

2018

UNDERSTANDING CARBOHYDRATE RECOGNITION MECHANISMS IN NON-CATALYTIC PROTEINS THROUGH MOLECULAR SIMULATIONS

Abhishek A. Kognole

University of Kentucky, abhishek.kognole@uky.edu

Author ORCID Identifier:

<https://orcid.org/0000-0002-6033-2344>

Digital Object Identifier: <https://doi.org/10.13023/ETD.2018.019>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Kognole, Abhishek A., "UNDERSTANDING CARBOHYDRATE RECOGNITION MECHANISMS IN NON-CATALYTIC PROTEINS THROUGH MOLECULAR SIMULATIONS" (2018). *Theses and Dissertations--Chemical and Materials Engineering*. 80.

https://uknowledge.uky.edu/cme_etds/80

This Doctoral Dissertation is brought to you for free and open access by the Chemical and Materials Engineering at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Chemical and Materials Engineering by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Abhishek A. Kognole, Student

Dr. Christina M. Payne, Major Professor

Dr. Thomas Dziubla, Director of Graduate Studies

UNDERSTANDING CARBOHYDRATE RECOGNITION MECHANISMS IN
NON-CATALYTIC PROTEINS THROUGH MOLECULAR SIMULATION

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Engineering
at the University of Kentucky

By

Abhishek Anil Kognole

Lexington, Kentucky

Co-Directors: Dr. Christina M. Payne, Professor of Chemical and Materials Engineering,
and Dr. Bradley J. Berron, Professor of Chemical and Materials Engineering,

Lexington, Kentucky

Copyright © Abhishek Anil Kognole 2017

ABSTRACT OF DISSERTATION

UNDERSTANDING CARBOHYDRATE RECOGNITION MECHANISMS IN NON-CATALYTIC PROTEINS THROUGH MOLECULAR SIMULATION

Non-catalytic protein-carbohydrate interactions are an essential element of various biological events. This dissertation presents the work on understanding carbohydrate recognition mechanisms and their physical significance in two groups of non-catalytic proteins, also called lectins, which play key roles in major applications such as cellulosic biofuel production and drug delivery pathways. A computational approach using molecular modeling, molecular dynamic simulations and free energy calculations was used to study molecular-level protein-carbohydrate and protein-protein interactions. Various microorganisms like bacteria and fungi secrete multi-modular enzymes to deconstruct cellulosic biomass into fermentable sugars. The carbohydrate binding modules (CBM) are non-catalytic domains of such enzymes that assist the catalytic domains to recognize the target substrate and keep it in proximity. Understanding the protein-carbohydrate recognition mechanisms by which CBMs selectively bind substrate is critical to development of enhanced biomass conversion technology. We focus on CBMs that target both oligomeric and non-crystalline cellulose while exhibiting various similarities and differences in binding specificity and structural properties; such CBMs are classified as Type B CBMs. We show that all six cellulose-specific Type B CBMs studied in this dissertation can recognize the cello-oligomeric ligands in bi-directional fashion, meaning there was no preference towards reducing or non-reducing end of ligand for the cleft/groove like binding sites. Out of the two sandwich and twisted forms of binding site architectures, twisted platform turned out to facilitate tighter binding also exhibiting longer binding sites. The exterior loops of such binding sites were specifically identified by modeling the CBMs with non-crystalline cellulose showing that high- and low-affinity binding site may arise based on orientation of CBM while interacting with non-crystalline substrate. These findings provide various insights that can be used for further understanding of tandem CBMs and for various CBM based biotechnological applications.

The later part of this dissertation reports the identification of a physiological ligand for a mammalian glycoprotein YKL-40 that has been only known as a biomarker in various inflammatory diseases and cancers. It has been shown to bind to oligomers of chitin, but there is no known function of YKL-40, as chitin production in the human body has never been reported. Possible alternative ligands include proteoglycans, polysaccharides, and fibers such as collagen, all of which make up the mesh comprising the extracellular matrix. It is likely that YKL-40 is interacting with these alternative polysaccharides or proteins within the body, extending its function to cell biological roles such as mediating cellular receptors and cell adhesion and migration. We considered the feasibility of polysaccharides, including cello-oligosaccharides, hyaluronan, heparan sulfate, heparin, and chondroitin sulfate, and collagen-like peptides as physiological ligands for YKL-40. Our simulation results suggest that chitohexaose and hyaluronan preferentially bind to YKL-40 over collagen, and hyaluronan is likely the preferred physiological ligand, as the negatively charged hyaluronan shows enhanced affinity for YKL-40 over neutral chitohexaose. Collagen binds in two locations at the YKL-40 surface, potentially related to a role in fibrillar formation. Finally, heparin non-specifically binds at the YKL-40 surface, as predicted from structural studies. Overall, YKL-40 likely binds many natural ligands *in vivo*, but its concurrence with physical maladies may be related to the associated increases in hyaluronan.

KEYWORDS: protein-carbohydrate interaction, cellulose, glycosaminoglycan, collagen, molecular dynamics, free energy perturbation.

Abhishek A. Kognole

08/15/2017

UNDERSTANDING CARBOHYDRATE RECOGNITION MECHANISMS IN NON-
CATALYTIC PROTEINS THROUGH MOLECULAR SIMULATION

By

Abhishek A. Kognole

Dr. Christina M. Payne

Co-Director of Dissertation

Dr. Bradley J. Berron

Co-Director of Dissertation

Dr. Thomas Dziubla

Director of Graduate Studies

September 25th, 2017

To my beloved parents

Acknowledgements

As I look back over my Ph.D., it was not about only the last 5 years, and I believe it has been a great wonderful journey since the day I started school. Needless to say, I could not have finished it without support and guidance of people in my life. First of all, my parents, they are the best parents one could ever ask for and their love, care and teachings have made me strong and keep me going. I cannot thank them enough! I am equally grateful to my sister, Akshata, who always pushes me to dream bigger but also, makes sure that my feet are on the ground. I am grateful to my grandparents for their unconditional love and the virtues I inherited from them that make me a better person. I have to thank my relatives for being such a kind and gracious family that always makes me proud and supports in my hard times. I wish to sincerely acknowledge all my teachers and mentors from my school and undergraduate college who are elemental architects of my scientific knowledge.

With all my heart, I would like to express my very great appreciation to my doctoral advisor, Dr. Christina M. Payne, for her exceptional guidance and unwavering support during my PhD. Since the day she welcomed me into her newly founded research group, she has made a long lasting impact on my scientific abilities and technical skills. I am grateful to her for providing amazing research opportunities and plenty of resources, and giving the freedom to pursue my research endeavors. I would like to offer my special thanks to Dr. Mats Sandgren from Swedish University of Agricultural Sciences for being a delightful host, a valuable mentor, and also presenting a great opportunity to connect with structural biologists. I would also like to thank my dissertation committee members, Dr. Brad Berron, Dr. Tate Tsang, and Dr. Matthew Gentry for their feedback and

recommendations on this research, and for dedicating their valuable time to my dissertation. I honestly appreciate the efforts made by the department staff, the DGS and the department head to not let the tedious paperwork come in the way of my research.

My past and present colleagues from our lab, Suvamay Jana, Amira (Yue Yu), and Japheth have always maintained a friendly environment to work. I would like to thank them for sharing their skills and knowledge with me. Particularly Jana, with whom, I have discussed a whole lot of scientific ideas and problems, as well as random general things in life. Thanks are not enough to appreciate the great friendships that I have been enjoying during my days here in Lexington. I shall be ever grateful to all my friends in Lexington, Vinod, Suraj, Saket, Prachi, Raghav, Priyesh, Ashish, Shreya, and specially Ishan for being the perfect roommate for last 5 years, and all the members of *Virginia Mandal* as we call ourselves (most of us live on Virginia Avenue). I genuinely cannot imagine surviving without all the fun that we had during festivals, trips, birthdays and almost everyday cooking great food and hanging out sharing every details of our lives. I would also like to thank Ravinder and Sai for their warm friendship and counsel.

Lastly, I would like to acknowledge the financial support provided by National Science Foundation (NSF) (CHE-1404849) for the carbohydrate binding module studies and Kentucky Science and Engineering Foundations (Grant KSEF-148-502-13-307) for the YKL-40 research. The NSF Extreme Science and Engineering Discovery Environment (XSEDE) and University of Kentucky provided the computational time used for this research.

Table of Contents

Acknowledgements	iii
Table of Contents	v
List of Tables	xi
List of Figures	xiii
List of Additional Files	xx
Chapter 1 – Introduction.....	1
1.1 Carbohydrate-binding proteins.....	1
1.2 Motivation.....	3
1.2.1 Sustainable cellulosic bioethanol.....	3
1.2.2 Lectin-mediated targeted drug delivery	5
1.3 Research background	6
1.3.1 Carbohydrate binding modules.....	6
1.3.2 Mammalian glycoprotein YKL-40	23
1.4 Outline of Dissertation.....	30

Chapter 2 – Computational methods	31
2.1 Pre-dynamics tasks	31
2.1.1 Homology modeling	32
2.1.2 Molecular Docking	35
2.2 Molecular dynamics (MD) Simulations	39
2.3 Free energy calculations	42
2.3.1 Free Energy Perturbation with Hamiltonian Replica Exchange Molecular Dynamics (FEP/ λ -REMD).....	42
2.3.2 Umbrella Sampling	45
Chapter 3 – Cello-oligomer binding dynamics and bi-directional binding phenomenon in Type B CBMs	48
3.1 Abstract.....	48
3.2 Introduction.....	49
3.3 Methods.....	53
3.3.1 Modeling of cello-oligomer in multiple orientations.....	53
3.3.2 MD simulation: Setup and parameters.....	54

3.3.3	Free Energy Calculation: FEP/ λ -REMD	57
3.4	Results and Discussion.....	59
3.4.1	Symmetry of the cellopentaose is critical to binding.....	59
3.4.2	Thermodynamic preference of cello-oligomer orientation	62
3.4.3	CfCBM4-1 hydrogen bonding	66
3.4.4	CfCBM4-1 dynamics	68
3.4.5	CfCBM4-2 dynamics	75
3.4.6	Evidence of bi-directional binding beyond <i>C. fimi</i> CBM4s	79
3.4.7	Bi-directional binding extends to family 17 and 28 CBMs	84
3.5	Conclusions.....	89
 Chapter 4 – Role of binding site architecture and recognition of non-crystalline		
cellulose in Type B Carbohydrate Binding Modules		91
4.1	Introduction.....	91
4.2	Methods and materials	94
4.2.1	Modeling protein-carbohydrate complexes	94
4.2.2	MD simulation parameters and protocols	100

4.2.3	Free energy calculations	102
4.3	Results and discussion	104
4.3.1	Role of binding site architecture in substrate recognition	104
4.3.2	Differentiation of high and low affinity binding sites non-crystalline cellulose	119
4.4	Conclusions.....	128
 Chapter 5 – Carbohydrate ligands of YKL-40: Binding mechanisms, thermodynamic preferences and surface binding ability.....		
5.1	Abstract.....	130
5.2	Introduction.....	131
5.3	Methods.....	136
5.3.1	Molecular Dynamics Simulation	136
5.3.2	Free Energy Calculations: FEP/ λ -REMD	140
5.4	Results and Discussion.....	144
5.4.1	Protein-polysaccharide binding in YKL-40.....	144
5.4.2	Putative heparin-binding site	147

5.4.3	Polysaccharide ligand binding affinity	150
5.4.4	Polysaccharide Binding Dynamics	153
5.4.5	Conformational changes in the YKL-40 binding site	162
5.5	Conclusions.....	164
 Chapter 6 – Protein-protein interactions of YKL-40: Identification and		
characterization of collagen binding sites.....		165
 6.1	Introduction.....	165
 6.2	Methods.....	168
6.2.1	Docking of collagen triple helix on YKL-40.....	168
6.2.2	Molecular Dynamics Simulation	170
6.2.3	Free Energy Calculations: Umbrella Sampling	171
 6.3	Results and Discussion.....	173
6.3.1	Protein-protein binding in YKL-40	173
6.3.2	Ligand Binding Dynamics and Comparison of Model Collagen-like Peptides. 173	
6.3.3	Collagen-like peptide binding affinity	185

6.4 Conclusions	187
Chapter 7 – Conclusions and future work.....	189
Appendix	198
A1 Supporting information related to CBMs.....	198
A1.1 Additional methods for Chapter 3.....	198
A1.2 Additional results	201
A2 Supporting information related to YKL-40.....	204
A2.1 Force-field parameterization for modeling heparin	204
References	207
Vita	229

List of Tables

Table 3.1 Binding free energies of cellopentaose to <i>Cj</i> CBM4-1 in two ligand orientations representing bi-directional binding. The solvation free energy of cellopentaose, ΔG_2 , is also tabulated as its three contributions – repulsion, dispersion, and electrostatics.....	63
Table 3.2 CBM structures with β -sandwich fold compared with <i>Cj</i> CBM4-1-RE (PDB ID 1GU3). ‘Same’ refers to the direction of ligand that is equivalent to that in 1GU3.....	81
Table 4.1 List of all the MD simulations performed in this study with length of MD simulations and free energy calculation method.....	95
Table 4.2 Distribution of free energy components of cellopentaose (G5) binding to CBMs at 300K and pH 7.....	107
Table 4.3 Percent occupancy of each hydrogen bond formed between the pyranose ring at each binding site and the surrounding protein residue over the 250 ns simulation....	112
Table 5.1 Simulations and calculations performed in the investigation of the binding of polysaccharides ligands to YKL-40.....	140
Table 5.2 Energetic components of the free energy of ligand binding to YKL-40. All values are in kcal/mol.....	150
Table 5.3 Hydrogen bonding pairs from polysaccharide-bound molecular dynamics simulations.....	159

Table 6.1 Simulations and calculations performed in the investigation of the binding of collagen ligands to YKL-40.....	171
Table 6.2 Interaction energies of YKL-40 residues with collagen peptides. The values are reported in terms of average interaction energy between major YKL-40 residues and collagen as a whole.....	179
Table 6.3 Hydrogen bonding pairs between YKL-40 and collagen model peptides at binding site A, including percentage occupancy, over 250-ns MD simulations.....	183

List of Figures

Figure 1.1 General classification of protein-carbohydrate interaction and groups of lectins based on their area of existence/function.....	2
Figure 1.2 Multi-modular glycoside hydrolase exhibiting a catalytic domain connected by linker peptides to carbohydrate binding modules.....	4
Figure 1.3 Classification of CBMs in three types (A, B and C) based on binding site topology and morphology of target substrate.....	9
Figure 1.4 Illustration of different possible morphologies of cellulose after biomass pretreatment.....	11
Figure 1.5 The study addresses carbohydrate recognition in six Type B CBMs, two from each of the three families – 4, 17 and 28 with two tandem CBMs.....	15
Figure 1.6 (A) YKL-40 (gray cartoon) aligned with <i>Serratia Marcescens</i> family 18 Chitinase A (cyan cartoon) illustrates structural similarity along with chito-oligomers (stick of respective color) bound in very similar way.....	26
Figure 1.7 Molecular composition of extracellular matrix. The glycosaminoglycans are most likely found as highly glycosylated parts of proteoglycans. Structural glycoproteins exhibit very little glycosylation.....	28
Figure 2.1 Homology modeling of <i>Bsp</i> CBM17 using the SWISS-MODEL.....	34
Figure 2.2 Docking of cello-oligomers through pairwise alignment.....	36

Figure 2.3 Transition from chitohexaose to cellobiohexaose using structural similarity and sidechain rebuilding with internal coordinates from topology database.....	37
Figure 2.4 Surface representation of YKL-40 and collagen triple helix illustrating the concave and convex patches.....	38
Figure 2.5 The thermodynamic pathway implemented in FEP/ λ -REMD to obtain ligand binding free energy.....	43
Figure 2.6 Scheme of replica distribution and exchange in FEP/ λ -REMD, as implemented in this dissertation.....	44
Figure 3.1 <i>Cf</i> CBM4-1 and <i>Cf</i> CBM4-2 ligand conformations considered in this study...	51
Figure 3.2 Binding site nomenclature for <i>Cf</i> CBM4-1. The CBM binds cellobiohexaose along five individual binding subsites perpendicular to the β -sheets forming the protein core. These subsites are numbered from 1 to 5.....	56
Figure 3.3 Thermodynamic cycle used to determine ligand binding free energy from FEP/ λ -REMD.....	58
Figure 3.4 Snapshots from the <i>Cf</i> CBM4-1-NRE' simulation at (a) 0 ns and (b) 2 ns.....	61
Figure 3.5 Calculated Gibbs free energy over 40 consecutive 0.1-ns calculations using FEP/ λ -REMD.....	64

Figure 3.6 Hydrogen bonding partners (dashed lines) at each subsite between side chains of cellopentaose (green and red sticks) and amino acids of <i>Cf</i> CBM4-1-RE (yellow, red, and blue sticks).....	67
Figure 3.7 <i>Cf</i> CBM4-1 hydrogen bonding behavior and protein-ligand dynamics from 250-ns MD simulations. (a) Root Mean Square Deviation (RMSD) of the <i>Cf</i> CBM4-1 protein backbone over the 250 ns simulation.....	70
Figure 3.8 Root mean square fluctuation (RMSF) of the protein backbone for (a) five <i>Cf</i> CBM4-1 systems and (b) three <i>Cf</i> CBM4-2 systems over 250 ns.....	72
Figure 3.9 (a) Comparison of binding site of <i>Cf</i> CBM4-1 (gray) and <i>Cf</i> CBM4-2 (salmon) illustrating substitutions of residues involved in cello-oligomer binding.....	76
Figure 3.10 Comparison of <i>Cf</i> CBM4-1 (gray surface) and <i>Cf</i> CBM4-2 (salmon surface) binding groove width.....	77
Figure 3.11 <i>Cf</i> CBM4-2 ligand dynamic measurements. (a) Average number of hydrogen bonds (b) RMSF of ligand on a per binding subsite. (c) Average total interaction energy (d) The average number of water molecules within 3.5 Å of each binding subsite of <i>Cf</i> CBM4-2.....	79
Figure 3.12 Family 15 CBM derived from <i>Pseudomonas cellulosa</i> xylanase Xyn10C, <i>Pc</i> CBM15 (purple cartoon), bound to xylopentaose (yellow and red sticks) aligned with <i>Cf</i> CBM4-1-RE (gray cartoon) bound to cellopentaose (green and red sticks).....	83

Figure 3.13 Structural alignment of <i>Cc</i> CBM17-RE (left; PDB 1J84) and <i>Cj</i> CBM28-NRE (right; PDB 3ACI) with <i>Cj</i> CBM4-1-RE (PDB 1GU3).....	85
Figure 3.14 Root mean square fluctuation (RMSF) of cellopentaose ligand from its average position over 250 ns trajectory calculated per binding subsite for all eight systems.....	87
Figure 3.15 Snapshots of cellopentaose (lines) at every 2.5 ns in the binding site of each CBM (gray cartoon) over the 250-ns simulations.....	88
Figure 4.1 CBMs (cartoon) from families 4, 17, and 28 with bound cello-oligomers (medium gray sticks). Binding site aromatic residues are shown in a dark gray stick representation.....	93
Figure 4.2 Cartoon illustration of the protein-carbohydrate complexes modeled in this study.....	97
Figure 4.3 Initial position of <i>Bsp</i> CBM28 in the forward orientation (after 500 ps of <i>NPT</i> equilibration) over the cellulose-I β microfibril with a middle chain of the top layer occupying the binding cleft of the CBM.....	100
Figure 4.4 Differences in the two binding site architectures of family 4, 17, and 28 CBMs, as illustrated through hydrophobic interactions (dark blue sticks and transparent surface) and hydrogen bonding (red sticks) with the cellopentaose ligand (light green and red sticks).....	105

Figure 4.5 Convergence of the free energy calculations of cellopentaose binding to CBMs over 20 consecutive windows of 0.1 ns using enhanced sampling method FEP/ λ -REMD.....	108
Figure 4.6 Root mean square fluctuation (RMSF) of the cellopentaose ligand from its average position in the clefts/grooves of representatives from family 4, 17, and 28 CBMs obtained from 250-ns MD simulation on a per-binding-subsite basis.....	109
Figure 4.7 Average change in solvent accessible surface area (Δ SASA) calculated using VMD over the 250 ns MD simulation trajectories of each CBM-cellopentaose system to compare the difference between sandwich and twisted platforms.....	115
Figure 4.8 Alignment of the twisted platform binding sites of CcCBM17-RE (top) and CjCBM28-NRE (bottom) with respect to the common pair of Trp residues.....	117
Figure 4.9 Average total interaction energy per binding subsite with the surrounding amino acid residues of CjCBM28 for a cellohexaose chain of the microfibril occupying the cleft in two different ways, A to F (red) and X to E (blue).....	119
Figure 4.10 RMSD of the CBM backbone bound to the model non-crystalline cellulose microfibril over 100 ns of MD simulation.....	121
Figure 4.11 Root mean square fluctuation (RMSF) of the backbone atoms of CcCBM17 (top) and BspCBM28 (bottom) in each ligand occupancy state.....	122
Figure 4.12 Total interaction energy between the substrate and each protein residue, averaged over the length of the MD simulations.....	124

Figure 4.13 Potential of mean force (PMF) in uncoupling (A) <i>CcCBM17</i> and (B) <i>BspCBM28</i> from non-crystalline cellulose.....	127
Figure 5.1 Monomeric units of the polysaccharides considered as potential physiological ligands of YKL- 40: cellohexaose, chitohexaose, heparan sulfate, heparin, hyaluronan, and chondroitin sulfate.....	133
Figure 5.2 Thermodynamic cycle used to determine ΔG with FEP/ λ -REMD method. ‘solv’ refers to the solvated state and ‘vac’ refers to the gas-phase state.....	141
Figure 5.3 Convergence of ΔG over 20 consecutive 0.1-ns free energy perturbation calculations using the FEP/ λ -REMD method.....	143
Figure 5.4 Relaxation of the polysaccharide ligands in the primary binding cleft of YKL-40. Each ligand is shown after a 100-ps equilibration.....	145
Figure 5.5 Hyaluronan in YKL-40 binding site at 0 ns (left) and at 250 ns (right) illustrating difference between V-shape conformations of hyaluronan.....	146
Figure 5.6 Snapshots from four independent MD simulations of heparin (white stick) binding to a putative heparin-binding site (blue surface) of YKL-40 (gray surface).....	149
Figure 5.7 (a) Root-mean-square deviation over 250-ns MD simulations and (b) root-mean-square fluctuation of YKL-40 without a ligand (apo) and bound to chitohexaose, cellohexaose, and hyaluronan.....	155

Figure 5.8 (a) RMSF of the polysaccharide ligands on a per-binding-subsite basis. (b) Average number of water molecules within 3.5 Å of each ligand monomer.....	157
Figure 5.9. Root mean square deviation of loop of residues 209 to 213 from the unusual configuration in apo YKL-40.....	163
Figure 6.1 Triple helical structure of collagen from crystal structure (PDB ID – 1CAG) that exhibits strands with repeating amino acid sequence of Gly-Pro-Hyp.....	166
Figure 6.2 Molecular shape complementarity docking calculations predict collagen-like peptides will bind to YKL-40 in two possible orientations.....	169
Figure 6.3 Root-mean-square deviation of collagen-like peptides over the course of 250-ns MD simulations at (a) collagen binding site A and (b) collagen binding site B.....	175
Figure 6.4 Native contact analysis of each collagen-like peptide model binding to YKL-40 at site A and at site B.....	176
Figure 6.5 Collagen binding with YKL-40. (a) binding site A (b) binding site B.....	181
Figure 6.6 Binding free energy obtained from umbrella sampling MD simulations of the YKL-40-collagen peptide systems.....	186
Figure 7.1 Visualization of residue-residue cross-correlation of tandem CBMs calculated based on RMSD of the protein backbone (α -carbon only).....	194
Figure 7.2 PCA analysis of preliminary MD simulation data. Left panels illustrate the clustering of conformers on a principle components 1 and 2 (PC1-PC2) space.....	195

List of Additional Files

Movie 3.1.mpg, 1.6 MB, Trajectory of *Cf*CBM4-1-RE' for first 25-ns showing the shifting of cellopentaose at around 8-ns. URL: <https://youtu.be/IHBIRffAjJw>

Movie 3.2.mpg, 1.6 MB, Trajectory of *Cf*CBM4-1-NRE' for first 25-ns showing the shifting of cellopentaose at around 2-ns. URL: <https://youtu.be/UcCXa1OkBrs>

Movie 6.1.mov, 5.8 MB, Trajectory of YKL-40 in complex with heparin where heparin was docked in the primary binding site. URL: https://youtu.be/u2RmwDxw_o8

Movie 6.2.mov, 3.8 MB, Trajectory of YKL-40 with starting position of heparin in the bulk with no initial interaction between them. URL: <https://youtu.be/HnOEzOAEh3o>

Movie 7.1.mov, 8.0 MB, Trajectory of YKL-40 binding to all four collagen peptide models at binding site A for 250-ns. URL: <https://youtu.be/A4XEy9c07Hg>

Chapter 1 – Introduction

1.1 Carbohydrate-binding proteins

Proteins are the most abundant macromolecules in mammals, and are the most diverse biomolecules in the living organisms [1, 2]. Proteins play many different roles in the chemical and biological events of a cell from its birth to death, functioning as enzymes, structural proteins, transporters, energy storage proteins, motor proteins, antibodies in immune response, and regulatory proteins, etc.. They are mostly known to work synergistically with either other proteins or biomolecules like carbohydrates [3]. Carbohydrates are the most abundant macromolecules in plant cells, where cell walls consist of long polysaccharides [4, 5]. However, various sizes of carbohydrates, oligosaccharides and short polysaccharides, are also found along the cell plasma membrane and in the extra-cellular matrix of all living organisms, both independently as well as in the form of protein-conjugates like glycoproteins. Naturally, the interaction between proteins and carbohydrates in biological processes has a significant role, reportedly in metabolic activities, cell recognition and signaling, catalysis, and inflammation [6-11]. Such interactions also affect industrial processes, and enzymatic degradation of cellulosic biomass to produce bioethanol is one of the most important process among them, as we will discuss further.

Carbohydrate-binding proteins can be primarily categorized into two broad classes: carbohydrate-active enzymes and catalytically inactive proteins. The latter class of this protein population is commonly referred to as lectins, which can recognize and bind carbohydrates with high specificity and high to moderate affinity, but without any catalytic activity [3, 12-14].

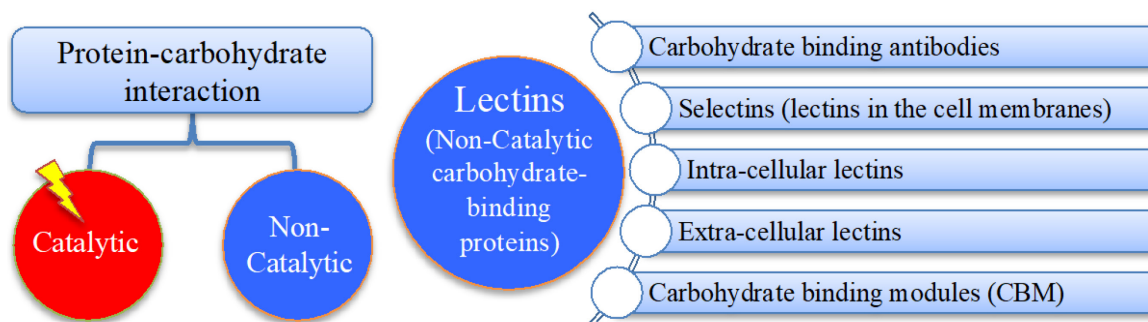


Figure 1.1 General classification of protein-carbohydrate interaction and groups of lectins based on their area of existence/function.

These non-catalytic carbohydrate-binding proteins, lectins, can be further divided into groups such as carbohydrate-binding antibodies, selectins (lectins in the cell membranes), intra-cellular lectins, extra-cellular lectins, and carbohydrate binding modules (CBMs) [3, 15, 16]. Every group of lectins exhibits further variation in structural and functional properties making it very challenging to generalize the overall non-catalytic protein-carbohydrate partnership. In this dissertation, I focus on case studies of two of these non-catalytic carbohydrate-binding proteins, investigating (1) the molecular-level recognition mechanisms of Type B carbohydrate binding modules that can differentially bind to oligomeric and non-crystalline cellulose, and (2) the binding mechanisms of the mammalian glycoprotein YKL-40, which is an extra-cellular lectin primarily known as a biomarker whose functionality remains largely unknown.

1.2 Motivation

1.2.1 Sustainable cellulosic bioethanol

Over the last few decades, it has been acknowledged that ethanol produced from lignocellulosic biomass has great potential to become an excellent alternative liquid fuel for the transportation sector [17-20]. Even though there are other upcoming green technologies like electric and fuel-cell vehicles, lignocellulosic biomass-derived ethanol is much needed, as unlike other technologies, it has great potential to achieve target reductions in greenhouse gas emissions [21-23]. In the journey toward developing liquid biofuels, researchers have studied various aspects of this field, pointing out advances as well as challenges in building a sustainable cellulosic bioenergy enterprise [24-27]. The biochemical conversion of cellulose to fermentable sugars is considered to be one of the most promising approaches, particularly when compared to other thermochemical routes [28, 29]. Lignocellulosic biomass comprises a majority of plant cell walls and can be biochemically converted to ethanol in five general steps: i) biomass handling ii) pretreatment, iii) enzymatic hydrolysis, iv) fermentation, and v) ethanol recovery. The enzymatic hydrolysis step is both a rate determining as well as expensive step [30, 31]. The highly crystalline nature of cellulose makes it recalcitrant to facile deconstruction and challenging for enzymes to access the strong glycosidic linkages connecting the soluble glucose monomers [29].

Nature uses multi-modular glycoside hydrolase (GH) enzymes to help overcome biomass recalcitrance. Multi-modular GHs can take many forms but generally consist of at least one catalytic domain (CD) appended by linker peptides to one or more carbohydrate binding modules (CBM) [32, 33] (Figure 1.2).

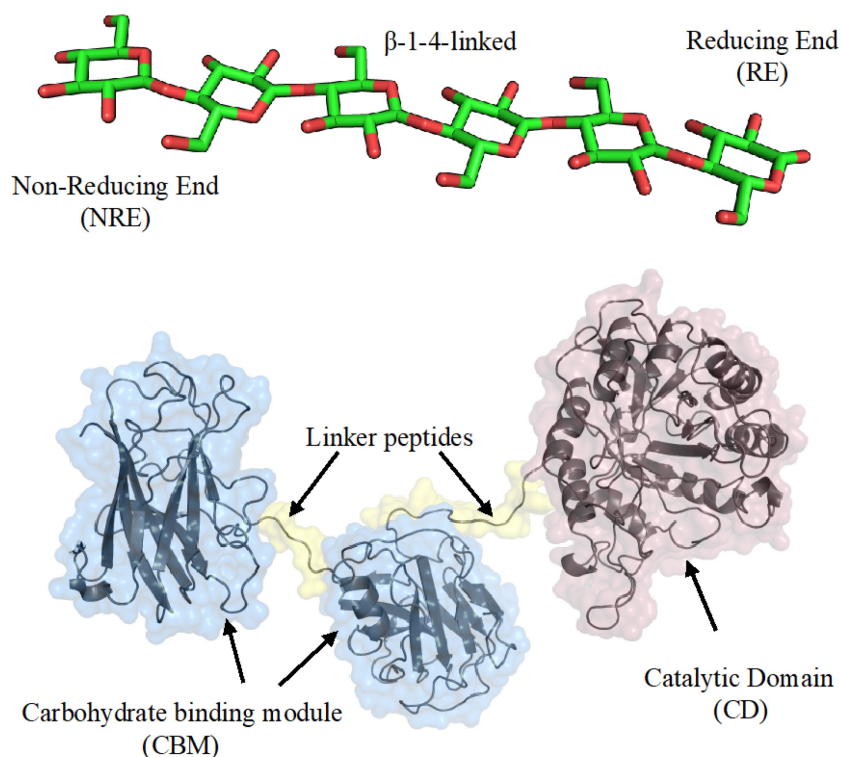


Figure 1.2 Stick representation of cello-oligomer showing β -1-4 glycosidic linkages between glucose monomers with reducing and non-reducing ends (top panel). Multi-modular glycoside hydrolase exhibiting a catalytic domain connected by linker peptides to carbohydrate binding modules (bottom panel).

The CD is responsible for catalytic activity, cleaving the glycosidic linkages of cellulose. The non-catalytic CBMs assist the CD in targeting the substrate and serves as the primary biological means of protein-carbohydrate recognition [16]. To attain the goal of efficient biomass conversion, a significant amount of prior research has focused on catalytic domains, as they are directly involved in the cleavage of glycosidic linkages and can offer obvious gains in enzymatic performance. On the other hand, researchers have just begun to explore and realize the carbohydrate recognition capabilities of CBMs in

optimization of efficient biomass conversion. The specificity of CBMs may also be harnessed for multitudes of other biotechnological applications, including, for example, bioprocessing, targeting, cell immobilization, protein engineering, diagnostics, and fiber modification [34, 35]. Therefore, we address fundamental questions surrounding how these carbohydrate binding modules specifically recognize their carbohydrate substrates as a means to develop enhanced CBM-based biotechnology.

1.2.2 Lectin-mediated targeted drug delivery

Given their substantial involvement in various biological processes, lectins have been extensively studied in medicinal chemistry as drug targets or as carriers to target and deliver drugs to their site of action [36, 37]. Vice versa, carbohydrates in their various forms (mono/oligo/poly-saccharide) have been used as drug carriers or labels and have also been the target of drug molecules, especially the glycosaminoglycans [38]. Accordingly, the intertwined relationship between lectins and carbohydrates is an invaluable asset in the field of targeted drug delivery [39].

Different cells are known to express different glycan arrays, and natural combination of monosaccharides with various sets of substitutions and isomorphs can result in a vast range of different chemical structures, each combination being a unique sugar code [36, 40]. In other words, malignant cells, like tumor cells, express different glycans or glycoproteins compared to their benign counterparts, making malignant cells potential targets of lectins [41]. Such appears to be the case with mammalian glycoprotein YKL-40, which is overproduced by the human body in conjunction with cancers and chronic inflammatory ailments, making it a well known biomarker of many

diseases [42]. However, experimental efforts to understand the biological role and mechanism of YKL-40 have been hindered by the difficulty associated with isolating the various contributions from the cellular environment. To date, it is unclear what YKL-40 interacts with in the human body, but it has been reported to bind *in vitro* to chito-oligosaccharides and collagen [43-45]. Better understanding of the different mechanisms by which such lectins can target specific carbohydrates can add a new dimension to engineering lectin-mediated drug delivery pathways and enhance our understanding of inflammatory disease and cancer progression. Here, we focus on identifying the potential physiological binding partners of this infamous biomarker lectin by screening the most likely carbohydrate ligand candidates found within the extra-cellular matrix and investigating the associated molecular-level binding mechanisms.

1.3 Research background

1.3.1 Carbohydrate binding modules

The catalytic domains of the carbohydrate-active enzymes, e.g., cellulases or cellulosome enzyme complexes, can have a single or multiple smaller CBM domains that aid in function; the domains are connected by linker peptides and typically hold complementary specificity towards carbohydrate substrates [33, 46-48] (Figure 1.2). Prior to 1999 [49], CBMs were known as cellulose binding domains (CBDs) because almost all of those initially characterized were specific to cellulose. As more enzymes with non-catalytic domains binding carbohydrates other than cellulose were observed through advances in biochemical and structural characterization (NMR and X-ray) techniques, the nomenclature shifted to a more general terminology of Carbohydrate Binding Modules (CBMs) [49]. There are three proposed functions of CBMs in biomass deconstruction: i)

maintaining proximity to substrates, ii) targeting specific regions, and iii) disrupting surface crystallinity [16, 50]. Experimental results confirm that maintaining proximity to substrate contributes to increasing enzyme concentration at the surface, resulting in enhanced enzymatic deconstruction of polysaccharides [51-58]. Their function in targeting distinct regions, specificity towards orientation of substrate [59-62], and apparent variable functional capacity on chemically invariant substrates [63-66] are appealing targets for enhanced biotechnology development and, accordingly, from the perspective of understanding fundamental protein-carbohydrate binding mechanisms. The disruption of substrate surface crystallinity by CBMs has been reported by relatively few biochemical studies [67-71], and similar results have not observed for other CBMs [50, 51, 56, 72]. Thus, this latter proposed CBM function has not been widely accepted as a general function by the community.

Finally, cellulosic substrates are comprised of glucose monomers linked together through β -1,4 glycosidic bonds and span a range of degree of polymerizations. β -1-4-glycan-specific CBMs appear to bind either crystalline or non-crystalline/amorphous and oligomeric cellulose [16]. Competition isotherms suggest CBMs do not compete for binding sites on these variable crystalline surfaces despite similar substrate specificities [66, 73]. These functional features, along with structural and sequence similarity, have led to the classification scheme of different CBMs [16].

1.3.1.1 Terminology of CBMs

Classification of CBMs is based on similarity in both protein sequence and function; CBM ‘families’ are defined according to the protein sequence and fold, while

‘types’ are illustrative of their functional activity [16]. CBMs largely appear to bind either crystalline or non-crystalline/amorphous and oligomeric substrates. As of August 2017, there are 81 CBM families categorized based on amino acid sequence in the Carbohydrate Active Enzymes (CAZy) database (<http://www.cazy.org/Carbohydrate-Binding-Modules.html>) [74]. The current protocol for abbreviation of a CBM from a given family is CBM#, where # is its family number. Also, the name of the native microorganism <Genus species> producing the CBM may be added as prefix, i.e., GsCBM#.

Although families are divided based on protein sequence, some protein folds are common across several families. The β -sandwich fold is the most common, shared by more than 30 families. The CBMs are also grouped into three types (A, B, and C) based on functional similarity and binding site topology (Figure 1.2). Type A CBMs consist of those with affinity towards crystalline substrate and have planar binding sites. Type B CBMs are specific to free-single-glycan chain polysaccharides, have groove or cleft-like binding site, and bind ‘internally’ on single free glycan chains. And Type C CBMs bind the termini of glycans with a simplified lectin-like binding site that can accommodate only mono-/di-/tri-saccharides at the terminal end of a glycan chain [16, 75] (Figure 1.3). Families generally belong to one of the types, for example CBMs from families 4, 17, and 28 are all Type B CBMs, but there are exceptions, e.g., in family 2 and 3, that illustrate the functional diversity of these carbohydrate-binding proteins and difficulty in developing a cohesive categorization scheme [76, 77]. Before going into the details of specific CBM families, it is also equally important to define the chemical nature of their substrate.

