ADAPTING MOLECULAR DOCKING TO STUDY SIALYLATED GLYCANS

by

HUIMIN HU

(Under the Direction of Robert J. Woods)

ABSTRACT

Sialic acids are some of the most important molecules of life because they often occupy the terminal position in glycans attached to proteins and lipids on cell membranes, and are involved in many biological and pathological recognition phenomena. These interactions are difficult to study experimentally, hence, prediction of the geometry and binding affinity of sialic acid-protein interactions is of great significance. Molecular docking algorithms aiming to be applicable to a broad range of ligands suffer reduced accuracy when applied to carbohydrates since they are unable to incorporate ligand-specific conformational energies. This thesis reports the development of a set Carbohydrate Intrinsic (CHI) energy functions that quantify the conformational properties of sialic acid linkages (Neu5Acα-2-3Gal and Neu5Acα-2-6Gal) in glycans, and incorporates these functions into the docking software AutoDock Vina.

INDEX WORDS: Carbohydrate docking, sialylated glycans, binding affinity, CHI energy function, Boltzmann factor, AutoDock Vina, Vina-Carb, glycosidic torsion angle.

ADAPTING MOLECULAR DOCKING TO STUDY SIALYLATED GLYCANS

by

HUIMIN HU

B.S., University of Science and Technology of China, 2013

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2015

ADAPTING MOLECULAR DOCKING TO STUDY SIALYLATED GLYCANS

by

HUIMIN HU

Major Professor:	Robert J. Woods
Committee:	Jeffrey L. Urbauer Ryan Hili

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2015

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to the following people:

My major professor, Dr. Robert J Woods, under whose advisement I have become a rigorous and independent researcher. There is no doubt that the skills and expertise I gained in his lab will take me very far in life.

Dr. Jeffrey L. Urbauer and Dr. Ryan Hili, for serving as my committee members and helping me throughout my graduate study.

A special thanks to Anita Nivedha, now, Dr. Nivedha, whose guidance in helping me become familiar with my project as well as the software and basic computation skills, have proven invaluable.

I would like to thank Mark Baine for his cheering me up during the tough times and putting me in my place when I got too full of myself. I would also like to express my gratitude to fellow lab members. Xiao, Mia, Arunima, Amika, and Lachele - hearty thanks to each of you for all you do.

Finally, I would like to acknowledge my parents. Without their love and support, I wouldn't be in a position today to follow my dreams.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

This thesis is mainly focused on computational docking of sialylated glycans. With the incorporation of carbohydrate intrinsic (CHI) energy functions of sialic acid linkages (Neu5Acα-2-3-Gal and Neu5Acα-2-6-Gal), we can better quantify the conformational energy of glycans, in order to improve the docking accuracy of sialylated glycan-protein systems. Chapter 2 describes the historical background, structural classification and significance of glycobiology in human health and disease. Chapter 3 covers the general introduction of docking, the docking program used in this study: AutoDock Vina, and the limitation of current docking program on carbohydrate docking and corresponding solution. Chapter 4 covers the structural and biological significance of sialic acids, methods being used in improving the docking accuracy and results analysis. In closing, Chapter 5 summarizes the conclusions of the work, acknowledges challenges and future work directions.

CHAPTER 2

CARBOHYDRATES AND GLYCANS

2.1 Historical Background and Overview

Carbohydrates comprise one of the four major classes (nucleic acids, proteins, carbohydrates and lipids) of organic molecules in living system, and are the most abundant organic molecules found in nature [1]. The term carbohydrate arose from the fact that most sugars have the empirical formula $C_nH_{2n}O_n$. Structurally, it is more accurate to define them as polyhydroxyaldehydes or polyhydroxyketones.

Synonyms for carbohydrate include saccharide and glycan. Glycans are defined by IUPAC as "compounds consisting of a large number of monosaccharides linked glycosidically" [2].

2.2 Classification of Carbohydrates/Saccharides

Monosaccharides are the smallest building block of glycans. A monosaccharide is a carbohydrate that cannot be hydrolyzed into a simpler unit. Free monosaccharides can exist in open chain or ring forms. Three characteristics are used to classified monosaccharides: the placement of carbonyl group, the number of atom it contains, and its chiral handedness. The monosaccharide is an aldose if the carbon group is an aldehyde, or it is a ketose if the carbonyl group is a ketone (See Figure 2.1). Monosaccharides with three carbon atoms are called trioses; those with four, five, six carbon atoms are called tetroses, pentoses, hexoses and so on. These two classifications are usually combined. i.e. glucose is an aldohexose.

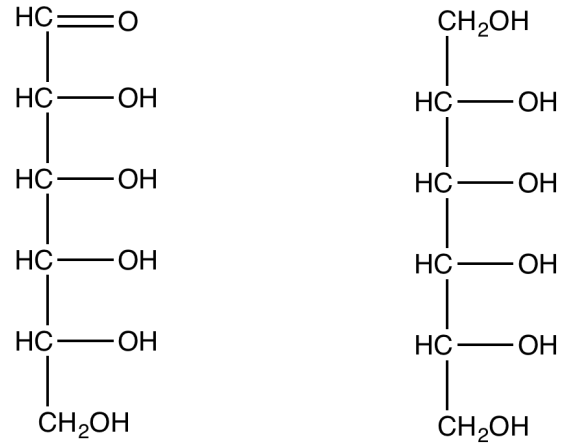


Figure 2.1 Example of six-carbon aldose (aldohexose) and six-carbon ketose (ketohexose)

With the exception of the terminal carbons, each hydroxyl-substituted carbon represents a chiral center, resulting in 2^n possible stereoisomer per monosaccharide, where n is the number of chiral centers. According to the orientation of the asymmetric carbon furthest carbonyl group, the assignment of D- or L- sugar is made. In a standard Fischer projection, if the hydroxyl group on the furthest carbonyl group is on the right the molecule is a D- sugar; otherwise it is an L- sugar.

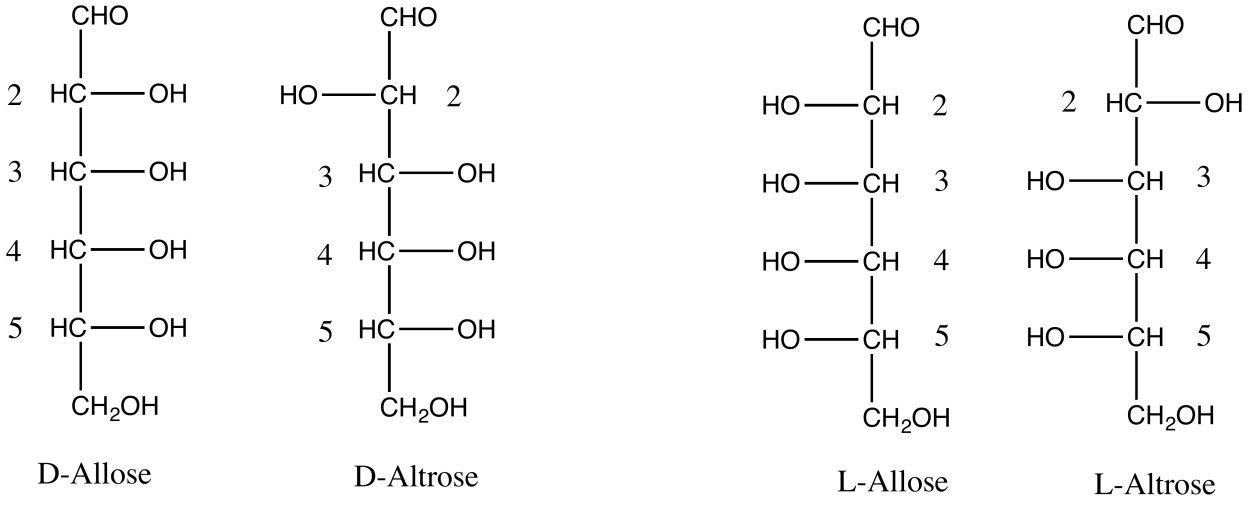


Figure 2.2 Two examples of Fischer projections for the D-aldohexose and L-aldohexose

During the conversion from linear form sugar to cyclic form sugar, the carbon atom containing the carbonyl oxygen, called the anomeric carbon, becomes a chiral center with two possible configurations: α anomer and β anomer (See Figure 2.3). In the α anomer, the hydroxyl-substituent on the anomeric carbon rests on the trans position of the ring from the CH_2OH side branch. In the β anomer, the hydroxyl-substituent on the anomeric carbon rests on the cis position of the ring from the CH_2OH side branch.

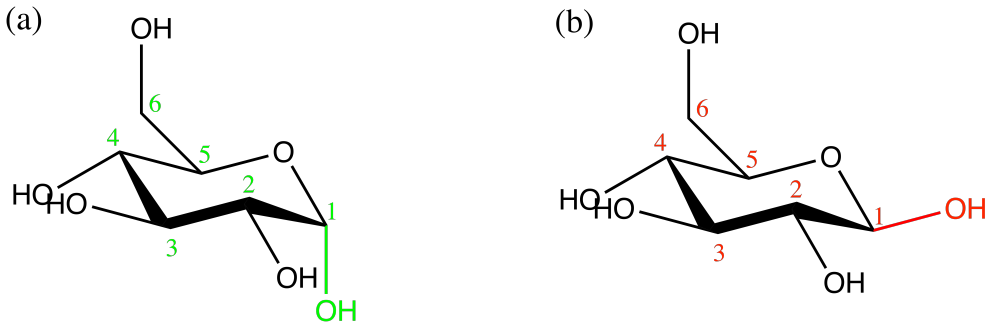


Figure 2.3 (a) α anomer of glucose; (b) β anomer of glucose

Two monosaccharides can combine by releasing a single water molecule via the condensation of the anomeric hydroxyl group of one monosaccharide with a hydroxyl group of a second monosaccharide. The C-O-C bridge formed between two monosaccharide is called glycosidic linkage (See Figure 2.4). For 1-2, 1-3 and 1-4 connections, ϕ ($O_{5'}$-$C_{1'}$-O_x-C_x) and ψ ($C_{1'}$-O_x-C_x-C_{x-1}) glycosidic torsion angles are used to describe the conformation of the linked monosaccharides. For 1-6 connections, an additional dihedral angle, ω (O_6-C_6-C_5-O_5), is defined. In the case of sialic acids, such as α-N-acetylneuraminic acid (αNeu5Ac), C_2 is the anomeric carbon, ϕ and ψ are alternatively defined as $O_{5'}$-$C_{2'}$-O_x-C_x and $C_{2'}$-O_x-C_x-C_{x-1}.

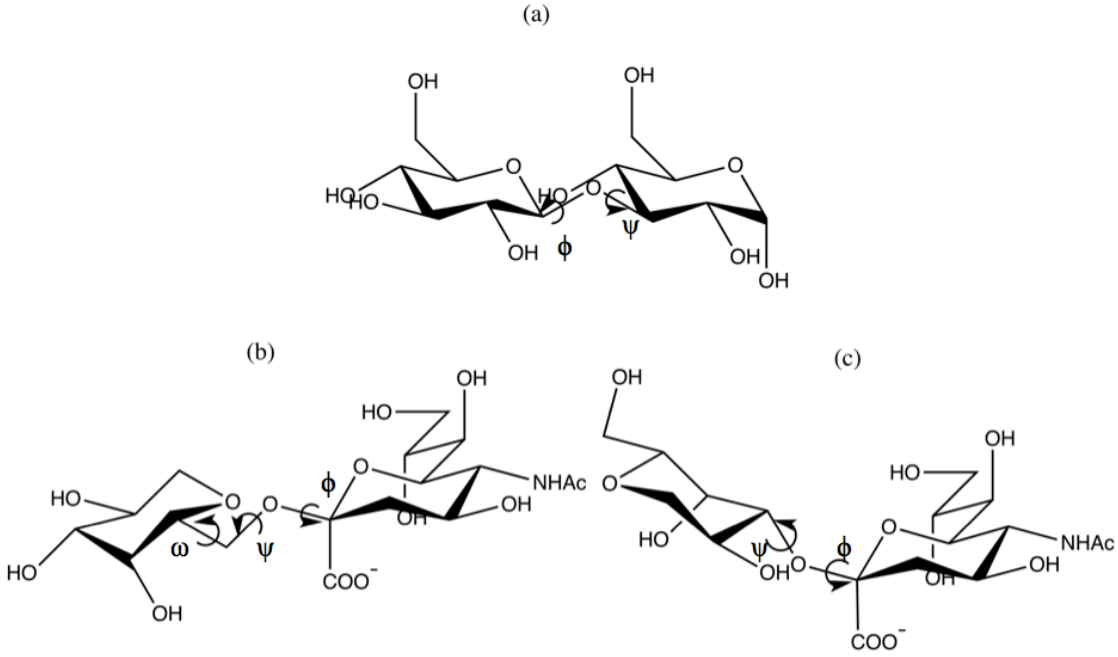


Figure 2.4 Given examples presented: (a) Glcβ-1-3-Glcα; (b) Neu5Acα-2-6-Galβ; (c) Neu5Acα-2-3-Galβ.

