DEVELOPMENT OF COMPUTATIONAL METHODS TO CHARACTERIZE CARBOHYDRATE-PROTEIN INTERACTIONS

by

AMIKA SOOD

(Under the Direction of Robert J. Woods)

ABSTRACT

Specific carbohydrate-protein interactions are crucial in numerous physiological processes, disruption of which has been implicated in many different diseases like cancer. This provides researchers an opportunity to utilize carbohydrates as biomarkers and targets for therapeutics for such diseases. There has been a tremendous surge in the research being conducted towards the development of techniques to analyze carbohydrates and their specificity and affinity for different proteins. However, owing to their complex three-dimensional structure, stereochemistry, low binding affinities and broad specificity, carbohydrates have proven to be challenging to study. Therefore, new techniques and improvements in the existing methodologies are required. Here, we show that the incorporation of experimental data into molecular modeling can be used as a powerful combination to gain an understanding of the structural features of proteins and carbohydrates leading to the specificity in their interactions. Firstly, hydroxyl radical protein footprinting (HRPF) was used to establish a relationship between the oxidation of amino acids exposed on the surface of a protein and their solvent accessible surface area (SASA). Oxidation, as well as SASA, are both directly proportional to the exposure of an amino acid to the solvent. This relationship was used to estimate SASA of residues of a protein in solution, which was then successfully utilized as a score to quantify the quality of models

generated through a molecular dynamics (MD) simulation and homology modeling. This relationship can also be used to study protein- carbohydrate interactions, which remains to be tested. Secondly, the functional groups of a monosaccharide essential for forming protein-carbohydrate interactions were identified by using co-crystal structures and per-atom binding energy analysis, which shows that not all chemically-equivalent functional groups are equally significant for binding. Lastly, the 3D structure of a group of monosaccharides was analyzed and it was observed that two monosaccharides can possess structural similarities depending on their alignment, which can be used to explain cross-reactivity between a protein and more than one carbohydrate.

INDEX WORDS: Protein-Carbohydrate interactions, GLYCAM, Molecular dynamics, Monte Carlo, Solvent accessibility surface area, MM-GBSA, MM-PBSA, cross-reactivity

# DEVELOPMENT OF COMPUTATIONAL METHODS TO CHARACTERIZE CARBOHYDRATE-PROTEIN INTERACTIONS

by

AMIKA SOOD

B.Tech, Vellore Institute of Technology, India, 2008

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

DEVELOPMENT OF COMPUTATIONAL METHODS TO CHARACTERIZE CARBOHYDRATE-PROTEIN INTERACTIONS

by

AMIKA SOOD

| | |
|---|---|
| Major Professor: | Robert J. Woods |
| Committee: | Natarajan Kannan |
| | David P. Landau |
| | Ying Xu |

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2016

This work is dedicated to my parents, for their unconditional love and their constant support and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

# CHAPTER 1

# INTRODUCTION

Proteins, polymers of amino acids and carbohydrates with monosaccharides as their basic unit comprise two of the most abundant biological molecules. Complex carbohydrates coat the surfaces of living cells and play an important role in a vast range of biological processes through glycan-binding proteins (GBPs) that recognize them as ligands. The research in this dissertation is focused on understanding and addressing the complexities of modeling protein-carbohydrate interactions. This subject was tackled from the aspects of both the protein and carbohydrate. Therefore, the work done here comprises of:

1. Development of a methodology to estimate solvent accessibility of residues in a protein, to quantify the quality of 3D models in the absence of experimental 3D structures.
2. Identify the functional groups in a carbohydrate involved in interactions with proteins.
3. Locate similarities in carbohydrate 3D structures to understand and predict their cross-reactivity.

These topics, including the review of their respective backgrounds and the methods applied to them, are presented as follows:

CHAPTER 2: PROTEIN-CARBOHYDRATE INTERACTIONS

Carbohydrates are ubiquitously expressed biomolecules, and their interactions with proteins are essential for cellular function. This chapter describes the structural complexities associated with carbohydrates and their significance.

CHAPTER 3: COMPUTATIONAL METHODS USED TO STUDY BIOMOLECULES

There are several computational and experimental methods that are used to study the structure and dynamics of biomolecules. Some of these methods, relevant to the research presented here, are discussed.

CHAPTER 4: MONITORING LARGE-SCALE PROTEIN CONFORMATIONAL CHANGES

This study highlights the potential of combining experimental HRPF analysis and SASA estimation in understanding protein flexibility and compare 3D structures. The results of this study were published in a peer-reviewed journal:

Poor TA, Jones LM, Sood A, Leser GP, Plasencia MD, Rempel DL et al. Probing the paramyxovirus fusion (F) protein-refolding event from pre- to postfusion by oxidative footprinting. Proceedings of the National Academy of Sciences of the United States of America. 2014 Jun 24;111(25).

CHAPTER 5: INTEGRATING MASS SPECTROMETRY FOOTPRINTING DATA IN PROTEIN STRUCTURE MODELING

This research employs SASA estimated from HRPF experiments as restraints in simulations of globular protein BPTI and to compare 3D models of a globular protein lysozyme, generated by molecular dynamics and homology modeling.

CHAPTER 6: QUANTIFYING FUNCTIONAL GROUP CONTRIBUTIONS TO UNDERSTANDING PROTEIN-CARBOHYDRATE AFFINITY

In this research, a carbohydrate binding protein called *Erythrina cristagalli* lectin (ECL) was used to identify the functional groups of its known ligands necessary for their affinity.

CHAPTER 7: MONOSACCHARIDE SIMILARITY ANALYSIS TO UNDERSTAND PROTEIN-CARBOHYDRATE SPECIFICITY

This study compares 3D features of monosaccharides to find similarities in their structures to help explain cross-reactivity.

CHAPTER 8: CONCLUSIONS AND FUTURE PROSPECTS

The major conclusions of the work are summarized, and possible future directions are provided.

# CHAPTER 2

## PROTEIN-CARBOHYDRATE INTERACTIONS

Carbohydrates are ubiquitous in nature as nearly all organisms synthesize and metabolize them, and they can also be referred to as sugars, oligo- or polysaccharides, or glycans. Glycans are commonly found covalently bound to other biomolecules such as proteins (glycoproteins) or lipids (glycolipids) on cell surfaces. Glycans that are covalently attached to a protein can alter its structure and function (1), by either preventing their interaction with the environment through blocking regions of the protein surface (2) or by hindering their dynamics because of their large mass (3). Even though they have a simple chemical formula, they can form complex 3-dimensional (3D) structures, because of their unique characteristics. The basic carbohydrate unit is called a monosaccharide and can exist in several ring forms, for example as furanose (five-membered ring structures) and pyranose (six-membered ring structures). Each pyranose has five positions available to form linkages with other monosaccharides, which allows carbohydrates to form branched structures (Figure 2.1). The exponential increase in the complexity of a disaccharide comprised of identical monosaccharide units can be emphasized by comparing it to an amino acid: two identical amino acids can form only a single dipeptide, but two identical monosaccharides can give rise to eleven different disaccharides (Figure 2.2) due to the availability of 5 different glycosidic linkages (1-1, 1-2, 1-3, 1-4 and 1-6) and the conformation of the anomeric carbon (α or β).

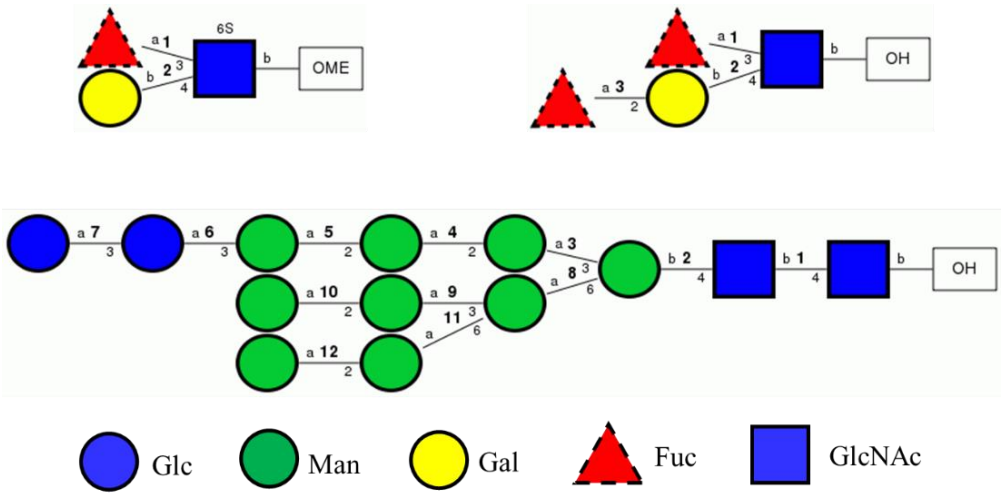



**Figure 2.1.** Examples of branched structures of carbohydrates. (images generated using oligosaccharide library available at www.glycam.org).

**Biological importance and applications**

Unlike proteins, carbohydrates are not encoded by the genome. Their assembly and degradation are dependent on the glycosyltransferases and glycosidases enzymes respectively, which are encoded by the genome. *In vivo*, carbohydrates are commonly found as glycoproteins and glycolipids, collectively known as "glycoconjugates". The activity of glycosyltransferases and glycosidases determines the glycosylation patterns of glycoconjugates. These glycoconjugates cover the cell surface, which places them in an optimal position to be able to communicate with the cell's environment. This makes them a target for glycan-binding proteins (GBPs), present at the interface between cells, tissues, and organs, to coordinate biological processes. The function of GBPs can be divided into two broad categories (endogenous or exogenous). GBPs in the first category recognize glycans from the same organism and are known to be involved in many physiologically important biological functions including embryonic development, differentiation, growth, cell-cell recognition, cell signaling, and metastasis (4-8).

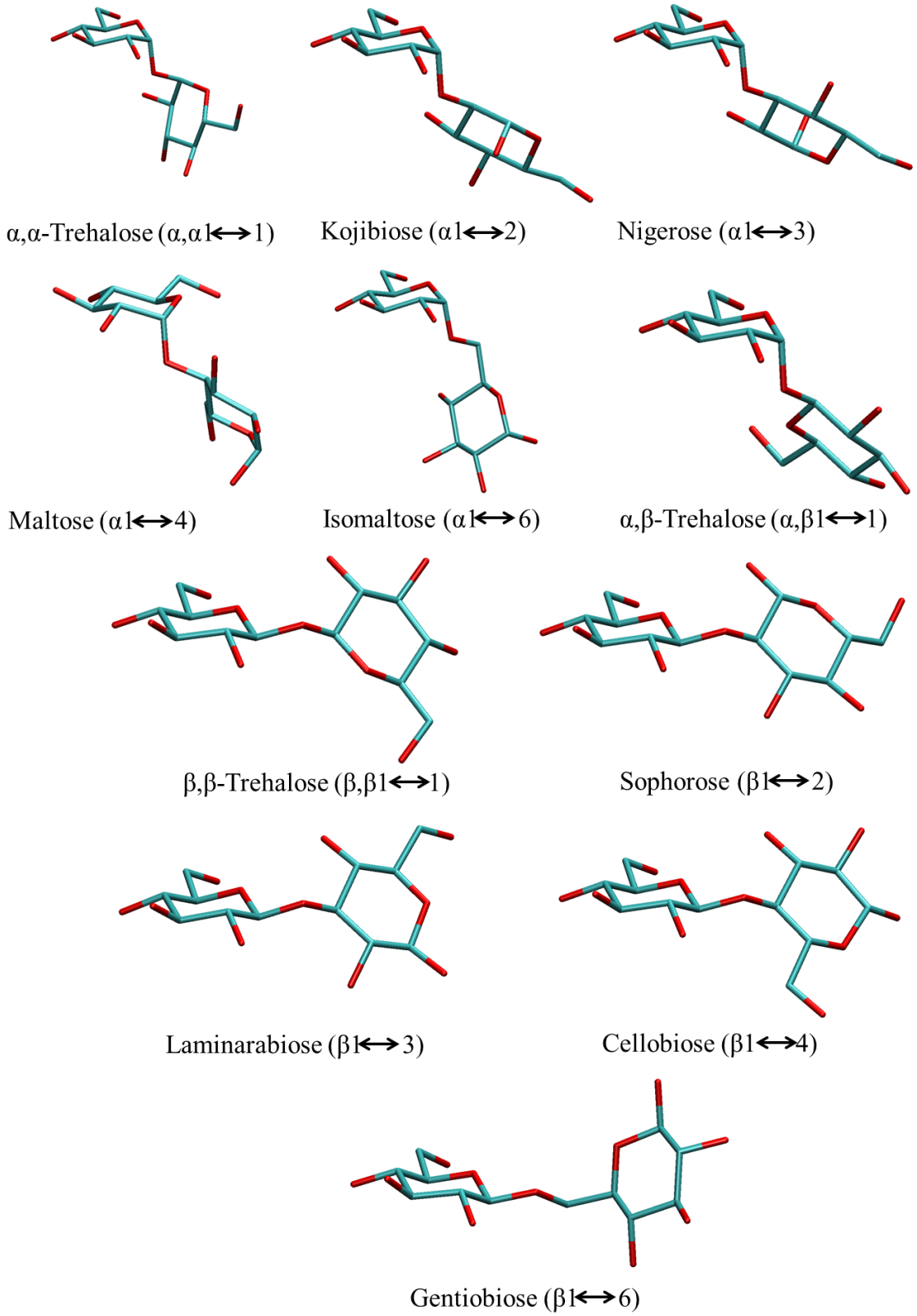



**Figure 2.2.** The eleven possible disaccharides of D-glucopyranose.

GBPs in the second category recognize glycans from a different organism. These consist mostly of pathogenic microbial adhesins, agglutinins, toxins, or host antibodies involved in host-pathogen interaction during infection, but some also facilitate symbiotic relationships (9-11). Alterations in the pattern of protein and cell surface glycosylation can lead to a number of diseases, from a range of cancers, rheumatoid arthritis to congenital diseases of glycosylation (12, 13). Therefore, the development of treatments aimed at targeting glycan-processing enzymes, or development of anti-bacterial vaccines specific for polysaccharides is an important step in drug development. Due to their specificity, some GBPs like lectins have found applications in various areas of science like medicine, clinical biology, agriculture, and biochemistry. They have been used to detect diseases, isolate glycoproteins and other carbohydrate containing molecules, and in staining and histochemistry of cells and tissues (14-16).

The specific nature of carbohydrate-GBP interactions has been exploited for drug targeting, i.e. to selectively deliver drugs to its intended site (17). This reduces the risk of unintended side effects, like in anti-cancer drugs. Moreover, carbohydrates can themselves be used as therapeutic agents. For example, a glucosaminoglycan heparin has been used as an anticoagulant for several decades. However, low tissue permeability, short serum half-life and poor stability makes them inadequate targets for oral drugs. To address these shortcomings, a new field of drug development called glycomimetics is emerging, which involves designing small molecules with bioactivity similar to carbohydrates that also show drug-like properties such as Oseltamivir (18). Oseltamivir is an orally administered antiviral medication used to treat influenza, designed by substituting exocyclic groups of sialic acid that were not required for affinity. Glycans can

perform a wide variety of these functions and find numerous applications owing to their stereochemistry and structure.

**Carbohydrate structure**

Monosaccharides contain one unit of aldehyde (aldose) or ketone (ketose) and can be classified as D- or L-isomers called enantiomers, based on the configuration of the penultimate carbon atom from the ketone or aldehyde group. Most monosaccharides exist in the D-configuration, with a few exceptions, like Fucose (19). Each monosaccharide has several chiral centers leading to multiple stereoisomers called epimers. In solution, they exist in equilibrium between cyclic and linear form, but the cyclic form is predominant. Cyclization adds another chiral center to the molecule forming two new diastereomers referred to as α (axial) or β (equatorial) (Figure 2.3) anomers. In their cyclic form, furanoses tend to adopt two different conformers called envelope and twist-boat, i.e. 20 conformations (Figure 2.4), while pyranoses prefer to adopt four distinct conformers called a chair, boat, skew boat and half chair. Pyranoses can also be found in envelope conformation while transitioning between different conformers, leading to a total of 38 (20) conformations (Figure 2.5). While most pyranoses favor the chair form in solution, iduronate, and glucuronate residues are known to be able to adopt multiple ring conformations.

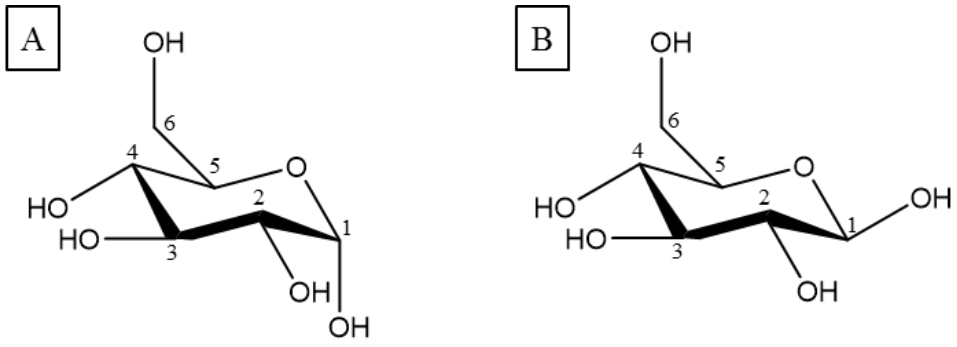



**Figure 2.3.** Anomeric configurations of glucopyranose. A. Axial (α) configuration at C-1. B. Equatorial (β) configuration.

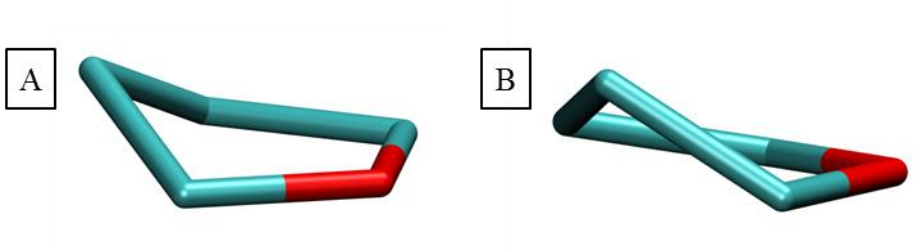



**Figure 2.4.** An example of the furanose envelope and furanose twist shapes. A. $^{2}E$ conformation. B. $^{2}T_{1}$ conformation.

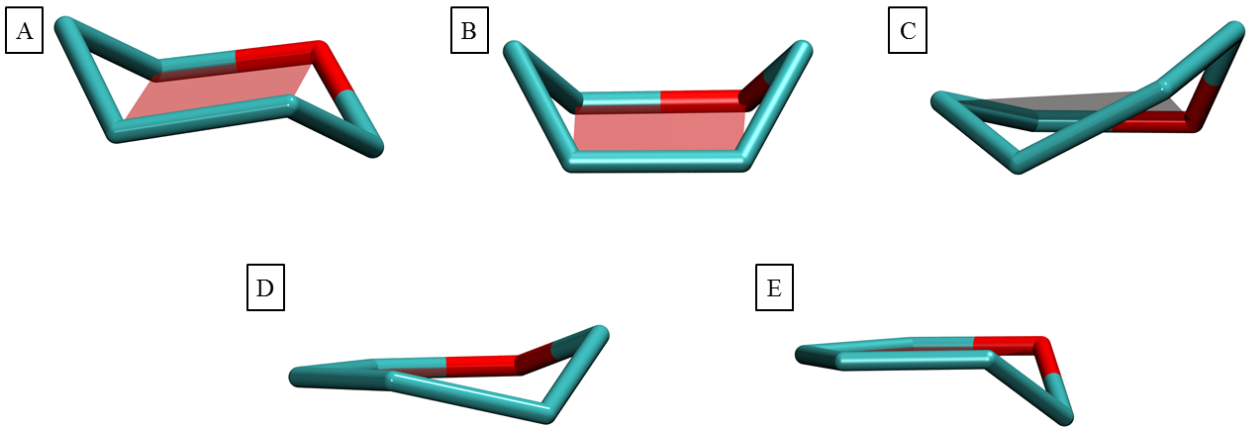



**Figure 2.5.** The cyclic conformations of pyranose. A. Chair ($^{4}C_{1}$). B. Boat ($^{1,4}B$). C. Skew-boat ($^{1}S_{3}$). D. Half-chair ($^{1}H_{2}$). E. Envelope ($E_{1}$).

Formation of disaccharides takes places via a condensation reaction between the anomeric hydroxyl group of one and the hydroxyl group of another, accompanied by the elimination of one water molecule forming a glycosidic bond between two monosaccharides (C-O-C bridge). The dihedral angles at the glycosidic linkage for 1-2, 1-3, and 1-4 connections are called φ (H1-C1-O-CX') and ψ (C1-O-CX'-HX'), where atoms marked with the prime symbol belong to the

residue on the reducing end (non-anomeric hydroxyl group). The glycosidic linkage for 1-6 connection has an additional dihedral angle called ω (O-C6'-C5'-H5'). The glycosidic linkages are a part of the backbone of a polysaccharide and determine its global structure.

A change in the linkage position or configuration can lead to profound changes in properties and functions. For example, maltose is a disaccharide made up of two glucose units linked by an alpha-1, 4-glycosidic bond, which forms a building block for starch. A disaccharide cellobiose, which forms a building block for cellulose is also made up of two glucose units, but they are linked by a beta-1, 4-glycosidic bond. While maltose can be easily digested by humans, we lack enzymes capable of breaking the beta-linkages, which is why we are unable to digest cellobiose. Similarly, yeast can be used to convert starch to ethanol, but does not act on cellulose.

**Carbohydrate chemistry and stereochemistry**

Relative to amino acids, monosaccharides are densely packed with polar exocyclic groups, and in solution, they can be indistinguishable from clusters of water molecules (21). However, unlike clusters of water molecules, monosaccharides have hydrophobic patches on both sides of the ring. It is this property that sets them apart from bulk solvent and introduces CH-π interactions (21). It has been implied that hydrophobic interactions are essential for affinity, and electrostatic interactions and hydrogen bonds provide selectivity to binding (22, 23). Carbohydrates also display stereoelectronic effects that can impact ring stability and linkage orientation. The anomeric effect is observed when there is a stabilizing interaction between the unshared electron pair on the ring oxygen and the σ* antibonding orbital of the C1–O1 bond in the axial orientation. Similarly, the exoanomeric biases the φ dihedral angles around the glycosidic bond into distinct rotational preferences, depending on the α/β configuration of the anomeric carbon involved (24). The orientation of the dihedral angle ψ is not influenced by the exo-anomeric

effect, but it is influenced by the steric constraints due to the attached rings (25). The dihedral angle ω, which does not involve a carbon atom located in the ring, shows three preferred orientations denoted tg, gg, and gt, where t and g represent trans and gauche, respectively (26). It is because of these structural and chemical properties that carbohydrates and glycans have a unique behavior compared to other biological molecules.

**Protein-carbohydrate interactions**

A single glycan may be involved in many different functions depending on its spatial and temporal expression. On the contrary, a given function might be carried out by several closely related glycan structures leading to cross-reactivity. However, despite these complications, the fundamental interactions between a carbohydrate and protein are similar.

*Electrostatic interactions*: Along with hydroxyl groups, sugars can also contain charged or polar groups such as carboxylate, acetyl, phosphoryl or sulfate, resulting in strong electrostatic properties. Hence, carbohydrate binding sites in proteins tend to have charged residues and/or ions (27). Charge-dependent binding interactions are regarded as major contributors to binding enthalpy (27).

*Hydrogen bonds*: GBPs like lectin require proper configuration of hydrogen bonds for a glycan to be able to bind, implying that hydrogen bonds impart specificity to these interactions. Mutations of hydrogen bonding partners can either inhibit binding or lead to a loss in affinity (28).

*Hydrophobic interactions*: CH-π interactions occur when CH groups on the hydrophobic face of the pyranoses interact with the π electron density of the aromatic ring. Even though these interactions are weak compared to hydrogen bonds, contributing between -0.5 to -0.8 kcal/mol to binding, they are important in stabilizing protein-carbohydrate interactions (29). The

hydrophobic face of the rings and the methyl moiety of the amido and acetamido sugars form hydrophobic interactions with aliphatic amino acid side chains.

*Entropy contributions*: Change in entropy is one of the main contributors to protein-carbohydrate binding. In solution, protein forms hydrogen bonds with water. For a carbohydrate to bind to a protein, these hydrogen bonds need to be broken, which incurs an enthalpic penalty. However, releasing tightly bound water molecules can result in favorable entropic changes (30). Changes in conformational entropy upon ligand binding also contribute significantly to the binding, as can be deduced from the analysis of thermodynamics of ligand binding to proteins, measured by isothermal titration calorimetry (ITC) (31). Ligand binding can cause the stiffening of both the ligand and the protein, which can contribute considerably to the free energy of binding. Therefore, designing conformationally restricted ligands (32, 33) and accounting for protein conformational entropy (34) can prove beneficial in predicting and comparing binding affinities.

**Challenges in studying carbohydrates**

Although carbohydrates have a great potential for future development of drugs and therapeutics (35), it is still challenging to exploit it due to a multitude of reasons. Unlike DNA, RNA and proteins, carbohydrates are not synthesized from a template, but by the combined action of multiple enzymes. This makes the in-vitro synthesis and amplification of complex carbohydrates difficult and tedious. Moreover, due to their high flexibility in solution, they are resistant to crystallization, leading to a deficiency in structural information. Owing to their structural heterogeneity, they can store a vast amount of information. This adds another level of complexity in the structure-function relationship of carbohydrates. Furthermore, carbohydrates typically have low binding affinities to GBPs in mili-molar to the micro-molar range. To perform biological functions, the strength of the interactions between carbohydrates and GBPs is

enhanced by multivalent binding, which leads to higher avidity. This is further complicated by a large number of techniques with different sensitivities that exist to quantify this affinity, producing data that is not always comparable.

The challenges are not limited to experimental methods, but also observed while modeling carbohydrate dynamics and interactions. The possibility of multiple ring conformations in solution, presence of branching, the internal flexibility of glycosidic linkage, and structurally and environmentally influenced glycosidic dihedral angle rotational preferences represent a unique set of structural and energetic features that can prove to be difficult to model accurately. As discussed earlier, carbohydrates possess highly polar exocyclic substituents causing complicated electrostatic features. For example, variations in conformations and stereoisomers of a single monosaccharide can lead to subtle variations in the spatial charge distributions in fixed charge force fields. Thus, these force fields require proper treatment of charge sets for both protein and ligand. Moreover, endo and exo-anomeric effects observed in carbohydrates that prefer sterically disfavored conformers should be accounted for by additional means. Due to their flexibility, adequately sampling all the conformations and configurations can prove to be computationally costly. Consequently, there has been an increased effort towards the development and improvement in the quality of computational tools specific for carbohydrates(36). However, sampling conformations is not the only obstacle in modeling protein-carbohydrate interactions. Estimating binding free energies can be a daunting task due to the intricate enthalpic and entropic relationship between protein, ligand, and solvent molecules. To overcome these bottlenecks, it becomes important to combine the knowledge from experimental and computational analysis to further our understanding. Some of these methods of analysis and current developments will be discussed in the next sections.

# CHAPTER 3

# COMPUTATIONAL METHODS USED TO STUDY BIOMOLECULES

## Molecular modeling

Molecular modeling includes a set of computational techniques used to replicate the structural and dynamic behavior of biological molecules. These techniques have found applications in fields of computational chemistry, computational biology, drug design etc. Most of the studies involve three steps i.e. choosing a method to model the interactions involved in the system, determining the type of calculation using these models to be performed based on the requirement of the study, followed by an analysis of these calculations. The two most commonly used methods to describe the inter- and intra-molecular interactions for molecular modeling are quantum mechanics (QM) and molecular mechanics (MM). QM utilizes complex mathematical formulations to explicitly model the electronic environment of each atom. Currently, this method is utilized for systems with a small number of atoms due to the computational expense required by these calculations. On the contrary, MM treats atoms as the individual basic unit drastically reducing the computational cost. This permits MM to be applied to much larger systems with biological relevance on much longer timescales. The potential energy of the molecular system is calculated using force fields.

## Classical mechanical force fields

A force field comprises of a mathematical formulation describing the potential energy function along with a set of parameters that can be adjusted to define the structural and dynamic behavior of a system. The individual components of a force field equation or potential energy function can

be divided as bonded and non-bonded interactions. The functional form of AMBER family of force fields is represented in equation 3.1 (37).

$$V_{Total} = \sum^{Bonds} \frac{1}{2K_r(r-r_0)^2} + \sum^{Angles} \frac{1}{2K_\theta(\theta-\theta_0)^2} + \sum^{Dihedrals} \frac{V_n}{2[1+\cos(n\Phi-\gamma)]} + \sum_{i<j}^{vdW} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right] + \sum_{i<j}^{Electrostatics} \frac{1}{4\pi\varepsilon_0\left[\frac{q_i q_j}{R_{ij}^2}\right]} \quad (3.1)$$

The potential energy function is calculated as a pair-wise summation over all the bonded and nonbonded atomic interactions. The bonded interactions include bonds, angles and dihedral angles that are formed by covalently attached two, three and four atoms respectively. The non-bonded interactions, represented by the last two terms in equation 3.1, comprise the vdW and electrostatic interactions between atoms separated by distance.

In classical mechanics, atoms are treated as balls attached to each other via a spring representing the covalent bonds between them. Therefore, a simple harmonic function describing elasticity like Hooke's law is used to model the dynamic behavior of bonds and angle. The equilibrium values are either based on crystallographic or QM optimized structures, or experimental diffraction data. The force constants are usually estimated using vibrational spectroscopy. It is important to note that by using Hooke's law, the bonds and angles are not allowed to break, at the same time atoms cannot form new bonds. Thus, CM MM is not used to study reaction mechanisms.

