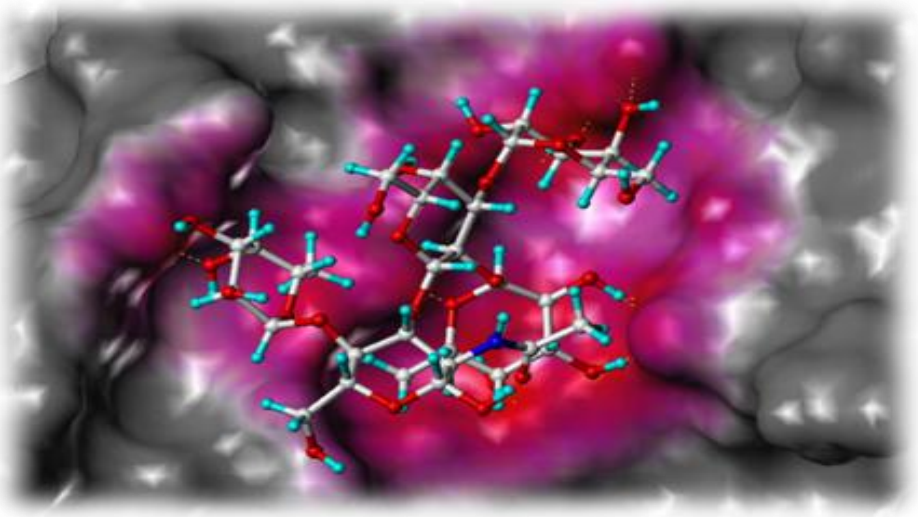
CHALMERS | UNIVERSITY OF GOTHENBURG



A study of norovirus-HBGA interactions

Computational studies on the interaction of norovirus surface protein with ABO- blood group active saccharides

Master of Science Thesis in Bioinformatics and System Biology

WAQAS NASIR

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
Göteborg, Sweden, December 2009

A study of norovirus-HBGA interactions
Computational studies on the interaction of norovirus surface protein with ABO- blood group active saccharides

Waqas Nasir

Examiner: Graham J.L. Kemp

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Cover:
Enlarged view of BLe^b seen in complex with VA387
with a distance threshold of 3 Å (Page 38)

Department of Computer Science and Engineering
Göteborg, Sweden December 2009

Acknowledgments

First of all, I feel obliged to pay huge gratitude to almighty ***God*** with all my capacities for providing me with adequate strength and capabilities to carry out this work.

I would then like to thank my supervisor ***Dr. Per-Georg Nyholm*** for his extensive support and encouragement throughout the course of this thesis work, for the great appreciation over little things and for several fruitful discussions which provided me with all the guidance that I needed to carry out this work. I thank him once again for his enormous support. I feel lucky to have found a supervisor as helpful as him.

A huge amount of gratitude goes to my examiner and co-supervisor ***Dr. Graham Kemp***. His attitude of always willing to help his students has found him numerous admirers like me. It was through his breathtaking lectures on structural bioinformatics that I chose this field as the area of research for my thesis work. I thank him for his support and encouragement throughout the course of this study and for healthy discussions and tips related to this work.

I would like to express many thanks to ***Goran Larson*** and ***Gustaf Rydell*** for several key discussions during the preparation of the manuscript, for all the valuable tips about the biology of this work and for always being open to help out.

I thank the team at Biognos AB, ***Chaitanya Koppisetty*** and ***Francesco Strino*** for all the guidance that they provided me with, particularly in the area of computational bio-chemistry. I thank them once again for always fetching the time for me whenever I needed any help and for providing me with adequate guidelines in connection to the present study. I especially thank Chaitanya for doing the Glide XP scoring for my thesis work.

I am grateful to ***Niklas Nordgren*** for always coming up with valuable comments and questions, for proofreading the whole thesis work chapter by chapter and for continuous moral support during the time of thesis.

In the last but by no means the least, I am enormously thankful with all my heart to my beloved wife ***Salmana*** for providing me with all the love and support always and for making things look so easy. I thank her once again for always being there.

Waqas Nasir

To my mother...

Abstract

Noroviruses are one of the major causes of non-bacterial gastroenteritis epidemics in humans. Different histo-blood group antigens (HBGA) bind to different norovirus strains and are considered as essential attachment factors on the host cell. Among the human HBGAs, the ABO structures and Lewis are involved in norovirus receptor binding. The norovirus strain VA387 is of dominating clinical importance. The surface protein of this strain in complex with the A-trisaccharide and the B-trisaccharide, both sharing the Fucα1-2Galβ moiety, has been crystallized. The α-fucose of the ligands is buried in the binding pocket and is observed to have strong hydrogen bonding interactions with the binding site and hence plays a key role in the binding.

The focus of this study is to understand the observed binding specificity for a series of oligosaccharides sharing the Fucα1-2Galβ moiety and then to predict their relative binding strength. The favoured conformations of the HBGAs were computed using the mm4-based force field method GLYGAL. The prediction of the favoured poses of the HBGAs in the binding site of VA387 was carried out using molecular dynamics simulations. The orientation of the fucose ring of the HBGAs in the binding site was assumed to be similar to the fucose in the case of the reported B-trisaccharide/VA387 complex. Calculations were carried out for type-1, type-2 and type-3 chains of HBGAs. The AMBER suite of programs was used for the molecular dynamics studies with implicit solvent.

The results from the molecular dynamics simulations for each complex were used to calculate the interaction score of the predicted complexes using the Glide XP scoring function. The major finding in the current study is the critically important role of the Fucα1,2 and the terminal Gal/GalNAc in the binding of VA387 with ABO-active saccharides. The results obtained from the present study give deeper insights into the structural details of the VA387 binding with HBGAs and are in good agreement with mutagenesis studies in which key protein residues for the binding of VA387 with ABO-active saccharides have been identified.

Table of contents

1. Introduction

Noroviruses are the major cause of acute non-bacterial gastroenteritis world-wide. Some genogroups of noroviruses infect humans with the disease lasting for 2-3 days characterized by nausea, vomiting, diarrhoea and abdominal cramps. The virus binds to histo-blood group antigens (HBGAs) in intestine of the host. The HBGAs are terminal structures of glycan chains of glycoprotein and glyco-lipid. The research reported here is focused on the use of computational techniques like molecular dynamics (MD) and Glide scoring to gain deeper insights into the structural details of binding of different human HBGAs to the norovirus surface protein.

1.1 Motivation

Among different strains of norovirus, VA387, which belongs to genogroup II and genetic cluster 4 (GII.4), is of dominating clinical importance because of its binding to secretor A, B and H blood group antigens. The norovirus surface protein VA387 of GII.4 strain has been crystallized with the A-trisaccharide and the B-trisaccharide in complex (Cao *et al*., 2007). The crystal structure demonstrated the binding of VA387 to the A- and B-trisaccharides but did not show anything about the binding specificity of different HBGAs to VA387. A year later, mutagenesis studies (Tan *et al*., 2008) however, reported the residues in the binding site of VA387 which were essential for binding to different HBGAs. Previously, the binding studies had shown that different HBGAs bind with different binding affinity to noroviruses (Huang *et al*., 2003; Huang *et al*., 2005) but the structural details of the binding of different HBGAs are not known.

Despite of these key developments in the area of norovirus research, a deeper understanding of the structure and dynamics of norovirus binding to different HBGAs is essential for the design of potential inhibitors to block the binding of HBGAs to noroviruses. Since crystallization of different surface proteins of noroviruses with different HBGAs is a tremendously expensive project both in terms of time and money, a computational approach has been used in the present project.

1.2 Aim

The aim of the present study is to predict the orientation of different HBGAs in the binding site of VA387 and to predict the energy of interaction for the different complexes. Furthermore, comparisons are made with experimental data. The results could be used in the design of potential inhibitors to eventually block the binding of the noroviruses with the carbohydrate ligands.

1.3 Overview of the thesis

The work reported here is organized mainly in two parts with the first part including the biological importance and background of noroviruses and the computational techniques that are used in the thesis work. The second part contains the methodology, results and discussion along with the conclusions and future expansions. The report consists of 5 chapters and two appendices. The organization of report is as follows;

Chapter 1 is an introduction. The chapter includes general introduction to the work that is reported here. The motivation and aims of the thesis work are also described in this chapter.

Chapter 2 is the background of the thesis work. The chapter contains 2 parts. First, the detailed biological background of the noroviruses is discussed which is then followed by a description of the computational techniques that are used.

Chapter 3 discusses the materials and methods that are used in the thesis work. The methodological considerations are described in chronological order for the ease of understanding and the chapter is summarized in the end by a flow chart showing all the methodological steps.

Chapter 4 is dedicated to the results obtained and the following discussion. The chapter gives a detailed description and explanation of results along with the limitations of the thesis work.

Chapter 5 concludes the thesis work with major findings in the current study.

Appendix A contains all the adiabatic maps for glycosidic linkages of the sugars considered in the present study.

Appendix B contains the table of interactions for all ligands considered in the present study with VA387 surface protein. The table was produced using the data from LIGPLOT.

2. Background

This chapter provides the reader with the background knowledge of the current study. The chapter is organized into two parts. The first part contains the biological background of the thesis work whereas the second part contains the theory behind the molecular dynamics (MD) simulations.

2.1 Noroviruses

2.1.1 The norovirus disease

The phrase "Hyperemesis hemis" or "winter vomiting disease" was first coined by Zahorsky in 1929. The illness referred to the sudden onset of vomiting and diarrhea that shoots up in the colder months (Zahorsky 1929). In 1972, after the examination of stools of volunteers through immune electron microscopy (IEM), Kapikian and others discovered the etiology of this disease - a virus which was called Norwalk virus. This is now classified as belonging to the genus *Norovirus* of the Caliciviridae family.

Noroviruses are the main cause of acute non-bacterial gastroenteritis world-wide (Huston *et al*., 2004). As mentioned before, the incidence peaks in the colder months and affects people of all ages. However, recent studies have demonstrated summer and spring outbreaks of norovirus disease, e.g. among children younger than 5 years (Lopman *et al*., 2003; Boga *et al*., 2004). The disease is characterized by a sudden and severe onset of nausea, vomiting, diarrhoea and possibly abdominal cramps and lasts for 2-3 days in general (Patel *et al*., 2009). Fever has also been reported in 37-45% of the cases (Wyatt *et al*., 1974; Kaplan *et al*., 1982). The disease normally resolves in 24 hours. Noroviruses can also result in deaths in immuno-compromised and weak persons (Dedman *et al*., 1998; Chadwick *et al*., 2000). Noroviruses are considered to be at least a contributing cause of death in about 200,000 cases annually.

2.1.2 Structural overview of the Norovirus capsid protein

Noroviruses were characterized as Calicivirus by successful cloning of the NoroVirus (NoV) genome from stool samples (Green *et al*., 2001) which marked the foundation of the molecular era of NoV. NoVs are positive-sense single-stranded RNA viruses. The positive-stranded poly-adenylated RNA genome has a protected cover of protein capsid, lacking lipid envelope. The capsid is mainly composed of Viral Protein 1 (VP1) and a few copies of Viral Protein 2 (VP2) which is a small structural protein. The recombinant expression of VP1 and VP2 assembles into a T=3 icosahedral Virus Like Particle (VLP) structure with 180 molecules arranged in 90 dimers of capsid protein. Each monomer capsid protein is divided into an N-Terminus arm that is in the interior of the VLP, a shell (S) domain which forms the icosahedral shell of the VLP and a P domain or protruding domain that emanates from the S-domain (Cao *et al*., 2007). The outer region of the VLP is formed by the P domain that is subdivided into P1 and P2 domains, where the P2 domain is the one that forms the outer most surface of the VLP (Figure 1). The P domain forms the dimer and the so called P particle (12 P dimers) on recombinant expression. The P domain has been shown to be involved in the binding to HBGAs with similar binding profiles in both the P particles and their parental VLPs (Tan *et al*., 2004; Tan and Jiang, 2005; Tan *et al*., 2006; Tan *et al*., 2008a). The S domain, that forms the S-particle, on the other hand is devoid of receptor binding functionality (Tan *et al*., 2004).

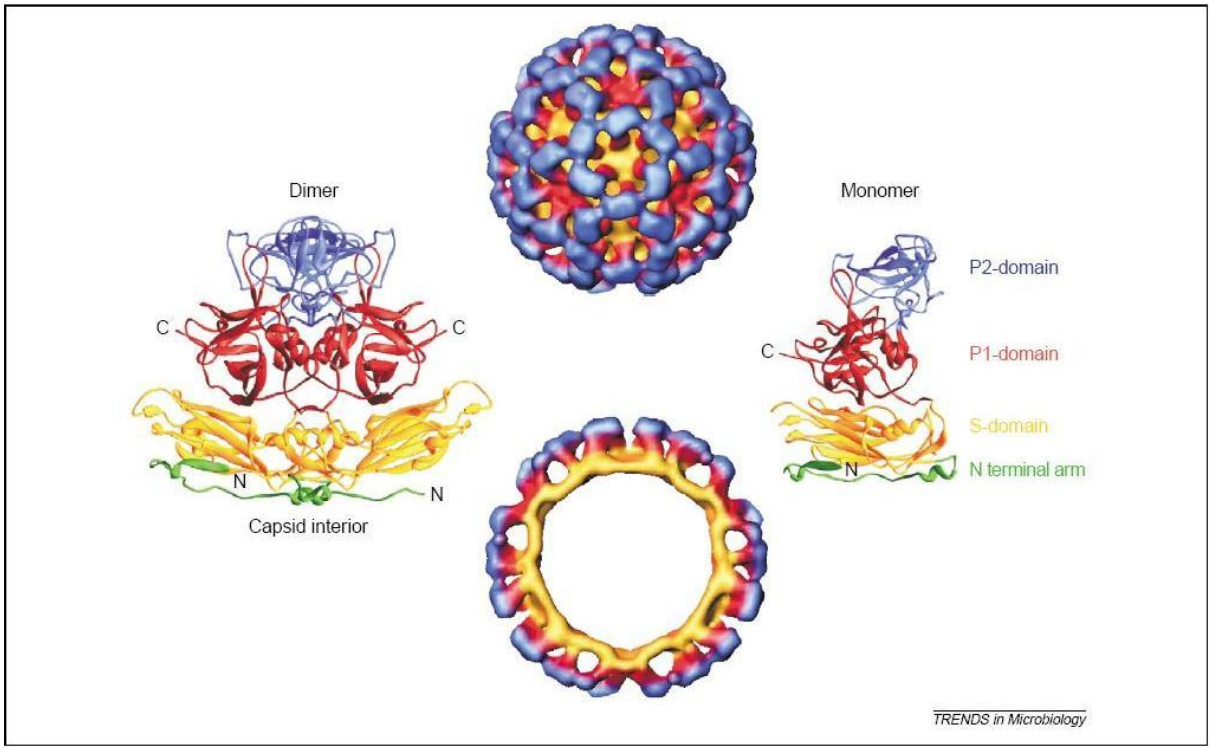


Figure 1 *The Norwalk VLP structure of the capsid viral protein. The surface representation (top) and the cross-section (bottom) of the VLP shows 90 dimers (left) arranged in T=3 icosahederal symmetry. The monomer (right) and the dimer representations display the arrangement of S and P domains along with the N-terminal arm. (From Hutson et al., 2004)*

2.1.3 Binding specificity of Norovirus strains

Over the years RT-PCR technology has made it possible to clone and sequence many strains of NoV from stool samples. NoVs are genetically highly diverse pathogens consisting of at least 5 genogroups and approximately 30 genotypes (Zheng *et al*., 2006). Genogroups I, II and IV, denoted by GI, GII and GIV, infect humans by causing acute gastroenteritis (Green *et al*., 2001; Estes *et al*., 2006; Tan and Jiang, 2008), whereas GIII infects cattle (Scipioni *et al*., 2008) and GV infects only immune-compromised mice (Wobus *et al*., 2006). The classification of the NoV strains, found to date, is shown in Figure 2.

Different NoVs are known to recognize human histo-blood group antigens (HBGAs) in a strain specific manner (Tan and Jiang 2005a, Tan and Jiang 2007). Eight distinct binding patterns have been described (Huang *et al*., 2003, Huang *et al*., 2005) based on 14 strains representing 13 genetic clusters of NoVs (Table 1). The binding patterns are further classified in two major groups, namely, the ABO- and the non-secretor or Lewis binding group. Huang *et al*., 2005 has also proposed a model based on this classification for the binding of NoVs (Figure 3).

The strains involved in the same binding pattern have been shown to be genetically related (Huang *et al*., 2005). For example, the VA387 and GrV strains of NoVs which recognize both A/B and O blood types belong to the same cluster GII4 and share 98% sequence identity in the capsid genes. Similarly, Norwalk and C59 are also at a nearest phylogenetic distance to each other and share the same binding pattern. However, there are cases where viruses belonging to the same genogroup have been shown to have different binding patterns and similarly viruses showing same binding patterns belonging to different genogroups. For

example, Norwalk and VA387 both bind to A and H epitopes but are in different gene clusters with distinct genogroup (Table 1).

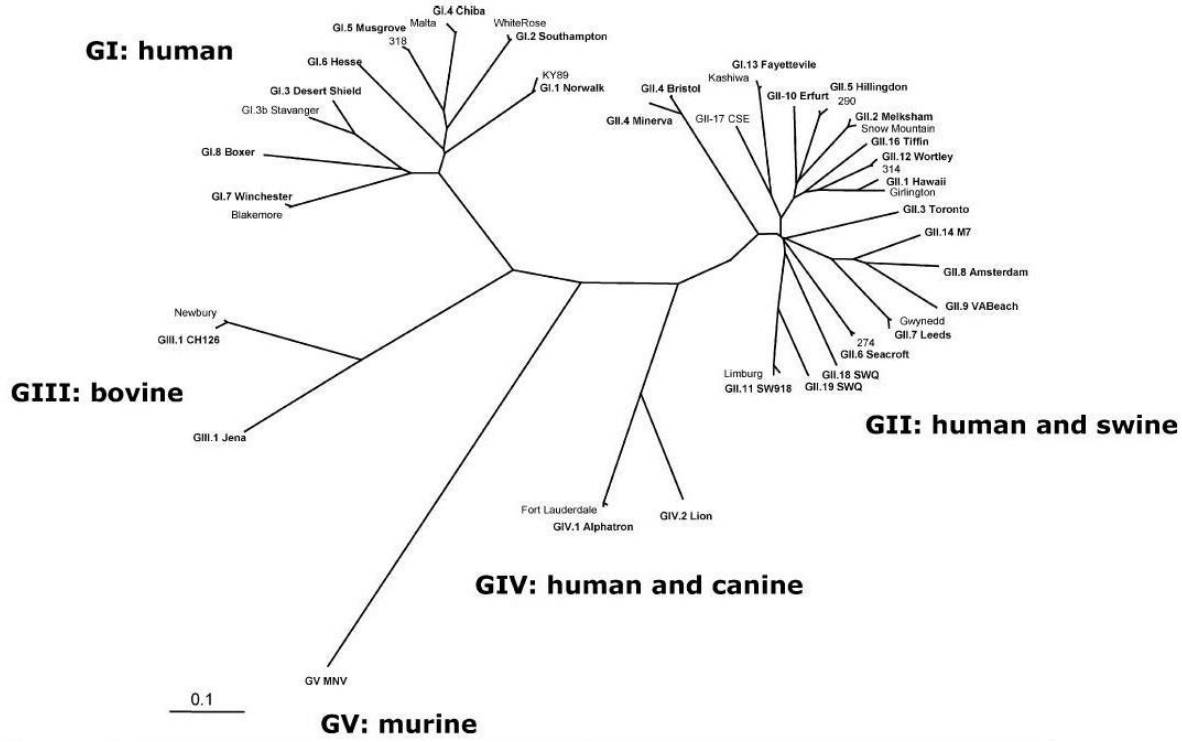


Figure 2 *Classification of NoV strains in 5 genogroups and ca. 32 genotypes. The GI, GII and GIV strains infect humans (from Patel et al., 2009).*

An interesting finding in the binding studies is the mutual exclusion of the secretor ABO- and non-secretor binding patterns, i.e. the strains involved in the A/B binding groups have shown binding to A and/or B and O saliva of secretors but not to the saliva of non-secretors (Table 1). Similarly, the strains in the Lewis group (non-secretor) have shown no binding to secretor A and/or B HBGA and weak binding to O type secretors. Therefore, human HBGAs are thought to be an important factor in the evolution process of noroviruses (Huang *et al*., 2005) and a recent mutagenesis study has claimed that the HBGA-binding interfaces are under a selection pressure by the human HBGAs (Tan *et al*., 2009) since human HBGAs are thought to be necessary for NoVs infection.

The NoV-HBGA interaction is a typical protein-carbohydrate interaction. NoVs are genetically highly diverse and the binding region in the capsid protein of NoVs is extremely sensitive even to point mutations. Studies have shown that a single amino acid change in the P domain of viral capsid protein can result in a different HBGA binding pattern (Tan *et al*., 2003).

The molecular basis of this binding specificity has been discussed elsewhere (Tan *et al*., 2008, Tan *et al*., 2009) based on *in vitro* studies. An attempt is made in the present study to further explain the binding data and extend the previous findings on the same molecular grounds with the focus on NoV GII.4 strain VA387.

Binding Patterns	Representative strains	Genogroup	Genetic Cluster	Targeted HBGAs			
				Secretors			Non-secretors
				A	B	O	Lewis
1	Norwalk	I	1	+++	-	++	-
	C59	I	2	+++	-	++	-
2	VA387	II	4	+++	+++	++	-
	GrV	II	4	+++	+++	++	-
3	HV	II	1	++	++	+	-
	MxV	II	3	+++	+++	+	-
	PiV	II	3	+++	+++	+	-
4	MOH	II	5	+++	+++	-	-
	BUDS	II	2	+++	+++	-	-
5	SMV	II	2	-	+++	-	-
6	VA207	II	9	-	-	+	++
7	Boxer	II	8	-	-	++	++
8	OIF	II	13	-	-	-	++

Table 1 *Different strains of Norovirus that have been described to have the stated binding patterns in terms of binding with ABO secretors and non-secretors. '+' indicates positive binding observed in assays with saliva and/or oligosaccharide conjugates, '++' and '+++' represents relatively higher binding. '-' is used to represent no binding. The data is from the studies conducted by Huang et al., 2005, except SMV that comes from Harrington et al., 2002.*

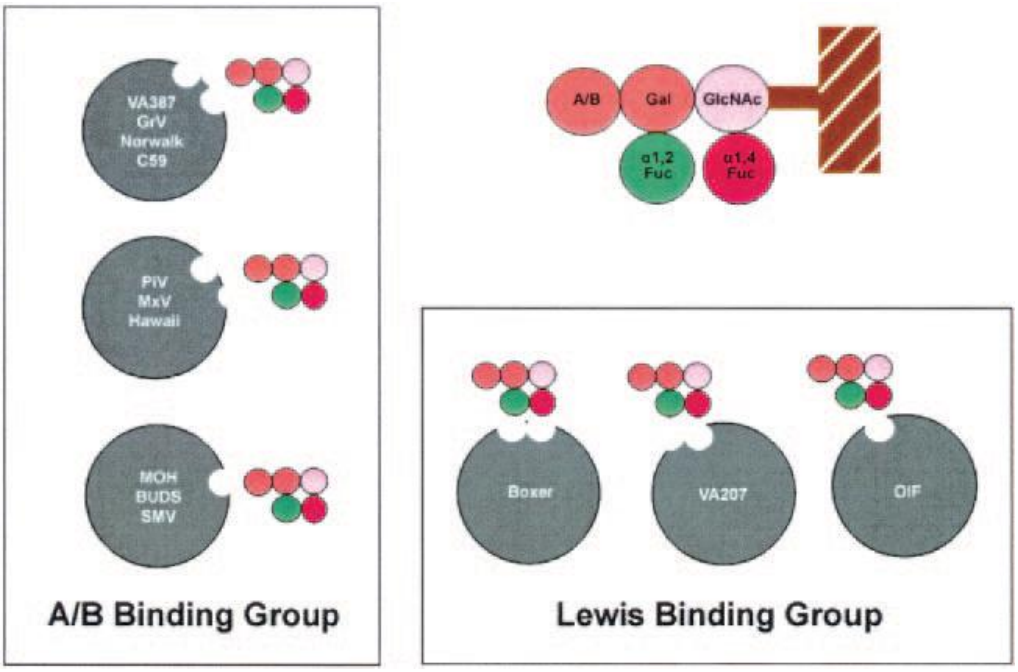


Figure 3 *Classification of binding patterns of NoV strains in two major groups, the A/B binding group and the Lewis binding group. The final structure in the form of a pentasaccharide (ALe^b or BLe^b) is shown on the top right of the picture. Thirteen NoV strains are indicated with the potential binding sites shown for each of the three HBGA epitopes on the capsid. (SMV is reported to recognize type B antigen only (Harrington et al., 2002)) (from Huang et al., 2005).*

2.1.4 The Lordsdale-like VA387 Norovirus strain

Among the strains of NoVs the VA387, a Lordsdale-like GII.4 strain of NoV, is of dominating clinical importance and is predominant in many countries (Ramirez *et al*., 2008; Siebenga *et al*., 2008; Tu *et al*., 2008; Verhoef *et al*., 2008). In the United States and several other countries, VA387 has been reported to cause 50 - 90% of all NoV associated outbreaks of acute gastroenteritis (Fankhauser *et al*., 1998; Vinje *et al*., 1996). The binding studies have suggested that VA387 has the ability to infect any secretor, which includes almost 80% of the European-derived population (Huang *et al*., 2003).