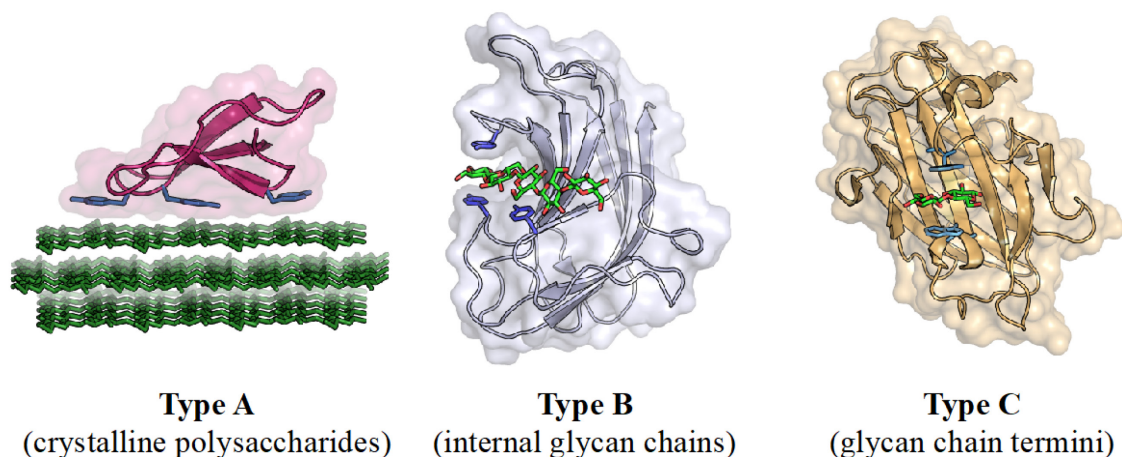


Figure 1.3 Classification of CBMs in three types (A, B, and C) based on binding site topology and morphology of target substrate. Type A CBMs have a planar binding site and target crystalline substrate. Type B CBMs have cleft or groove shaped binding sites and target single free glycan chains. Type C CBMs have a binding site that bind to glycan chain termini, i.e. reducing or non-reducing ends.

1.3.1.2 Substrates of CBMs - different target morphologies

Nature has developed plant cell walls as a complex network containing different high molecular weight polysaccharides to reinforce strength and provide protection against microbial and animal attack [29, 78]. These polysaccharides primarily include cellulose, hemicellulose, and pectin, in varying percentages depending on the plant. The secondary cell wall also contains lignin, which strengthens and waterproofs the cell wall. Cellulose, the most abundant biopolymer on Earth, is the unbranched polymer of repeating β -1-4-linked glucans and is the primary component of plant cell walls [4, 79]. Several polymer chains are stacked upon each other and are held together through a network of hydrogen bonds in either parallel or anti-parallel chains constituting

microfibrils of crystalline cellulose; such variations in chain directions and hydrogen bonding patterns between chains/sheets define various polymorphs of cellulose [80]. Two polymorphs, cellulose I α and I β , are naturally produced by plants and both exhibit only intralayer hydrogen bonding with parallel-oriented chains [81, 82]. Cellulose II and III are synthetic cellulose polymorphs that have been obtained through chemical pretreatment of cellulose I, and exhibit antiparallel chains with both intra- and inter-layer hydrogen bonding [83, 84]. The average number of monomeric glucose units in the polymeric chains is called as the degree of polymerization (DP) of that cellulosic substrate. The DP and polymorphism of substrate varies with source as well as pretreatment conditions of the biomass. For example, microcrystalline cellulose (e.g., Avicel) has a DP between 150 and 300, cotton and other plant fibers can have a DP in the range of 800-10,000, and secondary cell wall cellulose has a much higher DP, up to 15,000 [5, 85]. Apart from crystalline cellulose, other cellulosic forms, such as paracrystalline (pseudo-ordered), amorphous/non-crystalline, and soluble cellulose, have also been observed, although not essentially in its native structure [80]. This variation in crystallinity and polysaccharide construction is thought to significantly contribute to microbial recalcitrance (Figure 1.3), requiring the secretion of many different enzymes with quite varied specificity to effectively deconstruct plant cell wall components [29, 86, 87].

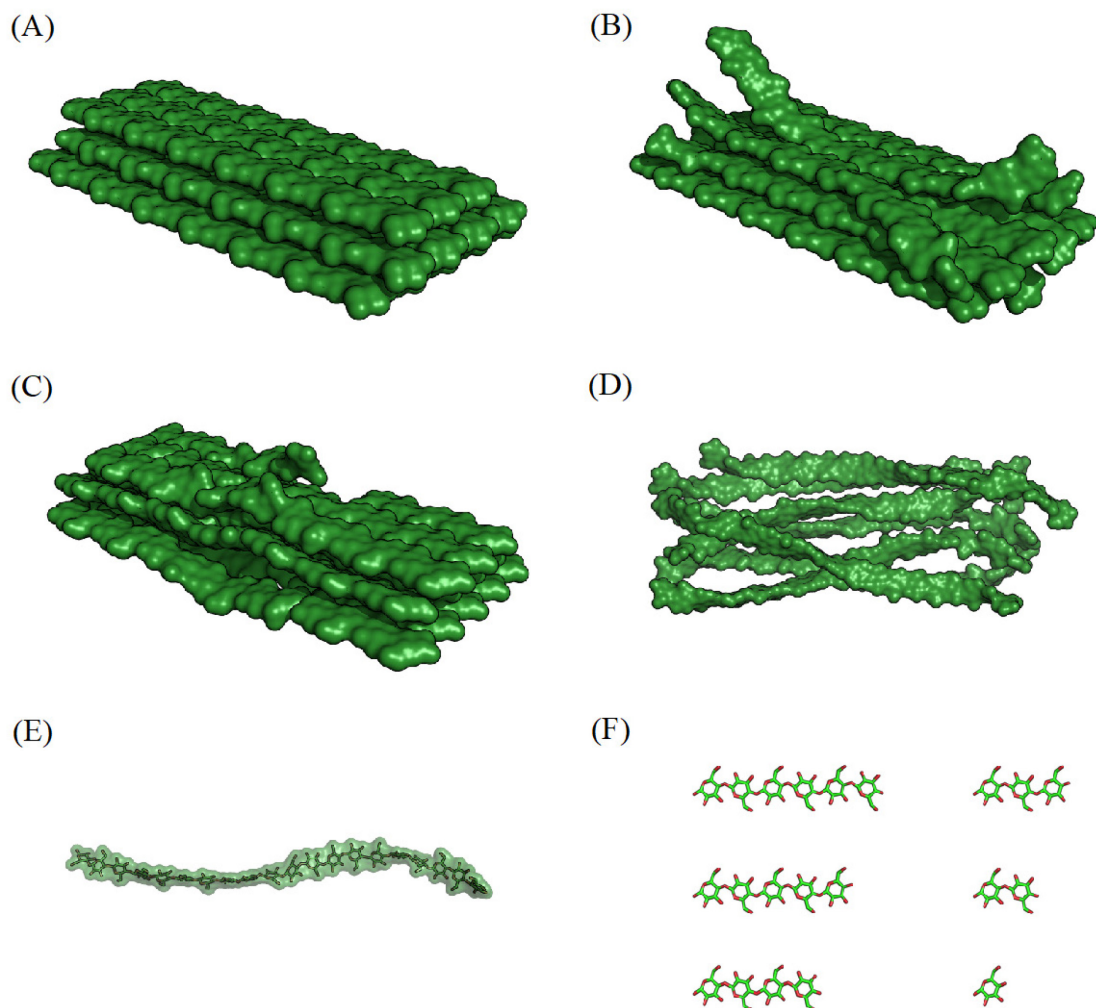


Figure 1.4 Illustration of different possible morphologies of cellulose after biomass pretreatment. (A) Surface representation of perfectly crystalline microfibril of cellulose I β . (B and C) Non-crystalline cellulose with disruptions in the microfibril; non-crystalline regions at the end and in the middle, respectively. (D) Highly amorphous bunch of polysaccharides of cellulose (E) Independent insoluble polysaccharide chain of cellulose (F) Soluble cello-oligosaccharides from celohexaose to monomeric glucose.

Crystalline cellulose is well characterized, including the network of hydrogen-bonding interactions, as the structure of both the native cellulose I α and I β allomorphs

and various other polymorphs have been identified through nuclear magnetic resonance (NMR) spectroscopy [88], atomic force microscopy (AFM) [89] and X-ray diffraction (XRD) [81, 82]. On the other hand, 3-dimensional structure determination of non-crystalline cellulose is not straightforward, leaving us with relatively little understanding as to the non-crystalline forms of cellulose that lie between truly crystalline and highly amorphous phases [90, 91]. Characterization of the amorphous cellulose regenerated from dissolution of microcrystalline cellulose in SO₂-diethylamine-dimethylsulfoxide with XRD, Fourier-transform infrared microscopy (FTIR) and differential scanning microscopy (DSC) could only confirm that it is a cellulose with a decreased degree of polymerization and crystallinity index [92]. Thus, the pretreated cellulosic biomass is thought to consist of a substantial amount of non-crystalline cellulose, including kinks or twists in microfibrils and/or voids, such as surface micropores, large pits, and capillaries, having either no specific structural properties or highly undistinguishable structural characteristics [33]. Soluble oligomers are the smallest forms of cellulose, known as cello-oligosaccharides, with a DP less than or equal to 6 (Figure 1.3) [93]. The origin of the plant biomass and pretreatment strategy are prime variables determining the target cellulose morphologies subjected to enzymatic hydrolysis. Certainly, cellulose crystallinity exists within a continuum, where clear delineations between regions are difficult to discern. However, with findings such as regional CBM specificity with non-crystalline cellulose substrates [64, 65], it is increasingly evident that CBMs hold the potential to probe such regions and provide insights to the structural complexity of the non-crystalline cellulose [94].

1.3.1.3 Cellulose-specific Type B CBMs

All Type B CBMs characterized up to now exhibit the most common protein fold among CBMs [16], termed the β -sandwich fold, which consists of two β -sheets each of which contain three to six β -strands. The characteristic groove-like nature of Type B CBM binding sites is akin to the active sites of glycoside hydrolase catalytic domains, where hydrogen bonding interactions along with aromatic stacking mechanisms are responsible for ligand binding [95]. The length of these binding sites can accommodate free-single-glycan chains, generally with at least 3 and a maximum of 6 monomers in a row. The depth of these grooves varies from being able to enclose the whole width of a pyranose ring ($> 6 \text{ \AA}$) to merely a shallow cleft ($1\text{-}2 \text{ \AA}$) [16]. The relatively solvent-exposed binding site allows these Type B CBMs to expand their specificity to a large range of substrate morphologies, excepting pure crystalline polysaccharides and very small sugars (mono-/di-/tri-saccharides).

The Type B CBMs from families 4, 17, and 28 have been shown to bind both soluble cello-oligomers and non-crystalline cellulose, but never crystalline cellulose [51, 52, 54, 73, 96-98]. Despite having the same β -sandwich fold and groove-like binding sites, CBMs from these three families have further individual characteristics that enable them to differentially bind non-crystalline cellulose in an uncompetitive mode [64, 99, 100]. Structural and thermodynamic studies of Type B CBMs have been conducted to gain insight into the CBM-carbohydrate recognition process, bringing to light additional questions as mentioned ahead. For example, variations in thermodynamic binding signatures for CBMs within these same families have been observed despite the strong structural and sequence similarities [64, 101, 102]. Mutagenic studies of key residues

comprising the binding site report individual contributions to binding affinities [99, 100], but comprehensive characterization of such contributions across several Type B CBMs is needed to identify their roles in the carbohydrate recognition process. An NMR study reported the counterintuitive finding that family 4 CBMs are able to bind oligomers in multiple orientations [61], although structural-level carbohydrate binding mechanisms were obscured; the crystal structures of ligand-bound CBMs from families 17 and 28 report only a single ligand orientation [62]. Additionally, the motivation for microbial evolution of these carbohydrate binding modules from individual entities to tandem systems is the long-term question that needed to be addressed (Details in Section 1.3.1.5). Formation of these multivalent carbohydrate-binding proteins can significantly enhance ligand-binding affinity relative to individual modules [51, 64, 102-104], although this is not a given [98].

Given all these open questions for consideration, along with the fact that we have an abundance of unexplained experimental phenomena, which could benefit from theoretical investigations, our focus lies on understanding the carbohydrate recognition mechanisms of family 4, 17, and 28 CBMs. We have conducted a comprehensive examination of carbohydrate recognition in six different Type B CBMs (two from each family 4, 17, and 28) having specificity for both cello-oligomers and non-crystalline cellulose, along with two tandem CBMs (Figure 1.5). One tandem CBM is a blend of two family 4 CBMs, whereas a second tandem CBM consists one from each family 17 and 28, as they would be naturally secreted by the associated microorganisms [105, 106]. Details of each CBM are described below based on the studies reporting their structural biology and characterizations of soluble cello-oligomeric binding. Further information

about the non-crystalline binding characterizations for these CBMs is then provided collectively in the next section.

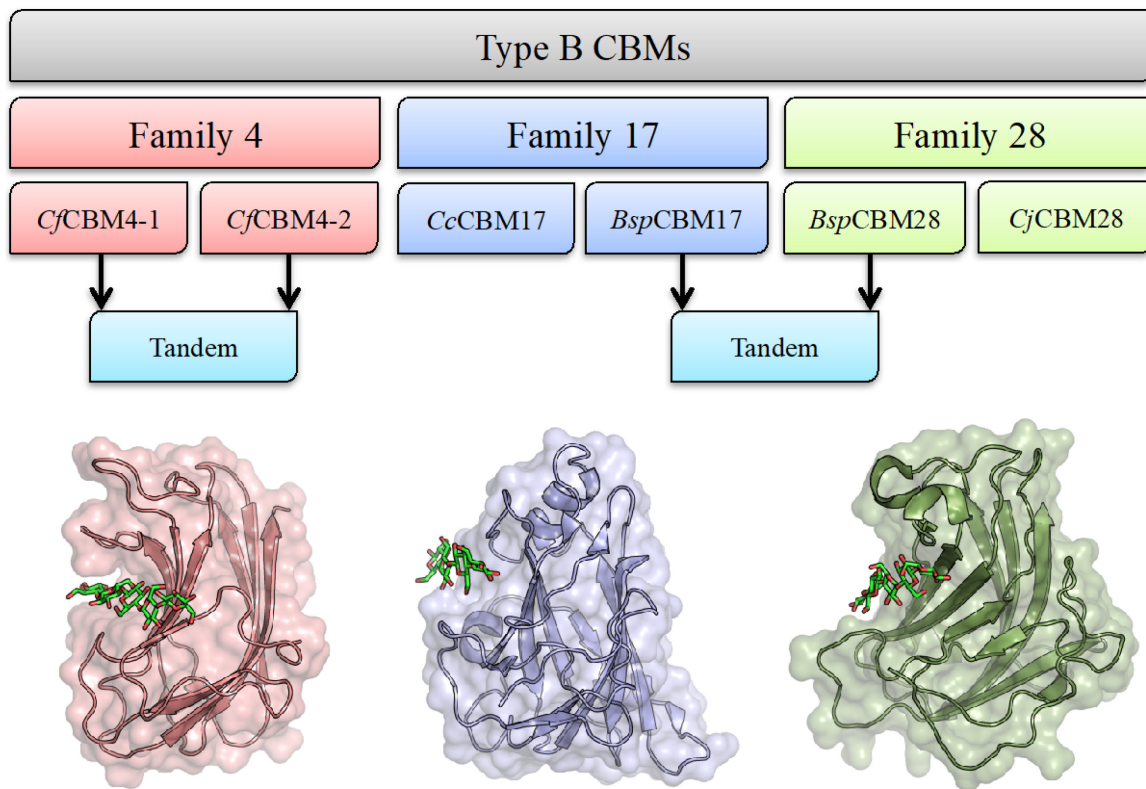


Figure 1.5 The study addresses carbohydrate recognition in six Type B CBMs, two from each of the three families – 4, 17, and 28. Two tandem CBMs were included in this work. The letters preceding the CBM# represent the name of bacteria that produces that CBM. *Cf* – *Cellulomonas fimi*, *Cc* – *Clostridium cellulovorans*, *Bsp* – *Bacillus sp. 1139*, *Cj* – *Clostridium josui*. Structures of *Cf*CBM4-1 (light pink, PDB 1GU3), *Cc*CBM17 (light blue, PDB 1J84) and *Cj*CBM28 (light green, PDB 3ACI) are shown for comparison.

1.3.1.3.1 Family 4 CBMs

*Cf*CBM4-1 and *Cf*CBM4-2, selected to represent family 4, are two N-terminal binding domains from *Cellulomonas fimi* β -1,4-glucanase C (CenC/Cel9B) [105]. The two CBMs naturally occur in tandem and are linked to the Cel9B catalytic domain through a four-amino acid peptide linker. Through various methods to detect carbohydrate-binding specificity, *Cf*CBM4-1 appears to preferentially bind cellulosic substrates, with cellopentaose and cellobiose binding with similar affinities [98, 107]. Quantitative evaluation of binding affinity by Isothermal Titration Calorimetry (ITC) suggested that *Cf*CBM4-1 binds cello-oligomers in an enthalpically-driven fashion, based on favorable change in enthalpy compensated by negative change in entropy [98]. Prevalent polar interactions, especially hydrogen bonding, appear to be central for specificity towards the cello-oligosaccharides [99, 108, 109]. Structural studies with both NMR spectroscopy and X-ray crystallography revealed that *Cf*CBM4-1 binds cello-oligomers in a binding groove formed on the face of its β -sandwich fold. This binding groove is comprised of oppositely facing hydrophobic residues sandwiching the relatively hydrophobic pyranoside rings [62, 108] (Figure 1.5). On the sides, several polar and hydrophilic residues are involved in binding, as identified through mutagenesis [99]; however, the role of orientation of cello-oligomers and their side chains in hydrogen bonding was unclear. We identify the critical binding modes with which cello-oligomers bind *Cf*CBM4-1 in a directionless fashion, details of which are discussed in Chapter 3.

In addition to being a tandem CBM partner, *Cf*CBM4-2 shares a very similar tertiary structure and sugar binding properties with *Cf*CBM4-1 [110]. Though very similar, the *Cf*CBM4-2 binding cleft is not identical to that of *Cf*CBM4-1. Detailed

comparison reveals that a few binding site residues that interact with the bound ligand in *Cf*CBM4-1 are substituted. The NMR structure captures *Cf*CBM4-2 in its apo state, and thus, the absence of ligand interactions could likely be the reason behind the larger width of its binding groove relative to that of *Cf*CBM4-1. Based on our molecular dynamics study of *Cf*CBM4-2 docked with cellopentaose in the binding groove, it is quite intriguing to observe that the width reduces to approximately that of *Cf*CBM4-1 during the equilibration period and remains almost the same over the remainder of the simulation (details in Chapter 3). We have further analyzed the variability in thermodynamic signature between these two family 4 CBMs in Chapter 4. The two CBM4s are naturally found in a side-by-side orientation, prevented from end-to-end orientation as a result of the lack of a flexible peptide linker [110]. We anticipate that this contributes, in part, to their apparent additive rather than cooperative binding when studied as tandem system.

1.3.1.3.2 Family 17 CBMs

This family of CBMs is represented by *Cc*CBM17, the C-terminal domain of *Clostridium cellulovorans* Cel5A solved by Notenboom et al. [100]. As the only CBM17 structure to date, *Cc*CBM17 has been the subject of numerous biochemical studies. *Cc*CBM17 appears to have optimal affinity towards cellohexaose, with a minimum binding requirement of cellotriose [97]. Although belonging to Type B, the binding site of *Cc*CBM17 is barely recognizable as groove, with a shallow depth of 1-2 Å compared to the 4-6 Å of family 4 CBMs. The ‘sandwich’ platform of CBM4s is also replaced by ‘twisted’ platform in CBM17s, such that two tryptophan residues are facing away from protein core to stack directly with the ligand pyranoside rings in addition to the polar

residue network running along the groove's sides. Aromatic residues are of supreme importance in these CBMs, as their mutation to alanine destroyed affinity of CcCBM17 for any tested ligand [100]. Up to 25-fold reductions in binding affinity with mutation of each polar residue also illustrates the crucial contribution of hydrogen bonding [100]. A computational alanine scan confirmed the importance, though not the role, of these polar residues [111]. This study also calculated ligand binding free energy of cellotetraose and cellohexaose bound to CcCBM17, but the accuracy of absolute binding free energies calculated using molecular mechanics/generalized Born surface area (MM/GBSA) are not reliable, and its application is typically restrained to relative ranking of binding affinities in pharmaceutical applications [112]. Here we use a robust, explicit solvent and enhanced sampling method, called as free energy perturbation with Hamiltonian replica exchange MD, recently used to calculate reliable and more accurate absolute binding free energies in relatively large protein-ligand system [113]. Initially, ΔH values from the thermodynamic study for cellotetraose binding in CcCBM17 suggest that it is an enthalpically-driven binding process, as they dominate over $T\Delta S$. However, the positive changes in both enthalpic and entropic terms with increasing length of bound cello-oligomer, with $\Delta\Delta H < T\Delta\Delta S$, means it is rather entropically-driven. The large negative change in heat capacity values, by comparison to those from family 4, are consistent with burial of significant non-polar surface and solvent reorganization, which is strongly correlated with change in entropy [100]. Such entropy-driven binding was also observed in another family 17 member from *Clostridium josui* (CjCBM17) [102].

BspCBM17 is also included in this work given its occurrence in the biochemically-characterized tandem CBM from *Bacillus* sp. 1139 Cel5A, though no

structure is currently available. *CcCBM17* has been used to draw parallels with *BspCBM17*, as they share 55% sequence similarity and 70% structural similarity and ligand binding residues are thought to be conserved [64]. Here, we used homology modeling to construct *BspCBM17* and investigate its binding functionality. Importantly, it has also been used to study characteristics of cello-oligomer binding in the tandem CBM, *BspCBM17/CBM28*. In spite of high fold similarity between family 17 and family 4, mode of carbohydrate recognition appears to be different across the two CBM families. As we point out, the differences in topology of their binding sites and thermodynamics are likely the driving force behind variation in binding (Chapter 4). Nature has evolved these CBMs to recognize various architectural regions on the cellulosic substrate.

1.3.1.3.3 Family 28 CBMs

The first representative of family 28 CBM was identified from *Bacillus* sp. 1139 Cel5A, a C- terminal carbohydrate-binding module (*BspCBM28*) [96]. A second, related CBM from *Clostridium josui* Cel5A (*CjCBM28*) is also considered in this study. Structural studies for both the CBMs illustrate their differences from the other two Type B CBM families. The binding site for family 28 CBMs is also wide and shallow, similar to that of *CcCBM17*, with one face of the cello-oligomer stacking with tryptophans and the other face exposed to solvent [101, 106]. There is one additional aromatic residue in the binding site of CBM28s compared to that of CBM17s. According to ITC results, *BspCBM28* binds to cellotetraose, cellopentaose, and cellohexaose in increasing order of binding affinity. Values for cellotetraose and cellopentaose suggest binding is driven enthalpically, but binding of cellohexaose exhibited a thermodynamic signature consistent with significant entropic contributions [96]. Contrasting with *BspCBM28*,

*Cj*CBM28 is reported to bind cello-oligosaccharides enthalpically, despite the fact that the two structures are very similar; although, the study did not include binding with cellobiose, limiting a complete comparison [102]. The recognition process depends upon key residues involved in binding of family 28 CBMs to cello-oligomers. However, no mutagenesis study of these proteins has been reported to date. Here, using a computational approach, we investigated the differences in oligosaccharide recognition for both family 28 CBMs and address many of the remaining questions posed by the experimental studies.

Comparing CBMs from all three families, it is apparent that family 4 has deep binding groove, whereas both family 17 and 28 have relatively shallow but similar binding clefts. Why this difference evolved for structurally related CBMs of seemingly the same function, at a superficial level, will be addressed. Additionally, even if relative binding affinities for cello-oligosaccharides of family 28 and 17 imply similar carbohydrate recognition mechanisms, the amino acids involved in binding are poorly conserved in two families [64]. Ligand bound structural evidence for CBMs from both family 17 and 28 show carbohydrates occupying the binding site in opposite directions [101], and again despite the similarities, this kind of variation likely relates to functional differences between the members and should be better understood from the molecular level.

1.3.1.4 Non-crystalline cellulose recognition by Type B CBMs

The interesting fact that these cellulose-specific Type B CBMs bind cello-oligomers and can selectively bind various forms of non-crystalline/amorphous cellulose

is one of the primary reasons we initiated our study investigating these proteins and their molecular-level substrate interactions. Experimentalists commonly use different forms of purified cellulose like Avicel, regenerated amorphous cellulose (RAC), and phosphoric acid swollen cellulose (PASC) to imitate specific cellulose microstructures. Avicel, although being known as microcrystalline cellulose, contains partially decrystallized regions, as binding of both *Cj*CBM4-1 and *Cj*CBM4-2 to Avicel was confirmed in one of the first biochemical studies of Type B CBMs [114]. The latter two cellulose model substrates have been specifically used as representatives of non-crystalline/amorphous forms of cellulose. *Cj*CBM4-1 was shown to bind 21-fold stronger to regenerated cellulose than to Avicel, although still with a higher affinity for cello-oligomers [98]. In contrast, the family 17 and 28 CBMs show a one order of magnitude affinity improvement towards Avicel and regenerated amorphous cellulose relative to cello-oligomers [99, 100]. Langmuir isotherms were not sufficient to describe this latter binding affinity data for *Cc*CBM17 and *Bsp*CBM28, and use of a two-binding-site model suggested both CBMs recognized two binding sites on the amorphous substrate differing in affinity, a high affinity site and a low affinity site [64]. Additionally, the CBMs did not compete for these two binding sites [64, 73], which is indicative of the presence of cellulose chains with structural features (e.g., conformation, proximity to other chains, or variations in solvation) that are distinguishable by the specific CBMs only. A recent study using fluorescent labels shows that two CBMs, one from each family 17 and family 28, bound two different non-crystalline regions of sweet potato roots [65]. In a further study from Boraston et al., high and low affinity binding sites were again observed, where binding to each was defined as an enthalpically-driven process [115]. Notably,

there was less loss of configurational entropy in the low affinity sites, and we hypothesize that they correspond to less crystalline or somewhat oligomeric nature of the ligand [115]. Although all these studies confirm the important role these Type B CBMs play in biomass degradation, they also raise fundamental questions about the targeting function of CBMs based on type, architecture, and thermodynamics. In Chapter 4, we propose that carbohydrate recognition mechanisms for cello-oligomers closely mimic those of the identified low affinity binding sites on non-crystalline cellulose. Likewise, CBM carbohydrate recognition of high affinity non-crystalline cellulose binding sites will proceed through adapted mechanisms, allowing the CBMs to discriminate between regions on the non-crystalline cellulose surface.

1.3.1.5 Carbohydrate recognition in tandem CBMs

Another confounding aspect of the Type B CBM carbohydrate recognition story is the evolutionary presence of tandem CBMs found in glycoside hydrolases. Both family 4 CBMs are found in tandem arrangement, and one from each family 17 and 28 comprise cooperatively acting tandem system (Figure 1.2). We anticipate that the specific carbohydrate recognition mechanisms belonging to individual CBMs ultimately relate back to the biological implications behind the evolutionary presence of tandem CBMs. The contrasting examples of *BspCBM17* and *BspCBM28* connected together in tandem in *Bacillus* sp. 1139 Cel5A that enhanced affinity of the system for non-crystalline cellulose by 10–100 fold relative to their individual affinities [106], and tandem-linked family 4 CBMs from *Cellulomonas fimi* CenC that merely additively improved affinity relative to separate domains [98] have driven us to learn more about this. Other than these affinity comparisons and a proposed two-step mechanism [116], much remains to be

explored about structural or dynamical mechanisms involved in binding cooperativity and effects of avidity on hydrolysis. The biological significance of cooperative affinity enhancement in tandem CBMs is unclear, with such cooperativity used by hyperthermophilic organisms to overcome weak binding in high temperature environments [104] also being exhibited by CBMs from mesophiles [64]. A potential alternative function of tandem CBMs is that they may fine-tune the binding specificity, as binding models suggest that low affinity binding sites will only be bound when all high affinity sites are completely occupied [64]. Our computational approach provides an avenue to investigate such questions.

1.3.2 Mammalian glycoprotein YKL-40

1.3.2.1 YKL-40: a biomarker

YKL-40, also known as chitinase 3-like 1 (CH3L1), is a mammalian glycoprotein implicated as a biomarker associated with progression, severity, and prognosis of chronic inflammatory diseases and a multitude of cancers [117-120]. The protein is overexpressed in many pathological conditions that involve connective tissue remodeling or increased deposition of connective tissue components. For example, increased levels of YKL-40 are reported in the blood serum of patients with rheumatoid and osteoarthritis, hepatic fibrosis, and asthma [121-125]. YKL-40 was first identified from the medium of human osteosarcoma cell line MG-63 [126]. The first three N-terminal amino acids of this protein, Tyrosine (Y), Lysine (K) and Leucine (L), and the molecular mass of 40 kDa are the basis of the name YKL-40. Many different types of cells including synovial, endothelial, epithelial, smooth muscle, and tumor cells produce YKL-40 *in vivo*, likely in response to environmental cues [42, 127-129]. Speculation as to biological function of

YKL-40 varies from both inhibiting and antagonizing collagen fibril formation as a result of injury or disease [45], as well as conferring drug resistance and increasing cell migration leading to progression of cancer [119], and protection from chitin-containing pathogens [43]. Though the association of YKL-40 with physical maladies is well-documented, identification of the physiological ligand of this lectin, and thus biological function, remains elusive.

1.3.2.2 Known structural and functional properties of YKL-40

Mammalian YKL-40 is quite similar to family 18 glycoside hydrolases (GH) based on high sequence homology with this well-conserved class of enzymes in the CAZy database [43, 44, 74]. It also holds 53 % sequence identity with the human macrophage chitinase (HCHT) [130]. The crystal structure of human YKL-40, also occasionally referred as human cartilage glycoprotein-39 (HCGP-39), was found to be similar to the crystal structure of human chito-triosidase [130], mouse YM1 [131], IDGF-2 from common fruit fly [132] and other family 18 GHs [133] (Figure 1.6.A). Structural analyses of YKL-40 and these enzymes reveal that they consist of a $(\beta\alpha)_8$ barrel (Figure 1.6.C), and in some cases, an extra α/β domain is inserted in one of the barrel loops [43, 44, 130]. Though very similar in binding site architecture to family 18 GHs, YKL-40 lacks catalytic activity due to substitution of the glutamic acid and aspartic acid at the end of the conserved DXXDXDXE motif typical of catalytically-active family 18 GHs, rendering YKL-40 a lectin – a non-catalytic sugar-binding protein. YKL-40 exhibits an *N*-glycosylation site at Asn60, where a disaccharide of *N*-acetyl glucosamine is attached (Figure 1.6.C) [43, 44]. Structural evidence suggests YKL-40 exhibits at least two functional binding regions, the primary binding cleft has nine binding subsites lined with

aromatic residues compatible with carbohydrate binding (Figure 1.6.B), where chito-oligomers have been shown to bind with YKL-40 [43]. A second putative heparin-binding site, located within a surface loop, has also been suggested (Figure 1.6.B), though *in vitro* binding affinity studies have been unable to conclusively demonstrate this [44]. Sequence identity analysis also suggest that there are hyaluronan-binding sites on the external surface of folded YKL-40 [134]; however, the structural evidence of these binding sites, again, has not been observed in any structural studies.

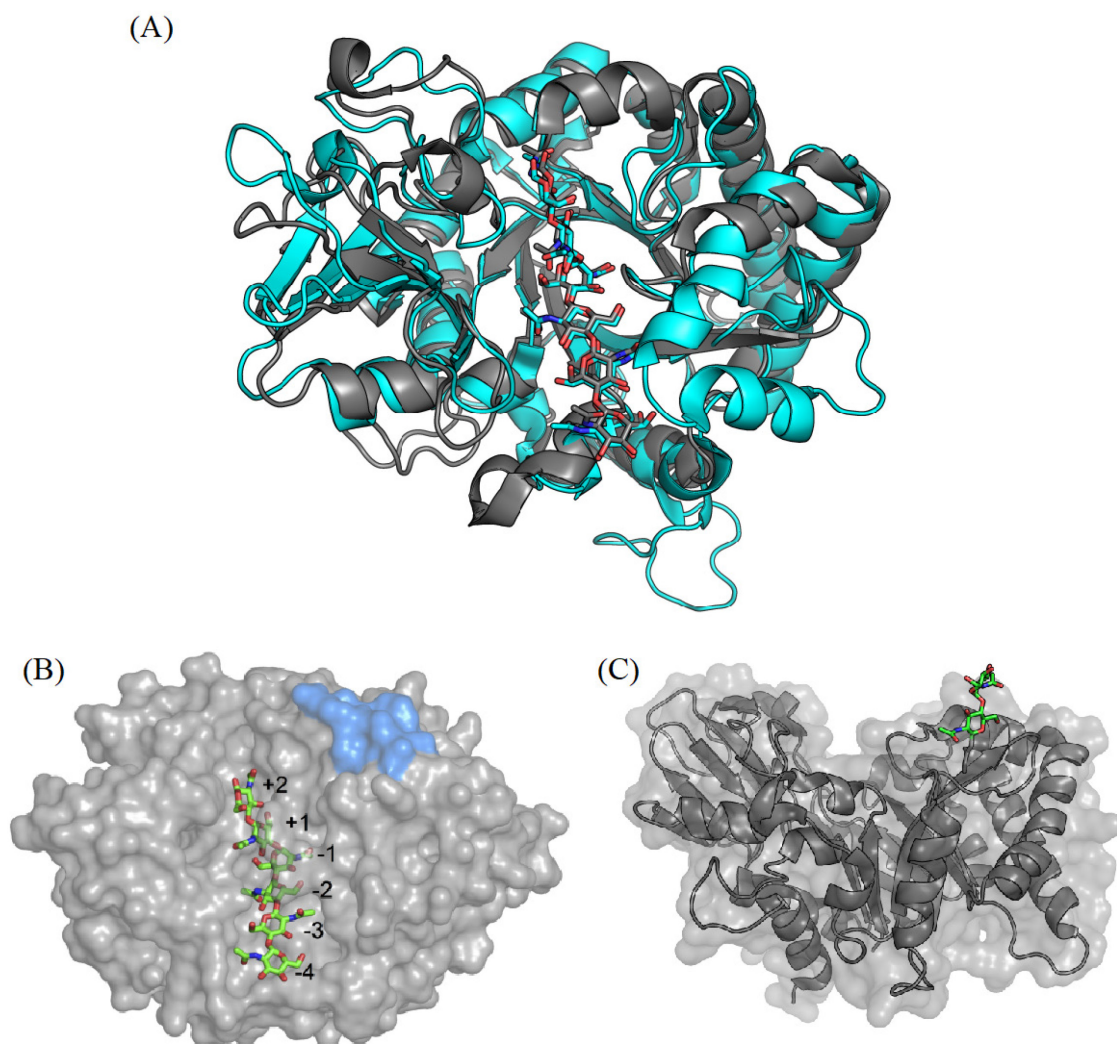


Figure 1.6 (A) YKL-40 (gray cartoon) aligned with *Serratia marcescens* family 18 Chitinase A (cyan cartoon) illustrating structural similarity and chito-oligomers (stick of respective color) binding similarity. (B) Surface representation of YKL-40 showing the binding cleft with a bound hexamer of chitin. Binding sites +2 through -4 are numbered. Sites -5, -6 and -7 have also been identified but are not shown. The putative heparin-binding site is shown in marine blue. (C) Side view of the structure of YKL-40 (gray cartoon) illustrating the $(\beta\alpha)_8$ barrel fold and *N*-linked glycosylation (green sticks). Transparent surface rendering shows the overall 'bean-shaped' nature of YKL-40.

1.3.2.3 Potential physiological ligands

Binding affinity and structural studies reveal that chito-oligosaccharides are a natural substrate [43, 44, 127]. In line with family 18 GHs, YKL-40 uniquely binds short and long chito-oligomers, indicating preferential site selection based on affinity [43]. Chitohexaose binding has also been purported to induce conformational changes in YKL-40 [43], though this has not been observed in all structural studies [44]. Lectin binding niches are widely believed to be “pre-formed” with respect to the preferred ligand, exhibiting little conformational change upon binding [13, 135]. Despite the apparent affinity, chitin is not a natural biopolymer within mammalian or bacterial cells, and the presence of chitin or chito-oligosaccharides in mammals is likely related to fungal infection [136]. The noted up-regulation of YKL-40 in response to inflammation lends credence to the argument that YKL-40 functions as part of the innate immune response in recognition of self from non-self [12, 127]. Although, high expression levels of YKL-40 in carcinoma tissues suggest function beyond the innate immune response may also exist [137, 138]. The extracellular matrix is comprised of a mesh of proteoglycans (protein-attached glycosaminoglycans (GAGs)), polysaccharides, and fibers, including collagen (Figure 1.7) [139]. An alternate theory to the pathogenic protection function is that a closely related polysaccharide, instead of chitin, plays the role of the physiological ligand in mediating cellular function [44].

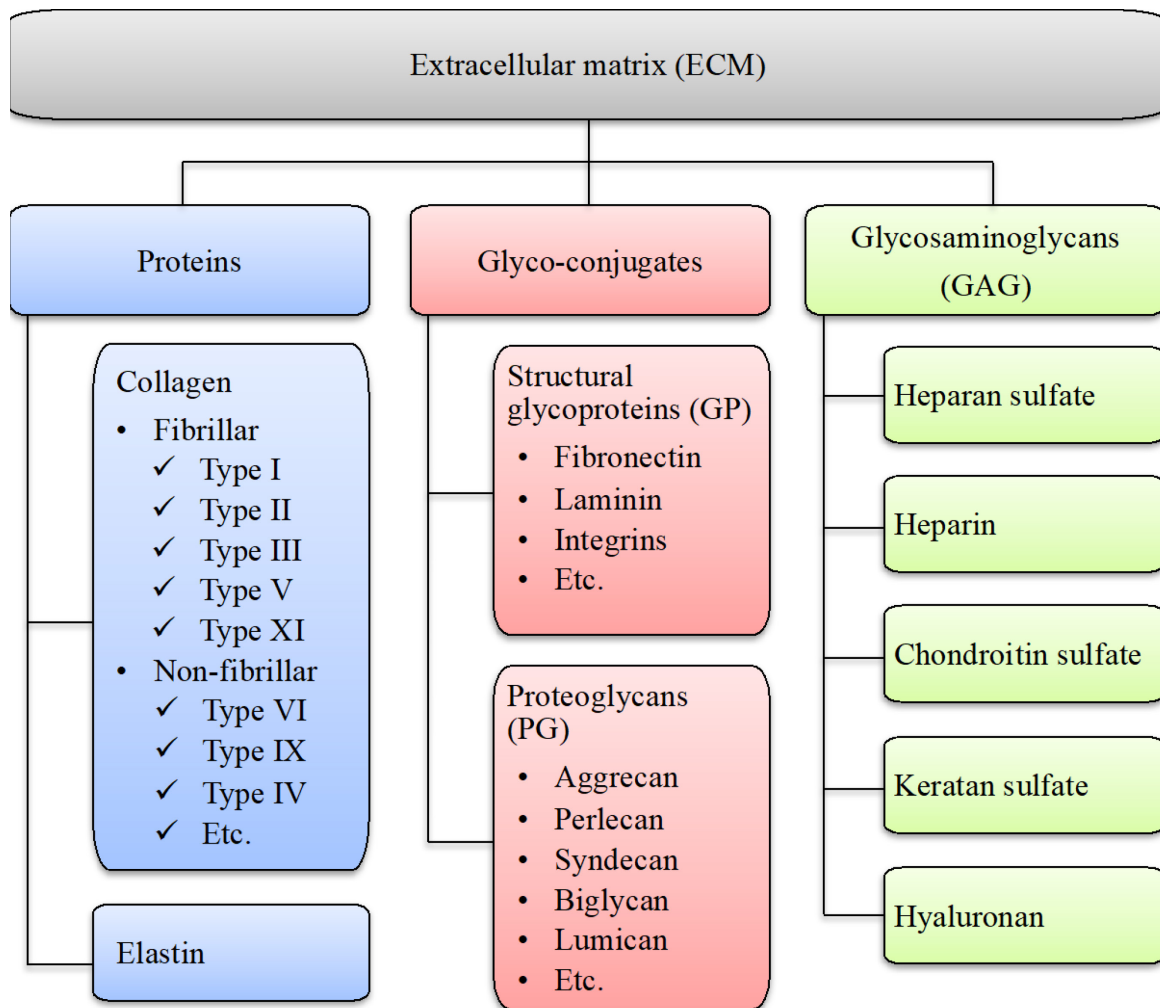


Figure 1.7 Molecular composition of extracellular matrix. The glycosaminoglycans are most likely found as highly glycosylated components of proteoglycans. Structural glycoproteins exhibit very little glycosylation.