The van der Waals (vdW) interactions are weak steric interactions between two atoms with both repulsive and attractive components, which are often modeled using the Lennard-Jones 12-6 potential. The repulsive component accounts for Pauli exclusion, which prevents atomic overlap. The attractive component accounts for the London dispersion forces that arise due to

instantaneous multipoles. The parameters involved are adjusted to reproduce pure liquid or crystal properties, such as enthalpies of vaporization or sublimation.

Electrostatic interactions occur in atoms due to their electric charge, which can be attractive or repulsive depending on the charges of the atoms involved (positive or negative). Electric charge is polarizable and dependent on the surrounding electrostatic environment. However, force fields like AMBER assign a fixed partial charge to atoms, as implementing charge polarizability in force fields like AMOEBA (38), increases the computational cost considerably. A common approach to estimating partial charges is to reproduce QM generated electrostatic potential (ESP).

The dihedral term is parametrized as a last step in the force field development and is used as a correction. There are several assumptions involved in the development of a force field, therefore the total potential energy of the system calculated using the other terms is not always sufficient. This is usually demonstrated by the failure of CM to reproduce the energetics of torsional rotation generated by QM. Thus, the torsion term can also be referred to as a quantum correction to the potential energy.

The parameters are usually developed by employing small molecular fragments with representative properties of the relevant system, which can then be combined and applied to larger molecules. The accuracy of the force field is then tested and validated by comparing the results of an MD simulation to the experimental data.

**Molecular dynamics (MD) simulation**

Molecular dynamics simulations are used to predict the time-dependent behavior of a molecular system of interacting particles. This technique has been applied to many systems including proteins, nucleic acids, lipids, and carbohydrates, to sample new conformations and to determine

thermodynamic averages of these molecules and their complexes (39-41). Classical mechanics (CM) equations of motion are applied following Newton's laws. Per Newton's first law, an object in motion stays in motion maintaining its speed and direction unless acted upon by an external force. And the second law characterizes the effect of the application of an external force on the motion of a particle as shown in the equation 3.2, which states that its acceleration is directly proportional to the external force, while it is inversely proportional to the mass of the particle. This implies that if the force acting on a particle is known; its acceleration can be calculated.

$$\vec{F} = m\vec{a} \qquad (3.2)$$

An additional equation (equation 3.3) (one dimension version) of force states that it is a gradient of potential energy with respect to the position of the atom. As the forces on an atom are arising due to its interactions with other atoms, this gradient can be estimated using the potential energy from force field equation to calculate the force, and in turn, the acceleration. Therefore, MD is a deterministic technique, which means if an initial set of positions and velocities are provided, the following time progression of the system is in principle completely determined.

$$F = -\frac{\partial V}{\partial x} \qquad (3.3)$$

In CM MD, time is considered in discrete intervals or regularly spaced instances Δt, which is predefined and depends on the underlying timescales of motion under study, for example, biological systems often use time steps of 1-2 fs (42). A common method used to integrate Newton's equations of motion over time to predict new positions and velocities at time t+Δt is the Verlet algorithm (equation 3.6) (43). Therefore, current position (x(t)) and acceleration (a(t)) along with position from the previous step (x(t-Δt)), which can be calculated using equations 3.4 and 3.5, are sufficient to predict the future (x(t + Δt)) atomic position. Before starting an MD

simulation, the initial atomic coordinates are static and do not possess information for previous positions. Therefore, commonly velocities are assigned to each atom in the first step of the simulation, selected randomly based on a Maxwell-Boltzmann distribution (a probability distribution characterizing particle speeds) appropriate to the simulation temperature.

$$x(t + \Delta t) = x(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 \qquad (3.4)$$

$$x(t - \Delta t) = x(t) - v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 \qquad (3.5)$$

$$x(t + \Delta t) = 2x(t) - x(t - \Delta t) + a(t)\Delta t^2 \qquad (3.6)$$

The entire process of MD simulation can be summarized in following steps:

1. The force field equation is evaluated to calculate the potential energy of the system.
2. Acceleration at current time t is computed using the position derivative of the potential energy.
3. Equation 3.6 is evaluated to predict future position x(t+Δt).
4. The time is incremented by Δt and the atom is moved to its new position.

This process is repeated in a loop till it reaches the desired predefined length of the simulation.

*MD simulation setup*

The initial structure for an MD simulation may be experimentally determined e.g. from X-ray crystallography or NMR, or a theoretical model e.g. from homology modeling. To mimic the environment of a biological system, simulations are usually performed in the presence of a solvent by means of two possible approaches. The first approach called implicit solvent employs a set of mathematical approximations to estimate the influence of bulk water as a continuum around the molecular surface. Although this method is computationally inexpensive, it is not able to capture some important features involving solute-solvent interactions like hydrogen bonds.

Therefore, a more accurate second approach of using explicit discrete water molecules is favored in biological systems, despite it being much more computationally expensive. There are many explicit water models designed to replicate different molecular and bulk water properties. Some of the widely used and extensively tested models are TIP3P (44), TIP4P (44), and TIP5P (45). To reduce overall complexity, these three models are designed to be rigid and do not undergo the internal motions of bond and angle stretching. After solvation, counter-ions are then added to neutralize the system.

Due to the deterministic nature of MD simulations, biological systems are typically subjected to energy minimization prior to the dynamical study, to ensure that the simulation proceeds from a reasonable area of phase space. Energy minimization attempts to locate the nearest local minimum and in process eliminates large interatomic forces from steric clashes and unrealistic geometries in the structure. Minimization after solvation adjusts the solute and solvent relative to each other. Some well-known minimization algorithms include Steepest descent (46) and Conjugate gradient (47). While steepest descent performs well for minimizing initial structures, it is slow to converge. On the other hand, the conjugate gradient is unstable far from a local minimum but converges quickly. Therefore, usually conjugate gradient minimization is preceded by steepest descent minimization to reduce computational cost. Geometry optimization is followed by equilibration to bring the system to the desired temperature and pressure.

**Monte Carlo (MC) sampling**

Unlike MD, Monte Carlo (MC) is a stochastic computational sampling technique that explores the energy surface by randomly probing the configuration space of the molecular system, generating samples from the Boltzmann distribution at a given temperature. Because it does not keep track of time, the time-dependent dynamical properties of a molecular system cannot be

derived from an MC simulation. It is widely used to search for low-energy structures of a protein and to estimate thermodynamic quantities over a conformation space. The partition function is used to calculate most of the statistical thermodynamic properties of a system in statistical physics. The canonical partition function (Equation 3.7) is applied in protein simulation studies, where $\beta = 1/k_BT$ with $k_B$ as the Boltzmann's constant and T as the temperature. The summation is over all possible conformations ($x_i$), and their potential energies ($E_i$) are calculated using a force field. The partition function can be approximated by Importance sampling, i.e. conformations with low energy are emphasized by sampling more frequently.

$$Z = \sum_i e^{-\beta E_i} \qquad (3.7)$$

The starting structure of an MC simulation is a static structure like MD. Each MC step consists of proposing a new structure by perturbing some degrees of freedom (DOFs), usually torsion angles. The new structure is accepted or declined based on an acceptance criterion such as frequently used Metropolis criterion (48), according to which the simulation moves to a new conformation with probability min, where $\Delta E$ is the difference in energy between the new and previous conformations. A new structure with lower potential energy is always accepted, while that with higher energy is accepted with a decreasing probability, as the energy barrier increases. The entire process of MC simulation can be summarized in following steps:

1. Generate new coordinates by perturbing the previous structure.
2. Compute the change in potential energy. If $\Delta E < 0$, accept the new coordinates.
3. If $\Delta E > 0$, generate a uniform random number *R* in the range [0,1] and calculate probability P.
4. Accept the new structure if P>*R*, decline them otherwise.

**Enhanced conformational sampling techniques**

The energy landscape of proteins is very rugged with multiple minima. This makes exhaustive sampling at lower temperatures extremely difficult, as a simulation can get stuck at a local minimum. To overcome this problem, a sampling technique is required that can efficiently explore a complex energy landscape. Some of these are discussed below.

*Parallel Tempering*

Parallel tempering or Replica Exchange (49) is one such method that aims to overcome this limitation by running parallel simulations, at a broad range of temperatures, and exchanging structures between these simulations after a fixed number of steps. This allows the system to escape metastable states when a simulation is at a higher temperature and relaxing it when it is exchanged at a lower temperature. This method can achieve protein folding at timescales that are substantially smaller than that for simulations at a fixed temperature. The efficiency of parallel tempering Monte Carlo can be optimized by maximizing the number of trips between two extreme temperatures. This is achieved by optimizing the distribution of temperature points used to run the simulation for a specific system. An iterative feedback method is used that concentrates the temperature points near the phase transition for that system or the ground state, thereby increasing sampling closer to that point.

In the parallel tempering Monte Carlo algorithm, N replicas of the system are simulated in parallel at N different temperatures ($T_1$, $T_2$. . . $T_N$). After a fixed number of Monte Carlo sweeps, the two neighboring replicas, i and i+1 with energy $E_i$ and $E_{i+1}$, and temperatures $T_i$ and $T_{i+1}$ respectively are exchanged with a probability,

$$p(E_i, T_i \rightarrow E_{i+1}, T_{i+1} = \min\left(1, e^{(\Delta\beta\Delta E)}\right) \qquad (3.8)$$

where, $\Delta\beta = \frac{1}{T_{i+1}} - \frac{1}{T_i}$, is the difference between the inverse temperatures, and $\Delta E = E_{i+1} - E_i$, is the difference in the energy of the two replicas.

The minimum temperature of the replicas is usually close to the room temperature, and the maximum temperature is high enough that the protein does not get stuck in a local minimum. The initial set of temperatures is a geometric progression between the minimum and the maximum temperature. The number of replicas closer to the square root of the number of the residues in a protein is believed to provide good sampling, but there is no set rule for that. The temperature set can be optimized using the protocol in reference (50).

*Simulated Annealing*

Another method for overcoming the multiple minima problem is simulated annealing (51). This method works on the assumption that for a protein the global minimum in free energy at room temperatures is the global minimum in potential energy. Therefore, simulated annealing tries to mimic the crystal growth process to find the global minimum in potential energy, by gradually lowering the temperature of a simulation from a high value, to a lower value where the simulation does not undergo significant changes. It needs to be made sure that the decrease in the temperature over the period of simulation is slow enough that the system stays in thermal equilibrium so that it does not get trapped in local minimum. Multiple simulations are needed, to increase the probability of finding the global minimum, with different starting structures or with a different random number seed.

*Multicanonical sampling*

Multicanonical sampling is a generalized ensemble method, where each state is weighted by a non-Boltzmann probability weight factor, to achieve a uniform or flat energy distribution of all the states, which allows a free random walk in energy space (E). This method can be used to

overcome metastability in first order phase transitions, as well as a multiple-minima problem in various systems. Using this 'density of states' method, each configuration with potential energy E, is updated with a weight, $w \propto g^{-1}(E) = e^{-S(E)}$ and $S(E) = \ln(g(E))$, where g(E) is the density. This results in a uniform energy distribution, $P(E) \propto w(E)g(E)$. Because the weights are not known, they are estimated by iterations of short preliminary runs.

*Parallel Wang-Landau*

Wang-Landau Sampling (52, 53) is a powerful technique which has found application in various areas of research, including protein modeling (54). Like multicanonical sampling, using this 'density of states' method, one can achieve a flat energy distribution by calculating the density of states (g(E)) of a system iteratively, by using a flatness criterion and a modification factor (f). First, g(E) is assumed to be uniform, usually, 1 for all configurations and f is typically e, and a histogram (H(E)) is introduced, which keeps track of all the visits to each energy level. The acceptance probability of exchange between previous ($E_1$) and new ($E_2$) conformation is $P_{acc} = \min(1, \frac{g(E_1)}{g(E_2)})$. After each visit the histogram is updated, so is the energy of the last visited conformation with f. after a flatness criterion is reached, the simulation starts from the beginning with new modification factor i.e. $\sqrt{f}$. This is repeated until f reaches a threshold ($10^{-6}$). The performance of this technique can be further improved by its parallelization. Vogel et al (55) describes a novel method of parallel WL based on replica exchange, which gives a remarkable speed-up without loss of accuracy. The entire energy range is divided into h windows with an overlap. Each energy window can have multiple walkers, with their own energy value and density of states $g_i(E(x))$, where i is the walker and E is the energy of conformation x, and H(E). These walkers undergo a replica exchange monte carlo, with an acceptance probability of exchange, $P_{acc} = \min[1, \frac{g_i(E(X))}{g_i(E(Y))} \frac{g_j(E(Y))}{g_j(E(X))}]$. Calculation of g(E) is also iterative in this case. After

reaching a flatness criterion, g(E) is averaged out and redistributed within the energy sub-window, before starting with the next iteration.

**Interaction energy calculation using the molecular mechanics–Poisson-Boltzmann/generalized Born surface area (MM-PB/GBSA) method**

Predicting the strength of carbohydrate-protein interactions accurately can be more challenging than reproducing their conformational behavior in solution. MM-PB/GBSA is an end-state post-processing method which allows high throughput calculation of free energies of binding by substituting explicit solvent with an implicit solvent (continuum dielectric model) and using a thermodynamic cycle depicted in figure 3.1, where P represents protein and L represents a ligand (56). It is usually determined as an average over multiple snapshots collected from an MD simulation, after removing all the explicit water molecules. The binding energy in this method is calculated as the equation 3.9 for every snapshot.

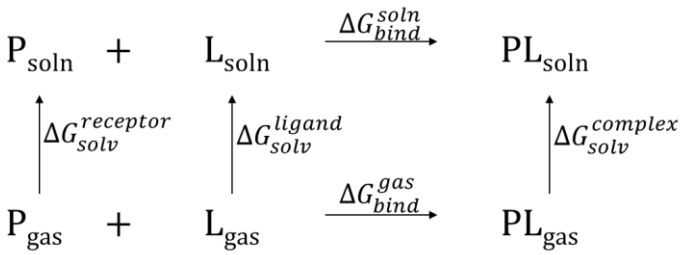



**Figure 3.1.** Thermodynamics cycle used by MM-PB/GBSA.

$$\Delta G_{bind} = G_{complex} - \left(G_{receptor} + G_{ligand}\right) \qquad (3.9)$$

The potential energy, G for each system is calculated as the sum of the gas phase potential energy which can be estimated using the force field equation ($E_{MM}$), the solvation free energy ($\Delta G_{Sol}$) and the conformational entropy (TS) (equation 3.10).

$$G = E_{MM} + G_{Solvation} - TS \qquad (3.10)$$

The solvation energy comprises of the electrostatic (polar contribution), and non-electrostatic solvation part (non-polar contribution). The electrostatic solvation is estimated using either GB or PB implicit solvent model, and the non-electrostatic solvation energy is estimated by solvent accessible surface area (SASA). The PB model is based on the Poisson continuum dielectric model for finding the electrostatic potential, combined with a Boltzmann distribution for finding the distribution of charges. Because GB model approximates PB method, it makes GB model more computationally efficient. In MM-PB/GBSA method, the solute with low dielectric lies within this continuum high dielectric model substituting water molecules, which is approximated by the non-polar component. Combining these equations, the binding free energy in solution involves energy changes accompanying the protein-ligand binding in the gas phase, the solvation of complex, protein and ligand independently, and the entropy changes upon binding (equation 3.11).

$$\Delta G_{bind}^{soln} = \Delta G_{bind}^{gas} + \Delta G_{solv}^{complex} - \left(\Delta G_{solv}^{receptor} + \Delta G_{solv}^{ligand}\right) - T\Delta S \qquad (3.11)$$

Although MM-PB/GBSA is routinely used to estimate binding free energies (57, 58), due to a number of assumptions inaccuracies can occur in these estimations. Implicit solvents can reproduce the behavior of bulk water, but not able to reproduce the behavior of water molecules that form a part of hydrogen bond network aiding the protein-ligand interaction. Because all the water molecules are removed for post-processing, the stabilizing effects of some the water molecules involved in binding will be neglected (59-61). Furthermore, PB and GB models can fail to estimate suitable polar and buried charges as the non-polar effect is represented by only a surface term.

CHAPTER 4

# MONITORING LARGE-SCALE PROTEIN CONFORMATIONAL CHANGES[1]

## Introduction

It is not always possible to determine high-resolution 3D structures of proteins and complexes using experimental techniques such as X-ray crystallography and NMR. Biomolecular surface mapping methods have emerged as a powerful substitute for characterizing protein-protein and protein-ligand interactions in such cases. Therefore, there has been a growing interest in alternate high throughput methods such as footprinting and molecular dynamics (MD). Moreover, x-ray structural data is unable to provide any information regarding the dynamics of side chains in solution, which can be obtained through an MD simulation. This can also prove important when employing homology models, to generate a realistic ensemble of side chain orientations. MD has been shown to characterize the relation between the per-residue degree of oxidation from footprinting and Solvent Accessibility Surface Area (SASA) (62). This relation can be used to better understand the dynamics and conformations of a protein. This will be illustrated by using a fusion protein from Parainfluenza virus that undergoes a large conformational change upon interaction with the host.

[1]Published as - Poor TA, Jones LM, Sood A, Leser GP, Plasencia MD, Rempel DL, Jardetzky TS, Woods RJ, Gross ML, Lamb RA. Probing the paramyxovirus fusion (F) protein-refolding event from pre-to postfusion by oxidative footprinting. *Proc. Natl. Acad. Sci. USA*. 2014 24;111(25):E2596-605.

Parainfluenza virus is an enveloped, negative-sense, single-stranded RNA virus (63) that recognizes (binds to a receptor) an appropriate target cell by variable attachment protein (Hemagglutinin-neuraminidase (HN), H, or G), and carries out invasion (fusion) by a more conserved fusion protein (F) (64-66). Both proteins are embedded in the lipid bilayer of the virus. For a successful infection, F protein undergoes irreversible large-scale and complicated conformational change, from prefusion structure to postfusion structure after activation by HN or heat (67-74). In prefusion conformation, the fusion protein is metastable, while in the postfusion state is highly stable. Apart from the prefusion and postfusion crystal structures, very little information is available about the intermediate structures during the fusion process or about the dynamics of this highly mobile protein in solution.

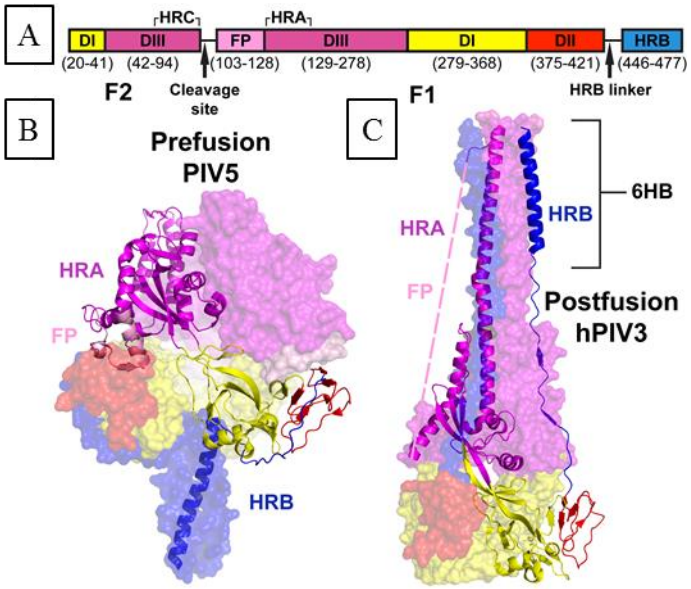



**Figure 4.1.** Organization of the trimeric, soluble PIV5 F protein. (A) The domain organization of PIV5 F-GCNt, with a unique color for each domain. Corresponding amino acid residues are noted below each segment. (B) The prefusion crystal structure of PIV5 F (PDB: 2B9B) and (C) the postfusion crystal structure of hPIV3 F (PDB: 1ZTM), colored per (A). In both (B) and (C), two of the trimers are represented as semi-transparent surfaces while the third trimer is depicted

as a ribbon cartoon. Highlighted structural elements include the heptad repeats A and B (HRA and HRB, respectively), the postfusion 6 helix bundle (6HB), and the hydrophobic fusion peptide (FP), which is disordered in the postfusion crystal structure. (Reprinted here with the permission of the publisher).

Theoretical SASA values were calculated and compared to the experimental fast photochemical oxidation of proteins (FPOP) coupled with high-resolution mass spectrometry for peptides of prefusion PIV5 F protein (PDB ID: 2B9B, Figure. 4.1B) and of a model based on the postfusion hPIV3 crystal structure (PDB ID: 1ZTM, Figure 4.1C). The average SASA (<SASA>) values were calculated for the peptides from 10 ns MD simulations. Most of the peptides compared show consistent trends in the <SASA> and FPOP oxidation for the peptides from prefusion-to-postfusion conformational change. While there are a few that exhibit greater-than-predicted FPOP oxidation in the prefusion state. The functional relevance of these regions will be further discussed.

**Methods**

*Preparation of pre- and postfusion protein structures*. A crystal structure of the PIV5 F protein in its prefusion conformation (PDB ID: 2B9B) was reported at 2.85 Å resolution and was used in this analysis (75). All the crystallized ligands (*N*-Acetyl-*D*-Glucosamine) and water molecules were removed. All histidine residues were considered as neutral with a hydrogen atom on the epsilon nitrogen. Because there is no structure available for the postfusion state of PIV5 F, a homology model was generated using the SWISS-MODEL (76) homology modeling server with the hPIV3 postfusion crystal structure (PDB ID: 1ZTM) as a template (22% identity, 65% similarity, and $E_{value} = 3 \times 10^{-115}$). A sequence 41 residues long and consisting of the

hydrophobic FP was unresolved in the hPIV3 F structure, which suggests that it does not adopt a defined conformation. To avoid biasing the model by introducing this peptide in a single possibly irrelevant conformation, this sequence was also omitted from the homology model, necessitating the use of backbone restraints in subsequent simulations. The terminal residues at the location of the missing loop were capped with *N*-methyl (NME) and *N*-acetyl (ACE) groups as appropriate. To check the quality of the modeled and template structures, Z-scores of the backbone conformations were calculated using WHATIF (77). Both the template and the model received acceptable Z-scores of 1.2 and 0.7, respectively. It is well established that class I fusion proteins share structural features without having high sequence similarity; therefore, the model generated was accepted as is and was used for further analysis.

*Energy Minimization*. Each protein structure was solvated in a truncated octahedral box of TIP3P water molecules (78), with counter ions ($Na^+$) were added to neutralize the charge, using the tLEAP module of AMBER. In the case of the prefusion structure, 97332 water molecules were required, whereas the postfusion required 110360. The simulations were performed using AMBER12 force field (79) with ff99SB parameters (80), with a cutoff for non-bonded interactions of 10 Å. To remove bad contacts, the system was minimized in two steps. Firstly, the energy of the water and ions was minimized while keeping all protein atoms restrained (500 kcal/mol $Å^2$). This was followed by energy minimization of the entire system. Each minimization was comprised of an initial phase of steepest descent method for 5000 steps, followed by conjugate gradient for 20000 steps. The resulting minimized structures were subjected to MD simulation performed with the pmemd.cuda version of AMBER12 (81).

*MD Simulation*. All the bonds involving hydrogen were constrained using the SHAKE algorithm, enabling an integration time step of 2 fs. Long-range electrostatic interactions were

treated with the Particle-Mesh Ewald algorithm (82), with a long-range non-bonded interaction cut-off set to 10 Å. The systems were heated from 5 K to 300 K over a span of 50ps, under NVT conditions employing the Berendsen thermostat with a coupling time constant of 1 ps. The simulation was then continued for 10 ns under NPT conditions with weak restraints on the backbone atoms (100 kcal/mol $Å^2$). The first 1 ns of this trajectory was discarded prior to analysis of the equilibrated data.

*Data analysis*. Solvent accessible surface area (SASA) values were computed with the NACCESS (83) program for snapshots collected every 10ps. Average values were computed from the total 900 snapshots. Error bars on SASA data graphs represent ±1 standard deviation.

*Unfolding simulations*. Ten different globular proteins were selected for unfolding i.e. hen egg-white lysozyme (PDB ID: 2LYZ), bovine trypsin (PDB ID: 2PTN), rat biliverdin reductase (PDB ID: 1GCU), Mg-chelatase cofactor GUN4 (PDB ID: 1Y6I), RdgB- inosine triphosphate pyrophosphatase (PDB ID: 1K7K), malonyl-CoA acyl carrier protein transacylase (PDB ID: 1MLA), pectate lyase C (PDB ID: 2PEC), human Pp2A phosphatase activator (PDB ID: 2IXM), TEM-1 beta-lactamase (PDB ID: 1ZG4) and xylanase 10A (PDB ID: 1E0W). To generate an ensemble of partially-unfolded structures, the energy-minimized structure of the unsolvated protein was heated during an MD simulation from 5 K to 1000 K over 10ns *in vacuo*. Snapshots were extracted from the simulation every 100ps.

**Results and Discussion**

*Solvent Accessibility Surface Area (SASA*) *analysis*. The per-residue SASA values were calculated from molecular dynamic simulations of available crystal structure data of PIV5 F-GCNt in the prefusion conformation and its homology modeled postfusion conformation using hPIV3 crystal structure. These per-residue SASA values were then evaluated at a per-peptide

level based on the 17 tryptic peptides of PIV5 F-GCNt in the prefusion and postfusion conformations. The changes in per-peptide SASA for both the conformations were compared to the changes in per-peptide oxidation, which can then be localized on the structures. The oxidation of amino acids due to free radicals is not only dependent on its accessibility to the solvent, but also its reactivity to free radical, which is different for different amino acids. Therefore, a direct inter-peptide comparison of raw values of oxidation and SASA is difficult. For the scope of this study, the evaluation will be limited to the net change from prefusion to postfusion for individual peptides, making the reasonable assumption that the total reactivities of each peptide are the same across the analysis. The net change in FPOP oxidation from prefusion to postfusion is calculated as $\%Oxidation_{prefusion}$ - $\%Oxidation_{postfusion}$, similarly, the net change in SASA is $SASA_{prefusion}$ - $SASA_{postfusion}$. When these are compared, different trends emerge for many of the peptides.

Of the tryptic peptides contained in both models, 10 out of the 17 peptides (highlighted in pink), demonstrate an FPOP labeling change that is different from what would be predicted by changes in the side chain SASA of the static crystal structures alone (e.g. prefusion > postfusion vs. prefusion = postfusion). Experimental FPOP values that deviate from what would be predicted from the side chain SASA calculations suggest that the crystal structure data do not accurately reflect the flexibility of one or both states. As 9 out of 10 of the disagreeing peptides have larger, more positive (prefusion > postfusion) ΔFPOP values, it suggests that the solvent accessibility calculations either underestimate the prefusion SASA or overestimate the postfusion SASA. Given that: 1) there is very little structural variation between the postfusion structures of multiple paramyxovirus F proteins; 2) the postfusion state is a high stability conformation that represents an energy minimum for the protein; and 3) there does exist variation between the two

prefusion crystal structures of paramyxovirus fusion proteins (PIV5 F and RSV F), we attribute most of the ΔFPOP/ΔSASA trend disagreement to an underestimation of the prefusion SASA calculation. The lopsided distribution of the 9 disagreeing peptides likely reflects the global decrease in flexibility of the postfusion state relative to the prefusion state. Further, FPOP labeling that deviates from expected SASA trends may highlight regions of greatest solvent accessibility and protein flexibility.

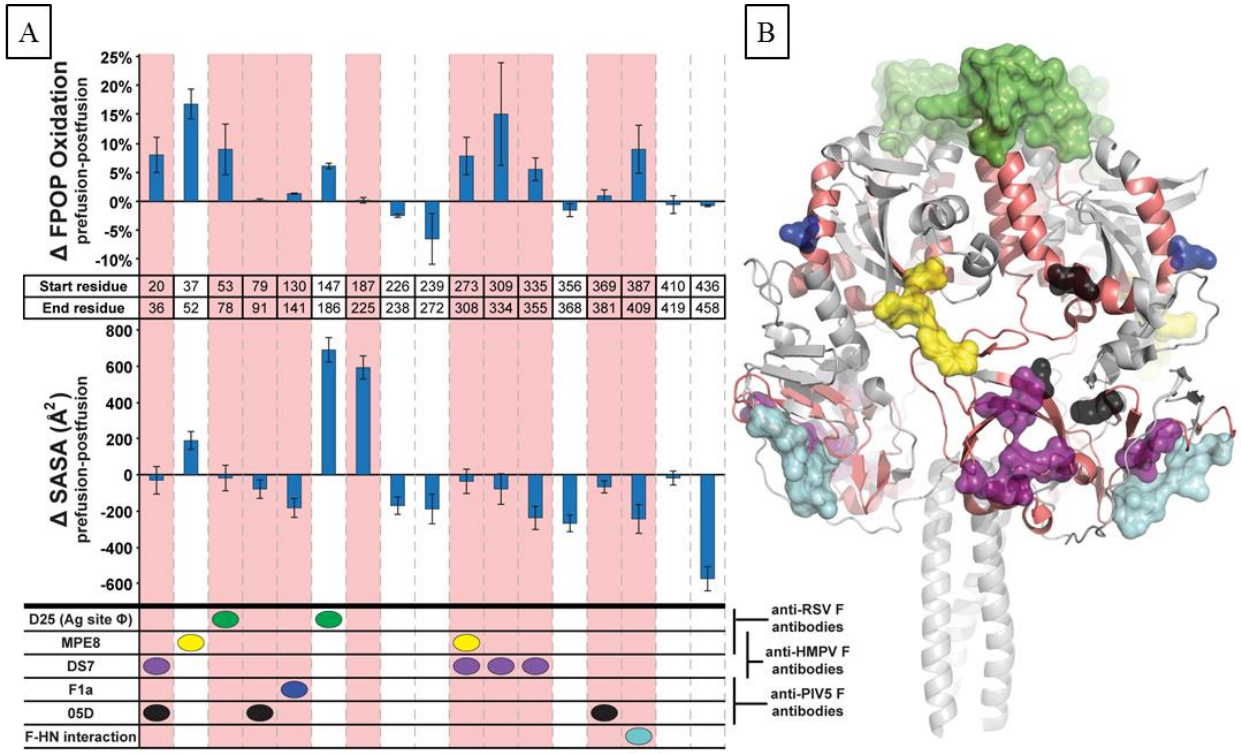



**Figure 4.2.** Changes in peptide FPOP oxidation and side chain SASA between the prefusion and postfusion states. (A) Postfusion FPOP or SASA values are subtracted from the corresponding prefusion FPOP or SASA values for the tryptic peptides that are common between the two states and graphed. The data represent the change in FPOP oxidation or side chain SASA between the prefusion and postfusion states. Peptides that exhibit a different ΔFPOP trend compared to ΔSASA are highlighted by pink backgrounds. The backbone of the cartoon trimer in (B) is colored similarly. At the bottom of (A), the epitopes of prefusion-specific, neutralizing

antibodies against various paramyxoviruses and the HN-interacting surface of PIV5 F are assigned to their homologous PIV5 F tryptic peptides. These important epitopes and interaction surfaces are represented as semi-transparent surfaces in (B), colored according to the ovals at the bottom of (A). The distribution of epitopes and surfaces to peptides that show greater-than-expected ΔFPOP values (relative to ΔSASA) is not random ($p$ = 0.0023, 1-tailed Fisher's exact test), suggesting a correlation between regions of increased prefusion FPOP oxidation and functionally important parts of the metastable prefusion trimer. (Reprinted here with the permission of the publisher).