Structural overview of the P protein dimer

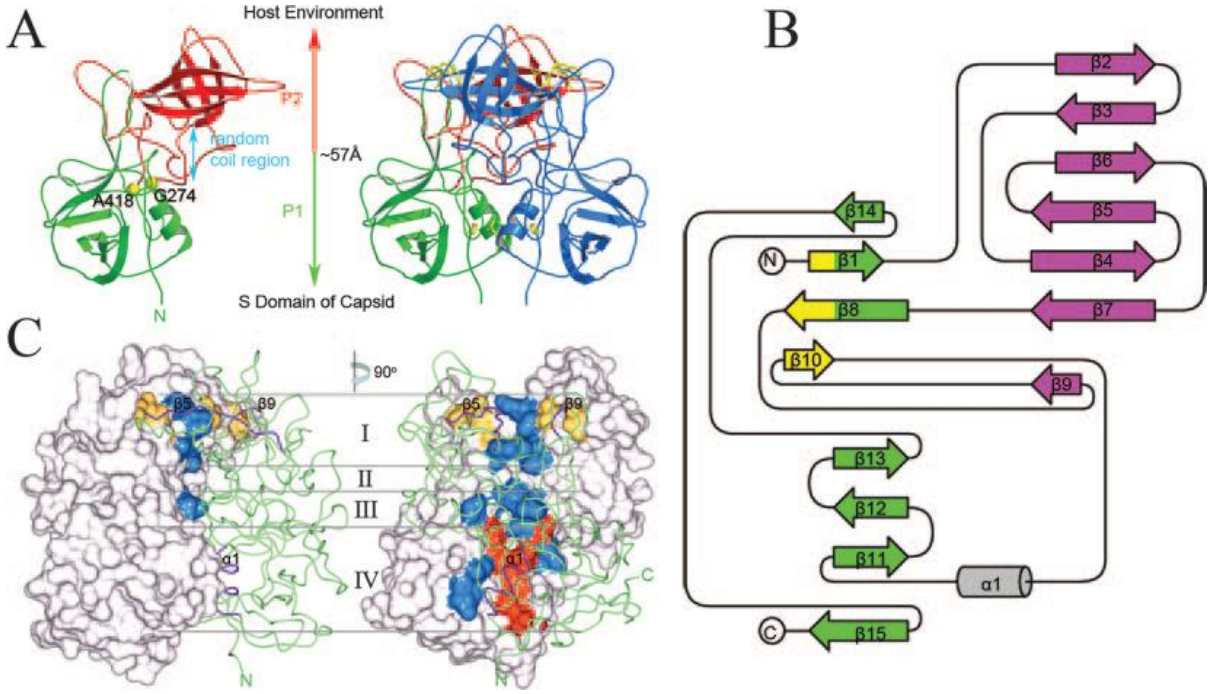


Figure 4 (*A) The monomer and the dimer of the VA387 P protein are shown in ribbon representation with the P1 domain in green and the P2 domain in red with yellow spheres showing ALA418 and GLY274. (B) The topology chart of the VA387 P protein with the arrows showing the direction of the β-strands and the cylinder showing the α-helix. β-strands in the same sheets have identical colouring. Four layers of dimeric interface along the diad axis are shown in C. The top layer has a hydrophobic centre and 2 hydrogen bond rich regions. Area around ARG345 in this layer forms an open cavity for receptor binding. The second and third layers contain a water tunnel and hydrophobic layers, respectively. The last layer is composed of a hydrophilic centre with 2 hydrophobic edges. Residues involved in hydrophobic, hydrophilic and hydrogen-bonding interactions are coloured blue, red and yellow, respectively (from Cao et al., 2007).*

The crystal structure of the VA387 P domain forms a homo-dimer with residues THR224 to GLY530 in each monomer. The P2 sub-domain (residues 275 to 417) is an insertion in the P1 sub-domain (residues 222 to 274 and 418 to 539[1]) between the residues GLY274 and

[1] Residues 214 to 223 and 531 to 539 could not be interpreted in the electron density map of the crystal structure (Cao *et al.*, 2007).

ALA418 (see figure 4). The P1 sub-domain has two β-sheets consisting of purely anti-parallel β-strands. The smaller β-sheet has the strand order of β1-β8-β10 whereas the strand order for the larger β-sheet is β14-β1-β8-β13-β12-β11-β15. The only well defined α-helix (residues 454 to 463) in the entire P domain is in the P1 sub-domain. The hydrophobic core of P1 consists of the two β-sheets and the α-helix along with the amino-terminal region. The P2 sub-domain contains an anti-parallel β-barrel of Greek-key topology with the strand order β2-β3-β6-β5-β4-β7 and a hydrophobic core. The P1 and P2 subdomains are supported with a random coil region. The P domain consists only 3% of α-helices and 29% of β-strands. Despite this low secondary structure content, the P domain is stable probably due to the high abundance of proline residues (26 in each monomer) along with the considerable interactions of localized crystal water molecules with this domain (Cao *et al*., 2007).

The carbohydrate binding site

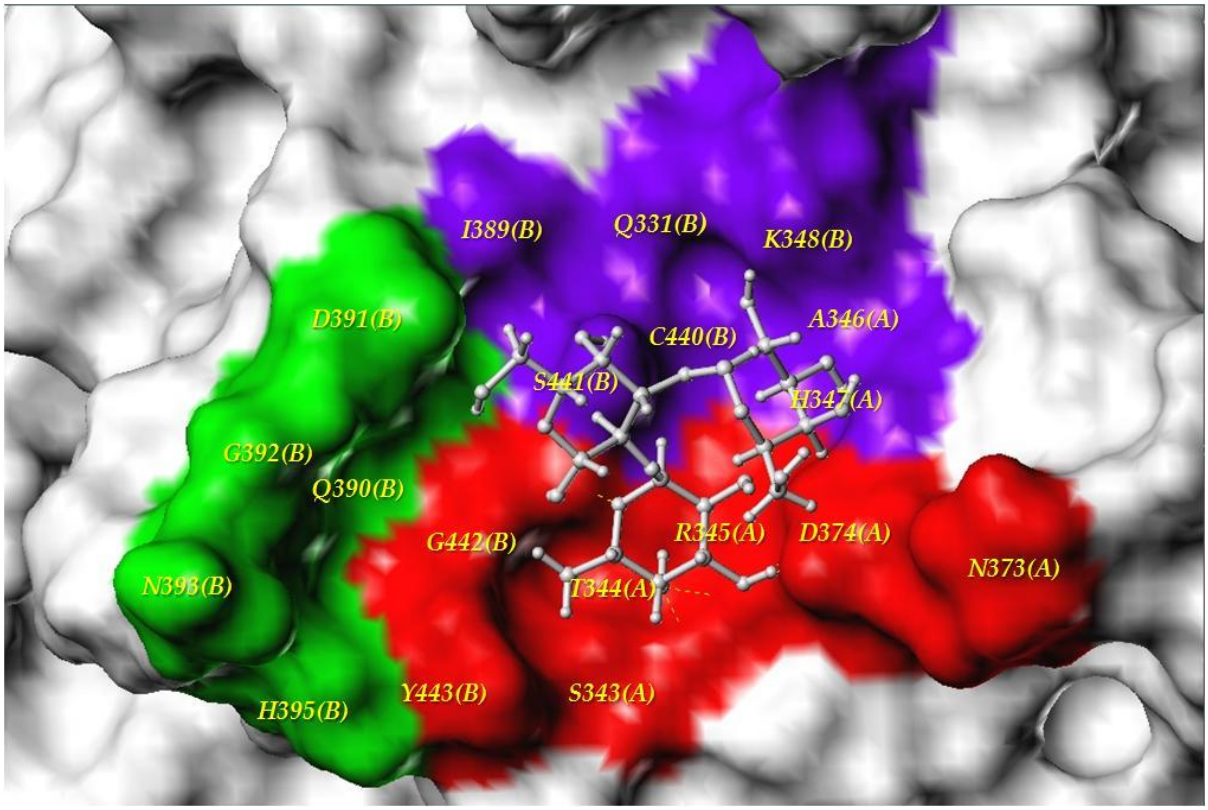


Figure 5 *The carbohydrate binding site of the VA387 norovirus surface protein according to the crystal structure in complex with the B-trisaccharide (Cao et al., 2007). Region coloured in red represents the residues involved in fucose binding. The purple and green regions represent residues which are potentially involved in binding with terminal Gal/GalNAc and the downstream extensions of HBGAs, respectively, as suggested by the mutagenesis studies (Tan et al., 2008).*

The crystal structures of recombinant P protein (2OBR) of VA387 in complex with A-trisaccharide[2] (PDB ID 2OBS) and in complex with B-trisaccharide (2OBT) have been published (Cao *et al*., 2007). These studies show the residues contributing to the binding site.

[2] It has been noted that the A-trisaccharide in the complex with norovirus VA387 surface protein has incorrect geometry (Koppisetty, C.A.K, Nasir, W., Rydell, G.E., Strino, F., Larson G. and Nyholm P.G. *Docking of histo-blood group ABO-active saccharides with the norovirus VA387 capsid protein can explain experimental binding data*. Manuscript in preparation).

The binding specificity of the A and B human HBGAs has been further investigated in a mutagenesis study (Tan *et al*., 2008).

The α-fucose saccharide ring is shared by secretor A-, B- and H-antigens, which have been shown to bind with VA387 *in vitro* (Huang *et al*., 2005). The crystal structure revealed that α-fucose of the A- and B-antigens has the most extensive interaction with the P protein of VA387 (Cao *et al*., 2007). The α-fucose binding site in VA387 is located at the dimer interface of the P protein suggesting that the dimer is important for the carbohydrate binding function (Cao *et al*., 2007). The cavity that interacts strongly with α-fucose is formed by the protruding β5 strand and the residues SER343, THR344, ARG345 and ASP374 in one monomer and SER441, GLY442 and TYR443 in the other according to the crystal structure (Cao *et al*., 2007), where THR344, ARG345 and TYR443 form the bottom of the binding pocket (Figure 5).

The residues THR344, ARG345, GLY442 and ASP374 are involved in strong hydrogen bonding interactions directly with the ligand. The phenol group of residue TYR443 exhibits van der Waals interactions with the methyl group of α-fucose monosaccharide. The residues CYS440, LYS348, SER441 and ASP391 in one monomer and ALA346 in the other monomer, not very close to the α-fucose binding site, are also identified by the crystal structure to be involved in water mediated hydrogen bonds between the B-trisaccharide and the P protein. Therefore, in total 10 residues were identified by the crystal structure to be involved in interactions between the P-dimer and the B-trisaccharide (Cao *et al*., 2007).

The mutagenesis study by Tan *et al*., published in the following year revealed some facts about the binding specificity of A- and B-antigens with the P-dimer of VA387. The only structural difference between the two antigens is the acetamido group in the A-antigen that is replaced in the B-antigen by a hydroxyl group. The mutations to ALA of the residues ILE389, GLN331 and LYS348, next to the α-fucose binding site close to acetamido group of the A-antigen, resulted in significantly reduced binding to A- but not to B-antigen. The pocket formed by these three residues is therefore thought to be interacting solely with the acetamido group of the A-antigen (Tan *et al*., 2008). The above three residues along with ALA346 and SER441 form the subsite binding to the α-galactose/α-N-acetylgalactosamine residue (Tan *et al*., 2008).

A third potential binding subsite has also been proposed (Cao *et al*., 2007; Tan *et al*., 2008) in the receptor binding interface of VA387 close to the α-fucose binding site. This pocket is formed by the residues GLN390, ASP391, GLY392, ASN393 and HIS395. This subsite is adjacent to the O1 atom of β-galactose ring in the crystal structure of the complex of VA387 P protein with B-trisaccharide and is thought to be involved in binding of extended ABH-antigens (Cao *et al*., 2007). However, the mutation of these residues with ALA have not resulted in any difference of binding affinities except for GLY392 and HIS395 in which case the P particles bound slightly weaker to the A-antigens but did not make any difference on binding to the B-antigens (Tan *et al*., 2008).

2.1.5 Human histo-blood group antigens (HBGAs)

ABO blood-group antigens were first discovered by Karl Landsteiner on human red blood cells about a century ago along with their respective antibodies. The antigens were more recently renamed to "Histo-blood group antigens" because of their presence in saliva and their expression in several tissues; especially the gut (Watkins *et al*., 1999; Ravn and Dabelsteen, 2000; Marionneau *et al*., 2001).

Histo-blood group antigens (HBGAs) are complex carbohydrate structures that are often found at the termini of glycan chains. These structures are determined by the inheritance of genes that encode glycosyltransferases with different functions. ABH(O), secretor and Lewis (genotype) gene families are of interest in the present study.

Biosynthesis of HBGAs

HBGAs are synthesized in a step-wise fashion by the addition of a monosaccharide ring to the precursor by the action of a set of glycosyltransferases. The type-1 (Galβ1-3GlcNAcβ1-R), type-2 (Galβ1-4GlcNAcβ1-R) and type-3 (Galβ1-3GalNAcα1-R) precursors are considered here. The glycosyltransferases are encoded by three genetic loci (the ABO, H and secretor [Se] loci).

The biosynthesis process starts with the addition of a fucose residue to the precursor with an α1,2 linkage by α1,2-fucosyltranserases. The resulting carbohydrate moiety is a H-antigen. In humans two fucosyltransferases are known to catalyse the synthesis of H-antigen. They are encoded by the H and Se loci. The H locus encodes α1,2-fucosyltransferase that is expressed in red cell precursors and it transfers the fucose to type-2 and type-4 precursors, whereas the Se locus encodes α1,2-fucosyltransferase that is expressed in epithelial cells and adds fucose to type-1 and type-3 precursors to produce H-antigen in the epithelial lining of lumen of gastrointestinal, respiratory and reproductive tracts and in salivary glands (Varki *et al*., 2008).

The ABO locus encodes glycosyltransferases which subsequently add an additional residue to the type-1, -2, -3 or -4 H-antigens to produce A or B blood group antigens. The A allele at the ABO locus is responsible for the synthesis of A-antigen and encodes α1,3-GalNAc-transferase that adds GalNAc monosaccharide to the H determinant with an α1,3-linkage. Alpha-1,3Gal-transferase, encoded by the B allele of the ABO locus, adds a galactose monosaccharide to the H determinant with an α1,3-linkage to produce the blood group B-antigen. The O allele at ABO locus encodes a functionally inactive glycosyltransferase and the H-antigen is therefore unmodified in this case. The individuals who do not produce A- or B-antigens have blood group type O with the genotype OO, whereas those who synthesize A- or B-antigens exclusively have the genotype AA (or AO) or BB (or BO). These individuals belong to blood groups A or B, respectively. Finally the individuals with both A and B antigens have the blood group type AB and genotype AB (Varki *et al*., 2008).

Along with the expression of ABO antigens on membrane proteins and glycolipids on the surface of red cells and in several tissues, some tissues also synthesize the soluble, secreted form of these molecules as glycans on secreted glycoproteins, glycolipids and free glycans. The expression of A-, B- or H-antigens in these tissues is the result of the action of α1,2-fucosyltransferase enzyme called FUT2 that is encoded by the Se locus. The individuals who are devoid of this enzyme are termed as non-secretors and are characterized by the fact that they cannot produce any A-, B- or H-antigens in soluble forms, since H-antigen is not expressed in secretory tissues. Therefore these antigens cannot be detected in their saliva.

The Lewis antigens are characterized by the glycans that carry α1,3-/α1,4-linked fucose residue (Figure 6). $Le^{a/x}$ antigens are produced by the action of α1,4/α1,3-fucosyltransferases on the type-1 or type-2 precursors respectively. The fucosyltransferases are encoded by the Lewis (Le) blood group locus. Similarly, the $Le^{b/y}$ antigens are synthesized by the combined action of α1,2 and α1,4/α1,3 fucosyltransferases encoded by Se and Le loci respectively.

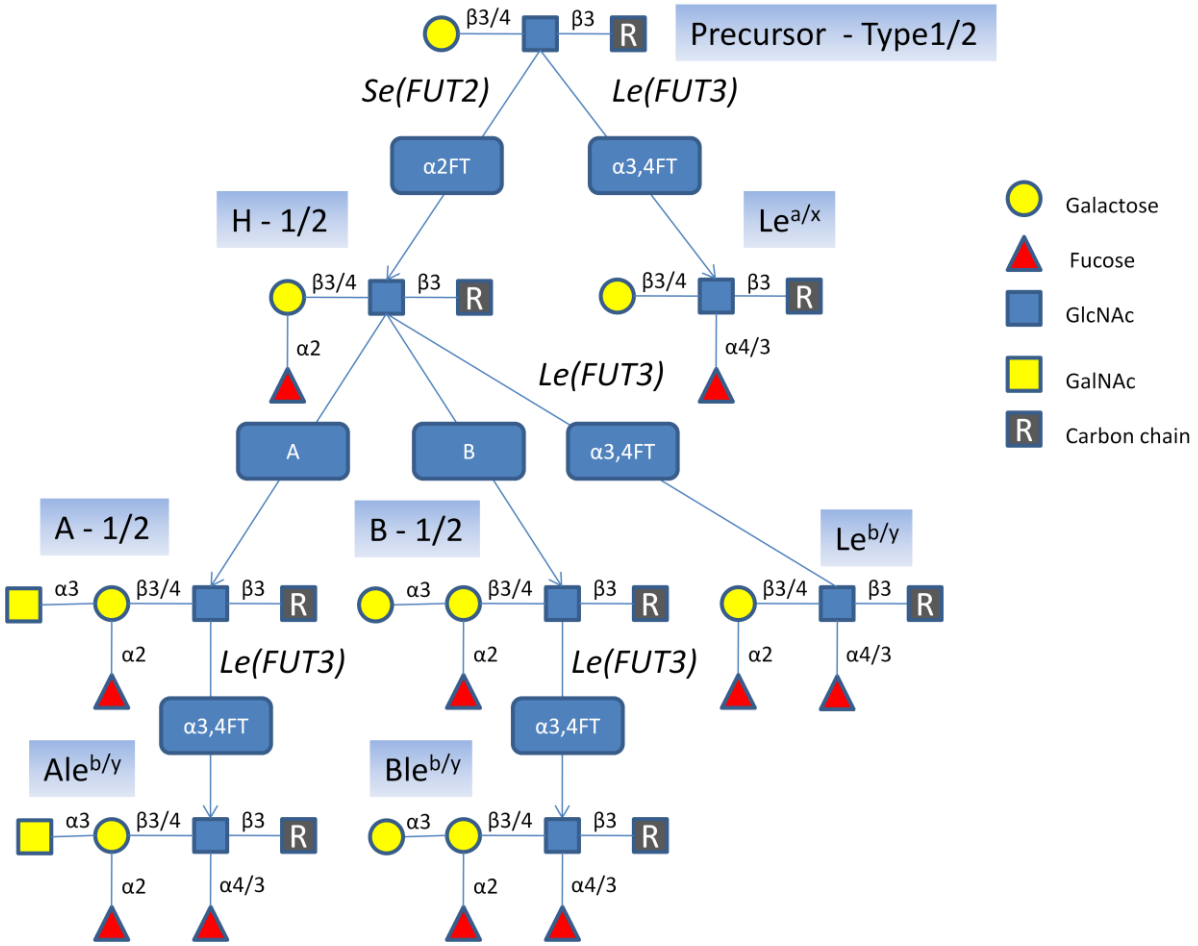


Figure 6 *Biosynthesis of HBGAs with type-1 or type-2 precursors.*

The final products of the biosynthetic pathway for HBGAs in case of type-1/2 precursors are ALe$^{b/y}$ and BLe$^{b/y}$ which are produced as a result of the addition of a α1,4-/α1,3-linked fucose residue to the terminal GlcNAc in A- or B-antigens, respectively. The antigens Le$^{b/y}$, ALe$^{b/y}$ and Ble$^{b/y}$ are termed as secretor positive Lewis structures if they are derived by the action of α1,2-fucosyltransferase encoded by the Se locus. Le$^{a/x}$, however, are termed as secretor negative Lewis structures because of the lack of Se encoded fucosyltransferase in their biosynthesis.

2.2 Molecular Dynamics

Over the last few decades, computer simulation methods have gained considerable popularity in the analysis of bio-molecular systems. Molecular dynamics (MD) is a form of computer simulation which is used to study the motion of atoms through classical physics (Newtonian Laws in particular), when the atoms or molecules are allowed to interact for a specified period of time. These simulations are then used to investigate structural, dynamic and thermodynamic properties of the biological molecule or protein-ligand complex in question. Moreover, they are also used in the determination of structures from x-ray crystallography and NMR (Stote *et al*., 1999).

MD simulation method is deterministic i.e. to know the velocity, position and acceleration of particles in the system at any given time $t \pm \Delta t$, the knowledge of the present state at time t is sufficient. The stochastic variant of these simulations for *many-body* or *N-body* systems is the Monte Carlo (MC) method, in which the configuration space is probed by trial moves of particles in the system and the so called *Metropolis Algorithm* is then used to determine the next state of the system depending upon the energy change between the two states (Sutmann, 1999).

2.2.1 Statistical Mechanics

The MD simulations produce data that are microscopic in nature i.e. the properties of motion like positions, velocities or acceleration of the atoms or particles. These microscopic data are then converted to macroscopic properties of the system like heat, energy etc. by the means of *statistical mechanics* which provides the rigorous mathematical expressions that are used to relate the system properties to the motion of particles in N-body system (Stote *et al*., 1999).

In order to connect or to relate the microscopic behaviour of the individual particles to the macroscopic properties of the system, *time independent* statistical averages are usually introduced that represent the observables from an experiment. These averages are termed as *ensemble* averages and correspond to the averages taken on an extremely large number of replicas of the system. These averages have shown considerable agreement with experiments.

To compute these averages theoretically one has to calculate all the states of the system through simulation followed by the integration to get the ensemble average, which is an extremely difficult process. In molecular dynamics the system states change sequentially in time. Therefore time averages are often calculated and are thought to be the substitute of these ensemble averages, according to the "ergodic hypothesis" which states;

$$Ensemble\ Average = Time\ Average$$

i.e. given sufficient period of time, the system will eventually pass through all possible states (Stote *et al*., 1999). This is the basic idea behind all molecular dynamics simulations.