The association of YKL-40 with ailments such as arthritis, fibrosis, and joint disease is suggestive of molecular-level interactions with connective tissue, and thus collagen [140-144]. Motivated by understanding the physiological role of YKL-40 in connective tissue remodeling and inflammation, Bigg *et al.* investigated association of YKL-40 with collagen types I, II, and III using affinity chromatography to confirm

binding to each type [45]. The authors report YKL-40 specifically binds to all three collagen types. Additionally, the authors used surface plasmon resonance (SPR) to confirm binding to Type I collagen. Unfortunately, the reported affinity constants were inconsistent across experiments as a result of aggregation. Nevertheless, the work clearly indicates YKL-40 is capable of binding collagen. With more than 28 types of collagen reported to exist in human body, it becomes critical to obtain molecular level insights to such protein-protein interactions, as it likely has profound physiological significance in understanding the functionality of YKL-40. However, this further confounds the question of mechanism when considering physiological ligands, as YKL- 40 is capable of binding both carbohydrates and proteins.

Proper consideration of the structure and chemical nature of the ligand relative to the YKL-40 binding site(s) enables us to envision the chemical design of potent binding partners for a target (in lectin-mediated drug delivery) or potential approaches to block lectins of medical importance (in infection, tumor spread, or inflammation). With this goal in the mind, we approach the tasks of identifying the physiological ligands of YKL-40 from a subset of six polysaccharides and glycosaminoglycans plus four triple helical collagen-like peptides and exploring the interactions characterizing the various aspects of functionality of this well-known, but mysterious lectin.

1.4 Outline of Dissertation

The overall theme of the dissertation is to identify and understand the interactions mediating protein-carbohydrate recognition at the molecular level. We examine these interactions in two different model protein-carbohydrate systems and address critical questions pertaining to their structure-function relationships through dynamics and affinity data. The insights we obtained here can be utilized in future studies in a wide range of applications.

As we study carbohydrate recognition in the Type B CBMs, we focus the investigation on two important objectives that highlight their story and role in nature:

- I. Cello-oligomer binding dynamics and bi-directional binding phenomenon in Type B CBMs (Chapter 3).
- II. Role of binding site architecture and high/low affinity binding on non-crystalline cellulose in Type B CBMs (Chapter 4).

In the quest to identify the physiological ligand of the multi-functional mammalian glycoprotein YKL-40, we divided the study into two parts:

- I. Carbohydrate ligands of YKL-40: Binding mechanisms, thermodynamic preferences and surface binding ability (Chapter 5).
- II. Protein-protein interactions of YKL-40: Identification and characterization of collagen binding sites (Chapter 6).

Chapter 2 – Computational methods

Our computational investigation of complex protein-carbohydrate systems to understand and solve many scientific questions related to their physical significance, biochemical behavior, and thermodynamic relationships required various computational methods. Classical molecular dynamics (MD) simulations and free energy calculations are two important methods that have been used in this study. Additionally, we have utilized tools prior to and after the primary, data-gathering calculations to build the models, run the simulations, and analyze the data. In this chapter, the theoretical background of these computational methods has been briefly explained to justify their application in this dissertation. Detailed protocols of implementation for methods used in a chapter have been described in respective methods section of each chapter.

2.1 Pre-dynamics tasks

Classical MD simulations of biomolecules have been used over the last few decades to predict structural and dynamical properties, microscopic interactions, and ultimately, calculate free energy profiles. To build and run an MD simulation for a particular molecular system, one needs to have dependable initial atomic positions, like a crystallographic structure or reliable homology model, well-tested, compatible force-fields, and a software package to numerically solve Newton's equations of motion. In this dissertation, we used available structural data for CBMs and YKL-40 from the Protein Data Bank (PDB) in most cases; however, we needed to perform homology modeling in the case of the *BspCBM17* structure and occasionally molecular docking to obtain the desired initial coordinates of the protein-carbohydrate and protein-protein complexes.

2.1.1 Homology modeling

As the term suggests, homology modeling, or comparative modeling, is a method to generate an atomic-resolution model of a target protein for which no structural data is available based on the sequence/structural similarity with a template protein for which atomic coordinates are reported using experimental techniques like X-ray crystallography or NMR spectroscopy. Various applications of homology modeling in the drug-discovery process have been reported, yielding critical insights about structural and mechanistic properties of proteins that are experimentally difficult to purify or crystallize [145, 146], e.g., G-protein-coupled receptors [147]. Out of all the proteins studied in this dissertation, we did not have the structural data for only one protein, *BspCBM17*. However, as described in Chapter 1, this CBM is an important part of this study as a representative of family 17 CBMs and as a component of the native tandem CBM construct of Cel5A from *Bacillus sp. 1139*.

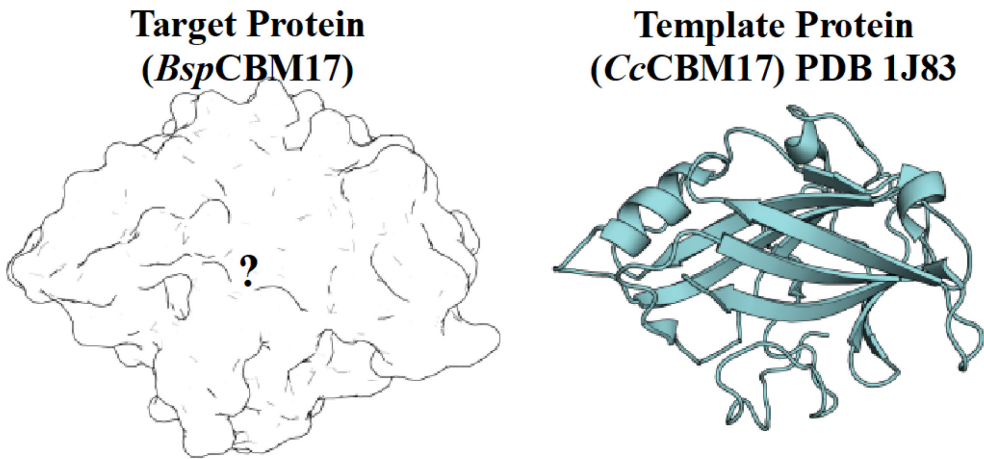
The general homology modeling protocol of a protein involves five general steps:

1. Search, identify, and select the template protein structure.
2. Align target sequence with the template sequence.
3. Build a preliminary model based on the structural information from the template.
4. Check for errors. Atom-atom overlaps, missing segments, etc.
5. Evaluate the model. Fold, ramachandran plot, stereochemistry, etc.

If the generated homology model does not satisfy the required quality criteria, one can go back to first step, choose the next best template and repeat steps 2 to 5. The detailed process of comparative protein modeling has been previously defined in the

literature [145]. We have used a fully automated protein structure homology-modeling server SWISS-MODEL that provides all the databases for protein sequence and structure, and tools for template selection, model building and structure quality evaluation with a simplified user interface and workflow [148-150]. It uses Qualitative Model Energy Analysis (QMEAN), a composite scoring function based on four different geometrical properties that provides both global (i.e., for the entire structure) and local (i.e., per residue) absolute quality estimates [151]. QMEAN4 < - 4 indicates very low quality. A general overview of homology modeling of the *BspCBM17* target structure from the *CcCBM17* template structure (PDB ID – 1J83), having 53% sequence identity, is illustrated in Figure 2.1.

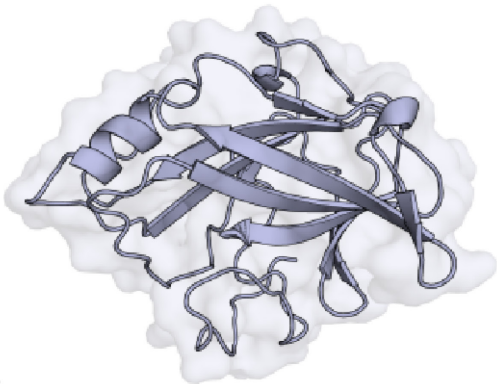
Template	Seq Identity	Oligo-state	Found by	Method	Resolution	Seq Similarity	Range	Coverage	Description
1j83.1.A	53.41	monomer	HHblits	X-ray	1.70Å	0.45	24 - 200	0.88	ENDO-1,4-BETA GLUCANASE ENG








Alignment

Target	VWVPEELSLSGEYVRARIKGVNYEPIDRTKYTKVLWDFNDGTQKQFGVNGDSPVEDVVIENEAGALKLSGLD--ASNDVS
1j83.1.A	-----QPTAPKDFSSGFWDFNDGTTQCGFVNPDSPITAINVENANNALKISNLNSKGSNDLS

Model building



QMEAN4	-2.34	
Cβ	-0.94	
All Atom	-2.57	
Solvation	-3.20	
Torsion	-0.66	

Model evaluation

Model #01	File	Built with	Oligo-State	Ligands	GMQE	QMEAN4
	PDB	ProMod Version 3.70.	MONOMER	None	0.75	-2.34

Figure 2.1 Homology modeling of *Bsp*CBM17 using the SWISS-MODEL. The QMEAN4 for this model was – 2.34 suggesting acceptable quality of the model.

2.1.2 Molecular Docking

When two macromolecules, such as a protein and a carbohydrate, are known through biochemical studies to form a complex, docking of one molecule to another can be utilized to understand physical interactions, binding mechanisms, and binding affinity between those two molecules; this presumes there is either sufficient structural evidence of similar associations from which to model or biochemical evidence identifying the protein binding site. There are multiple approaches to docking, and we implement three different docking methods in this dissertation, each of which is appropriate to the available experimental data and binding scenario. Our docking cases include two cases where the binding site is clearly identified from structural evidence (bound docking) and one case wherein a biochemical association suggests a binding site (unbound docking).

2.1.2.1 Docking of oligomeric ligands through pairwise alignment

In the case of Type B CBMs, we had a ‘bound docking’ situation where the cello-oligomeric ligand was common and CBMs were different. The general assumption behind this docking approach is that, with very high structural and chemical similarity of binding sites in CBMs from same family the oligosaccharide ligands would occupy approximately same position in the binding sites. For CBMs with only apo structures available, we transferred the coordinates of the cello-oligomer from a holo structure of a highly homologous CBM from the same family after a pairwise alignment. We used the DALI pairwise comparison tool to obtain the aligned structures [152]. After the alignment, we copied the cello-oligomer into the binding site of the target CBM, which was followed by an extensive vacuum minimization protocol in the MD simulation setup process. This docking methodology was used in the cases of *Cf*CBM4-2, *Bsp*CBM17

(apo structure obtained through homology modeling), and *Bsp*CBM28 where the cello-oligomer was docked from the available holo structures of *Cj*CBM4-1, *Cc*CBM17, and *Cj*CBM28, respectively. A general overview of this docking is shown in Figure 2.2.

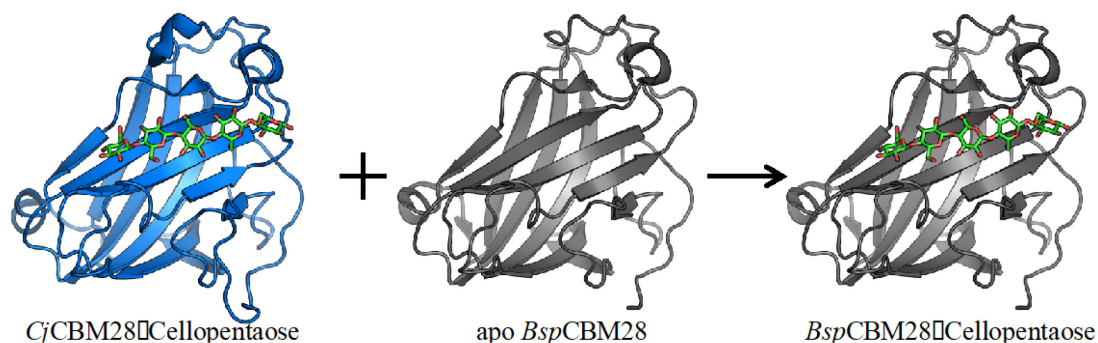


Figure 2.2 Docking of cello-oligomers through pairwise alignment. Illustrated here is an example of docking cellopentaose with apo *Bsp*CBM28 (PDB 1UWW) structure using the holo structure of *Cj*CBM28 (PDB 3ACI).

2.1.2.2 Oligosaccharide docking based on structural similarity

All the carbohydrate ligands considered in search of the potential physiological ligand of YKL-40 are composed of monomers with six-membered pyranose ring as the structural backbone. In contrast with CBMs, here we have a common protein with variable ligands. We utilized the structural similarity of the ligands to dock the monomeric units of the desired oligosaccharide at the most favorable positions based on the known coordinates of the bound chito-oligomer in the YKL-40 crystal structure. The chair conformation of six-membered pyranose ring, being mostly symmetrical, allows approximate positioning of the structurally similar monomer of the desired oligosaccharide. The sidechains of the ligand were then built by using the internal coordinate data (i.e., standard bond distances and angles for a given residue) provided in

the topology of the respective monomeric units. An example of cellohexaose docking based on chitohexaose coordinates is illustrated in Figure 2.3.

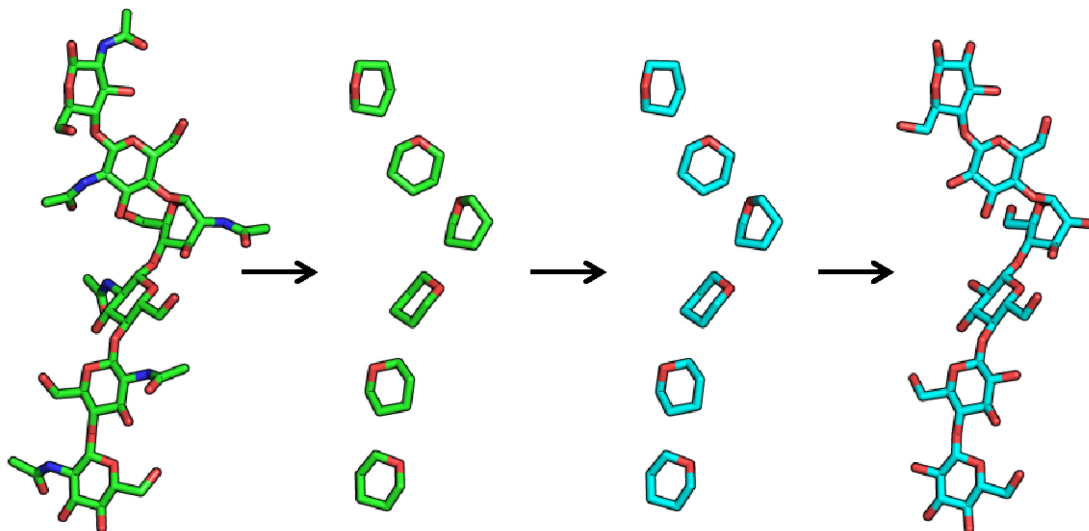


Figure 2.3 Transition from chitohexaose (green-red sticks) to cellohexaose (cyan-red sticks) using structural similarity and sidechain rebuilding with internal coordinates from topology database. This procedure is loosely referred to as ‘docking’ in this dissertation.

2.1.2.3 Protein-protein docking with shape complementarity

To identify the surface-binding site on YKL-40 for relatively large ligands like the triple helical collagen peptides, we used a different approach with ‘unbound docking’ of rigid molecules. The binding affinity between two molecules depends upon non-bonded interactions such as van der Waals and electrostatic contributions, but it is also necessary that their shapes are complementary to each other. Making the naïve assumption that the interacting molecules have rigid surfaces, matching local shape features, like local curvature maxima and minima, has been previously reported to correctly predict biomolecular association [153-155]. We implemented the shape

complementarity docking method by using the algorithm called PatchDock, which was inspired by object recognition and image segmentation techniques used in computer vision [156].

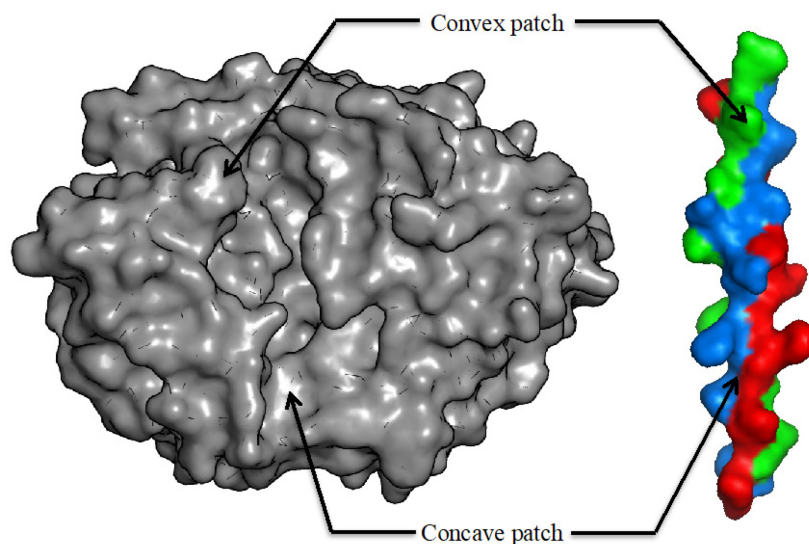


Figure 2.4 Surface representation of YKL-40 (gray) and collagen triple helix (multicolored) illustrating the concave and convex patches.

The algorithm has three major stages [154], explained here in brief:

1. **Molecular Shape Representation** – In this step, the molecular surface of the molecule is computed. Next, a segmentation algorithm is applied for detection of geometric patches (concave, convex and flat surface pieces). The patches are filtered, so that only patches with 'hot spot' residues are retained.
2. **Surface Patch Matching** – A hybrid of the Geometric Hashing and Pose-Clustering matching techniques are applied to match the patches detected in the previous step. Concave patches are matched with convex and flat patches with any type of patches.

3. Filtering and Scoring – The candidate complexes from the previous step are examined. All complexes with unacceptable penetrations of the atoms of the receptor to the atoms of the ligand are discarded. Finally, the remaining candidates are ranked according to a geometric shape complementarity score.

2.2 Molecular dynamics (MD) Simulations

MD simulations implemented in this study are a computational approach to predicting atomic-level interactions in a system of molecules based on time-dependent calculation of atomic positions using classical mechanics and empirically-derived forces between all the atoms in the system (i.e., force fields) [157, 158]. When such calculations are performed with very small time-steps (i.e., 1-2 femtoseconds) consecutively over a period of time to obtain a trajectory, the statistical analysis of this large data set of atomic positions in time can provide answers to questions of biological interest, such as non-catalytic protein-carbohydrate binding interactions. MD simulation of biomolecules has become ‘a computational microscope’ in molecular biology and is expected to have a significant impact on transformation of process of drug discovery [159, 160].

For N number of atoms in a system, the forces acting on each atom are governed by the potential energy function, $U(r^N)$, which is function of the position of each atom, r^N . The trajectory of the atoms is determined using an integrator, to solve Newton’s second law of motion (Equations 2.1 and 2.2). There are various versions of the Verlet Integrator [161, 162], ‘velocity Verlet’ [163, 164] and a ‘leapfrog form’ [165] being the most commonly used versions.

$$f_i = m_i \frac{\partial^2 \vec{r}_i}{\partial t^2} \quad \dots \text{Eq 2.1}$$

$$f_i = -\frac{\partial}{\partial r_i} U(r_1, r_1, r_1, r_1, \dots, r_N) \quad \dots \text{Eq 2.2}$$

where, f_i is the force acting upon atom i , m_i is the mass of the atom, a_i is the acceleration of the atom, and r is the position vector of the atom. $i=(1,2,3,\dots,N)$

The sum of bonded and non-bonded contributions represents the total potential energy of the system as a function of atomic coordinates (Equation 2.3). The bonded terms (Equation 2.4) include contributions from stretching of bonds (b) from equilibrium bond length (b_0), where K_b is the force constant; bending of angles (θ) from the equilibrium angle (θ_0), where K_θ is the force constant; rotation of dihedral angles (φ) with a phase shift (δ), where n is the periodicity of the dihedral angle and K_φ is the force constant; perturbation of improper angles (ω) from the equilibrium improper angle (ω_0), where K_ω is the force constant; the Urey-Bradley vibrational term (U_B, S); and the backbone torsional correction factor (CMAP, φ, ψ). The non-bonded terms (Equation 2.5) include summation, over pairs of atoms, of Coulombic interactions between point atomic charges (q_i and q_j) and the Lennard-Jones (LJ) 6-12 term, where ϵ_{ij}^{\min} represents the depth of the potential well, R_{ij}^{\min} is the distance at which the LJ potential reaches its minimum value, and r_{ij} is the interatomic distance between two atoms, i and j . The r_{ij}^{-12} term represents the short-range repulsive interaction, and the r_{ij}^{-6} term represents the long-range attractive/dispersive interaction of the LJ potential.

$$U(\vec{r}) = U_{\text{bonded}} + U_{\text{non-bonded}} \quad \dots Eq \ 2.3$$

$$U_{\text{bonded}} = \sum_{\text{bonds}} K_b (b - b_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\varphi (1 + \cos(n\varphi - \delta))$$

$$+ \sum_{\text{impropers}} K_\omega (\omega - \omega_0)^2 + \sum_{\text{Urey-Bradley}} K_{\text{UB}} (S - S_0)^2 + \sum_{\text{residues}} U_{\text{CMAP}}(\varphi, \psi) \dots Eq \ 2.4$$

$$U_{\text{non-bonded}} = \sum_{\text{van der Waals}} \epsilon_{ij}^{\text{min}} \left[\left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^6 \right] + \sum_{\text{coulombic}} \frac{q_i q_j}{4\pi\epsilon_0 \epsilon r_{ij}} \quad \dots Eq \ 2.5$$

The potential energy function, $U(\vec{r})$ in Equations 2.3, 2.4, and 2.5 [166], is the functional form of potential energy that can be used with the CHARMM36 all atom force-field, where the equilibrium values and force constants are implemented based on quantum mechanical calculations or experimental data. Detailed information about the algorithms to numerically solve the equations of motion, application of constraints, and periodic boundary conditions, and temperature and pressure control methods can be found in literature [167, 168]; also, details of our simulation protocols and parameter selections have been provided in the methods section of each chapter with corresponding references. In this dissertation, we have used CHARMM [166] and NAMD [169] as simulation software packages in association with VMD [170] and PyMOL [171] as molecular modeling and visualization tools. The latter two of which provide various modules to perform and analyze the MD simulations according to scientific needs.

2.3 Free energy calculations

While understanding the dynamic binding mechanisms of two biomolecules is important, knowing the free energy change associated with a binding event is also important, as it adds context to the characterization of ligand binding site dynamics and enables comparison of thermodynamic favorability in the case of multiple potential ligands. Free energy is a state function, and, although the difference in free energies of two states of a system is independent of the path chosen, one must pay attention to convergence and optimize use of computational resources. Here, we have used two enhanced-sampling free energy computation approaches to determine binding affinities.

2.3.1 Free Energy Perturbation with Hamiltonian Replica Exchange Molecular Dynamics (FEP/ λ -REMD)

FEP/ λ -REMD is an alchemical free energy technique with enhanced-sampling, i.e., replica exchange, to calculate relatively accurate absolute binding free energies of small molecules to proteins. A computationally inexpensive FEP simulation protocol was first developed by Deng and Roux, applicable to small ligands like benzene [172, 173]. With the availability of much faster and less expensive computational resources, Jiang et. al. [174, 175] modified this protocol by using FEP in association with Hamiltonian replica exchange MD to significantly improve the sampling and accelerate the convergence of computations, expanding application to a wider range of ligand molecules. The binding affinities of proteins for oligosaccharides calculated using this method are directly comparable to experimental binding free energies of the same systems measured using ITC.

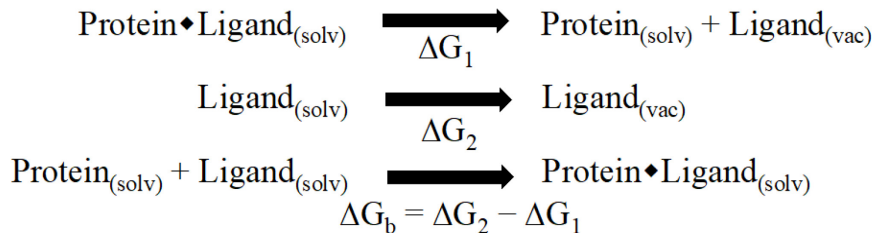


Figure 2.5 The thermodynamic pathway implemented in FEP/ λ -REMD to obtain ligand binding free energy. “Solv” refers to the solvated system and “Vac” refers to the vacuum.

The overall alchemical pathway used in this method (Figure 2.5) to calculate the absolute binding free energy (ΔG_b) between protein and ligand includes two independent steps: i) decoupling of the ligand interactions from the protein-ligand complex in solution and ii) decoupling of the ligand interactions from the solvated ligand without the protein. The difference between the free energy changes of these two steps gives ΔG_b . In each step of this pathway, the free energy change is calculated by gathering contributions from the distribution of the potential energy function (Equation 2.6) into non-bonded interactions and restraints.

$U = U_0 + \lambda_{rep} U_{rep} + \lambda_{disp} U_{disp} + \lambda_{elec} U_{elec} + \lambda_{rstr} U_{rstr}$	<i>... Eq. 2.6</i>
---	--------------------

where, U_0 is the potential energy of the system with totally non-interacting ligand

Thermodynamic coupling parameters, λ_{rep} , λ_{disp} , and λ_{elec} , that vary from ‘0’ to ‘1,’ representing transition from full interaction to full decoupling of repulsive, dispersive and electrostatic interactions, respectively, are uniformly distributed over the number of replicas dedicated to the type of interaction. Figure 2.6 illustrates the implementation of replica distribution in which we used 128 replicas in total, with 72 repulsive, 32

dispersive, and 24 electrostatic replicas. Each replica from the set of replicas within a given type occupies multiple processors, in a parallel/parallel mode. The probability of whether the λ -swap between the given two replicas will happen or not is calculated by a conventional Metropolis Monte Carlo algorithm [174]. In our study, the fourth contribution from restraints, usually having relatively little contribution (needed only in the 1st step to maintain the distance between the center of mass of protein and that of ligand), was not a part of the replica exchange protocol to reduce required computational time. Instead, MD simulations with λ_{rstr} distributed over 13 consecutive windows of 0.1 ns each, followed by numerical integration with Simpson's rule [173] were used to determine the contribution to free energy from added restraints.

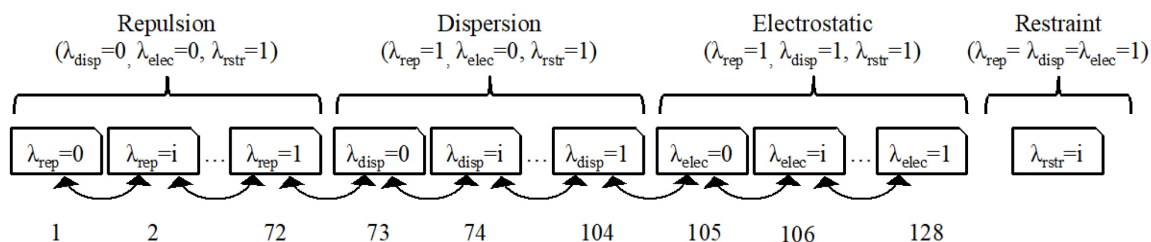


Figure 2.6 Scheme of replica distribution and exchange in FEP/ λ -REMD, as implemented in this dissertation. Each box represents an individual MD simulation with specified conditions of λ . The arrows represent the possible attempts of λ -swap with neighboring replicas after each replica has completed the MD steps specified by replica exchange frequency. Restraining contributions were not part of replica exchange in our case. The scheme has been adapted with permission from Jiang et. al. [175]. Copyright 2010 American Chemical Society.

For the three main contributions, we run these set of 128 replicas for 20 or more consecutive (but independent) windows of 0.1 ns each. After collecting the output potential energies from all the replicas and regrouping the values corresponding to λ -values for each type of interactions at the end of each window, multistate Bennett Acceptance Ratio (MBAR) method was used to determine the free energy changes for individual contributions along with their statistical uncertainties [176]. The convergence of the calculations was assured as the windows were analyzed sequentially to plot the time progression of the total free energy change. The calculation was continued until we observed converged output for 1 ns where 20 windows of 0.1-ns each were sufficient in most cases in this study. The final free energy change was reported as the average of the last 1 ns data. The error of the averaged free energy change was reported as 1 standard deviation. More specifics of use of this method in this dissertation are provided in the methods section of respective chapters. This free energy calculation method can now be implemented through a dedicated module in the widely used simulation software package, NAMD (version 2.12) [169], while we used a developer version in NAMD provided by Wei Jiang, Argonne National Laboratory.

2.3.2 Umbrella Sampling

In principle, umbrella sampling is merely several MD simulations conducted over the range of a pre-defined reaction coordinate (RC). A biasing potential term is added to the conventional potential energy function, directing the simulation to cross energy barriers and sample the conformational space typically inaccessible to unbiased classical MD [177, 178]. Umbrella sampling can be used to determine the free energy surface of the system, also referred to as potential of mean force (PMF), along that path of the

thermodynamic system from reference state to target state [179]. For calculation of the PMF, usually a harmonic restraint is used as the biasing potential, a function of reaction coordinate. This RC can be any observable like distance between protein and ligand ultimately defining the desired path between two states. Enhanced sampling is performed consecutively for windows covering the range of a selected RC that dictates the desired path.

$$V^{US}(\eta(\vec{r})) = \frac{1}{2} K_{US}(\eta(\vec{r}) - \eta_0)^2 \quad \dots \text{Eq. 2.7}$$

where, $V^{US}(\eta)$ is the biasing potential, K_{US} is the restraining force constant, η is the instantaneous RC, and η_0 is the equilibrium value of RC in given window.

$$U^{US}(\vec{r}) = U(\vec{r}) + V^{US}(\eta(\vec{r})) \quad \dots \text{Eq. 2.8}$$

where, $U^{US}(\vec{r})$ is new biased total potential energy function and $U(\vec{r})$ is old unbiased potential energy function

The range of RC is divided into specific values (or windows), and MD simulations are run for each window, collecting the instantaneous values of RC. Either weighted histogram analysis method (WHAM) [180, 181] or multistate Bennett acceptance ratio (MBAR) [176] analysis can be used to estimate unbiased probability distribution and ultimately build the PMF profiles [182]. With WHAM, errors need to be computed separately, which is usually done by a standard bootstrapping method [181, 183], while MBAR has a direct way to calculate errors [176]. The free energy of binding is then determined from the PMF by taking the difference between the free energy at RC

$= 1$ and $RC = 0$. For convergence and accuracy, two factors are closely observed, i)
equilibration of all umbrella-sampling windows where initial sampling data is discarded
ii) PMF profile near the ends of RC where minima/maxima/plateau is important for
accurate difference in free energy.

Chapter 3 – Cello-oligomer binding dynamics and bi-directional binding phenomenon in Type B CBMs

As the title suggests, Chapter 3 reports the bi-directional ligand binding phenomenon in all three families (4, 17 and 28) of Type B CBMs and cellopentaose binding dynamics in two family 4 CBMs. Most of this chapter, i.e. experiments related to family 4 CBMs, has been adapted with permission from Kognole and Payne [184], Copyright © 2015, Oxford University Press. As the experiments related to family 17 and 28 CBMs are going to be part of another journal article, we only add one subsection in this chapter to discuss the results relative to the topic of this chapter.

3.1 Abstract

Carbohydrate binding modules (CBMs) play significant roles in modulating the function of cellulases, and understanding the protein-carbohydrate recognition mechanisms by which CBMs selectively bind substrate is critical to development of enhanced biomass conversion technology. CBMs exhibit a limited range of specificity and appear to bind polysaccharides in a directional fashion dictated by the position of the ring oxygen relative to the protein fold. The two family 4 CBMs of *Cellulomonas fimi* Cel9B (*Cf*CBM4) are reported to preferentially bind cellulosic substrates. However, experimental evidence suggests these CBMs may not exhibit a thermodynamic preference for a particular orientation. We use molecular dynamics (MD) and free energy calculations to investigate protein-carbohydrate recognition mechanisms in *Cf*CBM4-1 and *Cf*CBM4-2 and to elucidate preferential ligand binding orientation. For *Cf*CBM4-1, we evaluate four cellopentaose orientations including that of the crystal structure and

three others suggested by NMR. These four orientations differ based on position of the ligand reducing end and pyranose ring orientations relative to the protein core. MD simulations indicate the plausible orientations reduce to two conformations. Calculated ligand binding free energy discerns each of the orientations is equally favorable. The calculated free energies are in excellent agreement with isothermal titration calorimetry measurements from literature. Through MD simulations we confirm the bi-directional binding of cellopentaose to four other Type B CBMs, two CBMs from family 17 and 28 each. These MD simulations further reveal the approximate structural symmetry of the oligosaccharides relative to the amino acids along the binding cleft plays a role in the promiscuity of ligand binding. A survey of ligand-bound structures insinuates this phenomenon may be characteristic of the broader class of proteins belonging to the β -sandwich fold.

3.2 Introduction

The multi-modular *Cellulomonas fimi* endoglucanase Cel9B (formerly CenC) exhibits tandem Type B, *N*-terminal CBMs, both of which belong to Family 4 [105]. The domains, *Cf*CBM4-1 (formerly CBD_{N1}) and *Cf*CBM4-2 (formerly CBD_{N2}), appear sequentially and additively bind amorphous cellulose [98]. *Cf*CBM4-1 is of historical significance as the first known soluble substrate-binding CBM, the discovery of which led to renewed interest in CBM function in general. Both *Cf*CBM4-1 and *Cf*CBM4-2 bind cellotetraose and cellopentaose with increasing affinity [98, 110], suggesting each binding cleft consists of 5 pyranose binding subsites formed by the parallel β -sheets of the β -sandwich (Figure 3.1); this was later confirmed when Boraston et al. solved the cellopentaose bound *Cf*CBM4-1 structure [62]. The affinity of *Cf*CBM4-1 and *Cf*CBM4-

2 for cello-oligomers is roughly the same for a given length, and isothermal titration calorimetry suggests binding of cello-oligomers to both CBM4s is enthalpically driven [98, 110]. This latter observation is consistent with the large population of potential hydrogen-bonding polar residues lining the binding cleft [99], compared to those from planar binding sites of Type A CBMs [16]. Despite the apparent similarities in specificity and binding mode, the two modules exhibit only 36% sequence identity with notable amino acid substitutions along the binding cleft. The binding cleft of *Cf*CBM4-2 is also noticeably wider than *Cf*CBM4-1 [62, 110]. We anticipated that direct comparison of the dynamics of cellopentaose-bound *Cf*CBM4-1 and *Cf*CBM4-2 will elucidate the fundamental interactions driving β -1,4-linked glucan specificity in Family 4 CBMs. Furthermore, these findings are likely to have broad applicability to other CBM families with β -sandwich folds.

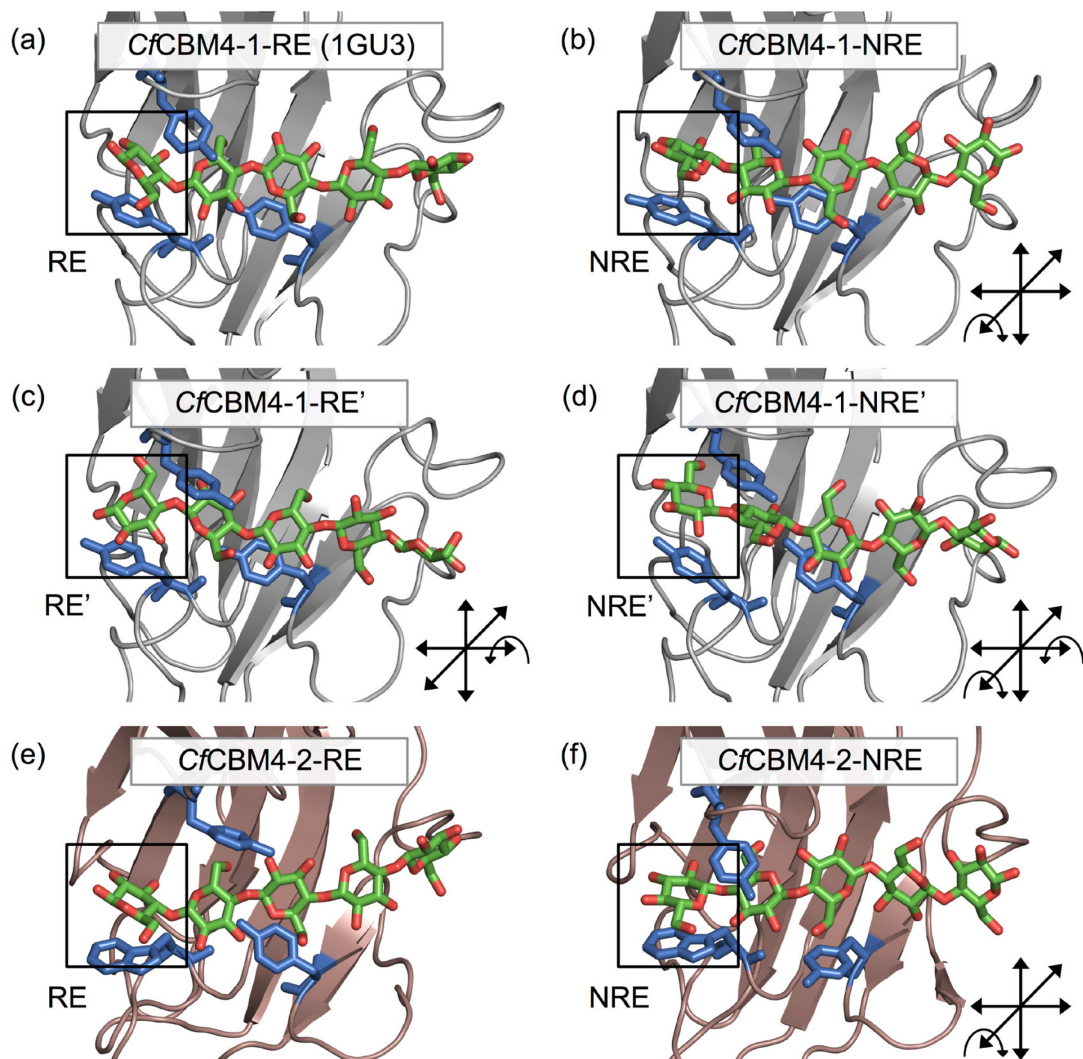


Figure 3.1 *Cj*CBM4-1 and *Cj*CBM4-2 ligand conformations considered in this study. *Cj*CBM4-1 is shown in gray cartoon. *Cj*CBM4-2 is shown in salmon cartoon, and cellopentaose is shown in green and red stick. (a) *Cj*CBM4-1-RE represents the ligand orientation of the *Cj*CBM4-1 structure (PDB 1GU3) with the reducing end (RE) in a left-to-right fashion, and (b) *Cj*CBM4-1- NRE illustrates the reverse, transverse axis transformation with the ligand oriented so the reducing end runs from right-to-left. (c) *Cj*CBM4-1-RE' represents the structural ligand orientation with the RE from left-to-right, but the cellopentaose has been rotated 180° about the length of the C1-C4 axis, locating

the hydroxymethyl groups out of register. (d) *Cj*CBM4-1- NRE' represents both the transverse axis rotation and the 180° C1-C4 rotation of the cellopentaose in the binding cleft. (e) *Cj*CBM4-2-RE represents the *Cj*CBM4-1 structural ligand orientation (PDB 1GU3) with the reducing end of the ligand running from left-to-right. (f) *Cj*CBM4-2-NRE represents the transverse axis transformation of cellopentaose so the reducing end runs from right-to-left.

