Computational Studies of Glycan Conformations in Glycoproteins

By

Sunhwan Jo

Submitted to the Department of Molecular Biosciences and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Chairperson: Wonpil Im

Roberto De Guzman

Scott Hefty

John Karanicolas

Krzysztof Kuczera

Mark Richter

Mario Rivera

Date defended: 04/12/2013

The Dissertation Committee for Sunhwan Jo certifies
that this is the approved version of the following dissertation :

Computational Studies of Glycan Conformations in Glycoproteins

Wonpil Im, Chairperson

Date approved: 04/12/2013

Abstract

N-glycans refer to oligosaccharide chains covalently attached to the side chain of asparagine (Asn) residues, and the majority of proteins synthesized in the endoplasmic reticulum (ER) are N-glycosylated. N-glycans can modulate the structural properties of proteins due to their close proximity to their parent proteins and their interactions between the glycan and the protein surface residues. In addition, N-glycans provide specific regions of recognition for cellular and molecular recognition. Despite their biological importance, the structural understanding of glycans and the impact of glycosylation to glycan or protein structure are lacking.

I have explored the conformational freedom of glycans and their conformational preferences in different environments using structural databases and computer simulations. First, I have developed an algorithm to reliably annotate a given atomic structure of glycans. This algorithm is important because many glycan molecules in the crystal structure database are misannotated or contain errors. Using the algorithm, a database of glycans found in the PDB is constructed and available to the public.

Second, the impact of glycosylation on the glycan conformation has been examined. Contrary to the common belief that the glycan conformations are independent to the protein structure, it appears that the protein structure can significantly affect the glycan structure upon glycosylation. This observation is significant because it may provide insight into protein-glycan interaction and opens up the possibility of a template-based glycan modeling approach.

Third, the differences in conformational preference between glycans in solution and in glycoproteins has been examined. Using molecular dynamics (MD) simulations, the conformational preference of N-glycan pentassacharide in solution is exhaustively studied. Surprisingly, the conformational distribution is dominated by a single major conformational state and several minor conformational states. The dominant conformational state adopts a more extended conformation, thus it appears that entropy plays an important role in determining the conformational state. On the other hand, in glycoproteins, glycans can interact with surrounding protein side chains and, as a result, several conformational states are more equally populated.

Based on these observations, a protocol is proposed for modeling the glycan portion of a known protein structure. It is typically more managable to acquire an atomic resolution structure or aglycoprotein (glycoprotein without glycan). In addition, the glycoform and the glycosylation site can be identified independently by mass spectrometry or NMR. The proposed modeling protocol assumes the glycosylation site, glycoform, and aglycoprotein structure are already known, and builds glycan structure models on top of the known aglycoprotein structure. The performance of the modeling protocol is greatly improved by using appropriate template structures. This protocol can be used to generate the initial model for MD simulations or refinement of low resolution models from experiments (small angle X-ray scattering and electron microscopy).

I dedicated this to my parents, Jungsoon Park and Byungryun Jo, and my wife, Sunyoung Kim.

Acknowledgements

First, I'd like to thank to my advisor Dr. Wonpil Im. For the last 7 years, under his guidance, I have learned how to do scientific research, and explore and enjoy the academic journey along the way. He provided more than enough freedom for me to do explore my own ideas, yet helped me focus on the goal through numerous and timely discussions. I deeply appreciate the support that he gave me over the years that I have spent in his lab.

I would also like to thank Dr. Roberto De Guzman for introducing me the exciting world of NMR and helping me to adopt in wet lab experiments. Through the exposure to the wet lab environment and handling the NMR machine, I was able to appreciate the amount of work that experimentalists have put into solve protein structures and to understand how proteins work. I appreciate the patience and encouragement he has shown regarding to my research.

Dr. Youngdo Won in Hanyang University, South Korea, first introduced me the theretical aspect of chemistry and allowed me to participate in his research while I was a undergraduate student. I fisrt learned about CHARMM, which is the software I used dilligently for my research, from him.

I am grateful to my defense committee, Dr. Scott Hefty, Dr. John Karanicolas, Dr. Krzysztof Kuczera, Dr. Mark Richter, and Dr. Mario Rivera for careful reading of this thesis and useful comments. I am also grateful to Dr. Edina Harsay for her useful comment and encouragement while I prepare for my pre-doctoral examination.

The work presented in here would not be possible without the help of numerous people from the lab and the people from other institution. I am very grateful to Dr. Heather Desaire for introducing me the world of glycobiology and Dr. Alex MacKerell Jr. for providing early access of the carbohydrate force field that was under development in his lab. I am also grateful to Dr. Jeffrey Klauda for his valuable comments and discussion on the structural analysis of glycan. It would be remiss not to mention Dr. Hui Sun Lee, Dr. Yifei Qi, Dr. Sryanta Mukherjee, Dr. Ambrish

Roy and talented undergraduate student (now in Chicago for his graduate study) Kevin Song for helping various tasks for the works presented in here.

During my undergraduate and graduate years in Lawrence, I have met many postdocs and students who helped my Ph.D. study directly or indirectly. I greatly appreciate Dr. Jinhyuk Lee, Dr. Taehoon Kim, Dr. Thenamalarchelvi Rathinavelan, Dr. Huan Rui, Dr. Kyuil Lee, Dr. Soohyung Park, Dr. Emilia Wu, Dr. Zhaowen Duan, Andrew Beaven, Dr. Jongcheol Jeong, George Li, Nathan Kern, Dr. Vidyashankara Iyer, Danielle Stuhlsatz, and Phillip Morris. I particularly thank Dr. Jinhyuk Lee, Dr. Taehoon Kim, and Dr. Soohyung Park for their helpful discussion to science and personal matter.

There are also many people who helped me on the NMR experiments and adjusting wet lab experience. It would have been very difficult for me to move forward my research without help of Dr. Dalian Zhong, Dr. Yu Wang, Dr. Fernando Estrada, Yan Xia, Srirupa Chatterjee, Sukanya Chaudhury, Kawaljit Kaur, Bryce Nordhues, and Dr. Asokan Anbandandam.

I am honored to have tremendous help from people in other institution as well. I am grateful to Dr. Richard Pastor, Dr. Jeffrey Klauda, Joseph Lim, Dr. Hyun-Suk Lim, Dr. Staley Opella, Dr. Francesca Marassi, Dr. Roger Koeppe, Dr. Taekyung Kwon, Dr. Thaddeus Bargiello, Dr. Benoit Roux, Dr. Wei Jiang, Dr. Kenno Vanommeslaeghe.

Finally, I would like to thank for my wife and my parents. None of the work that I have done would be possible without the countless support and tremendous encouragement from them. I've been with my wife, Sunyoung Kim, for last 10 years and the timeless affection and kindness she have shown to me is greatly appreciated. And special thanks to my sweet children, Heymin Jo and Alan Jo, and my sister in South Korea.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Biological roles of N-glycans and their biosynthesis

N-glycans refer to oligosaccharide chains covalently attached at the side chain of asparagine (Asn) residues in protein. The majority of proteins synthesized in the endoplasmic reticulum (ER) are N-glycosylated [7, 90, 136]. Once attached to a protein side chain, the roles of N-glycans are two-fold in general. First, the N-glycan structure can modulate the structural properties of a protein due to the close proximity and the interaction between the N-glycan and the protein. The protein secondary structure is expected to remain largely independent of the presence of N-glycans, but the glycosylation can affect the conformational preference of the peptide backbone and the rate of folding [21, 23, 44, 57, 77, 145]. Recent crystallographic studies revealed that the different glycoforms (N-linked oligosaccharides having different sequences) can also affect the overall shape of a protein complex and the binding affinity of the glycoprotein to its partner [31, 71, 89].

The other general biological role of N-glycans is to provide specific regions of recognition. The N-glycans on the surface of glycoproteins act as a “barcode” for the glycoprotein and allow other proteins to recognize the glycoprotein regardless of the sequence or the structure of the parent protein [1, 48]. The ability to recognize a variety of proteins based on N-glycans is essential to establish and maintain the subcellular localization of proteins in the higher organisms [121]. These

N-glycan "barcodes" are not only used in intracellular recognition but also used as specific ligands for cell-cell interactions. N-glycans are highly involved in intrinsic (recognizing glycans from same organisms or cell types) as well as extrinsic recognitions (regognizing glycans from different organisms or cell types), which are important in organ transplantation [35], cancer progression [25, 33, 76], immune response [117], host-pathogen interaction [2, 125], vaccine development [8, 47], and maintaining symbiotic relationships [136].

The N-glycosylation pathway in eukaryotes is well conserved across organisms and composed of two distinctive phases: N-glycan precursor synthesis and N-glycosylation [1, 48, 121, 136]. First, a glycan is assembled into a lipid-linked oligosaccharide (LLO) in step-wise addition of specific monosaccharides by various glycosyltransferases. The initial assembly of glycans onto a lipid carrier (dolichol-pyrophosphate; Dol-P) occurs in the cytosolic side of the ER membrane. Seven sugars are added to the lipid carrier (Dol-P-P-GlcNAc$_2$Man$_5$) before the complex is re-oriented to the luminal side of ER membrane by LLO specific flippase (see Figure 1.1). More sugars are added to the LLO molecule on the luminal side of ER to form the N-glycan precursor molecule (Dol-P-P-GlcNAc$_2$Man$_9$Glc$_3$). Then, when a peptide having a consensus sequence (Asn-X-Ser/Thr; X being any amino acid except proline) is synthesized and translocated into the ER lumen, oligosaccharyltransferase (OST) catalyzes *en bloc* transfer of the oligosaccharide moiety to the asparagine side chain of the nascent peptide.

Following the covalent attachment of N-glycan precursor, glucosidases in the ER remove the glucose cap at the end of N-glycan and mannosidase removes the terminal manose before the glycoprotein exit the ER. These reactions are facilitated by glycan-binding proteins, calnexin and calreticulin, and they determine whether the protein is properly folded. If the protein is misfolded, either a glucose residue is added at the terminal to provide additional time for folding, or two more terminal mannoses are removed, which is the signal for proteasomal degradation. Properly folded and trimmed glycoproteins are transported to the Golgi, and the N-glycans are further processed to become mature glycoproteins [136]. Maturation of N-glycans results in species-specific N-glycan sequences as well as complex-type glycan sequences (Figure 1.2).

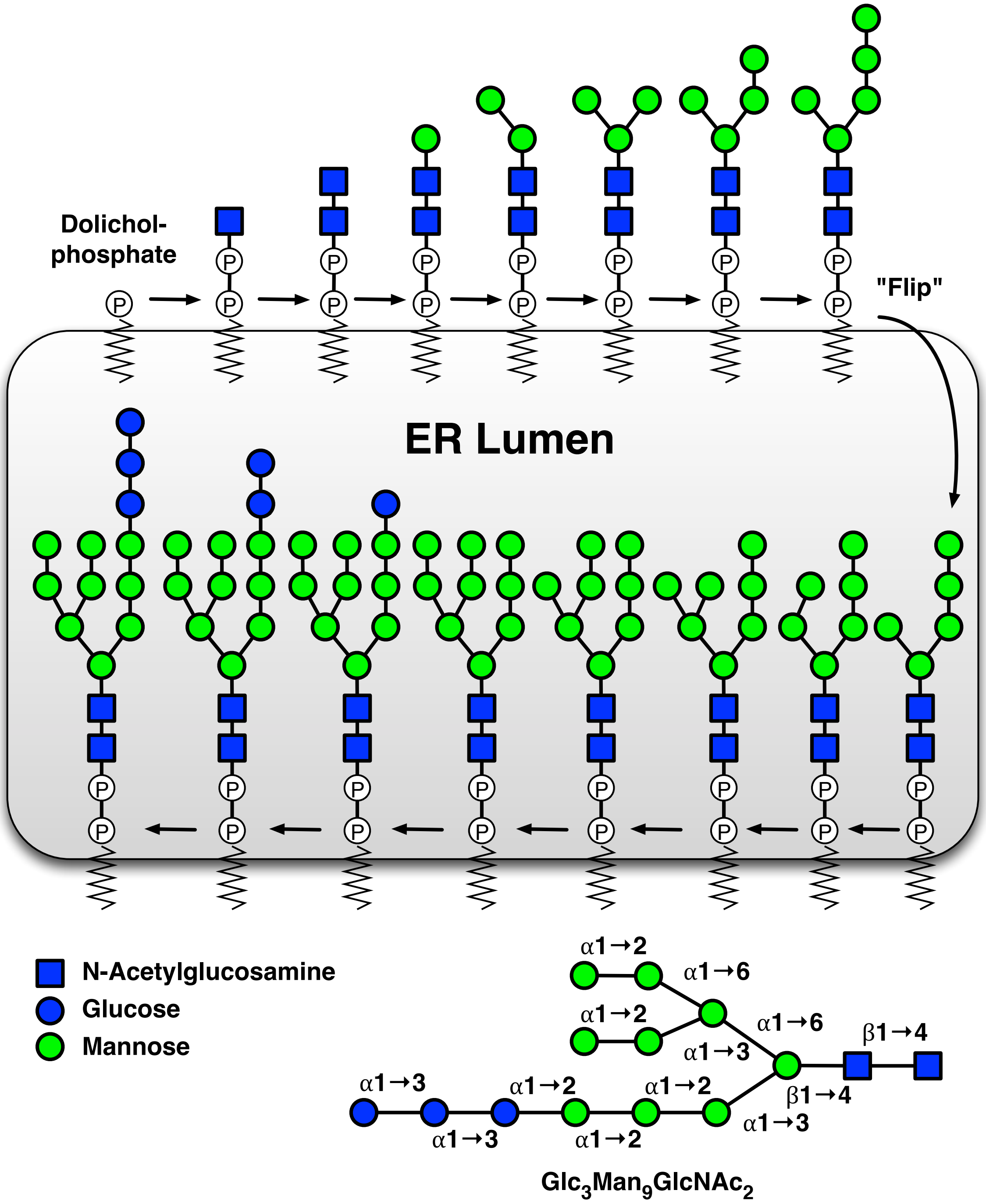


Figure 1.1: Synthesis of N-glycan precursor. Each arrow represents the addition of sugar catalyzed by different glycosyltransferases. The percursor intermediate is re-oriented to face the lumen side of the ER by a flippase (RTF1). (Figure is adapted and redrawn from [136])

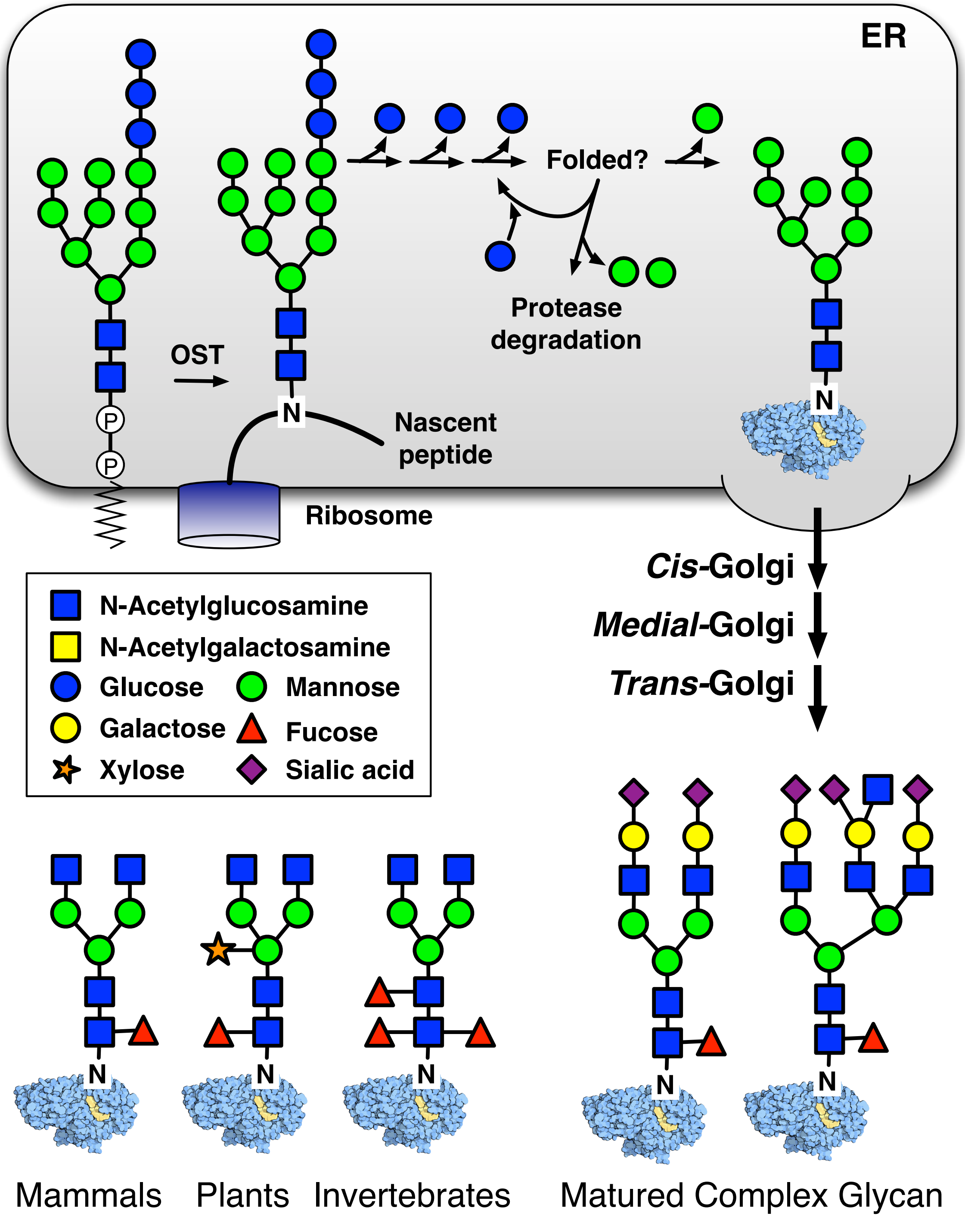


Figure 1.2: N-glycan precursor is transferred to a nascent peptide synthesized in the ER by oligosaccharyltransferase. Each arrow represents the addition or removal of sugar catalyzed by different glycosyltransferases. Glucosidases remove glucose caps in the precursor molecule and ER mannosidase removes the mannose from the N-glycan. Proteins in the ER (calnexin and calreticulin) determine whether the glycoproteins are properly folded. N-glycans are further processed and matured in Golgi. Several species-specific modifications are introduced in maturation process. (Figure is adapted and redrawn from [136])

1.2 Structural analysis of glycans

The basic structural unit of glycans is a monosaccharide. Two monosaccharides can be joined together by a glycosidic linkage between an anomeric carbon of one monosaccharide and a hydroxyl group of the other. The number of monosaccharides that commonly appear in glycoconjugates are limited. For example, there are only about 9 monosaccharides that commonly appear in glycoconjugates for vertebrates (see Figure 1.3) [136]. Although the number of monosaccharides is smaller than that of naturally occuring amino acids, the primary sequence of glycans can be much more diverse than the protein counterpart because a monosaccharide exists in two possible stereoisomers at the anomeric carbon (α or β), and there are four hydroxyl groups in a sugar that can accept a glycosidic linkage. In addition, more than one glycosidic linkage can be formed in one monosaccharide, resulting in a branched sequence.

Structural analysis of glycans could be analogous to that of proteins. A protein is a polypeptide chain comprised of linearly connected amino acids by peptide bonds. Each peptide bond has two rotatable bonds, and the peptide bond in the protein backbone is planar and rigid. Similarly, the oligosaccharide chain is comprised of monosaccharides, which are relatively rigid, joined by glycosidic linkages that have two or three rotatable bonds [111]. On the other hand, backbone atoms in polypeptides can form hydrogen bonds to form regular secondary structures [50, 107], such as helices or strands, whereas such regular structures are not found in glycans except in structural polysaccharides like cellulose.

Nonetheless, the torsion angle analysis of glycosidic linkages, similar to a Ramachandran plot for protein [108, 109], provides valuable insights on the preference of the glycan conformation. The degrees of freedom of these glycosidic linkages have been examined both experimentally and computationally [5, 32, 80, 81, 98, 119, 132, 140, 143]. The general consensus is that the glycosidic linkage is centered around a well known free energy basin. However, the width of the basin is larger than analogous ones for peptide bonds, and consequently, the number of internal degrees of freedom of glycosidic linkages is much larger than those of polypeptides. More studies on the conformational preference of large glycans are necessary to examine sequence dependent

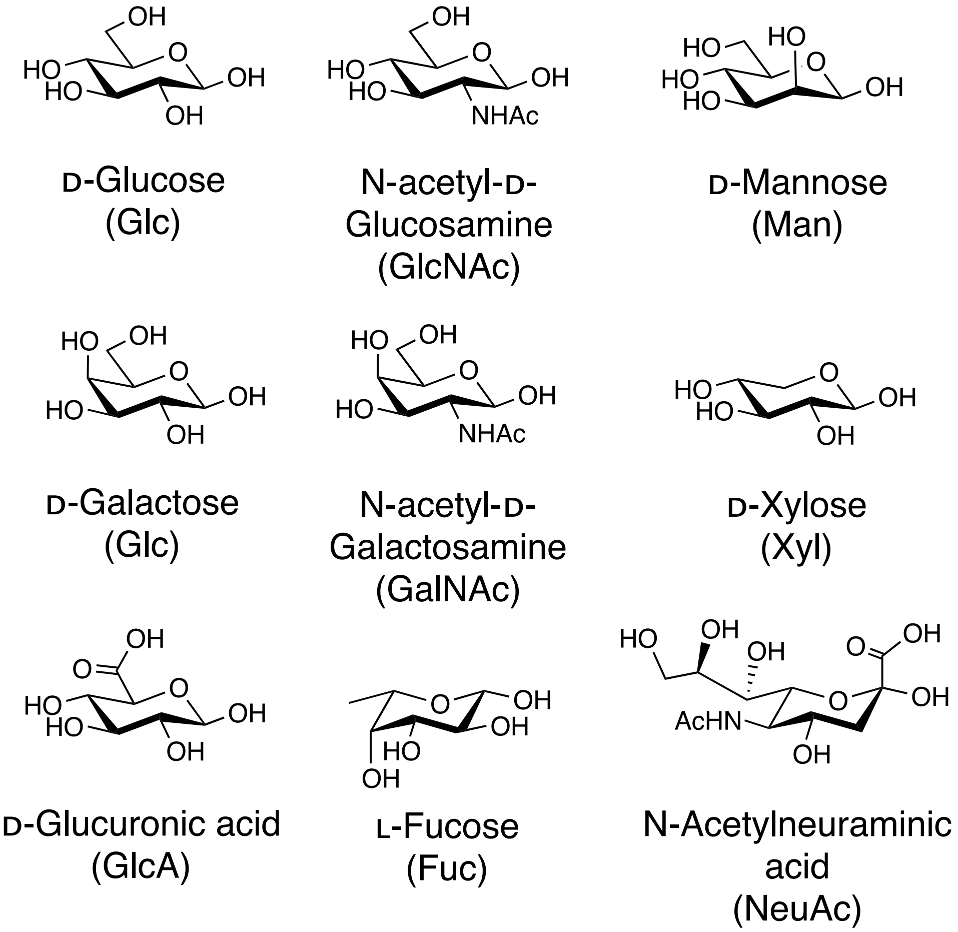


Figure 1.3: Common monosaccharides found in verterbrates. (Figure is adapted and redrawn from [136])

conformational preference of glycans.

1.3 Glycan conformation in solution and in the vicinity of protein

Traditionally, nuclear magnetic resonance (NMR) spectroscopy has been widely used to study the conformation of oligosaccharides, since crystallizations of oligosaccharides or glycoproteins are challenging. It is typically not easy to unambiguously derive distance information of neighboring glycan residues due to the crowded peaks, but NMR experiments can provide several valuable observables that can be used for structure determination of glycan, such as J-coupling or residual dipolar coupling (RDC) measurements. The conformation of an oligosaccharide is expected to be very flexible and inconsistent with a single conformer [143]. More recent NMR experiments have demonstrated biologically important glycans may have several different well-defined conformations that can readily undergo exchange between different conformations [5, 11, 132].

The N-glycan comes in close contact with protein surface residues, hence it has been of great interest whether the protein structure affects the N-glycan conformation and *vice versa*. An earlier NMR study about the conformational freedom of free oligosaccharide in solution and N-linked oligosaccharide concluded that the covalent attachment to the protein does not significantly affect the conformational freedom of the oligosaccharide [145]. However, it is well known that the carbohydrates in the vicinity of the protein can engage in specific interactions with protein side-chains, hence affecting the conformational freedom of the oligosaccharide [22, 26]. Thus, the conformational freedom of the N-glycan needs to be studied on a case-by-case basis. Structural change of the parent protein due to different glycoform sequences is also observed by systematically changing the glycoform [71]. These findings warrant more detailed studies of the interaction between the covalently linked N-glycan and the glycoprotein.

To gain a better understanding of conformational preference of oligosaccharides, it is essential to obtain atomic resolution structures in various environments. However, experimental determination of N-glycan conformations using X-ray crystallography or NMR is challenging due to flexibility of glycosidic linkages and the peak overlapping in NMR spectra [5, 78, 128, 143]. On the other hand, computational simulation studies of oligosaccharides can provide valuable insights on the conformational preference of oligosaccharides at the atomic level [96, 138]. Recent advances in carbohydrate force fields have been used to study diverse glycan sequences ranging from monosaccharides to polysaccharides, and have been shown to match experimental properties well [27, 42, 69]. Such a modeling approach may help refine the structural models from low-resolution experiments, e.g., small angle X-ray scattering or electron microscopy [39, 40].

In addition, preparing glycoprotein samples with homogeneous glycoform is extremely challenging because the biosynthesis of N-glycan is not controlled by genetic information. Therefore, producing glycoproteins with homogeneous glycoform typically requires either step-wise enzymatic deglycosylation from larger glycoforms [71, 88] or addition of N-glycan after producing the aglycoprotein using chemical synthesis [137]. Recently, *in vitro* glycosylation has been demonstrated, and the atomic structure was successfully determined using NMR [128], bringing exciting oppotunities in understanding the structural relationship betwwen the glycan and protein.

1.4 Outline of thesis

This dissertation mainly focuses on the structural analyses and computational studies of the conformational freedom of glycans in solution and in the vicinity of a protein. Several tools and computational algorithms were developed towards this goal. New knowledge discovered in this area can be used to gain general insight into the structural analysis of glycans in solution and in the vicinity of protein. Chapter 2 discusses the computational algorithm for reliable annotation of monosaccharides as well as oligosaccharide chains in a structure file solely based on the atomic coordinates and the connectivity to remove human errors in the annotation process. A database of glycan structures found in the Protein Data Bank (PDB) using the newly developed algorithm is discussed in Chapter 3. Then, the impact of protein structure on the conformation of glycans is disccused using the glycoconjugate crystal structures in the PDB in Chapter 4. Next, the conformational preferences of N-glycan core pentasaccharide in solution and in the vicinity of glycoprotein are examined in Chapter 5. Finally, in Chapter 6, the possibility of developing a computational protocol for prediction of N-glycan structure using template-based appraoch.

Chapter 2

Glycan Reader: Automated Sugar Identification and Simulation Preparation for Carbohydrates and Glycoproteins[1]

2.1 Introduction

Glycosylation is the most common post-translational modification process in proteins, and over half of all secreted proteins are expected to be glycosylated [7, 90, 136]. In addition to being a common protein appendage, glycans are also important in that they may alter protein structure and dynamics, and thus modify enzyme activity, protein-protein interactions, and the *in vivo* circulation half-life of protein pharmaceuticals [9, 92, 93, 117, 135]. Glycans are also involved in specific interaction with glycan-binding proteins and play a role in cellular or molecular recognition [117]. At this time, however, it is difficult to understand, on a case-by-case basis, which glycans are important components of protein function and specific recognition, and how to modify those glycans to optimize the protein properties of interest. To be able to predict a glycan's impact on the glycosylated protein's function and specific interaction with other proteins, it is critical to

[1] Reused from Jo, S., Song, K. C., Desaire, H., Mackerell, A. D., Jr and Im, W. *J. Comput. Chem.* **32**, pp 3135–3141 (2011) with the permission from *Wiley and Sons*

understand the structure and dynamics of glycans and the glycosylated protein.

The Protein Data Bank (PDB) is the largest database of biomolecular structures, [12] and, as of January 2011, the database contained about 71,000 entries. Among those entries, about 6% contain carbohydrate structures. Any type of biomolecular simulation begins with reading a protein structure into the simulation program. However, a task as simple as reading a PDB structure file into a molecular simulation, often becomes non-trivial when a carbohydrate is present due to the inconsistency and complexity existing in the PDB file format. Despite the efforts made to standardize nomenclature and data structures for representing carbohydrates in the PDB, the current naming convention does not unambiguously identify anomeric configurations; it also contains other limitations. For example, GLC and BGC refer to α-D-glucose and β-D-glucose for glucopyranose, but both GAL and GLA refer to α-D-galactose in the case of galactopyranose [30]. Such inconsistency in the nomenclature potentially leads to errors in annotating carbohydrates from the PDB. This problem is confounded by about 30% of carbohydrate structures in the PDB containing at least one error regarding the carbohydrate-type assignment [80]. Furthermore, there are cases where the entire carbohydrate chain is treated as a single residue, e.g., PDB:1AGM [80]. Thus, it is necessary to develop an algorithm that is able to automatically annotate carbohydrate structures based on their three-dimensional (3D) structures instead of relying on the PDB annotation.

There are several web-based toolsets for structural glycobiology presently available. The GLYCOSCIENCES.de web portal (`http://www.glycoscience.de`) [79, 81] and the Glycoconjugate Data Bank (`http://www.glycostructures.jp`) [94] offer convenient ways to automatically annotate carbohydrates in PDB files. In addition, a 3D model of a carbohydrate or glycoprotein structure can be generated through web-based tools such as SWEET (`http://www.glycosciences.de/modeling/sweet2/`) [15], GlyProt (`http://www.glycosciences.de/modeling/glyprot/`) [16], and Carbohydrate Builder (`http://glycam.ccrc.uga.edu/ccrc/carbohydrates/`) [37]. However, a significant effort is required to prepare the glycoprotein or protein/glycan complex system since the interfaces offer rather limited options for glycan structure generation and lack the ability to prepare the generated structures for biomolecular simulations.

Motivated by the above limitations and needs, we have developed *Glycan Reader* and its web-based interface (`http://www.charmm-gui.org/input/glycan`). *Glycan Reader* greatly simplifies the reading of PDB structure files with glycans by (i) detection of carbohydrate-like molecules, (ii) automatic annotation of carbohydrates based on their 3D structures, (iii) recognition of glycosidic linkages between carbohydrates as well as N-/O-glycosidic linkages to proteins, and (iv) generation of inputs for the biomolecular simulation program CHARMM [17] with proper glycosidic linkage setup. In addition, *Glycan Reader* is linked to other functional modules in the CHARMM-GUI (`http://www.charmm-gui.org`) [62], allowing users to easily generate protein/carbohydrate complexes or glycoprotein molecular simulation systems in solution or membrane environments and visualize the electrostatic potential on glycoprotein surfaces. These tools are useful for studying the impact of glycosylation on protein structure and dynamics. *Glycan Reader* utilizes the recently developed CHARMM carbohydrate force field, [41, 42, 46, 110] which includes a wide range of furanose and pyranose monosaccharides and glycosidic linkages including N-/O-glycosidic linkages to proteins and lipids.

In the next section, *Glycan Reader* and its web interface developments are described in detail. This is followed by some illustrations of *Glycan Reader*, such as PDB glycan statistics, electrostatic potential visualization on glycoprotein surfaces, and preparation of carbohydrate and glycoprotein simulation systems in both aqueous and membrane environments. Future directions of the *Glycan Reader* development project are then discussed briefly.

2.2 Methods

To annotate glycans in a given PDB file, *Glycan Reader* uses an algorithm that can detect carbohydrate-like molecules and assign correct carbohydrate types based on their molecular topology and 3D structures. The overall scheme in *Glycan Reader* is shown in Figure 2.1 and

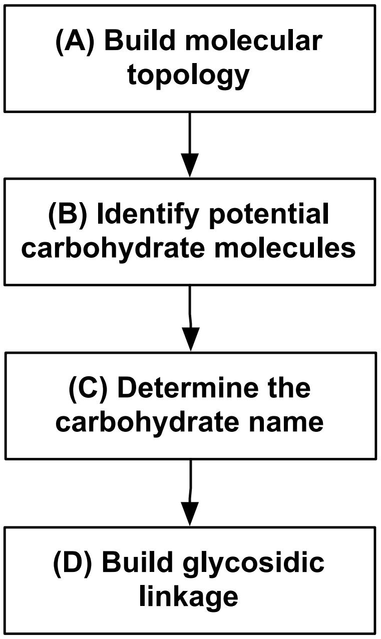

(A) Build molecular topology
(B) Identify potential carbohydrate molecules
(C) Determine the carbohydrate name
(D) Build glycosidic linkage

Figure 2.1: Overview of carbohydrate annotation procedure in *Glycan Reader*

illustrated in Figure 2.2. Molecular topologies are built based on the HETATM records and CONECT records in a PDB file, and any molecules that do not have carbohydrate-like topology are not considered. The chemical groups that are attached to the carbohydrate-like molecules are then examined to assign the correct carbohydrate type. Once the monomeric units are identified, glycosidic linkage types are determined.