Oligosaccharide refers to a saccharide polymer containing a small number (usually 3~9) of monosaccharides and polysaccharide is a long-chain carbohydrate made up of monosaccharides.

2.3 Biological Significance of Glycobiology

Generally, glycobiology is defined as the study of structure, biosynthesis and biology of saccharides (glycans) that are widely distributed in nature. Saccharides play numerous roles in living organisms. For example, polysaccharides can serve for the storage of energy and as structural components; the ribose, a 5-carbon monosaccharide, is a significant component of the genetic molecule known as RNA and the related deoxyribose is a component of DNA [3]. Saccharides on cell surfaces play key roles in biological recognition phenomena, such as

fertilization, immune recognition and pathogen adhesion [4]. Since the monosaccharide units can be coupled to each other in many different ways, variability is seen in saccharide structures, allowing them to function as signaling molecules, recognition molecules and adhesion molecules on complex cell surface. As a result, cell surface glycans are involved in many physiologically important functions that include normal embryonic development, differentiation, growth, contact inhibition, cell–cell recognition, cell signaling, host–pathogen interaction during infection, host immune response, disease development, intracellular trafficking and localization, rate of degradation and membrane rigidity [5].

CHAPTER 3

DOCKING: A COMPUTATIONAL METHOD TO PREDICT 3D STRUCTURES OF GLYCANS

3.1 Brief Introduction of Docking

Docking is a computational method that predicts the preferred orientation of one molecule to a second. Information of the preferred orientation can be used to estimate the binding affinity between two molecules and rank the most likely binding mode of the ligand. Docking can be defined as a "lock-and-key" problem, which was first presented by *Hermann Emil Fischer* in 1900 [6]. The protein can be thought as the "lock" and the ligand can be thought as the "key". The goal of docking is to find the correct orientation of the "key" which can open the "lock". However, since both the ligand and protein are flexible, docking can be described more accurately by the "induced-fit" model. First illustrated by *Daniel Edward Koshland* [7], "induced-fit" refers to the docking process that the ligand and the protein achieve an overall best-fit by adjusting their conformations.

As a computational method, docking is much faster and cheaper than the traditional X-ray crystallography to determine the three-dimensional structures of interacting molecules. Despite the uncertainty in the accuracy of docking, it is the only alternative when a molecule is resistant to crystallization, such as very flexible glycans. Therefore, docking plays a crucial role in predicting glycan-protein binding orientation and inhibiting glycan-protein becomes a developing area of drug design [8]. One of the main uses of docking in drug discovery is virtual

screening (VS), which is a computational technique used to search libraries of small molecules aiming to identify those structures which are most likely to bind to a drug target. By using VS, one can score, rank or filter a set of structures, helping to quickly screen a large database of potential drugs.

Another main usage of docking is to predict the geometry of the ligand in the protein-binding site, which is also called "pose prediction". With the awareness of where and how the ligand binds to the protein, first, it helps us understand the critical part of ligand for binding; second, it can be used to suggest changes to improve the binding affinity.

There are various approaches to docking based on the systems it is applied to. In the field of biochemistry, docking may be performed to study protein-ligand, protein-protein, and DNA-ligand interactions. As a result, many of docking programs have been developed, such as AutoDock, DOCK and GLIDE. Each of these docking programs uses different search algorithms and scoring functions.

3.2 Docking Program: AutoDock Vina

A docking program is made up of two main components - search algorithm and scoring function. First, the search algorithm generates the poses, orientations of particular conformations of the molecule in the binding site; second, the scoring function calculates a score or theoretical binding affinity for the generated poses. Given a score, the program is then able to rank the predicted orientations in terms of their strength of binding.

AutoDock Vina (Vina) [9] is a molecular docking program, especially effective for protein-ligand docking. Compared with the previous software (AutoDock 4 [10]), it achieves an approximately two orders of magnitude speed-up and a significant improvement of average

prediction accuracy from 49% to 72% [9]. (Accuracy is measured by root mean squared deviation (RMSD)).

A variety of stochastic global optimization approaches were explored during the development of Vina, such as genetic algorithm, particle swarm optimization, simulated annealing and others. Finally, the Iterated Local Search optimizer was selected [9]. In this algorithm, a set of steps consisting of a mutation and a local optimization are taken. An efficient quasi-Newton method, Broyden-Fletcher-Shanno (BFGS), is used for the local optimization in the implementation. Like other quasi-Newton optimization methods, BFGS uses not only the value of scoring function but also its gradient, i.e., the derivatives of the scoring function with regard to the position and orientation of the ligand, as well as the values of the torsions for the active rotatable bonds in the ligand and flexible residues.

Generally, a scoring function (See Equation 3.1) is developed by taking part of the interaction energy from traditional physical concepts and scaling it empirically to reproduce an experimental training set.

$$\Delta G_{binding} = \sum_{interactions} f_i E_i \tag{3.1}$$

The interactions (E_i) might include hydrogen bonds, electrostatic interactions, hydrophobic contacts and solvent exclusion volume, etc. As shown in the equation, each contribution has an adjustable weighting factor (f_i).

The Vina scoring function [9] includes six terms, with different weighting factors for each, derived from experimental database (See Table 3.1).

Table 3.1. AutoDock Vina scoring function terms and weights

Weight (f_i)	Terms (E_i)
-0.0356	$gauss_1$
-0.00516	$gauss_2$
0.840	Repulsion
-0.00351	Hydrophobic
-0.587	Hydrogen Bonding
0.0585	N_{rot}

The first three terms represent the steric interactions, which are identical for all atom pairs. The steric terms are as follows (See Equation 3.2, 3.3, 3.4. d is the distance between two atoms):

$$gauss_1(d) = e^{-(\frac{d}{0.5\text{Å}})^2} \tag{3.2}$$

$$gauss_2(d) = e^{-(\frac{d-3\text{Å}}{2\text{Å}})^2} \tag{3.3}$$

$$repulsion(d) = \begin{cases} d^2, & if\ d\ < 0 \\ 0, & if\ d \geq 0 \end{cases} \tag{3.4}$$

The hydrophobic term (See Equation 3.5) attempts to quantify the interaction between hydrophobic atoms (usually refers to carbon in carbohydrates, the atom which can not form a hydrogen bond with hydrogen):

$$hydrophobic(d) = \begin{cases} 0, & if\ d > 1.5\text{Å} \\ d - 0.5, & if\ 0.5\text{Å} \le d \le 1.5\text{Å} \\ 1, & if\ d < 0.5\text{Å} \end{cases} \quad (3.5)$$

The hydrogen bonding term (See Equation 3.6) is similar to the hydrophobic term and is applied to pairs of atoms potentially capable of forming hydrogen bonds:

$$hydrogen\ bonding(d) = \begin{cases} 0, & if\ d > 0\text{Å} \\ -\frac{10}{7}d, & if\ -0.7\text{Å} \le d \le 0\text{Å} \\ 1, & if\ d < -0.7\text{Å} \end{cases} \quad (3.6)$$

N_{rot} term stands for the number of active rotatable bonds between heavy atoms in the ligand and is used to estimate the entropic penalty that occurs when a ligand binds.

Vina was designed to be compatible with the file format used for previous AutoDock input structure files: PDBQT, which can be regarded as an extension of the PDB format. AutoDock Tools is used to prepare the files, to choose the search space and to view the results.

Usually, in preparing oligosaccharide ligand files, we allow the glycosidic torsion angles and the hydroxyl groups to be flexible and keep other bonds rigid (See Figure 3.1).

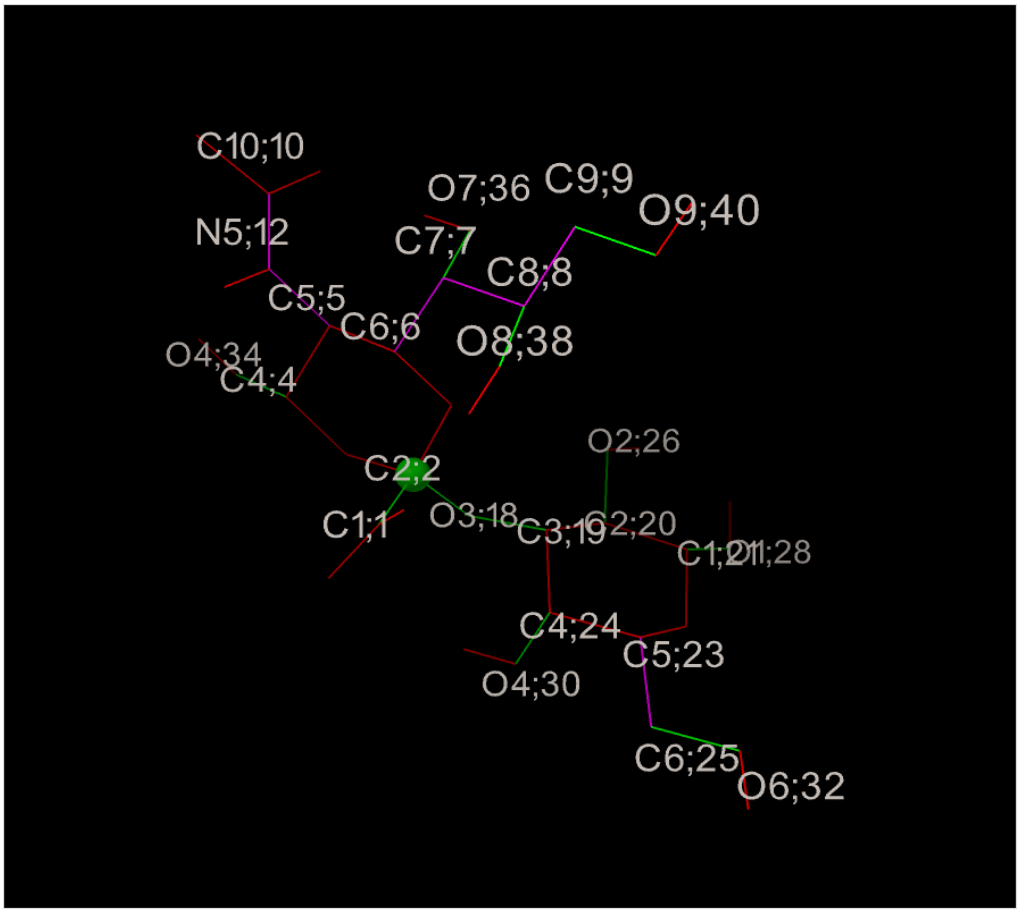


Figure 3.1 Preparation of a disaccharide ligand in AutoDock Tools. Green stands for the active bonds while pink and red stands for the inactive bonds.

After the preparation of files, typically, a grid box can be placed over the binding site to choose the search space (See Figure 3.2).