Interestingly, of the 9 peptides that show different ΔFPOP values relative to the ΔSASA, most contain the epitopes of prefusion-specific, neutralizing antibodies that have been discovered for a range of paramyxovirus F proteins. The epitopes of D25 (α-RSV F Fab) (84), MPE8 (broadly neutralizing against hRSV and hMPV) (85), DS7 (α-hMPV F Fab) (86), F1a (α-PIV5 F Mab) (87), 05D (α-PIV5 F Fab) are colored per the ovals at the bottom of Figure. 4.2A and mapped onto their homologous sequences in the prefusion PIV5 F atomic structure (Figure. 4.2B). Peptide 387-409 of PIV5 F (Figure. 4.2A,B, cyan surface) contains residues that have been implicated in the interaction of PIV5 F and HN (88) and MeV F and H (89) as well as being homologous with the RSV F antigenic site IV. The extensive overlap between antigenically or functionally significant regions of paramyxovirus fusion proteins and PIV5 F-GCNt peptides with greater-than-predicted levels of FPOP oxidation in the prefusion state is not random ($p$ = 0.0023, one-tailed Fisher's exact test) and suggests that there is a correlation between flexibility and functionality for these metastable proteins. Binding of neutralizing antibodies to these regions in paramyxovirus F proteins could either stabilize the flexibility of these dynamic

regions in the prefusion state or sterically interfere with the refolding event. These results demonstrate the importance of evaluating SASA to understand protein conformation and flexibility. Therefore, it is appropriate to further investigate the dependence of SASA on the type of amino acids.

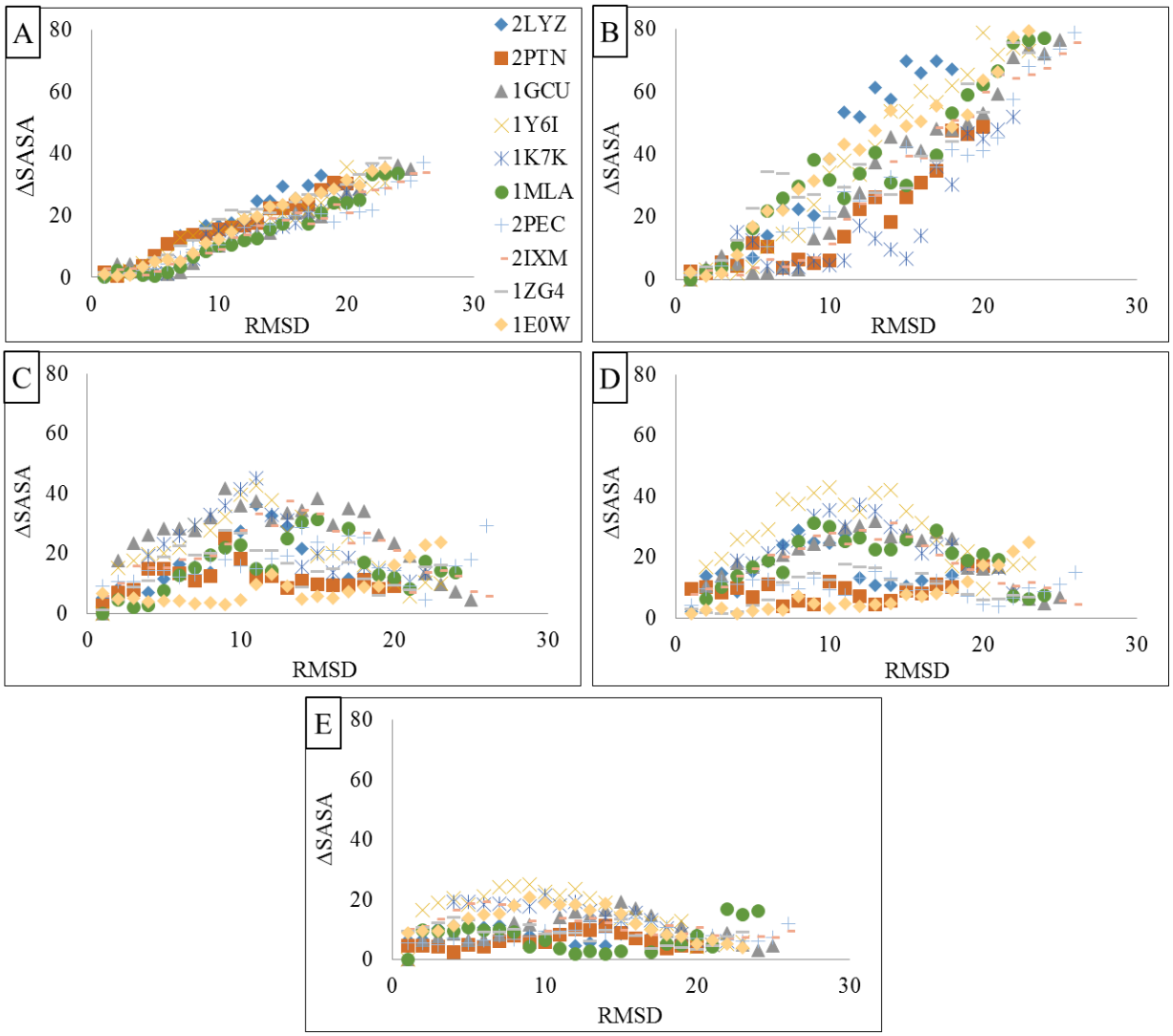



**Figure 4.3.** The relationship between ΔSASA and RMSD w.r.t the crystal structure as the protein unfolds. The ΔSASA values of non-polar (A) and aromatic residues (B) increases as the proteins unfold. The polar (C), positively charged (D) and negatively charged (E) residues do not behave in a similar fashion and lack a correlation with the RMSD.

*The relationship between all-atom RMSD and SASA*. In globular proteins, the hydrophobic residues are generally buried inside the structure, while the hydrophilic residues reside on the surface in contact with the solvent. Therefore, there is a relationship between the protein structure and solvent accessibility of different types of residues. To test this, ten different proteins were slowly unfolded and the solvent accessibility of all the residues was measured every 100ps. The effect of protein unfolding on SASA of the residues can be shown by plotting the all-atom RMSD between the crystal and the unfolded structure versus the absolute difference between their SASA (abs($SASA_{ref}$-$SASA_{current}$)), where $SASA_{ref}$ is the per-residue SASA value of the crystal and $SASA_{current}$ is the per-residue SASA value for snapshots from the unfolding simulation. As expected, non-polar (glycine, alanine, valine, leucine, methionine, isoleucine, and proline) and aromatic (phenylalanine, tyrosine, and tryptophan) residues show a direct correlation between the two (Figure 4.3A and B), i.e. as the protein unfolds their SASA starts to increase. While the SASA of the charged (lysine, arginine, histidine, aspartic acid and glutamic acid) and polar (serine, threonine, cysteine, asparagine, and glutamine) residues does not exhibit this relation (Figure 4.3C, D, and E).

**Conclusions**

This study highlights the need to view proteins as highly dynamic molecular machines. Correlation with the reported epitopes of neutralizing antibodies against other paramyxoviruses suggests that regions of viral fusion proteins that experience larger changes in FPOP labeling than predicted from the static crystal structure data may be important for proper protein function and, thus, make good targets for the development of neutralizing antibodies or small molecule inhibitors. We suggest that FPOP coupled with SASA results can provide the dynamic structural insights necessary for guiding more targeted and efficient approaches to anti-viral therapeutic

development. We also find a correlation between different types of amino acids and their SASA. Non-polar and aromatic residues tend to be buried inside the protein core, while charged and polar residues are relatively exposed. Due to no net difference in SASA of charged and polar residues of folded and unfolded conformation of a protein, they are not very informative about protein structure. On the other hand, as non-polar and aromatic residues show a direct correlation, in theory, they can be used as indicators for different conformations of proteins.

# CHAPTER 5

# INTEGRATING MS FOOTPRINTING DATA IN PROTEIN STRUCTURE MODELING

## Introduction

Almost all the biological processes are carried out by proteins, like replication, transcription, translation, metastasis and apoptosis and many other signal transduction pathways relevant to development and cell - cell communication. Protein interactions and functions are dependent on protein structure, therefore, understanding protein tertiary and quaternary structure is fundamental to understanding mechanisms of protein function. There are three main methods for obtaining 3D structures with varying levels of resolution i.e. X-ray crystallography, NMR spectrometry, and modeling. X-ray crystallography is a powerful tool for protein structure determination. For studying a protein using this technique, it first needs to be crystallized. Crystallization often takes place at environments far from biological relevance. This can result in regions of the protein with significantly different conformations in the crystal from those in solution (90). Moreover, obtaining crystals of a protein can be a difficult task. NMR spectrometry is a valuable tool for determining protein structures because it can be done in solution and does not require crystals. But, structures determined from insufficient restraints or misinterpreted data can be incorrect (91, 92). It is not a favored method to be used for larger proteins, but is suited to small proteins, typically smaller than 25 Da, as larger proteins show increased spectral complexity (93). The accuracy of computational methods for the determination of protein structures is varied (94). This is especially true when analyzing proteins with no known homologs. Low-resolution techniques can provide insight into the overall

secondary structure, solvent accessibility of the backbone or side chain etc. These techniques include absorption, fluorescence, and circular dichroism, chemical derivatization methods like biotinylation or acetylation, and mass spectroscopy-based protein footprinting, but they provide only sparse data and cannot be used to generate atomic coordinates, unlike high-resolution techniques. Due to all the reasons mentioned above, determination of the high-resolution structure of proteins and the refinement of low-resolution protein structure are long-standing challenges (95).

The accuracy of the computational methods can be improved by introducing experimental constraints for protein structures, and conversely, computational models can be used to interpret sparse experimental data. NMR restraints have been used to improve such predictions (96). Similarly, solvent accessibility can be employed as a constraint for certain amino acids, which help to define the solvent-accessible surface areas of the folded protein (97). Hydroxyl-radical protein footprinting (HRPF) is an emerging method of probing protein solvent accessibilities and mapping the surface of a protein. Hydroxyl radicals are highly reactive and covalently react with almost all the amino acid side chains, which are accessible on the surface of the protein, but with different intrinsic reactivities (98). The extent of hydroxyl radical oxidation depends in part on amino acid accessibility i.e. residues that are exposed to the solvent will react more readily than buried residues and on the reactivity of side chains. Hydroxyl radical footprinting coupled with MS has become increasingly popular as a labeling technique to probe intact protein structure, protein-protein interactions, protein folding, and protein-small molecule ligand interactions in solution (99-104). Based on their reactivity, side chains can be divided into three categories i.e. high, medium or low. The residues which are highly reactive or with medium reactivity are the most informative, and the ones with low reactivity do not show a linear relation between SASA

and percentage oxidation (62). A quantitative relationship between SASA and the magnitude of oxidation, for most informative amino acids, can be derived and used to estimate protein oxidation level with known 3D structure or to predict per residue SASA, given percentage oxidation. This combination of experimental assessment of side chain oxidation and theoretical estimation of SASA can be used to study conformational changes in proteins in solution and characterize ligand-binding. In this study, we present a computational method, which combines modeling and protein footprinting to obtain a high-resolution structure, as well as test the quality of generated models.

*Hydroxyl radical protein footprinting (HRPF)*

Protein footprinting involves the study of the surface of proteins by determining the solvent accessibility of the amino acid side chains. One of the ways of protein footprinting is to use chemical modification reagents, which react with side chains based on their reactivity. The hydroxyl radical is one such chemical modification reagent which has been used widely for this purpose. They act as good probes for solvent accessibility because of their comparable size to water, and nonspecific reactivity to several amino acid side chains, leading to good protein coverage. They also provide stable covalent modifications, which gives the user ample time for analysis. Protein solution oxidized with hydroxyl radicals is subjected to tryptic digestion and the spatial labeling patterns are analyzed with mass spectrometry. There are several methods of generating hydroxyl radicals with varying timescales for use in footprinting for example Fenton (101, 105-108) and Fenton-like reactions, Radiolysis of water via γ rays (109-112) or X-rays (113, 114) and UV Laser photolysis of $H_2O_2$. Hambly and Gross (115) (fast photochemical oxidation of proteins (FPOP)), and independently Aye and coworkers (116) developed a fast-labeling method to dissociate $H_2O_2$ using a laser beam, generating radicals in a nanosecond to

microsecond timescale. This technique has been demonstrated to obtain structural information of proteins that is fast and reliable.

The extent of oxidation of side chains by hydroxyl radicals is dependent on the solvent accessibility of the residues at the surface of a protein and on their chemical nature which determines their reactivity with hydroxyl radicals. It has also been observed that the oxidation has a small dependence on the local sequence as well (100). Sulfur containing and aromatic residues are the most reactive. The relative reactivity of the side chains is as follows: Cys > Met > Trp > Tyr > Phe > Cysteine > His > Leu , Ile > Arg, Lys, Val > Ser, Thr, Pro > Gln, Glu > Asp, Asn > Ala > Gly (117-119). As discussed above, there are several ways of generating hydroxyl radicals, which expose the protein to the radical for different lengths of time, but the chemical modifications resulting from the exposure are largely the same (117, 119, 120). Each amino acid has more than one competing mechanism of oxidizing via $HO^{\bullet}$ leading to different products (121). These mechanisms and major oxidation products have been identified and widely studied. Due to their detectable products, many aromatic, aliphatic, sulfur-containing, and charged residues are useful footprinting probes (122).

To extract information from this experiment, the protein solution is exposed to a series of a radical burst of different exposure times. The protein is cleaved using site-specific proteases to produce defined peptides. Peptides are separated by chromatography and to locate the residues modified by HRPF in each peptide, Tandem mass spectrometry is used. The peak area under the ion signal of the unmodified and modified peptide is compared, and the rate of modification of each peptide as a function of exposure time is calculated from dose response curve. The dose-response curve is plotted as the unmodified fractions versus X-ray exposure times. Pseudo-first-order function is used to fit these curves to get modification rate (123). To make sure of the

accuracy of the data it is important to make sure that the primary oxidation events are considered, and that all the different factors that affect the rate of oxidation are known, studied and accounted for. Due to multiple oxidation products, it is quite challenging to study the MS/MS spectra of each peptide, to determine the sites of oxidation. A lot of times, manual interpretation of the spectra is required.

Protein structures can get denatured or unfolded, if subjected to denaturants or if their side chains are modified. Therefore, upon oxidation, the native structure of a protein can unfold. Even a single oxidation event can trigger protein unfolding, which can expose residues excluded from solvent accessibility and cause increased oxidation events, resulting in misleading information (124-126). These events are affected by the time of exposure of the native structure of the protein to a denaturant, in this case, an oxidizing agent. So, for hydroxyl radical protein footprinting to be used as a reliable method for mapping protein structure, it should be done at shorter timescale than protein unfolding or other conformational changes.

Therefore, to study native protein structure, fast methods of generating radicals with low exposure times are preferred, like FPOP, which can produce radicals in a few nanoseconds**.** A 248 nm pulsed laser beam is used, to photodissociate $H_2O2$ into two hydroxyl radicals. To maximize the exposure of radicals to a small volume of protein, a flow system is designed such that the laser produces a small window of high flux light. The diffusing radical is dispersed through the protein solution reacting with the analyte as well as buffer components. The radical concentration achieved from a single pulse of a dilute peroxide solution (1% or less) is adequate to achieve significant levels of protein surface oxidation. Without any other additives, radicals require up to 100 microseconds to self-quench, long enough to allow super secondary structure unfolding. For quenching radicals in a shorter time, the appropriate scavenger is added like

glutamine or phenylalanine, which can bring down the exposure time of radicals to within a microsecond, as shown by kinetic analysis. This time is short enough that no large-scale protein motions take place and it can be argued if any super secondary structural changes take place.

Some proteins like therapeutic protein formulations require different components e.g. buffers, carrier proteins, to stay in their native structure. These extra components can compete for oxidation with hydroxyl radicals during the radical burst and scavenge them. This will ultimately lead to a decrease in oxidation footprint of the protein being studied to a different extent depending on their scavenging property and lower the apparent rate of oxidation. A reporter can be added to the solution of the formulations with and without the protein being studied, to correct for these components, and a concentration of radical is chosen such that the oxidation of the reporter is the same in both the solutions. Then, different formulations are compared for their scavenging properties (127).

The amount of labeling by hydroxyl radicals is dependent on the total concentration of protein and hydroxyl radicals. This becomes especially important in comparative studies like studying protein interactions or comparing proteins of different sizes, as it will lead to ambiguous results. Say protein A is 10 kDa and protein B is 20 kDa, and their 10 M solutions are prepared. If they are exposed to same concentration of hydroxyl radicals individually or in a solution with both the proteins, the solution with single protein will have higher level of oxidation, and the one with both the proteins will have lower oxidation level, not attributed to their shielding, but because of lower concentration of radicals available to cause the same level of oxidation. Similarly, if exposed to same radical concentration, protein A will have more oxidation than protein B, due to the available amount of radicals.

Hydroxyl radical protein footprinting along with mass spectrometry has developed into a powerful method to study structures and interactions of protein structure, protein-protein, and protein-ligand interaction interfaces. There have been rapid advancements in the field resulting in more sophisticated experiments. The understanding of the chemistry behind oxidation is continuously growing. Advances are also being made to make mass spectrometry techniques which are ever more sensitive to oxidation products that are difficult to detect. Other radicals are being tested for their application in protein footprinting. Further growth and understanding in this field can make it a high throughput technique.

**Computational Methods**

*Monte Carlo (MC) simulations:* Crystal structure of the globular protein Bovine Pancreatic Trypsin Inhibitor (BPTI, PDBID 3CI7) was used to perform *in vacuo* replica exchange MC simulations with ten temperatures in parallel. Initially, all waters of crystallization, as well as all the sulphate ions, were removed from the PDB structure. An all-atom force field called ECEPP/3 (128) was employed, using a simulation package called Simple Molecular Mechanics for Proteins (SMMP) (129). The package was modified to include $RMSD_{SASA}$ (equation 5.1) as a restraint at every step. The solvent accessibilities of non-polar and aromatic residues of the crystal structure were used as $SASA_{ref}$, calculated using Double Cubic Lattice (DCL) algorithm (130). The simulations were performed in two stages. The first stage employed a simulation of 100000 steps and the temperatures that were selected as a geometric progression between 270 K and 700 K. This stage was treated as an equilibration step. The temperatures in the next stage were selected using the feedback loop optimization (Table 5.1) (50) and the simulation was 200000 steps long.

$$RMSD_{SASA} = K\sqrt{\frac{\sum_n (SASA_{current} - SASA_{ref})^2}{n}} \qquad [5.1]$$

**Table 5.1.** Temperatures used for Parallel Tempering simulations.

| | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
|---|---|---|---|---|---|---|---|---|---|---|
| First | 700 | 578.56 | 525.99 | 478.19 | 434.74 | 395.24 | 359.32 | 326.67 | 296.99 | 270 |
| Second | 700 | 572.66 | 507.73 | 464.24 | 424.37 | 387.84 | 354.38 | 323.73 | 295.68 | 270 |

*Preparation of protein structures for MD:* Crystal structures of the globular proteins hen egg-white lysozyme (PDB ID: 2LYZ) (131) and horse heart myoglobin (PDB ID: 1YMB) (132) were used to test $RMSD_{SASA}$ as a score. Initially, all waters of crystallization were removed from both PDB structures, along with any ions, such as the sulfate molecule present in 1YMB. The heme group in myoglobin was retained, as it is required for protein stability. All histidine residues were considered neutral and protonated only at the epsilon nitrogen. The N- and C-terminal residues were capped with acetyl (ACE) and *N*-methylamine (NME) groups, respectively. The proteins were minimized before solvating them in TIP3P water molecules, with a buffer size of 10 Å, resulting in the addition of 11299 and 9711 water molecules in the case of lysozyme and myoglobin, respectively. Counter ions (8 $Cl^-$ ions) were added to neutralize the charge in lysozyme, using the tLEAP module of AMBER, no ions were required in the case of myoglobin.

*Minimization:* Energy minimization was performed with the SANDER module of AMBER12 (133) with ff12SB protein force field parameters (134). Prior to solvation, the proteins were subjected to 3000 steps of steepest descent and 2000 steps conjugate gradient minimization, *in vacuo* to relieve any steric collisions. After solvation, the energy of the water and ions was minimized while keeping all protein atoms restrained (500 kcal/mol-$Å^2$). The energy of the

entire system was subjected to 5000 steps of steepest descent minimization, followed by 20000 steps of conjugate gradient minimization.

*MD Simulation*. MD simulations of the energy-minimized systems were performed with the pmemd.cuda version of AMBER12 (81). All the bonds involving hydrogen atoms were constrained using the SHAKE algorithm (42), enabling an integration time step of 2 fs. Long-range electrostatic interactions, beyond a cut-off set to 10 Å, were treated with the Particle-Mesh Ewald algorithm (82). The systems were heated from 5 K to 300 K over a span of 50 ps, under NVT conditions employing the Langevin thermostat (135). The simulations were then continued for 30 ns under NPT conditions with weak restraints on the Cα atoms of the protein (10 kcal/mol-$Å^2$).

*Unfolding simulation*: To generate an ensemble of partially-unfolded structures, the energy-minimized structure of unsolvated Lysozyme was heated during an MD simulation from 5 K to 1000 K over 10 ns *in vacuo*. Snapshots were extracted from the simulation every ps. Over the course of the simulation, the all-atom RMSD of the conformations (with respect to the crystal structure) increased from 1.2 Å to 22.2 Å.

*SASA calculations*: The per-residue solvent accessible surface area (SASA) was computed with the NACCESS program (83). Average SASA values (<SASA>) were computed from a total of 1000 snapshots extracted at 30 ps intervals from the solvated MD simulations.

*SASA RMSD calculations*: The $RMSD_{SASA}$ values were calculated using equation 1, where $SASA_{current}$ is the per-residue SASA value for each residue in the model (obtained from MD or from homology modeling) and $SASA_{ref}$ is the SASA value for the same residue in the reference structure (either computed from the crystal structure or estimated from experimental HRPF data), and n is the total number of residues with SASA values.

*Homology models:* Homology models for Lysozyme were generated using the SWISS-MODEL homology modeling server (swissmodel.expasy.org) (76) for multiple template PDB structures with sequence identities that varied from 99% to as low as 37% with respect to Lysozyme (Table 1). Templates were selected that had at least 90% sequence coverage to ensure plausible fold structures. The all-atom RMSD values (relative to Lysozyme, PDB ID: 2LYZ) ranged from as 1.2 to 4.6 Å.

**Table 5.2.** Details of the homology models generated using SWISS MODEL.

| PDB ID | Sequence Coverage | Sequence Identity | All-Atom RMSD (Å) |
|---|---|---|---|
| 1LZE | 1 | 99.2 | 1.2 |
| 2GV0 | 1 | 69 | 1.6 |
| 2BQJ | 0.98 | 59.8 | 1.8 |
| 2Z2E | 0.99 | 50.8 | 2.4 |
| 4L41 | 0.95 | 36.9 | 3.1 |
| 3CB7 | 0.93 | 37.5 | 3.8 |
| 1GD6 | 0.92 | 42.9 | 4.2 |
| 1IIZ | 0.92 | 38.7 | 4.3 |
| 2RSC | 0.92 | 42.9 | 4.6 |

## Results and Discussion

*$RMSD_{SASA}$ as restraint:* This study has focused on testing and refining the process of including restraints based on the solvent accessible surface area (SASA) during Replica Exchange Monte Carlo simulation of proteins, using the package Simple Molecular Mechanics for Proteins (SMMP). To test the effect of the restraints, four simulations were performed with different restraint weights i.e. k in equation 5.1 (0, 1, 5 and 10 kcal/mol Å), starting from a partially

unfolded structure of the globular protein Bovine Pancreatic Trypsin Inhibitor (BPTI), generated by heating it at an increasing temperature from 5 K to 1000 K over 10 ns of gas phase molecular dynamics simulation. $SASA_{ref}$ values for BPTI were computed from the crystal structure (PDBID 3CI7), using the Double Cubic Lattice (DCL) algorithm. First, simulations were performed for 100,000 steps for equilibration for ten replicas at different temperatures (270 K to 700 K) chosen using geometric progression, with the restraints applied. Then, based on the data from the initial equilibration, new temperatures were chosen using feedback optimization, and production simulations were performed for 200,000 steps, with the restraints applied.

The results in Figure 5.1, suggest that a SASA restraint penalty of 5 kcal/mol narrowed the spread of backbone RMSD values, with a small (approximately 0.4 Å) reduction in the lowest RMSD values. Higher or lower restraint weights (10 or 1 kcal/mol Å) encouraged sampling of conformations that had significantly higher RMSD values, relative to the simulation performed with no restraints. When the restraint weight is high, it reduces the SASA penalty function, as expected; however, it does so at the expense of the structural correctness, as indicated by an increase on average in the backbone RMSD (figure 5.2, 5.3). These results indicate that SMMP can sample structures with biologically incorrect folds that nevertheless yield more accurate $RMSD_{SASA}$ values when the restraint weight is set at high values. The correlation between $RMSD_{SASA}$ and backbone RMSD in SMMP models support this observation. In the case of a weak restraint weight, a degradation of the structural correctness was also observed, for reasons yet to be determined.

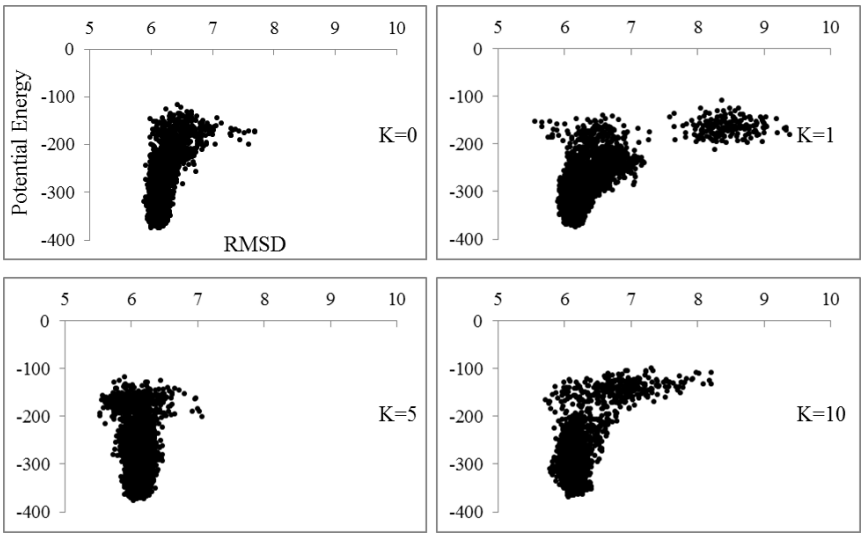



**Figure 5.1.** RMSD vs. potential energy for BPTI models generated by SMMP with no SASA restraint (k=0, top left) vs. SASA restraints set from (k=1, top right) 1 kcal/mol, (k=5, bottom left) 5 kcal/mol, or (k=10, bottom right) 10 kcal/mol.

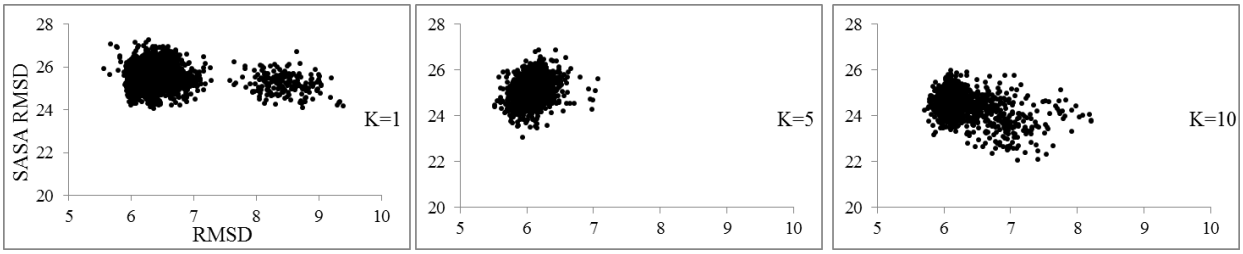



**Figure 5.2.** RMSD vs. SASA RMSD for BPTI models generated by SMMP with SASA restraints set from (k=1, left) 1 kcal/mol, (k=5, middle) 5 kcal/mol, or (k=10, right) 10 kcal/mol.