2.2.1 Automatic Detection and Assignment of Sugar Types

In the first step (Figs. 2.1A and 2.2A), *Glycan Reader* builds topologies of molecules in a PDB file based on both HETATM and CONECT records. Carbohydrate-like structures are identified by the presence of six- or five-membered rings that are composed of one oxygen atom and five or four carbon atoms depending on the size of the ring. Each potential carbohydrate molecule is further examined to identify the anomeric carbon by checking the carbon atoms connected to the ring oxygen to see if one of them has oxygen or nitrogen atom attached to it. If such an atom is found, the atom is designated as the anomeric carbon (C_1) and the rest of the ring constituents are re-numbered accordingly. However, in the case that no apparent anomeric carbon is found due

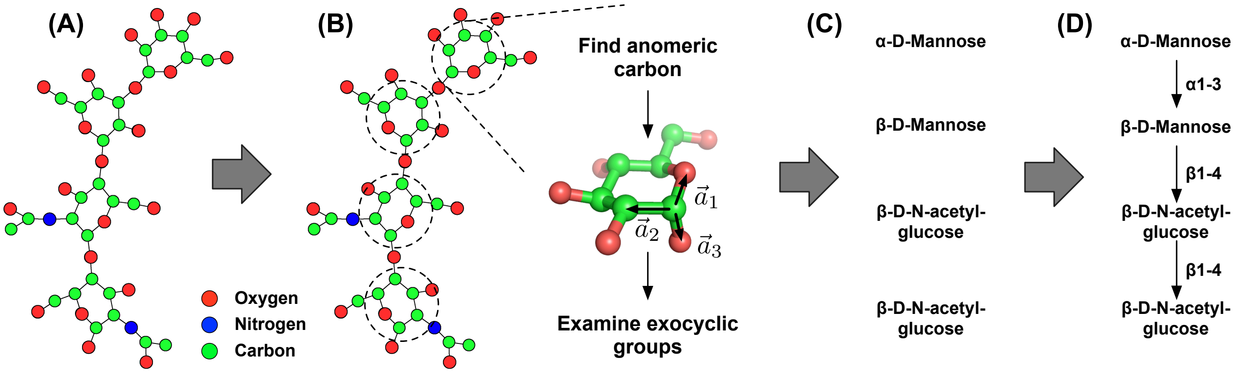


Figure 2.2: Illustration of carbohydrate annotation procedure in *Glycan Reader*. A) Molecular topology is built using HETATM and CONECT records in a PDB file. B) Potential carbohydrate molecules are examined for anomeric carbon, stereochemistry of each ring carbon atoms, and exocyclic groups. C) Carbohydrate type is annotated. D) Glycosidic linkages are assigned between monosaccharides.

to a lack of electron density or an error in the PDB structure, a carbon atom, that is connected to the ring oxygen and has an exocyclic carbon atom attached to it, is designated as C_5 (for six-membered ring) or C_4 (for five-membered ring), and the other carbon atom attached to the ring carbon is assigned as the anomeric carbon. This method will fail to properly detect the anomeric carbon if no exocyclic carbons are attached to the C_5 (for pyranose) or C_4 (furanose) atom, for example xylopyranose, however, this is a backup algorithm, which is only used when the oxygen atom attached to the anomeric carbon is not found.

Once the anomeric carbon is assigned, *Glycan Reader* determines the carbohydrate residue type. Carbohydrate monomers can be classified by examining the configuration of the hydroxyl group attached to each carbon atom in the ring. Therefore, *Glycan Reader* calculates the improper angle based on the angle difference between the C_n-O_n ($\mathbf{a_3}$) and a vector perpendicular to C_{n-1}-C_n ($\mathbf{a_1}$) and C_n-C_{n+1} ($\mathbf{a_2}$) as shown in Fig. 2.2B. The configuration is then compared with a pre-established look-up table to determine the carbohydrate residue type. Currently, *Glycan Reader* can recognize all pyranose and furanoses available in the recent version of CHARMM carbohydrate force field [41], and a few more carbohydrates that are not available at the date of publication (see Table 2.1 for the complete list of available carbohydrate types).

Derivatized carbohydrates, such as N-acetyl-glucosamine, N-acetyl-nuraminic acid, or

Chemical name	CHARMM residue name
D-Glucose	GLC
D-Altrose	ALT
D-Allose	ALL
D-Galactose	GAL
D-Gulose	GUL
D-Idose	IDO
D-Mannose	MAN
D-Talose	TAL
D-Xylose	XYL
L-Fucose	FUC
L-Rhamnose	RHM
D-Glucuronic acid	GLCA
L-Iduronic acid	IDOA
N-Acetyl-D-glucosamine	GLCNA
N-Acetyl-D-galactosamine	GALNA
N-Acetyl-D-nuraminic acid	NE5AC
Tetrahydropyran (THP)	THP2
Deoxyribose	DEO
Ribose	RIB
Arabinose	ARB
Lyxofuranose	LYF
Xylofuranose	XYF
Fructofuranose	FRu

Table 2.1: List of carbohydrates recognized by *Glycan Reader*

iduronic acid, are recognized by comparing the exocyclic chemical groups. The CHARMM carbohydrate force field [41, 42] provides separate residue definitions for such modification (e.g., acetylation, oxidation or deoxidation), and the residue names of such derivatized carbohydrates are renamed to the corresponding CHARMM residue names. When there is no residue definition available, e.g., no definition available for N-acetyl-mannosamine, *Glycan Reader* simply considers the residue as non-carbohydrate residue. As additional carbohydrate definitions become available in the CHARMM force field, they will be implemented in *Glycan Reader*.

2.2.2 Glycosidic Linkage Detection and Assignment

The anomeric position of each carbohydrate monomer is examined to check if the residue is connected to another carbohydrate by the glycosidic linkage. In our scheme, the root residue of a carbohydrate chain is simply assigned to a residue that has a free reducing end: for example, α-D-N-acetyl-glucose in Fig. 2.2D. N- or O-glycosylation is determined by cross-referencing the connected protein residue on the reducing end of the glycan chain; N-glycosylated when the reducing end is connected to ASN and O-glycosylated when the reducing end is connected to THR or SER. During the implementation, we frequently found incorrectly assigned bonds in glycan chains, which interfere with glycosidic linkage detection. For example, Figure 2.3A and 2.3B show incorrectly assigned bonds between neighboring residues possibly due to close proximity between two atoms, which forms a small ring structure and hinders the correct glycosidic linkage assignment. To assign glycosidic linkages reliably, each glycosidic linkage is re-examined to remove any chemical bonds that do not make chemical sense, e.g., oxygen atoms having three covalent bonds. On the other hand, there are some glycan chains that have missing glycosidic linkages (Fig. 2.3C). In such cases, *Glycan Reader* examines the distance between the anomeric carbon and the exocyclic oxygen on the neighboring residue; if it is in close proximity (e.g., < 2.5 Å), a glycosidic linkage is generated between the two residues. In rare occasions, covalent bonds with extreme bond lengths are present in the PDB (Fig. 2.3D); any chemical bonds that are longer than 5 Å will be removed in *Glycan Reader*. These error correction features have

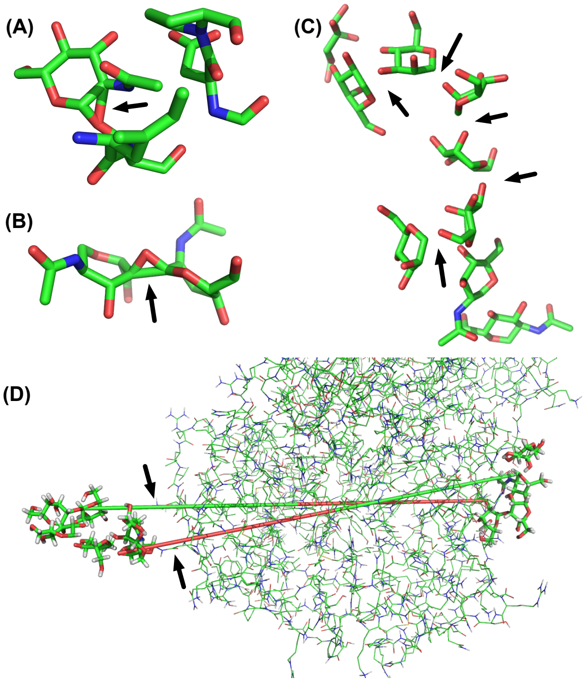


Figure 2.3: Examples of erroneous glycan chains in PDB. (A) Additional bond between the glycosidic oxygen and the ring oxygen (PDB:1Q5C) (B) Additional carbon-carbon bond between two residues (PDB:1BCR). (C) Missing glycosidic linkages (PDB:2H6O) (D) Incorrect connectivity between two atoms (PDB:1INH). Erroneous bonds are marked by arrows.

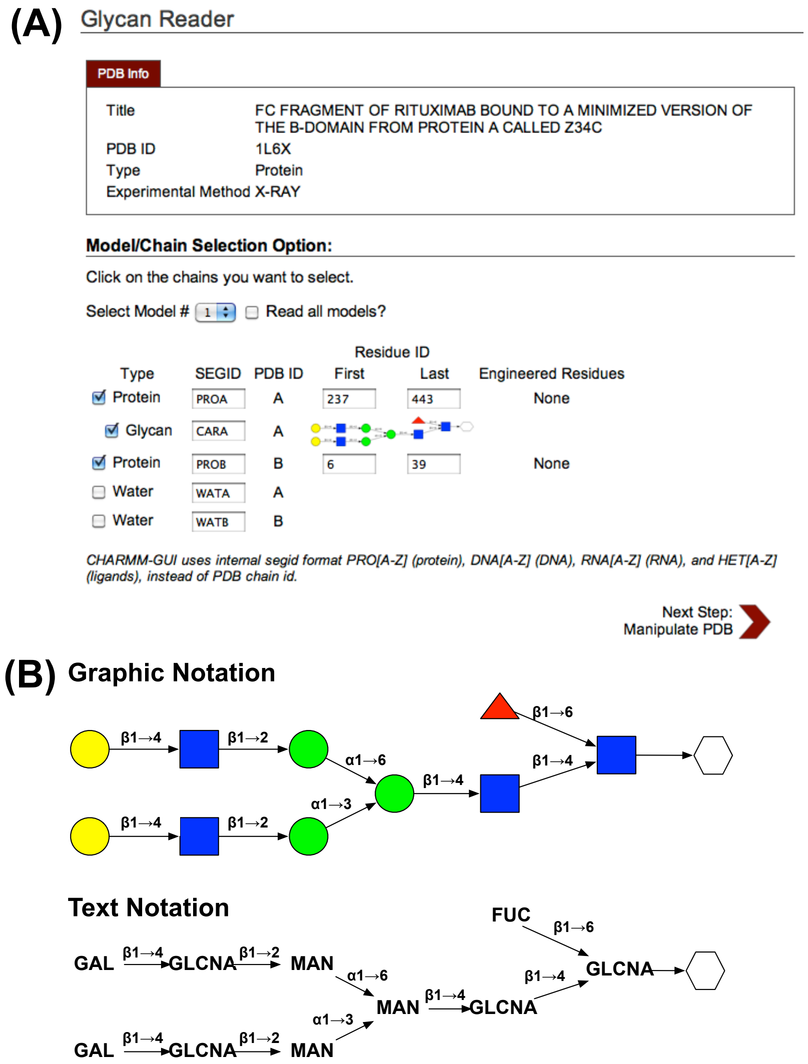
been tested by a number of internal testcases, however, users are always advised to make sure if the input structure is correct and the output from the *Glycan Reader* is as intended. In the case that a carbohydrate chain is connected to a non-carbohydrate molecule, the entire chain is ignored presently. For instance, PDB:1S0J contains a ligand molecule that is a derivative of sialic acid with a methylumbelliferyl moiety, and *Glycan Reader* classifies the molecule as a non-carbohydrate molecule. While currently not implemented in an automated fashion, the potential to treat such moieties using the CHARMM General Force Field is possible [134].

The CHARMM carbohydrate force field [41, 42] provides several linkage types for mixed pyranose and furanose compounds, such as sucrose, lactulose, melezitose, raffinose, kestose, 6-kestose, isomaltulose, planteose, and nystose. This is because it is not possible to use the same linkage type between pyranose and furanose due to different atom types. Therefore, *Glycan Reader* detects the presence of mixed pyranose and furanose compounds, and uses appropriate linkage types to make glycosidic linkages between the pyranose and furanose residues.

2.2.3 GUI Implementation of *Glycan Reader* and CHARMM Input Generation

Glycan Reader has been integrated into the CHARMM-GUI web interface [62]. The user can either specify the PDB ID or upload the PDB structure into the server to generate the carbohydrate or protein/carbohydrate complex structure. If a carbohydrate is detected, then the graphical representation of the carbohydrate chain sequence will be displayed and the user can select the carbohydrate chains that they want to initialize in CHARMM (see Fig. 2.4). CHARMM allows modification in chemical structures, e.g., disulfide bonds formation or phosphorylation using patch residues, and glycosidic linkages are generated using specific patch residues in CHARMM. The *Glycan Reader* web interface assigns the proper patches for glycosidic linkages and generates the CHARMM protein structure file (PSF) and coordinate files in both, PDB format and CHARMM specific coordinate format (CRD) files.

Currently, there are various patch residues available in the CHARMM carbohydrate force field to cover a range of carbohydrates including the majority found in eukaryotes [43]. For example, O-methyl-, octyl-, dodecyl-, phosphate, and sulfate groups can be added to the reducing end of a sugar, and those modifications are properly patched in the PSF generation step (see Table 2.2 for the complete list of patch residues available). However, other types of common derivatizations, such as deoxidation is not available, and, in such cases, the basic form of the carbohydrate molecule is used without modification and *Glycan Reader* informs the user. For example, if a user uploaded a structure of 2-deoxy glucose, a glucose molecule will be generated instead.



Glycan Reader

PDB Info

Title	FC FRAGMENT OF RITUXIMAB BOUND TO A MINIMIZED VERSION OF THE B-DOMAIN FROM PROTEIN A CALLED Z34C
PDB ID	1L6X
Type	Protein
Experimental Method	X-RAY

Model/Chain Selection Option:

Click on the chains you want to select.

Select Model # 1 □ Read all models?

Type	SEGID	PDB ID	Residue ID First	Residue ID Last	Engineered Residues
☑ Protein	PROA	A	237	443	None
☑ Glycan	CARA	A			
☑ Protein	PROB	B	6	39	None
□ Water	WATA	A			
□ Water	WATB	B			

CHARMM-GUI uses internal segid format PRO[A-Z] (protein), DNA[A-Z] (DNA), RNA[A-Z] (RNA), and HET[A-Z] (ligands), instead of PDB chain id.

Next Step: Manipulate PDB

Figure 2.4: Snapshot from CHARMM-GUI *Glycan Reader*. (A) When a glycan chain is found in a PDB, the sequence of the identified glycan chain is displayed. (B) When the sequence diagram is clicked, more detailed information on the glycan sequence is displayed in a popup window.

Modification	CHARMM patch residue
O-Methyl- at C_1	OME (pyranose)
	FOME (furanose)
Octyl- at C_1	OCT (pyranose)
Dodecyl- at C_1	DDM (pyranose)
Phosphate- at C_1	PH (THP)
Phosphate- at C_2	PH2 (THP)
Sulfate- at C_1	PH2 (THP)

The types of residues that are available for the modification are given in parenthesis.

Table 2.2: List of modifications recognized by *Glycan Reader*

2.3 Results and Illustrations

2.3.1 PDB Glycan Statistics

We have used *Glycan Reader* to analyze the entire PDB database to obtain the statistics on the available glycan-containing structures out of the total of 70,947 structures in the PDB as of January, 2011. The results are summarized in Figure 2.5. There are a total of 4,029 PDB structure files (6.0%) that have at least one glycan chain, yielding a total of 15,669 glycan chains. A total of 8,848 glycan chains (56%) are N-glycosylated, 688 glycan chains (4.3%) are O-glycosylated, and the rest (6,133 chains, 39%) exist as noncovalently-bound ligands. Figure 2.5A shows the number of PDB structures with glycans deposited each year into the PDB; despite that only 6% of PDB structures contain carbohydrate segments, the trend shows steady increase over time. Figure 2.5B shows the number of glycan chains as a function of glycan chain length (e.g. number of monosaccharides residues in a chain), illustrating that most glycan structures in PDB contain only one or two monosaccharides. The large number of shorter glycan chain could be due to the removal of glycans prior to structural studies or due to crystallization conditions. Our survey showed β-N-acetyl glucosamine (GlcNAc) is the most abundant monosaccharide (4,917 entries) and GlcNAc β(1→4) GlcNac β is the most abundant disaccharide (1,653 entries). The survey presented this work concerned about the number of PDBs with glycan chains present and the composition of carbohydrate molecules in the PDB database. With *Glycan Reader*, which allows one to conveniently and reliably recognize glycan chains, can be used for further studies on protein-carbohydrate interactions. To this end development of a PDB glycan database to retrieve any specific glycan structure is currently in progress.

2.3.2 Electrostatic Potential Visualization

Characterizing the electrostatic potential on a macromolecular surface is becoming a routine practice in structural biophysics [52]. Similarly, comparing the electrostatic representations with and without glycans could provide insights into the biological roles of glycans. For example,

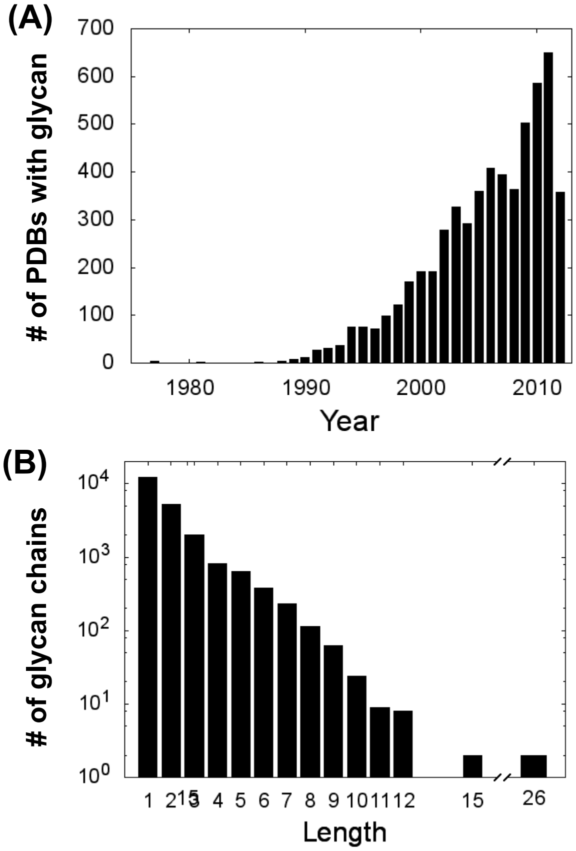


Figure 2.5: PDB glycan statistics. (A) The number of structures with glycans added to the PDB each year. (B) The number of glycan chains with respect to the glycan chain length.

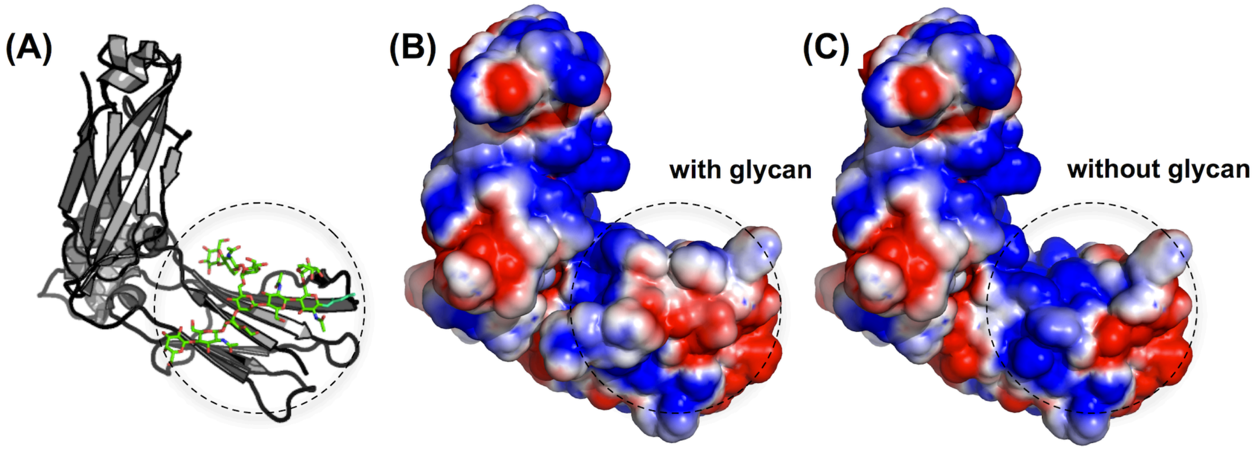


Figure 2.6: Molecular images of a glycoprotein and its electrostatic potential surface with and without glycan. (A) A molecular image of PDB:1L6X (constant region of immunoglobulin G1). (B) The electrostatic potential surface of the glycoprotein. (C) The electrostatic potential surface of the glycoprotein without glycan. The glycans on the surface are highlighted with dotted circles. Users can do various renderings of the images on the web using the MarvinSpace tools [20], or using PyMOL.

some carbohydrates, such as sialic acid or phosphorylated sugars, are negatively charged and their spatial distribution might be important for protein function. The CHARMM-GUI PBEQ-Solver (`www.charmm-gui.org/input/pbeqsolver`) calculates the electrostatic potential and solvation free energy of biomolecules by solving the Poisson-Boltzmann (PB) equation using the CHARMM PBEQ facility [56, 95, 114]. Using the web based electrostatic visualization interface, a user can quickly calculate the electrostatic potential of a glycoprotein or protein/glycan complex (Figure 2.6). Currently, a generic set of atomic radii (i.e., 2.3 Å for carbon, 1.8 Å for oxygen, and 2.3 Å for nitrogen) is used for carbohydrate molecules to calculate the electrostatic potential surface. Efforts are on-going in our laboratory to fine-tune such atomic radii allowing for their use in the context of implicit solvent models.

2.3.3 Simulations of Carbohydrates and Glycoproteins in Various Environments

MD simulations of biomolecules have become a common tool in the study of structural, dynamical, and energetic aspects of biological mechanisms [75]. The methodology for such simulations is well established and thus simulation input generation can be greatly simplified and automated once a PDB structure has been successfully read; this capability was the motivation for the widely-used CHARMM-GUI MD Simulator (`www.charmm-gui.org/input/mdsetup`) and CHARMM-GUI Membrane Builder (`www.charmm-gui.org/input/membrane`) [61, 64]. Unlike protein structures, each monosaccharide unit in a glycan chain is connected by different glycosidic linkages (i.e., specific patch residues), which needs to be correctly recognized in the simulation package. Moreover, glycan chains may be branched, which makes the residue numbering non-contiguous. Such complexity in a glycan structure complicates manual linkage building, making it susceptible to error. *Glycan Reader* is integrated with various modules in CHARMM-GUI, such as MD Simulator and Membrane Builder, which facilitate preparation of glycoproteins for simulations in solution or membrane embedded environments (Figure 2.7). In addition to CHARMM, files produced by *Glycan Reader* may be used directly to perform simulations in NAMD [106], and the capabilities exist to perform simulations using the CHARMM force fields in GROMACS[133] and AMBER [19].

2.4 Discussion

We have developed a web-based tool, *Glycan Reader* (`http://www.charmm-gui.org/input/glycan`), that can automatically identify and annotate carbohydrate based on atomic coordinates, atom types, and bonds in a PDB structure file. *Glycan Reader* reliably detects the carbohydrate molecules, assigns their configuration and identifies the glycosidic linkages between monosaccharaides. These capabilities will facilitate computational studies of glycoprotein or protein/glycan complexes. *Glycan Reader* may also be used during the determination of protein structures that contain carbohydrates to make sure the assignment of residue types and the chemical bonds are correctly done as intended. It is integrated into the CHARMM-GUI website as the *Glycan Reader* module and cross-linked to various modules in CHARMM-GUI. For example, one can use PBEQ Solver (`http://www.charmm-gui.org/input/pbeqsolver`) to

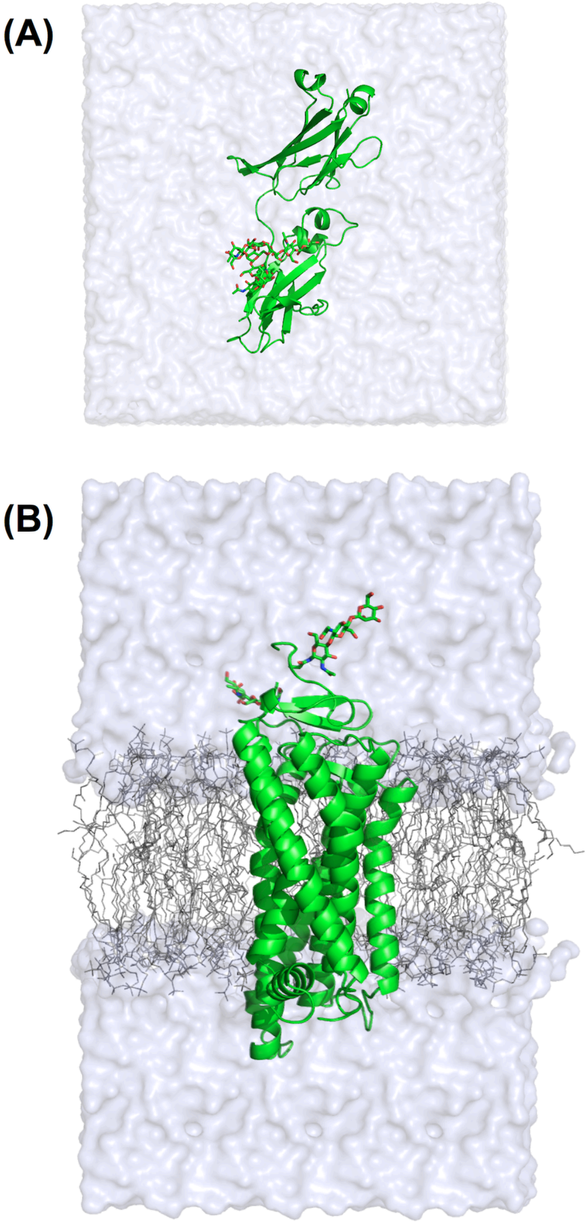


Figure 2.7: Images of glycoprotein simulation systems. (A) Constant region of immunoglobulin-1 (PDB:1L6X) in aqueous environment prepared using the CHARMM-GUI MD Simulator (B) Bovine rhodopsin (PDB:1GZM) in a lipid bilayer environment prepared using the CHARMM-GUI Membrane Builder.

calculate and visualize the electrostatic potential surface of glycoprotein. In addition, one can use MD Simulator (`http://www.charmm-gui.org/input/mdsetup`) or Membrane Builder (`http://www.charmm-gui.org/input/membrane`) to quickly generate MD simulation systems of glycoproteins and protein-carbohydrate complexes in aqueous or lipid bilayer environment.

Chapter 3

Glycan fragment database: a database of PDB-based glycan 3D structures[1]

3.1 Introduction

An oligosaccharide moiety in a glycoprotein, referred to as a glycan, comes in a diversity of sequences and structures, and specific interactions between carbohydrates and proteins are essential in many cellular events [105, 117, 143]. These events require molecular recognition of specific carbohydrate structures that seems to be sensitive to small differences in sequence. For instance, the carbohydrate structures found on a host cell receptor, which only differ by the sequence of the terminal sugar residues, are believed to be a major factor in determining the host range (e.g., swine, avian, or human) of influenza viruses [125, 127]. In addition, glycosyl transferases and glycosidases recognize specific sequence and spatially arranged oligosaccharide chain [122, 154]. Thus, understanding the conformation of carbohydrates will provide insight into the role of glycans in modulating many cellular events.

Analogous to protein structure, the structure of an oligosaccharide chain can be characterized by the torsion angles of glycosidic linkages between relatively rigid carbohydrate monomeric

[1] Reused from Jo, S. and Im, W. *Nucleic Acids Res.* **41**, pp D470–474. with the permission of *Oxford University Press*

units. Considerable efforts have been already made to characterize the potential energy surface of the peptide bond conformation, and the accessible torsion angles of a peptide are well known [50, 54, 107, 108, 109]. However, unlike proteins and peptides where the amino acid units are linearly linked together by the same peptide bonds, glycan chains can have branches and each monosaccharide unit can be linked by different types of glycosidic linkages. In addition, the lack of experimentally derived atomic structures of oligosaccharides in aqueous solution makes it difficult to characterize the accessible torsion angles of a particular glycosidic linkage.

Despite the difficulties involved in crystallization, the number of glycoprotein structures deposited in the Protein Data Bank (PDB) [13] has been steadily increasing [65, 80]. Although far from complete, glycan structures in the PDB can be used to study the accessible glycosidic torsion angles [78, 103, 104, 131]. Unfortunately, however, extracting structural information of glycans from the PDB is not trivial due to a lack of standardized nomenclature and the way the data is presented in the PDB [80, 143]. Recently, Säwén et al. analyzed the accessible glycosidic torsion angles of the "α-(1→2)-" linked mannose disaccharide using the PDB glycan structures, but they had to make considerable efforts to collect and filter out erroneous PDB entries [131].

In this work, we present the Glycan Fragment database (GFDB), a database of the glycosidic torsion angles derived from the PDB glycan structures. Carbohydrate structures in the PDB are recognized by *Glycan Reader*, an automatic sugar identification algorithm that we developed [65], instead of using the nomenclature presented in the PDB entries. The GFDB provides an intuitive glycan sequence search tool that allows the user to search complex glycan structures. After a glycan search is complete, each glycosidic torsion angle distribution of the searched glycan structures is displayed. In addition, the torsion angle distributions can be clustered to generate representative structures using the clustering analysis facility on the GFDB interface. To facilitate the conformational analysis of glycosidic linkages, the GFDB also provides various filters. In the following sections, we discuss how the glycan structural information was collected, how to search a glycan sequence, and how the search results are displayed. A stepwise guide about the GFDB is also provided in `http://www.glycanstructure.org/fragment-db`.

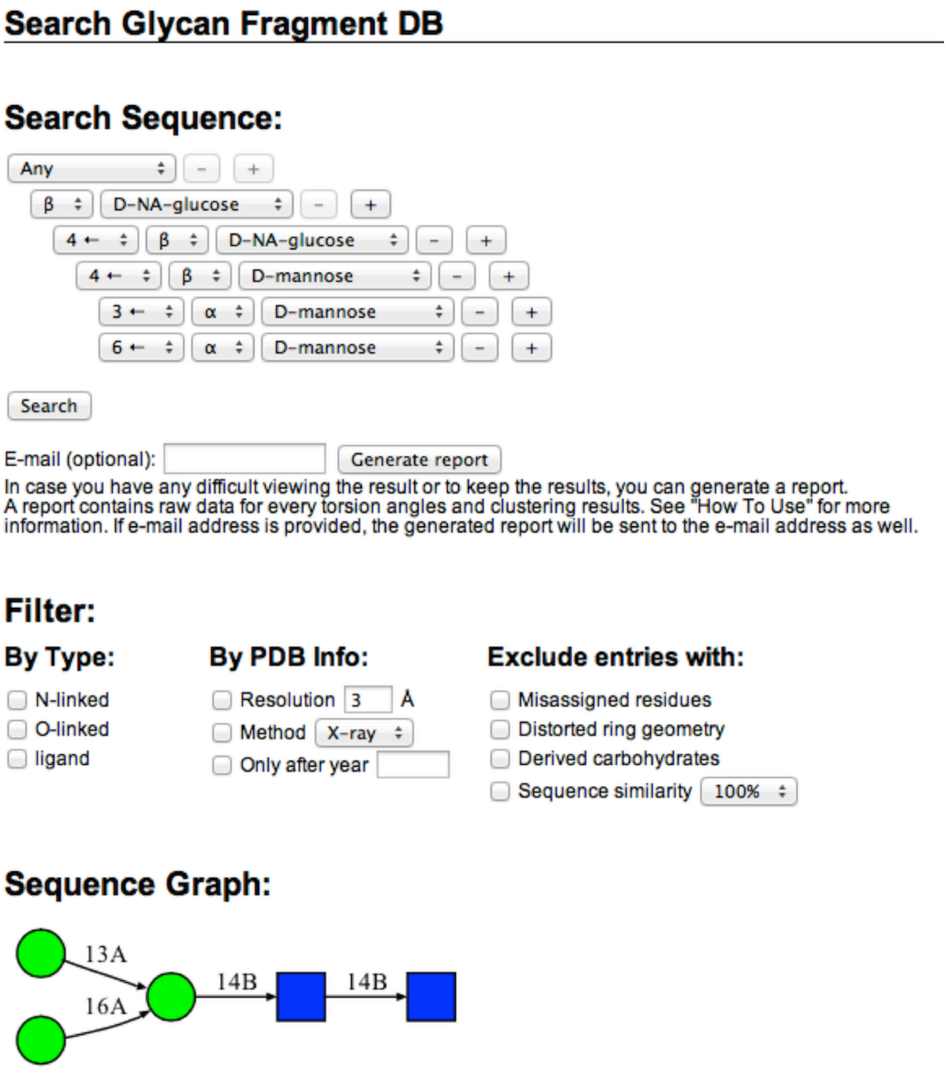
3.2 Glycan Fragment Database

To recognize the PDB entries that contain carbohydrate molecules, we used *Glycan Reader* for automatic sugar identification [65]. Briefly, in *Glycan Reader*, topologies of the molecules in the HETATM section of a PDB file are first generated using the atom connection information from the CONECT section. The carbohydrate candidate molecules (six-membered ring for a pyranose and five-membered ring for a furanose that are composed of only one oxygen and carbon atoms) are then identified. For each carbohydrate-like molecule, the chemical groups attached to each position of the ring and their orientations are compared with a pre-defined table to identify the correct chemical name for the carbohydrates. Glycan chains are constructed by examining the glycosidic linkages between the carbohydrate molecules that have chemical bonds between them.

Identified carbohydrate molecules are further analyzed and recorded in the GFDB. First, the residue name annotated in the PDB is compared with the molecular structure. The disparity of the residue name annotation in the PDB and the actual molecular structure is common [80]. Although *Glycan Reader* returns the correct carbohydrate names according to the molecular structures, such disparity could be a sign of potential error. Second, because a distorted ring geometry could mislead the interpretation of the glycosidic torsion angles, the geometry of the carbohydrate ring is calculated by virtual torsion angle definition [111] and recorded whether it is in chair conformation ($^1C_4$ or $^4C_1$) or not. Lastly, if the carbohydrate molecules have chemical groups (phosphate, sulfate, methyl, etc) attached in one of the hydroxyl groups, the carbohydrates are marked as derived carbohydrates in the GFDB. The entries that belong to these cases can be excluded from the search using the filtering options such as “Misassigned residues”, “Distorted ring geometry”, and “Derived carbohydrates” (see below and also Fig. 3.1).

3.3 Web Interface

The GFDB provides a glycan sequence search interface that allows the user to search complex glycan sequences (Fig. 3.1). The search interface provides a visual guide as the user builds a


Search Glycan Fragment DB

Search Sequence:

Any ⇕ − +
β ⇕ D-NA-glucose ⇕ − +
4 ← ⇕ β ⇕ D-NA-glucose ⇕ − +
4 ← ⇕ β ⇕ D-mannose ⇕ − +
3 ← ⇕ α ⇕ D-mannose ⇕ − +
6 ← ⇕ α ⇕ D-mannose ⇕ − +

Search

E-mail (optional): [ ] Generate report

In case you have any difficult viewing the result or to keep the results, you can generate a report. A report contains raw data for every torsion angles and clustering results. See "How To Use" for more information. If e-mail address is provided, the generated report will be sent to the e-mail address as well.

Filter:

By Type:
- □ N-linked
- □ O-linked
- □ ligand

By PDB Info:
- □ Resolution 3 Å
- □ Method X-ray ⇕
- □ Only after year [ ]

Exclude entries with:
- □ Misassigned residues
- □ Distorted ring geometry
- □ Derived carbohydrates
- □ Sequence similarity 100% ⇕

Sequence Graph:

Figure 3.1: The GFDB search interface

complex glycan query sequence, and the interface is compatible with any modern web browser with JavaScript capability. There is a report generation facility available to generate an archived report file that contains all the raw data for a given search as well as 3D structures based on the clustering analysis (see below); the user also can get the archived report file by e-mail. There are several filtering functions available (Fig. 3.1), which narrows the search results for specific needs, such as filters for only N-/O-glycosylated glycans, the resolution of the PDB entries, and/or the aforementioned three structural features (misassigned residues, distorted ring geometry, and derived carbohydrates).