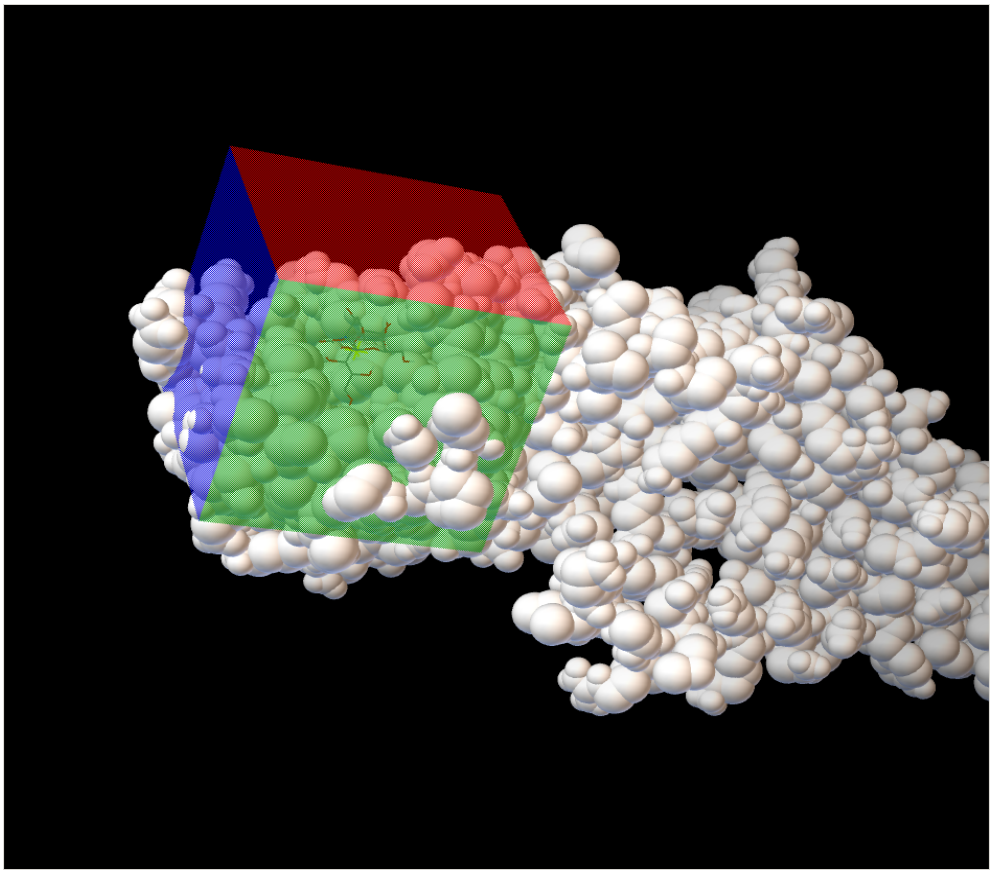


Figure 3.2 Choosing docking space by placing box

3.3 Motivation for the Development of Carbohydrate Intrinsic (CHI) Energy Function

In order to be applicable to various types of ligands, docking programs do not consider the specific conformational preferences of ligands. As illustrated in Table 3.1, the docking score only considers the interactions between protein and ligand. However, carbohydrates have strong conformational preference due to the gauche effect, the exo-anomeric effect and hyperconjugation, etc. Ignoring them lowers the accuracy of docking prediction.

Hence, there is a necessity to develop a set of additional scoring function to quantify the conformational energy in carbohydrates. With this goal, a set of Carbohydrate Intrinsic (CHI)

energy functions has been developed to account for the rotational glycosidic torsional angles in carbohydrates [11]Nivedha, Makeneni et al. (2013). Independent from their use in docking, CHI energy functions could also potentially be employed to evaluate the conformational energies of experimentally determined oligosaccharide structures. So far, CHI energy functions applicable to 1-2, 1-3, and 1-4 glycosidic linkages have been developed [11]. In this thesis, CHI energy functions are developed for the ϕ torsion angle associated with 2-3, and 2-6 linkages involving sialic acids. The conformational preferences of the phi angle in these linkages, like all phi-angles in oligosaccharides, are determined by the exo-anomeric effect. The exo-anomeric effect is a stereoelectronic effect that arises from interactions between the electron orbitals of the ring oxygen atoms and the exo-cyclic anomeric oxygen atom [12]. This effect results in phi-angle in sialic acids preferring only 2 out of the three possible staggered rotamers. In this thesis a set of CHI energy functions is developed to quantify the conformational energy of sialylated glycans, and their performance assessed by comparing the structures generated by docking sialylated oligosaccharides to protein receptors.

CHAPTER 4

SYSTEM UNDER STUDY: SIALIC ACIDS

4.1 Sialic Acids Structure and Biological Significance

Sialic acid is a generic term for the *N*- or *O*- substituted derivatives neuraminic acids, a monosaccharide with a nine-carbon backbone. It is also the name for the most common member of this group, N-acetylneuraminic acid (Neu5Ac) (See Figure 4.1).

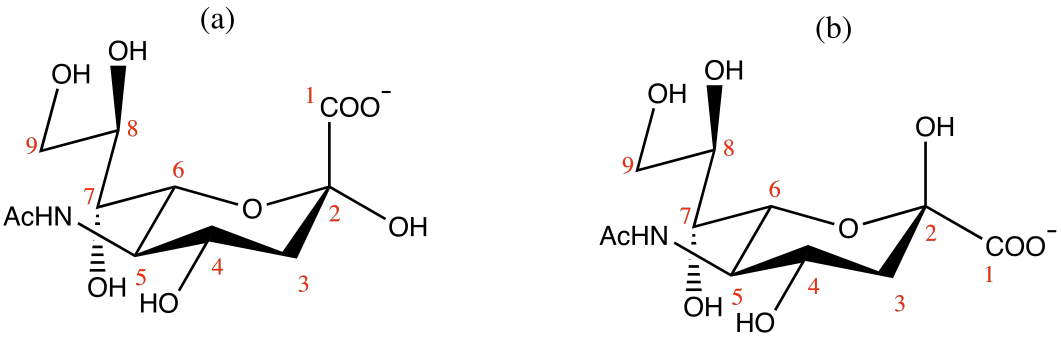


Figure 4.1 (a) α anomer of Neu5Ac; (b) β anomer of Neu5Ac

Widely distributed in animal tissues, sialic acid is one of the most important molecules involved in many biological and pathological phenomena [13]. Sialic acids are typically found linked to the 3- or 6- positions in Galactopyranose (Gal) or N-Acetyl-Galactopyranose (GalNAc) residues, but may also link to other sialic acids through the 7-, 8-, or 9- positions. The configuration of the sialic acid anomeric (C-2) carbon in these linkages is exclusively α, in contrast to most other linkages that can be either α or β.

Suggested by their high expression on outer cell membranes, on the interior of lysosomal membranes and on glycoproteins, sialic acids play an important role in the stabilization of molecules and membranes, as well as in modulating interactions with the environment [14]. Due to the relatively strong electronegative charge of the carboxylate group attached at C-2, sialic acids have the functions such as binding and transport of ions and drugs [15], stabilizing the conformation of protein [16] and enhancing the viscosity of mucins [17]. They can also act either as masks or recognition sites. In the first case, for example, penultimate galactose (Gal) residue can be masked by sialic acids. After sialic acids loss, molecules and cells can be bound via Gal-recognizing receptor [18]. On the other hand, sialic acids themselves can be served as ligand for a plenty of microbial and animal lectins [19] .

One of the popular glycoproteins which is responsible for binding the virus to cells with sialic acids on membrane is influenza hemagglutinin (HA), found on the surface of influenza virus. HA is a homotrimeric integral membrane glycoprotein, shaped like a cylinder, with three spherical heads containing the sialic acids binding sites.

In the first stages of infection by influenza, HA on the virus surface interacts with host cells by binding to sialylated glycans on the host cell surfaces. As mentioned above, the common linkages of sialic acid are to the C-3 and C-6 positions of galactose (See Figure 4.2).

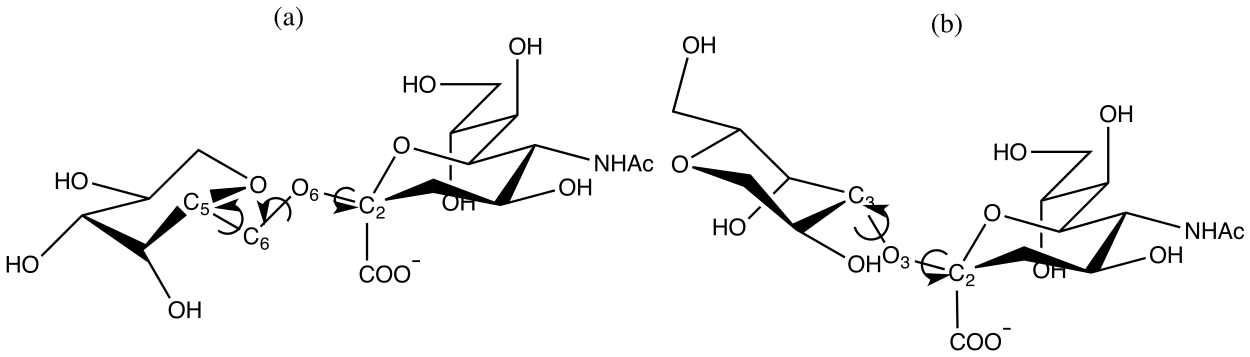


Figure 4.2 (a) Neu5Acα-2-6-Gal; (b) Neu5Acα-2-3-Gal

Avian influenza virus strains prefer to bind to sialic acids attached to galactose via an α(2,3) linkage (Neu5Acα-2-3-Gal) [20]. On the contrary, human influenza virus strains is likely to bind via an α(2,6) linkage(Neu5Acα-2-6-Gal) [21]. My research focuses on these two specific sialylated glycans.

4.2 Method: Developing CHI Energy Function

In contrast to previous work [22], where CHI energy functions were derived by fitting to theoretical (quantum mechanical) energy curves, here the CHI energy functions were developed based on the conformational distributions observed from experimental data from X-ray crystallography. As shown in Figure 4.3, the ϕ torsion angle in Neu5Acα-2-3-Gal in crystallographic complexes adopts two main conformations, at approximately ϕ = 60° and ϕ = 300° with a ratio of approximately 6:4.

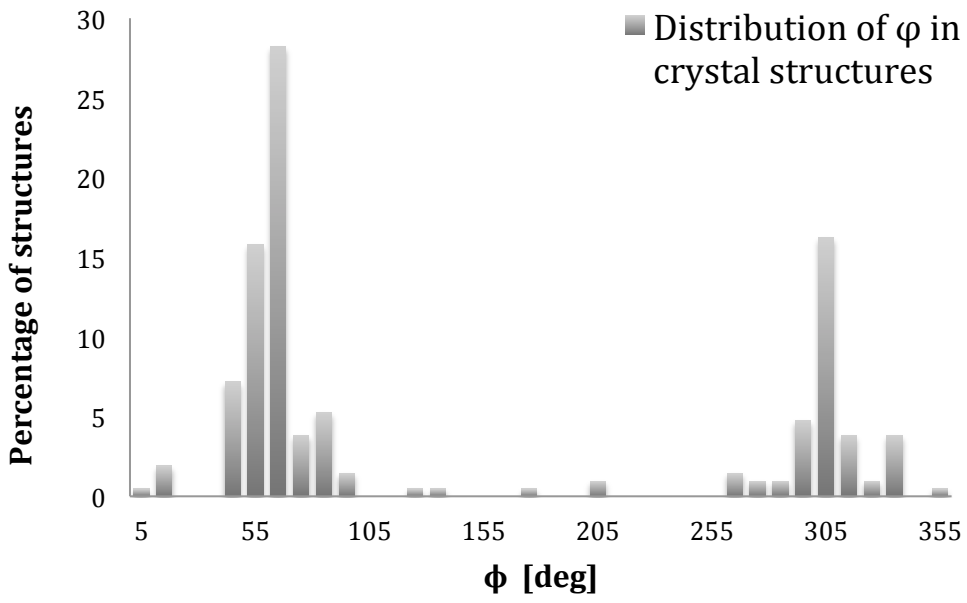


Figure 4.3 ϕ torsion angle distribution of Neu5Acα-2-3-Gal from X-ray data using the GlyTorsion tool at glycosience.de [23]

Similar to the developed CHI energy functions applicable to 1-2, 1-3 and 1-4 glycosidic linkages [11], quantum mechanical (QM) calculations were performed at first using Guassian09 [24] for ϕ torsion angle in Neu5Acα-2-3-Gal. Simplified structures based on analogs of tetrahydropyran (See Figure 4.4) were optimized at the B3LYP/6-31G++(2d, 2p) level of theory, and single-point energies calculated at the same level.

