 52.6%
C	ω3 ≈ 60 40.3%	ω3 ≈ 180 1.3 %	ω3 ≈ -60 58.4 %
D	ω4 ≈ 60 3.8%	ω4 ≈ 180 3.9%	ω4 ≈ -60 92.3%
2anti			
A	ω1 ≈ 60 67.4%	ω1 ≈ 180 1.5%	ω1 ≈ -60 31.1%
B	ω2 ≈ 60 11.9%	ω2 ≈ 180 0.9 %	ω2 ≈ -60 87.2 %
C	ω3 ≈ 60 55.9%	ω3 ≈ 180 3.0%	ω3 ≈ -60 41.1 %
D	ω4 ≈ 60 0.3%	ω4 ≈ 180 2.0%	ω4 ≈ -60 97.7%
3syn			
A	ω1 ≈ 60 23.7 %	ω1 ≈ 180 1.0%	ω1 ≈ -60 75.2%
B	ω2 ≈ 60 75.3%	ω2 ≈ 180 3.0%	ω2 ≈ -60 21.7 %
C	ω3 ≈ 60 42.3%	ω3 ≈ 180 1.4 %	ω3 ≈ -60 56.3%
D	ω4 ≈ 60 0.1%	ω4 ≈ 180 5.6%	ω4 ≈ -60 94.3%
3anti			
A	ω1 ≈ 60 86.8%	ω1 ≈ 180 1.0%	ω1 ≈ -60 12.2%
B	ω2 ≈ 60 39.2%	ω2 ≈ 180 1.4%	ω2 ≈ -60 59.4%
C	ω3 ≈ 60 51.9%	ω3 ≈ 180 2.5%	ω3 ≈ -60 45.7%
D	ω4 ≈ 60 18.7%	ω4 ≈ 180 7.0%	ω4 ≈ -60 74.3%
4syn			
A	ω1 ≈ 60 64.6%	ω1 ≈ 180 1.9%	ω1 ≈ -60 33.5%
B	ω2 ≈ 60 24.8%	ω2 ≈ 180 0.7%	ω2 ≈ -60 74.5%
C	ω3 ≈ 60 30.1%	ω3 ≈ 180 1.6%	ω3 ≈ -60 68.4 %
D	ω4 ≈ 60 --------	ω4 ≈ 180 1.5%	ω4 ≈ -60 98.5%
4anti			
A	ω1 ≈ 60 75.5%	ω1 ≈ 180 2.4%	ω1 ≈ -60 22.1%
B	ω2 ≈ 60 52.7%	ω2 ≈ 180 1.1 %	ω2 ≈ -60 46.2%
C	ω3 ≈ 60 12.3%	ω3 ≈ 180 0.5%	ω3 ≈ -60 87.2 %
D	ω4 ≈ 60 23.8%	ω4 ≈ 180 4.8%	ω4 ≈ -60 71.4%

Table 2S - Geometrical features of the five most visited conformations (families) of the synthetic nod factors **3** and **4** along 10 ns of MDs simulation. The families were classified according to the lipid flexibility. Letter code for the dihedral angles is written according to the Nomenclature section. The population of conformers is reported in per cent (P%).

	α1	**α2**	**α3**	**α4**	**α5**	**P%**
2syn						
Fam1	-60	-68	178	-63	-98	7.7
Fam2	5	68	178	63	88	6.1
Fam3	175	178	63	-178	88	5
Fam4	5	178	178	-178	-98	4.6
Fam5	175	178	178	63	-98	3.3
2anti						
Fam1	-165	178	-68	-68	108	14.9
Fam2	-45	-70	68	178	108	5
Fam3	-165	178	68	68	-103	4.1
Fam4	-165	178	-68	68	-103	4.1
Fam5	135	178	-68	-68	108	4
3syn						
Fam1	145	-68	-65	178	-178	11.9
Fam2	145	178	65	178	-178	10.1
Fam3	145	178	-65	178	-178	4.9
Fam4	-155	178	65	178	-178	4.4
Fam5	55	178	-65	178	-178	3.3
3anti						
Fam1	-135	63	63	-115	-145	17.5
Fam2	-45	-68	178	-115	155	6.4
Fam3	-135	178	-68	5	-145	5.2
Fam4	-45	-65	178	-115	-145	5.1
Fam5	-135	63	178	135	155	4.8

Table 3S – 10 ns MDs simulations: details for each simulated system.

	Solute Atom N.	**Water molecules N.**	**Na+**
N	152	2294	1
1	153	1836	1
2syn	153	1929	1
2anti	153	2705	1
3syn	155	1826	1
3anti	155	3291	1
4syn	151	2957	1
4anti	151	2005	1

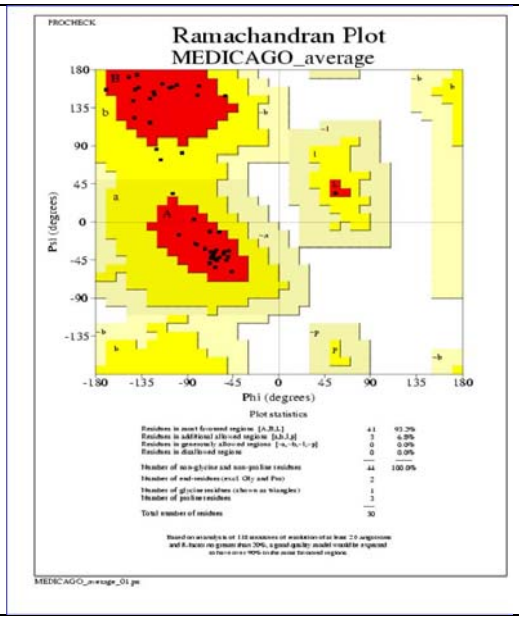
Table 4S – Radius of gyration (Rad.gyr) and the distance (Dist.) between the sulfate group at the carbohydrate reducing end and the extremity of the acyl chains of the native (**N**) and synthetic nod factors **1** - **4** along 10 ns of MDs simulation. Standard deviations are reported in brackets.

	N	**1**	**2**	**3**	**4**
Rad. Gyr. (Å)	7.6 (0.8)	7.3 (0.4)	7.8 (0.8)	7.5(0.8)	7.7 (0.8)
Dist. (Å)	14.6 (4.8)	12.0 (5.1)	14.6 (5.6)	12.9 (5.4)	15.3 (5.2)

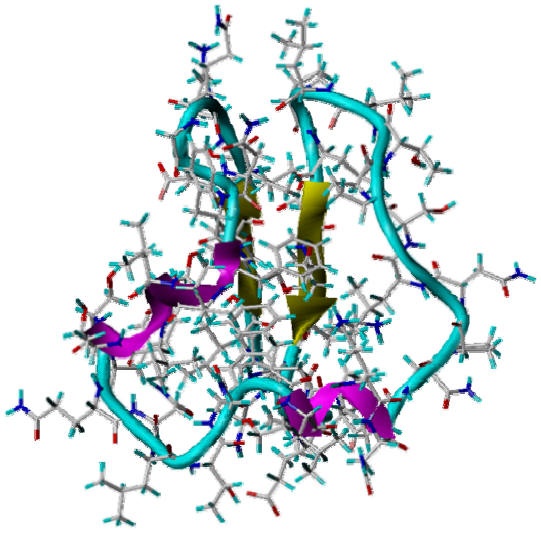
The following pages describe the modeling part of a manuscript that will include experimental studies on kinase domains of Nod receptors (manuscript in preparation, in collaboration with Dr. Julie Cullimore).

Homology model of the LYSM2 domain of NFP from Medicago truncatula.

The alignment of the LysM2 domain sequence from the plant *Medicago truncatula* obtained from NCBI protein database with the template the NMR structure of a LysM domain from *E. coli* MltD was obtained using FUGUE, the alignment tool implemented in Sybyl program (Tripos Associates, St. Louis, MO). The alignment was then manually refined to assure the correct superposition for the residues that characterize the secondary structure (For the alignment, refers to the paper *L.Mulder et al., Glycobiology, 2006*). Structurally conserved regions were built from the template whereas loops were modeled using libraries of fragments implemented in Sybyl whose geometries were optimized using the Amber force field. The model was then refined via energy minimization, and 20ps of production molecular dynamics simulation in explicit water. The structural average was then calculated, water molecules extracted and the whole structure subjected to 1000 steps of conjugate gradient minimization. The structure was then evaluated using the PROCHECK program. The model shows the typical βααβ secondary structure found in all the solved LysM domain structures (Fig.1). Further information is needed for the identification of the binding site.



A



B

Figure 1 – Ramachandran plot of the LysM2 model (A); 3D structure of the LysM2 (B)

Homology model of LYK3 kinase from Medicago truncatula

The aminoacid sequence of the kinase *LYK3*, obtained from NCBI protein database was used for an exhaustive PSI_BLAST search in the Protein Data Bank[1]. Among all kinases for which the tridimensional structure is known, the human serine/theonine kinase *IRAK4* shows a close homology relation to the kinase LYK3.

The homology model of LYK3 was built based on the protein crystal structure of IRAK4 (pdb code 2NRU) [2] using the Orchestrar modeling suite of SYBYL 7.3 (Tripos Associates, St. Louis, MO). The target-template sequence alignment (Fig. 1), constructed using CLUSTALW and imported in Sybyl, showed 37% of sequence identity[3] . The structural conserved regions were superposed on the coordinates of the most closely related residues of the template structure and modeled, leaving gaps for the most variable regions. These regions were modeled using libraries containing a set of protein fragments, fitting the new backbone coordinates on those of the initial model. Model building continued with the addition of amino acid side chains, whose torsion angles were defined by a rotamer library and optimized through few minimization steps.

The model was then subjected to a global optimization of the geometry to relieve molecular strain from side chain atom clashes, to correct bond length and angles and generally to move the structure towards a minimum. The model was then prepared for minimization, adding missing hydrogen atoms and charges according to the AMBER(ff99). The model was minimized in stages, using 4 as dielectric constant: firstly, only hydrogen atoms were minimized. This step was followed by a minimization of hydrogens and sidechains, a minimization of hydrogens, sidechains and backbone atoms except C-α, and a full minimization with no restraints. Phosphate groups where added on the specific residues, submitted a minimization with restraints on the protein. The stereo chemical properties of the model were then checked using PROCHECK: all residue positions, evaluating the ψ and φ angles between N-Cα and Cα-C, fall in acceptable ranges like reported for the template structure.

MDs simulations of LYK3 kinase from Medicago truncatula

MDs simulation in explicit water was used to check the stability of the model along 8ns of simulation and to evaluate the effect of phosphorylation on the structural dynamics of the activation loop of the enzyme.

The three-dimensional structures of the kinase LYK3 with and without the phosphorylated hydroxyl group of Thr170, Thr 167, Ser166 were created starting from the homology model previously built. The phosphorylated and unphosphorylated models were soaked in a cubic box of water molecules and subjected to energy minimization and 8ns long-duration molecular dynamics simulation using AMBER8 program (University of California). The coordinate and parameter files

for input were generated using the AMBER forcefield parm99 with parameters for phosphorilated residues taken from Craft et al[4].

The simulations were carried out using the particle mesh Ewald technique with 10Å non bonded cutoff and 2 fs integration time step[5]. The SHAKE algorithm was used to restrain all hydrogen-heavy atom bond lengths[6]. Water molecules were described by a TIP3 potential, and periodic boundary conditions were applied to the systems. After 7000 minimization steps including firstly the solvent molecules and then the whole systems, the models were heated from 10 to 300K. During this procedure, only water molecules were free to move. Harmonic constraints (force constant of 10.0 Kcal/mole/Å^2) were applied to the minimized models and simulated for 100 ps to prevent instabilities resulting from energy conservation due to initial bad contacts. An NPT ensemble simulation was used for the rest of the dynamics run: the temperature was maintained at 300K and the pressure at 1 atm using the Langevin piston. After another 100 ps of simulation, the force constant was decreased and the process continued for another 500 ps. The entire systems then moved freely for 7.3 ns.

Structure frames and properties of the simulated system were extracted using the PTRAJ module implemented in AMBER and illustrations were produced using CHIMERA software[7].

Results

The LYK3 Kinase Domain show Remarkable Structural Homology to the Human IRAK-4 kinase - Computational studies were carried out to provide new insights into the architecture of the LYK3 kinase. A homology model of the LYK3 kinase was built, using the crystal structure of the IRAK-4 kinase as template (pdb code 2NRU). This protein structure was determined in its active state, presenting phosphorylated residues in the activation loop. Based on the structure of the active IRAK-4, the model of LYK3 was built, including phosphate groups on the three previously identified residues Ser471, Thr472 and Thr475.

The model (Fig. 2A) shows the typical two-lobe structure of a kinase. The N-terminal lobe consists of the five stranded antiparallel ß-sheet and the prominent α helix, termed helix C (amino acids 355–369). As IRAK-4, LYK3 model contains a N-terminal extension with an alpha helix, helix B (amino acids 311-319) which packs against the ß-sheet. The characteristic glycine-loop (amino acids 329 – 334), which normally covers and anchors the non transferable phosphate of ATP, closely follows this region and it is located between the first two ß-strands. The N-terminal lobe is connected to the C-terminal one through a hinge sequence in which a tyrosine residue, known as “gatekeeper”, controls the ATP access. This residue is unique to the four related IRAK kinases in the human kinome; in IRAK-4, it interacts with the glutamate E233 from helix C, pulling the helix in to maintain the active orientation.

LYK3 contains the tyrosine gatekeeper (Tyr390), which establishes a hydrogen bond with the glutamate Glu362 (Fig.2B). These two residues are part of a stable hydrogen bond network since E362 also acts as hydrogen bond acceptor from the nitrogen of the phenylalanine Phe460 backbone in the conserved DFG motif. Thus, Glu362 plays a double role in the active conformation of the enzyme, promoting contacts with both the Phe390 gatekeeper residue and the Phe460 main chain. The C-terminal lobe is formed by several α helices, small flexible loops and the larger activation loop, which is often subject to regulation through phosphorylation (amino acids 467-480).

Activation Loop Phosphorylation Resembles that of IRAK-4 -Alignment of the activation loops of LYK3 with human IRAK-4 revealed a remarkable conservation between LYK3 and IRAK-4, particularly at the start and the end of the loop. In addition two of the three phosphorylated residues, identified in each protein, are conserved (Fig. 1). Computational studies were carried out to explain how phosphorylation can influence the structural dynamics of the activation loop of LYK3, by comparing the model of the unphosphorylated protein with a model in which the activation loop was triply phosphorylated at Ser471, Thr472 and Thr475. A Molecular Dynamics MDs approach was used to check the stability of the models in an explicit water environment and to evaluate the dynamic properties of the activation loop in the presence and absence of the three phosphorylated residues. The analysis of the energies for both models along the 7.3 ns simulations in explicit water reveals that the lowest energies were reached in the phosphorilated protein (Fig.3).

The overall geometries of the models remained stable along the simulations as deduced by the calculation of the root mean square deviations (RMSD) of the α-carbons with respect to their initial states (Fig.4). Superimposition of the unphosphorylated LYK3 structure extracted at the end of the dynamics simulation on the phosphorylated LYK3 further demonstrates that the overall domain structure is stable, with the two structures sharing the same 3D features and a RMSD of 1.1 Å(Fig.5A). Main differences were found in the activation loop (amino acids 467-480), in which the phosphorylated residues Ser471, Thr472 and Thr475 show a shift of 5.6 Å, 6.7 Å and 1.7 Å in the non phosphorylated model, respectively (Fig.5B).

The attention was then focused on the activation loop (amino acids 467-480). Detailed analysis of the activation loop in the phosphorylated model indicates that the phosphorylated Thr475 is part of a stable polar network (>60% occupancy along the simulation), involving salt bridges with Arg476, which precedes the catalytic residues of the activation loop, Arg440, located in the conserved HRD motif, and activation loop Lys464, the third residue downstream of the conserved DFG triplet, which is also positively charged in IRAK-4 (R334). During the simulation, the phosphorylated Thr472 loses its initial salt bridge with Lys497 that characterized the homology model (the starting structure of the MDs simulation) and points out of the protein like the residue pSer471: both residues expose the phosphate groups to the solvent environment (Fig. 2C). In the case of the unphosphorylated kinase,

the behaviour of the activation loop is less stable as indicated by the highest total energies of the system, resulting from the missing network of salt bridges involving Arg440 from the HRD motif and residues Arg476 and Lys464 (Fig.5B). Clearly, the absence of phosphorylated residues also influences the electrostatic properties of the activation loop, leading to a more electropositive region since the positively charged amino acids that characterize this part of the protein are not neutralized by the negative phosphate groups (Fig.5C).

References

1. Altschul, S. F.; Koonin, E. V., Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends in biochemical sciences* **1998,** 23, (11), 444-447.
2. Wang, Z.; Liu, J.; Sudom, A.; Ayres, M.; Li, S.; Wesche, H.; Powers, J. P.; Walker, N. P. C., Crystal structures of IRAK-4 kinase in complex with inhibitors: a serine/threonine kinase with tyrosine as a gatekeeper. *Structure* **2006,** 14, (12), 1835-1844.
3. Combet, C.; Blanchet, C.; Geourjon, C.; Deleage, G., NPS@: network protein sequence analysis. *Trends in biochemical sciences* **2000,** 25, (3), 147.
4. Craft, J. W.; Legge, G. B., An AMBER/DYANA/MOLMOL phosphorylated amino acid library set and incorporation into NMR structure calculations. *Journal of Biomolecular NMR* **2005,** 33, (1), 15-24.
5. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993,** 98, 10089-10092
6. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977,** 23, (3), 327-341
7. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **2004,** 25, (13), 1605-1612.

Figures

Figure 1 – CLUSTALW sequence alignment of the LYK3 aminoacid sequence and IRAK4 using default parameters of gap opening and extension penalties with BLOSUM62 matrix. The phosphorylation sites are indicated with the letter "P" whereas the gatekeeper is indicated with the letter "G". In the alignment, residues forming the activation loop are also underlined.

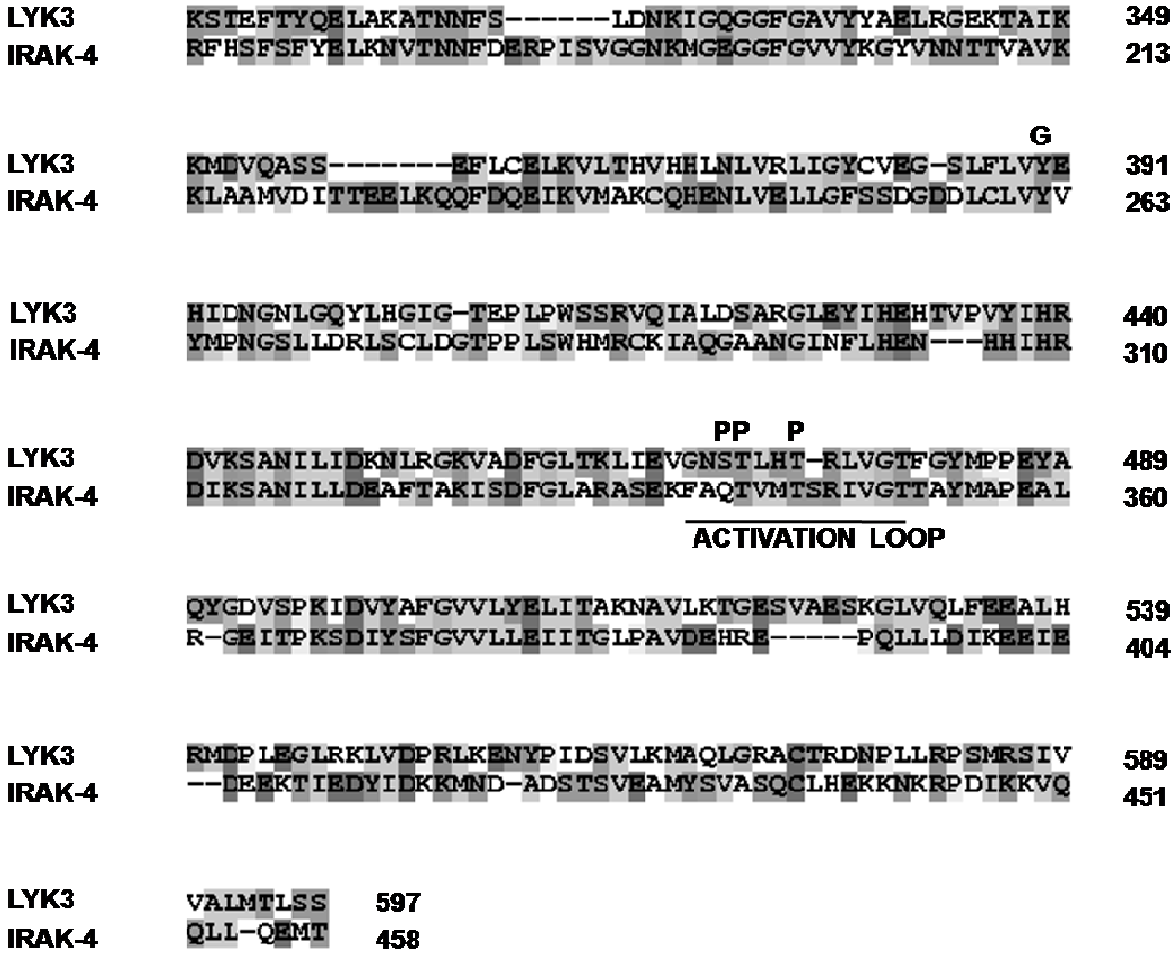
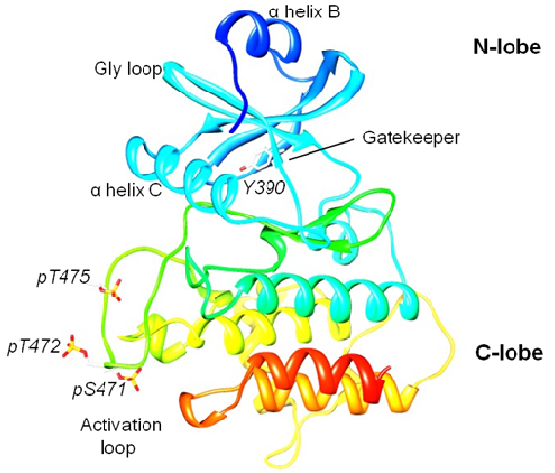
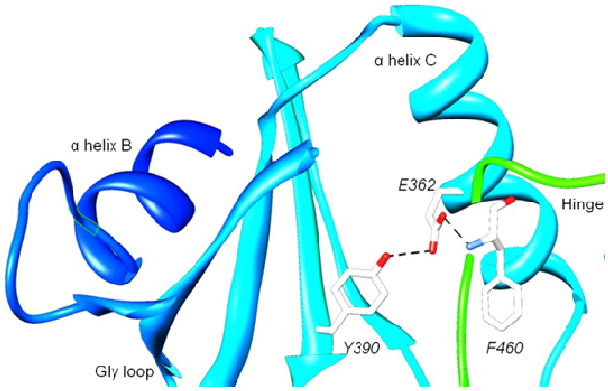


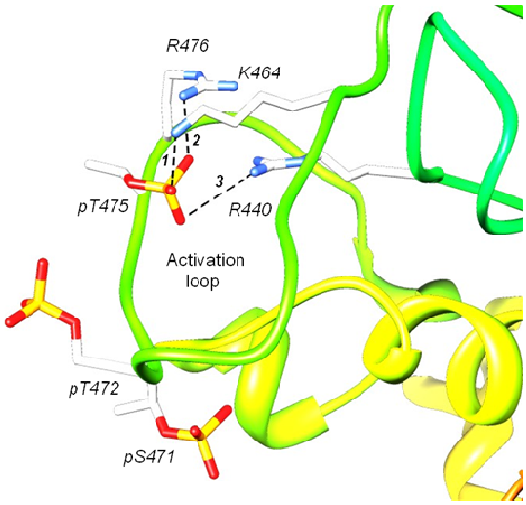
Figure 2 – Structure of the LYK3 kinase domain. A. Ribbon representation of the homology model of the LYK3 kinase. B. Details of the gatekeeper region of the LYK3 kinase C. Details of the activation loop region of the LYK3 kinase. The H bond 1,2 and 3 are stable along the simulations (>60% occupancy).



A



B



C

Figure 3 - Comparison between the total energies of the phosphorilated (line)/ unphosphorilated (dotted) systems during the 7.3ns MDs simulation production

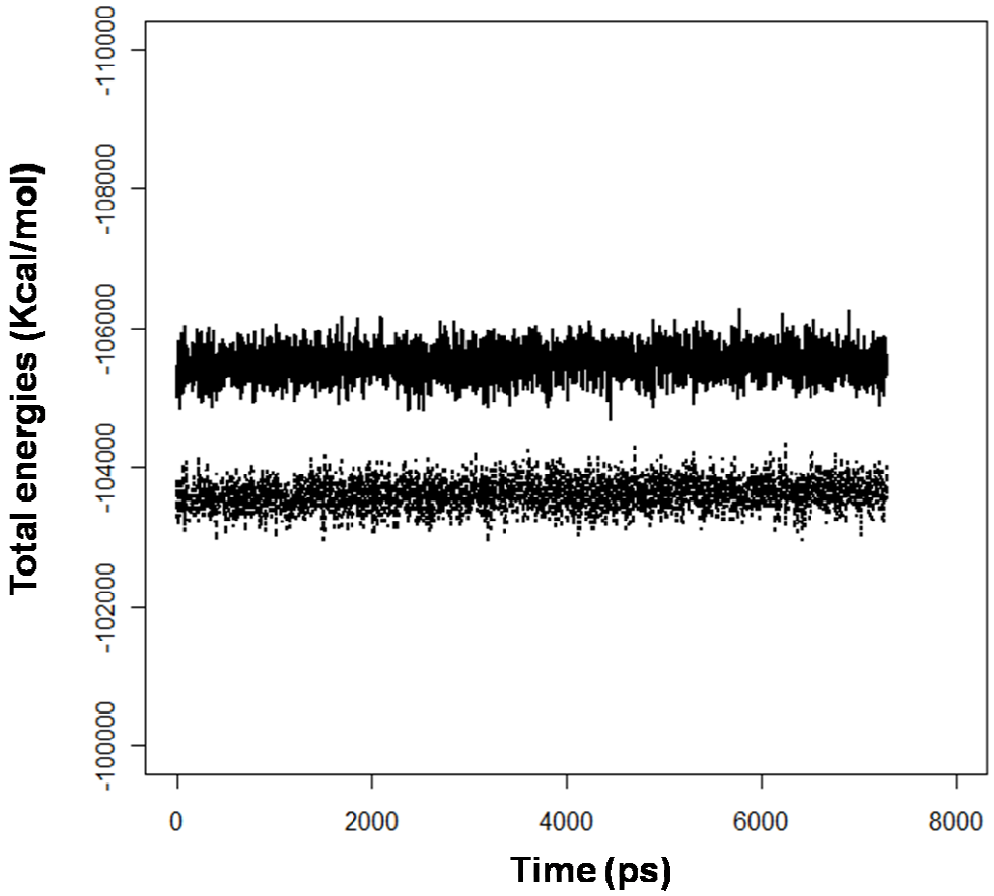
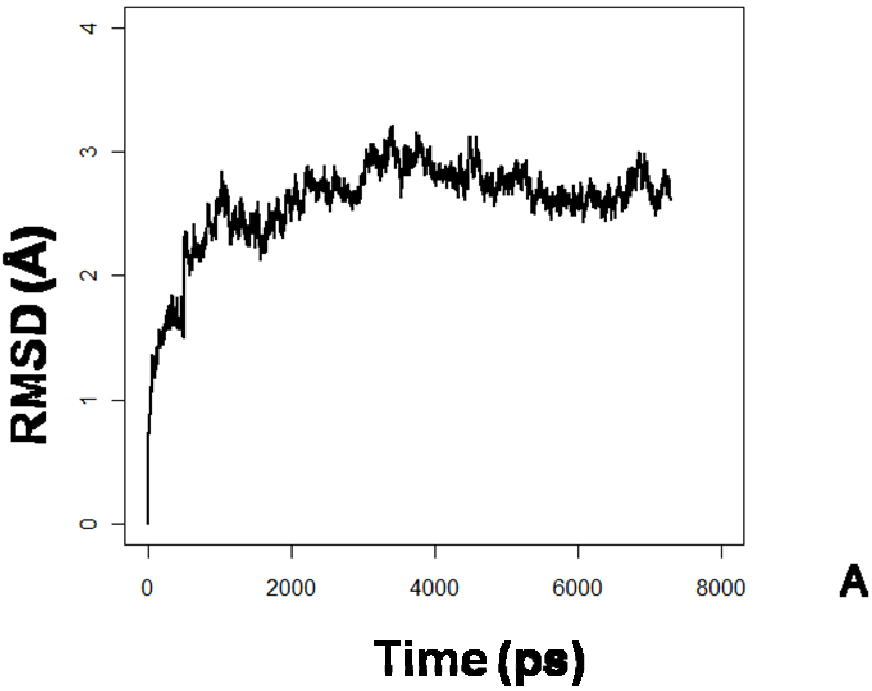


Figure 4 - αC RMSD as a function of time with respect to the first structure of the production stage for the phosphorylated (A) and unphosphorylated (B) LYK3 kinase domains along 7.3ns MDsxsimulation in explicit water.