When analyzing the glycosidic torsion angles in the PDB, it is important to understand that there are redundant PDB entries from the same or similar proteins. Without removing those redundant entries, it is possible to overestimate the preference of a certain conformation for a given glycan sequence. Although, redundancies in the PDB can be removed by post-processing the data obtained by the GFDB, the GFDB provides a preliminary filter option for removing such redundant protein entries for N-linked or O-linked glycan chains based on the sequence similarity of the parent protein.

After a glycan search is finished, the interface shows two torsion angle distributions side by side (Fig. 3.2): "exact match" and "fragment match" (Fig. 3.3). For the exact match, the GFDB first performs a sequence search to find the PDB entries that contain the glycan sequence identical to the query sequence, and the resulting torsion angle values for each glycosidic linkage are displayed to the user. On the other hand, the fragment search performs a search against the substructures (hence called fragments, Fig. 3.3) and returns the entries having at least one substructure that matches to the query sequence. This provides more samples for the torsion angle analysis. The torsion angle values from the fragment match always contain the exact match results. However, the fragment search results may not be the same as the exact match results because part of a glycan structure can adopt a different structure when it has extra intra- and intermolecular interactions. Therefore, the fragment match results implicitly include the influences from the nearby carbohydrate residues and different protein-carbohydrate interactions, so that one can assess the flexibility of a certain glycosidic linkage in the context of larger glycan chain by comparing the exact and fragment match results.

The glycosidic torsion angle definition in the GFDB is adopted from the crystallographic definition; O_5-C_1-O_1-C'_x (ϕ), C_1-O_1-C'_x-C'_{x-1} (ψ), and O_1-C'_6-C'_5-O'_5 (ω). The torsion angle between the first residue of the N-glycan chain and the side chain of the asparagine residue is defined as O_5-C_1-N'_{D2}-C'_G (ϕ) and C_1-N'_{D2}-C'_G-C'_B (ψ). The torsion angle between the first residue of the O-glycan chain and the side chain of the serine residue is defined as O_5-C_1-O'_G-C'_B (ϕ) and C_1-O'_G-C'_B-C'_A (ψ). For threonine, OG1 is used instead of OG. The atom names are based on the CHARMM topology.

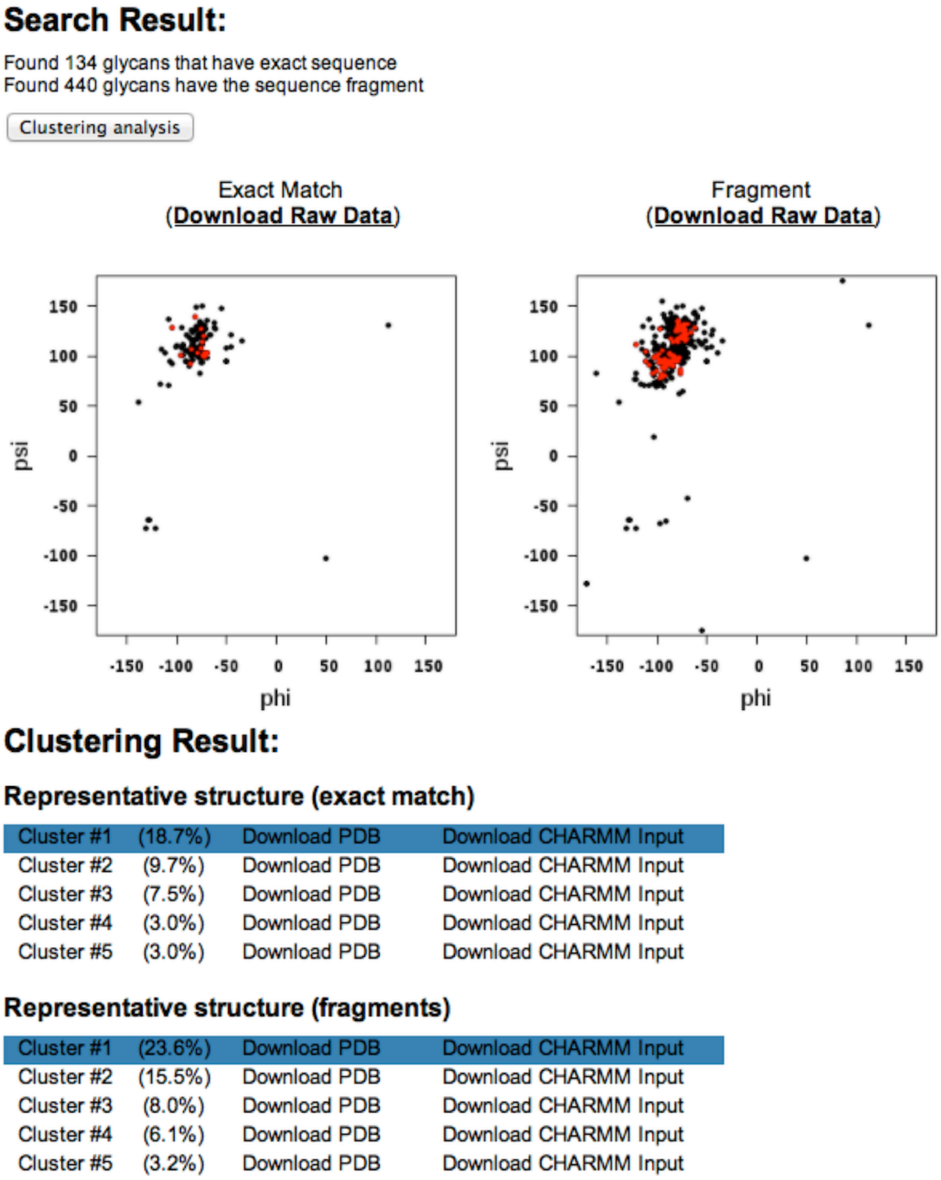
3.4 Clustering Analysis

Statistical analysis of the torsion angle values of a particular glycosidic linkage is useful to estimate the allowable conformations of glycan chains, but it is difficult to understand what would be the representative (or most probable) structures of the given glycan sequence among the available PDB glycan structures. To provide useful insight into the 3D glycan structure, the GFDB provides an option to perform clustering analysis of the torsion angle search results and produce the five most populated glycan structures.

The GFDB uses a simple clustering method to efficiently determine the members of each cluster. The pairwise torsion angle differences are first calculated:

$$d_{ij} = \sqrt{\frac{\sum_k \left(\phi_i^k - \phi_j^k\right)^2 + \left(\psi_i^k - \psi_j^k\right)^2}{N}} \tag{3.1}$$

where ϕ^k and ψ^k are the torsion angle values of the k-th glycosidic linkage, and i and j represent two glycan structures. ω torsion angle values are included only for glycosidic linkages that have three rotatable bonds. After the pairwise distance matrix of the searched glycan structures is calculated, the first cluster is identified with the maximum number of neighbors within a 30° cutoff



Search Result:

Found 134 glycans that have exact sequence
Found 440 glycans have the sequence fragment

Clustering Result:

Representative structure (exact match)

Cluster #1	(18.7%)	Download PDB	Download CHARMM Input
Cluster #2	(9.7%)	Download PDB	Download CHARMM Input
Cluster #3	(7.5%)	Download PDB	Download CHARMM Input
Cluster #4	(3.0%)	Download PDB	Download CHARMM Input
Cluster #5	(3.0%)	Download PDB	Download CHARMM Input

Representative structure (fragments)

Cluster #1	(23.6%)	Download PDB	Download CHARMM Input
Cluster #2	(15.5%)	Download PDB	Download CHARMM Input
Cluster #3	(8.0%)	Download PDB	Download CHARMM Input
Cluster #4	(6.1%)	Download PDB	Download CHARMM Input
Cluster #5	(3.2%)	Download PDB	Download CHARMM Input

Figure 3.2: An example of the search result for the query sequence in Figure 3.1. The glycosidic torsion angle distribution of a particular glycosidic linkage can be displayed by clicking the glycosidic linkage in ‘Sequence Graph’ in Figure 3.1. The clustering analysis of the glycan chain can be performed, and the representative structures from the five most populated clusters can be downloaded. The glycosidic torsion angle distribution of a selected cluster is shown in red.

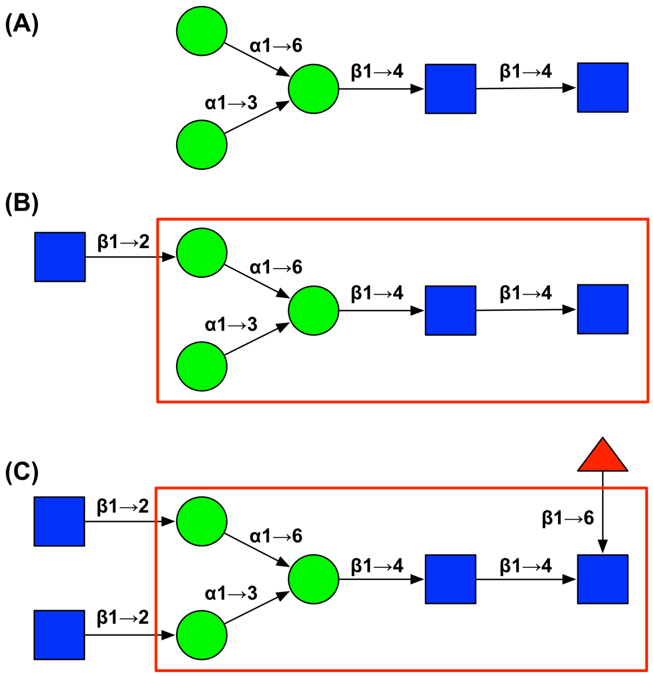


Figure 3.3: An example of the exact and fragment matches based on the query sequence in Figure 3.1. (A) The glycan sequence for the exact match results. (B and C) Examples of the glycan sequences for the fragment match results. The matched substructure is highlighted in the red rectangles. The sequence in (A) is also included in the fragment match results.

radius; the cutoff value was empirically determined. The second cluster is identified in the same manner after excluding the members that belong to the first cluster. The result of the five most populated clusters and the corresponding 3D glycan structure based on the centroid of each cluster are provided to the user along with the input files to generate the centroid glycan structures using the CHARMM biomolecular simulation program [17].

3.5 Discussion

There are several databases that provide information on glycan structures or sequences derived from the PDB (or from other experiments). Many of these databases, such as BCSDB [49], KEGG Glycan [45], and Glycoconjugate Data Bank [94], store only glycan sequence information, whereas the GFDB focuses on the 3D glycan structure. GlycoMaps DB [32] and GlyTorsion [81] provide torsion angle distributions of glycosidic linkages derived from computational calculations and from the PDB, respectively. Thus, the GlyTorsion database is the only database that can be directly compared to the GFDB. While the search interface of the GlyTorsion database is restricted to only one glycosidic linkage, the GFDB can search more complex glycan sequence with various filter functions and provide the clustering analysis and the representative structures from top five most populated clusters. These unique features in the GFDB allow researchers to collect complex glycan structural information easily and reliably.

As of August 2012, the GFDB contains 5,360 PDB entries that contain at least one carbohydrate molecule and 20,467 glycan chains. Among those glycan chains, 11,735 (57%) are N-linked glycan chains and 788 (4%) are O-linked glycans. And the remaining 7,944 (39%) exist as ligands. For the glycan structures with more than 2 carbohydrates, the hierarchical fragmentation identified a total of 81,370 fragment structures with 4,267 unique glycan sequences; a unique glycan sequence has more than 2 carbohydrates and is defined by the carbohydrate sequence and the glycosidic linkages. There are a total of 30,375 glycosidic torsion angles values available in the GFDB. By providing the straightforward search tool, the filtering functions, as well as the clustering analysis for the representative structures, we hope that the GFDB can help conformational analysis of various oligosaccharide chains and glycosidic linkages. The database will be updated quarterly and is freely available at `http://www.glycanstructure.org`.

Chapter 4

Restricted N-glycan Conformational Space in the PDB and Its Implication in Glycan Structure Modeling[1]

4.1 Introduction

Glycosylation represents one of the most important post-translational modifications [121, 136] and is ubiquitous in all domains of life. The glycosylation machinery is largely conserved in eukaryotes, and more than 50% of all eukaryotic proteins are expected to be glycosylated [7, 156]. An oligosaccharide moiety in a glycoprotein, referred to as a glycan, comes in a diversity of sequences and structures and is implicated in a vast array of biological processes [136]. The N-glycosylation pathway is the most common pathway in which an oligosaccharide is covalently attached to the side chain of asparagine [121]. In general, such an oligosaccharide appendage masks the protein surface, protecting the glycoprotein from degradation and nonspecific protein-protein interactions (reviewed in [48, 117, 142]). N-glycosylation also alters the biophysical properties in the vicinity of the glycosylation site and affects the folding rates and the thermal

[1] Reused from Jo, S., Lee, H. S., Skolnick, J., and Im, W. *PLoS Comp. Biol.* **9**, pp e1002946. with the permission under *Creative Commons License*

stability of the protein [123, 129]. Some N-linked oligosaccharides (N-glycans) are directly involved in specific molecular recognition events; e.g., lectins and antibodies can recognize specific N-glycans on viral envelope glycoproteins such as HIV gp120 [87, 100, 139, 147].

The impact of glycosylation on the structure of the parent protein and *vice versa* has been of great interest in structural glycobiology [21, 44, 91, 123, 143]. At this time, however, an understanding of which glycans are important components in protein function and how to modify these glycans to optimize the protein properties of interest remains an enigma. Therefore, knowledge of the structure and dynamics of N-glycans is central to understanding protein-carbohydrate recognition and its role in protein-protein interactions. An oligosaccharide chain is flexible in solution and has an ensemble of diverse conformations rather than a single well-defined structure [102, 132, 140]. The inherent flexibility of oligosaccharides often hinders crystallographic structure determination, and there are only a few crystal structures of oligosaccharides longer than 2-3 residues in the Cambridge Structure Database [3]. In contrast, there are many more crystal structures of glycoconjugates in the Protein Data Bank (PDB) [13], suggesting that the presence of protein residues may reduce the conformational freedom of oligosaccharides or even favor a certain conformation over others [78]. For example, the N-glycan conformations in the crystal structures of the Fc domain [24, 31, 55, 71, 83, 89, 97] exhibit remarkable similarity (Figure S4.1 in Supplementary Material), suggesting that the protein's structure around the glycan has an influence on the glycan's conformation.

The number of PDB entries containing carbohydrates has been steadily increasing, but obtaining the complete N-glycan structure remains challenging [78]. Mass spectrometric mapping of N-glycosylation sites is becoming common [156], providing information about glycosylation sites as well as the relative abundance of different glycoforms. In this context, computational modeling of N-glycan structures is an appealing approach to provide glycosylated protein structure models. In particular, a computational approach that can combine known glycoprotein structures and glycosylation information (i.e., glycosylation site, primary glycan sequence, and linkage information) would be very useful in a variety of applications in glycoscience. For successful template-based glycan structure modeling, it is essential to understand the conformational variability of an oligosaccharide chain when it is glycosylated. In addition, the influence of the protein residues around the glycosylation site can provide valuable insight into the design of new computational approaches that are optimized for glycoconjugates. Several structural database surveys have investigated the general features of N-glycosylation in terms of oligosaccharide and protein structures [16, 78, 81, 103, 104, 105, 143]. In these earlier studies, however, the oligosaccharide conformations were analyzed in terms of individual glycosidic torsion angles, making it difficult to recognize the actual structural variability of glycans en bloc. To the best of our knowledge, the conformational variability of N-glycans using the three-dimensional (3D) structures in the PDB has not been studied.

In this work, using the PDB crystal structures that contain N-glycans, we examined the conformational variability in various N-glycans. Using *Glycan Reader* [65], an automatic sugar recognition algorithm that we developed, all N-linked glycoprotein structures were obtained from the PDB and sorted by the N-glycan sequence. PDB entries with more than 3 Å resolution were excluded and N-glycan sequences with less than 20 PDB entries were also excluded, resulting in 35 N-glycan sequences (see the full list in Table S4.1 in Supplementary Material). Using random background conformations of each N-glycan sequence, the statistical significance of glycan structural similarity was estimated. The N-glycan structures in the PDB show statistically significant similarity when the local structure around the protein is conserved. When the local protein structures are different, N-glycan structures are not conserved, but their internal substructures appear to be strongly conserved due to the proximity to the protein. The results highlight the applicability of template-based approaches used in protein structure prediction to structure prediction and modeling of N-glycans of glycoproteins. Although the N-glycan sequences examined in this work mostly represent oligomannose-type glycans due to the limited numbers of crystal structures of complex- and hybrid-type glycans, the conclusions might be applicable to other glycoconjugates' glycan sequences.

4.2 Methods

4.2.1 N-Glycan structure dataset

Extracting structural information of glycans from the PDB is nontrivial due to a lack of standardized nomenclature and the way the data is presented in the PDB. To recognize the PDB entries that contain carbohydrate molecules, we used *Glycan Reader* for automatic sugar identification [65]. Briefly, in *Glycan Reader*, the topologies of the molecules in the HETATM section of a PDB file are first generated using the CONECT section of the PDB file, and the candidate carbohydrate molecules (a six-membered ring for a pyranose and a five-membered ring for a furanose that are composed of carbon atoms and only one oxygen atom) are identified. For each carbohydrate-like molecule, the chemical groups attached to each position of the ring and their orientations are compared with a pre-defined table to identify the correct chemical name for the carbohydrates. Glycan chains are constructed by examining the glycosidic linkages between the carbohydrate molecules that have chemical bonds between them. As of 2011 December, there were 2,517 PDB entries and 10,769 N-linked glycan chains in the RCSB database. The glycan fragment structure database, including the substructures of the original N-glycan chains, was generated, which resulted in a total of 48,568 N-glycan fragment chains.

From the N-glycan fragment database, we have collected glycan structures composed of more than 3 carbohydrate units. A glycan structure was excluded when its resolution was higher than 3 Å or when it had less than 20 structures in total, resulting in the 35 N-glycan fragment sequences listed in Table S4.1. A N-glycan structure pair is called "non-homologous" when the sequence similarity of the parent proteins is less than 30%. Because a glycoprotein can have multiple glycosylation sites in a single domain, if the distance between the backbone C_α atoms of the two glycosylated Asn residues is more than 10 Å after alignment of the glycoprotein chains using TM-align [153], the N-linked glycan structure pairs are considered "non-homologous" glycans. The rest of the N-glycan structure pairs are called "homologous" glycans. Figure 4.1 summarizes the protocol for building the N-glycan structure dataset.

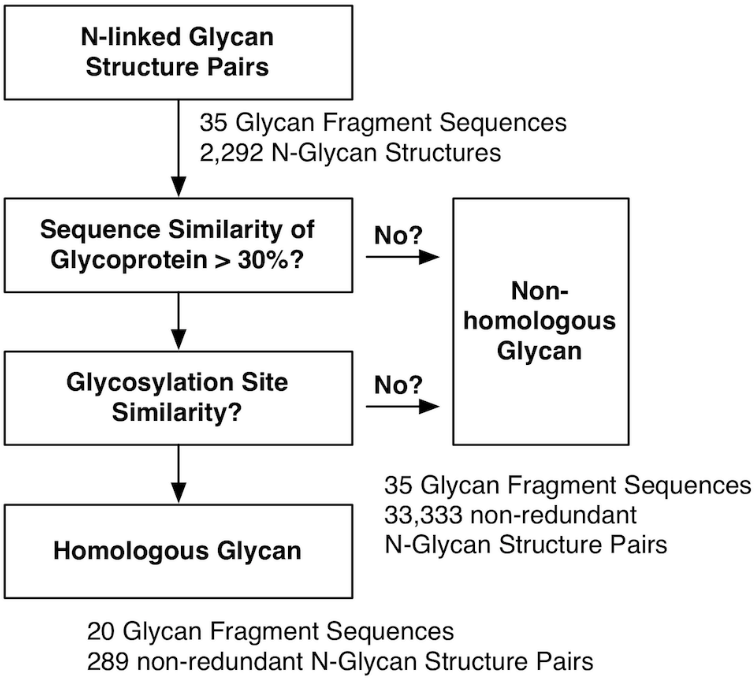

N-linked Glycan Structure Pairs
35 Glycan Fragment Sequences
2,292 N-Glycan Structures
Sequence Similarity of Glycoprotein > 30%?
No?
Non-homologous Glycan
Glycosylation Site Similarity?
No?
35 Glycan Fragment Sequences
33,333 non-redundant N-Glycan Structure Pairs
Homologous Glycan
20 Glycan Fragment Sequences
289 non-redundant N-Glycan Structure Pairs

Figure 4.1: Protocol for building the N-glycan structure dataset

4.2.2 Generation of random glycan conformation pool

To quantify the conformational variability of the PDB N-glycan structures, it is essential to know the upper bound of the conformational variability in a given oligosaccharide. In protein structural biology, the upper bound of conformational variability is estimated by using the non-homologous protein structure pool and sequence-independent structure alignment methods [14, 74, 146]. However, because such sequence-independent structure alignment methods are not available for oligosaccharides, it is difficult to estimate the upper bound of the conformational variability in oligosaccharides only using the crystal structures in the PDB.

Instead of using the crystal structures directly, a conformational pool that contains diverse conformations of a specific N-glycan sequence was generated as follows. For each of the 35 N-glycan sequences, a total of 1,000,000 glycan conformations were generated in an iterative fashion. The initial structures were generated by using the IC BUILD command in the CHARMM biomolecular simulation program [17] according to the glycan sequence. For each iteration, a glycosidic linkage was randomly selected and a new torsion angle value was also randomly chosen based on the accessible glycosidic torsion angles of the corresponding glycosidic linkage type. If the newly generated conformation had bad contacts with neighboring atoms, the conformation was rejected and the protocol was repeated until no bad contacts were found. If a conformation had no bad contacts, the conformation was recorded and the protocol repeated until 1,000,000 conformations were generated. A bad contact was defined by the CHARMM van der Waals energy higher than 10 kcal/mol. Accessible glycosidic torsion angle values were used rather than the values observed in the PDB because the number of observations is limited for certain types of glycosidic linkages. For example, Figure S4.6 in Supplementary Material shows the resulting glycosidic torsion angle distributions of the N-glycan core sequence using the accessible glycosidic torsion angle values, and Figure S4.7 shows the torsion angle values observed in the PDB, respectively.

To construct an accessible glycosidic torsion angle map, a total of 13 adiabatic (ϕ, ψ, ω) potential maps were constructed for each distinct glycosidic linkage type found in the 35 N-glycan sequences. For each glycosidic linkage type, a disaccharide connected by the corresponding glycosidic linkage type was generated by CHARMM [17], and the CHARMM carbohydrate force field [41, 42, 43] was used to evaluate the energy. The adiabatic map was generated by evaluating the energy over a grid of glycosidic torsion angles with a grid spacing of 5° resulting in a total of 373,248 grid points for ($1\rightarrow6$) linkages (ϕ, ψ, ω) and 5,184 grid points for the rest of the glycosidic linkages (ϕ, ψ). At each grid point, the conformations were minimized with the dielectric-screened Coulombic electrostatic and Lennard-Jones potential energy while the glycosidic torsion angles were restrained and a harmonic restraint potential was applied to the carbohydrate rings to prevent the distortion of the ring geometry. The generated adiabatic potential energy map was converted to a torsion angle probability map using the Boltzmann distribution. Finally, the resulting distribution was compared with the glycosidic torsion angles observed in the PDB using the Glycan Fragment DB [60], available at `http://www.glycanstructure.org`. The glycosidic torsion angle probability maps and the observations in the PDB matched well in general. However, the torsion angle probability map was clearly more restricted (data not shown). To remedy the restricted conformational space, glycosidic torsion angle pairs having probability above 0.0001 were considered “accessible”; this covers on average about 65% of the observed glycosidic torsion angles in the PDB.

4.2.3 Structural similarity of N-glycan and its statistical significance

The N-glycan structural similarity was measured by calculating pairwise RMSD in the following three different ways: First, the heavy atoms in the carbohydrate ring (C1, C2, C3, C4, C5, and O5) were used for the alignment of two N-glycan structures and in the RMSD calculation. Second, to examine the variability of N-glycan orientations with respect to the protein, the heavy atoms of glycosylated Asn residues were used to define the alignment, and then the Euclidean distance of the N-glycan structures was calculated using the carbohydrate ring heavy atoms. Third, many crystal structures only have a few residues at the glycosylation site due to difficulties associated with glycan crystal structure determination, and these partial glycan structures can be used to model the rest of a full glycan structure. To examine the efficacy of such an approach in obtaining a better N-glycan orientation with respect to the protein, the carbohydrate ring heavy atoms of the first two residues were used for the alignment of N-glycan structures, and then the Euclidean distance of the N-glycan structures was calculated using the ring heavy atoms excluding the first two residues.

The statistical significance of structural similarity between two glycan structures was estimated by comparing the structural similarity of 124,750 random glycan structure pairs for each N-glycan sequences. The structural similarity of random glycan structure pairs was calculated by the identical procedure described above. Using the statistical model, p-values of the corresponding structural similarity measure can be calculated and allows us to compare structural similarity across different sequences and length. Each RMSD distribution for each glycan sequence was modeled by the generalized extreme value distribution,

$$P(z) = \frac{1}{\sigma}\left[z(x)^{\xi+1}e^{-z(x)}\right] \tag{4.1}$$

where $z(x) = (1+\xi(x-\mu)/\sigma)^{-1/\xi}$. The variable x represents the RMSD of a structure pair; μ, σ, and ξ are the location, scale, and shape parameters, respectively. These parameters were obtained through the maximum likelihood estimates by the EVD package in R (`http://www.r-project.org`). 35 sets of determined parameters are given in Table S2 and the fitting results are shown in Figure S4.2. The resulting goodness of fit (χ^2) are generally good except for a few sequences. The correlation coefficients improved when more "liberal" protocols were used (e.g., the glycosidic linkage is not restricted and more tolerance to bad contact; data not shown). However, such protocols may produce unrealistic random glycan conformers and are not used in this work. The p-value of a glycan structure pair from the PDB having RMSD values smaller than the random glycan conformation background was calculated by

$$p - \text{value} = \begin{cases} e^{-z(x)}, & z \geq 0 \\ 1, & z < 0 \end{cases} \tag{4.2}$$

4.2.4 Local structure alignment and statistical significance

The local protein structures are defined for protein residues having any heavy atoms within 6 Å from any glycan heavy atom. The local protein structures were derived from the PDB structure files in our dataset, and the TM-align algorithm [153] was used to compare the structural similarity of a given local protein structure pair. Any local protein structures having less than 5 residues were excluded. The TM-scores calculated by TM-align are normalized by the length of the smaller structure. To estimate statistical significance, we have derived a random local protein structure pool using the N-linked glycoproteins in the PDB. Briefly, from the PDB, a non-redundant N-linked glycoprotein structure list having at least one carbohydrate residue and protein sequence similarity less than or equal to 30% were generated. A random local protein structure pool was derived from the protein residues having any heavy atom within 6 Å from any of the carbohydrate heavy atoms. The TM-align algorithm was used to calculate the distribution of TM-scores from the random local protein structure pairs. The calculated TM-scores were fit using the generalized extreme distribution (Eq. (1)), and the p-values of having TM-scores larger than the random background were estimated using Eq. (2).

Although there are several local structure alignment tools available [36, 70, 115], it was difficult to directly utilize them in this study because many of them are highly customized to specific domains, such as a protein-protein interface or protein-ligand interface. Thus, we used TM-align [153] to compare local structure similarity. Although TM-align is not designed to compare local structure similarity, it performed well in our internal testing and correctly found most homologous glycoproteins having similar local protein C_α structures; also see ref [73].

4.2.5 Structural similarity of internal substructure and the statistical significance

The residue distance is defined as the minimum number of glycosidic linkages between carbohydrate monomers, including the glycosidic linkage to Asn. For each of 35 N-glycan sequences, three types of internal substructures were generated; a) residue distance up to 3, b) residue distance up to 4, and c) residue distance up to 4, excluding residues linked by the 1-6 linkage. Then, the RMSD of substructures were measured after alignment using the carbohydrate ring atoms in the substructure. To estimate the statistical significance of the internal substructures, the random glycan internal structure pool was generated for each of three different types of substructures. The resulting random background distributions were fit using Eq. (1) and p-values were calculated using Eq. (2).

4.3 Results

Because glycan sequences have branches and different linkages between monomers, alignment of glycan structures with different sequences is challenging. Therefore, in this study, pairwise structure similarity is measured using the root-mean-squared deviation (RMSD) among glycan structures having the identical glycan sequence. Assuming that homologous protein structures share similar surface features, the structural similarity of glycans found on homologous proteins would provide insight into the influence of the protein structure on the N-glycan structure. Therefore, N-glycan structure pairs with the identical glycan sequence are designated as "homologous" or "non-homologous" depending on the sequence similarity of their parent protein (with a sequence similarity of 30% as a cutoff). Unless stated explicitly, highly homologous pairs (sequence similarity $\geq$ 90%) as well as redundant structure pairs were excluded from the analysis. There are a total of 289 homologous and 33,333 non-homologous glycan structure pairs in the final dataset (see Figure 4.1 and Methods for details). In this section, the N-glycan structural similarity is examined and its statistical significance is estimated using random background conformations of each N-glycan sequence (see Methods for details). The structural similarity of the N-glycans is then discussed in terms of the protein's structure as well as the structural rigidity of the oligosaccharide regions that are closer to the glycosylation site on the protein.

4.3.1 N-glycan structures on the surface of homologous proteins are significantly conserved

The structural similarities of the N-glycans are measured by calculating the glycan RMSD after alignment of the oligosaccharide structures using the carbohydrate ring heavy atoms. N-glycan structural similarity including their orientations with respect to the protein is discussed separately below. Figure 4.2 shows the RMSD distributions of the N-glycan structure pairs in the PDB and random conformation pool. Note that the RMSD is only measured between glycan structures having an identical sequence. The average RMSD of all PDB structural pairs are 1.4 ± 0.8 Å. The homologous and the non-homologous N-glycan structure pairs have RMSD values of 0.9 ± 0.8 Å and 1.4 ± 0.8 Å, respectively. Both the homologous and non-homologous N-glycans showed smaller RMSD values compared to those in the random glycan structure pool whose RMSD is 2.4 ± 0.8 Å (Figure 4.2A).

Measuring the structural similarity using RMSD is straightforward, but it is not an objective measure when comparing structures of different lengths and sequences due to its length

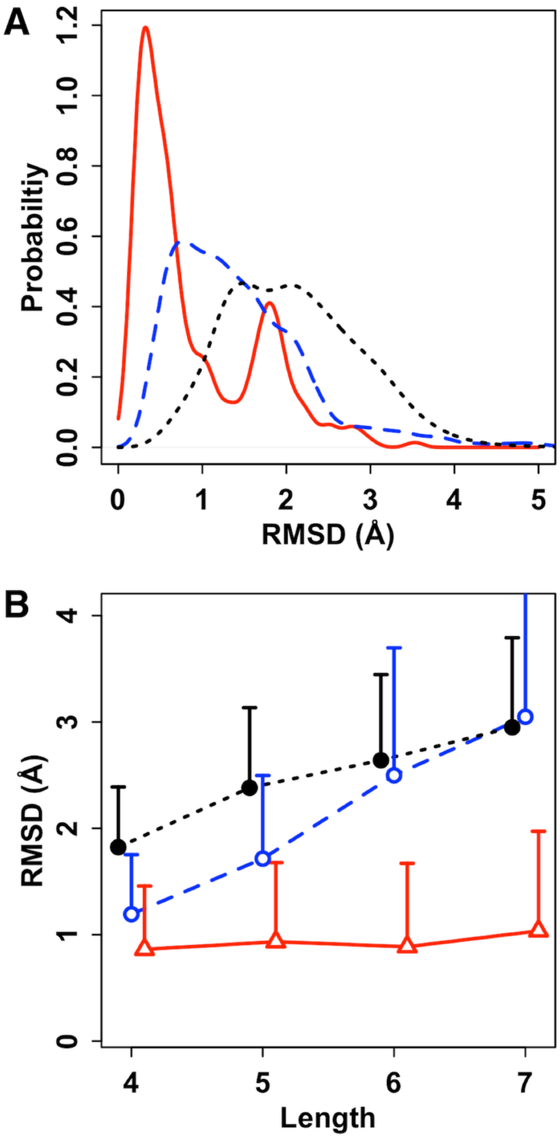


Figure 4.2: N-glycan structure similarity. (A) The RMSD distributions from the homologous (red), non-homologous (blue), and random glycan structure pairs (black). A 0.1-Å bin width was used. (B) Length dependence of average RMSD values from homologous (red), non-homologous (blue), and random glycan structure pairs (black). The length of a glycan chain is defined as the number of residues in the glycan chain. Error bars are the standard deviations and only the upper sides are displayed for clarity. Each data point is slightly shifted for clarity. Red and blue colors represent the homologous and non-homologous N-glycans, and the same color scheme is adopted throughout the figures unless stated otherwise.

N-glycan Length †	p-value	5×10^{-1}	1×10^{-1}	5×10^{-2}	1×10^{-2}
Overall	RMSD	2.4 Å	1.8 Å	1.5 Å	0.9 Å
7		2.9 Å	1.9 Å	1.6 Å	1.2 Å
6		2.6 Å	1.7 Å	1.5 Å	1.1 Å
5		2.3 Å	1.4 Å	1.2 Å	0.8 Å
4		1.8 Å	1.4 Å	1.0 Å	0.7 Å

† The N-glycan length is defined as the number of residues in the glycan chain

Table 4.1: Statistical significance of the RMSD values for the PDB N-glycan pairs

dependence. When the average RMSD values of the N-glycans are plotted against N-glycan length, i.e., the number of carbohydrate monomers (Figure 4.2B), a length dependence is observed for the random background and non-homologous glycan pairs, but homologous glycan pairs do not show such a length dependence. The smaller RMSD values of the homologous N-glycan structure pairs compared to the RMSD values of the non-homologous pairs indicate that the homologous N-glycan structures are more conserved than the non-homologous N-glycan structures.

Because our dataset contains different lengths of N-glycan sequences with different branching patterns (Table S4.1), we converted the RMSD values to their statistical significance (p-values) using the random background glycan structures (see Methods for details). By deriving the statistical significance using the random background having the identical N-glycan sequence, the length dependence is effectively removed. The generalized extreme value distribution (Eq. 1 in Methods) was used to estimate the statistical significance [59], and 35 sets of parameters were determined by fitting the generalized extreme value distribution to the original RMSD distribution of the random conformational pool of each glycan sequence (see the determined parameters in Table S4.2 and the fitting results in Figure S4.2). The calculated p-values (Eq. 2 in Methods) represent the probability of having randomly chosen two N-glycan structures whose RMSD is smaller than the random background. A list of p-values and the corresponding RMSD values averaged over different sequences are given in Table 4.1.