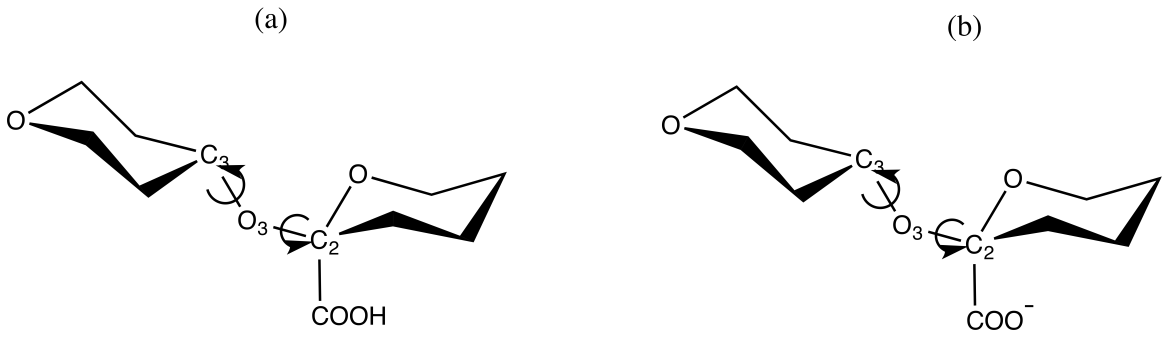


Figure 4.4 Two simplified models of Neu5Acα-2-3-Gal employed in QM optimization and single -point energy calculation: (a) with carboxylic acid; (b) with carboxylate

The ϕ-rotational energy profiles were computed at 15° increments, while other coordinates were allowed to be completely relaxed (See Figure 4.5).

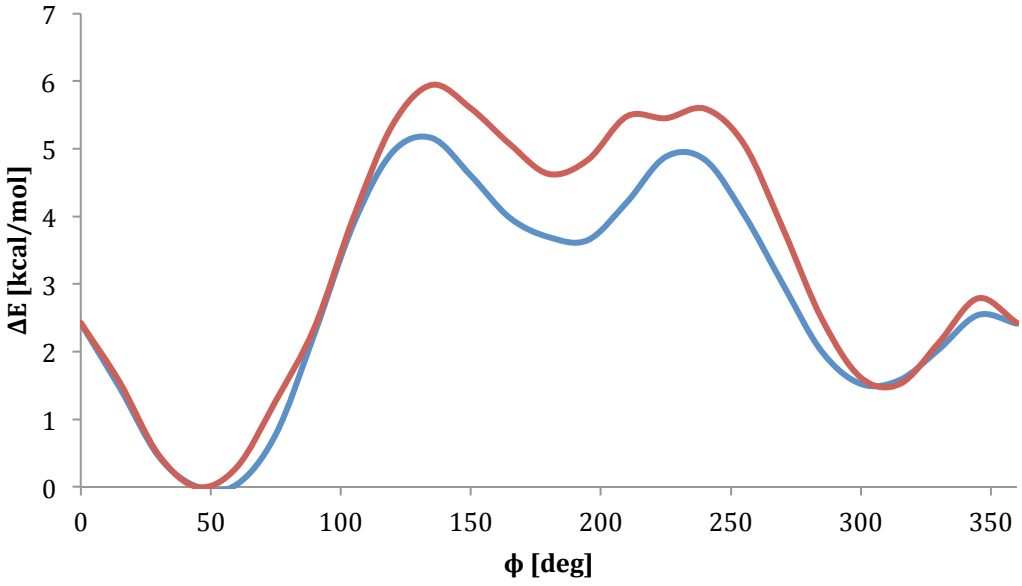


Figure 4.5 ϕ rotational energy curves for the two simplified models of Neu5Acα-2-3-Gal (See Figure 4.4): the red line is the analog of tetrahydropyran with carboxylic acid on the second position (Figure 4.4 (a)); the blue line is the analog of tetrahydropyran with carboxylate (Figure 4.4 (b)).

The energy difference of ϕ torsion angle in Neu5Acα-2-3-Gal is around 1.5kcal/mol between the two main conformations (ϕ = 60° and ϕ = 300°) according to QM calculations, which, however, would be equivalent to a ratio of 93 : 7, assuming a Boltzmann distribution. This ratio is not consistent with the crystallographic distribution, probably due to the incomplete treatment of water and counter ions in the quantum calculations. Therefore, an alternative method was employed to generate the energy curves. By assuming a Boltzmann distribution (See Equation 4.1), the populations of structures observed in the X-ray studies were employed to estimate the energy difference between the states near ϕ = 60° and ϕ = 300°.

A Maxwell-Boltzmann distribution is a probability distribution or frequency distribution of number of particles found in a given single-particle microstate. The distribution can be expressed in the form (See Equation 4.1):

$$\frac{N_i}{N} = \frac{\exp\left(-\frac{E_i}{kT}\right)}{\sum_j \exp\left(-\frac{E_j}{kT}\right)} \tag{4.1}$$

where i and j are indices of the single-particle micro states, N_i is the number of particles in the single-particle microstate i, N is the total number of particles in the system, E_i is the energy of microstate i, k is the Boltzmann constant. T is the thermodynamic temperature. The ratio of a Boltzmann distribution computed for two states is known as Boltzmann factor. If both i and j are indices of single-particle state (no degeneracy), the ratio only depends on the states' energy difference (See Equation 4.2):

$$\frac{N_i}{N_j} = \frac{\exp\left(-\frac{E_i}{kT}\right)}{\exp\left(-\frac{E_j}{kT}\right)} = \exp\left(\frac{E_j - E_i}{kT}\right) \tag{4.2}$$

Derived from the X-ray database of glycans containing Neu5Acα-2-3-Gal (See Figure 4.2), the ϕ-angle was binned into three intervals of 120° for all of the crystal structures (See Table 4.1).

Table 4.1 Number of crystal structures of glycans containing Neu5Acα-2-3-Gal falling into respective intervals of ϕ torsion angle

ϕ [deg]	**Number of crystal structures**
0~120	134
120~240	5
240~360	70

The median of each interval is 60°, 180° and 300°. By using the Boltzmann Factor, the energy difference was computed for each state. For example:

$$\frac{N_{60}}{N_{300}} = \frac{134}{70} = e^{\frac{\Delta E}{0.593}} \rightarrow \Delta E = 0.385 kcal/mol \quad (4.3)$$

In this analysis there is a small energy gap (0.385 kcal/mol) between the 60° and 300° states, and a larger one (1.95 kcal/mol) at 180°. In order to generate a penalty function around each of the states, a quadratic function was adopted as shown in Equation 4.4:

$$E_{SIA23}(x) = k(x-a)^2 + b \quad (4.4)$$

where *a* is the median of each interval *b* is the relative energy difference between each median and *k* is a constant that determines the width of the parabola and *x* is the ϕ torsion angle.
The resulting CHI energy function for Neu5Acα-2-3-Gal is as follows (See Equation 4.5):

$$\Delta E_{SIA23}(x) = \begin{cases} 0.0018(x-60)^2, & if\ x \in [0,120] \\ 0.0018(x-180)^2 + 1.95, & if\ x \in [120,240] \\ 0.0018(x-300)^2 + 0.385, & if\ x \in [120,240] \end{cases} \quad (4.5)$$

The experimental distribution of ϕ torsion angles in sialylated glycans-protein crystal structures in the PDB is considered a metric for comparison with the predicted CHI energy [11] (See Figure 4.6). This comparison matches the general fact that majority proteins that recognize sialylated glycans tend to select low energy conformations of glycosidic linkage [11].

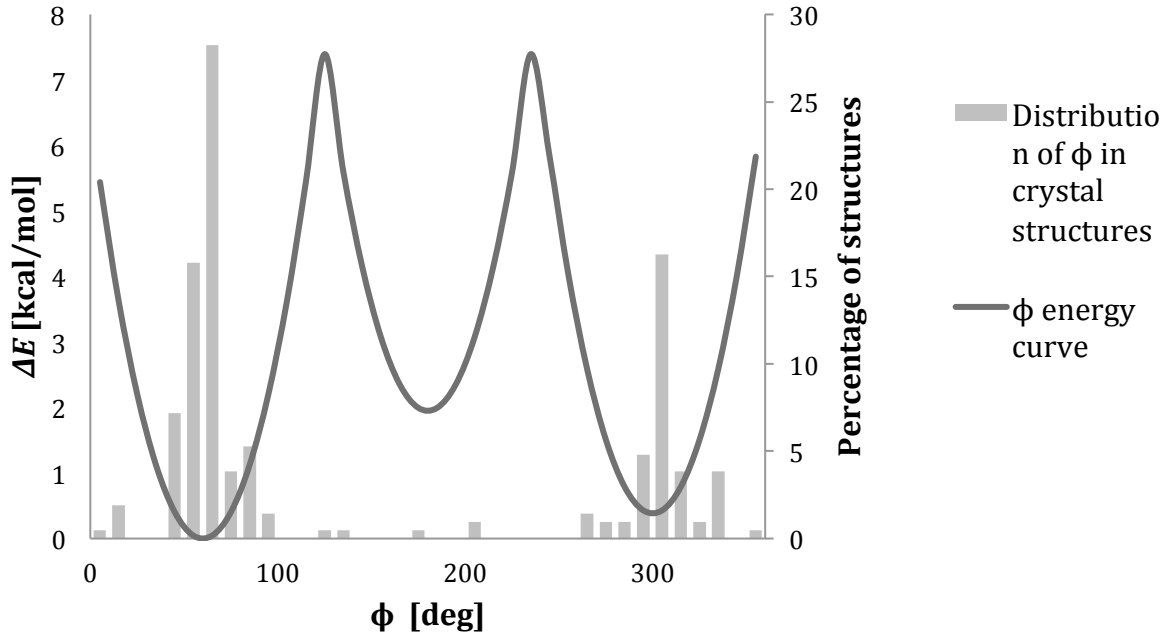


Figure 4.6 Comparison of the CHI energy function of Neu5Acα-2-3-Gal (solid line) to the ϕ distributions from experimental co-crystal structures (histogram)

Same protocol was used to generate the CHI energy function of glycans containing Neu5Acα-2-6-Gal (See Figure 4.7 and Table 4.2).