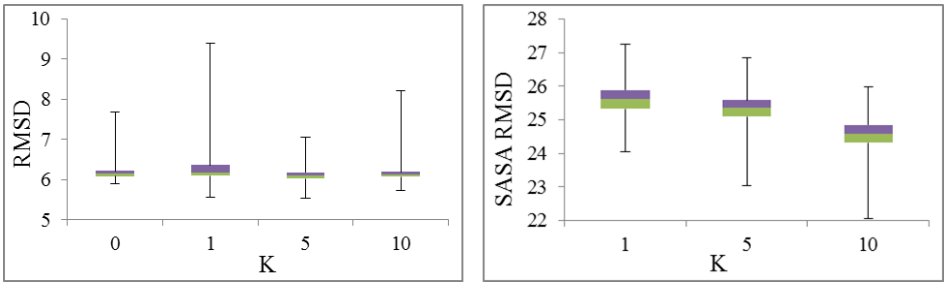



**Figure 5.3.** Box-plots showing variations in structural RMSD (left) and SASA RMSD (right) versus restraint weight (K). Despite the large standard deviations, the differences in the means are statistically significantly (as demonstrated by Kruskal–Wallis test for non-normal

distributions). The green box shows the variations from the 25$^{th}$ percentile to the median, while the purple box shows the variations from the 75$^{th}$ percentile to the median.

However, with a restraint weight of 5 kcal/mol, models with high backbone RMSDs were not observed compared to the simulations performed with either lower or higher restraint. At present, no structures were obtained with RMSD values below approximately 5.5 Å, suggesting that more sampling is required to conclude whether SASA restraints are effective at pushing low resolution (partially unfolded) protein towards a high resolution (crystal) structure. It should be noted that the restraint penalty is not applied in a pairwise manner (between residues), but is a global property of the protein structure. This feature likely results in many different conformations having similar SASA penalty values. Thus, the SASA penalty values may conflict with the force field potential energy, leading to a complex potential energy surface. This behavior likely impairs simulation convergence.

*$RMSD_{SASA}$ score*: An alternative approach was examined that uses $RMSD_{SASA}$ values as a scoring function, to assess the quality of a protein structure, relative to experimental SASA values, rather than attempting to employ the SASA values as restraints during a simulation. The experimental SASA values were generated using two different methods of normalizations to minimize the influence of the intrinsic reactivity of different amino acids. In the first case, because the amino acid types (especially less reactive amino acids to radicals) have a large influence on the accuracy of the prediction model for estimating SASA, only amino acids with certain reactivity to radicals (Trp, Tyr, Phe, His, Leu, and Ile) were used to build up the SASA prediction model. The sequence context, meaning the surrounding amino acids, from a natively folded protein can also play a prominent effect on the intrinsic reactivity of an amino acid. Therefore, in the second

case, the oxidation of heat denatured protein sample was used to normalize the predicted SASA values. In both the cases, the SASA prediction models were built using myoglobin, and these models were used to predict SASA of lysozyme.

To determine whether $RMSD_{SASA}$ scores can be used to characterize the quality of a modeled protein 3D structure, an unfolding simulation of the globular protein Lysozyme was performed by heating it at an increasing temperature from 5 K to 1000 K over 20 ns of gas phase molecular dynamics simulation. The simulation was initiated from the crystal structure (PDBID: 2lyz.pdb). Snapshots were extracted from the simulation every ps. The $RMSD_{SASA}$ values were calculated for every snapshot, with $SASA_{ref}$ values derived from the crystal structure, and separately, with $SASA_{ref}$ values estimated from experimental HRPF data, using the two SASA prediction models. As the protein unfolds, the $RMSD_{SASA}$ increases in both the cases (Figure 5.4).

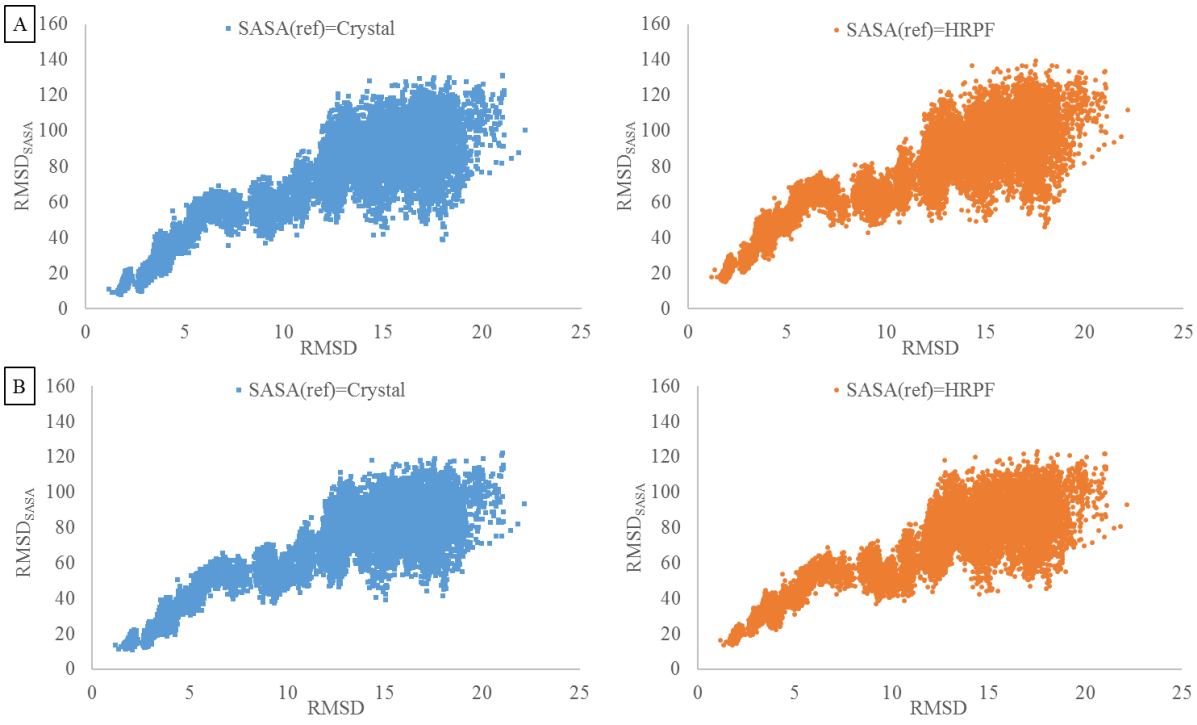



**Figure 5.4.** $RMSD_{SASA}$ calculated on structures obtained from an unfolding MD simulation. $RMSD_{SASA}$ calculated using the $SASA_{fold}$ from crystal structure in blue (Crystal) and $SASA_{fold}$

estimated from HRPF experiments in orange. A. The values calculated using the SASA prediction model that employs free amino acid oxidation for normalization. B. The values calculated using the SASA prediction model that employs oxidation of unfolded protein for normalization.

The similarity of the $RMSD_{SASA}$ values computed using the crystallographic SASA and HRPF SASA references indicate that the SASA values from the HRPF experiments are in good agreement with those from the crystallographic data, especially for prediction model normalized with the unfolded protein ($R^2$=0.7), versus prediction model normalized with free amino-acids ($R^2$=0.6). This observation is confirmed in Figure 5.5. The HRPF-derived SASA values, therefore, appear to be suitable to be used as reference values for $RMSD_{SASA}$ calculations. Using the first prediction model, a minimum $RMSD_{SASA}$ of approximately 8 $Å^2$ (crystal) or 15 $Å^2$ (HRPF) is exemplary of a well-folded protein. While with the second prediction model, the values change to 11 $Å^2$ and 16 $Å^2$.

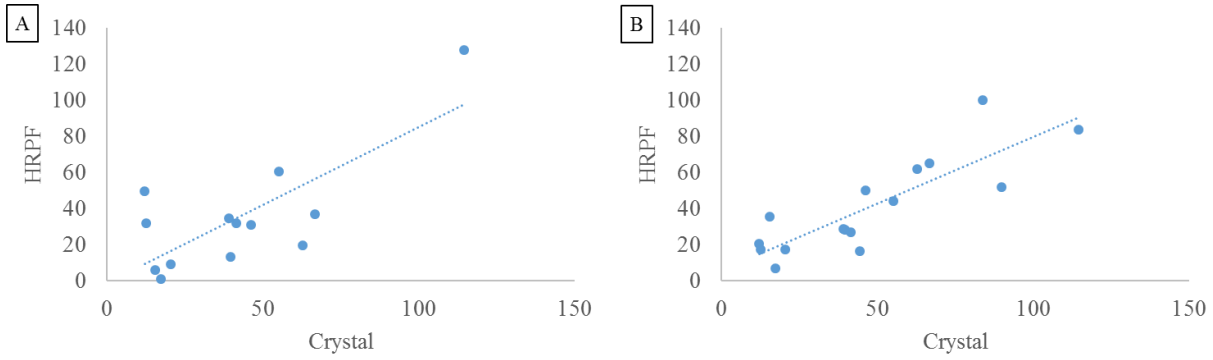



**Figure 5.5.** SASA computed using crystal structure vs. SASA estimated from HRPF experiment. A. The values calculated using the SASA prediction model that employs free amino acid oxidation for normalization. B. The values calculated using the SASA prediction model that employs oxidation of unfolded protein for normalization.

$RMSD_{SASA}$ values calculated for snapshots extracted from a simulation of folded lysozyme at 300K remain stable over the course of the simulation (Figure 5.6), indicating there is no variation in the quality of the conformations generated during this MD simulation. Note, the protein simulation is also stable as indicated by the average Cα RMSD (0.6 Å).

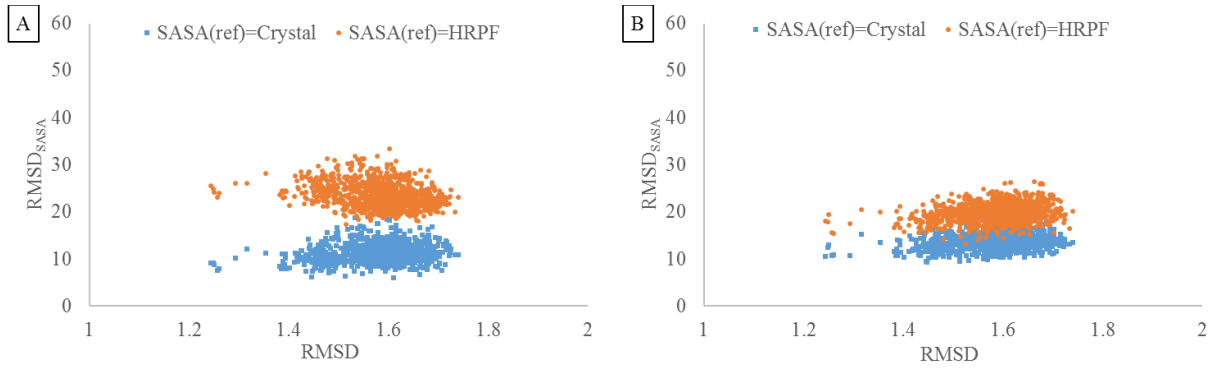



**Figure 5.6.** $RMSD_{SASA}$ calculated on structures obtained from an MD simulation of the crystal structure. A. The values calculated using the SASA prediction model that employs free amino acid oxidation for normalization. B. The values calculated using the SASA prediction model that employs oxidation of unfolded protein for normalization.

Subsequently, the $RMSD_{SASA}$ was used to rank the structures of lysozyme generated using homology modeling. Homology models were generated using the swiss-model server (Table 5.2). The all-atom RMSD of these structures with respect to the crystal structure varied from 1.2 to 4.6 Å. The $RMSD_{SASA}$ scores show a direct relationship with the increasing RMSD, with $R^2$ values greater than 0.83 using either of the prediction models (Figure 5.7). This is especially true for structures with RMSD greater than 2 Å, indicating the limitation of $RMSD_{SASA}$ to differentiate between structures close to the native structure.

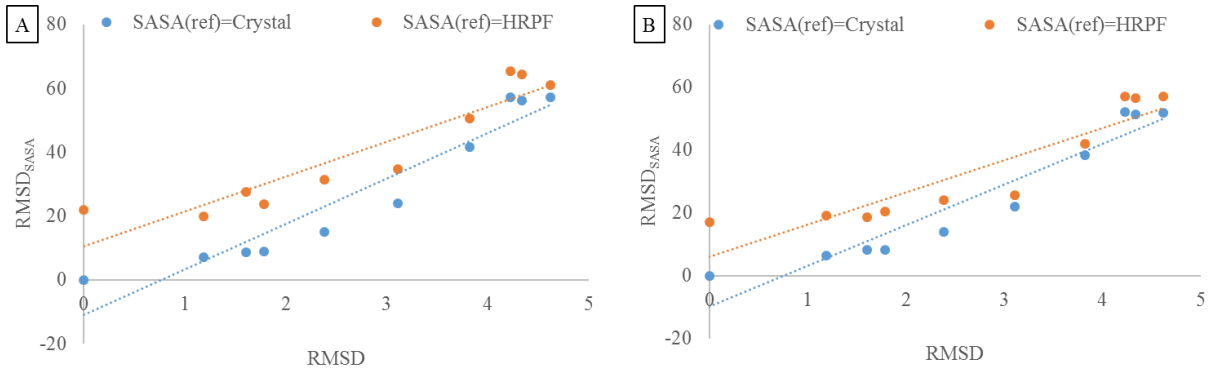



**Figure 5.7.** $RMSD_{SASA}$ used as a scoring function to rank the structure generated using homology modeling. A. The values calculated using the SASA prediction model that employs free amino acid oxidation for normalization. B. The values calculated using the SASA prediction model that employs oxidation of unfolded protein for normalization.

## Conclusion

The MC simulations show the potential use of $RMSD_{SASA}$ values as restraints, however, their applicability is hampered by the computational cost of determining SASA at every step. Further, unlike distance or angle values, per-residue SASA values are global properties, which means, to calculate the SASA for one residue, the whole molecule needs to be analyzed. The reliability of SASA as a restraint is dependent on the number of data points available. The number of conformations that can satisfy the constraint can be expected to increase with a decrease in the size of the data set.

$RMSD_{SASA}$ correlates with the structural RMSD of the protein (relative to the crystal structure) in an unfolding simulation of lysozyme. Two different prediction models derived from the relationship between oxidation and SASA of myoglobin were used to get the reference values of SASA. This suggests the capability of $RMSD_{SASA}$ to rank the conformations based on their proximity to the native structure. The minimum $RMSD_{SASA}$ value for a stably folded protein in

the MD analysis of the folded lysozyme appears to be approximately 8 $Å^2$ and 15 $Å^2$ for the two different prediction models, which remains stable throughout the course of this simulation. Similar results were obtained by scoring homology models of lysozyme, where we see a relationship between RMSD and $RMSD_{SASA}$. Therefore, $RMSD_{SASA}$ can be used to quantify the quality of protein models.

Notably, each side chain has been found to require a minimum level of exposure before it will be oxidized (62). By considering this minimal exposure requirement, with the known side chain reactivities, potential reporter groups may be identified. Knowledge of the expected reporter groups can be used to provide an estimate of the surface oxidation levels and is therefore of significance to the study of protein-protein and protein-ligand interactions. Although not all amino acid side chains react under the conditions presented here, the level of surface coverage, in terms of those residues that were exposed to solvent and that were not inert to oxidation was approximately 70%. This level of coverage is far greater than would be achieved using traditional chemical derivatization methods, such as biotinylation or acetylation, wherein only a few residues can act as reporter groups. Equally significant to good coverage is the ability to quantify the level of oxidation and relate that directly to per-residue <SASA> values. This ability significantly elevates the level of footprint resolution, which is key to the practical application of this method in characterizing protein complexes. Such quantification facilitates the identification of occluded surfaces and should provide a powerful tool for determining the 3D structures of complexes that are not amenable to analysis by traditional experimental structural methods.

# CHAPTER 6

# QUANTIFYING FUNCTIONAL GROUP CONTRIBUTIONS TO UNDERSTANDING PROTEIN-CARBOHYDRATE AFFINITY

## Introduction

The recognition of glycans present on cell surfaces as glycoconjugates lies at the heart of a number of biological processes in animals, plants, and microorganisms (136). Non-covalent glycan-protein interactions are involved in cellular adhesion, innate immunity, bacterial and viral infection, as well as plant defense mechanisms and other processes (137-142). Glycan-binding proteins (GBPs) such as lectins, adhesins, toxins, antibodies, carbohydrate-binding modules, are often multimers that possess the ability to crosslink cells, which is essential for cell signaling (143) and the disruption of recognition can lead to conditions such as delay in muscle fiber development. The multimeric structure of most carbohydrate-binding proteins serves also to enhance the apparent affinity of the binding processes through avidity effects (144). The affinity of monomeric carbohydrate protein interactions is typically weaker than uM, and yet the specificity appears to arise primarily from the structure of monomeric complexes (145).

Much of our understanding of carbohydrate recognition has come from crystallographic studies of plant lectins, because these proteins are often relatively stable, crystallize readily, and have a wide range of receptor specificities. More recently, glycan array screening has been widely applied to define specificity. However, the specificity of lectins (146) and anti-carbohydrate antibodies (147) can appear complex. Nevertheless, plant lectins have found widespread use as affinity reagents in the separation and characterization of oligosaccharides, and glycoconjugates

(148), and are often employed in staining and histochemistry of cells and tissues (15, 16, 149). For example, the legume lectin from *Erythrina cristagalli* (ECL) is widely used as a reagent for the detection of terminal galactopyranose (Gal) residues in glycans (its canonical specificity is for Gal), yet it also binds to *N*-acetylgalactosamine (GalNAc) and fucosylated Gal (Fucα1-2Gal). Although its function in the legume is unknown, understanding the complex specificity of lectins, such as ECL, is fundamental to the rational design of diagnostic and therapeutic agents that target specific glycans (150).

Numerous experimental methods have been used to quantify the affinity of GBP-carbohydrate interactions (including isothermal titration calorimetry (ITC), NMR spectroscopy, microscale thermophoresis (MST), biolayer interferometry (BLI), surface plasmon resonance (SPR), frontal affinity chromatography (FAC), and ELISA-based assays). Data from different experimental techniques can result in conflicting definitions of specificity, depending on the sensitivity of the method and on the presence or absence of avidity effects. This is particularly clear in the case of weak interactions, which may be observed by NMR (151) or MST (152), but not by glycan array screening (153, 154). Given the widespread use of glycan array screening, it has become the *de facto* method for defining the specificity of GBPs, and yet often requires amplification of the signal through multimerization of the protein analyte (155). Although glycan array screening is a high throughput method capable of screening hundreds of glycans, it is often unable to detect weak monomeric interactions and does not provide structural insights into the origin of the observed specificity and cross-reactivity. While site directed mutagenesis of the protein (156) or chemical modification of ligand (157) can be used to probe the mode of binding in the past, protein crystallography is by far the most widely used method to define the binding mode. However, crystallography often employs high ratios of ligand to a protein, and the ligand is

typically only a small fragment of the intact glycan, leading to questions as to the biological relevance of the co-complex (158). Given the high flexibility of glycans, it is not surprising then that these complex macromolecules are resistant to crystallization, making it difficult to determine the molecular structures for all but the simplest glycan fragments. Thus, experimental techniques alone can prove to be insufficient to understand the mechanism of low-affinity carbohydrate recognition. However, when these techniques are coupled with computational analyses, it can lead to an improved grasp of the underlying reasons behind the specificity of carbohydrate-protein interactions.

From a structural perspective, binding to the protein requires the carbohydrate to form interactions (hydrogen bonds, van der Waals contacts, hydrophobic contacts) that are specific in terms of geometry and charge complementarity. Discrimination between potential binders depends on differences in affinity, which depends on the strengths of individual interatomic (or inter-functional group) interactions. However, it is challenging to quantify these interactions experimentally, as any physical alteration to the protein (such as a point mutation) or to the ligand (such as a chemical modification) could perturb more than the local interaction, aside from the significant effort that may be required. Thus, an opportunity exists to exploit computational methods to estimate the energetic contributions made by individual interacting groups. There are a number of theoretical methods capable of estimating receptor-ligand affinities with varying levels of accuracy and computational cost (56), including thermodynamic integration (TI), free energy perturbation (FEP), and MM-PB/GBSA (molecular mechanics-Poisson–Boltzmann/Generalized Born surface area). While equilibrium methods such as TI and FEP are generally more accurate than end-point methods like MM-GBSA, achieving sufficient conformational sampling is only practical for TI/FEP calculations if the ligands differ only

slightly in structure; calculating the binding energy difference between ligands that differ by one or more monosaccharide is currently beyond the capability of TI/FEP. In contrast, MM-PB/GBSA methods are less size-limited, and by default are therefore the methods most widely applied for predicting the energetics of carbohydrate-protein complexes. MM-GBSA is known to not be able to reproduce experimental binding free energies, but it still shows a correlation with the experiments (159).

Here we perform molecular dynamics (MD) simulations of complexes of ECL with six ligands: lactose (160, 161) (Galβ1-4Glcβ, Lac, **1**), epi-lactose (Galβ1-4Manβ, Epilac, **2**), *N*-acetyllactosamine (Galβ1-4GlcNAcβ, LacNAc, **3**), *N,N*-diacetyllactosamine (GalNAcβ1-4GlcNAcβ , LacDiNAc, **4**), fucosylated lactose (Fucα1-2Galβ1-4Glcβ, FucLac, **5**) (160), and fucosylated *N*-acetyllactosamine (Fucα1-2Galβ1-4GlcNAcβ, FuclacNAc, blood group H trisaccharide, **6**). The MM-GBSA method is then used to compute absolute affinities, as well as inter-residue and inter-group interaction energies. This approach enables us to identify key components of the ligand that are responsible for the observed experimental specificity and to quantify their relative contributions. In addition, we report a novel crystal structure of ECL in complex with N-acetyllactosamine, and new experimental affinities for seven di- or trisaccharides. The results from the present analysis provide an explanation for the observed specificity of ECL and lead to insights that could be used in general to engineer lectins with new ligand specificities (162). From a theoretical perspective, the results also help to define the accuracy limitations of the computational methods.

**Materials and Methods**

*Crystallization.* A sample of ECL was dissolved in 100 mM NaCl, 20 mM HEPES pH 7.5, 0.1 mM $CaCl_2$ and 0.1 mM $MnCl_2$ to a concentration of ~ 7 mg/mL. About 1 hour prior to

crystallization, the solution of ECL was combined with the aqueous solution, 0.25 mM, of the particular ligand at a molar ratio of 1:10 (ECL:ligand). Crystals were grown by the vapor diffusion at 20-22 $^{0}$C using sitting drop method. For ECL with N-Acetyl-D-Lactosamine complex screening with QIAGEN's the JCSG Core I Suite resulted in diffraction quality crystals of pyramidal shape from several conditions: #10, 12, 13, 20, 22, and 31. The best crystals were obtained from either 0.2 M Calcium acetate hydrate, or Potassium Sodium tartrate and 20 % PEG 3350, corresponding conditions are # 20 and 22. The crystals grew 1 µL sitting drop Intelli-Plates. Co-crystals of ECL with epi-lactose were obtained from 10 µL drops in microbridges using well solutions containing 0.2 M Calcium Acetate, 0.1 M HEPES pH 7.5, 14-16 % PEG 3350.

*Data collection.* For both complexes X-ray crystallographic data were collected from frozen crystals at 100K. Prior to data collection crystals were placed in a cryoprotectant solution composed of 75% well solution and 25% glycerol and then flash cooled by immersion in liquid nitrogen. For ECL-N-Acetyl-D-Lactosamine complex diffraction data were collected using an ADSC Quantum 315r detector at the Advanced Photon Source (APS) on the ID19 beamline SBC-CAT to 1.9 Å resolution. For ECL-epi-lactose co-crystal crystallographic data were collected to 2.2 Å using a Rigaku HomeFlux system, equipped with a MicroMax-007 HF generator, Osmic VariMax optics, and an RAXIS-IV++ image-plate detector. X-ray diffraction data were collected, integrated and scaled using HKL3000 software suite (163). The structure was solved by molecular replacement using CCP4 suite (164). The structure of the binary complex of ECL with lactose (PDB ID 1UZY) (161) was used as a starting model with all waters, ligands including the *N*-linked glycosylated saccharide and metal ions removed. Refinement was completed using the *phenix.refine* program in the *PHENIX* (165) suite and the

resulting structure analyzed with molprobity (166). The structures were built and manipulated with program *Coot* (167), whereas the figures were generated using the *PyMol* molecular graphics software (v.1.5.0.3; Schrödinger LLC). A summary of the crystallographic data and refinement is given in Table S6.1.

*BLI binding experiment:* ECL (Cat#: L-1140, Vector Lab, Burlingame, CA, USA), **3** (Cat# A7791, Sigma-Aldrich, St. Louis, MO, USA), **1** (Cat#: 61339, Sigma-Aldrich, St. Louis, MO, USA), **2** (epi-Lac, Cat#: G0886, Sigma-Aldrich, St. Louis, MO, USA), **5** (2'FucLac, Cat#:OF06739, Carbosynth Limited, Berkshire, UK) and **7** (Cellobiose, Cat# 22150, Sigma-Aldrich, St. Louis, MO, USA) were purchased from their commercial resources. Biotinylated glycan Galβ1-4GlcNAcβ-$OCH_2CH_2CH_2NH$-biotin (LacNAc-biotin) was received as a gift from Dr. Nicolai Bovin. ECL was weighted and dissolved in the ECL buffer: 10 mM HEPES, 15 mM NaCl, 0.1 mM $CaCl_2$, and 0.1 mM $MnCl_2$ buffered at pH7.4, at 25°C.

*Protein BLI direct binding assay ($K_{D,surface}$):* Ligand LacNAc-biotin was loaded onto streptavidin biosensors (SA, Cat#: 18-5019, Pall ForteBio Corp., Menlo Park, CA, USA) at 1 μM for 1800s. Then the loaded LacNAc biosensors were dipped into 0.1μM EZ-link$^{TM}$ Hydrazide-Biocytin (biocytin, Cat#: 28020, Thermo Scientific, Rockford, IL, USA) for blocking the possible unoccupied biotin-SA binding sites for 1800s. The immobilization of ligand onto SA biosensors resulted in ~0.3nm as loading signal under this condition. ECL direct binding $K_D$ (LacNAc biosensor surface KD) was measured using a BioLayer Interferometer (BLI) Octet Red 96 system (Pall ForteBio Corp., Menlo Park, CA, USA) and data acquired using ForteBio Data Acquisition 8.2 software (Pall ForteBio Corp., Menlo Park, CA, USA). The protein direct binding experiment was performed at 600s for association and 1800s for dissociation in ECL buffer. ECL was prepared in two-fold serial dilution in ECL buffer from 0~50 μM, in the

replicates of three. Surface $K_D$ ($K_{D,surface\ LacNAc\ biosensor}$) was then calculated by ForteBio Data Analysis 8.2 software (Pall ForteBio Corp., Menlo Park, CA, USA) and Microsoft Office Excel 2011 (Microsoft, USA). Surface $K_D$ ($K_{D,surface\ LacNAc\ biosensor}$) was determined by 1:1 binding model from both steady state analysis and Scatchard plot (Figure S6.1) and resulted in 0.92 (STDEV: 0.02) µM of triplicates.

*Protein BLI inhibition assay ($IC_{50}$):* ECL protein was prepared at 2µM in ECL buffer in a large volume for protein inhibition assay. Eight compounds were tested in the inhibition assay including six inhibitors: **1**, **2**, **3**, **5**, **5**-$N_3$ (FucLac-$N_3$), **6**-$N_3$ (FucLacNAc-$N_3$), and a non-ECL binder **7**. All the compounds were prepared in two-fold serial dilution in ECL buffer from 0,1.25, 2.5, 5, 10, 20, 40, and 80mM. 100µL of 2µM ECL, 20µL of prepared inhibitor/non-binder at its concentration, and 80µL of ECL buffer were mixed and incubated at room temperature for 1hour. ECL inhibition assay was performed on Octet Red 96 at baseline time 120s, association time 600s, and dissociation time 1800s at shaker speed 1000RPM at room temperature, in replicates of three. $IC_{50}$ was calculated by using three-parameter dose-response inhibition model in GraphPad Prism 7 (GraphPad, La Jolla, CA, USA). The compounds **5** and **5**-$N_3$ result in similar $IC_{50}$ values, therefore, only the final values for **5** are reported (Table S6.2, Figure S6.2). this shows that the azide group attached to the compound **5**-$N_3$ does not affect binding. Hence, it can be assumed that the binding of **6** and **6**-$N_3$ will be similar as well.

*Solution $K_D$ conversion:* $IC_{50}$ is related to the equilibrium dissociation constants for the inhibitor and LacNAc biosensor competing binding to the ECL. When $IC_{50}$ of inhibitor and $K_D$ of LacNAc biosensor to ECL ($K_{D,\ surface\ LacNAc\ biosensor}$) were known, solution $K_D$ of inhibitor can be calculated from the equation: $IC_{50} = K_{D,solution\ inhibitor}\ (\ 1\ +\ [ECL]/K_{D,surface\ LacNAc\ biosensor}\ )$.