Figures 4.3A and 4.3B show the cumulative fraction of homologous and non-homologous glycans structure pairs as a function of their p-value. It is clear that about 67% of the homologous

N-glycan structure pairs have a statistically significant level ($p < 0.05$) of structural similarity, whereas about 36% of non-homologous N-glycan structure pairs have a statistically significant level of structural similarity. A correlation is also found between the sequence similarity of the glycoprotein and the structural similarity of the N-glycan (Figure S4.3). Specifically, about 81% and 91% of N-glycan structure pairs have statistically significant structure similarity ($p < 0.05$) when the parent proteins have sequence similarity greater than 50% and 60%, respectively. A similar analysis has been carried out independently using the global distance test (GDT) score [149] instead of RMSD, and the conclusion remains the same (Figure S4.4). Assuming that the proteins with similar sequences have similar surface features around the glycosylation site, such a high level of N-glycan structure similarity strongly indicates that the protein structure around the N-linked oligosaccharide plays an important role in determining the N-glycan structures.

Apparently, not all homologous glycans have significant structural similarity. Figure 4.4A shows an example of two homologous proteins, the Fc domain of IgG (PDB:2WAH) in green and the Fc domain of IgE (PDB:3H9Y) in cyan, which share a sequence similarity of about 50% and have significantly different glycan structures (RMSD of 2.9 Å and p-value of 0.6). The structures of these two homologous proteins around the glycosylation site are similar and well aligned. Notably, the structural difference of the N-glycans arises mainly from the terminal residues at the 1-6 branches (or 1-6 arm). The PDB:2WAH IgG-Fc domain is glycosylated with a different glycoform than typical IgG-Fc glycans whose 1-6 arm carbohydrates are tightly packed with the proteins [24, 31, 55, 71, 83, 89, 97]. This may explain such a different glycan conformation in PDB:2WAH.

There are some non-homologous N-glycan structure pairs that have a statistically significant level of structural similarity ($p < 0.05$). Visual inspection of several examples of non-homologous glycoproteins having similar N-glycan conformations shows no apparent similar protein surface features around the N-glycans. Figure 4.4B shows an example of two non-homologous glycoproteins, beta-galactosidase (PDB:3OG2) in green and the extracellular domain of the nicotinic acetylcholine receptor 1 subunit (PDB:2QC1) in cyan, having a significant level

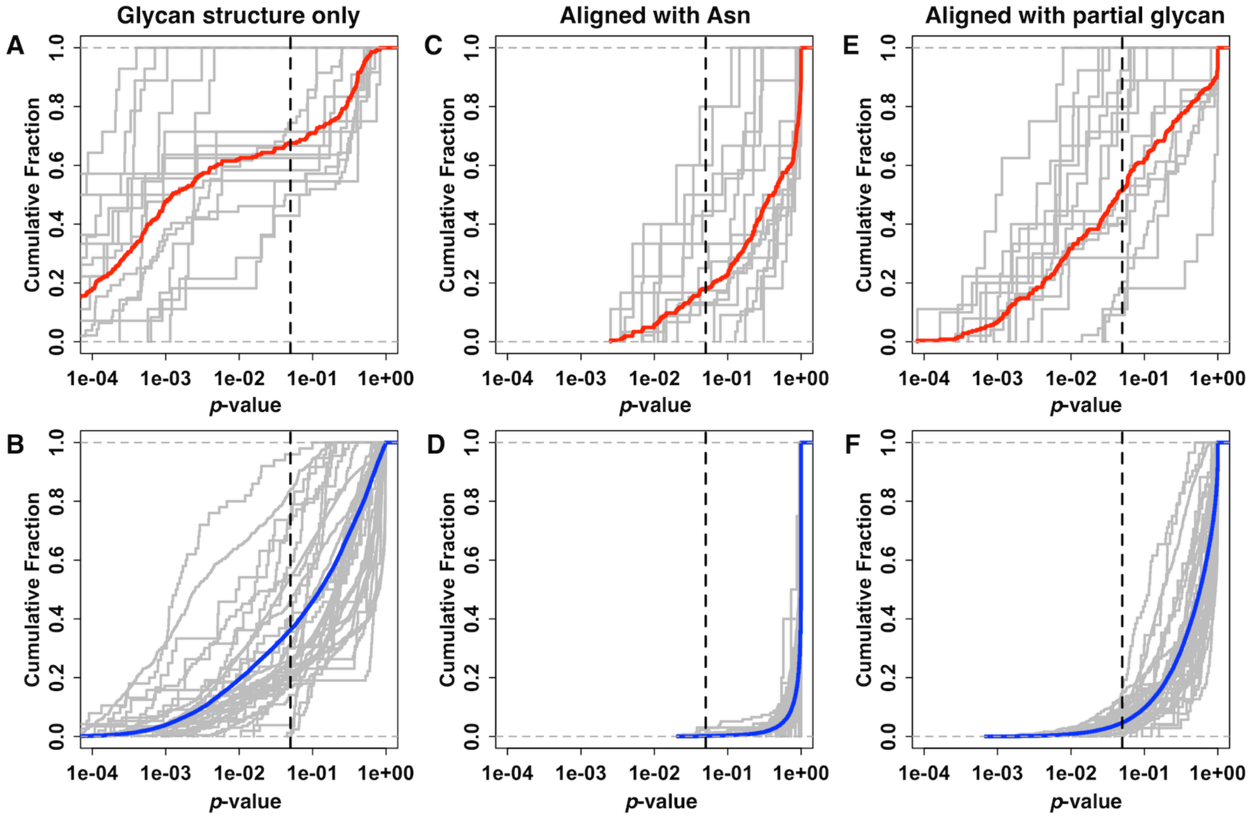


Figure 4.3: Cumulative fraction of glycan structure similarity using p-values. (A-B) Structural similarities of (A) homologous and (B) non-homologous glycans after alignment of glycan structures themselves. (C-D) Structural similarities of (C) homologous and (D) non-homologous glycans after alignment of glycosylated protein Asn residues. (E-F) Structural similarities of (E) homologous and (F) non-homologous glycans after alignment of the first two residues of the glycan chain. The gray lines in each plot represent the structural similarity of individual glycan sequences and the thick solid lines represent the average of cumulative fractions of all 35 N-glycan sequences. The vertical dotted line is drawn at a p-value of 0.05.

of structural similarity of the N-glycan (RMSD of 0.9 Å and p-value of 0.009). Nonetheless, the structure alignment of these two N-glycans results in a poor alignment of the parent proteins.

4.3.2 N-glycan orientations with respect to the protein are diverse even in homologous glycoproteins

The relative orientation of an oligosaccharide chain with respect to the parent protein can be affected by the Asn side chain conformation and the protein conformation in the vicinity of the glycosylation site. To examine N-glycan structural variability with respect to the parent protein, the heavy atoms of the glycosylated Asn residue were used for alignment, and then the Euclidean distances of the glycan portion were measured without further alignment. Figures 4.3C and 4.3D show the cumulative fraction of structure similarity of the homologous and non-homologous glycans aligned with glycosylated Asn residues. Clearly, structural similarity is greatly reduced when the Asn residues are used for the alignment. Given the fact that glycosylation has a bias towards turns and extended regions [103], it is not surprising that even homologous N-glycans show reduced structural similarity when the Asn residues are used for the alignment.

The observations so far indicate that a comparative modeling approach for N-glycan structures would successfully predict the N-glycan structure itself, especially when the homologous N-glycan templates are present in the PDB, but finding the global orientation of the glycan with respect to the protein would remain challenging. Such difficulties can be significantly alleviated when a partial glycan structure is available. In fact, there are large numbers of partial N-glycan structures available in the PDB, probably due to the removal of glycans prior to structural studies, due to crystallization conditions, or due to missing electron density resulting from flexible glycan structures. For example, as of 2011 December, there were 2,517 PDB entries and 10,769 N-linked glycan chains in the RCSB database; 84% (9,027 chains) had partial glycan structures with less than two carbohydrate units and 15% (1,394 chains) of such partial structures showed their parent protein sequence similarity less than 50%. Assuming that one can find such partial glycan structures, Figures 4.3E and 4.3F show the cumulative structural similarity of the N-glycans when

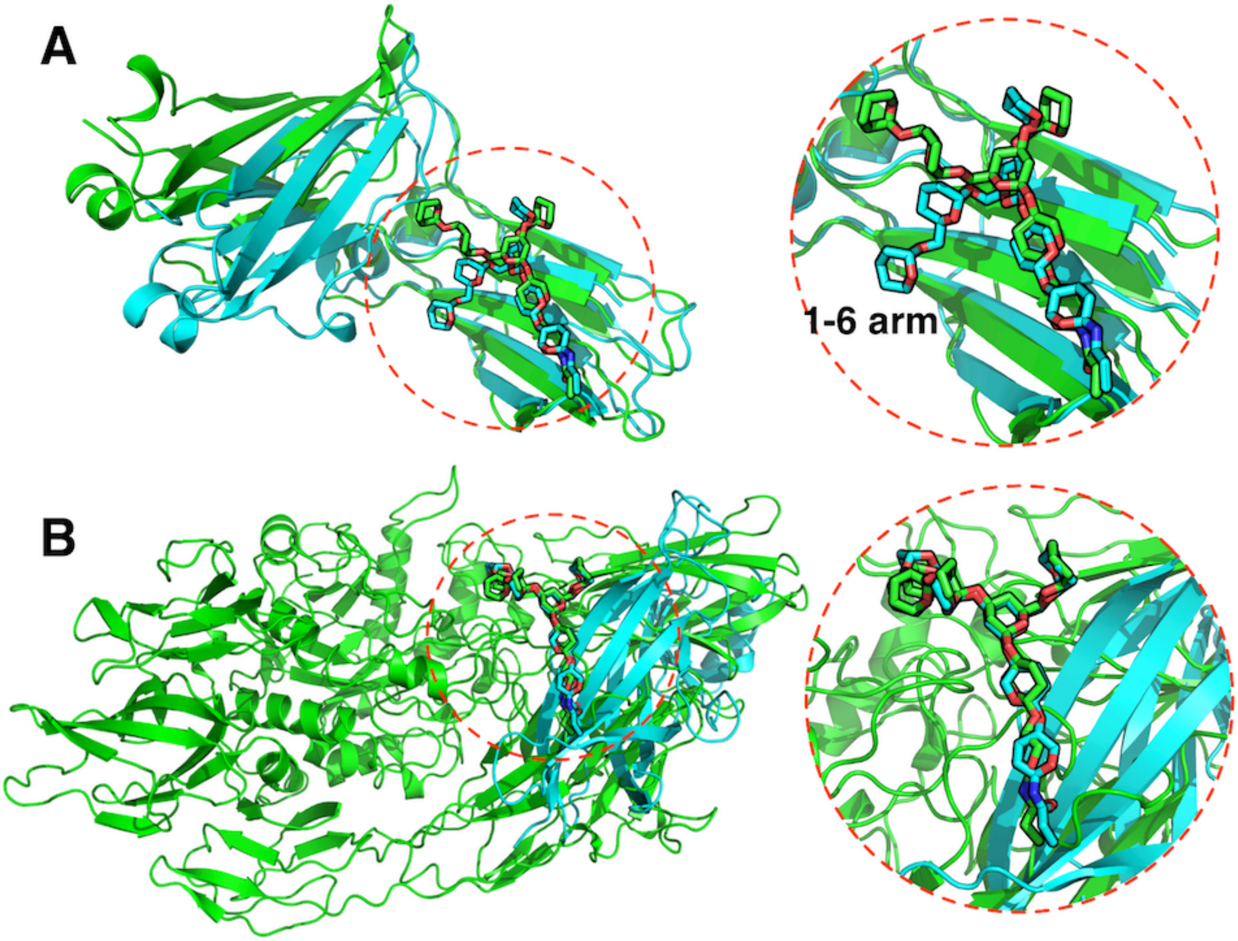


Figure 4.4: Examples of N-glycan structure pairs. (A) An example of homologous glycoproteins having dissimilar glycan structures. The IgG-Fc domain (PDB:2WAH) is drawn in green and the IgE-Fc domain (PDB:3H9Y) is drawn in cyan. The RMSD of the two oligosaccharides is 2.9 Å. Figures on the right-handed side are the detailed illustration around the N-linked oligosaccharides. Hydroxyl groups of the oligosaccharides are removed for clarity. (B) An example of non-homologous glycoproteins having similar glycan structures. The beta-galactosidase (PDB:3OG2) is drawn in green and the extracellular domain of the nicotinic acetylcholine receptor 1 subunit (PDB:2QC1) is drawn in cyan. The RMSD of the two oligosaccharides is 0.9 Å.

the first two carbohydrate units in the glycan chains are aligned. Both the structural similarities of the homologous and non-homologous N-glycan structures significantly increased, suggesting that the conformations of glycosylated Asn residues and the first few carbohydrates of the N-glycan are important in determining the N-glycan orientations.

4.3.3 The local structure around the glycoprotein influences the N-glycan conformation

What makes homologous N-glycan structures conserved compared to non-homologous N-glycans or random background? Possibly, the protein structures around the glycan may provide a steric barrier, thus restricting the conformational freedom of N-glycans nearby. In addition, specific protein-carbohydrate interactions may play an important role in favoring a certain conformation of the oligosaccharides. If local protein structure around the N-glycan is directly correlated with the N-glycan structure similarity, such information provides valuable criteria in N-glycan structure modeling.

Figure 4.5 shows the correlation between the local protein structure around the glycan chain and the N-glycan structure similarity. As expected, most homologous glycoproteins have similar local protein structures around the glycan chain. However, some homologous N-glycan structure pairs adopt significantly different conformations while their local protein structures are similar (p-RMSD > 0.05 and p-local < 0.01). Visual inspection of such structures shows that the structural differences are mainly due to the terminal residues, especially ones in the 1-6 branches, similar to the case in Figure 4.4A. The increased flexibility of the 1-6 linkage is not surprising because the 1-6 glycosidic linkage contains three rotatable torsional angles (compared to two for other glycosidic linkages), and the flexibility of the 1-6 linkage has been well documented by other experimental, computational, and structural database surveys [16, 98, 101, 112, 141].

To examine the flexibility of different regions of N-glycan structures, we have used the GDT chart [149]. Figure 4.6 shows two example N-glycan sequences and the corresponding GDT charts, where each bar represents an alignment of an N-glycan pair and the bar is colored according to how

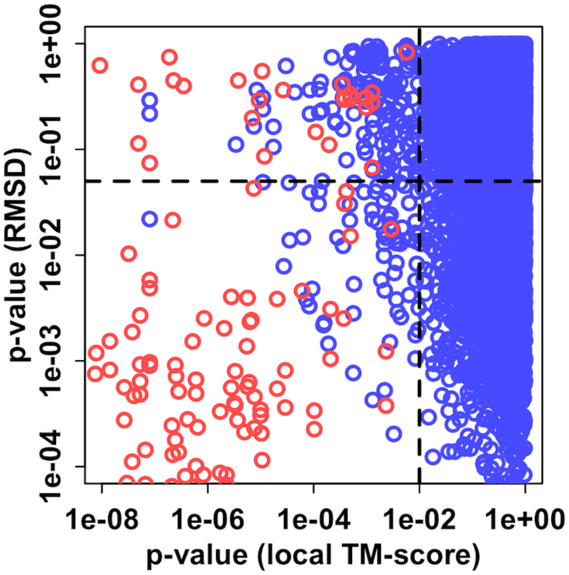


Figure 4.5: Correlation of local protein structure around the N-glycan and the N-glycan conformation. Red circles represent homologous glycan structure pairs and the blue circles are for non-homologous glycan structure pairs.

well a certain region of the sequence can be aligned each other. Clearly, the increased flexibility of terminal residues is apparent and, in particular, the residues in the 1-6 branches are even more flexible.

Non-homologous N-glycan structures in the PDB do not show a correlation with local protein structure around the glycan. There could be several factors responsible for this observation, and the accuracy of local protein structure alignment might be one important factor. To compare the similarity of local protein structure, TM-align [153] was used because the algorithm is general and performed well compared to other local structure algorithms available in our internal testing [73]. However, the TM-align algorithm was developed for comparison of global protein structure, and it is possible that the algorithm is insensitive to the structural similarities of the small number of residues around the glycan chain. Thus, further in-depth investigations with robust local structure algorithms are warranted.

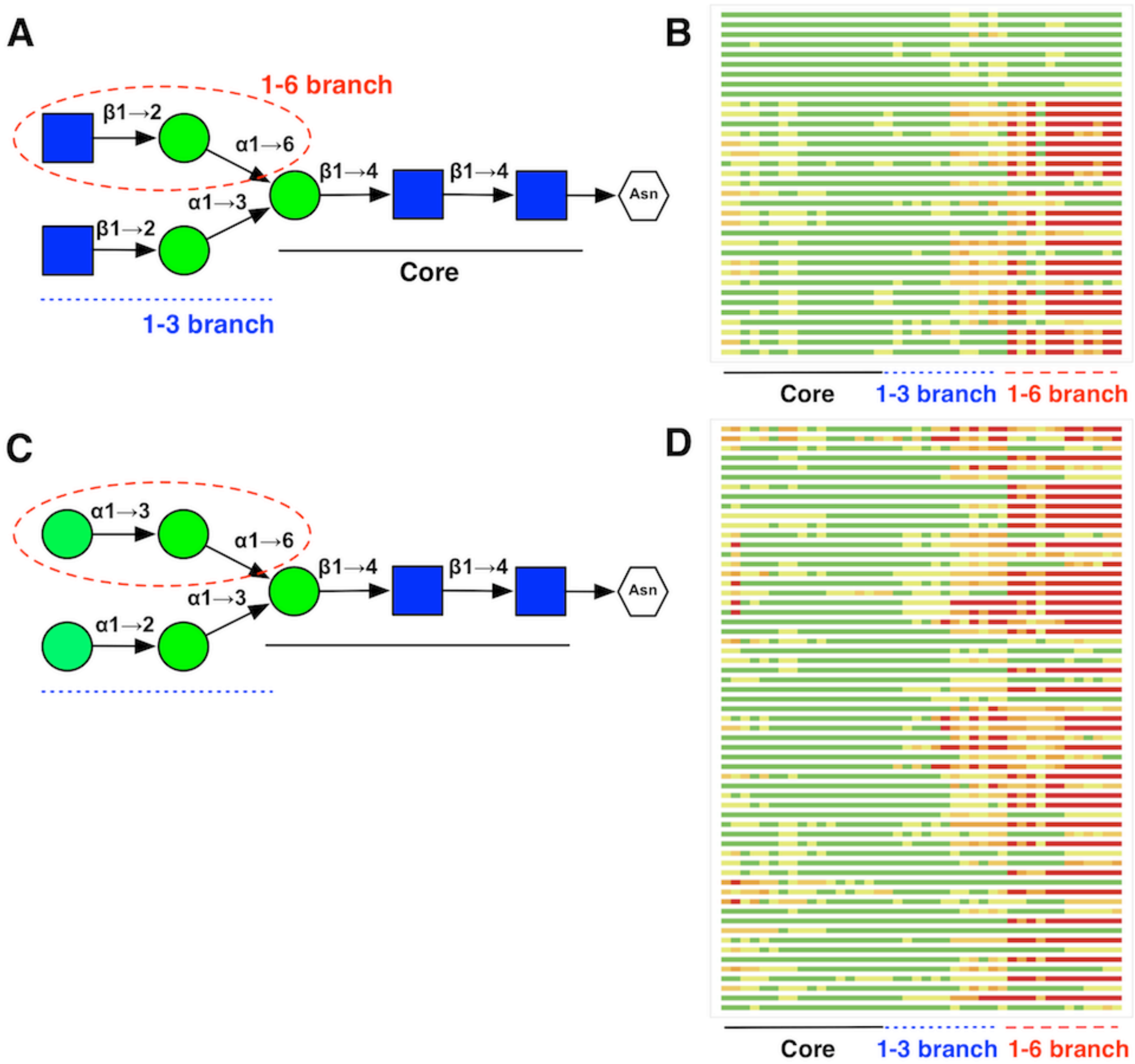


Figure 4.6: Structural flexibility within N-glycan chains. Two example N-glycan sequences (A and C) and the GDT charts (B and D) for the corresponding N-glycan structure pairs in the PDB. Each horizontal bars represents the distance deviation of carbohydrate ring atoms for different N-glycan structure pairs. Atoms superimposed below 1, 2, 3, and 4 Å are colored in green, light green, light orange, orange, and red, respectively. The atoms in the 1-6 branch of the sequence are aligned to be at the end of the GDT charts (highlighted with dashed red line).

4.3.4 Internal substructures of N-glycan structures in the PDB are conserved

The lack of correlation between the local protein structure and non-homologous glycan structures suggests that the gapless threading approach to N-glycan modeling would be inapplicable when no homologous templates are present. It was reported that the majority of glycosylation sites are found to be in convex or flat regions of the protein surface [103]. When the N-linked oligosaccharides are situated in such regions, the terminal residues of a long oligosaccharide may not be able to interact with the protein surface residues, and experience a smaller influence of the local protein environment. Thus, local protein structure around glycan chains might have a stronger impact on the first few residues of the glycan chain rather than on the global structure.

Internal substructure conservation can be visualized with the two examples in Figure 4.6, showing that the flexibility of the carbohydrate residues increases as the residues move away from the protein. In addition, a large increase in flexibility is observed after the 1-6 linkage, which is known to be flexible. If the N-glycan substructure is more conserved, a threading or fragment assembly approach could be useful to model the N-glycan structures. To quantify the conservation of internal substructures, we compared the structural similarity of the N-glycans as a function of glycan chain length from the protein. Figure 4.7A shows the average RMSD of N-glycan internal substructures containing only the residues within the given residue distance from the Asn residue of the parent protein. The conservation of the internal substructure is apparent up to 3 or 4 residues away from the Asn residue. Note that N-glycan sequences can have branches, and thus, there could be more residues in a substructure within a certain residue distance. For example, in the two examples in Figure 4.6, there are in fact 5 sugar residues at a residue distance of 4 from Asn.

To avoid the inherent length dependence of RMSD (i.e., a smaller substructure has a smaller RMSD), RMSD values for the substructures are converted to p-values using the random background. Figure 4.7B and 4.7C show the cumulative fraction of the substructure similarity for homologous and non-homologous N-glycans, respectively. About 80% and 60% of the substructure up to a residue distance of 3 (black curve) show significant structural similarity

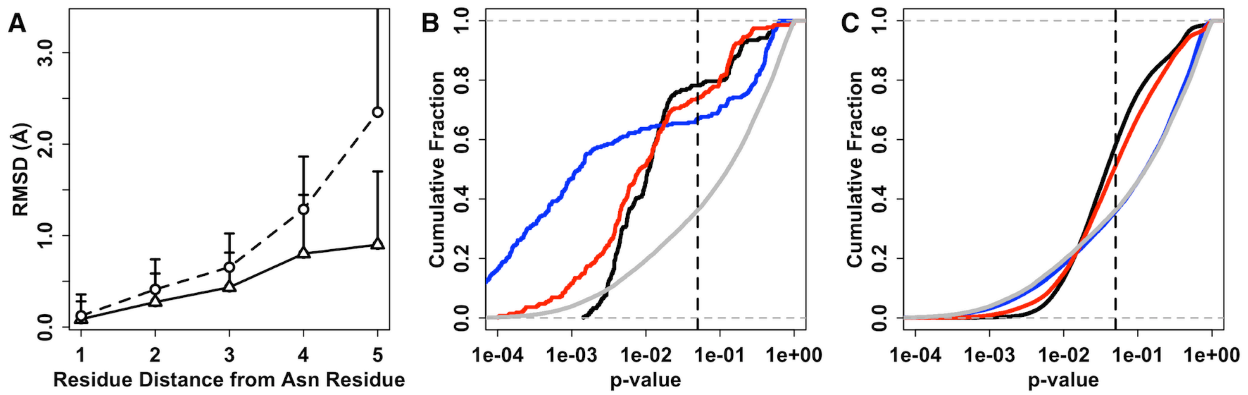


Figure 4.7: Structural similarity of N-glycan internal substructures. (A) RMSD of the internal substructures composed of residues within a certain distance from the protein. The distance is measured by the number of glycosidic linkages in a N-glycan chain including the glycosidic linkage to Asn. The lines are labeled for homologous (solid line) and non-homologous (dashed line). (B and C) Cumulative fraction of internal substructure similarity (p-value) for (B) homologous and (C) non-homologous glycans, respectively. The average substructure similarity of residues up to a distance of 3, 4, and 4 (without the 1-6 linkage) are colored in black, blue, and red, respectively. The gray curve represents the average substructure similarity of the overall N-glycan structure pairs.

for homologous and non-homologous N-glycans, respectively. The substructures are less conserved when residues up to a distance of 4 are included in the substructure (blue curve). As discussed above, due to its flexibility, the 1-6 linkage might contribute to the diversity of the N-glycan substructures more than other glycosidic linkages. Clearly, when structural similarity of substructures up to a residue distance of 4 is compared without residues linked by the 1-6 linkage (red curve), significant structural conservation is observed even for non-homologous N-glycans. This observation implies that the glycan residues closer to the protein surface have more restricted conformational space and conserved structures.

4.4 Discussion

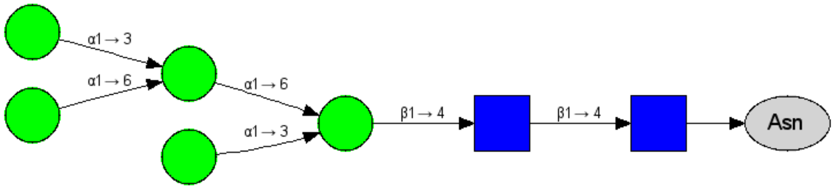
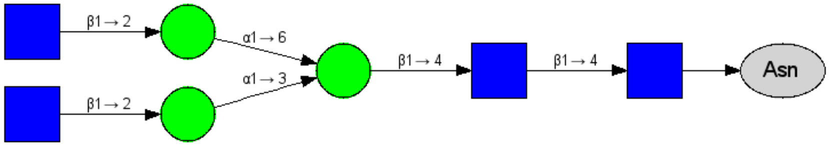
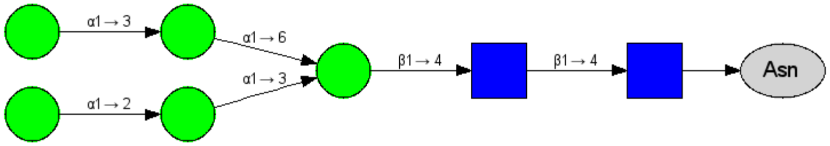
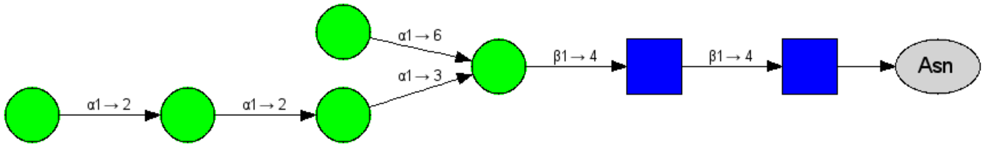
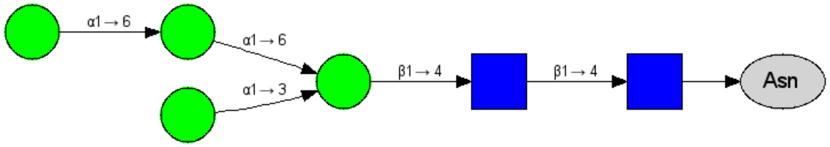
Elucidation of the factors influencing the conformational variability in N-glycans is essential to understand the dynamics of N-glycans and provides valuable insight into modeling and computational studies of the N-linked oligosaccharides. In this work, we have shown that the conformations of homologous N-glycans are restricted compared to the random background. About 67% of the homologous N-glycan pairs and 37% of the non-homologous N-glycan pairs show statistically significant level of structural similarity. Although excluded from the main analysis, more than 90% of highly homologous N-glycan structure pairs (protein sequence similarity $\geq$ 90%) show very significant structural similarity (Figure S4.7).

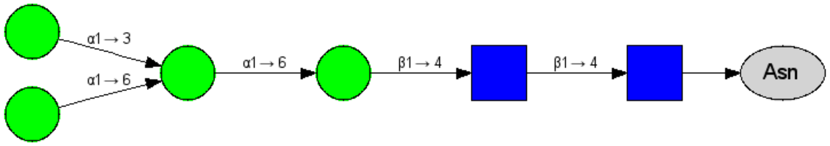
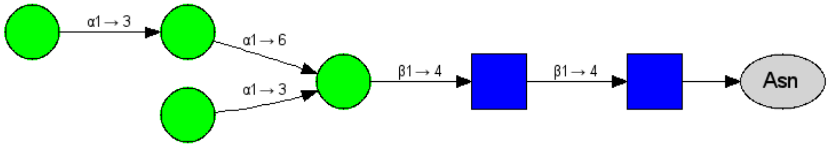
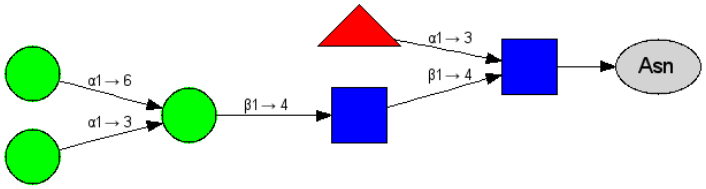
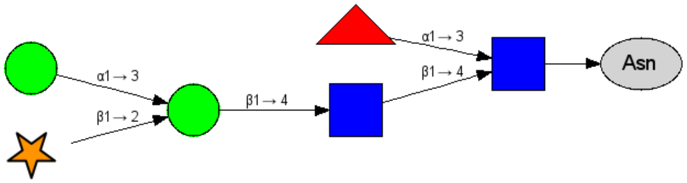
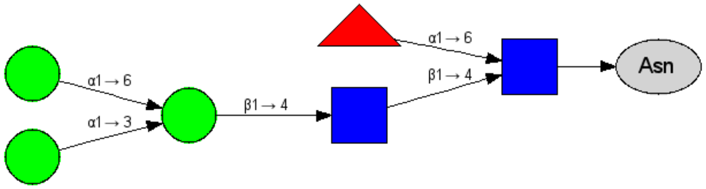
Why do homologous N-glycans have conserved conformations compared to the free oligosaccharides? First, protein-carbohydrate interactions may restrict the conformational freedom of the N-glycan. In addition, the shape of the local protein structure may also act as a non-specific steric barrier and restrict the N-glycans to adopt certain conformations. Lastly, crystallographic bias in the dataset could also play a role in conformational similarity of homologous N-glycan structures. Our dataset is composed of crystal structures of well-resolved N-glycan structures; hence, flexible N-glycan structures may not be included in our dataset.

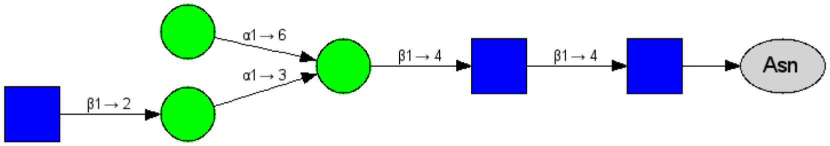
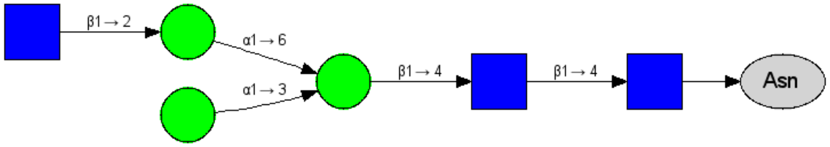
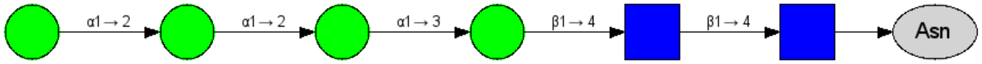
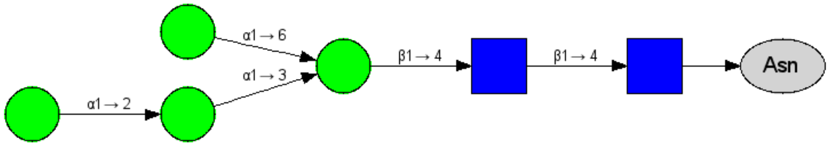
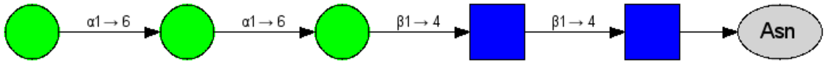
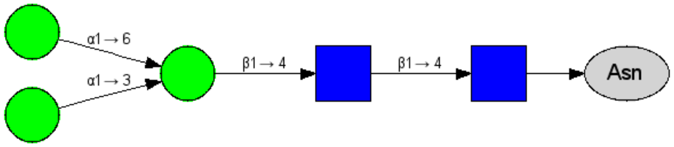
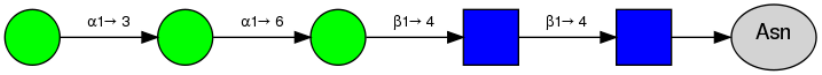
Despite the biological importance of N-glycans, understanding the structure and dynamics of N-glycans is currently lacking due to the difficulties in crystallization of glycoproteins and other experimental techniques. The high level of structural similarity among the N-glycan structures found on the surface of homologous proteins strongly indicates that the comparative modeling and threading approach used in protein structure prediction [10, 151, 152] might perform well in glycan structure modeling if appropriate templates are present. Despite the structural similarity of N-glycans on the homologous glycoproteins, the absolute orientation of N-glycan with respect to the glycosylated Asn residue may differ because the glycosylation site are often found on the loop regions.

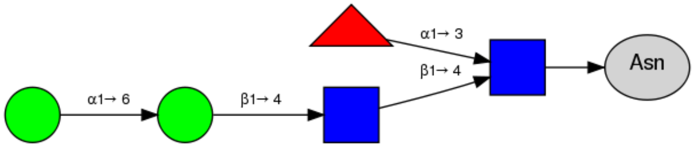
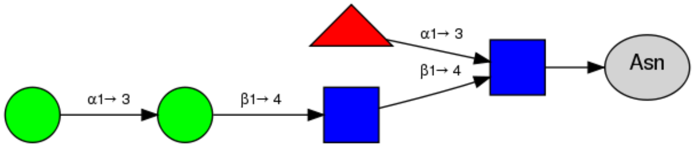
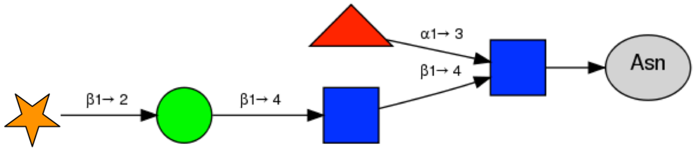
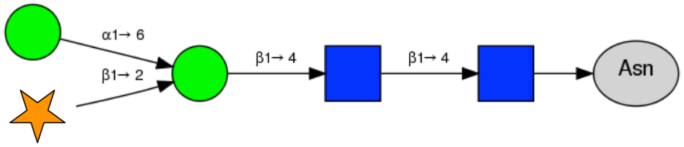
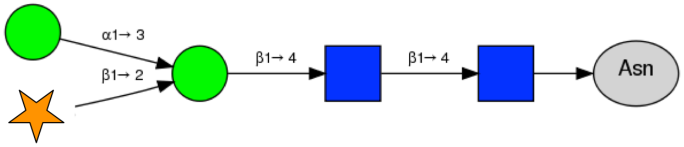
N-glycan modeling without good template structures appears to be challenging because of less conserved N-glycan structures found for non-homologous proteins. However, a higher level of internal substructure similarity exists even for non-homologous N-glycan pairs up to a residue distance of 4 without the 1-6 linkage. In fact, these carbohydrate structures that lie close to the protein are key determinants of the overall N-glycan orientation. Thus, a fragment assembly approach might perform well even without homologous N-glycans template structures because of this internal substructure conservation.

