Understanding Carbohydrate Recognition by Antiviral Lectins:

Applying Computational Methods to Protein−Carbohydrate Complexes

A Dissertation Presented

by

Yukiji Karen Fujimoto

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

December 2012

Stony Brook University

The Graduate School

Yukiji Karen Fujimoto

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

David F. Green – Dissertation Advisor
Associate Professor – Applied Mathematics & Statistics

Carlos Simmerling – Chairperson of Defense
Professor – Chemistry

Isaac Carrico – Third Member
Assistant Professor – Chemistry

Robert Haltiwanger – Outside Member
Professor – Biochemistry & Cell Biology

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Understanding Carbohydrate Recognition by Antiviral Lectins: Applying Computational Methods to Protein−Carbohydrate Complexes

by

Yukiji Karen Fujimoto

Doctor of Philosophy

in

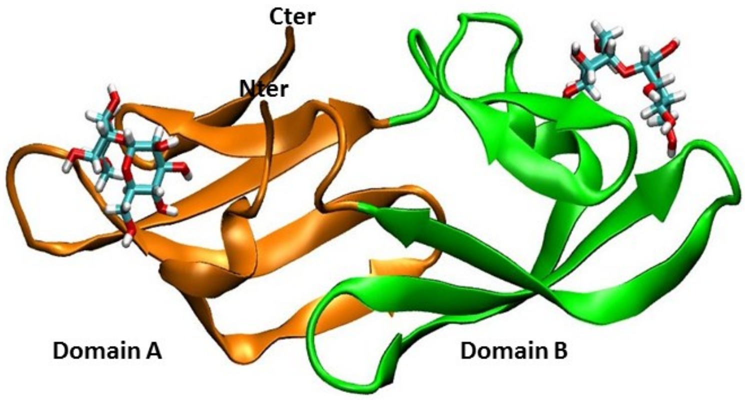
Chemistry

Stony Brook University

2012

Human immunodeficiency virus (HIV) infection of T-cells begins when the viral envelope glycoprotein, gp120, binds to CD4 receptors on the target cell surface. Over the past several years, proteins isolated from various prokaryotes have been shown to inhibit HIV cell entry by binding to gp120 and thus blocking the association with CD4. Lectins that bind to high-mannose oligosaccharides on gp120 are an attractive class of antiviral agents. While several of these have been quite well characterized both structurally and biochemically, there remain many open questions regarding their mechanism of inhibition. Among the best studied is cyanovirin-N (CVN), which is currently under clinical study for use as a topical prophylactic. Large-scale molecular dynamics simulations have identified important structural features of this system that are difficult to resolve experimentally, and binding free energies of a diverse set of oligosaccharide targets computed from these structural ensembles give remarkable agreement with experiment. Detailed decompositions of the binding free energies on a residue-by-residue basis have additionally identified several key interactions that define broad affinity for α-(1,2)-dimannose-containing sugars, as well as a number of determinants of specificity. These studies provide a deeper understanding of the mechanism of inhibitory activity. In addition, this work has provided a foundation for methodological improvements that allow us to more accurately capture the energetics of carbohydrate binding.

Frontispiece



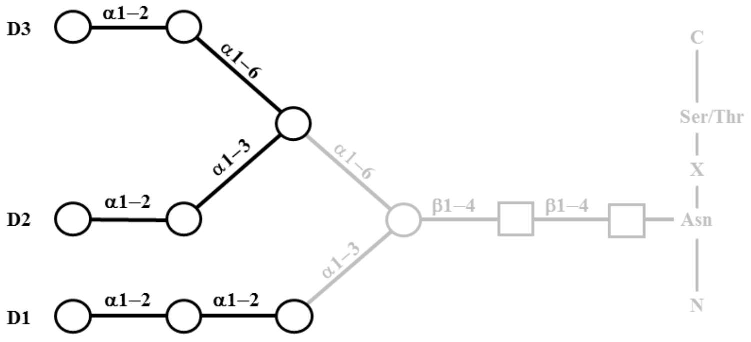


Table of Contents

Chapter 4

List of Figures

Chapter 1.

Chapter 2.

Chapter 3.

Chapter 4.

Chapter 5.

List of Tables

Chapter 1.

Chapter 2.

Chapter 3.

Chapter 4.

Chapter 5.

List of Abbreviations

AIDS	Acquired immunodeficiency syndrome
Asn	Asparagine
CBA	Carbohydrate−binding agent
ConA	Concanavalin A (lectin from the jack bean, *Canavalia ensiformis*)
cvdW	Continuum van der Waals
CVN	Cyanovirin-N (lectin from the cyanobacterium, *Nostoc ellipsosporum*)
CVNH	CVN homolog
Gal	Galactose
GB	Generalized Born
GalNAc	N-acetylgalactosamine
GlcNAc	N-acetylglucosamine
GRFT	Griffithsin (lectin from the red algae, *Griffithsia* spp.)
GSL	Gerardia lectin (lectin from the coral, *Gerardia savaglia*)
EM	Electron microscopy
ER	Endoplasmic reticulum
Fuc	Fucose
Glc	Glucose
gp41	Glycoprotein 41
gp120	Glycoprotein 120
HHA	Amaryllis lectin (lectin from the amaryllis bulb, *Hippeastrum hybrid*)
HIV	Human immunodeficiency virus
Man	Mannose
MVN	Microvirin
MD	Molecular dynamics
MM-PBSA	Molecular mechanics Poisson−Boltzmann surface area
MVL	Microcystis viridis lectin (lectin from the cyanobacterium, *Microcystis viridis*)
NMR	Nuclear magnetic resonance
NPA	Daffodil lectin (lectin from the daffodil, *Narcissus pseudonarcissus*)
PB	Poisson−Boltzmann

PDB	Protein Data Bank
SA	Sialic acid
Ser	Serine
SVN	Scytovirin (lectin from the cyanobacterium, *Scytonema varium*)
Thr	Threonine
vdW	van der Waals
Xyl	Xylose

Acknowledgments

Although only my name appears on the cover of this dissertation, I would never have been able to finish my thesis without the guidance and support of many people. My deepest gratitude goes to my advisor, Dr. David Green. I have been amazingly fortunate to have an advisor who provided me with an excellent atmosphere for doing research, and gave me guidance to recover when my steps faltered. I sincerely thank David for his understanding, patience, and most importantly, his friendship during my graduate studies at Stony Brook. I hope that one day I would become as good of a mentor to others as David has been to me.

I am also indebted to the members of my dissertation committee—Dr. Carlos Simmerling and Dr. Isaac Carrico. Their insightful comments and constructive criticisms at different stages of my research were thought-provoking. Special thanks go to Dr. Robert Haltiwanger, who was willing to participate in my final defense committee.

I would like to mention my deepest thanks to my other research mentors from years past—Dr. Mali Yin and Dr. Adele L. Boskey. Professor Yin served as a good mentor and gave me the encouragement to continue my interest in science. I want to express my gratitude to Dr. Boskey for giving me the opportunity to join her lab before I entered graduate school. She gave me the first of many advices and recommendations for starting my graduate study.

I am grateful to Dr. Bruce Tidor from MIT for making available the Integrated Continuum Electrostatics (ICE) software package (including the multigrid PBE solver and the continuum van der Waals software). I would also like to thank the staff at the Department of Chemistry and the Department of Applied Mathematics and Statistics for their various forms of support during my graduate study.

Many friends have helped me stay sane through these years. I am thankful to all the current and past members of the Green research group. Their support and care helped me overcome setbacks and stay focused on my graduate study. I greatly value their friendship and I deeply appreciate their belief in me. In particular, I would like to thank Vadim Patsalo. It was a pleasure to share the first office the Green lab had. We shared not only nice discussions during work, but also nice times during conferences. Ryan Terbush also deserves a special acknowledgment that began the work contained in this thesis.

Most importantly, none of this would have been possible without the endless support, love, and patience of my family and my husband, Chris Marraccino. This dissertation is dedicated to my grandfather, Koo-Hyon Lee. Although he is no longer here, he was undeniably the bedrock of my scientific endeavor since I was in elementary school. I know he would still cheer me on and stand by me through the next stages of my scientific career.

Next, I would like to thank the computational support of the NY Center for Computational Sciences at SBU/BNL (New York Blue) and the Stony Brook University Seawulf Cluster. Finally, I appreciate the financial support from the College of Engineering & Applied Sciences, and the Department of Applied Mathematics & Statistics (to DFG), and the National Institute of General Medical Science NIH/NIGMS R01 GM086199 (to DFG).

Curriculum Vitae

Education

Doctor of Philosophy in Chemistry

Stony Brook University *Aug 2006 – Dec 2012*
Stony Brook, NY 11794-3600
Dissertation Advisor: Prof. David F. Green, Dept. of Applied Mathematics & Statistics
Dissertation Title: Understanding Carbohydrate Recognition by Antiviral Lectins: Applying Computational Methods to Protein−Carbohydrate Complexes

Master of Science in Chemistry

Stony Brook University *Aug 2006 – Aug 2009*
Stony Brook, NY 11794-3600
Thesis Advisor: Prof. David F. Green, Dept. of Applied Mathematics & Statistics
Thesis Title: Using Computational Models to Understand Specific Carbohydrate Recognition by Cyanovirin-N

Bachelor of Arts in Liberal Arts

Sarah Lawrence College *Aug 1999 – May 2003*
Bronxville, NY 10708
Areas of Concentration: Chemistry, Art History

Research Positions

Stony Brook University Stony Brook, NY 11794
Laboratory of Prof. David F. Green *Apr 2007 – present*
Graduate Research Assistant

- Applying computational methods to study affinity and specificity in protein–carbohydrate interactions; specifically, studying the initial steps in the recognition of target cells by the HIV-1 virus, and understanding the role of glycosylation in the interactions of proteins
- Using large-scale molecular dynamics simulations, coupled with careful energetic analysis; and residue-by-residue decompositions of the binding free energies

Hospital for Special Surgery New York, NY 10021
Laboratory of Dr. Adele L. Boskey *Oct 2002 – Aug 2006*
Technician

- Studying the mechanisms of biologic calcification in health and disease – the role of noncollagenous extracellular matrix proteins in the regulation of bone and tooth mineralization
- Techniques for analysis include: the gel diffusion system for the study of *in vitro* mineralization, FT-IR microspectroscopy and imaging for the study of mineral and matrix properties, and wide angle x-ray diffraction

Teaching Position

Stony Brook University Stony Brook, NY 11794

Teaching Assistant: General Chemistry I (CHE 131) *Aug 2006 – Dec 2006*

Teaching Assistant: General Chemistry II (CHE 132) *Jan 2007 – May 2007*

- Held three recitation classes (three times a week; 80 minutes per class; ~20−30 students per class) to work on practice problems and have students ask questions about concepts and problems that they did not understand during lecture
- Held office hours to assist students in their understanding of the coarse material
- Proctored major exams

Awards

Chemistry Award for Outstanding Service, Stony Brook University	**2012**
Sigma Xi Travel Award, Stony Brook University	**2009**

Publications

1. **Fujimoto, Y.K.**, Green, D.F. Understanding Specific Carbohydrate Recognition by the Antiviral Lectin Cyanovirin-N. *Manuscript submitted.*

2. Gericke, A., Qin, C., Sun, Y., Redfern, R., Redfern, D., **Fujimoto, Y.**, Taleb, H., Butler, W.T., Boskey, A.L. Different forms of DMP1 play distinct roles in mineralization. *J Dent Res.* **2010,** *89*(4): 355-359.

3. Boskey, A., Frank, A., **Fujimoto, Y.**, Spevak, L., Verdelis, K., Ellis, B., Troiano, N., Philbrick, W., Carpenter, T. The PHEX transgene corrects mineralization defects in 9-month-old hypophosphatemic mice. *Calcif. Tissue Int.* **2009,** *84*(2): 126-137.

Presentations

Oral Presentations

1. **Fujimoto, Y.K.**; Green, D.F. "Understanding carbohydrate recognition by the antiviral lectin Cyanovirin-N." *Protein Engineering Mega Group meeting*, New York, NY. May 2010.

2. **Fujimoto, Y.K.**; Green, D.F. "Computational Models Explain the Oligosaccharide Specificity of Cyanovirin-N." *Applied Mathematics and Statistics Graduate Student Conference*, Stony Brook, NY. February 2008.

3. **Fujimoto, Y.**; Boskey, A.L. "Effects of Milk Osteopontin on the Amorphous Calcium Phosphate to Hydroxyapatite Transformation." *Annual Meeting and Exhibition of the American Association of Dental Research*, San Antonio, TX. March 2003.

Poster Presentations

1. **Fujimoto, Y.K.**; Green, D.F. "Improving the accuracy of computational models of protein-carbohydrate complexes." *Chemistry Research Day at Stony Brook University.* Stony Brook, NY. November 2011.

2. **Fujimoto, Y.K.**; Green, D.F. "Improving the accuracy of computational models of protein-carbohydrate complexes." *Institute of Chemical Biology & Drug Discovery Annual Symposium.* Stony Brook, NY. October 2011.

3. **Fujimoto, Y.K.**; Green, D.F. "Understanding carbohydrate recognition by the antiviral lectin cyanovirin-N." *New York Structural Biology Discussion Group – Winter Meeting,* New York, NY. January 2011.

4. **Fujimoto, Y.K.**; Green, D.F. "Understanding carbohydrate recognition by the antiviral lectin cyanovirin-N." *New York Theoretical & Computational Chemistry Conference,* New York, NY. January 2011.

5. **Fujimoto, Y.K.**; Green, D.F. "Understanding carbohydrate recognition by the antiviral lectin cyanovirin-N." *Chemistry Research Day at Stony Brook University,* Stony Brook, NY. November 2010.

6. **Fujimoto, Y.K.**; Green, D.F. "Understanding carbohydrate recognition by the antiviral lectin cyanovirin-N." *Institute of Chemical Biology & Drug Discovery Annual Symposium.* Stony Brook, NY. October 2010.

7. **Fujimoto, Y.K.**; Green, D.F. "Understanding carbohydrate recognition by the antiviral lectin cyanovirin-N." *New York Structural Biology Discussion Group – Summer Meeting,* Cold Spring Harbor, NY. August 2010.

8. **Fujimoto, Y.K.**; Green, D.F. "Understanding carbohydrate recognition by the antiviral lectin cyanovirin-N." *The 24th Annual Symposium of The Protein Society,* San Diego, CA. August 2010.

9. **Fujimoto, Y.K.**; Green, D.F. "Understanding carbohydrate recognition by the antiviral lectin cyanovirin-N." *Groups Studying the Structures of AIDS-Related Systems & Their Application to Targeted Drug Design.* Bethesda, MD. June 2010.

10. **Fujimoto, Y.K.**; Green, D.F. "Understanding specific carbohydrate recognition by the antiviral lectin cyanovirin-N." *Chemistry Research Day at Stony Brook University,* Stony Brook, NY. November 2009.

11. **Fujimoto, Y.K.**; Green, D.F. "Understanding specific carbohydrate recognition by the antiviral lectin cyanovirin-N." *Louis and Beatrice Laufer Center for Computational Biology and Genome Sciences Symposium,* Stony Brook, NY. October 2009.

12. **Fujimoto, Y.K.**; Green, D.F. "Understanding specific carbohydrate recognition by the antiviral lectin cyanovirin-N." *Institute of Chemical Biology and Drug Discovery Annual Symposium,* Stony Brook, NY. October 2009.

13. **Fujimoto, Y.K.**; Green, D.F. "Understanding specific carbohydrate recognition by the antiviral lectin cyanovirin-N." *New York Structural Biology Discussion Group,* Cold Spring Harbor, NY. August 2009.

14. **Fujimoto, Y.K.**; Green, D.F. "Understanding specific carbohydrate recognition by the antiviral lectin cyanovirin-N." *The 23rd Annual Symposium of The Protein Society,* Boston, MA. July 2009.

15. **Fujimoto, Y.K.**; Green, D.F. "Understanding and enhancing carbohydrate binding by the anti-viral lectin cyanovirin-N." *Chemistry Research Day at Stony Brook University,* Stony Brook, NY. November 2008.

16. **Fujimoto, Y.K.**; Green, D.F. "Understanding and enhancing carbohydrate binding by the anti-viral lectin cyanovirin-N." *Institute of Chemical Biology & Drug Discovery Annual Symposium,* Stony Brook, NY. October 2008.

17. **Fujimoto, Y.K.**; Green, D.F. "Understanding and enhancing carbohydrate binding by the anti-viral lectin cyanovirin-N." *The 22nd Annual Symposium of The Protein Society,* San Diego, CA. July 2008.

18. **Fujimoto, Y.K.**; Green, D.F. "Understanding and enhancing carbohydrate binding by the anti-viral lectin cyanovirin-N." *Groups Studying the Structures of AIDS-Related Systems & Their Application to Targeted Drug Design,* Bethesda, MD. June 2008.

19. **Fujimoto, Y.K.**; TerBush, R.; Green, D.F. "Carbohydrate recognition by anti-viral cyanobacterial proteins: computational approaches to understanding and design." *Chemistry Research Day at Stony Brook University,* Stony Brook, NY. November 2007.

20. **Fujimoto, Y.K.**; TerBush, R.; Green, D.F. "Carbohydrate recognition by anti-viral cyanobacterial proteins: computational approaches to understanding and design." *Institute of Chemical Biology & Drug Discovery Annual Symposium,* Stony Brook, NY. October 2007.

21. Green, D.F.; **Fujimoto, Y.K.**; TerBush, R. "Carbohydrate recognition by anti-viral cyanobacterial proteins: computational approaches to understanding and design." *The 21st Annual Symposium of The Protein Society,* Boston, MA. July 2007.

Professional Memberships

Protein Society – Graduate student member	**2007 – present**
American Chemical Society – Graduate student member	**2007 – present**
New York Academy of Sciences – Graduate student member	**2008 – present**

Chapter 1

Introduction to the Study of Protein−Carbohydrate Interactions

Carbohydrates are the most abundant class of biomolecules on Earth. They are produced during the process of photosynthesis, in which the energy from the sun is converted into chemical energy by combining carbon dioxide with water to form carbohydrates and molecular oxygen. During the last two decades of the nineteenth century, a great interest emerged in studying carbohydrates. One of the leading pioneers in the carbohydrate chemistry field, Emil Hermann Fischer, laid the foundation of carbohydrate terminology, which is still in use today. He was also recognized for his investigations of sugar and purine groups. Fischer introduced the "lock and key" hypothesis to interpret the action of an enzyme ("lock") on a substrate ("key") [1]. Over the years, the lock-and-key hypothesis had become one of the important features in cellular biology. During the last several decades, attention has been focused on recognition mediated by carbohydrates and lectins. There are many various types of lectins that differ in size and structure [2]. Fischer's monumental achievements inspired a whole generation of scientists studying carbohydrates.

In recent years, interest in chemical glycobiology has grown significantly because it is now well established that carbohydrates and carbohydrate conjugates are involved in cellular processes, including cell-cell interactions [3], viral host interactions [4], microbial pathogenesis [5, 6], and immune response [7] (Figure 1-1 [8]). With this growing interest for glycobiology also came an increased demand for tools to study carbohydrates. The biochemical properties of oligosaccharides make them a challenging class of molecules for conformational analysis.

Unlike proteins and nucleic acids, it is not possible to readily introduce point mutations in a sugar sequence. The analysis of carbohydrates differs in several ways from other molecules. In protein studies, much can be learned from modeling based on known X-ray or nuclear magnetic resonance (NMR) structures. In carbohydrate studies, there is a lack of data for carbohydrates from the inherent flexibility of many glycans—the more flexible a molecule is, the harder it is to induce crystallization. Although relatively rigid oligosaccharides exist (*e.g.* blood group determinants), it cannot be said that the properties of these compact branched structures are representative of the majority of oligosaccharides. Flexibility is not the only factor responsible for the resistance to crystal formation. Other contributing factors include how water molecules coordinate to glycans and the CHO−CHO interactions on "hydrophobic" side of ring.

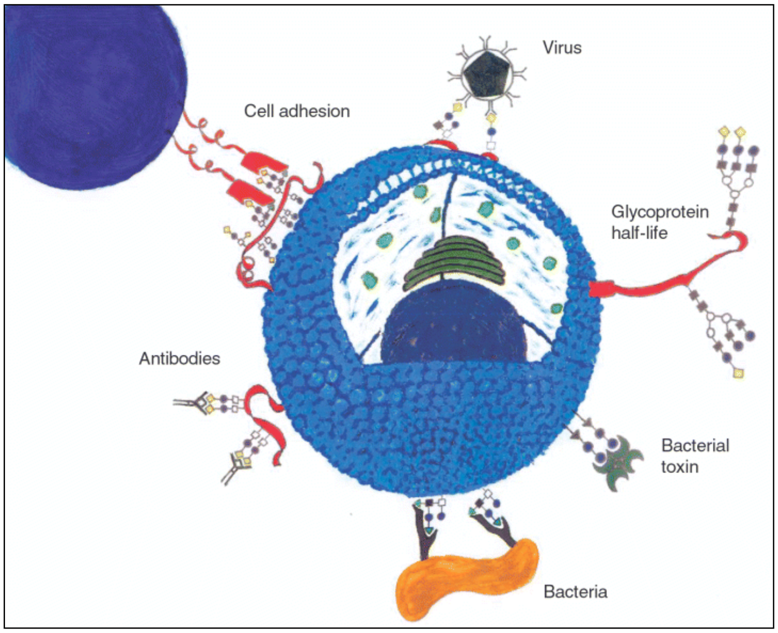


Figure 1- 1. **Schematic diagram portraying protein-carbohydrate interactions at the cell surface.** Red ribbons indicate that the sugar chains can be linked to proteins or anchored in the plasma membrane via a lipid *(Reprinted by permission from Nature Publishing Group: Immunology Cell Biology [8], copyright 2005)*.

In an effort to contribute to the glycobiology community, this thesis will explore a comprehensive set of computational methods for analysis in structural glycobiology, and the application of these tools to the glycobiology of HIV-1 infection. The long term aim is to understand the role of carbohydrates on the Env glycoproteins both in recognition of cellular targets and in recognition by the immune system. The following sections will address (1) a brief overview of glycosylation in biological systems, (2) introduce glycobiology in the context of HIV-1 virus, and (3) list the objectives of this dissertation.

1.1 Glycosylation in Biological Systems

Glycosylation is an important form of co- and post-translational modification. It is a process by which carbohydrates are chemically attached to proteins to form glycoproteins. It is a feature that enhances the functional diversity of proteins and influences their biological activity. Glycans undergo structural and functional roles in the membrane and secreted proteins [9]. The majority of proteins that are synthesized in the rough endoplasmic reticulum undergo glycosylation. Glycosylation is also present in the cytoplasm and nucleus as the O-GlcNAc modification. Glycosylation can be broadly divided into two categories—O-linked and N-linked glycosylation.

O-linked glycosylation refers to the carbohydrate moiety where it is covalently linked to the hydroxyl oxygen of serine and threonine residues of mammalian glycoproteins [9]. In addition, O-glycosylation also occurs as a primary modification of tyrosine and as a secondary modification of 5-hydroxylysine and 4-hydroxyproline [10]. O-linked glycosylation plays important roles in protein localization and trafficking, protein solubility, antigenicity and cell-cell interactions. O-linked glycans are built up in a stepwise fashion with sugars added incrementally. The most common type of O-glycosylation in secreted and membrane-bound mammalian proteins seems to be the addition of reducing terminal N-acetylgalactosamine (GalNAc). This type of O-linked glycan is also referred to as "mucin-type" glycan. There are also several types of non-mucin O-glycans, including α-linked O-fucose, β-linked O-xylose, α-linked O-mannose, β-linked O-GlcNAc (N-acetylglucosamine), α- or β-linked O-galactose, and α- or β-linked O-glucose glycans. Mucin glycoproteins are ubiquitous in mucous secretions on cell surfaces and in body fluids. Some cytoplasmic and nuclear proteins have simple O-linked glycans in which a single N-acetylglucosamine residue is linked to a serine or a threonine. This modification has been identified in a number of eukaryotes including plants and filamentous fungi. This type of O-linked glycosylation plays an important role in the modulation of the biological activity of intracellular proteins; in some proteins the same residue may be subject to competing phosphorylation and O-linked glycosylation [9].

The simplest mucin O-glycan is a single N-acetylgalactosamine residue linked to serine or threonine. The most common O-GalNAc glycan is Galβ-(1,3)GalNAc, and it is found in many glycoproteins and mucins. It is termed a core 1 O-GalNAc glycan because it forms the core of many longer, more complex structures. Another common core structure contains a branching N-acetylglucosamine attached to "core 1" and is termed "core 2". Core 2 O-GalNAc glycans are found in both glycoproteins and mucins from a variety of cells and tissues. Core 3 and core 4 O-GalNAc glycans have been found only in secreted mucins of certain mucin-secreting tissues, such as bronchi, colon, and salivary glands. Core structures 5–8 have an extremely restricted occurrence (*e.g.* core 5 have been reported in human meconium and intestinal adenocarcinoma tissue, core 6 are found in human intestinal mucin and ovarian cyst mucin, core 7 are shown in bovine submaxillary mucin, and core 8 has been reported in human respiratory mucin) [9]. All of the core structures can be sialylated. However, only cores 1–4 and core 6 have been shown to occur as extended, complex O-glycans that carry antigens such as the ABO and Lewis blood group determinants. The terminal structures of O-GalNAc glycans may contain fucose, galactose, N-acetylglucosamine, and sialic acid in α-linkages, N-acetylgalactosamine in both α- and β-linkages, and sulfate. Many of these terminal sugar structures are antigenic or represent recognition sites for lectins. In particular, the sialylated and sulfated Lewis antigens are ligands for selectins. Poly-N-acetyllactosamine units and terminal structures may also be found on N-glycans and glycolipids.

1.1.1 N-linked Glycosylation

N-linked glycosylation is the most common saccharide protein modification, and is the basis for the computational efforts described in this thesis. In N-linked glycosylation the carbohydrate moiety is attached to the amide nitrogen of the side chain of asparagine, when asparagine is part of the consensus sequence, Asn-X-Ser/Thr. Position "X" is referred to any amino acid other than proline. A proline residue there would prevent N-glycosylation [11]. N-glycosylation begins as a co-translational event in the endoplasmic reticulum (ER), where blocks of 14 sugars (including two N-acetylglucosamines, nine mannoses, and three glucoses) are first added to the polypeptide chain. After cleavage of 3 glucose and 1 mannose residues, the protein is transferred to the Golgi apparatus where the glycans lose a variable number of mannose residues and acquire a more complex structure [12]. There are three types of N-glycans: (1)

complex type N-glycans; (2) hybrid type N-glycans; and (3) high-mannose type N-glycans that are abundantly present on the envelope glycoprotein gp120 of HIV, but are rare on mammalian glycoproteins [12]. A few examples of the glycan structures are shown in Figure 1-2.

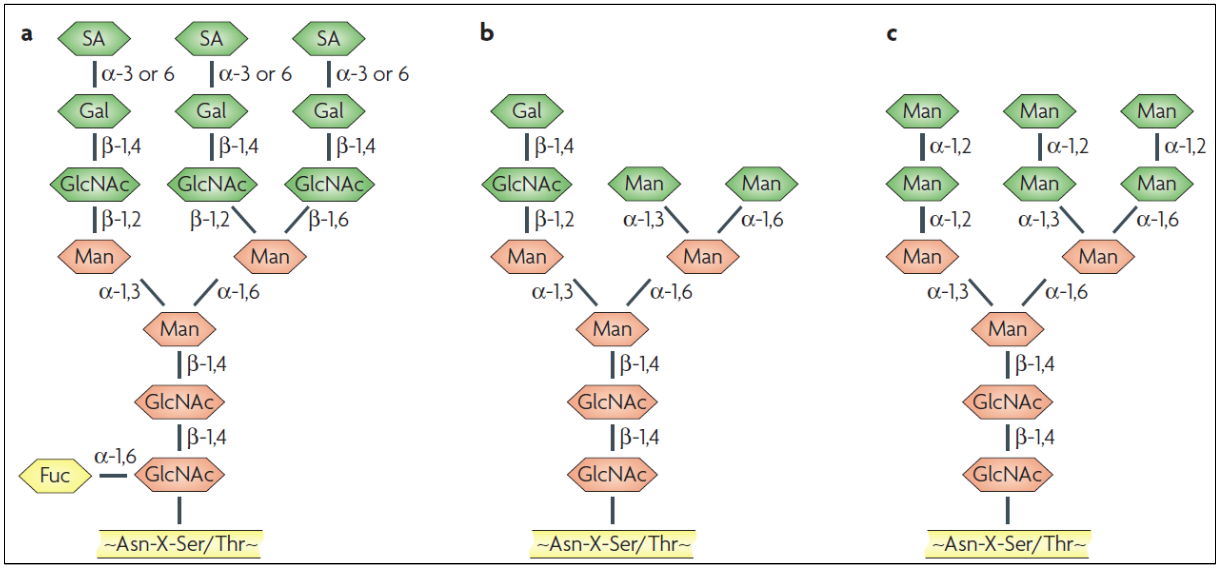


Figure 1- 2. **Representative structures of different N-glycan types.** All N-linked glycans are based on the common core pentasaccharide, $Man_3GlcNAc_2$ (shown in red). There are three main classes of N-linked glycan classes: (a) complex type N-glycans; (b) hybrid type N-glycans; and (c) high-mannose type N-glycans. High-mannose glycans contain mannose sugars and typically contain between five and nine mannose residues attached to the $GlcNAc_2$ core. Hybrid glycans are characterized as containing both mannose residues and substituted mannose residues with an N-acetylglucosamine linkage. Complex N-linked glycans differ from the high-mannose and hybrid glycans by adding GlcNAc residues at both α-(1,3) and α-(1,6) mannose sites. Complex glycans do not contain mannose residues apart from the core structure. Additional monosaccharides may occur in repeating GlcNAcβ-(1,4)Gal units. Complex glycans exist in bi-, tri- and tetra-antennary forms and make up the majority of cell surface and secreted N-glycans. Complex glycans commonly terminate with sialic acid residues. The name abbreviations are as follow: Asn, asparagine; Fuc, fucose; Gal, Galactose; GlcNAc, N-acetylglucosamine; Man, mannose; SA, sialic acid; Ser, serine; Thr, threonine; X, any amino acid except proline. *(Reprinted by permission from Nature Publishing Group: Nature Reviews Microbiology* [12], *copyright 2007).*

1.1.2 Interactions with Carbohydrate−Binding Agents

There are basically two types of carbohydrate−binding agents, (1) lectins, which are proteins that specifically recognize carbohydrate (glycan) structures, and (2) non-peptidic small-size agents that may have a good and often specific affinity for monosaccharide and/or oligosaccharide structures [12]. This dissertation focuses on the first type. Lectins are involved in many biological processes, among them host–pathogen interactions, cell–cell communication, induction of apoptosis, cancer metastasis and differentiation, targeting of cells, as well as recognizing and binding carbohydrates [13]. Currently, over 600 complexes of lectins with carbohydrates have been solved; most from plant sources [14]. Table 1-1 shows several examples of lectins that have anti-HIV activity.

Experimental studies have shown that the binding regions of carbohydrate−lectin complexes are mostly in the form of shallow clefts on the surface of the protein, where typically one or two segments of the ligand are bound [1]. Lectins can interact by way of hydrogen bonds, hydrophobic, electrostatic, and water-mediated interactions [15]. Hydrogen bonds are involved in providing affinity and specificity to protein–carbohydrate interactions [16]. They depend largely on interactions involving the hydroxyls of the carbohydrate. A sugar hydroxyl has the ability to interact with a protein both as a hydrogen bond donor and as an acceptor. As a hydrogen bond donor, the hydroxyl has rotational freedom around the C–OH torsional angle, and thus can often attain a strong linear bond with an acceptor group.

Table 1- 1. **Anti-HIV carbohydrate-binding proteins of a non-mammalian origin.**

Species	Lectin Name	Abbreviation	Carbohydrate Specificity
Cyanobacteria			
Nostoc ellipsosporum	Cyanovirin-N	CVN	Manα-(1,2)Man
Scytonema varium	Scytovirin	SVN	Manα-(1,2)Manα-(1,6)Man
Microcystis viridis	None	MVL	Manβ-(1,4)GlcNAc
Sea corals			
Gerardia savaglia	None	GSL	D-Man
Algae			
Griffithsia spp.	Griffithsin	GRFT	Man, Glc, GlcNAc
Plants			
Hippeastrum hybrid	Amaryllis lectin	HHA	Manα-(1,3)Manα-(1,6)Man
Narcissus pseudonarcissus	Daffodil lectin	NPA	Manα-(1,6)Man
Canavalia ensiformis	Jack bean lectin	ConA	Man>Glc>GlcNAc

The species listed above are a subset of many carbohydrate binding agents that have been isolated and characterized to show anti-HIV activity. The ">" symbol indicates a higher preference. The name abbreviations are as follow: Glc, glucose; GlcNAc, N-acetylglucosamine; Man, mannose.

Bidentate hydrogen bonds form when each of two adjacent hydroxyls of a monosaccharide interact with different atoms of the same amino acid (*e.g.* two oxygens from the carboxylate of glutamic or aspartic acid) [16]. Even though carbohydrates are highly polar molecules, the position of the hydroxyl groups can create hydrophobic regions on their surfaces, which can form contacts with hydrophobic side chains of the protein [17]. Sometimes contacts between the protein and ligand are mediated by water bridges [18]. These water–mediated interactions may be important for ligand recognition.

1.2 N-linked Glycans and the Glycobiology of HIV

One leading example of a protein–carbohydrate interaction comes from the human immunodeficiency virus (HIV-1). HIV is a member of the retrovirus family and it can hide for long periods of time in the body's cells. It attacks the T helper lymphocytes in the immune system. These cells play a crucial role in the immune system, by coordinating the actions of other immune system cells. A large reduction in the number of T helper cells seriously weakens the immune system. HIV infects the T helper cell because it has the protein CD4 on its surface, which HIV uses to attach itself to the cell before gaining entry. Once it has found its way into a cell, HIV produces new copies of itself, which can then go on to infect other cells. Over time, HIV infection leads to a severe reduction in the number of T helper cells available to help fight disease. It can take several years before the CD4 count declines to the point that an individual needs to begin an antiretroviral treatment. Without treatment, the CD4 count continues to decline to very low levels, at which point the individual is said to have progressed to AIDS. HIV is believed to have originated in primates in sub-Saharan Africa and transferred to humans early in the 20th century [19]. As of 2009, the Joint United Nations Program on HIV/AIDS and the World Health Organization estimated 34 million people worldwide were living with HIV. Of those, approximately 2.7 million people became newly infected and approximately 2.3 million people lost their lives to AIDS [20]. In the same report, the World Health Organization has estimated 1.2 million adults (15 and over) and children in the United States were infected with HIV [20].

As mentioned earlier, HIV is an enveloped retrovirus covered in glycoprotein (gp120), which is the major surface envelope glycoprotein of HIV-1 and contains 24 putative N-glycosylation sites [12]. The HIV-1 gp120 consists of approximately 30−40% high-mannose glycans (Figure 1-3). The N-linked glycans are critical to HIV infectivity, as they contribute to the proper folding of envelope glycoproteins and increased cyto-pathogenicity of the virus. Inhibition of N-linked glycosylation or downstream glycan processing has been proven to interfere with viral infectivity. Glycoprotein gp120 and integral membrane glycoprotein gp41 are attractive targets for developing therapies aimed to decrease HIV viral load in infected persons. Both reside in the external viral membrane as a trimer of three non-covalently associated heterodimers. This trimer plays an important role in HIV viral entry through interactions with CD4 molecules and chemokine receptors. The main steps in the viral entry process are (1) attachment of the viral gp120 to the CD4 T cell receptor, (2) binding of the gp120 to CCR5 or CXCR4 co-receptors and (3) fusion of the viral and cellular membranes. Disrupting this process interferes in viral infection.