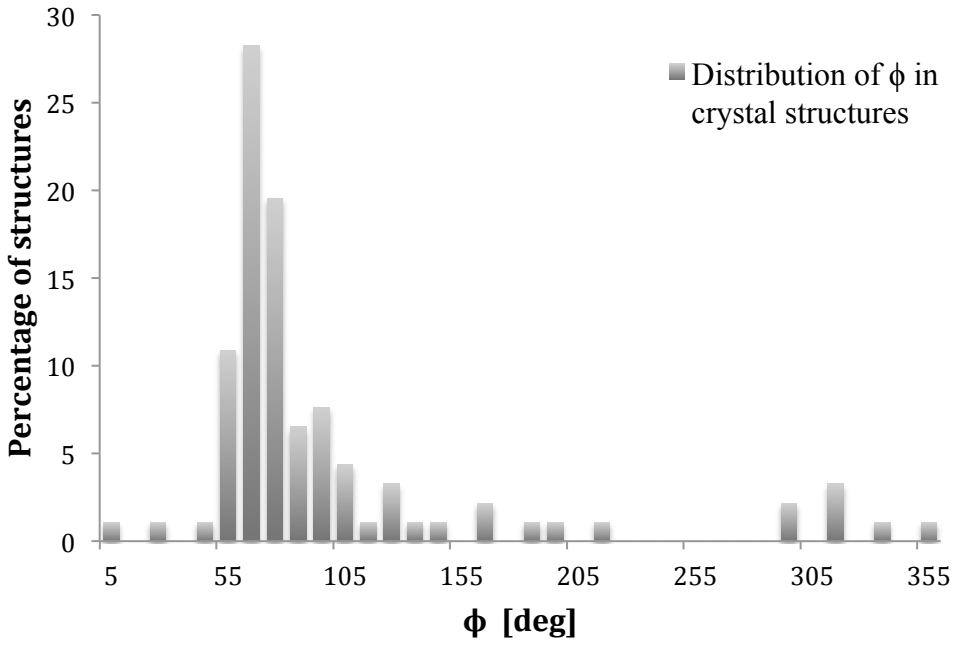


Figure 4.7 ϕ torsion angle distribution of Neu5Acα-2-6-Gal from X-ray data using the GlyTorsion tool at glycosience.de [23]

Table 4.2 Number of crystal structures of glycans containing Neu5Acα-2-6-Gal falling into respective intervals of ϕ torsion angle

ϕ [deg]	Number of crystal structures
0~120	75
120~240	10
240~360	7

The CHI energy function generated for glycans containing Neu5Acα-2-6-Gal is shown below (See equation 4.6).

$$\Delta E_{SIA26}(x) = \begin{cases} 0.0018(x-60)^2, & if\ x \in [0,120] \\ 0.0018(x-180)^2 + 1.19, & if\ x \in [120,240] \\ 0.0018(x-300)^2 + 1.41, & if\ x \in [120,240] \end{cases} \quad (4.6)$$

Energy curve for ϕ torsion angle in Neu5Acα-2-6-Gal is shown below (See Figure 4.8).

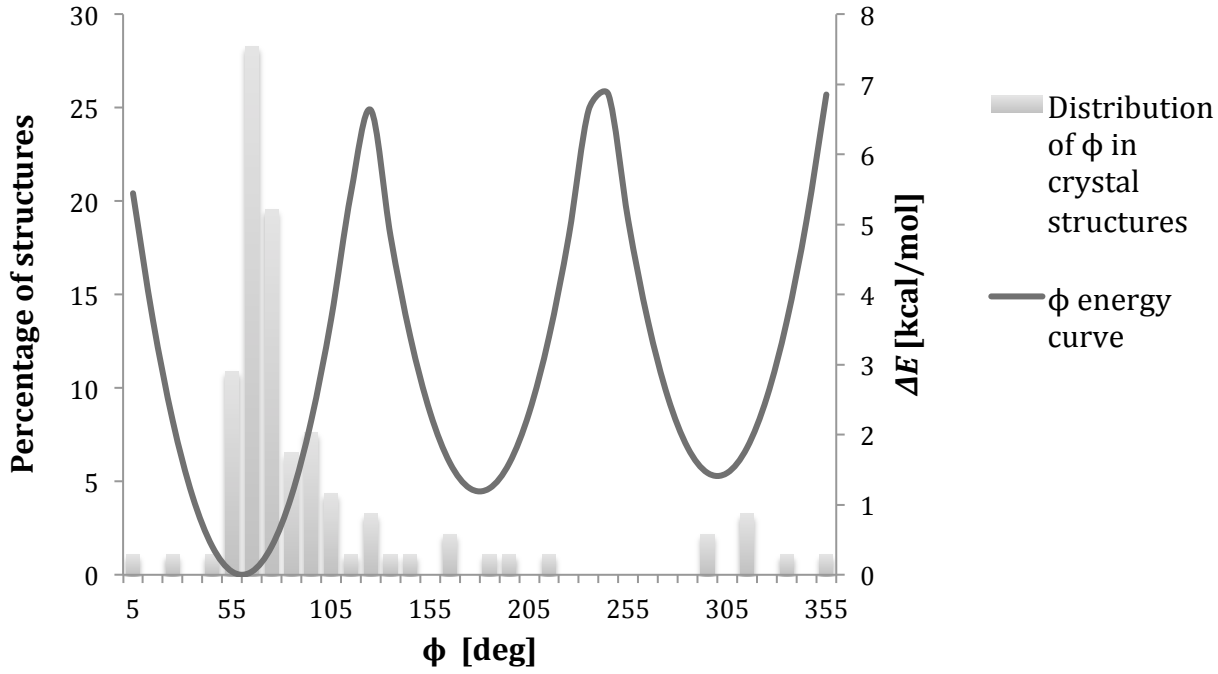
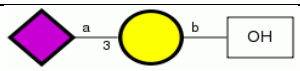
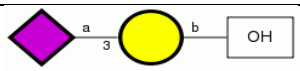
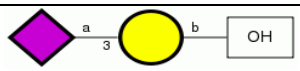
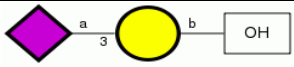
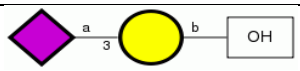
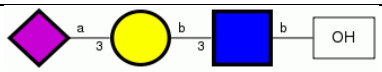
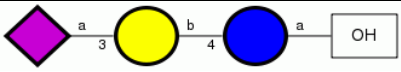
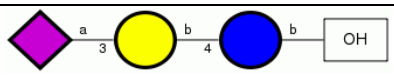
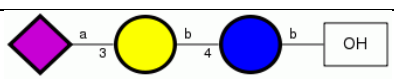


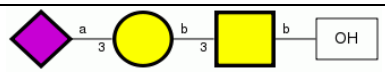
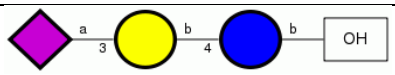
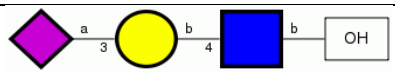
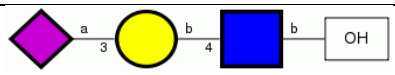
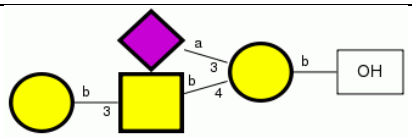
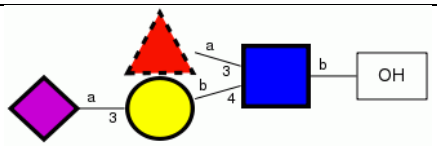
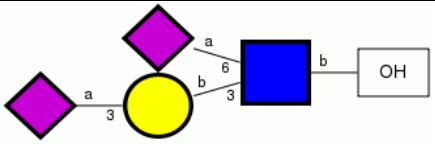
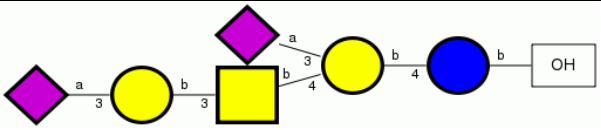
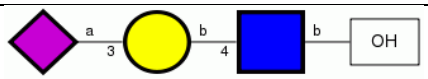
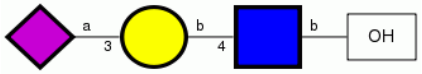
Figure 4.8 Comparison of the CHI energy function of Neu5Acα-2-6-Gal (solid line) to the ϕ distributions from experimental co-crystal structures (histogram)

4.3 Method: Systems Selection

Nineteen Neu5Acα-2-3-Gal-contained glycan-protein PDB systems and nine Neu5Acα-2-6-Gal-contained glycan-protein PDB systems were selected to test the performance of CHI energy function (See Table 4.3)

Table 4.3 (a) Neu5Acα-2-3-Gal-contained glycan-protein PDB systems and ligand sequences employed in the study. Graphic representation of ligands are generated by carbohydrate builder via glycam.org [25]

PDB systems	Ligand	Graphical Representation of ligand
1JSH	Neu5Acα-2-3-Galβ	
2JHD	Neu5Acα-2-3-Galβ	
2WBV	Neu5Acα-2-3-Galβ	
3UBQ	Neu5Acα-2-3-Galβ	
3ZPB	Neu5Acα-2-3-Galβ	
1JSN	Neu5Acα-2-3-Galβ-1-3-GlcNAcβ	
1S0I	Neu5Acα-2-3-Galβ-1-4-Glcα	
3AP7	Neu5Acα-2-3-Galβ-1-4-Glcβ	
3NV4	Neu5Acα-2-3-Galβ-1-4-Glcβ	

3RSJ	Neu5Acα-2-3-Galβ-1-3-GalNAcβ	
3S6X	Neu5Acα-2-3-Galβ-1-4-Glcβ	
4BH4	Neu5Acα-2-3-Galβ-1-4-GlcNAcβ	
4BH1	Neu5Acα-2-3-Galβ-1-4-GlcNAcβ	
2CHB	Galβ-1-3-GalNAcβ-1-4-[Neu5Acα-2-3]Galβ	
2Z8L	Neu5Acα-2-3-Galβ-1-4-[Fucα-1-3]GlcNAcβ	
1VPS	Neu5Acα-2-3-Galβ-1-3-[Neu5Acα-2-6]GlcNAcβ	
2VU9	Neu5Acα-2-3-Galβ-1-3-GalNAcβ-1-4-[Neu5Acα-2-3]Galβ-1-4Glcβ	
1RVX	Neu5Acα-2-3-Galβ-1-4-GlcNAcβ	
3VKO	Neu5Acα-2-3-Galβ-1-4-GlcNAcβ	

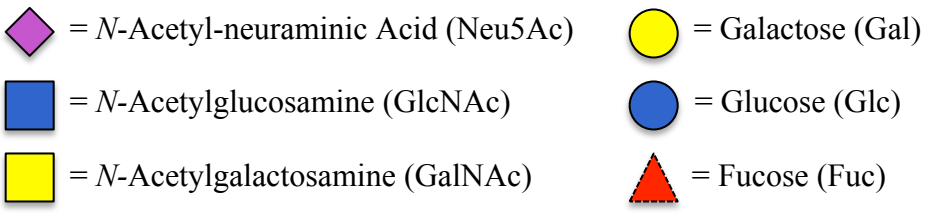
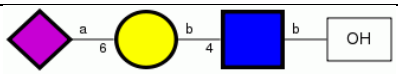
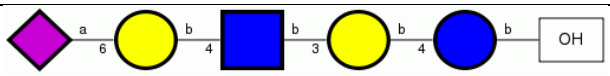
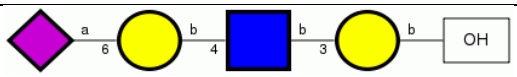
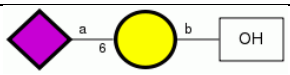
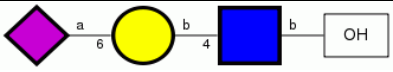
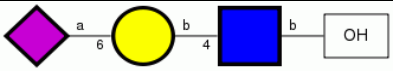
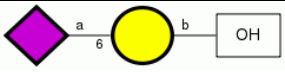
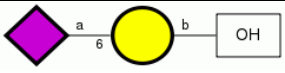
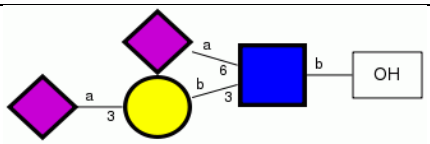


Table 4.3 (b) Neu5Acα-2-6-Gal-contained glycan-protein PDB systems and ligand sequences employed in the study. Graphic representation of ligands are generated by carbohydrate builder via glycam.org [25]

PDB systems	ligand	Graphic Representation of ligand
2YP3	Neu5Acα-2-6-Galβ-1-4-GlcNAcβ	
2YP4	Neu5Acα-2-6-Galβ-1-4-GlcNAcβ-1-3- Galβ-1-4-Glcβ	
3UBE	Neu5Acα-2-6-Galβ-1-4-GlcNAcβ-1-3-Galβ	
4BGX	Neu5Acα-2-6-Galβ	
4BH0	Neu5Acα-2-6-Galβ-1-4-GlcNAcβ	
4BH3	Neu5Acα-2-6-Galβ-1-4-GlcNAcβ	
4BSH	Neu5Acα-2-6-Galβ	
4LKK	Neu5Acα-2-6-Galβ	
1VPS	Neu5Acα-2-3-Galβ-1-3-[Neu5Acα-2-6]GlcNAcβ	

= *N*-Acetyl-neuraminic Acid (Neu5Ac)

= Galactose (Gal)

= *N*-Acetylglucosamine (GlcNAc)

= Glucose (Glc)

= *N*-Acetylgalactosamine (GalNAc)

= Fucose (Fuc)

4.4 Method: Pose Prediction Accuracy

Given a set of predicted oligosaccharides, accuracy is measured by root mean squared deviation (**RMSD**), including Pose RMSD **(PRMSD)** and Shape RMSD **(SRMSD)**, compared to known structures. PRMSD values are obtained by calculating the RMSD between the ring atoms of the docked ligand relative to the position of the same atoms in the co-crystal [26]. SRMSD values are obtained by calculating the RMSD between the respective ring atoms of the docked and crystal ligand after first superimposing them [26] (See Figure 4.9).