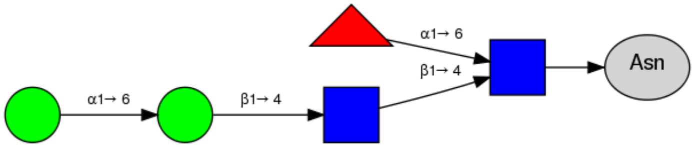
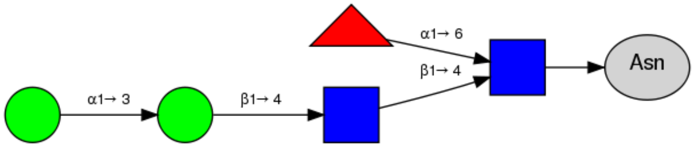
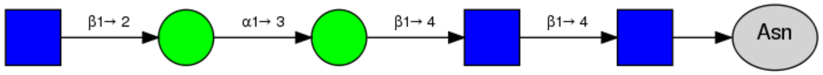
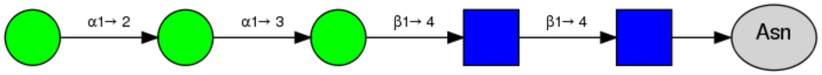
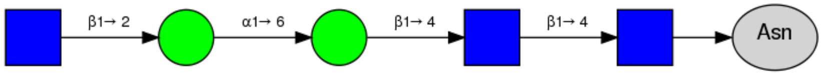
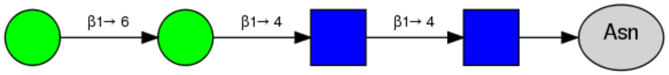
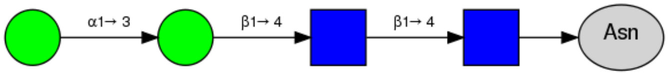
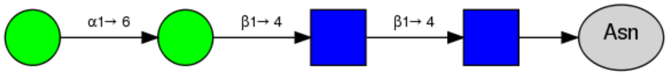
4.5 Supplementary Tables

Seq. #	Sequence	Length	# of homologous N-glycan paris	# of non-homologous N-glycan pairs
45		7	7	414
313		7	9	26
161		7	2	59
160		7	0	29
49		6	11	502

Seq. #	Sequence	Length	# of homologous N-glycan paris	# of non-homologous N-glycan pairs
47		6	7	651
46		6	8	680
330		6	0	19
328		6	0	3
319		6	1	40

Seq. #	Sequence	Length	# of homologous N-glycan paris	# of non-homologous N-glycan pairs
316		6	9	35
239		6	14	63
23		6	0	38
144		6	11	168
54		5	12	95
52		5	42	421
50		5	9	44

Seq. #	Sequence	Length	# of homologous N-glycan paris	# of non-homologous N-glycan pairs
336		5	0	83
335		5	0	116
334		5	0	24
332		5	0	23
331		5	0	29

Seq. #	Sequence	Length	# of homologous N-glycan paris	# of non-homologous N-glycan pairs
324		5	1	50
323		5	2	47
321		5	9	1230
25		5	14	5128
240		5	14	760
8		4	0	156
7		4	47	11578
58		4	52	9544

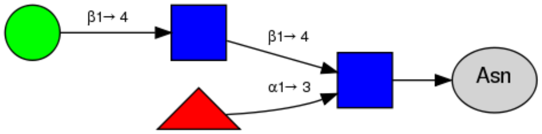
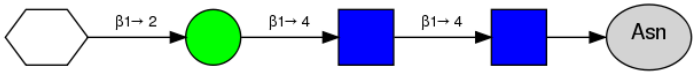
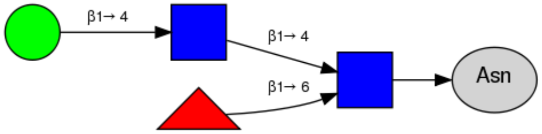
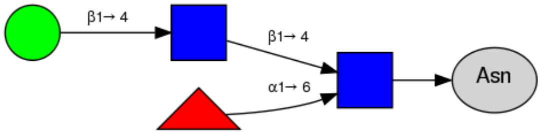
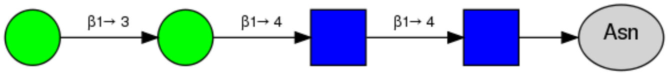
Seq. #	Sequence	Length	# of homologous N-glycan paris	# of non-homologous N-glycan pairs
341		4	2	389
337		4	1	93
201		4	2	52
200		4	3	496
150		4	0	248

Table S4.1: List of N-linked oligosaccharide sequences used in the structural similarity comparison. The nomenclature for glycan representation is adopted from [136]: blue square for N-acetyl glucose, green circle for mannose, red triangle for fucose, yellow star for xylose). The number of homologous and non-homologous N-glycan structure pairs (non-redundant) are given for each N-glycan sequence.

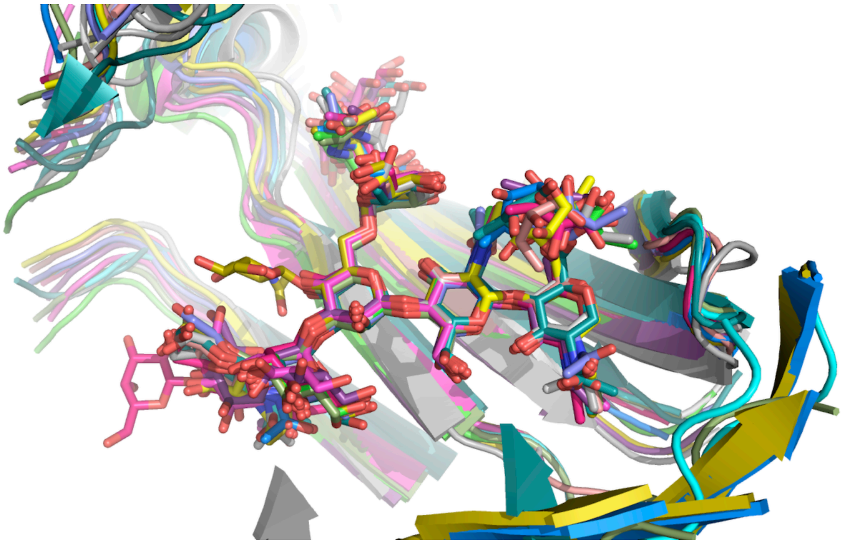


Figure S4.1: Overlay of the N-glycan core structures from the various IgG1 structures from the PDB. The PDB entries used in this overlay are 3AVE, 3AY4, 3C2S, 3D6G, 3DO3, 2DTS, 3FJT, 1H3X, 1I1A, 1I1C, 1L6X, 1OQO, 2QL1, 2RGS, 3SGJ, 3SGK, and 2VUO.

4.6 Supplementary Figures

Seq. #	μ	σ	ε	ξ^2
45	2.42	0.74	-0.10	1.4
313	2.60	0.89	-0.24	21.6
161	2.67	0.83	-0.18	1.1
160	2.63	0.78	-0.18	0.9
49	2.28	0.69	-0.14	0.1
47	2.73	1.04	-0.13	14.5
46	1.91	0.55	-0.10	0.19
330	2.54	0.60	-0.23	4.60
328	2.42	0.57	-0.30	6.70
319	2.68	0.78	-0.26	2.87
316	2.31	0.67	-0.18	0.30
239	2.11	0.62	-0.15	0.40
23	2.27	0.66	-0.14	0.08
144	2.14	0.77	-0.25	27.58
54	2.36	0.85	-0.24	7.16
52	1.84	0.54	0.15	0.25
50	2.21	0.82	-0.09	2.02
336	2.50	0.71	-0.24	3.76
335	2.07	0.56	-0.29	1.12
334	2.00	1.16	-0.51	507.15
332	1.84	0.53	-0.13	0.35
331	1.73	0.46	-0.23	0.17
324	2.59	0.83	-0.26	3.97
323	1.96	1.00	-0.37	263.86
321	1.92	0.55	-0.19	0.10
25	1.86	0.51	-0.22	0.45
240	2.21	0.75	-0.13	0.48
8	1.78	0.62	-0.11	1.01
7	1.50	0.42	-0.28	1.96
58	1.74	0.59	-0.14	0.57
341	1.65	0.49	-0.22	5.00
337	1.58	0.49	-0.25	1.51
201	1.56	0.53	-0.22	3.76
200	1.60	0.55	-0.24	4.02
150	1.53	0.48	-0.27	4.13

Table S4.2: Parameters for the generalized extreme value distributions

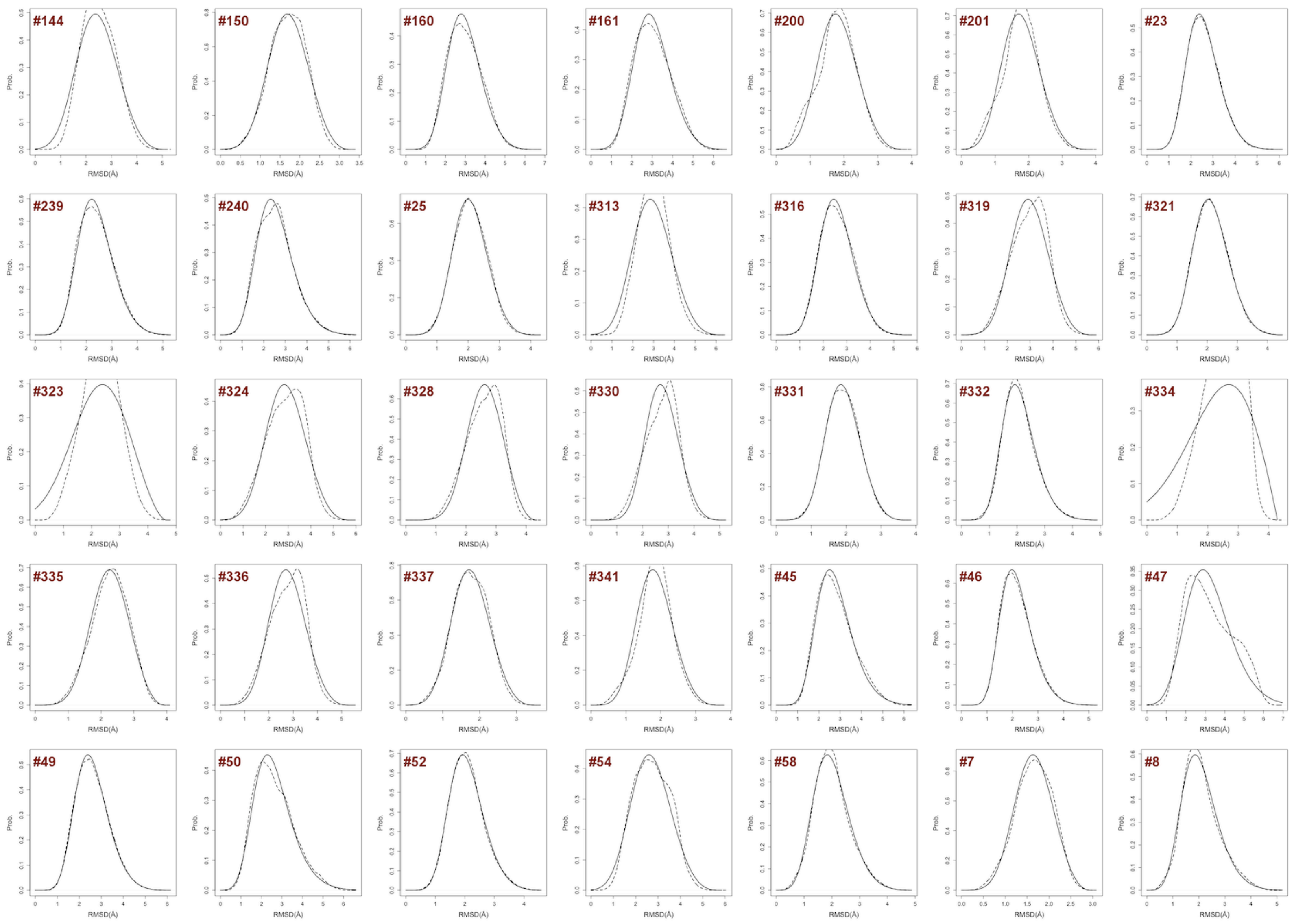


Figure S4.2: The comparison of the original RMSD distributions (dashed line) and the fitted generalized extreme distributions (solid line). The numbers on each plots represents the sequence identification number used in Table S4.1.

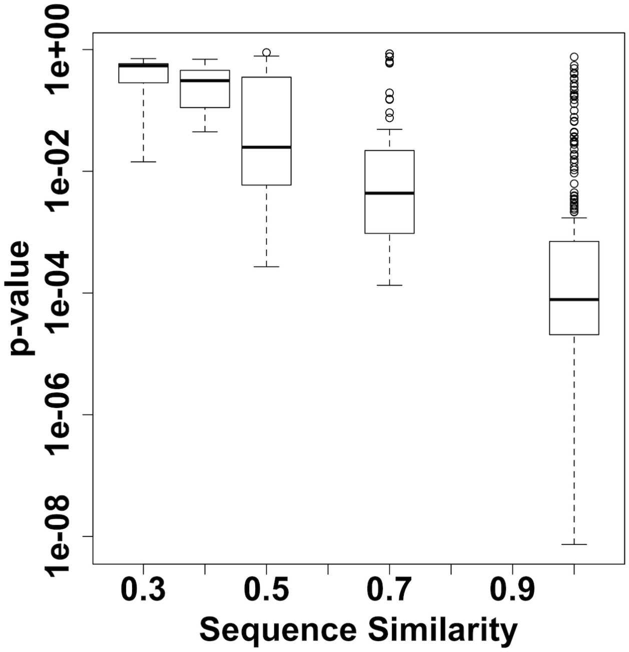


Figure S4.3: Correlation between the sequence similarity and the structural similarity (p-value). The box represents the range between the first and third quartiles of the distribution and the thick horizontal lines represent the median of the distribution. The open circles are outliers.

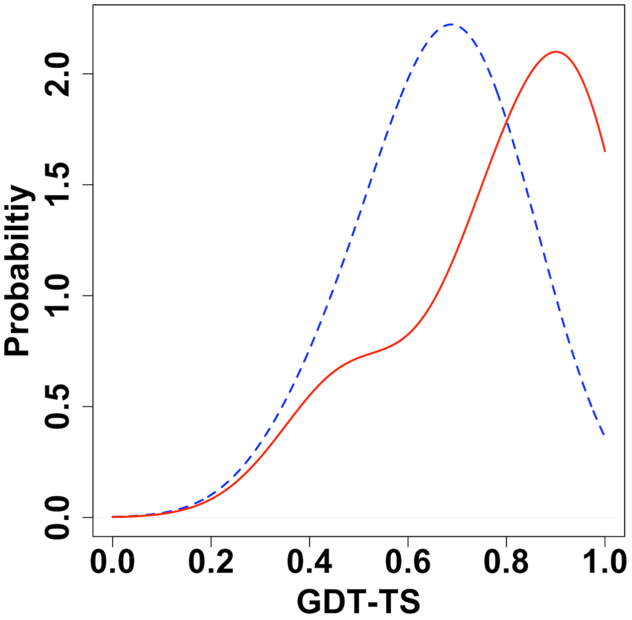


Figure S4.4: Structural similarity of N-glycans using the GDT-TS score. The GDT-TS score distributions from the homologous (red) and non-homologous (blue) structure pairs. A 0.1-Å bin width was used. The GDT-TS score is defined as GDT-TS = (P0.5 + P1 + P2)/3 where PX is the fraction of atoms that can be superimposed with corresponding cutoffs of X = 0.5, 1, and 2 Å.

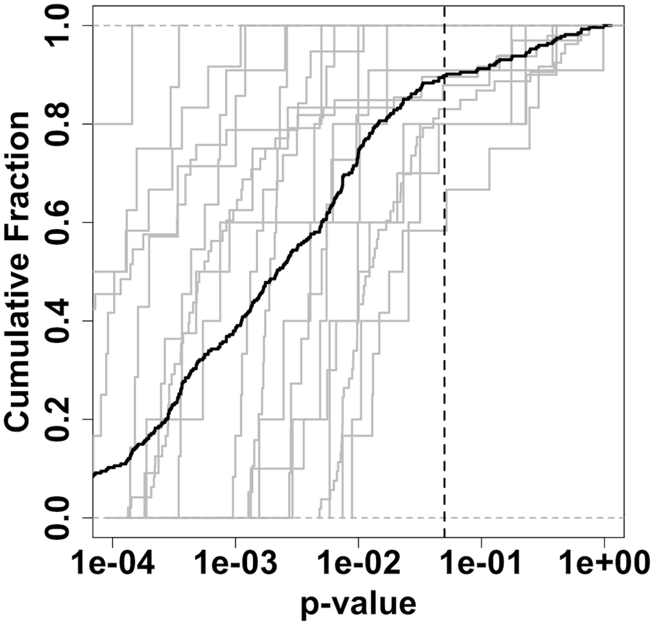


Figure S4.5: Cumulative fraction of structure similarity of N-glycan pairs whose parent proteins have sequence similarity greater than or equal to 90%.

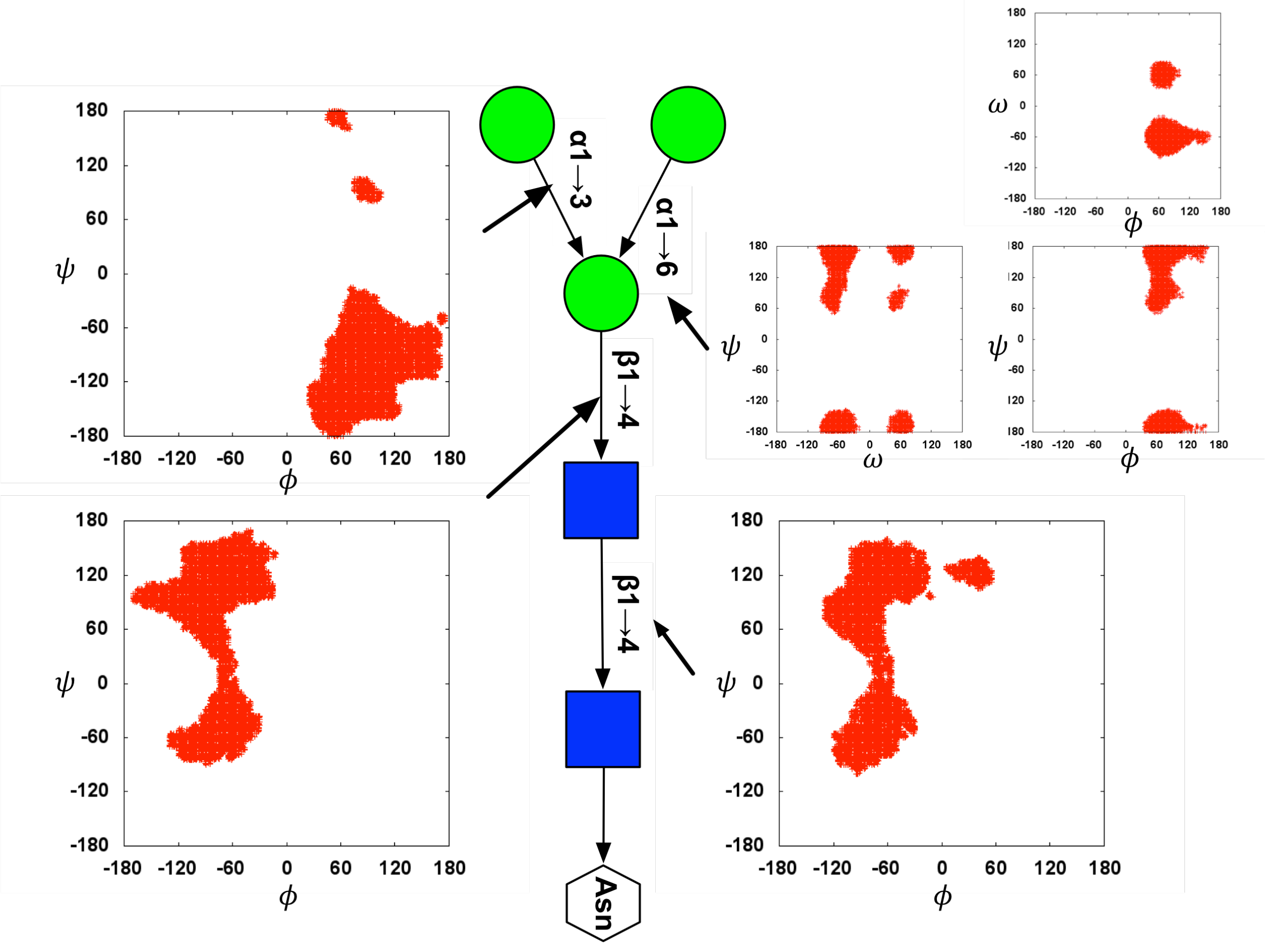


Figure S4.6: Glycosidic torsion angle distributions from the random glycan conformation pool for the N-glycan core sequence. 1,000,000 conformations were generated by assigning randomly chosen torsion angle values from the accessible torsion angles of the corresponding glycosidic linkage type. The following glycosidic torsion angle definitions are used; O5-C1-O1-C'x (ϕ), C1-O1-C'x-C'x-1 (ψ), and O1-C'6-C'5-O'5 (ω).

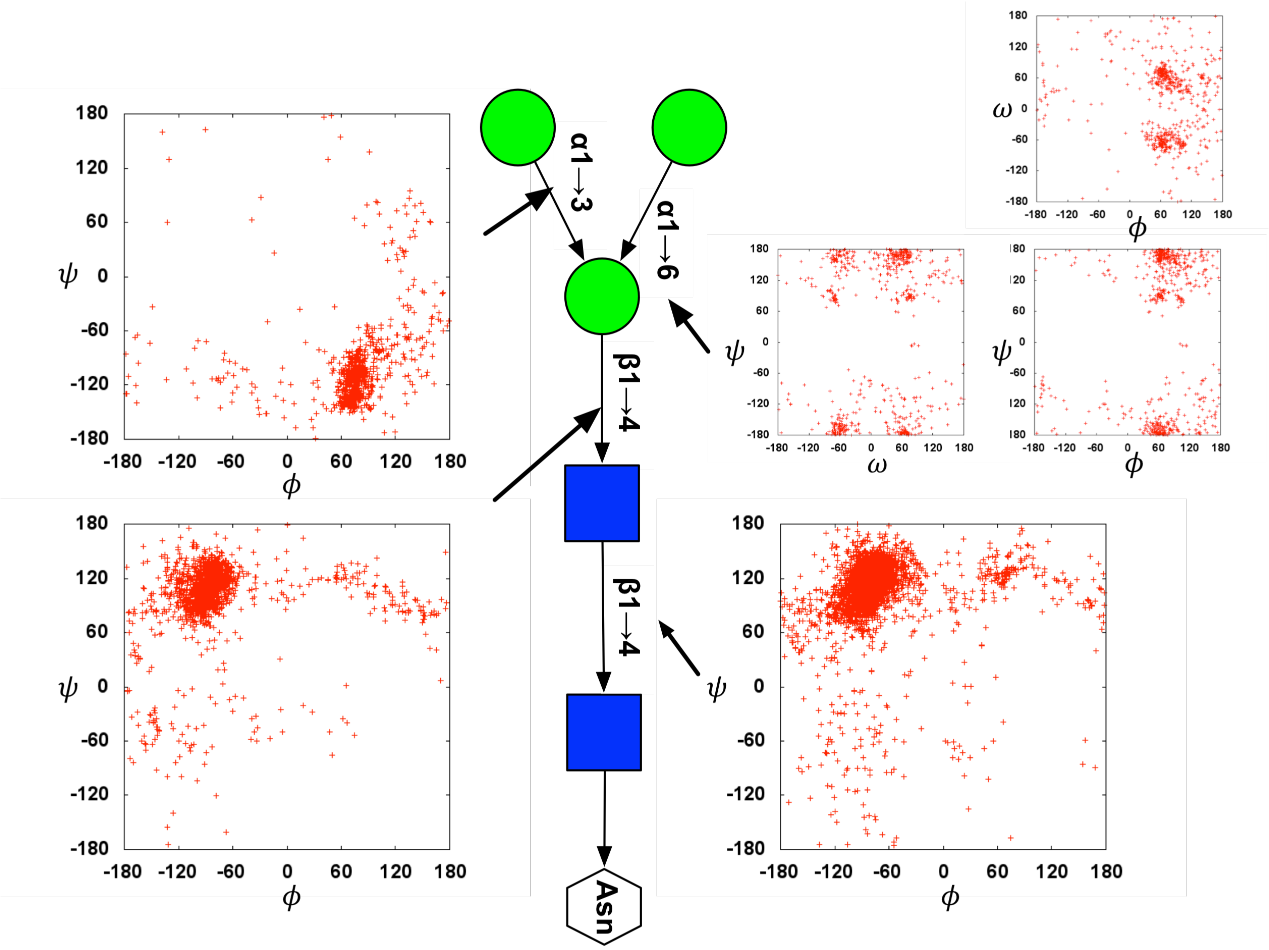


Figure S4.7: Glycosidic torsion angle distributions for the corresponding glycosidic linkage type (disaccharide) observed in the PDB. The Glycan-Fragment DB [2] was used to collect the glycosidic torsion angle distribution in the PDB. The following glycosidic torsion angle definitions are used; O5-C1-O1-C’x (ϕ), C1-O1-C’x-C’x-1 (ψ), and O1-C’6-C’5-O’5 (ω).

Chapter 5

Preferred Conformations of N-glycan Core Pentasaccharide in Solution and in Glycoproteins

5.1 Introduction

An oligosaccharide moiety in a glycoprotein, referred to as a glycan, comes in a variety of sequences and structures, and plays critical roles in a vast array of biological processes, such as protein quality control in ER [1, 48, 58, 72], protein trafficking [38, 85, 124], and stabilize protein structure [21, 22, 26, 44, 129]. These oligosaccharide moieties can be covalently attached to asparagine (Asn) side-chains of a nascent peptide being synthesized in the ER through the process known as glycosylation [121]. N-linked oligosaccharide moieties (N-glycans) initially have the same primary sequence, but are later processed by enzymes in the ER and the Golgi to become different glycoforms [1, 72]. In addition to being a simple appendage to a protein, many N-glycans are involved in molecular recognition in a sequence dependent manner [31, 122, 125, 127, 141, 143, 144, 154]. These recognition events require specific carbohydrate structures and are sensitive to small differences in carbohydrate sequence or conformation [87, 113, 126]. Thus, the understanding of conformational preference of N-glycans could provide valuable insight into the the mechanism and specificity of carbohydrate recognition events.

In general, N-glycans in glycoproteins are in close contact with protein surface, hence it has been of great interest whether the protein structure affects the N-glycan conformation or *vice versa* [21, 22, 26, 143, 145]. An earlier nuclear magnetic resonance (NMR) study about the conformational freedom of free oligosaccharides in solution and N-linked oligosaccharide concluded that the covalent attachment to the protein does not significantly affect the conformational freedom of the oligosaccharides [145]. However, it is well known that the carbohydrates in the vicinity of the protein can engage in a specific interaction with protein side-chains, which can affect the conformational freedom of oligosaccharides [22, 26]. Structural change of protein due to different glycoform sequences is also observed through systematic crystallization study [71]. A recent survey of crystal structures in the Protein Data Bank (PDB) also revealed that protein structure affects the conformations of N-glycans [63].

To gain a better understanding of the conformational preference of oligosaccharides, it is essential to obtain atomic resolution structures in various environments. However, experimental determination of oligosaccharide conformation using X-ray crystallography or NMR is challenging due to the flexibility of glycosidic linkage and the crowding of NMR spectra [5, 78, 128, 143]. Therefore, computational simulation studies of oligosaccharides can provide valuable insight into the conformational preference of oligosaccharides at the atomic level [96, 138]. Recent advances in carbohydrate force fields have been used to study diverse glycan sequences ranging from monosaccharides to polysaccharides, and the result thus far shown to match experimental properties well [27, 42, 69].

In this work, we have performed computational sampling of conformation of N-glycan core pentasaccharide (Man3GlcNAc2; Figure 5.1) in explicit water using molecular dynamics (MD) simulation (a total of 3.5 μs) and replica-exchange MD (REXMD) simulation (a total of 3.8 μs) [130]. Earlier computational studies of carbohydrates are often restricted to mono- or disaccharides due to computational resources [4, 68, 101, 119], but it is not clear whether the observations made

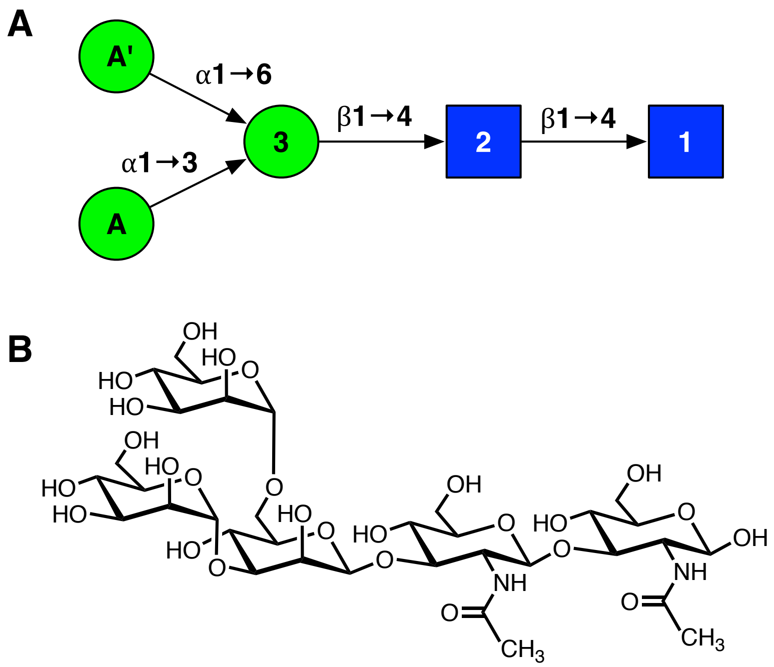


Figure 5.1: Pentasaccharide sequence used in this study. A) Symbolic notation and B) Chemical structure of the pentasaccharide.

in those study can be expanded to larger oligosaccharides due to non-neighbor interactions. In rare occasion, simulations of larger oligosaccharides were performed [5, 84, 140], but the simulation time was typically not long enough to produce well-converged conformational states for those oligosaccharide (< 50 ns). The pentasaccharide sequence used in this study is small enough to exhaustively sample its conformational states, but still big enough to investigate the presence of non-neighbor interaction.

Our aim is to utilize the simulation trajectory to characterize the conformational preference of the pentasaccharide in solution and the change of such preferences in the vicinity of glycoproteins using the crystal structures in PDB database [13]. We first examined the conformational variability of the oligosaccharide and its conformational preferences in solution. Then, the conformational preferences of the pentasaccharide in solution were compared with the pentasaccharide structures found in the vicinity of proteins. Finally, the correlation between hydrogen bond formation/deformation and change of conformational states were examined by transfer entropy analysis. The N-glcyan core pentassacharide sequence is found in virtually

	T1	T2	T3	T4	System Size	# Water
#1	(-85, 105)	(-91, 94)	(78, -112)	(75, 105, 60)	44 × 44 × 44	2,611
#2	(-77, 113)	(-80, 127)	(76, -107)	(65, 14, -64)	45 × 45 × 45	2,872
#3	(-80, 130)	(-77, 118)	(71, -133)	(64, 92, 73)	44 × 44 × 44	2,611
#4	(-82, 127)	(-39, 112)	(75, -142)	(97, 85, -71)	45 × 45 × 45	2,870
#5	(-73, 126)	(-91, 88)	(81, -98)	(138, 146, 37)	45 × 45 × 45	2,872

Table 5.1: Initial conformations of the pentasaccharide and the system setup for the MD simulation. T1, T2, T3, and T4 represents the glycosidic torsion angle (ϕ, ψ) between residue pair (1 and 2), (2 and 3), (3 and A), and (3 and A') in degree; residue names are based on Figure 5.1.

all N-linked glycosylated oligosaccharide chains [1, 72]. Thus a detailed understanding of its conformational preference and dynamics in solution and in the vicinity of protein will provide valuable insights for understanding of larger N-linked oligosaccharides in glycoproteins.

5.2 Methods

5.2.1 Computational Detail

Initial glycan conformations for the MD simulations were selected by using the Glycan Fragment Database (GFDB; `http://www.glycanstructure.org/fragment_db`) [60]. PDB entries with resolution higher than or equal to 3 Å were searched with various filters to remove distorted residues and redundant entries. Glycosidic torsion angle clustering analysis was performed using GFDB and the representative structures of the 5 largest clusters were selected as the initial conformations for the MD simulation (Table 5.1).

The selected initial structures were briefly minimized without water prior to the building the systems. The *Glycan Reader* and *Quick MD Setup* in CHARMM-GUI [62, 65] were used to build each initial MD simulation system. The system size was determined so that the resulting systems have at least a 12.5 Å water layer in each direction. The solvated simulation systems were briefly minimized while harmonic restraint was applied to the pentasaccharide molecule in the presence of water using CHARMM simulation software [17]. Each of minimized simulation systems was independently subjected to 700 ns of MD simulation at 300 K using NAMD simulation software

[106], which gave total simulation time of 3.5 μs in total. In addition, 100 ns of temperature replica exchange simulation with explicit water was performed using CHARMM and the MMTSB package [17, 28]. Total of 38 replicas were used (total simulation time 3.8 μs) to cover the temperature range from 300 K to 450 K. The initial configuration of the first MD simulation system was equilibrated at 1 bar using the NPT ensemble to determine the appropriate system size. The resulting snapshot was then duplicated and used as the initial configurations for each replica for REXMD simulation and the NVT ensemble was applied with system dimension of $42.9 \times 42.9 \times 42.9$ Å^3.