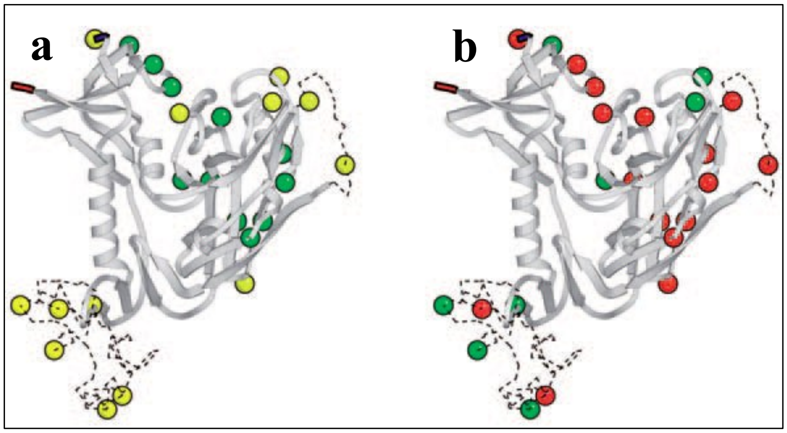


Figure 1- 3. **The HIV envelope glycoprotein gp120.** Ribbon diagrams showing the 24 putative N-glycosylation sites (colored circles) in the HIV-1 envelope glycoprotein gp120 [21, 22]**.** (a) High-mannose-type (green) and complex or hybrid type (yellow) glycans. (b) The red circles indicate the deleted N-glycosylation sites that appear under pressure from carbohydrate-binding agents (CBAs) (for example cyanovirin-N (CVN)) in more than 30 different mutant virus isolates. The green circles represent glycosylation sites that have not yet been found to be deleted under CBA pressure [22]. *(Reprinted by permission from Nature Publishing Group: Nature Reviews Microbiology* [12]*, copyright 2007).*

The prevalence of gp120 on the surface of HIV-1 makes it a target of immune response. A great deal of effort in vaccine design has focused on gp120 [23], but this approach is not always good [24]. The specific glycosylation pattern and structural composition of N-linked glycans is highly dependent on the host cell in which the virus was produced. Due to the high rate of mutation, HIV is able to optimize its interactions with various host proteins and pathways, thereby multiplying more quickly [25]. The virus ensures that the host cell survives until the replication cycle is completed, and—possibly even more damaging—may establish a stable latent form that supports the chronic nature of the infection. The complete suppression of the virus appears unlikely until effective methods are developed to purge these latent viral forms. More testing is needed to solve the mysteries of viral latency and replication [25].

Recently, there has been a growing interest in developing antiviral microbicides. Microbicides are substances that protect the body from infection by microorganisms such as bacteria, viruses, and fungi. They work by either destroying the microbes or preventing them from establishing an infection. Cyanovirin-N (CVN) is one of the microbicides being studied today. It was originally isolated from cultures of the cyanobacterium (blue green algae) *Nostoc ellipsosporum* [26]. CVN—a HIV−cell fusion blocker—was discovered in a National Cancer Institute (NCI) screening program for natural anti-HIV agents [27, 28]. CVN binds to the carbohydrates attached to the HIV envelope protein, and thus is an inhibitor of all strains of HIV. While CVN is being studied for the prevention of HIV infection, it does not cure HIV or AIDS.

1.3 Dissertation Objectives

One biological system the Green laboratory is focused on is the initial steps in the recognition of target cells by the HIV-1 virus. The Green lab is dedicated to advancing the field of glycobiology through the development of computational tools to aid in the ongoing investigations for understanding the role of glycosylation, in particular HIV glycobiology. The following dissertation focuses on the application and development of a comprehensive set of computational methods for analysis. In particular, we demonstrate the contribution of continuum electrostatic models and molecular dynamics simulations to dissect the protein−carbohydrate interactions. Using this method allows the separation of the individual contributions of various parts of the complex with their energies, which is inaccessible by experiments. Chapter 2 provides background to the theory and computational techniques used in this dissertation. The first project described in Chapter 3 involved simulations of complexes of CVN with a disaccharide and one trisaccharide model. The purpose of the study was to first validate the application of continuum electrostatic models to carbohydrate–protein association. In the next project shown in Chapter 4, we have extended the application used in Chapter 3 to various other saccharide models bound to CVN and developed a more robust procedure. After having confidence in the computational techniques used in the CVN system, Chapter 5 includes preliminary work done using the same protocol applied to another lectin—Microvirin (MVN). In addition, Chapter 5 includes work focusing on improvements to accurately capture the energetics of carbohydrate binding.

Chapter 2

Computational Methods to Study Protein−Carbohydrate Interactions

The range of systems that can be considered in molecular modeling is extremely broad, from simple isolated molecules to polymers and biological macromolecules (proteins and DNA). Computational studies of biological systems can play an important role complementary to experimental studies. These studies can separate individual contributions and energies of various parts of the complex in a way inaccessible to experiments. This chapter provides background to the theory and computational methods used in this dissertation.

2.1 Molecular Mechanics

In theory, quantum mechanical treatments should be the tool for a reliable description of a molecular system. However, this is not feasible for large macromolecules. A large number of particles must be considered and the calculations are time consuming. To overcome these obstacles, one method that has been particularly successful is molecular mechanics.

Molecular mechanics considers molecules as a collection of point particles, each with its own mass and bonds as springs with appropriate force constants [31]. Common force fields used for macromolecules include CHARMM [32, 33], AMBER [30], and OPLS [29, 34]. The CHARMM force field is used in this dissertation. CHARMM uses potential functions that approximate the total potential as a sum of bonded terms (stretching, angle, torsion, improper, and Urey-Bradley), plus potentials representing the non-bonded van der Waals, and electrostatic interactions [35, 36, 37]. This is shown in Figure 2-1 and Equation 2.1. In the bonded terms, bond stretching depends on the identity of the two atoms sharing the bond (l). The energy is approximated as a function of the coordinates of only the two atoms sharing the bond and the values of the constants k which depend on the types of the atoms that are involved. When CHARMM starts, it reads in all such force constants from a table of parameters. The energy of bond bending depends on the three atoms defining the angle (θ). The determination of the energy of bond twisting (torsion) requires four atoms to define the bond and the amount it is twisted (φ). This energy term usually allows a full 360° of rotation about a bond at normal temperatures, but introduces preferences in the angle that correspond to positions of minimum clash of the atoms bonded. Improper potentials are artificial potentials that are used to hold a group consisting of one central atom that is bonded to three others in a particular configuration. These structures are also known as improper dihedrals. By convention, the first of the four atoms listed is the central atom and ω is the angle between the plane containing atoms A, B, and C and the plane containing atoms B, C, and D. The energy constants used in these potentials are quite large, and so they serve to hold the atoms near the desired configuration. Finally, the Urey-Bradley component is the cross-term accounting for angle bending using 1, 3 non-bonded interactions (u) [38].

In addition to the terms discussed above that are transmitted by the covalent bonds between atoms, CHARMM includes van der Waals interactions and electrostatic interactions, both of which act at a distance. By definition, the non-bonded forces are only applied to atom pairs separated by at least three bonds. The van der Waals interactions are approximated with a standard 12-6 Lennard-Jones potential and the electrostatic energy with a Coulombic potential between two atoms, where q_i and q_j are the charges of the two atoms, r is their separation, and ε is the dielectric constant of the surrounding medium. Electrostatic and van der Waals interactions act over some distance and in principle, act between all pairs of atoms. Rather than compute their strength for every pair of atoms in a system, which would require an unmanageably large number of computations, CHARMM maintains and periodically updates a list of just those pairs of atoms that are sufficiently close together that they experience significant electrostatic or van der Waals interactions, and the two interactions are calculated only for atom pairs on the list. The contents of this non-bonded atoms list depends on the locations of atoms and on the approximations used in calculating the forces such as the non-bonded cutoff distance [38].

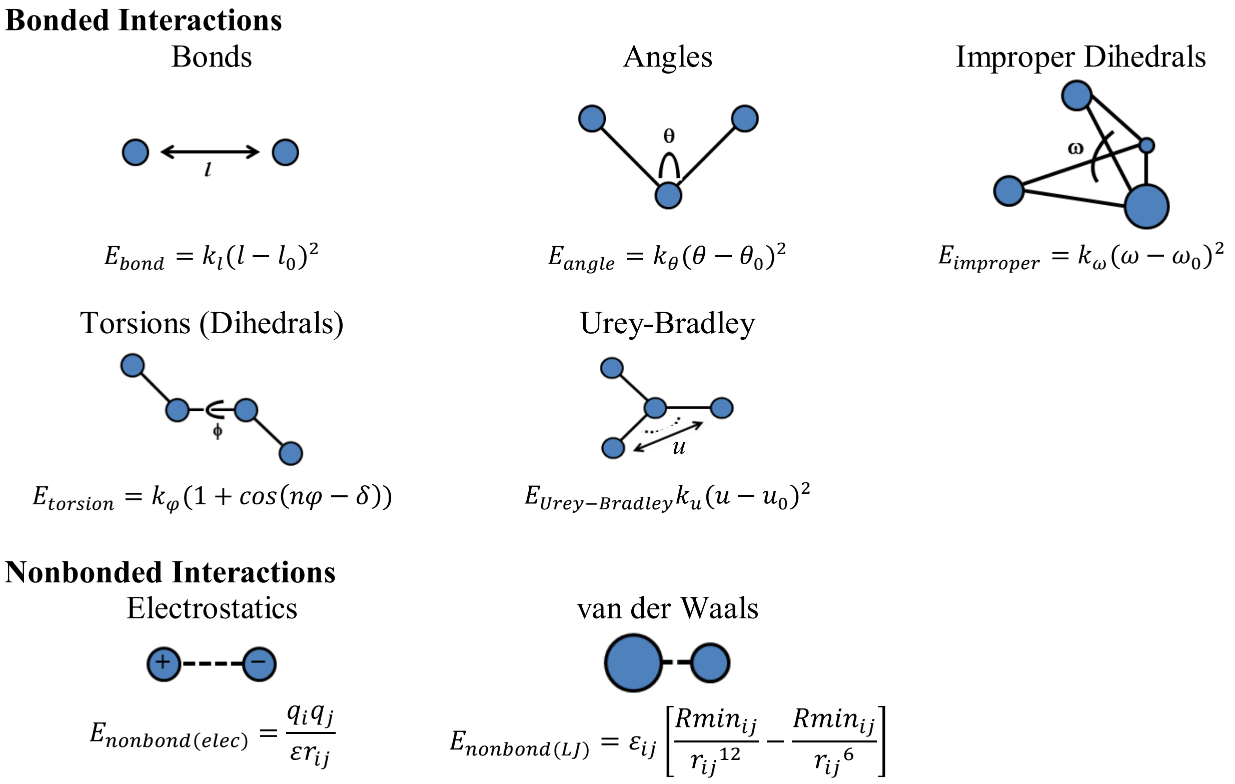

Bonded Interactions
Bonds
l
$E_{bond} = k_l(l - l_0)^2$
Angles
θ
$E_{angle} = k_\theta(\theta - \theta_0)^2$
Improper Dihedrals
ω
$E_{improper} = k_\omega(\omega - \omega_0)^2$
Torsions (Dihedrals)
ϕ
$E_{torsion} = k_\varphi(1 + cos(n\varphi - \delta))$
Urey-Bradley
u
$E_{Urey-Bradley} k_u(u - u_0)^2$
Nonbonded Interactions
Electrostatics
+
−
$E_{nonbond(elec)} = \frac{q_i q_j}{\varepsilon r_{ij}}$
van der Waals
$E_{nonbond(LJ)} = \varepsilon_{ij}\left[\frac{Rmin_{ij}}{{r_{ij}}^{12}} - \frac{Rmin_{ij}}{{r_{ij}}^{6}}\right]$

Figure 2- 1. **Schematic representations of bonded and nonbonded contributions to molecular mechanics.** Each type is labeled with an illustration and equation.

The form of the potential energy function used in CHARMM is given by the following equation:

$$V = \sum_{bonds} k_l\,(l - l_0)^2 + \sum_{angle} k_\theta\ (\theta - \theta_0)^2 + \sum_{improper} k_\omega(\omega - \omega_0)^2 + \sum_{torsions} k_\varphi\ \big(1 + \cos(n\varphi - \gamma)\big) + \sum_{Urey-Bradley} k_u\ (u - u_0)^2 + \sum_{nonbonds} \varepsilon\left[\left(\frac{Rmin_{ij}}{r_{ij}}\right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon r_{ij}} \tag{2.1}$$

In the next section, we focus on the conformational analysis of molecules, specifically protein–carbohydrate complexes. In this dissertation, we use molecular dynamics to study these complexes.

2.2 Molecular Dynamics

In the broadest sense, molecular dynamics is concerned with molecular motion. Motion is inherent to all chemical processes. Simple vibrations, like bond stretching and angle bending, give rise to IR spectra. Chemical reactions, hormone-receptor binding, and other complex processes are associated with many kinds of intra- and intermolecular motions. To complement experimental data, molecular dynamics can be used to understand more about complex molecular systems. Structures are available from database repositories (*e.g.* Protein Data Bank (PDB) [39]), which stores structures derived from experimental techniques such as electron microscopy (EM), X-ray crystallography, and nuclear magnetic resonance (NMR). The PDB was established in 1971 at Brookhaven National Laboratory and originally contained 7 structures. In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) became responsible for the management of the PDB, and currently has over 80,000 structures [39, 40]. The static view of a molecule from an X-ray crystal or NMR structure is just not enough to fully explain its biological role. Molecules are not frozen; the atoms continuously interact amongst themselves and environment. The motions can explain more about their structure and function, and serve as a starting point for molecular dynamics simulations.

Molecular dynamics simulations of macromolecules have provided many insights concerning the internal motions of these systems since the first protein was studied over 30 years ago [41, 42]. This simulation dealt with a folded globular protein, bovine pancreatic trypsin inhibitor (BPTI). Although the simulation was done in vacuum and lasted for only 9.2 ps, the results were instrumental in replacing our view of proteins as relatively rigid structures. During the subsequent years, a wide range of molecular dynamics simulations of proteins and nucleic acids were performed. They include the analysis of fluorescence depolarization of tryptophan residues [43], the effect of solvent and temperature on protein structure and dynamics [44], as well as using simulated annealing methods for X-ray structure refinement [45] and NMR structure determination [46]. With continuing advances in the methodology and the speed of computers molecular dynamics studies are being extended to larger systems [47], greater conformational changes, and longer time scales [48]. The results available today make clear that the applications of molecular dynamics will play an even more important role in our understanding of biology in the future; thus, the simulation of protein−carbohydrate complexes is an ideal system to study.

Water molecules surround most biological macromolecules (*e.g.* proteins and DNA), and because water molecules are polarized and make hydrogen bonds to other molecules, the energetics and motions within protein and DNA are critically dependent upon the surrounding water molecules. Therefore water is included in most calculations and simulations of proteins and DNA. Some groups on the surfaces of proteins hold water molecules quite tightly and generate structure in the water that persists for a thickness of several water molecules. Frequently to avoid any artifacts resulting from modeling with insufficient water, considerably more water is included. This means that some simulations of a protein may also include 10,000 to 200,000 water molecules. To minimize computation time it is important to represent water molecules just as simply as possible. On the other hand, since water is so important to the properties of biological macromolecules, it is important that the representation of water be as accurate as possible. Systems are normally simulated in a setting that represents its biological environment. For many cases, the molecular complex can be simulated in a solvent of water with sodium chloride ions or with a lipid bilayer. Molecular dynamics simulations can be done with explicit solvent molecules or by implicitly accounting for the effects of solvent. There are a variety of models for explicit water including TIP3P, TIP4P, TIP5P, SSD, SPC, and PPC [49]. The solvating water molecules are usually obtained from a suitable large box of water that has been previously equilibrated. The entire box of water is overlayed onto the macromolecule and those water molecules that overlap the protein are removed.

Currently the water molecules used in most CHARMM calculations are based on an approximation known as TIP3P [49]. In this, water is modeled with three point charges, two equal positive charges corresponding to the two partial charges on the hydrogen atoms and one negative charge (on the oxygen) equal in magnitude to the sum of the two positive charges. The van der Waals interactions involving a TIP3P water molecule are modeled with a Lennard-Jones potential centered at the oxygen atom and generally calculations are performed such that the oxygen-hydrogen bonds are held to a constant length and do not bend [38].

Molecular dynamic simulations involve a number of steps, and in the following section, an overview of how to prepare a simulation will be discussed.

2.2.1 Simulation Preparation

As mentioned earlier, an initial starting structure of the system needs to be selected. At this point, some care in choosing the structure is required. The initial arrangement of the structure often determines whether the simulation will be a success or failure. In many cases, the starting structure is obtained from X-ray crystallography or NMR and the coordinates of the structure can be found from the Protein Data Bank [39]. Once a structure is determined, atoms will be allowed to move with respect to one another. First, in energy minimization of a structure or system, positions of atoms are allowed to shift so as to bring the system energy to a lower value. An energy minimum is found by moving each atom down the potential until each is at a point where the energy can no longer be reduced by small movements [38].

The second source of movement is performing a molecular dynamics simulation. In such simulations atoms are given velocities appropriate to a chosen temperature and allowed to move in response to all the forces acting on them in paths determined by Newton's equations of motion. The study of the movements of proteins in molecular dynamics simulations may give more correct or more useful information about the structure and function of proteins than an examination of static structures. The studies clearly provide information about the motions and behavior of portions of proteins that range in size from side chains to domains.

One issue of concern in simulations is the possibility of anomalous effects generated by boundaries between a protein and the water molecule in which it is immersed and the rest of empty space, the surrounding vacuum. Periodic boundary conditions eliminate boundary effects, essentially by surrounding the molecule with space filling images of itself [38, 50]. This is done by mapping the edges of the volume being simulated back onto itself. In cubical geometry this is done by simulating the central cube and considering the surrounding cubes as identical images of the central cube. A number of different geometries completely fill space and can be used for calculations using periodic boundary conditions. These include cubes, rhombohedra, rhombic dodecahedra, truncated octahedra, and hexagonal prisms. When an atom or molecule passes out of the central cell, for example, by moving past the right boundary, it enters the right hand image box from the left (Figure 2-2). Of course, this same event happens in every one of the images of the central box. By these means no boundary exists, and surface effects are eliminated, but at the expense of increased computational load and occasional artifacts arising from the existence of images and interactions between image molecules and "real" molecules. The use of periodic boundary conditions is generally considered to best represent reality and many simulations are performed using periodic boundary conditions. Table 2-1 shows the major types of systems that can be simulated with CHARMM with various boundary conditions. The Newtonian equations of motion can be augmented by a friction term and random forces, which mimic the effect of collisions with a solvent. Since this is effectively equivalent to a heat bath, the method automatically controls the temperature. Due to the collisions, there is no conservation of rotation and translation.

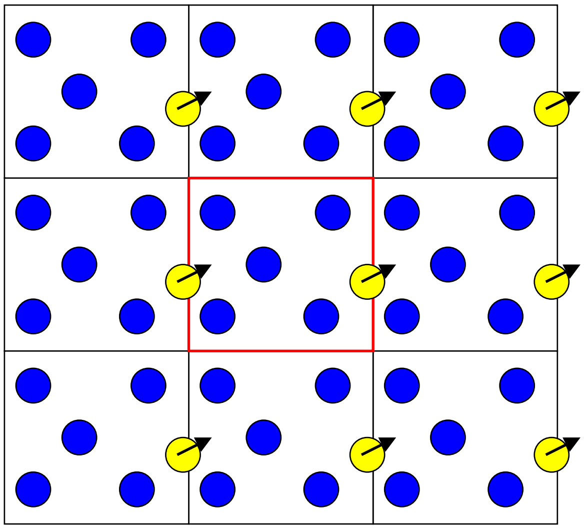


Figure 2- 2. **A representation of periodic boundary conditions.** The primary region (in red square) contains a solute (in yellow circle). The nearest-neighbor repeats this region (in black squares). The solute avoids interacting with its own image by having a cutoff of all non-bonded interactions be smaller than the distance between the edge of the solute and the nearest edge of its nearest image.

Table 2- 1. **Types of ensembles simulated by CHARMM.**

Ensemble	Constants	Boundary Conditions
Canonical	N, V, T	Fixed periodic boundary conditions
Microcanonical	N, V, E	Fixed periodic boundary conditions
Isothermic-isobaric	N, P, T	Periodic boundary conditions
Isoenthalpic-isobaric	N, P, H	Periodic boundary conditions
Langevin simulation	N, P, T or N, V, T	Periodic boundary conditions
	N, E	No boundary (infinite)

Listed are examples of systems that can be simulated with CHARMM under different conditions and various boundary conditions. The name abbreviations are as follow: N (number of particles); V (volume); P (pressure); E (total energy); T (temperature); H (enthalpy).

Starting a dynamics run with every atom at rest and every atom initially feeling a net force of zero (i.e. at energy minimum) is equivalent to starting the system at absolute zero. One of the main reasons, however, for doing molecular dynamics is permit study of systems under normal conditions. Thus, an energy minimized system needs to be warmed up to about 300 degrees Kelvin. In some cases it is possible to assign each atom a velocity near the value characteristic for an atom of that mass at 300 degrees and let the system evolve in time thereafter. Sometimes, however, this approach leads to the development of intolerable instabilities and CHARMM automatically stops the simulation. Therefore, usually systems are brought to the desired temperature by gradually incrementing each atom's velocity and then allowing time for this perturbation to die away before the next velocity increment. Often, systems are not completely equilibrated after these heating steps, and further evolution in the absence of heating is needed. Ordinarily the system then comes to equilibrium at a temperature slightly different from what is desired. After a further period of equilibration, the velocities are scaled so as to yield the desired target temperature. This process of equilibration followed by velocity scaling is performed until the system temperature remains constant at the desired value. Then the system is allowed to evolve in the absence of velocity adjustments. This is called the simulation or production phase. This can be from several hundred picoseconds to nanoseconds or even more now with the access to supercomputers. It is during the production phase that thermodynamic parameters can be calculated [38, 50].

As mentioned earlier, solvent plays an important role to the structure, dynamics, and function of biological macromolecules. When modeling biological systems with a solvent environment has been challenging. An explicit representation of the surrounding solvent can provide accurate treatment of solute−solvent interactions, but it typically increases the size of the system; therefore, it becomes computationally costly. Another disadvantage to the explicit solvent approach is entropy properties are difficult to determine accurately in simulations, since it is rarely clear to what extent all the important regions of phase space are sampled in the simulation [51]. To address this issue, an alternative method is applied to biological systems, called the continuum electrostatic model. In this approach, molecules are described as a set of point charges located at the center of the atoms in a region of low dielectric constant that is surrounded by a solvent treated as a region of high dielectric continuum. The electrostatic potential $\phi(r)$ is calculated by solving the linearized Poisson−Boltzmann (PB) equation (Equation 2.2),

$$\nabla \cdot [\varepsilon(r)\nabla\phi(r)] - \varepsilon(r)\kappa^2(r)\phi(r) = -4\pi\rho(r) \tag{2.2}$$

where $\varepsilon(r)$, $\kappa(r)$, and $\rho(r)$ are the dielectric constant, the Debye−Hückel screening factor, and the charge distribution of the solute, respectively. The different environments are characterized mainly by their dielectric response, with ε values ranging from 80 for water to 2−4 for the interior [52]. After the simulation has been completed, trajectories will be analyzed both structurally and energetically. The following section will discuss these methods.

2.3 Structural and Energetic Analysis from Molecular Dynamics Trajectory

The following sections give a brief introduction of the post-processing methods that will be covered in the subsequent chapters. Along with calculating the energies of a particular system, it is also important to visualize a molecular dynamics simulation. It generates a trajectory which is a set of configurations taken over a period of time. The program VMD [53] has been developed for interactive graphical display of molecular systems.

2.3.1 Binding Free Energy Calculations

Currently, there are over 80,000 crystallographic or solution structures of macromolecules available from the PDB. The rate of macromolecular structure determination continues to grow every year, particularly with the development of new techniques (*e.g.* high throughput X-ray crystallography). There have been many structure-based screening methods that have been developed to assist in identifying lead compounds that bind with reasonable affinity (low mM to nM range) to a particular target. Similarly, to aid lead optimization, computational methods capable of predicting binding affinities of similar compounds are required. In general, calculations of binding free energies play a significant role in correlating the structure and function of proteins. A commonly used method is the molecular mechanics Poisson–Boltzmann/surface area (MM/PBSA) [54]. This approach predicts the absolute binding free energy by combining molecular mechanics (MM) energy and solvation free energies with Poisson–Boltzmann calculations. This dissertation aims to calculate the binding free energy with the MM/PBSA approach.

Calculating free energies has been widely used with researchers working with molecular simulations. However, calculating free energies has been troublesome from the fact that accurate free energy results can be obtained if the contributions of all the populated states (including configurational, conformational, and vibrational states) are included in the calculation. To deal with such problems, relative free energies of rigid binding energies are calculated for a set of molecules that are computed by comparing the average ensemble energies extracted from bound state molecular dynamics and those of related conformations from the unbound state.

2.3.2 Energetic Decomposition of Binding

In many cases, refined tools are needed to understand the binding free energies in more detail. In order to understand the energy terms (i.e. intermolecular van der Waals, electrostatic, and solvent accessible surface area) that are between the binding partners, it is possible to break down the terms from various parts of the molecule. For proteins, every residue is divided into a backbone amino, backbone carbonyl, and a side chain group. For carbohydrates, we can divide based on functional group (such as each hydroxyl group).

Electrostatic interactions play a significant role in the structure and function of biological macromolecules. Hendsch, *et al.* studied the GCN4 leucine zipper [55] to investigate the electrostatic effects in protein binding using continuum calculations. To understand the role of electrostatic interactions in a complex, three types of calculations can be determined for every group. One term is the desolvation energy of an individual group, the second term can be defined as the solvent-screened Coulombic interactions between two groups in the bound state (direct interaction), and lastly, the difference in the solvent screening that intramolecular interactions experience in the bound and unbound states (indirect interactions). All three terms can be summed up to give the electrostatic contributions to binding. Within a group, the total of the desolvation, direct, and indirect terms gives an energy called the mutation free energy. It corresponds to the binding energy difference of the native complex and that of a complex with a specific group of interest that is substituted by a hydrophobic isostere. The positive or negative values indicate the unfavorable and favorable contributions to the binding affinity. This is a good way to determine the influences of single residues on the binding affinity and pinpoint hotspot residues. This information can be used as a guide for design. The van der Waals interactions also plays a role and two types of calculations can be determined—energy interactions from the protein or with the ligand.

This dissertation describes research projects involving the methods stated above. The project described in Chapter 3 involved simulations of the CVN bound to di- and one trimannose model. The purpose of this study is to validate the current techniques listed above.

Chapter 3

Computational Models Explain the Oligosaccharide Specificity of Cyanovirin-N

This chapter has been published as **Fujimoto, Y. K**.; TerBush, R. N.; Patsalo, V.; and Green, D. F. Computational models explain the oligosaccharide specificity of cyanovirin-N. *Prot. Sci.,* **2008**, 17, 2008–2014.

Author contributions: YKF performed research and analyzed data; RNT began the research; VP performed the implicit-solvent molecular dynamics; YKF and DFG wrote the paper.

Abstract

The prokaryotic lectin cyanovirin-N (CVN) is a potent inhibitor of HIV envelope-mediated cell entry, and thus is a leading candidate among a new class of potential anti-HIV microbicides. The activity of CVN is a result of interactions with the D1 arm of high-mannose oligosaccharides on the viral glycoprotein gp120. Here, we present computationally refined models of CVN recognition of the di- and trisaccharides that represent the terminal three sugars of the D1 arm by each CVN binding site. These models complement existing structural data, both from NMR spectroscopy and X-ray crystallography. When used with a molecular dynamics/continuum electrostatic (MD/PBSA) approach to compute binding free energies, these models explain the relative affinity of each site for the two saccharides. This work presents the first validation of the application of continuum electrostatic models to carbohydrate–protein association. Taken as a whole, the results both provide models of CVN sugar recognition and

demonstrate the utility of these computational methods for the study of carbohydrate-binding proteins.

3.1 Introduction

Since the discovery of the antiviral activity of legume lectins such as Concanavalin A more than 20 years ago [56], it has been known that one mechanism of blocking HIV-envelope-mediated cell entry is the association of a carbohydrate-binding protein with the oligosaccharides on the surface of the viral envelope. In particular, the outer envelope glycoprotein, gp120, is heavily glycosylated by N-linked carbohydrates, both of the high-mannose and complex types [21]. As this glycosylation plays a key role in viral avoidance of natural immune responses, inhibitors that work through specific interactions with these carbohydrates provide a unique opportunity to interfere with cellular infection in a manner that makes it difficult for viral resistance to evolve [57].

To date, numerous lectins from diverse sources have been identified as having virucidal activity against HIV, but among the best characterized is cyanovirin-N (CVN), originally isolated from the cyanobacterium, *Nostoc ellipsosporum* [26]. Both inhibition and calorimetric studies have determined that CVN specifically targets the α(1-2)-linked mannoses found on the D1 arm of high-mannose oligosaccharides; CVN contains two pseudo-symmetric binding domains which have differing affinities and specificities for various oligosaccharides [58, 59]. Several structures of CVN have also been solved, both in solution [60, 61] and crystal phases [62, 63]. Both its remarkable stability to denaturation and its potent antiviral activity have made CVN one of the leading candidates for use as a biopharmaceutical. Preclinical trials in primates have shown promise for the use of CVN as a topical agent to prevent sexual transmission of HIV [64, 65].

Despite the apparent wealth of data for this system, there remain open problems regarding the structure and energetics of specific sugar recognition. The crystal structures of CVN bound to each of two high-mannose oligosaccharides, Man_6 and Man_9, have been solved [63], but the carbohydrates are not fully resolved. While they are of reasonable resolution (2.4 and 2.5 Å), only one of the two binding sites is occupied, and the oligosaccharide structures deviate strongly from that expected; several mannoses are in the β configuration (where α anomers are expected), and numerous rings are in disfavored ring conformations. Overall, only the small portion of the oligosaccharide making the most intimate contact with the protein is particularly well structured. A solution structure with both sites bound to the disaccharide Manα-(1,2)Man is also available [61]. In this structure, however, few NOE constraints were available to accurately define the lower affinity site. Here, we present computationally refined models of α-(1,2)-linked di- (Man$_2$) and trimannose (Man$_3$), representative of the D1 arm of Man$_9$, bound in both sites. These models, combined with molecular dynamics and continuum electrostatic analysis, capture the observed specificity of binding with semi-quantitative accuracy.

3.2 Results and Discussion

Initially, explicit-solvent molecular dynamics simulations were carried out, beginning with the solution structure in complex with two mannose disaccharides (PDB 1iiy) [61]. These simulations quickly revealed instability in the lower affinity binding site; the disaccharide began to dissociate from cyanovirin-N within the first 500 ps of simulation (see Figure 3-1, top panels). In the higher affinity site, however, the sugar remains stably bound for as long as we have simulated (currently upwards of 20 ns). Similar behavior has been noticed by Margulis, using different parameters and simulation conditions [66], suggesting that the instability is not simply an artifact of the method.

The difference between the two sites in affinity for Man$_2$ is small (10-fold difference in K_a, as measured by a two-site fit to calorimetric data) [58]; thus, this behavior is not expected. The two binding sites of CVN are pseudo-symmetric, but the orientation of the sugar in each site of the structure differs noticeably; the sugar in the low-affinity site makes less intimate contact with some binding-site residues (see Figure 3-2, top panels). A higher symmetry model of dimannose binding in the low-affinity site was thus considered. The protein backbone of the high affinity binding site was superimposed on that of the lower affinity site, and the sugar from the high affinity site placed in the lower affinity pocket. The structure was then briefly minimized while all protein residues further than 4.0 Å (minimum distance) from the sugar were held fixed. Molecular dynamics simulation from this starting structure shows similar stability to that of the high-affinity site. As an additional test, this procedure was repeated in reverse—the low-affinity site was superimposed on the high affinity site, and the low-affinity sugar orientation placed in the high-affinity pocket. The binding site was briefly relaxed, and then subjected to molecular dynamics. This model shows the same instability as was first observed in the lower affinity site. These data are displayed in Figure 3-1 (top panels). However, it remains possible that the observed stability of the refined structure is due to the constraints used in the simulation, which involved a sphere of solvent placed around each binding site, with the protein atoms outside this sphere held fixed.