NMR analysis of nitroxide spin-labeled cello-oligomer derivatives also put forth the intriguing, though somewhat controversial, hypothesis that *Cj*CBM4-1 and *Cj*CBM4-2 are capable of binding a cello-oligomer in a multi-directional fashion [61]. Johnson et al. examined association of 2,2,6,6-tetramethylpiperidine-1-oxyl-4-yl (TEMPO) labeled cellotriose and cellotetraose with the individual *Cj*CBM4-1 and *Cj*CBM4-2 domains [61]. At the time of this study, structural resolution of a ligand bound CBM4 was unavailable, and NMR techniques were a complementary approach toward understanding ligand binding in lieu of crystallographic evidence. Determination of ¹H and ¹⁵N chemical shifts confirmed labeling did not significantly affect affinity, and paramagnetic relaxation studies further revealed the nitroxide label could lie at either end of the binding clefts. However, relative occupancies were not determined as a means to suggest a “more favorable” conformation. The multi-directional binding observation is interesting because it is counter to intuition. Polysaccharides exhibit a large dipole along the length of the polymer as a result of several factors including the parallel orientation of chains, the asymmetric pyranose ring oxygen atom, and the chemical polarity of the individual chains [185, 186]; for the cellopentaose ligand, the dipole moment is approximately 12 D. On the surface, it seems such a dipole would preclude multi-directional binding, as

proteins would likely evolve in such a way as to most effectively hydrogen bond with the oligomer in a given direction. Boraston et al. reached a similar conclusion upon solution of the cellopentaose-bound *Cf*CBM4-1 structure in 2002 [62]. The structure captured the cellopentaose with one hydrophilic edge of the sugar pointed in toward the binding groove and the other edge exposed to solvent. Unambiguous electron density pointed to a single thermodynamically favorable conformation occupying the 5 subsites of the binding cleft. Nevertheless, the authors left open the possibility that serendipitous crystal packing interactions may have resulted in binding the least favorable cellopentaose orientation. Of course, the ability to multi-directionally bind of cello-oligomers would be significantly advantageous in engineered cellulase or cellulosomal constructs, allowing the CBMs to target amorphous cellulose from virtually any angle. Thus, determining whether this capability does in fact exist in a thermodynamically equivalent capacity and how multi-directional CBM4 substrate binding is accomplished promises to inform future biotechnological development.

3.3 Methods

3.3.1 Modeling of cello-oligomer in multiple orientations

We use molecular dynamics (MD) simulations to explicitly examine cello-oligomer binding mechanisms in Family 4 CBMs. MD simulations of eight total systems representing the various ligand configurations of *Cf*CBM4-1 and *Cf*CBM4-2 were conducted for 0.25 μ s. Six systems representing possible variations in binding cleft occupation were examined in addition to the two unbound proteins. Figure 3.1 illustrates the corresponding case/system abbreviation used throughout this study. The *Cf*CBM4-1 systems were constructed based on the 1GU3 Protein Data Bank (PDB) structure [62],

and the *Cf*CBM4-2 systems were constructed from the 1CX1 PDB structure [110]. As discussed above, the *Cf*CBM4-1 structure features a bound cellopentaose ligand, which was used here as the basis for investigation of ligand dynamics and directionality preference. Four ligand orientations bound to *Cf*CBM4-1 were considered representing: (1) the structural orientation (*Cf*CBM4-1-RE); (2) a reversed ligand orientation where the non-reducing end of the cellopentaose occupies the original reducing end position of the structural conformation and symmetry of the glucopyranose side chains is maintained (*Cf*CBM4-1-NRE); (3) a rotation of the structural cellopentaose conformation about C1-C4 axis so the opposite hydrophilic edge faces inward to the protein, effectively locating a C5 hydroxymethyl group where the C3 hydroxyl previously existed (*Cf*CBM4-1-RE'); and (4) a transverse axis reversal along with the C1-C4 rotation (*Cf*CBM4-1-NRE'). As the *Cf*CBM4-2 NMR structure does not contain a ligand, the bound *Cf*CBM4-2 systems were prepared by aligning *Cf*CBM4-2 to *Cf*CBM4-1 protein backbones and docking the cellopentaose to the *Cf*CBM4-2 structure. Two ligand orientations in *Cf*CBM4-2 were considered, representing the orientation of the 1GU3 structure (*Cf*CBM4-2-RE) and the transverse axis transformation (*Cf*CBM4-2-NRE). The unbound *Cf*CBM4-1 and *Cf*CBM4-2 systems were also considered to understand the contributions of ligand binding to protein dynamics. A detailed description of simulation construction is provided ahead.

3.3.2 MD simulation: Setup and parameters

All protein components of *Cf*CBM4-1 and *Cf*CBM4-2 simulations were constructed from crystal structures, manually docking the cellopentaose ligands as necessary through secondary structure alignment with crystal structure of *Cf*CBM4-1 that

already exhibits bound cellopentaose. The *Cf*CBM4-1 simulations were constructed from the 1GU3 PDB structure, in which *Cf*CBM4-1 binds cellopentaose in the binding cleft. The nomenclature used in this study reflects the orientation of the ligand captured in 1GU3 structure [62]; we have defined this as the “reducing end” conformation of the bound ligand (*Cf*CBM4-1-RE), numbering the pyranose moieties from 1 to 5 accordingly (Figure 3.2). The “non-reducing end” conformation (*Cf*CBM4-1-NRE) was prepared from this same structure. To reverse the ligand direction, the coordinates of the heavy ring atoms were retained from the 1GU3 structure, and atom types were reassigned so as to locate the pyranose ring oxygen at the opposite end of the cleft (Figure 3.1). CHARMM was used to reconstruct the remaining hydrogens and primary alcohol groups from internal coordinate tables [166]. The reducing end and non-reducing end cellopentaose conformations bound to *Cf*CBM4-2, *Cf*CBM4-2-RE and *Cf*CBM4-2-NRE, respectively, were constructed by docking the *Cf*CBM4-1-RE and *Cf*CBM4-1-NRE ligands through structural alignment with the 1CX1 PDB structure [110]. To prepare the *Cf*CBM4-1-RE’ and *Cf*CBM4-1-NRE’ conformations, the coordinates of the pyranose ring heavy atoms were again renamed such that the ligand was rotated along the longitudinal axis relative to *Cf*CBM4-1-RE and *Cf*CBM4-1-NRE, respectively. CHARMM was used to reconstruct remaining hydrogen and primary alcohol groups. Protonation states of the titratable residues were determined using H++ and manual inspection of the protein environment (i.e., possible salt bridge formation) [187-190]. PyMOL and VMD were used for structural alignment and visualization [170, 171].

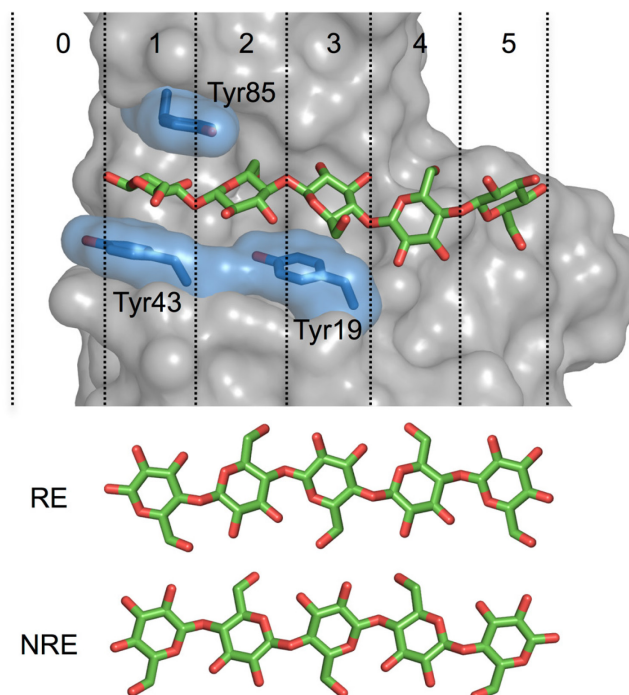


Figure 3.2 Binding site nomenclature for *CfCBM4-1*. The CBM binds cellopentaose along five individual binding subsites perpendicular to the β -sheets forming the protein core. These subsites are numbered from 1 to 5. Here, we define an additional “binding subsite,” 0, for discussion of MD simulations of *CfCBM4-1-RE*’ and *CfCBM4-1-NRE*’. Subsite 0 represents a completely solvent exposed pyranose ring of the cellopentaose chain. The bottom panel illustrates the symmetry of a cello-oligomer oriented in the opposite direction. The primary alcohol groups remain in approximately the same location regardless of direction.

All constructed systems were vacuum minimized, solvated with water, neutralized with sodium ions, and minimized again. The minimized systems were then heated to 300 K and density equilibrated in CHARMM. After equilibration, the systems were simulated for 250 ns at 300 K in the NVT ensemble using NAMD [169]. All simulations used the CHARMM36 force-field with CMAP correction for proteins [166, 191, 192], and the

CHARMM36 carbohydrate force-field for the cellopentaose ligands [193-195]. The modified TIP3P force-field was applied to water molecules [196, 197]. Analysis of the 250 ns MD simulations included: determination of the RMSD and RMSF of the protein backbones, the RMSF of cellopentaose on a per binding subsite basis, the hydrogen bonding and interaction energies of each glucose residue with protein, and average solvation of the ligand on per binding subsite basis.

3.3.3 Free Energy Calculation: FEP/ λ -REMD

We quantitatively examined thermodynamic preference of ligand directionality through a computational determination of absolute binding free energy. An enhanced sampling free energy methodology, Free Energy Perturbation with Hamiltonian Replica Exchange Molecular Dynamics (FEP/ λ -REMD), was used to calculate the affinity of cellopentaose to *Cf*CBM4-1 [174]. We considered two cases representing the structural orientation and the transverse axis rotation, *Cf*CBM4-1-RE and *Cf*CBM4-1-NRE, respectively. The remaining two *Cf*CBM4-1 ligand orientations, rotations about the C1-C4 axis, were excluded from free energy calculations as the ligands significantly shifted along the length of the binding cleft over the course of the MD equilibration simulations and no longer represented the intended conformational state.

This free energy calculation protocol couples free energy perturbation with Hamiltonian replica-exchange molecular dynamics to enhance Boltzmann sampling [173, 174]. The calculations were performed by decoupling the potential energy into four separate contributions scaled according to coupling parameters, defined mathematically by Jiang et al [174]. In short, the contributions to overall free energy included the shifted

Weeks-Chandler Anderson repulsive and dispersive components, ΔG_{repu} and ΔG_{disp} , respectively, and the electrostatics contribution, ΔG_{elec} . Additionally, contributions from an applied restraining potential, where necessary, were considered, ΔG_{rstr} . We used the thermodynamic cycle in Figure 3.3 to arrive at the free energy of binding a cellopentaose to *Cj*CBM4-1. The cycle consisted of two separate sets of calculations: (1) decoupling the bound cellopentaose from the solvated *Cj*CBM4-1 and (2) decoupling the solvated cellopentaose from solution. The difference between the two values is the standard binding free energy, ΔG_b . The restraining potential was used only in the first leg of the cycle, decoupling cellopentaose from *Cj*CBM4-1.

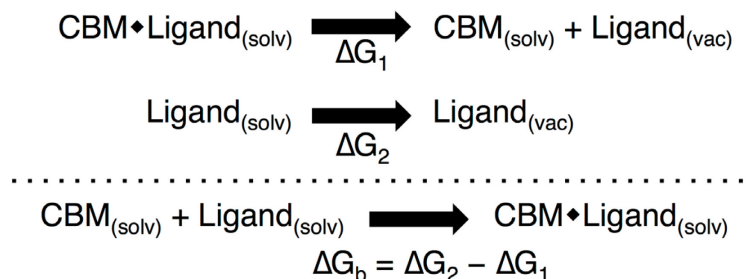


Figure 3.3 Thermodynamic cycle used to determine ligand binding free energy from FEP/ λ - REMD. In this case, “CBM” is *Cj*CBM4-1 and “ligand” is cellopentaose. The subscripts “solv” and “vac” refer to the solvated and vacuum (or decoupled) systems, respectively.

The free energy calculations were constructed from 4 ns snapshots from the explicitly solvated *Cj*CBM4-1-RE and *Cj*CBM4-1-NRE MD simulations. The absolute binding free energy was determined from 40 consecutive 0.1 ns calculations, where the first 1 ns data was discarded as equilibration. The simulations used a set of 128 replicas (72 repulsive, 24 dispersive, and 32 electrostatic) with an exchange frequency of 1/100

steps (every 0.1 ps). The *Cf*CBM4-1/cellopentaose systems included a positional restraint defined by the distance of the center of the mass of the ligand to the center of mass of the protein. This restraint bias during the decoupling of cellopentaose from the solvated complex to vacuum was determined by numerical integration with Simpsons' rule [173]. The output energies collected during simulation were post-processed using the Multistate Bennett Acceptance Ratio (MBAR) to calculate the free energies and statistical uncertainty of the individual repulsive, dispersive, and electrostatic contributions [176]. Finally, summation of all the four contributions gives total free energy change for each leg of the thermodynamic pathway (i.e., ΔG_1 and ΔG_2). The binding free energy of cellopentaose to *Cf*CBM4-1 is the difference between the free energy of decoupling of solvated cellopentaose from solution and the free energy of decoupling of bound cellopentaose from *Cf*CBM4-1 (i.e., $\Delta G_b = \Delta G_2 - \Delta G_1$; Figure 3.3). As described above, the standard deviation of these values over the 3 ns data collection period, which were combined using error propagation rules, is reported as the final binding free energy error. Convergence was determined by monitoring the time evolution of the free energy calculations. Additional details are provided in Appendix 1.

3.4 Results and Discussion

3.4.1 Symmetry of the cellopentaose is critical to binding

Four possible cellopentaose conformations occupying the *Cf*CBM4-1 binding groove were investigated as potential multi-directional binding forms. Two of these conformations, *Cf*CBM4-1-RE' and *Cf*CBM4-1-NRE', were constructed so as to test the suitability of the binding groove to accommodate larger carbohydrate side chain groups, such as the hydroxymethyl group, regardless of binding subsite. The nomenclature of

different binding subsites for *Cf*CBM4-1 is illustrated in Figure 3.2. These two systems, constructed by rotating the ligand around its longitudinal axis (Figures 3.1c and 3.1d), place the cellopentaose off register by one binding subsite compared to the structurally bound ligand. Acceptance of the latter two ligand conformations would require each of the binding subsites to consist of semi-redundant hydrogen bonding residues in every binding subsite.

MD simulations indicate the *Cf*CBM4-1 binding groove will not accept the cellopentaose with the hydroxymethyl group arbitrarily located along the groove. This result is immediately evident from visualization of both the *Cf*CBM4-1-RE' and *Cf*CBM4-1-NRE' trajectories (Movie 3.1 and 3.2). From the *Cf*CBM4-1-NRE' trajectory, we observe the cellopentaose shift longitudinally across the groove within 2 ns of the 250 ns simulation (Figure 3.4, Movie 3.2). The displacement of the cellopentaose exposes a glucopyranose moiety to solvent, external to the binding groove, leaving only four moieties bound in the groove. For the purposes of describing ligand dynamics going forward, we have numbered this external "binding subsite" as "0" (Figure 3.2). An equivalent shift occurred at 8 ns in the *Cf*CBM4-1-RE' simulation (Movie 3.1). As described in the Methods, each of these starting configurations was extensively minimized in a stepwise fashion, significantly reducing the possibility that unfavorable molecular contacts influenced the ability of the cellopentaose to occupy the alternative binding site. Additionally, each of these simulations was independently repeated varying the random number seed, and the same shift of the cellopentaose across the binding groove was observed. In the remaining two cellopentaose conformations, *Cf*CBM4-1-RE and *Cf*CBM4-1-NRE, this displacement was not observed.

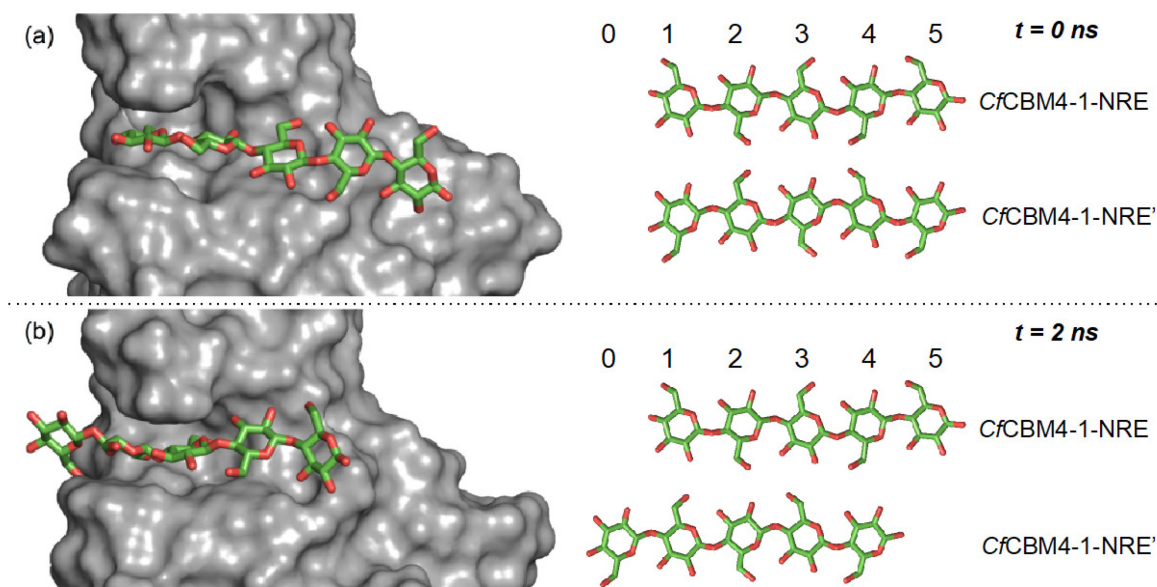


Figure 3.4 Snapshots from the *Cj*CBM4-1-NRE' simulation at (a) 0 ns and (b) 2 ns. The protein is shown in gray surface, and the ligand is shown in green and red stick. The snapshots illustrate the ligand, initially out of register from the structurally bound position, naturally sliding to the more energetically favorable position, defined by the position of the hydroxymethyl side chain facing into the binding cleft. Schematic for this sliding is illustrated on the panels to right in comparison with *Cj*CBM4-1-NRE.

Johnson et al. suggested that the approximate structural symmetry of oligosaccharides accounts for the ability of the protein to bind the cello-oligomer regardless of directionality [61]. That is to say, upon reversing the cellopentaose within the binding site, the hydroxymethyl group occupies roughly the same position as the hydroxyl group, which may allow for similar hydrogen bonding. Rotating the cellopentaose, as in the *Cj*CBM4-1-RE' and *Cj*CBM4-1-NRE' cases, effectively disrupts this structural symmetry. Positioning the ligand so that the hydrogen-bonding side chains are no longer occupying symmetrically similar locations, the cellopentaose is no longer

able to make the hydrogen bonds necessary to bind within the active site, as we will show through explicit characterization of hydrogen bonding. Naturally, the ligand was displaced by one glucopyranose moiety as it readjusted its side chains similar to *CfCBM4-1-NRE* or *CfCBM4-1-RE*. After the cellopentaose reached an equilibrium position, the *CfCBM4-1-NRE'* and *CfCBM4-1-RE'* cases were approximately equivalent to *CfCBM4-1-NRE* and *CfCBM4-1-RE*, respectively.

In the upcoming sections, we discuss the results of the *CfCBM4-1-NRE'* and *CfCBM4-1-RE'* by comparing the equilibrium position of the cellopentaose, with four protein-bound bound moieties and a singular “external” moiety in subsite 0. Furthermore, as the *CfCBM4-1-RE'* and *CfCBM4-1-NRE'* cases are approximately equivalent to *CfCBM4-1-RE* and *CfCBM4-1-NRE*, respectively, we did not perform free energy calculations on the former two cases.

3.4.2 Thermodynamic preference of cello-oligomer orientation

A primary question we have sought to address by this study is whether *CfCBM4-1* has a thermodynamic preference for a given bound cello-oligomer conformation given the inconclusive nature of experimental approaches to date. We used FEP/ λ -REMD to calculate the free energy of binding a cellopentaose ligand to the *CfCBM4-1* binding groove in two different orientations, *CfCBM4-1-RE* and *CfCBM4-1-NRE*, having narrowed down putative binding conformations using MD. The free energy of binding cellopentaose to *CfCBM4-1* in either the *CfCBM4-1-RE* or *CfCBM4-1-NRE* conformation was approximately equal. As shown in Table 3.1, the binding free energies were within error at -4.52 ± 1.29 kcal mol⁻¹ and -5.86 ± 1.51 kcal mol⁻¹ for *CfCBM4-1-*

RE and *Cf*CBM4-1-NRE, respectively. The repulsive, dispersive, electrostatics, and restraining potential contributions are provided individually. The free energies of each step in the thermodynamic cycle, ΔG_1 and ΔG_2 , were obtained by summing these contributions. Error calculation is explained ahead.

Table 3.1 Binding free energies of cellopentaose to *Cf*CBM4-1 in two ligand orientations representing bi-directional binding. The solvation free energy of cellopentaose, ΔG_2 , is also tabulated as its three contributions – repulsion, dispersion, and electrostatics.

	ΔG_b (kcal mol ⁻¹)	ΔG_{repu} (kcal mol ⁻¹)	ΔG_{disp} (kcal mol ⁻¹)	ΔG_{elec} (kcal mol ⁻¹)	ΔG_{rstr} (kcal mol ⁻¹)
Cellopentaose	-	68.08 ± 0.38	- 61.81 ± 0.12	- 66.28 ± 0.33	-
<i>Cf</i> CBM4-1-RE	- 4.52 ± 1.29	73.86 ± 1.10	- 78.91 ± 0.19	- 59.21 ± 0.55	- 0.28
<i>Cf</i> CBM4-1-NRE	- 5.86 ± 1.51	74.25 ± 1.15	- 78.89 ± 0.29	- 61.29 ± 0.6	0.06
<i>Cf</i> CBM4-1 Experimental ^a	- 5.24 ± 0.91	-	-	-	-

^a. Tomme, Creagh [98]

The corresponding error values represent standard deviations over the final 30 of 40 intervals, i.e., the final 3 ns of 4 ns total. The free energy over the course of the 4 ns calculation, in 100 ps intervals, is given in Figure 3.5. The error of the binding free energy, ΔG_b , was obtained by taking the square root of the sum of the squared standard deviations of the free energy of decoupling cellopentaose from *Cf*CBM4-1 and the

cellopentaose solvation free energy, ΔG_1 and ΔG_2 . Error calculations based on statistical correlation of the data for each 100 ps interval are reported in the Supplementary Data (Figure 3.5). We have chosen to report the standard deviation here, as this represents the larger of the two values. Progress toward convergence was assessed by monitoring the time evolution of the free energy calculation (Figure 3.5). The effect of replica-exchange frequency on the sampling and convergence of the binding free energy in the case of *Cf*CBM4-1-RE was also considered (Details are provided in Appendix 1).

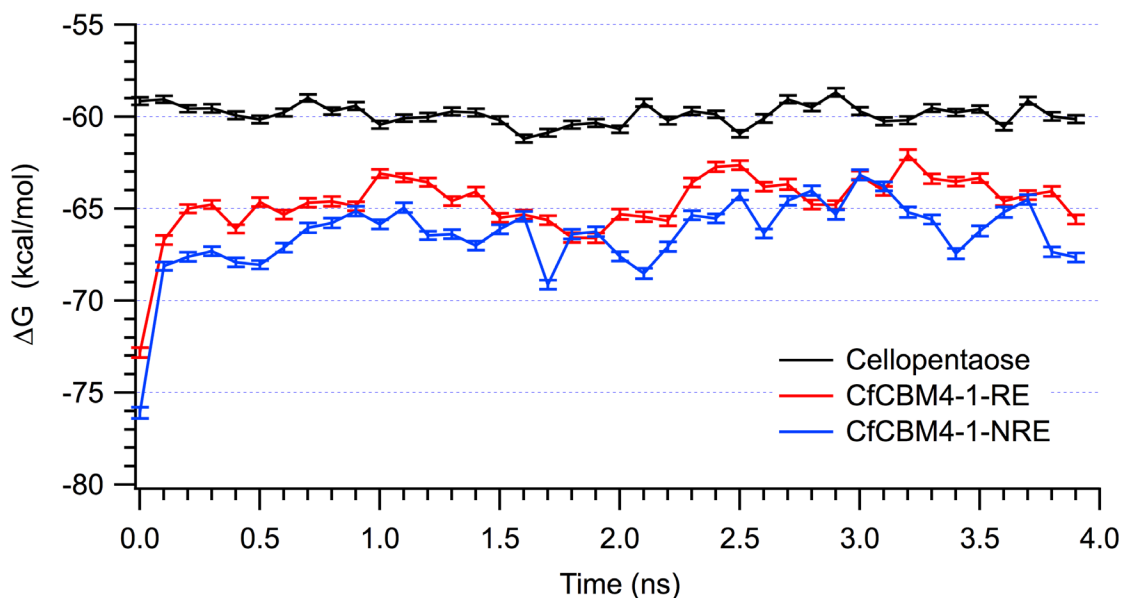


Figure 3.5 Calculated Gibbs free energy over 40 consecutive 0.1-ns calculations using FEP/ λ -REMD. The difference between the average value for either *Cf*CBM4-1-RE or *Cf*CBM4-1-NRE and the cellopentaose solvation free energy represents the binding free energy.

The calculated binding free energies were in excellent agreement with a previously measured value obtained by isothermal titration calorimetry (ITC) at 35°C [98]. The reported value of cellopentaose binding to *Cf*CBM4-1 in pure water at 35°C is -5.24 ± 0.91 kcal mol⁻¹. As ITC does not provide structural-level resolution of ligand binding, the experimental binding free energy likely represents the ensemble of both putative binding conformations. Considering the accuracy of both ITC and free energy calculations [198, 199], the difference between the two is relatively insignificant. Calculated free energies of binding cellopentaose to *Cf*CBM4-1 support the hypothesis that *Cf*CBM4-1 possesses the ability to bi-directionally bind cello-oligomers. Our findings of approximate thermodynamic equality are in line with the original Johnson et al. study using TEMPO-labeled cello-oligomers coupled with NMR to observe ligand binding [61]. The crystallographic structure, later captured by Boraston, et al., temptingly suggests that *Cf*CBM4-1 binds cellopentaose in a single, thermodynamically favorable orientation relative to the binding cleft [62]. Boraston et al. describe how the distance-dependent nature of the NMR spin-labeling analysis prohibits calculation of relative occupancy of each of the ligand binding conformations, dismissing the possibility that bi-directional binding represents anything more than a low-occupancy state. While free energy calculations also suffer from the inability to capture the statistical likelihood of a given orientation, the equality of the free energy of binding cellopentaose in either the *Cf*CBM4-1-RE or the *Cf*CBM4-1-NRE conformation suggests occupancy of each state is equally likely and the captured structural orientation was a result of circumstance or experimental conditions of crystallography.

3.4.3 CfCBM4-1 hydrogen bonding

The cellopentaose ligand formed approximately the same number of hydrogen bonds in each binding subsite regardless of the direction of the ligand (Figure 3.6). VMD was used to determine the average number of hydrogen bonds formed per pyranose ring and side chain with the surrounding protein [170]. The criteria used to define a hydrogen bond was a 3.0 Å donor-acceptor distance and a 20° angle cutoff. The number of hydrogen bonds a ring formed was determined for each frame of the trajectory and averaged over the 250-ns length. Hydrogen bonding primarily occurred with subsites 1 through 3, where Arg75, Gln124, Gln128, Asn50, and Asn81 were the primary residues participating in hydrogen bonding.

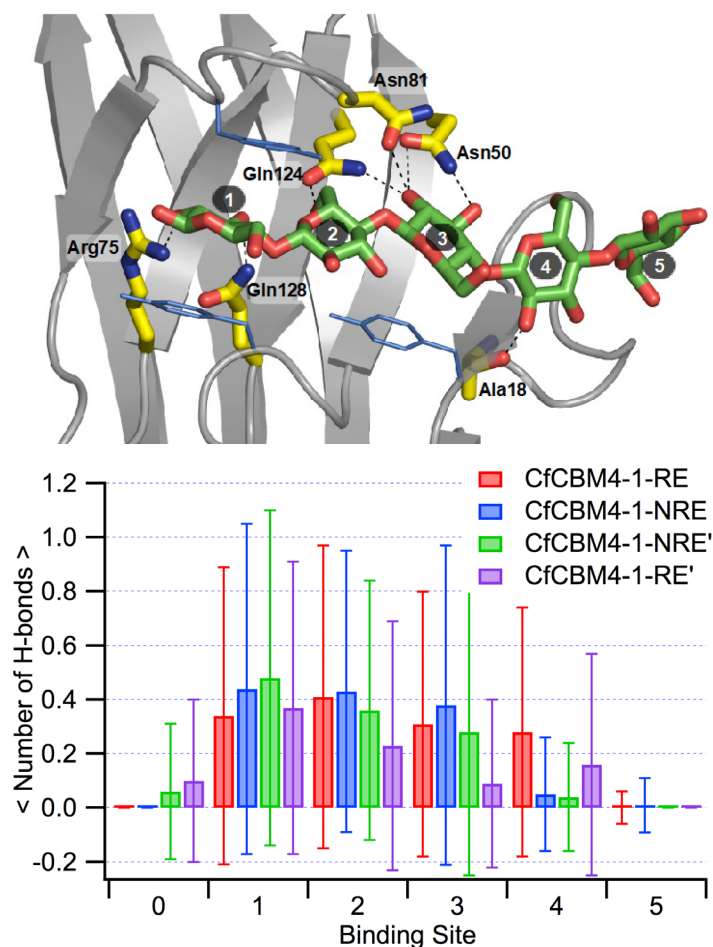


Figure 3.6 Hydrogen bonding partners (dashed lines) at each subsite between side chains of cellopentaose (green and red sticks) and amino acids of *Cf*CBM4-1-RE (yellow, red, and blue sticks). The *Cf*CBM4-1 backbone is shown in gray cartoon. To aid in viewing, the aromatic residues of binding cleft are shown in thin, marine blue lines. Binding subsites are numbered in balloons. Lower panel - Average number of hydrogen bonds (H-bonds) formed between the pyranose ring and the surrounding protein of *Cf*CBM4-1 binding site. Error bars represent one standard deviation.

Detailed analysis of hydrogen bonding over the course of MD simulations defined the primary hydrogen bonding partners in the *Cf*CBM4-1 binding subsites responsible for acceptance of a bi-directionally bound cellopentaose (Figure 3.6). In subsite 1, Arg75 and Gln128 hydrogen bond with the secondary hydroxyl groups of the pyranose ring. Gln124 generally bonds with the primary hydroxyl group of the pyranose ring in subsite 2, while occasionally hydrogen bonding with the secondary hydroxyl group of the subsite 3 pyranose. Asn50 and Asn81 hydrogen bond with the secondary hydroxyl groups of the subsite 3 pyranose. The protein surrounding subsite 4 rarely participated in hydrogen bonding with the pyranose ring, but when a hydrogen bond was formed, Ala18 was the partner. This specificity for primary and secondary hydroxyl groups can only be fulfilled by the orientation of cellopentaose in *Cf*CBM4-1-RE and *Cf*CBM4-1-NRE, resulting from the symmetry of the ligands (Figure 3.2). When cellopentaose occupies the binding site as initialized in *Cf*CBM4-1-RE' and *Cf*CBM4-1-NRE', these hydrogen bonding partners were inaccessible, and thus, the ligand shift by one binding subsite accommodates formation of hydrogen bonds with the protein. The binding subsites of *Cf*CBM4-1 do not appear to have redundant hydrogen bonding partners that would allow binding of the *Cf*CBM4-1-RE' and *Cf*CBM4-1-NRE' ligand conformations.

3.4.4 CfCBM4-1 dynamics

Molecular dynamics simulations further support the feasibility of bi-directional ligand binding. Examination of the five *Cf*CBM4-1 and three *Cf*CBM4-2 molecular dynamics simulations described above reveals remarkably similar dynamic behavior among the *Cf*CBM4-1-RE and *Cf*CBM4-1-NRE conformations and the *Cf*CBM4-2-RE and *Cf*CBM4-2-NRE conformations. Furthermore, the dynamics of the *Cf*CBM4-1-RE'

and *CfCBM4-1-NRE'* conformations corresponded to the *CfCBM4-1-RE* and *CfCBM4-1-NRE* dynamics, respectively, following translocation as described above. To evaluate dynamic similarity, we applied a host of simulation trajectory analyses including: analysis of protein and ligand flexibility measured through the root mean square deviation (RMSD) and root mean square fluctuation (RMSF), non-bonded interaction energy measurements, and degree of ligand solvation.

Evaluation of the RMSD of the protein backbone over the course the 250-ns simulation illustrates the relative stability of the *CfCBM4-1-RE* and *CfCBM4-1-NRE* conformations (Figure 3.7a). The RMSD was calculated for each of the five *CfCBM4* simulations, using the last coordinates of the 1-ns equilibration simulation as the reference coordinates. The RMSD of the protein backbones in the *CfCBM4-1-RE* and *CfCBM4-1-NRE* simulations were extraordinarily well behaved, deviating little over the course of the simulation. This particular result suggests the opposite ligand conformation did not adversely affect the protein structure, and the binding site was capable of accommodating the ligand without a significant structural rearrangement. When the ligand was rotated around the longitudinal axis, as in the cases of *CfCBM4-1-RE'* and *CfCBM4-1-NRE'*, the dynamics of the protein backbone reflected the translocation of the ligand to the equivalent *CfCBM4-1-RE* and *CfCBM4-1-NRE* positions. The RMSD deviated significantly from that of the initial position as the protein rearranged the ligand, and the last glucose binding subsite remained unoccupied. The *CfCBM4-1-RE'* and *CfCBM4-1-NRE'* simulations eventually reached an equilibrium similar to that of the ligand free *CfCBM4-1*.

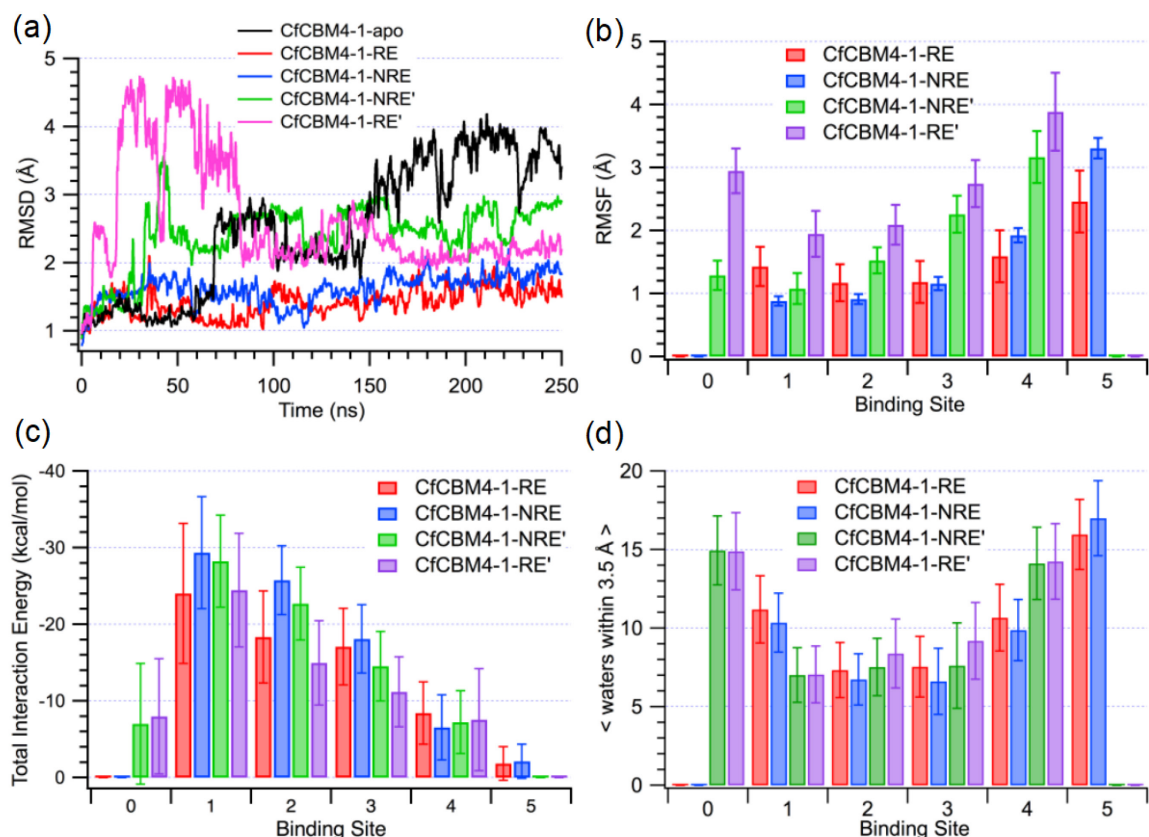


Figure 3.7 *CfCBM4-1* hydrogen bonding behavior and protein-ligand dynamics from 250-ns MD simulations. (a) Root Mean Square Deviation (RMSD) of the *CfCBM4-1* protein backbone over the 250 ns simulation. The RMSD reference structure is the last frame of the 1-ns equilibration simulation, which is why RMSD does not start at 0 Å at 0 ns. (b) Root Mean Square Fluctuation (RMSF) of the cellopentaose ligand on the per-binding-site basis. Error bars were determined from block averaging over 2.5 ns blocks of data. (c) Average total interaction energy (sum of van der Waals and electrostatic contributions) of each pyranose ring with the surrounding protein. Error bars represent one standard deviation. (d) The average number of water molecules within 3.5 Å of each binding subsite of *CfCBM4-1*. Error bars represent one standard deviation.