2.2.2 Basic Principle – Classical Mechanics

The molecular dynamics simulation method is based on Newton's second law of motion,

$$F = ma$$

where F is the force exerted, m is the mass and a is the acceleration on the particle. The knowledge of the force on the particle is enough to determine the acceleration of each particle in the system. The consequent integration of equation of motion can then provide the information for the *trajectory* that describes the acceleration, velocity and position of the

particles as they vary with time. The average values of the properties of the system can then be calculated from this trajectory (Stote *et al*., 1999).
Manipulating the equation of motion, one can easily find that for the initial position x_0, initial velocity v_0 and acceleration a, the value of position x at time t, is given by the following equation;

$$x = at^2 + v_0 t + x_0$$

Similarly the equation of motion incorporates the potential energy of the system in the following way;

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2}$$

where V, m_i and r_i are the potential energy of the system, the mass and the position of the particle, respectively. Therefore the derivative of the potential can be related with the changes in position of particles as a function of time.

The acceleration is given as the gradient of potential with respect to the position of the particle.

$$a = -\frac{1}{m}\frac{dE}{dr}$$

To calculate the trajectory one only needs initial positions of atoms, initial distribution of velocities and the acceleration that is calculated through the gradient of potential as shown above. The initial positions can be determined from experimental structures, usually x-ray or NMR studies (Stote *et al*., 1999). The velocities are determined as the random distribution with the magnitude conforming to the required temperature and satisfying the condition that the overall momentum of the system is zero (Stote *et al*., 1999).

i.e.

$$P = \sum_{i=1}^{N} m_i v_i = 0$$

The velocities v_i are chosen randomly from Maxwell-Boltzmann or Gaussian distributions. The temperature is calculated from the velocities through the following relation (Stote *et al*., 1999);

$$T = \frac{1}{3N} \sum_{i=1}^{N} \frac{P_i}{2m_i}$$

2.2.3 The Force Field

The combination of mathematical *expressions* and *parameters* which give the complete picture of the system in the form of energy and force on each particle is often termed as a *force field* (Sutmann 2002; Ponder and Case, 2003). Over the last couple of decades different force fields have been developed which include AMBER, CHARMM, OPLS, MM4, etc. The most common functional form of the force field, that is used as a basis for the potential almost in every case of MD simulations, contains some of the terms listed in figure 1 and is of the following shape;

$$E_{total} = E_{bonded} + E_{non-bonded}$$

The term E_{bonded} contains the following sum;

$$E_{bonded} = E_{bond-stretch} + E_{angle-bend} + E_{rotate-along-bond}$$

Whereas the term $E_{non-bonded}$ can be expressed as;

$$E_{non-bonded} = E_{van-der-waals} + E_{electrostatic}$$

The force fields used in *classical* MD simulations are based more or less on the above structure and the extension of these basic terms of the force fields is made according to the resources at hand. There is always a trade-off between the computational cost (CPU time for example) and the accuracy of the force field. The more *detailed* and less approximated the terms of a force field are, the greater time it will take for the simulations. Typically an MD *production run* on a protein in water for one nano-second may take about a quarter of a year on single CPU to complete (Sutmann 2002).

The general terms in the force field as described above, are bond stretching, angle bending, torsional and possibly improper dihederal terms along with the expressions for non-bonded interactions (van der Waals, Coulomb potential etc.). Central to the modelling of the force field is the *interaction model* for the potential. The interactions can be modelled through pair-wise potentials where only a pair of atoms is considered making the calculations very simple and reducing the complexity of the force field dramatically. On the other hand, the multi-body interactions are often very hard to implement due to the complexity involved in calculating the interaction partners. Another important distinction is made based on short-range or long-range interactions. In cases where the long-range interactions are either weak or counter balancing; their effect can be neglected. However, in cases where there is no counter-effect, like gravitational potential, their effect might not be neglected. To model the long-range interactions accurately every particle in the system has to be considered which makes the problem $O(N^2)$ complex. Therefore approximations of long-range interactions are required in most cases.

A classification of MD simulations is made on the approximation methods for modelling the inter-atomic interactions. Classical Molecular Dynamics, i.e. the one used in this work, uses the approximation method discussed above, where the interactions are modelled as pair-wise or multi-body, long-range or short-range etc depending upon the computational resources at hand. These potential terms are often "fixed", and therefore are usually calculated in advance. *Ab initio* molecular dynamics simulations are characterized by the fact that the force on the nuclei of each atom are computed through electronic structure calculations and therefore the corresponding inter-atomic interactions are no longer approximated through fixed potentials (Marx and Hutter, 2000).

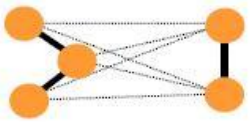
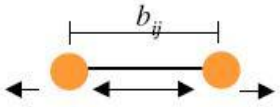
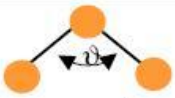
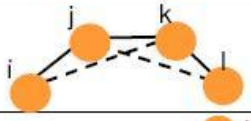
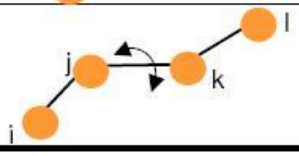
6-9 van der Waals 6-12 van der Waals Electrostatic		$u_{ij}(r_{ij}) = 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{9} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right)$ $u_{ij}(r_{ij}) = 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right)$ $u_{ij}(r_{ij}) = \frac{q_i q_j}{r_{ij}}$
Quadratic bond-stretching Morse bond-stretching		$u_{ij}^{s}(r_{ij}) = \frac{1}{2}k_{ij}(r_{ij} - b_{ij})^2$ $u_{ij}(r_{ij}) = k\left(1 - e^{-a(r_{ij} - r_0)}\right)^2$
Bond-bending		$u_{ij}^{b}(\vartheta_{ijk}) = \frac{1}{2}k_{ijk}(\vartheta_{ijk} - \vartheta_{ijk}^{0})^2$
Improper dihedrals		$u_{ij}^{id}(\xi_{ijkl}) = \frac{1}{2}k_{ijkl}(\xi_{ijkl} - \xi_0)^2$
Proper dihedrals		$u_{ij}^{pd}(\varphi_{ijkl}) = k_{\varphi}(1 + \cos(n\varphi_{ijkl} - \varphi_0))$

Table 2 *General terms in a force field with their graphical illustration and expressions. (From Sutmann 2002).*

Force field modifications – The implicit solvent

In bio-molecular systems, the whole system usually comprises of a solute, typically a protein or a protein-ligand complex, and water molecules as solvent. Solvent interactions are thought to be very important in determining the structure and function of proteins, e.g. water mediated interactions sometimes play a key role in protein-ligand interactions. The inclusion of explicit water molecules to provide a detailed description of the system is usually preferable in terms of quality of the simulation. However, describing the system with explicit solvent molecules drastically increases the time for simulation (AMBER 10 user's manual).

In cases where the interactions involving solvent molecules are not critically important and the focus is purely on solute molecules, the solvent interactions are usually approximated by the incorporation of a mean force potential for solvation interactions to accelerate the computation. The mean force potential is based on the average of degrees of freedom for solvent molecules and is often called continuum or implicit solvent (AMBER 10 user's manual).

In implicit solvent simulations the inter- and intra-molecular solute interactions are calculated by the molecular dynamics force field as discussed earlier. The solute-solvent and solvent-solvent interactions are approximated by a mean-field approximation by the use of a model for implicit solvent. Common examples of such models include Generalized Born/Surface Area (GB/SA) (Tsui and Case, 2001; Sosa *et al*., 2001) and Analytical Linearized Poisson Boltzmann (ALPB) (Sigalov *et al*., 2005; Sigalov *et al*., 2006) models. The solvation free energy of the molecule can be decomposed into the electrostatic and non-electrostatic terms.

In the GB/SA model, the electrostatic term is approximated by an analytical formula that assumes every atom in the molecule to be represented by a sphere whose interior is uniformly filled with the material of dielectric constant 1 whereas; the surrounding solvent has a higher dielectric (e.g. 80 for water at 300K) (Still *et al.*, 1990; Srinivasan *et al.*, 1999). In the AMBER implementation of GB/SA, the non-electrostatic part of the solvation energy is taken to be proportional to the total solvent accessible surface area (SA) of the molecule. The proportionality constant is derived from experimental values for small non-polar molecules (Amber 10.0 user manual). The GB/SA approach, used in the present study, is the most widely used implicit solvent technique.

2.2.4 The Integration Algorithms

Numerical methods are often used in MD simulations to calculate the potential energy function, as the number of particles in a general simulation is extremely large. It is impossible to calculate the properties of these particles analytically and the equations of motion therefore can only be solved numerically.

There are several integration algorithms used to integrate the equations of motion (Stote *et al*., 1999) for example;

- Verlet algorithm.
- Leap-Frog Algorithm.
- Velocity Verlet.
- Beeman's Algorithm.

In general, all the integration algorithms use Taylor's series expansion to approximate positions, velocities and accelerations of the particles/atoms in the system.

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \cdots$$
$$v(t+\delta t) = v(t) + a(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \cdots$$
$$a(t+\delta t) = a(t) + b(t)\delta t + \frac{1}{2}c(t)\delta t^2 + \cdots$$

where r is the position, v is the velocity and a is the acceleration of the atoms or particles. Here, v is the first derivative with respect to time, a is the second and so on (Stote *et al*., 1999).

The details of each and every integration algorithm are left for the sake of brevity. The interested reader is encouraged to consult the referred literature.

Based on the above model for MD simulations, an MD simulation program has to incorporate the following;

1. An interaction model, i.e. if the particles (atoms, molecules etc) interact pair-wise only or there are other interactions possible. Usually, for simplicity the pair-wise model is used and a pair-wise potential is then calculated which greatly reduces the complexity of the program.
2. An integrator or an efficient integration algorithm is needed which propagates particle position and velocity from one state to the next and so on.

3. A statistical *ensemble* must be chosen where certain thermodynamic properties like pressure, temperature and the number of atoms are controlled. Common examples are the micro-canonical ensemble (NVE – number of atoms, volume and energy fixed), the canonical ensemble (NVT) etc.

These three selections define the molecular dynamics simulations in a nutshell (Sutmann 2002).

2.2.5 Application Area

Classical Molecular Dynamics address a very broad range of problems, which include properties of liquids, defects in solids, fracture, surface properties etc (Sutmann, 2002). The main application area with which we are concerned, however, is the study of biomolecular systems containing proteins, carbohydrates and nucleic acids. Within this area, massive research throughput has been observed concerning protein-ligand and protein-protein interaction studies. The present work can be considered as the application of molecular dynamics for protein-ligand interaction studies.

2.2.6 Molecular Dynamics – Amber

The term "Amber" is an acronym for *Assisted Model Building with Energy Refinement.* It refers to a set of molecular mechanics force fields for molecular simulations and a package of molecular simulation programs for molecular dynamics distributed by UCSF (University of California, San Francisco). The history of Amber dates back to the late 1970's when a single program for energetically refined model building was developed. Today, AMBER contains a group of programs for doing molecular mechanics and molecular dynamics calculations particularly on systems with proteins, nucleic acids and carbohydrates. AMBER version 8.0 is used in the present study for simulations (Case *et al*., 2004; Pearlman *et al*., 1995).

The AMBER Force Fields

The general form of AMBER force field that is used as a basis for all its variants (ff94, ff99, ff99SB, ff03, etc) can be written as (Ponder and Case, 2003);

$$V(x) = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta_i - \theta_0)^2 + \sum_{torsions} k_\phi[\cos(n\phi + \delta) + 1] + \sum_{\substack{nonbond \\ pairs}} \left[\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{{r_{ij}}^{12}} - \frac{C_{ij}}{{r_{ij}}^{6}}\right]$$

After the development and extensive usage of ff94 for several years some problems, like over-stabilization of α-helices, were reported in the parameter set of this force field. This led to the development of different force fields in AMBER (ff96, ff99 etc). These force fields, however, also suffered from the problem of inadequate balance for the secondary structure of proteins. An attempt was made to address this problem in particular and a new parameter set called ff99SB was introduced (Hornak *et al*., 2006). ff99SB achieved better balance of secondary structure elements and was shown to be in better agreement with the experimental

data than its predecessors. The parameter set ff99SB has been used in the force field for the present study.

The AMBER suite of Programs

The AMBER suite of programs can be categorized into three main classes, which give the overview of the whole system of calculations and analyses.

Preparation programs

The main preparation programs for AMBER are *Leap* and *Antechamber*. The former is used to create the files for the simulation when the components of the systems are in the standard library for the AMBER force field being used, whereas the later is used to create non-standard residues that are not included in the library. The main use of Leap is the construction of the residues, assignment of force field terms along with associated parameters and solvating the system.

The topology (prmtop) file created with leap contains all the data that is required by the subsequent simulation programs for calculating forces and energies. The data include masses, list of bonds, angles, dihedrals, force field parameters etc. The coordinate file generated by leap contains only the coordinates of the atoms in the system.

Nucleic Acid Builder (NAB) and the AMBER/GLYCAM configuration tool (http://glycam.ccrc.uga.edu/ccrc/biombuilder/biomb_index.jsp) can be used with AMBER for initial model building of nucleic acid and carbohydrates, respectively.

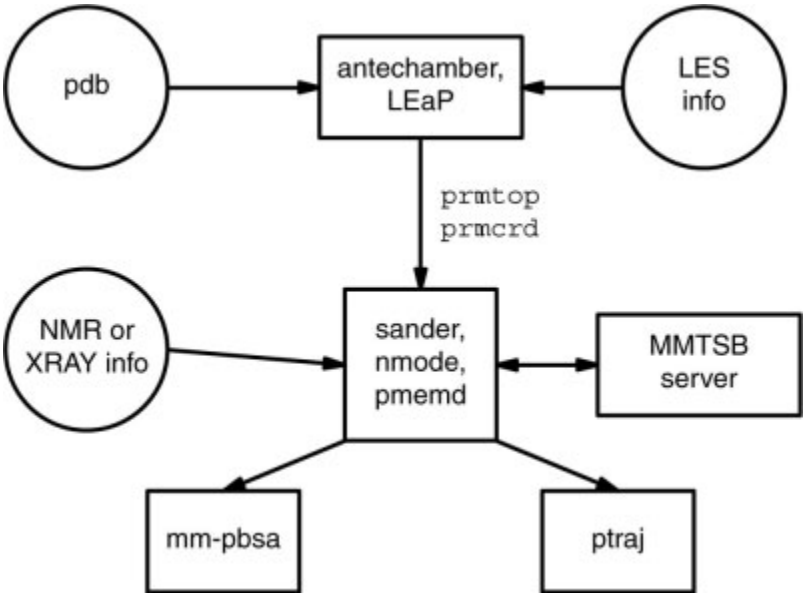


Figure 7 *Basic information flow in AMBER (from the AMBER 10 user manual).*

Simulation programs

SANDER *(Simulated Annealing with NMR Derived Energy Restraints):*

The SANDER module is responsible for processing energy minimization calculations, molecular dynamics and energy refinements on biomolecular systems. The result of preparatory programs is given as an input to SANDER, which reads the information from topology and coordinate files.

PMEMD *(Particle Mesh Ewald Molecular Dynamics):*

The PMEMD module is a recently produced reimplementation of SANDER for performance improvement in connection to parallel processing and speed. It is limited to

non-periodic simulations and does not support some features that are in SANDER, but does support some very common ones that are used more frequently on small systems.

Analysis Programs

The analysis of AMBER trajectory files produced with SANDER consists of two parts; the structural analysis and the analysis based on energies. The modules in AMBER that support these two functionalities are PTRAJ and MM-PBSA.

PTRAJ is a general purpose utility for structural analysis of bio-molecular systems. It is used to analyse and process the trajectory and coordinate files from SANDER, where the analyses might include carrying out superimposition, clustering analysis of hydrogen bonds, calculating fluctuations in bonds, angles or dihedrals, correlation functions etc.

MM-PBSA is a perl script used to do analysis based on energies for the trajectory files produced by SANDER. It automates the energy analysis of these trajectories based on the Poisson-Boltzmann continuum solvation model.

3. Materials and methods

The chapter provides detailed description of the methodology adapted in the present study. The list of HBGAs is shown in the first subsection along with the preparation of input files for subsequent minimization and molecular dynamics simulations. The following subsection discusses the MD standardized settings for all the ligands considered in the present study. A flow chart is then shown to summarize the complete methodology.

3.1 HBGAs considered in the present study

The structures shown in Table 3 were considered for the prediction of docking poses in complex with VA387. The HBGAs that were studied all depend upon the secretor gene for their biosynthesis and contain the common Fucα1-2Galβ linkage. All of the structures considered in the present work, including the crystal structures with A- and B-trisaccharides, have shown binding *in vitro* (Huang *et al*., 2003; Huang *et al*., 2005). As stated earlier crystal structures are available only for the A- and B-trisaccharides in complex with VA387 (2OBT, 2OBS).

Ligands (HBGAs)	**Formulae**
A-tri (A tri-saccharide)	GalNAcα3(Fucα2)Galβ
H-1 (H type-1 chain)	Fucα2Galβ3GlcNAcβ
H-2 (H type-2 chain)	Fucα2Galβ4GlcNAcβ
H-3 (H type-3 chain)	Fucα2Galβ3GalNAcα
Le^b	Fucα2Galβ3(Fucα4)GlcNAcβ
A-1 (A type-1 chain)	GalNAcα3(Fucα2)Galβ3GlcNAcβ
A-3 (A type-3 chain)	GalNAcα3(Fucα2)Galβ3GalNAcα
ALe^b	GalNAcα3(Fucα2)Galβ3(Fucα4)GlcNAcβ
B-1 (B type-1 chain)	Galα3(Fucα2)Galβ3GlcNAcβ
BLe^b	Galα3(Fucα2)Galβ3(Fucα4)GlcNAcβ

Table 3 *Histo-blood group antigens considered in the present study along with their empirical formulae.*

3.2 Preparation of initial structures

The global minimum energy conformations of all the ligands, except the ones in crystal structures, were predicted with the GLYGAL program, an in-house developed software that uses a genetic algorithm search method for conformational searches (Nahmany *et al*., 2005). The B-trisaccharide was obtained from the crystal structure (PDB 2OBT), whereas A-trisaccharide could not be used from the crystal structure due to serious geometry errorsof the GalNAcalpah1-3Gal moiety (bond angle of glycosidic linkage ≈ 165^o and planar geometry of C3). The A-trisaccharide was therefore modelled from B-trisaccharide of the crystal structure by the addition of an acetamido group to the terminal Gal of the sugar.

The fucose ring in all the structures with Fucα1-2Galβ linkage was superimposed on the fucose ring of B-trisaccharide in complex with VA387, using SYBYL (Tripos Inc., St Louis, USA). The protein part of 2OBT was used as the initial structure of protein for all the complexes. These protein-sugar complexes in pdb file format were then used as input for *Leap.*

3.3 Preparation of input protein-sugar complexes

All the initial structures in pdb file format were used as input for *Leap*, the preparation program for AMBER. The graphical interface of xleap was used. The force fields ff99SB (Hornak *et al*., 2006) and Glycam06 (Kirschner *et al*., 2008) were specified for protein and sugar in the complexes, respectively. Separate units were created for sugars for each of the 10 protein-ligand complexes. The protein was the same for all the complexes and therefore the same unit was used for every complex. The protein and sugar units were then merged to get a single unit for every complex. The hydrogen atoms were added through xleap. The topology and coordinate files were then created through Leap to be used as input for minimization and molecular dynamics simulations by SANDER.

3.4 Energy minimization

All the MD simulations and minimizations were carried out using the SANDER program of the molecular dynamics package AMBER 8 (Case *et al*., 2004). The generalized Born solvation model (Onufriev *et al*., 2004) was used for both the minimization and MD simulations. A total of 200-500 cycles of minimization were carried out on each system before the MD runs. The first half of the cycles used the steepest descent algorithm whereas the conjugate gradient algorithm was employed in the last half of the cycles. Almost all of the structures showed energy convergence with an energy decrease of <1 kcal/mol between subsequent steps. Position constraints of 50 kcal/mol and 500 kcal/mol, respectively, were applied on main chain protein atoms and the fucose ring of the sugar. The rest of the atoms were free to move during the minimization cycles.

3.5 Molecular dynamics simulations

SANDER was also used for the molecular dynamics equilibration and production runs. The system was first equilibrated gradually from 0 to 300 K for 10 ps with 1 fs time steps. The position constraints on the fucose ring of the sugars during equilibration and production run were relaxed from 500 kcal/mol/Å^2 to 50 kcal/mol/Å^2. This equilibration was then followed by 250 ps of production run at 300 K. During the production run the side chains of protein residues that were within 10 Å from the centre of mass of ligand were allowed to move whereas position constraints of 50 kcal/mol/Å^2 were applied to the protein main-chain atoms in that area. All other protein atoms were frozen with position constraints of 50 kcal/mol/Å^2.

During both the equilibration and production runs the SHAKE procedure was used to constrain all solute bonds containing at least one hydrogen atom. Temperature regulation was realized by the use of Langevin Dynamics with a collision frequency of 1.0 ps^{-1} and a non-bonded cut-off of 12 Å was used to truncate the non-bonded pairs. The snapshots of the trajectory were taken at every 1 ps.

3.6 Analysis of output files

The molecular dynamics trajectory files produced by SANDER contain all the detailed structural information about the system under study. Ptraj, the analysis program for AMBER, was used to fetch this information and to do the structural analysis for the complexes studied in the present work. Mass-weighted root mean square deviation (RMSd) was calculated for sugar rings in all 250 frames (corresponding to 1 ps each) with the first frame used as reference. Since the α-fucose ring was restrained during the whole course of the production dynamics, it was not included in the calculation of RMSd. Dihederals for all glycosidic linkages were calculated using the Ptraj program of AMBER. A small perl script was used to read out the thermodynamic parameter values (energy, pressure, temperature, volume etc) for the system from all 250 output files.

GLYGAL was used to plot the glycosidic dihederal angles as dots superimposed on the MM4 adiabatic maps.

3.7 GLIDE scoring[3]

The last 6 snapshots with an interval of 10ps from the MD simulations were used for scoring of interaction energies using Glide XP. The snapshots were first minimized using Macromodel with the OPLS2001 force field because of the dependence of Glide XP scoring function parameters on the OPLS force field. For each protein-ligand complex a grid with the grid size of 19 Å centred on the ligand was calculated. The Glide XP "refine and score" calculations were then performed with rigid protein. The sugars were minimized in place. A maximum of 1000 iterations of conjugate gradient in OPLS2001 force field were used for refinement of each protein-ligand complex and the resulting scores were calculated with Glide XP and finally averaged.

3.8 Identification of protein-ligand interactions

The final 6 snapshots for each ligand after OPLS minimization were then studied through the LIGPLOT program to generate protein-ligand interaction diagrams. The PDBSUM (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/) website was used to produce the LIGPLOT results. The unions of protein-ligand interactions found in the resulting diagrams (6 for each protein-ligand complex) were taken to obtain the final results.

3.9 Summary

The following figure (Figure 8) summarizes the complete methodology adapted in the present study.

[3] The Glide scoring was performed by Chaitanya Koppisetti at Biognos AB.