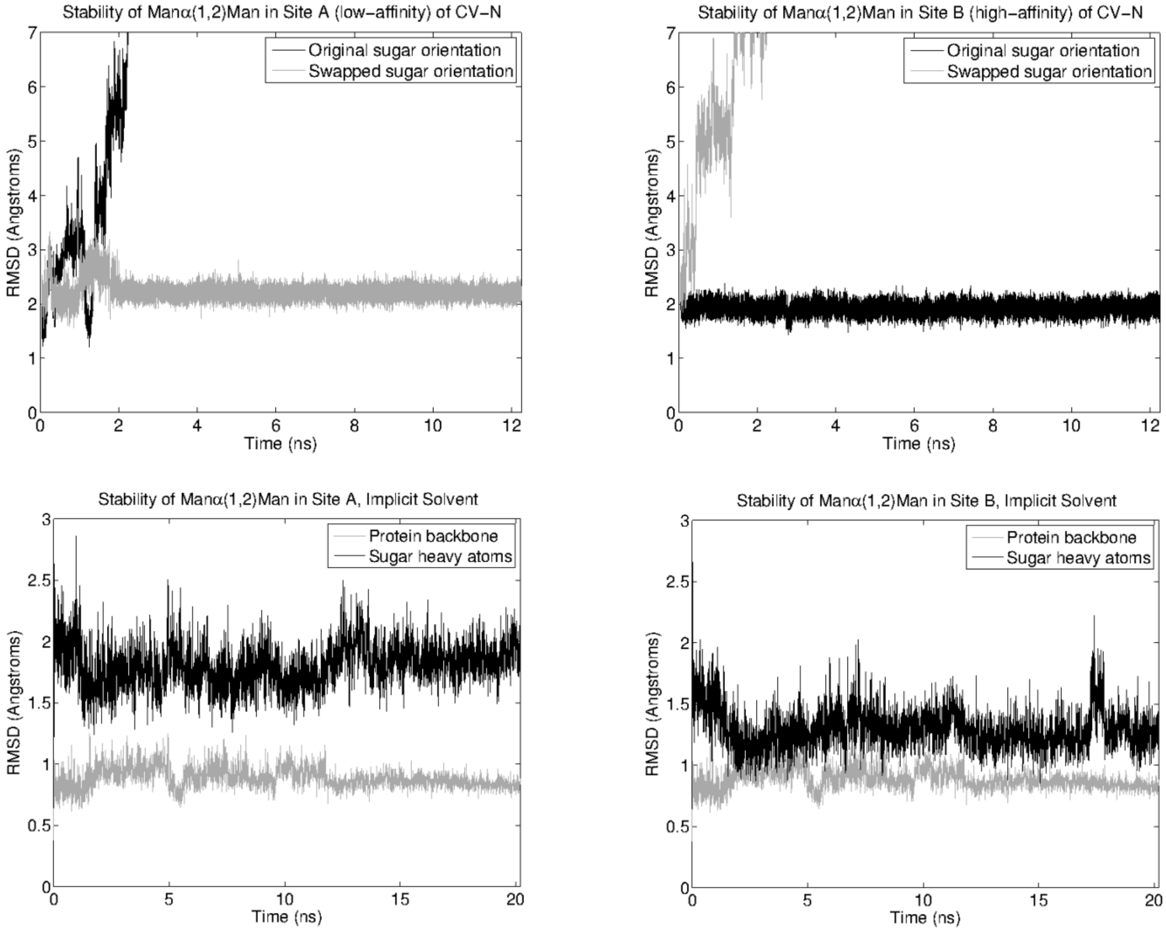
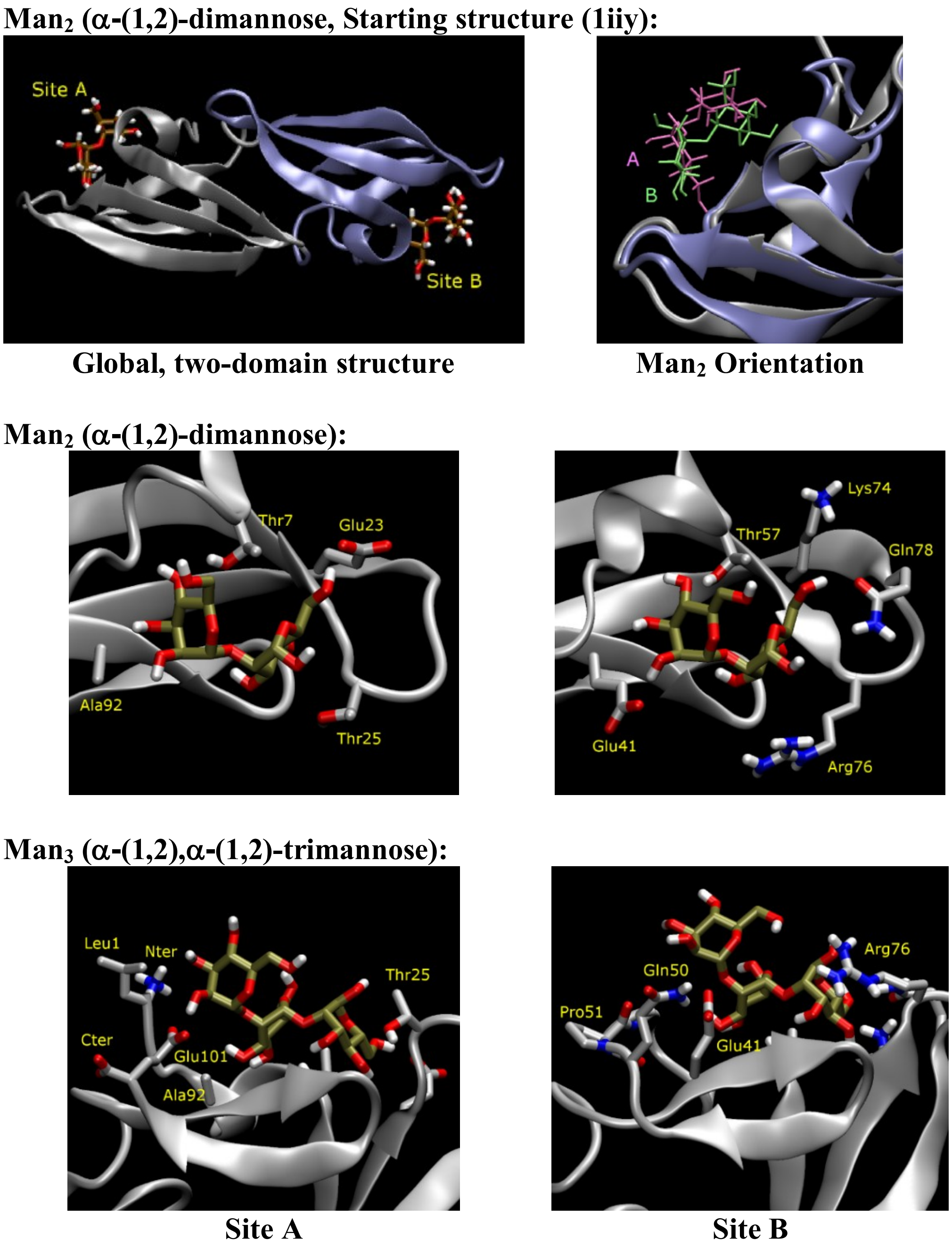


Figure 3- 1. **Stability of Man_2 in each CVN binding site.** *(Top)* The root-mean-square deviation (RMSD) of sugar-heavy atoms, relative to the initial coordinates, is shown over 22 ns of explicit-solvent molecular dynamics. Trajectories beginning from the published structure are in black; those beginning with the sugar placed in the orientation found in the other site are shown in gray. A vertical bar indicates the end of the equilibration phase; all analysis was done on frames following this point. *(Bottom)* The RMSD of both protein backbone atoms (in black) and sugar-heavy atoms (in gray), relative to the initial coordinates, are shown over 30 ns of implicit solvent (Generalized Born) molecular dynamics.



Man_2 (α-(1,2)-dimannose, Starting structure (1iiy)):

Global, two-domain structure

Man_2 Orientation

Man_2 (α-(1,2)-dimannose):

Man_3 (α-(1,2),α-(1,2)-trimannose):

Site A

Site B

Figure 3- 2. **Structures of di- and trimannose bound to CVN.** *(Top left)* The overall structure of CVN, showing two sugar binding sites in pseudosymmetric domains (domain A in gray, domain B in blue). *(Top right)* A superposition of the two binding sites of CVN, showing the difference in orientation of the original Man_2 model in the low-affinity site (pink/gray) from the structure of the high-affinity site (green/blue). *(Middle and bottom)* A representative frame from each dynamic simulation is displayed, with protein in gray and sugar in bronze. All amino acid side chains involved in significant electrostatic interactions are shown. The viewpoint perspective of Man_3 is rotated roughly 90° from that of Man_2. Figures generated with VMD [53].

To evaluate the possibility, an unconstrained simulation using the Generalized Born implicit solvent model was run, beginning with the preferred sugar orientation in each site. The sugars remain stably bound throughout this simulation (Figure 3-1, bottom panels), while similar simulations with the nonpreferred starting orientations show rapid dissociation. Taken in combination, these results strongly suggest that the source of the instability is the initial orientation of the sugar.

Fewer NOEs were observed experimentally for the lower affinity site than for the higher, and thus the published model was not uniquely determined by experimental constraints [61]. To further test the validity of the new model, we tracked the distances of all atoms involved in observed intermolecular NOEs throughout the first 10 ns of molecular-dynamics simulation (see Table 3-1); all experimentally observed contacts remain within 6.0 Å for the majority of the simulation, and most remain within 5.0 Å or less. In comparison, several contacts in the published model are beyond the largest constraint distance (6.0 Å). While the timescale of the simulation is much lower than that observed in the NOESY experiment, these data demonstrate the consistency of the refined model with available data.

These structures begin to explain the key determinants of oligosaccharide specificity in the two sites of CVN, as shown in Figure 3-2 (middle panels). In particular, a strong electrostatic interaction in the high-affinity site (between Glu41 and Man_2 OH2) is absent in the low-affinity site, where the corresponding residue is an alanine (Ala92). There are other ways in which the two binding sites vary, but the overall differences in interactions are less pronounced. A hydrogen bond made between Glu23 and Man1 OH6 in Site A is replaced by a similar interaction with Gln78 in Site B; these residues are not in equivalent positions, but make equivalent interactions. Thr25 in Site A is replaced by Arg76 in Site B, but neither makes close, specific interactions with the sugar. Interestingly, the Glu/Ala variation at sites 41/92 also may explain an intriguing feature of CVN—α-(1,2)-linked trimannose binds to these sites with a specificity reversed from that of the dimannose. That is, Site A (low affinity for dimannose) binds the trimannose with higher affinity than does Site B (high-affinity for dimannose). With an additional α(1–2)-linkage extending from Man_2, the favorable hydrogen bond made by Glu41 in Site B would be lost, as the donating hydroxyl would become part of the glycosidic bond; in Site A the lack of an interaction with this hydroxyl would more easily accommodate this change.

Table 3- 1. **Persistence of observed NOEs in simulation of refined structure.**

Sugar atom	CVN atom	d[a]	Occ.[b]	τ[c]	CVN atom	d[a]	Occ.[b]	τ[c]
Man$_1$:								
H2	I94-C$_{\delta}$H$_3$	6.0 (6.8)	1.70	1.37				
H4	T7-C$_{\gamma}$H$_3$	5.0 (5.2)	1.43	0.92				
H5	T7-C$_{\gamma}$H$_3$	5.0 (2.3)	0.96	3.56	T25-C$_{\alpha}$H$_1$	4.0 (2.4)	1.00	3333
H6b	T25-C$_{\alpha}$H$_1$	4.0 (2.9)	0.98	12.5				
Man$_2$:								
H2	A92-C$_{\beta}$H$_3$	4.0 (5.4)	1.05	3.81	I94-C$_{\delta}$H$_3$	4.0 (5.0)	1.16	0.90
H3	A92-C$_{\beta}$H$_3$	5.0 (5.5)	0.97	3.66	I94-C$_{\delta}$H$_3$	6.0 (5.2)	1.12	0.90
H4	K3-C$_{\delta}$H$_2$	4.0 (3.4)	1.11	17.2	Q6-C$_{\beta}$H$_2$	5.0 (4.2)	1.45	1.77
	T7-C$_{\gamma}$H$_3$	5.0 (4.5)	1.40	1.06				
H5	T7-C$_{\beta}$H$_1$	5.0 (5.2)	1.00	175	T7-C$_{\gamma}$H$_3$	4.0 (3.0)	2.60	2.16
H6a	K3-C$_{\delta}$H$_2$	6.0 (6.7)	0.70	0.95	Q6-C$_{\beta}$H$_2$	4.0 (7.2)	1.29	1.93
	Q6-C$_{\gamma}$H$_2$	5.0 (5.3)	1.38	1.85	T7-C$_{\beta}$H$_1$	4.0 (6.4)	0.62	4.00
	T7-C$_{\gamma}$H$_3$	4.0 (4.4)	1.98	1.97				
H6b	K3-C$_{\delta}$H$_2$	6.0 (5.5)	0.59	0.96	Q6-C$_{\beta}$H$_2$	5.0 (5.7)	1.48	4.78
	T7-C$_{\beta}$H$_1$	5.0 (7.3)	0.82	1.32	T7-C$_{\gamma}$H$_3$	4.0 (5.1)	2.33	2.17

[a] Persistent contact cutoff from MD (distance in published structure, [61]) in Å.

[b] Fraction of simulation in which contact is made (sum of all hydrogens in group).

[c] Average lifetime of contact, in ps.

To investigate these further, models of Man$_3$ in each site were generated by extending the dimannose by one unit. A new glycosidic bond formed either with OH2 of Man$_2$ or OH1 of Man$_1$ would result in the same α(1–2),α(1–2)-linked trimannose; the structure of Man$_9$—accurate in this region—indicates that the first of these configurations is preferred [63], and thus this was used in model building. Again, molecular dynamics simulations were performed to assess stability, and the sugar in both binding sites remained stably bound throughout. Representative structures are shown in Figure 3-2 (bottom). These structures support the analysis above: The incorporation of Man$_2$ OH2 into the glycosidic bond makes Glu41 unable to interact favorably with the sugar in Site B. As well, two new interactions are made in Site A, explaining the increased affinity of this site for the trisaccharide: Glu101 (the C-terminal residue) receives a hydrogen bond from OH2 of the third sugar; and the charged N terminus donates one to OH3. In Site B, neither of these interactions is made: Gln50 (analogous to Glu101) is involved in satisfying Glu41 with an intramolecular interaction, and thus makes no interactions with the sugar; and a proline-containing loop replaces the chain termini, eliminating any hydrogen-bonding potential.

While the explanations based on the structural models seem reasonable, consideration of a bound state alone can be misleading, as the energetics of binding involves differences between bound and unbound states. Desolvation effects, in particular, have been clearly established as essential contributors to affinity and specificity [67, 68]. To assess this, the binding free energies of each complex were computed, accounting for solvent with a Poisson–Boltzmann/Surface Area (PBSA) model [69, 70] recently optimized for use with carbohydrates [71]; these energies were combined with van der Waals interactions and averaged over one thousand frames extracted from the MD simulations (the MD/PBSA approach) [54].

To assess convergence of these simulations, averages were also computed over the first and last halves of the trajectory independently, and with 5- and 10-fold coarser sampling; all these ensembles gave statistically equivalent results (see Figures 3-3 and 3-4, and Tables 3-2 and 3-3), suggesting strong convergence of the results, both in terms of length of simulation and sampling frequency.

Table 3- 2. **Convergence of semi-rigid binding free energies.**

Averaging	Sampling	α-(1,2)-Man_2		α-(1,2), α-(1,2)-Man_3	
Domain	Period	Site A	Site B	Site A	Site B
2-12 ns	200 ps	−28.24 ± 0.27	−31.14 ± 0.36	−33.21 ± 0.39	−29.10 ± 0.45
2-12 ns	100 ps	−28.13 ± 0.18	−31.32 ± 0.25	−33.41 ± 0.26	−29.14 ± 0.31
2-12 ns	20 ps	−27.98 ± 0.09	−31.31 ± 0.11	−33.34 ± 0.12	−29.14 ± 0.14
12-22 ns	200 ps	−27.92 ± 0.25	−30.65 ± 0.48	−33.81 ± 0.45	−28.84 ± 0.45
12-22 ns	100 ps	−28.03 ± 0.19	−31.18 ± 0.29	−33.92 ± 0.29	−28.92 ± 0.32
12-22 ns	20 ps	−28.01 ± 0.09	−31.20 ± 0.11	−33.99 ± 0.12	−28.96 ± 0.14
2-22 ns	200 ps	−28.08 ± 0.18	−30.89 ± 0.29	−33.51 ± 0.30	−28.97 ± 0.32
2-22 ns	100 ps	−28.08 ± 0.13	−31.25 ± 0.19	−33.66 ± 0.20	−29.03 ± 0.22
2-22 ns	20 ps	−27.99 ± 0.07	−31.26 ± 0.08	−33.67 ± 0.09	−29.05 ± 0.10

Each energy value is an average binding-free energy, computed over a set of evenly-spaced snapshots from a dynamic trajectory. Averages are given for two consecutive 10 ns intervals, as well as the combined 20 ns simulation. Sampling was done every 20, 100, and 200 ps, corresponding to 500, 100 and 50 frames per 10 ns. Errors are computed as the standard error of the mean for each set. All energies in kcal/mol.

These calculations give three energetic terms: the intermolecular van der Waals interactions made in the bound state (ΔG^{vdw}), the hydrophobic solvation energy ($\Delta G^{h\phi}$, computed as proportional to the solvent-accessible surface area buried upon binding), and the electrostatic contribution (ΔG^{elec}, including both the solvent-screened intermolecular interactions and the cost of desolvating each molecule upon binding). The sum of these terms gives what may be termed a "semirigid" binding free energy; energies are averaged over multiple conformations of both sugar and protein, but the structural ensembles of the unbound state are identical to those of the bound state.

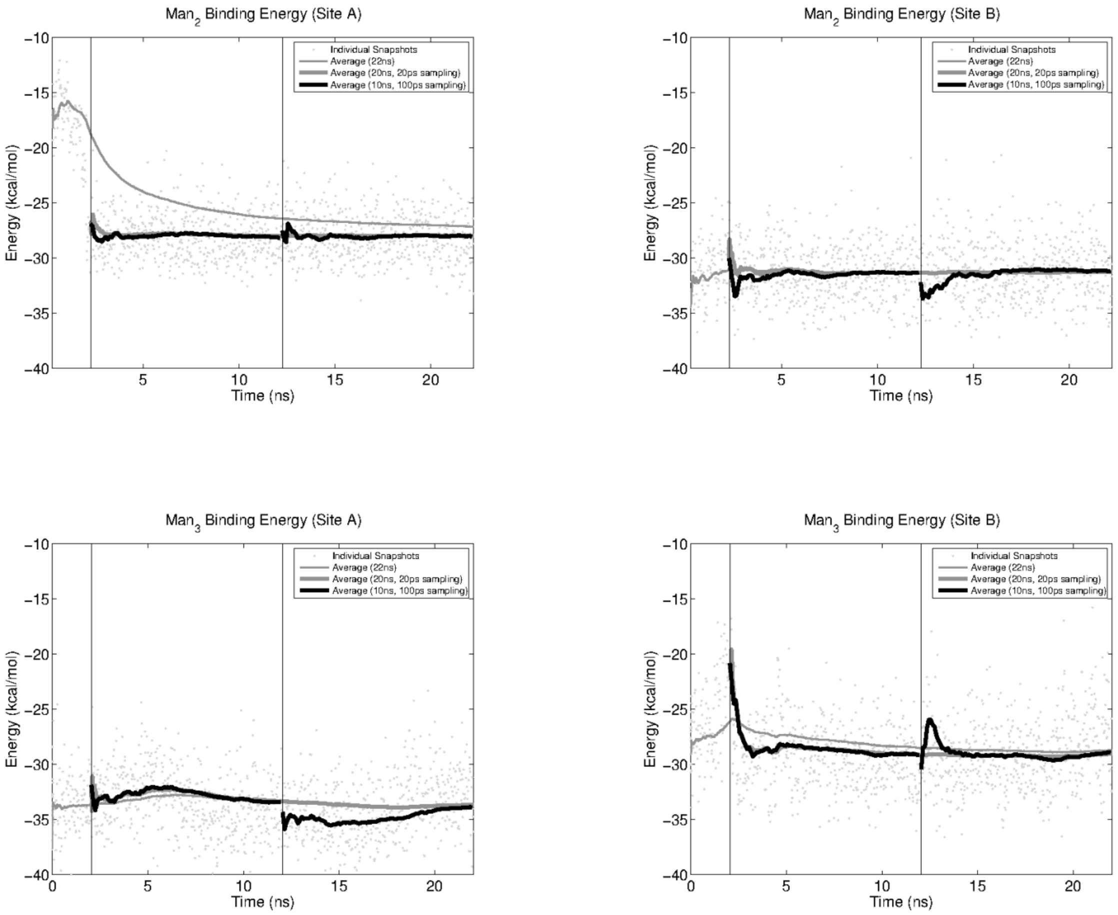


Figure 3- 3. **Convergence of computed semi-rigid binding free energies.** Grey points are the rigid-body binding free energies computed for individual snapshots, sampled every 20 ps. The grey lines show the cumulative average of these points, with the thin line including 2 ns of equilibration, and the thick line excluding this. The black lines are cumulative averages computed over each 10 ns interval, with a sampling of 100 ps.

Table 3- 3. **Convergence of configurational state free energies.**

Averaging	Sampling	α-(1,2)-Man$_2$		α-(1,2),α-(1,2)-Man$_3$	
Domain	Period	Site A	Site B	Site A	Site B
Unbound-state configurations:					
2-12 ns	200 ps	39.06 ± 0.45	39.41 ± 0.61	52.03 ± 0.70	54.94 ± 0.80
2-12 ns	100 ps	39.40 ± 0.42	39.15 ± 0.42	53.24 ± 0.57	54.79 ± 0.55
2-12 ns	20 ps	39.36 ± 0.20	39.51 ± 0.21	54.35 ± 0.25	54.45 ± 0.26
12-22 ns	200 ps	39.56 ± 0.64	39.52 ± 0.56	56.31 ± 0.73	55.09 ± 0.81
12-22 ns	100 ps	39.26 ± 0.46	38.88 ± 0.41	54.66 ± 0.54	55.39 ± 0.56
12-22 ns	20 ps	39.47 ± 0.20	39.12 ± 0.21	54.23 ± 0.25	54.19 ± 0.25
2-22 ns	200 ps	39.31 ± 0.39	39.47 ± 0.41	54.17 ± 0.55	55.01 ± 0.57
2-22 ns	100 ps	39.33 ± 0.31	39.02 ± 0.29	53.95 ± 0.40	55.09 ± 0.39
2-22 ns	20 ps	39.42 ± 0.14	39.31 ± 0.15	54.29 ± 0.18	54.32 ± 0.18
Bound-state configurations:					
2-12 ns	200 ps	40.93 ± 0.60	42.26 ± 0.57	55.52 ± 0.75	56.87 ± 0.86
2-12 ns	100 ps	41.22 ± 0.42	42.11 ± 0.37	55.60 ± 0.50	57.08 ± 0.58
2-12 ns	20 ps	41.71 ± 0.20	42.15 ± 0.19	55.87 ± 0.22	57.69 ± 0.27
12-22 ns	200 ps	41.77 ± 0.65	40.89 ± 0.72	56.36 ± 0.80	58.06 ± 0.75
12-22 ns	100 ps	41.55 ± 0.45	41.82 ± 0.47	55.94 ± 0.55	58.04 ± 0.57
12-22 ns	20 ps	41.54 ± 0.19	41.93 ± 0.20	55.71 ± 0.24	58.18 ± 0.26
2-22 ns	200 ps	41.35 ± 0.44	41.57 ± 0.46	55.94 ± 0.54	57.46 ± 0.57
2-22 ns	100 ps	41.39 ± 0.31	41.96 ± 0.30	55.77 ± 0.37	57.56 ± 0.41
2-22 ns	20 ps	41.62 ± 0.14	42.04 ± 0.14	55.79 ± 0.16	57.93 ± 0.19

Each energy value is an average binding-free energy, computed over a set of evenly-spaced snapshots from a dynamic trajectory. Averages are given for two consecutive 10 ns intervals, as well as the combined 20 ns simulation. Sampling was done every 20, 100, and 200 ps, corresponding to 500, 100 and 50 frames per 10 ns. Errors are computed as the standard error of the mean for each set. All energies in kcal/mol.

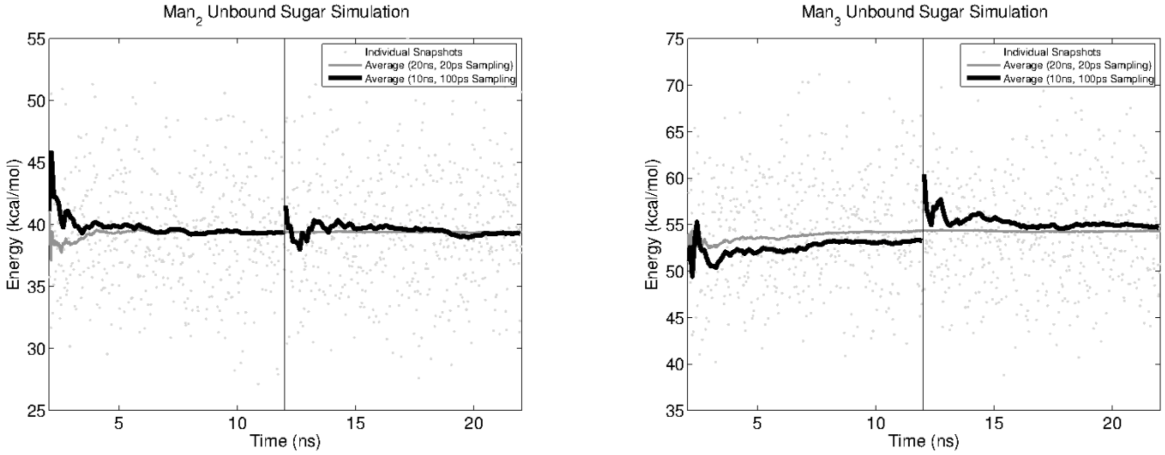


Figure 3- 4. **Convergence of computed unbound-state free energies.** Grey points are the total single-point free energies computed for individual snapshots, sampled every 20 ps over a 20 ns trajectory. The grey line shows the cumulative average of these points. The black lines are cumulative averages computed over each 10 ns interval, with a sampling of 100 ps.

The unbound ensembles will not, in actuality, be identical to those of the bound complex, and thus several additional terms contribute to the true binding free energy. Two of these are strain energies—the energetic cost of perturbing the ensemble of structures found in the unbound state into that found in the bound state—with a contribution both from the sugar and from the protein. Strain energies are easily computed by running additional simulations of the unbound state; the difference in total energy (bond, angle, and dihedral strain, intramolecular van der Waals and Coulombic interactions, and solute–solvent interactions) between the unbound ensemble and the ensemble of structures taken from the bound state (with the binding partner removed) is the strain. This contribution was computed for each oligosaccharide (ΔG^{str}); the constraints used in the simulations led to artifacts in the protein strain term, which was thus neglected (see Table 3-4). Conceptually, this approach considers binding as a two-step process, in which the ligand ensemble first is perturbed into an ensemble that is "pre-formed" for binding (the carbohydrate strain energy), followed by rigid-body binding of each member of this ensemble to the corresponding member of the protein ensemble (semirigid binding energy).

Table 3- 4. **Free energies of di- and tri-mannose binding to CVN.**

Absolute semi-rigid binding free-energies:						
	ΔG^{vdw}	$\Delta G^{h\phi}$	ΔG^{des}_{CVN}	ΔG^{des}_{carb}	ΔG^{int}	ΔG^{bnd}
Man_2, A	−19.3 ± 0.1	−2.1 ± 0.0	+8.9 ± 0.1	+11.4 ± 0.1	−27.0 ± 0.3	−28.0 ± 0.1
Man_2, B	−20.2 ± 0.1	−2.2 ± 0.0	+9.6 ± 0.2	+12.5 ± 0.1	−30.8 ± 0.4	−31.3 ± 0.1
Man_3, A	−20.6 ± 0.1	−2.7 ± 0.0	+14.8 ± 0.1	+13.5 ± 0.1	−38.8 ± 0.3	−33.7 ± 0.1
Man_3, B	−24.4 ± 0.1	−2.4 ± 0.0	+9.9 ± 0.2	+11.5 ± 0.1	−23.4 ± 0.4	−28.9 ± 0.1
Carbohydrate strain energies:						
	$\Delta G^{internal}$	ΔG^{vdw}	$\Delta G^{h\phi}$	ΔG^{des}	ΔG^{elec}	ΔG^{str}
Man_2, A	+2.0 ± 0.3	−0.3 ± 0.1	0.0 ± 0.0	−2.0 ± 0.0	+2.5 ± 0.1	+2.2 ± 0.2
Man_2, B	+2.7 ± 0.3	−0.5 ± 0.1	0.0 ± 0.0	−2.0 ± 0.0	+2.5 ± 0.1	+2.6 ± 0.2
Man_3, A	+0.9 ± 0.3	+0.1 ± 0.1	0.0 ± 0.0	−3.6 ± 0.0	+4.1 ± 0.1	+1.5 ± 0.2
Man_3, B	+3.5 ± 0.3	+0.7 ± 0.1	−0.1 ± 0.0	−1.8 ± 0.0	+1.3 ± 0.1	+3.6 ± 0.3
Relative binding free energies:						
	$\Delta\Delta G^{vdw}$	$\Delta\Delta G^{h\phi}$	$\Delta\Delta G^{elec}$	$\Delta\Delta G^{bnd}$	$\Delta\Delta G^{str}$	$\Delta\Delta G^{comp}$
Site B – Site A:						
Man_2	−0.9 ± 0.1	−0.1 ± 0.0	−2.3 ± 0.1	−3.3 ± 0.1	+0.4 ± 0.2	−2.8 ± 0.2
Man_3	−3.8 ± 0.1	+0.3 ± 0.0	+8.2 ± 0.1	+4.7 ± 0.1	+2.1 ± 0.2	+6.9 ± 0.3
Man_3 – Man_2:						
Site A	−1.3 ± 0.1	−0.7 ± 0.0	−3.7 ± 0.1	−5.7 ± 0.1	−0.7 ± 0.3	−6.4 ± 0.3
Site B	−4.2 ± 0.1	−0.2 ± 0.0	+6.8 ± 0.1	+2.3 ± 0.1	+1.0 ± 0.3	+3.3 ± 0.4

All energies are in kcal/mol, computed as ensemble averages over 20 ns trajectories with 20 ps samples. Errors are the standard error of the mean.

Additionally, there are terms related to solute entropy (solvent entropy is included implicitly in the PBSA model), both due to configuration/vibrational flexibility of each solute as well as the translational and rotational degrees of freedom of each component. Configurational entropy is most often estimated using normal modes (as in the first applications of the MM/PBSA approach) or quasi-harmonic analysis, with the latter having an advantage of being derived from the same simulation as other energetic values [72]. There are significant assumptions made in both these approaches, however, and thus the values are best considered estimates. Entropic contributions to binding from the carbohydrate estimated with a quasiharmonic analysis are roughly 0.5 kcal/mol, and contributions to relative binding free energies are even smaller (<0.2 kcal/mol, see Table 3-5).

Table 3- 5. **Entropies of di- and tri-mannose binding to CVN.**

	$-TS_{state}$			$-T\Delta S_{bind}$		$-T\Delta\Delta S_{bind}$	
	Man_2	Man_3		Man_2	Man_3		
Site A	−0.76	−1.16	Site A	+0.43	+0.64	Site B – Site A:	
Site B	−0.76	−1.24	Site B	+0.43	+0.56	Man_2	0.00
Unbound	−1.20	−1.80				Man_3	−0.08
						$Man_3 - Man_2$:	
						Site A	+0.21
						Site B	+0.13

Computed from a quasi-harmonic analysis of 20 ns trajectories with 0.5 ps sampling.

These values have been neglected in subsequent analysis, but their inclusion would not change the results in any significant way. The small contributions may initially be nonintuitive, if one expects significant flexibility in the free sugars. However, analysis of glycosidic bond dihedrals throughout the unbound simulation suggests a highly restricted conformational space, even in the unbound state (see Figure 3-5); the small additional constriction upon binding incurs a relatively small entropic penalty. Of course, it is entirely possible that more complete conformational sampling of the unbound state requires much longer timescales than considered here, a caveat in all simulations. In keeping with the semirigid approximation for the protein (to avoid artifacts from constraints), protein entropy was neglected.

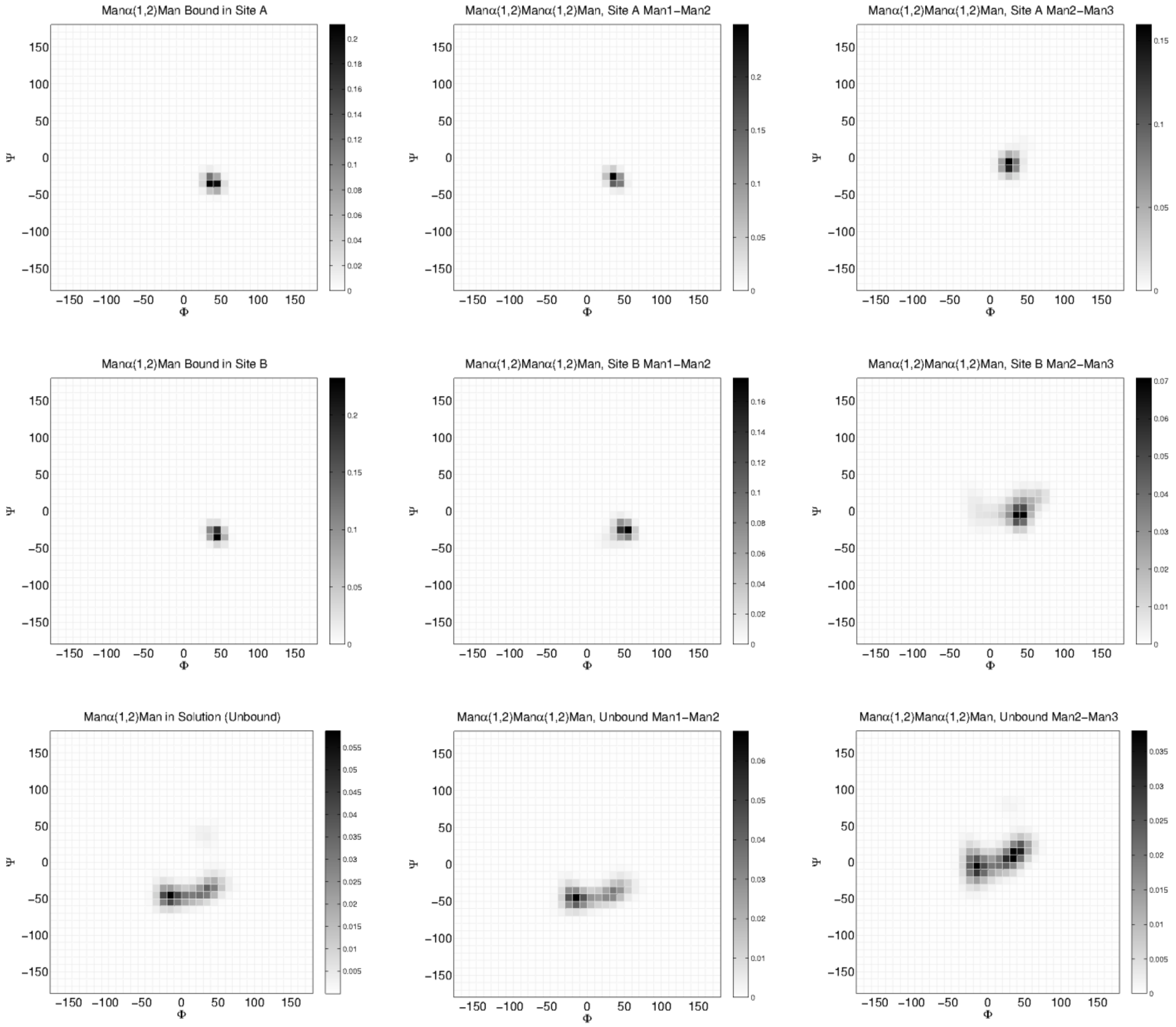


Figure 3- 5. **Glycosidic dihedral distributions for Man$_2$ and Man$_3$.** Boxes are shaded according to the fraction populated during a 20 ns simulation (0.5 ps sampling). ϕ corresponds to the C1-O2-C2-H2 dihedral; ψ corresponds to the H1-C1-O2-C2 dihedral. The first column contains values for the sole glycosidic link of Man$_2$, the second column values for the first (Man$_1$-Man$_2$) glycosidic link of Man$_3$ and the third column values of the second (Man$_2$-Man$_3$) glycosidic link of Man$_3$.

The terms neglected with a semirigid (protein) receptor model certainly contribute to the absolute binding free energies, and may contribute to relative energies as well. However, the qualitative agreement of the results with experiment suggests that these approximations are reasonable. In Site B, Man$_2$ and Man$_3$ make largely the same contacts with the protein, with the exception of a single hydrogen bond to E41; thus, the protein strain and entropy of these two states may be expected to be quite similar; in Site A, Man$_3$ makes additional contacts with the free N and C termini, and thus a larger unfavorable contribution might be expected than for Man$_2$

binding in this site, or for Man$_3$ binding in Site B. Site B additionally contains a large, flexible arginine at position 76, replaced with Thr 25 in Site A; this could lead to an increased unfavorable contribution in Site B for both sugars. Quantifying these effects is beyond the scope of this work, but the general trends should be considered in considering the following results.

The relative affinities of each state (presented in Table 3-6) are all computed with reasonable accuracy, further supporting the validity of the structural models; no attempt has been made here to modify our energetic model to give improved agreement with experiment. In all cases, the trends are computed correctly, both trimer/dimer specificity in each site and the relative affinities of either ligand in the two sites. However, the magnitude of the differences are somewhat overpredicted, a matter worthy of further consideration; for the two favorable values, the strain and entropic penalties paid by the protein on binding might be expected to modulate these toward lower magnitudes. Considering each energetic term suggests a dominant role for electrostatic interactions in defining the specificity at each site, as only this term shows the same trend as the overall free-energy differences. This is consistent with the structural analysis, in which all major differences involved electrostatic interactions.

Table 3- 6. **Relative free energies of binding.**

	$\Delta\Delta G^{vdw}$	$\Delta\Delta G^{h\phi}$	$\Delta\Delta G^{elec}$	$\Delta\Delta G^{str}$	$\Delta\Delta G^{comp}$	$\Delta\Delta G^{expt}$
ΔΔG Site A − Site B						
Man$_2$	−0.9 (0.1)	−0.1 (0.0)	−2.3 (0.1)	+0.4 (0.2)	−2.8 (0.2)	−1.5
Man$_3$	−3.8 (0.1)	+0.3 (0.0)	+8.2 (0.1)	+2.1 (0.2)	+6.9 (0.3)	+1.7
ΔΔG Man$_3$ − Man$_2$						
Site A	−1.3 (0.1)	−0.7 (0.0)	−3.7 (0.1)	−0.7 (0.3)	−6.4 (0.3)	−1.4
Site B	−4.2 (0.1)	−0.2 (0.0)	+6.5 (0.1)	+1.0 (0.3)	+3.3 (0.4)	+1.8

Van der Waals, hydrophobic surface burial, electrostatics, carbohydrate strain, total computed, and experimental [58, 59], all in kcal/mol. Errors are the standard error of the mean for the ensemble averages.