Similarly, the RMSF of the protein backbones indicate the average protein structure of each ligand bound *Cf*CBM4-1 was generally unaffected by the ligand's conformation (Figure 3.8a). In fact, the absence of ligand appeared to impact the protein more than any of the ligand conformations. Figure 3.8a illustrates that the aromatic residues along the binding cleft, Tyr43 and Tyr85, and key hydrogen bonding residues, Asn50 and Asn81, fluctuated significantly in the absence of a ligand. These fluctuations potentially contributed to the increase in RMSD of the apo structure, and gradually, the backbone of the apo CBM became more flexible. This may be a mechanism by which the CBM makes the binding site more accessible to ligands. In general, the N- and C-terminal domain RMSF values were significantly higher than the core of the protein domain. While high terminal domain fluctuation is an expected behavior in nearly all proteins, we mention this to add the caveat that *Cf*CBM4-1 has been simulated without a bound calcium ion. The 1GU3 structure does not contain a resolved metal ion [62], but Johnson et al. have reported that *Cf*CBM4-1 coordinates calcium binding through residues Thr8, Gly30, and Asp142, where Thr8 and Asp142 comprise the N- and C-terminus, respectively. This lack of coordinating bonds tying together the termini leads to higher relative fluctuation, as can be seen in the RMSD of the *Cf*CBM4-1-apo at 70 ns (Figure 3.7a). However, the calcium ion and coordinating residues are located on the surface of *Cf*CBM4-1, directly opposite the binding cleft, and the lack of a calcium ion has no affect on binding affinity [200]. Thus, we chose to simulate the protein without the calcium ion in accordance with the structure.

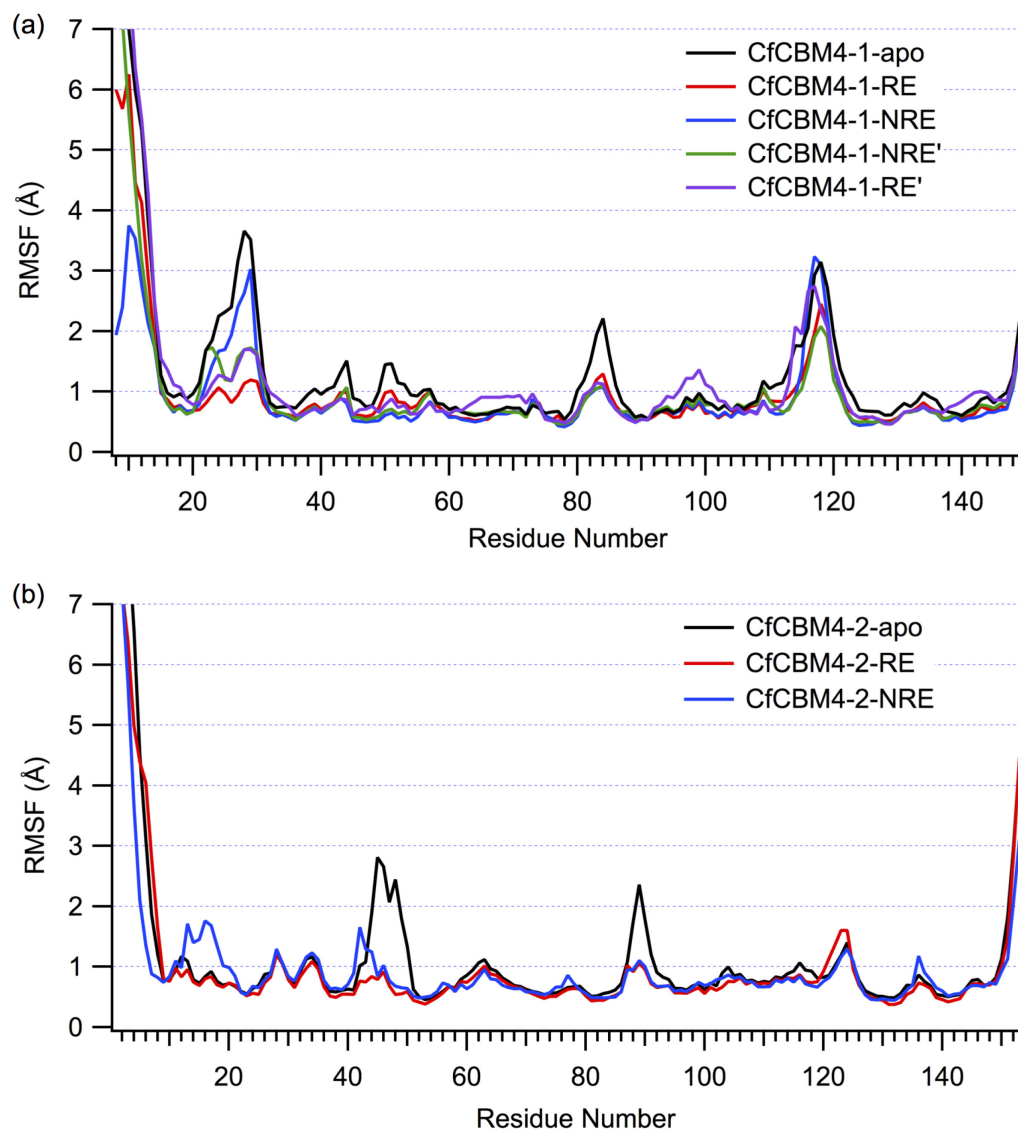


Figure 3.8 Root mean square fluctuation (RMSF) of the protein backbone for (a) five *CfCBM4-1* systems and (b) three *CfCBM4-2* systems over 250 ns.

The flexibility of the *CfCBM4-1-RE* and *CfCBM4-1-NRE* ligands, as measured by RMSF of the pyranose ring atoms, was equivalent within error. The RMSF of the ligand is a determination of the average position of the ring atoms over the course of the entire simulation and is delineated on a per-binding subsite basis (Figure 3.7a). As

previously described, the *Cf*CBM4-1-RE' and *Cf*CBM4-1-NRE' ligands shifted out of register very early in the MD simulation to positions approximating the side chain and ring positioning of the *Cf*CBM4-1-RE and *Cf*CBM4-1-NRE ligands, respectively. The "0" binding subsite represents a solvent exposed pyranose ring, external to the cleft. Otherwise, the RMSF as a function of binding subsite (Figure 3.7a) illustrates equivalent positions along the cleft, where *Cf*CBM4-1-RE has the same ring and side chain orientation as *Cf*CBM4-1-RE'. Along the entirety of the cleft, the *Cf*CBM4-1-RE and *Cf*CBM4-1-NRE pyranose rings fluctuated within error of each other, suggesting the cleft accommodates each ligand with equal favorability. Though equivalent in position in the 1 to 4 binding subsites, the *Cf*CBM4-1-RE' and *Cf*CBM4-1-NRE' ligands fluctuated more than the fully bound ligands. The solvent exposed pyranose rings had a much larger range of motion (Movies 3.1 and 3.2), uninhibited by the protein cleft, and this translated into increased fluctuation along the entirety of the four bound pyranose rings.

The degree of solvation within the binding cleft was unaffected by the ligand conformation (Figure 3.7d). For a given trajectory frame, the number of water molecules within 3.5 Å of the pyranose ring of binding subsite was determined. This value was averaged for each binding subsite over the entire 250-ns trajectory. The average value is a numerical estimate of the degree to which any pyranose ring is exposed to the water solvent. The degree of solvation of any given binding subsite was within one standard deviation of that of any of the various ligand conformations. This is consistent with the notion that *Cf*CBM4-1 is capable of binding the cellopentaose ligand in both the *Cf*CBM4-1-RE and *Cf*CBM4-1-NRE directions.

The total interaction energy of the protein with the pyranose rings of cellopentaose reveals aromatic stacking interactions were maintained with both faces of the pyranose rings along the cleft. Electrostatic and van der Waals components of the non-bonded interactions were calculated over the 250-ns MD simulations. The same non-bonded interaction cutoffs used in producing the simulations were applied in the data analysis. For computational efficiency, the interaction energy analysis was conducted using a culled dataset, 2,500 equally-spaced frames rather than the 25,000 frames collected. The total interaction energy, the sum of the two components, was highest in binding subsites 1 through 3, reflecting the availability of hydrogen bonding partners and aromatic stacking interactions relative to subsites 4 and 5. As with other dynamic analyses, the total interaction energy was generally unaffected by the direction of the bound ligand (Figure 3.7c). The residues along the binding cleft maintained a similar degree of contact with the pyranose rings and were equally capable of maintaining stacking interactions with either face of the pyranose ring. From perturbations of ^1H chemical shifts upon cellotetraose binding, Johnson et al. reported that, despite the multitude of aromatic residues lining the binding cleft, only Tyr19 and Tyr85 were directly involved in aromatic stacking interactions [201]. From this determination of interaction energies, we endorse addition of Tyr43 to the list of stacking aromatic residues [62], as the interaction energy of Tyr43 with the ligand is of similar order as Tyr85.

3.4.5 CfCBM4-2 dynamics

Johnson et al. also made the case that CfCBM4-2 was capable of binding TEMPO-labeled cello-oligomers with the label at either end of the cleft [61]. To date, a ligand-bound structure of this homologous CBM4 structure has not been reported, though the similarity in fold enabled docking of the 1GU3 cellopentaose and the modeled CfCBM4-1-NRE cellopentaose to the CfCBM4-2 structure. MD simulations of the two conformations were performed to elucidate the molecular interactions governing ligand binding and the possibility of bi-directional binding.

CfCBM4-2 differs structurally from CfCBM4-1. The cleft of CfCBM4-2 appears to be wider and several ligand binding residues are substituted (e.g., CfCBM4-1 residues Gly87, Trp49, His132, and Ser23 appear as Asn81, Tyr43, Gln128 and Val17, respectively, in CfCBM4-2; Figure 3.9a) [110]. However, MD simulation suggests the apparent widened cleft of CfCBM4-2 may be an artifact of the structural study conditions. When CfCBM4-2 was docked with cellopentaose in the binding groove, the cleft width reduced, approximately matching that of CfCBM4-1 (Figure 3.10). The reduction in cleft width occurred quickly, during equilibration, and the protein remained in close contact with the ligand over the remainder of the simulation. The NMR structure captured CfCBM4-2 in its ligand-free state [110], and thus, the absence of ligand interactions is the likely reason behind the larger binding groove width relative to that of CfCBM4-1. The RMSD of the CfCBM4-2 protein backbone reflects the protein rearrangement that occurs when the binding cleft closes around the bound ligand, eventually equilibrating around 3.5 Å (Figure 3.9b). The ligand-free CfCBM4-2 exhibited a great deal more flexibility in the loops surrounding the cleft (Figure 3.8b), suggesting

flexibility in the cleft as an acquisition mechanism. The *N*-terminus of the *Cf*CBM4-2-RE structure underwent a conformational change around 220 ns, as indicated by the change in RMSD, but this does not affect ligand binding.

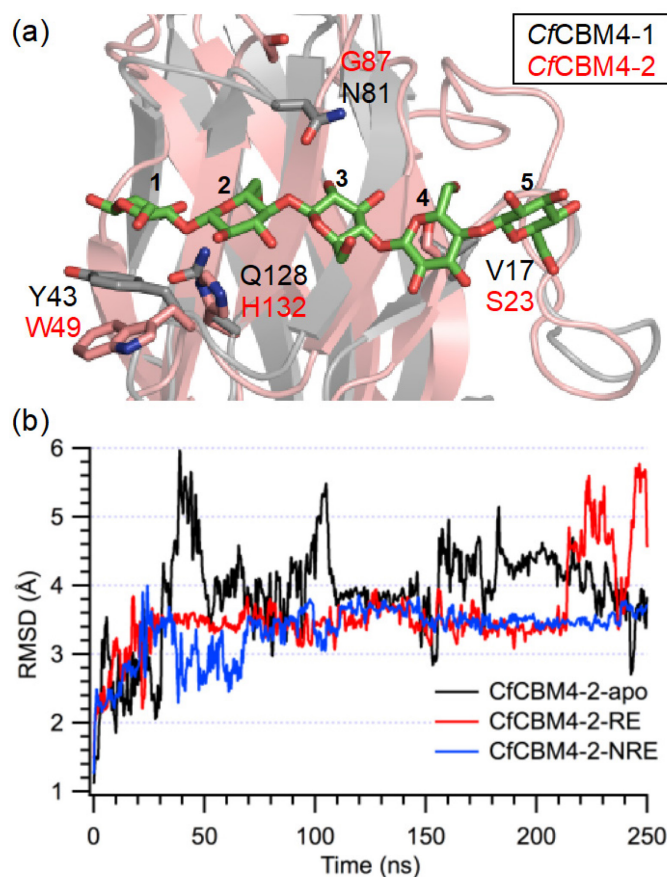


Figure 3.9 (a) Comparison of binding site of *Cf*CBM4-1 (gray) and *Cf*CBM4-2 (salmon) illustrating substitutions of residues involved in cello-oligomer binding. *Cf*CBM4-1 residues are labeled in black letters, and residues in the same position in *Cf*CBM4-2 are labeled in red. The binding subsites are numbered. (b) RMSD of the *Cf*CBM4-2 protein backbone over the 250-ns simulation. The RMSD reference structure is the last frame of the 1-ns equilibration simulation.

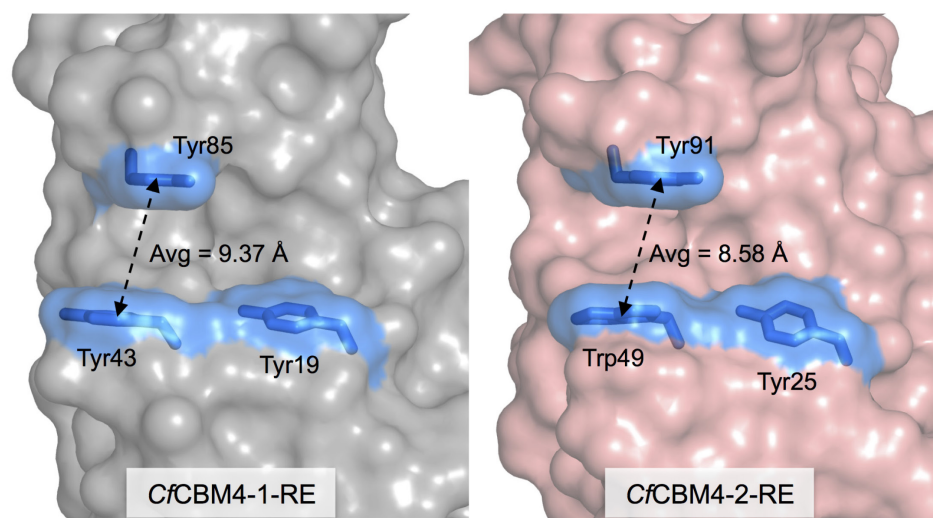


Figure 3.10 Comparison of *CfCBM4-1* (gray surface) and *CfCBM4-2* (salmon surface) binding groove width. The average distance between center of mass of side chains of two aromatic residues (blue sticks) forming sandwich platform in both CBMs was measured over 250 ns (25,000 frames). For *CfCBM4-1*, this average distance was 9.37 Å and for *CfCBM4-2* this average distance was 8.58 Å. The < 1 Å difference between these averages is insignificant in context of groove width. The *CfCBM4-2* binding groove width, measured from Tyr91 to Trp49, was 15.27Å in the apo structure, from which the simulation was initialized.

As with *CfCBM4-1*, dynamic measurement associated with ligand binding and hydrogen bond formation suggest *CfCBM4-2* is capable of bi-directional binding. Again, we have compared the average number of hydrogen bonds formed between the protein and this pyranose rings of a given binding subsite, the RMSF of ligand along the cleft, and the total interaction energy of the ligand with the protein on a per-binding-site basis as function of ligand direction in the *CfCBM4-2* cleft. All of these measures were expected to be the same for the two conformations, indicating both ligands are equally

stable in the *Cj*CBM4-2 cleft. Despite the significant sequence variation in the two clefts, the number of hydrogen bonds formed between a given pyranose ring and the *Cj*CBM4-2 binding site did not vary significantly upon reorientation of the ligand (Figure 3.11a). In general, each binding subsite formed one intermittent hydrogen bond with the protein over the course of the simulation. This is similar to the behavior of *Cj*CBM4-1 (Figure 3.6), implying the substituted residues play equivalent roles in ligand binding. The RMSF of the ligand was approximately the same in each binding subsite irrespective of where the reducing end resided (Figure 3.11b). The *Cj*CBM4-2-NRE pyranose in binding subsite 5 was unable to maintain stable interactions with any surrounding protein residues, accounting for the slight deviation from the *Cj*CBM4-2-RE ligand behavior. However, the total interaction energy of the pyranose ring in a given *Cj*CBM4-2 binding subsite was the same regardless of ligand direction (Figure 3.11c). These dynamic measures support the hypothesis that the *Cj*CBM4-2 can bind cello-oligomers in at least two different conformations. As we will describe, we further posit the ability to bi-directionally bind carbohydrate oligomers may be common to the β -sandwich protein fold.

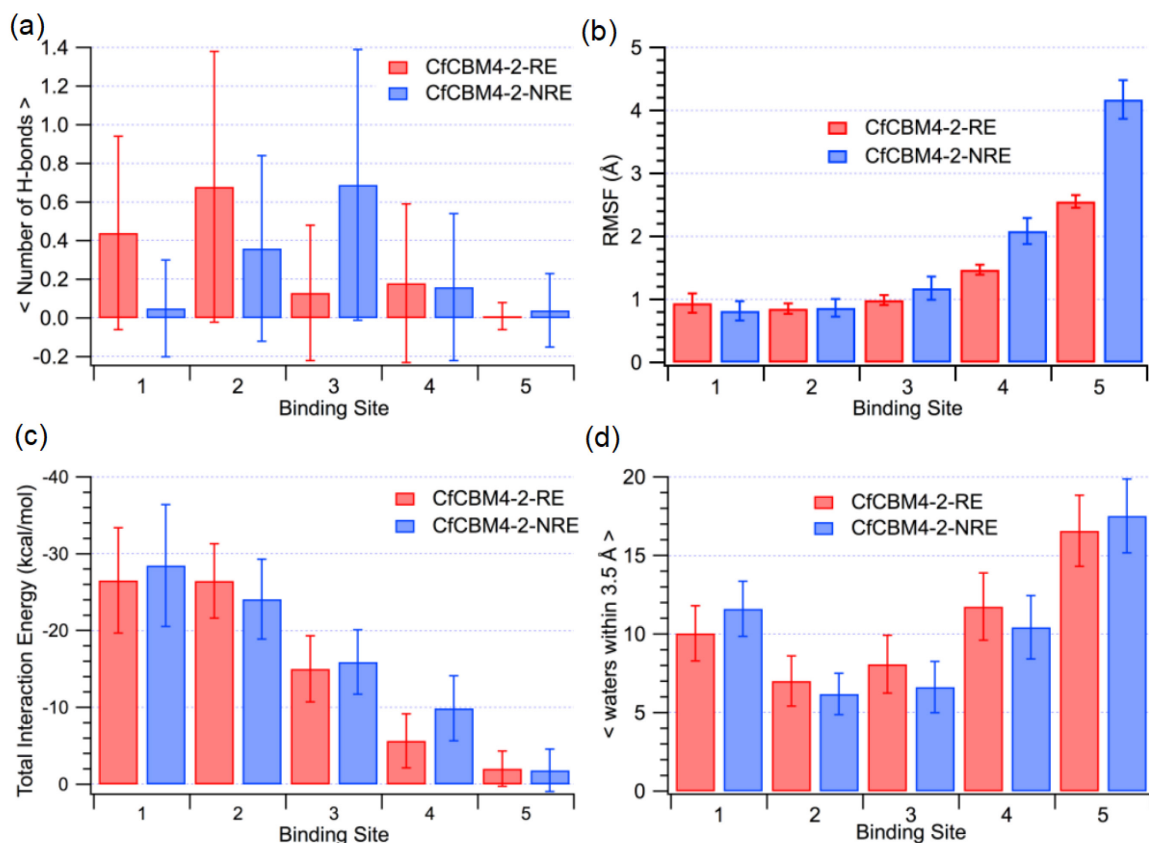


Figure 3.11 *Cj*CBM4-2 ligand dynamic measurements. (a) Average number of hydrogen bonds formed between each of the five pyranose rings and side chains in each binding subsite with the surrounding protein (b) RMSF of ligand on a per binding subsite basis for *Cj*CBM4-2 systems. Error bars were calculated using block averaging over 2.5 ns (c) Average total interaction energy of each pyranose with the surrounding protein. Error bars represent one standard deviation. (d) The average number of water molecules within 3.5 Å of each binding subsite of *Cj*CBM4-2. Error bars represent one standard deviation.

3.4.6 Evidence of bi-directional binding beyond *C. fimi* CBM4s

Bi-directional cello-oligomer binding is likely a phenomenon common to CBM4s and the broader class of β -sandwich CBMs. While structural resolution of cello-oligomers in two different orientations of the same CBM4 binding cleft does not

currently exist, our computational results combined with the NMR studies from Johnson et al. strongly suggest both *Cf*CBM4-1 and *Cf*CBM4-2 demonstrate bi-directional binding capabilities, despite significant differences in sequence similarity. As further evidence of bi-directional binding in CBM4s, a computational docking study of a *Clostridium thermocellum* CBM4, part of the cellulosomal cellobiohydrolase A construct, found this particular CBM4 was likely to bind a cellohexaose in a direction opposite that of the cellobiose bound in the reported crystal structure (PDB ID 3K4Z) [202].

*Cf*CBM4-1 adopts a characteristic β -jelly roll fold, which belongs to the larger family of β -sandwich structures [16]. As CBMs are classified in the CAZy database (Carbohydrate Active Enzyme Database; <http://www.cazy.org>) according to sequence and structural similarity, all CBM4s belong to the β -sandwich protein fold [74]. Further, the β -sandwich fold is common among other CBM families and is noted for its broad specificity [16]. At the writing of this manuscript, 29 of the 69 CBM families documented in CAZy exhibit a form of the β -sandwich fold, with a remarkable relative diversity of sequence. Accordingly, we hypothesized that bi-directional binding has been previously observed in these structurally-related CBMs, but that it had perhaps not been recognized as such. In such a comparison, one must be cognizant that β -sandwiches can have two binding sites, one on the face of the β -sheets and the other on the edge of the β -sheets [16]. Of the 29 β -sandwich CBM families, 10 families had deposited structures with a glycan bound at the same binding site as that of *Cf*CBM4-1-RE (i.e., on the face of the β -sheets). A total of 34 glycan-bound CBM structures, representing 10 of the 29 β -sandwich CBM families, were available for examination (Table 3.2). Using the Dali Web Server (http://ekhidna.biocenter.helsinki.fi/dali_lite/start)[152] to structurally align the 34

structures with *Cf*CBM4-1 (PDB code 1GU3), we examined the conformation of the ligands within the CBM binding clefts.

Table 3.2 CBM structures with β -sandwich fold compared with *Cf*CBM4-1-RE (PDB ID 1GU3). ‘Same’ refers to the direction of ligand that is equivalent to that in 1GU3.

CBM Family	PDB ID	Protein Name	Ligand Direction
4	1GU3	endo- β -1,4-glucanase C (CenC) (Cel9B)	Same
4	2Y64	xylanase (XynI;Xyn1;RmXyn10A)(Xyn10A)	Same
4	2Y6G	xylanase (XynI;Xyn1;RmXyn10A)(Xyn10A)	Same
4	2Y6K	xylanase (XynI;Xyn1;RmXyn10A)(Xyn10A)	Same
4	2Y6L	xylanase (XynI;Xyn1;RmXyn10A)(Xyn10A)	Same
4	3K4Z	cellobiohydrolase (CbhA;Cthe_0413) (Cbh9A)	Same
4	1GUI	laminarinase (Lam;TmLam;TM0024) (Lam16)	Same
6	1UY0	endo- β -1,4-glucanase B (CELB) (Cel5A; Cel5B)	OPPOSITE
6	1UYX	endo- β -1,4-glucanase B (CELB) (Cel5A; Cel5B)	OPPOSITE
6	1UYY	endo- β -1,4-glucanase B (CELB) (Cel5A; Cel5B)	OPPOSITE
6	1UZ0	endo- β -1,4-glucanase B (CELB) (Cel5A; Cel5B)	OPPOSITE
6	2CDO	endo- β -agarase I / B (AgaB; Sde_1175) (Aga16B)	Same
6	2CDP	endo- β -agarase I / B (AgaB; Sde_1175) (Aga16B)	Same
9	1I82	xylanase A (XynA;Tm0061) (Xyl10A)	OPPOSITE
15	1GNY	xylanase F / 10C (Xyl10C;CJA_3066) (Xyn10C)	OPPOSITE
15	1US2	xylanase F / 10C (Xyl10C;CJA_3066) (Xyn10C)	OPPOSITE
16	2ZEX	β -mannanase A (ManA; CelA)(Man5A)	Same
16	2ZEY	β -mannanase A (ManA; CelA)(Man5A)	Same

16	3OEA	β -mannanase A (ManA; CelA)(Man5A)	Same
16	3OEB	β -mannanase A (ManA; CelA)(Man5A)	Same
17	1J84	endo- β -1,4-glucanase 5A (EngF) (Cel5A)	Same
27	1PMH	β -mannanase (ManA;CsMan26)	Same
27	1OF4	β -mannanase (ManB;TM1227) (Man5)	Same
27	1OH4	β -mannanase (ManB;TM1227) (Man5)	Same
28	3ACG	endo- β -1,4-glucanase 5A (CelA) (Cel5A)	OPPOSITE
28	3ACH	endo- β -1,4-glucanase 5A (CelA) (Cel5A)	OPPOSITE
28	3ACI	endo- β -1,4-glucanase 5A (CelA) (Cel5A)	OPPOSITE
29	1GWL	non-catalytic protein 1 (Ncp1)	Same
29	1GWM	non-catalytic protein 1 (Ncp1)	Same
29	1OH3	non-catalytic protein 1 (Ncp1)	Same
29	1W8T	non-catalytic protein 1 (Ncp1)	Same
29	1W8U	non-catalytic protein 1 (Ncp1)	Same
39	3AQX	β -1,3-glucan recognition protein (betaGRP;Gnbp3)	OPPOSITE
39	3AQZ	β -1,3-glucan recognition protein (BGBP;Gnbp3)	OPPOSITE

Visualization of the glycan-bound β -sandwich fold CBM structures reveals apparent promiscuity in binding. The examined CBMs bind not only C6 sugars but also C5 sugars; the sugars were often bonded through a variety of glycosidic linkages as well. Further, multi-directional binding along the binding cleft appeared often across the observed β -sandwich CBM structures. Of the 34 structures examined, 22 displayed a ligand in the same bound conformation as the 1GU3 structure (i.e., *Cf*CBM4-1-RE); 12 ligands appeared in the opposite conformation corresponding to the modeled *Cf*CBM4-1-NRE conformation. As an example of the latter, we illustrate a family 15 CBM derived

from *Pseudomonas cellulosa* xylanase Xyn10C (PDB code 1GNY)[203] aligned with CfCBM4-1 (PDB code 1GU3) in Figure 3.12. Family 6 CBMs exhibit bi-directional ligand binding within the same family. Binding of glycans in β -sandwich CBMs makes use of standard aromatic stacking interactions common among carbohydrate binding proteins [204, 205], but we anticipate bi-directional binding is a consequence of the evolutionary diversity of the protein fold [206], resulting in conveniently-spaced hydrogen bonding partners along the cleft. While 34 structures is too small a sample to draw conclusions relative to frequency of conformational occupancy, this evaluation indicates bi-directional binding occurs more frequently than acknowledged and offers new possibilities in the development of cellulosic biotechnology.

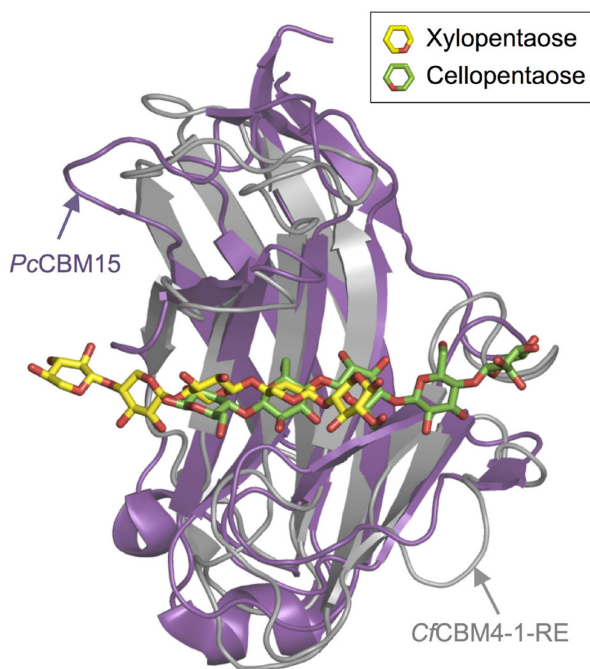


Figure 3.12 Family 15 CBM derived from *Pseudomonas cellulosa* xylanase Xyn10C, PcCBM15 (purple cartoon), bound to xylopentaose (yellow and red sticks) aligned with CfCBM4-1-RE (gray cartoon) bound to cellopentaose (green and red sticks).

3.4.7 Bi-directional binding extends to family 17 and 28 CBMs

We have found that family 4 CBMs showed no thermodynamic preference towards a given longitudinal orientation of cello-oligomers (i.e., the oligomers can bind ‘bi-directionally’ with the reducing end of the chain at either end of the cleft); moreover, structural comparison of all 29 available (as of June 2015) ligand-bound CBM structures exhibiting a β -sandwich fold revealed ligand binding in opposite directions in many other β -sandwich CBM families [184]. We hypothesize bi-directional binding may be feature Type B CBMs developed as an evolutionary advantage, given that bi-directional binding could increase the probability of binding events up to 2-fold. Within the scope of this dissertation, we investigate this bi-directional binding phenomenon in family 17 and 28 CBMs. According to Table 3.2, we already know that crystal structures of family 17 CBMs show same orientation of ligands while crystal structures of family 28 CBMs exhibit the ligand in opposite orientation (Figure 3.13). Although there are architectural differences in binding sites of family 4 (Sandwich platform) and, family 17 and 28 CBMs (Twisted platform; discussed in next chapter), the approximate symmetry of cello-oligomeric ligands and redundancy of available hydrogen bonding partners in the cleft are the determining factors in bi-directionality, which is transferable over the architectures. To consider bi-directional binding within the twisted platform CBMs, we investigated the binding dynamics of four CBMs from families 17 and 28 (*CcCBM17*, *BspCBM17*, *BspCBM28*, and *CjCBM28*) with the cellopentaose ligand bound in both possible orientations in the binding grooves. The homology modeling was used to build *BspCBM17* as its crystal structure is unavailable. Apo crystal structure was available to

build *Bsp*CBM28. Details of these methods and docking of cellopentaose in opposite direction for all CBMs has been discussed in the methods section of next chapter.

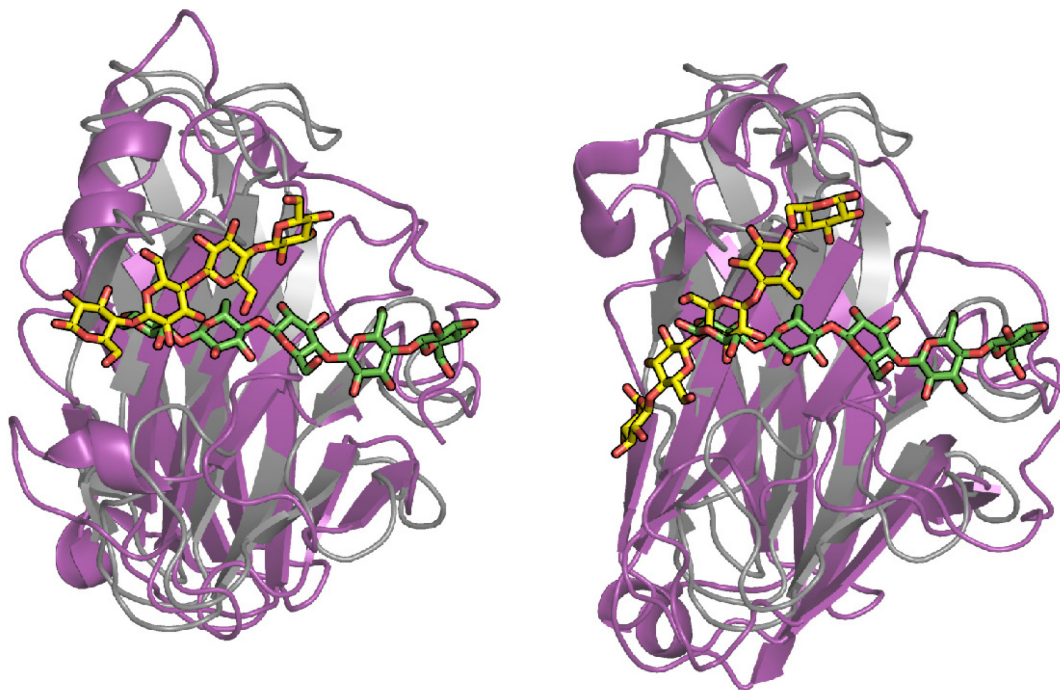


Figure 3.13 Structural alignment of *Cc*CBM17-RE (left; PDB 1J84) and *Cj*CBM28-NRE (right; PDB 3ACI) with *Cj*CBM4-1-RE (PDB 1GU3). Both *Cc*CBM17-RE and *Cj*CBM28-NRE are shown in purple cartoon with cello-oligomer in yellow sticks. The common structure of *Cj*CBM4-1-RE is shown in gray cartoon with cello-oligomer in green sticks.

In all eight simulation cases, the bound cellopentaose ligand maintained continuous interaction with the CBM binding surface over the entire 250-ns simulations, indicating that the binding sites of these family 17 and 28 CBMs can generally accommodate cello-oligomers bi-directionally. The RMSF of ligand in the binding site provides a quantitative measure of stability of the interactions (Fig. 6), and while all four

CBMs can accommodate the ligand bi-directionally, not all of them exhibit fully stabilized protein-ligand interactions. *Cc*CBM17-RE, *Bsp*CBM28-NRE and *Cj*CBM28-NRE bind the cello-oligomer with relatively little fluctuation about the average (~ 1 Å). In the remaining five cases, though the cellopentaose ligands maintain contact with the CBM binding grooves, we observed sliding of the cellopentaose ligand along the binding site, which is reflected in the increased RMSF. We previously observed cellopentaose sliding within the CBM4 binding sites, however, the oligomers moved only a single subsite in either direction to rearrange the primary hydroxyl groups within the groove, as a result of the purposeful perturbation of ligand orientation (Section 3.4.1). The sliding observed in *Bsp*CBM17-RE, by two subsites or a cellobiose unit, maintains the primary and secondary hydroxyl group positions within a given subsite, which is suggestive of a functional mechanism rather than merely alleviation of steric hindrance. A cluster of snapshots (every 2.5 ns) from each simulation has been provided in Figure 3.15 illustrating this phenomenon. In case of *Bsp*CBM17-NRE, between 85 ns to 100 ns, the cellopentaose is slides by one subsite, but an accompanying flip around the longitudinal axis maintains the original hydroxyl group orientation within the groove. Again, these results suggest the family 17 and 28 CBMs feature extended binding sites capable of binding cellohexasaccharide or longer oligomers. These simulations provide sufficient evidence to support the hypothesis that cellulose specific CBMs from all three families hold the characteristics of a binding site that favors bi-directional binding to cello-oligosaccharides irrespective of its overall architecture. On the other hand, this difference in binding site architecture, being evolved within the same type of CBMs, does play a role in recognition of substrate as discussed in the next chapter.

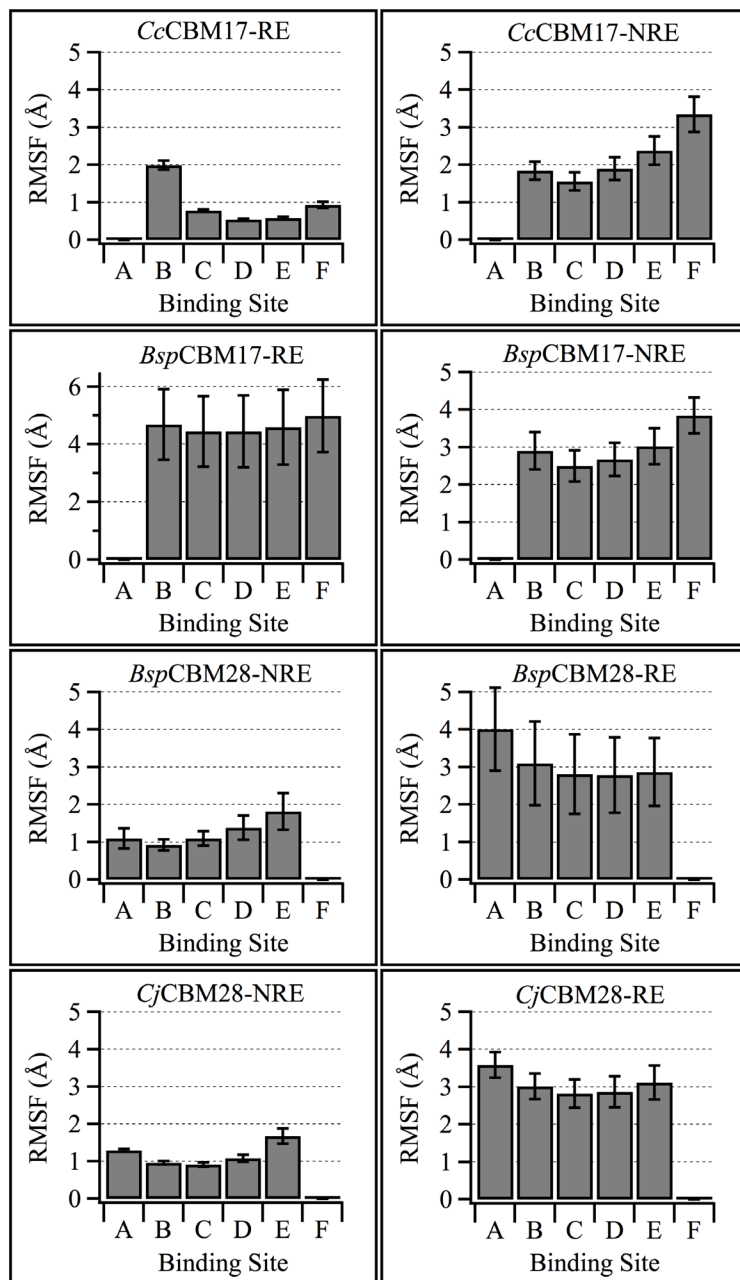


Figure 3.14 Root mean square fluctuation (RMSF) of cellopentaose ligand from its average position over 250 ns trajectory calculated per binding subsite for all eight systems. The Error bars were calculated as a standard deviation of block averages of RMSFs with block size 2.5 ns each.