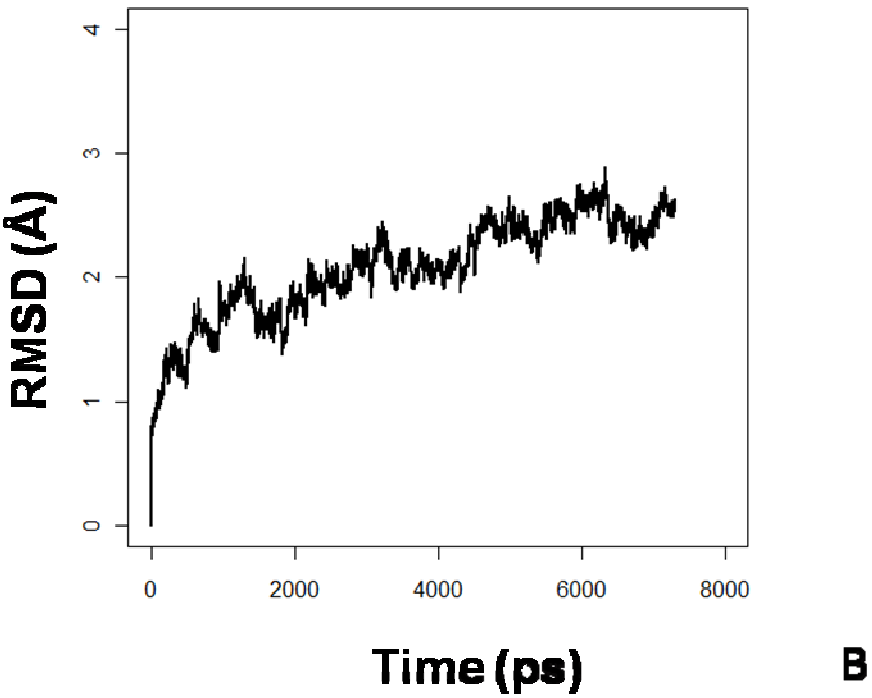
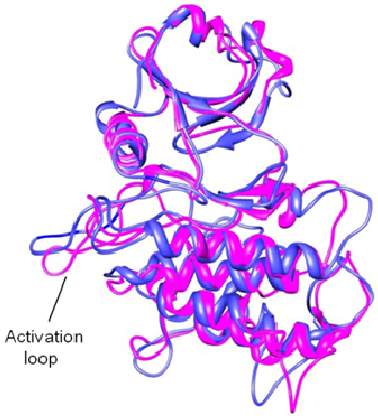
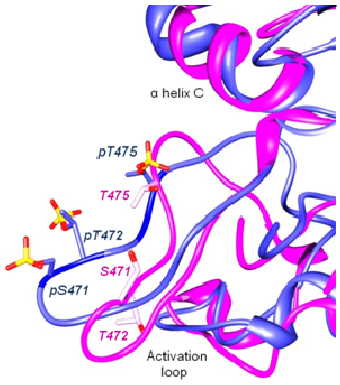


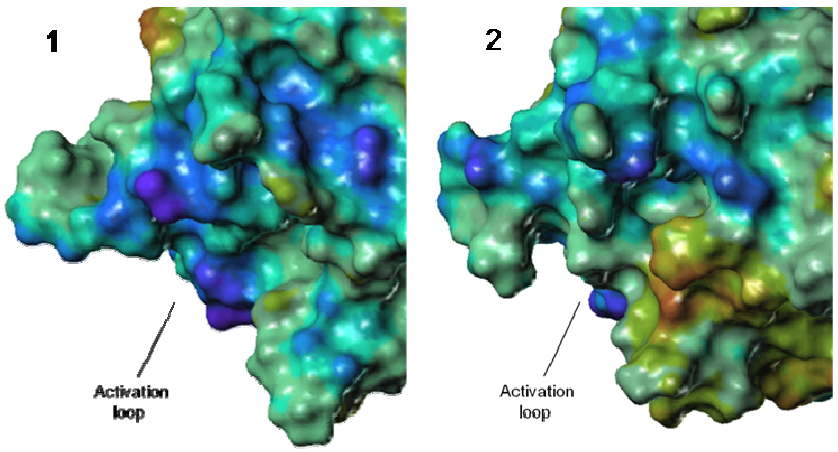
Figure 5 - Ribbon representation of the superimposed phosphorylated (blue ribbon) and unphosphorylated (magenta ribbon) LYK3 structures at the end of the MDs simulations. B. Details of the conformational changes at the active loop region level in the phosphorylated (blue ribbon) and unphosphorylated (magenta ribbon) LYK3 model. C. Electrostatic potential surface maps of the activation loop of phosphorylated (1) and unphosphorylated LYK3 kinase (2) The most positive electrostatic potential is coloured in red whereas the most negative is represented in violet.



A



B



C

SECTION 4
CONCLUSIONS & REFERENCES

4 Conclusions & References

4.1 General conclusions

Carbohydrate-protein systems of relevant biological interest have been studied by means of classical computational techniques such as molecular docking, molecular dynamics simulations and homology modeling.

Different flexible docking approaches have been tested on calcium dependent lectins. These proteins contain metals in the binding pocket that directly coordinate sugars. Among the docking program used in this work, the commercial package Glide slightly outperformed the other academic docking engines, DOCK and Autodock, in terms of RMSD calculated from the experimental structures. However, the best results indicated by the Glide docking score did not always correspond to the best solution, making difficult a reliable docking evaluation when the crystallographic pose is unknown. The academic program AutoDock and the clustering histograms created after each docking run always estimated the right carbohydrate position according to the experimental information with low RMSD values. Autodock has demonstrated to be a user-friendly program that can be adapted to particular situations as in the case of metallo-proteins. Further improvements could be performed by optimizing the docking parameters for the calcium ion. Also results would be more statistically reliable by significantly increasing the number of runs or the number of lectins considered in the data set. The proven capability of modern algorithms to mimic this particular metal dependent binding mode is encouraging: screening of large databases of molecules could be carried out using these methodologies to determine therapeutically active compounds and to evaluate the agreement between computed energies and known affinities.

The docking approach described for calcium mediated protein-carbohydrate interactions using Autodock has been successively implied for predicting the binding mode of oligosaccharides on the human lectin langerin. It has been shown that langerin can bind HIV-1 and HIV-2 gp120 proteins, supporting gp120 glycan recognition and blocking the virus propagation. Previous experimental published data indicated that the high-mannose type N glycan, abundantly expressed on gp120 and its derivative oligosaccharide are among the best ligands for langerin. Consequently, two linear mannosidic fragments were considered. The

docking calculation was not limited to the calcium-binding site but included the whole monomer. Two different binding sites were identified: a calcium-dependent binding site and a calcium independent binding site. The last one is a very hydrophilic region, corresponding to calcium sites in most C-type lectins, contains many residues favorable to carbohydrate binding and has indeed been reported as a secondary mannose-binding site in the recent x-ray structure of langerin. The orientation of the oligosaccharides in the binding sits of langerin is similar to the ones recently observed for the monosaccharide mannose in langerin and, in general, in mannose-binding proteins, validating the prediction in silico. Modeling of ectodomain of langerin also suggested its trimeric shape, whose structural features were in good agreement with the experimental values determined by size exclusion chromatography and sedimentation velocity. All the information derived from this work can be eventually used as a starting point for the conception of glycomimetics that enhance the biological function of the protein target.

The atomic bases of the tetramer PA-IL, *Pseudomonas aeruginosa* lectin I, were elucidated using a combination of experimental and computational data. Previous studies demonstrated that PA-IL has affinity towards aGal-bearing oligosaccharides that represent the tail of sphingolipids expressed on erythrocytes and endothelial cells. For instance, PA-IL agglutinates human erythrocytes that bear the B epitope, terminating by αGal1–3Gal. The lectin is able to differentiate P-positive (P1 and Pk both bearing αGal1–4Gal) and P-negative (p) red blood cells. αGal1–4Gal was recently proposed as a more general cancer-related marker. In the first structural work on PA-IL, docking and MDs simulations confirmed the occurrence of a specific hydrogen bond network for both αGal1–4ßGalOMe and αGal1–3ßGalOMe disaccharides. Indeed, the presentation of glycoconjugates on the cell surface strongly influences the binding. Higher affinity could occur if the molecules have the appropriate presentation for multivalency. Sphingolipids were built from the information derived by crystallography and modeling. However, only the αGal1–4ßGalOMe binding mode was compatible with a correct parallel orientation of the lipid tails in the close binding sites of the PA-IL tetramer. The multivalency features highlighted in silico can justify and confirm the major affitity of PA-IL towards αGal1–4ßGalOMe found by glycanarray essays. The attention was then focused on the interactions between PA-IL and three digalactosides different only by the linkage position. The interactions were evaluated in terms of structure

and energy. Molecular dynamics simulation was a useful tool for the analysis of the solvent in PA-IL binding site, allowing the identification of structural waters. This study revealed the stability of a structural bridge water molecule always present along the simulations in the three disaccharide/PA1L systems. The qualitative approach MM-PBSA was used for predicting the trend of the energies of binding, in which the common structural water molecule was included in the calculations. The predicted energetic features were in excellent agreement with the experimetal data. PA-IL showed a preference for the 1-2 linked disaccharides: in the resulting interaction, the enthalpy term overpassed the entropy cost, resulting in a stronger affinity constant. These works provide key information for understanding how specific carbohydrates bind to the PA-1L binding site. Molecular docking and molecular dynamics simulations data, in agreement with structural data derived from crystallography, can be considered a valid method for the prediction of the behavior of different ligands in PA1L binding site. The identification of structural water molecules in the PA-IL/disaccharide interactions can lead to the conception of ligands that can occupy the bridging water site with a synthetic side chain in order to gain in affinity. Considering the multivalency of lectins, the synthesis of multivalent ligands can be also considered a classical way for gaining avidity. The detailed knowledge about structure and energetics of the PA-IL binding site together with a multivalency approach can be considered the route for the design of antiadhesive compounds against infections by *Pseudomonas aeruginosa*. The accurate characterization of this lectin can be useful not only in drug design but also in medicine and diagnostics, since this lectin can be used as tool for cell typing, tumor targeting and in the xenotransplantation research.

The symbiosis between legumes and bacteria is fundamental for reaching and mantaining a physiological equilibrium between plants and soil since nitrogen, the main product of this process, is necessary for plants'growth. Nodulation factors are the bacterial signals of this process. They are flexible molecules, lipochitoligosaccharides, which may adopt a variety of shapes in water solution. The three dimensional structures of natural and synthetic Nodulation factors were characterized using a combination of experimental NMR and MDs simulation data. The results of these studies indicated that chemical modifications influence the spatial disposition of the lipid chain and the shape that they can assume, while the carbohydrate portion displays a relatively stable dynamic behaviour. The lipid moieties have

a considerable degree of freedom and can assume rather different orientations in equilibrium. The classification of the acyl chain conformations in solution gives information about the preferential disposition of these molecules, influenced by an explicit water environment. Different biological activities have been reported for the different analogues. From the results of the MDs simulation of Nod factors analogs and the respective native compound, an incoming protein receptor would initially be expected to recognize the carbohydrate scaffold. The high flexibility of the acyl moieties could play an important role in modulating the recognition process and the biological response, decreasing accessible portions of the oligosaccharide to the binding site. The different behavior of the lipid moieties elucidated in this study could explain and support the different biological activity of these compounds. The receptor binding site, probably located on extracellular domains of plant kinases, can clearly influence the preferred conformation of these molecules. However, in absence of structural data concerning the receptor, these studies might be useful for explaining the different biological activity of these molecules.

The lysin motif domain-containing receptor-like kinase-3 (RLK-LYK3) of the legume *Medicago truncatula* is considered one of the possible Nod factor receptor involved in the symbiosis induction. It shows 37% amino acid sequence identity with the human Interleukin-1 receptor-associated kinase-4 (IRAK-4), over the kinase domains. Using the structure of this animal kinase as a template, homology modelling revealed that the plant RLK contains structural features particular to this group of kinases, including the tyrosine gatekeeper, the N-terminal extension helix B and a pattern of basic and serine/threonine residues in the activation loop. Amino acid substitutions in conserved residues showed that kinase activity of LYK3 is essential for its biological role in the establishment of the root nodule nitrogen-fixing symbiosis with rhizobial bacteria. The kinase domain of LYK3 has dual Ser/Thr and Tyr specificity and mass spectrometry analysis identified six serine, eight threonine and potentially one tyrosine residue as autophosphorylation sites *in vitro*. Three activation loop Ser/Thr residues are required for full kinase and biological activities and homology modeling identified Thr475 as the prototypical phosphorylated residue, whereas Thr472 may be involved in substrate access. A threonine in the juxtamembrane region and two threonines in the C-terminal lobe of the kinase domain are important for biological but not kinase activity. The structure-function similarities identified between LYK3 and IRAK-4 may be

more widely applicable to plant RLKs/RLCKs and could give additional information about the kinase mechanism.

4.2 Conclusions générales

Les cibles biologiques glucides-protéines ont été étudiées par des techniques classiques de modélisation moléculaire, l'amarrage moléculaire (*docking*), la dynamique moléculaire et la modélisation par homologie.

Différentes méthodes d'amarrage moléculaire ont été testées sur des lectines dépendantes du calcium. Ces protéines contiennent des métaux dans la poche de reconnaissance qui coordonnent directement les glucides. Parmi les programmes utilisés dans ce travail, Glide était le plus performant comparé aux autres programmes de *docking*, DOCK et Autodock, en considerant les valeurs de RMSD calculées à partir des structures expérimentales. Cependant, les résultats indiqués par le meilleur *score* de *docking* ne représentent pas, dans tous les cas, les meilleures solutions, en rendant difficile une évaluation fiable de résultats lorsque la pose cristallographique du ligand est inconnue. La méthode d'évaluation statistique finale du programme Autodock a toujours indiquée la bonne position des glucides dans le site de reconnaissance, en accord avec les données expérimentales. Autodock est un programme convivial qui peut être adapté à des situations particulières comme dans le cas des métallo-protéines. D'autres améliorations techniques pourraient être réalisées en optimisant les paramètres d'amarrage moléculaire pour le calcium. De plus, les résultats seraient statistiquement plus fiables en augmentant significativement le nombre d'évaluations énergétiques ou le nombre de lectines prises en compte. La capacité des algorithmes modernes de *docking* d'imiter ce mode particulier de liaison dépendant du calcium est encourageante: le *screening* de grandes bases de données moléculaires pourraient être réalisées avec ces méthodes, en permettant d'isoler des composés thérapeutiquement actifs, d'évaluer et de prédire les énergies de liaison.

Autodock a été successivement utilisé pour prédire des possibles modes de liaison entre oligosaccharides et Langerine, une lectine humaine dépendante du calcium. Cette protéine peut interagir et se lier avec les épitopes gp120 des virus VIH-1 et VIH-2, en bloquant la propagation virale. Des données expérimentales précédemment publiées ont indiqué que les

glucides exprimés sur la surface de gp120 sont constitués par des résidus de mannose. En conséquence, deux fragments linéaires ont été construits et utilisés pour les calculs de *docking*. Le monomère de la Langerine a été pris en considération pour la recherche de possibles sites de reconnaissance glucidiques. Deux différents sites de *binding* ont été identifiés: un site dépendant du calcium et un autre, metallo-indépendant. Ce dernier on se trouve dans une zone très hydrophile qui correspond à des sites dépendants du calcium dans la plupart des lectines qui contiennent des nombreux résidus qui favorisent la liaison avec les glucides. Ce site a été récemment signalé comme site de reconnaissance secondaire dans la structure cristallographique de la Langerine. L'orientation des oligosaccharides dans le site de liaison dépendant du calcium de la Langerin est similaire à l'orientation récemment observée pour le mannose dans la Langerine et, en général, dans les protéines liant le mannose, qui valide la prédiction *in silico*. La modélisation de la forme trimérique de la Langerine propose un modèle tridimensionnel dont les caractéristiques structurelles sont en bon accord avec les valeurs expérimentales. Les résultats de ce travail peuvent être utilisés comme point de départ pour la conception de glycomimétiques qui peuvent incrémenter la fonction biologique de la protéine cible.

Les bases atomiques de la lectine PA-IL de *Pseudomonas aeruginosa* ont été élucidées en combinant données expérimentales et computationnelles. Des études ont démontré que PA-IL a une affinité pour les résidus αGal qui représentent la partie terminale de sphingolipides, molécules exprimées sur la surface des érythrocytes et des cellules endothéliales. Dans le premier travail structurel sur la lectine PA-IL, des calculs de *docking* et des simulations de dynamique moléculaire ont confirmé la présence d'un réseau de liaisons hydrogène spécifiques pour les disaccharides αGal1-4ßGalOMe et αGal1-3ßGalOMe. Sphingolipides ont été construits à partir des informations obtenues par cristallographie et par modélisation. Toutefois, seulement le mode de liaison du sphingolipide qui contient le disaccharide aGal1-4ßGalOMe présentait des caractéristiques de multivalence: ce résultat a pu justifier et confirmer la plus haute affinité de PA-IL vers αGal1-4ßGalOMe déterminée *in vitro*. Les interactions entre PA-IL et trois différents digalactosides ont été successivement étudiées *in silico*, en termes de structure et d'énergie. L'analyse du rôle du solvant dans le site de liaison a révélé la présence d'une molécule d'eau structurale dans les trois systèmes disaccharide/PA-IL. L'approche qualitative MM-PBSA a été utilisée pour prédire la

tendance des énergies de liaison. Lorsque la molécule d'eau structurale était incluse dans les calculs, les valeurs énergiques étaient en excellent accord avec les données expérimentales. PA-IL a montré sa préférence pour des disaccharides avec une liaison 1-2. Ces travaux fournissent des informations utiles pour comprendre la manière dont des glucides spécifiques se lient à PA-IL. Les méthodes computationnelles peuvent être considérées comme des outils fiables pour la prédiction du comportement de différents ligands dans le site de liaison de cette lectine. L'identification d'une molécule d'eau structurale dans les interactions PA-IL/disaccharide peut être utile pour la génération de ligands avec une chaîne latérale synthétique qui occupe l'espace de la molécule d'eau afin de gagner en affinité. La synthèse de ligands multivalents peut aussi être considérée comme un moyen pour gagner en avidité. Les connaissances détaillées de la structure et de la thermodynamique de la lectine PA-IL peuvent aider à la conception de molécules antiadhésives contre les infections par *Pseudomonas aeruginosa* et à la production d'outils diagnostics.

La symbiose entre les légumineuses et les bactéries est fondamentale pour établir un équilibre physiologique entre les plantes et le sol car l'azote, le principal produit de ce processus, est nécessaire pour la croissance des plantes. Les facteurs de Nodulation, facteurs Nod, sont les signaux bactériens. Ce sont des molécules flexibles, lipochitoligosaccharides, qui peuvent adopter une variété de formes en solution aqueuse. Les structures tridimensionnelles des facteurs de Nodulation naturels et synthétiques ont été caractérisés en utilisant une combinaison de méthodes expérimentales, RMN, et computationnelles, dynamique moléculaire. Les résultats de ces études ont indiqué que les modifications chimiques influencent la disposition spatiale de la chaîne lipidique, tandis que la partie glucidique affiche un comportement relativement stable. Les fractions lipidiques ont un haut degré de liberté et peuvent montrer des orientations assez différentes à l'équilibre. A partir des résultats de dynamique moléculaire, on peut imaginer qu'un récepteur serait initialement capable de reconnaître la partie glucidique des facteurs Nod. La grande flexibilité des fractions lipidiques pourrait jouer un rôle important dans la modulation du processus de reconnaissance, en diminuant la partie glucidique accessible au site de liaison. Les différents comportements des fractions lipidiques élucidés dans cette étude pourraient expliquer la différente activité biologique de ces composés. Les sites de reconnaissance des

facteurs Nod, probablement situés dans les domaines extracellulaires des kinases végétales, peuvent clairement influencer la conformation préférentielle de ces molécules. Toutefois, en absence de données concernant la structure du récepteur, ces études pourraient être utiles pour expliquer la différente activité biologique de ces molécules.

La kinase RLK-LYK3 de la légumineuse *Medicago truncatula* est considérée comme un des récepteurs des facteurs Nod impliqués dans l'induction de la symbiose. Cette protéine présente 37% d'identité de séquence avec une kinase humaine, IRAK-4. En utilisant la structure de cette kinase comme gabarit, la modélisation par homologie a révélé que la kinase LYK3 a des caractéristiques structurelles similaires à ce groupe de kinases animales. Les mutations d'acides aminés particuliers ont demontré que l'activité de cette kinase est essentielle dans la symbiose. Trois résidus Ser / Thr sont nécessaires pour obtenir une complète activité kinasique et biologique: la modélisation par homologie a identifié le résidu Thr475 comme le résidu phosphorylé essentiel pour l'activité, lorsque Thr472 peut être impliqué dans l'accès du substrat. Les similitudes identifiées entre LYK3 et IRAK-4 peuvent aussi donner plus d'informations sur le mécanisme d'action des kinases des plantes.