Structure-guided mutagenesis and design are important tools, both for exploring function and for the development of complexes with enhanced affinity. For these to be successful, however, accurate models are essential. The models presented here thus provide an important reference for future work on cyanovirin-N. Ensembles of snapshots from the MD trajectories, as well as a minimized representative structure of each refined complex, are available from the authors.

Additionally, the results validate the use of the MD/PBSA approach in the study of carbohydrate-binding proteins. While this method has been used extensively in the study of protein–protein and protein–small molecule interactions, to date there have been few applications to carbohydrates. This work strongly motivates the pursuit of future studies on protein–carbohydrate recognition by these approaches.

3.3 Materials and Methods

3.3.1 Construction of Increased Symmetry Binding Models

The initial structure for all calculations was the solution structure of cyanovirin-N (CVN) bound to Manα(1–2)Man (PDB 1iiy) [61]. CVN has two pseudo-symmetric binding domains; in order to construct binding models based on symmetry, the backbone atoms of equivalent residues in each site were superimposed by an RMSD fit. The coordinates of the sugar in the site were then replaced with those from the superimposed structure; visual analysis indicated that no major steric clashes were introduced by this procedure. This initial placement was then subjected to a short minimization, with all protein residues >4.0 Å from the sugar held fixed, to alleviate small clashes. These manipulations were done using the CHARMM software package (version 32B1) [32].

3.3.2 Explicit-solvent Molecular Dynamics

Molecular dynamics calculations were performed with CHARMM, using the Param22 protein force field [35], the Carbohydrate Solution Force Field (csff) [73], and a TIP3P water model [49]. A 15.0 Å radius sphere of water molecules was centered on each ligand; waters with oxygens within 2.8 Å of any solute heavy atom were then removed. After 10 ps of simulation with fixed sugar and protein atoms, these steps were repeated to fill voids in the solvent. For further dynamics, all atoms outside the sphere were fixed, and a spherical boundary potential maintained waters in the droplet. Protein atoms in the outer 2 Å of the sphere were harmonically restrained (10 kcal/mol/Å^2) and subjected to Langevin dynamics with a moderate friction coefficient (10 ps^{-1}); waters in the outer 2 Å were unrestrained, but subjected to Langevin dynamics with a high friction coefficient (62 ps^{-1}). Atoms in the inner 13 Å radius sphere were propagated with Newtonian dynamics. The SHAKE algorithm was used to fix all bonds involving hydrogens, and the time step of integration was 2 fs. The system was equilibrated for 2 ns, following which 20 ns of dynamics were collected for further study.

3.3.3 Implicit-solvent Molecular Dynamics

Implicit-solvent molecular dynamics simulations were carried out with the CHARMM software package and force field parameters as detailed above. The GBSW module was used to provide Generalized Born-based solvation terms [74], using atomic radii optimized for this approach [71, 75]. No constraints were applied, other than the use of the SHAKE algorithm to fix all bonds involving hydrogens. Simple Newtonian dynamic was used, with a time step of 2 fs.

3.3.4 Computation of Binding Free Energies

Binding free energies were computed with an MD/PBSA model [54]. The explicit-solvent MD trajectories were sampled every 20 ps, for a total of one thousand frames per system. Rigid-body binding free energies were computed for each snapshot, and the results averaged over all frames. Electrostatic contributions were computed for each snapshot as the difference between the total electrostatic free energies of the bound complex and the two unbound components. These were obtained by solution of the linearized Poisson–Boltzmann (PB) equation [69], using a multigrid finite difference solver [76] distributed with the Integrated Continuum Electrostatics (ICE) software package [77, 78]. A solute dielectric constant of 2.0 and a solvent dielectric constant of 80.0 were used; the dielectric boundary was defined by the molecular surface using a 1.4 Å radius probe, with radii optimized for this purpose [71, 75]. The ionic strength was set to 0.145 M, with a 2.0 Å ion-exclusion layer. A 65^3-unit grid was used with overfocusing boundary conditions (the longest dimension of the molecule occupying first 23%, then 92%, and then 184% of one edge of the grid). Boundary potentials extracted from the previous calculation were used in each case, with Debye–Hückel boundary potentials used for the initial calculation. Energetic contributions from atoms falling outside the finest grid were taken from the middle-resolution grid. The total binding energy for each snapshot was the sum of the electrostatic contribution, the intermolecular van der Waals energy and a term proportional to the solvent accessible surface area buried on binding. The area was computed with CHARMM, using a 1.4 Å probe radius, and the energetic contribution was given by $\Delta G^{h\phi} = 0.005\Delta A + 0.86$ kcal/mol [70].

3.3.5 Computation of Sugar Strain Energies

Sugar strain energies were computed by comparing the total ensemble-averaged energies of sugar conformations extracted from bound-state molecular dynamics and those for conformations from unbound simulations. Electrostatic solvation free energies were computed with a Poisson–Boltzmann model, as the difference between a system with solvent dielectric constant of 80 (ionic strength of 0.145 M) and solute dielectric constant of 2, and a system of uniform dielectric constant of 2 (zero ionic strength). For these calculations, the largest grid contained the whole sugar (92% of the longest dimension). Hydrophobic solvation free energies were estimated with a term proportional to the total solute surface area, as described above. These energies were added to the molecular-mechanics energy of the solute, including all bonded terms (bonds, angles, dihedrals), intramolecular van der Waals, and intramolecular Coulombic (in uniform dielectric of 2) interactions. The sugar strain energy is given by the difference in the ensemble average of the total energy of the sugar in conformations extracted from the complex simulation and the similar average for conformations from a simulation of the free sugar. Thus, these values correspond to the energetic cost of perturbing the unbound conformational ensemble into the ensemble that is capable of binding, in a fully solvated context. Sugar entropies were computed from the same trajectories using the QUASI module of the CHARMM software package [72], using the average structure as a reference.

3.4 Conclusion

The results validate the use of the MD/PBSA approach in the study of carbohydrate-binding proteins. While this method has been used extensively in the study of protein–protein and protein–small molecule interactions, to date there have been few applications to carbohydrates. This work strongly motivates the pursuit of future studies on protein–carbohydrate recognition by these approaches. In the next chapter, the same methods will be implemented to other saccharide models bound to CVN.

Chapter 4

Carbohydrate Recognition by the Antiviral Lectin Cyanovirin-N

This chapter corresponds to a manuscript in preparation.

Author contributions: YKF performed research and analyzed data; YKF and DFG wrote the manuscript.

Abstract

Cyanovirin-N is a cyanobacterial lectin with potent antiviral activity, and has been the focus of extensive pre-clinical investigation as a potential prophylactic for the prevention of the sexual transmission of the human immunodeficiency virus (HIV). Here we present a detailed analysis of carbohydrate recognition by this important protein, using a combination of computational methods, including extensive molecular dynamics simulations and Molecular-Mechanics/Poisson−Boltzmann/Surface-Area (MM/PBSA) energetic analysis. The simulation results strongly suggest that the observed tendency of wildtype CVN to form domain-swapped dimers is the result of a previously unidentified *cis*-peptide bond present in the monomeric state. The energetic analysis additionally indicates that the highest-affinity ligand for CVN characterized to date (α-Man-(1,2)-α-Man-(1,2)-α-Man) is recognized asymmetrically by the two binding sites. Finally, we are able to provide a detailed map of the role of all binding site functional groups (both backbone and side chain) to various aspects of molecular recognition: general affinity for cognate ligands, specificity for distinct oligosaccharide targets and the asymmetric recognition of α-Man-(1,2)-α-Man-(1,2)-α-Man. Taken as a whole, these results

complement past experimental characterization (both structural and thermodynamic) to provide the most complete understanding of carbohydrate recognition by CVN to date. The results also provide strong support for the application of similar approaches to the understanding of other protein−carbohydrate complexes.

4.1 Introduction

Nearly 25 years ago, Lifson and colleagues, identified mannose-specific carbohydrate binding proteins (most notably legume lectins such as Concanavalin A) as potential inhibitors of human immunodeficiency virus (HIV) viral cell fusion [56]. The outer envelope glycoprotein of HIV (gp120) is heavily glycosylated, with between 20 and 28 *N*-linked glycosylation sites occupied in various natural viral isolates; glycosylation consists of both high-mannose and complex oligosaccharide subtypes, in roughly equal proportions [21]. Antiviral lectins act by binding these carbohydrates and thus interfering with some aspect of cellular recognition and/or membrane fusion; the specific mechanism of inhibition may involve either direct blocking of CD4 receptor or CXCR4/CCR5 co-receptor binding, or interference in required conformational changes associated with fusion [12, 79, 80].

More recently, numerous lectins from diverse sources have been found to have antiviral activity, many with much greater potency and specificity than the plant lectins. As a therapeutic target, the carbohydrates of the viral envelope are attractive for several reasons. First, the glycosylation plays an essential role in helping the virus avoid detection by the humoral immune response. The carbohydrates are added by the enzymatic systems of the host cell, and reduced levels of glycosylation have been associated with increased susceptibility to immune-system recognition; inhibitors that work through specific interactions with the sugars may thus be less-susceptible to the evolution of viral resistance [57]. Secondly, the remarkable density of carbohydrates on the surface of gp120 distinguishes it from naturally-occurring human glycoproteins, which typically have many fewer sites of glycosylation, and thus non-specific interactions may be avoided. Finally, while the lack of oral bioavailability of protein therapeutics can make small molecule drugs more attractive, topical application as a prophylactic virucide does not suffer from this concern [81, 82]. Pre-clinical trials for use of lectins as a topical agent to prevent sexual transmission of HIV in simian models have shown significant promise [64, 65].

Beyond the inhibition of cellular infection by HIV, protein−carbohydrate interactions play key roles in a vast array of biology. Many human retroviruses are heavily glycosylated in much the same manner as HIV, and, in fact, lectins with anti-HIV activity are often active against a diverse range of viruses; including the Ebola and Herpes Simplex Virus [83, 84]. Additionally, bacterial pathogens often display cell-surface carbohydrates distinct from those on eukaryotic glycoproteins, and many cancer cells are characterized by unique glycosylation patterns. Thus, specific, high-affinity lectins could have potential application both as antimicrobial agents as well as for cancer-cell targeting.

Despite their importance, the study of protein−carbohydrate interactions has greatly lagged that of other biomolecular complexes. In particular, many computational methods that have shown great success in understanding protein−protein, protein−nucleic acid, and protein−small molecule interactions, have seen only limited application to carbohydrate recognition. As a good example, the combination of explicit solvent molecular dynamics with Poisson−Boltzmann-based calculations of free energies has been demonstrated as a powerful approach for predicting relative binding free energies and for dissecting energies into contributions from individual chemical groups [85−90]. As well as facilitating the understanding of natural systems, computational approaches provide unique opportunities for molecular design; as a perfect example, detailed studies on protein systems, coupled with robust computational models and innovative algorithms, have made the computational design and engineering of proteins and their complexes a reality [91−96]. Robust protocols for the simulation of protein−carbohydrate systems thus promise to be an enabling technology, opening up a wide range of potential applications to these important systems.

Among the best characterized of the antiviral lectins is cyanovirin-N (CVN), originally isolated from the cyanobacterium, *Nostoc ellipsosporum* [26]. Under physiological conditions, CVN is a small, monomeric protein, containing two pseudo-symmetric binding domains with differing affinities for specific oligosaccharides, but that both bind exclusively to sugars containing an α-(1,2)-linked mannobiose substructure [58−61]. Interestingly, at high concentrations CVN has a tendency to form domain-swapped dimers, and this form is the only state found in the crystal phase [62, 63]; mutations that preferentially stabilize both the monomeric state and the dimer have been identified. Thermodynamic, structural, and *in vivo* efficacy studies have all suggested preferential recognition of the D1 arm of high mannose sugars [58, 59, 63], and multivalent interactions appear necessary for antiviral potency. Despite the apparent wealth of data for this system, a complete understanding of the mechanism of CVN's antiviral activity has been hindered by the inherent complexities in both synthetic chemistry and structural biology of carbohydrates. Among the questions that remain to be answered are: what are the determinants of specific carbohydrate recognition by CVN, what is the mechanism by which multivalent interactions by CVN lead to potent antiviral activity, and can CVN be engineered into a better virucidal agent? Here we present a comprehensive answer to the first of these questions, building on a computational framework for modeling protein−carbohydrate interactions, which we have previously demonstrated as a promising approach [90]. The success of computational methods in explaining the structural determinants of carbohydrate recognition provides strong motivation for the use of similar approaches to address the remaining questions.

The minimal carbohydrate recognized by CVN is an α-(1,2)-linked mannobiose, a moiety found on the terminal arms of high-mannose, *N*-linked oligosaccharides. A structure of the largest of these (Man-9) is shown in Figure 4-1, with α-Man-(1,2)-Man moieties highlighted (red); all high-mannose oligosaccharides are derived from Man-9 by varying degrees of glycolytic processing of one or more of the three arms (D1−3). The thermodynamics of binding of trisaccharides corresponding to each of the three arms to the antiviral lectin Cyanovirin-N has been previously characterized by Bewley and co-workers, who demonstrated that the two CVN binding sites show distinct specificities for these targets [59].

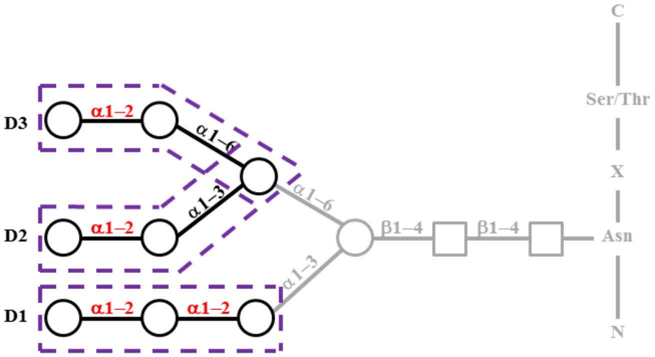


Figure 4- 1. **A representative structure of a high mannose oligosaccharide (Man$_9$GlcNAc$_2$).** The arms are labelled D1–3. The glycosidic linkages labelled in red represent the α-(1,2) dimannose common to the tip of all three arms. The trimannose structures were computationally built by extending the dimannose anchor by one unit (dashed box).

Structural insight into carbohydrate recognition by CVN has been limited by experimental challenges in working with oligosaccharide ligands, as well as the intrinsic propensity of CVN to form domain-swapped dimers when crystalized [62]. An NMR-derived structure of monomeric CVN in complex with mannobiose has been solved, [61] as have crystal structures of the domain-swapped dimer bound to Man-9 and to a smaller Man-9 fragment, Man-6 [63]. However, in the domain-swapped dimer structure, only a single binding site is occupied, and much of the sugar is poorly resolved. The work here provides a mechanistic look between the structural and thermodynamical studies, both by molecular dynamics and continuum electrostatic analysis.

4.2 Results

4.2.1 Structural Models of Trisaccharide Recognition

Initial models of the three trisaccharides highlighted in Figure 4-1 (purple dashed boxes) were constructed from the structure of CVN in complex with α-Man-(1,2)-α-Man by extending the dimannose by one monosaccharide; the starting disaccharide structures for this procedure were computationally-refined models based on the solution (NMR) structure, as previously described [90]. To construct an α-(1,2)-α-(1,2)-linked trimannose with this procedure, one may consider either the (new) monosaccharide or the (existing) disaccharide to be the reducing sugar. The two approaches result in chemically identical trisaccharides, but different bound state geometries; both of these were considered. Each trisaccharide was modeled in the two binding sites both independently and in a doubly-bound form; each model was then subjected to explicit-solvent molecular-dynamics simulation.

4.2.2 Backbone Fluctuations in Domain B Binding Site Loop

In our first set of simulations, domain A behaved stably and gave consistent results (both in structural and energetic terms) in the doubly and singly-bound models, while domain B showed a fundamental lack of stability, with the carbohydrate dissociating from the protein in many cases. Visual analysis localized the structural plasticity of domain B to the “hinge” region of residues 50 through 54 (Figure 4-2). In the starting NMR structure (Figure 4-3, structure A), the backbone amide protons S52 and N53 clash (1.56 Å H−H distance), resulting in a highlystrained configuration; during the simulations, two alternate conformations were sampled that relieved this strain. In all of the simulations, a crank-shaft-like rotation of the peptide bond between S52 and N53 was observed (Figure 4-3, structure B: the movement involves a concerted rotation of the ϕ dihedral of S52 and the ψ dihedral of N53). While the energetic barrier between this state and the starting structure was low (as characterized by a 60 ps average lifetime of the initial state), it was not particularly stable; in several cases transitions back to the initial state were observed, and visual analysis linked sugar dissociation to this state.

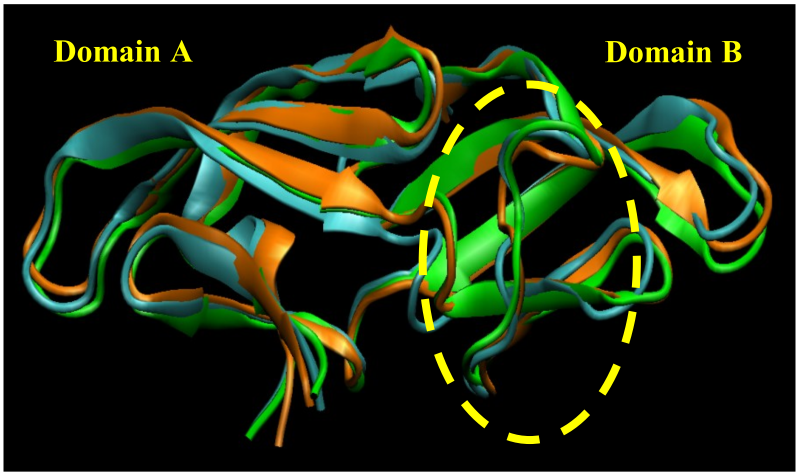


Figure 4- 2. **Overlay image of three snapshots of CVN (no sugar is present in image).** The hinge region (residues 50−54) is highlighted by a dashed circle. Snapshot from simulation of a doubly bound CVN with Manα(1–2)Man (orange); snapshot from a simulation of a singly bound CVN with Manα(1–2)Man (cyan); and a simulation from an unbound CVN (green).

To further characterize this alternate conformation, a second set of simulations were carried out, beginning with trisaccharide structures modeled into the alternate protein structure. These simulations showed only moderate backbone flexibility in the hinge region, with roughly 60° fluctuations in the ϕ dihedral of S52 and the ψ dihedral of N53, but in every case the sugar disassociated within 100 ns. Visual analysis of the structures provides a clear rationale for this behavior—the backbone carbonyl of N53 makes a hydrogen-bonded interaction with the hydroxyl at position 3 of the non-reducing sugar in the NMR structure, but this is lost (and replaced by an unfavorable interaction with the backbone amide proton) in the alternate configuration.

4.2.3 Identification of *Cis* Peptide Bond in Domain B

In a single simulation, a second transition—involving a *trans*- to *cis*-isomerization of the peptide bond between P51 and S52—was observed (Figure 4-3, structure C). Unsurprisingly, the barrier to this transition is very high, and only a single transition was seen in a total of about 3 μs of simulation. However, the structure after the transition remained stable for the duration of the simulation (100 ns). In order to further assess the stability of this conformer, models of all trisaccharides were again constructed in this background, and subjected to the same simulation protocol. In every case, the structure remained remarkably stable, with only thermal fluctuations around a single structure in the hinge region. Additionally, all oligosaccharide ligands remained stably bound throughout the simulations, which have been carried out to beyond 200 ns.

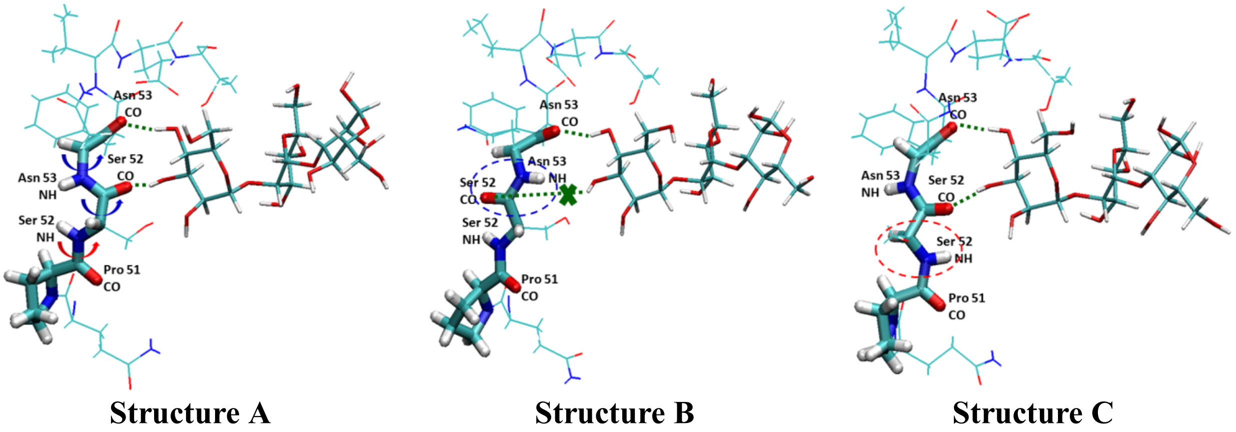


Figure 4- 3. **Alternate backbone configurations in hinge region.** (A) Monomeric NMR structure of wt CVN (PDB: 1IIY) (B) Crank-shaft movement of the middle region (C) *Cis*-trans isomerization movement in the lower region.

4.2.4 Calculation of Trisaccharide Binding Free Energies

Simulations of each sugar (doubly-bound, as well as singly-bound in each site) were carried out for 200 ns from a *cis*-peptide-containing starting structure, and rigid-body binding free energies were computed for 1500 evenly spaced (every 100 ps) snapshots using an MM-PBSA model. In order to minimize potential bias from the initial structure, the first 50 ns of each simulation were excluded from further analysis (see Appendix A for time dependence of computed energies). When average binding energies computed this way (over the last 150 ns of simulation) were compared to those computed over the entire 200 ns simulation, the effect was less than 1 kcal/mol for 87.5% (14/16) of the simulations (and less than 0.5 kcal/mol in 75% (12/16)); in the remaining two cases (both for Domain B), excluding the first 50 ns reduced the computed binding free energies by 1.2 kcal/mol. Overall binding free energies (ΔG^{bind}) were further broken down into contributions from electrostatics (ΔG^{elec}), from bound-state van der Waals interactions (ΔG^{vdW}), and from non-polar solute-solvent interactions ($\Delta G^{h\phi}$). The electrostatic term includes contributions from loss of favorable interactions with solvent in the bound state (relative to the unbound state), as well as from solvent-screened Coulombic interactions made in the bound state; the hydrophobic solvation term is directly proportional to the solvent-accessible surface area buried upon binding. For all computed energies, results are presented as ensemble-averaged values over all 1500 snapshots; autocorrelation analysis (Appendix A) suggests that binding free energies computed from all these snapshots are statistically independent.

Contributions from individual side chains, backbone amino and backbone carbonyl groups (or groupings of these components), ΔG^{group}, were similarly computed from ensemble averages of electrostatic, van der Waals, and non-polar solvation terms. The electrostatic term was computed as the difference in computed binding free energy between the native sequence and a hypothetical mutant in which only the groups under consideration have been mutated to hydrophobic isosteres; this has been termed the "mutation energy" in previous work on continuum electrostatic component analysis. The van der Waals term is simply the sum of the bound-state (intermolecular) van der Waals interactions made by the atoms of the group of interest, and the hydrophobic solvation term is directly related to the change in solvent-accessible surface area of the group. It should be noted that while the van der Waals and surface-area dependent terms are additive within a group of multiple components, the electrostatic term is not; simply adding the electrostatic terms of individual components would double-count any electrostatic interactions between them.

4.2.5 Asymmetric Recognition of D1-arm Trisaccharides in Domains A and B

Table 4-1 details the computed binding free energies for α-Man-(1,2)-α-(1,2)-α-Man (the trisaccharide corresponding to the D1 arm of high-mannose oligosaccharides) in each of two possible binding orientations discussed above; we term the orientation in which the anomeric hydroxyl of the original disaccharide is untouched as the "internal" binding mode, and that where this hydroxyl is linked to the third sugar as the "terminal" binding mode. In domain A, the internal binding mode is preferred by about 3.0 kcal/mol; in domain B, on the other hand, it is the terminal mode that is preferred (by upwards of 8 kcal/mol). In both cases, the differences result entirely from the electrostatic contribution to the binding free energies.

Table 4-1. **Asymmetric recognition of α-Man-(1,2)-α-(1,2)-α-Man.**

	ΔG^{bind}	=	ΔG^{elec}	+	ΔG^{vdw}	+	$\Delta G^{h\phi}$
Domain A							
Terminal	–31.7 (0.1)		–7.4 (0.1)		–21.8 (0.1)		–2.5 (0.0)
	–30.5 (0.1)		–6.5 (0.1)		–21.5 (0.1)		–2.5 (0.0)
Internal	–34.3 (0.1)		–10.0 (0.1)		–21.6 (0.1)		–2.7 (0.0)
	–34.0 (0.1)		–10.6 (0.1)		–20.7 (0.1)		–2.7 (0.0)
Domain B							
Terminal	–37.5 (0.1)		–10.9 (0.1)		–24.1 (0.1)		–2.6 (0.0)
	–35.7 (0.1)		–9.4 (0.1)		–23.7 (0.1)		–2.6 (0.0)
Internal	–28.1 (0.1)		–0.9 (0.1)		–24.7 (0.1)		–2.5 (0.0)
	–27.5 (0.1)		+0.2 (0.1)		–25.2 (0.1)		–2.6 (0.0)

All energies are in kcal/mol; for each value, the results of both singly and doubly-bound simulations are provided as the first and second row, respectively.

The determinants of these differences can be narrowed down to a small number of key interactions that differ between the two binding modes in each site; these are detailed in Table 4-2. In domain A, glutamate 101 makes a strongly favorable interaction with the last sugar in the internal binding mode; the charged N-terminus similarly makes a more moderately favorable interaction in this state, while both interactions are absent when the trisaccharide is bound in the terminal mode. In domain B, on the other hand, only slightly favorable interactions with glutamine 50 are made in the internal mode, and the free N-terminus is replaced with (non-interacting) proline 51. Additionally, glutamate 41 in domain B makes favorable interactions in the context of terminal-mode binding, but unfavorable interactions (for a net difference of over 5 kcal/mol) in the internal mode. Finally, both threonine 25 (domain A) and its corresponding residue in domain B (arginine 76) make more favorable interactions in the terminal model. For T25 this is primarily a van der Waals effect, and serves to slightly offset the enhanced stabilization of the internal mode in domain A by E101 and the N-terminus. For R76 on the other hand, a combination of both electrostatic and van der Waals interactions act to stabilize the terminal mode in domain B, reinforcing the destabilization of the internal mode by E41. It should be noted, however, in the context of larger oligosaccharides, interactions between T25 and R76 are lost with additional carbohydrates.

Table 4- 2. **Energetic determinants of internal vs. terminal recognition.**

	$\Delta\Delta G^{group}$	$\Delta\Delta G^{elec}$	$\Delta\Delta G^{vdw}$	$\Delta\Delta G^{h\phi}$
L1/P51 Backbone Amino				
A (L1) Terminal	−0.3/ −0.2	−0.1/ 0.0	−0.2/ −0.2	0.0/ 0.0
A (L1) Internal	−1.5/ −1.6	−1.4/ −1.4	−0.1/ −0.1	0.0/ −0.1
B (P51) Terminal	0.0/ 0.0	0.0/ 0.0	0.0/ 0.0	0.0/ 0.0
B (P51) Internal	0.0/ +0.1	+0.1/ +0.2	−0.1/ −0.1	0.0/ 0.0
E101/Q50 Side Chain				
A (E101) Terminal	−0.2/ −0.4	−0.1/ −0.3	−0.1/ −0.1	0.0/ 0.0
A (E101) Internal	−4.0/ −4.3	−5.5/ −6.1	+1.5/ +1.8	0.0/ 0.0
B (Q50) Terminal	−0.1/ −0.1	0.0/ 0.0	−0.1/ −0.1	0.0/ 0.0
B (Q50) Internal	−0.7/ −0.8	−0.5/ −0.6	−0.2/ −0.2	0.0/ 0.0
A92/E41 Side Chain				
A (A92) Terminal	−0.4/ −0.4	+0.1/ +0.1	−0.5/ −0.5	0.0/ 0.0
A (A92) Internal	−0.8/ −0.8	+0.1/ +0.1	−0.9/ −0.9	0.0/ 0.0
B (E41) Terminal	−2.1/ −2.0	−2.6/ −2.4	+0.6/ +0.5	−0.1/ −0.1
B (E41) Internal	+3.0/ +3.8	+4.8/ +6.0	−1.8/ −2.0	0.0/ −0.2
T25/R76 Side Chain				
A (T25) Terminal	−2.4/ −2.5	−0.4/ −0.4	−1.8/ −1.8	−0.2/ −0.3
A (T25) Internal	−1.2/ −1.4	−0.1/ −0.2	−1.0/ −1.1	−0.1/ −0.1
B (R76) Terminal	−9.5/ −9.3	−4.3/ −4.1	−4.8/ −4.8	−0.4/ −0.4
B (R76) Internal	−3.8/ −4.5	−0.8/ −1.0	−2.9/ −3.3	−0.1/ −0.2
Subtotal				
A Terminal	−3.1/ −3.2	−0.3/ −0.3	−2.6/ −2.6	−0.2/ −0.3
A Internal	−5.7/ −6.2	−5.1/ −5.7	−0.5/ −0.3	−0.1/ −0.2
B Terminal	−11.5/ −11.2	−6.7/ −6.3	−4.3/ −4.4	−0.5/ −0.5
B Internal	−0.9/ −0.8	+4.2/ +5.2	−5.0/ −5.6	−0.1/ −0.4
Remainder				
A Terminal	−28.6/ −27.3	−7.1/ −6.2	−19.2/ −18.9	−2.3/ −2.2
A Internal	−28.6/ −27.8	−4.9/ −4.9	−21.1/ −20.4	−2.6/ −2.5
B Terminal	−26.1/ −24.5	−4.2/ −3.1	−19.8/ −19.3	−2.1/ −2.1
B Internal	−27.2/ −26.8	−5.1/ −5.0	−19.7/ −19.6	−2.4/ −2.2

All energies are in kcal/mol; for each entry, the two values given are those from the singly and doubly-bound simulations, respectively.

When the contributions of these four groups are taken together (Table 4-2, Subtotal), they account almost entirely for the differences between the two binding modes in both sites. While a significant fraction of the overall binding energies (about 27 kcal/mol) come from other groups

(Table 4-2, Remainder), the contribution of these remaining groups is near equal in both binding sites and with both binding modes. Figure 4-4 shows the structural view of these determinants.

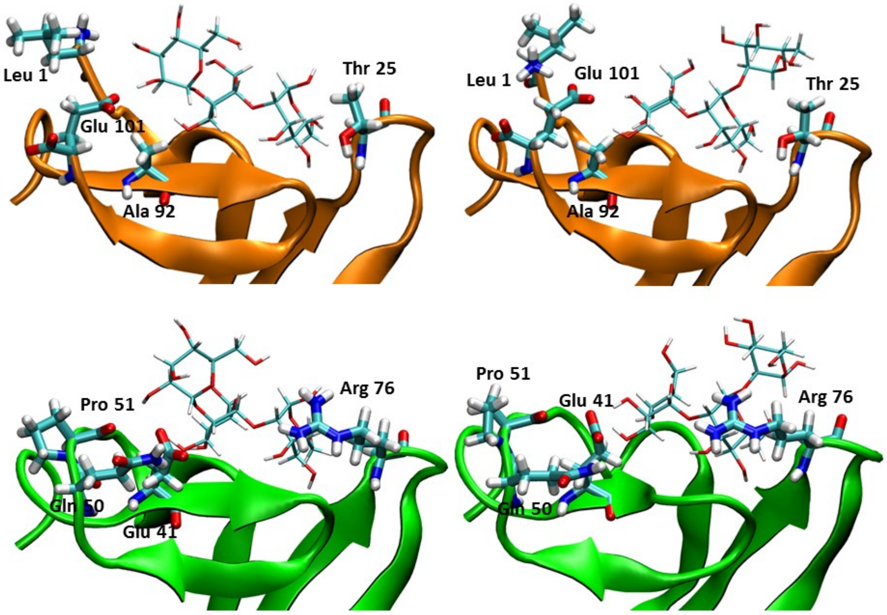


Figure 4- 4. **Structural view of the energetic determinants of internal vs. terminal recognition.** Residues in domain A are found in the orange ribbon structures and residues in domain B are found in the green ribbon structures.

4.2.6 Near Quantitative Prediction of Relative Binding Free Energies

Table 4-3 shows the computed affinities for each of the three sugars in each binding site; for α-Man-(1,2)-α-Man-(1,2)-α-Man the values given are for the preferred binding mode in each site. The two computed values for each sugar/binding site combination (from singly and doubly-bound simulations) are strongly consistent, with the calculations agreeing to within 1.0 kcal/mol for all but one case. As the structure is quite rigid, and the binding sites are separated by approximately 40 Å, no cooperativity in binding is expected.

Figure 4-5 shows these same values plotted against the experimentally-determined binding free energies of Bewley and co-workers [59]. The computed energies show remarkable correlation with the experimental values for five of the six cases; these data lie along a best-fit line with slope of 0.77. Interestingly, the electrostatic component correlates even more strongly with the experimental values, with a best-fit slope of 0.91, and deviation of the single outlier from this line is less than for the total energy. This outlier, α-Man-(1,2)-α-(1,2)-α-Man bound in domain B, also shows the least consistency between the values computed for the singly and doubly bound states, with a difference of 1.8 kcal/mol.

Table 4- 3. **Overall energetics of trisaccharide binding.**

	ΔG^{bind}	=	ΔG^{elec}	+	ΔG^{vdw}	+	$\Delta G^{h\phi}$
Domain A							
α-(1,2), α-(1,2), α-Man$_3$[b]	−34.3 (0.1)		−10.0 (0.1)		−21.6 (0.1)		−2.7 (0.0)
	−34.0 (0.1)		−10.6 (0.1)		−20.7 (0.1)		−2.7 (0.0)
α-(1,2), α-(1,3), α-Man$_3$	−29.5 (0.1)		−5.7 (0.1)		−21.4 (0.1)		−2.4 (0.0)
	−30.2 (0.1)		−6.9 (0.1)		−21.0 (0.1)		−2.4 (0.0)
α-(1,2), α-(1,6), α-Man$_3$	−30.3 (0.1)		−7.3 (0.1)		−20.6 (0.1)		−2.4 (0.0)
	−31.3 (0.1)		−7.2 (0.1)		−21.5 (0.1)		−2.5 (0.0)
Domain B							
α-(1,2), α-(1,2), α-Man$_3$	−37.5 (0.1)		−10.9 (0.1)		−24.1 (0.1)		−2.6 (0.0)
	−35.7 (0.1)		−9.4 (0.1)		−23.7 (0.1)		−2.6 (0.0)
α-(1,2), α-(1,3), α-Man$_3$	−33.2 (0.1)		−8.0 (0.1)		−22.7 (0.1)		−2.4 (0.0)
	−32.2 (0.1)		−7.6 (0.1)		−22.2 (0.1)		−2.4 (0.0)
α-(1,2), α-(1,6), α-Man$_3$	−32.5 (0.1)		−7.6 (0.1)		−22.4 (0.1)		−2.5 (0.0)
	−33.0 (0.1)		−8.2 (0.1)		−22.2 (0.1)		−2.6 (0.0)

All energies are in kcal/mol; for each value, the results of both singly and doubly-bound simulations are provided as the first and second row, respectively. [b] For Domain A, the α-(1,2), α-(1,2)-linked sugar is bound in the internal orientation; all other sugars are bound in the terminal mode.