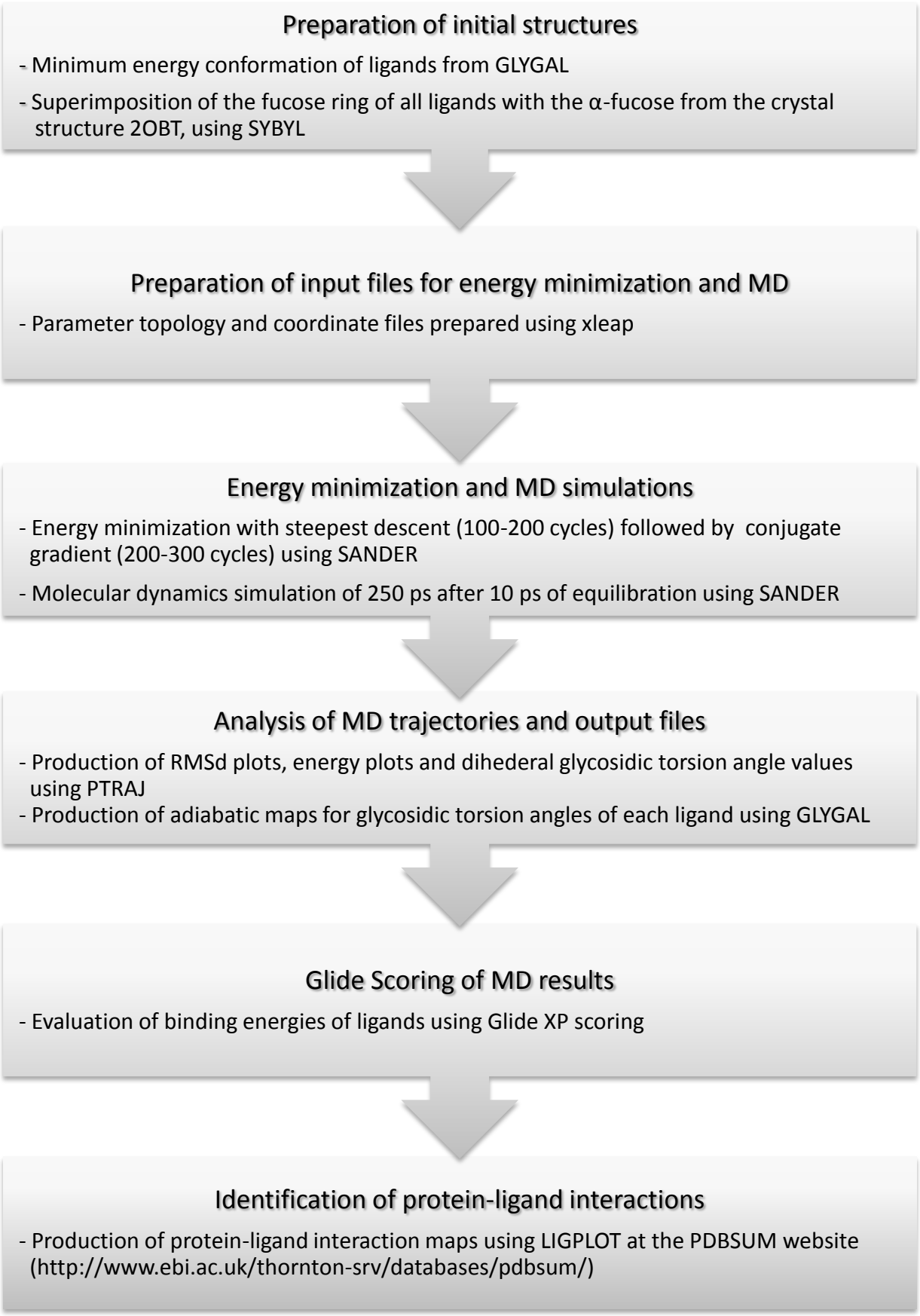


Figure 8 *Summary of the methodological steps for the present study.*

4. Results and Discussion

This chapter discusses the quantitative as well as qualitative results obtained from the present work. The mass-weighted RMSd plot and the energy plot for each ligand are shown along with the adiabatic maps for the glycosidic linkages. The final docking poses, as a result of 250 ps MD production run, are also shown followed by the discussion on the protein-ligand interactions which are found to be in agreement with the published mutagenesis data (Tan *et al*., 2008). After the detailed elucidation of protein-ligand interactions on atomic level, the chapter is concluded by listing, in a nutshell, the main findings of the present work.

4.1 Initial fit of ligands in the VA387 binding site

The structures of the complexes after superimposition on the fucose of crystal structure and a brief minimization in Sybyl were inspected. The type-1 and type-3 structures displayed a very good fit without any clashes with the binding site residues of VA387. This is shown in figure 9.

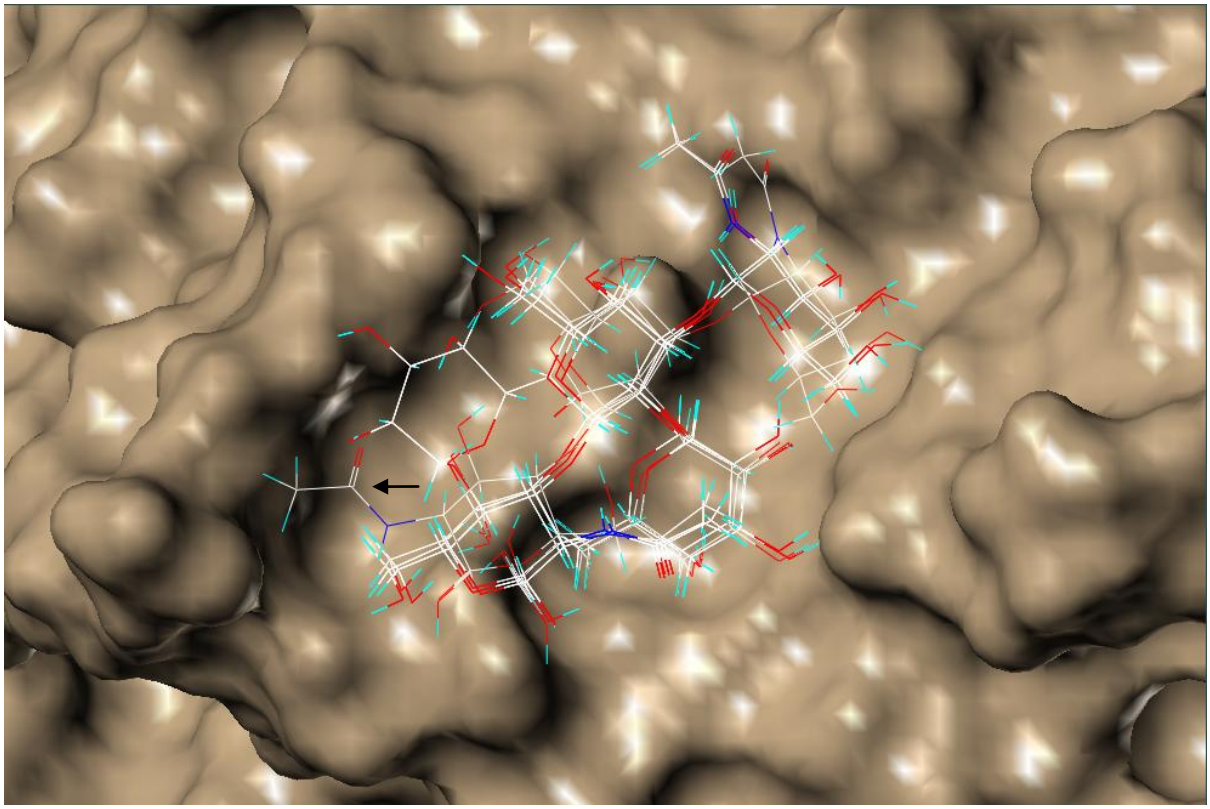


Figure 9 *Cluster view of the initial structures of all the ligands, particularly type-1 and type-3, considered in the present study. Protein is shown as a brown coloured surface whereas ligands are represented as wireframe model with CPK colouring. All the type-1 and type-3 initial structures have similar fit whereas H-2, the only type-2 structure considered for MD, shows the acetamido group of GlcNAc to be in the opposite orientation (indicated by arrow).*

4.1.1 The type-2 structures

The type-2 structures (H-2, A-2, ALe^y) showed some clashes with D391 and Q390 in the initial conformation. These clashes were resolved after AMBER minimization but the fit was still not as good as for the type-1 and type-3 structures. It is the NAc-group of the GlcNAc

residue at the reducing end of the sugar in type-2 structures that clashes with protein residues as shown in figure 10.

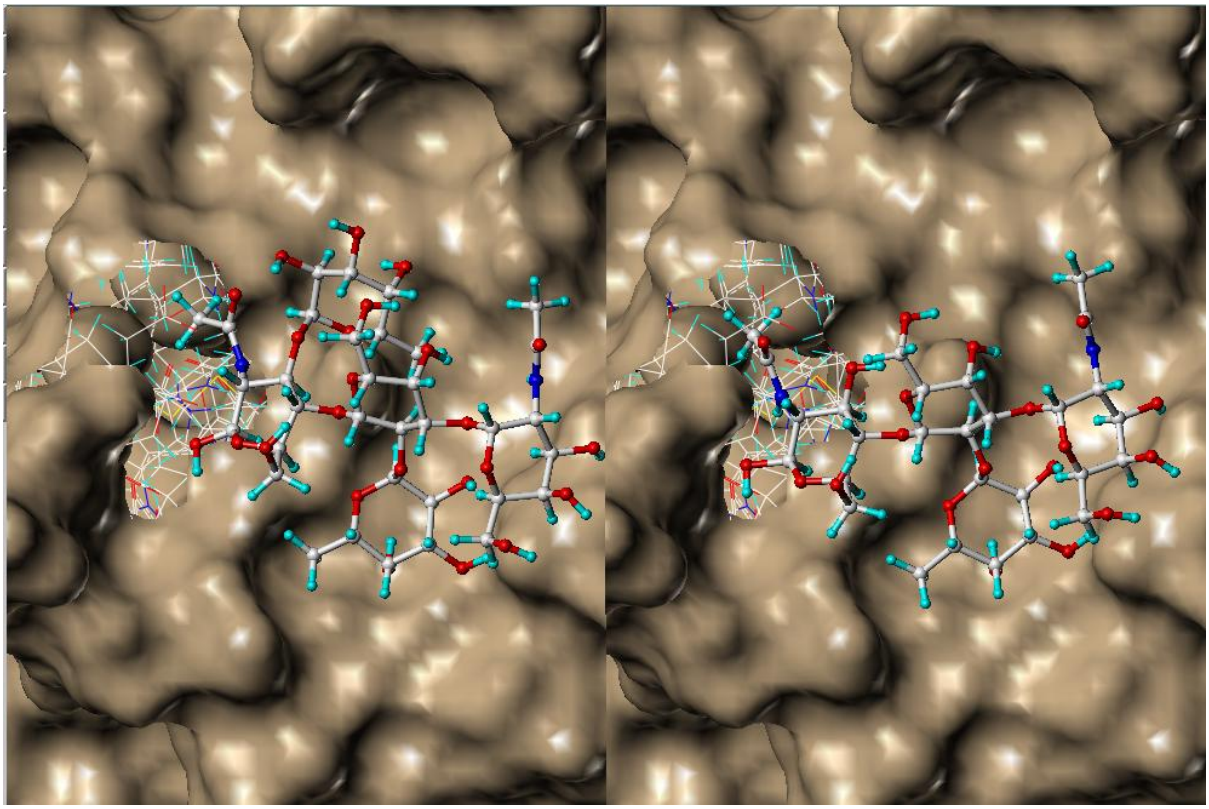
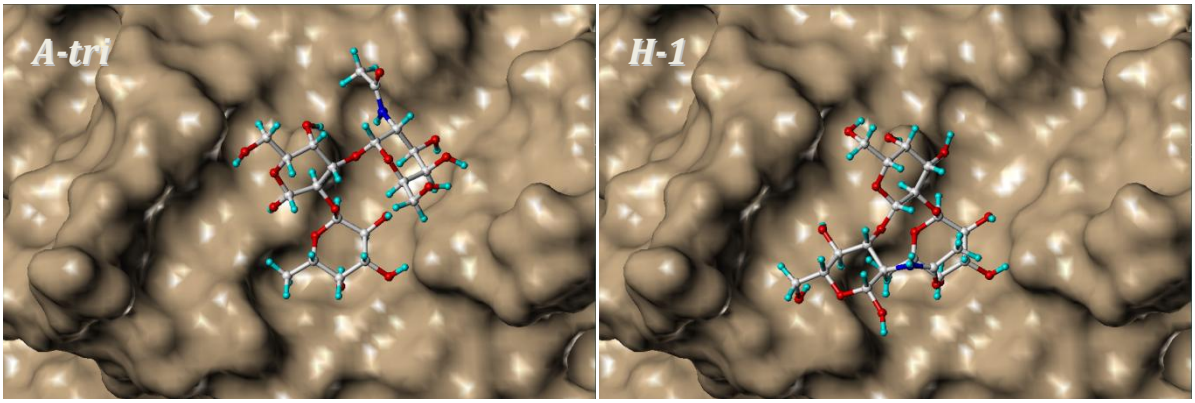


Figure 10 *ALe^y (left) and A-2 (right) initial structures in complex with VA387 binding site. The protein is coloured brown with surface representation whereas the ligands are represented as ball-and-stick model with CPK colouring. To show the clashes the residues Q390 and D391 are not included in the surface drawing. The acetamido group of GlcNAc at the reducing end of the sugar clashes with Q390 and D391.*

4.1.2 Starting structures for MD simulations

The final structures from minimization were used as initial structures for MD (Figure 11). Of the type-2 structures only H-2 was used since preliminary MD runs on A-2 and ALe^y produced the trajectories which were not in agreement with the Φ/Ψ adiabatic energy maps for glycosidic linkages.



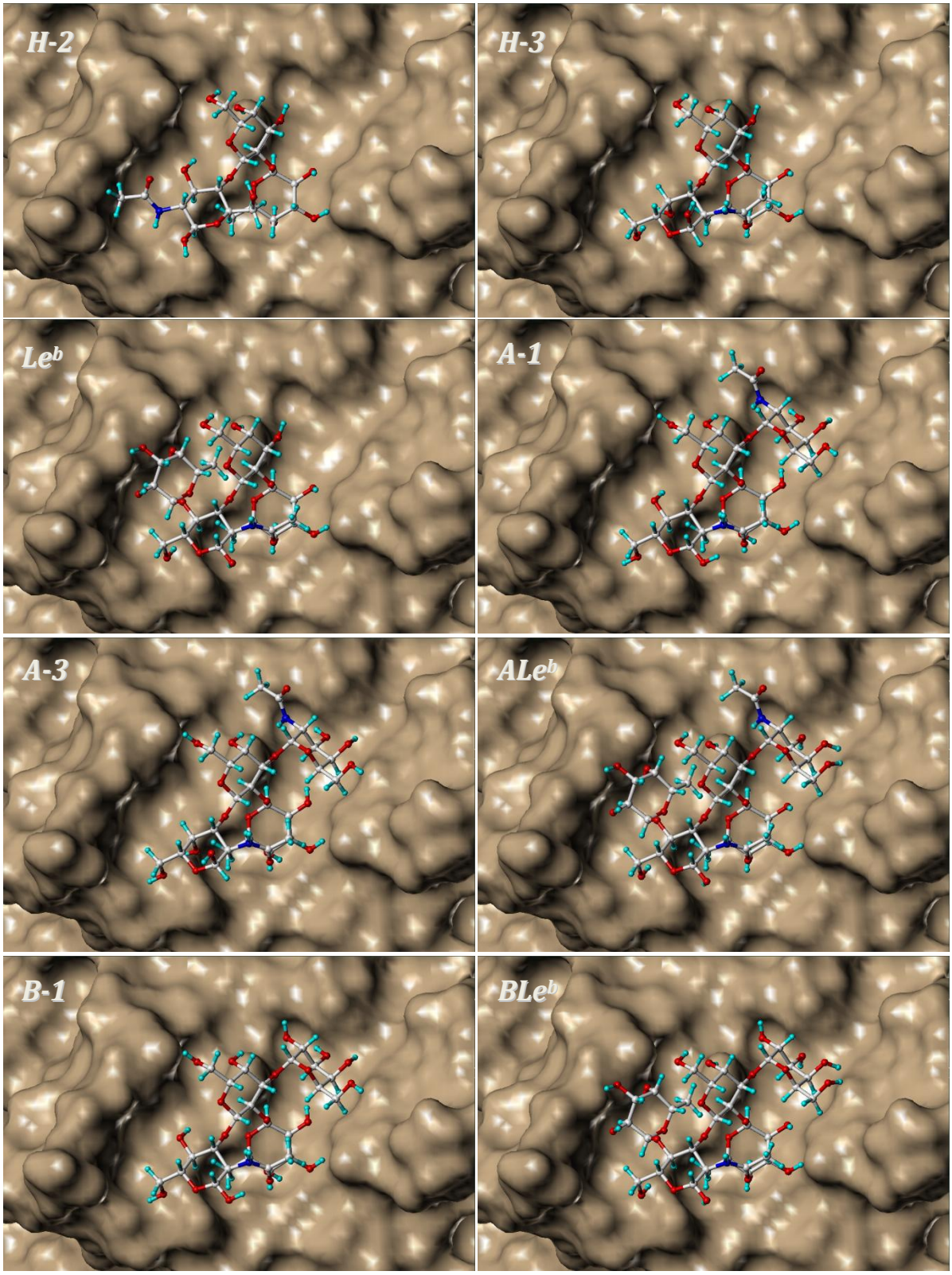


Figure 11 *Individual initial structures for MD simulations considered in the present study. Protein is shown as brown coloured surface whereas the ligands are represented as ball-and-stick models with CPK colouring. All the ligands illustrated fit nicely without steric clashes.*

4.2 Molecular dynamics simulations

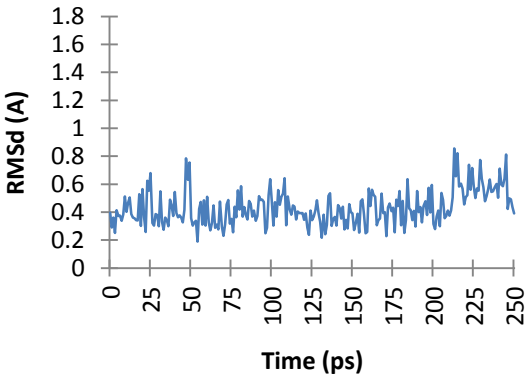
Molecular dynamic simulations of 250 ps for all 10 ligands (Table 4) in complex with the VA387 crystal structure displayed stable trajectories over time. The mass weighted RMSd plots of all the ligands, not accounting for the restrained fucose ring, are displayed along with the energy plots to show the stability of the system over the course of the simulations.

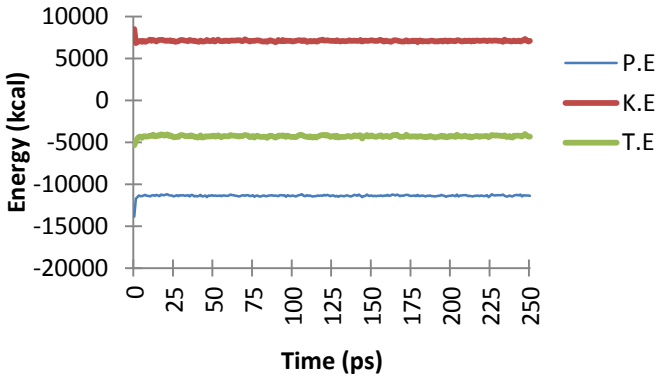
Ligand	Mean RMSd (Std. Dev)
A-tri	0.4275(0.1248)
H-1	0.6983(0.1297)
H-2	0.7107(0.1111)
H-3	0.7092(0.1024)
Le[b]	0.6124(0.1306)
A-1	0.7340(0.1540)
A-3	0.7457(0.1171)
ALe[b]	0.6515(0.1073)
B-1	0.7337(0.1481)
BLe[b]	0.6692(0.1048)

Table 4 *Mean RMSd and standard deviations are shown for all the ligands considered in the present study. The RMSd only includes heavy atoms in sugar residues other than the restrained fucose.*

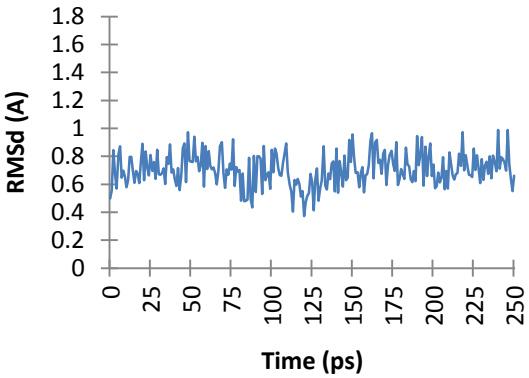
The results of the molecular dynamics simulations show that the system has been reasonably stable over 250 ps of production dynamics. The adiabatic maps of the glycosidic linkages in each ligand were also produced with GLYGAL and are shown in appendix A.

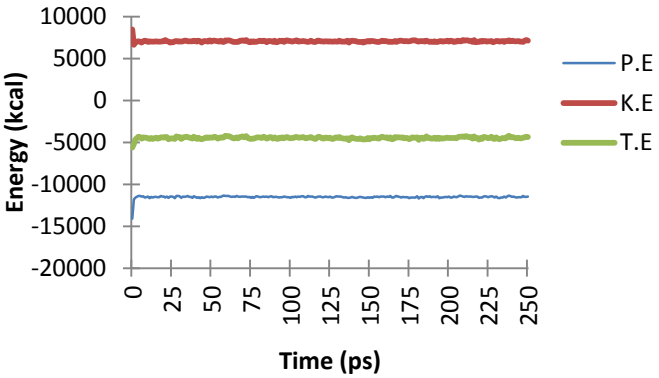
A-tri



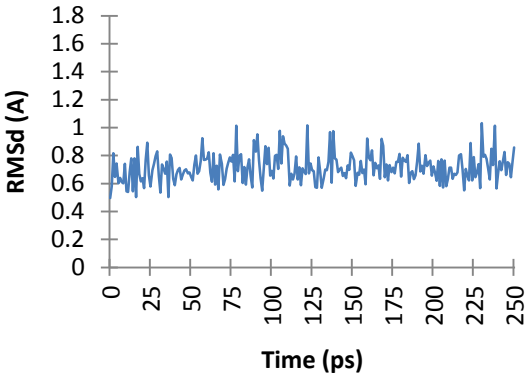


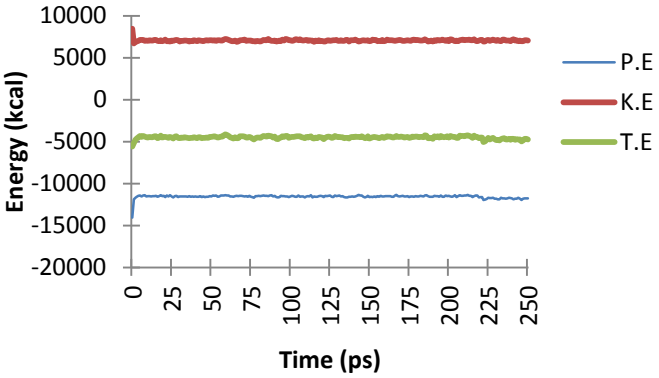
H-1


RMSd (A)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 25 50 75 100 125 150 175 200 225 250
Time (ps)

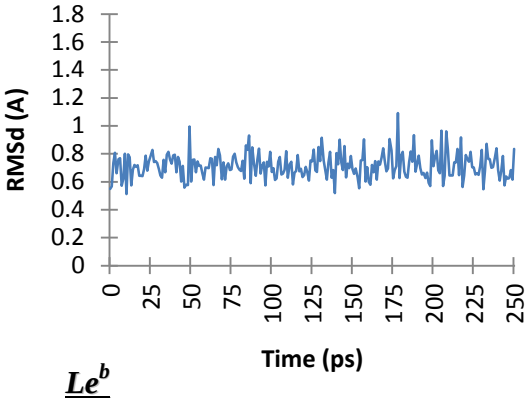

Energy (kcal)
10000
5000
0
-5000
-10000
-15000
-20000
0 25 50 75 100 125 150 175 200 225 250
Time (ps)
P.E
K.E
T.E

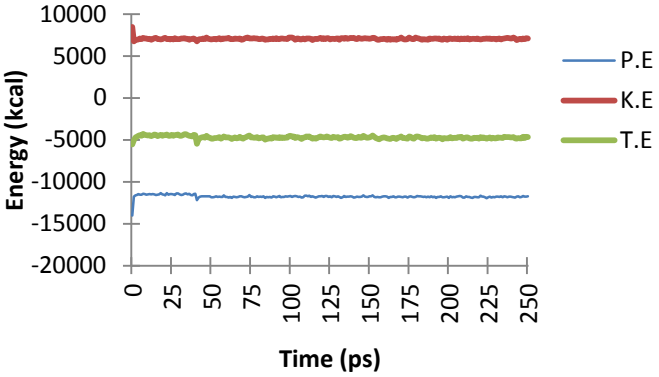
H-2


RMSd (A)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 25 50 75 100 125 150 175 200 225 250
Time (ps)