All simulations were performed using CHARMM C36 carbohydrate force field [42] and TIP3P water model [66]. The van der Waals interactions were smoothly switched off between 10 and 12 Å by a forced-based switching function. Long-range electrostatic interactions were calculated using the particle-mesh Ewald (PME) method [148]. An interpolation order of 6 and a direct space tolerance of 10^{-6} were used for the PME method. A time-step of 2 fs was used with the SHAKE algorithm [118]. For the CHARMM simulations, the temperature was held constant with the Hoover thermostat [53] and the pressure was maintained at 1 bar with the Nose-Hoover piston [6]. For the NAMD simulations, Langevin dynamics was used to maintain constant temperatures for each system, while the Nose-Hoover Langevin-piston algorithm [29, 86] was used to maintain constant pressure at 1 bar.

5.2.2 Conformational variability

The conformational variability is measured as the pair-wise RMSD distribution using the RMSDYN module in CHARMM [17]. To calculate the pair-wise RMSD distribution, a set of conformations was selected from the trajectory and the RMSDs were calculated for each pair of conformations using all non-hydrogen atoms for alignment. For MD simulation, conformations were selected from the aggregated trajectories every 2.5 ns, which resulted in 1,400 conformers. For REXMD simulation, conformations were selected every 100 ps, which resulted in 1,000 conformers. To estimate the upper limit of the conformational variability of the pentasaccharide, a random conformation pool was built based on the protocol used in ref [63]. Briefly, a random conformation pool of 1,000,000 conformations was built in iterative fashion. For each iteration, new torsion angle value was assigned to a randomly selected glycosidic linkage, and the new conformation was accepted if it did not have any bad contacts. Torsion angle values were selected among the pre-calculated accessible torsion angles based on the adiabatic map of the corresponding glycosidic linkage. For the pair-wise RMSD distribution of random glycan conformation pool, conformations generated every 1,000th iteration were extracted, which resulted in 1,000 conformers.

5.2.3 Selection of PDB entries for comparison of conformational preference

Oligosaccharide structures or disaccharide structures in the PDB were selected using the GFDB. Various filters available in GFDB were used to refine the selection and to remove potentially erroneous entries. For example, PDB entries determined by X-ray crystallography with resolution equal to or higher than 3 Å were only searched. The glycan chains with distorted carbohydrate structures or inaccurate residue name annotation were excluded. In addition, redundant PDB entries were removed to prevent overrepresent certain conformations. However, it is not straightforward to remove redundancy in the case of N-glycan. Here, we follow the protocol adopted in [63], which uses the glycoprotein sequence similarity to identify redundant entries. N-glycans that are attached on homologous glycoproteins (>70% sequence similarity) were also excluded. The sequence similarity provided by the PDB was used to determine the sequence similarity.

5.2.4 Coarse-graining of conformational state using glycosidic torsion angle

The torsion angle distribution from the high-temperature REXMD simulation was used to identify a set of initial basins as shown in Figure 5.2. For each glycosidic linkage, several well-defined basins were readily identifiable by examining the torsion angle distribution. Once the basins were roughly identified, basins were refined by assigning the torsion angles observed during the

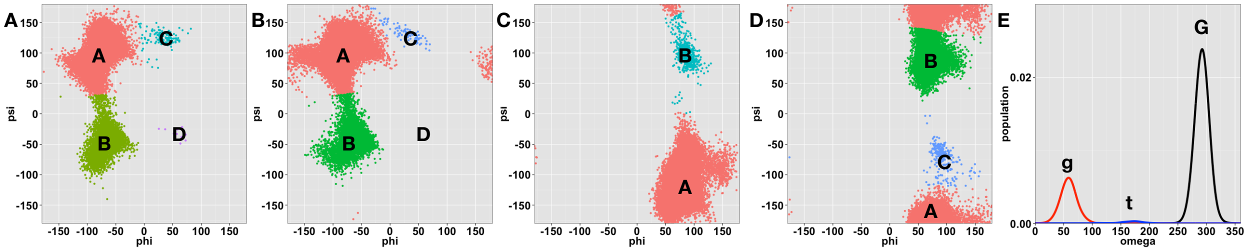


Figure 5.2: Glycosidic torsion angle distribution from REXMD simulation at 450 K and the assignment of torsion angle states for each glycosidic linkage. (A) GlcNAc β(1→4) GlcNAc, (B) Man β(1→4) GlcNAc, (C) Man α(1→3) Man, (D) Man α(1→6) Man, and (E) Omega torsion angle of the Man α(1→6) Man linkage.

simulations to the nearest basin in an iterative fashin using k-medoid algorithm [99]. The glycosidic torsion angle definition was adopted from the crystallographic definition: O_5-C_1-O_1-C'_x (ϕ), C_1-O_1-C'_x-C'_{x-1} (ψ), and O_1-C'_6-C'_5-O'_5 (ω). The angular distance metric [34] was used to preserve the periodicity of torsion angle between two torsion angles in the clustering algorithm.

We denote torsional state for each glycosidic linkage based on the size of the basin ("A" refers to the largest basin and the "B" refers to the second largest basin, and so on), with the omega torsion angle as an exception. For the omega torsion angle, the basins are named after the well-known staggered rotameric states of the omega torsion angle: G (gauche-gauche), g (gauche-trans), and t (trans-gauche). Note that the k-medoid algorithm sometimes does not preserve the initial basin assignment when the basin is too small and such basins were manually assigned (e.g., basin C and basin D for the first glycosidic linkage). By combining the torsion angle states, the conformation of pentasaccharide can be described with 5-letter notation, starting from the anomeric carbohydrate residue. For example, "AAAAG" indicates the each glycosidic (ϕ, ψ) torsion angles adopted their largest basin, and the omega torsion angle adopted gauche-gauche orientation.

5.2.5 Transfer entropy (TE) between conformational state and the formation or deformation of hydrogen bonds

Transfer entropy is a measure that quantifies the information flow from the past of one time series $y(t)$ to the future of another time series $x(t)$ [120]. In the present work, the following form is used

$$
\begin{aligned}
\mathrm{TE}_{y\rightarrow x} &= H\left(x_{t+1}|x_t^{(k)}\right) - H\left(x_{t+1}|x_t^{(k)}, y_t^{(l)}\right) && (5.1)\\
&= H\left(x_{t+1}, x_t^{(k)}\right) + H\left(x_t^{(k)}, y_t^{(l)}\right) - H\left(x_{t+1}, x_t^{(k)}, y_t^{(l)}\right) - H\left(x_t^{(k)}\right) && (5.2)
\end{aligned}
$$

where k and l are the embedding dimensions that are the number of steps to be included from the past of time series $x(t)$ and $y(t)$. $H(x) = -\sum p(x_i)\log p(x_i)$ is Shannon entropy, where $p()$ is the probability of one state and the summation is over all possible combinations of states. $H(|)$ is conditional Shannon entropy. Due to finite sample size of the time series, two irrelevant series can have non-zero (statistically insignificant) TE. To remove this bias, the shuffling method has been used to calculate the effective transfer entropy ($\mathrm{TE}^{\mathrm{eff}}$) is defined as below [67, 82].

$$
\mathrm{TE}_{y\rightarrow x}^{\mathrm{eff}} = \mathrm{TE}_{y\rightarrow x} - \frac{1}{N}\sum_{n=1}^{N} \mathrm{TE}_{y^{\mathrm{shuffled}}\rightarrow x} \tag{5.3}
$$

where N, the number of shuffling, was set to 500 for all calculations in this study. Using the effective TE, a normalized directional index can be derived as

$$
D_{y\rightarrow x} = \frac{\mathrm{TE}_{y\rightarrow x}^{\mathrm{eff}}}{H\left(x_{t+1}|x_t^{(k)}\right)} - \frac{\mathrm{TE}_{x\rightarrow y}^{\mathrm{eff}}}{H\left(y_{t+1}|y_t^{(l)}\right)} \in [-1, 1] \tag{5.4}
$$

where $H\left(x_{t+1}|x_t^{(k)}\right)$ and $H\left(y_{t+1}|y_t^{(l)}\right)$ are the maximal TE. A positive D value indicates information flow from y to x or y drives x, and *vice versa* for a negative value. For two completely irrelative time series, $D_{y\rightarrow x}$ and $\mathrm{TE}^{\mathrm{eff}}$ are 0. Both k and l were set as 1 and only the $D_{y\rightarrow x}$ values larger than 0.1 and having p-value larger than 0.05 were taken into further analysis.

For TE analysis, two time series of instantaneous conformational state and hydrogen bond between atom pairs were generated based on the trajectories of standard MD simulations. For instantaneous conformational state, the torsional angle state definition for each glycosidic linkage is used as described in Result section. The hydrogen bond was defined as the distance between the donor and the acceptor below 2.8 Å and the angle below 120°.

5.3 Results

5.3.1 Convergence of glycosidic torsion angle distribution

The average acceptance ratio of replica exchange in REXMD simulation was 37.4% and the random walk of replica in the temperature space was very efficient (Figure S5.1) as multiple travels between the lowest and highest temperatures were observed. These results demonstrates reliable sampling of the simulation system during the REXMD simulation. In addition, we examined the torsion angle distribution of glycosidic linkages from standard MD and REXMD simulations (Figure S5.2 and S5.3). The glycosidic torsion angle distribution from simulations started at different initial conformation are well-converged to each other and to those derived from REXMD simulation at 300 K. The convergnece was also examined by the relative population of dominant conformational states (Figure S5.4). 100-ns block averages of the population of conformational states show a relatively stable conformational state population distribution over the simulation timescale, suggesting that most conformational states exchange within 100 ns, except a few long-lived conformational states (see below). The bias in the conformational state population due to the initial configuration of simulation system is quickly resolved within 50 ns of simulation. These observations suggest that the sampling of current simulation is robust and the results presented below are based on the aggregated trajectories unless explicitly stated otherwise.

5.3.2 Conformational variation of the pentasaccharide in solution

It is assumed that oligosaccharides in solution are flexible, but how flexible are they? Here, the general conformational variability was measured by pair-wise RMSD distribution (Figure 5.3). Conformational variability of the pentasaccharide in solution at 300 K appeared to be around 1–3 Å in terms of RMSD. Although the frequencies of sampled conformations in building the pair-wise RMSD distribution were different in MD simulation and REXMD simulation (more frequent in REXMD), the resulting distribution at 300 K agree well each other. The conformational variability of the pentasaccharide in solution at 300 K is smaller compared to the variability at higher

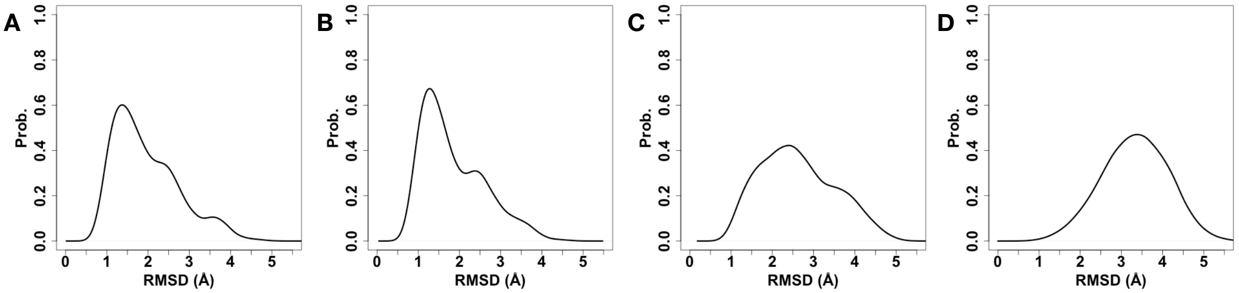


Figure 5.3: Conformational variability of the pentasaccharide in solution. Pair-wise RMSD distribution is calculated from the (A) standard MD simulation, (B) REXMD simulation at 300 K, (C) REXMD simulation at 450 K, and (D) random conformation pool.

temperature or when when compared to random glycan conformation, where the variability is around 2–4 Å. In addition, the presence of several peaks in the conformational variability suggests the existence of several well-defined conformational states.

5.3.3 Conformational preference of the pentasaccharide in solution

To gain further insight, we defined conformational states using glycosidic torsion angles. The pentasaccharide used in this study has 5 glycosidic linkages (omega angle for 1-6 linkage was separated for clarity), and thus it was relatively straightforward to identify basins from the torsion angle distribution. Such a description of conformational states using dihedral angles is common in protein/peptide conformational analysis [18].

The torsion angle distribution from the high-temperature REXMD simulation was used to identify a set of initial basins as shown in Figure 5.2. For each glycosidic linkage, several well-defined basins were readily identifiable by examining the torsion angle distribution. The basins from each glycosidic linkages resulted in total 4 × 4 × 2 × 3 × 3 = 288 possible conformation states, but not every states are visited in the simulation. In fact, only 63 states were visited in REXMD simulation at 450 K, while 22 and 42 states were visited in REXMD simulation at 300 K and standard MD simulations, respectively. Surprisingly, the conformational states are very restricted to only a several states (Table 5.2). For example, state "AAAAG" accounts for more than 75% of the total simulation trajectories, and the 10 largest conformational states accounts for over

	MD		REXMD (300 K)		REXMD (450 K)		PDB	
1	AAAAG	75.5%	AAAAG	78.7%	AAAAG	50.0%	AAAAG	23.2% (22)
2	AAAAg	7.8%	AAAAg	7.9%	AAABG	13.5%	AAABG	23.2% (22)
3	AAABG	6.3%	AAABG	5.3%	AAAAg	10.8%	AAAAg	22.1% (21)
4	ABAAG	3.0%	ABAAG	3.4%	AAACg	5.4%	AAACg	6.3% (6)
5	BAAAG	2.8%	BAAAG	2.4%	AAACG	4.9%	AAACG	4.2% (4)
6	AAABg	1.2%	AAABg	1.3%	AAABG	3.6%	AAABG	4.2% (4)
7	ABAAg	0.6%	ABAAg	0.3%	AAABt	2.6%	AAABt	3.2% (3)
8	ABABG	0.6%	BAABG	0.2%	AABBG	2.1%	AABBG	3.2% (3)
9	AABAG	0.2%	BABAG	0.1%	AAAAt	1.1%	AAAAt	3.2% (3)
10	BAAAg	0.2%	BAAAg	0.1%	AABAt	0.9%	AABAt	1.1% (1)
Sum		99.3%		99.7%		94.8%		93.7% (89)

Table 5.2: Conformational preferences of the pentasaccharide in solution and in the vicinity of protein. The numbers in the paranthesis refers the number of PDB entries.

99% of the conformations visited in the MD simulation.

To validate the assignment of conformational states using glycosidic torsion angle distribution, conformational variability of each state was compared. From each conformational state, 1000 conformers were arbitrarily chosen from the trajectory and the pair-wise RMSD distribution was calculated (Figure 5.4). The overall RMSD of conformations belong to the same conformational state was about 1-2 Å. Only a single peak is present in the distribution, suggesting the conformations in the same state are well grouped. The largest conformational state in REXMD simulation at 300 K matches well with the standard MD simulation and suggests robust conformational sampling during the simulations.

Based on the populations of each conformational state, the free energy difference between the most populated state (*AAAAG*) and the second most populated state (*AAAAg*) is about $\Delta G = -k_B T \log[P_1/P_2] = 1.38\text{kcal/mol}$. When we compared the representative structures of some of the largest conformational states (Figure 5.5), it appears that the state *AAAAG* has more extended conformations whereas the state *AAAAg* has a conformations that are folded back onto itself. In addition, because the terminal residue is folded back to itself, the state "AAAAg" has more potential interaction partners. In fact, on average the *AAAAg* has 2.4 ± 1.0 (direct) and 2.8 ± 1.9 (water mediated) hydrogen bonds while *AAAAG* has 1.5 ± 0.8 (direct) and 2.4 ± 1.7 (water

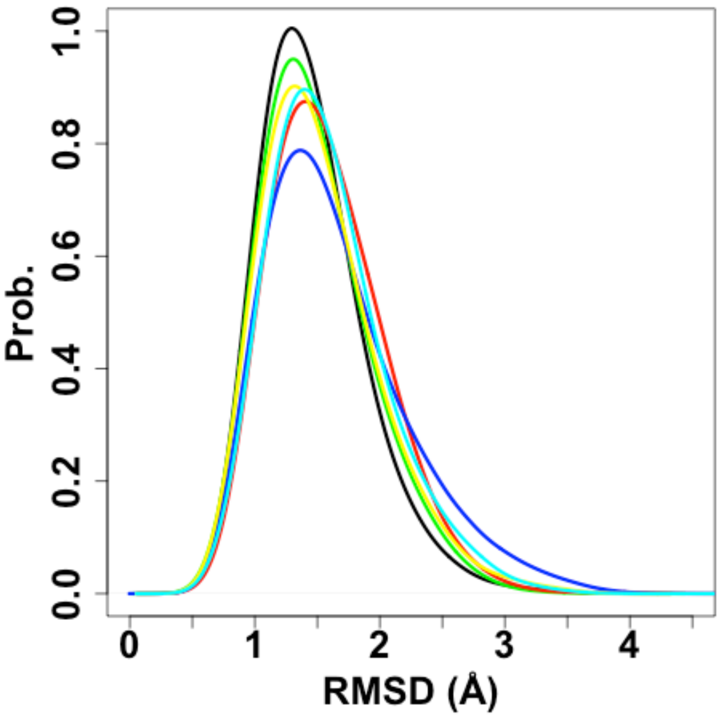


Figure 5.4: Conformational variability of major conformational states. Each colored line represents a pair-wise RMSD distribution from 1,000 conformers belonging to the same conformational states.

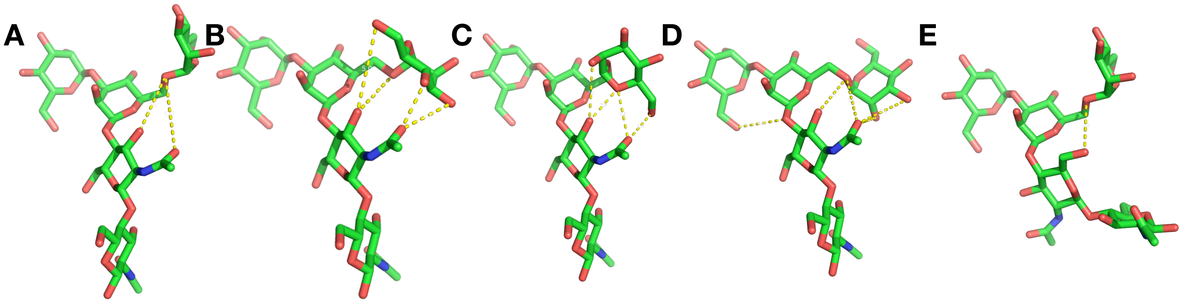


Figure 5.5: Representative pentasaccharide conformations in solution from the 5 major conformational states. Each conformations are generated by the average structure from the state A) AAAAG, B) AAAAg, C) AAABG, D) ABAAG, and E) BAAAG.

mediated).

The number of hydrogen bonds and the free energy differences suggest that the preference of the most populated states (*AAAAG*) must be entropically favorable. Indeed, it's well known that, in the case of polymers, the extended conformations are entropically more favorable than the ones that are folded back onto itself [116]. In the polymer model, the entropy increases as the end-to-end distance between the chain increases. The radius of gyration of state *AAAAG* is slightly larger (6.4 Å vs. 6.0 Å) than the state *AAAAg*, which supports the idea that the state *AAAAG* is entropically more favorable than the state *AAAAg*.

It is interesting to note that NMR experiments [51] and a recent MD simulation study [96] shows significantly increased fold-back conformation in a larger N-glycans. Overall, these observation suggests that there could be a competition between the entropic and enthalpic contribution. For example, in a smaller N-glycan, the number of interactions is not enough to favors the fold-back conformation and entropic contribution dominates. However, as the number of sugars increases as there are more number of interactions can compensate for the loss of entropy upon fold-back conformation.

5.3.4 Conformational preference of the pentasaccharide in glycoprotein

Here, we examined the change of conformational preference of the pentasaccharide in the vicinity of glycoprotein using the crystal structures. We have used GFDB to select 88 non-redundant PDB entries that have N-glycan chain whose sequence starts with the pentasaccharide used in this study. Although the number of crystal structures is not large, a significant shift in the conformational preference of the pentasaccharide in the vicinity of protein is observed (Table 5.2). Compared to the conformational preferences in solution, the state *AAAAG*, which occupies more than 70% of trajectories, is only observed in 23% of the glycoconjugate crystal structures and is not a dominant conformational state. In addition, the conformational states that are not favorable in solution occurred more frequently in the glycoconjugate crystal structure.

Figure 5.6 shows several examples of glycoconjugate crystal structures having three major conformational states. In those examples, numerous contacts between proteins and the pentasaccharide are observed. The interaction between the glycans and the proteins appears to play significant role in stablizing the conformational state which would be otherwise unfavorable in solution. These observations suggest that the interaction between the oligosaccharide and proteins can compensate unfavorable conformations. In addition, crystal contact also appears to be important in stablizing the unfavorable interaction. Although the number of observations are limited, we've found several examples of a stablizing interaction with neighboring crystal units (Fig. 5.7).

Typically, the first two residues of N-glycan have extensive interaction with surrounding protein side chains [103]. Interestingly, the first two residues in the N-glycan pentasaccharide have not visited the less favorable conformational state in glycoconjugate crystals while the residues at the termini are more flexible. This suggests that the residues closer to the protein have limited degree of conformational freedom. Similar observations were made in the recent survey of N-glycan structures in PDB [63]. It should be noted that we have used crystal structures to compare the conformational preference but crystal structure itself may reduce the apparent dynamics of glycan. Sampling bias can be removed by having a large number of crystal structures, however,

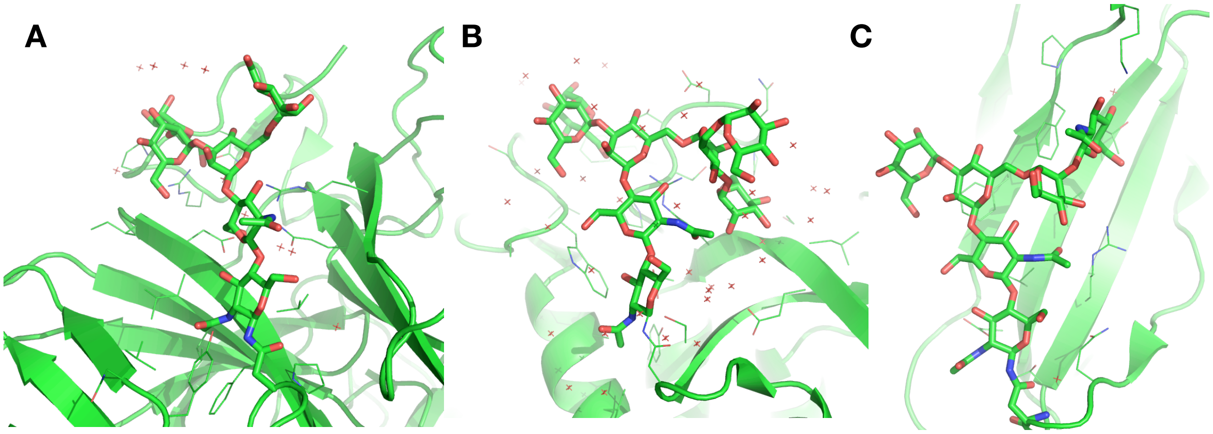


Figure 5.6: Examples of glycoconjugate structures. Each panel shows examples of three major glycoconjugates in conformational states, A) AAAAG (PDB:3PPS), B) AAABG (PDB:3GLY), and C) AAAAg (PDB:2DTS). Protein structure is drawn in cartoon representation and the protein side chains withint 5 Å distance from the N-glycan chain are drawn as lines. The crystal waters are drawn as red points.

the number of crystal structures available is limited, so care must be taken to interpret crystal structure observations.

5.3.5 Causal relationship between hydrogen bonding and the conformation exchange

The hydrogen bond is known to play important role in determining the conformation of oligosaccharides [4, 5, 26, 96, 140]. The N-glycan core pentasaccharide has several hydrogen bond donors and acceptors. Here, we examine the role of hydrogen bonds in solution conformation of N-glycan core pentasaccharide. Figure 5.8 shows the hydrogen bond pattern between the pentasaccharide residues. Strong direct and water mediated hydrogen bonds are observed between neighboring residues. Hydrogen bonds between non-neighboring residues are not common, but somewhat strong hydrogen bonds between the residue 2 and A' are observed.

Some hydrogen bonds appeared to be tightly associated with different conformational states. Figure 5.9 shows an example of highly correlated hydrogen bond and dihedral angle. There is consensus that hydrogen bonds between the oligosaccharide residues are important, but, it is not clear whether these hydrogen bonds are responsible for the formation of specific conformational

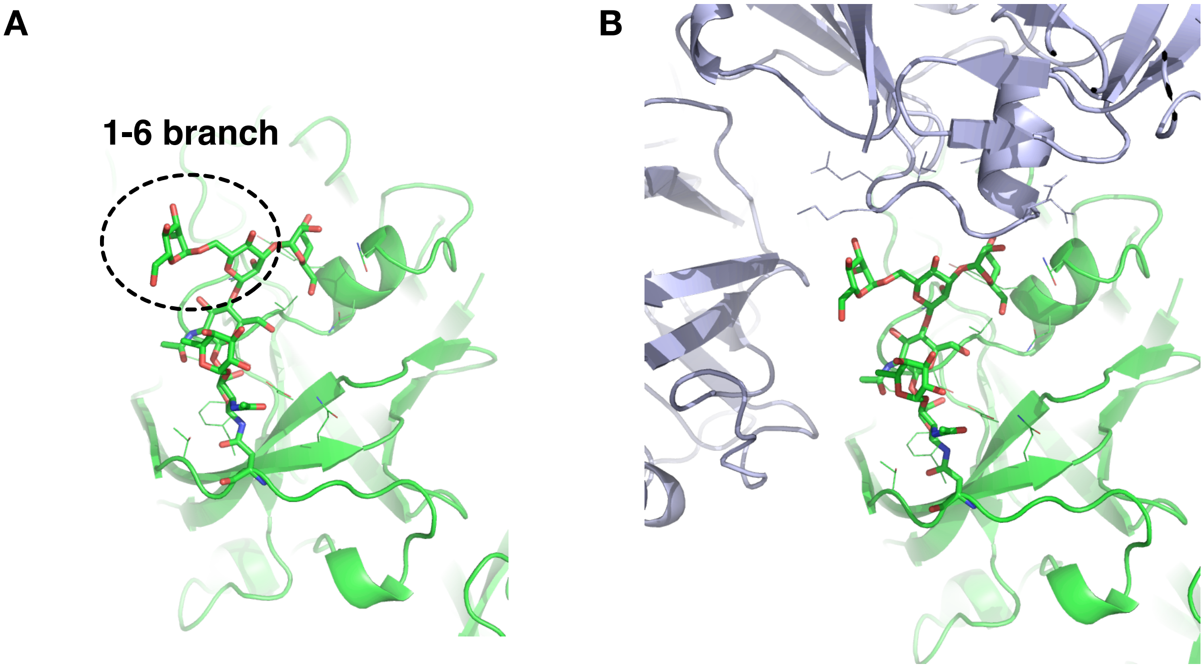


Figure 5.7: Examples of glycoconjugate structure having N-glycan adopting unfavorable conformational states in solution (PDB:1B5F). A) The N-glycan 1-6 branch is extended away from the protein, yet adopted an unfavorable conformation. B) The N-glycan 1-6 branch is involved in interaction with the neighboring crystal units (drawn in purple). Protein structure is drawn in cartoon representation and the protein side chains within 5 Å from the N-glycan chain are drawn as lines. The crystal waters are drawn as red points.

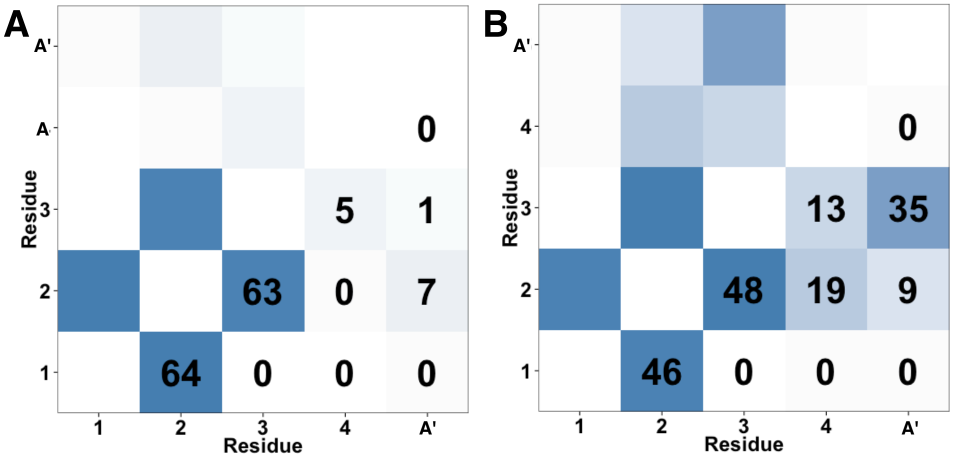


Figure 5.8: Hyrogen bonding pattern of N-glycan pentasaccharide in solution. A) Direct hydrogen bond and B) Water mediated hydrogen bond. The number and the color in each squares represent the occupancy of the hydrogen bond as a percent of the total trajectory.

H-bond	#1	#2	#3	#4	#5
2:HO3-3:O6	-0.18 (0.017)	-0.17 (1e-11)	-0.14 (2e-7)	-	-0.15 (9e-13)
2:O3-A';:HO4	-0.32 (0.02)	-0.28 (1e-8)	-0.35 (0.04)	-	-0.36 (2.8e-7)
2:O3-W-A':HO6	-0.32 (1e-17)	-0.26 (1e-17)	-0.31 (1e-17)	-0.34 (1e-17)	-0.31 (1e-17)
2:O2-W-A':HO2	-0.33 (3e-5)	-0.35 (0.0002)	-	-0.32 (9e-11)	-
2:O3-W-A':HO5	-0.37 (1e-5)	-0.37 (6e-11)	-0.39 (1e-17)	-0.37 (1e-17)	-0.38 (1e-17)

Table 5.3: Transfer entropy between time series of hydrogen bond and the conformational state. The magnitutde of each number represents the how strongly one time-series is "driving" the other time-series and the numbers in paranthesis refers to the *p*-value of the TE. Each column is from different, independent simulation trajectory. The H-bond between two atom is designated as (residue):(atom)-(water bridge)-(residue):(atom).

state. In other words, does hydrogen-bond formation/deformation drives the conformation change? We used information theoretic transfer entropy (TE) to quantify the causal relationship between the time series of conformational state change and the hydrogen bonding formation.

Surprisingly, our TE analysis showed that hydrogen bonds are not responsible for conformational change. Rather, the change of rotameric state showed a stronger causal relationship for the formation and deformation of hydrogen bond in the pentasaccharide. Table 5.3 shows the TE between the conformational state and the hydrogen-bond formation/deformation between specific atom pairs. The TE values are bound between -1 and 1, with 1 meaning the former time series drives the latter one and -1 meaning the opposite. A TE value of zero indicates no causal relationship between the two time series.

The negative transfer entropies from the simulation trajectories suggest that the hydrogen bond formation/deformation is driven by the change of glycosidic torsional state changes. In other words, the rotameric state of glycosidic linkage drives the formation/deformation of hydrogen bond. This, in turn, suggests that the hydrogen bond in glycan is important for maintaining the conformational state. It would be interesting to see how this relationship changes in a larger oligosaccharide, since the cooperative hydrogen bonding may still exists in larger oligosaccharide.

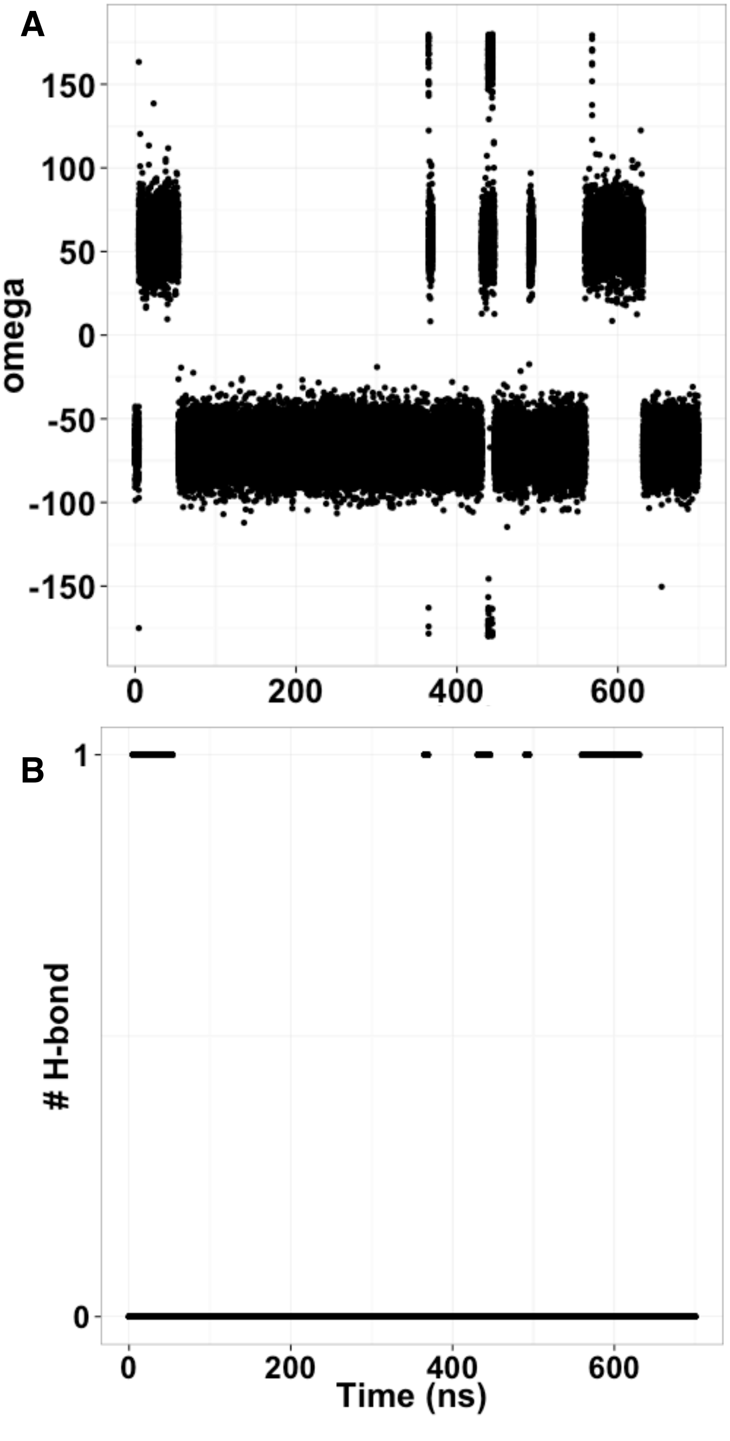


Figure 5.9: An example of a correlatin between a hydrogen bond and different conformational state. A) A time series of omega torsion angle from MD simulation #2. B) A time series of hydrogen bond between atom 2:O3 and A':HO4. from same trajectory.