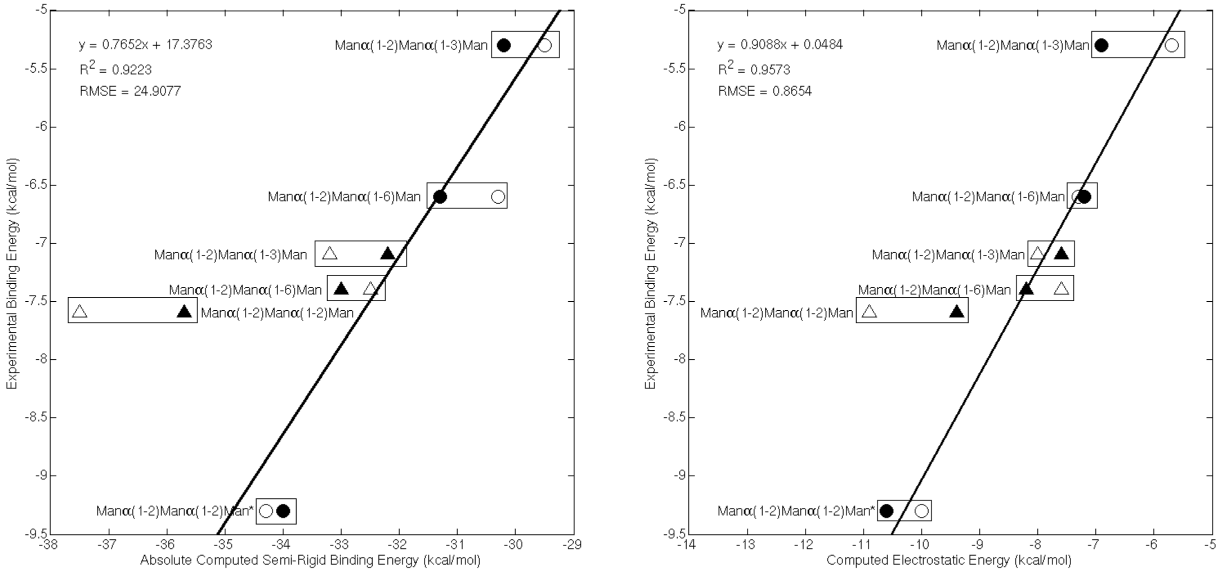


Figure 4- 5. **Comparison of computed and experimental binding free energies.** Open shapes represent simulations from singly bound models and closed shapes represent simulations from doubly bound models; circles represent domain A region and triangles represent domain B; black line represents the linear best-fit line for trisaccharide data (excluding the terminal orientation of Manα-(1,2)Manα-(1,2)Man).

4.2.7 Conserved Backbone Interactions are Primary Source of Affinity

A large fraction of the binding affinity for all three sugars in both sites can be attributed to a common set of interactions, primarily involving backbone contacts; the contributions of these groups are detailed in Table 4-4 and Figure 4-6. Overall, these groups account for roughly −20 kcal/mol of overall affinity, with about −16 kcal/mol of this from electrostatics and −4 kcal/mol from van der Waals contacts. In terms of electrostatic contributions, this is a more favorable contribution than the total overall electrostatic contribution to affinity. For van der Waals interactions, on the other hand, these groups contribute a much smaller fraction.

Table 4- 4. **Energetic determinants of general affinity for cognate sugars.**

	$\Delta\Delta G^{group}$	$\Delta\Delta G^{elec}$	$\Delta\Delta G^{vdw}$	$\Delta\Delta G^{h\phi}$
Subtotal Backbone				
Domain A (2−4)	[−6.5, −5.8]	[−6.5, −5.2]	[−1.3, +0.7]	[−0.1, 0.0]
Domain B (52−54)	[−5.3, −5.1]	[−4.8, −4.4]	[−0.7, −0.4]	[0.0, 0.0]
Domain A (23−24)	[−4.7, −4.1]	[−3.8, −2.9]	[−1.2, −0.8]	[0.0, 0.0]
Domain B (74−75)	[−5.0, −3.6]	[−4.3, −2.6]	[−1.1, −0.7]	[0.0, 0.0]
Domain A (92−95)	[−9.9, −9.5]	[−7.8, −7.2]	[−2.7, −1.8]	[−0.1, 0.0]
Domain B (41−44)	[−10.5, −8.4]	[−8.1, −5.5]	[−2.9, −2.2]	[−0.1, 0.0]
Q6/E56 Side Chain				
Domain A (Q6)	[−1.2, −1.1]	[−0.1, −0.1]	[−1.0, −0.9]	[−0.1, 0.0]
Domain B (E56)	[−1.1, −0.9]	[−0.5, −0.3]	[−0.7, −0.5]	[−0.1, 0.0]
T7/T57 Side Chain				
Domain A (T7)	[−3.7, −3.4]	[−1.6, −1.5]	[−2.0, −1.7]	[−0.2, −0.2]
Domain B (T57)	[−3.6, −3.2]	[−1.7, −1.4]	[−1.7, −1.4]	[−0.2, −0.1]
Subtotal Side Chain				
Domain A (6−7)	[−4.9, −4.5]	[−1.8, −1.6]	[−3.0, −2.6]	[−0.3, −0.3]
Domain B (56−57)	[−4.6, −4.2]	[−2.1, −1.8]	[−2.3, −2.1]	[−0.3, −0.2]
Subtotal				
Domain A (Overall)	[−24.9, −23.7]	[−17.9, −17.2]	[−7.4, −5.7]	[−0.5, −0.3]
Domain B (Overall)	[−24.6, −20.6]	[−18.6, −13.5]	[−6.8, −5.3]	[−0.3, −0.2]

All energies are in kcal/mol; for each entry, the range of values seen over all simulations is given.

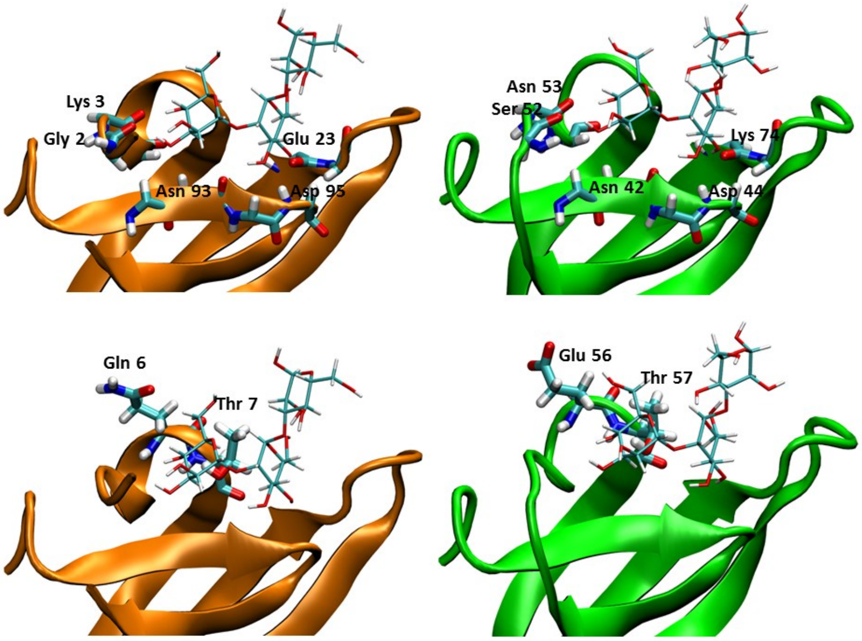


Figure 4- 6. **Structural view of the energetic determinants of general affinity for cognate sugars.** Residues in domain A are found in the orange ribbon structures and residues in domain B are found in the green ribbon structures.

4.2.8 A Small Number of Residues Mediate Affinity Differences Between Domains

A small number of groups make consistent interactions within each domain, but show significant differences between the two binding sites (Table 4-5 and Figure 4-7). One of these is Ala92/Glu41, previously noted to contribute significantly to the differences in the terminal versus internal binding orientation of α-Man-(1,2)-α-(1,2)-α-Man; in Domain B, Glu 41 makes strongly favorable electrostatic interactions will all sugars (bound in a terminal orientation), while Ala 92 in domain A contributes almost nothing to the affinity. Similarly, Gln78 in Domain B makes a moderate electrostatic interaction, as well as favorable van der Waals contact, for a net favorable contribution of about 1 kcal/mol; the corresponding amino acid in Domain A is Gly 27, with no side chain to make interactions with the sugar. On the other hand, Glu 23 in Domain A makes much stronger electrostatic interactions (about −3.0 kcal/mol) than does the equivalent residue in Domain B (Lys 74, between −0.6 and −0.9 kcal/mol), although these differences are modulated somewhat by opposing differences in van der Waals contact.

Table 4- 5. **Determinants of differences in affinity for terminal mode.**

	$\Delta\Delta G^{group}$	$\Delta\Delta G^{elec}$	$\Delta\Delta G^{vdw}$	$\Delta\Delta G^{h\phi}$
A92/E41 Side Chain				
Domain A (A92)	[−0.4, −0.3]	[+0.1, +0.1]	[−0.5, −0.4]	[0.0, 0.0]
Domain B (E41)	[−2.2, −1.9]	[−2.7, −2.4]	[+0.8, +0.5]	[−0.1, −0.1]
G27/Q78 Side Chain				
Domain A (G27)	[0.0, 0.0]	[0.0, 0.0]	[0.0, 0.0]	[0.0, 0.0]
Domain B (Q78)	[−1.3, −1.2]	[−0.7, −0.5]	[−0.8, −0.4]	[−0.1, 0.0]
E23/K74 Side Chain				
Domain A (E23)	[−2.7, −2.5]	[−3.1, −2.8]	[+0.5, +0.3]	[−0.1, 0.0]
Domain B (K74)	[−1.9, −1.7]	[−0.9, −0.6]	[−1.1, −1.0]	[0.0, 0.0]
L1/P51 Backbone Carbonyl				
Domain A (L1)	[−1.8, −0.2]	[−1.9, +0.1]	[−0.3, +0.1]	[0.0, 0.0]
Domain B (P51)	[−0.2, −0.1]	[−0.1, 0.0]	[−0.1, −0.1]	[0.0, 0.0]
Subtotal				
Domain A	[−4.8, −3.2]	[−4.9, −2.7]	[−0.5, +0.2]	[−0.1, 0.0]
Domain B	[−5.5, −5.1]	[−4.2, −3.9]	[−1.3, −1.0]	[−0.2, −0.1]

All energies are in kcal/mol; for each entry, the range of values seen over all simulations is given

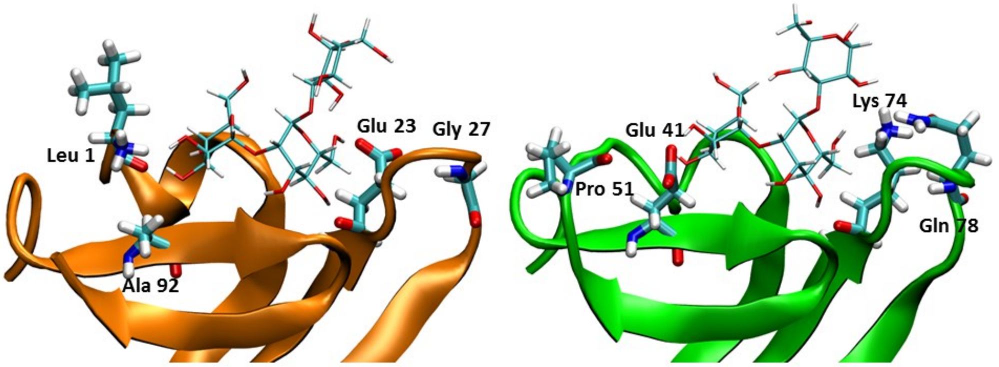


Figure 4- 7. **Structural view of the energetic determinants of differences in affinity for terminal mode.** Residues in domain A are found in the orange ribbon structures and residues in domain B are found in the green ribbon structures.

4.2.9 One Pair of Interacting Residues Dominate Specific Sugar Recognition in Domain B

A single pair of interacting residues (Arg 76 and Asp 44) in Domain B make strong contributions to binding that are not seen in Domain A (Table 4-6). In all cases, the side chain of Arg 76 interacts with the "third" mannose that extends from the core dimannose binding site. In particular, Arg 76 makes a strong favorable electrostatic and van der Waals interactions, especially with the α-Man-(1,2)-α-(1,2)-α-Man model. In this model, Arg 76 makes this interaction because Asp 44 makes a salt bridge interaction with Arg 76. These residues interact for about 80% of the simulation, while in the α-Man-(1,2)- α- (1,3)- α-Man it was about 40% and α-Man-(1,2)-α-Man-(1,6)-α-Man was about 20%. In the other two models, Arg 76 alone would interact with the carbohydrate and Asp 44 would move away towards the solvent, and vice versa.

Table 4- 6. **Interactions of the R76/D44 salt-bridge and its equivalent.**

	$\Delta\Delta G^{group}$	=	$\Delta\Delta G^{elec}$	+	$\Delta\Delta G^{vdw}$	+	$\Delta\Delta G^{h\phi}$
R76 Side Chain (B)							
α-(1,2), α-(1,2), α-Man$_3$	−9.5/−9.3		−4.3/−4.1		−4.8/−4.8		−0.4/−0.4
α-(1,2), α-(1,3), α-Man$_3$	−6.0/−5.1		−1.7/−1.4		−4.1/−3.5		−0.2/−0.2
α-(1,2), α-(1,6), α-Man$_3$	−4.6/−7.3		−0.7/−2.4		−3.7/−4.4		−0.2/−0.5
D44 Side Chain (B)							
α-(1,2), α-(1,2), α-Man$_3$	+0.4/ 0.0		+1.0/+0.4		−0.6/−0.4		0.0/ 0.0
α-(1,2), α-(1,3), α-Man$_3$	+0.0/−0.8		+0.5/−0.8		−0.5/ 0.0		0.0/0.0
α-(1,2), α-(1,6), α-Man$_3$	−0.9/−3.8		−0.8/−4.7		−0.1/+0.9		0.0/0.0
Domain B Subtotal							
α-(1,2), α-(1,2), α-Man$_3$	−8.2/−8.4		−2.4/−2.8		−5.4/−5.2		−0.4/−0.4
α-(1,2), α-(1,3), α-Man$_3$	−5.2/−4.7		−0.4/−1.0		−4.6/−3.5		−0.2/−0.2
α-(1,2), α-(1,6), α-Man$_3$	−4.6/−8.9		−0.6/−4.9		−3.8/−3.5		−0.2/−0.5
T25 Side Chain (A)							
α-(1,2), α-(1,2), α-Man$_3^b$	−1.2/−1.4		−0.1/−0.2		−1.0/−1.1		−0.1/−0.1
α-(1,2), α-(1,3), α-Man$_3$	−2.1/−2.4		−0.2/−0.5		−1.7/−1.7		−0.2/−0.2
α-(1,2), α-(1,6), α-Man$_3$	−2.4/−2.0		−0.4/0.0		−1.7/−1.8		−0.3/−0.2
D95 Side Chain (A)							
α-(1,2), α-(1,2), α-Man$_3^b$	+0.2/+0.3		+0.7/+0.8		−0.5/−0.5		0.0/0.0
α-(1,2), α-(1,3), α-Man$_3$	+0.5/+0.4		+1.0/+0.9		−0.5/−0.5		0.0/0.0
α-(1,2), α-(1,6), α-Man$_3$	+0.4/+0.6		+0.9/+1.0		−0.5/−0.4		0.0/0.0
Domain A Subtotal							
α-(1,2), α-(1,2), α-Man$_3^b$	−1.1/−1.3		+0.5/+0.4		−1.5/−1.6		−0.1/−0.1
α-(1,2), α-(1,3), α-Man$_3$	−1.9/−2.3		+0.5/+0.1		−2.2/−2.2		−0.2/−0.2
α-(1,2), α-(1,6), α-Man$_3$	−2.2/−1.7		+0.3/+0.7		−2.2/−2.2		−0.3/−0.2

All energies are in kcal/mol; for each entry, the two values given are those from the singly and doubly-bound simulations, respectively. [b] For Domain A, the α-(1,2), α-(1,2)-linked sugar is bound in the internal orientation; all other sugars are bound in the terminal mode.

4.3 Discussion

4.3.1 Monomeric CVN Contains a *Cis*-Peptide Bond in the Hinge Region

One of the most significant results of the simulations presented here is the strong suggestion that monomeric Cyanovirin-N contains a *cis*-peptide bond in the hinge region. As *cis*-peptide bonds are rare, these observations warrant some discussion. First, in the original NMR structure determination, minimal constraints were available in this region, and thus the backbone structure of the hinge was not rigorously defined (C. Bewley, personal communication); it is not surprising that, in absence of explicit restraints, annealing would not sample the *cis*-conformer. Secondly, the existence of a P51−S52 *cis*-peptide bond provides a structural and energetic explanation for an interesting observation involving a CVN variant containing a Pro to Gly mutation at position 51, previously characterized by Gronenborn and colleagues. This mutation has been shown to preferentially stabilize the monomeric form of CVN; the monomeric melting temperature is increased by 10 degrees, and the P51G mutant crystalizes exclusively as a monomer. Typically, mutations to glycine destabilize proteins, unless they are made in regions requiring violations of the typically allowed region of Ramachandran space, due to increased entropy in the unfolded state. Taken as a whole, these observations lead us to the conclusion that wild-type, monomeric CVN likely contains a *cis*-peptide bond between Pro51 and Ser52.

4.3.2 The D1 Arm is Asymmetrically Recognized by Domains A and B

Oligosaccharide recognition by CVN involves very interesting differences in the preferences of each domain for binding α-(1,2)-linked mannobiose and a trimannose with two α-(1,2)-linkages; Bewley and co-workers observed that while the disaccharide binds domain B tighter (by roughly 10-fold in affinity) than domain A, the trisaccharide preferentially binds domain A by a similar amount. The simulations presented here provide direct insight into this specificity-switch, with an interesting structural mechanism.

The α-(1,2), α-(1,2) trimannose could potentially bind in one of two binding modes both maintaining the α-(1,2)-linked mannobiose in the same orientation—with the third sugar extending either from the reducing or non-reducing end of the disaccharide. In the context of a larger oligosaccharide, the first would correspond to recognition of the last two sugars on the D1 arm as the mannobiose anchor, and thus we term this the “terminal” binding mode. The second, on the other hand, would correspond to recognition of the second and third sugars from tip of D1 as the anchor, and thus this mode is termed the “internal” orientation. In the crystal structure of Man-9 bound to the domain-swapped dimer, only site A is occupied, and an internal binding mode is observed.

Consistent with the crystal structure, the calculations indicate a strong preference of domain A for the internal binding mode, but notably show an even stronger bias against the internal mode in domain B; the net binding energy is 3.5 kcal/mol more favorable for the internal mode in site A and over 8 kcal/mol less favorable for the internal mode in domain B. The source of these preferences are entirely electrostatic in nature, with differences in buried surface nearly zero and differences in van der Waals interactions slightly opposing the net difference. In domain A, the internal binding mode makes strongly favorable electrostatic interactions (−10.6 kcal/mol), compared with more moderately favorable interactions (−6.5 kcal/mol) for the terminal mode. In site B, net electrostatic contributions are reduced from a (favorable) contribution of −9.4 kcal/mol for the terminal binding mode to essentially zero (+0.2 kcal/mol) for the internal mode. The results are entirely consistent between both doubly- and singly-bound models. These data are calculated based on semi-rigid binding free energies, averaging energies over 150 bound-state configurations, but with no explicit consideration of the unbound states. However, in comparing binding modes for a single ligand, the unbound state is identical, and thus no error arises from this neglect. Additionally missing is an assessment of bound state configurational entropy, of both the protein and carbohydrate.

The origins of this asymmetry in recognition have a clear basis in structure. In site A, several groups make favorable interactions in the internal binding mode that are absent (or weaker) in the terminal binding mode. Chief among these is the side chain of Glu101, which makes a strongly favorable interaction with the hydroxyl at position C2 of the third monosaccharide in the internal mode (and makes no interactions in the terminal mode). Additionally, the free N-terminus makes a moderately favorable interaction with the hydroxyl at C2.

4.3.3 Determinants of Binding Affinity and Specificity

Overall, all the trisaccharide models in both sites captured several determinants that contributed to the binding affinity and specificity. In particular, a large portion of the affinity in both sites (about −24 kcal/mol) came from backbone contacts making interactions with the core α-Man-(1,2)-Man disaccharide. Only two side chain groups (Domain A: Gln6 and Thr7; Domain B: Glu56 and Thr57) make interactions with the same disaccharide. In this case, the electrostatic energies (about −17 kcal/mol) dominated the binding affinities in both sites, while the van der Waals energies make moderate interactions (about −7 kcal/mol).

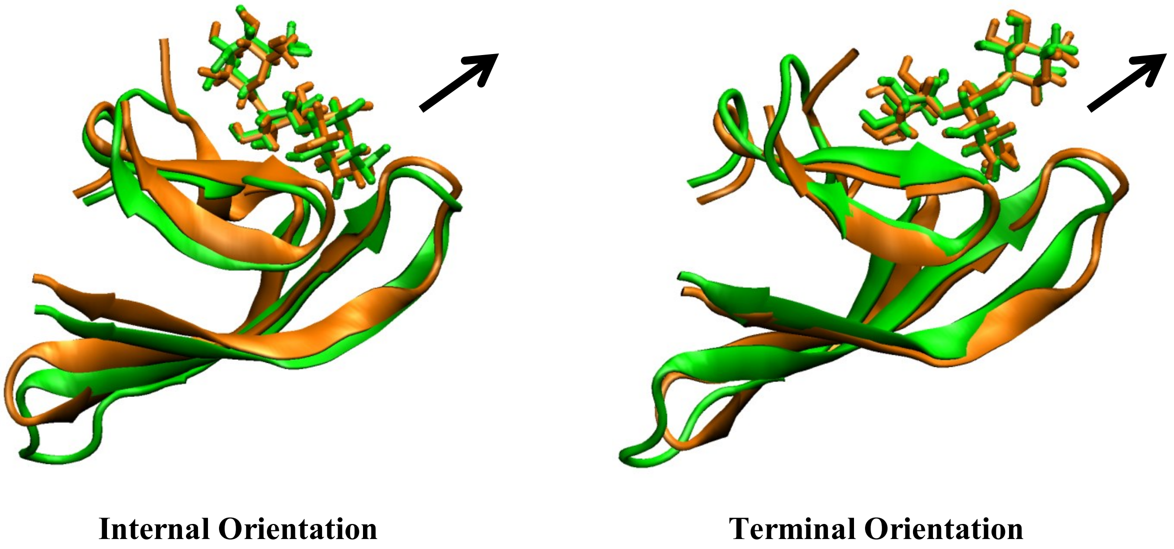
A few residues make interactions within each site; however, its corresponding residue in the other site shows energetic differences. These residues include: side chain groups of Ala 92/Glu 41, Gly 27/Gln 78, and Glu 23/Lys 74. Glu 41 and Gln 78 make interactions with the carbohydrates in Domain A, but its corresponding residues in Domain A, Ala 92 and Gly 27, do not make interactions. In Domain A, Glu23 made more overall favorable energy than its corresponding residue in Domain B, Lys 74.

Of all the residues that make interactions with the carbohydrates, only one pair appears to dominate energetically in Domain B, Arg 76 and Asp 44. When there is a salt-bridge interaction being made between Asp 44 and Arg 76, it makes a strong interaction with the third mannose of the trisaccharide models. This is strongly seen in the α-Man-(1,2)-α-Man-(1,2)-Man model. This is less seen in the other trisaccharide models. In Domain A, its corresponding residues, Thr 25 and Asp 95, make a moderate contribution to the overall energetics.

4.4 Methods

4.4.1 Construction of CVN Complexes with Trisaccharides

The initial structure for all simulations originated from the last snapshot taken from an earlier study where the solution NMR structure of CVN bound to α-Man-(1,2)-α-Man (PDB 1IIY) was modified and simulated in a droplet of water [90]. Briefly, the backbone atoms of equivalent residues in each site were superimposed (by minimizing the backbone heavy-atom root-mean square deviation, RMSD), and the coordinates of the sugar in domain A were then replaced with those from the superimposed structure to construct binding models of increased symmetry.



Internal Orientation **Terminal Orientation**

Figure 4- 8. **Internal vs. terminal anchor placement in Manα-(1,2)Manα-(1,2)Man.** A glycosidic bond can be formed either with OH2 of the dimannose anchor (internal) or with OH1 (terminal). The crystal structure of Man9 bound in domain A [63] suggests that the internal orientation is preferred for domain A; no data is available for domain B. Energetically, the internal orientation is preferred in domain A and the terminal orientation is preferred in domain B. The black arrows represent the direction of the Man$_9$GlcNAc$_2$ structure.

Structures of three distinct trimannoses representing the three arms of Man9 (α-Man-(1,2)-α-Man-(1,2)-α-Man, α-Man-(1,2)-α-Man-(1,3)-α-Man, and α-Man-(1,2)-α-Man-(1,6)-α-Man) were built by extending our dimannose model by one unit from the anomeric carbon of the reducing sugar. In addition, an alternate structure of α-Man-(1,2)-α-Man-(1,2)-α-Man was constructed by extending the dimannose by one unit from C2 of the non-reducing sugar (Figure 4-8). All these manipulations were done using the CHARMM software package, [33] and default conformations were used for the newly-built portions of each molecule. A short minimization (100 steps) was performed on all the newly built structures to avoid any clashes. Three models were constructed in each case, two 1:1 (protein:sugar) complexes with a single sugar bound to each of the two binding sites, as well as a 1:2 complex with both binding sites occupied.

4.4.2 Molecular Dynamics Simulations

Explicit-solvent molecular dynamics simulations were performed using the CHARMM [33] and NAMD [97] computer programs, using PARAM22 (protein) [35] and CSFF (carbohydrate) [73] parameter sets and the TIP3P water model [49]. Pre- and post-processing of all complexes was done with CHARMM, while production simulations were done using NAMD. Each complex was solvated in a box of water with a minimum of 10 Å between any solute atom and the box edge in all directions. Randomly selected water molecules were replaced with about 14 sodium and 11 chloride ions, the number of ions added was chosen to match physiological ionic strength (145 mM) and to obtain a net zero charge for the system (CVN has a formal charge of −3.0e). A total of 200 ns was simulated for each complex (using a 2 fs time step) under NPT ensemble conditions (P=1 atm, T=300 K) with periodic boundary conditions and particle-mesh Ewald (PME) for long-range electrostatics. Short-range interactions were cut off at 12 Å, and bonds involving hydrogens were held fixed using SHAKE.

4.4.3 Calculation of Binding Free Energies

Binding free energies were computed with a Molecular Mechanics/Poisson−Boltzmann Surface Area (MM/PBSA) model. The first 50 ns of each simulation were excluded from the analysis to ensure adequate equilibration of each system. The explicit-solvent MD trajectories were sampled every 100 ps, for a total of 1500 snapshots per trajectory. The total binding energy for each snapshot was computed as the sum of a Poisson−Boltzmann-based electrostatic contribution (ΔG^{elec}), the intermolecular van der Waals energy (ΔG^{vdW}), and a term proportional to the solvent accessible surface area buried on binding; the area was computed with CHARMM, using a 1.4 Å probe radius, and the energetic contribution was given by ($\Delta G^{h\phi}$) = $0.005\Delta A+0.86$ kcal/mol for each snapshot. These energies were then averaged all frames (Tables 4-1 and 4-3). Standard errors of the mean were computed using an effective number of independent frames; this value was extracted from an autocorrelation analysis of the energetic time scales.

To assess convergence of these simulations, we first looked at ensembles from domain A. The slope of the running averages were nearly at zero after 50 ns. For consistency, 50 ns was the cut-off mark in domain B.

4.4.4 Continuum Electrostatic Calculations

The linearized Poisson−Boltzmann equation was solved using a multi-grid finite-difference solver distributed with the ICE software suite (courtesy of B. Tidor), using standard protocols [76, 77, 78]. Charges were taken from the PARAM22 and CSFF parameter sets for consistency with the molecular dynamics simulations. The dielectric boundary was set as the molecular surface generated with a 1.4 Å radius probe, and a 2.0 Å ion exclusion layer was used; the surfaces were generated using radii optimized specifically for use in continuum electrostatic calculations [71, 75]. The internal and external dielectric constants were set to 2 and 80, respectively, and the (monovalent) ionic strength was set to 145 mM. Boundary conditions were computed using a 3-step focusing procedure on a 129^3-unit cubic grid, with the molecule occupying first 23%, then 92%, and finally 184% of the grid. Boundary conditions at each level were taken from the previous calculation, with Debye−Hückel potentials used at the boundary of the lowest level. The highest-resolution grid was centered on the oligosaccharide, and potentials at atoms falling off this grid were taken from the middle-resolution calculation. Electrostatic contributions to the binding free energies were computed as the sum of a desolvation penalty for both the protein and the sugar and a bound-state, solvent-screened interaction.

4.4.5 Component Analysis

Group-wise component analysis was done using the Integrated Continuum Electrostatic (ICE) package (courtesy of B. Tidor), [77, 78] again using standard protocols [55, 99. 100]. Each protein residue was partitioned into three groups—backbone carbonyl, backbone amino and side chain—and the sugars were partitioned into one group per hydroxyl. For each group, the desolvation penalty, indirect (intramolecular) interactions, and direct (intermolecular) interactions were computed. The sum of these is the mutation energy, equivalent to the difference in energy between the natural system and a hypothetical mutant with that group replaced with a hydrophobic isostere (in the context of all other groups in their natural state). Additionally, the solvent-accessible surface area (and corresponding energy) of each group was computed, as were the pair-wise intermolecular van der Waals interaction energies between all groups.

4.5. Conclusion

Current force-fields are able to reasonably reproduce affinity differences in sugar binding by the two domains of CVN, in the context of an MM/PBSA model with optimized radii for the continuum electrostatic calculations. We have been able to explain a number of important features of this protein. First among these is that it seems likely that CVN contains a rare *cis*-peptide bond in the monomeric state, which may contribute to its tendency to form domain-swapped dimers under some conditions. Additionally, we have demonstrated that the two domains of CVN, while highly homologous in sequence, recognize certain oligosaccharide targets with distinct binding modes. The results from component analysis identified several key interactions throughout all the carbohydrate models.

However, challenges remain for addressing the sugar models that make non-conserved contacts with the protein. As the most notable outlier, the computed binding free energy of Manα-(1,2)Manα-(1,2)Man in Domain B deviates significantly from the linear fit, being computed as overly favorable (shown in Figure 4-5). In this case, it appears to be more favorable due to one pair dominating energetically in Domain B—Arg 76 and Asp 44. After calculating the side chain entropies, one of the largest differences observed came from Arg 76 in the doubly and singly bound simulations of the Manα-(1,2)Manα-(1,2)Man in Domain B. This suggests a possible entropic penalty missing from the computed Manα-(1,2)Manα-(1,2)Man binding energy. In addition, just using bound state energies for van der Waals contributions can lead to size-dependent bias. In the following chapter, we will discuss the inclusion of solute-solvent vdW interactions (with a continuum approach) that improves the predictive power of the calculations. Furthermore, we will discuss the contribution from entropy when specific interactions affect the bound-state flexibility of the protein or the carbohydrate, as well as calculating carbohydrate strain energies.

Chapter 5

Improving the Accuracy of Computational Models of Protein−Carbohydrate Complexes

Molecular recognition is extremely important for biological function. It involves specific, noncovalent formation of receptor−ligand, antigen−antibody, protein−DNA, and protein−carbohydrate complexes, which are crucial for normal cell functioning. Therefore, accurate prediction of binding affinities is very helpful. Thus far, we have used large-scale molecular dynamics simulations, coupled with energetic analysis; and a residue-by-residue decomposition of the binding free energies to identify key determinants of affinity and specificity. In the last decades, a variety of methods have been developed to calculate affinities. Naturally, the goal is to obtain the highest possible accuracy with the least amount of effort. With the aim to improve efficiency and maintain accuracy, we recently started to focus on further improvements to capture the energetics of carbohydrate binding. The three areas we will be focusing on are: solute−solvent van der Waals interaction energy, conformational entropy, and strain energies.

5.1 Solute−solvent van der Waals Interaction Energy

An important step in computational modelling is the calculation of binding free energies. Hydration plays an important role in every process occurring in aqueous solution and has an effect on the thermodynamic processes involving the breakage and formation of noncovalent bonds. Explicit solvent models provide the most detailed description of the hydration phenomena. However, these models are computationally demanding because of the large number of atoms involved. Implicit solvent models offer an attractive alternative to explicit solvent models. In a typical solvent model, the solvation free energy is decomposed into a nonpolar and an electrostatic component (Equation 5.1) [101].

$$\Delta G^{solv} = \Delta G^{polar} + \Delta G^{nonpolar} \qquad (5.1)$$

The nonpolar component corresponds to the free energy of hydration of the uncharged solute and the electrostatic component corresponds to the free energy of turning on the solute's partial charges. The electrostatics is calculated either with Poisson–Boltzmann or Generalized Born continuum models. The structure and properties of proteins in water are highly influenced by hydrophobic interactions [102]. Hydrophobic interactions also play a key role in the mechanism of ligand binding to protein [103].

The nonpolar solvation free energy is often computed from a linear relationship between the nonpolar free energy and the solute surface area (Equation 5.2),

$$\Delta G^{nonpolar} = \gamma A + b \qquad (5.2)$$

where A is the solute surface area; γ is the surface tension proportionality constant, which represents the contribution to the solvation free energy per unit surface area; and b is the free energy of hydration for a point solute. The value of the surface tension constant with the nonpolar solvation surface area models vary. They can range from 5 cal/mol/Å^2 to 138 cal/mol/Å^2 because there are various definitions of solute surface area (*e.g.* van der Waals surface, molecular surface, or solvent accessible surface).

The solvation properties of hydrophobic groups are determined by the volume and shape of the solvent volume and van der Waals interaction with the solvent. The hydrophobic nonpolar hydration free energy is shown as (Equation 5.3),

$$\Delta G^{np} = \Delta G^{cav} + \Delta G^{vdW} \qquad (5.3)$$

where ΔG^{cav} is the cavity hydration free energy, defined as the hydration free energy due to excluded volume effects, and ΔG^{vdW} is the free energy of the solute−solvent van der Waals dispersion interactions. The cavity formation depends on the size and shape of the cavity, which includes parameters such as volume and surface area. The solute solvent dispersion term depends on the density, location, and the nature of the solute atoms that are placed in the cavity.

In this section, the focus will be on the solute–solvent dispersion term. A variety of work has been done indicating deficiencies in the surface-area approach to the nonpolar term of solvation [104, 105, 106, 107, 108]. One limitation of a surface-area approach is that atoms that contribute little or no solvent-exposed surface area can in fact interact favorably with solvent. Another limitation is that surface-area approaches tend to use a single energy per unit area independent of atom type, or a parameterized value for each of hydrophilic and hydrophobic surface area, but do not account for the full chemical diversity of atom types and their different parameterizations within molecular mechanics force fields.

Levy *et al.* [104] found an efficient approach in a continuum solvent model to calculate the solute–solvent van der Waals dispersion energy which is able to reproduce the results from explicit solvent simulations. The vdW interactions with solvent can be computed as a sum of integrals over the solvent region, where each integral models the vdW energy to one atom in a solute molecule [109, 110, 111]. In the next section, we introduce the continuum van der Waals model followed by examples of showing how the continuum van der Waals changes the calculated outcome.