4.3 References

1. Guo, Z., Boons,G-J., *Carbohydrate Based Vaccines and Immunotherapies*. Wiley Series in Drug Discovery and Development; 2009

2. Varki, A. C., R; Esko,J;Freeze,H; Stanley, P; Bertozzi, C; Hart, G; Etzler, M *Essentials of glycobiology*. 2 ed.; 2008

3. Ohtsubo, K.; Marth, J. D., Glycosylation in cellular mechanisms of health and disease. *Cell* **2006,** 126, (5), 855-867

4. Richard, E.; Vega, A. I.; Pérez, B.; Roche, C.; Velázquez, R.; Ugarte, M.; Pérez-Cerdá, C., Congenital disorder of glycosylation Ia: new differentially expressed proteins identified by 2-DE. *Biochemical and Biophysical Research Communications* **2009,** 379, (2), 267-271

5. Hennet, T., From glycosylation disorders back to glycosylation: What have we learned? *BBA-Molecular Basis of Disease* **2009,** 1792, (9), 921-924

6. Freeze, H. H., Update and perspectives on congenital disorders of glycosylation. *Glycobiology* **2001,** 11, (12), 129-143

7. Brooks, S. A.; Carter, T. M.; Royle, L.; Harvey, D. J.; Fry, S. A.; Kinch, C.; Dwek, R. A.; Rudd, P. M., Altered glycosylation of proteins in cancer: what is the potential for new anti-tumour strategies. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry)* **2008,** 8, (1), 2-21

8. Fukuda, M., Possible roles of tumor-associated carbohydrate antigens. *Cancer research* **1996,** 56, (10), 2237–2244

9. Xu, Y.; Sette, A.; Sidney, J.; Gendler, S. J.; Franco, A., Tumor-associated carbohydrate antigens: A possible avenue for cancer prevention. *Immunology and cell biology* **2005,** 83, (4), 440-448

10. Arnold, J. N.; Wormald, M. R.; Sim, R. B.; Rudd, P. M.; Dwek, R. A., The impact of glycosylation on the biological function and structure of human immunoglobulins. *J Annual Review of Immunology* **2007,** 25, 21-50

11. Dobos, K. M.; Khoo, K. H.; Swiderek, K. M.; Brennan, P. J.; Belisle, J. T., Definition of the full extent of glycosylation of the 45-kilodalton glycoprotein of Mycobacterium tuberculosis. *Journal of bacteriology* **1996,** 178, (9), 2498–2506

12. Campbell, B. J.; Finnie, I. A.; Hounsell, E. F.; Rhodes, J. M., Direct demonstration of increased expression of Thomsen-Friedenreich (TF) antigen in colonic adenocarcinoma and ulcerative colitis mucin and its concealment in normal mucin. *Journal of Clinical Investigation* **1995,** 95, (2), 571-576

13. Campbell, B. J.; Yu, L. G.; Rhodes, J. M., Altered glycosylation in inflammatory bowel disease: a possible role in cancer development. *Glycoconjugate journal* **2001,** 18, (11), 851-858

14. Brooks, S. A., Strategies for Analysis of the Glycosylation of Proteins: Current Status and Future Perspectives. *Molecular Biotechnology* **2009,** 43, (1), 76-88

15. Abbott, K. L.; Aoki, K.; Lim, J. M.; Porterfield, M.; Johnson, R.; O'Regan, R. M.; Wells, L.; Tiemeyer, M.; Pierce, M., Targeted glycoproteomic identification of biomarkers for human breast carcinoma. *Journal of Proteome Research* **2008,** 7, (4), 1470-1480

16. Block, T. M.; Comunale, M. A.; Lowman, M.; Steel, L. F.; Romano, P. R.; Fimmel, C.; Tennant, B. C.; London, W. T.; Evans, A. A.; Blumberg, B. S., Use of targeted glycoproteomics to identify serum glycoproteins that correlate with liver cancer in woodchucks and humans. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** 102, (3), 779–784

17. Dube, D. H.; Bertozzi, C. R., Glycans in cancer and inflammation potential for therapeutics and diagnostics. *Nature Reviews Drug Discovery* **2005,** 4, (6), 477-488

18. Karlsson, K. A., Animal glycolipids as attachment sites for microbes. *Chemistry and physics of lipids* **1986,** 42, (1-3), 153-172

19. Karlsson, K. A., Animal glycosphingolipids as membrane attachment sites for bacteria. *Annual review of biochemistry* **1989,** 58, (1), 309-350

20. Sharon, N., Carbohydrates as future anti-adhesion drugs for infectious diseases. *BBA-General Subjects* **2006,** 1760, (4), 527-537

21. Wagner, R.; Matrosovich, M.; Klenk, H. D., Functional balance between haemagglutinin and neuraminidase in influenza virus infections. *Reviews in medical virology* **2002,** 12, (3), 159-166

22. Wellens, A.; Garofalo, C.; Nguyen, H.; Van Gerven, N.; Slättegård, R.; Hernalsteens, J. P.; Wyns, L.; Oscarson, S.; De Greve, H.; Hultgren, S., Intervening with Urinary Tract Infections Using Anti-Adhesives Based on the Crystal Structure of the FimH–Oligomannose-3 Complex. *PLoS One* **2008,** 3, (4), e2040

23. Lucas, A. H.; Apicella, M. A.; Taylor, C. E.; Gellin, B.; Modlin, J. F., Carbohydrate moieties as vaccine candidates. *Clinical Infectious Diseases* **2005,** 41, (5), 705-712

24. Jennings, H., Further approaches for optimizing polysaccharide-protein conjugate vaccines for prevention of invasive bacterial disease. *The Journal of infectious diseases* **1992,** 165, 156-159

25. Aragao, K. Etudes structure-fonction de lectines (DiscI et DiscII) de Dictyostelium discoideum. Joseph Fourier university, Grenoble (France), 2008

26. Rao, V. S. R.; Qasba, P. K.; Chandrasekaran, R.; Balaji, P. V., *Conformation of carbohydrates*. CRC: 1998

27. Lindhorst, T. K., *Essentials of carbohydrate chemistry and biochemistry*. Wiley-Vch 2007

28. Juaristi, E.; Cuevas, G., Recent studies of the anomeric effect. *Tetrahedron* **1992,** 48, (24), 5019-5087

29. Box, V. G. S., The anomeric effect of monosaccharides and their derivatives. Insights from the new QVBMM molecular mechanics force field. *Heterocycles* **1998,** 48, (11), 2389-2417

30. Takahashi, O.; Yamasaki, K.; Kohno, Y.; Ohtaki, R.; Ueda, K.; Suezawa, H.; Umezawa, Y.; Nishio, M., The anomeric effect revisited. A possible role of the CH/n hydrogen bond. *Carbohydrate research* **2007,** 342, (9), 1202-1209

31. Cremer, D.; Pople, J. A., General definition of ring puckering coordinates. *Journal of the American Chemical Society* **1975,** 97, (6), 1354-1358

32. Ferro, D. R.; Provasoli, A.; Ragazzi, M.; Casu, B.; Torri, G.; Bossennec, V.; Perly, B., Conformer populations of L-iduronic acid residues in glycosaminoglycan sequences. *Carbohydrate research* **1990,** 195, (2), 157-167

33. Marchessault, R. H.; Perez, S., Conformations of the hydroxymethyl group in crystalline aldohexopyranoses. *Peptide Science* **1979,** 18, (9), 2369-2374

34. Kirschner, K. N.; Woods, R. J., Solvent interactions determine carbohydrate conformation. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** 98, (19), 10541-10545

35. Gabius, H. J., *The sugar code: fundamentals of glycosciences*. Wiley-VCH: 2009

36. Lemieux, R. U.; Koto, S., The conformational properties of glycosidic linkages. *Tetrahedron* **1974,** 30, (13), 1933-1944

37. Kjellen, L.; Lindahl, U., Proteoglycans: structures and interactions. *Annual review of biochemistry* **1991,** 60, (1), 443-475

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

39. Chester, M. A., Nomenclature of glycolipids. *Pure and Applied Chemistry* **1997,** 69, 2475-2487

40. Engelsen, S. B.; Perez, S., Unique similarity of the asymmetric trehalose solid-state hydration and the diluted aqueous-solution hydration. *J. Phys. Chem. B* **2000,** 104, (39), 9301-9311

41. Reynolds, M.; Fuchs, A.; Lindhorst, T. K.; Perez, S., The hydration features of carbohydrate determinants of Lewis antigens. *Molecular Simulation* **2008,** 34, (4), 447-460

42. Braccini, I.; Perez, S., Molecular basis of Ca2+-induced gelation in alginates and pectins: the egg-box model revisited. *Biomacromolecules* **2001,** 2, (4), 1089-1096

43. Sharon, N.; Lis, H., *Lectins*. Kluwer Academic Pub: 2003

44. Lis, H.; Sharon, N., Lectins: Carbohydrate-Specific Proteins That Mediate Cellular Recognition. *Chem. Rev* **1998,** 98, (2), 637-674

45 Reynolds, M. Synthesis and biological evaluation of gold glyconanoparticles: ligands for studying multivalence effects. Joseph Fourier University, Grenoble (France), 2009

46. Ogata, S.; Muramatsu, T.; Kobata, A., Fractionation of glycopeptides by affinity column chromatography on concanavalin A-Sepharose. *Journal of Biochemistry* **1975,** 78, (4), 687-696

47. Pokorna, M.; Cioci, G.; Perret, S.; Rebuffet, E.; Kostlanova, N.; Adam, J.; Gilboa-Garber, N.; Mitchell, E. P.; Imberty, A.; Wimmerova, M., Unusual Entropy-Driven Affinity of Chromobacterium violaceum Lectin CV-IIL toward Fucose and Mannose†,‡. *Biochemistry* **2006,** 45, (24), 7501-7510

48. Naismith, J. H.; Emmerich, C.; Habash, J.; Harrop, S. J.; Helliwell, J. R.; Hunter, W. N.; Raftery, J.; Kalb, A. J.; Yariv, J., Refined structure of concanavalin A complexed with methyl aD-mannopyranoside at 2.0 A resolution and comparison with the saccharide-free structure. *Acta Crystallogr.* **1994,** 50, 847-858

49. Cioci, G.; Mitchell, E. P.; Chazalet, V.; Debray, H.; Oscarson, S.; Lahmann, M.; Gautier, C.; Breton, C.; Perez, S.; Imberty, A., B-propeller Crystal Structure of Psathyrella velutina Lectin: An Integrin-like Fungal Protein Interacting with Monosaccharides and Calcium. *Journal of molecular biology* **2006,** 357, (5), 1575-1591

50. Wimmerova, M.; Mitchell, E.; Sanchez, J. F.; Gautier, C.; Imberty, A., Crystal structure of fungal lectin. *Journal of Biological Chemistry* **2003,** 278, (29), 27059-27067

51. Elgavish, S.; Shaanan, B., Structures of the Erythrina corallodendron lectin and of its complexes with mono-and disaccharides1. *Journal of molecular biology* **1998,** 277, (4), 917-932

52. Delbaere, L. T. J.; Vandonselaar, M.; Prasad, L.; Quail, J. W.; Wilson, K. S.; Dauter, Z., Structures of the lectin IV of Griffonia simplicifolia and its complex with the Lewis b human blood group determinant at 2.0 Å resolution. *J Mol Biol* **1993,** 230, 950-965

53. Cioci, G. Etude structure - fonction de glycoconjugues et de lectines bacteriennes et fongiques. Joseph Fourier University, Grenoble (France), 2006

54. Becker, O. M., MacKerell, A.; Roux, B.; Watanabe, M. , *Computational biochemistry and biophysics*. Dekker M: New York, 2001

55. Höltje, H. D.; Sippl, W.; Rognan, D.; Folkers, G., *Molecular modeling: basic principles and applications*. Wiley-Vch 2008

56. Mackerell, A. D., Empirical force fields for biological macromolecules: overview and issues. *Journal of computational chemistry* **2004,** 25, (13), 1584-1604

57. Ponder, J. W.; Case, D. A., Force fields for protein simulations. *Advances in protein chemistry* **2003,** 66, 27-86

58. Eklund, R. Computational analysis of carbohydrates:dynamical properties and interactions. Stockholm University, Stockholm (Sweden), 2005

59. Walker, R. C. The development of a QM/MM based linear response method and its application to proteins. Imperial College London, London (U.K.), 2003

60. Bultinck, P., *Computational medicinal chemistry for drug discovery*. Dekker M.: New York, 2004

61. Allinger, N. L.; Yuh, Y. H.; Lii, J. H., Molecular mechanics. The MM3 force field for hydrocarbons. *Journal of the American Chemical Society* **1989,** 111, (23), 8551-8566

62. Morse, P. M., Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Physical Review* **1929,** 34, (1), 57-64

63. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., 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

64. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **1993,** 97, (40), 10269-10280

65. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollmann, P. A., Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society* **1993,** 115, (21), 9620-9631

66. Hill, T. L., Steric Effects. I. Van der Waals Potential Energy Curves. *The Journal of Chemical Physics* **1948,** 16, 399-404

67. Kranjc, A. Predicting structural determinants and ligand poses in proteins involved in neurological diseases: bioinformatics and molecular simulation studies International School for Advanced Studies (SISSA/ISAS), Trieste (Italy), 2009

68. Zhang, Y., Progress and challenges in protein structure prediction. *Current opinion in structural biology* **2008,** 18, (3), 342-348

69. Madhusudhan, M. S.; Marti-Renom, M. A.; Eswar, N.; John, B.; Pieper, U.; Karchin, R.; Shen, M. Y.; Sali, A., *The Proteomics Protocols Handbook*. Totowa, NJ, 2005; p 831–860

70. Tsai, C. S., *An introduction to computational biochemistry*. Wiley-Liss: 2002

71. Fortune', A. Techniques de Modélisation Moléculaire Appliquées à l'Etude et à l'Optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance. Joseph Fourier University, Grenoble (France), 2006

72. Rohl, C. A.; Baker, D., De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc* **2002,** 124, (11), 2723-2729

73. Kim, D. E.; Chivian, D.; Baker, D., Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research* **2004,** 32, 526-531

74. Mosimann, S.; Meleshko, R.; James, M. N. G., A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins: Structure, Function, and Bioinformatics* **1995,** 23, (3), 301-317

75. Doolittle, R. F., Searching through sequence databases. *Methods in enzymology* **1990,** 183, 99-110

76. Pearson, W. R., Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology* **1990,** 183, 63-98

77. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *Journal of molecular biology* **1990,** 215, (3), 403-410

78. Needleman, S. B.; Wunsch, C. D., A general method applicable to search for similarities in the amino acid sequences of the two proteins. *J. Mol. Biol* **1970,** 48, 443-453

79. Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C., A model of evolutionary change in proteins. *Atlas of protein sequence and structure* **1978,** 5, (Suppl 3), 345–352

80. Henikoff, S.; Henikoff, J. G., Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **1992,** 89, (22), 10915-10919

81. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **1997,** 25, (17), 3389-3402

82. Thompson, J. D., Higgins, D.G. and Gibson, T.J. , CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* **1994,** 22, 4673 - 4680

83. Shi, J.; Blundell, T. L.; Mizuguchi, K., FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology* **2001,** 310, (1), 243-257

84. Montalvao, R. W.; Smith, R. E.; Lovell, S. C.; Blundell, T. L., CHORAL: a differential geometry approach to the prediction of the cores of protein structures. *Bioinformatics* **2005,** 21, (19), 3719-3725

85. Deane, C. M.; Blundell, T. L., CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Science: A Publication of the Protein Society* **2001,** 10, (3), 599-612

86. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C., The penultimate rotamer library. *Proteins Structure Function and Genetics* **2000,** 40, (3), 389-408

87. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M., PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **1993,** 26, (2), 283-291

88. Vriend, G., WHAT IF: A molecular modeling and drug design program. *Journal of Molecular Graphics* **1990,** 8, (1), 52-56

89. Lengauer, T.; Rarey, M., Computational methods for biomolecular docking. *Current opinion in structural biology* **1996,** 6, (3), 402-406

90. Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design* **2001,** 15, (5), 411-428

91. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **1998,** 19, (14), 1639-1662

92. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem* **2004,** 47, (7), 1739-1749

93. Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. L., OPLS all-atom force field for carbohydrates. *Journal of computational chemistry* **1997,** 18, (16), 1955-1970

94. Muegge, I.; Martin, Y. C., A General and Fast Scoring Function for Protein- Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem* **1999,** 42, (5), 791-804

95. Cramer, C. J., *Essentials of computational chemistry: theories and models*. John Wiley & Sons: New York, 2004

96. Verlet, L., Computer 'Experiments' on Classical Fluids. I. Thermodynamical Propertier of Lennard-Jones Fluid. *Mol. Phys. Rev* **1967,** 159, 98-103

97. Hockney, R. W., The potential calculation and some applications. *Methods in Computational Physics* **1970,** 9, 136-211

98. Case, D. A.; Cheatham Iii, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of computational chemistry* **2005,** 26, (16), 1668-1688

99. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977,** 23, (3), 327-341

100. Gilson, M. K.; Sharp, K. A.; Honig, B. H., Calculating the electrostatic potential of molecules in solution: method and error assessment. *Journal of Computational Chemistry* **1988,** 9, (4), 327-335

101. Wojciechowski, M.; Lesyng, B., Generalized Born model: Analysis, refinement, and applications to proteins. *Journal of physical chemistry. B, Condensed matter, materials, surfaces, interfaces, & biophysical chemistry* **2004,** 108, (47), 18368-18376

102. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983,** 79, 926-935

103. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993,** 98, 10089-10092

104. Mishra, N. K. Computational study on lectin-carbohydrate complexes. National Centre for Biomolecular Research, Brno (Czech Republic), 2009

105. Retegan, M. Etudes de systèmes organométalliques et biologiques par des méthodes hybrides mécanique quantique/mécanique moléculaire. Université Joseph Fourier, Grenoble, 2009

106. DeMarco, M. L.; Woods, R. J., Structural glycobiology: A game of snakes and ladders. *Glycobiology* **2008,** 18, (6), 426-440

107. Brandsdal, B. O.; Osterberg, F.; Almlof, M.; Feierberg, I.; Luzhkov, V. B.; Aqvist, J., Free energy calculations and ligand binding. *Advances in protein chemistry* **2003,** 66, 123-158

108. Gilson, M. K.; 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

109. Sitkoff, D.; Sharp, K. A.; Honig, B., Accurate calculation of hydration free energies using macroscopic solvent models. *The Journal of Physical Chemistry* **1994,** 98, (7), 1978-1988

110. Srinivasan, J.; Cheatham Iii, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A., Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate- DNA Helices. *J. Am. Chem. Soc* **1998,** 120, (37), 9401-9409

111. Gohlke, H.; Case, D. A., 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

112. Basdevant, N.; Weinstein, H.; Ceruso, M., Thermodynamic Basis for Promiscuity and Selectivity in Protein- Protein Interactions: PDZ Domains, a Case Study. *J. Am. Chem. Soc* **2006,** 128, (39), 12766-12777

113. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W. E. I., Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res* **2000,** 33, 889-897

114. Woods, R. J., Computational carbohydrate chemistry: what theoretical methods can tell us. *Glycoconjugate journal* **1998,** 15, (3), 209-216

115. Takahashi, N.; Yagi, H.; Kato, K., Comprehensive Glycoscience In Elsevier: 2007; Vol. 2

116. Sorin, E. J.; Pande, V. S., Empirical force-field assessment: the interplay between backbone torsions and noncovalent term scaling. *Journal of computational chemistry* **2005,** 26, (7), 682-690

117. Fadda, E.; Woods, R.J., Molecular simulations of carbohydrates and protein–carbohydrate interactions: motivation, issues and prospects**.** *Drug Discovery Today* **2010**, in press

118. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J., GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry* **2008,** 29, (4), 622-655

119. Tessier, M. B.; DeMarco, M. L.; Yongye, A. B.; Woods, R. J., Extension of the GLYCAM06 biomolecular force field to lipids, lipid bilayers and glycolipids. *Molecular Simulation* **2008,** 34, (4), 349-364

120. Dejoux, A.; Cieplak, P.; Hannick, N.; Moyna, G.; Dupradeau, F. Y., AmberFFC, a flexible program to convert AMBER and GLYCAM force fields for use with commercial molecular modeling packages. *Journal of Molecular Modeling* **2001,** 7, (11), 422-432

121. Biarnes, X.; Nieto, J.; Planas, A.; Rovira, C., Substrate Distortion in the Michaelis Complex of Bacillus 1, 3-1, 4-beta-Glucanase:insight from first principles molecular dynamics simulations *Journal of Biological Chemistry* **2006,** 281, (3), 1432-1441

122. Basma, M.; Sundara, S.; Calgan, D.; Vernali, T.; Woods, R. J., Solvated ensemble averaging in the calculation of partial atomic charges. *Journal of computational chemistry* **2001,** 22, (11), 1125-1137

123. Woods, R. J.; Khalil, M.; Pell, W.; Moffat, S. H.; Smith, V. H., Derivation of net atomic charges from molecular electrostatic potentials. *Journal of Computational Chemistry* **1990,** 11, (3), 297-310

124. Woods, R. J.; 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

125. Lins, R. D.; Hunenberger, P. H., A new GROMOS force field for hexopyranose-based carbohydrates. *Journal of computational chemistry* **2005,** 26, (13), 1400-1412

126. Schuler, L. D.; Daura, X.; Van Gunsteren, W. F., An improved GROMOS 96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry* **2001,** 22, (11), 1205-1218

127. Schuler, L. D.; Van Gunsteren, W. F., On the choice of dihedral angle potential energy functions for n-alkanes. *Molecular Simulation* **2000,** 25, (5), 301-319

128. Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Hermans, J., Interaction models for water in relation to protein hydration. *Intermolecular forces* **1981**, 331-333

129. Gandhi, N. S.; Mancera, R. L., Free energy calculations of glycosaminoglycan-protein interactions. *Glycobiology* **2009,** 19, (10), 1103-1115

130. Guvench, O.; Greene, S. N.; Kamath, G.; Brady, J. W.; Venable, R. M.; Pastor, R. W.; Mackerell Jr, A. D., Additive empirical force field for hexopyranose monosaccharides. *Journal of Computational Chemistry* **2008,** 29, (15), 2543-2564

131. Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell Jr, A. D., CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *Journal of Chemical Theory and Computation* **2009,** 5, (9), 2353-237

132. Hatcher, E. R.; Guvench, O.; MacKerell Jr, A. D., CHARMM Additive All-Atom Force Field for Acyclic Polyalcohols, Acyclic Carbohydrates, and Inositol. *J. Chem. Theory Comput* **2009,** 5, (5), 1315-1327

133. Hatcher, E.; Guvench, O.; MacKerell Jr, A. D., CHARMM Additive All-Atom Force Field for Aldopentofuranoses, Methyl-aldopentofuranosides, and Fructofuranose. *The Journal of Physical Chemistry B* **2009,** 113, (37), 12466-12476

134. MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B-Condensed Phase* **1998,** 102, (18), 3586-3616

135. Mackerell, A. D.; Feig, M.; Brooks, C. L., 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

136. MacKerell Jr, A. D.; Banavali, N.; Foloppe, N., Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000-2001,** 56, (4), 257-265

137. Kony, D.; Damm, W.; Stoll, S.; Van Gunsteren, W. F., An improved OPLS-AA force field for carbohydrates. *Journal of Computational Chemistry* **2002,** 23, (15), 1416-1429

138. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc* **1996,** 118, (45), 11225-11236

139. Lii, J. H.; Allinger, N. L., The MM 3 force field for amides, polypeptides and proteins. *Journal of Computational Chemistry* **1991,** 12, (2), 186-199

140. Allinger, N. L.; Li, F.; Yan, L.; Tai, J. C., Molecular mechanics(MM3) calculations on conjugated hydrocarbons. *Journal of Computational Chemistry* **1990,** 11, (7), 868-895

141. Clark, M.; Cramer, R. D.; Van Opdenbosch, N., Validation of the general purpose tripos 5. 2 force field. *Journal of Computational Chemistry* **1989,** 10, (8), 982-1012

142. Imberty, A.; Hardman, K. D.; Carver, J. P.; Perez, S., Molecular modelling of protein-carbohydrate interactions. Docking of monosaccharides in the binding site of concanavalin A. *Glycobiology* **1991,** 1, (6), 631-642

143. Perez, S.; Meyer, C.; Imberty, A., Practical tools for accurate modeling of complex carbohydrates and their interactions with proteins. *Modeling of Biomolecular Structures and Mechanisms, A. Pullman, J. Jortner, and B. Pullman, eds (Dordrecht, The Netherlands: Kluwer Academic Publishers)* **1995**, 425–454

144. French, A. D., Rigid- and relaxed-residue conformational analyses of cellobiose using the computer program MM 2. *Biopolymers* **1988,** 27, (9), 1519-1525

145. French, A., Comparisons of rigid and relaxed conformational maps for cellobiose and maltose. *Carbohydrate Research* **1989,** 188, 206-211

146. Mikros, E.; Labrinidis, G.; Pérez, S., Conformational Analysis of C-Disaccharides Using Molecular Mechanics Calculations. *Journal of Carbohydrate Chemistry* **2000,** 19, (9), 1319-1349

147. Braccini, I.; Grasso, R. P.; 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

148. Nyholm, P. G.; Mulard, L. A.; Miller, C. E.; Lew, T.; Olin, R.; Glaudemans, C. P. J., Conformation of the O-specific polysaccharide of Shigella dysenteriae type 1: molecular modeling shows a helical structure with efficient exposure of the antigenic determinant {alpha}-L-Rhap-(1-> 2)-{alpha}-D-Galp. *Glycobiology* **2001,** 11, (11), 945-955

149. Koca, J., Computer program CICADA: travelling along conformational potential energy hypersurface. *Journal of molecular structure. Theochem* **1994,** 308, 13-24

150. Koca, J., Travelling through conformational space: an approach for analyzing the conformational behaviour of flexible molecules. *Progress in biophysics and molecular biology* **1998,** 70, (2), 137-173

151. Perez, S., Conformational analysis and flexibility of carbohydrates using the CICADA approach with MM3. *Journal of Computational Chemistry* **1995,** 16, (3), 296-310

152. Casset, F.; Imberty, A.; du Penhoat, C. H.; Koca, J.; Pérez, S., Validation of two conformational searching methods applied to sucrose: Simulation of NMR and chiro-optical data. *Journal of Molecular Structure: THEOCHEM* **1997,** 395, 211-224

153. Cioci, G.; Rivet, A.; Ko a, J.; Perez, S., Conformational analysis of complex oligosaccharides: the CICADA approach to the uromodulin O-glycans. *Carbohydrate research* **2004,** 339, (5), 949-959

154. Rosen, J.; Robobi, A.; Nyholm, P. G., The conformations of the O-specific polysaccharides of Shigella dysenteriae type 4 and Escherichia coli O159 studied with molecular mechanics (MM3) filtered systematic search. *Carbohydrate research* **2004,** 339, (5), 961-966

155. Rosen, J.; Miguet, L.; Perez, S., Shape: automatic conformation prediction of carbohydrates using a genetic algorithm. *Journal of Cheminformatics* **2009,** 1, (16), 1-7

156. Strino, F.; Nahmany, A.; Rosen, J.; Kemp, G. J. L.; 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

157. Nahmany, A.; Strino, F.; Rosen, J.; Kemp, G. J. L.; Nyholm, P. G., The use of a genetic algorithm search for molecular mechanics (MM3)-based conformational analysis of oligosaccharides. *Carbohydrate research* **2005,** 340, (5), 1059-1064

158. Kalos, M. H.; Whitlock, P. A., *Monte carlo methods*. Wiley-VCH: 2008

159. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N., Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **1953,** 21, (6), 1087-1092

160. Spieser, S.; Mazeau, K.; Brochier, M. C.; Gey, C.; Utille, J. P.; Taravel, F. R., Conformational properties of a cyclic oligosaccharide: cyclotrikis-(1 6)-[-D-glucopyranosyl-(1 4)- -D-glucopyranosyl]. *Glycoconjugate journal* **1998,** 15, (5), 511-521

161. Mazeau, K.; Moine, C.; Krausz, P.; Gloaguen, V., Conformational analysis of xylan chains. *Carbohydrate research* **2005,** 340, (18), 2752-2760

162. Monteiro, M. A.; Slavic, D.; St. Michael, F.; Brisson, J. R.; MacInnes, J. I.; Perry, M. B., The first description of a (1 6) -D-glucan in prokaryotes:(1 6)-D-glucan is a common component of Actinobacillus suis and is the basis for a serotyping system. *Carbohydrate Research* **2000,** 329, (1), 121-130

163. Ghose, A. K.; Jaeger, E. P.; Kowalczyk, P. J.; Peterson, M. L.; Treasurywala, A. M., Conformational searching methods for small molecules. I. Study of the SYBYL Search Method. *Journal of Computational Chemistry* **1993,** 14, (9), 1050-1065

164. Kirkpatric, S.; Gelatt, C. D.; Vecchi, M. P., Optimization by simulated annealing. *Science* **1983,** 220, (4598), 671-680

165. Corzana, F.; Motawia, M. S.; Du Penhoat, C. H.; Perez, S.; Tschampel, S. M.; Woods, R. J.; Engelsen, S. B., A hydration study of (1/4) and (1/6) linked a-glucans by comparative 10 ns molecular dynamics simulations and 500-MHz NMR. *Journal of computational chemistry* **2004,** 25, (4), 573-586

166. Vishnyakov, A.; Widmalm, G.; Kowalewski, J.; Laaksonen, A., Molecular Dynamics Simulation of the [alpha]-d-Manp-(1 3)-[beta]-d-Glcp-OMe Disaccharide in Water and Water/DMSO Solution. *J. Am. Chem. Soc* **1999,** 121, (23), 5403-5412

167. Jackson, T. A.; Robertson, V.; Imberty, A.; Auzanneau, F. I., The flexibility of the LeaLex Tumor Associated Antigen central fragment studied by systematic and stochastic searches as well as dynamic simulations. *Bioorganic & Medicinal Chemistry* **2009,** 17, (4), 1514-1526

168. Naidoo, K. J.; Brady, J. W., The application of simulated annealing to the conformational analysis of disaccharides. *Chemical Physics* **1997,** 224, (2-3), 263-273

169. Groves, P.; Offermann, S.; Rasmussen, M. O.; Cañada, F. J.; Bono, J. J.; Driguez, H.; Imberty, A.; Jiménez-Barbero, J., The relative orientation of the lipid and carbohydrate moieties of lipochitooligosaccharides related to nodulation factors depends on lipid chain saturation. *Organic & Biomolecular Chemistry* **2005,** 3, (8), 1381-1386