5.4 Conclusion

Despite the biological importance, understanding of glycan conformation and its implication on the protein structure, dynamics, and function are lacking. Here we have used standard MD simulation (total of 3.5 μs) and REXMD simulation (total of 3.8 μs) to exhaustively sample conformational preference and compared the change of conformational preference in the vicinity of protein.

The conformational variability of the pentasaccharide appeared to be limited in solution compared to the ones from high temperature REXMD simulation or random glycan conformation models. More detailed analysis on the preference of pentasaccharide conformation showed a single major dominant conformation (>70%) and a several minor conformational (~5–8%). The 1-6 linkage appears to bring the most conformational diversity since it can completely extend or fold-back to itself.

The conformational distribution appears to be determined by the competition between the entropy and enthalpy. In the major conformational state, the 1-6 linkage extended into the solvent and is entropically favorable whereas in the minor states the 1-6 linkage fold-back onto itself. The fold-back conformation have slightly more interaction partners, but, the added interactions appear insufficient to overcome the entropic penalty. From the other NMR experiments and computational studies, the conformational preference regarding extended and fold-back conformation changes in sequence dependent manner suggesting the entropy-enthalpy compensation plays important role in conformational preference of oligosaccharide in solution.

We have used crystal structures of glycoconjugate to examine the conformational preference of the pentasaccharide when they are glycosylated. Glycosylated pentasaccharide in the crystal structure database showed significnat shifts in conformational distribution and several conformational states are equally probable (~20%). The increased conformational preferences for the states that are not favorable in solution are typically accompanied by interactions with proteins or interactions with crystal partners. Although care must be taken to interpret the data since the number of crystal structures is limited, the results suggest that the glycans in the vicinity of protein may have significantly different conformational preference due to the interaction with

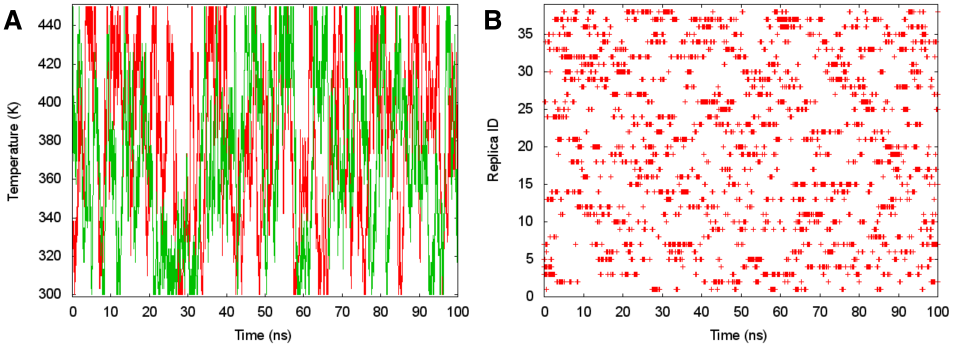


Figure S5.1: Efficient random walk across temperature space in the REXMD simulation. (A) Time series of temperature exchange of two arbitrarily chosen replicas. Reid line is for replica #1 and green line is replica #38. (B) Time series of different replicas visiting at temperature 300 K.

protein. This suggests that modeling of oligosaccharides in solution and glycosylated forms must take into account the environment.

We have found several hydrogen bonds that are tightly associated with conformational state changes. We have examined whether the hydrogen bond formation/deformation drives the conformational change using transfer entropy. In pentasaccharide, hydrogen bonds do not contribute the change of torsional states. Therefore, it appears the hydrogen bonds play a role in maintaining the conformational state rather than driving the change of conformational state. It would be interesting to examine a larger oligosaccharide since it may have more hydrogen bonds involved in, and cooperatively induce, conformational change. The hydrogen bonds formation driven by the conformational states may have implication in maintaining the conformational states.

5.5 Supplementary Figures

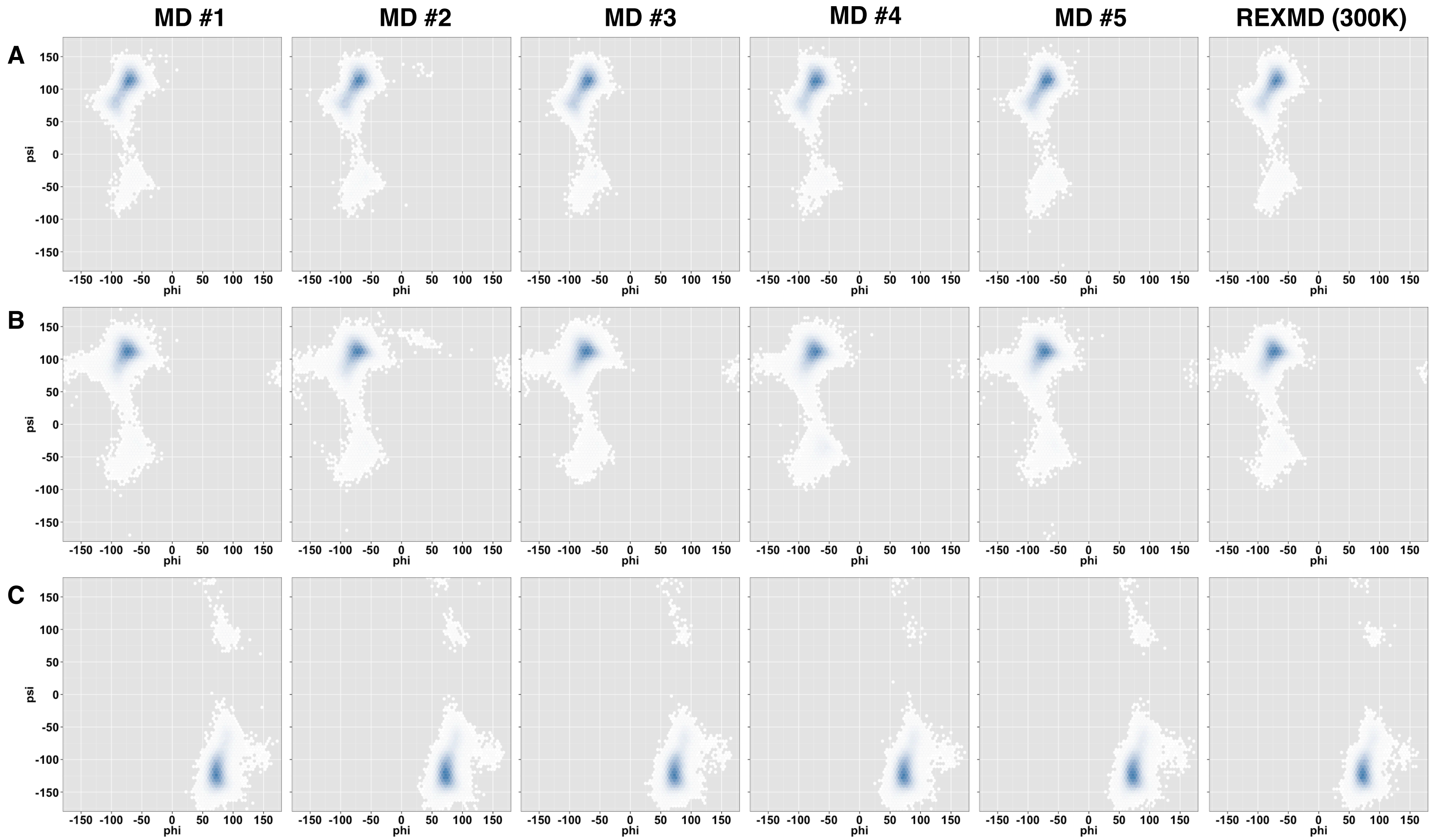


Figure S5.2: Torsion angle distribution for the first three glycosidic linkages. (A) GlcNAc β(1→4) GlcNAc, (B) Man β(1→4) GlcNAc, and (C) Man α(1→3) Man.

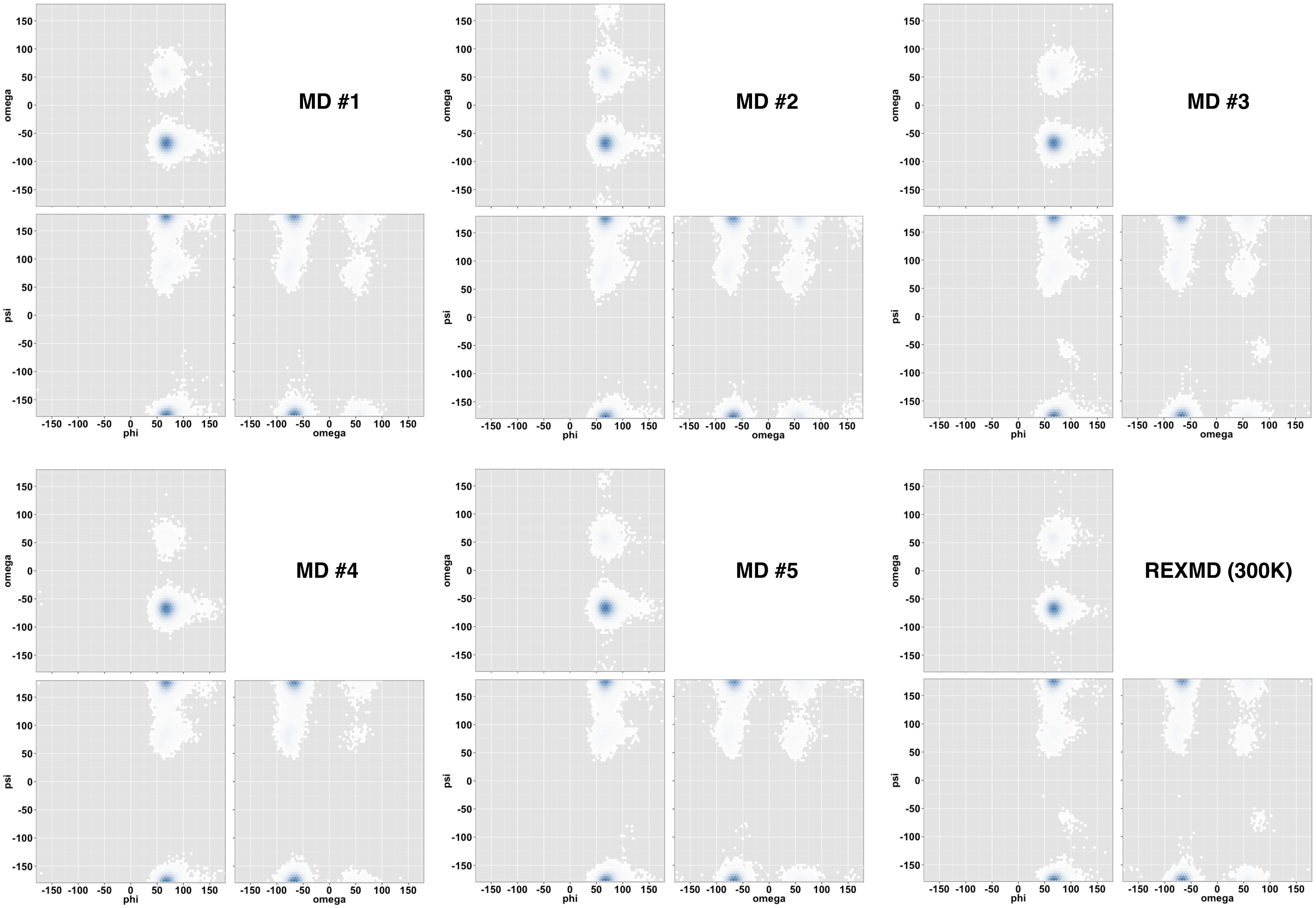


Figure S5.3: Torsion angle distribution for the Man $\alpha(1\rightarrow6)$ Man glycosidic linkage.

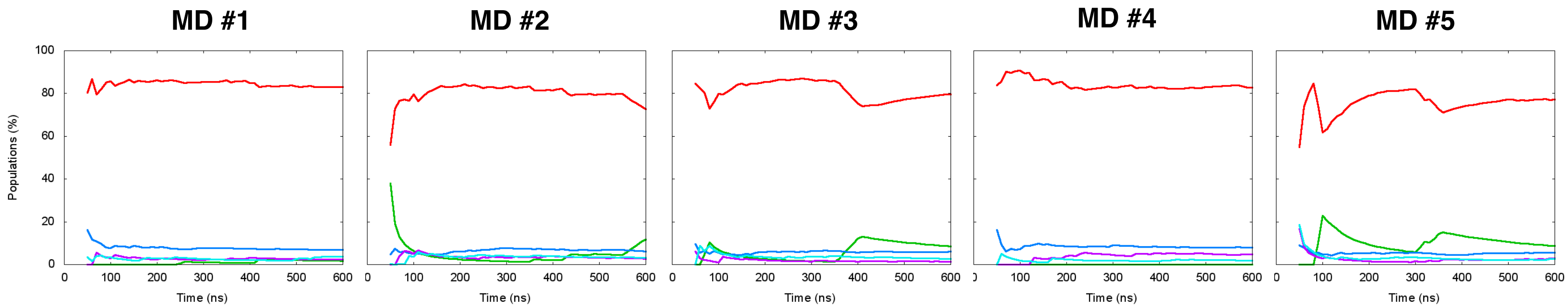


Figure S5.4: Cumulative average of conformational state population for the five largest conformational states. Each colored line represents the population of a particular conformational state.

Chapter 6

Application of homology modeling approach to N-glycan structure prediction

6.1 Introduction

One of the important biological roles of N-glycan is the molecular recognition and several N-glycans that are important in molecular recognition have been reported to date. Because molecular recognition is sensitive to small changes in glycan sequence and structure [31, 125, 126], it is important to understand the structural relationship between the N-glycan and the glycoprotein. However, such understanding remains a grand challenge mainly due to difficulties in preparing glycoprotein sample with homogeneous glycoform and solving the atomic resolution structure of N-glycans using crystallography or nuclear magnetic resonance (NMR). On the other hand, computational approach, particularly molecular dynamics (MD) simulation, can provide structure of glycan and glycoprotein in atomic detail if robust sampling is obtained. Recent advances in carbohydrate force field and efficient simulation software allow robust sampling of glycan conformation possible [27].

Although computational approach is a viable option in studying the structure and dynamics of glycan and glycoconjugate, the main challenge is the lack of known structure of glycan or glycoconjugates. Currently there is a large gap between the number of known glycoprotein and the number of solved glycoprotein structures. For example, 50% of all eukaryotic proteins are expected to be glycosylated [7, 156], while the PDB has only about 500 glycoprotein structural entries that are glycosylated (protein sequence similarity < 0.5). Solving the structure of aglycoprotien is more managable than solving the structure of glycoconjugate, thus methodology that can reliably model the glycan portion of glycoconjugate on top of a known protein structure may provide a reasonable model for further understanding of the role of glycans in biological processes.

The current stage of glycan structure modeling based on glycosylation information (glycosylation site and glycan primary sequence) is rudimentary at best, compared to the mature field of protein structure prediction/modeling. Since protein structure prediction and modeling have shown impressive successes and become mature [10, 150, 152, 155], it is natural to follow a similar paradigm for glycan structure prediction and modeling. However, unlike proteins where each amino acid is linearly connected by identical peptide bonds, each monomeric unit in a glycan chain is connected by different glycosidic linkages, e.g., $\alpha 1 \rightarrow 6$, $\beta 1 \rightarrow 4$, etc., and branched sequences are common. These features greatly increase the possible sequence and structure space and make it difficult to define and search homologous sequences and structures.

Due to high flexibility, it is not straightforward to build a reliable glycan model by simply connecting sugars based on most favorable glycosidic torsion angles of the corresponding disaccharides. Thus, applying template-based modeling approach to glycan structure modeling is appealing. The major challenge in applying the template-based approach is identifying "good" templates because the performace of the template-based approach depends on the quality of template structures. We have conducted a survey of N-glycans in glycoconjugate crystal structures and found that the N-glycan structures are sigficantly conserved across homologous proteins [63]. This suggests that the glycoprotein sequence similarity can be used to identify "good" N-glycan templates.

In this work, we have applied template-based modeling approach to predict glyan structure. A protocol for identifying homologous template is proposed and the quality of the produced structures is compared with simple modeling protocol, which generates structures by assembling favorable glycosidic torsion angles. For simplicity, no scoring function is used during modeling. Finally, potential improvements are discussed.

6.2 Methods

6.2.1 Basic glycan structure modeling protocol

The basic idea of glycan conformation sampling is based on the Monte Carlo sampling of glycosidic linkage torsion angle. Each glycosidic torsion angle movement is performed by assigning a new glycosidic torsion angle value for the selected glycosidic linkage. In general, new glycosidic torsion angle is selected from an "accessible" torsion angle region for the corresponding glycosidic linkage. The "accessible" torsion angle region of a glycosidic linkage is defined in the previous work [63]. Briefly, for each glycosidic linkage, adiabatic potential map was calculated by systematically changing the torsion angle of the corresponding glycosidic linkage with disaccharide in vacuum. The CHARMM carbohydrate force field [41, 42, 43] was used to evaluate the energy. The generated adiabatic potential energy map was converted to a torsion angle probability map using the Boltzmann distribution. A set of glycosidic torsion angle pairs having probability above 0.0001 were considered "accessible".

For a glycan sequence comprised of N residues and M terminal residues (an oligosaccharide chain can have more than one terminal residues due to branching) and L branches, a single time unit consists of N attempts of glycosidic torsion angle movement, M attempts of movements for the residues at each terminal, L attempts of movements for the residues at the beginning of each branches, and one attempt of movement for the residue at the beginning of N-glycan chain. Each trial movement is accepted or rejected according to Metropolis acceptance criteria. Before energy evaluation, a trial conformation that has steric collisons of glycan-glycan or glycan-protein is rejected.

6.2.2 Template-based glycan structure modeling protocol

template-based glycan structure modeling protocol is based on the basic glycan structure modeling protocol described above, but utilizes the template glycan structures found in the PDB. To identify the template structures, the protocol requires two pieces of information; desired N-glycan sequence and the sequence of the glycoprotein. The protocol first query the glycan fragment database [60] to find the list of PDB entries containing the matching glycan sequence. Query glycan sequence is converted into graph representation, so that not only the sequences match exactly but also the ones that match partially can be found. Second, from the list of PDB entries, only the ones containing protein sequence that is homologous to the query sequence are selected based on the sequence similarity (similarity > 50%). When no homologous PDB entries are found, the protocol either use non-homologous PDB entries or falls back to basic glycan structure modeling protocol.

Once the list of template structure is identified, sampling is performed. Basic procedure of sampling is identical, but, prior to each single time unit, one of the template structures is selected and the glycosidic torsion angle values from the template is applied and marked as 'fixed'. The glycosidic torsion angles marked as 'fixed' are only allowed to change ± 5° and the other glycosidic torsion angles are free to change. The rest of the procedure is identical.

6.3 Results

We have selected one of IgG1 domain crystal structures, PDB:1L6X, as a test case because it has several homologous glycan structures available in the PDB. To test the performance of template-based modeling protocol, we have generated 10,000 conformations using both the template-based glycan structure modeling protocol and the basic glycan structure modeling protocol. In template-based glycan structure modeling protocol, the PDB entries having sequence similarity of glycoprotein greater than or equal to 90% were excluded. The performance is measured in terms of root-mean-square distance (RMSD) with respect to the crystal structure.

Figure 6.1 shows the RMSD distribution of the conformations generated by the proposed

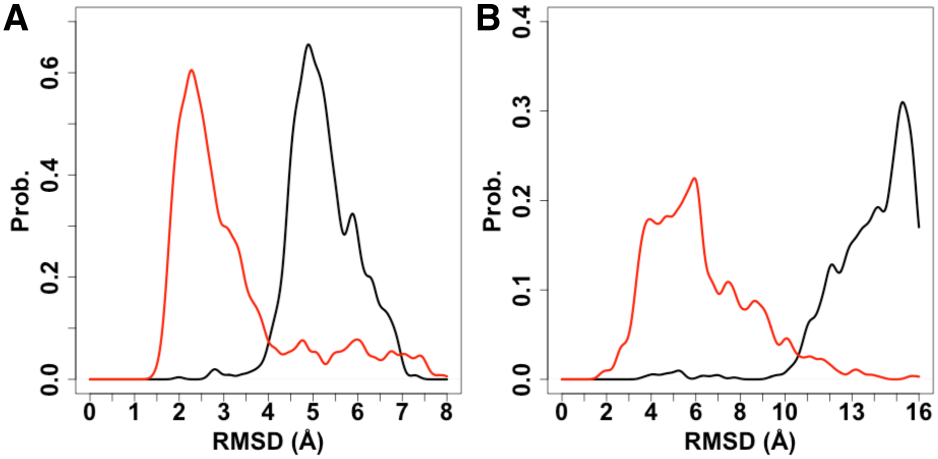


Figure 6.1: Performance of template-based structure prediction protocols. A) RMSD of N-glycan after alignment of glycan. B) RMSD of N-glycan after alignment of protein. The red lines represent the structures generated by template-based modeling protocol and the black lines represent the structures generated by basic modeling protocol.

modeling protocols. Clearly, the template-based structure modeling protocol performed much better than simple modeling protocol. When only the glycan structures are compared, template-based modeling protocol produced 12% and 61% of structures having RMSD less than 2 Å and 3 Å, respectively, with respect to the crystal structure, while the basic modeling protocol have not produced any structures having RMSD less than 3 Å.

While the glycan structure itself appeared to be successfully modeled by the template-based modeling protocol, however, the orientation of the N-glycan with respect to the glycoprotein performed seems more challenging. The majority of structures predicted by the template-based structure modeling protocol have 3-6 Å RMSD after the alignment of protein. While the result is still better than the basic modeling approach, some improvements are needed for the method to be useful. One potential reason for the poor performance could be the flexibility of glycosylated protein residues. Glycosylated residues are preferentially found at the loop region [103], thus, simply applying the torsion angles of glycosylated residues found in the PDB may reduce the performance of the modeling and more robust sampling at the glycosylated region may be necessary.

6.4 Conclusion and outlook

Despite the importance of glycan in several biological processes, our understanding of glycans remains elusive. Determining the structure of glycoconjugate is one of the challenging tasks. Due to recent advancement in mass spectrometry and structure biology, the atomic resolution structures of aglycoprotein and the glycosylation information are more accessible. Thus, a computational methodology that can reliably model the glycan conformation on top of protein structure will be useful to provide a reasonable model for further understanding of the role of glycan in important biological processes. Here, we have proposed a protocol for modeling of glycoconjugate using the known crystal structures and compared their performances.

Based on the benchmark, the template-based modeling approach performed much better than the simple modeling protocol. This is not unexpected because of the conserved glycan structures among the homologous glycoproteins, but the protocol presented here is the first demonstration of template-based modeling approach can be successfully applied in glycan modeling, to our knowledge. The glycan structure itself appeared to be successfully modelled, however, predicting the orientation with respect to protein may needs improvement. The majority of structures predicted by the template-based structure modeling protocol have 3–6 Å RMSD after the alignment of protein.

There are several improvements can be made in the protocol proposed here. First, the proposed protocols are not using any energy function but only excluded volume. The performance could be improved by adopting appropriate energy function during the modeling process. Secondly, the glycan structure with respect to the protein appears to be more challenging. Thus, additional sampling of glycan orientation with respect to the protein may improve the overall accuracy of the modeling.

In the future, more improvements and extensive benchmarks are warranted to prove the robustness of the proposed protocol. Nevertheless, the protocol proposed here can be useful to produce a reasonable initial model for MD simulations or refinement of low resolution model from experiments, such as small angle X-ray scattering and electron microscopy.

References

[1] Aebi, M., Bernasconi, R., Clerc, S., & Molinari, M. (2010). N-glycan structures: recognition and processing in the ER. *Trends Biochem. Sci.*, 35(2), 74–82.

[2] Akira, S., Uematsu, S., & Takeuchi, O. (2006). Pathogen recognition and innate immunity. *Cell*, 124(4), 783–801.

[3] Allen, F. H. & Taylor, R. (2004). Research applications of the Cambridge Structural Database (CSD). *Chem. Soc. Rev.*, 33(8), 463–475.

[4] Almond, A. (2005). Towards understanding the interaction between oligosaccharides and water molecules. *Carbohydr. Res.*, 340(5), 907–920.

[5] Almond, A., Petersen, B. O., & Duus, J. Ø. (2004). Oligosaccharides implicated in recognition are predicted to have relatively ordered structures. *Biochemistry*, 43(19), 5853–5863.

[6] Andersen, H. C. (1980). Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72(4), 2384.

[7] Apweiler, R., Hermjakob, H., & Sharon, N. (1999). On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, 1473(1), 4–8.

[8] Astronomo, R. D. & Burton, D. R. (2010). Carbohydrate vaccines: developing sweet solutions to sticky situations? *Nature Reviews Drug discovery*, 9(4), 308–324.

[9] Baenziger, J. U. (1985). The role of glycosylation in protein recognition. Warner-Lambert Parke-Davis Award lecture. *Am J Pathol*, 121(3), 382–91.

[10] Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, 294(5540), 93–6.

[11] Barb, A. W. & Prestegard, J. H. (2011). NMR analysis demonstrates immunoglobulin G N-glycans are accessible and dynamic. *Nature Chem. Biol.*, 7(3), 147–153.

[12] Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., & Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, 58(Pt 6 No 1), 899–907.

[13] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, 28(1), 235–242.

[14] Betancourt, M. R. & Skolnick, J. (2001). Universal similarity measure for comparing protein structures. *Biopolymers*, 59(5), 305–309.

[15] Bohne, A., Lang, E., & von der Lieth, C. W. (1999). SWEET - WWW-based rapid 3D construction of oligo- and polysaccharides. *Bioinformatics*, 15(9), 767–8.

[16] Bohne-Lang, A. & von der Lieth, C.-W. (2005). GlyProt: in silico glycosylation of proteins. *Nucleic Acids Res.*, 33(Web Server issue), W214–9.

[17] Brooks, B. R., Brooks, C. L., Mackerell, Jr, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., & Karplus, M. (2009). CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, 30(10), 1545–1614.

[18] Buchete, N.-V. & Hummer, G. (2008). Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B*, 112(19), 6057–6069.

[19] Case, D. A., Cheatham, T. E., I., Darden, T., Gohlke, H., Luo, R., Merz, K. M., J., Onufriev, A., Simmerling, C., Wang, B., & Woods, R. J. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16), 1668–88.

[20] ChemAxon Ltd (2007). MarvinSpace.

[21] Chen, M. M., Bartlett, A. I., Nerenberg, P. S., Friel, C. T., Hackenberger, C. P. R., Stultz, C. M., Radford, S. E., & Imperiali, B. (2010). Perturbing the folding energy landscape of the bacterial immunity protein Im7 by site-specific N-linked glycosylation. *Proc. Natl. Acad. Sci. USA*, 107(52), 22528–22533.

[22] Culyba, E. K. E., Price, J. L. J., Hanson, S. R. S., Dhar, A. A., Wong, C.-H. C., Gruebele, M. M., Powers, E. T. E., & Kelly, J. W. J. (2011). Protein native-state stabilization by placing aromatic side chains in N-glycosylated reverse turns. *Science*, 331(6017), 571–575.

[23] Davis, J. T., Hirani, S., Bartlett, C., & Reid, B. R. (1994). 1H NMR studies on an Asn-linked glycopeptide. GlcNAc-1 C2-N2 bond is rigid in H2O. *J. Biol. Chem.*, 269(5), 3331–3338.

[24] Deisenhofer, J. (1981). Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from Staphylococcus aureus at 2.9- and 2.8-Å resolution. *Biochemistry*, 20(9), 2361–2370.

[25] Dube, D. H. & Bertozzi, C. R. (2005). Glycans in cancer and inflammation — potential for therapeutics and diagnostics. *Nature Reviews Drug discovery*, 4(6), 477–488.

[26] Ellis, C. R., Maiti, B., & Noid, W. G. (2012). Specific and nonspecific effects of glycosylation. *J. Am. Chem. Soc.*, 134(19), 8184–8193.

[27] Fadda, E. & Woods, R. J. (2010). Molecular simulations of carbohydrates and protein-carbohydrate interactions: motivation, issues and prospects. *Drug Discov. Today*, 15(15-16), 596–609.

[28] Feig, M., Karanicolas, J., & Brooks, C. L. (2004). MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graphics Modell.*, 22(5), 377–395.

[29] Feller, S. E., Zhang, Y., Pastor, R. W., & Brooks, B. R. (1995). Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.*, 103(11), 4613.

[30] Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M., & Westbrook, J. (2004). Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, 20(13), 2153–5.

[31] Ferrara, C., Grau, S., Jäger, C., Sondermann, P., Brünker, P., Waldhauer, I., Hennig, M., Ruf, A., Rufer, A. C., Stihle, M., Umaña, P., & Benz, J. (2011). Unique carbohydrate-carbohydrate interactions are required for high affinity binding between FcgammaRIII and antibodies lacking core fucose. *Proc. Natl. Acad. Sci. U.S.A.*, 108(31), 12669–12674.

[32] Frank, M., Lütteke, T., & von der Lieth, C. W. (2007). GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res.*, 35(Database issue), 287–290.

[33] Fuster, M. M. & Esko, J. D. (2005). The sweet and sour of cancer: glycans as novel therapeutic targets. *Nature Reviews Cancer*, 5(7), 526–542.

[34] Gaile, G. L. & Burt, J. E. (1980). *Directional Statistics*. Geo Abstracts Limited.

[35] Galili, U. (2001). The α-Gal epitope (Galα1-3Galβ1-4GlcNAc-R) in xenotransplantation. *Biochimie*, 83(7), 557–563.

[36] Gao, M. & Skolnick, J. (2010). iAlign: a method for the structural comparison of protein-protein interfaces. *Bioinformatics*, 26(18), 2259–2265.

[37] Group, W. (2005-2011). GLYCAM Web (http://www.glycam.org).

[38] Guo, Y., Feinberg, H., Conroy, E., Mitchell, D. A., Alvarez, R., Blixt, O., Taylor, M. E., Weis, W. I., & Drickamer, K. (2004). Structural basis for distinct ligand-binding and targeting properties of the receptors DC-SIGN and DC-SIGNR. *Nat. Struct. Mol. Biol.*, 11(7), 591–598.

[39] Guttman, M., Kahn, M., Garcia, N. K., Hu, S.-L., & Lee, K. K. (2012). Solution structure, conformational dynamics, and CD4-induced activation in full-length, glycosylated, monomeric HIV gp120. *J. Virol.*, 86(16), 8750–8764.

[40] Guttman, M., Weinkam, P., Sali, A., & Lee, K. K. (2013). All-atom ensemble modeling to analyze small-angle x-ray scattering of glycosylated proteins. *Structure*, 21(3), 321–331.

[41] Guvench, O., Greene, S. N., Kamath, G., Brady, J. W., Venable, R. M., Pastor, R. W., & Mackerell Jr, A. D. (2008). Additive empirical force field for hexopyranose monosaccharides. *J. Comput. Chem.*, 29(15), 2543–2564.

[42] Guvench, O., Hatcher, E. R., Venable, R. M., Pastor, R. W., & MacKerell, Alexander D, J. (2009). CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J Chem Theory Comput*, 5(9), 2353–2370.

[43] Guvench, O., Mallajosyula, S. S., Raman, E. P., Hatcher, E. R., Vanommeslaeghe, K., Foster, T. J., Jamison, F. W., & Mackerell, Jr, A. D. (2011). CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate-protein modeling. *Journal of Chemical Theory and Computation*, 7(10), 3162–3180.

[44] Hanson, S. R., Culyba, E. K., Hsu, T.-L., Wong, C.-H., Kelly, J. W., & Powers, E. T. (2009). The core trisaccharide of an N-linked glycoprotein intrinsically accelerates folding and enhances stability. *Proc. Natl. Acad. Sci. U.S.A.*, 106(9), 3131–3136.

[45] Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K. F., Ueda, N., Hamajima, M., Kawasaki, T., & Kanehisa, M. (2006). KEGG as a glycome informatics resource. *Glycobiology*, 16(5), 63R–70R.

[46] Hatcher, E., Guvench, O., & MacKerell, Alexander D, J. (2009). CHARMM additive all-atom force field for aldopentofuranoses, methyl-aldopentofuranosides, and fructofuranose. *J Phys Chem B*, 113(37), 12466–12476.

[47] Hecht, M.-L., Stallforth, P., Silva, D. V., Adibekian, A., & Seeberger, P. H. (2009). Recent advances in carbohydrate-based vaccines. *Current opinion in chemical biology*, 13(3), 354–359.

[48] Helenius, A. & Aebi, M. (2004). Roles of N-linked glycans in the endoplasmic reticulum. *Annu. Rev. Biochem.*, 73, 1019–1049.

[49] Herget, S., Toukach, P. V., Ranzinger, R., Hull, W. E., Knirel, Y. A., & von der Lieth, C. W. (2008). Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC structural biology*, 8, 35.

[50] Ho, B. K., Thomas, A., & Brasseur, R. (2003). Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Protein science : a publication of the Protein Society*, 12(11), 2508–22.

[51] Homans, S. W., Dwek, R. A., Boyd, J., Mahmoudian, M., Richards, W. G., & Rademacher, T. W. (1986). Conformational transitions in N-linked oligosaccharides. *Biochemistry*, 25(20), 6342–6350.

[52] Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, 268, 1144–1149.

[53] Hoover, W. G. (1985). Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3), 1695.

[54] Hovmoller, S., Zhou, T., & Ohlson, T. (2002). Conformations of amino acids in proteins. *Acta crystallographica. Section D, Biological crystallography*, 58(Pt 5), 768–76.

[55] Idusogie, E. E., Presta, L. G., Gazzano-Santoro, H., Totpal, K., Wong, P. Y., Ultsch, M., Meng, Y. G., & Mulkerrin, M. G. (2000). Mapping of the C1q binding site on rituxan, a chimeric antibody with a human IgG1 Fc. *J. Immunol.*, 164(8), 4178–4184.

[56] Im, W., Beglov, D., & Roux, B. (1998). Continuum solvation model: Electrostatic forces from numerical solutions to the Poisson-Bolztmann equation. *Comput Phys Comm*, 111, 59–75.

[57] Imperiali, B. & Rickert, K. W. (1995). Conformational implications of asparagine-linked glycosylation. *Proceedings of the National Academy of Sciences of the United States of America*, 92(1), 97–101.

[58] Ito, Y., Hagihara, S., Matsuo, I., & Totani, K. (2005). Structural approaches to the study of oligosaccharides in glycoprotein quality control. *Curr. Opin. Struct. Biol.*, 15(5), 481–489.

[59] Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81, 158–171.

[60] Jo, S. & Im, W. (2013). Glycan fragment database: a database of PDB-based glycan 3D structures. *Nucleic Acids Res.*, 41(D1), D470–4.

[61] Jo, S., Kim, T., & Im, W. (2007). Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS ONE*, 2(9), e880.

[62] Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.*, 29(11), 1859–1865.