5.1.1 Continuum van der Waals Model

Levy *et al* [104] expresses the solute–solvent van der Waals interaction energy as a volume integral over the solvent (Equation 5.4),

$$U_{vdW} = \sum_{i=1}^{n} \int_{solvent} \rho_w u_{vdW}^{(i)} \left(|r' - r_i|\right) dr' \tag{5.4}$$

where *n* is the number of solute atoms, ρ_w is the bulk number density of water, $u_{vdW}^{(i)}(r)$ is van der Waals potential due to atom *i* at distance *r* from the solute's i^{th} atomic center r_i. The van der Waals potential is modeled using the Lennard-Jones 6-12 function (Equation 5.5),

$$u_{vdW}^{(i)}(r) = \frac{A^{(i)}}{r^{12}} - \frac{B^{(i)}}{r^6} \tag{5.5}$$

The solvent interface that is used comes from Lee and Richards definition of the solvent-accessible surface as the boundary of a union of spheres (shown in Figure 5-1). Each sphere corresponds to one atom in the solute molecule, and each sphere radius is equal to the corresponding atom's vdW radius plus the radius of a spherical probe molecule. This defines how close the probe center can approach the solute atoms. This surface definition is chosen because the van der Waals potential is defined as a function of the distance between atom centers. The boundary between protein and solvent is taken as the solvent-accessible surface, which is the closest approach of the center of a probe sphere rolled over the protein set of spherical atoms. However, the radius of this probe sphere (commonly taken to be 1.4 Å to approximate a water molecule) is left as a free parameter. This parameter of the continuum van der Waals interaction model is used to balance the assumptions that the solvent region is of constant density and approaches the protein to a discrete, solvent-accessible surface.

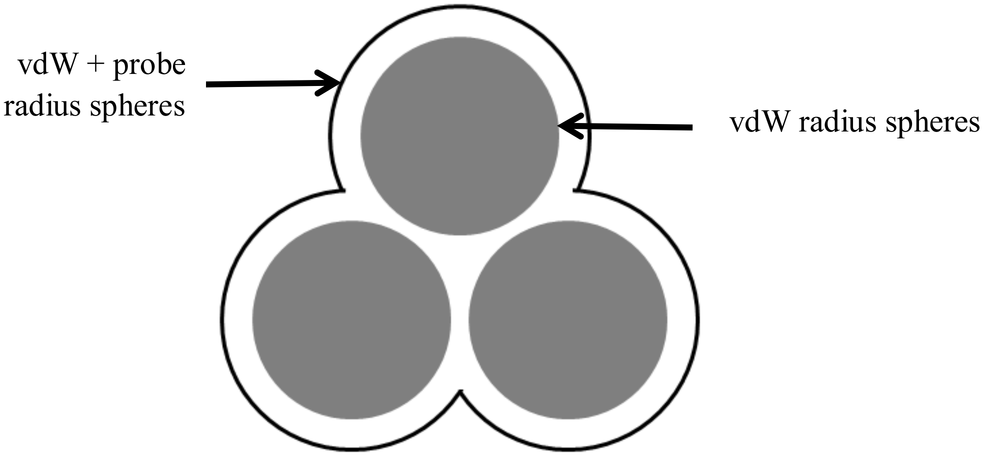


Figure 5- 1. **Illustration of the solvent-accessible surface as the boundary of a union of spheres.**

5.1.2 Application of Continuum van der Waals Calculations to CVN Complexes

The continuum van der Waals formulation was tested on all CVN systems bound to various saccharide models using the PARAM22 parameters, a TIP3P water model, and a molecular surface probe radius of 1.4 Å. The calculations were performed on 150 snapshots. Each snapshot was separated 1 ns apart. Table 5-1 shows a list of continuum van der Waals calculations computed from the protein and carbohydrate components of the CVN complex, the intermolecular van der Waals calculations taken from each simulation, and the sum between the intermolecular and continuum van der Waals calculations. The values in the $cvdW^{prot}$ and $cvdW^{carb}$ are positive because in the unbound state, the protein and sugar make more favorable van der Waals interactions with solvent. The magnitude is slightly higher in the $cvdW^{carb}$ compared to the $cvdW^{prot}$ because there are more interactions made between solvent and the carbohydrate molecule than with the protein.

Table 5- 1. **Decomposition of the continuum van der Waals interaction energies.**

	cvdW^{prot}	cvdW^{carb}	cvdW^{total}	Inter. vdW	Sum
Domain A					
α-(1,2)Man$_2$	+7.0/ +7.1	+12.5/ +12.4	+19.5/ +19.5	−18.8/ −18.9	+0.7/ +0.6
α-(1,2), α-(1,2)Man$_3$	+9.4/ +9.4	+14.6/ +14.7	+24.0/ +24.1	−21.6/ −20.7	+2.4/ +3.4
α-(1,2), α-(1,3)Man$_3$	+8.5/ +8.6	+13.6/ +13.6	+22.1/ +22.2	−21.4/ −21.0	+0.7/ +1.2
α-(1,2), α-(1,6)Man$_3$	+8.6/ +8.8	+13.5/ +13.7	+22.1/ +22.5	−20.6/ −21.5	+1.5/ +1.0
Domain B					
α-(1,2)Man$_2$	+7.2/ +7.0	+12.3/ +12.4	+19.5/ +19.4	−19.2/ −20.6	+0.3/ −0.6
α-(1,2), α-(1,2)Man$_3$	+9.1/ +9.3	+13.8/ +13.7	+22.9/ +23.0	−24.1/ −23.7	−1.2/ −0.7
α-(1,2), α-(1,3)Man$_3$	+8.9/ +9.0	+13.7/ +13.6	+22.6/ +22.6	−22.7/ −22.2	−0.1/ +0.4
α-(1,2), α-(1,6)Man$_3$	+9.1/ +9.3	+13.7/ +13.7	+22.8/ +23.0	−22.4/ −22.2	+0.4/ +0.8

All energies are in kcal/mol; for each entry, the two values given are those from the singly and doubly-bound simulations, respectively.

As shown in Figure 5-2 (left panel), the computed dimannose energies show a destabilization with a constant offset of approximately 4.5 kcal/mol from the best-fit line, suggesting a systematic error for the smaller ligands. As the most notable outlier, the computed binding free energy of Manα-(1,2)Manα-(1,2) in domain B deviates significantly from the linear fit, being computed as overly favorable. The addition of the total cvdW calculations were applied to the computed binding energy results (Figure 5-2, right panel), the cvdW values helped lower the magnitude closer to the experimental binding energy results. The most noteworthy effect came from the dimannose models. By adding the cvdW term, it significantly stabilizes the computed dimannose energies towards the trimannose energies.

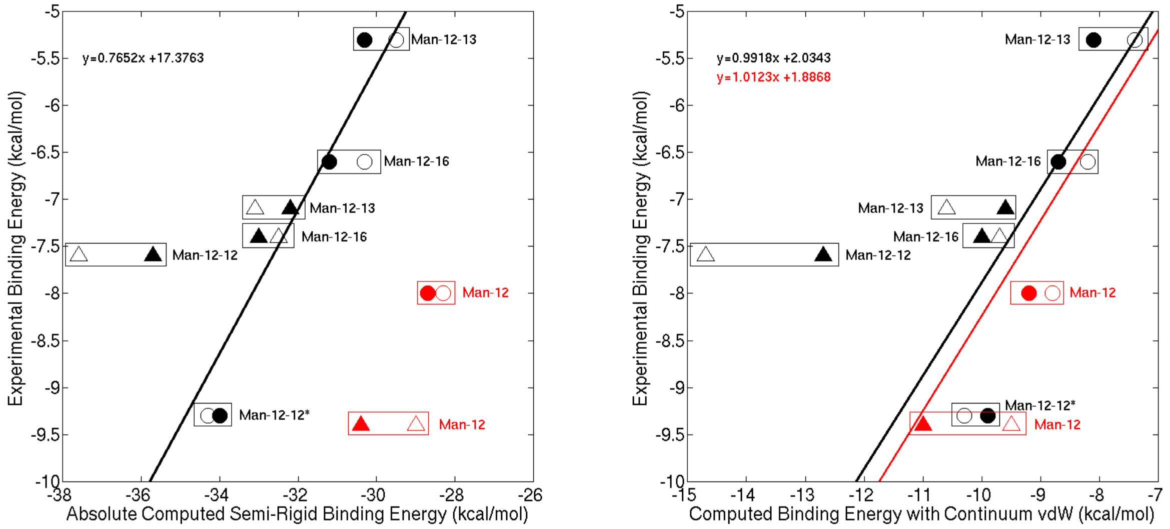


Figure 5- 2. **Comparison of computed vs. experimental binding free energies.** (Left) The comparison of energies *without* cvdW. (Right) The comparison *with* cvdW. Open shapes represent simulations from singly bound models and closed shapes represent simulations from doubly bound models; circles represent domain A region and triangles represent domain B; black line represents the linear best-fit line for trisaccharide data (excluding the terminal orientation of Manα-(1,2)Manα-(1,2)Man). The red line represents the linear best-fit line with the dimannose models.

In chapter 4, we were able to breakdown the electrostatic and van der Waals energies on every protein residue by its backbone amino, backbone carbonyl, and side chain groups. We can apply that procedure to the continuum van der Waals calculations. The intermolecular vdW energy decomposition of each protein residue is shown in Figure 5-3 (top panel). Figure 5-3 (bottom panel) also highlights the energy decomposition from the sum of the intermolecular vdW and continuum vdW. Every residue was divided into three groups—amino groups, carbonyl groups, and side chain groups. Overall, adding the cvdW energy for each component decreases the magnitude of the energies—in almost all cases, values are now within ± 1.0 kcal/mol. Prior to the addition of the cvdW calculations, many components had energy values greater than ± 2.0 kcal/mol. Several other cases show the cancellation of positive and negative terms when the cvdW is added. These values correspond to a contribution toward improvement in binding energy calculations. The increased interactions appear to prefer larger amino acids. One of the largest changes came from Arg 76 for all models in both doubly and singly bound simulations. It lowered the magnitude of energy by 1.5−2.0 kcal/mol.

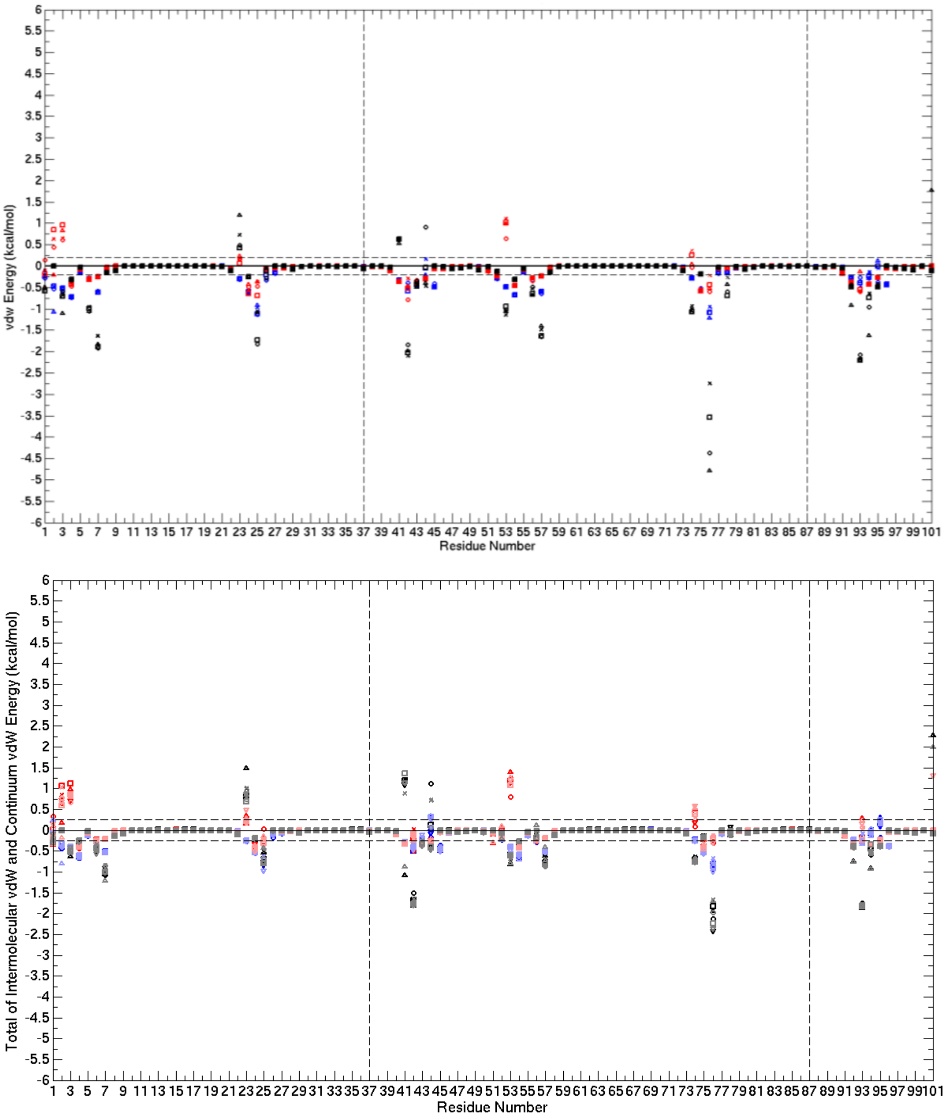


Figure 5- 3. **Comparison between the intermolecular results and the combined intermolecular and cvdW results.** *(Top panel)* Decomposition of the intermolecular van der Waals results. *(Bottom panel)* Decomposition of the combined intermolecular and cvdW results. × Man-12; △ Man-12-12*; ▽ Man-12-12; □ Man-12-13; ○ Man-12-16.

We observe just using bound state energies for van der Waals contributions can lead to size-dependent bias; the inclusion of solute-solvent vdW interactions (with a continuum approach) improves the predictive power of the calculations.

5.2 Calculation of Conformational Entropy

The binding affinity of a ligand for its protein partner depends on the balance between intermolecular interactions between the binding partners, desolvation of the binding partners, and the conformational entropy changes. The entropy change arises from changes in configurational fluctuations of the ligand, changes in configurational fluctuations of the protein, and the translational and rotational freedom of the ligand with respect to the protein. Changes in the entropy of molecules are believed to contribute importantly to the free energies of conformational change and binding. A reliable method of computing entropy could provide valuable insights into the mechanism of such processes. However, calculating entropy is traditionally a challenging problem, especially for complex molecules. One approach has been the quasiharmonic approximation [112], which involves a molecular dynamics simulation and computing the covariance matrix of atomic coordinates. The quasiharmonic method was successfully applied to simple systems with only one highly occupied energy well, and fluctuations were analyzed in a system of internal bond, angle, torsion coordinates. However, in recent studies this method does not yield accurate estimates of entropy when dealing with larger complexes. Quasiharmonic analysis has been used to study the motions in torsion-angle space for side chains. However, the torsional fluctuations of the main chain atoms restrict the range of side chain fluctuations [113]. The present work aims to analyze the conformational entropy for the protein side chains, and the conformational entropy for the carbohydrates' torsions around glycosidic bonds and hydroxyl groups.

Once again, we used the singly and doubly bound CVN systems for this analysis. In this work, the method used to estimate the entropies was a histogram-based approach. To calculate entropy from each simulation, torsion angles were extracted every picosecond and grouped into 10° bins. For carbohydrates, the frequencies in each bin (p_i) were used to calculate the entropy from the torsion angles of each glycosidic bond and for every hydroxyl group. For proteins, the frequencies in each bin (p_i) were used to calculate the entropy from the torsion angles of each side chain. The entropy equation for each case is shown below (Equation 5.6),

$$S = -RT \sum p_i \ln p_i \qquad (5.6)$$

where R is the gas constant and T is temperature (300 K).

In calculating the entropy for the protein side chains, there was one outlier from the trisaccharide analysis, and that came from the Manα-(1,2)Manα-(1,2)Man simulations. The largest difference observed came from Arg 76 in the doubly and singly bound simulations of Manα-(1,2)Manα-(1,2)Man (highlighted in red box in Figure 5-4). This suggests a possible entropic penalty missing from the computed Manα-(1,2)Manα-(1,2)Man binding energy for Domain B.

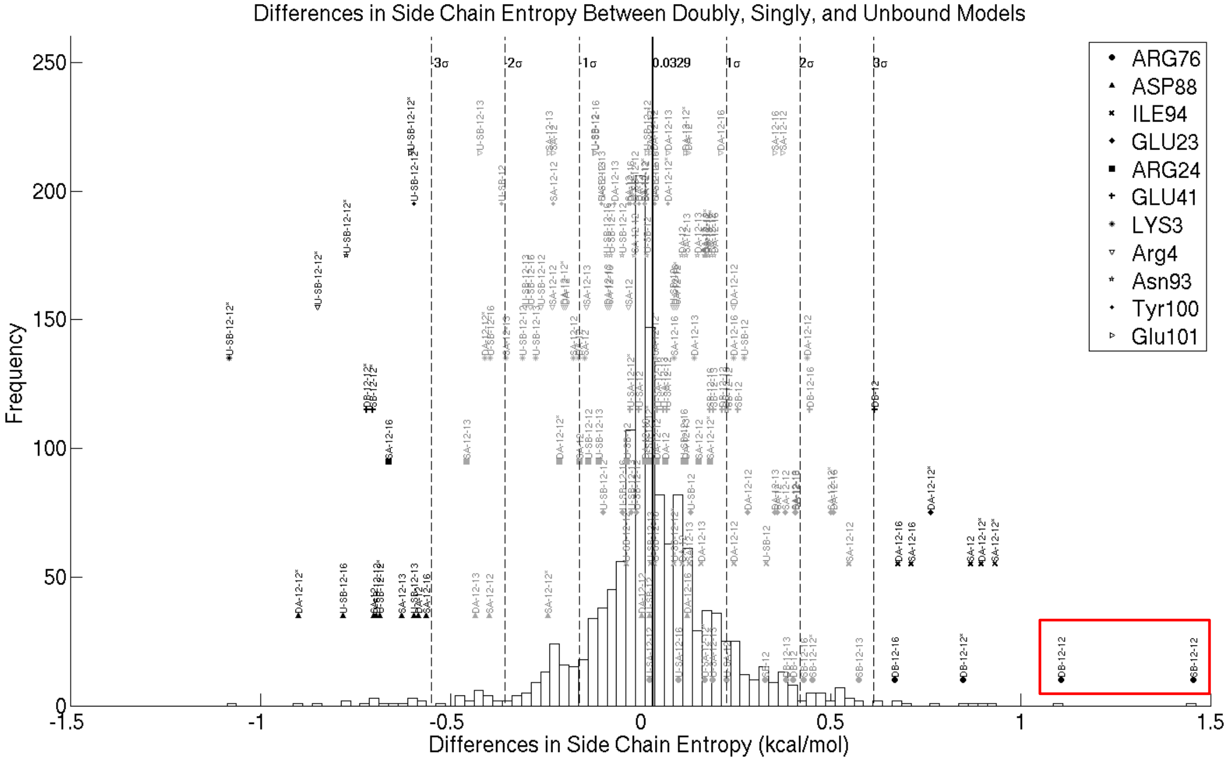


Figure 5- 4. **Changes in side-chain entropy on binding**. The distribution of differences in computed side-chain entropy between the bound and unbound states is shown as a histogram; the heavy vertical line indicates the mean of the distribution, and the vertical dashed lines correspond to ±σ, ±2σ and ±3σ (where σ is the standard deviation of the distribution). All side chains for which any value was outside 3σ are labeled, with the outliers in black and values within 3σ in grey. Points are labeled (U-)[S/D][A/B]-12-yy(*), where S/D denotes results from a singly- or doubly-bound simulation, A/B denotes the domain, yy denotes the type of linkage for the second glycosidic bond and * denotes binding in the internal (as opposed to terminal) mode; a preceding U indicates that the residue is from the unbound domain of in a singly-bound simulation. Only a few residues show consistent deviations from the bulk distribution, with Arg 76 (domain B) a notable outlier in the context of α-Man-(1,2)-α-Man-(1,2)-α-Man binding.

Table 5- 2. **Carbohydrate entropy – Domain A.**

	α(1,2)Man$_2$	α(1,2)-α(1,2)Man$_3$	α(1,2)-α(1,3)Man$_3$	α(1,2)-α(1,6)Man$_3$
1,2	+0.6/ +0.6	+0.5/ +0.5	+0.6/ +0.6	+0.6/ +0.6
1,2/1,3/1,6	ND/ ND	+0.3/ +0.4	−0.1/ −0.2	+0.4/ 0.0
C5/C6 (1)	+0.1/ +0.1	−0.1/ +0.1	+0.1/ +0.1	+0.1/ +0.1
C5/C6 (2)	+1.2/ +1.1	+1.2/ +1.1	+1.2/ +1.2	+1.3/ +1.2
C5/C6 (3)	ND/ ND	−0.1/ +0.1	−0.1/ −0.1	ND/ ND
OH2 (1)	+0.2/ −0.2	ND/ ND	+0.1/ −0.3	+0.3/ +0.1
OH3 (1)	+0.8/ +0.7	+0.8/ +0.8	+0.7/ +0.7	+0.8/ +0.7
OH4 (1)	+0.6/ +0.6	+0.6/ +0.6	+0.6/ +0.6	+0.6/ +0.6
OH3 (2)	+0.6/ +0.8	+0.8/ +0.8	+0.7/ +0.7	+0.8/ +0.8
OH4 (2)	0.0/ +0.1	+0.2/ +0.3	+0.3/ +0.3	+0.5/ +0.3
SUM	+4.2/ +3.8	+4.7/ +4.8	+4.2/ +3.6	+5.5/ +4.2

All energies are in kcal/mol, the two values given are those from the singly and doubly-bound simulations, respectively. In the first column, 1,2; 1,3; 1,6 refers to the glycosidic linkage; C5/C6 (1) refers to the hydroxymethyl group around Man$_1$; C5/C6 (2) refers to the hydroxymethyl group around Man$_2$; C5/C6 (3) refers to the hydroxymethyl group around Man$_3$; OH2 (1) refers to the hydroxyl group from the C2 position of Man$_1$; OH3 (1) refers to the hydroxyl group from the C3 position of Man$_1$; OH4 (1) refers to the hydroxyl group from the C4 position of Man$_1$; OH3 (2) refers to the hydroxyl group from the C3 position of Man$_2$; OH4 (2) refers to the hydroxyl group from the C4 position of Man$_2$.

Table 5- 3. **Carbohydrate entropy – Domain B.**

	α(1,2)Man$_2$	α(1,2)-α(1,2)Man$_3$	α(1,2)-α(1,3)Man$_3$	α(1,2)-α(1,6)Man$_3$
1,2	+0.5/ +0.5	+0.5/ +0.6	+0.5/ +0.5	+0.5/ +0.6
1,2/1,3/1,6	ND/ ND	+0.5/ +0.7	0.0/ +0.1	+0.2/ −0.1
C5/C6 (1)	+0.3/ +0.2	+0.2/ +0.2	+0.2/ +0.2	+0.2/ +0.2
C5/C6 (2)	+0.4/ +0.4	+1.0/ +1.2	+0.9/ +0.9	+0.9/ +0.9
C5/C6 (3)	ND/ ND	+0.6/ +0.8	+0.1/ +0.1	ND/ ND
OH2 (1)	+0.3/ +0.3	+0.2/ +0.2	+0.2/ +0.3	+0.2/ +0.2
OH3 (1)	+0.7/ +0.7	+0.7/ +0.7	+0.7/ +0.7	+0.6/ +0.7
OH4 (1)	+0.7/ +0.7	+0.7/ +0.8	+0.7/ +0.7	+0.6/ +0.7
OH3 (2)	+0.5/ +0.3	+0.6/ +0.8	+0.5/ +0.5	+0.3/ +0.5
OH4 (2)	+0.1/ 0.0	+0.1/ +0.5	+0.3/ +0.2	+0.1/ +0.3
SUM	+3.4/ +3.0	+5.3/ +6.4	+4.1/ +4.2	+3.8/ +4.0

All energies are in kcal/mol, the two values given are those from the singly and doubly-bound simulations, respectively. In the first column, 1,2; 1,3; 1,6 refers to the glycosidic linkage; C5/C6 (1) refers to the hydroxymethyl group around Man$_1$; C5/C6 (2) refers to the hydroxymethyl group around Man$_2$; C5/C6 (3) refers to the hydroxymethyl group around Man$_3$; OH2 (1) refers to the hydroxyl group from the C2 position of Man$_1$; OH3 (1) refers to the hydroxyl group from the C3 position of Man$_1$; OH4 (1) refers to the hydroxyl group from the C4 position of Man$_1$; OH3 (2) refers to the hydroxyl group from the C3 position of Man$_2$; OH4 (2) refers to the hydroxyl group from the C4 position of Man$_2$.

In calculating the entropy for various regions of the carbohydrates, many values that were conserved throughout all the models in both domains, correspond to contacts made in the core α-(1,2) dimannose unit. The largest values observed came from regions where it makes direct contact with a residue in the binding pocket (see Table 5-2 and 5-3). For example, in domain A, the largest entropy value came from the hydroxymethyl group of Man_2 and that was consistent throughout all the models. This is due to a consistent interaction with the side chain of Glu 23. In domain B, the largest magnitude was seen in the model with the hydroxymethyl group of Man_2 from Manα-(1,2)Manα-(1,2)Man. Interactions with Arg 76 lead to less flexibility of this region in the bound state, and thus a corresponding increase in entropy. Again, this helps explain the computed over-stabilization of this molecule. However, the data above cannot be simply added to the computed binding energies. When looking at the sum of the entropies in both tables, there is a range of values throughout all the models, and it appears it is due to taking the difference taken from the unbound state. Even taking the difference from the complete unbound state to all unbound regions from singly bound simulations, the values vary (Tables 5-4 and 5-5). In Table 5-5, if the difference between the unbound region in a singly bound simulation and the complete unbound state is taken, one could assume the value to be zero. However, this is not the case. Instead, we see a range of positive and negative values. Since many regions of CVN are highly dynamic, it raises the possibility of imperfect sampling of the various states; longer simulations could help address this question. Nevertheless, these data were computed by assuming that all the side chains behave independently, that backbone configurations do not contribute, and without consideration of the flexibility of the carbohydrate. As a result, they should not be taken as a quantitative prediction of configurational entropic effects, nor should they be directly added to the semi-rigid binding free energies presented here. Further work will be needed to develop quantitative assessments of these contributions.

Table 5- 4. **The sum of all residue side chain entropies after the differences for each residue was taken between the complete unbound state and the bound state.**

	α-(1,2)Man_2	α-(1,2),α-(1,2)Man_3	α-(1,2),α-(1,3)Man_3	α-(1,2),α-(1,6)Man_3
Domain A	+1.0/ +1.3	+2.7/ +1.1	+0.1/ +2.5	−1.1/ +5.5
Domain B	+3.2/ +3.2	+5.5/ +5.2	+2.8/ +3.6	+2.6/ +3.7

All energies are in kcal/mol, the two values given are those from the singly and doubly-bound simulations, respectively.

Table 5- 5. **The sum of all residue side chain entropies after the differences for each residue was taken between the complete unbound state and the unbound side from the singly bound simulations.**

	α-(1,2)Man_2	α-(1,2),α-(1,2)Man_3	α-(1,2),α-(1,3)Man_3	α-(1,2),α-(1,6)Man_3
Domain A	−1.4	+6.9	+1.8	+0.8
Domain B	−0.5	−1.0	−1.6	−1.0

All energies are in kcal/mol.

5.3 Carbohydrate Strain Energies

Another energy component missing from the binding free energy calculation is the strain energy. Strain energies are easily computed by running additional simulations of the unbound state; the difference in total energy (bond, angle, and dihedral strain, intramolecular van der Waals and Coulombic interactions, and solute–solvent interactions) between the unbound ensemble and the ensemble of structures taken from the bound state (with the binding partner removed) is the strain. It corresponds to the energetic cost of perturbing the unbound conformational ensemble into the ensemble that is capable of binding. The carbohydrate strain energy is given by the difference in the ensemble average of the total energy of the sugar in conformations extracted from the complex simulation and the average for conformations from a simulation of the free sugar. The protein strain energies were also calculated by taking the difference in the ensemble average of the total energy of the protein in conformations extracted from the complex simulation and the average for conformations from a simulation of the free protein; however, it produced large values. Just like the entropy calculations, the protein strain calculations cannot be additive due to a lack of convergence. This calculation requires an exhaustive sampling of the protein conformational space and currently this has not been tested.

Table 5-6 shows the calculated results from the carbohydrate strain energies. The similarity of the computed values for most cases suggests reasonable convergence of the bound-state simulations; although the Man-12-16 in Domain B is a notable outlier to this trend.

Table 5- 6. **Carbohydrate strain energies.**

	$\Delta G^{internal}$	ΔG^{vdW}	$\Delta G^{h\phi}$	ΔG^{des}	ΔG^{elec}	ΔG^{str_A}
Domain A						
α-(1,2)Man$_2$	+1.3/ +1.4	+0.1/ +0.1	0.0/ 0.0	−2.7/ −2.5	+3.3/ +3.0	+2.0/ +2.0
α-(1,2)-α-(1,2)Man$_3$	+1.3/ +1.3	+0.1/ −0.2	0.0/ 0.0	−2.8/ −3.0	+8.2/ +7.0	+6.8/ +5.1
α-(1,2)-α-(1,3)Man$_3$	+0.8/ +1.1	+0.3/ +0.3	0.0/ 0.0	−2.5/ −2.5	+3.5/ +3.6	+2.1/ +2.4
α-(1,2)-α-(1,6)Man$_3$	−1.3/ +1.2	−0.5/ −0.2	−0.1/ 0.0	−3.0/ −3.2	+8.7/ +5.5	+3.8/ +3.3
Domain B						
α-(1,2)Man$_2$	+2.4/ +2.2	+0.1/ −0.1	0.0/ 0.0	−1.8/ −1.5	+1.8/ +1.5	+2.5/ +2.1
α-(1,2)-α-(1,2)Man$_3$	+1.4/ +1.3	−0.2/ 0.0	0.0/ 0.0	−1.6/ −2.2	+3.4/ +4.2	+3.1/ +3.4
α-(1,2)-α-(1,3)Man$_3$	+2.2/ +2.3	−0.1/ 0.0	0.0/ 0.0	−2.0/ −2.1	+1.8/ +2.1	+1.8/ +2.4
α-(1,2)-α-(1,6)Man$_3$	+0.1/ +2.1	−0.5/ −0.2	−0.1/ 0.0	−2.0/ −2.2	+7.7/ +3.8	+5.8/ +3.5

All energies are in kcal/mol. The two values given are those from the singly and doubly-bound simulations, respectively.

Carbohydrate strain calculations appear to play a role, but questions remain about the convergence of the unbound state results in both carbohydrates and protein. These terms are difficult to accurately compute because there is a lack of convergence due to transitions between glycosidic-bond dihedrals and protein side chain torsions that are rarely seen. The molecular dynamic simulations performed here may not be able to capture all degrees of freedom in these regions, thus cannot accurately capture the energetics.

The objective of this section was to define a protocol to accurately and efficiently estimate the contributions to the protein−carbohydrate calculations. Although the current data shows noise (likely due to a lack of convergence) which limits the value in making quantitative predictions; it still does help explain where certain energetic differences are coming from. Hence, further detailed investigation of these methodologies is ongoing in our laboratory.

5.4 Application of Current Methods to a New Lectin Model

In an earlier chapter, it was mentioned that one way to prevent HIV infection was to target the glycan shield of surface envelope proteins for HIV to inhibit membrane fusion and infection [12]. The initial steps leading to viral entry include binding of the HIV surface envelope glycoprotein gp120 to cellular receptors CD4 and CXCR4 or CCR5, followed by gp41-mediated membrane fusion [114]. This display of glycans functions as a barrier to protect the virus from recognition by the human immune system [57]. To exploit this barrier as a therapeutic target, carbohydrate-binding proteins, known as lectins, have emerged as promising anti-HIV agents [12, 65]. One that has been mentioned earlier in this dissertation was cyanovirin-N. While CVN was originally thought to be an orphan lectin with little homology to any other known protein family [60], a family of CVN homologs (CVNH), has been described [115, 116]. Some members of this family are found in multicellular ascomycetous fungi and in ferns and share a 3-dimensional fold [116]. A CVNH of the toxin-producing cyanobacterium, *Microcystis aeruginosa* also binds high mannose-type glycans and is involved in cell–cell attachment of Microcystis [117]. Defining the glycan binding specificity and mode of action for virucidal lectins may help to develop new therapeutic approaches directed at combating viral infections.

The cyanobacterial protein, microvirin (MVN) [117] is an attractive candidate for microbicide development; when compared with its well-studied homolog cyanovirin-N (CVN), it is reported to show comparable potency in HIV-1 neutralization assays but with notably reduced toxicity profiles [118]. Recently, Shahzad-ul-Hussan, *et al.* [119] has solved the solution structure of MVN free and in complex with its ligand Man-α-(1,2)Man, and compared specificity with CVN. They also showed by NMR and analytical ultracentrifugation that MVN is monomeric in solution and demonstrate by NMR that Man-α-(1,2)Man-terminating carbohydrates interact with a single carbohydrate-binding site. Although MVN is a member of the CVN family of lectins, it possesses distinct structural characteristics compared with CVN that accounts for its low toxicity and narrow antiviral profile. We will implement the same methods mentioned in earlier chapters to the MVN system and use these techniques to understand the structural and energetic differences between CVN and MVN.

The initial structure for all simulations originated from a solution structure of MVN bound to dimannose in a single binding pocket (PDB 2YHH), as shown in Figure 5-5.

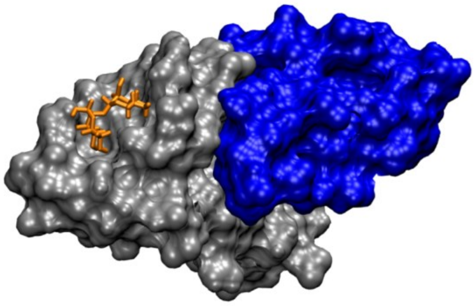


Figure 5- 5. **Figure of dimannose bound to MVN.** Domain A (gray) contains residues 38−93 and domain B (blue) contains residues 1−37 and 94−108.

The reducing mannopyranose ring is positioned in a deep pocket formed by two β turns (one around residues 81–84 and residues 44–47). It has been shown the hydrophobic interactions govern the binding to this site by the methyl group on M83 making van der Waals contacts with H2, H3, and H5 protons of the sugar ring. There is another methyl group positioned at T59 making van der Waals contacts with H4 and H6 of one ring and H5 of another ring. The carbohydrate moiety can only be bound to one site because proline introduces a steric clash with the dimannose. Histidine is facing away from the binding site, losing hydrogen bond interactions with the mannose. Residues 101-104 in domain B obstruct the binding site space (residues 45-48 in domain A position away from the binding site) as shown in Figure 5-6.

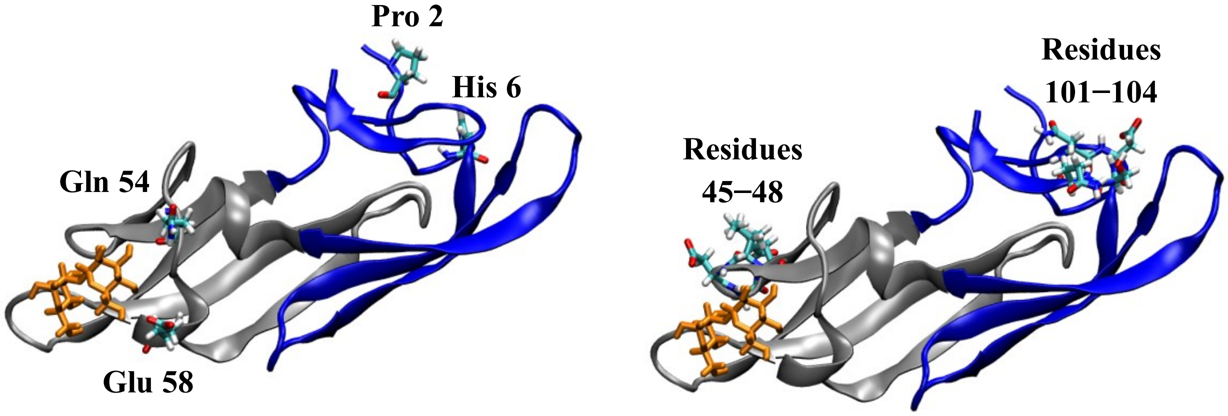


Figure 5- 6. **Structures of MVN bound to Manα-(1,2)Man showing the inability of Manα-(1,2)Man to bind through domain B.** Domain A (gray) contains residues 38−93 and domain B (blue) contains residues 1−37 and 94−108. The dimannose (orange licorice) is bound in domain A. Residues shown in the figures depict the inability of Manα-(1,2)Man to bind through domain B. In domain B, Pro 2 introduces steric clash with the disaccharide and is positioned on the polar face of the mannopyranose ring, and the side chain of His 6 is directed away from the binding site removing hydrogen bonding interactions with the terminal ring. The corresponding residues in domain A (Gln 54 and Glu 58) do not encounter this. In addition residues 101–104 of domain B block the space that is used for carbohydrate binding and the corresponding residues in domain A do not (residues 45-48).