170. Homans, S. W.; Forster, M., Application of restrained minimization, simulated annealing and molecular dynamics simulations for the conformational analysis of oligosaccharides. *Glycobiology* **1992,** 2, (2), 143-151

171. Kozar, T.; von der Lieth, C. W., Efficient modelling protocols for oligosaccharides: from vacuum to solvent. *Glycoconjugate journal* **1997,** 14, (8), 925-933

172. Orozco, M.; Alhambra, C.; Barril, X.; López, J. M.; Busquets, M. A.; Luque, F. J., Theoretical methods for the representation of solvent. *Journal of Molecular Modeling* **1996,** 2, (1), 1-15

173. Andersson, C.; Balling Engelsen, S., The mean hydration of carbohydrates as studied by normalized two-dimensional radial pair distributions. *Journal of Molecular Graphics and Modelling* **1999,** 17, (2), 101-105

174. Engelsen, S. B.; Monteiro, C.; Hervé de Penhoat, C.; Pérez, S., The diluted aqueous solvation of carbohydrates as inferred from molecular dynamics simulations and NMR spectroscopy. *Biophysical chemistry* **2001,** 93, (2), 103-127

175. Masukawa, K. M.; Kollman, P. A.; Kuntz, I. D., Investigation of neuraminidase-substrate recognition using molecular dynamics and free energy calculations. *J. Med. Chem* **2003,** 46, (26), 5628-5637

176. Abu Hammad, A. M.; Afifi, F. U.; Taha, M. O., Combining docking, scoring and molecular field analyses to probe influenza neuraminidase–ligand interactions. *Journal of Molecular Graphics and Modelling* **2007,** 26, (2), 443-456

177. Wang, N. X.; Zheng, J. J.; Inc, E. B.; Charities, A., Computational studies of H5N1 influenza virus resistance to oseltamivir. *Protein Sci* **2009,** 18, 707–715

178. Chachra, R.; Rizzo, R. C., 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

179. Du, Q. S.; Wang, S. Q.; Huang, R. B.; Zhang, D. W.; Chou, K. C., Insights from investigating the interaction of oseltamivir (Tamiflu) with neuraminidase of the 2009 H1N1 swine flu virus. *Biochemical and Biophysical Research Communications* **2009,** 386, (3), 432–436

180. Lawrenz, M.; Wereszczynski, J.; Amaro, R.; Walker, R.; Roitberg, A.; McCammon, J. A., Impact of calcium on N1 influenza neuraminidase dynamics and binding free energy. *Proteins: Structure, Function, and Bioinformatics* **2010,** in press

181. Ricard-Blum, S.; Feraud, O.; Lortat-Jacob, H.; Rencurosi, A.; Fukai, N.; Dkhissi, F.; Vittet, D.; Imberty, A.; Olsen, B. R.; van der Rest, M., Characterization of endostatin binding to heparin and heparan sulfate by surface plasmon resonance and molecular modeling: role of divalent cations. *Journal of Biological Chemistry* **2004,** 279, (4), 2927-2936

182. Verli, H.; Guimarães, J. A., Insights into the induced fit mechanism in antithrombin–heparin interaction using molecular dynamics simulations. *Journal of Molecular Graphics and Modelling* **2005,** 24, (3), 203-212

183. Wimmerová, M.; Mishra, N. K.; Pokorná, M.; Ko a, J., Importance of oligomerisation on Pseudomonas aeruginosaLectin-II binding affinity. In silico and in vitro mutagenesis. *Journal of Molecular Modeling* **2009,** 15, (6), 673-679

184. Mishra, N. K.; Kulhánek, P.; Snajdrová, L.; Petrek, M.; Imberty, A.; Koca, J., Molecular dynamics study of Pseudomonas aeruginosa lectin-II complexed with monosaccharides. *PROTEINS-NEW YORK-* **2008,** 72, (1), 382-392

185. Bowman, A. L.; Grant, I. M.; Mulholland, A. J., QM/MM simulations predict a covalent intermediate in the hen egg white lysozyme reaction with its natural substrate. *Chemical Communications* **2008,** 2008, (37), 4425-4427

186. Nimlos, M. R.; Matthews, J. F.; Crowley, M. F.; Walker, R. C.; Chukkapalli, G.; Brady, J. W.; Adney, W. S.; Cleary, J. M.; Zhong, L.; Himmel, M. E., Molecular modeling suggests induced fit of Family I carbohydrate-binding modules with a broken-chain cellulose surface. *Protein Engineering Design and Selection* **2007,** 20, (4), 179-187

187. Crowley, M. F.; Uberbacher, E. C.; Iii, C. L. B.; Walker, R. C.; Nimlos, M. R.; Himmel, M. E. In *Developing improved MD codes for understanding processive cellulases*, Journal of Physics: Conference Series, 2008; Institute of Physics Publishing: 2008; pp 12049-12012056

188. Blobaum, J. L. O. The role of Id2 (inhibitor of differentiation/ DNA binding) in dendritic cell development in steady state and inflammation. University of Hamburg, Hamburg (Germany), 2009

189. Bell, D.; Young, J. W.; Banchereau, J., Dendritic cells. *Advances in immunology* **1999,** 72, 255-305

190. de Witte, L.; Nabatov, A.; Pion, M.; Fluitsma, D.; de Jong, M.; de Gruijl, T.; Piguet, V.; van Kooyk, Y.; Geijtenbeek, T. B. H., Langerin is a natural barrier to HIV-1 transmission by Langerhans cells. *Nature Medicine* **2007,** 13, (3), 367-371

191. Valladeau, J.; Ravel, O.; Dezutter-Dambuyant, C.; Moore, K.; Kleijmeer, M.; Liu, Y.; Duvert-Frances, V.; Vincent, C.; Schmitt, D.; Davoust, J., Langerin, a novel C-type lectin specific to Langerhans cells, is an endocytic receptor that induces the formation of Birbeck granules. *Immunity* **2000,** 12, (1), 71-81

192. Birbeck, M. S.; Breathnach, A. S.; Everall, J. D., An electron microscope study of basal melanocytes and high-level clear cells (Langerhans cells) in vitiligo. *J Invest Dermatol* **1961,** 37, 51-64

193. McDermott, R.; Bausinger, H.; Fricker, D.; Spehner, D.; Proamer, F.; Lipsker, D.; Cazenave, J. P.; Goud, B.; de la Salle, H.; Salamero, J., Reproduction of Langerin/CD207 traffic and Birbeck granule formation in a human cell line model. *Journal of Investigative Dermatology* **2003,** 123, (1), 72-77

194. Stambach, N. S.; Taylor, M. E., Characterization of carbohydrate recognition by langerin, a C-type lectin of Langerhans cells. *Glycobiology* **2003,** 13, (5), 401-410

195. Feinberg, H.; Powlesland, A. S.; Taylor, M. E.; Weis, W. I., Trimeric Structure of Langerin. *The Journal of biological chemistry* **2010**, in press

196. Weis, W. I.; Drickamer, K., Structural basis of lectin-carbohydrate recognition. *Annual review of biochemistry* **1996,** 65, (1), 441-473

197. Weis, W. I.; Drickamer, K., Trimeric structure of a C-type mannose-binding protein. *Structure* **1994,** 2, (12), 1227-1240

198. Merad, M.; Ginhoux, F.; Collin, M., Origin, homeostasis and function of Langerhans cells and other langerin-expressing dendritic cells. *Nature Reviews Immunology* **2008,** 8, (12), 935-947

199. Fahrbach, K. M.; Barry, S. M.; Ayehunie, S.; Lamore, S.; Klausner, M.; Hope, T. J., Activated CD34-derived Langerhans cells mediate transinfection with human immunodeficiency virus. *The Journal of Virology* **2007,** 81, (13), 6858-6868

200. de Jong, M.; Vriend, L.; Theelen, B.; Taylor, M.; Fluitsma, D.; Boekhout, T.; Geijtenbeek, T., C-type lectin Langerin is a [beta]-glucan receptor on human Langerhans cells that recognizes opportunistic and pathogenic fungi. *J Molecular Immunology* **2010**, in press

201. Lyczak, J. B.; Cannon, C. L.; Pier, G. B., Establishment of Pseudomonas aeruginosa infection: lessons from a versatile opportunist*. *Microbes and infection* **2000,** 2, (9), 1051-1060

202. Collins, F. S., Cystic fibrosis: molecular biology and therapeutic implications. *Science* **1992,** 256, (5058), 774-779

203. Lyczak, J. B.; Cannon, C. L.; Pier, G. B., Lung infections associated with cystic fibrosis. *Clinical microbiology reviews* **2002,** 15, (2), 194-222

204. Blanchard, B. Etudes structurales et fonctionnelles de lectines et adhésines chez Pseudomonas aeruginosa. Joseph Fourier university Grenoble (France), 2009

205. Arora, S. K.; Ritchings, B. W.; Almira, E. C.; Lory, S.; Ramphal, R., The Pseudomonas aeruginosa flagellar cap protein, FliD, is responsible for mucin adhesion. *Infection and immunity* **1998,** 66, (3), 1000-1007

206. Gilboa-Garber, N., Lectins of Pseudomonas aeruginosa: Properties, biological effects, and applications. *Microbial Lectins and Agglutinins: Properties and Biological Activity* **1986**, 255–269

207. Vallet, I.; Olson, J. W.; Lory, S.; Lazdunski, A.; Filloux, A., The chaperone/usher pathways of Pseudomonas aeruginosa: identification of fimbrial gene clusters (cup) and their involvement in biofilm formation. *Proceedings of the National Academy of Sciences* **2001,** 98, (12), 6911-6916

208. Tielker, D.; Hacker, S.; Loris, R.; Strathmann, M.; Wingender, J.; Wilhelm, S.; Rosenau, F.; Jaeger, K. E., Pseudomonas aeruginosa lectin LecB is located in the outer membrane and is involved in biofilm formation. *Microbiology* **2005,** 151, (5), 1313-1323

209. Gilboa-Garber, N.; Mizrahi, L.; Garber, N., Purification of the galactose-binding hemagglutinin of Pseudomonas aeruginosa by affinity column chromatography using sepharose. *FEBS letters* **1972,** 28, (1), 93-95

210. Liu, Z. J., Tempel, W., Lin, D., Karaveg, K., Doyle, R.J., Rose, J.P., Wang, B.C. In *Structure determination of P. aeruginosa lectin-1 using single wavelength anomalous scattering data from native crystals*, Am.Cryst.Assoc.,Abstr. Papers (Annual Meeting), 29, 99, 2002; 2002.

211. Garber, N.; Guempel, U.; Belz, A.; Gilboa-Garber, N.; Doyle, R. J., On the specificity of the-galactose-binding lectin (PA-I) of Pseudomonas aeruginosa and its strong binding to hydrophobic derivatives of-galactose and thiogalactose. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1992,** 1116, (3), 331-333

212. Kirkeby, S.; Moe, D., Analyses of Pseudomonas aeruginosa Lectin Binding to -Galactosylated Glycans. *Current microbiology* **2005,** 50, (6), 309-313

213. Lanne, B.; Cîopraga, J.; Bergström, J.; Motas, C.; Karlsson, K. A., Binding of the galactose-specificPseudomonas aeruginosa lectin, PA-I, to glycosphingolipids and other glycoconjugates. *Glycoconjugate Journal* **1994,** 11, (4), 292-298

214. Gilboa-Garber, N.; Sudakevitz, D.; Sheffi, M.; Sela, R.; Levene, C., PA-I and PA-II lectin interactions with the ABO (H) and P blood group glycosphingolipid antigens may contribute to the broad spectrum adherence ofPseudomonas aeruginosa to human tissues in secondary infections. *Glycoconjugate Journal* **1994,** 11, (5), 414-417

215. Chen, C. P.; Song, S. C.; Gilboa-Garber, N.; Chang, K. S.; Wu, A. M., Studies on the binding site of the galactose-specific agglutinin PA-IL from Pseudomonas aeruginosa. *Glycobiology* **1998,** 8, (1), 7-15

216. Cioci, G.; Mitchell, E. P.; Gautier, C.; Wimmerová, M.; Sudakevitz, D.; Pérez, S.; Gilboa-Garber, N.; Imberty, A., Structural basis of calcium and galactose recognition by the lectin PA-IL of Pseudomonas aeruginosa. *FEBS letters* **2003,** 555, (2), 297-301

217. Blanchard, B.; Nurisso, A.; Hollville, E.; Tétaud, C.; Wiels, J.; Pokorná, M.; Wimmerová, M.; Varrot, A.; Imberty, A., Structural basis of the preferential binding for globo-series glycosphingolipids displayed by Pseudomonas aeruginosa lectin I. *Journal of molecular biology* **2008,** 383, (4), 837-853

218. Chemani, C.; Imberty, A.; De Bentzmann, S.; Pierre, M.; Wimmerova, M.; Guery, B. P.; Faure, K., Role of LecA and LecB lectins in Pseudomonas aeruginosa-induced lung injury and effect of carbohydrate ligands. *Infection and immunity* **2009,** 77, (5), 2065-2076

219. Winzer, K.; Falconer, C.; Garber, N. C.; Diggle, S. P.; Camara, M.; Williams, P., The Pseudomonas aeruginosa lectins PA-IL and PA-IIL are controlled by quorum sensing and by RpoS. *Journal of Bacteriology* **2000,** 182, (22), 6401-6012

220. Boteva, R. N.; Bogoeva, V. P.; Stoitsova, S. R., PA-I lectin from Pseudomonas aeruginosa binds acyl homoserine lactones. *Biochimica et Biophysica Acta (BBA)-Proteins & Proteomics* **2005,** 1747, (2), 143-149

221. Stoitsova, S. R.; Boteva, R. N.; Doyle, R. J., Binding of hydrophobic ligands by Pseudomonas aeruginosa PA-I lectin. *Biochimica et Biophysica Acta (BBA)-General Subjects* **2003,** 1619, (2), 213-219

222. Landry, R. M.; An, D.; Hupp, J. T.; Singh, P. K.; Parsek, M. R., Mucin-Pseudomonas aeruginosa interactions promote biofilm formation and antibiotic resistance. *Molecular microbiology* **2006,** 59, (1), 142-151

223. Rebiere-Huët, J.; Martino, P. D.; Hulen, C., Inhibition of Pseudomonas aeruginosa adhesion to fibronectin by PA-IL and monosaccharides: involvement of a lectin-like process. *Canadian journal of microbiology* **2004,** 50, (5), 303-312

224. Diggle, S. P.; Stacey, R. E.; Dodd, C.; Cámara, M.; Williams, P.; Winzer, K., The galactophilic lectin, LecA, contributes to biofilm development in Pseudomonas aeruginosa. *Environmental microbiology* **2006,** 8, (6), 1095-1104

225. Wu, H.; Song, Z.; Hentzer, M.; Andersen, J. B.; Molin, S.; Givskov, M.; Hoiby, N., Synthetic furanones inhibit quorum-sensing and enhance bacterial clearance in Pseudomonas aeruginosa lung infection in mice. *Journal of Antimicrobial Chemotherapy* **2004,** 53, (6), 1054-1065

226. Laughlin, R. S.; Musch, M. W.; Hollbrook, C. J.; Rocha, F. M.; Chang, E. B.; Alverdy, J. C., The key role of Pseudomonas aeruginosa PA-I lectin on experimental gut-derived sepsis. *Annals of surgery* **2000,** 232, (1), 133-142

227. Garber, N.; Guempel, U.; Gilboa-Garber, N.; Royle, R. J., Specificity of the fucose-binding lectin of Pseudomonas aeruginosa. *FEMS Microbiology Letters* **1987,** 48, (3), 331-334

228. Mitchell, E. P.; Sabin, C.; S najdrová, L.; Pokorná, M.; phanie Perret, S.; Imberty, A., 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. *Biol* **2002,** 9, 918-921

229. Loris, R.; Tielker, D.; Jaeger, K. E.; Wyns, L., Structural basis of carbohydrate recognition by the lectin LecB from Pseudomonas aeruginosa. *Journal of molecular biology* **2003,** 331, (4), 861-870

230. Perret, S.; Sabin, C.; Dumon, C.; Pokorná, M.; Gautier, C.; Galanina, O.; Ilia, S.; Bovin, N.; Nicaise, M.; Desmadril, M., Structural basis for the interaction between human milk oligosaccharides and the bacterial lectin PA-IIL of Pseudomonas aeruginosa. *Biochemical Journal* **2005,** 389, (2), 325-337

231. Mewe, M.; Tielker, D.; Schönberg, R.; Schachner, M.; Jaeger, K. E.; Schumacher, U., Pseudomonas aeruginosa lectins I and II and their interaction with human airway cilia. *The Journal of Laryngology and Otology* **2006,** 119, (08), 595-599

232. Letters, E., Interactions between soil and tree roots accelerate long-term soil carbon decomposition. *Ecology Letters* **2007,** 10, 1046-1053

233. Peoples, M. B.; Craswell, E. T., Biological nitrogen fixation: investments, expectations and actual contributions to agriculture. *Plant and Soil* **1992,** 141, (1), 13-39

234. Limpens, E.; Bisseling, T., Signaling in symbiosis. *Current opinion in plant biology* **2003,** 6, (4), 343-350

235. Geetanjali, N. G., Symbiotic nitrogen fixation in legume nodules: process and signaling. A review. *Agron. Sustain. Dev.* **2007,** 27, (1), 59 - 68

236. Rüdiger, H.; Gabius, H. J., Plant lectins: occurrence, biochemistry, functions and applications. *Glycoconjugate journal* **2001,** 18, (8), 589-613

237. Reli, B.; Perret, X.; Estrada-Garcia, M. T.; Kopcinska, J.; Golinowski, W.; Krishnan, H. B.; Pueppke, S. G.; Broughton, W. J., Nod factors of Rhizobium are a key to the legume door. *Molecular microbiology* **2006,** 13, (1), 171-179

238. Brewin, N. J., Plant cell wall remodelling in the Rhizobium–legume symbiosis. *Critical Reviews in Plant Sciences* **2004,** 23, (4), 293-316

239. Dénarié, J.; Debellé, F.; Promé, J. C., Rhizobium lipo-chitooligosaccharide nodulation factors: signaling molecules mediating recognition and morphogenesis. *Annual Review of Biochemistry* **1996,** 65, (1), 503-535

240. D'Haeze, W.; Holsters, M., Nod factor structures, responses, and perception during initiation of nodule development. *Glycobiology* **2002,** 12, (6), 79-104

241. Cullimore, J.; Lefebvre, B.; J.F., A.; A., B.; Bono, J. J.; Rouge, P.; Samain, E.; Driguez, H.; Imberty, A.; Untergasser, A.; Geurts, R.; Gadella, W. J.; Canada, J.; Barbero, J. J., *Biological Nitrogen fixation: Towards Poverty Alleviation*. Springer Science: 2008

242. Roche, P.; Debellé, F.; Maillet, F.; Lerouge, P.; Faucher, C.; Truchet, G.; Dénarié, J.; Promé, J. C., Molecular basis of symbiotic host specificity in Rhizobium meliloti: nodH and nodPQ genes encode the sulfation of lipo-oligosaccharide signals. *Cell* **1991,** 67, (6), 1131-1143

243. Ardourel, M.; Demont, N.; Debellé, F.; Maillet, F.; De Billy, F.; Promé, J. C.; Dénarié, J.; Truchet, G., Rhizobium meliloti lipooligosaccharide nodulation factors: different structural requirements for bacterial entry into target root hair cells and induction of plant symbiotic developmental responses. *The Plant Cell* **1994,** 6, (10), 1357-1374

244. Geurts, R.; Fedorova, E.; Bisseling, T., Nod factor signaling genes and their function in the early stages of Rhizobium infection. *Current opinion in plant biology* **2005,** 8, (4), 346-352

245. Limpens, E.; Franken, C.; Smit, P.; Willemse, J.; Bisseling, T.; Geurts, R., LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. *Science* **2003,** 302, (5645), 630-633

246. Radutoiu, S.; Madsen, L. H.; Madsen, E. B.; Felle, H. H.; Umehara, Y.; Grønlund, M.; Sato, S.; Nakamura, Y.; Tabata, S.; Sandal, N., Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature* **2003,** 425, (6958), 585-592

247. Arrighi, J. F.; Barre, A.; Ben Amor, B.; Bersoult, A.; Soriano, L. C.; Mirabella, R.; de Carvalho-Niebel, F.; Journet, E. P.; Gherardi, M.; Huguet, T., The Medicago truncatula lysine motif-receptor-like kinase gene family includes NFP and new nodule-expressed genes. *Plant physiology* **2006,** 142, (1), 265-279

248. Kaku, H.; Nishizawa, Y.; Ishii-Minami, N.; Akimoto-Tomiyama, C.; Dohmae, N.; Takio, K.; Minami, E.; Shibuya, N., Plant cells recognize chitin fragments for defense signaling through a plasma membrane receptor. *Proceedings of the National Academy of Sciences* **2006,** 103, (29), 11086-11097

249. Bateman, A.; Bycroft, M., The structure of a LysM domain from E. coli membrane-bound lytic murein transglycosylase D (MltD) 1. *Journal of Molecular Biology* **2000,** 299, (4), 1113-1119

250. Bielnicki, J.; Devedjiev, Y.; Derewenda, U.; Dauter, Z.; Joachimiak, A.; Derewenda, Z. S., B. subtilis ykuD protein at 2.0 Å resolution: Insights into the structure and function of a novel, ubiquitous family of bacterial enzymes. *Proteins* **2006,** 62, (1), 144-151

251. Mulder, L.; Lefebvre, B.; Cullimore, J.; Imberty, A., LysM domains of Medicago truncatula NFP protein involved in Nod factor perception. Glycosylation state, molecular modeling and docking of chitooligosaccharides and Nod factors. *Glycobiology* **2006,** 16, (9), 801-809

252. Radutoiu, S.; Madsen, L. H.; Madsen, E. B.; Jurkiewicz, A.; Fukai, E.; Quistgaard, E. M. H.; Albrektsen, A. S.; James, E. K.; Thirup, S.; Stougaard, J., LysM domains mediate lipochitin–oligosaccharide recognition and Nfr genes extend the symbiotic host range. *The EMBO Journal* **2007,** 26, (17), 3923-3935

253. Ohnuma, T.; Onaga, S.; Murata, K.; Taira, T.; Katoh, E., LysM Domains from Pteris ryukyuensis Chitinase-A. *Journal of Biological Chemistry* **2008,** 283, (8), 5178-5187

SECTION 5
ANNEXES

5 Annexes

5.1 Annex I – R scripts

Molecular dynamics calculations required the development of programs for particular analysis. The programs were developed and written in R (http://www.r-project.org/) for post-processing snapshots extracted with the Ptraj utility

5.2 Annex IA – Bridging water residence time

```
#
#Script for evaluating the residence time of bridging water molecules taking into
account 10000 snapshots from MDs simulation
#
#Type "water analysis" followed by the numbers associated to the atoms involved in a
potential water bridge
#
water_analysis<-function(n,nb)
{
matrice = matrix (0,10000,100)
cont=0
for (i in seq(1,50000,5))
{
cont=cont+1 #tiene il conto dei file
print (i)
#
#Introducing the path in which snapshots are stored
#
c = read.table (paste ("G:/RESIDENCE_TIME/GAL13GAL/snapshot.pdb.",i, sep=""))
d = as.numeric (c [n,6:8])
bkp =  c
z = length (c[,1])
e = sqrt ( (array (d[1],z) - c[,6])^2 + (array (d[2],z) - c[,7])^2 + (array (d[3],z)
- c[,8])^2 )
w = c[e < 3.6,]
matrice [ cont, 1: length (w[,2])] = w [,2]
}
k=sort(array(matrice[matrice>0]))
x = k [!duplicated (k)]
matrice2 = matrix (0, 10000, length(x))
cont=0
for (s in seq(1,50000,5))
{
cont=cont+1
print(s)
for (y in 1: length (x))
if (sum (x[y] == matrice [cont,])== 1)
matrice2 [cont, y] = 1}
sum  (matrice2 [,1])
presenze = colSums(matrice2)
name=array(NA,length(x))
name2=array(NA,length(x))
for (g in 1:length(x))
{name[g]=paste(c[c[,2]==x[g],4])
name2[g]=paste(c[c[,2]==x[g],3])
}
finale = data.frame (atomo = x, presenze = presenze,name=name,name2=name2)
finale=finale[(finale$name=="WAT")&(finale$name2=="O"),]
#
#Type "finale" if  you want to know the number of water molecules between the atoms
#and the number associated to each water molecule
#
matriceb = matrix (0,10000,100)
contb=0
for (ib in seq(1,50000,5) )
{
contb=contb+1 #tiene il conto dei file
print (ib)
cb = read.table (paste ("G:/RESIDENCE_TIME/GAL13GAL/snapshot.pdb.",ib, sep=""))
db = as.numeric (cb [nb,6:8]) #contiene le 3 coord dell'atomo n
bkpb =  cb  #non si sa mai, meglio una copia della matrice c
zb = length (cb[,1])  #dice quanti atomi ci sono in tutto il file
eb = sqrt ( (array (db[1],zb) - cb[,6])^2 + (array (db[2],zb) - cb[,7])^2 + (array
```

```
(db[3],zb) - cb[,8])^2 )
wb = cb[eb < 3.6,] # tiene a memoria i dati solo degli atomi piu vicini di 3.5
matriceb [ contb, 1: length (wb[,2])] = wb [,2]
}
kb=sort(array(matriceb[matriceb>0]))
xb = kb [!duplicated (kb)]
matrice2b = matrix (0, 10000, length(xb))
contb=0
for (sb in seq(1,50000,5))
{
contb=contb+1
print(sb)
for (yb in 1: length (xb))
if (sum (xb[yb] == matriceb [contb,])== 1)
matrice2b [contb, yb] = 1}
sum  (matrice2b [,1])
presenzeb = colSums(matrice2b)
nameb=array(NA,length(xb))
name2b=array(NA,length(xb))
for (gb in 1:length(xb))
{nameb[gb]=paste(cb[cb[,2]==xb[gb],4])
name2b[gb]=paste(cb[cb[,2]==xb[gb],3])
}
finale2b = data.frame (atomo = xb, presenze = presenzeb,name=nameb,name2=name2b)
finale2b=finale2b[(finale2b$name=="WAT")&(finale2b$name2=="O"),]
finaletot=c(finale$atomo,finale2b$atomo)
finaletot2=finaletot[duplicated(finaletot)]
tracking=matrix(NA,10000,length(finaletot2))
for (ici2 in 1:length(finaletot2))
{contc=0
a1=which(x==finaletot2[ici2])
a2=which(xb==finaletot2[ici2])
for (ici in seq(1,50000,5))
{contc=contc+1
tracking[contc,ici2]=matrice2[contc,a1]*matrice2b[contc,a2]
}
}
tracking[1:100,]
track<<-tracking
sommario=data.frame(atomo=finaletot2,presenze=colSums(tracking,na.rm=TRUE))
sommario2=data.frame(atomo=finaletot2,presenze_percentuali=colSums(tracking,na.rm=TR
UE)/length(sort(tracking[,1]))*100)
sommario3=data.frame(atomo="TOTALE",presenze=sum(sommario$presenze))
sommario4<<-rbind(sommario,sommario3)
print("Variabili in output: track e sommario4")
}
#
#Type "sommario4" for obtaining the residence time of the water molecules between
the two input atoms
#
```