*Molecular Dynamics*: Crystal structures of ECL in complex with **1**, **2**, **3** and **5**, along with the 3D models of **4** and **6** in complex with ECL were used for performing MD simulations. GLYCAM-Web server (www.glycam.org) was used to generate 3D structures of **4** and **6**, which were then superimposed on **3** and **5** respectively to get the complex structures. All the waters of crystallization and ions were retained, while the N-glycosylated sugar at N113 was removed from the crystal structures. The missing protons were added to all the structures in the presence of crystal waters using a tool provided by AMBERTOOLS called reduce. These structures were then minimized *in vacuo* to get rid of steric clashes if present by steepest descent minimization for 5000 steps followed by 20000 steps of conjugate gradient minimization. The charges in the systems were neutralized by adding counter ions (6 $Na^+$ ions) and truncated octahedral solvent box of pre-equilibrated TIP3P explicit water molecules was employed to solvate them using the tLEAP module provided by the AMBER suite of programs. The water molecules are allowed to equilibrate around the solute by keeping the solute atoms restrained (500 kcal/mol-$Å^2$) while performing a steepest descent minimization for 5000 steps and conjugate gradient minimization for 20000 steps. The next stage of minimization was performed without any restraints using the same steps involved in the previous stage. They were then heated from 5 K to 300 K over a span of 50ps, under NVT conditions followed by a 1ns equilibration under NPT conditions with weak restraints on the Cα atoms in the protein backbone (10 kcal/mol-$Å^2$) with pmemd.cuda version of AMBER14. The MD simulations were performed under the same conditions as equilibration for 100 ns.

*Binding affinity and entropy calculations*: Five different parametrizations ($GB^{HCT}$, igb=1; $GB_1^{OBC}$, igb=2; $GB_2^{OBC}$, igb=5; GB$n_1$, igb=7; GB$n_2$, igb=8) of Molecular Mechanics-Generalized Born Solvent Accessible Surface Area (MM/GBSA), and Molecular Mechanics-

Poisson Boltzmann Solvent Accessible Surface Area (MM/PBSA) using mbondi radii were employed to estimate binding affinities of all the six complexes. These calculations were carried out on 30,000 snapshots extracted evenly from 30ns of MD simulation using a single trajectory method with the MMPBSA.py.MPI module of AMBER.

Quasi harmonic (QH) entropies were extrapolated to an infinitely long simulation period by fitting a linear regression curve to entropy as a function of inverse simulation period (168) (Figure 6.1). Three different sets of snapshots were used from a 100ns simulation to get three different extrapolated entropies, which were then averaged. The cpptraj module provides a functionality to calculate QH entropy of a system. To get the net entropy, protein and ligand entropies were subtracted from the entropy of the complex.

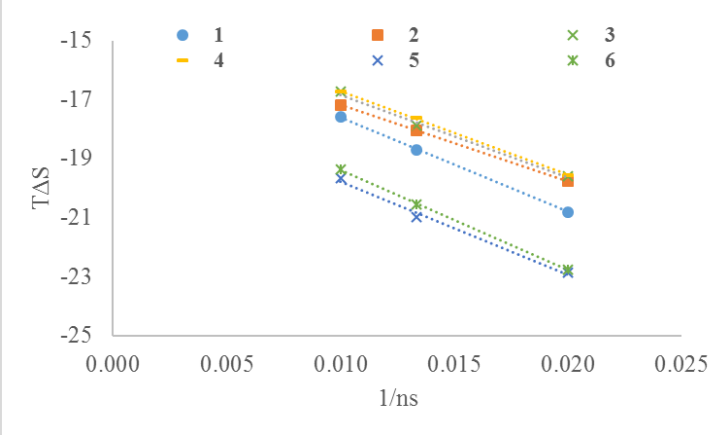



**Figure 6.1.** Extrapolation of quasi-harmonic entropy to infinite time for all the ligands.

Normal mode (NM) entropy calculations were performed using the MMPBSA.py.MPI module. As normal mode analysis is computationally very costly, it was performed using 100 snapshots from the simulation (169). A single calculation using 250 snapshots from a simulation of ECL in complex with **1** results in entropy values (-19.2 kcal/mol) comparable to the entropy calculations

from 250 snapshots (-19.0 kcal/mol), suggesting that a sample of 100 snapshots is sufficient for these calculations.

**Results and Discussion**

*Specificity of ECL*: ECL is a Gal/GalNAc specific legume lectin with Galβ1-4GlcNAc as the preferred binding motif. A number of experimental studies have been performed to determine and compare the affinity of ECL for various monosaccharides and sugars. Thermodynamic studies performed here using Biolayer Interferometry compare well with reported values obtained by Isothermal Titration Calorimetry (ITC), and show that lactose (Galβ1-4Glcβ**, 1**), Epi-lactose (Galβ1-4Manβ, Epi-Lac, **2**), and fucosylated lactose (Fucα1-2Galβ1-4Glcβ, FucLac, **5**) are equivalent binders while the introduction of an N-acetyl moiety into the Glc residue enhances affinity, as in N-acetyllactosamine (Galβ1-4GlcNAcβ, LacNAc, **3**) and 2′-Fucosyl-N-acetyllactosamine (Fucα1-2Galβ1-4GlcNAcβ, FucLacNAc, Blood group H trisaccharide, **6)** (Table 6.1). Neither Cellobiose (Glcβ1-4 Galβ, **7**) nor Maltose (Glcα1-4 Galβ, **8**) show any measurable affinity for ECL. Interestingly, data from glycan array screening of ECL indicates that **1** and **5** are non-binders, while only **3**, **6**, and GalNAcβ1-4GlcNAcβ (LacDiNAc, **4**) are binders (170). The false negative binding observed in the glycan array data for **1** and **5** may indicate the relative weakness of the binding of these ligands and suggests a need for caution when employing glycan array screening to define glycan-binding specificity for low affinity ligands.

While affinity measurements can indicate which regions of the ligand may be important for binding, a detailed rationalization can best be obtained from examination of the 3D structures of the complexes. Conversely, 3D structures alone can provide at best only a qualitative guide to the impact of any given intermolecular interaction on the affinity of the ligand. Computational

simulations, employing accurate 3D structures, can permit structure-function relationships to be derived that include the critical contributions from molecular motion, solvation, and entropy. Here we report the first structures for ECL complexed with **2** and **3**, enabling MD simulations (100 ns) to be performed on ECL bound to **1**-**6**. The data from the simulations were then compared to the crystallographic data and found to reproduce the majority of the observed inter-molecular interactions over the course of the simulations. Having validated the structural accuracy of the simulated data, interaction energies were computed for each system, with the goal of quantifying not only the contributions made by each monosaccharide but also the contributions made by each interacting chemical moiety (NAc groups, OH groups, ring atoms, etc). This approach permits both an assessment of the accuracy and utility of simulations in developing glycan structure-function relationships, as well as an opportunity to define the limitations and weaknesses of the MM-GBSA energy estimation method.

**Table 6.1.** Thermodynamic parameters determined by Titration Microcalorimetry.

| | $K_D$ | $\Delta G$ (kcal/mol) | Ref |
|---|---|---|---|
| **1** | 0.32 (0.02) | -4.83 (0.04) | -4.9 (0.2)[1], -4.8 (0.0)[2] |
| **2** | 0.21 (0.01) | -5.08 (0.02) | |
| **3** | 0.08 (0.01) | -5.66 (0.04) | -5.5 (0.1)[1] |
| **5** | 0.22 (0.01) | -5.04 (0.06) | -4.8 (0.0)[2] |
| **6** | 0.032 (0.01) | -6.21 (0.14) | |

[1] Experiments performed at 27°C by Gupta et.al. (1996).
[2] Experiments performed at 25°C by Svensson et. al. (2002).

*Structural basis of ligand recognition:* All ECL crystal structures indicate that there is only one carbohydrate binding site per monomer, which is characterized by a shallow groove. All the

ligands occupy the same binding site with Gal and Glc residues residing in equivalent positions in each of the complexes. The fucosyl residue in **5**, and the *N*-acetyl group in **3** form additional hydrogen bonds and van der Waals contacts with the protein, relative to **1**. It is notable that despite the presence of presumably favorable interactions with the fucosyl residue, the affinity of **5** is not significantly different than **1**, suggesting a need to examine the interaction energies in detail. Assuming that all of the known ligands bind ECL in a similar fashion with Gal in the binding pocket, 3D models of **4** and **6** in complex with ECL were created. 3D structures for **4** and **6** were retrieved from the GLYCAM-Web server (www.glycam.org), and models for their complexes with ECL were generated by superimposing the coordinates for the ring atoms on to those present in the complex with **3**.

To examine and compare the stabilities and strengths of the interactions of each of the ligands with ECL, each complex was subjected to molecular dynamics (MD) simulation (100 ns) in the presence of explicit water, using the AMBER/GLYCAM (133, 171) force field. The ligand-protein complexes remained stable over the course of the simulations (average ligand displacement RMSD: **1** = 0.85 Å, **2** = 0.99 Å, **3** = 0.79 Å, **4** = 0.96 Å, **5** = 0.77 Å, **6** = 0.81 Å), which signified that the trajectories were equilibrated and appropriate for further analysis. Consistent with the crystal structures (Figure 6.2), each of the ligands formed stable hydrogen bonds between O3 and O4 hydroxyl group of Gal residue and D89, N133, and A218, during the simulation (Table 6.2). In **5** and **6**, the Fuc-O2 group maintained its hydrogen bond with the side chain of Asn133. A hydrogen bond between the O3 group in the terminal reducing residue (Glc, Man, GlcNAc) in **1**-**6** was also observed but found to be significantly more stable in the case of GlcNAc. Although a hydrogen bond is present between Gal-O3 and Gly107 in all the crystal structures, it was not highly occupied over the course of the simulations. Similarly, the hydrogen

bond between Fuc-O4 and Tyr108 in **5**, present in the crystal structure, only formed occasionally during the simulation.

*Quantification of molecular contributions to affinity:* The strength of these interactions was quantified by performing an MM-GBSA and MM-PBSA analysis of the MD simulations. In addition to contributions from direct interactions (van der Waals and electrostatics), the energies generated this way also include estimates of desolvation free energy and entropy. Conformational entropies were estimated using a quasi-harmonic (QH) approach, which employs a covariance analysis of the changes in atomic fluctuations that occur upon ligand binding to predict the entropy changes (172), and normal mode (NM) vibrational analysis (173), which estimates the entropic contributions for binding resulting from changes in the frequencies associated with bond stretching and angle bending.

In agreement with the experimental data, and independent of the five desolvation parameterizations evaluated, **1** and **2** were always ranked the weakest binders, and displayed essentially equivalent interaction energies (Table 6.3). All the MM-GBSA desolvation models ranked **5** amongst the best binders, in disagreement with experiment, but in all the models **6** was correctly ranked as the highest affinity ligand. In contrast, MM-PBSA desolvation model correctly ranked **5** along with **1** and **2** amongst the weakest binders, and **6** was also correctly ranked as the strongest binder. Overall none of the MM-GBSA models could correctly rank all the ligands. On the other hand, MM-PBSA model could correctly rank every ligand. As expected (174), incorporation of QH entropy reduced the magnitude of the interaction energies but did not lead to an improvement in the ranking of the relative affinities of the ligands (Table 6.4).

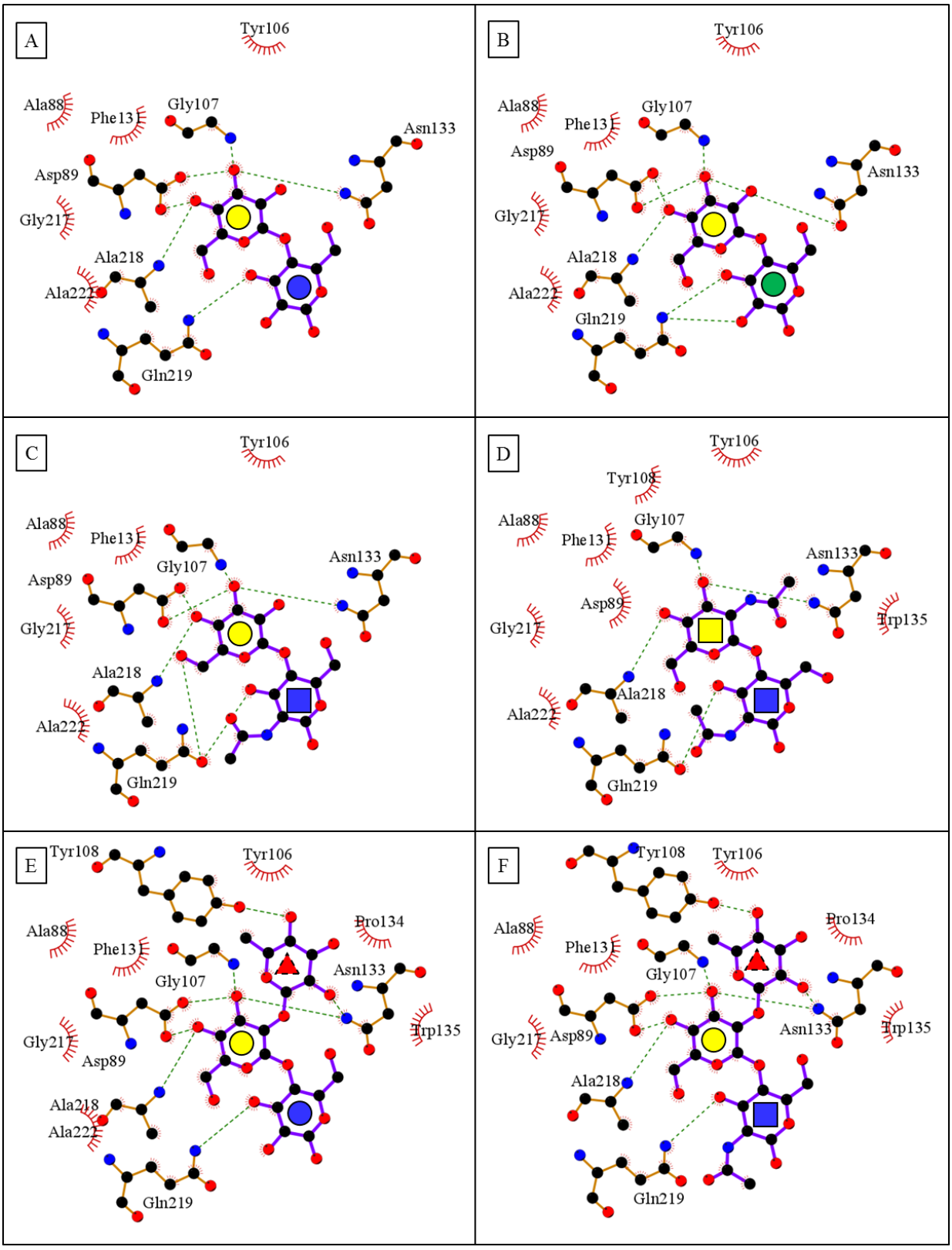

A
Tyr106
Ala88
Phe131
Gly107
Asn133
Asp89
Gly217
Ala218
Ala222
Gln219
B
Tyr106
Ala88
Phe131
Gly107
Asp89
Asn133
Gly217
Ala218
Ala222
Gln219
C
Tyr106
Ala88
Phe131
Asn133
Gly107
Asp89
Gly217
Ala218
Ala222
Gln219
D
Tyr106
Tyr108
Ala88
Gly107
Phe131
Asn133
Asp89
Trp135
Gly217
Ala218
Ala222
Gln219
E
Tyr108
Tyr106
Pro134
Ala88
Phe131
Asn133
Gly107
Trp135
Gly217
Asp89
Ala218
Ala222
Gln219
F
Tyr108
Tyr106
Pro134
Ala88
Phe131
Gly107
Gly217
Asp89
Asn133
Trp135
Ala218
Gln219

**Figure 6.2.** The contacts between the ECL protein and the ligands **1** to **6** represented from A to F.

**Table 6.2.** Hydrogen bonds present in the crystal structure and in the MD simulation.

| | | | Distance[1] | | | |
|---|---|---|---|---|---|---|
| Ligand | Protein residue | Ligand residue | Crystal | MD | Occupancy | Interaction energy[2] |
| **1** | Asp89-Oδ2 | Gal-O4 | 2.6 | 2.6 (0.1) | 1.0 | -3.7 (2.3) |
| **2** | | Gal-O3 | 2.6 | 2.8 (0.1) | 1.0 | -3.2 (2.7) |
| **3** | | Gal-O3 | 2.6 | 2.6 (0.1) | 1.0 | -3.9 (2.3) |
| **5** | | Gal-O4 | 2.6 | 2.6 (0.1) | 1.0 | -3.5 (2.3) |
| Modelled structures | | | | | | |
| **4** | | GalNAc-O3 | | 2.6 (0.1) | 1.0 | -3.0 (3.5) |
| **6** | | Gal-O4 | | 2.6 (0.1) | 1.0 | -3.5 (2.3) |
| **1** | Asp89-Oδ1 | Gal-O3 | 2.7 | 2.7 (0.1) | 1.0 | -3.3 (2.7) |
| **2** | | Gal-O4 | 2.7 | 2.6 (0.1) | 1.0 | -3.8 (2.3) |
| **3** | | Gal-O4 | 2.6 | 2.7 (0.1) | 1.0 | -3.1 (2.7) |
| **5** | | Gal-O3 | 2.7 | 2.7 (0.1) | 1.0 | -2.8 (3.4) |
| Modelled structures | | | | | | |
| **4** | | GalNAc-O4 | | 2.7 (0.1) | 1.0 | -3.4 (3.8) |
| **6** | | Gal-O3 | | 2.7 (0.1) | 1.0 | -2.7 (2.6) |
| **1** | Asn133-Nδ2 | Gal-O3 | 2.85 | 2.9 (0.1) | 0.8 | -2.6 (0.9) |
| **2** | | Gal-O3 | 4.0 | 3.0 (0.1) | 0.9 | -2.8 (0.8) |
| **3** | | Gal-O3 | 3.1 | 3.0 (0.1) | 0.9 | -2.8 (0.8) |
| **5** | | Gal-O3 | 2.85 | 3.0 (0.1) | 1.0 | -3.4 (3.4) |
| Modelled structures | | | | | | |
| **4** | | GalNAc-O3 | | 3.0 (0.1) | 1.0 | -2.8 (2.9) |
| **6** | | Gal-O3 | | 3.0 (0.1) | 1.0 | -3.4 (0.8) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **1** | Ala218-N | Gal-O4 | 3.05 | 3.0 (0.1) | 1.0 | -2 (0.5) |
| **2** | | Gal-O4 | 3.15 | 3.1 (0.1) | 1.0 | -2.0 (0.5) |
| **3** | | Gal-O4 | 3.05 | 3.2 (0.2) | 1.0 | -1.8 (0.6) |
| **4** | | GalNAc-O4 | | 3.0 (0.1) | 1.0 | -2.2 (2.2) |
| Modelled structures | | | | | | |
| **4** | | GalNAc-O4 | | 3.0 (0.1) | 1.0 | -2.2 (2.2) |
| **6** | | Gal-O4 | | 3.1 (0.1) | 1.0 | -2.0 (0.5) |
| **1** | Gly107-N | Gal-O3 | 3.0 | 3.0 (0.1) | 0.3 | -1.5 (0.9) |
| **2** | | Gal-O3 | 3.0 | 3.0 (0.1) | 0.3 | -1.5 (0.4) |
| **3** | | Gal-O3 | 2.9 | 3.0 (0.1) | 0.4 | -1.6 (0.9) |
| **5** | | Gal-O3 | 3.0 | 3.0 (0.1) | 0.4 | -1.7 (1.9) |
| Modelled structures | | | | | | |
| **4** | | GalNAc-O3 | | 3.1 (0.2) | 0.1 | -1.3 (1.5) |
| **6** | | Gal-O3 | | 3.0 (0.1) | 0.4 | -1.7 (0.9) |
| **1** | Gln219-Nε2 | Glc-O3 | 3.1 | 4.0 (1.1) | 0.2 | -1.3 (2.1) |
| **2** | | ManO3 | 3.0 | 3.9 (0.9) | 0.3 | -1.5 (2.3) |
| **3** | | GlcNAc-O3 | 2.9 | 3.4 (0.8) | 0.7 | -2.7 (2.2) |
| **5** | | Glc-O3 | 3.1 | 4.0 (1.0) | 0.3 | -1.5 (2.7) |
| Modelled structures | | | | | | |
| **4** | | GlcNAc-O3 | | 3.9 (1.4) | 0.6 | -2.3 (3.1) |
| **6** | | GlcNAc-O3 | | 3.2 (0.6) | 0.8 | -3.0 (3.4) |
| **5** | Asn133-Nδ2 | Fuc-O2 | 2.7 | 3.1 (0.3) | 0.6 | -3.9 (4.5) |
| Modelled structure | | | | | | |
| **6** | | Fuc-O2 | | 3.0 (0.2) | 0.8 | -4.5 (2.3) |
| **5** | Tyr108-OH | Fuc-O4 | 3.0 | 4.4 (0.8) | 0.1 | -0.8 (1.4) |
| Modelled structures | | | | | | |
| **6** | | Fuc-O4 | | 4.0 (0.8) | 0.3 | -1.2 (1.8) |

[1]Å, with standard deviations in parentheses.

[2]kcal/mol, with standard deviations in parentheses.

However, it does lead to an improvement in the $R^2$ values for all MM-GBSA models. Use of NM entropies instead of QH brought the binding energies into closer agreement with the experimental data, both in terms of the magnitudes and relative affinities, particularly with the $GB^{HCT}$, $GB_1{}^{OBC}$, $GB_2{}^{OBC}$, and GB$n_2$ desolvation models. The inclusion of NM entropies significantly improved the relative affinity of fucosylated ligand **5**, ranking it comparable to ligands **1** and **2**. However, opposed to the experiment, inclusion of NM entropies with GB$n_1$ and PBSA desolvation models placed fucosylated ligand **6** along with **1** and **2** as a weak binder, resulting in loss of correlation between the experimental and theoretical ranking of ligands. The binding free energies calculated using the MM-GBSA desolvation models $GB^{HCT}$, $GB_1{}^{OBC}$, and GB$n_2$ desolvation model along with NM entropies result in the best correlation with the experiment ($R^2 = 0.87$). Among all the models used, MM-PBSA performs the best in ranking the ligands, however, the inclusion of entropies, especially NM entropy, leads to a decrease in correlation. The difference between the QH and NM entropies is notable and suggests that the 100 ns time scale is insufficient to capture the low frequency motions, such as stiffening of the backbone, that may occur upon ligand binding. This would likely impact the accuracy of the QH values, more than the NH values, as in the latter method those frequencies are directly computed, whereas in the former, the changes must be observed during the simulation. Additional features, such as the absence of explicit solvent molecules in the NM analyses can affect the computed values (175).

**Table 6.3.** Binding free energies from MM-GBSA calculation[1].

| | Expt. | $GB^{HCT}$ | $GB_1^{OBC}$ | $GB_2^{OBC}$ | $GB_{n1}$ | $GB_{n2}$ | PBSA |
|---|---|---|---|---|---|---|---|
| **1** | -4.83 | -27.17 | -30.17 | -33.07 | -35.73 | -26.39 | -13.44 |
| | (0.04) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| **2** | -5.08 | -28.25 | -30.66 | -33.59 | -36.68 | -26.53 | -14.02 |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) | (0.01) |
| **3** | -5.66 | -30.74 | -32.67 | -35.48 | -37.34 | -29.11 | -19.35 |
| | (0.04) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| **4** | | -31.48 | -32.24 | -34.52 | -36.68 | -28.12 | -15.28 |
| | n.d. | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.02) |
| **5** | -5.04 | -36.46 | -37.69 | -41.27 | -43.15 | -33.68 | -12.30 |
| | (0.06) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.02) |
| **6** | -6.21 | -42.29 | -41.56 | -44.96 | -44.36 | -38.44 | -22.85 |
| | (0.14) | (0.02) | (0.02) | (0.02) | (0.03) | (0.02) | (0.03) |
| | $R^2$ | 0.55 | 0.50 | 0.46 | 0.32 | 0.54 | 0.98 |

[1]kcal/mol, with standard deviations in parentheses.

**Table 6.4.** Binding free energies from MM-GBSA calculation employing quasi-harmonic entropies ($\Delta G_{QH}$) and normal mode entropies ($\Delta G_{NM}$)[1].

| | Expt. | $GB^{HCT}$ | | $GB_1^{OBC}$ | | $GB_2^{OBC}$ | | $GBn_1$ | | $GBn_2$ | | PBSA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $S_{QH}$ | $S_{NM}$ | $S_{QH}$ | $S_{NM}$ | $S_{QH}$ | $S_{NM}$ | $S_{QH}$ | $S_{NM}$ | $S_{QH}$ | $S_{NM}$ | $S_{QH}$ | $S_{NM}$ |
| **1** | -4.83 | -12.79 | -8.17 | -15.80 | -11.18 | -18.69 | -14.07 | -21.36 | -16.73 | -12.02 | -7.39 | 0.94 | 5.56 |
| | (0.04) | (0.01) | (0.87) | (0.01) | (0.87) | (0.01) | (0.87) | (0.01) | (0.87) | (0.01) | (0.87) | (0.01) | (0.87) |
| **2** | -5.08 | -13.65 | -7.77 | -16.05 | -10.12 | -18.99 | -13.12 | -22.07 | -16.20 | -11.93 | -6.05 | 0.59 | 6.46 |
| | (0.02) | (0.03) | (1.00) | (0.03) | (1.00) | (0.03) | (1.00) | (0.03) | (1.00) | (0.03) | (1.00) | (0.03) | (1.00) |
| **3** | -5.66 | -16.76 | -11.32 | -18.69 | -13.25 | -21.50 | -16.06 | -23.37 | -17.93 | -15.13 | -9.69 | -5.37 | 0.07 |
| | (0.04) | (0.03) | (0.95) | (0.03) | (0.95) | (0.03) | (0.95) | (0.03) | (0.95) | (0.03) | (0.95) | (0.03) | (0.95) |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4** | | -17.60 | -11.00 | -18.36 | -11.76 | -20.64 | -14.04 | -22.80 | -16.2 | -14.24 | -7.64 | -1.40 | 5.20 |
| | n.d. | (0.02) | (0.98) | (0.02) | (0.98) | (0.02) | (0.98) | (0.02) | (0.98) | (0.02) | (0.98) | (0.03) | (0.98) |
| **5** | -5.04 | -19.82 | -10.44 | -21.05 | -11.67 | -24.63 | -15.25 | -26.51 | -17.13 | -17.04 | -7.66 | 4.34 | 13.72 |
| | (0.06) | (0.02) | (1.04) | (0.02) | (1.04) | (0.02) | (1.04) | (0.02) | (1.04) | (0.02) | (1.04) | (0.03) | (1.04) |
| **6** | -6.21 | -23.68 | -15.81 | -20.44 | -15.09 | -23.47 | -18.49 | -19.71 | -17.89 | -19.33 | -11.97 | -6.83 | 3.63 |
| | (0.14) | (0.03) | (1.16) | (0.03) | (1.16) | (0.03) | (1.16) | (0.03) | (1.16) | (0.03) | (1.16) | (0.03) | (1.16) |
| | $R^2$ | 0.67 | 0.87 | 0.65 | 0.87 | 0.60 | 0.82 | 0.47 | 0.65 | 0.69 | 0.87 | 0.80 | 0.29 |

[1]kcal/mol, with standard deviations in parentheses.

The origin of the variations in absolute affinity arising from the desolvation model can be illustrated by a subset of per-residue interactions, in the case of **1** (Table 6.5). Each model results in similar (within approximately 2.1 kcal/mol) estimates for the interaction energies that do not involve hydrogen-bonds (Phe 131, Tyr 106). For polar-neutral hydrogen bonds (Asn 133, Gly 107) the interaction energies are generally favorable but vary according to the desolvation model up to approximately 1.6 kcal/mol. Most significantly, the strength of the only interaction with a charged side chain (Asp 89) is predicted to range from -7.8 to + 8.4 kcal/mol. This latter observation clearly points to an important source of uncertainty in the choice of desolvation model. An indication of the overall variation in the energies is provided by Z-scores (Table 6.6) for each of the per-residue interactions and indicates that the $GB_1$ $^{OBC}$ model is in closest agreement with the average of all the models. While this doesn't imply that the $GB_1$ $^{OBC}$ model is the optimal choice, it provides a basis to state that it is a representative GBSA model, enabling us to select it for further analysis.

**Table 6.5.** The impact of desolvation model on per-residue interaction energies[1].

| ECL Residue | $GB^{HCT}$ | $GB_1^{OBC}$ | $GB_2^{OBC}$ | GB*n1* | GB*n2* | PBSA | Mean | Standard Deviation |
|---|---|---|---|---|---|---|---|---|
| ASN 133 | -1.91 | -1.08 | -1.03 | -0.28 | -0.74 | -2.15 | -1.2 | 0.71 |
| ASP 89 | -1.44 | -4.8 | -6.23 | -7.83 | 3.56 | 8.45 | -1.38 | 6.29 |
| GLY 107 | -1.22 | -0.67 | -0.6 | -0.13 | -1.05 | -1.73 | -0.9 | 0.55 |
| PHE 131 | -2.17 | -2.46 | -2.64 | -2.63 | -2.33 | -0.54 | -2.13 | 0.8 |
| TYR 106 | -2.26 | -1.55 | -1.52 | -1.46 | -1.95 | -2.86 | -1.93 | 0.55 |

[1]kcal/mol

**Table 6.6.** Z-scores[1] for per-residue interaction energies as a function of the desolvation model

| ECL Residue | $GB^{HCT}$ | $GB_1^{OBC}$ | $GB_2^{OBC}$ | $GBn_1$ | $GBn_2$ | PBSA |
|---|---|---|---|---|---|---|
| ASN 133 | -1.0 | 0.2 | 0.2 | 1.3 | 0.6 | -1.3 |
| ASP 89 | 0.0 | -0.5 | -0.8 | -1.0 | 0.8 | 1.6 |
| GLY 107 | -0.6 | 0.4 | 0.5 | 1.4 | -0.3 | -1.5 |
| PHE 131 | -0.1 | -0.4 | -0.6 | -0.6 | -0.3 | 2.0 |
| TYR 106 | -0.6 | 0.7 | 0.8 | 0.9 | 0.0 | -1.7 |

[1]Z-score = (observed value – Mean)/Standard Deviation

*Quantification of per-residue contributions to affinity:* Amino acids making significant interactions with the ligand were identified on the basis of their individual contributions to the total interaction energy, and confirmed all of the expected interactions (Figure 6.3). In addition, stabilizing non-polar (van der Waals) contacts were observed between the Fuc residue and Y106, Y108, P134, and W135, which have been noted from analyses of the crystal structure. Non-polar contacts were also observed in the presence of GlcNAc residue, stabilizing the interaction of Q219 with the ligand by over 0.5 kcal/mol. While the presence of the GalNAc residue

introduced van der Waals contacts with N133, it also introduced electrostatic repulsion, reducing the overall contribution of N133 to the binding. The significance of some of these residues (A88G, Y106A, FI31A, A218G, D89A, N133A, and Q219A among others) has been examined by performing mutation studies on a closely related protein called *Erythrina corallodendron* lectin (ECorL) (176). From the perspective of the ligand, the Gal/GalNAc residues were found to be the main contributors to binding, accounting for more than 80% of the interaction energy in all cases. In **5** and **6**, the fucosyl residue contributed 4% and 7%, fully consistent with the observation that fucosylation impacts the affinity only marginally. The Glc and Man residues contributed less than 4%, while the presence of NAc group in the GlcNAc residue brings its contribution up to over 8.5% in **3**, **4** and **6** (Figure 6.4).