Table 5- 7. **Computed energetic results compared to experimental binding analysis for MVN.**

Carbohydrate	ΔG^{vdw}	$\Delta G^{h\phi}$	ΔG^{elec}	ΔG^{cvdW}	ΔG^{comp}
α-(1,2)Man$_2$	−17.6 (0.1)	−2.1 (0.1)	−7.4 (0.1)	+19.2	−7.9 (0.1)
α-(1,2)-α-(1,2)Man$_3$	−20.7 (0.1)	−2.4 (0.1)	−6.9 (0.1)	+22.5	−7.5 (0.1)
α-(1,2)-α-(1,6)Man$_3$	−20.5 (0.1)	−2.4 (0.1)	−6.5 (0.1)	+22.2	−7.2 (0.1)

All energies are in kcal/mol.

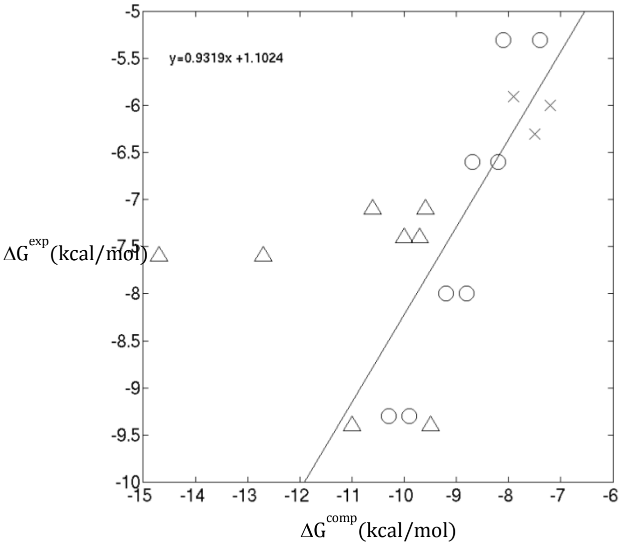


Figure 5- 7. **Comparison of computed and experimental binding free energies with continuum van der Waals calculations for CVN and MVN.** ○, CVN (Domain A); △ CVN (Domain B); ×, MVN. The black line represents the linear best-fit line for all data, excluding the terminal orientation of Manα-(1,2)Manα-(1,2)Man).

Structures of disaccharide and two trimannoses (Manα-(1,2)Manα-(1,2)Man, and Manα-(1,2)Manα-(1,6)Man) were built by extending the dimannose model by one unit from the anomeric carbon of the reducing sugar. All these manipulations were done using the CHARMM software package, [33] and default conformations were used for the newly-built portions of each molecule. A short minimization (100 steps) was performed on all the newly built structures to avoid any clashes. The binding affinities for each MVN complex were calculated and results are shown in Table 5-7 and Figure 5-7. An addition to the binding affinity calculation, an electrostatic and van der Waals component analysis were performed.

When the MVN results were compared to the CVN analysis, the data for the CVN models were energetically more favorable. To start evaluating the MVN complexes and comparing them to the CVN complexes, we first looked at components that had favorable energies greater than −0.2 kcal/mol, and those residues are listed in Tables 5-8 and 5-9. In the analysis, many components that were energetically conserved in all the CVN models were also conserved in the corresponding residues in the MVN systems.

If we first look at the electrostatic determinants, components that were energetically conserved in both CVN and MVN models mainly came from the backbone atoms and one side chain. These residues mainly made interactions with the core dimannose molecule (backbone:

Asn 42, Asp 44, Ser 52, Asn 53, Lys 74; side chain: Thr 57). The components that contributed to CVN being more favorable than MVN are from side chain groups: Glu 41, Asp 44, Arg 76, and Gln 78. The corresponding residues in the MVN system made either no or unfavorable interactions. There was only one component that made more favorable interaction in MVN (Asp 53) and not in CVN (Pro 51). The van der Waals component analysis showed more residues that contributed to the binding energy. The van der Waals energetics were similar in both systems and the largest contributors in both systems came from side chain groups (Lys 74, Arg 76, and Gln 78).

Although this analysis is in its preliminary stages, we can start to see the similarities and differences between the two systems energetically and structurally. The residues that contribute more in CVN than MVN come from charged side chain groups, while in MVN it is not. Further work needs to be done to fully understand these models.

Table 5- 8. **CVN and MVN determinants – electrostatic energy.**

Residue	α-(1,2)Man$_2$	α-(1,2), α-(1,2)Man$_3$^{b}	α-(1,2), α-(1,6)Man$_3$
Amino			
N42/ N44	−1.5/ −1.4	−1.4/ −1.3	−1.5/ −1.4
D44/ D46	−1.6/ −1.9	−2.2/ −2.2	−2.0/ −2.2
F54/ F56	−0.3/ −0.3	−0.3/ −0.3	−0.3/ −0.3
T75/ T82	−0.3/ −0.4	−0.5/ −0.4	−0.4/ −0.4
Carbonyl			
E41/ G43	−0.7/ −1.0	−0.7/ −0.9	−0.7/ −0.9
N42/ N44	−1.0/ −1.5	−1.9/ −1.8	−1.4/ −1.7
V43/ I45	−0.8/ −0.8	−0.9/ −1.0	−0.8/ −1.0
D44/ D46	−0.5/ −0.5	−0.6/ −0.7	−0.6/ −0.7
S52/ F54	−1.2/ −1.3	−1.3/ −1.3	−1.1/ −1.2
N53/ N55	−2.8/ −2.7	−2.8/ −2.7	−2.8/ −2.6
E56/ E58	−0.3/ −0.3	−0.3/ −0.3	−0.3/ −0.3
T57/ T59	−0.2/ −0.3	−0.3/ −0.3	−0.3/ −0.3
K74/ Q81	−2.3/ −2.9	−3.4/ −2.8	−2.8/ −2.7
T75/ T82	−0.6/ −0.4	−0.3/ −0.3	−0.4/ −0.3
Side chain			
E41/ G43	−2.3/ 0.0	−2.6/ 0.0	−2.6/ 0.0
D44/ D46	−2.9/ +0.7	−0.8/ +1.0	−0.8/ +0.9
P51/ D53	0.0/ −2.8	0.0/ −2.8	0.0/ −2.8
E56/ E58	−0.4/ −0.5	−0.3/ −0.4	−0.5/ −0.3
T57/ T59	−1.7/ −1.7	−1.7/ −1.5	−1.7/ −1.6
R76/ M83	−0.8/ −0.1	−4.3/ −0.2	−0.7/ −0.2
Q78/ G85	−0.3/ 0.0	−0.7/ 0.0	−0.5/ 0.0

All energies are in kcal/mol. ^{b} The α-(1,2), α-(1,2)Man$_3$-linked sugar is bound in the terminal orientation, and the two values given are those from the CVN and MVN simulations, respectively.

Table 5- 9. **CVN and MVN determinants – van der Waals energy.**

Residue	α-(1,2)Man$_2$	α-(1,2), α-(1,2)Man$_3$[b]	α-(1,2), α-(1,6)Man$_3$
Amino			
E41/ G43	−0.3/ −0.3	−0.3/ −0.3	−0.3/ −0.3
N42/ N44	−0.4/ +0.1	−0.6/ −0.1	−0.4/ +0.1
V43/ I45	−0.3/ −0.1	−0.2/ −0.4	−0.3/ −0.1
G45/ G47	−0.4/ −0.4	−0.6/ −0.5	−0.4/ −0.5
S52/ Q54	−0.3/ −0.4	−0.3/ −0.4	−0.3/ −0.4
N53/ N55	−0.5/ −0.4	−0.5/ −0.5	−0.5/ −0.5
F54/ F56	−0.7/ −0.7	−0.7/ −0.7	−0.7/ −0.7
E56/ E58	−0.3/ −0.3	−0.3/ −0.3	−0.3/ −0.3
T57/ T59	−0.6/ −0.6	−0.6/ −0.6	−0.6/ −0.6
K74/ Q81	−0.3/ −0.3	−0.3/ −0.3	−0.3/ −0.3
T75/ T82	−0.5/ −0.6	−0.6/ −0.6	−0.5/ −0.6
R76/ M83	−0.9/ −0.9	−1.3/ −1.2	−0.9/ −1.2
Carbonyl			
E41/ G43	−0.4/ −0.4	−0.4/ −0.4	−0.5/ −0.4
V43/ I45	−0.4/ −0.4	−0.4/ −0.4	−0.5/ −0.4
D44/ D46	−0.3/ −0.3	−0.3/ −0.3	−0.5/ −0.3
F54/ F56	−0.5/ −0.5	−0.6/ −0.5	−0.5/ −0.5
T75/ T82	−0.5/ −0.5	−0.6/ −0.7	−0.5/ −0.5
R76/ M83	−0.2/ −0.2	−0.6/ −0.5	−0.5/ −0.5
Side chain			
N42/ N44	−2.0/ −2.1	−2.1/ −2.2	−2.0/ −2.0
V43/ I45	−0.3/ −0.7	−0.5/ −0.8	−0.4/ −0.9
D44/ D46	+0.6/ −0.4	−0.6/ −0.5	−0.1/ −0.5
N53/ N55	−1.1/ −0.7	−1.0/ −0.6	−1.0/ −0.7
F54/ F56	−0.3/ −0.3	−0.3/ −0.3	−0.3/ −0.3
E56/ E58	−0.6/ −0.5	−0.5/ −0.5	−0.5/ −0.5
T57/ T59	−1.5/ −1.8	−1.7/ −1.9	−1.7/ −1.7
K74/ Q81	−1.0/ −0.3	−1.1/ −0.5	−1.0/ −0.6
R76/ M83	−2.7/ −1.9	−4.8/ −3.8	−3.4/ −3.5
Q78/ G85	−0.4/ 0.0	−0.5/ 0.0	−0.3/ 0.0

All energies are in kcal/mol. [b] The α-(1,2),α-(1,2)Man$_3$-linked sugar is bound in the terminal orientation, and the two values given are those from the CVN and MVN simulations, respectively.

5.5 Conclusion

While the results presented here go a long way towards explaining the details of specific carbohydrate recognition by CVN and MVN, it does provide near quantitative reproduction of relative binding free energies in many cases. However, the results are not perfect. In the CVN model, Manα-(1,2)Manα-(1,2)Manα bound in domain B is a notable outlier. The computed stabilization of this complex is dominated by contributions from a single amino acid, Arg 76.

Arginine is one of most naturally occurring amino acids, with four dihedral degrees of freedom; as a result, the formation of persistent interactions by an arginine may be expected to be accompanied by significant entropic costs. These data strongly suggest that side chain entropy is a major contribution to the overprediction of the stability of this complex. The computed binding energies from the models gave reasonable accuracy to experimental trend. Component analysis is a useful technique to find where the observed interactions are coming from. The inclusion of solute-solvent vdW interactions improves the predictive power of the calculations.

Chapter 6

General Conclusions

As stated in the introduction of this dissertation, the characterization of the 3-dimensional structures of oligosaccharides is particularly challenging for traditional experimental methods. Computational methods and, in particular, molecular dynamics simulations provide complementary tools to augment both X-ray and NMR data and are particularly well suited to the characterization of the structure and dynamics of glycans and glycoconjugates. Molecular simulation methods provide a basis for interpreting sparse experimental data and for independently predicting conformational and dynamic properties of glycans.

One important issue affecting the quality of MD simulations is conformational sampling. In the early years of biomolecular simulations, MD simulations were limited to small biological systems and ran for very short time frames. From the standpoint of today's technical achievements, advances in computer technology and software algorithms enable us today to sample the conformational space of biomolecular systems on the order of hundreds of nanoseconds. With the availability of NY Blue, large-scale molecular dynamics simulations were performed and the studies presented in this dissertation provided examples of validating the application of continuum electrostatic models to carbohydrate–protein association. We have been able to explain a number of important features of CVN. With the availability of a supercomputer, we were able to observe a rare *cis*-peptide bond in the monomeric state.

Carbohydrate–protein interactions are intrinsically more dynamic than many other protein–ligand complexes, and their affinity and specificity arises from electrostatic, hydrogen bonding, and hydrophobic interactions between the protein, the solvent, and the carbohydrate, which result in significant changes in both enthalpy and entropy upon binding. We have demonstrated that the two domains of CVN, while highly homologous in sequence, recognize certain oligosaccharide targets with distinct binding modes. The results from component analysis identified several key interactions throughout all the carbohydrate models.

Although it has been technically possible to simulate the structure of a carbohydrate–protein complex for more than a decade, the accurate prediction of binding affinity remains a challenging task. In this dissertation, we were able to demonstrate the contribution of continuum electrostatic models and molecular dynamics simulations to dissect the protein−carbohydrate interactions. The observed variations in binding affinity were mainly due to electrostatic effects. The electrostatic analysis showed important information for affinity and specificity for the various CVN complexes. Although the computational approach to calculating the binding free energies gave values significantly larger than the experimental results, the observed trends were present. The electrostatic analysis for all preferred models in domain B revealed strong electrostatic effects from Glu 41, which is known to be important for specificity. The electrostatic interactions from the N and C termini of the preferred Manα-(1,2)Manα-(1,2)Man in domain A gave a favorable affinity over the other models in that domain.

In general, there might be significant changes in entropy and indeed, entropic contributions have been implicated as a major factor in determining carbohydrate-binding affinity. In addition to entropy, solvent plays an important role. To answer all this, this dissertation begins to take account for these factors. Although the quantitative calculations still need further development, the present method allows us to begin to understand more where the energetic differences are coming from and obtain values closer to the experimental results. Once again, these methods allowed us to observe more of the effects coming from Arg 76.

In the future, we can apply the methods used in chapter 4 to other lectins involved with HIV infection, as well as studying larger saccharide molecules (*e.g.* gp120) and get a better understanding on how the mechanism work between protein and carbohydrates. In addition, we can take the information from the electrostatic analysis and apply it towards the design of mutants to alter affinity and specificity.

Bibliography

[1] N. Sharon and H. Lis, "Lectins as cell recognition molecules," *Science,* vol. 246, p. 227–234, 1989.

[2] N. Sharon, "Lectin-carbohydrate complexes of plants and animals: an atomic view," *Trends Biochem. Sci.,* vol. 246, p. 221–226, 1993.

[3] T. B. H. Geijtenbeek, R. Torensma, S. J. van Vliet, G. C. F. van Duijnhoven, G. J. Adema, Y. van Kooyk and C. G. Figdor, "Identification of DC-SIGN, a novel dendritic cell–specific ICAM-3 receptor that supports primary immune responses," *Cell,* vol. 100, p. 575–585, 2000.

[4] C. F. Zinecker, B. Striepen, H. Geyer, R. Geyer, J. F. Dubremetz and R. T. Schwarz, "Two glycoforms are present in the GPI-membrane anchor of the surface antigen I (P30) of Toxoplasma gondii," *Mol. Biochem. Parasit.,* vol. 116, p. 127–135, 2001.

[5] M. D. Disney and P. H. Seeberger, "The use of carbohydrate microarrays to study carbohydrate–cell interactions and to detect pathogens," *Chem. Biol.,* vol. 11, p. 1701–1707, 2004.

[6] D. S. Newburg, G. M. Ruiz-Palacios and A. L. Morrow, "Human milk glycans protect infants against enteric pathogens," *Annu. Rev. Nutr.,* vol. 25, p. 37–58, 2005.

[7] D. A. Calarese, C. N. Scanlan, M. B. Zwick, S. Deechongkit, Y. Mimura, R. Kunert, P. Zhu, M. R. Wormald, R. L. Stanfield, K. H. Roux, J. W. Kelly, P. M. Rudd, R. A. Dwek, H. Katinger, D. R. Burton and I. A. Wilson, "Antibody domain exchange is an immunological solution to carbohydrate cluster recognition," *Science,* vol. 300, p. 2065–2071, 2003.

[8] J. Holgersson, A. Gustafsson and M. E. Breimer, "Characteristics of protein–carbohydrate interactions as a basis for developing novel carbohydrate–based antirejection therapies," *Immunol. Cell Biol.,* vol. 83, p. 694–708, 2005.

[9] A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze , P. Stanley , C. R. Bertozzi, G. W. Hart and M. E. Etzler , Essentials of Glycobiology, 2nd ed., Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2009.

[10] T. H. T. Christlet and K. Veluraja, "Database analysis of O-glycosylation sites in proteins," *Biophys. J.,* vol. 80, p. 952–960, 2001.

[11] Y. Gavel and G. von Heijne, "Sequence differences between glycosylated and nonglycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering," *Protein Eng.,* vol. 3, p. 433–442, 1990.

[12] J. Balzarini, "Targeting the glycans of glycoproteins: a novel paradigm for antiviral therapy," *Nat. Rev. Microbiol.,* vol. 5, p. 583–597, 2007.

[13] Y. Ni and I. Tizard, "Lectin–carbohydrate interaction in the immune system," *Vet. Immunol. Immunop.,* vol. 55, p. 205–223, 1996.

[14] E. Bettler, A. Imberty, R. Loris and A. Rivet, "Lectines 3D structure of lectins," Centre National de la Recherche Scientifique (Cermav-CNRS), [Online]. Available: http://www.cermav.cnrs.fr/lectines/.. [Accessed 2012].

[15] H. Lis and N. Sharon, "Lectins: carbohydrate–specific proteins that mediate cellular recognition," *Chem. Rev.,* vol. 98, p. 637–674, 1998.

[16] F. A. Quiocho, "Probing the atomic interactions between proteins and carbohydrates," *Biochem. Soc. Trans.,* vol. 21, p. 442–448, 1993.

[17] Y. Bourne, P. Rouge and C. Cambillaun, "X-ray structure of a biantennary oligosaccharide-lectin complex refined at 2.3-A resolution," *J. Biol. Chem.,* vol. 267, p. 197–203, 1992.

[18] R. Loris, S. Phillipe and L. Wyns, "Conserved waters in legume lectin crystal structures," *J. Biol. Chem.,* vol. 269, p. 26722–26733, 1994.

[19] F. Gao, E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cumminsk, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp and B. H. Hahn, "Origin of HIV-1 in the chimpanzee pan troglodytes," *Nature,* vol. 397, p. 436–441, 1999.

[20] World Health Organization (WHO), [Online]. Available: http://www.who.int/hiv/data/en/. [Accessed 2012].

[21] C. K. Leonard, M. W. Spellman, L. Riddle, R. J. Harris, J. N. Thomas and T. J. Gregory, "Assignment of interchain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in Chinese hamster ovary cells," *J. Biol. Chem.,* vol. 265, p. 10373–10382, 1990.

[22] P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski and W. A. Hendrickson, "Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody," *Nature,* vol. 393, p. 648–659, 1998.

[23] S. Zolla-Pazner and T. Cardozo, "Structure–function relationships of HIV-1 envelope sequence–variable regions provide a paradigm for vaccine design," *Nat. Rev. Immunol.,* vol. 10, no. 7, p. 527–535, 2010.

[24] J. R. Mascola and D. C. Montefiori, "The role of antibodies in HIV vaccines," *Annu. Rev. Immunol.,* vol. 28, p. 413–444, 2010.

[25] P. Volberding, Global HIV/AIDS Medicine, Philadelphia: Elsevier, 2008.

[26] M. R. Boyd, K. R. Gustafson, J. B. McMahon, R. H. Shoemaker, B. R. O'Keefe, R. J. Gulakowski, L. Wu, M. I. Rivera, C. M. Laurencot, M. J. Currens, J. H. Cardellina, 2nd, R. W. Buckheit, Jr., P. L. Nara, L. K. Pannell, R. C. Sowder, 2nd and L. E. Henderson, "Discovery of cyanovirin-N, a novel human immunodeficiency virus-inactivating protein that binds viral surface envelope glycoprotein gp120: potential applications to microbicide development," *Antimicrob. Agents Ch.,* vol. 41, p. 1521–1530, 1997.

[27] M. J. Currens, R. J. Gulakowski, J. M. Mariner, R. A. Moran, R. W. Buckheit, Jr., K. R. Gustafson, J. B. McMahon and M. R. Boyd, "Antiviral activity and mechanism of action of calanolide A against the human immunodeficiency virus," *J. Pharmacol. Exp. Ther.,* vol. 279, p. 645–651, 1996.

[28] M. J. Currens, J. M. Mariner, R. A. Moran, J. B. McMahon and M. R. Boyd, "Kinetic analysis of inhibition of HIV-1 reverse transcriptase by calanolide A," *J. Pharmacol. Exp. Ther.,* vol. 279, p. 652–661, 1996.

[29] W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids," *J. Am. Chem. Soc.,* vol. 118, p. 11225–11236, 1996.

[30] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *J. Am. Chem. Soc.,* vol. 117, p. 5179–5197, 1995.

[31] A. D. Mackerell, Jr., "Empirical force fields for biological macromolecules: overview and issues," *J. Comput. Chem.,* vol. 25, p. 1584–1604, 2004.

[32] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *J. Comput. Chem.,* vol. 4, p. 187–217, 1983.

[33] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im and K. Kuczera, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.,* vol. 30, p. 1545–1614, 2009.

[34] W. L. Jorgensen and J. Tirado-Rives, "The OPLS force field for proteins: energy minimizations for crystals of cyclic peptides and crambin," *J. Am. Chem. Soc.,* vol. 110, p. 1657–1666, 1988.

[35] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom and Reiher, "All-atom empricial potential for molecular modeling and dynamics studies of proteins," *J. Phys. Chem. B.,* vol. 102, p. 3586–3616, 1998.

[36] N. Foloppe and A. MacKerell, Jr., "All-atom empirical force field for nucleic acids: parameter optimization based on small molecule and condensed phase macromolecular target data," *J. Comput. Chem.,* vol. 21, p. 86–104, 2000.

[37] A. D. MacKerell, Jr. and N. Banavali, "All-atom empirical force field for nucleic acids: application to molecular dynamics simulations of DNA and RNA in solution," *J. Comput. Chem.,* vol. 21, p. 105–120, 2000.

[38] R. Schleif, Analysis of Protein Structure and Function: A Beginner's Guide to CHARMM, Baltimore: Johns Hopkins University, 2006.

[39] "Protein Data Bank," [Online]. Available: http://www.rcsb.org/ .

[40] T. L. Blundell, S. C. Harrison, R. M. Stroud, S. Yokoyama, L. E. Kay, M. G. Rossmann, H. M. Berman, B. Kobilka, J. M. Thornton and D. Cowburn, "Celebrating structural biology," *Nat. Struct. Mol. Biol.,* vol. 18, p. 1304–1316, 2011.

[41] J. A. McCammon, B. R. Gelin and M. Karplus, "Dynamics of folded proteins," *Nature,* vol. 267, p. 585–590, 1977.

[42] M. Karplus, "Molecular dynamics of biological macromolecules: a brief history and perspective," *Biopolymers,* vol. 68, p. 350–358, 2003.

[43] T. Ichiye and M. Karplus, "Fluorescence depolarization of tryptophan residues in proteins: a molecular dynamics study," *Biochemistry,* vol. 22, p. 2884–2893, 1983.

[44] W. Nadler, A. T. Brunger, K. Shulten and M. Karplus, "Molecular and stochastic dynamics of proteins," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 84, p. 7933–7937, 1987.

[45] A. T. Brunger and M. Karplus, "Molecular dynamics simulations with experimental restraints," *Acc. Chem. Res.,* vol. 24, p. 54–61, 1991.

[46] A. T. Brunger, G. M. Clore, A. M. Gronenborn and M. Karplus, "Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 83, p. 3801–3805, 1986.

[47] J. Ma, P. B. Sigler and Z. Xu, "A dynamic model for the allosteric mechanism of GroEL," *J. Mol. Biol.,* vol. 302, p. 303–313, 2000.

[48] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen and S. Piana, "Atomic–level characterization of the structural dynamics of proteins," *Science,* vol. 330, p. 341–346, 2010.

[49] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey and M. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys. ,* vol. 79, p. 926–935, 1983.

[50] A. R. Leach, Molecular Modelling Principles and Applications, 2nd ed., Harlow: Pearson Education Limited, 2001.

[51] C. J. Cramer and D. G. Truhlar, "Continuum solvation models: classical and quantum mechanical implementations," in *Rev. Comp. Ch.*, New York, VCH Publishers, 1995, p. 1–73.

[52] M. Feig, "Implicit solvation based on Generalized–Born theory in different dielectric environments," *J. Chem. Phys.,* vol. 120, p. 903–911, 2004.

[53] W. Humphrey, A. Dalke and K. Schulten, "VMD—visual molecular dynamics," *J. Mol. Graph.,* vol. 14, p. 33–38, 1996.

[54] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman and D. A. Case, "Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices," *J. Am. Chem. Soc.,* vol. 120, p. 9401–9409, 1998.

[55] Z. S. Hendsch and B. Tidor, "Electrostatic interactions in the GCN4 leucine zipper: substantial contributions arise from intramolecular interactions enhanced on binding," *Prot. Sci.,* vol. 8, p. 1381–1392, 1999.

[56] J. Lifson, S. Coutre´, E. Huang and E. Engleman, "Role of envelope glycoprotein carbohydrate in human immunodeficiency virus (HIV) infectivity and virus-induced cell fusion," *J. Exp. Med.,* vol. 164, p. 2101–2106, 1986.

[57] C. N. Scanlan, J. Offer, N. Zitzmann and R. A. Dwek, "Exploiting the defensive sugars of HIV-1 for drug and vaccine design," *Nature,* vol. 446, p. 1038–1045, 2007.

[58] C. A. Bewley and S. Otero-Quintero, "The potent anti-HIV protein cyanovirin-N contains two novel carbohydrate binding sites that selectively bind to Man8D1D3 and Man9 with nanomolar affinity: Implications for binding to the HIV envelope protein gp120," *J. Am. Chem. Soc.,* vol. 123, p. 3892–3902, 2001.

[59] C. A. Bewley, S. Kiyonaka and I. Hamachi, "Site-specific discrimination by cyanovirin-N for a-linked trisaccharides comprising the three arms of Man8 and Man9," *J. Mol. Biol.,* vol. 322, p. 881–889, 2002.

[60] C. A. Bewley, K. R. Gustafson, M. R. Boyd, D. G. Covell, A. Bax, G. M. Clore and A. M. Gronenborn, "Solution structure of cyanovirin-N, a potent HIV–inactivating protein," *Nat. Struct. Biol.,* vol. 5, p. 571–578, 1998.

[61] C. A. Bewley, "Solution structure of a cyanovirin-N:Manα1–2Mana complex: structural basis for high-affinity carbohydrate–mediated binding," *Structure,* vol. 9, p. 931–940, 2001.

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

[63] I. Botos, B. R. O'Keefe, S. R. Shenoy, L. K. Cartner, D. M. Ratner, P. H. Seeberger, M. R. Boyd and A. Wlodawer, "Structures of the complexes of a potent anti-HIV protein cyanovirin-N and high mannose oligosaccharides," *J. Biol. Chem.,* vol. 277, p. 34336–34342, 2002.

[64] C. C. Tsai, P. Emau, Y. Jiang, B. P. Tian, W. R. Morton, K. R. Gustafson and M. R. Boyd, "Cyanovirin-N as a topical microbicide prevents rectal transmission of SHIV89.6P in macaques," *AIDS Res. Hum. Retrov.,* vol. 19, p. 535–541, 2003.

[65] C. C. Tsai, P. Emau, Y. Jiang, M. B. Agy, R. J. Shattock, A. Schmidt, W. R. Morton, K. R. Gustafson and M. R. Boyd, "Cyanovirin-N inhibits AIDS virus infections in vaginal transmission models," *AIDS Res. Hum. Retrov.,* vol. 20, p. 11–18, 2004.

[66] C. J. Margulis, "Computational study of the dynamics of mannose disaccharides free in solution and bound to the potent anti-HIV virucidal protein cyanovirin," *J. Phys. Chem. B.,* vol. 109, p. 3639–3647, 2005.

[67] F. B. Sheinerman, R. Norel and B. Honig, "Electrostatic aspects of protein–protein interactions," *Curr. Opin. Struct. Biol.,* vol. 10, p. 153–159, 2000.

[68] T. Lazaridis, "Binding affinity and specificity from computational studies," *Curr. Org. Chem. ,* vol. 6, p. 1319–1332, 2002.

[69] M. K. Gilson and B. Honig, "Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis," *Proteins,* vol. 4, p. 7–18, 1988.

[70] D. Sitkoff, K. A. Sharp and B. Honig, "Accurate calcuation of hydration free energies using macroscopic solvent models," *J. Phys. Chem.,* vol. 98, p. 1978–1988, 1994.

[71] D. F. Green, "Optimized parameters for continuum solvation calculations with carbohydrates," *J. Phys. Chem. B.,* vol. 112, p. 5238–5249, 2008.

[72] M. Karplus and J. N. Kushick, "Method for estimating the configurational entropy of macromolecules," *Macromolecules,* vol. 14, p. 325–332, 1981.

[73] M. Kuttel, J. W. Brady and K. J. Naidoo, "Carbohydrate solution simulations: producing a force field with experimentally consistent primary alcohol rotational frequencies and populations," *J. Comput. Chem.,* vol. 23, p. 1236–1243, 2002.

[74] W. Im, M. S. Lee and C. L. Brooks, III, "Generalized–Born model with a simple smoothing function," *J. Comput. Chem.,* vol. 24, p. 1691–1702, 2003.

[75] M. Nina, D. Beglov and B. Roux, "Atomic radii for continuum electrostatic calculations based on molecular dynamics free energy simulaitons," *J. Phys. Chem. B.,* vol. 101, p. 5239–5248, 1997.

[76] M. D. Altman and B. Tidor, *MultigridPBE–software for computation and display of electrostatic potentials,* Boston: MIT, 2003.

[77] D. F. Green and B. Tidor, "Evaluation of electrostatic interactions," in *Current Protocols in Bioinformatics*, G. E. Petsko, Ed., New York, John Wiley & Sons, Inc., 2003, p. 8.3.1–8.3.16.

[78] D. F. Green, E. Kangas, Z. S. Hendsch and B. Tidor, *ICE—Integrated Continuum Electrostatics,* Boston: MIT, 2000.

[79] M. T. Esser, T. Mori, I. Mondor, Q. J. Sattentau, B. Dey, E. A. Berger, M. R. Boyd and J. D. Lifson, "Cyanovirin-N binds to gp120 to interfere with CD4-dependent human immunodeficiency virus type 1 virion binding, fusion, and infectivity but does not affect the CD4 binding site on gp120 or soluble CD4-induced conformational changes in gp120," *J. Virol.,* vol. 73, p. 4360–4371, 1999.

[80] B. Dey, D. L. Lerner, P. Lusso, M. R. Boyd, J. H. Elder and E. A. Berger , "Multiple antiviral activities of cyanovirin-N: blocking of human immunodeficiency virus type 1 gp120 interaction with CD4 and coreceptor and inhibition of diverse enveloped viruses," *J. Virol.,* vol. 74, p. 4562–4569, 2000.

[81] I. McGowan, "Microbicides: a new frontier in HIV prevention," *Biologicals,* vol. 34, p. 241–255, 2006.

[82] P. J. Klasse, R. Shattok and J. P. Moore, "Antiretroviral drug-based microbicides to prevent HIV-1 sexual transmission," *Annu. Rev. Med.,* vol. 59, p. 455–471, 2008.

[83] L. G. Barrientos, B. R. O'Keefe, M. Bray and A. Sanchez, "Cyanovirin-N binds to the viral surface glycoprotein, GP1,2 and inhibits infectivity of Ebola virus," *Antivir. Res.,* vol. 58, p. 47–56, 2003.

[84] V. Tiwari, S. Y. Shukla and D. Shukla , "A sugar binding protein cyanovirin-N blocks herpes simplex virus type-1 entry and cell fusion," *Antivir. Res.,* vol. 84, p. 67–75, 2009.

[85] I. Massova and P. A. Kollman, "Computational alanine scanning to probe protein–protein interactions: a novel approach to evaluate binding free energies," *J. Am. Chem. Soc.,* vol. 121, p. 8133–8143, 1999.

[86] G. Archontis, T. Simonson and M. Karplus, "Binding free energies and free energy components from molecular dynamics and Poisson–Boltzmann calculations. Application to amino acid recognition by aspartyl-tRNA synthetase," *J. Mol. Biol.,* vol. 306, p. 307–327, 2001.

[87] H. Gohlke and D. A. Case, "Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf," *J. Comput. Chem.,* vol. 25, p. 238–250, 2004.

[88] J. Chocholousova and M. Feig, "Implicit solvent simulations of DNA and DNA-protein complexes: agreement with explicit solvent versus experiment," *J. Phys. Chem. B.,* vol. 110, p. 17240–17251, 2006.

[89] B. Strockbine and R. C. Rizzo, "Binding of antifusion peptides with HIV gp41 from molecular dynamics simulaitons: quantitative correlation with experiment," *Proteins: Struct., Func., Bioinf.,* vol. 67, p. 630–642, 2007.

[90] Y. K. Fujimoto, R. N. TerBush, V. Patsalo and D. F. Green, "Computational models explain the oligosaccharide specificity of cyanovirin-N," *Prot. Sci.,* vol. 17, p. 2008–2014, 2008.

[91] B. I. Dahiyat and S. L. Mayo, "De novo protein design: fully automated sequence selection," *Science,* vol. 278, p. 82–87, 1997.

[92] C. A. Sarkar, K. Lowenhaupt, T. Horan, T. C. Boone, B. Tidor and D. A. Lauffenbuger, "Rational cytokine design for increased lifetime and enhanced potency using pH-activated hisitine switching," *Nat. Biotechnol.,* vol. 20, p. 908–913, 2002.

[93] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard and D. Baker, "Design of a novel globular protein fold with atomic-level accuracy," *Science ,* vol. 302, p. 1364–1368, 2003.

[94] L. L. Looger, M. A. Dwyer, J. J. Smith and H. W. Hellinga, "Computational design of receptor and sensor proteins with novel functions," *Nature,* vol. 423, p. 185–190, 2003.

[95] J. Ashworth , J. J. Havranek, C. M. Duarte, D. Sussman, R. J. Monnat, Jr., B. L. Stoddard and D. Baker, "Computational redesign of endonuclease DNA binding and cleavage specificity," *Nature,* vol. 441, p. 656–659, 2006.

[96] D. F. Green, A. T. Dennis, P. S. Fam, B. Tidor and A. Jasanoff, "Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide," *Biochemistry,* vol. 45, p. 12547–12559, 2006.

[97] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale and K. Schulten, "Scalable molecular dynamics with NAMD," *J. Comput. Chem.,* vol. 26, p. 1781–1802, 2005.

[98] D. F. Green and B. Tidor, *Current Protocols in Bioinformatics*, G. E. Petsko, Ed., New York, John Wiley & Sons, Inc., 2003, p. 8.3.1–8.3.16.

[99] D. F. Green and B. Tidor, "Design of improved protein inhibitors of HIV-1 cell entry: optimization of electrostatic interactions at the binding interface," *Proteins: Struct., Funct., Bioinf.,* vol. 60, p. 644–657, 2005.