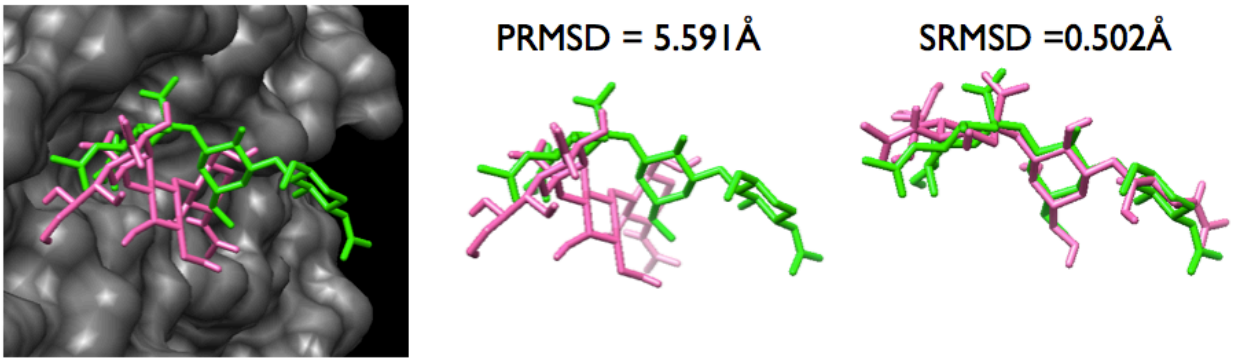


Figure 4.9 PRMSD and SRMSD calculation. Green is the crystal ligand and Pink is the docked ligand

SRMSD provides an indication of the correctness of the 3D shape of the ligand, but says nothing about its orientation in the binding site. PRMSD is more frequently used in measuring the accuracy of pose prediction. The structures with PRMSD below or equal to 2.0 Å are considered acceptable [11].

4.5 Method: Incorporation of CHI Energy Functions in AutoDock Vina

Ten independent docking runs were performed for each selected systems (See Table 4.3) with (Vina-Carb) and without (Vina) the incorporation of CHI energy function for sialylated glycans in AutoDock Vina separately. Each test generated 20 ligand poses and ranked them according to the calculated binding affinity. The PRMSD and SRMSD for each pose in each run were calculated using Chimera for Vina and Vina-Carb [22]. The averaged PRMSDs and SRMSDs of the first acceptable pose (or, if no acceptable pose was generated, the pose with the lowest PRMSD) were calculated for each selected sialylated glycan-protein systems.

4.6 Results Analysis and Assessment

In fifteen of the nineteen Neu5Acα-2-3-Gal-contained cases, the averaged PRMSD decreased (as well as SRMSD) when the CHI-energy functions were presented (See Figure 4.10). Similarly, in eight of the nine Neu5Acα-2-6-Gal-contained glycan-protein cases, better poses were obtained using Vina-Carb (See Figure 4.11).

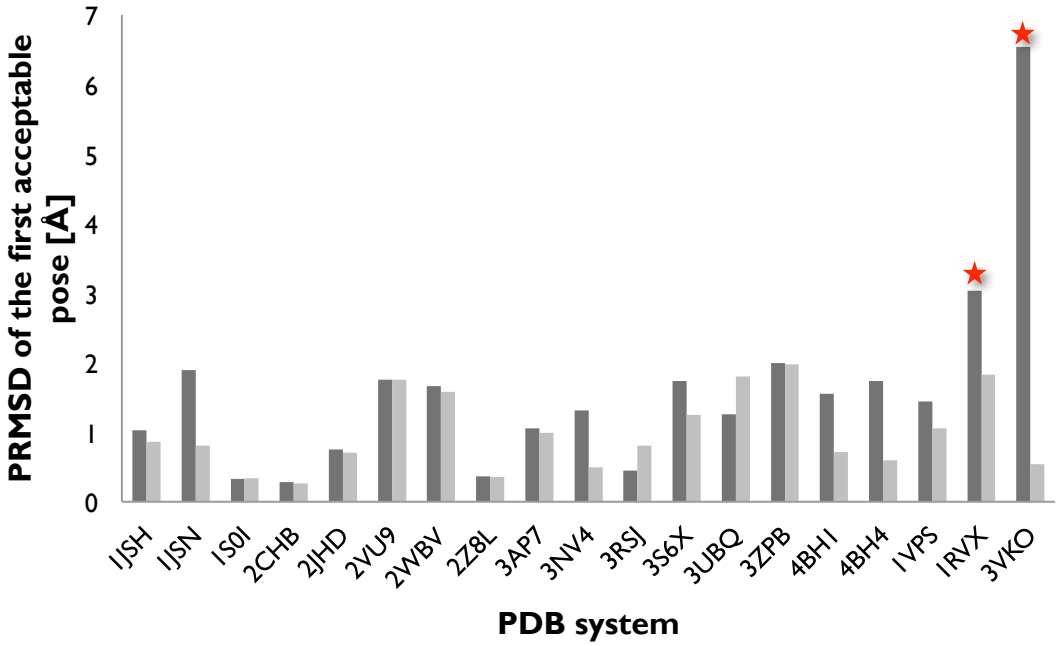


Figure 4.10 (a) averaged PRMSDs of the first acceptable pose for Neu5Acα-2-3-Gal-contained glycan-protein systems from AutoDock Vina, (dark gray) and Vina-Carb (light gray). The★ indicates that no acceptable pose was generated, and so the minimum PRMSD is reported.

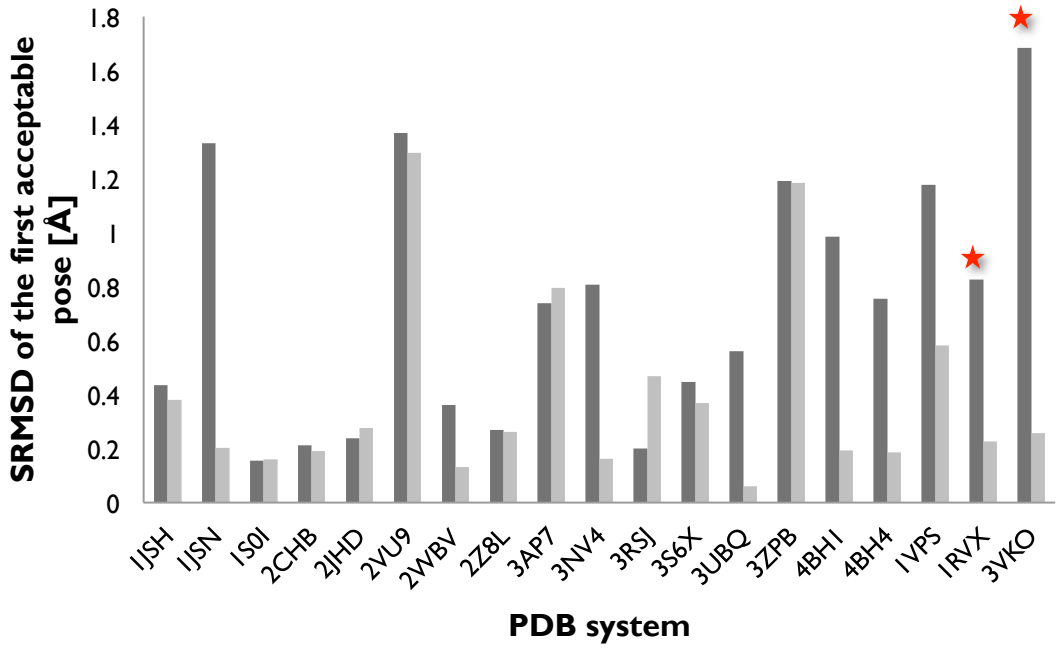


Figure 4.10 (b) averaged SRMSDs of the first acceptable pose for Neu5Acα-2-3-Gal-contained glycan-protein systems from AutoDock Vina, (dark gray) and Vina-Carb (light gray). The ★ indicates that no acceptable pose was generated, and so the SRMSD is reported for the minimum PRMSD pose.

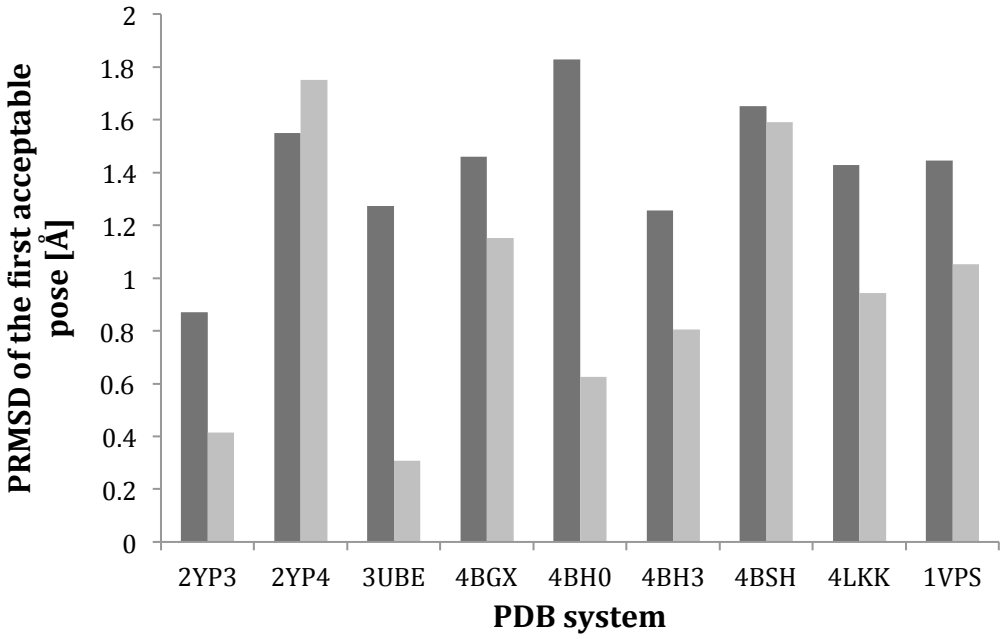


Figure 4.11 (a) averaged PRMSDs of the first acceptable pose for Neu5Acα-2-6-Gal-contained glycan-protein systems from AutoDock Vina, (dark gray) and Vina-Carb (light gray)

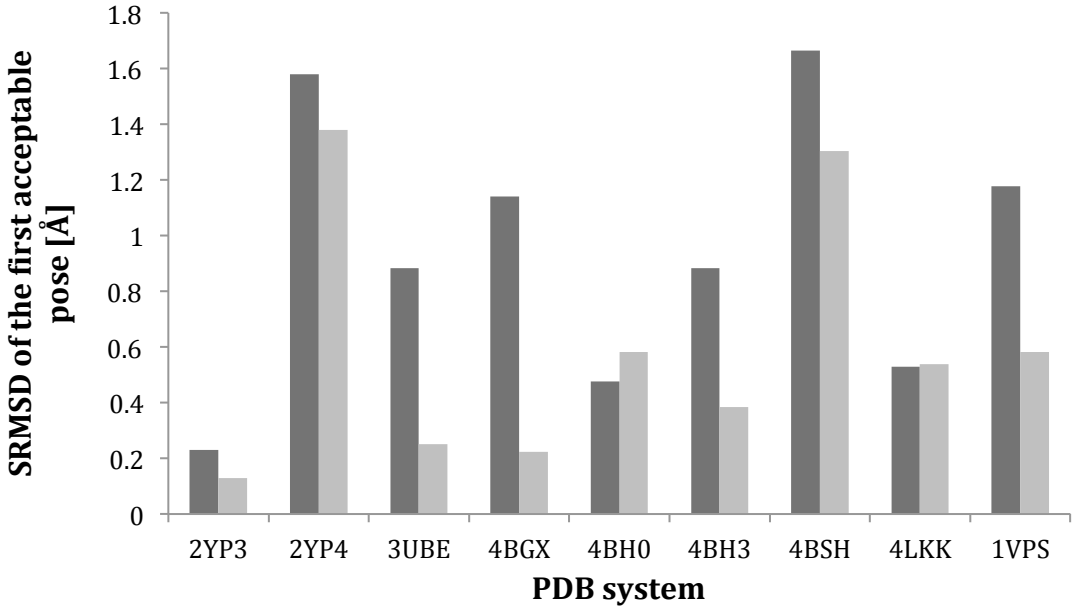


Figure 4.11 (b) averaged SRMSDs of the first acceptable pose for Neu5Acα-2-6-Gal-contained glycan-protein systems from AutoDock Vina, (dark gray) and Vina-Carb (light gray)

An example of system where inclusion of the CHI-energy functions led to an improvement in the PRMSD is presented in Figure 4.12. This was the general outcome, however, in two cases (3RSJ and 3UBQ) Vina resulted in a slightly lower PRMSD than Vina-Carb. The cases of 3UBQ and 3RSJ are illustrated in Figure 4.13 and Figure 4.14 separately.