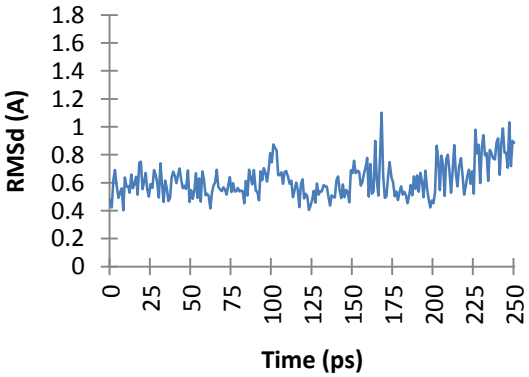

Energy (kcal)
10000
5000
0
-5000
-10000
-15000
-20000
0 25 50 75 100 125 150 175 200 225 250
Time (ps)
P.E
K.E
T.E

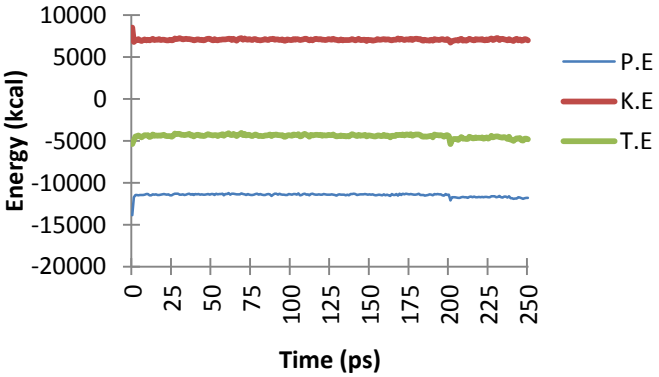
H-3


RMSd (A)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 25 50 75 100 125 150 175 200 225 250
Time (ps)

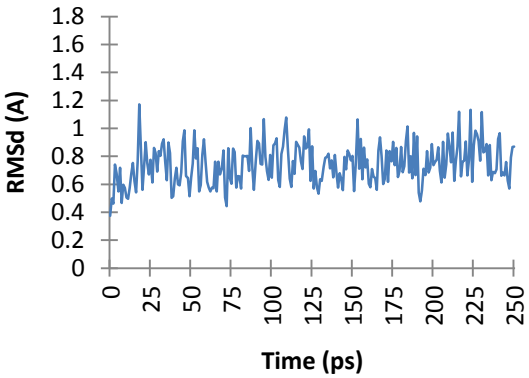

Energy (kcal)
10000
5000
0
-5000
-10000
-15000
-20000
0 25 50 75 100 125 150 175 200 225 250
Time (ps)
P.E
K.E
T.E

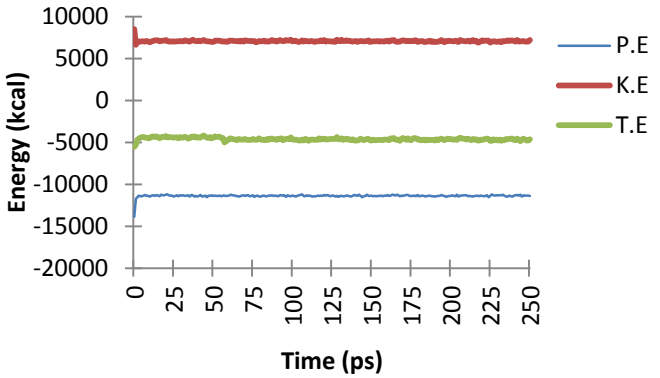
Le^{b}


RMSd (A)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 25 50 75 100 125 150 175 200 225 250
Time (ps)

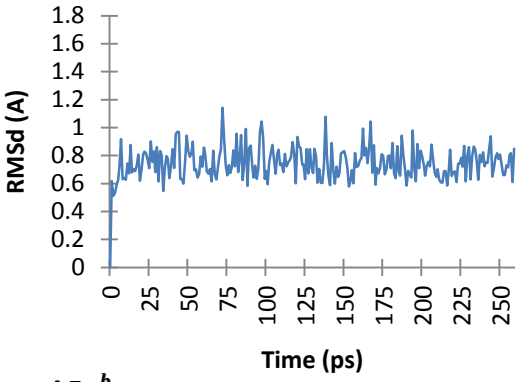

Energy (kcal)
10000
5000
0
-5000
-10000
-15000
-20000
0 25 50 75 100 125 150 175 200 225 250
Time (ps)
P.E
K.E
T.E

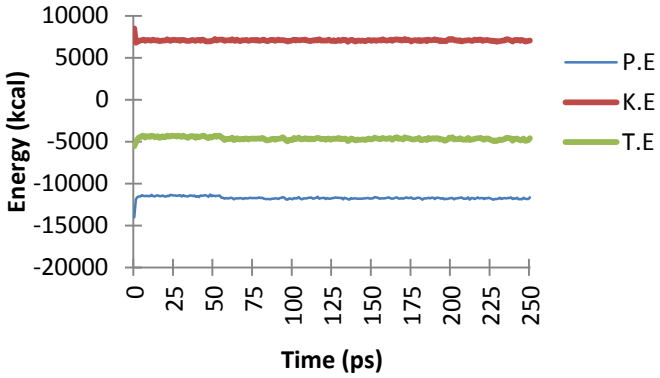
A-1


RMSd (A)
Time (ps)

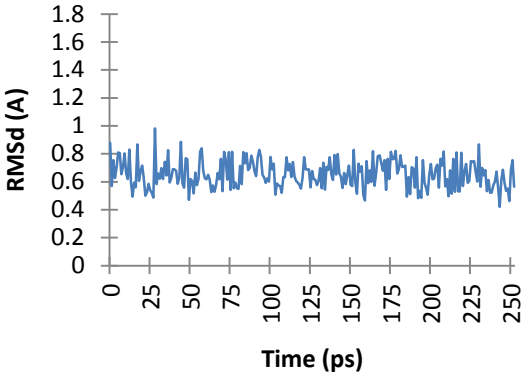

Energy (kcal)
Time (ps)
P.E
K.E
T.E

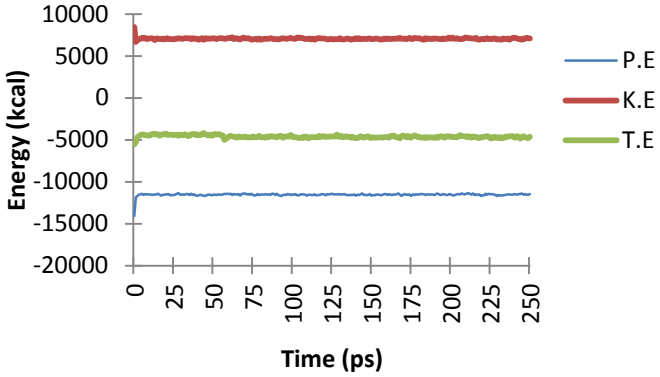
A-3


RMSd (A)
Time (ps)

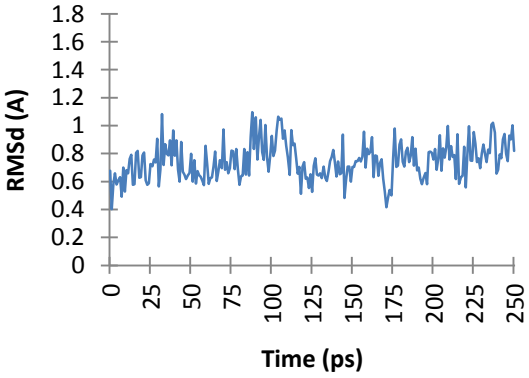

Energy (kcal)
Time (ps)
P.E
K.E
T.E

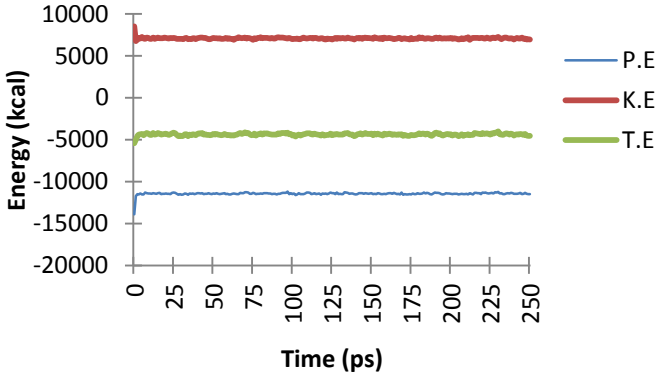
ALe^{b}


RMSd (A)
Time (ps)


Energy (kcal)
Time (ps)
P.E
K.E
T.E

B-1


RMSd (A)
Time (ps)


Energy (kcal)
Time (ps)
P.E
K.E
T.E

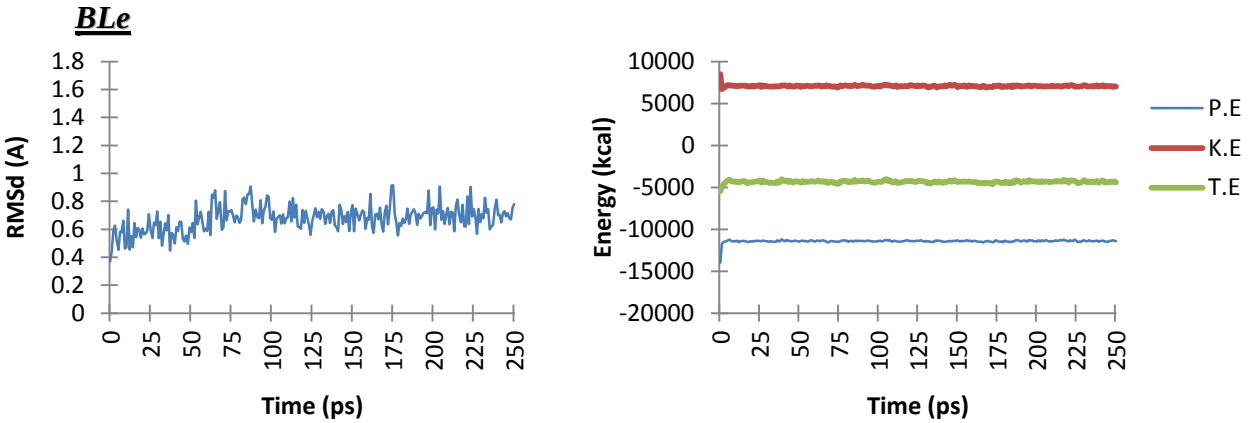


Figure 12 *The RMSd and energy plots, including potential energy (P.E), kinetic energy (K.E) and total energy (T.E) of the system for each ligand. These plots show stable trajectories over 250 ps for each ligand.*

4.2.1 The length of molecular dynamics simulations

Equilibration of 10 ps and production dynamics of 250 ps were performed using AMBER 8.0. The length of the simulations was relatively shorter as compared to the normal MD simulations performed in protein-carbohydrate interaction studies (typically a few nano-seconds). The reason for reduced length of production dynamics in this case was the restrained fucose. This reduced the possible number of conformations and orientations that could be sampled through the MD simulations in contrast to a free unrestrained MD run. For some of the ligands (A-1, Le^b) the length of the production dynamics was extended by about 100 ps. This did not result in any substantial difference from the observed behaviour during the 250 ps of simulation. The RMSd plots shown also illustrate that the systems were fairly stable over the 250 ps of production dynamics.

4.3 GLIDE scoring of MD results

The results from GLIDE scoring performed on snapshots of the final part of MD are summarized in Table 5. The averages are shown in the right column. According to the scores calculated with GLIDE, ALe^b, A-1, BLe^b and B-1 are the best binders with scores < -8 kcal/mol. The lowest score was calculated for BLe^b. B structures showed about 1.5 kcal/mol weaker scores than the corresponding A structures suggesting that the acetamido group of the terminal GalNAc does not strengthen the binding. Further studies are implicated to investigate the structural details of the interactions of the acetamido group of the A-structures in these complexes.

A-tri and Le^b have intermediate GLIDE scores whereas the H-1, H-2 and H-3 structures have the highest interaction scores (> -6.2). This is in agreement with experimental observations. The fucose monosaccharide, in the same orientation as in the crystal structure, has a score of -5.9 kcal/mol. The small difference between the fucose monosaccharide and the H structures suggests that strong interactions are made by the fucose monosaccharide, as described by the crystal structure.

Ligand	200ps	210ps	220ps	230ps	240ps	250ps	Average XP score (kcal/mol)
Fucose monosaccharide	-6.7	-6.6	-6.5	-6.6	-6.6	-6.5	-6.6 -5.9 (crystal)
A-trisaccharide	-6.6	-9.7	-6.7	-9.0	-9.3	-7.8	-8.2
B-trisaccharide	-8.9	-10.6	-8.3	-7.6	-6.6	-9.4	-8.6
H-1	-7.6	-7.2	-7.1	-4.3	-7.2	-8.0	-6.8
H-2	-4.9	-6.7	-5.5	-5.8	-5.2	-5.4	-5.6
H-3	-6.7	-6.7	-6.1	-6.7	-7.	-6.6	-6.7
Le^b	-8.2	-6.7	-9.4	-7.4	-6.9	-7.4	-7.7
A-1	-8.5	-9.5	-7.5	-9.6	-10.1	-9.9	-9.2
A-3	-7.6	-10.7	-6.6	-10.2	-8.1	-9.1	-8.7
ALe^b	-10.4	-9.4	-9.5	-9.5	-9.5	-10.6	-9.8
B-1	-10.2	-8.4	-11.0	-12.7	-9.9	-8.3	-10.1
BLe^b	-12.7	-9.7	-12.6	-11.0	-10.7	-11.7	-11.4

Table 5 *Interaction scores for all the ligands calculated using Glide XP scoring. The last 6 snapshots from the production dynamics with an interval of 10ps are considered. The final score is the average of all 6 scores for the ligand.*

Similarly, the substantial difference in scores between the H-structures and the A/B-structures (ca 3 kcal/mol) illustrate the role of terminal Gal/GalNAc, the only difference between the H- and A/B-antigens, which is responsible for enhanced binding affinity of A/B-structures as compared to the H-antigens.

4.4 The final docking poses

The protein-ligand interaction diagrams for all ligands from Glide scoring were produced using the LIGPLOT program from the PDBSUM website (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/). The final results obtained from the present work could be studied by classifying the ligands in four different categories as follows.

1. The H-antigens
2. The A-antigens
3. The B-antigens
4. The secretor gene dependent Lewis structures

The results for each of these groups are discussed at length in the following subsections. The labelling of the pictures in these sections represents the interacting residues found as a result of the union taken over all 6 interaction diagrams for each of 10 ligands considered. The interactions with the fucose in the binding site are the same as in the crystal structure because of the position restraints on fucose throughout the course of the simulation.

4.4.1 The H-antigens

H-1, H-2 and H-3 were considered for Glide XP scoring and for the production of interaction diagrams.

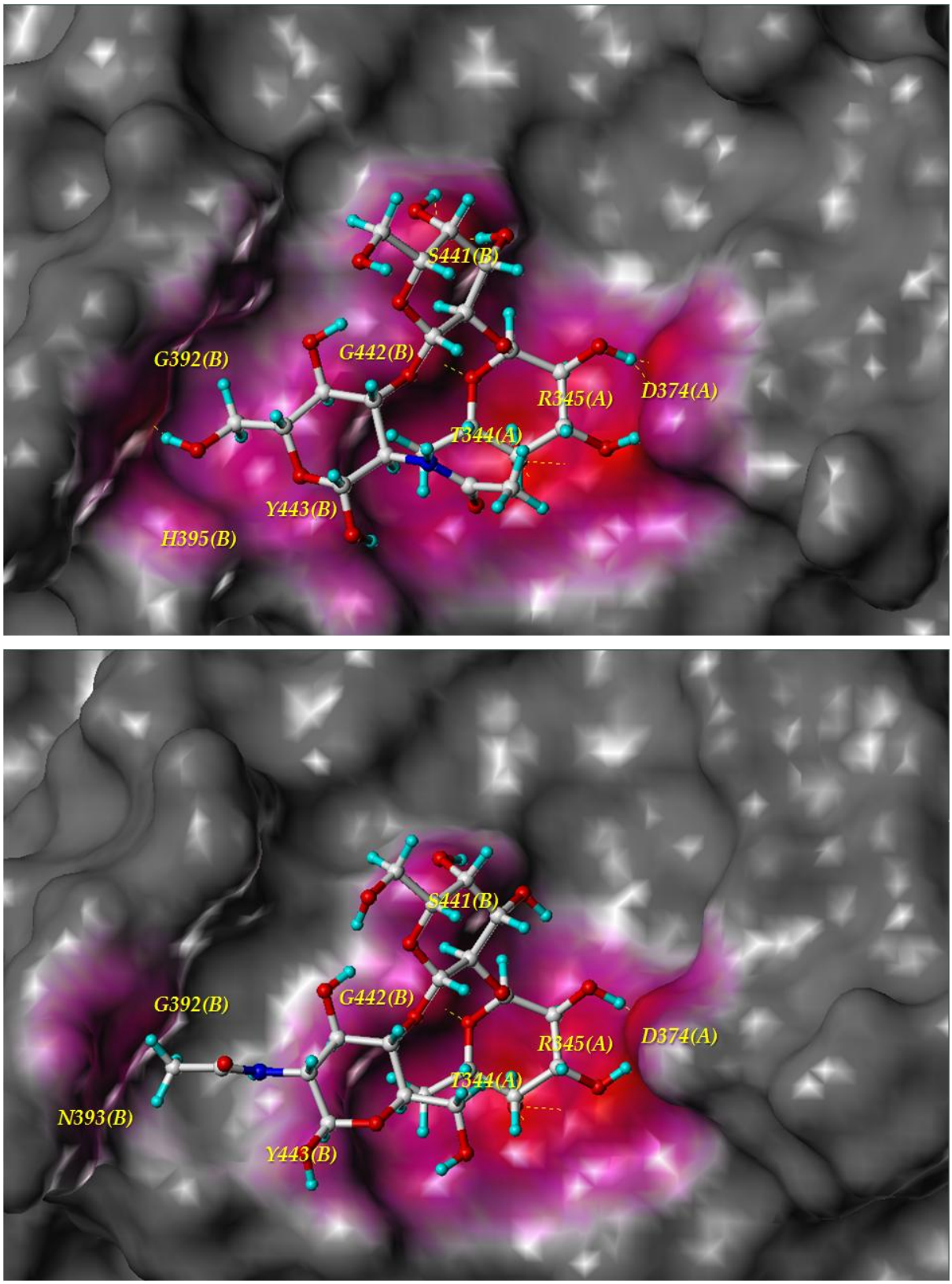


Figure 13 *The enlarged view of H-1 (top) and H-2 (bottom) in complex with VA387 with a distance threshold of 3 Å. Only the residues that are found to be interacting as per LIGPLOT analysis are labelled. Protein is shown as MOLCAD surface whereas the ligand is displayed as ball-and-stick model.*

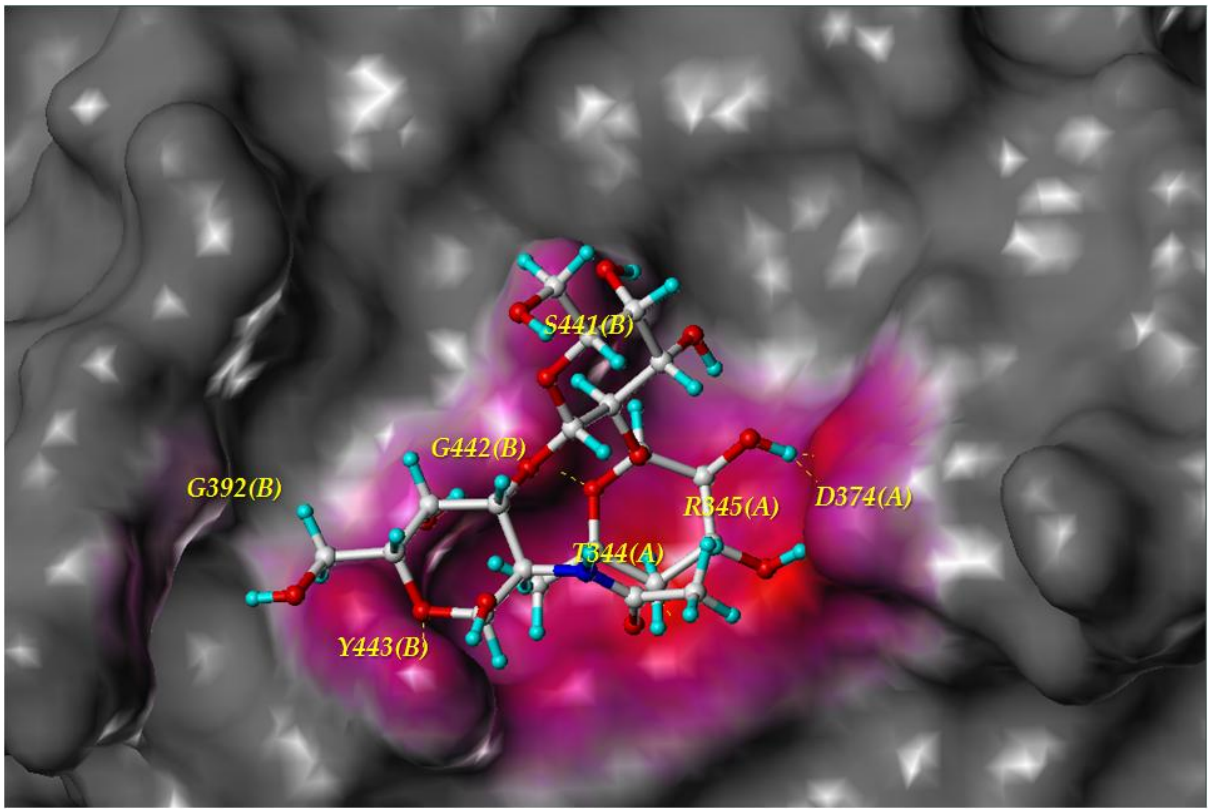


Figure 14 *The enlarged view of H-3 in complex with VA387 with a distance threshold of 3 Å. Only the residues that are found to be interacting as per LIGPLOT analysis are labelled. Protein is shown as MOLCAD surface whereas the ligand is displayed as ball-and-stick model.*

The weak binding affinity for H-structures might be explained through the smallest number of interactions among all the ligands found with LIGPLOT. Apart from the fucose residue in the binding pocket, the GlcNAcβ residue at the reducing end of the sugar makes some interactions with the residues Y443, G392 and Q390 of protein. The O1 atom of GlcNAcβ, at the reducing end of the sugar in case of H-1 (Figure 13) is close to the protein surface whereas in case of H-3 it points towards the bulk region outside protein (Figure 14). This might explain the relatively stronger binding affinity of type-3 conjugates compared to type-1 conjugates (Huang *et al*., 2005).

The only type-2 structure that is considered in the present study is H-2 (Figure 13). The clear difference in the orientation of NAc-group in the GlcNAcβ residue at the reducing end of sugar, visible from the picture, is responsible for the hydrophobic interactions between the methyl group of GlcNAc and the side chain of N393. The GlcNAcβ residue in H-1, on the other hand, has a hydrogen bonding interaction with protein.

In a nutshell, the H-antigens show the smallest number of interactions with the protein residues in the binding site and are scored as the weakest binders among all the antigens that were considered for the present study. Moreover, the critical role of secretor fucose for the binding affinity of H-antigens is also evident from the protein-ligand interaction diagrams.