5.3 Annex IB – Proton-proton average distances

```
#
#Average proton-proton interdistance along a simulation.10000 snapshots are taken
into account
#
matrice = matrix (0,10000,100)
cont=0
for (i in seq(1,50000,5) )
{
cont=cont+1 #tiene il conto dei file
print (i)
#
#Introducing the path in which the snapshots are stored
#
c = read.table (paste ("C:/Users/user/Desktop/Alessandra/snapshot.pdb.",i, sep=""),
skip=1)
d=c[(substring (c[,3],1,1) =="H")|(substring (c[,3],2,2) =="H"),]

e =(d[,6:8])
if (cont==1)
f = dist (e)
if (cont>1)
f = f + dist(e)
}
f = f/10000
#
#Introduce the distance cut-off for the calculation
#
g = which (f<4.5)
cont2=0
posx=0
posy=0
for (j in 1:(length(e[,1])-1)) #righe
for (i in (j+1) : length(e[,1])) #colonne
{cont2=cont2+1
posx[cont2]=i
posy[cont2]=j}
final=data.frame(atomo1=posx[g], nome1=d[posx[g],3], atomo2=posy[g],
nome2=d[posy[g],3], dist=f[g])
#
#Type "final" for obtaining a matrix in which the interdistance are stored
#
```

5.4 Annex IC – Conformation analysis of Nodulation factor lipid chains

```
#
#Script for the geometrical analysis of NF lipid chain.
#For this script is necessary to prepare a table in Excel that contains all the
dihedal angle values
#of the 5 dihedrals collected along the simulation (10000 snapshots were considered)

#
#Path in which the table is saved
#
nomefile="C:/Users/user/Desktop/NOD_FACTORS/table.txt"
x=read.table(nomefile)
y=x[seq(1,50000,5),]
z=y
#z[z<0]=z[z<0]+360
#
#Print in a eps file the dihedral angle distribution of each dihedral
#
postscript("C:/Users/user/Desktop/NOD_FACTORS/distribution.eps")
par(mfrow=c(3,2))
r=seq(-180,180,10)
a1=hist(z[,1], xlim=c(-180,180), freq=FALSE, ylim=c(0,0.03),breaks=r)
box()
a2=hist(z[,2], xlim=c(-180,180), freq=FALSE, ylim=c(0,0.03),breaks=r)
box()
a3=hist(z[,3], xlim=c(-180,180), freq=FALSE, ylim=c(0,0.03),breaks=r)
box()
a4=hist(z[,4], xlim=c(-180,180), freq=FALSE, ylim=c(0,0.03),breaks=r)
box()
a5=hist(z[,5], xlim=c(-180,180), freq=FALSE, ylim=c(0,0.03),breaks=r)
box()
dev.off()
#
#
#
peak_total=matrix(NA,5,3)
for (i in 1: 5)
{ peak=c(NA,NA,NA)
if (i ==1)
        {a=a1}
        if (i ==2)
        {a=a2}
        if (i ==3)
        {a=a3}
        if (i ==4)
        {a=a4}
        if (i ==5)
        {a=a5}
#
peak[1]=a$mids[which(a$density==max(a$density))]
b1=a$density
        maxint=max(b1)
if (peak[1]>60)
{b1[(a$mids>(peak[1]-60))&(a$mids<(peak[1]+60))]=0}
if (peak[1]<60)
{b1[(a$mids<(peak[1]+60))]=0
 b1[(a$mids>(300+peak[1]))]=0}
if (peak[1]>300)
{b1[(a$mids>(peak[1]-60))]=0
b1[(a$mids<(peak[1]-300))]=0}
#
peak[2]=NA
if (max(b1)>maxint*0.2)
{
```

```
peak[2]=a$mids[which(b1==max(b1))]
        maxint2=max(b1)
        if (peak[2]>60)
{b1[(a$mids>(peak[2]-60))&(a$mids<(peak[2]+60))]=0}
if (peak[2]<60)
{b1[(a$mids<(peak[2]+60))]=0
b1[(a$mids>(300+peak[2]))]=0}
if (peak[2]>300)
{b1[(a$mids>(peak[2]-60))]=0
b1[(a$mids<(peak[2]-300))]=0}
#
peak[3]=NA
        if (max(b1)>maxint2*0.2)
        {
        peak[3]=a$mids[which(b1==max(b1))]}}
        peak_total[i,]=peak
}
#
#If you type "peak_total" the peaks derived from the histrograms are translated in
numbers - check point
#
#The code starts to cluster the lipid conformations (ALL dihedral angle values are
connected)
#
totpaths=prod(length(sort(peak_total[1,]))*length(sort(peak_total[2,]))*length(sort(
peak_total[3,]))*length(sort(peak_total[4,]))*length(sort(peak_total[5,])))
#
paths=matrix(NA,totpaths,5)
cont=1
{for (i1 in 1:length(sort(peak_total[1,])))
        {
                for (i2 in 1:length(sort(peak_total[2,])))
                {
                        for (i3 in 1:length(sort(peak_total[3,])))
                        {
                                for (i4 in 1:length(sort(peak_total[4,])))
                                {
                                        for (i5 in 1:length(sort(peak_total[5,])))
                                        { print(paste(i1,i2,i3,i4,i5))
                                                paths[cont,1]=peak_total[1,i1]
                                                paths[cont,2]=peak_total[2,i2]
                                                paths[cont,3]=peak_total[3,i3]
                                                paths[cont,4]=peak_total[4,i4]
                                                paths[cont,5]=peak_total[5,i5]
                                                cont=cont+1
                                        }}}}}}
#
contapaths=matrix(0,10000,(totpaths))
sde=matrix(NA,10000,totpaths)
tolleranza=60
for (i in 1: 10000)
{print(i)
        #
        for (j in 1: (totpaths))
        {
                d=paths[j,]
                c=z[i,]
                #
                if (min(abs(c(c[1,1]-360,c[1,1],c[1,1]+360)-d[1]))<tolleranza)
                        if
(min(abs(c(c[1,2]-360,c[1,2],c[1,2]+360)-d[2]))<tolleranza)
                                if
(min(abs(c(c[1,3]-360,c[1,3],c[1,3]+360)-d[3]))<tolleranza)
```

lipids.txt

```
(min(abs(c(c[1,4]-360,c[1,4],c[1,4]+360)-d[4]))<tolleranza)
                                                 if
(min(abs(c(c[1,5]-360,c[1,5],c[1,5]+360)-d[5]))<tolleranza)
                                                 {

sde[i,j]=(min(abs(c(c[1,1]-360,c[1,1],c[1,1]+360)-d[1])))^2

sde[i,j]=sde[i,j]+(min(abs(c(c[1,2]-360,c[1,2],c[1,2]+360)-d[2])))^2

sde[i,j]=sde[i,j]+(min(abs(c(c[1,3]-360,c[1,3],c[1,3]+360)-d[3])))^2

sde[i,j]=sde[i,j]+(min(abs(c(c[1,4]-360,c[1,4],c[1,4]+360)-d[4])))^2

sde[i,j]=sde[i,j]+(min(abs(c(c[1,5]-360,c[1,5],c[1,5]+360)-d[5])))^2
                                                        contapaths[i,j]=1
}
                                            }}
for ( i in 1:10000)
{contapaths[i,]=0
 contapaths[i,][which(sde[i,]==min(sde[i,],na.rm=TRUE))]=1
 }
#
d6=colSums(contapaths)
#
#if you type "d6" you  can unerstand the frequency of conformers in a particular
class
#
d7=sort(d6,index.ret=TRUE)
d8=cumsum(rev(d7$x))/10000
d9=length(d8[d8<0.8])
d10=d7$ix[length(d7$ix):(length(d7$ix)-d9+1)]
d11=d7$x[length(d7$ix):(length(d7$ix)-d9+1)]
d12=d7$x/10000
#
#
#Graphical rapresentation of the 5 most visited conformations along the simulations
(eps file)
#
postscript("C:/Users/user/Desktop/LIPID/NR175/175.eps")
plot(1,1,cex=0,xlim=c(1,5),ylim=c(0,360))
for (i in 1: 5) #length(d10)
{points(paths[d10[i],]+i*5,col=rainbow (5)[i],lwd=3, type = "l")
}
dev.off()
#
#
#if you want to list all the families found in the search with the respective
dihedral values, type "paths"
#if you want to find the number corresponding to the most populated families, type
"d10"
#if you type "d12", the average frequency in % of the most populated families is
given
#
```

5.5 Annex II - Tutorials

During the course of my PhD I wrote easy tutorials for people coming to the laboratory for short periods of time who wanted to start modeling carbohydrate-protein interactions. The tutorials are available on line in the CERMAV website and are herein reported.

5.5.1 Annex IIA - MM3: calculation of simple and adiabatic maps

OBJET :

Ce mode opératoire a pour objet la description des étapes à suivre pour créer une carte PHI-PSI en utilisant le champ de force MM3

DOMAINE D'APPLICATION :

Ensemble du personnel de l'unité de recherche CERMAV

VOCABULAIRE :

Anglais

DIFFUSION :

Intranet qualité : SOURCE

DOCUMENTS DE REFERENCE :

PR-13 : Système d'information

Indice	Date	Auteur	Nature de la révision	
B	27/02/08	A. NURISSO	Modifications suite à réalisation de cartes (atom types)	
A	01/07/04	A. RIVET	Création	
			Vérification	Approbation
Signé par :				
Fonction :				
Date :				
Visa :				

The creation of a φ/ ψ maps follow four main steps

- Generating mm3 files
- Running the conformational search using mm3 force field
- Building the phi-psi maps
- Visualization of isocontour maps

All the files needed for this calculation are saved in the directory: /usr/people/rivet/DEV/carte_phi_psi. Copy and save all the files in a new directory in which all files will be saved.

GENERATING MM3 FILES

Use Sybyl to create a "*.mol2*" file of your carbohydrate (ex: galactose-galactose.mol2).
Save your file in your directory, that must contain the program "*mol2-mm3*".

In the Unix window, type "mol2-mm3" followed by the name of your file without the extension "*.mol2*"

> mol2-mm3 galactose-galactose

A file "*.xyz*" is generated and saved in your directory. Check this file and change its extension to obtain a ".mm3" file.

> mv galactose-galactose.xyz galactose-galactose.mm3
> nedit galactose-galactose.mm3

Warning: check the atom types! In the MM3 manual there is a table in which you could find the description of every MM3 atom type. Each one is identified by a number that always precedes the atom name in the ".mm3" file. For example "1" represents the carbon atom sp3, "6" the oxygen atom sp3, "5" the apolar hydrogen and "21" the polar one in hydroxyl functional groups.
Between each atom and its coordinates remember to type a letter: "C" for the carbon atoms, "H" for hydrogen atom, "O" for oxygen atom. You can find the correct letters to add in the MM3 manual.

RUNNING THE CONFORMATIONAL SEARCH

Check the presence in your directory of the following files:
- KONST.mm3,
- RDPARA,
- minifind,
- mm3-scan
- your ".mm3" file

And type

> mm3-scan

The program asks some information about your file, like the number of structures to scan (in this case 1), the name of your file (that you have to type without the ".mm3" extension), the glycosidic linkage atoms that you want to analyze (the program now requests the identification of the atoms that describe the torsion in terms of number. You can identify the numbers referred to the torsion angle opening the

stating mol2 file) and the scan step of the conformational search (e.i. 20°).

The program starts to run. If something weird happens, check the ".mm3" file, find the errors and run the program again.

BUILDING THE PHI-PSI MAPS

A home program called "uniphipsi" is used to build the maps and must be located in your directory. Type

```
> uniphipsi
```

and answer to the questions, knowing that you are working with "*.mm3*" files ("m") and energy files ("0" is the right answer referred to the number of atoms in your molecule) and that you want a simple energy map ("0" will be your choice in every question related to the map features).

You have to type the name of your energy file created in the previous step (ex: galactose-galactose.nrg) followed by the degree of scan step used (20), the limits for the psi and phi angles for a better visualisation of the energy minima in the map (0 – 360). Remember to use a high energy cut-off: the program will take into account more energy values related to each φ/ ψ couple.

Choose the output file typing ("x") and the name the output file, written with the right extension "*.xfb*" (ex: galactose-galactose.xfb) because, lately, you will use the program Xfarbe that requires the xfb extension. You can also write the title of your work that will be printed on your map and select the number of isocontours that will be displayed (I usually prefer to visualize 10 isocontours in the map).

If you are interested in knowing φ/ ψ values and the related calculated energies you can edit the "*.nrg"* file.

```
> nedit galactose-galactose.nrg
```

Once you have created a set of simple maps for a particular compound, you might want to obtain an adiabatic map. For this purpose, use the program mm3-scan and answer "1" to the question referred to the type of maps you want to calculate. Afterwards, type the number of simple maps to superpose and type the name of the energy files to superpose. The step increment is always required (20).At the end, type the name of the file that will contain a summary of the data extracted from your energy files and choose the name of the output file in a ".xfb" extension.

VISUALIZATION OF ISOCONTOUR MAPS

"xfb" program is in your directory: it allows the visualization of the results of your conformational search (simple or adiabatic map). Type

```
> xfb galactose-galactose.xfb
```

The isocontour plot is created and can be saved in an eps file. For more information about Xfarbe program, you can press the middle button of the mouse and Xfarbe menu will appear.

5.5.2 Annex IIB - POLYS: how to build complex oligosaccharides

OBJET :

Ce mode opératoire a pour objet d'apporter des informations supplémentaires au manuel POLYS
POLYS est utilisé par la constitution de polysaccharides

DOMAINE D'APPLICATION :

Ensemble du personnel de l'unité de recherche CERMAV

VOCABULAIRE :

Anglais

DIFFUSION :

Intranet qualité : SOURCE
Classeur Qualité du service informatique

DOCUMENTS DE REFERENCE :

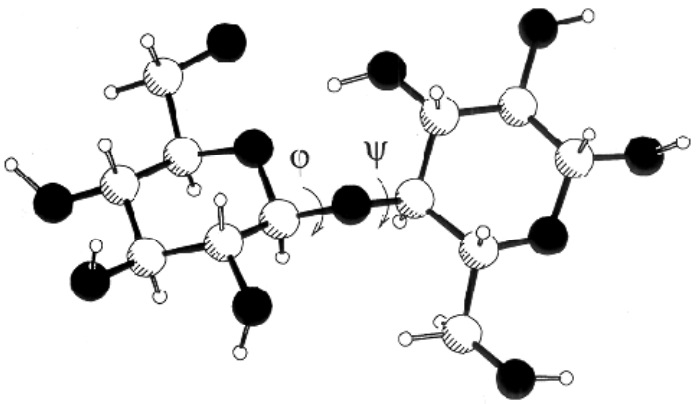
1) POLYS-Manual from SØren Balling Engelsen (1993/1995)
2) Engelsen, S.B., Cros, S., Mackie, W. and Perez, S. (1996) A Molecular Builder for Carbohydrates: Application to Polysaccharides and Complex Carbohydrates. Biopolymers, 39, 417-433.
3) Haxaire, K, Braccini, I, Milas, M, Rinaudo, M. and Pérez, S (2000) Conformational behavior of hyaluronan in relation to its physical properties as probed by molecular modeling. Glycobiology,10, 6, 587-594

Indice	Date	Auteur	Nature de la révision	
A	5/02/09	NURISSO DE CASTRO	Création	
			Vérification	Approbation
Signé par :				
Fonction :				
Date :				
Visa :				

This manual could be used in order to better understand the POLYS-Manual from SØren Balling Engelsen (1993/1995).
POLYS is a software created for building polysaccharides, complex carbohydrates (linear and branched ones) and also for generating and optimizing helical structures (also double, triple...helices). All polysaccharides can be built starting from monosaccharide-units listed in the mono bank (utils/SOFTW/polys/mono).

POLYS: CONVENTIONS&DEFINITIONS

Dihedral angles



φ : O5 – C1 – O1 - C_x
ψ : C1 – O1 – C_x – $C_{(x+1)}$
ω : O5 - C5 – C6 – O6
τ : C1 – O1 – C_x

For a linkage 1 →6

φ : O5 – C1 – O1 - C_x
ψ : C1 – O1 – C_x – $C_{(x-1)}$
ω : O1 – C6 – $C_{(x-1)}$ – $C_{(x-2)}$
τ : C1 – O1 – C_x

How to introduce monosaccharide units in order to build a polysaccharide

Ex. 1: [<sugar> (linkage; φ; ψ; ω)]r
Ex. 2: [<sugar> (linkage; φ; ψ; ω) <sugar>]r (linkage; φ; ψ; ω)
Ex. 3: [<sugar> (linkage; φ; ψ; ω) <sugar> (linkage; φ; ψ; ω)]r (linkage; φ; ψ; ω)

POLYS considers **Unit** all the residues in the square brackets and it will repeat their coordinates and geometries **r** times.

<sugar>: name of the file referred to a specific monosaccharide from the mono bank (type it without the extension .x)

linkage: definition of the type of linkage between two residues (ex. 1 : 2;1 : 3;1 : 4). Please, note the blank space

φ; ψ; ω: definition of the dihedral angles. The definition of ω is optional

Helical parameters

Doo or **"O - - - O":** oxygen-oxygen distance between residues per repeat

H or **h:** distance between two adjacent Units

Residues per repeat: 1/3 of the total number of sugar residues defined by **[unit] *r**

Residues per helical turn (n): number of **units** per helical turn

Helical rotation per repeat (na): degree of helical rotation per repeat (na*n=360°)

Helical repeat extension: glycosidic oxygen-oxygen distance between residues per repeat + 1

Helical repeat advancement: projection of the residues per repeat on the helix axis. This distance could be calculated according to the following equation: [H*residues per repeat]/number of residues per Unit

Helical fiber repeat (n*h): the definition of n*h is wrong. In order to obtain the helical fiber repeat we need to multiply the number reported here for the number of sugar residues that define the Unit.

EX.1: GENERATE A POLYSACCHARIDE

In this example a simple polysaccharide will be built. This polysaccharide is formed by 6 units of β-D-Glucopyranose with a linkage 1 →4 with phi and psi angles of 280.0 and 210.0 respectively.

- polys
- INIT #Starting the polys session
- SET title "molecule" #Give a title to your molecule (optional)
- PRIMARY # Generate the polysaccharide connectivity table
- [<bDGlcp> (1 : 4;280.0;210.0)]6
- STOP
- BUILD
- GENERATE internal #Generate and write in the output file the topology of the polysaccharide
- GENERATE internal rotbond
- WRITE internal rotbond
- WRITE coord PDB “molecule.pdb” #Generate a pdb file of the polysaccharide
- WRITE coord SYBYL “molecule.mol2” #Generate a mol2 file of the polysaccharide
- END

EX.2: GENERATE A POLYSACCHARIDE

In this example a polysaccharide will be built. This polysaccharide is formed by 6 disaccharidic units (Glcpβ 1→3 Galpβ) with phi and psi angles of 280.0 and 80.0 respectively.

- polys
- INIT #Starting the polys session
- SET title "molecule" #Give a title to your molecule (optional)
- PRIMARY #Generate the polysaccharide connectivity table
- [<bDGlcp> (1 : 3;280.0;80.0) <bDGalp>]6
- STOP
- BUILD
- GENERATE internal #Generate and write in the output file the topology of the polysaccharide
- GENERATE internal rotbond
- WRITE internal rotbond
- WRITE coord PDB “molecule.pdb” #Generate a pdb file of the polysaccharide
- WRITE coord SYBYL “molecule.mol2” #Generate a mol2 file of the polysaccharide
- END

EX.3: GENERATE A HELIX

In this example the hyaluronic acid will be build. Hyaluronan is constituted from the disaccharide repeating unit [4)-β-D-GlcpA-(1→3)-β-D-GlcpNAc-(1→) $_n$. The number of unit repeats must be at least 3. In order to obtain a reliable result, start using r = 3.

- polys
- INIT
- SET title "molecule" #Give a title to your molecule (optional)
- PRIMARY #Generate the polysaccharide connectivity table
- [<bDGlcpA> (1 : 3;-74.5;117.0) <bDGlcpNAc>]3 (1 : 4;-74.3;-117.9)
- STOP
- BUILD
- GENERATE internal #Generate and write in the output file the topology of the polysaccharide
- GENERATE internal rotbond
- WRITE internal rotbond
- HELIX # Calculation of helical parameters

The output of the program should look like this:

```
--------------------------------------------------------------------------------
POLYS:10 > HELIX
HELIX

Calculation of Helical Parameters

Note: 3 repeats must have been build
Note: A table of rotatable bonds must have been created

Main chain repeat segment: 3 --> 4  (offset = 0)

Input for helical calculations:
PHI[21]: O_5 - C_1 - O_1 - C_4  ==  -74.30
PSI[14]: C_1 - O_1 - C_4 - C_5  == -117.87
TAU:     C_1 - O_1 - C_4 ==  116.50

---HELICAL-PARAMETERS--------------------------
Residues per repeat                :    2
Residues per helical turn       (n):    3.18
Helical rotation per repeat    (na):  113.26
-----------------------------------------------
Helical repeat extension  (O- - -O):    9.79 A
Helical repeat advancement      (h):   -9.12 A
Helical fiber repeat          (n*h):  -14.50 A
-----------------------------------------------
```

- HELIX optimize nfold=X #Optimization of the torsion angles φ and ψ to obtain the nfold equal to #the defined X integer value (in this case X=3)

The output of the program should look like this:

```
--------------------------------------------------------------------------------
POLYS:11 > HELIX optimize nfold=3
HELIX optimize nfold=3

Calculation of Helical Parameters

        Note: 3 repeats must have been build
        Note: A table of rotatable bonds must have been created

Main chain repeat segment: 3 --> 4  (offset = 0)

Input for helical calculations:
        PHI[21]: O_5 - C_1 - O_1 - C_4  ==  -74.30
        PSI[14]: C_1 - O_1 - C_4 - C_5  == -117.87
        TAU:     C_1 - O_1 - C_4 ==  116.50

---HELICAL-PARAMETERS--------------------------
Residues per repeat                :    2
```

```
Residues per helical turn       (n):    3.18
Helical rotation per repeat    (na):  113.26
-----------------------------------------------
Helical repeat extension  (O- - -O):    9.79 A
Helical repeat advancement      (h):   -9.12 A
Helical fiber repeat          (n*h):  -14.50 A
-----------------------------------------------

Orthogonal Optimization of Helical Structure:

Search for nearest iso-(n=3.00000) contour
Relaxed main chain dihedrals:

        ROT[21]: C_1[69] - O_1[77] -->  -74.300
        ROT[14]: O_1[30] - C_4[52] --> -117.870
        ROT[15]: C_1[49] - O_1[55] -->  -74.500
        ROT[20]: O_1[55] - C_3[71] -->  117.020

---- 1-------- 2-------- 3-------- 4---- || -----doo----------H-------N-------
   -74.300  -117.870   -74.500   117.020 ||      9.794      -9.122   3.1784509
   -74.310  -117.877   -74.506   117.014 ||      9.794      -9.122   3.1780650
   -74.326  -117.887   -74.517   117.004 ||      9.794      -9.122   3.1774407
   -75.888  -118.946   -75.526   116.049 ||      9.813      -9.187   3.0839404
   -78.416  -120.658   -77.160   114.503 ||      9.843      -9.279   2.9432926
------------
   -77.371  -119.951   -76.485   115.142 ||      9.830      -9.240   3.0016521
   -77.371  -119.951   -76.485   115.142 ||      9.830      -9.240   3.0016521
   -77.371  -119.951   -76.485   115.142 ||      9.830      -9.240   3.0016521
   -77.371  -119.951   -76.485   115.142 ||      9.830      -9.240   3.0016521
   -77.371  -119.951   -76.485   115.142 ||      9.830      -9.240   3.0016521

Refined main chain vector:

        ROT[21]: C_1[69] - O_1[77] -->  -77.371
        ROT[14]: O_1[30] - C_4[52] --> -119.951
        ROT[15]: C_1[49] - O_1[55] -->  -76.485
        ROT[20]: O_1[55] - C_3[71] -->  115.142
--------------------------------------------------------------------------------
```

Please note that with the optimize command, polys will read the previous output of the HELIX command and it will optimizes the initial dihedrals to generate an helical structure that will fit with the requested nfold (in this case nfold=3)

- HELIX # Calculation of the new optimized helical parameters
- WRITE coord PDB “molecule.pdb” #Generate a pdb file of the polysaccharide
- WRITE coord SYBYL “molecule.mol2” #Generate a mol2 file of the polysaccharide
- END

To generate a long helical structure of the polysaccharide, run POLYS again, and set the optimized dihedral angles found in the previous run, in this case [(1 : 4;-77.371; -119.951) and (1 : 3;-76.485;115.142)]. HELIX and HELIX optimize commands are not necessary anymore.
To calculate “H” please refer to the value calculated previously (h=9.240).

5.5.3 Annex IIC - Autodock3: set up a docking run

OBJET :

Ce mode opératoire a pour objet la prise en main rapide d'Autodock version 3.05

Autodock est une suite d'exécutables et de scripts permettant d'obtenir des solutions de docking d'un ligand flexible sur une cible rigide. Autodock est particulièrement adapté à l'étude de petits ligands (comprenant 28 pivots au maximum) en interaction avec des protéines.

DOMAINE D'APPLICATION :

Ensemble de l'unité de recherche CERMAV

VOCABULAIRE :

Anglais

DIFFUSION :

Intranet qualité : SOURCE

DOCUMENTS DE REFERENCE :

1) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function Morris, GM, Goodsell, DSHalliday, RS, Huey, R, Hart, WE, Belew, RK, Olson, AJ. Journal of Computational Chemistry,19(14), 1639-1662, 1998
2) Techniques de Modélisation Moléculaire Appliquées à l'Etude et à l'Optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance, Fortune' A., 2006 (thèse – CERMAV)
3) Autodock3 Manual

Indice	Date	Auteur	Nature de la révision	
B	27/02/08	A. NURISSO	Modifications	
A	15/06/04	A. RIVET	Création	
			Vérification	Approbation
Signé par :				
Fonction :				
Date :				
Visa :				

1. Introduction

The program AutoDock was developed to provide an automated procedure for predicting the interaction of ligands with biomacromolecular targets.
A rapid energy evaluation is achieved by pre-calculating atomic affinity potentials for each atom type in the substrate molecule. In the AutoGrid procedure the protein is embedded in a three-dimensional grid and a probe atom is placed at each grid point. The energy of interaction of this single atom with the protein is assigned to the grid point. An affinity grid is calculated for each type of atom in the substrate, typically carbon, oxygen, nitrogen and hydrogen, as well as a grid of electrostatic potential. The time to perform an energy calculation using the grids is proportional only to the number of atoms in the substrate, and is independent of the number of atoms in the protein.
The docking simulation is carried out using one of a number of possible search methods. In each method the protein is static while the substrate molecule performs a random walk in the space around the protein, free to translate of its center of gravity, to rotate around each of its flexible internal dihedral angles.
The original AutoDock supported only one search method, the *Metropolis* method, also known as Monte Carlo simulated annealing. This version shows also a search method that is an implementation of a modified genetic algorithm which is called Lamarkian genetic algorithm.