Vina was unable to generate an acceptable pose (PRMSD ≤ 2.0 Å) for the trisaccharide ligand in 3VKO, whereas the first acceptable pose generated by Vina-Carb had an averaged PRMSD = 0.542 Å.

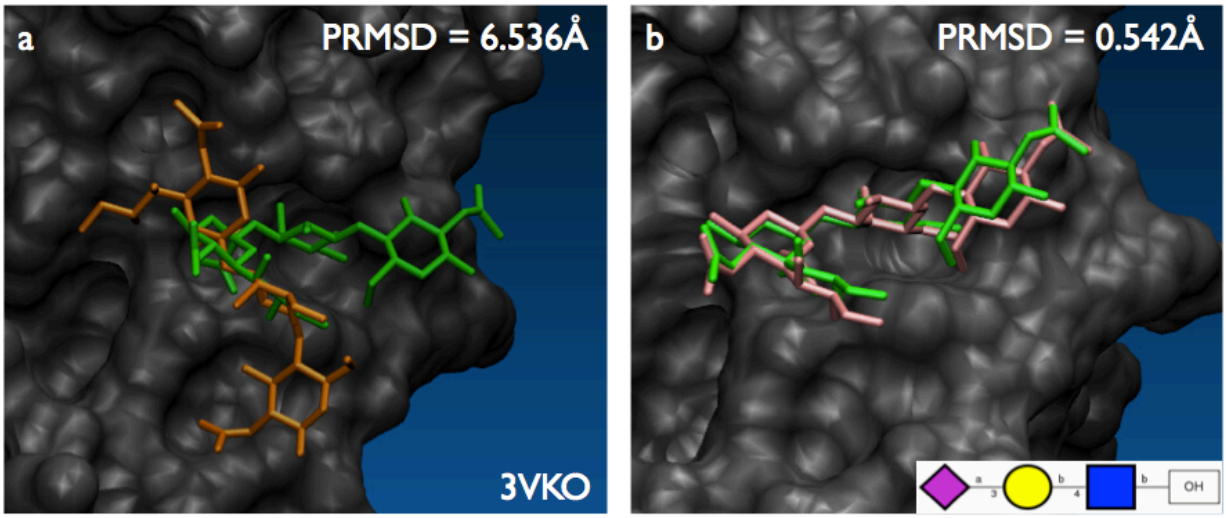


Figure 4.12 (a) AutoDock Vina lowest PRMSD pose for 3VKO from Vina (orange) compared to the crystal ligand (green). (b) First acceptable pose (which was also the lowest PRMSD pose) from Vina-Carb (pink) compared to the crystal ligand (green)

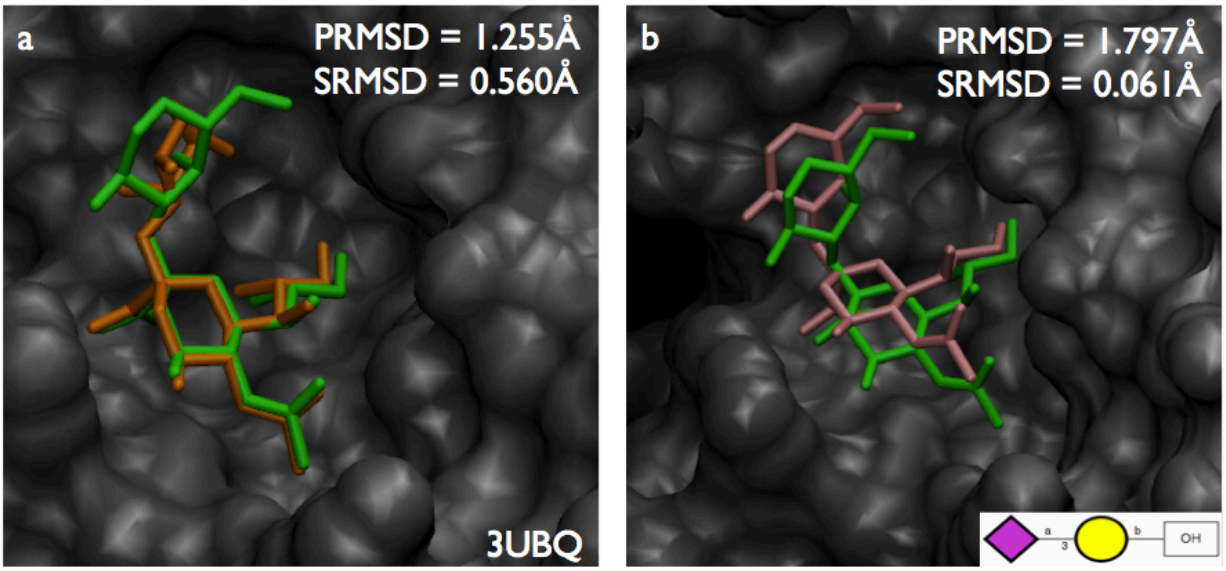


Figure 4.13 (a) AutoDock Vina first acceptable pose for 3UBQ from Vina (orange) compared to the crystal ligand (green). (b) First acceptable pose from Vina-Carb (pink) compared to the crystal ligand (green)

As presented in Figure 4.13, inclusion of CHI energy terms in AutoDock Vina resulted in a slightly increased PRMSD because the glycosidic torsion angles of docked ligand in Vina are distorted (ϕ = 352.15°, ψ = 211.72°), which means conformational unfavorite. Therefore, in Vina-Carb, similar docked structure cannot be found due to the incorporation of CHI energy functions. However, though better poses haven't been found in Vina-Carb, the averaged PRMSD is still within the acceptable range (PRMSD $\leq$ 2Å). What's more, improved shape has been found in Vina-Carb, suggesting the accurate prediction of glycosidic torsion angles (crystal: ϕ = 290.85°, ψ = 225.22°; docked: ϕ = 293.84°, ψ = 219.32°).

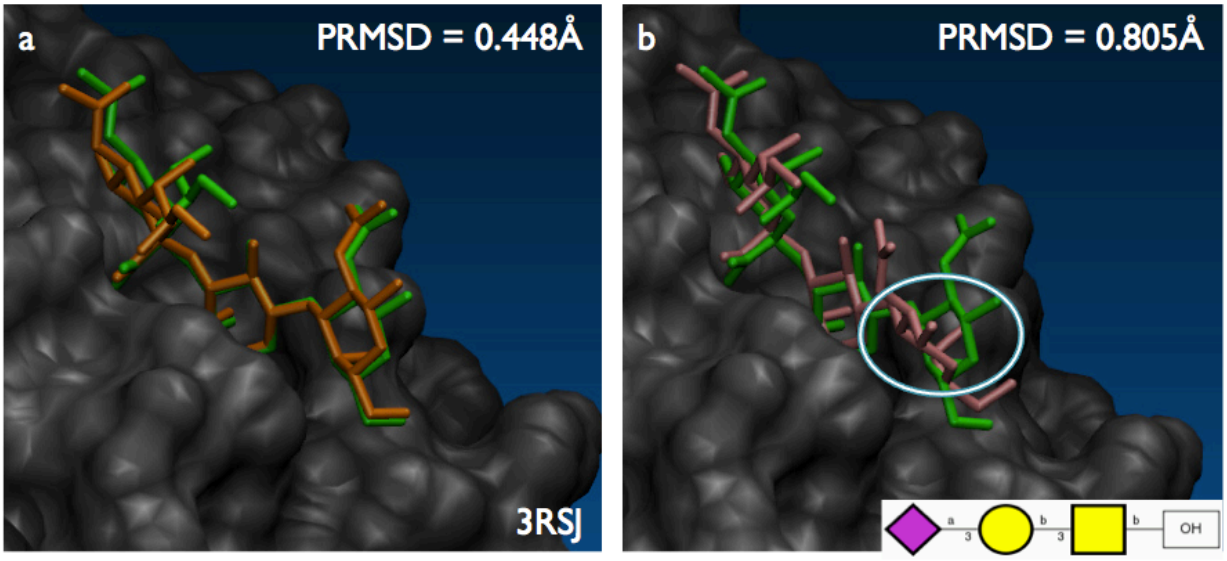


Figure 4.14 (a) AutoDock Vina first acceptable pose for 3RSJ from Vina (orange) compared to the crystal ligand (green). (b) First acceptable pose from Vina-Carb (pink) compared to the crystal ligand (green)

As illustrated in Figure 4.14, Vina-Carb led to a slightly increased PRMSD than Vina, mainly due to the conformational difference of glycosidic torsion angles (white circle in Figure 4.13 (b)) of docked (ϕ = 292.06°, ψ = 227.34°) and crystal (ϕ = 269.60°, ψ = 200.03°) ligands. This circled glycosidic linkage is a β-1-3eqatorial linkage, for which the conformational energy (crystal: E_ϕ = 1.05 kcal/mol, $E_\psi \approx 0$ kcal/mol; docked: $E_\phi \approx 0$ kcal/mol, $E_\psi \approx 0$ kcal/mol) was quantified by the previous developed CHI-energy functions. In Vina-Carb, the previous developed CHI-energy functions have a default of CHI-energy-cutoff = 2.0 kcal/mol, which means that each glycosidic torsion angle that possesses an energy less than or equal to 2.0 kcal/mol will be considered as 0 kcal/mol. However, in order to test the newly developed CHI-energy functions for Neu5Acα-2-3Gal and Neu5Acα-2-6Gal, no CHI-energy-cutoff was employed, resulting in a higher penalty to linkages that deviate from the theoretical minimum values. This particularly affects the

estimation of E_{ϕ}. By employing a CHI-energy-cutoff specific to each linkage type, this issue can be fixed.

The main role of CHI energy functions is to quantify the conformational energy of carbohydrates. As illustrated in Figure 4.15, Vina didn't show any conformational preference for ligands while Vina-Carb led to concentrated distributions, which match experimental co-crystal structures.