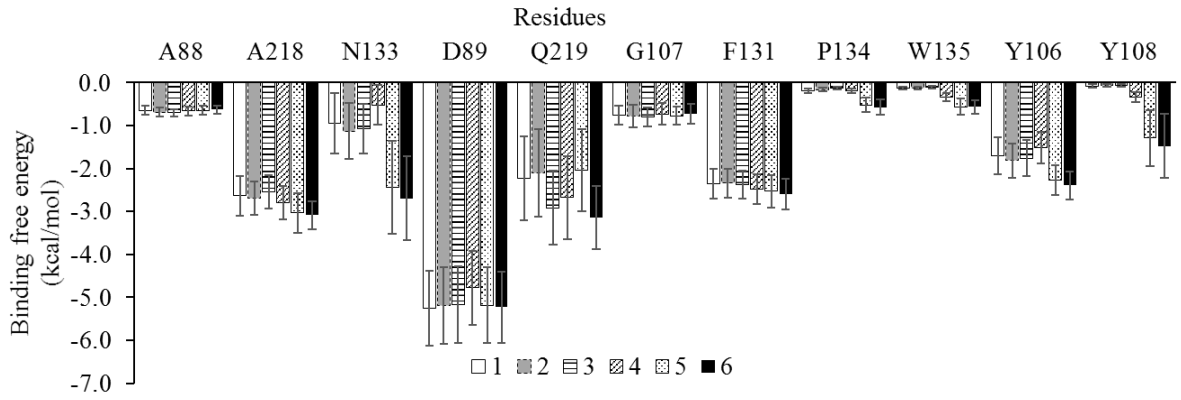



**Figure 6.3.** The binding free energy contribution of amino acids making significant interactions with the ligand.

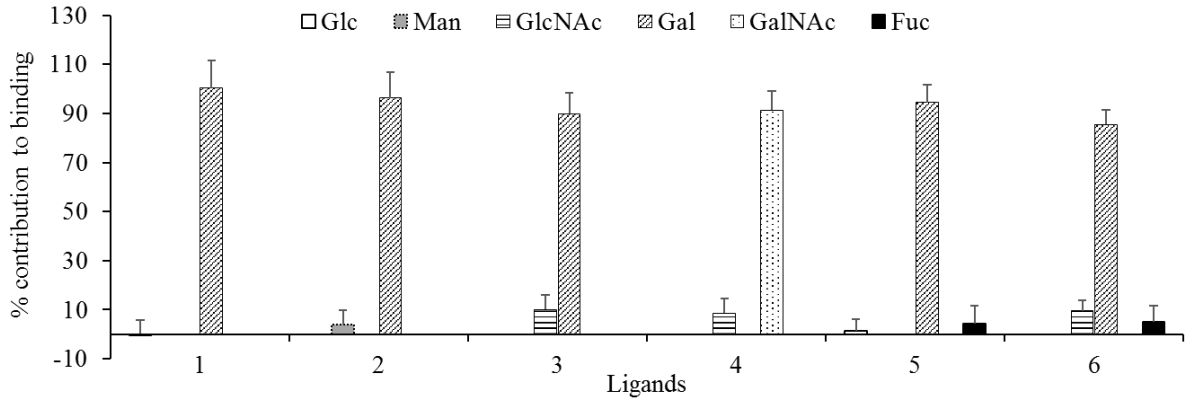



**Figure 6.4**. The percentage contribution of all the ligands on a per-residue basis.

*Quantification of per-functional group contributions to affinity:* Per-atom decomposition of the interaction energy, and categorizing it on the basis of the per-exocyclic group, revealed which of these groups were most involved in the interaction. This analysis showed that the main contribution to binding came from electrostatic interactions with the O3 and O4 groups (O3 over 20%, O4 over 14%) along with van der Waals contacts from the framework atoms (FW) of the Gal/GalNAc residue (over 30%). The N-acetyl group stabilized the interaction by contributing about 1 kcal/mol to the binding. It was observed that some groups are crucial for the protein-ligand interaction (O3, O4 group and framework of Gal/GalNAc residue), while some enhance this interaction (NAc) and others do not participate (such as O6 and O2 groups of Gal/GalNAc and Glc/GlcNAc/Man residues) (Figure 6.5). This provides an objective method to quantify features of the ligand that are critical for binding. Based on these observations it can be deduced that the conformation of the groups contributing most to the binding, defines the minimum 3D motif required for that protein-ligand interaction. Therefore, these 3D motifs in sugar residues can be denoted as the Minimum Binding Determinants (MBD).

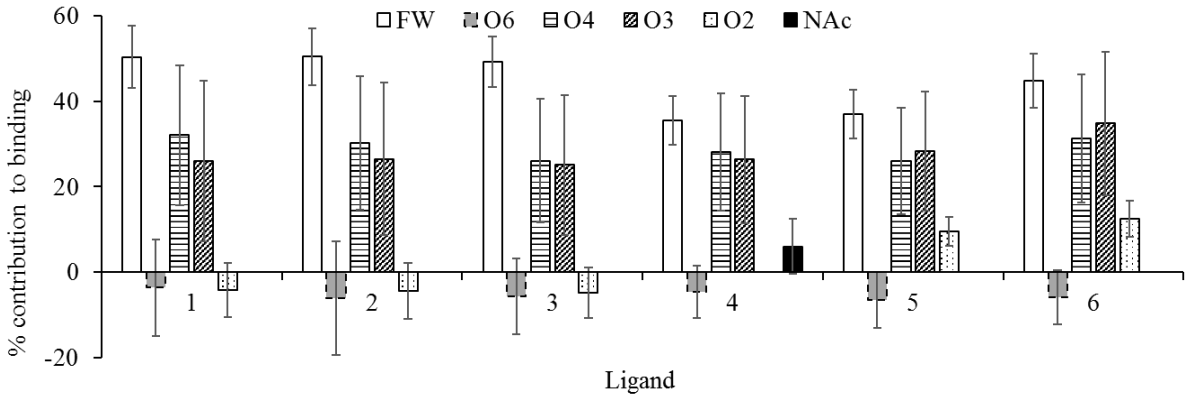



**Figure 6.5.** The percentage contribution of the functional groups of Gal residue in all the ligands.

The lack of participation of the Glc residue explains why replacing it in Lactose (**1**) with its O2 epimer i.e. Mannose in Epi-Lactose (**2**) results in their equivalent binding affinities. Similarly, modifying the O2 group of Gal residue should not affect the binding interaction as it does not make a significant contribution if the modification does not introduce any clashes. This was observed in **4, 5** and **6** where O2 hydroxyl of Gal residue was replaced with N-acetyl group and fucosyl groups. Conversely, modification of groups with a high contribution (O3 and O4 groups of Gal residue) should significantly affect the binding. For example, replacing Gal residue with its O4 epimer i.e. Glucose, resulting in Cellobiose (Glcβ1-4 Galβ, **7**) and Maltose (Glcα1-4 Galβ, **8**) should hamper its interaction. This was proven by Biolayer Interferometry performed here which was unable to detect any binding of **7** and **8** with ECL, hence validating the hypothesis that not all groups in a bound ligand participate in the binding.

**Conclusion**

Through a combination of experimental and computational analysis, the study provides insight into the features that lead to carbohydrate specificity of using the lectin ECL and its six known ligands as an example system. The results from binding free energy analyses, employing

different desolvation models, along with entropy calculations, indicate that, for agreement with the experiments, improvements need to be made in the current desolvation models. It would likely be beneficial to include carbohydrate-protein interactions to re-calibrate the current GB/PB methods (177). We see a large variation in the per-residue decomposition of binding energy for charged residue ASP89, with a difference of 16.2 kcal/mol, which leads to an ambiguity in the choice of desolvation model. The addition of QH entropies improves the correlation of the MM-GBSA models, but they did not converge even after 100 ns of simulation and had to be extrapolated. The nmode entropies improve the correlation even further and correctly rank the ligands in most cases, but they are computationally expensive and were performed on 100 snapshots from the entire simulation. Nonetheless, by decomposing the binding free energy on a per-residue basis, the MM-GBSA calculations could identify and rank key residues responsible for the protein-ligand interactions. Thus, it was possible to locate the functional groups in each ligand that were responsible for the specificity of these ligands. Based on the range of strengths of their interactions, the functional groups could be characterized as critical, stabilizing, or non-interacting. Critical groups are essential for achieving measurable affinity while stabilizing groups improve the strength of the binding. As expected, non-interacting groups can be replaced, if doing so does not introduce unfavorable van-der Waals or electrostatic repulsions. The ability to rank the functional groups in terms of their importance to binding can be used to design novel ligands and can aid in explaining the specificity and affinity of different ligands for a protein.

# CHAPTER 7

# MONOSACCHARIDE SIMILARITY ANALYSIS TO UNDERSTAND PROTEIN-CARBOHYDRATE SPECIFICITY

## Introduction

Carbohydrates (oligo- and polysaccharides, glycans) comprise a structurally diverse group of biopolymers, which participate in a multitude of biological processes, a number of which involve recognition by specific glycan-binding proteins (GBPs). The ability of GBPs to recognize specific glycans is essential for organisms to carry out their physiological or pathological processes. They are known to participate in cell signaling, cell adhesion, endocytosis, immune response, hemostasis, host-pathogen interactions among other roles in the functions of many cells (178). Any disease that disrupts the cellular glycosylation machinery can alter the ensemble of glycans displayed on the cell surface, hence glycans can also be markers for diseases, such as liver, ovarian and pancreatic cancer, which show elevated levels of fucosylation (179).

Understanding glycan specificity is, therefore, a crucial component in glycobiology, but also essential for the design and development of carbohydrate-specific reagents, such as antibodies.

Glycan specificity is typically defined in terms of the shortest glycan sequence that is found among the ensemble of glycans that bind to a given GBP. This sequence is often referred to as the minimal binding determinant (MBD). Notably, glycan specificity studies have also revealed (146, 147) that some GBPs bind to glycans that appear to contain unrelated MBDs; for example, the lectin Wheat Germ Agglutinin (WGA) binds to both N-acetylglucosamine (GlcNAc) or sialic acid (Neu5Ac). Such cross-reactivities lead to uncertainties in the canonical definitions of GBP

specificity and have led lectins and even anti-carbohydrate antibodies to be described as displaying broad or complex specificity. If the specificity of such reagents cannot be well defined or understood it limits their potential utility in diagnostic or therapeutic applications.

Challenges in interpreting carbohydrate specificity arise in part from the assumption that specificity can be defined uniquely by the residues (including the inter-residue linkages) that make up the oligosaccharide sequence of the MBD. Such a nomenclature-based definition fails to identify the precise pharmacophore (the subset of underlying 3D structural features within the MDB that bind directly to the GBP) responsible for the binding affinity. Because of the structural similarities among monosaccharides (many of which are diasteriomers) it is possible, and even common, for a GBP to be able to interact with the same pharmacophore among multiple monosaccharides. For example, the pharmacophore in Neu5Ac associated with the binding of WGA consists of the *N*-acetyl and O4-hydroxyl groups. In the case of GlcNAc, the same pharmacophore can be created by the *N*-acetyl and O3-hydroxyl groups (Figure 7.1).

The location (in terms of 3D structure) of the pharmacophore within the glycan will also impact the extent to which the pharmacophore will continue to be recognized by a given GBP. Again, in the case of WGA, as the Neu5Ac residues are typically present on the termini of glycan branches, WGA will recognize the Neu5Ac residues in a broad range of glycans. In contrast, the GlcNAc residues in glycans are often present in non-terminal positions, and for this reason, WGA recognizes only a sub-set of glycans that contain GlcNAc. The dependence of recognition on the context of the MBD (or more precisely on the pharmacophore) within the glycan further complicates the interpretation of glycan specificity (145). As a prerequisite to understanding the context-dependence of glycan recognition, or conversely to determine the basis for cross reactivity, it is essential to be able to identify the relevant pharmacophore (180). While the

pharmacophore that leads to cross-reactivity can be obvious (as in the case of WGA), its discovery necessitates abandoning the traditional representations of monosaccharides and focusing not on residue descriptors but on the 3D structure.


**Figure 7.1.** Monosaccharides GlcNAc (left) and Neu5Ac (right) showing a shared pharmacophore (red).

In glycan-protein complexes, not all exocyclic groups in the carbohydrate are involved in the interaction. The groups that participate depend on the orientation of the monosaccharide in the binding site, and on the configuration of the exo-cyclic groups. Any monosaccharide that presents comparable interacting groups appropriately, and does not introduce sterical collisions with the protein is a potential ligand (181). Any alteration of the interacting groups, will potentially adversely affect or eliminate recognition, while other non-interacting moieties may be altered with little affect. Therefore, to be able to identify and predict binding specificity, a new approach to comparing structural similarities among monosaccharides is required.

Monosaccharides are structurally diverse, as they can exist as D- or L-isomers and as α- or β-anomers. Moreover, a change in anomer configuration or linkage position can lead to molecules with dissimilar biological attributes, which further increases their complexity. Oligo and polysaccharides vary widely in size and shape, and because of their flexibility, can adopt multiple conformations in solution. Owing to their complexity, generating an unambiguous and

consistent representation of glycans can be problematic, as it needs to consist of information from monomers involved to three-dimensional (3D) shapes and linkages. Currently, there are several representations which incorporate individual monosaccharide units (182), their isomeric and anomeric state, linkages and ring structure (183). The specificity of GBPs is defined based on these representations.

Simple 1D and 2D representations like SMILES strings (184), Sybyl Line Notation (185), InChI (186) and WURCS (183) are routinely used to encode molecular structure and have proven to be a powerful tool for ligand comparison. However, these become quite complicated for carbohydrates and as monosaccharides are cyclic, there can be multiple equally valid notations for one structure. The rules to derive a valid SMILES string can often be rather arduous to encode. A simple representation encoding the 3D features of carbohydrates can be used to compare them and score their similarities, that can be used explain and predict the cross reactivity of GBPs. A linear representation of carbohydrate monosaccharides based on the historic Fischer Projection representation was developed, which can in principle be employed to automatically detect sub-structure motifs. A scoring function was introduced to assess similarities and locate like exocyclic groups. The significance of the scoring function will be illustrated by comparing the ligands of GBPs known to show cross reactivity and to predict novel ligands.

**Methods**

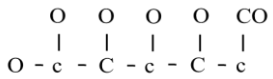
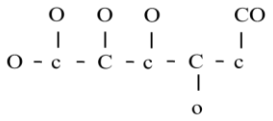
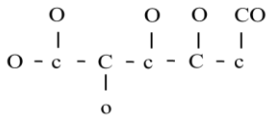
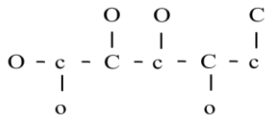
*Atom-based (3D) representation of Monosaccharides.* A novel representation, specific for monosaccharides, based on Fischer and Haworth projections called Enhanced Haworth-Fischer (EHF) projection was introduced. This representation is capable of incorporating 3D structural features of individual monosaccharides and can be used to compare them. Combined with

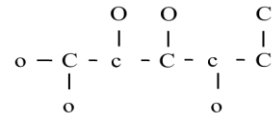
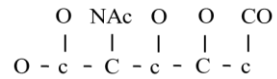
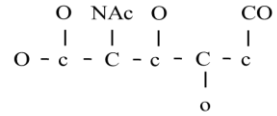
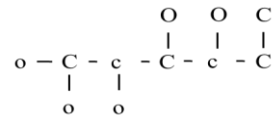
experimental and binding energy data, it can communicate the information required to specify the location of the atom-based 3D motifs involved in protein-glycan interactions. As it is independent of current naming conventions, representing motifs by their names or symbols to define specificity can be avoided.

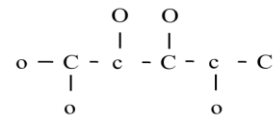
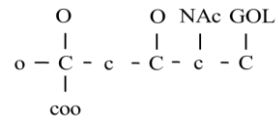
*Rules to generate the representation*. First, the anomeric carbon was located, and the rest of the ring atoms were noted in the clockwise direction. Atoms above the average plane of the ring were denoted by capital letters, and the ones below the plane were denoted by lowercase letters. Uppercase or lowercase letters were used to represent the names for exocyclic atoms depending on whether they were equatorial or axial respectively. The equatorial groups were denoted above and the axial groups were denoted below the ring atoms they were bonded to. A monosaccharide can be represented in two different ways depending on how it is viewed. An 180° rotation around an axis will change the sequence of atoms in a clockwise direction and change the location of ring atoms with respect to the average plane. Using these rules, this representation can also be linearized resembling SMILES notation. The ring atoms are separated by an underscore, and the associated exocyclic groups are separated by a hyphen. Some of the monosaccharides with their SMILES notations and EHF representations are listed in Table 7.1. This specialized notation can represent the chair form, which is the predominant conformation of pyranoses and is capable of distinguishing subtle structural variations.

**Table 7.1.** Carbohydrate representations. The numbers in the second column indicate these representations as 1 – SMILES string; 2 – WURCS; 3 – InChI; 4 – EHF; 5 – Linear EHF.

| Monosaccharide | | Representations |
|---|---|---|
| β-D-Glc | 1 | C([C@@H]1[C@H]([C@@H]([C@H]([C@@H](O1)O)O)O)O)O |

| | | |
|---|---|---|
| | 2 | 1.0/1,0/[12122h\|1,5] |
| | 3 | 1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6-/m1/s1 |
| | 4 |  |
| | 5 | O_c-O_C-O_c-O_C-O_c-CO |
| β-D-Gal | 1 | C([C@@H]1[C@@H]([C@@H]([C@H]([C@@H](O1)O)O)O)O)O |
| | 2 | 1.0/1,0/[12112h\|1,5] |
| | 3 | 1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3+,4+,5-,6-/m1/s1 |
| | 4 |  |
| | 5 | O_c-O_C-O_c-O_C-o_c-CO |
| β-D-Man | 1 | C([C@@H]1[C@H]([C@@H]([C@@H]([C@@H](O1)O)O)O)O)O |
| | 2 | 1.0/1,0/[11122h\|1,5] |
| | 3 | 1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5+,6-/m1/s1 |
| | 4 |  |
| | 5 | O_c-O_C-o_c-O_C-O_c-CO |
| α-D-Fuc | 1 | C[C@@H]1[C@@H]([C@@H]([C@H]([C@H](O1)O)O)O)O |
| | 2 | 1.0/1,0/[22112m\|1,5] |
| | 3 | 1S/C6H12O5/c1-2-3(7)4(8)5(9)6(10)11-2/h2-10H,1H3/t2-,3+,4+,5-,6+/m1/s1 |
| | 4 |  |
| | 5 | O_c-o_C-O_c-O_C-o_c-C |
| α-L-Fuc | 1 | C[C@H]1[C@H]([C@H]([C@@H]([C@@H](O1)O)O)O)O |
| | 2 | 1.0/1,0/[11221m\|1,5] |
| | 3 | 1S/C6H12O5/c1-2-3(7)4(8)5(9)6(10)11-2/h2-10H,1H3/t2-,3+,4+,5-,6+/m0/s1 |

| | | |
|---|---|---|
| | 4 |  |
| | 5 | o_C-o_c-O_C-O_c-o_C-C |
| β-D-GlcNAc | 1 | CC(=O)N[C@@H]1[C@H]([C@@H]([C@H](O[C@H]1O)CO)O)O |
| | 2 | 1.0/1,0/[12122h\|1,5\|2*NCC/3=O] |
| | 3 | 1S/C8H15NO6/c1-3(11)9-5-7(13)6(12)4(2-10)15-8(5)14/h4-8,10,12-14H,2H2,1H3,(H,9,11)/t4-,5-,6-,7-,8-/m1/s1 |
| | 4 |  |
| | 5 | O_c-O_C-NAC_c-O_C-O_c-CO |
| β-D-GalNAc | 1 | CC(=O)N[C@@H]1[C@H]([C@H]([C@H](O[C@H]1O)CO)O)O |
| | 2 | 1.0/1,0/[12112h\|1,5\|2*NCC/3=O] |
| | 3 | 1S/C8H15NO6/c1-3(11)9-5-7(13)6(12)4(2-10)15-8(5)14/h4-8,10,12-14H,2H2,1H3,(H,9,11)/t4-,5-,6+,7-,8-/m1/s1 |
| | 4 |  |
| | 5 | O_c-O_C-NAC_c-O_C-o_c-CO |
| α-L-Rha | 1 | C[C@H]1[C@@H]([C@H]([C@H]([C@@H](O1)O)O)O)O |
| | 2 | 1.0/1,0/[12211m\|1,5] |
| | 3 | 1S/C6H12O5/c1-2-3(7)4(8)5(9)6(10)11-2/h2-10H,1H3/t2-,3-,4+,5+,6+/m0/s1 |
| | 4 |  |
| | 5 | o_C-o_c-o_C-O_c-O_C-C |
| β-L-Ara | 1 | C1[C@@H]([C@@H]([C@H]([C@H](O1)O)O)O)O |
| | 2 | 1.0/1,0/[2211h\|1,5] |
| | 3 | 1S/C5H10O5/c6-2-1-10-5(9)4(8)3(2)7/h2-9H,1H2/t2-,3-,4+,5-/m0/s1 |

| | | |
|---|---|---|
| | 4 |  |
| | 5 | o_C-o_c-O_C-O_c-o_C |
| α-D-Neu5Ac | 1 | CC(=O)N[C@@H]1[C@H](C[C@@](O[C@H]1[C@@H]([C@@H](CO)O)O)(C(=O)O)O)O |
| | 2 | 1.0/1,0/[a2d21122h\|2,6\|2*O\|5*NCC/3=O] |
| | 3 | 1S/C11H19NO9/c1-4(14)12-7-5(15)2-11(20,10(18)19)21-9(7)8(17)6(16)3-13/h5-9,13,15-17,20H,2-3H2,1H3,(H,12,14)(H,18,19)/t5-,6+,7+,8+,9+,11+/m0/s1 |
| | 4 |  |
| | 5 | o_C-coo-O_c_C-O_c-NAC_C-GOL |

1 – SMILES string; 2 – WURCS; 3 – InChI; 4 – EHF; 5 – Linear EHF.

*Maximum similarity score.* A scoring function was developed to compare monosaccharides using the new notation and to locate and quantify structural similarities in them. Two monosaccharides can be aligned in multiple ways, with each alignment resulting in a different set of structural similarities (Figure 7.2). The aim of the maximum similarity score is to find an alignment which results in the highest structural resemblance.

*Rules to generate the score.* All possible alignments are scored by first matching the ring atoms based on their location with respect to the average plane of the ring and if the atoms themselves are a match. Then the orientation of all the exocyclic groups is compared and scored at the aligned ring positions. Each match of location, ring atoms and exocyclic groups gets a score of +1, leading to a maximum possible similarity score of 22 if the exact same monosaccharides are aligned. The scores and positions of an alignment that result in the maximum similarity score are reported for some of the monosaccharide pairs (Table 7.2 and 7.3). In Figure 7.2, the maximum

score is observed with an alignment with 4 leading to a score of 16. These alignments can then be used to explain and predict cross reactivity. It can also be used to generate models for known cross reactivity, but unknown structures. It is important to note that because different alignments can lead to different conformational similarities, and multiple alignments can lead to the same score, the maximum similarity is not an essential explanation for cross reactivity.

**Example applications**

*Explain cross-reactivity*: The ability of a GBP to bind multiple glycans, leading to cross-reactivity, arises from the presence of the same pharmacophore in those glycans. Therefore, to explain cross-reactivity, it becomes essential to locate the pharmacophore responsible for binding. Six GBPs (Concanavalin A (ConA), M-ficolin, *Pseudomonas aeruginosa*-II lectin (PA-II L), Rhamnose-binding lectin (RBL), *Sambucus nigra* lectin (SNL) and Wheat Germ agglutinin (WGA)) with known cross-reactivity and complex crystal structures with those ligands were selected to test the predicted alignments of the monosaccharides, by comparing them to the binding modes of the ligands in the crystal structures (Figure 7.3). In all the cases, maximum similarity score could predict the similarities in those ligands, and thus their binding modes, which are detailed in Table 7.4. The similar positions are denoted by the exocyclic group followed by their ring position for ligand 1 and then for ligand 2. For example, Concanavalin A (ConA) is a legume lectin known to specifically interact with α-D-mannosyl and α-D-glucosyl groups, that are epimers at position C2 (Figure 7.3A). The alignment of these monosaccharides that results in the maximum similarity is when they match at all the positions except at C2, earning them a score of 20.

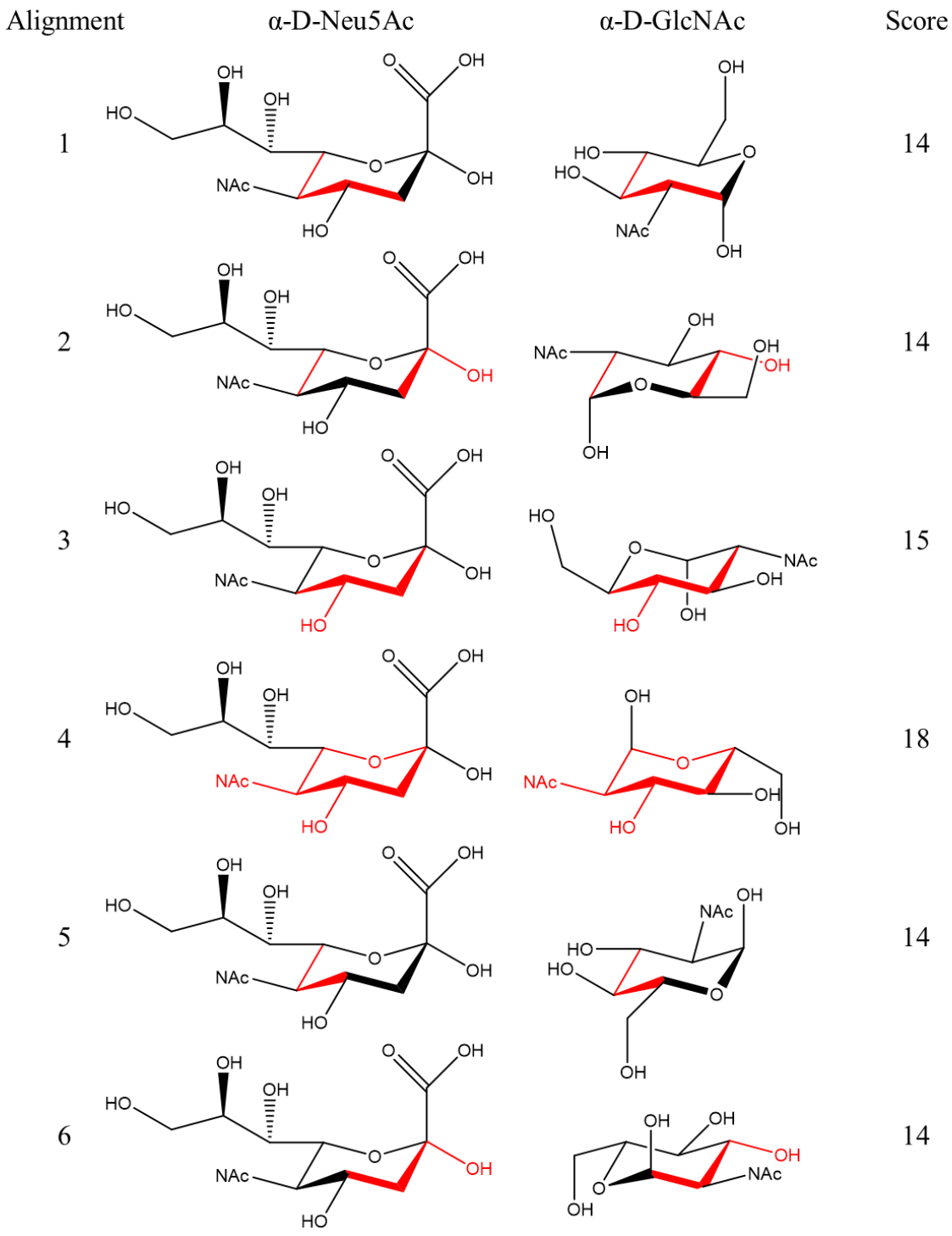



**Figure 7.2.** Different alignments of α-D-Neu5Ac and α-D-GlcNAc. These alignments are scored by first matching the ring atoms based on their location with respect to the average plane of the ring and if the atoms themselves are a match. Then the orientation of all the exocyclic groups at

both axial and equatorial positions is compared and scored at the aligned ring positions the maximum similarity score of the above aligned conformations are (alignment-score) 1-14, 2-14, 3-15, 4-18, 5-14, 6-14.