2. Standard procedure

- Preparation of macromolecule
- Preparation of ligand
- Preparation of grid parameter file for Autogrid program
- Preparation of docking parameter file for Autodock program
- Running Autogrid
- Running Autodock
- Visualization and analysis of results

3. Detailed procedure

3.1 Preparation of macromolecule

Use Sybyl for this procedure. Load the structure and delete all hydrogens, ligand, water molecules and ions if they are present and not essential for the docking. The macromolecule first needs polar hydrogens to be added

Sybyl > Biopolymer > Prepare Structure > Add Hydrogens… Essential

Change the name of all polar hydrogens in “X” editing the file or using the Sybyl tool

Sybyl > Build/Edit > Modify > Atom > Name

Now add the apolar hydrogens

Sybyl > Biopolymer > Prepare Structure > Add Hydrogens… All

Assign the partial atomic charges to the macromolecule

Sybyl > Biopolymer > Prepare Structure > Load Charges > Biopolymer > Kollman All

Optimize the position of the hydrogen atoms creating an aggregate that includes all atoms of macromolecule without hydrogens and running a minimization using TRIPOS forcefield

Sybyl > Build/Edit > Aggregates… > New (All Difference > Atom Types > H)

Sybyl > Compute > Minimize > Ok

Save the protein in .mol2 format (macro.mol2), and then convert into PDBQ and PDBQS format

%> ***cnvmol2topdbq macro.mol2 > macro.pdbq***

%> ***addsol macro.pdbq macro.pdbqs***

3.1 Preparation of ligand

Add hydrogens to all atoms in the ligand, ensuring their valences are completed.
This can be done using Sybyl. Make sure that the atom types are correct and the position of the ligand is closed to the macromolecule but not superposed before adding hydrogens. Next, assign partial atomic charges to the molecule choosing the right method and save in mol2 format (ligand.mol2).

Sybyl > Compute > Charges …

Normally AutoDock considers ligands with just one type of hydrogen, namely polar hydrogens. Polar hydrogens can be defined here as those bonded to heteroatoms like nitrogen and oxygen, while non-polar hydrogens are bonded to carbon atoms.
If you want to model non-polar hydrogens as well, you would need a separate map in the next steps for such hydrogens. You could use the atom name code “h” for non-polar hydrogens, and “H” for polar hydrogens: edit the ligand.mol2 file and changing the atom name and type of hydrogens. Remember that hydrogen bonds are frequently important in ligand binding, so I suggest to take into account this distinction.
Create the ligand PDBQ file using deftors program in order to define any torsions that you want to be explored during the docking (Label the ligand with “Atom ID” or atom serial numbers in Sybyl. This will help in assigning the atoms):

%> **deftors ligand.mol2**

3.2 Preparation of the grid parameter file for Autogrid program

Create the GPF (grid parameter file)

%> **mkgpf3 ligand.pdbq macro.pdbqs**

Since you have two different hydrogen types in the ligand, you *must* specify the appropriate parameters modifying the AutoGrid parameter file. You have to use 12-6 Lennard-Jones parameters for non-polar hydrogens, 12-10 for polar hydrogens in order to distinguish them.
Pairwise atomic interaction energy parameters are always given in blocks of 7 lines, in the order: C, N, O, S, H, X, M. X and M are “spare” atom types: in this case X identifies the polar hydrogens of the macromolecule.
You can find here an example of .gpf file in which correct Lennard-Jones parameters are shown; in this case the ligand is formed by 6 different atom types: C, O, N, S, H, h. The macromolecule has also a fundamental calcium ion defined as M atom type.
Check your .gpf file, edit and modify it according to the values shown here (something is changed respecting the default gpf file. Remember that you have to add all information about the “h” atom type. Here you can find the appropriate Lennard-Jones parameters that I have calculated for you following the procedure described in the Autodock3 manual.).

```
receptor macro.pdbqs                #macromolecule
gridfld  macro.maps.fld                #grid_data_file
npts     60 60 60                 #num.grid points in xyz
spacing  .375                      #spacing (Angstroms)
gridcenter -1.095 -2.278 3.349           #xyz-coordinates or "auto"
types CONSHh                       #atom type names
smooth 0.500                             #store minimum energy within radius (Angstroms)
map macro.C.map                       #filename of grid map
nbp_r_eps  4.00 0.0222750 12  6          #C-C lj
nbp_r_eps  3.75 0.0230026 12  6          #C-N lj
nbp_r_eps  3.60 0.0257202 12  6          #C-O lj
nbp_r_eps  4.00 0.0257202 12  6          #C-S lj
nbp_r_eps  3.00 0.0081378 12  6          #C-H lj
nbp_r_eps  3.00 0.0081378 12  6          #C-X lj
nbp_r_eps  3.33 0.0385668 12  6          #C-M lj
sol_par 12.77 0.6844               #C atomic fragmental volume, solvation param.
constant 0.000                       #C grid map constant energy
map macro.O.map                       #filename of grid map
nbp_r_eps  3.60 0.0257202 12  6          #O-C lj
nbp_r_eps  3.35 0.0265667 12  6          #O-N lj
nbp_r_eps  3.20 0.0297000 12  6          #O-O lj
nbp_r_eps  3.60 0.0297000 12  6          #O-S lj
nbp_r_eps  2.60 0.0093555 12  6          #O-H lj
nbp_r_eps  1.90 0.3280000 12 10          #O-X hb
nbp_r_eps  2.93 0.0445332 12  6          #O-M lj
sol_par  0.00 0.0000               #O atomic fragmental volume, solvation param.
constant 0.236                       #O grid map constant energy
map macro.N.map                       #filename of grid map
nbp_r_eps  3.75 0.0230026 12  6          #N-C lj
nbp_r_eps  3.50 0.0237600 12  6          #N-N lj
nbp_r_eps  3.35 0.0265667 12  6          #N-O lj
nbp_r_eps  3.75 0.0265667 12  6          #N-S lj
nbp_r_eps  2.75 0.0084051 12  6          #N-H lj
nbp_r_eps  1.90 0.3280000 12 10          #N-X hb
nbp_r_eps  3.08 0.0398317 12  6          #N-M lj
sol_par  0.00 0.0000               #N atomic fragmental volume, solvation param.
constant 0.000                       #N grid map constant energy
map macro.S.map                       #filename of grid map
nbp_r_eps  4.00 0.0257202 12  6          #S-C lj
nbp_r_eps  3.75 0.0265667 12  6          #S-N lj
nbp_r_eps  3.60 0.0297000 12  6          #S-O lj
nbp_r_eps  4.00 0.0297000 12  6          #S-S lj
nbp_r_eps  3.00 0.0093555 12  6          #S-H lj
nbp_r_eps  2.50 0.0656000 12 10          #S-X hb
nbp_r_eps  3.33 0.0445335 12  6          #S-M lj
sol_par  0.00 0.0000               #S atomic fragmental volume, solvation param.
constant 0.000                       #S grid map constant energy
map macro.H.map                       #filename of grid map
```

```
nbp_r_eps  3.00 0.0081378 12  6            #H-C lj
nbp_r_eps  1.90 0.3280000 12 10            #H-N hb
nbp_r_eps  1.90 0.3280000 12 10            #H-O hb
nbp_r_eps  2.50 0.0656000 12 10            #H-S hb
nbp_r_eps  2.00 0.0029700 12  6            #H-H lj
nbp_r_eps  2.00 0.0029700 12  6            #H-X lj
nbp_r_eps  2.33 0.0140826 12  6            #H-M lj
sol_par  0.00 0.0000               #X atomic fragmental volume, solvation param.
constant 0.118                       #X grid map constant energy
map macro.h.map                       #filename of grid map
nbp_r_eps  3.00 0.0081378 12  6            #h-C lj
nbp_r_eps  2.75 0.0084645 12  6            #h-N lj
nbp_r_eps  3.00 0.0093555 12  6            #h-O lj
nbp_r_eps  2.60 0.0093555 12  6            #h-S lj
nbp_r_eps  2.00 0.0029700 12  6            #h-H lj
nbp_r_eps  2.00 0.0029700 12  6            #h-X lj
nbp_r_eps  2.33 0.0140826 12  6            #h-M lj
sol_par  0.00 0.0000               #h atomic fragmental volume, solvation param.
constant 0.118                       #h grid map constant energy
elecmap macro.e.map                #electrostatic potential map
dielectric -0.1146                 #<0,distance-dep.diel; >0,constant
#fmap macro.f.map                  #floating grid
# gpf3gen.awk 3.0.4 #
```

Create a PDB file from the grid parameter file in order to show how big and where the grid box will be when AutoGrid calculates the grid maps. You can use this “box molecule” to help you in refining the center (the center of the box is always the ligand by default) and the number of grid points in the grid maps editing the .gpf file, before you run AutoGrid.

%> mkbox macro.gpf >! macro.gpf.box.pdb

3.3 Preparation of the docking parameter file for Autogrid program

Create the DPF (docking parameter file)

%> mkdpf3 ligand.pdbq macro.pdbqs

Edit the docking parameter file and modify it according to the file shown here. In particular, check the “types” adding h atom type and modify the internal non bonded parameters.
You can also change the parameters for the conformational search method. In this example default parameters are shown. See the manual for major details.

```
seed    time pid # for random number generator
types CONSHh                              # atom type names
fld macro.maps.fld                        # grid data file
map macro.C.map                           # C-atomic affinity map file
map macro.O.map                           # O-atomic affinity map file
map macro.N.map                           # N-atomic affinity map file
map macro.S.map                           # S-atomic affinity map file
map macro.H.map                           # H-atomic affinity map file
map macro.h.map                           # h-atomic affinity map file
map macro.e.map                           # electrostatics map file

move ligand.pdbq                          # small molecule file
about -3.320 -4.902 6.267                 # small molecule center
```

```
# Initial Translation, Quaternion and Torsions
tran0 random                                # initial coordinates/A or "random"
quat0 random                                # initial quaternion or "random"
ndihe 12                                    # number of initial torsions
dihe0 random                                # initial torsions

torsdof 0 0.3113                            # num. non-Hydrogen torsional DOF & coeff.
#ligand_is_not_inhibitor   # uncomment if small molecule is substrate or T.S.

# Initial Translation, Quaternion and Torsion Step Sizes and Reduction Factors
tstep 2.0                                   # translation step/A
qstep 50.0                                  # quaternion step/deg
dstep 50.0                                  # torsion step/deg
trnrf 1.                                    # trans reduction factor/per cycle
quarf 1.                                    # quat reduction factor/per cycle
dihrf 1.                                    # tors reduction factor/per cycle

# Hard Torsion Constraints
#hardtorcon 1 -180. 30. # constrain torsion, num., angle(deg), range(deg)

# Internal Non-Bonded Parameters
intnbp_r_eps  4.00 0.0222750  12  6         #C-C lj
intnbp_r_eps  3.60 0.0257202  12  6         #C-O lj
intnbp_r_eps  3.75 0.0230026  12  6         #C-N lj
intnbp_r_eps  4.00 0.0257202  12  6         #C-S lj
intnbp_r_eps  3.00 0.0081378  12  6         #C-H lj
intnbp_r_eps  3.00 0.0081378  12  6         #C-h lj
intnbp_r_eps  3.20 0.0297000  12  6         #O-O lj
intnbp_r_eps  3.60 0.0297000  12  6         #O-S lj
intnbp_r_eps  1.90 0.3280000  12 10         #O-H hb
intnbp_r_eps  3.00 0.0093555  12  6         #O-h lj
intnbp_r_eps  3.50 0.0237600  12  6         #N-N lj
intnbp_r_eps  3.35 0.0265667  12  6         #N-O lj
intnbp_r_eps  3.75 0.0265667  12  6         #N-S lj
intnbp_r_eps  2.75 0.0084051  12  6         #N-h lj
intnbp_r_eps  1.90 0.3280000  12 10         #N-H hb
intnbp_r_eps  4.00 0.0297000  12  6         #S-S lj
intnbp_r_eps  2.50 0.0656000  12 10         #S-H hb
intnbp_r_eps  3.00 0.0093555  12  6         #S-h lj
intnbp_r_eps  2.00 0.0029700  12  6         #H-H lj
intnbp_r_eps  2.00 0.0029700  12  6         #H-h lj
intnbp_r_eps  2.00 0.0029700  12  6         #h-h lj

#intelec 0.1146                             # calculate internal electrostatic energy

# Simulated Annealing Parameters
#rt0 616.                                   # SA: initial RT
#rtrf 0.95                                  # SA: RT reduction factor/per cycle
#linear_schedule                           # SA: do not use geometric cooling
#runs 10                                    # SA: number of runs
#cycles 50                                  # SA: cycles
#accs 100                                   # SA: steps accepted
#rejs 100                                   # SA: steps rejected
#select m                                   # SA: minimum or last

# Trajectory Parameters (Simulated Annealing Only)
#trjfrq 100                                 # trajectory frequency
#trjbeg 1                                   # start trj output at cycle
#trjend 50                                  # end trj output at cycle
#trjout ligand.trj                          # trajectory file
#trjsel E                                   # A=acc only;E=either acc or rej

#watch ligand.watch.pdb                     # real-time monitoring file

outlev 1                                                # diagnostic output level
```

```
# Docked Conformation Clustering Parameters for "analysis" command
rmstol 1.0                                  # cluster tolerance (Angstroms)
rmsref ligand.pdbq                          # reference structure file for RMS calc.
#rmsnosym                                   # do no symmetry checking in RMS calc.
write_all                                   # write all conformations in a cluster

extnrg 1000.                                # external grid energy
e0max 0. 10000                              # max. allowable initial energy, max. num. retries


# Genetic Algorithm (GA) and Lamarckian Genetic Algorithm Parameters (LGA)
ga_pop_size 50                              # number of individuals in population
ga_num_evals 250000                         # maximum number of energy evaluations
ga_num_generations 27000                    # maximum number of generations
ga_elitism 1                                #  num.  of  top  individuals  that  automatically
survive
ga_mutation_rate 0.02                       # rate of gene mutation
ga_crossover_rate 0.80                      # rate of crossover
ga_window_size 10                           # num. of generations for picking worst individual
ga_cauchy_alpha 0                           # ~mean of Cauchy distribution for gene mutation
ga_cauchy_beta 1                            #  ~variance  of  Cauchy  distribution  for  gene
mutation
set_ga                                       # set the above parameters for GA or LGA

# Local Search (Solis & Wets) Parameters (for LS alone and for LGA)
sw_max_its 300 # number of iterations of Solis & Wets local search
sw_max_succ 4 # number of consecutive successes before changing rho
sw_max_fail 4 # number of consecutive failures before changing rho
sw_rho 1.0  # size of local search space to sample
sw_lb_rho 0.01                              # lower bound on rho
ls_search_freq 0.06 # probability of performing local search on an indiv.
set_psw1                                    # set the above pseudo-Solis & Wets parameters

# Perform Dockings
#do_local_only 50                           # do only local search
#do_global_only 10 # do only global search (traditional GA)

#simanneal  # do as many SA runs as set by the "runs" command above

ga_run 10                                   # do this many GA or LGA runs

# Perform Cluster Analysis
analysis                                    # do cluster analysis on results
```

Parameters usually changed for the improvement of a docking run are underlined.

3.4 Running Autogrid

Use AutoGrid to calculate the grid maps

%> autogrid3 -p macro.gpf -l macro.glg &

3.5 Running Autodock

Perform the dockings using AutoDock

%> autodock3 -p ligand.macro.dpf -l ligand.macro.dlg &

3.6 Visualization and analysis of results

At the end of a docking job in which more than one run was performed, the program outputs a histogram of clusters and their energies. Look in the ligand.macro.dlg file and search the word “HISTOGRAM” (see the example).

```
CLUSTERING HISTOGRAM
____________________

_____________________________________________________________________________
     |            |     |            |     |
Clus | Lowest     | Run | Mean       | Num | Histogram
-ter | Docked     |     | Docked     | in  |
Rank | Energy     |     | Energy     | Clus|    5    10   15   20   25   30   35
_____|____________|_____|____________|_____|____:____|____:____|____:____|____:___
   1    |                 -7.91    |       95    |                -7.57    |       50
########################################################
   2 |      -7.84 | 100 |      -7.70 |    4 |####
   3 |      -7.72 |  22 |      -7.29 |    8 |########
   4 |      -7.66 |  43 |      -7.66 |    1 |#
   5 |      -7.60 |  59 |      -7.60 |    1 |#
   6 |      -7.55 |  80 |      -7.31 |    3 |###
   7 |      -7.53 |  32 |      -7.38 |    3 |###
   8 |      -7.43 |  14 |      -7.28 |    2 |##
   9 |      -7.40 |  56 |      -7.26 |    2 |##
  10 |      -7.32 |  10 |      -7.32 |    1 |#
  11 |      -7.32 |  33 |      -7.17 |    4 |####
  12 |      -7.29 |  26 |      -7.29 |    1 |#
|_____|____________|_____|____________|_____|_____________________________________
```

The clustering of docked conformations is determined by the rms tolerance specified in Å by the “rmstol” keyword. The best conformation from each cluster (that with the lowest energy) is identified by a number (run column). To visualize docking results in a molecular modelling program, use “get-docked” command, to create a PDB formatted file. It will be called “ligand.macro.dlg.pdb” and will contain all the docked conformations output by AutoDock in the ligand.macro.dlg file.

%> get-docked ligand.macro.dlg

Sybyl is used for the visualization and analysis of docking results (Sybyl tutorial is also available on line).

3.7 Save time…

If you want to submit some jobs during the night you have to prepare a file (ex: commande.txt) in which the AutoDock command is the only thing written (ex: autodock3 -p ligand.macro.dpf -l ligand.macro.dlg &). Choose the time to run the job and type:

%> **at -f *commande.txt* 01:00 nov 27**

The job will run during the night, starting at 1:00 a.m.

5.5.4 Annex IID – MDs simulations of carbohydrate-protein interactions

OBJET :

Ce mode opératoire a pour objet l'utilisation du logiciel AMBER 10 en dynamique moléculaire

DOMAINE D'APPLICATION : Ensemble de l'unité de recherche CERMAV

VOCABULAIRE : anglais

DIFFUSION : Source

DOCUMENTS DE REFERENCE :

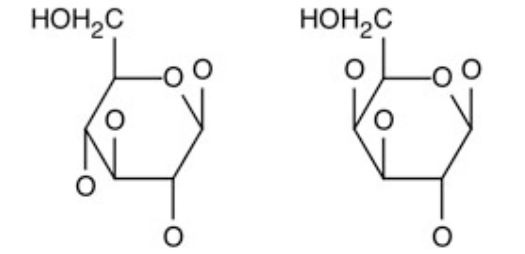
1) AMBER10 manual
2) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., Kollman, P. A., Journal of the American Chemical Society, 117(19) 5179-5197 – 1995

Indice	Date	Auteur	Nature de la révision	
A	20/08/09	NURISSO	Création	
			Vérification	Approbation
Signé par :				
Fonction :				
Date :				
Visa :				

AMBER 10: MOLECULAR DYNAMICS of a PROTEIN-CARBOHYDRATE COMPLEX

1) Introduction

The material for this tutorial was taken from the AMBER Glycomodelling workshop attended at Westminster University on April 2009 and modified in order to simplify the protocol.
The purpose of this tutorial is to setup and run a basic molecular dynamics system of a protein / carbohydrate complex taking into account the system lactose bound (Galactose + Glucose) to Galactin I.



a-D-glucose a-D-galactose

2) PDB Modifications

Downoad the pdb file (PDB: 1SLT). This system is actually a dimer. However, just one of the monomers will be simulated in order to keep the calculation expense manageable.
Extract just the A conformer of the Galectin (residues 2 to 134 Chain A) form the pdb and save this as **galectin_A_1slt.pdb.**
Extract the LactosNAC (The NDG and GAL residues) and save this as **lac_A_1slt.pdb**.

3) LAC Setup

For the LactosNAC, the GLYCAM06 force field will be used. This will require some modifications to the atom names in the pdb in order to match the naming used in the force field files.
Go to www.glycam.com and build the lactose with the carbohydrate builder. A pdb file, **glycam.pdb**, will be created, containing correct Glycam atom types and coordinates. At this point, transfer the coordinates of the file **lac_A_1slt.pdb** in the new **glycam.pdb** file. You can help yourself visualizing both molecules in Sybyl. Save the file as **lac_A_1slt_glycam_corrected.pdb**

The next step is to run this pdb file through tleap in order to create a pdb file with the protons added. We will also create prmtop and inpcrd files at the same time without water molecules.

mk_lac_1.lpcmd

```
# create lacnac amber input files from pdb file

source leaprc.GLYCAM_06
source leaprc.ff99SB

lacnac = loadpdb lac_A_1slt_glycam_corrected.pdb

# charge amber_seq
saveamberparm lacnac lacnac.prmtop lacnac.inpcrd
savepdb lacnac lac_A_1slt_glycamH.pdb

quit
```

> $AMBERHOME/exe/tleap -f mk_lac_1.lpcmd

This will produce the following files: **lac_A_1slt_glycamH.pdb, lacnac.prmtop, lacnac.inpcrd**

4) Galectin Setup

The next stage is to setup the Galectin protein itself.
The first few steps consist of cleaning up the pdb file to make it suitable for AMBER. We will start from the **galectin_A_1slt.pdb** we created above.

Step 1
There are several OCS (cysteine sulfonic acid) residues in this pdb file which are not part of the native form of the enzyme. These are the result of oxidation of cysteine residues. Hence we should change these to be the native Cysteine's before running any simulations. Grepping for OCS shows that these are residues 16, 88 and 130.

>grep OCS galectin_A_1slt.pdb

Change 'HETATM' to 'ATOM.
Change the residue names to CYS.
Delete atom entries OD1 and OD2 for those residues.

Step 2
Next we need to identify any disulphide bonds and modify the CYS residues as necessary.
Residues 16 and 88 are close enough to form a disulphide bond. Thus we need to change the names of these two residues from CYS to CYX in the pdb file since we do not want leap to protonate the sulphur atoms.
This gives the final modified pdb file: **galectin_A_1slt_mod.pdb**
Step 3

The final step is to create the input files for AMBER. We do this with leap.

mk_galectin_1.lpcmd

```
# create lacnac amber input files from pdb file
```

```
source leaprc.GLYCAM_06
source leaprc.ff99SB

glct = loadpdb galectin_A_1slt_mod.pdb

bond glct.16.SG glct.88.SG

saveamberparm  glct glct.prmtop glct.inpcrd
savepdb glct glct_H.pdb

quit
```

>$AMBERHOME/exe/tleap -f mk_galectin_1.lpcmd

This will produce the files: **glct.prmtop**, **glct.inpcrd**, **glct_H.pdb**

5) Complex Setup

We are now ready to build the complex.
The first step is to build a single pdb file containing the protein + ligand with a TER card between them. Final file: **glct_lac_0.pdb**
Next we run this through leap to produce both unsolvated and solvated topology and inpcrd files.

mk_cmplx.lpcmd

```
# create lacnac amber input files from pdb file

source leaprc.GLYCAM_06
source leaprc.ff99SB

glct = loadpdb glct_lac_0.pdb

bond glct.15.SG glct.87.SG

saveamberparm glct glct_lac.prmtop glct_lac.inpcrd
savepdb glct glct_lac.pdb

solvateoct glct TIP3PBOX 12.0 0.75
addions glct Na+ 0
saveamberparm glct glct_lac_wat.prmtop glct_lac_wat.inpcrd
savepdb glct glct_lac_wat.pdb

quit
```

>$AMBERHOME/exe/tleap -f mk_cmplx.lpcmd

Output files: **glct_lac.prmtop**, **glct_lac.inpcrd**, **glct_lac.pdb**, **glct_lac_wat.prmtop**,

glct_lac_wat.inpcrd, glct_lac_wat.pdb

6) Equilibrate Complex

The step 6 is to minimize heat and then equilibrate the complex in solution. We will do this in a number of stages consisting of some initial minimization with restraints, followed by a longer minimization. We will then heat the system over 20ps using NVT. We use NVT here instead of NPT to avoid problems with instabilities caused by inaccurate pressure calculation at low temperature. Then we will run NPT at 300K to equilibrate the density and then switch to a loosely coupled NVT ensemble to run 2ns of production MD.

Step 1 - Minimize with Restraints

We will initially carry out minimization with some weak restraints on the backbone of the protein and on the carbons of the carbohydrate.

min0.in

```
Minimization restraining Backbone
 &cntrl
   imin=1, maxcyc=200, ntb=1, cut=8,
   ntpr=5, ntr=1,
   restraint_wt=5.0, restraintmask='@C1,C2,C3,C4,C5,C6,CA,C,N'
 /
```

>$AMBERHOME/exe/sander -O -i min0.in -o min0.out -p glct_lac_wat.prmtop –c glct_lac_wat.inpcrd –ref glct_lac_wat.inpcrd -r min0.rst

Output files: **min0.out, min0.rst**

Step 2 - Minimize Entire System

Next we will minimize the entire system.

min1.in

```
Minimization restraining Backbone
 &cntrl
   imin=1, maxcyc=2000, ncyc=200, ntb=1, cut=8,
   ntpr=20, ntr=0,
 /
```

>$AMBERHOME/exe/sander -O -i min1.in -o min1.out -p glct_lac_wat.prmtop -c min0.rst -r min1.rst

Output files**: min1.out, min1.rst**

Step 3 - Heat System

The next step is to heat the system over approximately 20ps. We will do this at constant volume using

the Langevin thermostat and using the NMR weight restraint function to linearly ramp the target temperature over the first 15ps. During this time we will also keep the backbone weakly restrained since even with the minimization we run a risk of being in a high energy minimum which could distort the protein structure.

heat.in

```
Equilibrate restraining Backbone
 &cntrl
  imin=0, irest=0, ntx=1,
  ntb=1, cut=8.0,
  ntf=2, ntc=2,
  nstlim=10000, dt=0.002,
  ntpr=100, ntwx=200,
  ntr=1, restraintmask="@C,CA,N,O,H", restraint_wt=5,
  ntt=3, gamma_ln=1., temp0=300.0, tempi=0.0,
  nmropt=1,
 /
 &wt type='TEMP0', istep1=0, istep2=7500,
  value1=0.0, value2=300.0 /
 &wt type='TEMP0', istep1=7501, istep2=10000,
  value1=300.0, value2=300.0 /
 &wt type='END' /
```

>$AMBERHOME/exe/sander -O -i heat.in -o heat.out -p glct_lac_wat.prmtop -c min1.rst -ref min1.rst -r heat.rst -x heat.mdcrd bzip2 -v heat.mdcrd

Output files: **heat.out, heat.rst, heat.mdcrd.bz2**

Run the process_mdout.pl script (available on line) to look at the energies, temperature etc... Additionally you can calculate RMSD's with Ptraj to check the behavior of the dynamics.