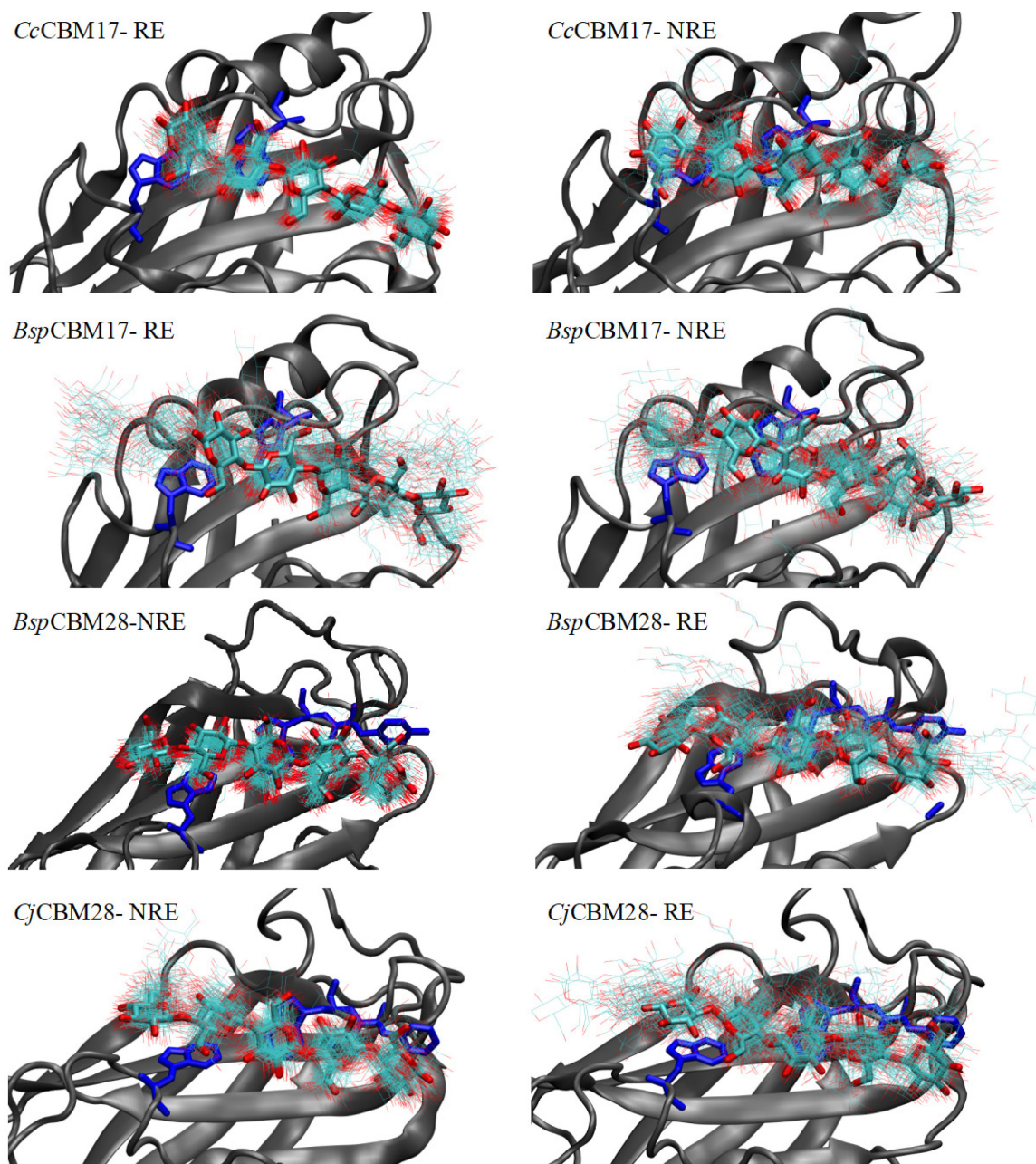


Figure 3.15 Snapshots of cellopentaose (lines) at every 2.5 ns in the binding site of each CBM (gray cartoon) over the 250-ns simulations. The position of cellopentaose at 0 ns is shown in thick cyan stick representation. The aromatic residues along the binding site are shown in dark blue stick representation.

This section of MD simulations indicates that the modeling of protein-carbohydrate complexes that involve homology modeled protein structures or has ligands docked based on structural alignment could use longer simulation times, most likely in microseconds, to get well-stabilized non-bonded protein-ligand interactions. We also confirmed that extensive stepwise minimization and equilibration are highly essential in the simulation setup, as without it we observed dissociation of ligands from these open cleft binding sites at initial stages.

3.5 Conclusions

MD simulations and free energy calculations have enabled us to investigate the molecular-level contributions to cellopentaose binding in protein-carbohydrate systems that have eluded structural resolution techniques. Our results support the original Johnson et al. hypothesis that *C. fimi* CBM4s are capable of binding cello-oligomers with the reducing end of the pyranose at either end of the binding cleft. Free energy calculations are remarkably comparable to experimental ITC measurement and go beyond experiment in enabling delineation between conformational populations. MD simulations reveal abundant hydrogen bonding partners, in near 1:1 parity, exist along the binding cleft, so that regardless of direction, the pyranose ring primary and secondary alcohol are capable of maintaining a hydrogen bond with relevant partners from the interior of the cleft. MD simulations of *Cf*CBM4-2 extend these observations to loosely related (36% sequence similarity) familial representatives. Observation of the dynamic markers indicative of a stably-bound ligand again suggest that *Cf*CBM4-2 is capable of binding cellopentaose in a bi-directional fashion. This observation appears to be not limited to CBM4s, but rather, many carbohydrate-binding proteins bearing the β -sandwich fold, which currently include

29 additional CBM families, may bind pyranose rings irrespective of direction. Out of those 29 CBM families, we further confirm the bi-directional binding phenomenon for family 17 and 28 CBMs that have been categorized as Type B CBMs along with family 4 CBMs but differ in shape of binding site.

Chapter 4 – Role of binding site architecture and recognition of non-crystalline cellulose in Type B Carbohydrate Binding Modules

Chapter 4 reports the characteristics of two different binding site platforms in Type B CBMs and non-crystalline cellulose binding in family 17 and 28 CBMs. Copyright © Abhishek A. Kognole 2017.

4.1 Introduction

CBMs are structurally diverse proteins, binding with many different types of carbohydrate polymorphs and morphologies. To capture this diversity, CBMs have been divided into both families and types based on protein sequence and functional similarity, respectively [16, 75]. Currently, this nomenclature defines function as the ability to target particular substrate crystallinities, as CBMs appear to bind either crystalline or non-crystalline/amorphous and oligomeric substrates. Type A CBMs are specific for crystalline substrates and exhibit a complementary planar binding site lined with aromatic residues [207, 208]. Type B and C CBMs are only subtly different from each other, with both types binding oligosaccharides and non-crystalline/amorphous substrates in clefts or grooves. Type C CBMs have been also shown to bind crystalline cellulose [59]. However, Type B CBMs are capable of binding at any point along the length of the substrate, and Type C CBMs are limited to the end of the oligomer. The underlying protein features enabling this distinction are difficult to define.

The most common protein fold among Type B CBMs is the β -sandwich fold, the proteins of which uniquely recognize not only different kinds of carbohydrates but also varying degrees of polymerization, from the smallest of oligosaccharides to amorphous

substrates [16]. This suggests the β -sandwich fold is a versatile architecture that allows relatively minor variations in sequence and, accordingly, chemical properties of the binding cleft/groove to determine carbohydrate binding specificity. Moreover, despite similar substrate specificities and, in some cases, similar measured affinities, some Type B CBMs appear to uncompetitively discriminate between binding sites on variable crystallinity surfaces [73, 208]. Attempts to experimentally characterize non-crystalline/amorphous cellulose have revealed few details of specific structural properties, only that it is cellulose with a decreasing degree of polymerization and crystallinity index [92]. Non-crystalline/amorphous cellulose derived from pretreatment of native crystalline cellulose could be composed of anything from variable-length polysaccharide chains to only partially decrystallized substrate. Thus, the ability to recognize both soluble oligomers and non-crystalline/amorphous cellulose is a key aspect of Type B CBM functionality.

Cellulose-specific Type B CBMs, including those from families 4, 17, and 28, each with the β -sandwich fold (Figure 4.1), have been shown to bind both soluble cello-oligomers and non-crystalline cellulose [51-54, 56, 96-98]. Additionally, adsorption isotherms suggest families 17 and 28 individually recognize ‘high’ and ‘low’ affinity binding sites on representative non-crystalline cellulose substrates [64]. These studies also reveal that family 17 and 28 Type B CBMs exhibit higher affinities towards non-crystalline cellulose than toward oligomeric substrates [64, 100]. Oligomeric substrates of *Cellulomonas fimi* CBM4-1 and CBM4-2 (*Cf*CBM4-1 and *Cf*CBM4-2, respectively) also appear to bind cello-oligomers bi-directionally, with the reducing end of the pyranose ring at either end of the cleft; there is positive, but limited, evidence of this phenomenon

being common among β -sandwich CBMs [61, 184]. Collectively, the data imply that these Type B CBMs are discriminating between the various available binding sites on the non-crystalline carbohydrate surface, but there is not necessarily a directional preference within the binding site.

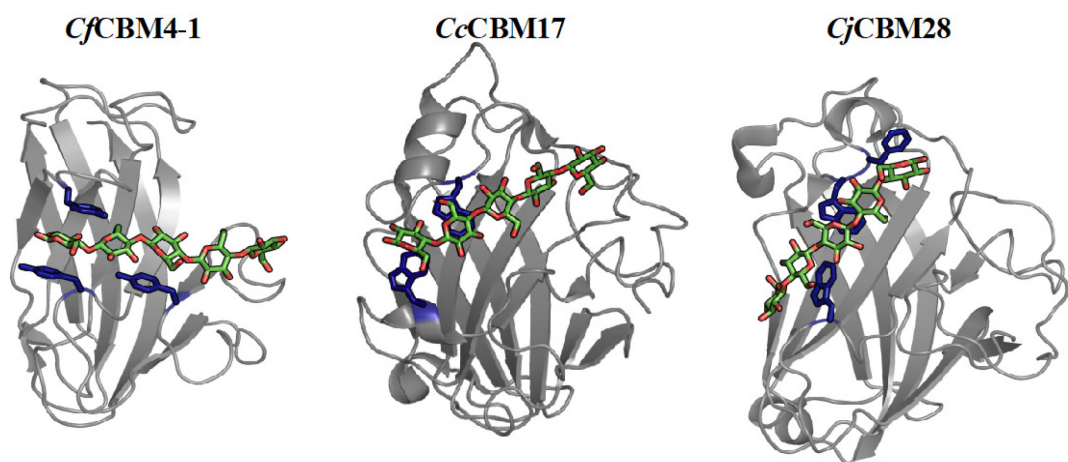


Figure 4.1 CBMs (cartoon) from families 4, 17, and 28 with bound cello-oligomers (medium gray sticks). Binding site aromatic residues are shown in a dark gray stick representation. The structures, *Cellulomonas fimi* CBM4-1, *Clostridium cellulovorans* CBM17, and *Clostridium josui* CBM28, were obtained from crystal structures with PDB IDs 1GU3, 1J84, and 3ACI, respectively. After structural alignment of the β -sandwich proteins, the family 4 and 17 CBM cello-oligomer is bound in same direction, with the reducing end toward the left of the figure, whereas the family 28 CBM's cello-oligomer is oriented in the opposite direction.

To gain a molecular-level understanding of how these three families of Type B CBMs discriminate between binding soluble oligomeric and non-crystalline/amorphous substrates, we implemented a computational approach to describe the differences in binding behavior and affinities within and among the CBM families. From molecular dynamics (MD) simulations, we explore the role of binding site architecture. Free energy perturbation with Hamiltonian replica exchange MD (FEP/ λ -REMD) and umbrella sampling MD was used to examine bidirectional ligand binding ability and apparent binding modes in non-crystalline substrate recognition. At each step of our study, we compare the computational results with available experimental data to assess the validity of our observations and to translate observations to practice.

4.2 Methods and materials

4.2.1 Modeling protein-carbohydrate complexes

Two representative CBMs from each of the three CBM families, 4,17, and 28, were selected to gain an understanding of the variations in protein-carbohydrate binding within and across the families. The selected representatives were *Cellulomonas fimi* CBM4-1 and CBM4-2 (CfCBM4-1 and CfCBM4-2), *Clostridium cellulovorans* CBM17 (CcCBM17), *Bacillus sp. 1139* CBM17 and CBM28 (*Bsp*CBM17 and *Bsp*CBM28), and *Clostridium josui* CBM28 (*Cj*CBM28). The representatives were selected on the basis that they are characterized as cellulose-specific Type B CBMs and are shown to have affinity for both oligomeric and non-crystalline cellulose. CBMs with available structural data were preferred to be able to successfully apply computational techniques.

The protein-carbohydrate systems were modeled in the following configurations (Figure 4.2): (A) CBMs with the cello-oligosaccharide bound in the orientation observed

in the crystallographic structure, (B) CBMs with the oligosaccharide bound in the opposite direction of the structural orientation (i.e., with the reducing end of the sugar longitudinally rotated to the opposite end of the groove), and (C) CBMs bound with a partially decrystallized cellulose microfibril, approximating non-crystalline cellulose, in both the structural and reverse orientations. Additionally, each of the CBM representatives was modeled without a bound ligand for comparison, totaling 16 unique molecular models (Table 4.1).

Table 4.1 List of all the MD simulations performed in this study with length of MD simulations and free energy calculation method.

Group	CBM	Substrate	System	Simulation Time	Free Energy Calculation
Apo	<i>CcCBM17</i>	-	<i>CcCBM17</i>	250 ns	-
	<i>BspCBM17</i>	-	<i>BspCBM17</i>	250 ns	-
	<i>BspCBM28</i>	-	<i>BspCBM28</i>	250 ns	-
	<i>CjCBM28</i>	-	<i>CjCBM28</i>	250 ns	-
A	<i>CcCBM17</i>	Cellopentaose	<i>CcCBM17</i> -RE	250 ns	FEP/ λ -REMD
	<i>BspCBM17</i>	Cellopentaose	<i>BspCBM17</i> -RE	250 ns	FEP/ λ -REMD
	<i>BspCBM28</i>	Cellopentaose	<i>BspCBM28</i> -NRE	250 ns	FEP/ λ -REMD
	<i>CjCBM28</i>	Cellopentaose	<i>CjCBM28</i> -NRE	250 ns	FEP/ λ -REMD
B	<i>CcCBM17</i>	Cellopentaose	<i>CcCBM17</i> -NRE	250 ns	-
	<i>BspCBM17</i>	Cellopentaose	<i>BspCBM17</i> -NRE	250 ns	-
	<i>BspCBM28</i>	Cellopentaose	<i>BspCBM28</i> -RE	250 ns	-

	<i>Cj</i> CBM28	Cellopentaose	<i>Cj</i> CBM28-RE	250 ns	-
C	<i>Cc</i> CBM17	Cellulose microfibril	<i>Cc</i> CBM17-F	100 ns + 100 ns	Umbrella Sampling
	<i>Cc</i> CBM17	Cellulose microfibril	<i>Cc</i> CBM17-R	100 ns + 100 ns	Umbrella Sampling
	<i>Bsp</i> CBM17	Cellulose microfibril	<i>Bsp</i> CBM17-F	100 ns + 100 ns	Umbrella Sampling
	<i>Bsp</i> CBM17	Cellulose microfibril	<i>Bsp</i> CBM17-R	100 ns + 100 ns	Umbrella Sampling

Explicitly solvated models of each CBM were developed from Protein Data Bank (PDB) structures or via homology modeling. *Cj*CBM4-1 and *Cj*CBM4-2 models, in the apo and cellopentaose-bound states, were previously constructed [184]. *Cc*CBM17 was constructed from the 1J84 PDB structure, which features cellotetraose bound in the groove [100]. Similarly, *Bsp*CBM28 was constructed from the 1UWW PDB structure, having no bound ligand [106], and *Cj*CBM28 was constructed from the 3ACI PDB structure, featuring cellopentaose [101]. With no available crystal structure for *Bsp*CBM17, we used homology modeling, with *Cc*CBM17 as a template, to build the protein model [148, 150]; the two proteins are quite similar, having 55% sequence similarity and 70% structural similarity. For comparative purposes, we modeled the CBM-bound cello-oligomers as cellopentaose; an additional beta-D-glucose residue was constructed near the end of the *Cc*CBM17 groove, and the cellopentaose ligand was docked with *Bsp*CBM17 and *Bsp*CBM28 by structural alignment with their homologous family member using Dali pairwise alignment tool [152]. These four systems represent the oligomer-bound CBMs exhibiting the structural orientation, *Cc*CBM17-RE, *Bsp*CBM17-RE, *Bsp*CBM28-NRE, and *Cj*CBM28-NRE (Figure 4.2A).

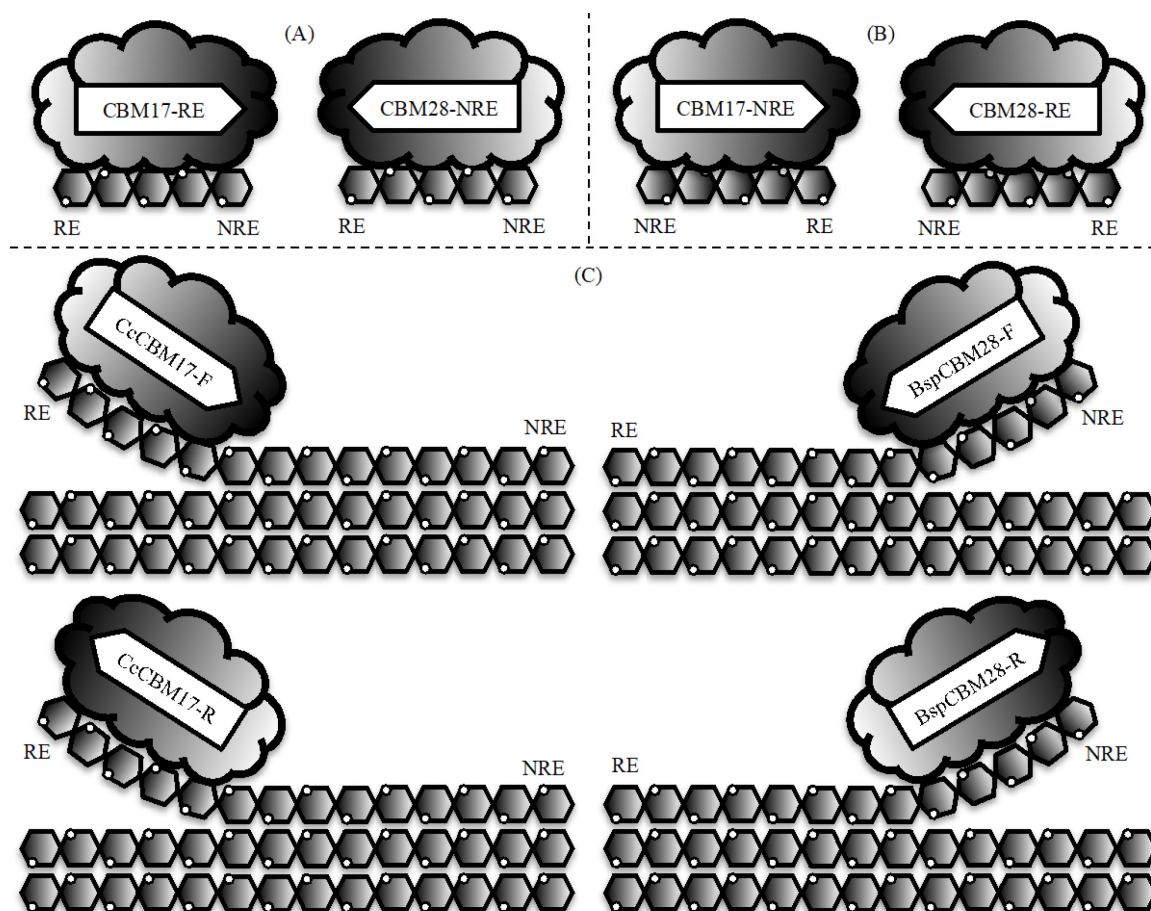


Figure 4.2 Cartoon illustration of the protein-carbohydrate complexes modeled in this study. CBMs from family 17 and 28 were modeled with cellopentaose bound in the (A) crystallographic structure orientation and (B) with the reducing end of the pyranose ring at the opposite end of the groove from the structural orientation. (C) CBMs were also bi-directionally bound with partially decrystallized cellulose I β microfibrils, approximating non-crystalline cellulose substrates. RE = reducing end; NRE = non-reducing end.

To investigate the bi-directional binding phenomenon in family 17 and 28 CBMs (Figure 4.2B), we rotated the ligand from the structural orientation longitudinally along the ligand, as described for *Cj*CBM4-1 and *Cj*CBM4-1 [184]. Cellopentaose was docked in the opposite direction of that captured in the crystal

structures by assuming the mean position of the pyranose ring heavy atoms must reside in approximately the same position regardless of direction. The approximate symmetry of the pyranose chair conformation enables this by merely exchanging the ring atom coordinates. CHARMM internal coordinate data was then used to establish the coordinates of the remaining sidechain atoms [193-195]. Extensive stepwise minimization of the ligand and the protein system was conducted before and after solvation to remove any deformation or bad contacts. These four systems, representing the “opposite” orientation, have been named *CcCBM17-NRE*, *BspCBM17-NRE*, *BspCBM28-RE*, and *CjCBM28-RE* for reference here.

We hypothesize high affinity CBM-binding occurs when the CBM associates with amorphous or non-crystalline cellulose via partially decrystallized oligomeric chains decorating the top layers of degraded cellulose microfibrils (i.e., whiskers). Here, the partially decrystallized microfibril model used to represent amorphous/non-crystalline cellulose was adapted from the three-layer cellulose I β model used in previous cellulose decrystallization studies [209, 210]. The five-pyranose long decrystallized segment was aligned with the cellopentaose from the equilibrated oligomeric systems described above using PyMOL (Figure 4.3). We docked two CBMs, a representative from both families 17 and 28 selected based on the availability of experimental affinity data for later comparison, in both ligand orientations such that we explore both possible interactions between these CBMs and non-crystalline cellulose. When aligned with each other or with *CfCBM4-1-RE* (Figure 3.13), *CcCBM17* and *BspCBM28* appear to bind their cello-oligomeric ligands in opposite orientations, relative to the directionality of the core β -sheets. Assuming the structural orientations represent thermodynamically-preferred

recognition modes, we docked the *CcCBM17* on the cellulose reducing end and *BspCBM28* on the cellulose non-reducing end and refer to them as *CcCBM17-F* and *BspCBM28-F* (i.e., ‘forward binding mode’). A second set of systems were prepared with the CBMs in the ‘reverse binding mode,’ exploring both bi-directional binding and additional CBM-substrate recognition mechanisms. These ‘reverse’ systems are referred as *CcCBM17-R* and *BspCBM28-R* (Figure 4.2C). System construction was followed by extensive minimization and 1-ns of *NPT* equilibration to ensure the stability of the modeled protein-carbohydrate interaction and reduce solvation effects. During heating, equilibration, and production MD, the lower layer of the cellulose microfibril was restrained by applying harmonic restraints to the pyranose ring atoms; the CBMs and all other atoms of the systems were free of restraints. Protein alignment and ligand docking by alignment was carried out using PyMOL [171] and Dali pairwise comparison version 3.1 [152].

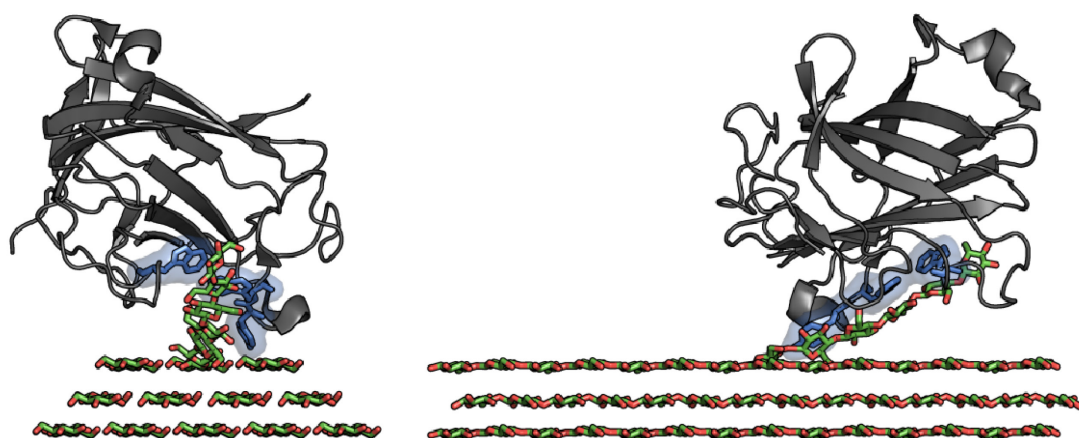


Figure 4.3 Initial position of *BspCBM28* in the forward orientation (after 500 ps of *NPT* equilibration) over the cellulose-I β microfibril with a middle chain of the top layer occupying the binding cleft of the CBM. The front view (left) and left-side view (right) illustrate the CBM (gray cartoon), its aromatic residues in the shallow binding cleft (blue sticks with transparent surface), and the cellulose microfibril (green sticks with red oxygens). A similar setup approach was used for the other three cases.

4.2.2 MD simulation parameters and protocols

The CHARMM36 force-field with CMAP corrections was used to simulate all proteins [191, 192], and carbohydrates were modeled with the CHARMM36 carbohydrate force-field [193-195]. Water molecules were represented by the modified TIP3P force-field [196, 197]. Ions were modeled based on the force-field by Beglov and Roux [211]. After acquiring the atomic coordinates from crystal structures (*CcCBM17* – 1J84, *BspCBM28* – 1UWW & *CjCBM28* – 3ACI) and homology modeling (*BspCBM17*), the pKa values of the CBMs' titratable residues were determined at pH 7.0 using H++ web server [189]. Visual inspection revealed additional residues to be

protonated, including Asp200 in *CcCBM17*, Asp72 in *BspCBM17*, Asp184 in *BspCBM28*, and Asp198 in *CjCBM28*. Sixteen different molecular dynamics (MD) simulations were constructed using CHARMM [166]. The systems, containing CBMs, crystallographic waters, calcium ions, and ligands (cellopentaose or microfibril), were constructed in vacuum and minimized for 1000 steps of Steepest Descent (SD) and 1000 steps of adopted basis Newton-Raphson (ABNR) with a tolerance of 0.01 for the average gradient. The vacuum-minimized systems were then solvated in explicit water, where the apo CBMs and CBMs bound with cellopentaose were solvated in a $70 \text{ \AA} \times 70 \text{ \AA} \times 70 \text{ \AA}$ cubic box ($\sim 35,000$ atoms), and the CBMs bound with the cellulose microfibril were solvated in a $110 \text{ \AA} \times 80 \text{ \AA} \times 110 \text{ \AA}$ orthorhombic box. To neutralize the system charge, sodium or chloride ions were added by replacing random waters with the ions. The solvated systems were then subjected to extensive stepwise minimization: 2000 steps of SD with the protein and ligand fixed, 2000 steps of SD with only the protein heavy atoms fixed, and 10000 steps of SD and 10000 steps of ABNR (tolerance 0.01) with no restraints. The minimized systems were heated from 100 K to 300 K in 50 K increments over 20 ps, and then equilibrated for 500 ps in the *NPT* ensemble at 300 K and 1 atm. The Nosé-Hoover thermostat and barostat were used to control temperature and pressure in CHARMM [212, 213].

For the data collection (production) MD, in the *NVT* ensemble, the apo and oligomeric systems were simulated for 250 ns seconds, while the CBM-microfibril systems were simulated twice (independently) for 100 ns each. These simulations were carried out at 300 K using NAMD 2.10 [169]. The Langevin thermostat was used to control temperature [214], and the SHAKE algorithm was used to fix the bond distances

of all hydrogen atoms [215]. Non-bonded interactions were truncated with a cutoff distance of 10 Å, a switching distance of 9 Å, and a non-bonded pair list distance of 12 Å. Long range electrostatics were described using the Particle Mesh Ewald (PME) method with 6th order b-spline, a Gaussian distribution of 0.320 Å, and a 1 Å grid spacing [216]. The velocity Verlet multiple time-stepping integration scheme was used to evaluate non-bonded interactions every 1 time step, electrostatics every 3 time steps, and 6 time steps between atom reassignments. All simulations used a 2-fs time step.

4.2.3 Free energy calculations

We calculated the absolute free energies of binding cellopentaose to CBMs for all three families using an enhanced sampling free energy method, FEP/ λ -REMD. FEP/ λ -REMD is an enhanced sampling free energy methodology developed by Jiang, Hodoscek [174], which we have previously implemented for protein-carbohydrate systems obtaining good agreement with experimental data [113, 184, 217]. For two different systems, the CBM-cellopentaose complex in solvent and solvated cellopentaose, the non-bonded interactions of cellopentaose with the rest of the system were systematically turned off to obtain the change in free energy. This free energy calculation protocol was implemented using dedicated module in NAMD [169]. The non-covalent interaction between the CBM and cellopentaose was distributed into repulsive, dispersive, electrostatic, and restraining components over 128 replicas. The total change in free energy of binding was then calculated as the aggregate of ΔG_{repu} , ΔG_{disp} , ΔG_{elec} , and ΔG_{rstr} . The difference between the free energy of ‘disappearing’ cellopentaose from the CBM groove into vacuum and the solvation free energy of cellopentaose gives the absolute free energy of binding a solvated ligand to a solvated protein. Convergence of

the free energy values was determined by Multistate Bennett Acceptance Ratio (MBAR) analysis method [176] and can depend on whether the model was prepared from crystal structure or homology model. Free energy calculations using models implementing ligand docking or homology modeling included additional restraining forces to improve convergence. For direct comparison, the FEP/ λ -REMD calculations conducted here comply with the specifications outlined in our earlier study of family 4 CBMs [184]; accordingly, all methodological details are identical.

Umbrella sampling MD was used to determine the potential of mean force (PMF) of decoupling the CBM from the model non-crystalline surface into the solvent, from which we can estimate the free energy of binding. The distance between the projection of the center of mass of the CBM and the projection of center of mass of the lower layer of the cellulose microfibril on the Z-axis served as the reaction coordinate. This distance was gradually increased by 15 Å in 0.5 Å increments, or 31 windows, until the non-bonded interaction between the protein and substrate no longer existed. The biasing force along the reaction coordinate was applied using collective variables during the 10 ns MD of each window in NAMD [169]. To assist strictly perpendicular movement of the CBM relative to the microfibril surface, the distance between the same pair of projections on the X- and Y-axes was restrained as a constant. The harmonic restraint on the ring atoms of the lower layer of the microfibril was maintained throughout sampling. A force constant of 10 kcal/mol was used to maintain the collective variables to their specified values. In *CcCBM17-F*, the pyranose ring immediately prior to the decrystallized chain was harmonically restrained to the cellulose surface preventing further decrystallization as the CBM was pulled away. Last 5 ns data was used in construction of potential mean

force at each window discarding first 5 ns to account for equilibration. The reaction coordinates were normalized to represent the change in distance (i.e. 0 Å to 15 Å). The calculation of potential mean force profile and error analysis was performed using MBAR analysis [176].

4.3 Results and discussion

4.3.1 Role of binding site architecture in substrate recognition

The three CBM families, 4, 17, and 28, share the same β -sandwich protein fold but exhibit key differences in binding site architectures/platforms. As they all belong to the Type B classification, the binding site generally conforms to either a cleft or groove capable of accommodating a single glycan chain. However, structural examine reveals the family 4 CBMs exhibit much deeper binding clefts relative to the more open grooves of family 17 and 28 CBMs, which we expect plays a critical role in substrate recognition mechanisms. Both *Cj*CBM4-1 and *Cj*CBM4-2 display aromatic residues lining the cleft and whose hydrophobic surfaces face each to sandwich the substrate pyranose rings between them (Figure 4.4). The oligomeric substrate is enveloped in a 4 to 5 Å-deep cleft with its pyranose ring perpendicular to the CBM surface [62]. Family 17 and 28 CBM binding grooves also display aromatic residues, although they are positioned side-by-side with their hydrophobic surfaces exposed to the solvent. Additionally, these aromatic residues are not exactly aligned in parallel planes, as in Type A CBMs, but, rather, comprise a shallow 1 to 2 Å groove with a ‘twisted’ polysaccharide-binding platform [100, 101].

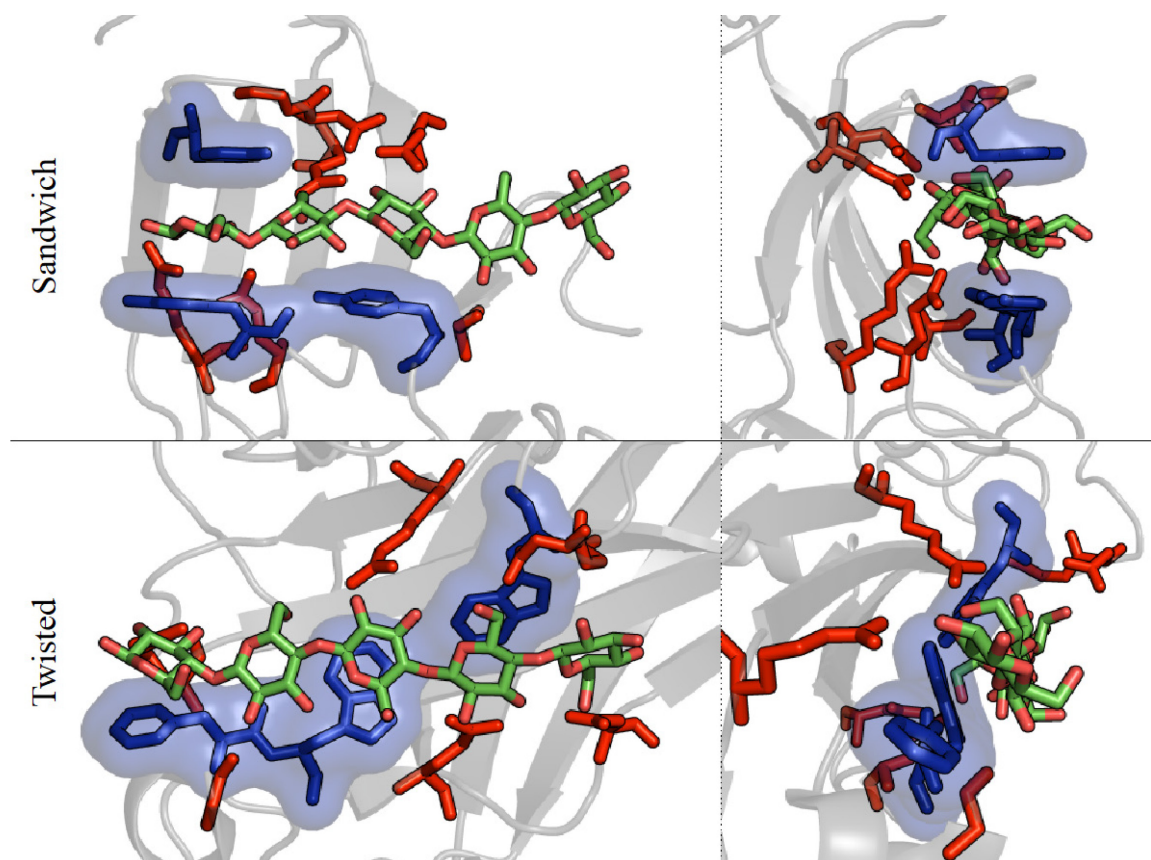


Figure 4.4 Differences in the two binding site architectures of family 4, 17, and 28 CBMs, as illustrated through hydrophobic interactions (dark blue sticks and transparent surface) and hydrogen bonding (red sticks) with the cellopentaose ligand (light green and red sticks). The front view (top left) and side view (top right) of the *Cj*CBM4-1 binding site with bound cellopentaose clearly show the sandwich platform and deep cleft with one-sided hydrogen bonding of the ligand. The front view (bottom left) and side (bottom right) of the *Cj*CBM28 binding site with bound cellopentaose show a twisted surface platform and shallow groove with hydrogen bonding partners available on both sides of ligand.

The significance of individual hydrophobic aromatic residues and polar residues in both family 17 and 28 CBMs has been examined in prior experimental studies [100, 102, 111]; however, binding affinity studies suggest that, despite structural and sequence similarity, thermodynamic binding signatures are not always consistent within members of the same family [96, 101]. We have previously discussed the similarities and differences within the two family 4 CBMs for ligand binding dynamics and thermodynamic preference [184]. In this section, we focus on comparing and contrasting oligomeric ligand binding modes and affinity across the two different binding platforms of CBM4s and CBM17 and 28s, ‘sandwich’ and ‘twisted,’ respectively, as well as within and across the three Type B CBM families.

Here, we select *Cj*CBM4-1 and *Cj*CBM4-2, having sandwich platforms, and *Cc*CBM17-RE and *Cj*CBM28-NRE, having twisted platforms for comparison, as experimental binding affinities and structures have been determined for each. Reported affinities for cellopentaose of each of the four CBMs are -5.24 ± 0.9 kcal/mol [98], -5.80 ± 0.005 kcal/mol [110], -5.80 ± 0.03 kcal/mol [100], and -7.7 ± 0.6 kcal/mol [102], respectively. Additional cellopentaose affinities have been reported for several of these CBMs, though experimental conditions vary making direct comparison challenging [96, 97, 218]. For the same four CBM·cellopentaose systems, we calculated binding affinity using FEP/ λ -REMD under conditions identical to experiment, at 300 K and pH 7.0. We found that *Cj*CBM4-1 and *Cj*CBM4-2 exhibited affinities for cellopentaose at -4.51 ± 1.30 kcal/mol [184] and 5.41 ± 1.38 kcal/mol. *Cc*CBM17-RE and *Cj*CBM28-NRE exhibited affinities for cellopentaose at -6.9 ± 0.9 kcal/mol and -6.3 ± 0.7 kcal/mol, respectively, and were more favorable than affinities of CBM4s. Detailed distribution of

free energy components, including charge, dispersion, van der Waals, and restraining contributions (Table 4.2), and illustration of calculation convergence (Figure 4.5) has been provided in Supplementary Material. These subtle thermodynamic preferences of different platforms are definitely one of the factors that play a role in building the recognition mechanisms targeted towards specific substrates. Further in this study, we address our hypothesis that this tighter binding in twisted platform is evolutionary feature of family 17 and 28 CBMs that allow them to preferably recognize non-crystalline cellulose over cello-oligomers.

Table 4.2 Distribution of free energy components of cellopentaose (G5) binding to CBMs at 300K and pH 7. All values are in kcal/mol. Errors for ΔG_b° represent one standard deviation.

System	ΔG_b°	ΔG_{rep}	ΔG_{disp}	ΔG_{elec}	ΔG_{rstr}
^a <i>Cj</i> CBM4-1 + G5	-4.51 ± 1.30	73.54 ± 0.19	-78.87 ± 0.05	-59.18 ± 0.15	0.29
<i>Cj</i> CBM4-2 + G5	-5.41 ± 1.38	81.19 ± 0.33	-81.01 ± 0.06	-67.59 ± 0.17	^b 2.03
<i>Cc</i> CBM17 + G5	-6.94 ± 0.91	76.09 ± 0.19	-77.27 ± 0.05	-67.81 ± 0.18	^b 2.05
<i>Cj</i> CBM28 + G5	-6.26 ± 0.74	75.74 ± 0.19	-72.55 ± 0.08	-69.34 ± 0.17	-0.11

^a Data obtained from Kognole and Payne [184]

^b Harmonic restraints were applied to rings atoms of ligand.