**Table 7.2.** The maximum similarity scores for a pair of monosaccharides.

| | β-D-Glc | β-D-Gal | β-D-Man | β-D-Fuc | β-L-Fuc | β-D-GlcNAc | β-D-GalNAc | β-L-Rha | β-D-Xyl | β-L-Ara |
|---|---|---|---|---|---|---|---|---|---|---|
| β-D-Gal | 20 | | | | | | | | | |
| β-D-Man | 20 | 18 | | | | | | | | |
| β-D-Fuc | 19 | 21 | 17 | | | | | | | |
| β-L-Fuc | 18 | 17 | 20 | 18 | | | | | | |
| β-D-GlcNAc | 21 | 19 | 20 | 18 | 18 | | | | | |
| β-D-GalNAc | 19 | 21 | 18 | 20 | 16 | 20 | | | | |
| β-L-Rha | 18 | 20 | 17 | 20 | 18 | 17 | 19 | | | |
| β-D-Xyl | 21 | 19 | 19 | 20 | 18 | 20 | 18 | 18 | | |
| β-L-Ara | 17 | 15 | 19 | 15 | 19 | 17 | 15 | 15 | 17 | |
| β-D-Neu5Ac | 17 | 17 | 16 | 17 | 16 | 18 | 18 | 17 | 17 | 17 |

**Table 7.3.** The aligned positions for monosaccharide pairs observed when an alignment results in the maximum similarity score.

| | β-D-Glc | β-D-Gal | β-D-Man | β-D-Fuc | β-L-Fuc | β-D-GlcNAc | β-D-GalNAc | β-L-Rha | β-D-Xyl | β-L-Ara |
|---|---|---|---|---|---|---|---|---|---|---|
| D-Gal-b | O1O1; O2O2; O3O3; CO5CO5 | | | | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| D-Man-b | O1O1; O3O3; O4O4; CO5CO5 | O1O1; O3O3; CO5CO5 | | | | | | | |
| D-Fuc-b | O1O1; O2O2; O3O3 | O1O1; O2O2; O3O3; o4o4 | O1O1; O3O3; | | | | | | |
| D-GlcNAc-b | O1O1; O3O3; O4O4; CO5CO5 | O1O1; O3O3; CO5CO5 | O1O1; O3O3; O4O4; CO5CO5 | O1O1; O3O3; | O3O3; O2O4; | | | | |
| D-GalNAc-b | O1O1; O3O3; CO5CO5 | O1O1; O3O3; o4o4; CO5CO5 | O1O1; O3O3; CO5CO5 | O1O1; O3O3; o4o4; | O3O3; | O1O1; NAC2N AC2; O3O3; CO5CO5 | | | |
| L-Rha-b | O3O3; O2O4 | o4o2; O3O3; O2O4 | O3O1; o2o2; O1O3 | o4o2; O3O3; O2O4; | O1O1; O3O3; C5C5; | O3O3 | o4o2; O3O3 | | |
| D-Xyl-b | O1O1; O2O2; O3O3; O4O4 | O1O1; O2O2; O3O3 | O1O1; O3O3; O4O4; | O1O1; O2O2; O3O3; C5C5 | O3O3; O2O4; | O1O1; O3O3; O4O4; | O1O1; O3O3 | O4O2; O3O3 | |
| L-Ara-b | O4O2; O3O3 | O2O1; O3O2 | O4O2; O3O3; o2o4 | O2O1; O3O2 | O2O2; O3O3; o4o4 | O4O2; O3O3; | O3O3; | O3O3 | O4O2; O3O3 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| D-Neu5Ac-b | O3O3 | O3O3 | O3O3 | O3O3 | O3O3 | O3O3; NAC2NAC4 | O3O3; NAC2NAC4 | O3O3 | O3O3 | o1o1; O3O3 |

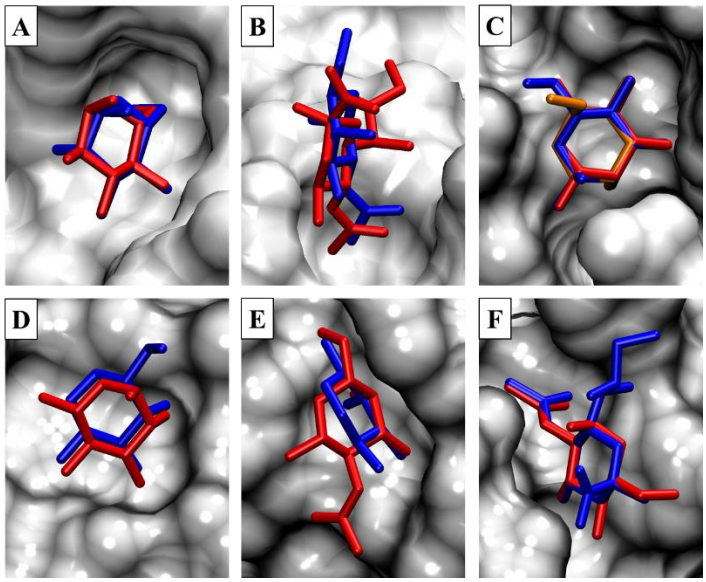



**Figure 7.3.** Aligned structures of GBPs with known cross reactivity and crystal structures with their ligands in the binding pocket. A. ConA with αDGlc in red and α-D-Man in blue. B. M-ficolin with β-D-Neu5Ac in red and α-D-GalNAc in blue. C. PA-II L with α-L-Fuc in blue, α-D-Man in red and β-L-Ara in orange. D. RBL with β-D-Gal in blue and αLRha in red. E. SNA with β-D-GalNAc in blue and α-D-Fuc in red. F. WGA with α-D-Neu5Ac in blue and α-D-GlcNAc in red.

**Table 7.4.** GBPs with known cross-reactivity and structures observe the same alignment as predicted by maximum similarity score.

| GBP | Ligand 1 (PDB ID) | Ligand 2 (PDB ID) | Aligned positions |
|---|---|---|---|

| | | | |
|---|---|---|---|
| ConA (Figure 7.2A) | α-D-Glc (1GIC) | α-D-Man (1I3H) | o1o1;O3O3;O4O4;CO5CO5; |
| M-ficolin (Figure 7.2B) | α-D-GalNAc (2JHI) | β-D-Neu5Ac (2JHL) | O3O3;NAC2NAC4; |
| PA-II L (Figure 7.2C) | α-D-Man (1OUR) | α-L-Fuc (1OXC) | O4O2;O3O3;o2o4; |
| PA-II L (Figure 7.2C) | α-D-Man (1OUR) | β-L-Ara (2BOJ) | O4O2;O3O3;o2o4; |
| RBL (Figure 7.2D) | α-L-Rha (2ZX2) | β-D-Gal (2ZX4) | o2o4;O3O3;O4O2; |
| SNA (Figure 7.2E) | α-D-Fuc (3CAH) | β-D-GalNAc (3CA3) | O3O3;o4o4; |
| WGA (Figure 7.2F) | α-D-GlcNAc (2JHI) | α-D-Neu5Ac (2CWG) | O3O3;NAC2NAC4; |

*Predict binding modes*: There are GBPs with known cross-reactivity that lack crystal structures or other experimental data detailing the interactions involved. The maximum similarity score can be used to model the binding based on the structures available for at least one of the ligands. The models were created for four such proteins i.e. *Amaranthus caudatus* lectin (ACL), family 9 Carbohydrate-binding module from *Thermotoga maritima* Xylanase 10A (CBM9-2), *Helix pomatia* lectin (HPL) and *Ricinus communis* agglutinin I (RCA120). All these proteins have a complex crystal structure available for at least one of their known ligands, and the models were generated based on these existing structures (Table 7.5, Figure 7.4). Such as CBM9-2 is known to show an affinity for D-glucose, D-galactose, and D-xylose (187), while its only complex structure available is bound with D-glucose. Therefore, its binding mode with D-xylose was predicted based on this available structure (Figure 7.4B).

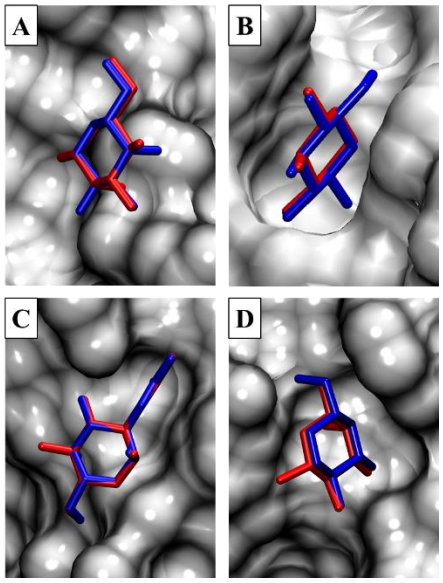



**Figure 7.4.** GBPs with known cross reactivity and crystal structure with ligand in the binding pocket, along with a modeled ligand with the unknown crystal structure. All crystal ligands are in blue and modeled ligands in red. A. ACL with β-D-Gal and β-D-Man. B. CBM-9 with β-D-Glc and β-D-Xyl. C. HPL with α-D-GalNAc and α-D-GlcNAc. D. RCA120 with α-D-Gal and β-L-Rha.

**Table 7.5.** Predicted alignments of known ligands based on maximum similarity score.

| | crystal ligand (PDB ID) | modeled ligand | position |
|---|---|---|---|
| ACL (Figure 7.3A) | β-D-Gal (1GIC) | β-D-Man | O1O1;O3O3;CO5CO5; |
| CBM-9 (Figure 7.3B) | β-D-Glc (1I8A) | β-D-Xyl | O1O1;O2O2;O3O3;O4O4; |
| HPL (Figure 7.3C) | α-D-GalNAc (2CCV) | α-D-GlcNAc | o1o1;NAC2NAC2;O3O3;CO5CO5; |
| RCA120 (Figure 7.3D) | α-D-Gal (3RTI) | β-L-Rha | o4o2;O3O3;O2O4; |

*Predict cross-reactivity*: Based on the above observations, the maximum similarity score was used to predict cross-reactivity of P domain of norovirus, that interacts with histo-blood group antigens (HBGAs) with L-fucose in the binding pocket. According to the maximum similarity

score, it can be hypothesized that this protein can also bind sialic acid in its binding pocket, with the alignment at exocyclic hydroxyls attached to C3 in fucose and C4 in sialic acid, and the ring atoms (Figure 7.5).

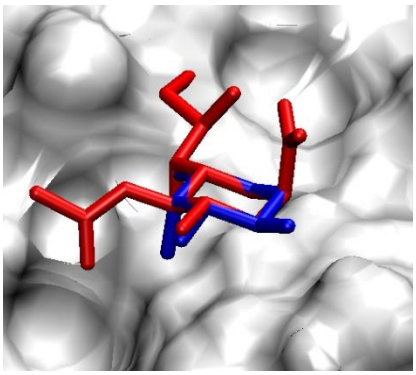


**Figure 7.5.** The co-crystal structure P domain of norovirus with fucose (blue) (PDB ID: 4OPO) and predicted sialic acid ligand (red).

## Conclusions

The 3D features of different monosaccharides were analyzed and compared, leading to a novel representation specific for monosaccharides, which was then used to quantify similarities in a pair of monosaccharides by assigning them a maximum similarity score, based on their alignment. The new representation is simple and can account for the complexity of 3D structures of carbohydrates. It focuses on the most favored conformation of sugars i.e. the chair form, however, there is a potential to include the features of other conformations as well. The maximum similarity score was successful to locate structural similarities, that help in explaining cross-reactivity observed in several GBPs, and predicting them.

# CHAPTER 8

# CONCLUSIONS AND FUTURE PROSPECTS

Experimental methods employed in biomolecular research have proven tedious and expensive to implement for structure determination and dynamics. This is especially true when it comes to carbohydrates as they present unique challenges because of their complex structure and stereochemistry. The combination of experimental techniques and computational analysis has emerged as a powerful methodology to examine their behavior and interactions with proteins.

In the first part of this study, hydroxyl radical protein footprinting (HRPF) combined with molecular dynamics (MD) and solvent accessible surface area (SASA) estimation was used to establish relationship models between amino acid accessibility and oxidation. The residues exposed to the solvent can get oxidized by hydroxyl radicals, while the ones buried inside the core of the globular protein are shielded from this reaction. The SASA of the residues that are oxidized can then be estimated from the established relationship models. These estimated SASA values can be used to test the accuracy of protein conformations generated through MD or homology or comparative modeling. The research presented here verifies the efficiency of such an analysis by calculating $RMSD_{SASA}$ of homology models with respect to the known crystal structure of lysozyme. The models close to the crystal structure show lower $RMSD_{SASA}$ and vice versa. This information can also be extended to studying protein-protein and protein-ligand interactions, as the residues in the binding pocket are shielded from the hydroxyl radicals. The results of this study provide an innovative way of quantifying the quality of protein models,

which can lead to improved accuracy of protein structure prediction, and a better estimate of position and orientation of a ligand when it is bound to a protein receptor.

While the previous part of the study was focused on the analysis of proteins, in the next section, the role of carbohydrates in their interactions with proteins was examined. The main objective was to gain insight into the structural features of carbohydrates leading to the observed cross-reactivity in their interactions with Glycan binding proteins (GBPs). The free binding energy analysis of (ECL) and its interaction with its known ligands, on a per-residue basis, followed by the decomposition of free binding energy at the per-functional group, revealed that all the exocyclic groups of a monosaccharide do not contribute to binding. There are groups that are necessary, some enhance the interaction, while others are non-participating. The mutation of non-participating groups, if it does not introduce any steric clashes, does not hinder the protein-carbohydrate interaction. On the other hand, the mutation of groups critical for the interaction can lead to a loss in binding. This type of analysis can be applied to any known co-crystal structures or modeled complexes. By employing a novel carbohydrate representation and comparing different monosaccharides, we find that based on their alignment, monosaccharides can share various structural similarities. As we know that only certain parts of the ligand are involved in binding, these similarities can be used to explain and predict cross-reactivity. Understanding the molecular interactions and conformational similarities of functional carbohydrates can lead to the rational design of glycomimetics, and to the development of libraries of molecules sharing structural similarities with various monosaccharides.

Despite tremendous developments in the field of molecular modeling of carbohydrates and proteins, there are still limitations to the available computational power and the underlying assumptions. The achievement of convergence while performing an MD simulation, has been the

subject of debate, even after an exceptional increase in the computing power and improvements in sampling techniques. The conformations sampled during a simulation depends on the starting structure and precautions need to be taken to ensure that the orientations of atoms form relevant interactions. A small change such as differences in protonation states of histidine can lead to large variations. If a crystal structure is being used as a starting structure, like in the present study, it is recommended to employ programs such as Reduce (provided by AMBERTOOLS) that are capable of optimizing orientations of adjustable groups (ASN, GLN, and HIS side chain orientation), optimize the protonation state of HIS, and add and adjust missing hydrogens to these structures. Water molecules are known to be involved in protein-ligand interactions, therefore it is important to retain crystallized water molecules to study these systems. The process of crystallization can lead to unfavorable contacts, which need to be addressed, consequently, minimization is a necessary step before a simulation. The type of solvent used also affects the course of simulations. While implicit water models are faster, they are not able to mimic all their properties, therefore, even though they are computationally expensive and require multiple considerations as discussed further, explicit water models are favored for accuracy.

Implicit solvent models are widely used to calculate the desolvation energies for the binding site to predict binding affinities of protein-ligand complexes. There is a trade-off between precision and accuracy, as to perform this analysis all the water molecules are usually removed, even the ones stabilizing the protein-ligand interaction. Including water molecules can improve the accuracy at the expense of computational time, and can lead to large variations in the outcomes. The accuracy of this method depends on the quality of the implicit solvent models as well. The interactions energies are also sensitive to the force fields and their parameters, which calculate the electrostatic and van der Waals contributions. As observed in Chapter 6, the presence of

charged residues leads to a large variation in the binding affinity contribution between different implicit models, resulting in uncertainty in the choice of models.

In spite of these limitations, the current study can act as a road map for other investigators to understand the underlying structural features of proteins and carbohydrates that lead to their specificity and can be extended to other biological systems. Along with advancing the fundamental knowledge of interactions involved in protein-carbohydrate binding, this study also provides tools for improved prediction and qualification of models.

REFERENCES


1. Mitra N, Sinha S, Ramya TN, Surolia A. N-linked oligosaccharides as outfitters for glycoprotein folding, form and function. Trends in biochemical sciences. 2006;31(3):156-63.

2. Petsko GA, Ringe D. Control of Protein Function. In: Lawrence E, Robertson M, editors. Protein Structure and function: New Science Press Ltd.; 2004.

3. Lee HS, Qi Y, Im W. Effects of N-glycosylation on protein conformation and dynamics: Protein Data Bank analysis and molecular dynamics simulation study. Scientific reports. 2015;5:8926.

4. Akiyama SK, Yamada SS, Yamada KM. Analysis of the role of glycosylation of the human fibronectin receptor. Journal of Biological Chemistry. 1989;264(30):18011-8.

5. Sharon N, Lis H. Lectins as cell recognition molecules. Science. 1989;246(4927):227.

6. Opdenakker G, Rudd PM, Ponting CP, Dwek RA. Concepts and principles of glycobiology. The FASEB Journal. 1993;7(14):1330-7.

7. Zheng M, Fang H, Tsuruoka T, Tsuji T, Sasaki T, Hakomori S. Regulatory role of GM3 ganglioside in alpha 5 beta 1 integrin receptor for fibronectin-mediated adhesion of FUA169 cells. Journal of Biological Chemistry. 1993;268(3):2217-22.

8. Zachara NE, Hart GW. Cell signaling, the essential role of O-GlcNAc! Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids. 2006;1761(5–6):599-617.

9. Disney MD, Seeberger PH. The Use of Carbohydrate Microarrays to Study Carbohydrate-Cell Interactions and to Detect Pathogens. Chemistry & Biology.11(12):1701-7.

10. Karlsson KA. Bacterium-host protein-carbohydrate interactions and pathogenicity. Biochemical Society Transactions. 1999;27(4):471.

11. Sharon N. Carbohydrates as future anti-adhesion drugs for infectious diseases. Biochimica et Biophysica Acta (BBA) - General Subjects. 2006;1760(4):527-37.

12. Freeze HH. Update and perspectives on congenital disorders of glycosylation. Glycobiology. 2001;11(12):129R-43R.

13. Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA. The Impact of Glycosylation on the Biological Function and Structure of Human Immunoglobulins. Annual Review of Immunology. 2007;25(1):21-50.

14. Brooks SA, Carter TM, Royle L, Harvey DJ, Fry SA, Kinch C, et al. Altered Glycosylation of Proteins in Cancer: What Is the Potential for New Anti-Tumour Strategies. Anti-Cancer Agents in Medicinal Chemistry. 2008;8(1):2-21.

15. Ching CK, Black R, Helliwell T, Savage A, Barr H, Rhodes JM. Use of lectin histochemistry in pancreatic cancer. Journal of Clinical Pathology. 1988;41(3):324-8.

16. Ambrosi M, Cameron NR, Davis BG. Lectins: tools for the molecular understanding of the glycocode. Organic & Biomolecular Chemistry. 2005;3(9):1593-608.

17. Davis BG, Robinson MA. Drug delivery systems based on sugar-macromolecule conjugates. Curr Opin Drug Discov Devel. 2002;5(2):279-88.

18. Kim CU, Lew W, Williams MA, Liu H, Zhang L, Swaminathan S, et al. Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site: design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity. Journal of the American Chemical Society. 1997;119(4):681-90.

19. Rao VSR, Qasba PK, Balaji PV, Chandrasekaran R. Conformation of Carbohydrates. Amsterdam, The Netherlands: Harwood Academic; 1998.

20. Ionescu AR, Bérces A, Zgierski MZ, Whitfield DM, Nukada T. Conformational Pathways of Saturated Six-Membered Rings. A Static and Dynamical Density Functional Study. The Journal of Physical Chemistry A. 2005;109(36):8096-105.

21. Kubik S. Carbohydrate recognition: A minimalistic approach to binding. Nat Chem. 2012;4(9):697-8.

22. Rini JM. Lectin structure. Annual review of biophysics and biomolecular structure. 1995;24(1):551-77.

23. Brandl M, Weiss MS, Jabs A, Sühnel J, Hilgenfeld R. CH··· π-interactions in proteins. Journal of molecular biology. 2001;307(1):357-77.

24. Wong C-H. Carbohydrate-based drug discovery: John Wiley & Sons; 2003.

25. Jeffrey G, Pople J, Binkley Jt, Vishveshwara S. Application of ab initio molecular orbital calculations to the structural moieties of carbohydrates. 3. Journal of the American Chemical Society. 1978;100(2):373-9.

26. Kirschner KN, Woods RJ. Solvent interactions determine carbohydrate conformation. Proceedings of the National Academy of Sciences. 2001;98(19):10541-5.

27. Fadda E, Woods RJ. Molecular simulations of carbohydrates and protein–carbohydrate interactions: motivation, issues and prospects. Drug discovery today. 2010;15(15):596-609.

28. Drickamer K. Engineering galactose-binding activity into a C-type mannose-binding protein. 1992.

29. Laughrey ZR, Kiehna SE, Riemen AJ, Waters ML. Carbohydrate− π interactions: What are they worth? Journal of the American Chemical Society. 2008;130(44):14625-33.

30. De Lucca GV, Erickson-Viitanen S, Lam PY. Cyclic HIV protease inhibitors capable of displacing the active site structural water molecule. Drug Discovery Today. 1997;2(1):6-18.

31. Olsson TS, Williams MA, Pitt WR, Ladbury JE. The thermodynamics of protein–ligand interaction and solvation: insights for ligand design. Journal of molecular biology. 2008;384(4):1002-17.

32. Chia-en AC, Chen W, Gilson MK. Ligand configurational entropy and protein binding. Proceedings of the National Academy of Sciences. 2007;104(5):1534-9.

33. Mobley DL, Dill KA. Binding of small-molecule ligands to proteins:"what you see" is not always "what you get". Structure. 2009;17(4):489-98.

34. Baron R, McCammon JA. (Thermo) dynamic role of receptor flexibility, entropy, and motional correlation in protein–ligand binding. ChemPhysChem. 2008;9(7):983-8.

35. Magnani JL, Ernst B. Glycomimetic drugs-a new source of therapeutic opportunities. Discovery medicine. 2009;8(43):247-52.

36. Foley BL, Tessier MB, Woods RJ. Carbohydrate force fields. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2012;2(4):652-97.

37. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. Journal of the American Chemical Society. 1995;117(19):5179-97.

38. Shi Y, Xia Z, Zhang J, Best R, Wu C, Ponder JW, et al. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. Journal of Chemical Theory and Computation. 2013;9(9):4046-63.

39. Dror RO, Dirks RM, Grossman JP, Xu H, Shaw DE. Biomolecular Simulation: A Computational Microscope for Molecular Biology. Annual Review of Biophysics. 2012;41(1):429-52.

40. Wereszczynski J, McCammon JA. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. Quarterly reviews of biophysics. 2012;45(1):1-25.

41. Piana S, Klepeis JL, Shaw DE. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Current Opinion in Structural Biology. 2014;24:98-105.

42. Ryckaert J-P, Ciccotti G, Berendsen HJ. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics. 1977;23(3):327-41.

43. Verlet L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. Physical Review. 1967;159(1):98-103.

44. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics. 1983;79(2):926.

45. Mahoney MW, Jorgensen WL. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. The Journal of Chemical Physics. 2000;112(20):8910.

46. Arfken G. The method of steepest descents. Mathematical methods for physicists. 1985;3:428-36.

47. Hestenes MR, Eduard S. Methods of conjugate gradients for solving linear systems. J Res Natl Bur Stand 1952;49:409-36.

48. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics. 1953;21(6):1087-92.

49. Hukushima K, Nemoto K. Exchange Monte Carlo Method and Application to Spin Glass Simulations. Journal of the Physical Society of Japan. 1996;65(6):1604-8.

50. Katzgraber HG, Trebst S, Huse DA, Troyer M. Feedback-optimized parallel tempering Monte Carlo. Journal of Statistical Mechanics: Theory and Experiment. 2006;2006(03):P03018.

51. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by Simulated Annealing. Science. 1983;220(4598):671.

52. Landau DP, Tsai S-H, Exler M. A new approach to Monte Carlo simulations in statistical physics: Wang-Landau sampling. American Journal of Physics. 2004;72(10):1294-302.

53. Wang F, Landau D. Efficient, multiple-range random walk algorithm to calculate the density of states. Phys Rev Lett. 2001;86(10):2050.

54. Gervais C, Wüst T, Landau DP, Xu Y. Application of the Wang–Landau algorithm to the dimerization of glycophorin A. The Journal of Chemical Physics. 2009;130(21):215106.

55. Vogel T, Li YW, Wüst T, Landau DP. Generic, hierarchical framework for massively parallel Wang-Landau sampling. Phys Rev Lett. 2013;110(21):210603.

56. Hadden JA, Tessier MB, Fadda E, Woods RJ. Calculating Binding Free Energies for Protein–Carbohydrate Complexes. In: Lütteke T, Frank M, editors. Glycoinformatics. New York, NY: Springer New York; 2015. p. 431-65.

57. Liu Z, Zhang Y. Molecular dynamics simulations and MM–PBSA calculations of the lectin from snowdrop (Galanthus nivalis). Journal of Molecular Modeling. 2009;15(12):1501.

58. Yao J, Nellas RB, Glover MM, Shen T. Stability and Sugar Recognition Ability of Ricin-like Carbohydrate Binding Domains. Biochemistry. 2011;50(19):4097-104.

59. Wang J, Tan C, Tan Y-H, Lu Q, Luo R. Poisson-Boltzmann solvents in molecular dynamics simulations. Communications in Computational Physics. 2008;3:1010-31.

60. Kadirvelraj R, Foley BL, Dyekjær JD, Woods RJ. Involvement of Water in Carbohydrate-Protein Binding: Concanavalin A Revisited. Journal of the American Chemical Society. 2008;130(50):16933-42.

61. Kadirvelraj R, Grant OC, Goldstein IJ, Winter HC, Tateno H, Fadda E, et al. Structure and binding analysis of Polyporus squamosus lectin in complex with the Neu5Acα2-6Galβ1-4GlcNAc human-type influenza receptor. Glycobiology. 2011;21(7):973-84.

62. Charvátová O, Foley BL, Bern MW, Sharp JS, Orlando R, Woods RJ. Quantifying Protein Interface Footprinting by Hydroxyl Radical Oxidation and Molecular Dynamics Simulation: Application to Galectin-1. Journal of the American Society for Mass Spectrometry. 2008;19(11):1692-705.

63. Lamb R, Parks G. Paramyxoviridae: the viruses and their replication, p 1449–1496. Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE (ed), Fields virology. Lippincott Williams & Wilkins, Philadelphia, PA; 2007.

64. Lamb RA, Jardetzky TS. Structural basis of viral invasion: lessons from paramyxovirus F. Current opinion in structural biology. 2007;17(4):427-36.

65. Russell CJ, Luque LE. The structural basis of paramyxovirus invasion. Trends in Microbiology.14(6):243-6.

66. Bossart KN, Fusco DL, Broder CC. Paramyxovirus Entry. In: Pöhlmann S, Simmons G, editors. Viral Entry into Host Cells. New York, NY: Springer New York; 2013. p. 95-127.

67. Aguilar HC, Matreyek KA, Filone CM, Hashimi ST, Levroney EL, Negrete OA, et al. N-Glycans on Nipah Virus Fusion Protein Protect against Neutralization but Reduce Membrane Fusion and Viral Entry. Journal of Virology. 2006;80(10):4878-89.

68. Bishop KA, Hickey AC, Khetawat D, Patch JR, Bossart KN, Zhu Z, et al. Residues in the Stalk Domain of the Hendra Virus G Glycoprotein Modulate Conformational Changes Associated with Receptor Binding. Journal of Virology. 2008;82(22):11398-409.

69. Plemper RK, Hammond AL, Gerlier D, Fielding AK, Cattaneo R. Strength of Envelope Protein Interaction Modulates Cytopathicity of Measles Virus. Journal of Virology. 2002;76(10):5051-61.

70. Melanson VR, Iorio RM. Amino Acid Substitutions in the F-Specific Domain in the Stalk of the Newcastle Disease Virus HN Protein Modulate Fusion and Interfere with Its Interaction with the F Protein. Journal of Virology. 2004;78(23):13053-61.

71. Bose S, Zokarkar A, Welch BD, Leser GP, Jardetzky TS, Lamb RA. Fusion activation by a headless parainfluenza virus 5 hemagglutinin-neuraminidase stalk suggests a modular mechanism for triggering. Proceedings of the National Academy of Sciences. 2012;109(39):E2625-E34.

72. Bose S, Welch BD, Kors CA, Yuan P, Jardetzky TS, Lamb RA. Structure and Mutagenesis of the Parainfluenza Virus 5 Hemagglutinin-Neuraminidase Stalk Domain Reveals a Four-Helix Bundle and the Role of the Stalk in Fusion Promotion. Journal of Virology. 2011;85(24):12855-66.

73. Welch BD, Yuan P, Bose S, Kors CA, Lamb RA, Jardetzky TS. Structure of the Parainfluenza Virus 5 (PIV5) Hemagglutinin-Neuraminidase (HN) Ectodomain. PLoS Pathog. 2013;9(8):e1003534.

74. Yuan P, Thompson TB, Wurzburg BA, Paterson RG, Lamb RA, Jardetzky TS. Structural Studies of the Parainfluenza Virus 5 Hemagglutinin-Neuraminidase Tetramer in Complex with Its Receptor, Sialyllactose. Structure.13(5):803-15.