Step 4 - Equilibrate System

The next step is to equilibrate the density of the system by running approximately 200ps of NPT dynamics.

equil0.in

```
Equilibrate density
 &cntrl
  imin=0, irest=1, ntx=5,
  ntb=2, ntp=1, cut=8.0,
  ntf=2, ntc=2,
  nstlim=100000, dt=0.002,
  ntpr=500, ntwx=500,
  ntr=0, ntt=3, gamma_ln=1., temp0=300.0, /
```

>$AMBERHOME/exe/sander -O -i equil0.in -o equil0.out -p glct_lac_wat.prmtop -c heat.rst -r equil0.rst -x equil0.mdcrd

Output files: **equil0.out**, **equil0.rst**, **equil0.mdcrd**

Before proceeding further we should check that the density has equilibrated correctly before we switch to the final run where we will switch back to NVT.

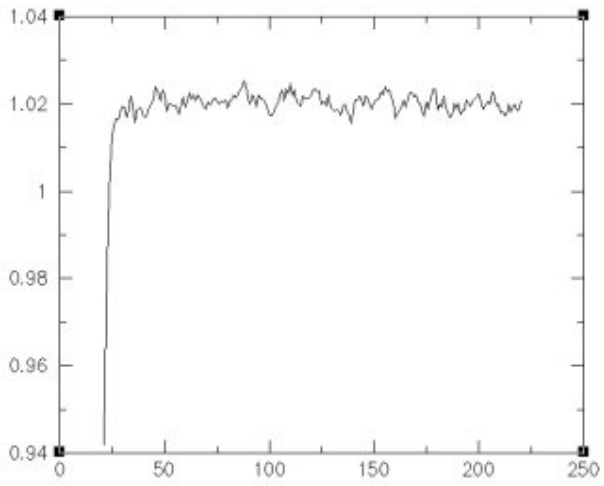
>mkdir analysis

>cd analysis

>chmod 777 process_mdout.pl

>process_mdout.pl equil0.out

>xmg DENSITY.dat



As you can see the density appears to have equilibrated implying that we are okay to proceed with the production run.

Step 5 - Production MD

The final stage is to run production MD. We will do this using the NVT ensemble but this time with the Berendsen thermostat and a very weak coupling constant of 10 ps. This provides minimal perturbation of the system similar to running in the NVE ensemble. We will run a total of 2 ns of MD which should be sufficient to fully equilibrate our system. We will also turn on wrapping (iwrap=1) to force sander to always write the coordinates of molecules within the central box. This negates the need for us to image later and avoids problems with water molecules diffusing beyond the coordinate limits of the restart file during long timescale MD. The resulting trajectory will be used as the basis for the calculations of free energy of binding.

prod_2ns.in

```
Production 2ns NVT
 &cntrl
   imin=0, irest=1, ntx=5,
   ntb=1, ntp=0, cut=8.0,
   ntf=2, ntc=2,
   nstlim=1000000, dt=0.002,
   ntpr=1000, ntwx=1000,
   ntr=0,
   ntt=1, tautp=10.0, temp0=300.0,
   iwrap=1,
 /
```

>$AMBERHOME/exe/sander -O -i prod_2ns.in -o prod_0.0-2.0ns.out -p glct_lac_wat.prmtop -c equil0.rst -r prod_2ns.rst -x prod_0.0-2.0ns.mdcrd

Output files: **prod_0.0-2.0ns.out**, **prod_0.0-2.0ns.mdcrd**, **prod_2ns.rst**

You can analyze the complex using PTRAJ module of AmberTools. A useful tutorial can be found on line (http://ambermd.org/tutorial/ptraj/)

5.6 Annex III – ARTICLE VII: *in silico* studies of hyaluronic acid

Conformational studies on hyaluronic acid where carried out using molecular dynamics simulations and NMR residual dipolar couplings. The 3.5 ns MD simulations in explicit water showed that hyaluronic acid adopts two arrangements which can be described by three- or four-folded left-handed helix matching the results found in previous studies. The importance of conformational studies and the informatics tools used for the accomplishment of this work have been described in detail (§2.3.2) .This work has been accomplished in collaboration with Dr. Cristina De Castro, from the University of Naples and J.J.Barbero's group, in Madrid.

Glycobiology Advance Access published April 29, 2010

Insights on the Conformational Properties of Hyaluronic Acid by using NMR Residual Dipolar Couplings and MD simulations

Valentina Gargiulo,[1] Maria A. Morando,[2] Alba Silipo,[1] Alessandra Nurisso,[3] Serge Pérez,[4] Anne Imberty,[3] F. Javier Cañada,[2] Michelangelo Parrilli,[1] Jesús Jiménez-Barbero,[2] Cristina De Castro[1]*

[1]Department of Organic Chemistry and Biochemistry, University of Napoli Federico II, Complesso Universitario Monte Sant'Angelo, Via Cintia 4, 80126 Napoli, Italy;

[2]Chemical and Physical Biology, Centro de Investigaciones Biológicas, C.S.I.C., Ramiro de Maeztu 9, 28040 Madrid, Spain;

[3]Centre de Recherche sur les Macromolécules Végétales –CERMAV CNRS-UPR5301 (affiliated with Grenoble Université and ICMG), 601 rue de la chimie, BP53, 38041 Grenoble cedex 9, France;

[4]ESRF, 6 rue Jules Horowitz, 38000 Grenoble, France

Corresponding author: tel. +39081674124, fax. +39081674393, e-mail: decastro@unina.it

Supplementary Data: Table S1, XCluster, Figure S1, Figure S2, Figure S3 and Figure S4, are available as Supplementary Data. This material is available online at http://glycob.oxfordjournals.org/.

Keywords: Molecular Dynamics / POLYS / RDC / three-folded helix / XCluster

Abstract.

The conformational features of hyaluronic acid, a key polysaccharide with important biological properties, have been determined through the combined used of NMR spectroscopy and molecular modeling techniques. A decasaccharide fragment of sodium hyaluronate was submitted to 3.5 ns of molecular dynamics in explicit water environment form. The same decasaccharide was prepared by hyaluronidase digestion for the experimental study. The approach consisted in the measurements of NMR Residual Dipolar Coupling which were used to filter the Molecular Dynamics data, by retaining those structures which were in agreement with the experimental observations. Further analysis of the new conformer ensemble (HA_{RDC}) and clustering the molecules with respect to their overall length led to seven representative structures, which were described in terms of their secondary motifs, namely the best fitting helix geometry. As result, this protocol permitted to assess that hyaluronic acid can adopt two different arrangements, which can be described by a three- or four-folded left-handed helix, with a higher occurrence of the first one.

Introduction

Hyaluronic acid (HA) is a linear anionic polysaccharide characterized by a disaccharide repeating unit [4)-D-β-GlcA-(1→3)-D-β-GlcNAc-(1→], and by a high molecular weight (10^5-10^7 Dalton). It belongs to the glycosaminoglycans family like chondroitin sulfate, dermatan sulfate, heparan sulfate, heparin, and keratan sulfate. Unlike these polysaccharides, HA is not sulfated and it is not linked to a core protein. It is present in the extracellular matrix (ECM) of higher animals but is also a significant component of some bacterial polysaccharides.

The original description dates 70 years ago, when HA was classified as an inert space-filler polysaccharide. However, the discovery of a large number of hyaluronan binding proteins (hyaloadherins) revealed that HA takes part in many biologically important processes (Day and Prestwich 2002). In the ECM at the surface of eukaryotic cells, HA interacts with a large array of cell-surface receptors and it plays a key role in the activation of various intracellular signaling cascades, of importance in both physiological and pathological conditions. These abilities are also related to its size: high-molecular weight hyaluronan is found in healthy tissues, while HA breakdown products signal that injury has occurred, and activate the recruitment of monocytes and lymphocytes in the wound site, and the expression of inflammatory cytokines (Stern et al. 2006).

Consequently, hyaluronic acid is a polysaccharide constantly under study, and many efforts have been dedicated to the description of its three-dimensional structure, a key parameter for the elucidation of its rheological properties (Hardingham 2004) or its interaction mechanism with hyaloadherins (Day and Mascarenhas 2004). The detailed structure of this macromolecule has been defined by crystallographic data on polysaccharide fibers and films (Sheehan and Atkins 1983). Moreover, the combined use of molecular modeling and NMR spectroscopy has also permitted to elucidate some conformational features (Almond et al 2006). The crystallographic analyses have revealed that HA oligomers exhibit a regular helical conformation whose features depend on the nature of the counterion, and on the pH;

generally, the found conformations can be depicted as left-handed helices, with three- or four-folded periodicity (Sheehan and Atkins 1983). Molecular modeling demonstrated that the interconversion between the different helix shapes only requires limited conformational modifications at the glycosidic linkages with low associated energy costs (Haxaire et al 2000). The NMR spectroscopy analysis in solution yielded to the same results, suggesting that the preferential conformation of the molecule was the four-folded helix (Almond et al 2006).

At this point, it must be evidenced that the NMR-based conformational analyses in solution performed so far have employed NOEs and scalar couplings to deduce the conformational information. Given the chemical nature of HA, a linear polysaccharide, these NMR parameters can only provide local structural information, at short range. NOEs sample interproton distances shorter than 4-5 Å, while scalar couplings provide torsion angle restraints only for groups of atoms separated by three bonds. Therefore, the direct information available only focuses on a small area of the compound under study.

In recent years, it has been proposed that this "classical" NOE/*J*-based approach can be complemented by measuring Residual Dipolar Couplings (RDC). This parameter can be measured by placing the molecule in a weakly orienting medium, field aligned, and permits to measure the orientations of the vectors joining two NMR-active nuclei with respect to the magnetic field. In this manner, global structural information throughout the molecule can be extracted (Lipsitz and Tjandra 2004), and used for conformational analysis as successfully reported for different molecules, including carbohydrates and even glycosaminoglycan oligosaccharides as dermatan (Silipo et al 2007) and heparin (Hricovíni et al 2009).

For water-soluble compounds, the most used aligning media are bicelles, micelles, bacteria phages, cellulose crystals and *n*-alkyl-poliPEG-alcohol mixtures (Yan and Zartler 2005). For all of them, the alignment effect is the result of steric and electrostatic interactions of the solutes with the medium, which hopefully do not modify their conformational behavior. When neutral media are used, the electrostatic contribution can be neglected, so that the alignment

reflects the asymmetries in the shape of the molecule and its tensor of inertia encodes the structural information (Azurmendi and Bush 2002).

In this work, RDC data for a hyaluronic acid decasaccharide (HA_{10}) have been collected in anisotropic conditions, using a neutral crystalline medium. The obtained RDC values have been used to restrain the number of conformer generated by MD simulations. The conformer ensembles matching the RDC measurements were analyzed through different tools (MSpin 2009, MacroModel XCluster 2009, and POLYS (Engelsen et al 1996)), leading to two different helical structures.

Results and Discussion

Experimental NMR data acquisition.- HA was depolymerized by enzymatic treatment (Tranchepain et al 2006) and the target decasaccharide (Figure 1) was isolated by chromatographic purification.

The NMR analysis of HA_{10} was first performed under isotropic conditions, via 1D and 2D NMR. The buffer used in the experiment is phosphate buffer (PB, 10 mM) resulting in the sodium hyaluronate form. The spectral analysis led to the complete assignment of the ^{1}H and ^{13}C chemical shifts (Table 1 in Supporting Information): signals belonging to the reducing (figure 1, unit **L**) and not reducing (figure 1, unit **A**) ends of the oligomer were distinct from those inside the molecule (units **B**-**I**), which collapsed in equivalent sets of resonances. The found values were in agreement with those reported for shorter oligomers (Blundell et al 2006). The analysis of the coupled gHSQC also led to the determination of the $^{1}J_{C,H}$ values for each carbon signal.

Then, in order to measure the RDC ($^{1}D_{C,H}$) values, a ternary mixture constituted by PB, pentaethylene glycol mono-octyl ether (C_8E_5), and *n*-octanol was employed as solvent for HA_{10} (Rückert and Otting 2000).

Estimation of the residual dipolar coupling carbon-proton constants ($^1D_{C,H}$) was performed through the measurement of the splitting of the corresponding ^{1}H-^{13}C correlations in the coupled-HSQC spectra recorded under both conditions, isotropic (PB) and anisotropic (the medium described above). Then, the $^1D_{C,H}$ constants (Table 1) were calculated as the difference between the measured splittings in the anisotropic ($^1J_{CH} + {}^1D_{CH}$) and isotropic ($^1J_{CH}$) conditions. In the anisotropic medium, GlcNAc *C*-6 residual dipolar coupling could not be evaluated due to the overlapping with the crystalline medium signals; similarly GlcA *C*-2 and GlcNAc *C*-5 signals partially overlapped, precluding the estimation of their $^1D_{CH}$ values. Therefore, $^1D_{C,H}$ were calculated for *C*-1, *C*-3, *C*-4 and *C*-5 of the GlcA units, and for *C*-1, *C*-2, *C*-3 and *C*-4 of the GlcNAc moieties (Table 1). As described below, these experimental values were compared with those estimated with MSpin (MSpin 2009) for the conformational ensemble obtained via MD simulations.

MD simulations.- With the use of Amber program (Case et al 2006) together with Glycam06e force field (Woods et al 1995), a 3.5 ns molecular dynamics trajectory was obtained for HA decasaccharide in explicit water in the presence of neutralizing sodium cations. A preliminary control of the quality of the MD data was performed by analyzing the energy parameters of the system (potential, kinetic and total energy, plus temperature, density and other physical data, as gathered in Figure S1, in supporting information) and the stability of the conformational ensemble could be checked.

The MD data were further validated by additional analysis. It was assessed that all the pyranose rings maintained the $^4C_1$ conformation, as indicated by the trajectories of the intraresidual H-1/H-3 and H-1/H-5 distances, which were constant around 2.7 and 2.5 Å, respectively. The amidic HN proton adopted always the typical *anti* orientation with respect to H-2 of GlcNAc. Similarly, for the *N*-acetyl moiety, the carbonyl group was always *anti* with respect to the amidic proton, whereas the carboxylate plane was almost parallel to the C-5/H-5 bond. The torsion angle values of all glycosidic linkages display limited variation around the

main energy minimum as defined previously by the energy map of each disaccharide linkage (Haxaire et al 2000). The Φ value is almost always in agreement with the *exo*-anomeric effect (Lemieux et al 1979), as indicated by the scattered plots (Fig 2 and figure S2 in supporting information). Only some brief visit to an anti-conformation are observed for the (1→3) linkage of residue I close to the reducing end disaccharide.

In this context, the averaged *inter*-residual distances between the protons around the glycosidic linkages were consistent with the experimentally observed NOEs (data not shown). All the above controls demonstrated the accuracy of the protocol and the employed force field. Thus, the conformation ensemble, referred as HA_{TOT}, was considered for further analysis.

RDC refinement of the conformer population obtained by Molecular Dynamics.- The MSpin software, through the application of the TRAMITE procedure (Azurmendi and Bush 2002), was used to select those MD-based conformers with expected RDC close to those experimentally measured. The comparison of the simulated RDC with those provided experimentally (Table 1) permitted to rate each molecule, as well as the total and partial MD ensembles, with a quality factor Q. Low Q values indicated a close agreement between the theoretical and experimental RDC values.

The whole MD simulation responded with an averaged Q factor of 0.58. The analysis showed the occurrence of conformers with expected RDC far from the experimental values (Q up to 0.96). Therefore, a Q value of 0.5 was selected and used as threshold to filter out those frames whose estimated RDCs were not consistent with the observed ones. The new Q factor, for the new MD ensemble dubbed HA_{RDC}, was 0.41. The HA_{RDC} ensemble contained 571 frames (33% of the original data). RDCs were simulated on this new pool of structures and a better agreement with the experimental values was found as expected (Table 1), finally the new ensemble, HA_{RDC}, was used for further analysis.

Comparison of the total and RDC-filtered dynamic data.-In order to understand the effect of the RDC filtering on the simulation data, HA_{TOT} and HA_{RDC} were compared, paying attention to the behavior of the Φ, Ψ dihedrals and to the head-to-tail distance (referred as DIS).

Considering HA_{TOT} first, for a given type of linkage, either 1→3 or 1→4, the Φ/Ψ scattered maps (figure 2 and S2) were similar and showed one main large distribution; additionally, the time-dependent trajectories (figure 3 and S3) showed that the behavior of each angle was not correlated with the other ones: actually, each Φ (or Ψ) moved within its allowed values, without persisting in any specific state, and with a time behavior apparently unrelated to that of the other dihedrals.

The head-to-tail distance (DIS) was considered as well. This value was estimated by measuring the distance between the anomeric oxygen of the reducing GlcNAc **L** (figure 1) and the oxygen at position four of the GlcA, at the non reducing terminus of the molecule (figure 1, unit A). Thus, the DIS trajectory (figure 4) showed that this geometry parameter spanned from 38 to 52 Å.

Then, the Φ/Ψ scattered graphics and trajectories from HA_{RDC} were superimposed with those from HA_{TOT} (figures 2, 3, S2 and S3); it can be observed that this graphical comparison did not reveal any particular difference among the two sets of data. Apparently, each RDC filtered dihedral distribution matched the original sets of values.

However, the effect of the RDC filter in terms of geometry changes appeared clearly when the DIS was considered (figure 4): in this case, the distance variation was restricted to a narrower range and varied among 44 and 52 Å. Oligomers with DIS < 44 Å never met the experimental selection criteria, whereas only a discrete number of those with DIS within the new range were kept. This dramatic selection reflected the fact that only specific combinations of the glycosidic torsions responded to the experimental RDC constraints.

Clustering of the RDC-filtered ensemble.-As described above, RDC filtering of HA_{TOT} permitted to reduce the number of conformers to be considered for further analysis to one third of the original ensemble. Nevertheless, additional simplification was achieved by clustering the HA_{RDC} ensemble with the XCluster program (MacroModel XCluster 2009). As described in the Experimental Section and in Supporting Information, the selection was guided by the Reordering Entropy graphic (figure 5) and the Maximum Size profile (figure 6). This protocol allowed the selection of cluster level 515, which contained 57 clusters (see methods section), seven of which represented the 70% of the RDC-allowed conformers. Inspection of the DIS variations within each cluster (Supporting Information, figure 4) showed the XCluster efficiency for the grouping procedure. The selection of this reduced (70%) set of conformers restricted DIS variation to a narrower range, from ca. 46 to 50 Å. Indeed, this procedure cut out the less populated species at the edges of the DIS distribution, and resulted in a fairly limited conformational freedom for HA_{10} head-to-tail distance range of only 4 Å around the mean value of 48 Å.

In a further step, the analysis was focused by just considering the averaged structure of each cluster, which was built by setting the glycosidic dihedral angles to the averaged values calculated for the particular cluster considered (Table 2).

Secondary structure analysis with POLYS.-In order to further compare the distinct geometrical features of each cluster, the helical conformational features of each averaged oligomer were calculated using the POLYS program (Engelsen et al 1996), which permits to evaluate the number of repeating units per helical repeat, *n*, and the axial rise, *h*.

The *n* values found before optimization were never integer numbers (Table 3), indicating that the corresponding conformation was in between a 3_2 (*n*=3, 2 is the number of residues of the repeating unit) and a four-folded (or 4_2) helix; *h* values (Table 3) were always negative disclosing the occurrence of left-handed structures.

This initial evaluation also pointed out that for increasing DIS values of the oligosaccharide, as in cls_61, the three-folded character of the helix was more pronounced, while the four-folded motif displayed the opposite trend. In a further step, the dihedral angles of each oligosaccharide were optimized by fitting the molecule to a regular three- or four-folded helix. Importantly, the optimized Φ/Ψ values (Table 3) fell within the dihedral range (for both the averaged value and the associated error) defined in Table 2, thus confirming that both the three- and four-folded helix geometries were possible. As extension of the previous indications pointed out above, the extended conformations of the oligomer adopted the three folded helix geometry (Table 3), whereas those with a more compact arrangement (with smaller DIS values), were accompanied by switching from the three- to a four-folded helix geometry.

Quantification of three- versus four-folded helical structures.-On the basis of the above results, both 3_2 and 4_2 helices occurred in the HA_{RDC} ensemble. Therefore, quantification of the relative participation of the two different geometries, referred as HA_{n3} and HA_{n4}, was performed using the superimposition utility of Maestro (Maestro 2009). In particular, the divergence of the coordinates between each conformer of the HA_{RDC} ensemble and the selected model, HA_{n3} or HA_{n4}, was expressed as a Root Mean Square Deviation (RMSD) value, with the smallest values associated to a higher degree of structural similarity to the three or four helix model. RMSD values were then visualized as shown in figure 7. The two curves displayed a different growing rate, with that for HA_{n3} being the slower. This information permitted to confirm that the best structural homology was that between the HA_{RDC} data and HA_{n3}.

Also, the number of conformers with the “best” similarity to the considered model (the three or four helix model) was also derived. Different values of RMSD were used as selection filter, and the number of the oligomers within a given limit is reported in table 4.

The employment of small RMSD values (< 0.6) selected relatively few conformers with a regular *n*-folded geometry, so it was not considered. On the other hand, the selection of higher RMSD values (as 1.0) permitted to define a more populated group, but more divergent from the model structure. Thus, an intermediate RMSD value of 0.75 was selected (figure 8) after visual inspection of the superimposition between a structure with a given RMSD value and its reference model. The 0.75 RMSD value represented a fair compromise, since the selected number of conformers was reasonable, while a good structural similarity with the model structure was maintained. As a result, the number of structures with a "pure" helical conformation within the RDC ensemble followed a HA_{n3}: HA_{n4} = 29 : 7 ratio, or an approximate 4 : 1 ratio, favoring the three-folded helix.

It should be noticed that the pure three-folded helical conformer accounted only for the 5.1% of the HA_{RDC} ensemble (Table 4). Nevertheless, this number is only apparently small, since it reflects the probability for nine glycosidic junctions (or eighteen dihedrals) to adopt, at the same time, the distinctive values for the three-folded helix structure.

In any case, the analysis carried out herein shows that there is a major population of conformers which conform with the experimental results, namely that possesses a pronounced three-folded helix character, or that within each molecule, most of the glycosidic linkages adopt torsion angle values close to a three-folded helix geometry, with a 4 : 1 proportion with respect to those characterizing the four-folded motif.

Conclusions - The conformation of hyaluronic acid has been studied through the combined used of NMR spectroscopic data and molecular modeling techniques.

Form the experimental side, RDC measurements have been selected as experimental restraints, since they provide global structure information and thus have the potential to characterize the overall shape of the molecules (Tjandra and Bax 1997). Parallel to the experimental NMR data acquisition, a hyaluronate decasaccharide was simulated in explicit water and the obtained MD data were filtered with the experimental restrains. The analysis of

the new conformer ensemble was simplified by clustering the obtained frames with respect to their overall length: this process led to seven representative structures, which were further analyzed and described in terms of their secondary motifs, and their ability to best fit certain helix geometry. As a result, hyaluronic acid can adopt two different arrangements described by a three- or four-folded left-handed helix: the first motif is more elongated and occurs more frequently than the second one.

Indeed, the conformation of hyaluronic acid can be described by considering two different geometries, represented by HA_{n3} and HA_{n4}. Both forms coexist in solution and interconvert into each other. The observation of the DIS trajectory from the HA_{RDC} data (figure 4) points out the continuous variation of the total length of the oligomer. The molecule spends more time in an extended conformation, adopting a major three-folded helix geometry. However, this motif is not kept continuously and is replaced by a four-folded motif, with a more compact and less elongated arrangement of the molecule.

These information are of paramount importance in the study and comprehension of hyaluronan – protein interactions and evidence how subtle variations in the dihedral angles are accompanied from a switch from one geometry to another, like from a three- to a four-folded helix.

On the light of this evidence, it is clear that the modulation of HA conformation allows it to accomplish the vast plethora of roles in which it is involved.

Material and Methods

Isolation of the decasaccharide $HA_{10.}$ -Commercial hyaluronic acid (Hyaluronic acid sodium salt from *Streptococcus equi*, Biochemica, 53747), was partially depolymerized using bovine testicular hyaluronidase (type I-S, Sigma, H3506), producing oligosaccharide of various length. The enzymatic digestion was performed using the conditions reported to obtain HA fragments with a molar mass ranging between 2000 and 5000 g/mol (Tranchepain et al 2006).

The polymer was dissolved in 0.25 M $NaNO_3$, 5 mM Na_2HPO_3, pH 4 to a final concentration of 5 mg/mL, and treated with Hyaluronidase (1 mg of enzyme for 10 mg of HA, at 37°C for 5 hours). The enzyme was then denatured, boiling the solution for 15 minutes.

The resulting mixture of oligosaccharides was then fractionated by SEC using a AcA202 ultrogel (115 x 1.5 cm, flow 0.2 mL/min, eluent: 0.25 M acetic acid and 0.28 M pyridine in water). The fractions collected between 640 and 680 minutes were recovered, reduced under-vacuum, and desalted on Sephacryl G-10 (95 x 1.5 cm, flow 0.23 mL/min, H_2O as eluent). MALDI analysis indicated a purity of 98% of the decasaccharide.

All the chromatographic separations were monitored online with a refractive index refractometer (K-2310 Knauer).

NMR spectroscopy in isotropic conditions.- The decasaccharide (15 mg) was dissolved in PB 10 mM, pH 7 in D_2O, and transferred into the NMR tube. ^1H and ^1H-^{13}C NMR experiments were performed on a Bruker DRX-600 spectrometer equipped with a reverse probe. Spectra were calibrated with respect to internal acetone (δ_H = 2.225 ppm; δ_C = 31.45 ppm) and recorded at 288 K in order to shift downfield the residual HOD peak and to observe the signal of the β-reducing GlcNAc end. For all the homonuclear spectra, experiments were measured with data sets of 2048 x 512 points, a mixing time of 200, and 120 ms was employed for NOESY, and TOCSY, respectively. Each data matrix was zero-filled in both dimensions to give a matrix of 4096 x 2048 points, and was resolution-enhanced in both dimensions by a shifted sine-bell function before Fourier transformation.

The HSQC experiment was measured using a data set of 2048 x 512 points, 16 scans were acquired for each t_1 value, and pulse sequence was optimized for a 140 Hz coupling constant. During processing, matrix was extended to 4096 x 1024 points by forward linear prediction extrapolation.

The coupled-HSQC was performed, and processed in the same conditions listed above, but 40 scans were acquired.

NMR spectroscopy in anisotropic conditions.- The saccharide was dissolved in a mixture of PB 10 mM, pH 7 in D_2O, pentaethylene glycol mono-octyl ether (C_8E_5), and *n*-octanol, prepared as reported (Rückert and Otting 2000).

This liquid crystalline system was selected because it is not charged, it is fairly insensitive to pH, poorly sensitive to salts, it displays a poor binding ability to macromolecules, and it is stable under a wide range of temperatures (0-40°C). In addition, the RDC splitting induced from the medium depends on the composition and not on the specific temperature used, as other employed systems.