[63] Jo, S., Lee, H. S., Skolnick, J., & Im, W. (2013). Restricted N-glycan Conformational Space in the PDB and Its Implication in Glycan Structure Modeling. *PLoS Comp. Biol.*, 9(3), e1002946.

[64] Jo, S., Lim, J., Klauda, J., & Im, W. (2009). CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. *Biophys. J.*, 97, 50–58.

[65] Jo, S., Song, K. C., Desaire, H., Mackerell, Jr, A. D., & Im, W. (2011). Glycan Reader: automated sugar identification and simulation preparation for carbohydrates and glycoproteins. *J. Comput. Chem.*, 32(14), 3135–3141.

[66] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2), 926.

[67] Kamberaj, H. & van der Vaart, A. (2009). Extracting the causality of correlated motions from molecular dynamics simulations. *Biophys. J.*, 97(6), 1747–55.

[68] Kirschner, K. N. & Woods, R. J. (2001). Solvent interactions determine carbohydrate conformation. *Proc. Natl. Acad. Sci. U.S.A.*, 98(19), 10541–10545.

[69] Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., & Woods, R. J. (2008). GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.*, 29(4), 622–655.

[70] Konc, J. & Janezic, D. (2010). ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, 26(9), 1160–1168.

[71] Krapp, S., Mimura, Y., Jefferis, R., Huber, R., & Sondermann, P. (2003). Structural analysis of human IgG-Fc glycoforms reveals a correlation between glycosylation and structural integrity. *J. Mol. Biol.*, 325(5), 979–989.

[72] Lederkremer, G. Z. (2009). Glycoprotein folding, quality control and ER-associated degradation. *Curr. Opin. Struct. Biol.*, 19(5), 515–523.

[73] Lee, H. S. & Im, W. (2012). Identification of Ligand Templates using Local Structure Alignment for Structure-Based Drug Design. *Journal of Chemical Information and Modeling*.

[74] Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. U.S.A.*, 95(11), 5913–5920.

[75] Levy, R. M. & Gallicchio, E. (1998). Computer simulations with explicit solvent: Recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu Rev Phys Chem*, 49, 531–567.

[76] Liu, F.-T. & Rabinovich, G. A. (2005). Galectins as modulators of tumour progression. *Nature Reviews Cancer*, 5(1), 29–41.

[77] Live, D. H., Kumar, R. A., Beebe, X., & Danishefsky, S. J. (1996). Conformational influences of glycosylation of a peptide: a possible model for the effect of glycosylation on the rate of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 93(23), 12759–12761.

[78] Lütteke, T. (2009). Analysis and validation of carbohydrate three-dimensional structures. *Acta Crystallogr. D Biol. Crystallogr.*, 65(Pt 2), 156–168.

[79] Lütteke, T., Bohne-Lang, A., Loss, A., Goetz, T., Frank, M., & von der Lieth, C. W. (2006). GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology*, 16(5), 71R–81R.

[80] Lütteke, T., Frank, M., & von der Lieth, C.-W. (2004). Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr. Res.*, 339(5), 1015–1020.

[81] Lütteke, T., Frank, M., & von der Lieth, C.-W. (2005). Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res.*, 33(Database issue), D242–6.

[82] Marschinski, R. & Kantz, H. (2002). Analysing the information flow between financial time series - An improved estimator for transfer entropy. *Eur. Phys. J. B*, 30(2), 275–281.

[83] Martin, W. L., West, A. P., Gan, L., & Bjorkman, P. J. (2001). Crystal structure at 2.8 Å of an FcRn/heterodimeric Fc complex: mechanism of pH-dependent binding. *Mol. Cell*, 7(4), 867–877.

[84] Martin-Pastor, M. & Bush, C. A. (2000). Conformational studies of human milk oligosaccharides using (1)H-(13)C one-bond NMR residual dipolar couplings. *Biochemistry*, 39(16), 4674–4683.

[85] Martinez-Fleites, C., Macauley, M. S., He, Y., Shen, D. L., Vocadlo, D. J., & Davies, G. J. (2008). Structure of an O-GlcNAc transferase homolog provides insight into intracellular glycosylation. *Nat. Struct. Mol. Biol.*, 15(7), 764–765.

[86] Martyna, G. J., Tobias, D. J., & Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys.*, 101(5), 4177.

[87] McLellan, J. S., Pancera, M., Carrico, C., Gorman, J., Julien, J.-P., Khayat, R., Louder, R., Pejchal, R., Sastry, M., Dai, K., O'Dell, S., Patel, N., Shahzad-ul Hussan, S., Yang, Y., Zhang, B., Zhou, T., Zhu, J., Boyington, J. C., Chuang, G.-Y., Diwanji, D., Georgiev, I. I., Kwon, Y. D., Lee, D., Louder, M. K., Moquin, S., Schmidt, S. D., Yang, Z.-Y., Bonsignori, M., Crump, J. A., Kapiga, S. H., Sam, N. E., Haynes, B. F., Burton, D. R., Koff, W. C., Walker, L. M., Phogat, S., Wyatt, R., Orwenyo, J., Wang, L.-X., Arthos, J., Bewley, C. A., Mascola, J. R., Nabel, G. J., Schief, W. R., Ward, A. B., Wilson, I. A., & Kwong, P. D. (2011). Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature*, 480(7377), 336–343.

[88] Mimura, Y., Sondermann, P., Ghirlando, R., Lund, J., Young, S. P., Goodall, M., & Jefferis, R. (2001). Role of oligosaccharide residues of IgG1-Fc in Fc gamma RIIb binding. *J. Biol. Chem.*, 276(49), 45539–45547.

[89] Mizushima, T., Yagi, H., Takemoto, E., Shibata-Koyama, M., Isoda, Y., Iida, S., Masuda, K., Satoh, M., & Kato, K. (2011). Structural basis for improved efficacy of therapeutic antibodies on defucosylation of their Fc glycans. *Genes Cells*, 16(11), 1071–1080.

[90] Moens, S. & Vanderleyden, J. (1997). Glycoproteins in prokaryotes. *Arch. Microbiol.*, 168(3), 169–75.

[91] Molinari, M. (2007). N-glycan structure dictates extension of protein folding or onset of disposal. *Nat. Chem. Biol.*, 3(6), 313–320.

[92] Morell, A. G., Gregoriadis, G., Scheinberg, I. H., Hickman, J., & Ashwell, G. (1971). The role of sialic acid in determining the survival of glycoproteins in the circulation. *J. Biol. Chem.*, 246(5), 1461–7.

[93] Morelle, W. & Michalski, J. C. (2005). Glycomics and mass spectrometry. *Curr. Pharm. Des.*, 11(20), 2615–45.

[94] Nakahara, T., Hashimoto, R., Nakagawa, H., Monde, K., Miura, N., & Nishimura, S.-I. (2008). Glycoconjugate Data Bank: Structures–an annotated glycan structure database and N-glycan primary structure verification service. *Nucleic Acids Res.*, 36(Database issue), D368–71.

[95] Nina, M., Beglov, D., & Roux, B. (1997). Atomic Radii for Continuum Electrostatics Calculations based on Molecular Dynamics Free Energy Simulations. *J. Phys. Chem. B*, 101, 5239–5248.

[96] Nishima, W., Miyashita, N., Yamaguchi, Y., Sugita, Y., & Re, S. (2012). Effect of bisecting GlcNAc and core fucosylation on conformational properties of biantennary complex-type N-glycans in solution. *J. Phys. Chem. B*, 116(29), 8504–8512.

[97] Oganesyan, V., Damschroder, M. M., Leach, W., Wu, H., & Dall'Acqua, W. F. (2008). Structural characterization of a mutated, ADCC-enhanced human Fc fragment. *Mol. Immunol.*, 45(7), 1872–1882.

[98] Olsson, U., Säwén, E., Stenutz, R., & Widmalm, G. (2009). Conformational flexibility and dynamics of two (1-6)-linked disaccharides related to an oligosaccharide epitope expressed on malignant tumour cells. *Chem. Eur. J.*, 15(35), 8886–8894.

[99] Park, H.-S. & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.*, 36(2).

[100] Pejchal, R., Doores, K. J., Walker, L. M., Khayat, R., Huang, P.-S., Wang, S.-K., Stanfield, R. L., Julien, J.-P., Ramos, A., Crispin, M., Depetris, R., Katpally, U., Marozsan, A., Cupo, A., Maloveste, S., Liu, Y., McBride, R., Ito, Y., Sanders, R. W., Ogohara, C., Paulson, J. C., Feizi, T., Scanlan, C. N., Wong, C.-H., Moore, J. P., Olson, W. C., Ward, A. B., Poignard, P., Schief, W. R., Burton, D. R., & Wilson, I. A. (2011). A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science*, 334(6059), 1097–1103.

[101] Perić-Hassler, L., Hansen, H. S., Baron, R., & Huenenberger, P. H. (2010). Conformational properties of glucose-based disaccharides investigated using molecular dynamics simulations with local elevation umbrella sampling. *Carbohydr. Res.*, 345(12), 1781–1801.

[102] Petrescu, A. J., Butters, T. D., Reinkensmeier, G., Petrescu, S., Platt, F. M., Dwek, R. A., & Wormald, M. R. (1997). The solution NMR structure of glucosylated N-glycans involved in the early stages of glycoprotein biosynthesis and folding. *EMBO J.*, 16(14), 4302–4310.

[103] Petrescu, A.-J., Milac, A.-L., Petrescu, S. M., Dwek, R. A., & Wormald, M. R. (2004). Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology*, 14(2), 103–114.

[104] Petrescu, A. J., Petrescu, S. M., Dwek, R. A., & Wormald, M. R. (1999). A statistical analysis of N- and O-glycan linkage conformations from crystallographic data. *Glycobiology*, 9(4), 343–352.

[105] Petrescu, A.-J., Wormald, M. R., & Dwek, R. A. (2006). Structural aspects of glycomes with a focus on N-glycosylation and glycoprotein folding. *Curr. Opin. Struct. Biol.*, 16(5), 600–607.

[106] Phillips, J. C., Braun, R., Wang, W., Gumbart, J. C., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26(16), 1781–1802.

[107] Porter, L. L. & Rose, G. D. (2011). Redrawing the Ramachandran plot after inclusion of hydrogen-bonding constraints. *Proc. Natl. Acad. Sci. USA*, 108(1), 109–113.

[108] Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7, 95–9.

[109] Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Protein Chem.*, 23, 283–438.

[110] Raman, E. P., Guvench, O., & MacKerell, Alexander D, J. (2010). CHARMM additive all-atom force field for glycosidic linkages in carbohydrates involving furanoses. *J. Phys. Chem. B*, 114(40), 12981–12994.

[111] Rao, V. S. R. (1998). *Conformation of carbohydrates*. Australia: Harwood Academic Publishers.

[112] Re, S., Miyashita, N., Yamaguchi, Y., & Sugita, Y. (2011). Structural diversity and changes in conformational equilibria of biantennary complex-type N-glycans in water revealed by replica-exchange molecular dynamics simulation. *Biophys. J.*, 101(10), L44–6.

[113] Rose, D. R. (2012). Structure, mechanism and inhibition of Golgi a-mannosidase II. *Curr. Opin. Struct. Biol.*, (pp. 1–5).

[114] Roux, B. (1997). The influence of the membrane potential on the free energy of an intrinsic protein. *Biophys. J.*, 73, 2980–2989.

[115] Roy, A. & Zhang, Y. (2012). Recognizing Protein-Ligand Binding Sites by Global Structural Alignment and Local Geometry Refinement. *Structure*.

[116] Rubinstein, M. & Colby, R. H. (2003). *Polymer Physics*. Oxford: Oxford University Press.

[117] Rudd, P. M., Wormald, M. R., & Dwek, R. A. (2004). Sugar-mediated ligand–receptor interactions in the immune system. *Trends Biotech.*, 22(10), 524–530.

[118] Ryckaert, J.-P., Ciccotti, G., & Berendsen, H. J. C. (1977). Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.*, 23(3), 327–341.

[119] Salisburg, A. M., Deline, A. L., Lexa, K. W., Shields, G. C., & Kirschner, K. N. (2009). Ramachandran-type plots for glycosidic linkages: Examples from molecular dynamic simulations using the Glycam06 force field. *J. Comput. Chem.*, 30(6), 910–921.

[120] Schreiber, T. (2000). Measuring information transfer. *Phys. Rev. Lett.*, 85(2), 461–4.

[121] Schwarz, F. & Aebi, M. (2011). Mechanisms and principles of N-linked protein glycosylation. *Curr. Opin. Struct. Biol.*, 21(5), 576–582.

[122] Shah, N., Kuntz, D. A., & Rose, D. R. (2008). Golgi alpha-mannosidase II cleaves two sugars sequentially in the same catalytic site. *Proc. Natl. Acad. Sci. USA*, 105(28), 9570–9575.

[123] Shental-Bechor, D. & Levy, Y. (2009). Folding of glycoproteins: toward understanding the biophysics of the glycosylation code. *Curr. Opin. Struct. Biol.*, 19(5), 524–533.

[124] Shi, X. & Elliott, R. M. (2004). Analysis of N-linked glycosylation of hantaan virus glycoproteins and the role of oligosaccharide side chains in protein folding and intracellular trafficking. *J. Virol.*, 78(10), 5414–5422.

[125] Shinya, K., Ebina, M., Yamada, S., Ono, M., Kasai, N., & Kawaoka, Y. (2006). Avian flu: influenza virus receptors in the human airway. *Nature*, 440(7083), 435–436.

[126] Siebert, H.-C., André, S., Lu, S.-Y., Frank, M., Kaltner, H., van Kuik, J. A., Korchagina, E. Y., Bovin, N., Tajkhorshid, E., Kaptein, R., Vliegenthart, J. F. G., von der Lieth, C.-W., Jiménez-Barbero, J., Kopitz, J., & Gabius, H.-J. (2003). Unique conformer selection of human growth-regulatory lectin galectin-1 for ganglioside GM1 versus bacterial toxins. *Biochemistry*, 42(50), 14762–14773.

[127] Skehel, J. J. & Wiley, D. C. (2000). Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu. Rev. Biochem.*, 69, 531–569.

[128] Slynko, V., Schubert, M., Numao, S., Kowarik, M., Aebi, M., & Allain, F. H. T. (2009). NMR structure determination of a segmentally labeled glycoprotein using in vitro glycosylation. *J. Am. Chem. Soc.*, 131(3), 1274–1281.

[129] Solá, R. J., Rodríguez-Martínez, J. A., & Griebenow, K. (2007). Modulation of protein biophysical properties by chemical glycosylation: biochemical insights and biomedical implications. *Cell. Mol. Life Sci.*, 64(16), 2133–2152.

[130] Sugita, Y. & Okamoto, Y. (2000). Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.*, (329), 261–270.

[131] Säwén, E., Massad, T., Landersjö, C., Damberg, P., & Widmalm, G. (2010). Population distribution of flexible molecules from maximum entropy analysis using different priors as background information: application to the ϕ, ψ-conformational space of the α-(1→2)-linked mannose disaccharide present in N- and O-linked glycoproteins. *Org. Biomol. Chem.*, 8(16), 3684–3695.

[132] Säwén, E., Stevensson, B., Östervall, J., Maliniak, A., & Widmalm, G. (2011). Molecular Conformations in the Pentasaccharide LNF-1 Derived from NMR Spectroscopy and Molecular Dynamics Simulations. *J. Phys. Chem. B*, 115(21), 7109–7121.

[133] Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *J. Comput. Chem.*, 26(16), 1701–18.

[134] Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., & MacKerell, A D, J. (2010). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31(4), 671–690.

[135] Varki, A. (1993). Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, 3(2), 97–130.

[136] Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., & Etzler, M. E. (2009). *Essentials of glycobiology*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press, 2nd edition.

[137] Watt, G. M., Lund, J., Levens, M., Kolli, V. S. K., Jefferis, R., & Boons, G.-J. (2003). Site-Specific Glycosylation of an Aglycosylated Human IgG1-Fc Antibody Protein Generates Neoglycoproteins with Enhanced Function. *Chem. Biol.*, 10(9), 807–814.

[138] Wehle, M., Vilotijevic, I., Lipowsky, R., Seeberger, P. H., Varon Silva, D., & Santer, M. (2012). Mechanical Compressibility of the Glycosylphosphatidylinositol (GPI) Anchor Backbone Governed by Independent Glycosidic Linkages. *J. Am. Chem. Soc.*

[139] Weis, W. I. & Drickamer, K. (1996). Structural basis of lectin-carbohydrate recognition. *Annu. Rev. Biochem.*, 65, 441–473.

[140] Woods, R. J., Pathiaseril, A., Wormald, M. R., Edge, C. J., & Dwek, R. A. (1998). The high degree of internal flexibility observed for an oligomannose oligosaccharide does not alter the overall topology of the molecule. *Eur. J. Biochem.*, 258(2), 372–386.

[141] Wooten, E. W., Bazzo, R., Edge, C. J., Zamze, S., Dwek, R. A., & Rademacher, T. W. (1990). Primary sequence dependence of conformation in oligomannose oligosaccharides. *Eur. Biophys. J.*, 18(3), 139–148.

[142] Wormald, M. R. & Dwek, R. A. (1999). Glycoproteins: glycan presentation and protein-fold stability. *Structure*, 7(7), R155–60.

[143] Wormald, M. R., Petrescu, A.-J., Pao, Y., Glithero, A., Elliott, T., & Dwek, R. A. (2002). Conformational studies of oligosaccharides and glycopeptides: Complementarity of NMR, X-ray crystallography, and molecular modelling. *Chem. Rev.*, 102(2), 371–386.

[144] Wormald, M. R., Rudd, P. M., Harvey, D. J., Chang, S. C., Scragg, I. G., & Dwek, R. A. (1997). Variations in oligosaccharide-protein interactions in immunoglobulin G determine the site-specific glycosylation profiles and modulate the dynamic motion of the Fc oligosaccharides. *Biochemistry*, 36(6), 1370–1380.

[145] Wormald, M. R., Wooten, E. W., Bazzo, R., Edge, C. J., Feinstein, A., Rademacher, T. W., & Dwek, R. A. (1991). The conformational effects of N-glycosylation on the tailpiece from serum IgM. *Eur. J. Biochem.*, 198(1), 131–139.

[146] Xu, J. & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26(7), 889–895.

[147] Yang, F., Bewley, C. A., Louis, J. M., Gustafson, K. R., Boyd, M. R., Gronenborn, A. M., Clore, G. M., & Wlodawer, A. (1999). Crystal structure of cyanovirin-N, a potent HIV-inactivating protein, shows unexpected domain swapping. *J. Mol. Biol.*, 288(3), 403–412.

[148] York, D. M., Darden, T. A., & Pedersen, L. G. (1993). The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the Ewald and truncated list methods. *J. Chem. Phys.*, 99(10), 8345.

[149] Zemla, A., Venclovas, C., Moult, J., & Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins*, Suppl 3, 22–29.

[150] Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.*, 18(3), 342–348.

[151] Zhang, Y. (2009). Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.*, 19(2), 145–155.

[152] Zhang, Y. & Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. U.S.A.*, 101(20), 7594–7599.

[153] Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, 33(7), 2302–2309.

[154] Zhong, W., Kuntz, D. A., Ember, B., Singh, H., Moremen, K. W., Rose, D. R., & Boons, G.-J. (2008). Probing the substrate specificity of Golgi alpha-mannosidase II by use of synthetic oligosaccharides and a catalytic nucleophile mutant. *J. Am. Chem. Soc.*, 130(28), 8975–8983.

[155] Zhou, H. & Skolnick, J. (2007). Ab initio protein structure prediction using chunk-TASSER. *Biophys. J.*, 93(5), 1510–1518.

[156] Zielinska, D. F., Gnad, F., WiSniewski, J. R., & Mann, M. (2010). Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell*, 141(5), 897–907.

Appendix A

List of Publications

1. Jo, Sunhwan, Song, Kevin C., Desaire, Heather, Mackerell, Alexander D., Jr and Im, Wonpil Glycan Reader: automated sugar identification and simulation preparation for carbohydrates and glycoproteins. *J. Comput. Chem.* 32, 3135–3141 (2011).

2. Jo, Sunhwan and Im, Wonpil Glycan fragment database: a database of PDB-based glycan 3D structures. *Nucleic Acids Res.* 41, D470–4 (2013).

3. Jo, Sunhwan, Lee, Hui Sun, Skolnick, Jeffrey and Im, Wonpil Restricted N-glycan Conformational Space in the PDB and Its Implication in Glycan Structure Modeling. *PLoS Computational Biology* 9, e1002946 (2013).

Appendix B

List of All Publications

1. Jo, S., Lee, H. S., Skolnick, J. & Im, W. Restricted N-glycan Conformational Space in the PDB and Its Implication in Glycan Structure Modeling. PLoS Computational Biology 9, e1002946 (2013).

2. Jo, S., Jiang, W., Lee, H. S., Roux, B. & Im, W. CHARMM-GUI Ligand Binder for Absolute Binding Free Energy Calculations and Its Application. J. Chem. Inf. Model. 53, 267–277 (2013).

3. Jo, S. & Im, W. Glycan fragment database: a database of PDB-based glycan 3D structures. Nucleic Acids Res. 41, D470–4 (2013).

4. Zhong, D. et al. The Salmonella Type III Secretion System Inner Rod Protein PrgJ Is Partially Folded. J. Biol. Chem. 287, 25303–25311 (2012).

5. Lee, H. S., Jo, S., Lim, H.-S. & Im, W. Application of Binding Free Energy Calculations to Prediction of Binding Modes and Affinities of MDM2 and MDMX Inhibitors. J. Chem. Inf. Model. (2012). doi:10.1021/ci3000997

6. Kwon, T. et al. Molecular dynamics simulations of the cx26 hemichannel: insights into voltage-dependent loop-gating. Biophys. J. 102, 1341–1351 (2012).

7. Im, W., Jo, S. & Kim, T. An ensemble dynamics approach to decipher solid-state NMR observables of membrane proteins. Biochim. Biophys. Acta 1818, 252–262 (2012).

8. Lee, K. I. et al. Web interface for brownian dynamics simulation of ion transport and its applications to beta-barrel pores. J. Comput. Chem. 33, 331–339 (2012).

9. Jo, S., Song, K. C., Desaire, H., Mackerell, A. D., Jr & Im, W. Glycan Reader: automated sugar identification and simulation preparation for carbohydrates and glycoproteins. J. Comput. Chem. 32, 3135–3141 (2011).

10. Kim, T., Jo, S. & Im, W. Solid-State NMR Ensemble Dynamics as a Mediator between Experiment and Simulation. Biophys. J. 100, 2922–2928 (2011).

11. Jo, S. & Im, W. Transmembrane helix orientation and dynamics: insights from ensemble dynamics with solid-state NMR observables. Biophys. J. 100, 2913–2921 (2011).

12. Lee, J. H. et al. Novel Pyrrolopyrimidine-based α-helix mimetics: cell-permeable inhibitors of proteins-protein interactions. J. Am. Chem. Soc. 133, 676-679 (2011).

13. Jo, S., Rui, H., Lim, J. B., Klauda, J. B. & Im, W. Cholesterol flip-flop: insights from free energy simulation studies. J. Phys. Chem. B 114, 13342–13348 (2010).

14. Jo, S., Lim, J. B., Klauda, J. B. & Im, W. CHARMM-GUI Membrane Builder for mixed bilayers and its application to yeast membranes. Biophys. J. 97, 50–58 (2009).

15. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. J. Comput. Chem. 29, 1859–1865 (2008).

16. Jo, S., Vargyas, M., Vasko-Szedlar, J., Roux, B. & Im, W. PBEQ-Solver for online visualization of electrostatic potential of biomolecules. Nucleic Acids Res. 36, W270–5 (2008).

17. Jo, S., Kim, T. & Im, W. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. PLoS ONE 2, e880 (2007).

Appendix C

License Terms and Conditions

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Feb 18, 2013

This is a License Agreement between Sunhwan Jo ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3091951016358
License date	Feb 18, 2013
Licensed content publisher	John Wiley and Sons
Licensed content publication	Journal of Computational Chemistry
Licensed content title	Glycan reader: Automated sugar identification and simulation preparation for carbohydrates and glycoproteins
Licensed copyright line	Copyright © 2011 Wiley Periodicals, Inc.
Licensed content author	Sunhwan Jo,Kevin C. Song,Heather Desaire,Alexander D. MacKerell,Wonpil Im
Licensed content date	Aug 3, 2011
Start page	3135
End page	3141
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Total	**0.00 USD**
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a "Wiley Company") or a society for whom a Wiley Company has exclusive publishing rights in relation to a particular journal (collectively WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your Rightslink account (these are available at any time at http://myaccount.copyright.com)

Terms and Conditions

1. The materials you have requested permission to reproduce (the "Materials") are protected by copyright.

2. You are hereby granted a personal, non-exclusive, non-sublicensable, non-transferable, worldwide, limited license to reproduce the Materials for the purpose specified in the licensing process. This license is for a one-time use only with a maximum distribution equal to the number that you identified in the licensing process. Any form of republication granted by this licence must be completed within two years of the date of the grant of this licence (although copies prepared before may be distributed thereafter). The Materials shall not be used in any other manner or for any other purpose. Permission is granted subject to an

appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Material. Any third party material is expressly excluded from this permission.

3. With respect to the Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Materials without the prior permission of the respective copyright owner. You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Materials, or any of the rights granted to you hereunder to any other person.

4. The Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc or one of its related companies (WILEY) or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto.

5. NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

6. WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

7. You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

8. IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

9. Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

10. The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

11. This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

12. Any fee required for this permission shall be non-refundable after thirty (30) days from receipt.

13. These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

14. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

15. WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

16. This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

17. This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

Wiley Open Access Terms and Conditions

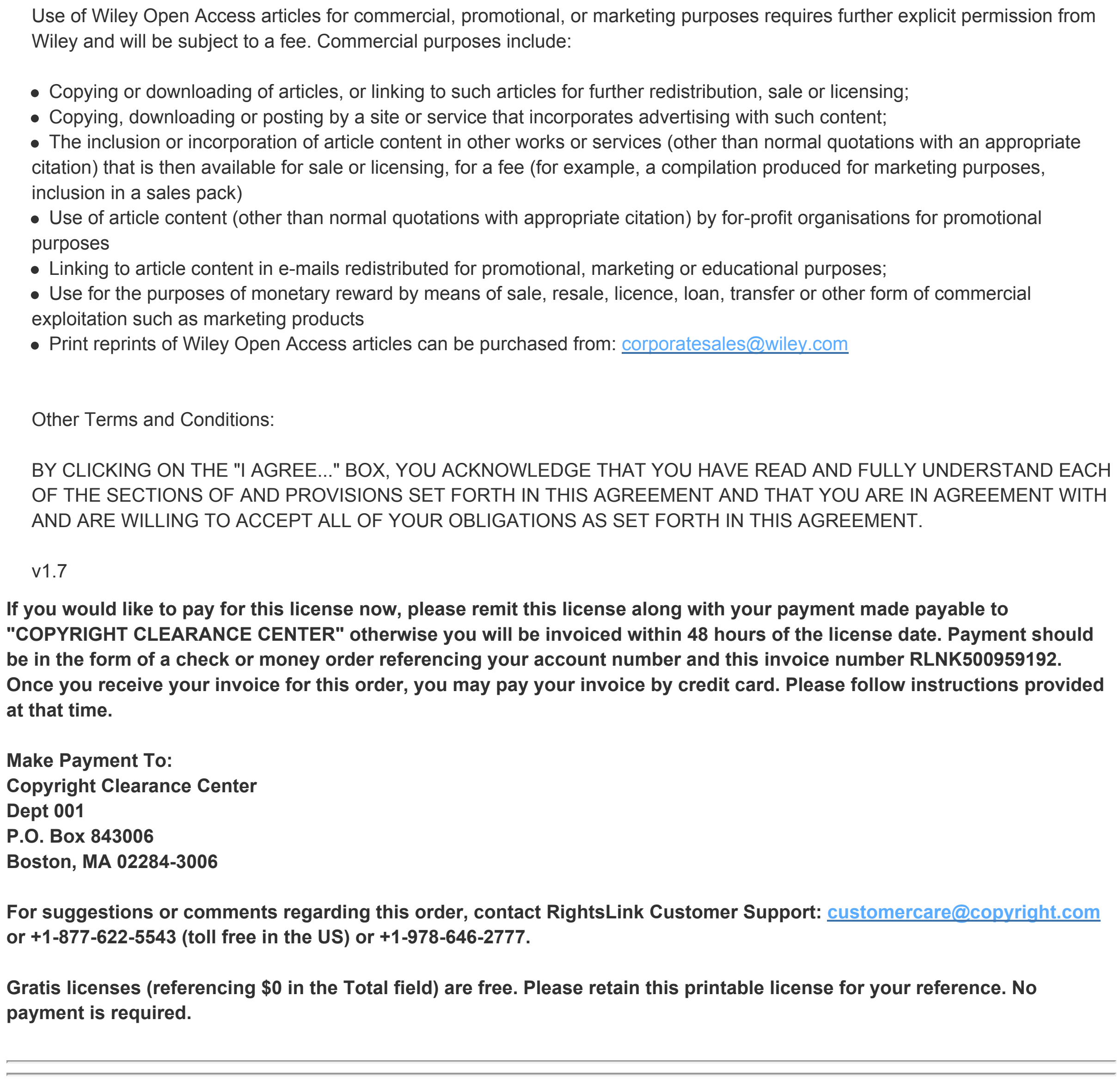
All research articles published in Wiley Open Access journals are fully open access: immediately freely available to read, download and share. Articles are published under the terms of the Creative Commons Attribution Non Commercial License. which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes. The license is subject to the Wiley Open Access terms and conditions:
Wiley Open Access articles are protected by copyright and are posted to repositories and websites in accordance with the terms of the Creative Commons Attribution Non Commercial License. At the time of deposit, Wiley Open Access articles include all changes made during peer review, copyediting, and publishing. Repositories and websites that host the article are responsible for incorporating any publisher-supplied amendments or retractions issued subsequently.
Wiley Open Access articles are also available without charge on Wiley's publishing platform, **Wiley Online Library** or any successor sites.

Use by non-commercial users

For non-commercial and non-promotional purposes individual users may access, download, copy, display and redistribute to colleagues Wiley Open Access articles, as well as adapt, translate, text- and data-mine the content subject to the following conditions:

- The authors' moral rights are not compromised. These rights include the right of "paternity" (also known as "attribution" - the right for the author to be identified as such) and "integrity" (the right for the author not to have the work altered in such a way that the author's reputation or integrity may be impugned).
- Where content in the article is identified as belonging to a third party, it is the obligation of the user to ensure that any reuse complies with the copyright policies of the owner of that content.
- If article content is copied, downloaded or otherwise reused for non-commercial research and education purposes, a link to the appropriate bibliographic citation (authors, journal, article title, volume, issue, page numbers, DOI and the link to the definitive published version on Wiley Online Library) should be maintained. Copyright notices and disclaimers must not be deleted.
- Any translations, for which a prior translation agreement with Wiley has not been agreed, must prominently display the statement: "This is an unofficial translation of an article that appeared in a Wiley publication. The publisher has not endorsed this translation."

Use by commercial "for-profit" organisations



Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee. Commercial purposes include:

- Copying or downloading of articles, or linking to such articles for further redistribution, sale or licensing;
- Copying, downloading or posting by a site or service that incorporates advertising with such content;
- The inclusion or incorporation of article content in other works or services (other than normal quotations with an appropriate citation) that is then available for sale or licensing, for a fee (for example, a compilation produced for marketing purposes, inclusion in a sales pack)
- Use of article content (other than normal quotations with appropriate citation) by for-profit organisations for promotional purposes
- Linking to article content in e-mails redistributed for promotional, marketing or educational purposes;
- Use for the purposes of monetary reward by means of sale, resale, licence, loan, transfer or other form of commercial exploitation such as marketing products
- Print reprints of Wiley Open Access articles can be purchased from: corporatesales@wiley.com

Other Terms and Conditions:

BY CLICKING ON THE "I AGREE..." BOX, YOU ACKNOWLEDGE THAT YOU HAVE READ AND FULLY UNDERSTAND EACH OF THE SECTIONS OF AND PROVISIONS SET FORTH IN THIS AGREEMENT AND THAT YOU ARE IN AGREEMENT WITH AND ARE WILLING TO ACCEPT ALL OF YOUR OBLIGATIONS AS SET FORTH IN THIS AGREEMENT.

v1.7

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500959192. Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Feb 18, 2013

This is a License Agreement between Sunhwan Jo ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3091951193405
License date	Feb 18, 2013
Licensed content publisher	Oxford University Press
Licensed content publication	Nucleic Acids Research
Licensed content title	Glycan fragment database: a database of PDB-based glycan 3D structures:
Licensed content author	Sunhwan Jo, Wonpil Im
Licensed content date	01/01/2013
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Glycan
Publisher of your work	n/a
Expected publication date	May 2013
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD
Terms and Conditions	

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources

which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oxfordjournals.org

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

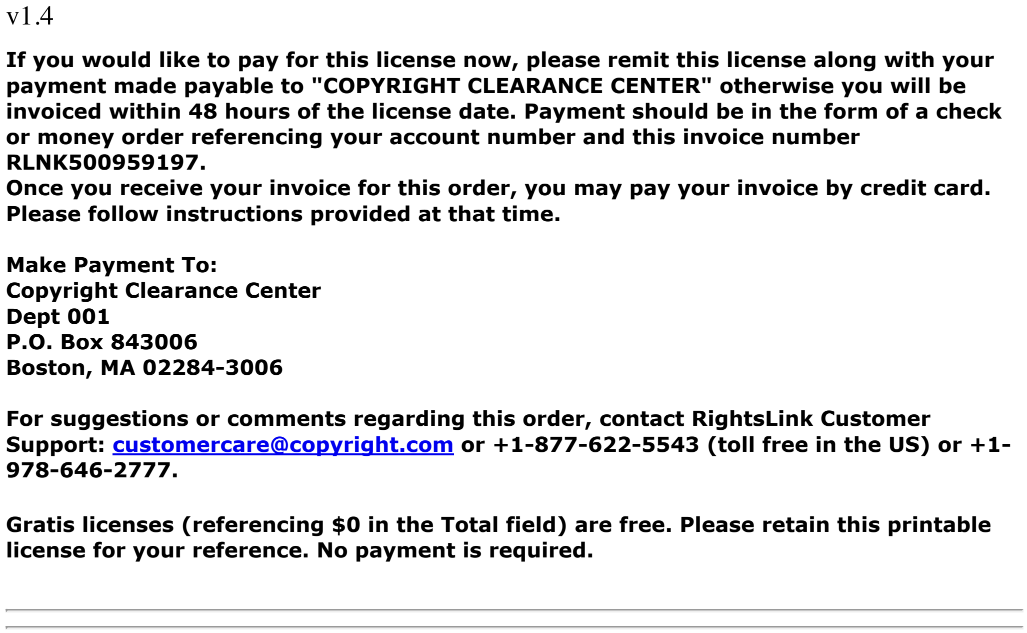
8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:



v1.4

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500959197.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.