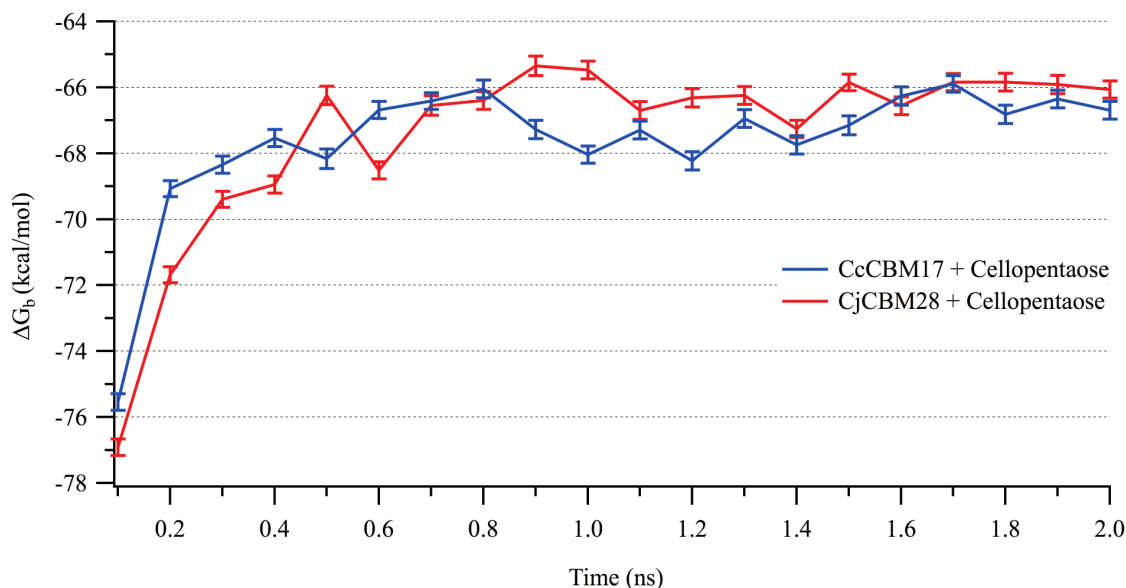


Figure 4.5 Convergence of the free energy calculations of cellopentaose binding to CBMs over 20 consecutive windows of 0.1 ns using enhanced sampling method FEP/ λ -REMD.

MD simulations provide additional insight into the binding free energy calculations, revealing that *CcCBM17*-RE and *CjCBM28*-NRE form more stable non-covalent interactions with the cellopentaose ligand than either family 4 CBM. From the 250-ns MD trajectories, we calculated the root mean square fluctuation (RMSF) of the ligand on a per-binding-subsite basis (Figure 4.6); error was estimated by block averaging over 2.5 ns blocks. This value describes how much a given pyranose ring fluctuates from its average position over the course of a simulation. Collectively, as well as in nearly every binding site, the pyranose rings within the *CjCBM4*-1 and *CjCBM4*-2 binding cleft fluctuate more than that of either *CcCBM17*-RE or *CjCBM28*-NRE, indicating the latter two ligands form more protein-carbohydrate contacts and are, likely, more tightly bound as we will show below. Moreover, the lower RMSF combined with

higher binding affinity in the twisted platforms suggests that the unfavorable entropic penalty is compensated by enthalpic contributions, especially hydrogen bonding, as discussed ahead.

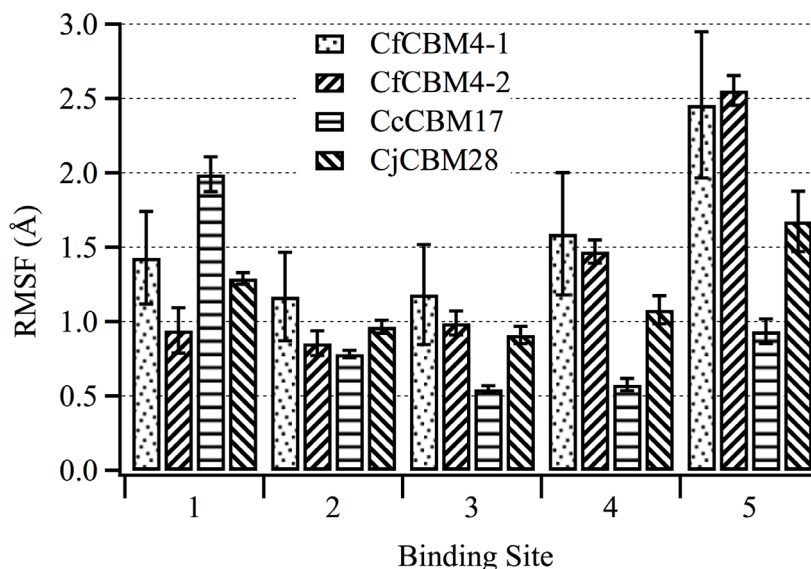


Figure 4.6 Root mean square fluctuation (RMSF) of the cellopentaose ligand from its average position in the clefts/grooves of representatives from family 4, 17, and 28 CBMs obtained from 250-ns MD simulation on a per-binding-subsite basis. Error was calculated from block averaging with block sizes of 2.5 ns. The binding site nomenclature with subsites 1 to 5 is assigned from reducing end to non-reducing end of cellopentaose; refer Figure 4.1.

There are three aromatic residues in the binding sites of the CBM4s and CBM28s, while CBM17s display only two, so the contribution to ligand binding from hydrophobic stacking interactions is not platform-dependent, varying by family. Rather, hydrogen bonding interactions appear to be a key determinant in affinity differences between the

two binding site architectures. The average number of hydrogen bonds formed between a given pyranose ring with the side chains of the surrounding protein was determined using VMD; detailed analysis of hydrogen bonding over the course of the MD simulations identified the primary hydrogen bonding partners in all the CBM-oligomer interactions. The average number of hydrogen bonds formed per binding site was calculated from the 250-ns MD trajectories, where a hydrogen bond was defined as two polar atoms having a donor-acceptor distance of $< 3.0 \text{ \AA}$ and a 20° cutoff angle. Table 4.3 shows the hydrogen bonding pairs from the calculations along with percent occupancy of each pair, where occupancy refers to the percent of the simulation during which the hydrogen bond was formed. While the CBMs with the same binding site architecture exhibit comparable hydrogen bonding, the total number of hydrogen bonds formed with the twisted platform was almost 100% higher than that of the sandwich platform. Total percent occupancy of 100% indicates that at any given time of simulation there is, on average, at least one live hydrogen bond between the ligand and protein. Along the twisted platform, there are one or more additional hydrogen bonding partners, accounting for an additional 1-2 kcal/mol of binding free energy for the whole binding site [219, 220]. *CcCBM17-RE* and *CjCBM28-NRE* form more hydrogen bonds with cellopentaose than either *CfCBM4-1* and *CfCBM4-2*, fitting with our conjecture that the loss of conformational entropy in ligand binding is compensated with enthalpic contributions to free energy. This difference in hydrogen bonding can be justified by analysis of the positioning of partner amino acid residues along the binding site. In the sandwich platform, where the ligand is approximately perpendicular to protein surface, primary and secondary hydroxyl groups of only one edge of cellopentaose chain contact the CBM and the other edge is exposed

to solvent (Figure 4.4). In contrast, the cellopentaose bound in the twisted platform hydrogen bonds with partners on both sides of the groove. In CBM4s, there are relatively few hydrogen bonding partners available at binding subsite 5, but in the case of *CcCBM17*-RE and *CjCBM28*-NRE, each binding subsite exhibits at least one residue capable of hydrogen bonding.

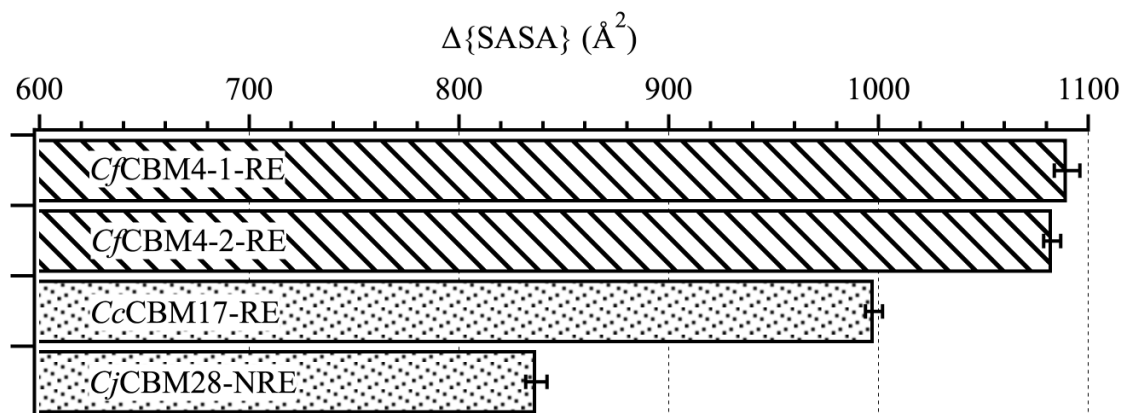
Table 4.3 Percent occupancy of each hydrogen bond formed between the pyranose ring at each binding site and the surrounding protein residue over the 250 ns simulation. Data are shown in decreasing order of occupancy. Pairs with occupancy lower than 1% are not shown. BGC is an acronym for β -D-glucose. A hydrogen bond was defined as two polar atoms having a donor-acceptor distance of $< 3.0 \text{ \AA}$ and a 20° cutoff angle.

	<i>Cf</i> CBM4-1			<i>Cf</i> CBM4-2		
	Donor	Acceptor	Occupancy	Donor	Acceptor	Occupancy
	ARG75-Side	BGC1-Side	26.86%	ARG81-Side	BGC1-Side	43.83%
Sandwich Platform	BGC3-Side	ASN81-Side	25.70%	BGC2-Side	GLN128-Side	40.02%
	BGC4-Side	ALA18-Main	25.14%	HSE132-Side	BGC2-Side	26.11%
	BGC2-Side	TYR43-Main	16.61%	BGC4-Side	LEU24-Main	7.83%
	BGC2-Side	GLN124-Side	15.37%	BGC4-Side	SER23-Side	6.96%
	GLN128-Side	BGC1-Side	6.54%	BGC3-Side	SER23-Side	5.24%
	GLY82-Main	BGC2-Side	5.11%	ASN56-Side	BGC3-Side	3.98%
	GLN124-Side	BGC3-Side	2.23%	SER23-Side	BGC3-Side	3.27%
	BGC4-Side	ASN50-Side	1.90%	SER23-Side	BGC4-Side	2.86%
	ASN50-Side	BGC3-Side	1.57%	GLN128-Side	BGC2-Side	1.47%
	BGC2-Side	ASN81-Main	1.43%	-	-	-
	BGC3-Side	GLN124-Side	1.11%	-	-	-
	GLN124-Side	BGC2-Side	1.04%			
	Total		130.61%	Total		141.57%

Twisted Platform	<i>CcCBM17-RE</i>			<i>CjCBM28-NRE</i>		
	Donor	Acceptor	Occupancy	Donor	Acceptor	Occupancy
	BGC3-Side	ASP54-Side	63.20%	ARG83-Side	BGC3-Side	65.98%
	BGC4-Side	GLN129-Side	59.43%	ARG178-Side	BGC1-Side	54.69%
	ARG92-Side	BGC2-Side	37.35%	BGC4-Side	GLN131-Side	45.93%
	BGC2-Side	ASP54-Side	27.80%	BGC5-Side	ASP76-Side	26.08%
	GLN129-Side	BGC3-Side	20.06%	BGC2-Side	GLY127-Main	23.77%
	ASN185-Side	BGC4-Side	15.67%	GLY77-Main	BGC5-Side	7.54%
	ASN137-Side	BGC1-Side	14.64%	BGC5-Side	ASP135-Side	3.14%
	ASN52-Side	BGC3-Side	9.25%	BGC4-Side	ASP76-Side	2.68%
	THR184-Side	BGC5-Side	2.64%	GLN131-Side	BGC4-Side	1.67%
	ARG92-Side	BGC1-Side	2.15%	TRP129-Side	BGC4-Side	1.57%
	BGC5-Side	THR184-Main	1.54%	TRP78-Main	BGC5-Side	1.12%
	BGC1-Side	ASN137-Side	1.08%	-	-	-
	Total		254.81%	Total		234.17%

Average change in solvent accessible surface area (SASA) upon ligand binding (Figure 4.7) reveals that the sandwich platform buried more solvent exposed surface area upon binding than the twisted platform, though the latter was more solvent exposed initially. The average change in SASA was calculated over 2500 frames of MD simulation, taking the difference between summation of average SASA of apo CBMs and average SASA of solvated cellopentaose and average SASA of respective CBM-cellopentaose complexes (Figure 4.7). The mean change in SASA is lower for twisted

platform CBMs than sandwich platform CBMs, with less of a change in SASA observed for *Cj*CBM28-NRE than *Cc*CBM17-RE. The extra aromatic residue (Phe128) in the *Cj*CBM28 binding groove, being the most obvious difference within the twisted platforms of family 17 and 28 CBMs, appears to contribute to this difference, but it also suggests that having an aromatic residue may not always contribute to higher change in SASA when compared to sandwich platform CBMs that also have three aromatic residues. Solvent-exposed residues along the twisted platforms do not appear to retain ordered water molecules proximal to the aromatic side chains when there is no bound ligand; upon ligand binding, additional water molecules were retained at the protein-carbohydrate interface, as hydroxyl groups of cello-oligosaccharides enable solvent reorganization [101]. Thus, there is limitation to use of mean change in SASA, which is only a quantitative measure, and cannot be directly correlated to entropic contribution through solvent reorganization. However, the larger change in SASA for sandwich platform CBMs reflects a conformational change upon ligand binding. As observed in the MD simulations, the deep cleft of the two family 4 CBMs narrows over time as it sandwiches the cello-oligomer and excludes water from the hydrophobic core of the protein. The twisted platform on the other hand does not appear to implement this sandwiching mechanism and may, rather, prefer sliding along the polysaccharide chain more freely that would involve less change in SASA.



$$\Delta\{\text{SASA}\} = \text{Avg}\{\text{SASA}(\text{Prot})\} + \text{Avg}\{\text{SASA}(\text{Lig})\} - \text{Avg}\{\text{SASA}(\text{Prot}+\text{Lig})\}$$

Figure 4.7 Average change in solvent accessible surface area (ΔSASA) calculated using VMD over the 250 ns MD simulation trajectories of each CBM-cellopentaose system to compare the difference between sandwich (lined pattern) and twisted (dotted pattern) platforms. The error bars represent the standard deviation (SD) of the mean.

Overall, MD simulation results, especially hydrogen bonding patterns, suggest that the difference in cleft architecture (i.e., twisted vs. sandwich) greatly contributes to differences in affinity and, likely, protein-carbohydrate recognition mechanism. The recognition mechanism of the oligomeric ligand by these two different architectures is readily distinguishable based on the binding affinity and hydrogen-bonding pattern. It is tempting to suggest variations in molecular-level behavior, such as these, are a result of evolutionary necessity, where each binding site architecture is uniquely suited for targeting regional substrate features [64, 100].

To further differentiate oligomeric recognition mechanisms between *CcCBM17* and *CjCBM28*, we compared ligand binding dynamics at each binding subsite (Figure 4.8). Despite the apparent similarity in binding site architecture, the two CBMs feature

cello-oligomers bound in opposite directions in their crystal structures (i.e., with the reducing end oriented at a different end of the groove when structurally aligned) [100, 101]. To enable comparison, the binding subsites of the two CBMs were structurally aligned, and a letter-based subsite nomenclature was invoked based on the two common solvent-exposed Trp residues (Figure 4.8). The RMSF and hydrogen bonding evaluations reported above follow the numbered binding subsite nomenclature from crystal structure publications, as a cumulative comparison across the platforms. Alignment and renaming binding subsites (A to F), as previously implemented by Tsukimoto, Takada [101], reveals that four common binding subsites (B, C, D, and E) are occupied by cellopentaose in *Cc*CBM17-RE (B to F) and *Cj*CBM28-NRE (A to E).

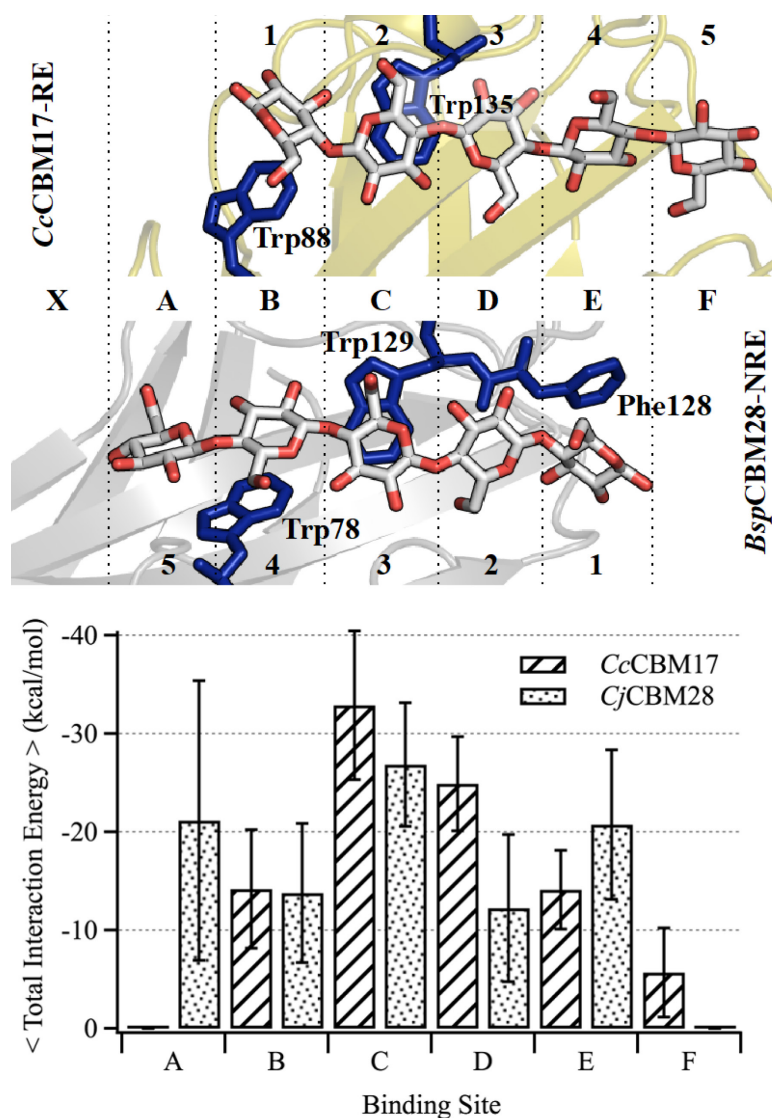


Figure 4.8 Alignment of the twisted platform binding sites of *CcCBM17*-RE (top) and *CjCBM28*-NRE (bottom) with respect to the common pair of Trp residues (dark sticks). The new common naming of binding subsites (letters) is given in between the panel, and the original nomenclature (numbers) is given above and below the cartoon representations. (B) Average total interaction energy of the pyranose rings with the surrounding amino acid residues, on a per-subsite-basis, of *CcCBM17*-RE and *CjCBM28*-NRE calculated from the 250-ns trajectory. Error bars represent 1 SD.

The average total interaction energy of protein with the cellopentaose ligand was determined from the 250-ns trajectory on a per-binding-subsite basis. The interaction energy distribution was very similar for CBM17 and CBM28, in the binding subsites B and C that reside along the hydrophobic face of pair of Trp residues common to both CBMs (Figure 4.8). Difference arises in the binding subsites as the extra aromatic residue in family 28 CBMs (Phe128 in *Cj*CBM28 and Tyr118 in *Bsp*CBM28) that can provide hydrophobic stacking interaction at binding site E. Nevertheless, we observe little difference at subsites D and E in total interaction energy calculation that accounts for both van der Waals and electrostatic interactions. Based on overall analysis oligomeric binding affinities, collective nature of hydrogen bonding and total interaction energy at twisted platform, one can agree that these two families with same platforms exhibit very similar binding mechanisms for oligomeric ligands except the difference in mean change in SASA. CBM17s and CBM28s are reported to bind oligomers as long as cellohexaose [100] [96], and it is apparent that, for CBM17s, the sixth subsite would be A, while for CBM28s, the sixth sugar can be accommodated in either F or X. As *Cc*CBM17 and *Cj*CBM28 are known to bind non-crystalline substrates as well, it is possible there exist secondary binding subsites for chains even longer than cellohexaose. Accordingly, we docked cellohexaose with *Cj*CBM28 in two orientations, occupying subsites A to F and X to E, and conducted 100-ns MD simulations; these simulations showed that both subsite X and F functionally interact with the ligand, although X had a higher interaction than subsite F (Figure 4.9). Extended binding sites may play a critical role in recognition of non-crystalline substrates, as we will discuss ahead.

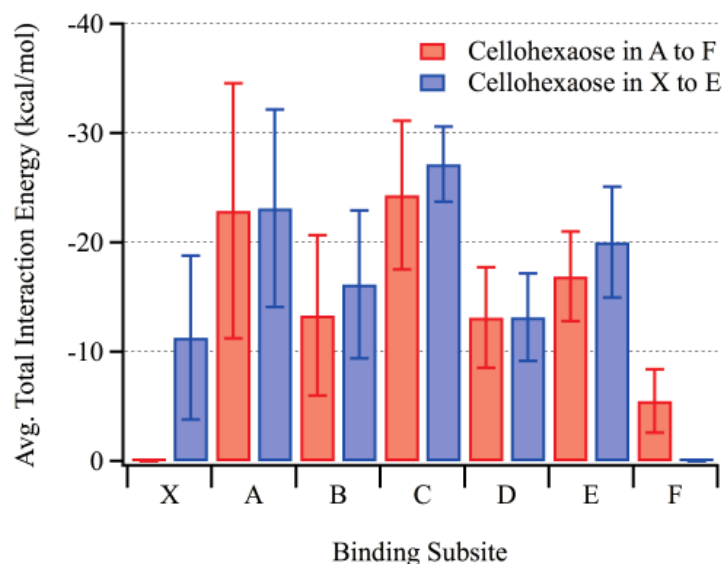


Figure 4.9 Average total interaction energy per binding subsite with the surrounding amino acid residues of *Cj*CBM28 for a cellohexaose chain of the microfibril occupying the cleft in two different ways, A to F (red) and X to E (blue). Values were calculated over the entire 100-ns trajectory. The error bars represent one standard deviation.

4.3.2 Differentiation of high and low affinity binding sites non-crystalline cellulose

Structural characterizations of many carbohydrate active enzymes focus strictly on the interactions occurring in the carbohydrate binding site or catalytic active sites, while protein surface residues or secondary binding sites may be just as important to functionality [221]. Type B CBMs are reported to bind both cello-oligomers and non-crystalline/amorphous cellulose, covering a broad range of polymeric structural diversity and suggesting recognition processes may involve interactions beyond the primary binding site. Interestingly, CBMs from families 17 and 28 appear to bind non-crystalline cellulose with high and low binding affinities, as determined from isothermal titration calorimetry (ITC) data, and the two families do not compete with each other for

carbohydrate binding sites [64, 65]. We further explore both the concept of bi-directional binding and the high/low binding affinity phenomena of family 17 and 28 CBMs on non-crystalline cellulose by modeling representative Type B CBMs bound with a model non-crystalline substrate in multiple orientations. At the nanoscale, we propose a partially decrystallized cellulose I β microfibril sufficiently represents the interaction of a CBM with non-crystalline cellulose, which retains a significant degree of crystallinity. Additionally, given our above insights into family 17 and 28 CBM members (i.e., that there is relatively little difference in oligomeric binding dynamics between members of the same family), we modeled only four representative CBM·microfibril systems: one CBM from each family attached to the decrystallized chain, or ‘whisker,’ in two possible orientations, forward and reverse. Details of this have been provided in methods section above.

Fully atomistic MD simulations were used to explore the primary modes of Type B carbohydrate recognition with respect to non-crystalline cellulose. All atomic interactions were unbiased except for the lower layer of the cellulose microfibril, which was harmonically restrained to prevent excessive fraying and further decrystallization. In all four cases, *CcCBM17-F*, *CcCBM17-R*, *BspCBM28-F*, and *BspCBM28-R*, the CBM·non-crystalline cellulose complexes stabilized in a global minimum state in each of the 100-ns MD simulations, illustrated by the rapid plateau in the protein backbone RMSD over time (Figure 4.10). Throughout the simulation, most CBMs bind all five pyranose moieties of the whisker along the twisted binding sites in the fully decrystallized state; in the case of *BspCBM-F*, the fifth pyranose ring closest to the cellulose surface partially re-annealed into the microfibril, which is not unexpected [222].

Comparing RMSF of the CBM backbone when bound to either an oligomer or non-crystalline cellulose reveals that ligand binding stabilized the protein (lower RMSF); unbound CBM RMSFs exhibited larger fluctuations near binding site residues in both CBMs. Only *Bsp*CBM28-RE, bound with the rotated cellopentaose, showed large protein backbone fluctuations (Figure 4.11), resulting from ligand movement along the binding groove discussed in Chapter 3 (Figure 3.15).

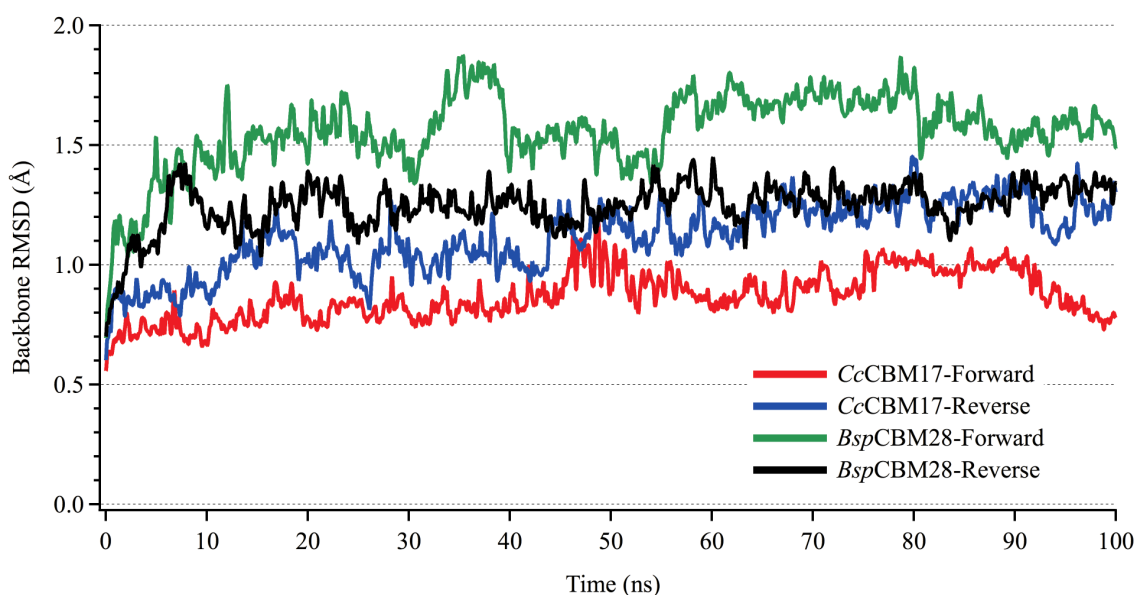


Figure 4.10 RMSD of the CBM backbone bound to the model non-crystalline cellulose microfibril over 100 ns of MD simulation (10000 frames captured at every 0.01 ns). RMSD was determined with respect to the coordinates of each respective CBM at 0 ns.

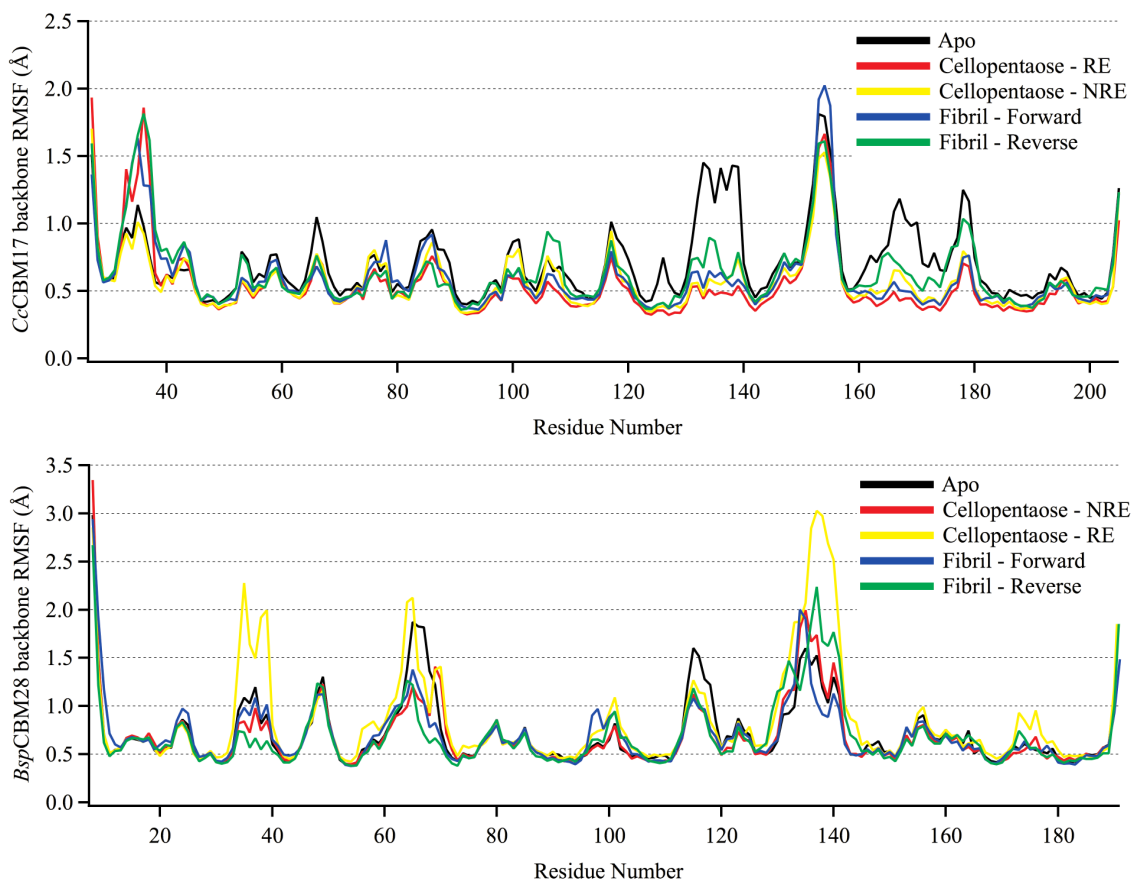


Figure 4.11 Root mean square fluctuation (RMSF) of the backbone atoms of CcCBM17 (top) and BspCBM28 (bottom) in each ligand occupancy state.

Comparing these four simulations and the oligomer-bound simulations above, we identified molecular-level factors contributing to substrate recognition in each family with respect to variation in substrate and orientation. The interaction energy of each CBM residue with the substrate was determined by averaging the calculation over trajectories, for all CBM·substrate systems (Figure 4.12). For both family 17 and 28 CBMs, the average interaction of a given CBM residue with the substrate is independent of direction of cellopentaose ligand in the binding site, which is, again, consistent with bi-directional ligand binding. The hydrophobic-stacking aromatic residues and hydrogen bonding partners of the CBM·cellopentaose systems, as discussed above, produce substantial

favorable interaction energies (< -5 kcal/mol). These same residue·substrate interactions exist when the CBM is bound with non-crystalline cellulose. However, additional protein residues along the CBM surface also appear to be involved in binding non-crystalline cellulose (Figure 4.12), as revealed from the rather significant new interactions formed in regions where the CBM·cellopentaose systems produce no such interactions.

While it is clear that protein surface residues play an auxiliary role in non-crystalline cellulose binding, each CBM and orientation relative to the cellulose surface results in a unique set of protein·substrate interactions to amplify non-crystalline cellulose binding affinity over oligomeric affinity. In the case of *CcCBM17-F*, two peptide loops adjacent to the binding groove, residues 30-35 and 95-106, interact with cellulose as a result of their proximity to the cellulose surface in this ‘forward’ orientation. Most residues in these loops are polar residues, including Pro31, Lys32, Asp33, Asp96, Gln100, Ser101, Asn103, and Tyr105, and serve to anchor the CBM over the microfibril through additional hydrogen bonding. In the case of *CcCBM17-R*, Asp81, Asn86, and Asn137 produce new, large electrostatic interactions between the CBM and substrate. Also, aromatic residues like Trp88 produced more favorable interaction energies in the reverse orientation, while interacting loops in the forward orientation played no role at all. Similarly, for *BspCBM28-F*, the family 28 CBM lost hydrogen bond interactions between the ligand and Arg73 in the binding groove (subsite 3) and Gln112 (subsite 4); however, new hydrogen bond interactions with residues in loop 65-68 were formed. The *BspCBM28-R* orientation exhibited more consistent interaction patterns, with no loss of affinity contributors and formation of additional favorable interactions between cellulose and residues in loops 66-68 and 115-130. Ultimately, it seems each

orientation of a given CBM relative to the cellulose surface produces a specific set of substrate interactions that enhance non-crystalline cellulose binding relative to oligomeric binding.

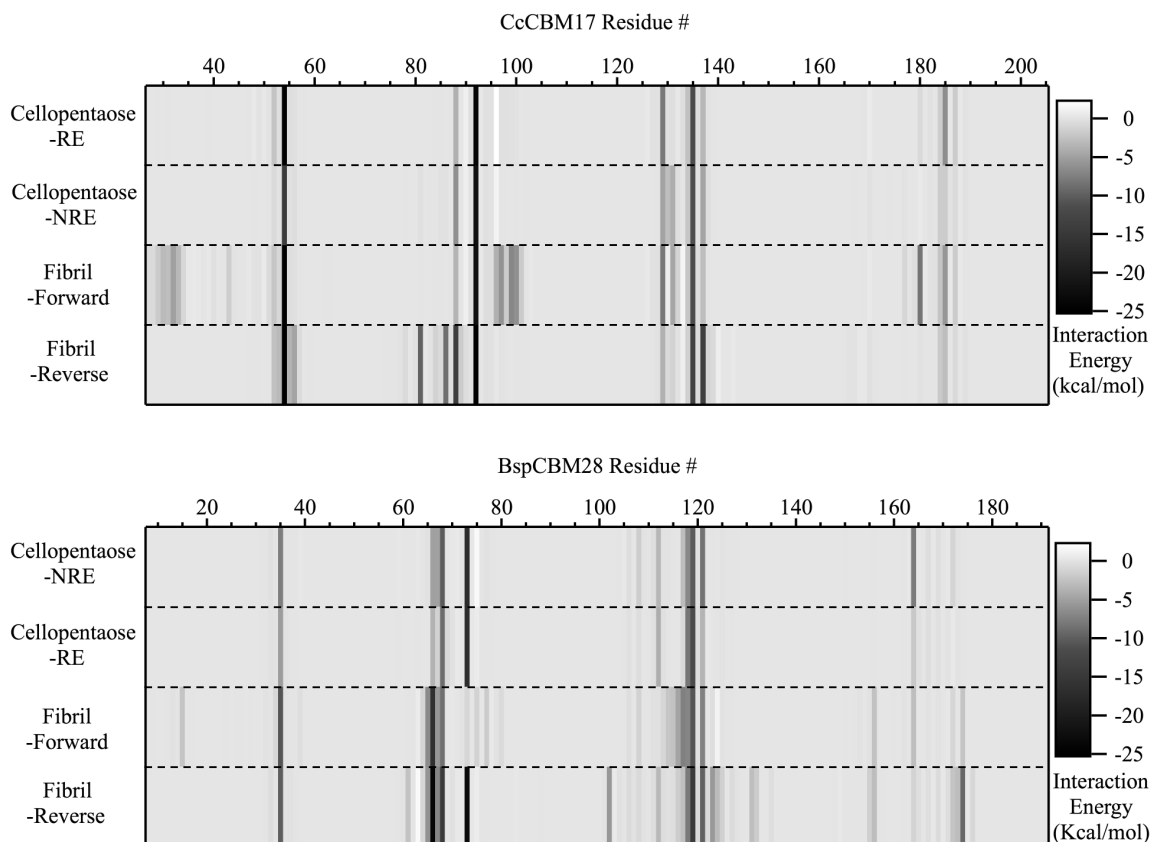


Figure 4.12 Total interaction energy between the substrate and each protein residue, averaged over the length of the MD simulations. The *CcCBM17* (top) and *BspCBM28* (bottom) residue numbers are shown along the x-axis. The simulation case label is given at left, four cases for each family 17 and 28 CBM. The magnitude of the interaction energy between a given residue and the bound ligand, as indicated in the case name, is shown in grayscale. Favorable interactions are more negative and, thus, darker. In cellopentaose binding, ligand direction does not affect CBM·cellopentaose interactions,

as redundant protein residues along the binding groove maintain association with cellopentaose. In non-crystalline cellulose binding, the CBM protein surface interacts with the surrounding carbohydrate, in both forward and reverse orientations, to enhance binding affinity; the new protein-carbohydrate interactions are unique for each CBM and each direction.

To thermodynamically characterize the effects of orientation and substrate crystallinity on family 17 and 28 binding, we calculated binding affinities from the potential of mean force (PMF), or work required, to separate the CBMs from the non-crystalline cellulose substrate. We used umbrella sampling MD to disassociate the CBM from non-crystalline cellulose, pulling the CBMs away from the substrate perpendicularly. Sampling simulations were started from equilibrated 100-ns MD simulation snapshots of each CBM·non-crystalline cellulose complex. For all four cases, the corresponding PMFs indicate binding affinities are higher for non-crystalline cellulose than for oligomeric ligands in respective CBMs (Figure 4.13); this result aligns with our hypothesis that the higher affinity binding sites described in experimental binding studies corresponds to CBM·non-crystalline cellulose binding and lower affinity binding sites correspond to CBM binding in oligomeric or highly decrystallized regions.

The PMF provides both a binding free energy and a quantitative view of the CBM dissociation process from a non-crystalline substrate (Figure 4.13). The free energy of binding non-crystalline cellulose is determined from the difference between the free energy at the beginning (0 Å) and end (15 Å) of the reaction coordinate. For both *CcCBM17* and *BspCBM28*, the orientation of the CBM relative to the surface affects

binding affinity, favoring the forward orientation in CBM17 and the reverse orientation with CBM28. Additionally, there is a significant difference in affinity between the two high-affinity orientations of each CBM family; *Cc*CBM17-F binds with the highest affinity, 23.0 ± 1.1 kcal/mol, and *Bsp*CBM28-R binds with an affinity equivalent to 15.9 ± 0.8 kcal/mol. Combined with the knowledge that these two CBM families do not competitively bind non-crystalline cellulose [64], our results suggest that CBMs from these two families are capable of recognizing cellulose binding sites based on binding orientations relative to the substrate. The difference between the affinity of CBM17 and CBM28 for non-crystalline cellulose may be correlated to the qualitative difference in the surface interactions that contribute to the affinity as well as fortuitous compatibility of CBM17s than CBM28s with proposed non-crystalline cellulose model. General surface topology around oligomeric binding site of CBM28 Decrystallized edge chain morphology could be one of the other cases of non-crystalline cellulose that are preferred by CBM28s over CBM17s.