After optimization for different ratios of the three components, the weight percentage for the ratio C_8E_5/water was of 3%, and the molar ratio of C_8E_5 to *n*-octanol was 0.87 w/w. Under these conditions, the resulting quadrupolar splitting of the mixture in the ^{2}H-NMR spectrum was of 14.8 Hz, while the ^{1}H spectrum kept narrow signals for the sugar resonances.

The coupled-HSQC experiments were performed using a data set of 8192 x 256 points acquiring 140 scans for each t_1 value. During processing, matrix was extended to 4096 x 1024 points by forward linear prediction extrapolation. For both the isotropic and anisotropic conditions, three coupled-HSQC experiments were performed. The obtained coupling constants were averaged. All NMR spectra were acquired, transformed, and analyzed with Topspin 2.1 program.

MSPIN refinement.- The RDC experimental data were analyzed, and the alignment tensor determined with the MSpin software 1.0.1. This program calculates the alignment tensor and the theoretical RDC value for both a single conformer or for a conformational ensemble. In this last case, the calculated RDCs represent the averaged values over the whole population. At the end of the process, each structure is rated with a number, dubbed the quality factor Q. Low Q values indicate a close agreement between the theoretical and the experimental values.

For a conformation ensemble, the program reports the averaged Q value. Thus, the 1750 molecular frames, obtained through a MD simulation in explicit solvent, were loaded together with the experimental RDC values in the format requested by the employed software. The computation algorithm selected for the RDC calculation was TRAMITE (Azurmendi and Bush 2002), which assumes that the vectors system of the alignment tensor has the same orientation as that of the inertia vector.

MD of the HA decasaccharide.- The decasaccharide was constructed using the building facility offered at the Glycam web page (Woods group (2005-2010)), which provides the final coordinates in the PDB format. Charges, bond, angle, and torsion paramenters were taken from the Glycam06 force field (Woods et al 1995). Carboxylate groups were set parallel to the H-5/C-5 bond, as suggested from the deposited X-Ray structures at the Protein Data Bank (as 2HYA or 2BVK). The molecule was then treated with the Xleap module of Amber9: the global charge of the system was neutralized by adding five sodium ions, and the whole molecule was placed within a 11 Å octahedral TIP3P water box (with 6859 solvent molecules, for more information the reader can refer to Amber manual).

System coordinates were then saved and used for the successive calculations using the Sander module implemented in Amber9, namely: solvent minimization with strong constrains to the solute (constant volume, 500 SD iterations followed by 500 PRCG iterations), minimization of the whole system (constant volume, 1500 SD and 1500 PRCG iterations), heating of the system up to 288 K with weak constraints to the solute (SHAKE protocol to the C-H bonds, constant volume, 100 ps) system equilibration at 288 K (constant pressure, 100 ps), and the final producing MD simulation at 288 K, at constant pressure, for a total time of 3.5 ns, and collecting 1750 frames. The last MD simulation was divided into 7 steps of 500 ps each. Each step was concatenated to the previous one and, at the end, all the frames were mounted together on the basis of their increasing simulation time. The simulation was performed using

periodic boundary conditions and applying the particle-mesh Ewald approach in order to introduce long-range electrostatic effects for a cutoff of 10 Å.

Conversion from Amber to Maestro coordinates was done by a home-made script.

XCluster.- The key information on the use of XCluster program, necessary to understand the selection criterion of the clustering level 515 are reported in the Supporting Information. Creation of the XCluster input file was performed through the Maestro graphical interface, but the calculation was performed by starting the program directly. This approach was necessary since Maestro facilitated the selection of the conformer family and the setting up of the different clustering options, but did not allowed the possibility to select the desired distance as criterion. The problem was overcome by creating a temporary input XCluster file with three atoms (those defining DIS and a third one), deleting the extra atom in the script with a text editor, and directly running the new script with XCluster. Finally, selection of the appropriate cluster level was performed by combining the information from the Reordering Entropy with those from the Maximum Size graphic (figures 5 and 6, respectively). Figure 6 represents the number of conformers present in the most populated cluster found at each clustering level. In the present case, it showed that the cls_levels after the entropy minimum, namely from 465 to 516, contained one main cluster with always the same number of conformers (127). The main difference between the different cls_levels (465-516) was related to the number of clusters contained (figure 6), that decreased from 107 to 56, respectively. As result of these analyses and in order to reduce the number of clusters to be analyzed, the cls_level 515 was selected. It contained 57 clusters, with the following ones more populated (number of conformers in parenthesis): cls_level_515_1 (127), cls_level_515_3 (92), cls_level_515_6 (61), cls_level_515_13 (42), cls_level_515_61 (31), cls_level_515_58 (25) and cls_level_515_20 (23). These seven clusters represented the 70% of the total HA_RDC conformers and for simplicity, they will be referred as cls_1, cls_3, etc.

POLYS.- In order to evaluate the best fitting helix for hyaluronan oligomers, the POLYS monosaccharide database was implemented with the Glycam coordinates of the single residues, GlcA and GlcNAc. For our target molecule, POLYS evaluated the best fitting helix conformation which was described through the parameters: *h*, the axial rise per repeating unit, and *n*, the number of unit per helical turn.

On the basis of the Φ/Ψ values of the repeating unit of the oligosaccharide, the program returned a preliminary *n* value, which indicated the number of repeating units per helical repeat. In this initial query, *n* index was never an integer number (Table 3), but it was comprised between 3 and 4, indicating that the oligosaccharide might fit both in a three- and a four-folded helix. In the successive optimization procedure, POLYS adjusted each dihedral in order to fit the overall conformation of the oligomer in a regular conformation, either three- or four-folded.

Superimposition analysis.-This last analysis was performed by considering two model structures, one with a well-defined three-folded helix geometry, named HA_{n3}, and HA_{n4}. The two oligomers were built using the dihedral parameters calculated from the more abundant cluster cls_1 (Table 3, similar and representative of cls_3 as well), and included in a Maestro project, together with the HA_{RDC} data ensemble. Then, the superimposition utility of Maestro was used to compare each model geometry to the HA_{RDC} group of structures. The atoms selected for the comparison were those defining the glycosidic junctions of each disaccharide pair (O_5, C_1, O_1, C_n and C_{n+1}). Then, the fit between the coordinates of one selected conformer and the model structure was expressed as Root Mean Square Deviation (RMSD), with small values indicating that the homology degree between the target and the model geometry conformers was fair.

Acknowledgments. CDC visiting at CIB was supported by a travel fellowship from the School of Sciences and Technology – University of Naples Federico II. The groups at Madrid

and Grenoble thank the MICINN of Spain (Grant CTQ2009–08536) and EU (MRTN2006–035546) for funding.

Abbreviations: Extra Cellular Matrix ECM, gradient Heteronuclear Single Quantum Correlation gHSQC, Hyaluronic Acid or hyaluronate HA, Nuclear Overhauser Effect NOE, Phosphate Buffer PB, Protein Data Bank PDB, Residual Dipolar Coupling RDC, Root Mean Square Deviation RMSD

References.

Almond A, deAngelis PL, Blundell CD. 2006. Hyaluronan: The Local Solution Conformation Determined by NMR and Computer Modeling is Close to a Contracted Left-handed 4-fold Helix. *J. Mol. Biol*. 358: 1256-1269.

Azurmendi HF, Bush A. 2002. Tracking Alignment from the Moment of Inertia Tensor (TRAMITE) of biomolecules in neutral dilute liquid crystal solutions. *J. Am. Chem. Soc*. 124: 2426-2427.

Blundell AD, Reed MAC, Almond A. 2006. Complete assignment of hyaluronan oligosaccharides up to hexasaccharides. *Carbohydr. Res*. 341: 2803-2815.

Case DA, Darden TA, Cheatham TE, Simmerling III CL, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, Roitberg A, Seabra G, Wong K.F, Paesani F, Wu X, Brozell S, Tsui V, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Beroza P, Mathews DH, Schafmeister C, Ross WS, Kollman PA. 2006, AMBER 9, University of California, San Francisco.

Day AJ, Prestwich GD. 2002. Hyaluronan-binding Proteins: Tying up the Giant. *J. Biol. Chem*. 277: 4585-4588.

Day RM, Mascarenhas MM. 2004 Signal Transduction Associated with Hyaluronan. In: Garg HG, Hales CA, editors. *Chemistry and Biology of Hyaluronan.* Amsterdam: Elsevier Ltd. p. 153-188.

Engelsen SB, Cros S, Mackie W, Perez S. 1996. Molecular Builder for Carbohydrates: Application to Polysaccharides and Complex Carbohydrates. *Biopolymers*, 39: 417-433.

Hardingham T. 2004. Solution properties of Hyaluronan. In: Garg HG, Hales CA, editors. *Chemistry and Biology of Hyaluronan.* Amsterdam: Elsevier Ltd. p. 1-19.

Haxaire K, Braccini I, Milas M, Rinaudo M, Pérez S. 2000. Conformational behavior of hyaluronan in relation to its physical properties as probed by molecular modeling. *Glycobiology*, 10: 587-594.

Jin L, Hricovíni M, Deakin JA, Lyon M, Uhrín D. 2009. Residual dipolar coupling investigation of a heparin tetrasaccharide confirms the limited effect of flexibility of the iduronic acid in the molecular shape of heparin. *Glycobiology*, 19: 1185-1196.

Lemieux RU, Koto S, Voisin D. 1979. The Exo-anomeric Effect. In: Szarek WA, Horton D, editors. *Anomeric Effect: Origin and Consequences.* Washington DC: Amer. Chem. Soc. ACS symposium Series Vol. 87, p. 17-29.

Lipsitz RS, Tjandra N. 2004. Residual dipolar couplings in NMR structure analysis. *Annu. Rev. Biophys. Biomol. Struct*. 33: 387-413.

MacroModel XCluster, version 9.7, Schrödinger, LLC, New York, NY, 2009.

Maestro, version 9.0, Schrödinger, LLC, New York, NY, 2009.

MSpin 1.0.1. program information at: http://mestrelab.com/Products/MSpin/Details.html, 2009

Rückert M, Otting G. 2000. Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *J. Am. Chem. Soc*. 122: 7793-7797.

Sheehan JK, Atkins EDT. 1983. X-ray fiber diffraction study of conformational changes in hyaluronate induced in the presence of sodium, potassium and calcium cations. *Int. J. Biol. Macromol*. 5: 215-221.

Silipo A, Zhang Z, Cañada FJ, Molinaro A, Linhardt RJ, Jiménez-Barbero J. 2007. Conformational Analysis of a Dermatan Sulfate-Derived Tetrasaccharide by NMR, Molecular Modeling, and Residual Dipolar Couplings. *ChemBioChem*, 9: 240-252.

Stern R, Asari AA, Sugahara KN. 2006. Hyaluronan fragments: an information-rich system. *Eur. J. Cell. Biol*. 85: 699-715.

Tjandra N, Bax A. 1997. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*. 278: 1111–1114.

Tranchepain F, Deschrevel B, Courel MN, Levasseur N, Le Cerf D, Loutelier-Bourhis C, Vincent JC. 2006. A complete set of hyaluronan fragments obtained from hydrolysis catalyzed by hyaluronidase: application to studies of hyaluronan mass distribution by simple HPLC devices. *Anal. Chem*. 348: 232-242.

Yan J, Zartler ER. 2005. Application of residual dipolar couplings in organic compounds. *Magn. Reson. Chem*. 43: 53-64.

Woods RJ, Dwek RA, Edge CJ, Fraser-Reid B.1995. Molecular mechanical and molecular dynamical simulations of glycoproteins and oligosaccharides. 1. GLYCAM_93 parameter development. *J. Phys. Chem*. 99: 3832-3846.

Woods group. (2005-2010) Glycam Web. Complex Carbohydrate Research Center, University of Georgia, Athens, GA. (http://www.glycam.com).

Legends to figures:

Figure 1: structure of HA decasaccharide, all residues are D and glycosidic linkages are β configured.

Figure 2: scattered graphics and projections of Φ,Ψ values of two representative glycosidic junctions of the hyaluronic decasaccharide. Dihedral angles from the total dynamic simulation (HA_{TOT}) are black, those from the HA_{RDC} ensemble are gray. Φ and Ψ are defined as O_5-C_1-O-C_n and C_1-O-C_n-C_{n+1}, respectively.

Figure 3: trajectories of selected Φ,Ψ values of two representative glycosidic junctions of the hyaluronic decasaccharide. Dihedral angles from HA_{TOT} are black, those from HA_{RDC} ensemble are grey. Φ and Ψ are defined as O_5-C_1-O-C_n and C_1-O-C_n-C_{n+1}, respectively. Abscissa unit is time (ps).

Figure 4: DIS trajectory HA_{TOT} (black) and from HA_{RDC} ensemble (gray).

Figure 5: reordering entropy graphic obtained analyzing HA_{RDC} ensemble with Xcluster. Clustering was performed setting DIS parameter as criterion.

Figure 6: Maximum Size graphic obtained analyzing RDC filtered conformer ensemble with Xcluster. Clustering was performed using DIS parameter as criterion.

Figure 7: overimposition of the RMSD values obtained comparing HA_{RDC} ensemble to HA_{n3} (black) or HA_{n4} (gray) models

Figure 8: overimposition of the model compound (cyan), HA_{n3} (left) or HA_{n4} (right), with three conformers from the RDC filtrated ensemble (orange), each with a different RMSD value with respect to the model compound.

Tables:

Table 1: Experimental RDC values calculated for the hyaluronic decasaccharide (structure in figure 1). $^1D_{CH}$ are calculated as difference between the splitting measured in anisotropic ($^1J_{C,H} + {}^1D_{C,H}$) and isotropic ($^1J_{C,H}$) conditions. The theoretical values are also given and were estimated by applying MSpin software on HA_{TOT} and HA_{RDC} conformational ensembles.

GlcA residue	**Atom pair**	**Exp. $^1D_{CH}$[a]**	**HA_{TOT} $^1D_{CH}$**	**HA_{RDC} $^1D_{CH}$**	**GlcNAc residue**	**Atom pair**	**Exp. $^1D_{CH}$[a]**	**HA_{TOT} $^1D_{CH}$**	**HA_{RDC} $^1D_{CH}$**
A[b]	C_1-H_1	6.2 ± 1.0	1.1	5.5	**B**	C_1-H_1	9.5 ± 2.0	10.9	12.6
	C_3-H_3	13.5 ± 0.9	3.1	8.6		C_2-H_2	13.0 ± 1.3	10.5	12.2
	C_4-H_4	13.5 ± 1.3	-0.6	5.2		C_3-H_3	15.8 ± 1.5	10.3	11.9
	C_5-H_5	6.4 ± 0.7	1.9	7.0		C_4-H_4	13.1 ± 1.3	10.3	12.1
C	C_1-H_1	6.2 ± 1.0	5.8	6.9	**D**	C_1-H_1	9.5 ± 2.0	10.5	12.3
	C_3-H_3	13.5 ± 0.9	8.2	9.9		C_2-H_2	13.0 ± 1.3	10.0	12.0
	C_4-H_4	13.5 ± 1.3	5.7	7.4		C_3-H_3	15.8 ± 1.5	9.6	11.6
	C_5-H_5	6.4 ± 0.7	7.1	8.7		C_4-H_4	13.1 ± 1.3	10.5	12.1
E	C_1-H_1	6.2 ± 1.0	6.0	7.4	**F**	C_1-H_1	9.5 ± 2.0	11.1	12.6

	C_3-H_3	13.5 ± 0.9	8.3	10.0		C_2-H_2	13.0 ± 1.3	10.7	12.2
	C_4-H_4	13.5 ± 1.3	5.6	7.3		C_3-H_3	15.8 ± 1.5	10.2	11.6
	C_5-H_5	6.4 ± 0.7	7.2	8.7		C_4-H_4	13.1 ± 1.3	10.9	12.4
G	C_1-H_1	6.2 ± 1.0	7.4	8.2	**H**	C_1-H_1	9.5 ± 2.0	9.2	11.7
	C_3-H_3	13.5 ± 0.9	9.1	10.7		C_2-H_2	13.0 ± 1.3	9.0	11.5
	C_4-H_4	13.5 ± 1.3	6.4	8.2		C_3-H_3	15.8 ± 1.5	8.7	11.3
	C_5-H_5	6.4 ± 0.7	8.2	9.6		C_4-H_4	13.1 ± 1.3	9.7	11.9
I	C_1-H_1	6.2 ± 1.0	6.9	8.5	**L**[c]	C_1-H_1	8.4 ± 2.0	6.8	10.5
	C_3-H_3	13.5 ± 0.9	8.2	10.3		C_3-H	13.0 ± 1.3	6.3	10.2
	C_4-H_4	13.5 ± 1.3	6.2	8.2		C_4-H_4	15.8 ± 1.5	6.1	9.9
	C_5-H_5	6.4 ± 0.7	7.6	9.3		C_5-H_5	13.1 ± 1.3	7.3	10.9

[a] GlcA (or GlcNAc) residues of the decasaccharide gave rise to equivalent NMR signals and $^1D_{CH}$ values, as well.

[b] non reducing end of the oligomer

[c] β reducing end of the oligosaccharide

Table 2. Analysis of the seven clusters obtained by analyzing the RDC filtered MD data with the XCluster program. The averaged Φ/Ψ values for each type of glycosidic junction and their RMSD values (italic) are reported together with the averaged DIS values and the number of conformers within each group.

	$\Phi_{1\rightarrow3}$	$\Psi_{1\rightarrow3}$	$\Phi_{1\rightarrow4}$	$\Psi_{1\rightarrow4}$	DIS	No. Conf.	No. Conf. %
Cls_13	-73,57	116,32	-74,07	-118,55	46,64	42	7,36
	10,04	*14,33*	*9,97*	*14,85*	*0,12*	--	--
Cls_20	-74,79	116,50	-73,38	-116,69	46,93	23	4,03
	10,07	*14,14*	*9,99*	*13,49*	*0,05*	--	--
Cls_1	-75,99	114,71	-73,78	-119,63	47,46	127	22,24
	11,20	*16,78*	*11,57*	*14,79*	*0,22*	--	--
Cls_3	-76,17	114,31	-73,92	-119,98	48,13	92	16,11
	10,77	*12,83*	*11,42*	*14,25*	*0,15*	--	--
Cls_6	-76,63	114,50	-74,83	-123,26	48,63	61	10,68
	10,18	*11,65*	*10,81*	*13,35*	*0,12*	--	--
Cls_58	-77,51	113,62	-74,62	-126,24	49,20	28	4,90
	11,19	*12,72*	*10,27*	*14,10*	*0,08*	--	--
Cls_61	-78,95	111,21	-74,38	-126,25	49,71	31	5,43
	12,17	*12,19*	*10,71*	*12,10*	*0,07*	--	--

Table 3: Calculated *n* values by POLYS for each cluster before the optimization process and the new values, together with the axial raise of the repeating unit *h*, obtained after minimization of both three- and four-folded helix.

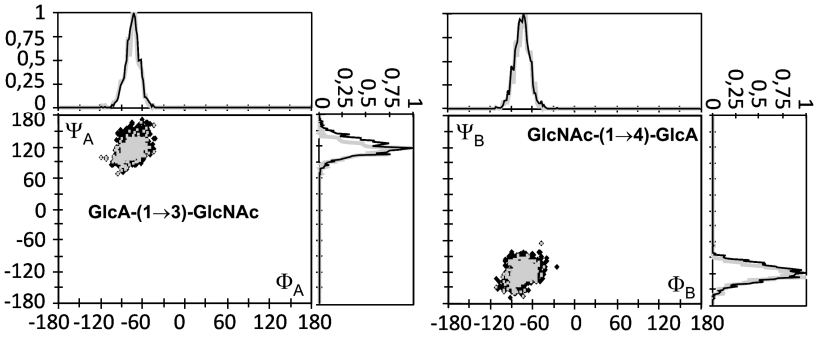
			Values obtained with *n* = 3					**Values obtained with *n* = 4**				
	n	***h***	h_3	$\Phi_{1\to3}$	$\Psi_{1\to3}$	$\Phi_{1\to4}$	$\Psi_{1\to4}$	h_4	$\Phi_{1\to3}$	$\Psi_{1\to3}$	$\Phi_{1\to4}$	$\Psi_{1\to4}$
Cls_13	3,44	-9.79	-9,95	-78,3	112,2	-80,1	-123,3	-9,53	-68,9	120,40	-68,2	-113,8
Cls_20	3,48	-9.77	-9,95	-79,9	112,0	-79,9	-121,8	-9,54	-70,5	120,2	-68,0	-112,4
Cls_1	3,33	-9.88	-9,99	79,6	111,5	-78,3	-123,3	-9,60	-70,3	119,8	-66,8	-113,8
Cls_3	3,31	-9.90	-10,00	-79,6	111,3	-78,1	-123,4	-9,61	-70.2	119.6	-66.6	-113.9
Cls_6	3,23	-9.95	-10,02	-79,2	112,2	-77,0	-125,9	-9,64	-69,8	120,6	-65,7	-116,4
Cls_58	3,11	-10.00	-10,03	-78,8	112,4	-76,1	-127,6	-9,67	-69,4	121,0	-65,1	-118,0
Cls_61	3,04	-10.06	-10,07	-79,4	110,8	-74,9	-126,8	-9,72	-70,0	119,4	-64,1	-117,2

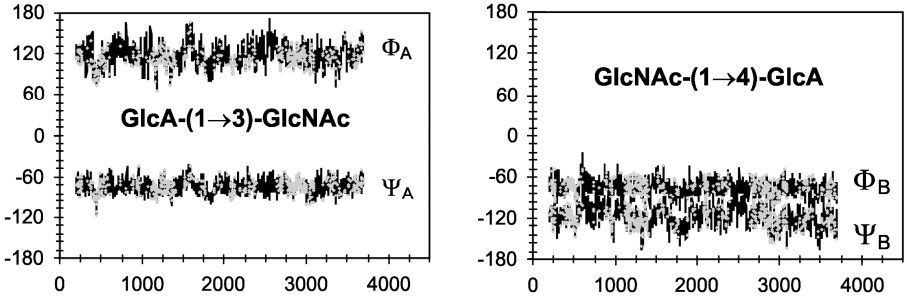
Table 4: Quantification of the population of conformers with respect to the three-fold or four-fold helix models and with respect to the different values of the RMSD filter. Percentages (in parenthesis) are calculated with respect to the HA_{RDC} MD ensemble.

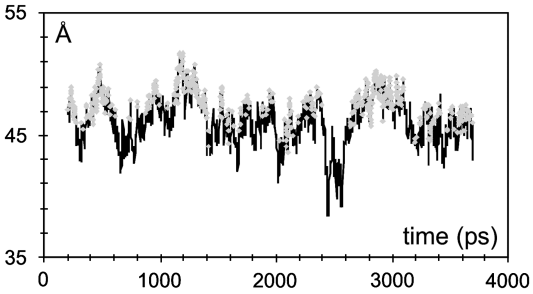
Model structure	**RMSD**		
	0.6	**0.75**	**1.0**
HA_{n3}	7	29 (5.1%)	143 (25%)
HA_{n4}	1	7 (1.2%)	56 (9.8%)

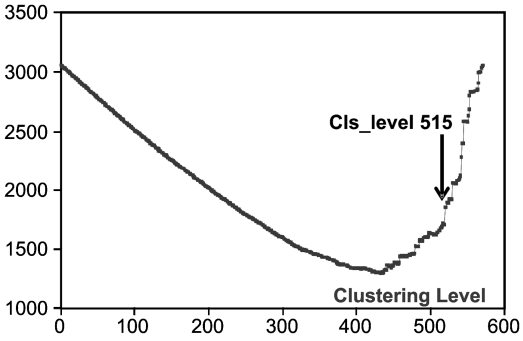
GlcA→3GlcNAc→4GlcA→3GlcNAc→4GlcA→3GlcNAc→4GlcA→3GlcNAc→4GlcA→3GlcNAc

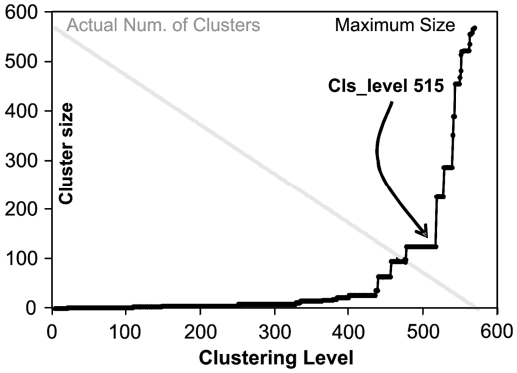
A B C D E F G H I L

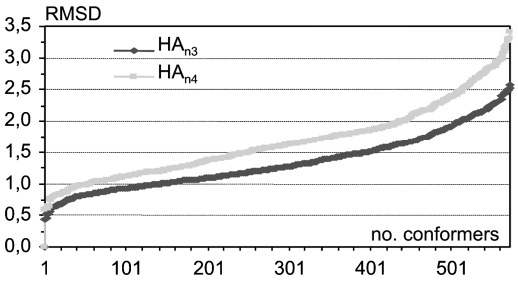


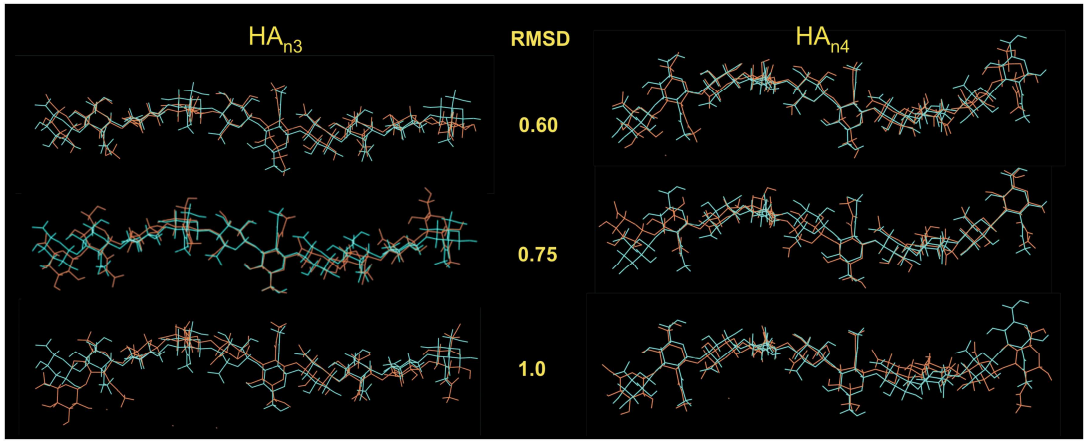



55
Å
45
35
time (ps)
0
1000
2000
3000
4000


3500
3000
2500
2000
1500
1000
Cls_level 515
Clustering Level
0
100
200
300
400
500
600


600
500
400
300
200
100
0
Actual Num. of Clusters
Maximum Size
Cls_level 515
Cluster size
0
100
200
300
400
500
600
Clustering Level


RMSD
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
HA$_{n3}$
HA$_{n4}$
no. conformers
1
101
201
301
401
501


HA$_{n3}$
RMSD
HA$_{n4}$
0.60
0.75
1.0