4.4.2 The A-antigens

The modelled A-trisaccharide shows strong interactions between the terminal GalNAcα and protein residues in the binding site (Figure 15). The mutagenesis studies (Tan *et al*., 2008),

however, have shown some mutations to be crucial for binding to A-antigens. These include K348, Q331 and I389. The mutations of these residues to alanine have resulted in significantly reduced binding to A-antigens. The present study reported here has shown these residues to be involved in strong hydrogen bonding and hydrophobic interactions with the terminal GalNAcα residue of A-tri and A-1 (Figure 15, 16 and Appendix B). Thus the modelled A-trisaccharide complex is in agreement with mutagenesis results. Chaitanya Koppisetty (at Biognos AB) had previously shown that the crystal structure of VA387/A complex has serious geometry errors in the GalNAcα-Galβ disaccharide moiety[4].

It therefore appears that the failure to relate the crystal structure to the mutational data is at least partially due to these geometry errors.

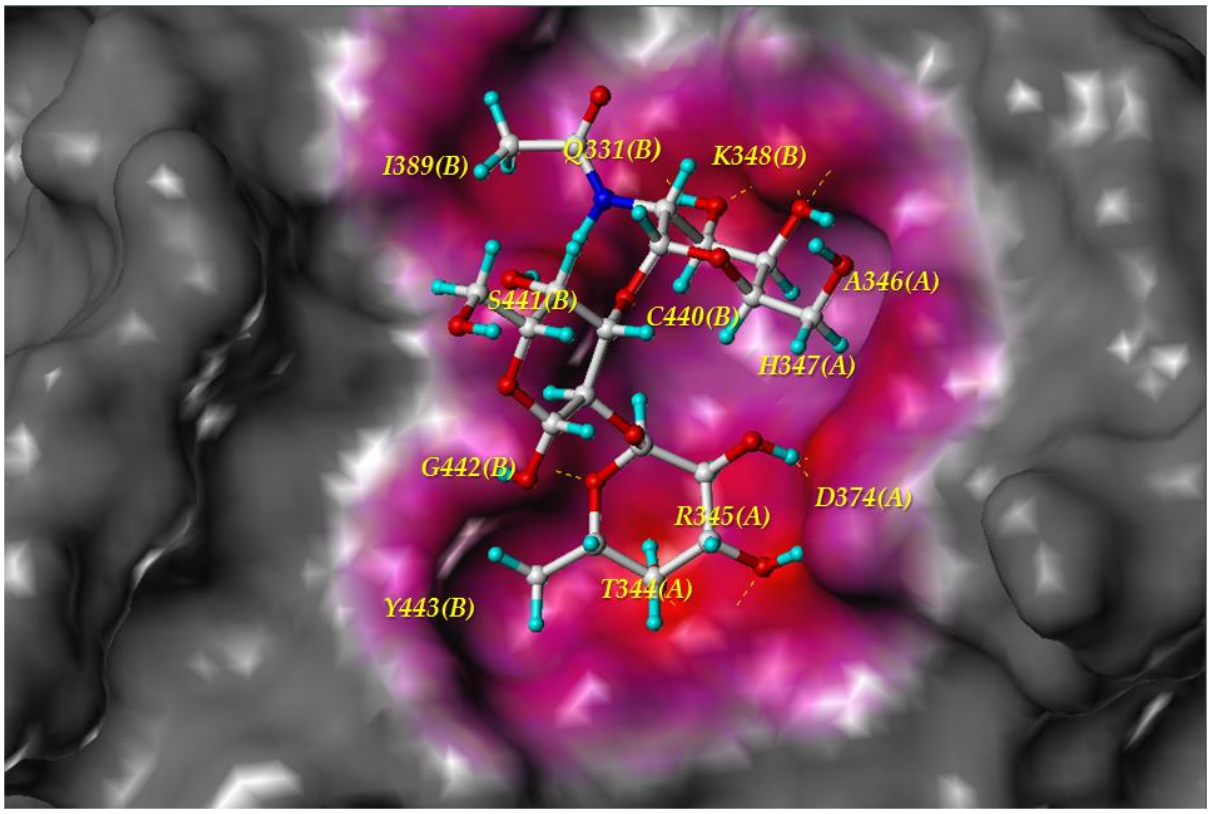


Figure 15 *The enlarged view of A-trisaccharide from MD simulations in complex with VA387 with a distance threshold of 3 Å. Only the residues that are found to be interacting as per LIGPLOT analysis are labelled. Protein is shown as MOLCAD surface whereas the ligand is displayed as ball-and-stick model.*

The difference between the A-1/3 and H-1/3 is only the presence of terminal GalNAcα in the A-antigens. The increased number of interactions found in A-antigens is due to the presence of this terminal GalNAcα residue which clearly explains the better Glide score and stronger binding affinity in case of A-antigens in contrast to the H-antigens (Huang *et al*., 2003; Huang *et al*., 2005). The GlcNAcβ residue at the reducing end of sugar is found to have the same interactions as in H-1.

[4] Koppisetty, C.A.K, Nasir, W., Rydell, G.E., Strino, F., Larson G. and Nyholm P.G. *Docking of histo-blood group ABO-active saccharides with the norovirus VA387 capsid protein can explain experimental binding data*. Manuscript in preparation

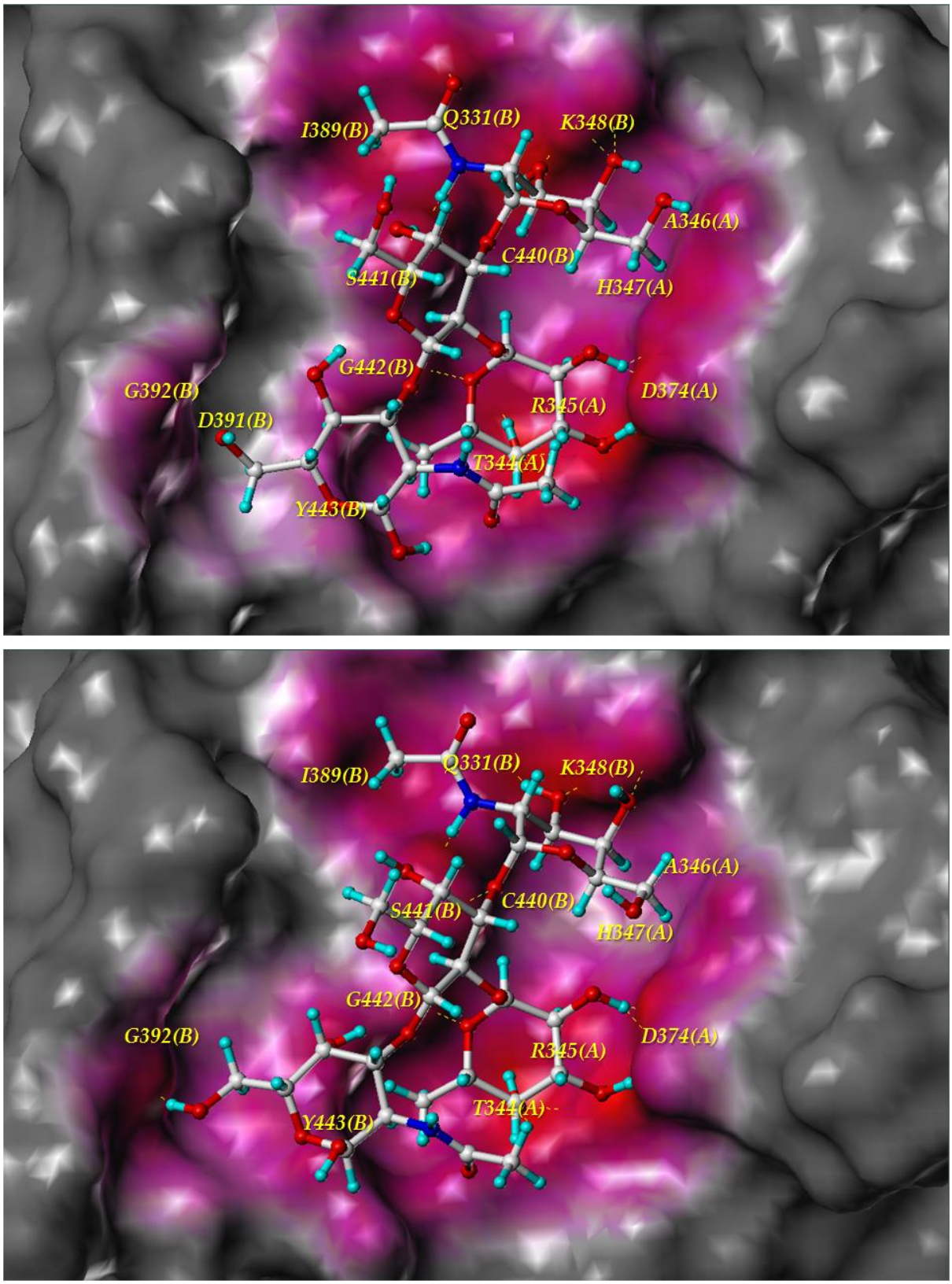


Figure 16 *Enlarged view of A-1 (top) and A-3 (bottom) seen in complex with VA387 with a distance threshold of 3 Å. Only the residues that are found to be interacting as per LIGPLOT analysis are labelled. Protein is shown as a MOLCAD surface whereas the ligand is displayed as ball-and-stick model.*

4.4.3 The B-antigens

B type-1 was the only ligand considered in this category. The terminal GalNAcα in A-1 is replaced by terminal Galβ in B-1, the only difference between the two. The absence of acetamido group at the terminal residue of B-1 results in the loss of a couple of interactions but surprisingly in a somewhat lower Glide score (Figure 17). This might be due to slight clashing in the case of the A-structures. The terminal Galα in B-1 is found to be interacting with K348 and Q331. The mutations of K348, Q331 and I389 have no effect on B-antigens but reduce binding of A-antigens significantly. The loss of interactions with I389 is in complete agreement with this fact but the interactions with K348 and Q331 could not be explained on a structural basis. K348 makes hydrogen bonds with O3 and O4 atoms of the terminal Galα/GalNAcα residue, whereas Q331 interacts with the O2 atom of the same sugar residue according to our model. Therefore the mutations Q331A and K348A are only in partial agreement with mutational mutagenesis data.

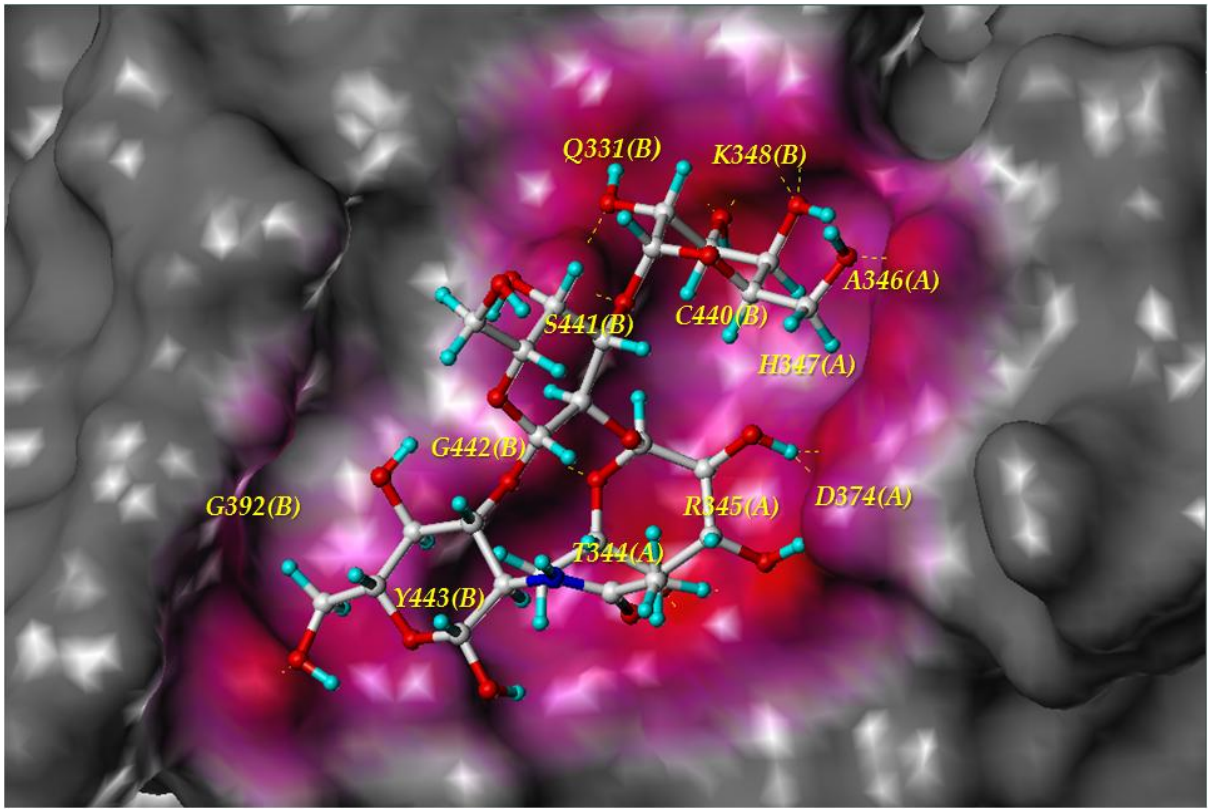


Figure 17 *Enlarged view of B-1 seen in complex with VA387 with a distance threshold of 3 Å. Only the residues that are found to be interacting as per LIGPLOT analysis are labelled. Protein is shown as a MOLCAD surface whereas the ligand is displayed as ball-and-stick model.*

The Galβ in the Fucα1-2Galβ linkage and the GlcNAcβ at the reducing end found the same interactions as in case of A-1 or H-1. The interactions of the fucose were also retained to be the same as in the crystal structure.

4.4.4 The secretor gene dependent Lewis structures

This group contains Le^b, ALe^b, and BLe^b. All the Lewis structures are characterized by the α1,4-linked fucose, the Lewis epitope, to the GlcNAcβ at the reducing end of the sugar. This internal fucose when added to H-1, A-1 and B-1 produces Le^b, ALe^b and BLe^b structures, respectively. The internal fucose gives rise to the hydrogen bonding interactions with the residues Q390, D391 and G392. This is also in agreement to the mutagenesis studies which suggested that this binding region might be involved in the interactions with the downstream extensions of HBGAs (Tan *et al*., 2008). The mutations of the residues Q390, D391 and G392 to alanine resulted in decreased binding affinity to A- and B- antigens in saliva assays which include also the A- and B- structures with downstream extensions.

The terminal Galα/GalNAcα residue in ALe^b/BLe^b showed the same interactions with the binding site as in A-1/B-1. BLe^b did not show the interactions with Q331 and I389 whereas ALe^b was involved in strong hydrogen bonding interactions in this region. Similarly the GlcNAcβ residue in all the Lewis structures was found to be interacting with Y443. In the Lewis negative structures (structures lacking Lewis epitope) the GlcNAcβ at the reducing end of the sugar was found to be interacting with G392 and Q390, whereas in Lewis positive structures these interactions are replaced by the α1,4-linked fucose (Figure 18, 19).

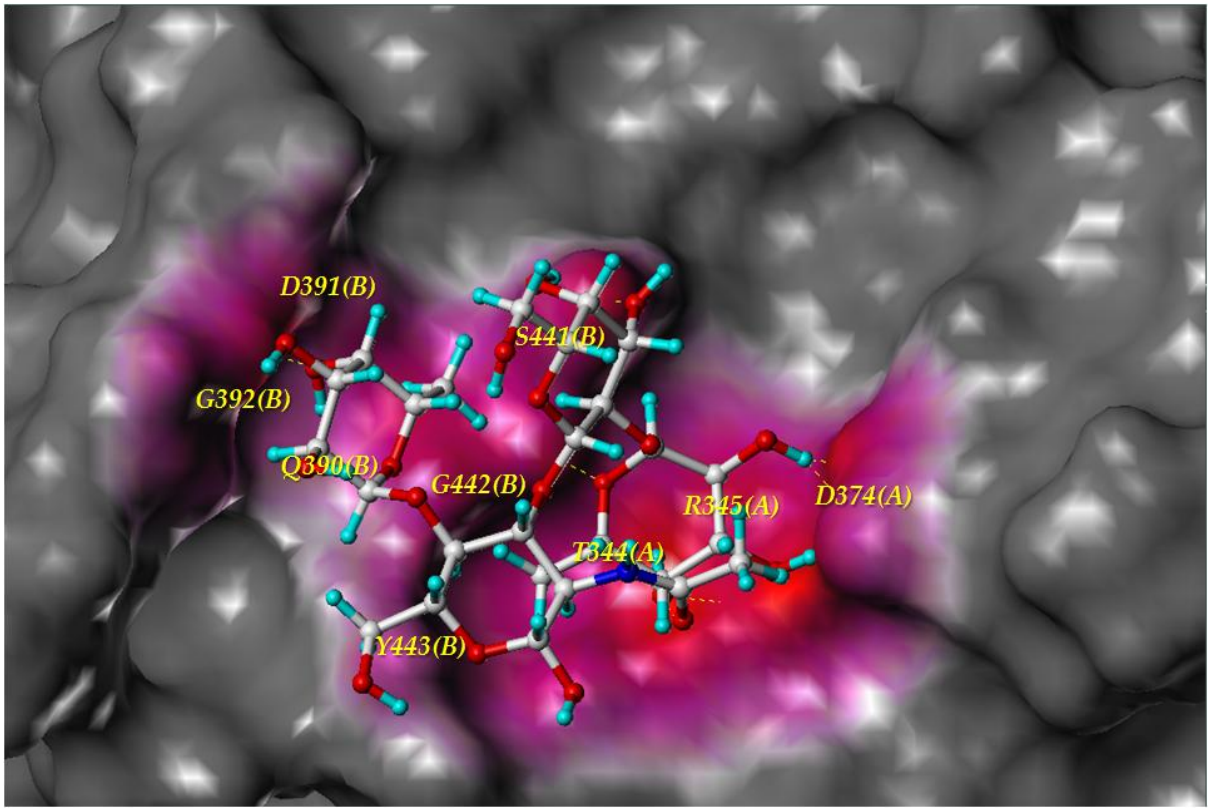


Figure 18 *Enlarged view of Le^b seen in complex with VA387 with a distance threshold of 3 Å. Only the residues that are found to be interacting as per LIGPLOT analysis are labelled. Protein is shown as a MOLCAD surface whereas the ligand is displayed as ball-and-stick model.*

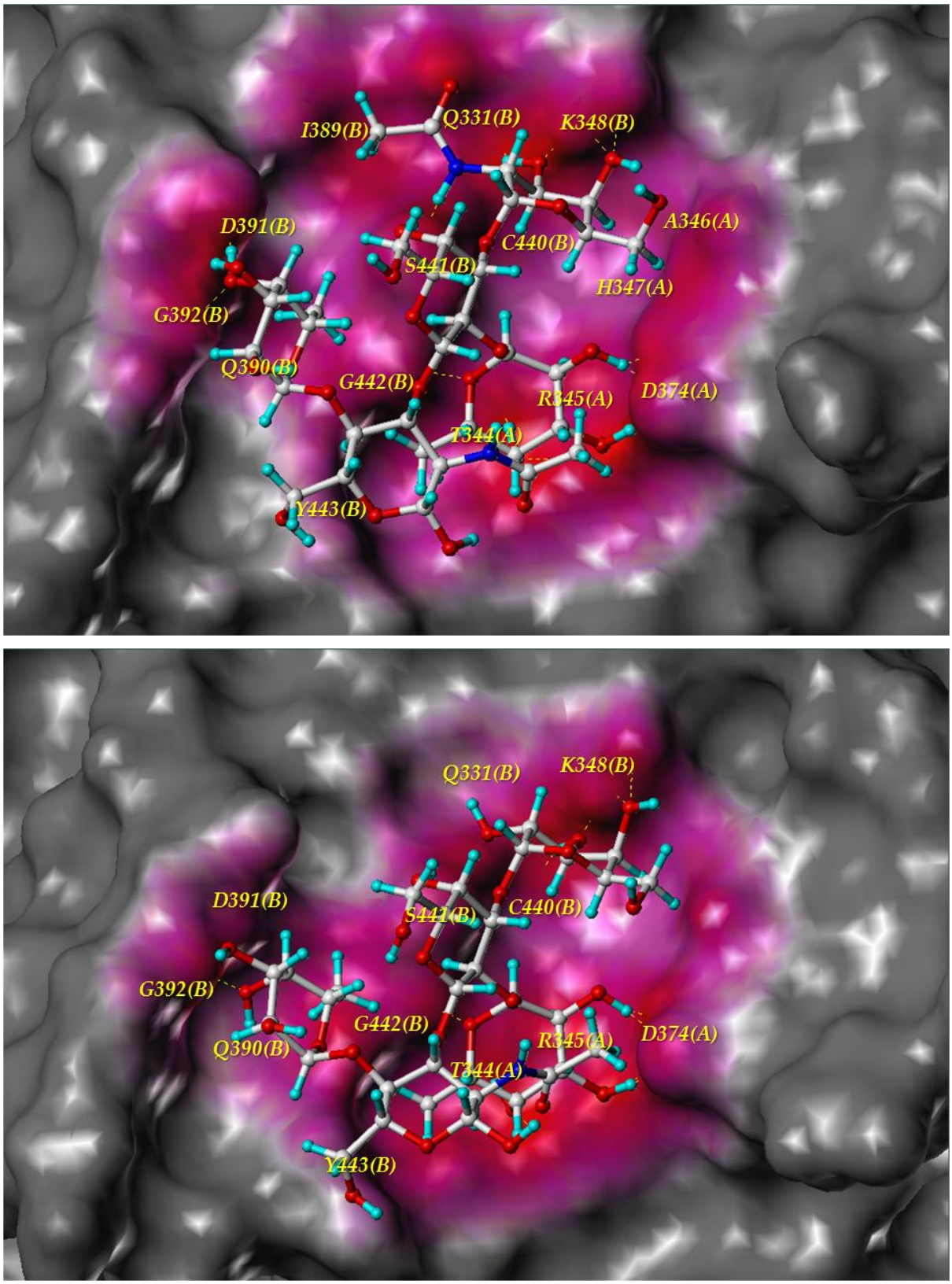


Figure 19 *Enlarged view of ALe^b (top) and BLe^b (bottom) seen in complex with VA387 with a distance threshold of 3 Å. Only the residues that are found to be interacting as per LIGPLOT analysis are labelled. Protein is shown as a MOLCAD surface whereas the ligand is displayed as ball-and-stick model.*

5. Conclusions

According to the present study a variety of structures from H-1/2/3, the A/B tri-saccharide, to the penta-saccharides, ALe^b and BLe^b, could be accommodated in the binding site. The interactions found in the protein-ligand complexes for all the ligands were in good agreement with the crystal structure (Cao *et al*., 2007), the binding studies (Huang *et al*., 2005) and the mutagenesis study (Tan *et al*., 2008). Thus the interesting observation is that the binding site can accommodate a number of different natural saccharides without steric hindrance.

The low Glide XP score of the fucose monosaccharide in the binding site illustrates the key role of fucose in the binding of VA387 to HBGAs. This can also explain the dependence of VA387 strain on the secretor gene that encodes for the fucosyl transferase which adds the Fucα1-2 to the precursor. Similarly, the numerous interactions found with the terminal GalNAcα/Galα along with substantial differences in Glide scores between H- and A/B-antigens suggested that the terminal GalNAcα and Galα in A- and B- structures, respectively, play a role in binding. Furthermore, the Glide XP scores are in fairly good agreement with the available binding data.

The erratic crystal structure of A-trisaccharide complex did not show any protein contacts with the terminal GalNAcα and was not in agreement with the mutagenesis studies. It is of great importance that the corrected A-trisaccharide protein complex is in agreement with the mutational studies.

The B-structures were also docked successfully in agreement to the crystal structure and resulted in slightly higher interaction score as compared to the corresponding A-structures. This suggests that the acetamido group in the terminal GalNAcα of A-antigens does not play a crucial role in binding to A-antigens in terms of binding affinity.

5.1 Future studies

The present work will be extended to include explicit water MD simulations. The crystal structure of VA387/B-trisaccharide complex (Cao *et al.*, 2007) has reported a couple of water mediated hydrogen bonds in the binding site. Explicit water calculations should be focused on the role of water molecules in the binding of HBGAs to the VA387 surface protein.