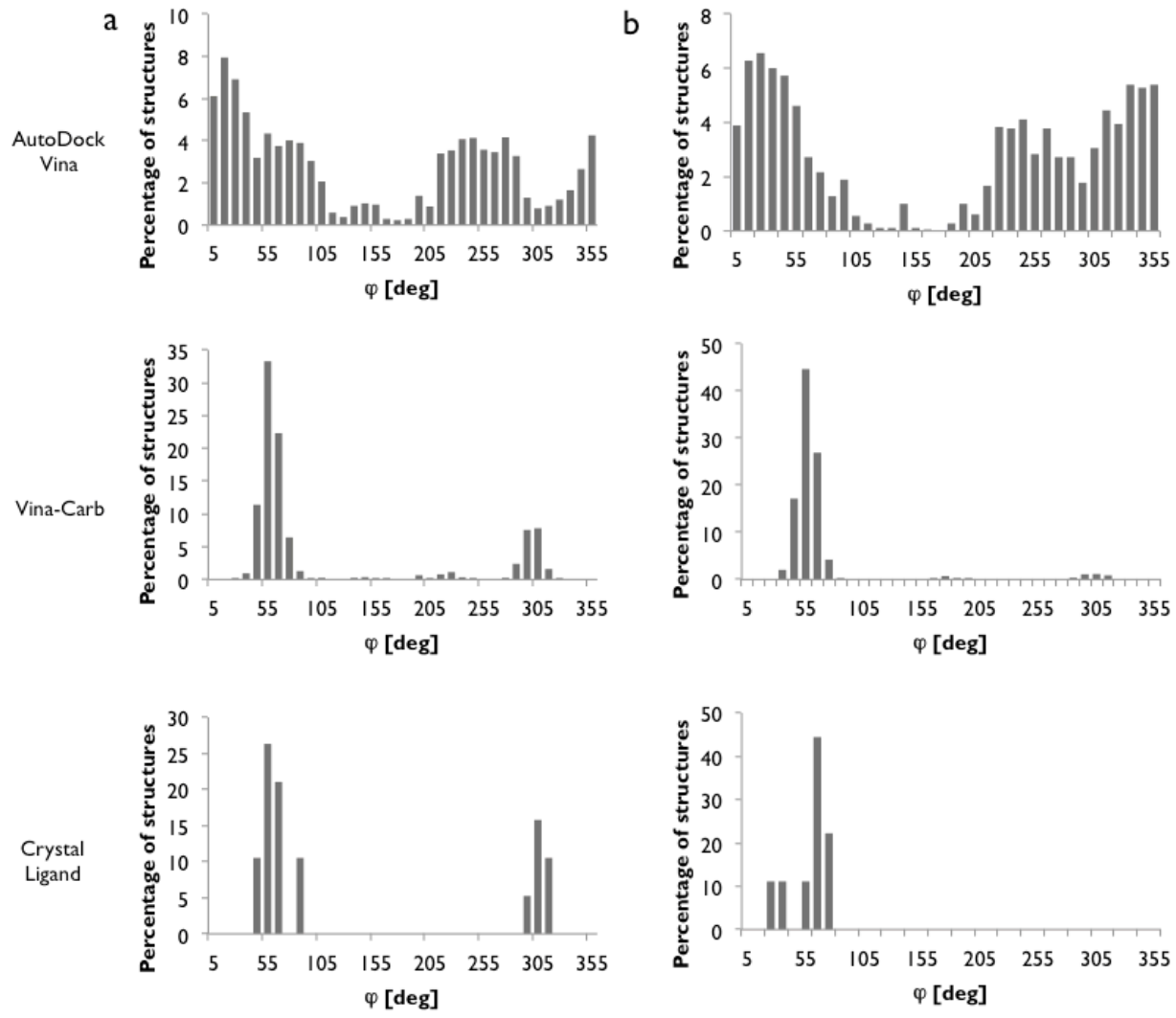


Figure 4.15. (a) ϕ torsion angle distribution of Neu5Acα-2-3-Gal in AutoDock Vina, Vina-Carb and crystal structures; (b) ϕ torsion angle distribution of Neu5Acα-2-6-Gal in AutoDock Vina, Vina-Carb and crystal structures

CHAPTER 5

CONCLUSION AND RECOGNIZED CHANLLENGES

5.1 Conclusion

Intrinsic energy terms (CHI energy functions) for sialylated glycans, which quantify the relative conformational energy of the most common linkages of sialic acids (Neu5Acα-2-3Gal and Neu5Acα-2-6Gal), have been developed. These terms help to address one of the challenges encountered in flexible carbohydrate docking. Better poses have been predicted (PRMSD decreased) via Vina-Carb in fifteen of the nineteen (79%) cases of Neu5Acα-2-3-Gal-contained glycan-protein systems, and in eight of the nine (89%) cases of Neu5Acα-2-6-Gal-systems. In two (1RVX and 3VKO) of the fifteen cases of Neu5Acα-2-3Gal-contained glycan-protein systems, no acceptable poses (PRMSDs > 2.0 Å) were generated by Vina, however, including the CHI-energies led to acceptable poses (PRMSDs ≤ 2.0 Å) being generated using Vina-Carb for both of the systems. In conclusion, including the CHI energy terms in AutoDock Vina (Vina-Carb) enabled the selection of poses based on ligand conformations; increasing the chances of finding the correct pose among all output poses generated.

5.2 Recognized Challenges

Apart from the CHI-energy-cutoff mentioned in chapter 4.6, there are two recognized challenges in optimizing the CHI energy functions of sialic acid linkages (Neu5Acα-2-3-Gal and Neu5Acα-2-6-Gal). One of the challenges is that the sample size of the crystal structures is

limited. There are only 209 Neu5Acα-2-3-Gal-contained glycan-protein systems and 92 Neu5Acα-2-6-Gal-contained glycan-protein systems in the PDB database, which cannot be considered as a large database to derive energy curve empirically. Especially for the intervals containing less than 10 crystal structures (i.e. interval that $\phi \in [120,240]$ in Neu5Acα-2-3-Gal-contained glycan-protein systems. Table 4.1), a small number change could result in a comparative large change in calculating relative bottom energy, and additionally, not all the crystal structures are accurate. Another challenge worth discussing is the broadness of each parabolic curve. In building the energy curve, a constant k = 0.018 (See Equation 3.5) was used for three intervals in both systems. However, take Figure 4.6 for example, the parabolic shape should probably be broader over the intervals with $\phi \in [0,120]$ and $\phi \in [240, 360]$ than that with $\phi \in [120,240]$ because they are more structural preferred.

The first limitation cannot be fixed, but more and more co-crystal structures of sialylated glycan-proteins will be discovered in the future. Accuracy of prediction can be improved by recalculating and updating the CHI energy functions. The second limitation can be fixed by adjusting the constant k. Instead of using a same constant for all the intervals, different k_1, k_2 and k_3 can be employed in different intervals to adjust the broadness of each parabola.

REFERENCES

[1] Ghazarian, H., B. Idoni and S. B. Oppenheimer (2011). "A Glycobiology Review: Carbohydrates, Lectins and Implications in Cancer Therapeutics." *Acta Histochem.* **113**(3): 236-247.

[2] "IUPAC Gold Book - Glycans"

[3] French, A. D. and J. W. Brady (1990). *Computer Modeling of Carbohydrate Molecules*. Washington, D.C., American Chemical Society.

[4] Acquotti, D., L. Cantu, E. Ragg and S. Sonnino (1994). "Geometrical and Conformational Properties of Ganglioside GalNAc-G_{D1a}, IV^{4}GalNAcIV^{3}Neu5AcII^{3}Neu5AcGgOse$_4$Cer." *Eur. J. Biochem.* **225**(1): 271-288.

[5] Haltiwanger, R. S. and J. B. Lowe (2004). "Role of glycosylation in development." *Annu Rev Biochem* **73**: 491-537.

[6] Jorgensen WL (1991). "Rusting of the lock and key model for protein-ligand binding." *Science* **254** (5034): 954–5.

[7] Wei BQ, Weaver LH, Ferrari AM, Matthews BW, Shoichet BK (2004). "Testing a flexible-receptor docking algorithm in a model binding site". *J. Mol. Biol.* **337** (5): 1161–82.

[8] Alonso, H., A. A. Bliznyuk and J. E. Gready (2006). "Combining Docking and Molecular Dynamic Simulations in Drug Design." *Med. Res. Rev.* **26**(5): 531-568.

[9] Trott, O. and A. J. Olson (2010). "AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading." *J. Comput. Chem.* **31**(2): 455-461.

[10] http://autodock.scripps.edu/wiki/Introduction

[11] Nivedha, A. K., S. Makeneni, B. L. Foley, M. B. Tessier and R. J. Woods (2013). "Importance of ligand conformational energies in carbohydrate docking: Sorting the wheat from the chaff." *J Comput Chem*.

[12] Alvarado, E., T. Nukada, T. Ogawa and C. E. Ballou (1991). "Conformation of the Glucotriose Unit in the Lipid-Linked Oligosaccharide Precursor for Protein Glycosylation." *Biochem.* **30**(4): 881-886

[13] Schauer, R. (1985). "Sialic Acids and Their Role as Biological Masks." *TIBS* **10**(9): 357-360.

[14] Crocker, P. R., J. C. Paulson and A. Varki (2007). "Siglecs and their roles in the immune system." *Nature Reviews Immunology* **7**: 255-266.

[15] Ferraris, O. and B. Lina (2008). "Mutations of neuraminidase implicated in neuraminidase inhibitors resistance." *Journal of Clinical Virology* **41**(1): 13-19.

[16] Chavas, L. M. G., C. Tringali, P. Fusi, B. Venerando, G. Tettamanti, R. Kato, E. Monti and S. Wakatsuki (2005). "Crystal Structure of the Human Cytosolic Sialidase Neu2: EVIDENCE FOR THE DYNAMIC NATURE OF SUBSTRATE RECOGNITION." *Journal of Biological Chemistry* **280**(1): 469-475.

[17] Matsuda, K., A. Niitsuma, M. K. Uchida and T. Suzuki-Nishimura (1994). "Inhibitory effects of sialic acid- or N-acetylglucosamine-specific lectins on histamine release induced by

compound 48/80, bradykinin and a polyethylenimine in rat peritoneal mast cells." *Jpn J Pharmacol* **64**(1): 1-8.

[18] Suzuki, Y., T. Ito, T. Suzuki, R. E. Hollland Jr., T. M. Chambers, M. Kiso, H. Ishida and Y. Kawaoka (2000). "Sialic Acid Species as a Determinant of the Host Range of Influenza A Viruses." *Journal of Virology* **74**(24): 11825–11831.

[19] Guo, H. B., A. Nairn, K. Harris, M. Randolph, G. Alvarez-Manilla, K. Moremen and M. Pierce (2008). "Loss of expression of N-acetylglucosaminyltransferase Va results in altered gene expression of glycosyltransferases and galectins." *FEBS Lett* **582**(4): 527-535.

[20] Ning, Z. Y., X. T. Wu, Y. F. Cheng, W. B. Qi, Y. F. An, H. Wang, G. H. Zhang and S. J. Li (2012). "Tissue distribution of sialic acid-linked influenza virus receptors in beagle dogs." *J Vet Sci* **13**(3): 219-222.

[21] Connor, R. J., Y. Kawaoka, R. G. Webster and J. C. Paulson (1994). "Receptor Specificity in Human, Avian, and Equine H2 and H3 Influenza Virus Isolates." Virology **205**(1): 17-23.

[22] Nivedha, A. K., Thieker F. David, Woods, R. J. (2015). "Vina-Carb: Improving Glycosidic Angles During Carbohydrate Docking." *J. Chem. Theory Comput.* **(Accepted)**.

Schauer, R. (1985). "Sialic Acids and Their Role as Biological Masks." *TIBS* **10**(9): 357-360.

[23] http://glycosciences.de/tools/glytorsion/

[24] Frisch, M., G. Trucks, H. B. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, B. Mennucci and G. Petersson (2009). "Gaussian 09, Revision A. 02, Gaussian." *Inc., Wallingford, CT* **200**.

[25] http://glycam.org/tools/molecular-dynamics/oligosaccharide-builder

[26] Borrelli, K. W., B. Cossins and V. Guallar (2009). "Exploring Hierarchical Refinement Techniques for Induced Fit Docking with Protein and Ligand Flexibility." *Journal of Computational Chemistry* **31**(6): 1224-1235.