75. Yin H-S, Wen X, Paterson RG, Lamb RA, Jardetzky TS. Structure of the parainfluenza virus 5 F protein in its metastable, prefusion conformation. Nature. 2006;439(7072):38-44.

76. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. 2006;22(2):195-201.

77. Vriend G. WHAT IF: A molecular modeling and drug design program. Journal of Molecular Graphics. 1990;8(1):52-6.

78. Jorgensen WL, Madura JD. Solvation and conformation of methanol in water. J Am Chem Soc. 1983;105:1407-13.

79. Case D, Darden T, Cheatham Iii T, Simmerling C, Wang J, Duke R, et al. AMBER 10. University of California, San Francisco. 2008;32.

80. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins: Structure, Function, and Bioinformatics. 2006;65(3):712-25.

81. Götz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. Journal of Chemical Theory and Computation. 2012;8(5):1542-55.

82. Darden T, York D, Pedersen L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. The Journal of Chemical Physics. 1993;98(12):10089.

83. Hubbard SJ, Thornton JM. Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London. 1993;2(1).

84. McLellan JS, Chen M, Leung S, Graepel KW, Du X, Yang Y, et al. Structure of RSV Fusion Glycoprotein Trimer Bound to a Prefusion-Specific Neutralizing Antibody. Science (New York, NY). 2013;340(6136):1113-7.

85. Corti D, Bianchi S, Vanzetta F, Minola A, Perez L, Agatic G, et al. Cross-neutralization of four paramyxoviruses by a human monoclonal antibody. Nature. 2013;501(7467):439-43.

86. Wen X, Krause JC, Leser GP, Cox RG, Lamb RA, Williams J, et al. Structure of the Human Metapneumovirus Fusion Protein with Neutralizing Antibody Identifies a Pneumovirus Antigenic Site. Nature structural & molecular biology. 2012;19(4):461-3.

87. Randall RE, Young DF, Goswami KKA, Russell WC. Isolation and Characterization of Monoclonal Antibodies to Simian Virus 5 and Their Use in Revealing Antigenic Differences between Human, Canine and Simian Isolates. Journal of General Virology. 1987;68(11):2769.

88. Bose S, Heath CM, Shah PA, Alayyoubi M, Jardetzky TS, Lamb RA. Mutations in the Parainfluenza Virus 5 Fusion Protein Reveal Domains Important for Fusion Triggering and Metastability. Journal of Virology. 2013;87(24):13520-31.

89. Apte-Sengupta S, Negi S, Leonard VHJ, Oezguen N, Navaratnarajah CK, Braun W, et al. Base of the Measles Virus Fusion Trimer Head Receives the Signal That Triggers Membrane Fusion. The Journal of Biological Chemistry. 2012;287(39):33026-35.

90. Cross TA, Sharma M, Yi M, Zhou H-X. Influence of solubilizing environments on membrane protein structures. Trends in biochemical sciences. 2011;36(2):117-25.

91. Barone G, Gomez-Paloma L, Duca D, Silvestri A, Riccio R, Bifulco G. Structure Validation of Natural Products by Quantum-Mechanical GIAO Calculations of 13C NMR Chemical Shifts. Chemistry–A European Journal. 2002;8(14):3233-9.

92. Kobayashi N, Iwahara J, Koshiba S, Tomizawa T, Tochio N, Güntert P, et al. KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR structure studies. Journal of biomolecular NMR. 2007;39(1):31-52.

93. Frueh DP, Goodrich AC, Mishra SH, Nichols SR. NMR methods for structural studies of large monomeric and multimeric proteins. Current opinion in structural biology. 2013;23(5):734-9.

94. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round x. Proteins: Structure, Function, and Bioinformatics. 2014;82(S2):1-6.

95. Misura KMS, Baker D. Progress and challenges in high-resolution refinement of protein structure models. Proteins: Structure, Function, and Bioinformatics. 2005;59(1):15--29.

96. Xu Y, Xu D, Crawford OH, Einstein JR. A computational method for NMR-constrained protein threading. J Comput Biol. 2000;7(3-4):449-67.

97. Xu Y, Xu D. Protein threading using PROSPECT: Design and evaluation. Proteins: Structure, Function, and Bioinformatics. 2000;40(3):343--54.

98. Huang W, Ravikumar KM, Chance MR, Yang S. Quantitative mapping of protein structure by hydroxyl radical footprinting-mediated structural mass spectrometry: a protection factor analysis. Biophysical journal. 2015;108(1):107-15.

99. Sharp JS, Guo J-t, Uchiki T, Xu Y, Dealwis C, Hettich RL. Photochemical surface mapping of C14S-Sml1p for constrained computational modeling of protein structure. Analytical biochemistry. 2005;340(2):201-12.

100. Sharp JS, Becker JM, Hettich RL. Analysis of Protein Solvent Accessible Surfaces by Photochemical Oxidation and Mass Spectrometry. Analytical Chemistry. 2004;76(3):672-83.

101. Sharp JS, Becker JM, Hettich RL. Protein surface mapping by chemical oxidation: Structural analysis by mass spectrometry. Analytical Biochemistry. 2003;313(2):216-25.

102. Goldsmith SC, Guan J-Q, Almo SC, Chance MR. Synchrotron protein footprinting: a technique to investigate protein-protein interactions. Journal of Biomolecular Structure and Dynamics. 2001;19(3):405-18.

103. Chance MR. Unfolding of apomyoglobin examined by synchrotron footprinting. Biochemical and biophysical research communications. 2001;287(3):614-21.

104. Kiselar JG, Janmey PA, Almo SC, Chance MR. Visualizing the Ca2+-dependent activation of gelsolin by using synchrotron footprinting. Proceedings of the National Academy of Sciences. 2003;100(7):3942-7.

105. Fenton HJH. LXXIII.-Oxidation of tartaric acid in presence of iron. Journal of the Chemical Society, Transactions. 1894;65(0):899-910.

106. Shcherbakova I, Mitra S, Beer RH, Brenowitz M. Fast Fenton footprinting: a laboratory-based method for the time-resolved analysis of DNA, RNA and proteins. Nucleic Acids Research. 2006;34(6):e48-e.

107. Shcherbakova I, Brenowitz M. Monitoring structural changes in nucleic acids with single residue spatial and millisecond time resolution by quantitative hydroxyl radical footprinting. Nat Protocols. 2008;3(2):288-302.

108. Berens C, Streicher B, Schroeder R, Hillen W. Visualizing metal-ion-binding sites in group I introns by iron(II)-mediated Fenton reactions. Chemistry & Biology. 1998;5(3):163-75.

109. Franchet-Beuzit J, Spotheim-Maurizot M, Sabattier R, Blazy-Baudras B, Charlier M. Radiolytic footprinting. .beta. Rays, .gamma. photons, and fast neutrons probe DNA-protein interactions. Biochemistry. 1993;32(8):2104-10.

110. Armstrong RC, Swallow AJ. Pulse- and Gamma-Radiolysis of Aqueous Solutions of Tryptophan. Radiation Research. 1969;40(3):563-79.

111. Kopoldova J, Hrncir S. Gamma-radiolysis of aqueous solution of histidine. Journal of biosciences. 1977;32(7/8):482-7.

112. Winchester RV, Lynn KR. X- and γ-radiolysis of Some Tryptophan Dipeptides. International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine. 1970;17(6):541-8.

113. Maleknia SD, Ralston CY, Brenowitz MD, Downard KM, Chance MR. Determination of Macromolecular Folding and Structure by Synchrotron X-Ray Radiolysis Techniques. Analytical Biochemistry. 2001;289(2):103-15.

114. Guan J-Q, Vorobiev S, Almo SC, Chance MR. Mapping the G-Actin Binding Surface of Cofilin Using Synchrotron Protein Footprinting. Biochemistry. 2002;41(18):5765-75.

115. Hambly DM, Gross ML. Laser Flash Photolysis of Hydrogen Peroxide to Oxidize Protein Solvent-Accessible Residues on the Microsecond Timescale. Journal of the American Society for Mass Spectrometry. 2005;16(12):2057-63.

116. Aye TT, Low TY, Sze SK. Nanosecond Laser-Induced Photochemical Oxidation Method for Protein Surface Mapping with Mass Spectrometry. Analytical Chemistry. 2005;77(18):5814-22.

117. Xu G, Takamoto K, Chance MR. Radiolytic Modification of Basic Amino Acid Residues in Peptides: Probes for Examining Protein−Protein Interactions. Analytical Chemistry. 2003;75(24):6995-7007.

118. Xu G, Chance MR. Radiolytic Modification of Sulfur-Containing Amino Acid Residues in Model Peptides: Fundamental Studies for Protein Footprinting. Analytical Chemistry. 2005;77(8):2437-49.

119. Takamoto K, Chance MR. RADIOLYTIC PROTEIN FOOTPRINTING WITH MASS SPECTROMETRY TO PROBE THE STRUCTURE OF MACROMOLECULAR COMPLEXES. Annual Review of Biophysics and Biomolecular Structure. 2006;35(1):251-76.

120. Xu G, Chance MR. Radiolytic Modification of Acidic Amino Acid Residues in Peptides: Probes for Examining Protein−Protein Interactions. Analytical Chemistry. 2004;76(5):1213-21.

121. Xu G, Chance MR. Radiolytic Modification and Reactivity of Amino Acid Residues Serving as Structural Probes for Protein Footprinting. Analytical Chemistry. 2005;77(14):4549-55.

122. Xu G, Chance MR. Hydroxyl Radical-Mediated Modification of Proteins as Probes for Structural Proteomics. Chemical Reviews. 2007;107(8):3514-43.

123. Mendoza VL, Vachet RW. Probing Protein Structure by Amino Acid-Specific Covalent Labeling and Mass Spectrometry. Mass spectrometry reviews. 2009;28(5):785-815.

124. Sharp JS, Sullivan DM, Cavanagh J, Tomer KB. Measurement of multisite oxidation kinetics reveals an active site conformational change in Spo0F as a result of protein oxidation. Biochemistry. 2006;45(20):6260-6.

125. Sharp JS, Tomer KB. Analysis of the oxidative damage-induced conformational changes of apo-and holocalmodulin by dose-dependent protein oxidative surface mapping. Biophysical journal. 2007;92(5):1682-92.

126. Venkatesh S, Tomer KB, Sharp JS. Rapid identification of oxidation-induced conformational changes by kinetic analysis. Rapid Communications in Mass Spectrometry. 2007;21(23):3927--36.

127. Watson C, Sharp JS. Conformational Analysis of Therapeutic Proteins by Hydroxyl Radical Protein Footprinting. The AAPS Journal. 2012;14(2):206-17.

128. Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, et al. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. The Journal of Physical Chemistry. 1992;96(15):6472-84.

129. Eisenmenger F, Hansmann UH, Hayryan S, Hu C-K. An enhanced version of SMMP—open-source software package for simulation of proteins. Computer physics communications. 2006;174(5):422-9.

130. Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M. The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. Journal of Computational Chemistry. 1995;16(3):273-84.

131. Diamond R. Real-space refinement of the structure of hen egg-white lysozyme. Journal of molecular biology. 1974;82(3):371IN5375-374IN11391.

132. Evans SV, Brayer GD. High-resolution study of the three-dimensional structure of horse heart metmyoglobin. Journal of molecular biology. 1990;213(4):885-97.

133. Case D, Darden T, Cheatham III T, Simmerling C, Wang J, Duke R, et al. AMBER 12; University of California: San Francisco, 2012. There is no corresponding record for this reference.

134. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. Journal of chemical theory and computation. 2015;11(8):3696-713.

135. Loncharich RJ, Brooks BR, Pastor RW. Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N′-methylamide. Biopolymers. 1992;32(5):523-35.

136. Wang B, Boons G-J. Carbohydrate Recognition: Biological Problems, Methods, and Applications. Wang B, editor: Wiley; 2011. 448 pages p.

137. Lasky LA. Selectins: interpreters of cell-specific carbohydrate information during inflammation. Science. 1992;258(5084):964.

138. Perillo NL, Pace KE, Seilhamer JJ, Baum LG. Apoptosis of T cells mediated by galectin-1. Nature. 1995;378(6558):736-9.

139. Haltiwanger RS, Lowe JB. Role of Glycosylation in Development. Annual Review of Biochemistry. 2004;73(1):491-537.

140. Brown GD, Gordon S. Immune recognition: A new receptor for [beta]-glucans. Nature. 2001;413(6851):36-7.

141. Cobb BA, Kasper DL. Coming of age: carbohydrates and immunity. European Journal of Immunology. 2005;35(2):352-6.

142. Nakahara S, Raz A. Biological Modulation by Lectins and Their Ligands in Tumor Progression and Metastasis. Anti-cancer agents in medicinal chemistry. 2008;8(1):22-36.

143. Elola M, Wolfenstein-Todel C, Troncoso M, Vasta G, Rabinovich G. Galectins: matricellular glycan-binding proteins linking cell adhesion, migration, and survival. Cellular and Molecular Life Sciences. 2007;64(13):1679-700.

144. Monsigny M, Mayer R, Roche A-C. Sugar-lectin interactions: sugar clusters, lectin multivalency and avidity. Carbohydrate letters. 2000;4:35-52.

145. Grant OC, Smith HM, Firsova D, Fadda E, Woods RJ. Presentation, presentation, presentation! Molecular-level insight into linker effects on glycan array screening data. Glycobiology. 2014;24(1):17-25.

146. Manimala JC, Roach TA, Li Z, Gildersleeve JC. High-Throughput Carbohydrate Microarray Analysis of 24 Lectins. Angewandte Chemie International Edition. 2006;45(22):3607-10.

147. Manimala JC, Roach TA, Li Z, Gildersleeve JC. High-throughput carbohydrate microarray profiling of 27 antibodies demonstrates widespread specificity problems. Glycobiology. 2007;17(8):17C-23C.

148. Cummings RD. [6] Use of lectins in analysis of glycoconjugates. Methods in Enzymology. Volume 230: Academic Press; 1994. p. 66-86.

149. Ambepitiya Wickramasinghe IN, de Vries RP, Weerts EAWS, van Beurden SJ, Peng W, McBride R, et al. Novel Receptor Specificity of Avian Gammacoronaviruses That Cause Enteritis. Journal of Virology. 2015;89(17):8783-92.

150. Tessier MB, Grant OC, Heimburg-Molinaro J, Smith D, Jadey S, Gulick AM, et al. Computational Screening of the Human TF-Glycome Provides a Structural Definition for the Specificity of Anti-Tumor Antibody JAA-F11. PLoS ONE. 2013;8(1):e54874.

151. Sauter NK, Bednarski MD, Wurzburg BA, Hanson JE, Whitesides GM, Skehel JJ, et al. Hemagglutinins from two influenza virus variants bind to sialic acid derivatives with millimolar dissociation constants: a 500-MHz proton nuclear magnetic resonance study. Biochemistry. 1989;28(21):8388-96.

152. Xiong X, Coombs PJ, Martin SR, Liu J, Xiao H, McCauley JW, et al. Receptor binding by a ferret-transmissible H5 avian influenza virus. Nature. 2013;497(7449):392-6.

153. Yang H, Carney PJ, Chang JC, Villanueva JM, Stevens J. Structural analysis of the hemagglutinin from the recent 2013 H7N9 influenza virus. Journal of virology. 2013:JVI. 01854-13.

154. Xu R, de Vries RP, Zhu X, Nycholat CM, McBride R, Yu W, et al. Preferential recognition of avian-like receptors in human influenza A H7N9 viruses. Science. 2013;342(6163):1230-5.

155. Cao Z, Partyka K, McDonald M, Brouhard E, Hincapie M, Brand RE, et al. Modulation of glycan detection on specific glycoproteins by lectin multimerization. Analytical chemistry. 2013;85(3):1689-98.

156. Vyas NK. Atomic features of protein-carbohydrate interactions. Current Opinion in Structural Biology. 1991;1(5):732-40.

157. Lee YC, Lee RT. Carbohydrate-protein interactions: basis of glycobiology. Accounts of chemical research. 1995;28(8):321-7.

158. Xu R, McBride R, Nycholat CM, Paulson JC, Wilson IA. Structural characterization of the hemagglutinin receptor specificity from the 2009 H1N1 influenza pandemic. Journal of virology. 2012;86(2):982-90.

159. Sun H, Li Y, Tian S, Xu L, Hou T. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. Physical Chemistry Chemical Physics. 2014;16(31):16719-29.

160. Svensson C, Teneberg S, Nilsson CL, Kjellberg A, Schwarz FP, Sharon N, et al. High-resolution Crystal Structures of Erythrina cristagalli Lectin in Complex with Lactose and 2′-α-l-Fucosyllactose and Correlation with Thermodynamic Binding Data. Journal of Molecular Biology. 2002;321(1):69-83.

161. Turton K, Natesh R, Thiyagarajan N, Chaddock JA, Acharya KR. Crystal structures of Erythrina cristagalli lectin with bound N-linked oligosaccharide and lactose. Glycobiology. 2004;14(10):923-9.

162. Imamura K, Takeuchi H, Yabe R, Tateno H, Hirabayashi J. Engineering of the glycan-binding specificity of Agrocybe cylindracea galectin towards α (2, 3)-linked sialic acid by saturation mutagenesis. Journal of biochemistry. 2011;150(5):545-52.

163. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. HKL-3000: the integration of data reduction and structure solution–from diffraction images to an initial model in minutes. Acta Crystallographica Section D: Biological Crystallography. 2006;62(8):859-66.

164. Collaborative CP. The CCP4 suite: programs for protein crystallography. Acta crystallographica Section D, Biological crystallography. 1994;50(Pt 5):760.

165. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D: Biological Crystallography. 2010;66(2):213-21.

166. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic acids research. 2007;35(suppl 2):W375-W83.

167. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallographica Section D: Biological Crystallography. 2010;66(4):486-501.

168. Schlitter J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. Chemical Physics Letters. 1993;215(6):617-21.

169. Xue W, Yang Y, Wang X, Liu H, Yao X. Computational study on the inhibitor binding mode and allosteric regulation mechanism in hepatitis C virus NS3/4A protein. PloS one. 2014;9(2):e87077.

170. Grant OC, Tessier MB, Meche L, Mahal LK, Foley BL, Woods RJ. Combining 3D structure with glycan array data provides insight into the origin of glycan specificity. Glycobiology. 2016;26(7):772-83.

171. Kirschner KN, Yongye AB, Tschampel SM, González-Outeiriño J, Daniels CR, Foley BL, et al. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. Journal of computational chemistry. 2008;29(4):622-55.

172. Karplus M, Kushick JN. Method for estimating the configurational entropy of macromolecules. Macromolecules. 1981;14(2):325-32.

173. Case DA. Normal mode analysis of protein dynamics. Current Opinion in Structural Biology. 1994;4(2):285-90.

174. Gohlke H, Case DA. Converging free energy estimates: MM-PB (GB) SA studies on the protein–protein complex Ras–Raf. Journal of computational chemistry. 2004;25(2):238-50.

175. Genheden S, Kuhn O, Mikulskis P, Hoffmann D, Ryde U. The normal-mode entropy in the MM/GBSA method: effect of system truncation, buffer region, and dielectric constant. Journal of chemical information and modeling. 2012;52(8):2079-88.

176. Adar R, Sharon N. Mutational Studies of the Amino Acid Residues in the Combining Site of Erythrina corallodendron Lectin. European Journal of Biochemistry. 1996;239(3):668-74.

177. Eid S, Saleh N, Zalewski A, Vedani A. Exploring the free-energy landscape of carbohydrate–protein complexes: development and validation of scoring functions considering the binding-site topology. Journal of computer-aided molecular design. 2014;28(12):1191-204.

178. Varki A. Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins. Nature. 2007;446(7139):1023-9.

179. An HJ, Kronewitter SR, de Leoz MLA, Lebrilla CB. Glycomics and disease markers. Current opinion in chemical biology. 2009;13(5):601-7.

180. Ardá A, Blasco P, Varón Silva D, Schubert V, André S, Bruix M, et al. Molecular recognition of complex-type biantennary N-glycans by protein receptors: a three-dimensional view on epitope selection by NMR. Journal of the American Chemical Society. 2013;135(7):2667-75.

181. Sood A, Ji Y, Gerlits OO, Woods RJ. Quantifying functional group contributions to understanding protein-carbohydrate affinity. In preparation.

182. Varki A, Cummings RD, Aebi M, Packer NH, Seeberger PH, Esko JD, et al. Symbol nomenclature for graphical representations of glycans. Glycobiology. 2015;25(12):1323-4.

183. Tanaka K, Aoki-Kinoshita KF, Kotera M, Sawaki H, Tsuchiya S, Fujita N, et al. WURCS: the Web3 unique representation of carbohydrate structures. Journal of chemical information and modeling. 2014;54(6):1558-66.

184. Anderson E, Veith GD, Weininger D. SMILES, a line notation and computerized interpreter for chemical structures: US Environmental Protection Agency, Environmental Research Laboratory; 1987.

185. Ash S, Cline MA, Homer RW, Hurst T, Smith GB. SYBYL line notation (SLN): A versatile language for chemical structure representation. Journal of chemical information and computer sciences. 1997;37(1):71-9.

186. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI-the worldwide chemical structure identifier standard. Journal of cheminformatics. 2013;5(1):1.

187. Boraston AB, Creagh AL, Alam MM, Kormos JM, Tomme P, Haynes CA, et al. Binding specificity and thermodynamics of a family 9 carbohydrate-binding module from Thermotoga maritima xylanase 10A. Biochemistry. 2001;40(21):6240-7.

# APPENDIX

# PROBING THE PARAMYXOVIRUS FUSION (F) PROTEIN-REFOLDING EVENT FROM PRE- TO POSTFUSION BY OXIDATIVE FOOTPRINTING


## Abstract

To infect a cell, the *Paramyxoviridae* family of enveloped viruses relies on the coordinated action of a receptor-binding protein (variably HN, H, or G) and a more conserved metastable fusion protein (F) to effect membrane fusion and allow genomic transfer. Upon receptor binding, HN (H or G) triggers F to undergo an extensive refolding event to form a stable postfusion state. Little is known about the intermediate states of the F refolding process. Here, a soluble form of parainfluenza virus 5 F was triggered to refold using temperature and was footprinted along the refolding pathway using fast photochemical oxidation of proteins (FPOP). Localization of the oxidative label to solvent-exposed side chains was determined by high-resolution MS/MS. Globally, metastable prefusion F is oxidized more extensively than postfusion F, indicating that the prefusion state is more exposed to solvent and is more flexible. Among the first peptides to be oxidatively labeled after temperature-induced triggering is the hydrophobic fusion peptide. A comparison of peptide oxidation levels with the values of solvent-accessible surface area calculated from molecular dynamics simulations of available structural data reveals regions of the F protein that lie at the heart of its prefusion metastability. The strong correlation between the regions of F that experience greater-than-expected oxidative labeling and epitopes for neutralizing antibodies suggests that FPOP has a role in guiding the development of targeted

therapeutics. Analysis of the residue levels of labeled F intermediates provides detailed insights into the mechanics of this critical refolding event.

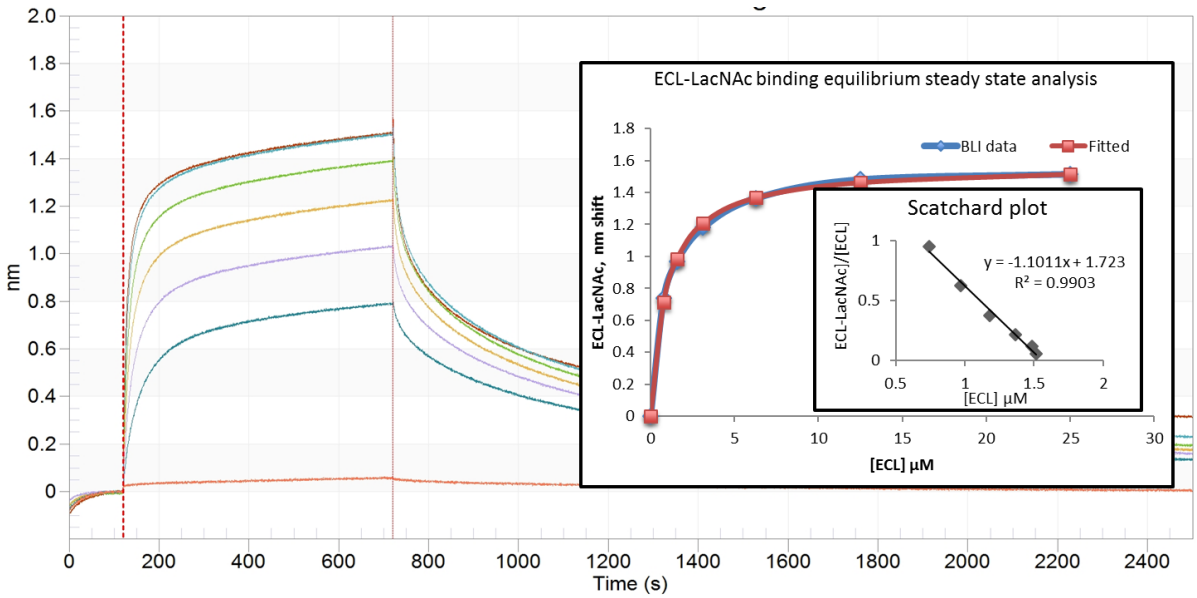



**Figure S6.1.** BLI sensorgram of ECL direct binding to LacNAc on SA biosensors and the KD resulting from steady state analysis and scatchard plot analysis.

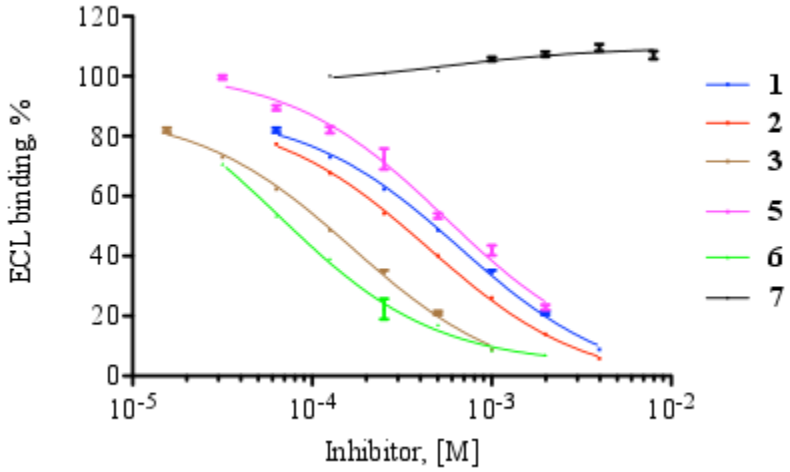



**Figure S6.2.** $IC_{50}$ of oligosaccharides inhibiting ECL binding to LacNAc on SA biosensors.

**Table S6.1.** X-ray crystallographic data-collection and refinement statistics.

| | ECL-**2** | | ECL-**3** |
|---|---|---|---|
| Beamline/Facility | Rigaku HomeLab/ORNL | HighFlux | SBC-CAT 19ID/APS |
| Space group | $P6_5$ | | $P6_5$ |

| | | |
|---|---|---|
| Cell dimensions: | | |
| a, b, c (Å) | 134.950, 134.950, 81.794 | 134.669, 134.669, 81.212 |
| α, β, γ (°) | 90, 90, 120 | 90, 90, 120 |
| Resolution (Å) | 40.00-2.20 (2.28-2.20) | 44.08-1.90 (1.93-1.90) |
| No. reflections measured | 42764 (4263) | 65144 (3227) |
| $R_{merge}$ | 0.085 (0.496) | 0.068 (0.461) |
| $I / \sigma I$ | 13.1 (2.1) | 38.4 (4.4) |
| Completeness (%) | 98.8 (98.8) | 99.2 (98.1) |
| Redundancy | 3.3 (3.1) | 6.7 (5.8) |
| *R*work / *R*free | 0.1814 / 0.2042 | 0.2217 / 0.2636 |
| No. atoms (non-H) | 4142 | 4274 |
| Water | 296 | 394 |
| R.m.s.d. bonds (Å) | 0.003 | 0.007 |
| R.m.s.d. bond angles (°) | 0.684 | 1.188 |

**Table S6.2** $IC_{50}$ of all carbohydrate candidates.

| Ligand | IC50 (mM) |
|---|---|
| **1** | 0.66 (0.04) |
| **2** | 0.44 (0.01) |
| **3** | 0.17 (0.01) |
| **5** | 0.49 (0.05) |
| **6** | 0.07 (0.01) |
| **7** | 0.00 (0.00) |

LIST OF PUBLICATIONS

1. Poor TA, Jones LM, Sood A, Leser GP, Plasencia MD, Rempel DL, Jardetzky TS, Woods RJ, Gross ML, Lamb RA. Probing the paramyxovirus fusion (F) protein-refolding event from pre-to postfusion by oxidative footprinting. Proceedings of the National Academy of Sciences. 2014 Jun 24;111(25):E2596-605.
2. Ng S, Lin E, Kitov PI, Tjhung KF, Gerlits OO, Deng L, Kasper B, Sood A, Paschal BM, Zhang P, Ling CC. Genetically encoded fragment-based discovery of glycopeptide ligands for carbohydrate-binding proteins. Journal of the American Chemical Society. 2015 Apr 16;137(16):5248-51.
3. Xie B, Sood A, Woods RJ, Sharp JS. Quantitative protein topography measurements by fast protein-hydroxyl radical chemistry and mass spectrometry. In preparation.
4. Sood A, Ji Y, Gerlits OO, Woods RJ. Quantifying functional group contributions to understanding protein-carbohydrate affinity. In preparation.
5. Sood A, Foley BL, Woods RJ. Monosaccharide similarity analysis to understand protein-carbohydrate specificity. In preparation.