Further studies are also planned to be focused on type-2 structures to explain their reported binding (Huang *et al.*, 2003; Huang *et al.*, 2005) to noroviruses.

Attempts should be made to design ligands that interfere with the binding of the α-fucose residue.

6. Bibliography

Boga, J. A., S. Melón, I. Nicieza, I. De Diego, M. Villar, F. Parra, and M. De Oña. 2004. *Etiology of sporadic cases of pediatric acute gastroenteritis in Asturias, Spain, and genotyping and characterization of norovirus strains involved*. J Clin Microbiol 42:2668-2674.

Cao, S., Z. Lou, M. Tan, Y. Chen, Y. Liu, Z. Zhang, X. C. Zhang, X. Jiang, X. Li, and Z. Rao. 2007. *Structural basis for the recognition of blood group trisaccharides by norovirus*. J Virol 81:5949-5957.

Case, D., T. Darden, T. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, K. Merz, B. Wang, D. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J. Caldwell, W. Ross, and P. Kollman. 2004. *Amber 8*. University of California, San Francisco.

Chadwick, P. R., G. Beards, D. Brown, E. O. Caul, J. Cheesbrough, I. Clarke, A. Curry, S. O'Brien, K. Quigley, J. Sellwood, and D. Westmoreland. 2000. *Management of hospital outbreaks of gastro-enteritis due to small round structured viruses*. J Hosp Infect 45:1-10.

Dedman, D., H. Laurichesse, E. O. Caul, and P. G. Wall. 1998. *Surveillance of small round structured virus (SRSV) infection in England and Wales, 1990-5*. Epidemiol Infect 121:139-149.

Estes, M. K., B. V. Prasad, and R. L. Atmar. 2006. *Noroviruses everywhere: has something changed?*. Curr Opin Infect Dis 19:467-474.

Fankhauser, R. L., J. S. Noel, S. S. Monroe, T. Ando, and R. I. Glass. 1998. *Molecular epidemiology of "norwalk-like viruses" in outbreaks of gastroenteritis in the United States*. J Infect Dis 178:1571-1578.

Green, K., R. Chanock, and A. Kapikian. 2001. *Human caliciviruses*. Fields Virology 1:841-874.

Hornak, V., R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. 2006. *Comparison of multiple amber force fields and development of improved protein backbone parameters*. Proteins 65:712-725.

Huang, P., T. Farkas, S. Marionneau, W. Zhong, N. Ruvoën-Clouet, A. L. Morrow, M. Altaye, L. K. Pickering, D. S. Newburg, J. LePendu, and X. Jiang. 2003. *Noroviruses bind to human ABO, Lewis, and secretor histo-blood group antigens: identification of 4 distinct strain-specific patterns*. J Infect Dis 188:19-31.

Huang, P., T. Farkas, W. Zhong, M. Tan, S. Thornton, A. L. Morrow, and X. Jiang. 2005. *Norovirus and histo-blood group antigens: demonstration of a wide spectrum of strain specificities and classification of two major binding groups among multiple binding patterns*. J Virol 79:6714-6722.

Hutson, A. M., R. L. Atmar, and M. K. Estes. 2004. *Norovirus disease: changing epidemiology and host susceptibility factors*. Trends Microbiol 12:279-287.

Jiang, X., D. Graham, K. Wang, and M. Estes. 1990. *Norwalk virus genome cloning and characterization*. Science 250:1580-1583.

Kaplan, J. E., R. Feldman, D. S. Campbell, C. Lookabaugh, and G. W. Gary. 1982. *The frequency of a norwalk-like pattern of illness in outbreaks of acute gastroenteritis*. Am J Public Health 72:1329-1332.

Kirschner, K., A. Yongye, S. Tschampel, J. González-Outeiriño, C. Daniels, B. Foley, and R. Woods. 2008. *Glycam06: A generalizable biomolecular force field. Carbohydrates*. J Comput Chem 29:622-655.

Lopman, B. A., M. Reacher, C. Gallimore, G. K. Adak, J. J. Gray, and D. W. G. Brown. 2003. *A summertime peak of "winter vomiting disease": surveillance of noroviruses in England and Wales, 1995 to 2002*. BMC Public Health 3:13.

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

Marx, D., and J. Hutter. 2000. *Ab initio molecular dynamics: theory and implementation*. NIC Series 1:301-449.

Nahmany, A., F. Strino, J. Rosen, G. Kemp, and P. Nyholm. 2005. *The use of a genetic algorithm*

search for molecular mechanics (MM3)-based conformational analysis of oligosaccharides. Carbohydr. Res 340:1059-1064.

Noble, W. S.. 2006. *What is a support vector machine?*. Nature Biotechnology 24:1565 1567.

Onufriev, A., D. Bashford, and D. Case. 2004. *Exploring protein native states and large scale conformational changes with a modified generalized Born model*. Proteins 55:383-394.

Patel, M. M., A. J. Hall, J. Vinjé, and U. D. Parashar. 2009. *Noroviruses: acomprehensive review*. J Clin Virol 44:1-8.

Pearlman, D., D. Case, J. Caldwell, W. Ross, T. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. 1995. *AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules*. Comp. Phys. Commun 91:1-41.

Ponder, J., and D. Case. 2003. *Force fields for protein simulations*. Advances in Protein Chemisty 66:27-85.

Ramirez, S., G. M. Giammanco, S. De Grazia, C. Colomba, V. Martella, and S. Arista. 2008. *Genotyping of GII.4 and GIIb norovirus RT-PCR amplicons by RFLP analysis*. J Virol Methods 147:250-256.

Ravn, V., and E. Dabelsteen. 2000. *Tissue distribution of histo-blood group antigens*. APMIS 108:1-28.

Scipioni, A., A. Mauroy, J. Vinjé, and E. Thiry. 2008. *Animal noroviruses*. Vet J 178:32 45.

Siebenga, J., A. Kroneman, H. Vennema, E. Duizer, M. Koopmans. 2008. *Food-borne viruses in europe network report: the norovirus GII.4 2006b (for US named minerva-like, for Japan kobe034-like, for UK v6) variant now dominant in early seasonal surveillance*. Euro Surveill 13:.

Sigalov, G., A. Fenley, and A. Onufriev. 2006. *Analytical electrostatics for biomolecules: beyond the generalized Born approximation*. J. Chem. Phys. 124:124902.

Sigalov, G., P. Scheffel, and A. Onufriev. 2005. *Incorporating variable environments into the generalized Born model*. J. Chem. Phys. 122:09451.

Sosa, C., T. Hewitt, M. Lee, and D. Case. 2001. *Vectorization of the generalized Born model for molecular dynamics on shared-memory computers*. J. Mol. Struct. (Theochem) 549:193-201.

Srinivasan, J.; Trevathan, M.W.; Beroza, P.; Case, D.A. *Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects*. Theor. Chem. Acc., 1999, 101, 426–434.

Still, W.C.; Tempczyk, A.; Hawley, R.C.; Hendrickson, T. *Semianalytical treatment of solvation for molecular mechanics and dynamics*. J. Am. Chem. Soc., 1990, 112, 6127– 6129.

Stote, R., A. Dejaegere, D. Kuznetsov, L. Falquet. 1999. *Theory of molecular dynamics simulations. Swiss institute of bioinformatics*. MD Tutorial Version 1.0:.

Sutmann, G.. 2002. *Classical molecular dynamics*. NIC Series 10:211-254.

Tan, M., J. Meller, and X. Jiang. 2006. *C-terminal arginine cluster is essential for receptor binding of norovirus capsid protein*. J Virol 80:7322-7331.

Tan, M., M. Xia, S. Cao, P. Huang, T. Farkas, J. Meller, R. S. Hegde, X. Li, Z. Rao, and X. Jiang. 2008. *Elucidation of strain-specific interaction of a GII-4 norovirus with HBGA receptors by site-directed mutagenesis study*. Virology 379:324-334.

Tan, M., M. Xia, Y. Chen, W. Bu, R. S. Hegde, J. Meller, X. Li, and X. Jiang. 2009. *Conservation of carbohydrate binding interfaces: evidence of human HBGA selection in norovirus evolution*. PLoS One 4:e5058.

Tan, M., P. Fang, T. Chachiyo, M. Xia, P. Huang, Z. Fang, W. Jiang, and X. Jiang. 2008a. *Noroviral P particle: Structure, function and applications in virus-host interaction*. Virology 382:115-123.

Tan, M., P. Huang, J. Meller, W. Zhong, T. Farkas, and X. Jiang. 2003. *Mutations within the P2 domain of norovirus capsid affect binding to human histo-blood group antigens: evidence for a binding pocket*. J Virol 77:12562-12571.

Tan, M., R. S. Hegde, and X. Jiang. 2004. *The P domain of norovirus capsid protein forms dimer and binds to histo-blood group antigen receptors*. J Virol 78:6233-6242.

Tan, M., and X. Jiang. 2005. *The P domain of norovirus capsid protein forms a subviral particle that binds to histo-blood group antigen receptors*. J Virol 79:14017-14030.

Tan, M., and X. Jiang. 2005a. *Norovirus and its histo-blood group antigen receptors: an answer to a historical puzzle*. Trends Microbiol 13:285-293.

Tan, M., and X. Jiang. 2007. *Norovirus-host interaction: implications for disease control and prevention*. Expert Rev Mol Med 9:1-22.

Tan, M., and X. Jiang. 2008. *Norovirus gastroenteritis, increased understanding and future antiviral options*. Curr Opin Investig Drugs 9:146-151.

Tsui, V., and D. Case. 2001. *Theory and applications of the generalized Born solvation model in macromolecular simulations*. Biopolymers (Nucl. Acid. Sci.) 56:275-291.

Tu, E. T., R. A. Bull, G. E. Greening, J. Hewitt, M. J. Lyon, J. A. Marshall, C. J. McIver, W. D. Rawlinson, and P. A. White. 2008. *Epidemics of gastroenteritis during 2006 were associated with the spread of norovirus GII.4 variants 2006a and 2006b*. Clin Infect Dis 46:413-420.

Turcios-Ruiz, R. M., P. Axelrod, K. St John, E. Bullitt, J. Donahue, N. Robinson, and H. E. Friss. 2008. *Outbreak of necrotizing enterocolitis caused by norovirus in a neonatal intensive care unit*. J Pediatr 153:339-344.

Varki, A., D. Richard, D. Jeffrey, H. Hudson, S. Pamela, W. Horace, R. Carolyn, W. Gerald, and E. Marilynn. *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2009.

Verhoef, L., E. Depoortere, I. Boxman, E. Duizer, Y. van Duynhoven, J. Harris, C. Johnsen, A. Kroneman, S. Le Guyader, W. Lim, L. Maunula, H. Meldal, R. Ratcliff, G. Reuter, E. Schreier, J. Siebenga, K. Vainio, C. Varela, H. Vennema, M. Koopmans. 2008. *Emergence of new norovirus variants on spring cruise ships and prediction of winter epidemics*. Emerg Infect Dis 14:238-243.

Vinjé, J., and M. P. Koopmans. 1996. *Molecular detection and epidemiology of small round-structured viruses in outbreaks of gastroenteritis in the Netherlands*. J Infect Dis 174:610-615.

Watkins, W.. 1999. *A half century of blood-group antigens research. some personal recollections.* Trends glycosci. Glycotechnol 11:391-411.

Wobus, C. E., L. B. Thackray, and H. W. 4. Virgin. 2006. *Murine norovirus: a model system to study norovirus biology and pathogenesis*. J Virol 80:5104-5112.

Wyatt, R. G., R. Dolin, N. R. Blacklow, H. L. DuPont, R. F. Buscho, T. S. Thornhill, A. Z. Kapikian, and R. M. Chanock. 1974. *Comparison of three agents of acute infectious nonbacterial gastroenteritis by cross-challenge in volunteers*. J Infect Dis 129:709-714.

Zahorsky, J.. 1929. *Hyperemesis hemis or winter vomiting disease*. Arch Pediatr 46:391 395.

Zheng, D., T. Ando, R. L. Fankhauser, R. S. Beard, R. I. Glass, and S. S. Monroe. 2006. *Norovirus classification and proposed strain nomenclature*. Virology 346:312-323.

Appendices

A. Adiabatic Maps

The appendix lists down the adiabatic maps of the glycosidic linkages for the ligands considered in the present work. The total of 28 glycosidic linkages were observed in the 10 ligands considered. The energy plots/adiabatic maps for these 28 linkages are shown below. The plots were produced using GLYGAL.

The H-antigens

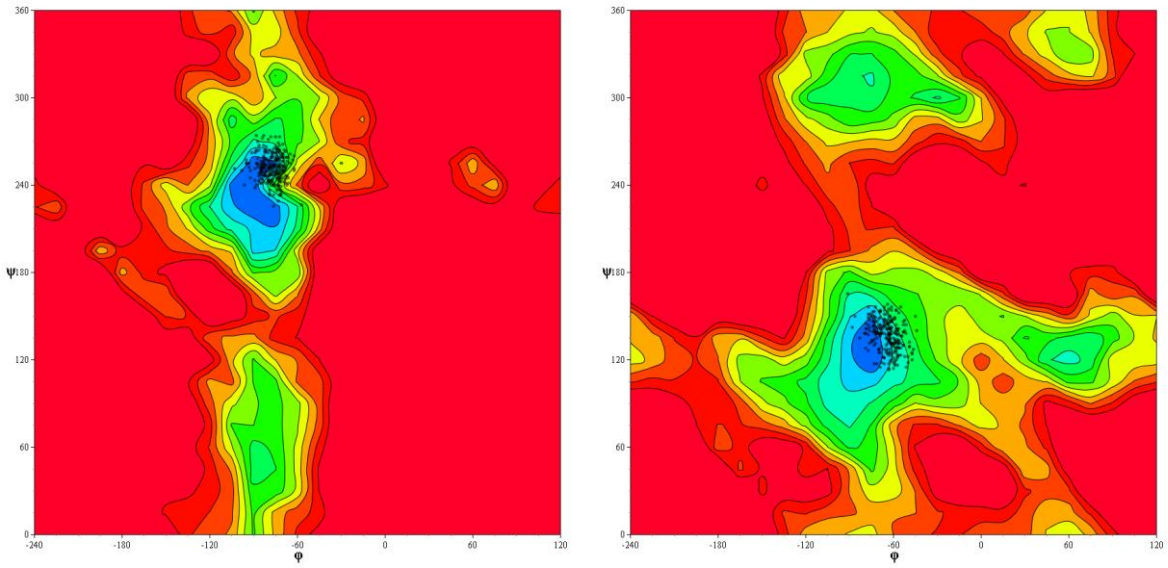


Figure 20 *The adiabatic maps for dihedrals in* ***H-1****. The energy plots for Fucα1-2Galβ (left) and Galβ1-3GlcNAcβ (right) glycosidic linkages are shown.*

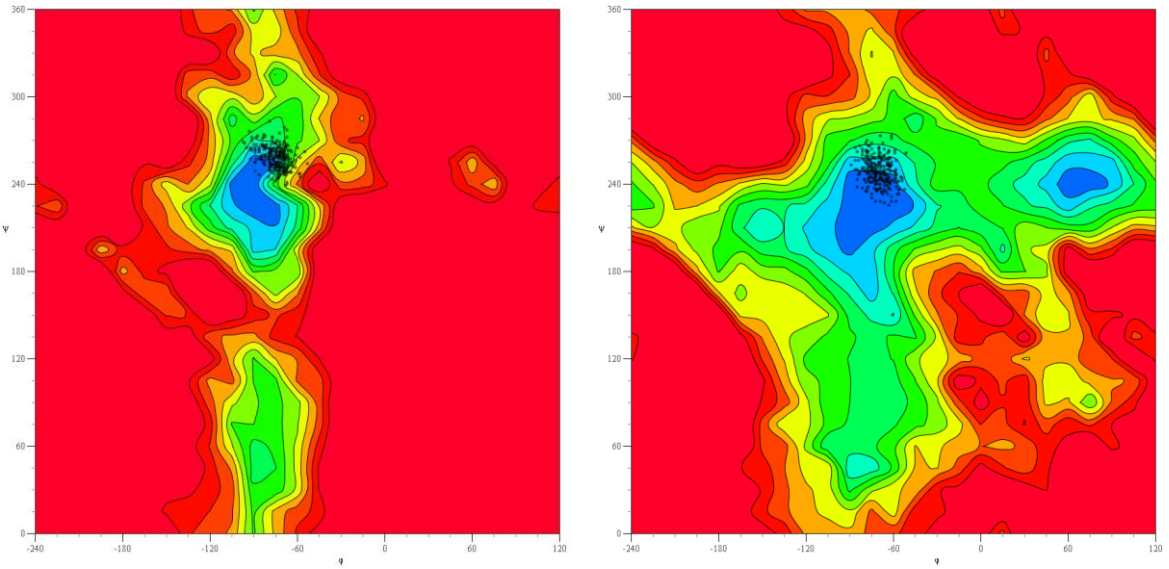


Figure 21 *The adiabatic maps for dihedrals in* ***H-2****. The energy plots for Fucα1-2Galβ (left) and Galβ1-4GlcNAcβ (right) glycosidic linkages are shown.*

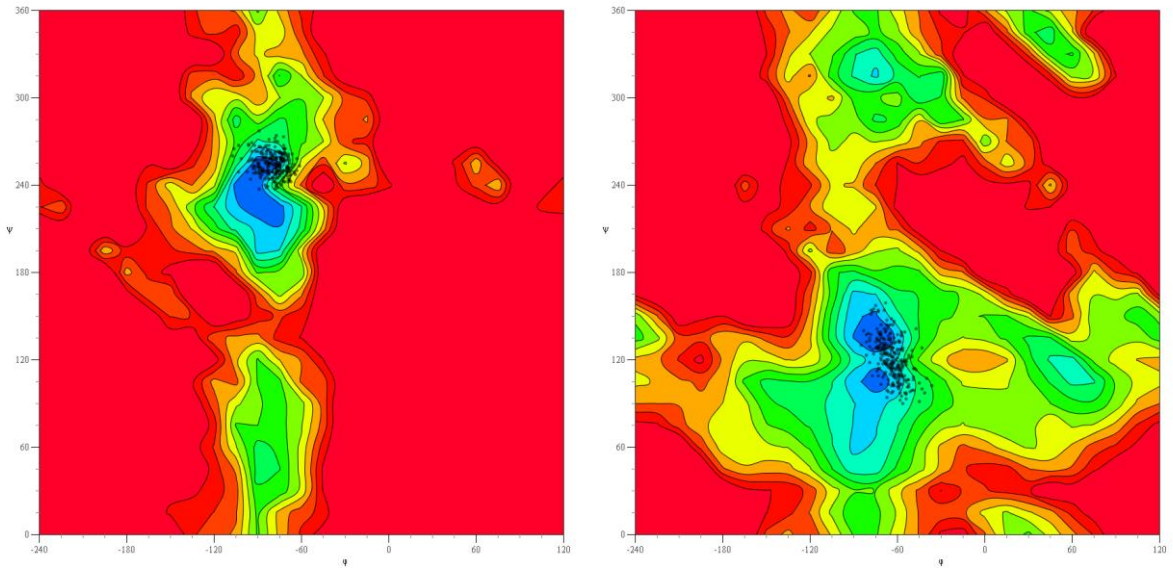


Figure 22 *The adiabatic maps for dihedrals in* ***H-3****. The energy plots for Fucα1-2Galβ (left) and Galβ1-3GalNAcα (right) glycosidic linkages are shown.*

The A-antigens

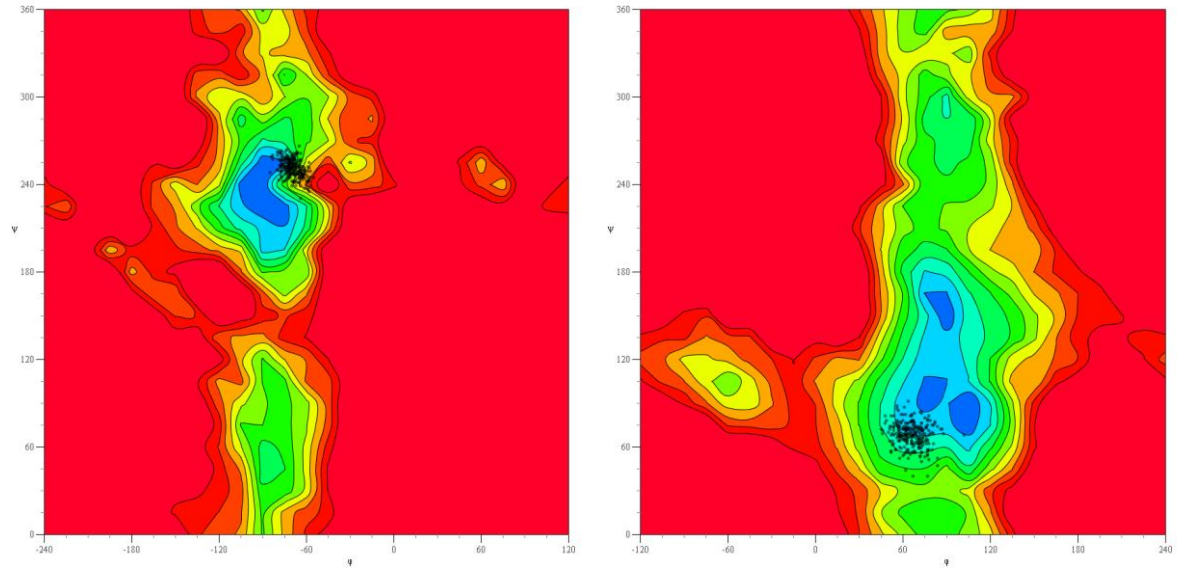


Figure 23 *The adiabatic maps for dihedrals in* ***A-tri****. The energy plots for Fucα1-2Galβ (left) and GalNAcα1-3Galβ (right) glycosidic linkages are shown.*

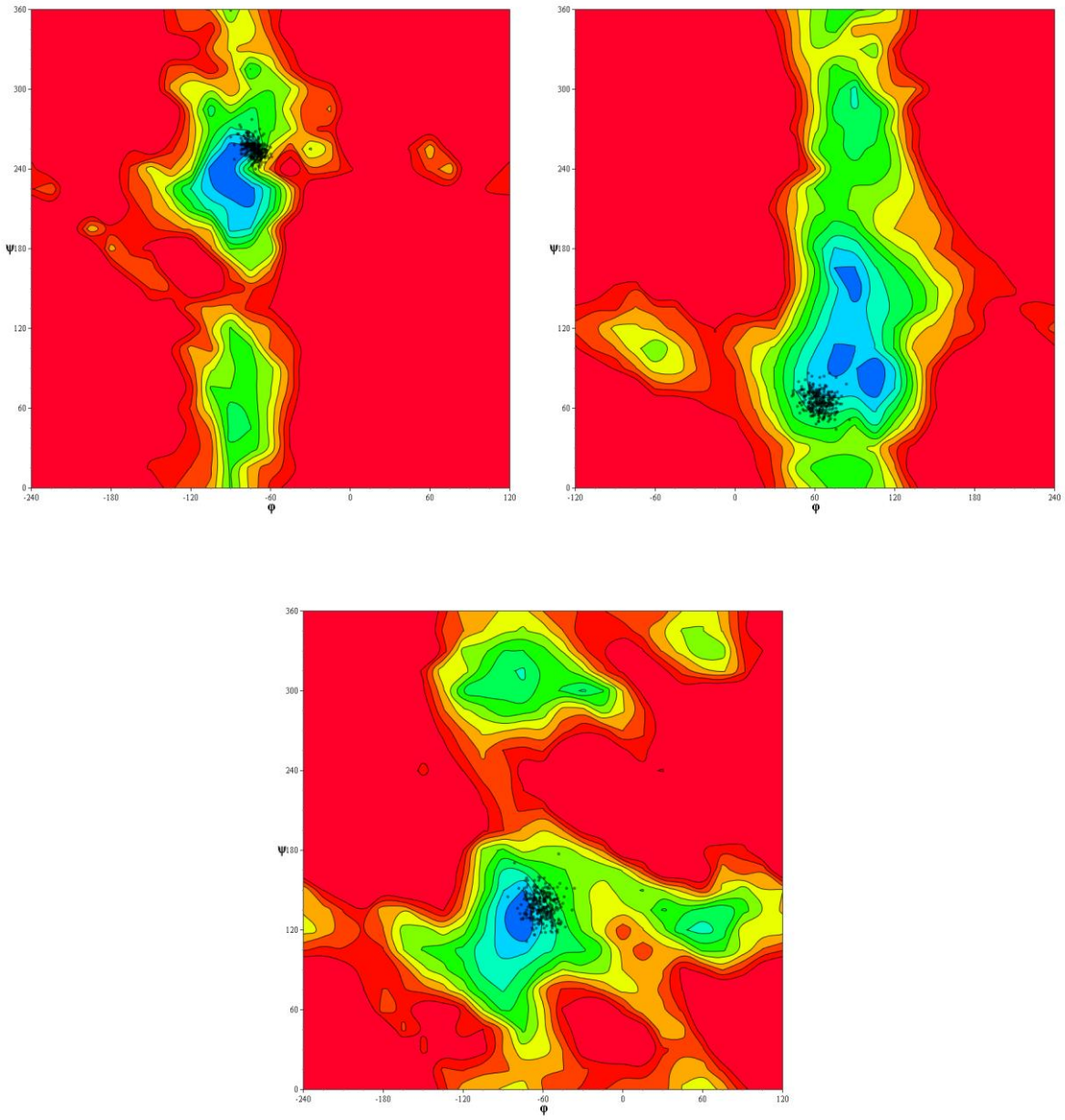


Figure 24 *The adiabatic maps for dihedrals in* ***A-1****. The energy plots for Fucα1-2Galβ (top left), GalNAcα1-3Galβ (top right) and Galβ1-3GlcNAcβ (bottom) glycosidic linkages are shown.*