[100] N. Carrascal and D. F. Green, "Energetic decomposition with the Generalized–Born and Poisson–Boltzmann solvent models: lessons from association of G-protein components," *J. Phys. Chem. B.,* vol. 114, p. 5096–5116, 2010.

[101] B. Roux and T. Simonson, "Implicit solvent models," *Biophys. Chem.,* vol. 78, p. 1–20, 1999.

[102] K. Dill, "Dominant forces in protein folding," *Biochemistry–U.S.,* vol. 29, p. 7133–7155, 1990.

[103] X. Siebert and G. Hummer, "Hydrophobicity maps of the N-peptide coiled coil of HIV-1 gp41," *Biochemistry,* vol. 41, p. 2956–2961, 2002.

[104] R. M. Levy, L. Y. Zhang, E. Gallicchio and A. K. Felts, "On the nonpolar hydration free energy of proteins: surface area and continuum solvent models for the solute−solvent interaction energy," *J. Am. Chem. Soc.,* vol. 125, p. 9523–9530, 2003.

[105] J. A. Wagoner and N. A. Baker, "Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 103, p. 8331–8336, 2006.

[106] E. Gallicchio, L. Y. Zhang and R. M. Levy, "The SGB/NP hydration free energy model based on the surface Generalized–Born solvent reaction field and novel nonpolar hydration free energy estimators," *J. Comput. Chem.,* vol. 23, p. 517–529, 2002.

[107] J. W. Pitera and W. F. van Gunsteren, "The importance of solute–solvent van der Waals interactions with interior atoms of biopolymers," *J. Am. Chem. Soc.,* vol. 123, p. 3163–3164, 2001.

[108] J. Wagoner and N. A. Baker, "Solvation forces on biomolecular structures: a comparison of explicit solvent and Poisson–Boltzmann models," *J. Comput. Chem.,* vol. 25, p. 1623–1629, 2004.

[109] J. P. Bardham, M. D. Altman, S. M. Lippow, B. Tidor and J. K. White, "A curved panel integration technique for molecular surfaces," *NSTI-Nanotech.,* vol. 1, p. 512–515, 2005.

[110] J. P. Bardham, M. D. Altman, D. J. Willis, S. M. Lippow, B. Tidor and J. K. White, "Numerical integration techniques for curved–element discretizations of molecule–solvent interfaces," *J. Chem. Phys.,* vol. 127, pp. 014701-1–014701-18, 2007.

[111] S. M. Lippow and B. Tidor, *Continuum van der Waals software,* Boston: MIT.

[112] C. E. Chang, W. Chen and M. K. Gilson, "Evaluating the accuracy of the quasiharmonic approximation," *J. Chem. Theory Comput.,* vol. 1, p. 1017–1028, 2005.

[113] D. M. LeMaster, "NMR relaxation order parameter analysis of the dynamics of protein side chains," *J. Am. Chem. Soc.,* vol. 121, p. 1726–1742, 1999.

[114] D. C. Chan and P. S. Kim, "HIV entry and its inhibition," *Cell,* vol. 93, p. 681–684, 1998.

[115] R. Percudani, B. Montanini and S. Ottonello, "The anti-HIV cyanovirin-N domain is evolutionarily conserved and occurs as a protein module in eukaryotes," *Proteins,* vol. 60, p. 670–678, 2005.

[116] L. M. Koharudin, A. R. Viscomi, J.-G. Jee, S. Ottonello and A. M. Gronenborn, "The evolutionary conserved family of cyanovirin-N homologs: structures and carbohydrates specificity," *Structure,* vol. 16, p. 570–584, 2008.

[117] J. C. Kehr, Y. Zilliges, A. Springer, M. D. Disney, D. D. Ratner, C. Bouchier, P. H. Seeberger, N. T. de Marsac and E. Dittmann, "A mannan binding lectin is involved in cell–cell attachment in a toxic strain of Microcystis aeruginosa," *Mol. Microbiol.,* vol. 59, p. 893–906., 2005.

[118] D. Huskens, G. Férir, K. Vermeire, J. C. Kehr, J. Balzarini, E. Dittmann and D. Schols, "Microvirin, a Novel α(1,2)-mannose-specific lectin isolated from microcystis aeruginosa, has anti-HIV-1 activity comparable with that of cyanovirin-N but a much higher safety profile," *J. Biol. Chem.,* vol. 285, p. 24845–24854, 2010.

[119] S. Shahzad-ul-Hussan, E. Gustchina, R. Ghirlando, G. M. Clore and C. A. Bewley, "Solution structure of the monovalent lectin microvirin in complex with Manα(1-2)Man provides a basis for anti-HIV activity with low toxicity," *J. Biol. Chem.,* vol. 286, p. 20788–20796, 2011.

[120] *MATLAB R2010b (Version 7.11.0.584),* The MathWorks, Inc..

[121] National Institute of Allergy and Infectious Diseases (NIAID), [Online]. Available: http://www3.niaid.nih.gov/. [Accessed 2012].

[122] "Centre de Recherches sur les Macromolécules Végétales,," Centre National de la Recherche Scientifique (Cermav-CNRS), [Online]. Available: http://www.cermav.cnrs.fr/lectines/.. [Accessed 2012].

[123] D. R. Burton, "Structural biology: images from the surface of HIV," *Nature,* vol. 441, p. 817–818, 2006.

[124] M. Chaplin, "Water Models," [Online]. Available: http://www.lsbu.ac.uk/water/models.html. [Accessed 2012].

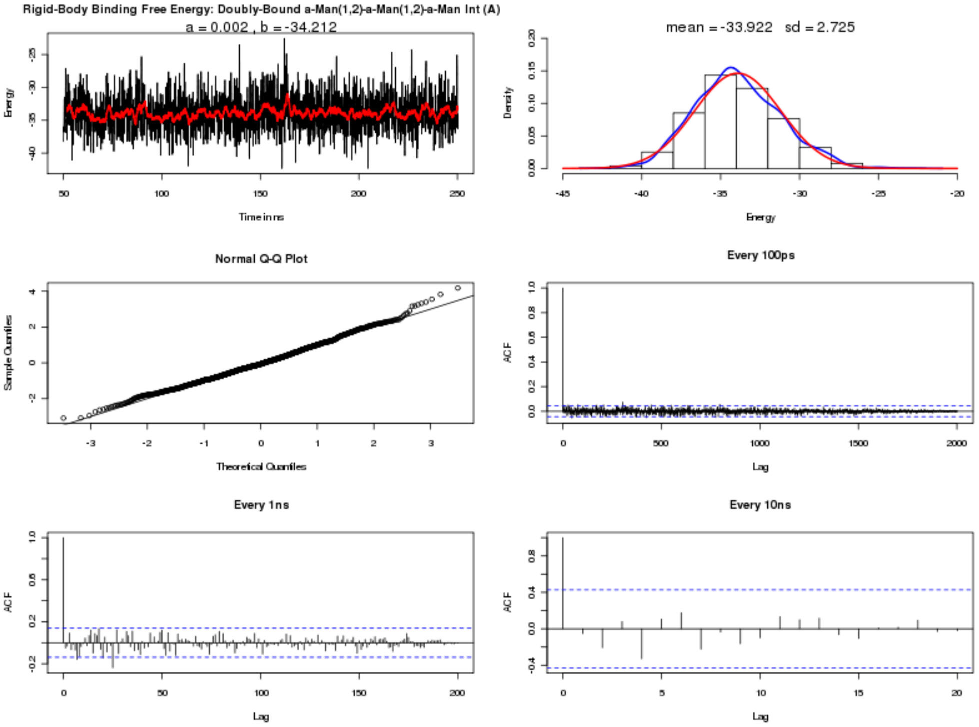
[125] K. Hornik, "R," [Online]. Available: http://www.r-project.org/. [Accessed 2012].

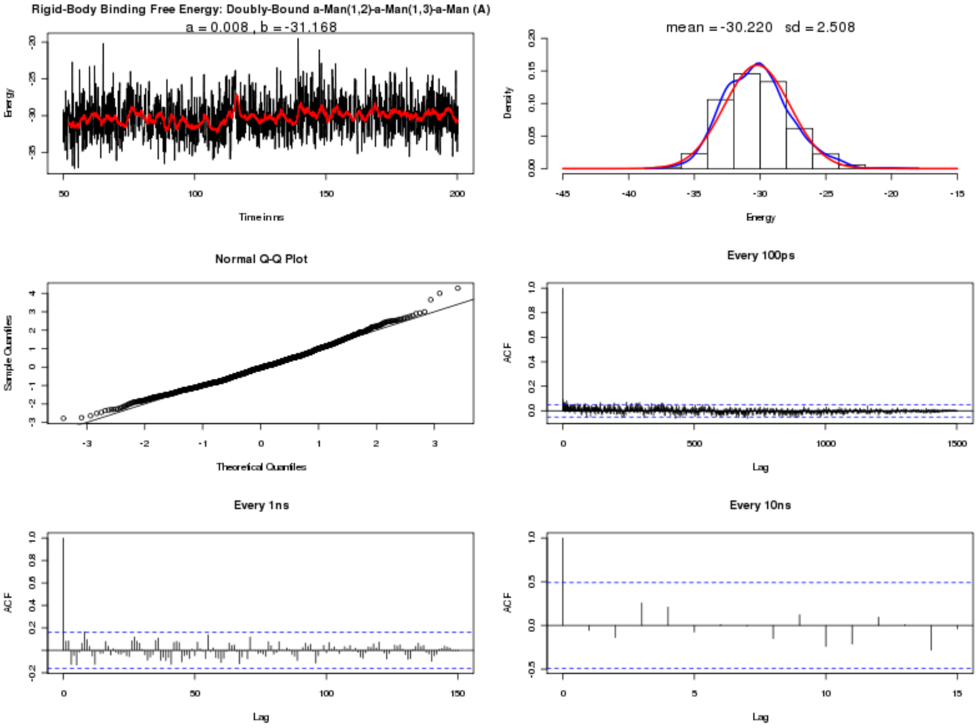
Appendix A.

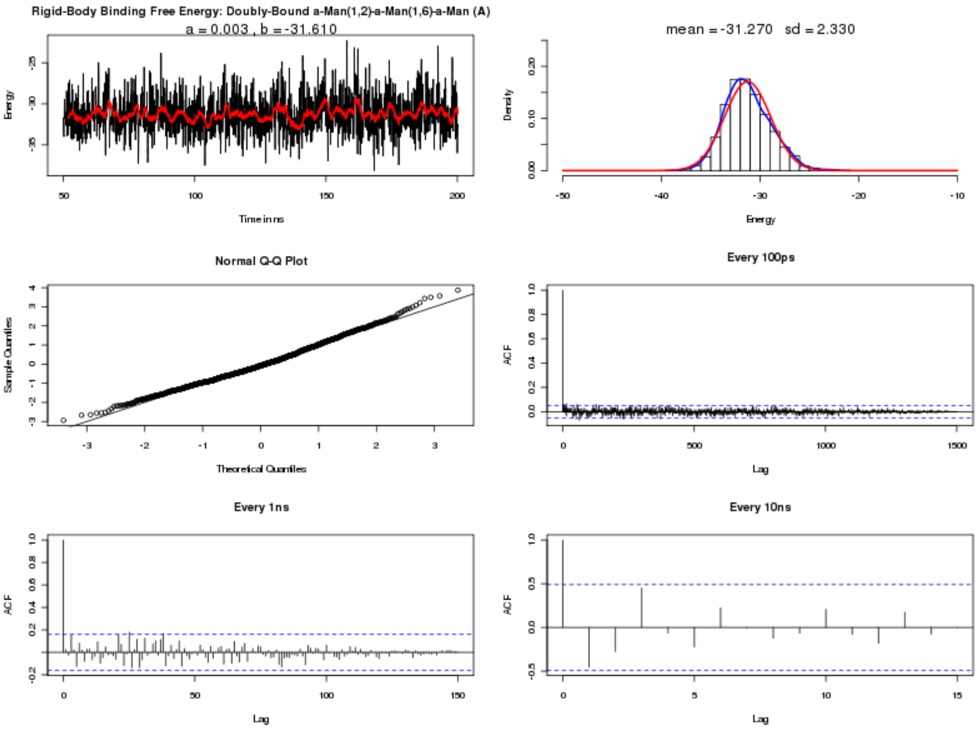
Time Dependent Computed Energies for Trisaccharide−CVN Bound Models

The following sections show figures of time-dependent plots of computed energies from doubly and singly bound simulations in both domain A and domain B. All figures are based on simulations starting from 50 ns. *(Top left)* A plot of energy versus time for every snapshot (1500 snapshots). The red line represents the running average of all 1500 snapshots. The slope (a) and the y-intercept (b) is computed above the plot. *(Top right)* A histogram of the data. The blue line represents the density of the data and the red line represents the normal distribution of the data. The average of the energy and standard deviation is computed above the histogram. *(Middle left)* A plot of the quantile-quantile. *(Middle right)* A plot of the autocorrelation for every 100 ps. *(Bottom left)* A plot of the autocorrelation for every 1 ns. *(Bottom right)* A plot of the autocorrelation for every 10 ns.

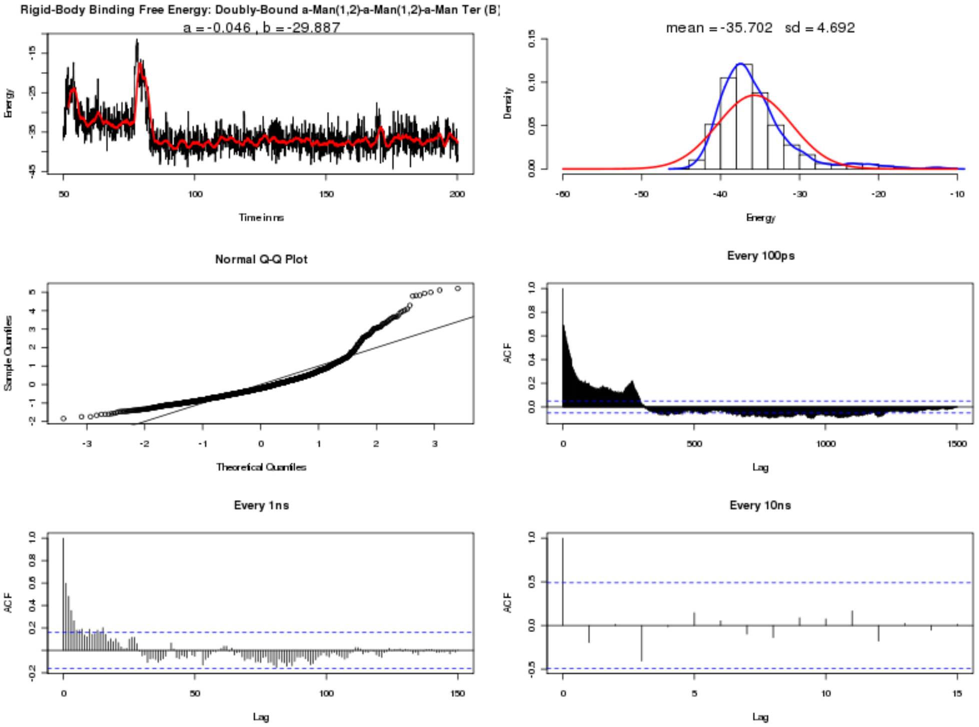
A.1 Doubly Bound Simulations (Domain A).

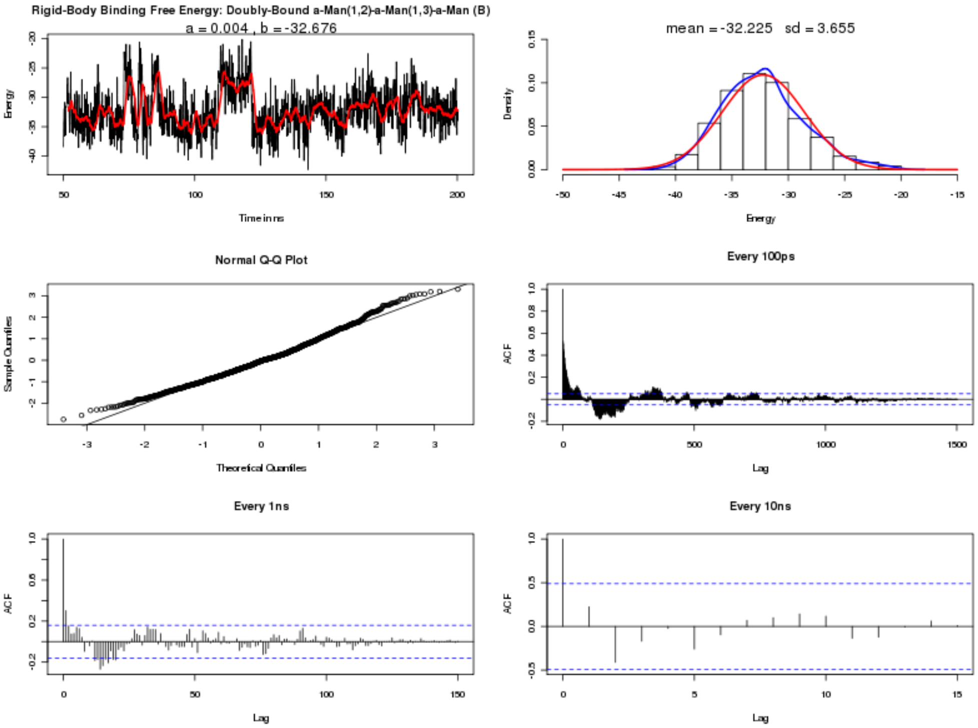


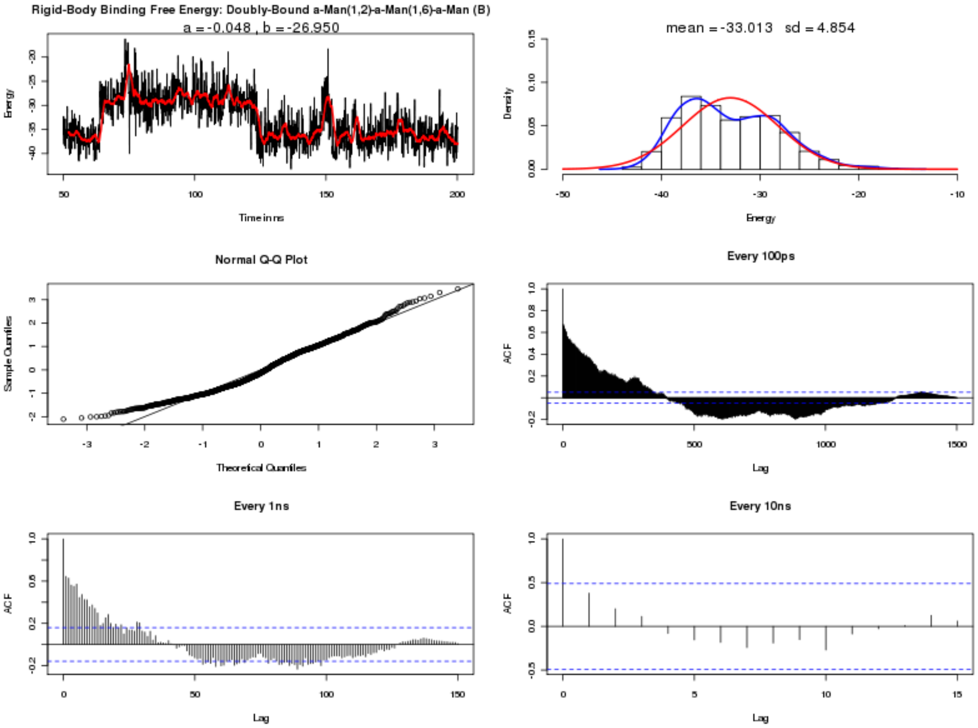




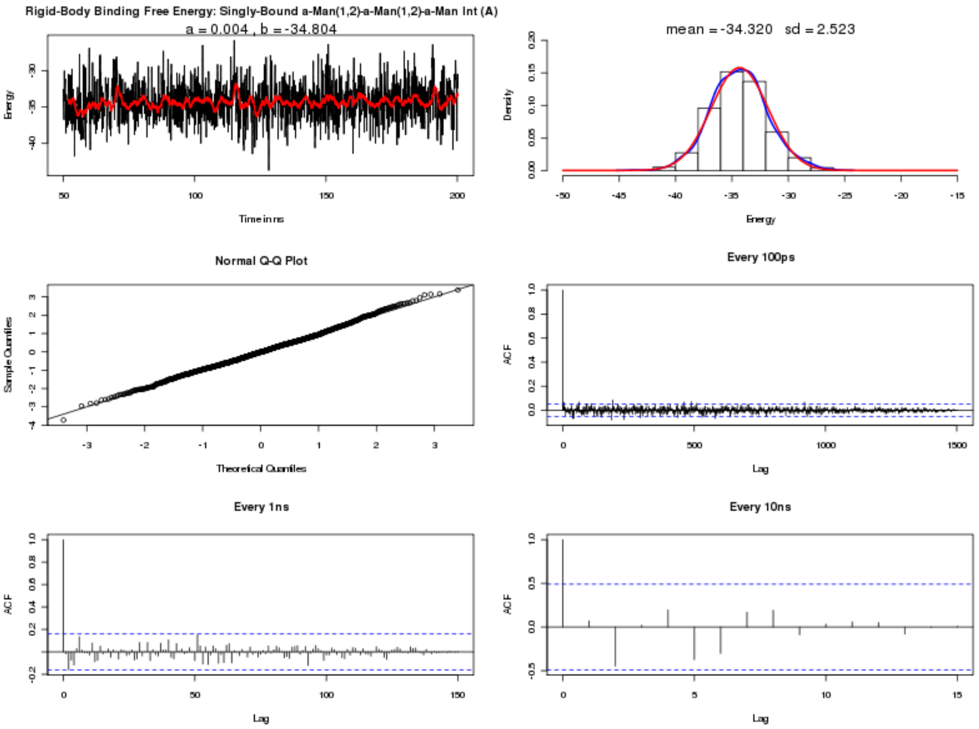
A.2 Doubly Bound Simulations (Domain B).

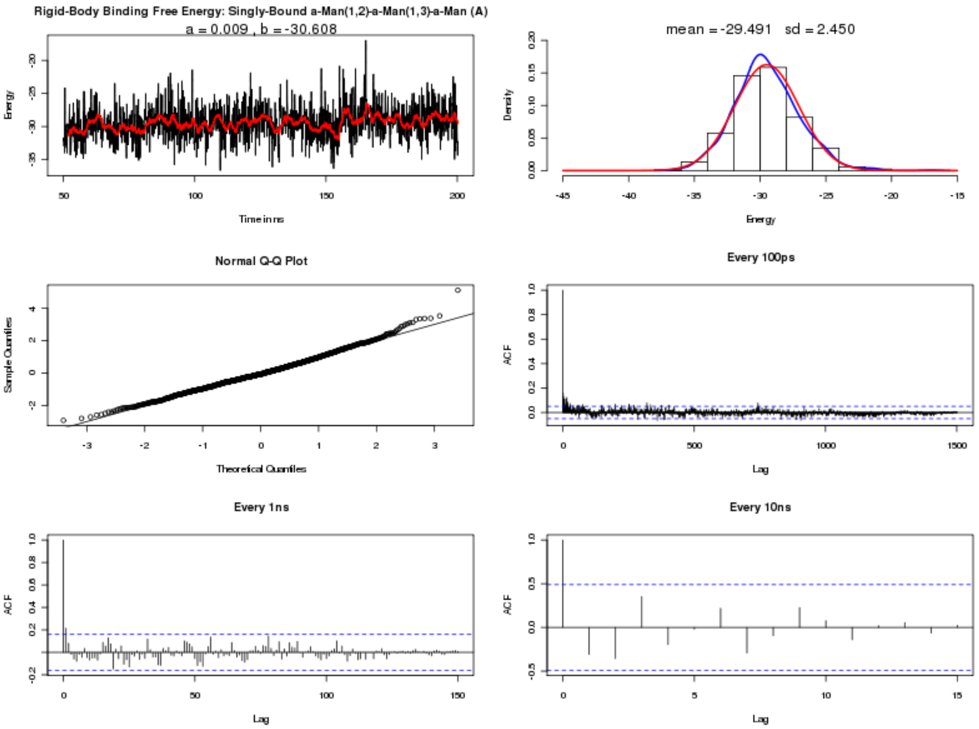


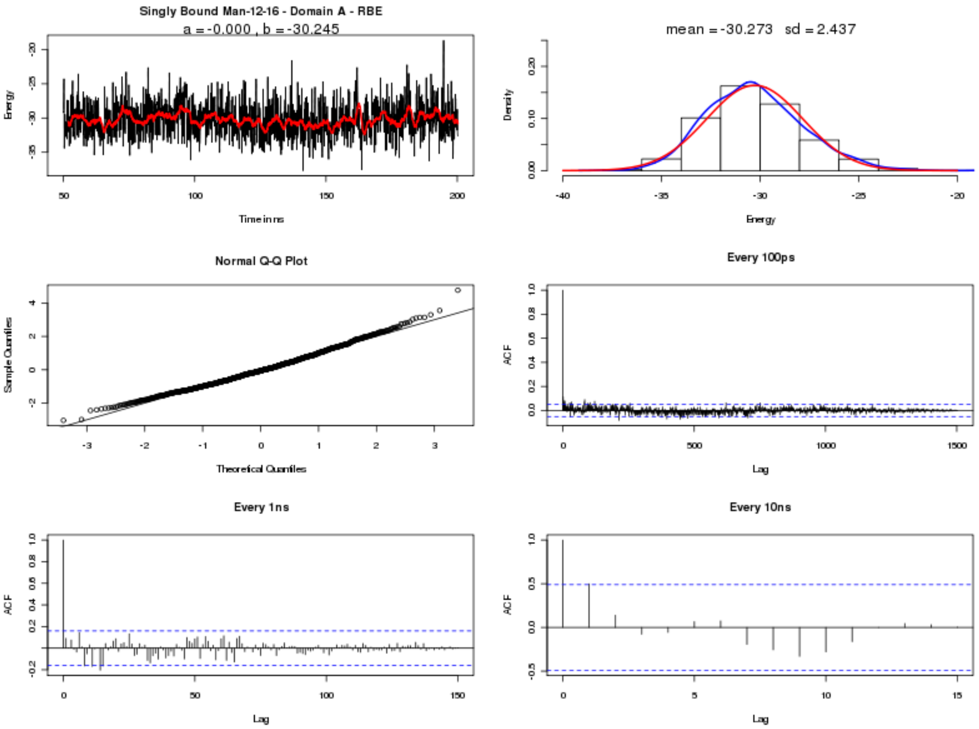

Rigid-Body Binding Free Energy: Doubly-Bound a-Man(1,2)-a-Man(1,3)-a-Man (B)
a = 0.004 , b = -32.676
Energy
Time in ns
-20
-25
-30
-35
-40
50
100
150
200
mean = -32.225   sd = 3.655
Density
Energy
0.15
0.10
0.05
0.00
-50
-45
-40
-35
-30
-25
-20
-15
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
3
Every 100ps
ACF
Lag
1.0
0.8
0.6
0.4
0.2
-0.2
0
500
1000
1500
Every 1ns
ACF
Lag
1.0
0.6
0.2
-0.2
0
50
100
150
Every 10ns
ACF
Lag
1.0
0.5
0.0
-0.5
0
5
10
15


Rigid-Body Binding Free Energy: Doubly-Bound a-Man(1,2)-a-Man(1,6)-a-Man (B)
a = -0.048 , b = -26.950
Energy
Time in ns
-20
-25
-30
-35
-40
50
100
150
200
mean = -33.013   sd = 4.854
Density
Energy
0.15
0.10
0.05
0.00
-50
-40
-30
-20
-10
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
3
Every 100ps
ACF
Lag
1.0
0.8
0.6
0.4
0.2
-0.2
0
500
1000
1500
Every 1ns
ACF
Lag
1.0
0.6
0.2
-0.2
0
50
100
150
Every 10ns
ACF
Lag
1.0
0.5
0.0
-0.5
0
5
10
15

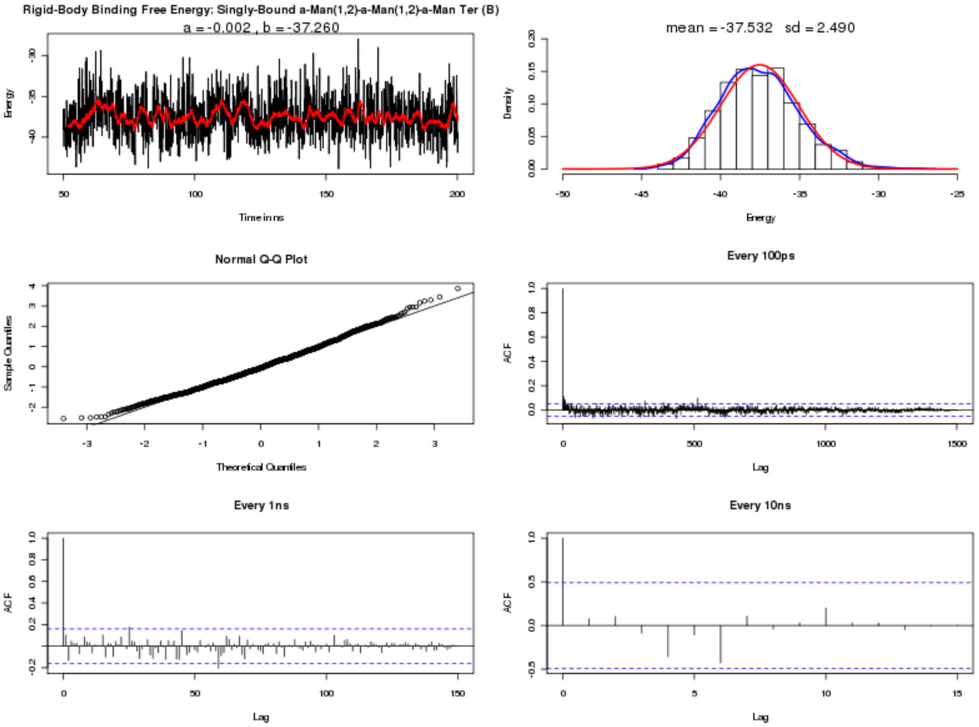
A.3 Singly Bound Simulations (Domain A).

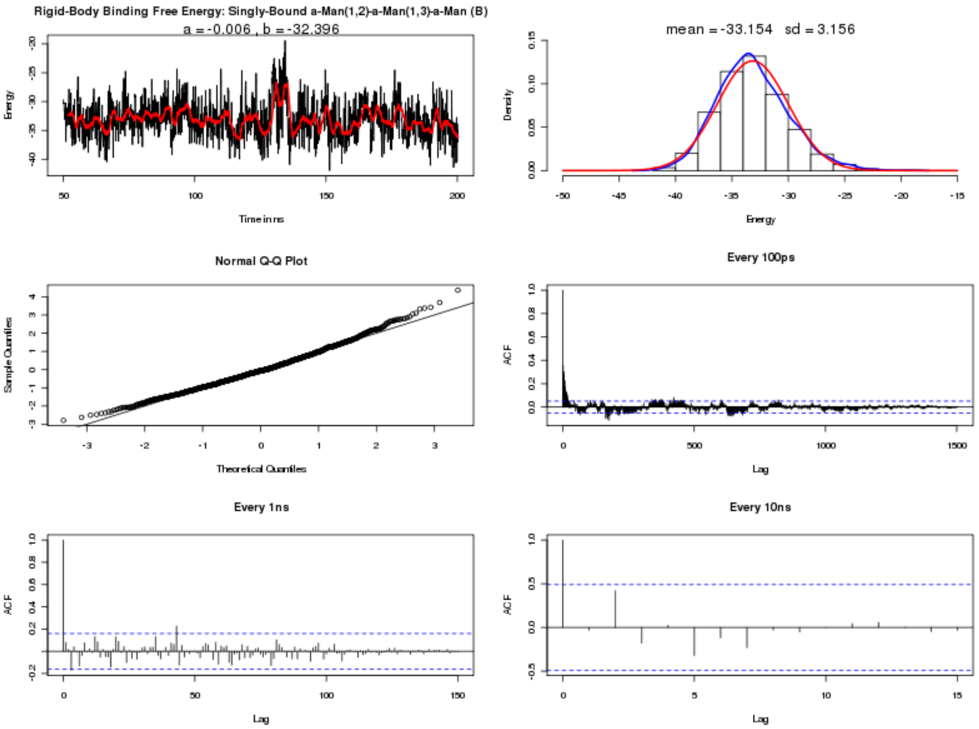


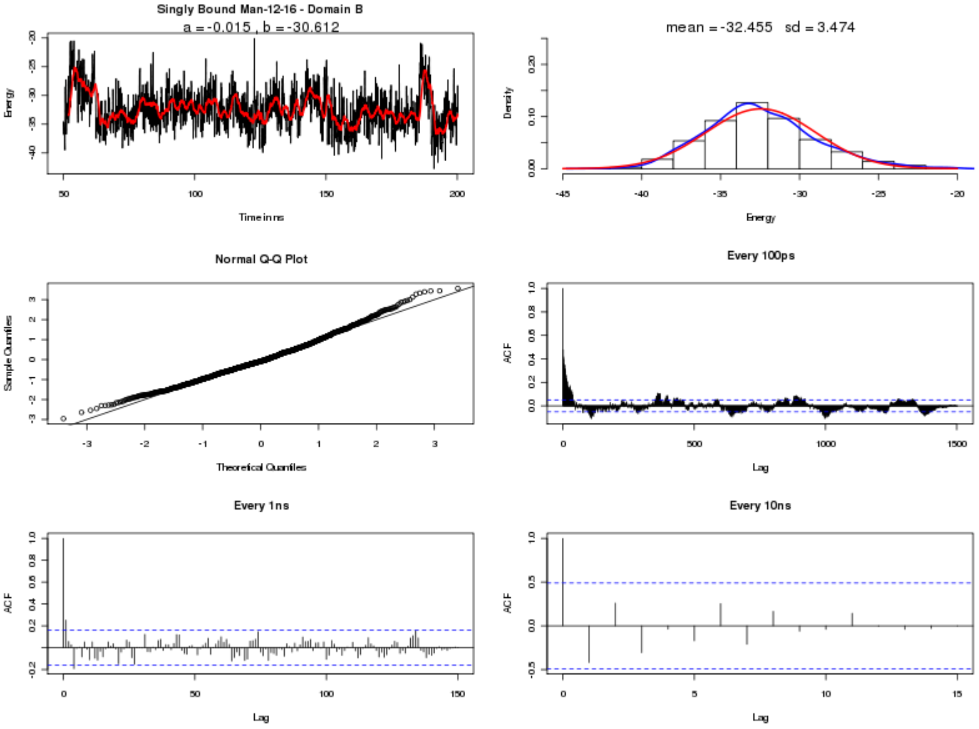




A.4 Singly Bound Simulations (Domain B).




Rigid-Body Binding Free Energy: Singly-Bound a-Man(1,2)-a-Man(1,3)-a-Man (B)
a = -0.006 , b = -32.396
Energy
-20
-25
-30
-35
-40
50
100
150
200
Time in ns
mean = -33.154   sd = 3.156
Density
0.15
0.10
0.05
0.00
-50
-45
-40
-35
-30
-25
-20
-15
Energy
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Every 100ps
ACF
Lag
0
500
1000
1500
Every 1ns
ACF
Lag
0
50
100
150
Every 10ns
ACF
Lag
0
5
10
15


Singly Bound Man-12-16 - Domain B
a = -0.015 , b = -30.612
Energy
-20
-25
-30
-35
-40
50
100
150
200
Time in ns
mean = -32.455   sd = 3.474
Density
0.20
0.10
0.00
-45
-40
-35
-30
-25
-20
Energy
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Every 100ps
ACF
Lag
0
500
1000
1500
Every 1ns
ACF
Lag
0
50
100
150
Every 10ns
ACF
Lag
0
5
10
15