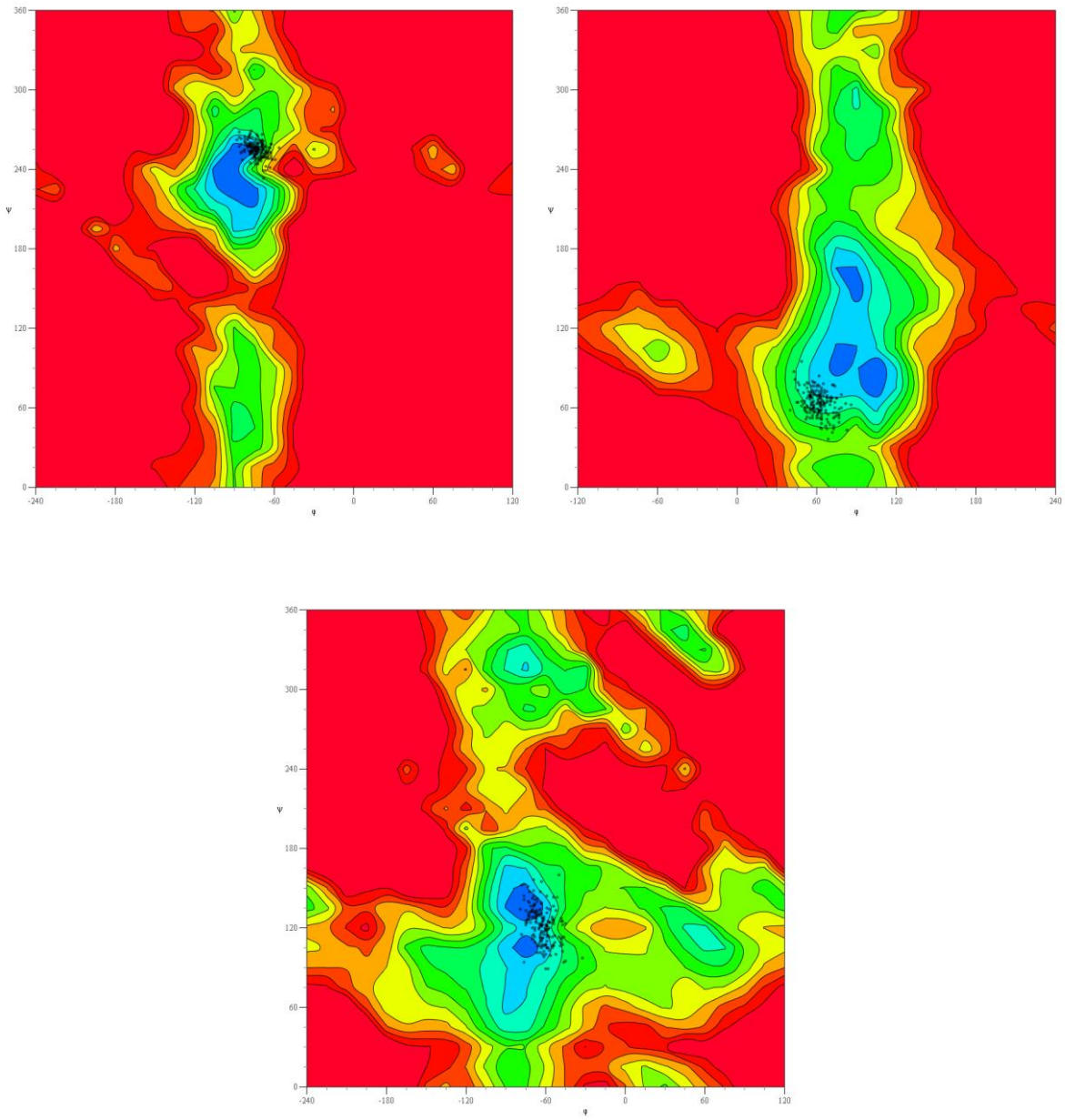


Figure 25 *The adiabatic maps for dihedrals in* ***A-3****. The energy plots for Fucα1-2Galβ (top left), GalNAcα1-3Galβ (top right) and Galβ1-3GalNAcα (bottom) glycosidic linkages are shown.*

The B-antigen

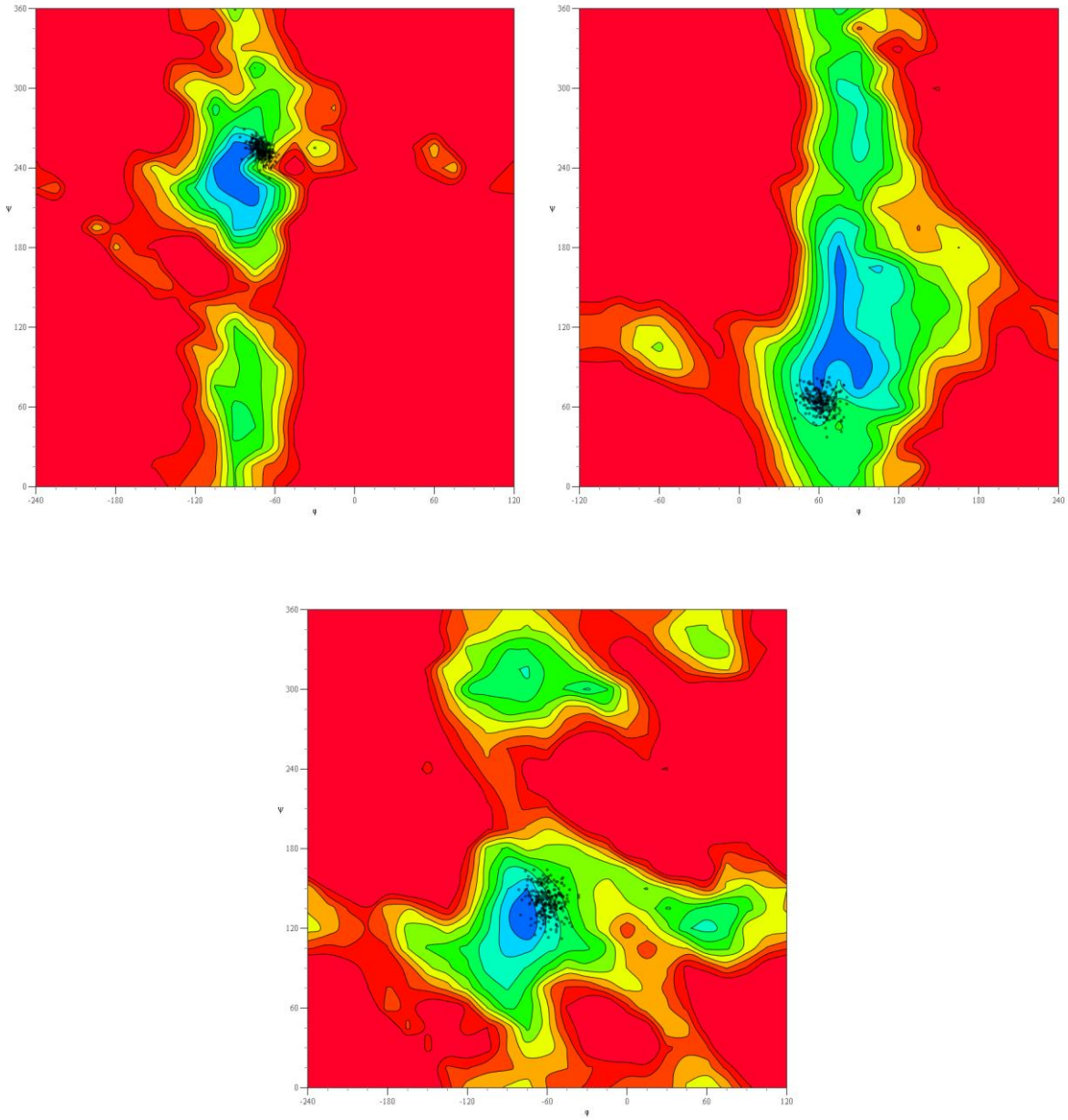


Figure 26 *The adiabatic maps for dihedrals in* ***B-1****. The energy plots for Fucα1-2Galβ (top left), Galα1-3Galβ (top right) and Galβ1-3GlcNAcβ (bottom) glycosidic linkages are shown.*

The secretor gene dependent Lewis structures

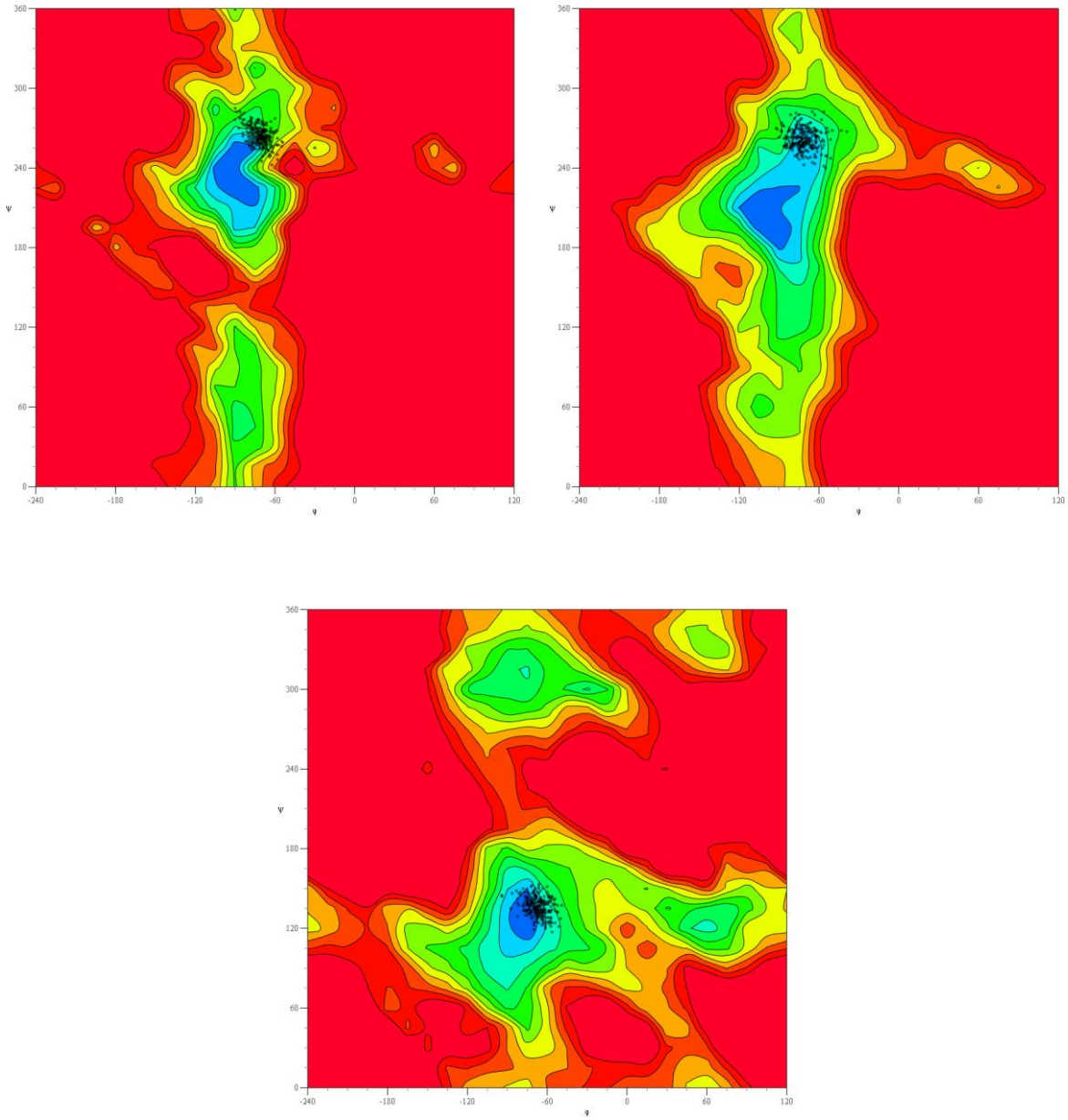


Figure 27 *The adiabatic maps for dihedrals in* ***Le^b****. The energy plots for Fucα1-2Galβ (top left), Fucα1-4GlcNAcβ (top right) and Galβ1-3GlcNAcβ (bottom) glycosidic linkages are shown.*

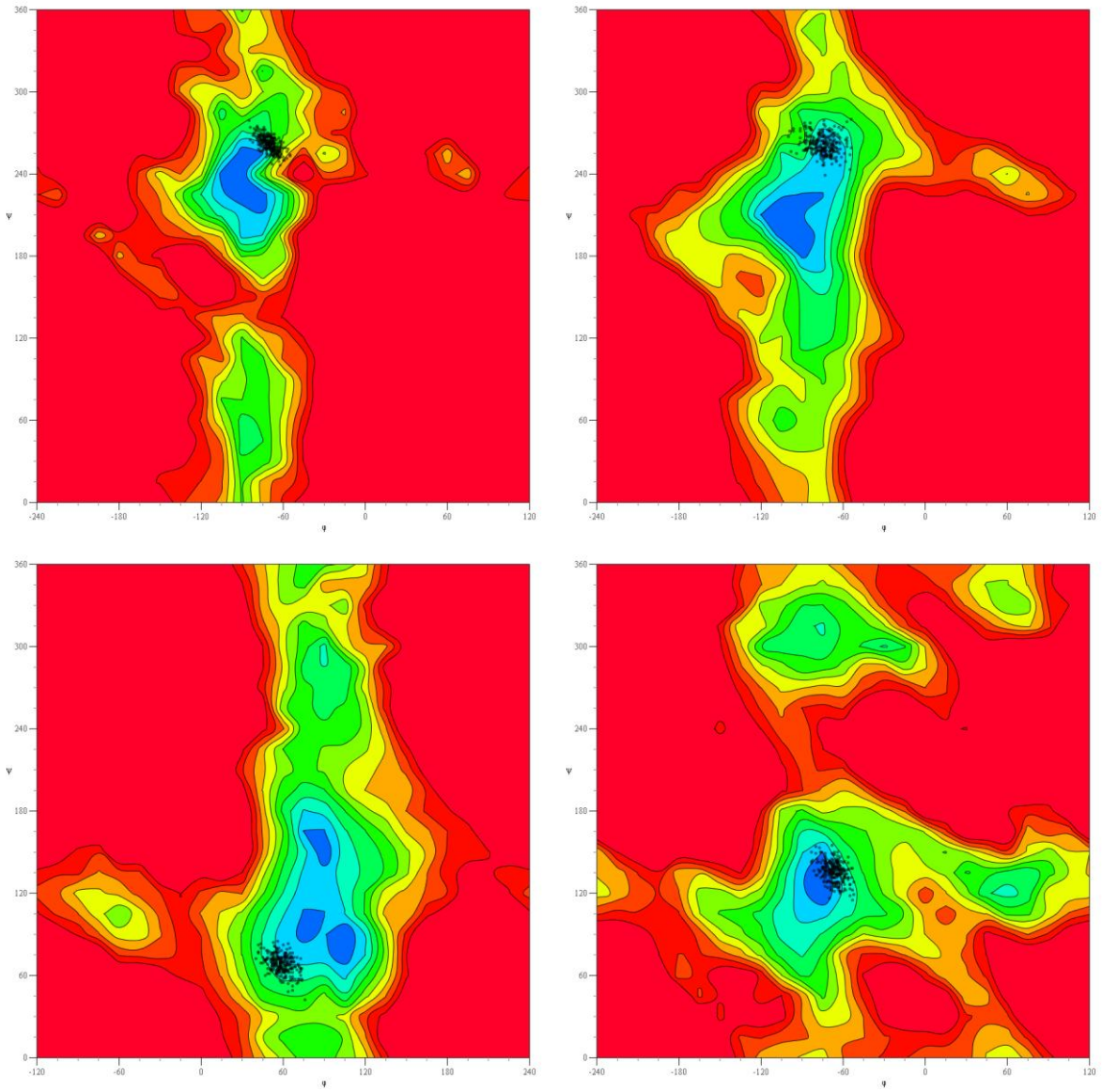


Figure 28 *The adiabatic maps for dihedrals in* ***ALe^b****. The energy plots for Fucα1-2Galβ (top left), Fucα1-4GlcNAcβ (top right), GalNAcα1-3Galβ (bottom left) and Galβ1-3GlcNAcβ (bottom right) glycosidic linkages are shown.*

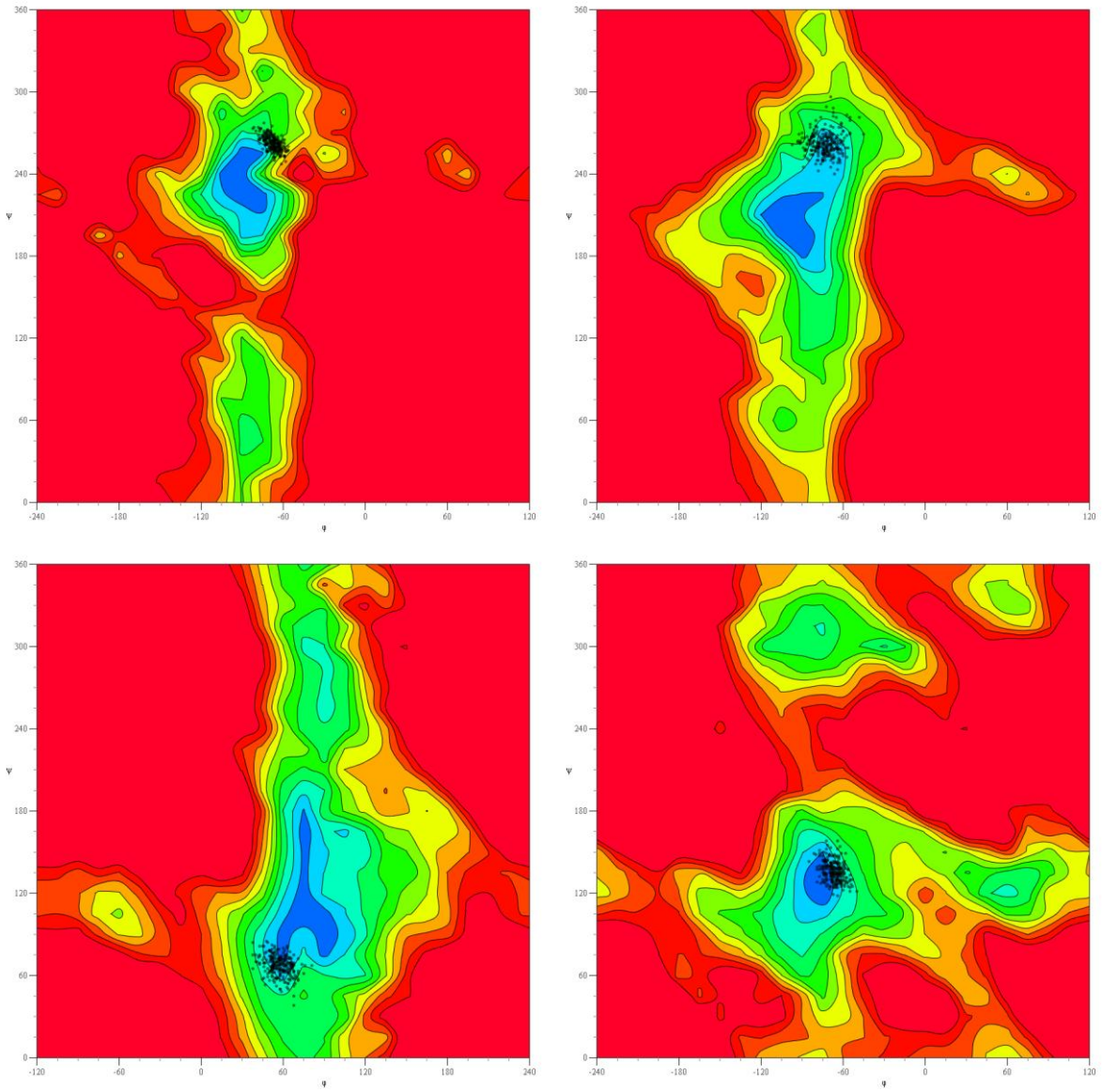


Figure 29 *The adiabatic maps for dihedrals in* ***BLe^b****. The energy plots for Fucα1-2Galβ (top left), Fucα1-4GlcNAcβ (top right), Galα1-3Galβ (bottom left) and Galβ1-3GlcNAcβ (bottom right) glycosidic linkages are shown.*

B. Table of predicted VA387/carbohydrate interactions

	Saccharide	Q_{390} (B)[1]	D_{391} (B)[1]	G_{392} (B)[1]	A_{346} (A)[2]	H_{347} (A)[2]	Q_{331} (B)[2]	K_{348} (B)[2]	I_{389} (B)[2]	C_{440} (B)[2]	N_{393} (B)	H_{395} (B)	S_{441} (B)	Y_{443} (B)
A-tri	Galβ												3	
	GalNAcα				4	6	6	6	6	6			5	
H-1	Galβ												6	
	GlcNAcβ			5								1		2
H-2	Galβ												5	
	GlcNAcβ			4							5			
H-3	Galβ												6	
	GalNAcα			3										3
Le^b	Galβ												6	
	Fucα	3	6	6										
	GlcNAcβ													
A-1	GalNAcα				4	6	5	6	6	5			5	
	Galβ												4	
	GlcNAcβ		1	5										1
A-3	GalNAcα1,3				3	3	5	6	6	3			6	
	Galβ												2	
	3GalNAcα			4										1
ALe^b	GalNAcα				1	6	6	6	5	6			6	
	Galβ													
	Fucα	6	6	6										
	GlcNAcβ													
B-1	Galα				3	4	5	6		4			6	
	Galβ												6	
	GlcNAcβ			5										3
BLe^b	Galα						1	6		2			6	
	Galβ												3	
	Fucα	4	6	6										
	GlcNAcβ													3

Table 6 The interactions of protein residues with saccharide residues of the ligands are shown for the input files used for Glide scoring of docking poses. The numbers represent the number of snapshots (out of 6) in which the interaction was observed.

[1] Residues interacting mainly with sugar residues at the reducing terminus.

[2] Residues interacting mainly with the non-reducing terminal sugar residues.