The model non-crystalline substrate simulated in this study represents a subset of cellulose morphologies that are very close to crystalline substrate, and the calculated free energies correspond to association constants as high as 10^{12} mol⁻¹, which are not detectable by experimental methods such as ITC. The reported high affinity cellulose binding sites for *Cc*CBM17 and *Bsp*CBM28 on regenerated cellulose, from ITC, were -8.41 ± 0.32 and -8.28 ± 0.35 kcal/mol, respectively [64] and while these values are much lower than those calculated from PMFs, it is plausible that the experimental affinities correspond to a range of other cellulose morphologies more amorphous in nature than the model non-crystalline substrate. Nevertheless, taken qualitatively together with the

calculated and experimental values of cellopentaose binding to *CcCBM17* and *BspCBM28*, our results offer promising evidence that high and low affinity non-crystalline cellulose binding sites correspond to degree of substrate crystallinity. In other words, these family 17 and 28 CBMs appear to bind cellulose with a higher degree of crystallinity with greater affinity than small, oligomeric substrates.

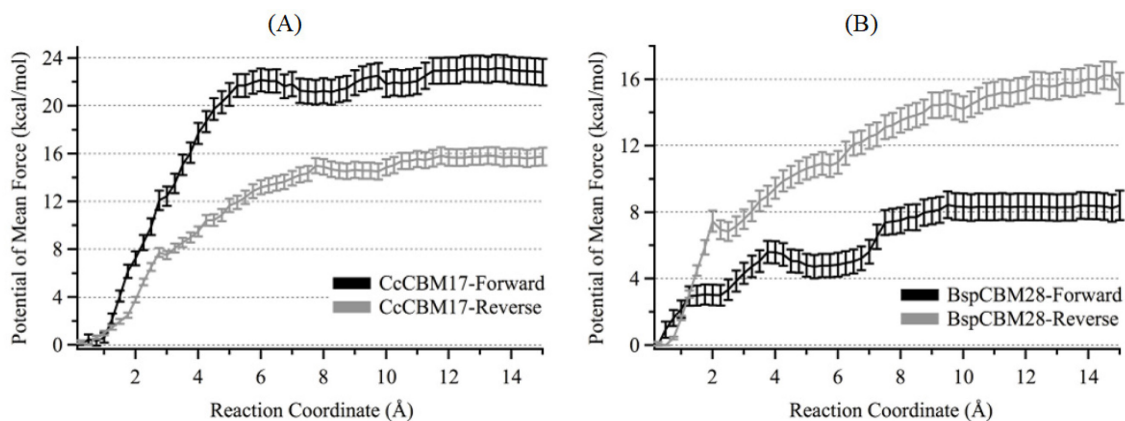


Figure 4.13 Potential of mean force (PMF) in uncoupling (A) *CcCBM17* and (B) *BspCBM28* from non-crystalline cellulose. Umbrella sampling MD was conducted over 30 0.5-Å-windows using the projection of the distance vector on the *z*-axis as the reaction coordinate.

Finally, dissociation appears to occur in two separate events along the PMF profile (Figure 4.13), with an initial exertion of work to decouple the CBM from the substrate surface and a final extrication of the polymeric chain from the CBM binding groove. The CBM bound with non-crystalline cellulose must initially overcome the strong electrostatic interactions and hydrogen bonds formed between the CBM protein surface and the cellulose surface. After the exterior of the CBM was free of the cellulose surface, the final amount of work required to dissociate the CBM was associated with

overcoming both van der Waals interactions between with the aromatic residues and pyranose rings and several hydrogen bonds formed with the substrate along the length of the groove. Combined with our MD simulation results above, the increase in affinity observed in binding *CcCBM17* and *BspCBM28* with non-crystalline cellulose with appears to be directly related to the additional protein·carbohydrate interactions mediated by residues exterior to the CBM binding groove.

4.4 Conclusions

With better resolution of thermodynamic affinities and detailed analysis of protein-carbohydrate interactions like hydrogen bonding for two different binding platforms within same type of CBMs, it is evident that binding site architecture has profound effect on CBM functionality in recognizing carbohydrate substrates. Comparison of twisted platform in two different CBM families, 17 and 28, showed similarity in oligomeric ligand binding dynamics and certainly provided sufficient rationale towards their extended binding sites. Consequently, we have addressed the many questions raised by Boraston et. al. in regards to mechanisms of Type B CBM·non-crystalline cellulose binding, expanding upon experimental observations identifying enthalpic interactions as dominant in non-crystalline substrate recognition by *CcCBM17* and *BspCBM28* [115]. Specifically, we identified individual contributions to thermodynamics parameters, revealing that the gain in enthalpy in binding non-crystalline cellulose over oligomers results from direct contact of the CBM exterior with the cellulose substrate. We also provided insights into how family 17 and 28 CBMs could uncompetitively bind non-crystalline cellulose, despite having very similar binding specificities and protein structure. The question of specifically assigning CBM·cellulose

binding affinities to non-crystalline substrate binding sites remains, hinging on future experimental efforts to structurally characterize non-crystalline cellulose of increasingly amorphous nature. This study also provides the basis for our future investigations of glycoside hydrolases linked with tandem CBMs, as the two family 4 CBMs (*Cf*CBM4-1 and *Cf*CBM4-2) and the two *Bacillus sp. 1139* family 17 and 28 CBMs (*Bsp*CBM17 and *Bsp*CBM28) are natural tandem constructs appended to β -1,4-endoglucanases. We anticipate the results toward understanding Type B CBM oligomeric and non-crystalline recognition mechanisms will advance our understanding of how protein-protein interactions and inter-module networking determines additive or cooperative binding in tandem systems and why organisms secrete multi-modular enzymes with seemingly redundant CBM domains.

Chapter 5 – Carbohydrate ligands of YKL-40: Binding mechanisms, thermodynamic preferences and surface binding ability

In Chapter 5, we report molecular-level investigation of YKL-40s binding sites for carbohydrate ligands like chito-oligomer and determine the most likely physiological binding partner. This chapter has been adapted with permission from Kognole and Payne [223], Copyright © 2017, American Society for Biochemistry and Molecular Biology.

5.1 Abstract

YKL-40 is a non-catalytic mammalian glycoprotein and biomarker associated with progression, severity, and prognosis of chronic inflammatory diseases and a multitude of cancers. Despite this well-documented association, conclusive identification of the lectin's physiological ligand, and accordingly, biological function, has proven experimentally difficult. From experiments, YKL-40 has been shown to bind chito-oligosaccharides; however, the natural production of chitin by the human body has not yet been documented. Possible alternative ligands include proteoglycans, polysaccharides, and fibers such as collagen, all of which make up the mesh comprising the extracellular matrix. It is likely that YKL-40 is interacting with these alternative polysaccharides or proteins within the body, extending its function to cell biological roles such as mediating cellular receptors and cell adhesion and migration. Here, we consider the feasibility of polysaccharides, including cello-oligosaccharides, hyaluronan, heparan sulfate, heparin, and chondroitin sulfate as potential physiological ligands for YKL-40. Molecular dynamics (MD) simulations resolve the molecular-level recognition mechanisms, as several of these potential ligands appear to bind YKL-40 in modes

analogous to chito-oligosaccharides. Further, we calculate the free energy of binding of the hypothesized ligands to YKL-40 to address thermodynamic preference relative to chito-oligosaccharides. Our results suggest that chitohexaose and hyaluronan preferentially bind to YKL-40, and hyaluronan is likely the preferred physiological ligand, as the negatively charged hyaluronan shows enhanced affinity for YKL-40 over neutral chitohexaose. Finally, heparin non-specifically binds at the surface of YKL-40, as predicted from structural studies. Overall, YKL-40 likely binds many natural ligands *in vivo*, but its concurrence with physical maladies may be related to associated increases in hyaluronan.

5.2 Introduction

Significance of YKL-40 as a biomarker in various malignancies and structural properties of this chitinase-3-like-1 protein have been described in the general introduction (Section 1.3.2). In this chapter, we specifically focus on the polysaccharide binding sites of YKL-40 and exploring oligosaccharides of different glycosaminoglycans (GAGs) that can bind to YKL-40. Despite the structural similarity between chito-oligosaccharides and the GAG monomers, little evidence of polysaccharide binding beyond the original structural studies exists [43, 44]. In fact, we are aware of only one other study focusing on the molecular-level mechanism of carbohydrate binding in YKL-40 [224]. From a bioinformatics and structural comparison of YKL-40 to a similar chitinase, mammary gland protein-40, the authors propose an oligosaccharide binding mechanism that involves tryptophan-mediated gating of the primary carbohydrate binding site [224, 225]. Though in lieu of a dynamics-based investigation, little can be concluded about the binding mechanism of YKL-40 ligands other than chito-

oligosaccharides, and conformational changes relative to binding are inaccessible. From protein purification techniques, namely heparin-Sepharose chromatography, we also know that YKL-40 reversibly binds heparin [44, 128, 226]; however, affinity data for this interaction does not exist. Based on the interaction with heparin, it is reasonable to hypothesize heparan sulfate glycosaminoglycans, existing as part of the extracellular matrix construct, are a potential physiological ligand. Visual inspection of the protein structure initially suggested heparan sulfate fragments might be easier to accommodate within the carbohydrate binding site than heparin itself [44]. It follows that other structurally similar carbohydrate fragments would bind with similar affinity in a comparable mechanism.

Understanding the mechanism and affinity by which YKL-40 binds ligands is crucial to our comprehension of its physiological function. This knowledge will serve as a foundation for future campaigns toward rational development of a potent antagonists enabling cell biological study and addressing YKL-40 as a therapeutic target. To accomplish this goal, we must describe the molecular-level mechanisms governing the interaction of YKL-40 with polysaccharides and quantitatively evaluate affinity. In this study, we used classical molecular dynamics (MD) simulations to differentiate modes of ligand recognition and specificity. Using free energy perturbation with replica exchange molecular dynamics (FEP/ λ -REMD), we quantitatively determined affinities overcoming the experimental difficulties encountered thus far. As polysaccharide physiological ligands, we considered several options; provided below is a brief description of each carbohydrate ligand considered, as well as justification for consideration.

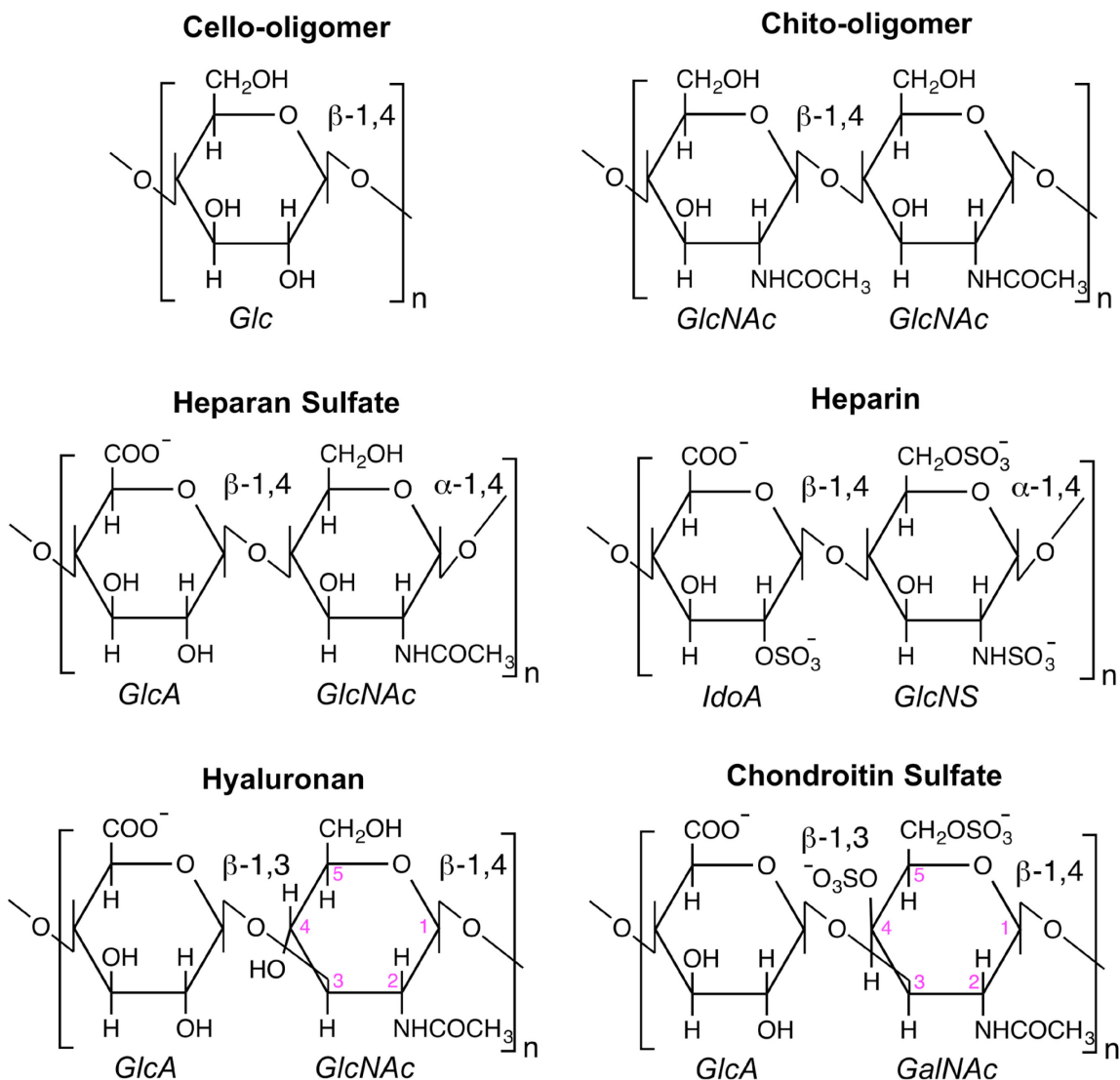


Figure 5.1 Monomeric units of the polysaccharides considered as potential physiological ligands of YKL- 40: cellohexaose, chitohexaose, heparan sulfate, heparin, hyaluronan, and chondroitin sulfate. The chito- oligomer is a polymer of β -1,4-linked GlcNAc monomers. Heparan sulfate was modeled as a β -1,4, α -1,4- linked chain of GlcA and GlcNAc. Heparin was represented as the β -1,4, α -1,4-linked oligomer of GlcA and GlcNS. Hyaluronan and chondroitin sulfate are β -1,3, β -1,4-linked oligomers; the former consists of GlcA and GlcNAc, and the latter consists of GlcA and GalNAc. Glc – β -D-

glucose; GlcNAc – N-acetyl- α -D-glucosamine; GlcA – β -D-glucuronic acid; IdoA – α -D-iduronic acid; GlcNS – N-sulfo- α -D- glucosamine; GalNAc – N-acetyl- β -D-galactosamine.

Chito-oligomer

After cellulose, chitin is the second most abundant naturally occurring biopolymer on earth [4], and is comprised of repeating N-acetyl- β -D-glucosamine (GlcNAc) monomeric units connected by β -1,4 glycosidic linkages (Figure 5.1). Based on the experimental evidence of *in vitro* binding to YKL-40 [43], the chito-oligomers were included as a control for comparison with other carbohydrates. Additionally, structural data is available for YKL-40 bound to chito-hexaose, which was used to as base in our computational modeling [43, 44].

Cello-oligomer

The central ring of the monomer, a six-membered pyranose, is common to a number of carbohydrates including glucose, the monomer of cellulose. Given this chemical similarity with chitin, as well as the general presence of glucose in mammalian cells as a form of energy, a hexameric cello-oligomer was also examined as a potential physiological ligand, despite its unlikely presence among mammalian GAGs.

Heparan sulfate and heparin

As described earlier, YKL-40 binds heparin, and thus, likely also binds heparan sulfate. Heparan sulfate, a less sulfated form of heparin, is a polysaccharide found in

abundance in the ECM and at the cell surface [227]. Heparan sulfate is constructed from a repeating disaccharide of β -D-glucuronic acid and N-acetyl- α -D-glucosamine (Figure 5.1). Of all the glycosaminoglycans, heparan sulfate is the most structurally complex. At least 24 different combinations of the disaccharide monomer exist, with differences arising as a result of variation in both isomer and degree of side chain sulfation [228]. Additionally, the heparan sulfate polysaccharide can exhibit both sulfated (NS) and unsulfated (NA) domains. Physiologically, the unsulfated disaccharide β -D-glucuronic acid – (1,4) N-acetyl- α -D-glucosamine is the most prevalent form of heparan sulfate [228]. Focusing on the most relevant physiological ligands, we examined the fully sulfated form heparin and the completely unsulfated form heparan sulfate.

Chondroitin sulfate

Chondroitin sulfate is also a glycosaminoglycan prevalent in mammals reportedly known to have various functions as cell surface receptors, as extracellular signaling molecules, in sulfation-mediated neuronal plasticity, and in myogenic differentiation/regeneration [229]. The primary structural units of chondroitin sulfate are a repeating β -D-glucuronic acid and N-acetyl- α -D-galactosamine disaccharides connected by alternating β -1,3 and β -1,4 glycosidic linkages (Figure 5.1). As with heparan sulfate, chondroitin sulfate exists in variably sulfated types along with its stereoisomer dermatan sulfate [229, 230]; we have selected the 4,6-O-disulfated GalNAc variant of chondroitin sulfate polysaccharide as our candidate based on its dominant existence over other forms in human aggrecan preparations isolated from knee cartilages [231].

Hyaluronan

Hyaluronan is a particularly interesting glycosaminoglycan relative to this study because of two main reasons, first being the fact that chito-oligosaccharides are precursors to hyaluronan synthesis *in vivo* [232-234], and secondly it is seen that hyaluronan is also up-regulated in similar malignancies just like YKL-40 which we discuss the details ahead in this chapter. The structural relationship of these two molecules is such that binding mechanisms were expected to be similar at alternating binding sites. Hyaluronan is a polysaccharide of a repeating β -D-glucuronic acid and N-acetyl- β -D-glucosamine disaccharides connected by alternating β -1,3 and β -1,4 glycosidic linkages (Figure 5.1). As with heparan sulfate, hyaluronan is also a glycosaminoglycan comprising the extracellular matrix and plays critical role in stabilization of cartilage matrix [235]. At extracellular pH, the carboxyl groups of glucuronic acid are fully deprotonated giving the ligand an overall negative charge under typical physiological conditions [236, 237].

5.3 Methods

5.3.1 Molecular Dynamics Simulation

MD simulations were constructed starting from the chitohexaose-bound YKL-40 structure deposited by Houston et al. (PDB ID 1HJW) [43]. The apo simulation simply removed the chito-oligomer from the primary binding cleft. As crystal structures of YKL-40 bound to other polysaccharides are not available, we used the structural similarity of polysaccharides as the basis for modeling the remaining polysaccharides in this investigation. In the case of cellohexaose, hyaluronan, heparan sulfate, heparin, and chondroitin sulfate, we located the central ring atoms of the ligand backbone in the same

location as that of the original chitohexaose. Appropriate pyranose side chains and glycosidic linkages (Figure 5.1) were added using CHARMM internal coordinate tables to construct the remainder of the sugar residue [166]. All polysaccharides were described using the CHARMM36 carbohydrate force field [193-195]. The missing force-field parameters for N-sulfated glucosamine (GlcNS) in heparin were developed using the Force-field Toolkit (ffTK) plugin for VMD [170, 238]. More details of this force-field parameterization are provided in Appendix 2.

Protonation states of all the titratable residues were determined according to the corresponding pKa values calculated by the H++ web server [189]. The protein, structural waters, and ligands were constructed in a vacuum using CHARMM [166]. The system was minimized for 1000 steps in vacuum using the Steepest Descent (SD) algorithm followed by another 1000 steps of minimization with the adopted basis Newton-Raphson (ABNR) algorithm. This procedure reduces the number of bad contacts prior to solvation of the solute. The polysaccharide systems were solvated in $100 \text{ \AA} \times 100 \text{ \AA} \times 100 \text{ \AA}$ cubic boxes. Sodium or chloride ions were added to the solution to ensure overall charge neutrality. For neutral ligands, six chloride ions were required to neutralize the charge of YKL-40 titratable residues. The charged ligands, hyaluronan (-3), heparan sulfate (-12), and chondroitin sulfate (-9), required 3 chloride ions, 6 sodium ions, and 3 sodium ions for charge neutrality, respectively. After solvation, the systems were minimized again in the following sequence: 1000 steps of SD with the protein and ligand restrained, 1000 steps of SD with only the protein restrained, and 2000 steps of SD and 2000 steps of ABNR with no harmonic restraints. Extensive minimization, up to 10000 steps of SD, was carried out for systems bound to highly sulfated polysaccharides and

collagen. The solvated and minimized systems were then equilibrated prior to production MD simulations. The systems were heated from 100 K to 300 K in 50-K increments over 20 ps in the canonical ensemble. The system density was then equilibrated in the *NPT* ensemble at 300 K and 1 atm (101325 Pa) for 100 ps. The Nosé-Hoover thermostat and barostat were used to control temperature and pressure in CHARMM [212, 213].

Production MD simulations of 250 ns were performed in the canonical ensemble at 300 K using NAMD [169]. Temperature was controlled using Langevin thermostat [214]. The SHAKE algorithm was used to fix the bond distances to all hydrogen atoms [215]. Non-bonded interactions were truncated with a cutoff distance of 10 Å, a switching distance of 9 Å, and a non-bonded pair list distance of 12 Å. Long range electrostatics were described using the Particle Mesh Ewald method with a 6th order b-spline, a Gaussian distribution width of 0.320 Å, and a 1 Å grid spacing [216]. The velocity Verlet multiple time-stepping integration scheme was used to evaluate non-bonded interactions every 1 time step, electrostatics every 3 time steps, and 6 time steps between atom reassignments. All simulations used a 2-fs time step. The CHARMM36 force field with the CMAP correction [166, 191, 192] was used to describe YKL-40. The polysaccharides were described using the CHARMM36 carbohydrate force field [193-195]. Water was modeled using the TIP3P force field [196, 197]. All simulations used explicit solvent.

A complete list of simulations and calculations performed to meet the objectives of this study is given in Table 5.1. The length of each MD simulation is also given, as not all simulation lengths were the same; several of the hypothesized ligands dissociated

from the binding cleft, and the simulation was halted to conserve computational resources. The free energy calculations performed are also indicated. If a ligand did not remain in the binding cleft throughout the entirety of the MD simulation, a free energy calculation was not performed.

In addition to these protein-carbohydrate complexes, oligo-saccharides were solvated in water separately, without YKL-40. These ligand-only simulations were required as input to the free energy calculations. Several additional system configurations beyond those originally proposed were also developed, as described below, in order to study the effect of ligand position on conformational changes and to understand the statistical significance of observed interactions with the putative heparin-binding subsite.

Table 5.1 Simulations and calculations performed in the investigation of the binding of polysaccharides ligands to YKL-40.

Case No.	System	MD simulation	Free Energy Calculation
1	Apo YKL-40	250 ns	--
2	YKL-40 + chitohexaose	250 ns	FEP/ λ -REMD
3	YKL-40 + cellohexaose	250 ns	FEP/ λ -REMD
4	YKL-40 + hyaluronan	250 ns	FEP/ λ -REMD
5 ^a	YKL-40 + heparin (fully sulfated)	50 ns	--
6	YKL-40 + heparan sulfate (unsulfated)	50 ns	--
7	YKL-40 + chondroitin sulfate	50 ns	--

^a Four YKL-40 + heparin systems were constructed: two with heparin initially in the primary polysaccharide binding cleft and two with heparin initially located in bulk solution.

5.3.2 Free Energy Calculations: FEP/ λ -REMD

Free energy perturbation with Hamiltonian replica-exchange molecular dynamics (FEP/ λ -REMD) was used to calculate the absolute free energy of binding the polysaccharide ligands to YKL-40 [174, 175]. This protocol uses Hamiltonian replica-exchange as a means of improving the Boltzmann sampling of free energy perturbation calculations. The parallel/parallel replica exchange MD algorithm in NAMD was implemented as recently described [113, 169]. The free energy calculations performed

using this approach were accomplished through two separate sets of free energy calculations following the thermodynamic cycle illustrated in Figure 5.2. To obtain each binding free energy, ΔG , the bound carbohydrate ligand was first decoupled from the solvated protein-carbohydrate complex to determine ΔG_1 . The second calculation entailed decoupling the solvated oligosaccharide from solution into vacuum to obtain ΔG_2 . The difference between the two values, $\Delta G_2 - \Delta G_1$, gives the absolute free energy of binding the given ligand to YKL-40.

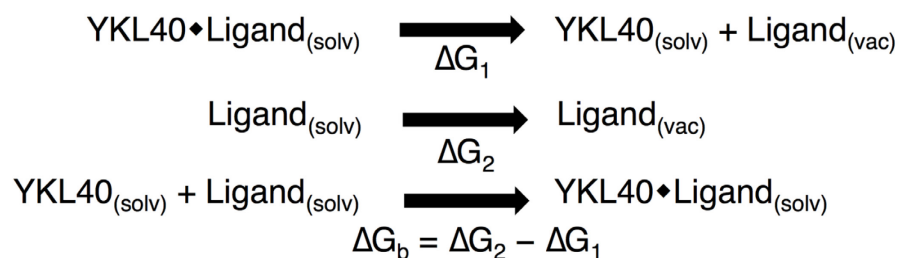


Figure 5.2 Thermodynamic cycle used to determine ΔG with FEP/ λ -REMD method. ‘solv’ refers to the solvated state and ‘vac’ refers to the gas-phase state.

In each free energy calculation, five separate terms contribute to the potential energy of the system: the non-interacting ligand potential energy, repulsive and dispersive contributions to the Lennard-Jones potential, electrostatic contributions, and the restraining potential. In each calculation, the ligand was decoupled from the system by thermodynamic coupling parameters controlling the non-bonded interaction of the ligand with the environment. The parameters decoupled the ligand in a four-stage process, wherein the coupling parameters defined replicas that were exchanged along the length of the alchemical pathway. This decoupling, as reported shortly ahead, has been described in detail previously [113]. A total of 128 FEP replicas were used (72

dispersive, 24 repulsive, and 32 electrostatic), and a conventional Metropolis Monte Carlo exchange criterion governed the swaps throughout the replica exchange process [175]. The free energy of binding was determined from 20 consecutive, 0.1-ns simulations of each corresponding system, where the first 1 ns of data was discarded as equilibration. The oligosaccharide ligands were restrained in the ligand-binding pose using a harmonic restraint on the distance between the center of mass of the protein and the center of mass of the ligand. The harmonic restraint force constant was 10 kcal/mol/Å². This restraint bias was removed from the free energy calculation according to the approach outlined by Deng and Roux [173]. Multistate Bennett Acceptance Ratio (MBAR) was used to determine electrostatic, repulsive, and dispersive contributions to free energy [176]. Standard deviation of the final 1 ns free energy values serves as the error estimate. All simulation parameters in the free energy calculations mimic those described in the *MD simulations* section. The progress towards the convergence of free energy calculations for cellohexaose, chitohexaose and hyaluronan systems are shown in Figure 5.3.

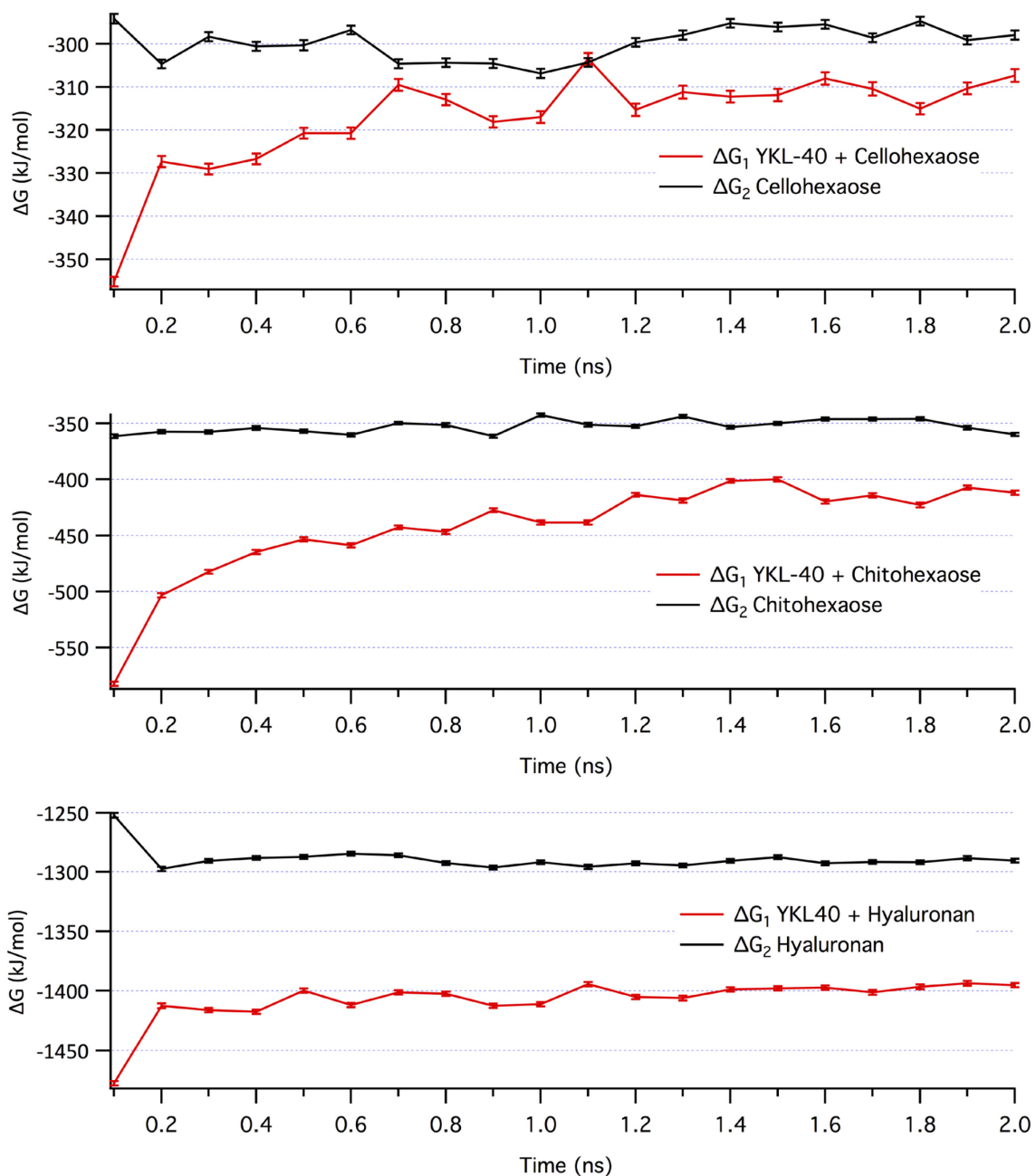


Figure 5.3 Convergence of ΔG over 20 consecutive 0.1-ns free energy perturbation calculations using the FEP/ λ -REMD method.

5.4 Results and Discussion

5.4.1 Protein-polysaccharide binding in YKL-40

MD simulation suggests that of the six polysaccharide oligomers investigated, only three bind in a stable fashion in the primary carbohydrate binding site of YKL-40. The three potential polysaccharide physiological ligands at this site include chitohexaose, cellohexaose, and hyaluronan. In the section that follows, we will describe the dynamics of chitohexaose, cellohexaose, and hyaluronan binding to YKL-40. The remaining three ligands – heparin, heparan sulfate, and chondroitin sulfate – were dislodged from the binding site over the course of MD simulations. The α -1,4 glycosidic linkages in heparin and heparan sulfate, instead of β -1,4, modifies the relative orientation of disaccharide monomers from that of the chito-oligosaccharide. The NMR solution structure of heparin (PDB ID 1HPN) shows that the relaxed conformation is semi-helical [239], which cannot be feasibly accommodated in the conserved, narrow carbohydrate binding site of YKL-40. Heparan sulfate suffers from similar steric constraints posed by the relaxation driving force. The bulky sulfated side chains of heparin introduce further steric hindrance, and in the case of heparin and chondroitin sulfate, unfavorably strong electrostatic interactions resulting from negatively charged moieties inconveniently located along the cleft (i.e., without a co-located, oppositely charged protein residue) eject the ligands from the cleft.

In the cases of heparin, heparan sulfate, and chondroitin sulfate, the ligands quickly “relax” from the initial wide “V-shape” conformation as they are expelled from the cleft by charge- and steric-based effects. Relaxation of the sugar from the initial binding pose is sufficient to initiate loss of critical non-bonded interactions along with a subsequent reduction in affinity (Figure 5.4). Within 25 ns, heparin, heparan sulfate, and

chondroitin sulfate were expelled from the cleft into bulk solution. Each of the three ligands capable of binding with the primary binding cleft maintained the -1 boat conformation over the entire simulation. Chitohexaose and cellohexaose remained in the binding cleft over the entire 250-ns MD simulation, while maintaining the initial wide “V-shape.” Hyaluronan developed a sharp “V-shape” within a few nanoseconds and maintained this conformation within the binding cleft for the remainder of the simulation (Figure 5.5); this is primarily due to variation in glycosidic linkage, where hyaluronan exhibits a β -1,3 linkage within the monomer instead of the β -1,4 linkage of cello- and chitohexaose. Also, comparison of the equilibrated chitohexaose- and hyaluronan-bound structures disabuses one of the notion that similar binding mechanisms exist at alternate binding sites, as only the -1 site pyranose appears to maintain similar sidechain orientation.

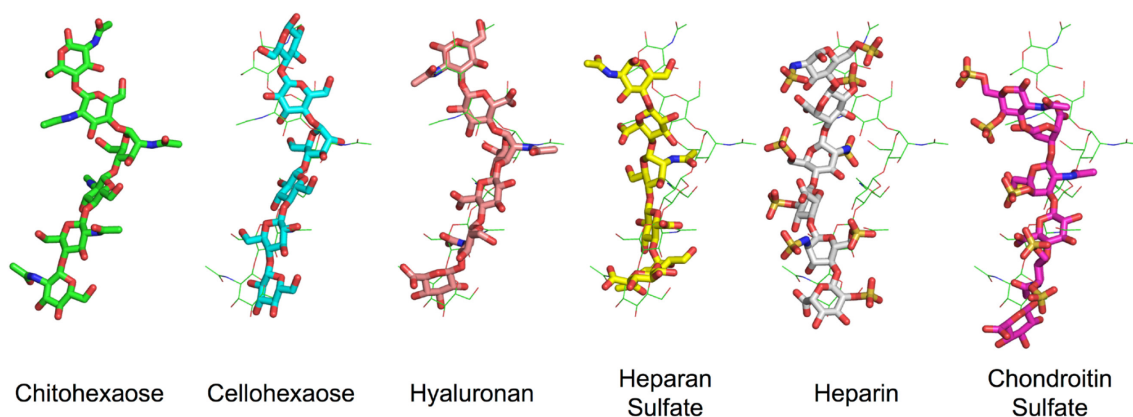


Figure 5.4 Relaxation of the polysaccharide ligands in the primary binding cleft of YKL-40. Each ligand is shown after a 100-ps equilibration in a thick stick representation. For comparison, the chito-oligomer, in its equilibrated conformation, is shown in thin green lines behind each oligosaccharide. The YKL-40 protein has been aligned such that each

oligosaccharide is oriented in the same manner; though, YKL-40 is not shown for visual clarity. Heparan sulfate, heparin, and chondroitin sulfate relax significantly from the initial distorted conformation.

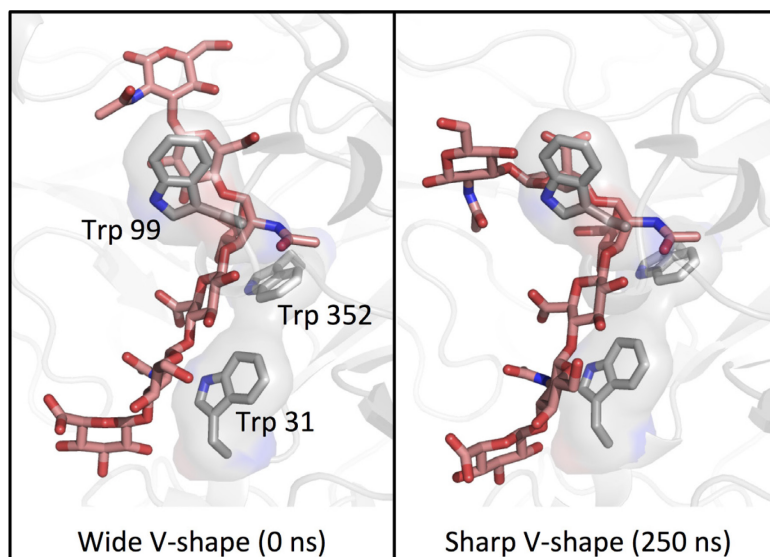


Figure 5.5 Hyaluronan in YKL-40 binding site at 0 ns (left) and at 250 ns (right) illustrating difference between V-shape conformations of hyaluronan.

The native distorted conformation is characteristic of glycoside hydrolase pyranose binding behavior in the -1 site (Figure 5.4) [80]. In solution, polysaccharide pyranose moieties adopt the energetically favorable chair conformation [240]; however when bound to an enzyme, the active sites of catalytically-active glycoside hydrolases distort the pyranose ring in the -1 binding subsite into a less energetically favorable conformation, such as a boat or skew conformation [241-244], priming the substrate for hydrolytic cleavage. Interestingly, the chitohexaose ligand bound in the primary binding site of YKL-40 exhibits a boat conformation despite not being catalytically active [43]. This suggests that the sugar distortion in the -1 binding site contributes to ligand binding

as well catalysis, as there is no evolutionary requirement to overcome an activation energy barrier in a catalytically-inactive lectins. A recent study of a homologous chitinase suggested that -1 pyranose relaxation reduces binding affinity and affords the ligand more flexibility and entropic freedom [245], which is consistent with our findings from the 250-ns MD here.

5.4.2 Putative heparin-binding site

Despite the fact that the heparin oligomer could not be accommodated by the YKL-40 binding cleft, MD simulations do suggest that the oligomer interacts with the surface of YKL-40 at a putative heparin-binding site (Figure 1.6B). After ejection from the primary binding site, the oligomer spontaneously binds to the YKL-40 heparin-binding site (Movie 6.1). To address the significance of this unanticipated event, we performed three additional independent MD simulations of the YKL-40/heparin system: one with a new random number seed, though in the same configuration, and two additional simulations with the ligand randomly placed in solution (Movie 6.2). In each case, the heparin oligomers were capable of finding and binding to a group of charged residues at the surface of YKL-40 (Figure 5.6); these were the basic residues of a putative heparin-binding site, GRRDKQH, at position 143-149. Interestingly, this domain follows a consensus protein sequence – XBBXBX, where B is a basic residue and X is any non-basic amino acid – that is noted for its ability to recognize polyanions like heparin [246]. In all four cases, heparin recognized the binding site within 25 ns of MD simulation (Figure 5.6), occasionally visiting other moderately basic surface locations before localizing around the GRRDKQH motif. The strong electrostatic interaction arose from the dynamic formation of salt-bridges between either the sulfate or the carboxyl groups of

the heparin oligosaccharide and the side chains of the basic amino acids. Coupled with experimental observation of heparin affinity, our MD simulations suggest a non-specific, surface-mediated binding interaction between YKL-40 and the extensively sulfated heparin oligomer [43, 44]. While the unsulfated variant, heparan sulfate, did not visit the heparin-binding site, chondroitin sulfate also attached to the putative heparin-binding site in a similar fashion to heparin. Given the chemical similarity of these glycosaminoglycans, i.e., highly sulfated and negatively charged, we anticipate the XBBXB motif may also routinely appear in chondroitin-binding proteins.

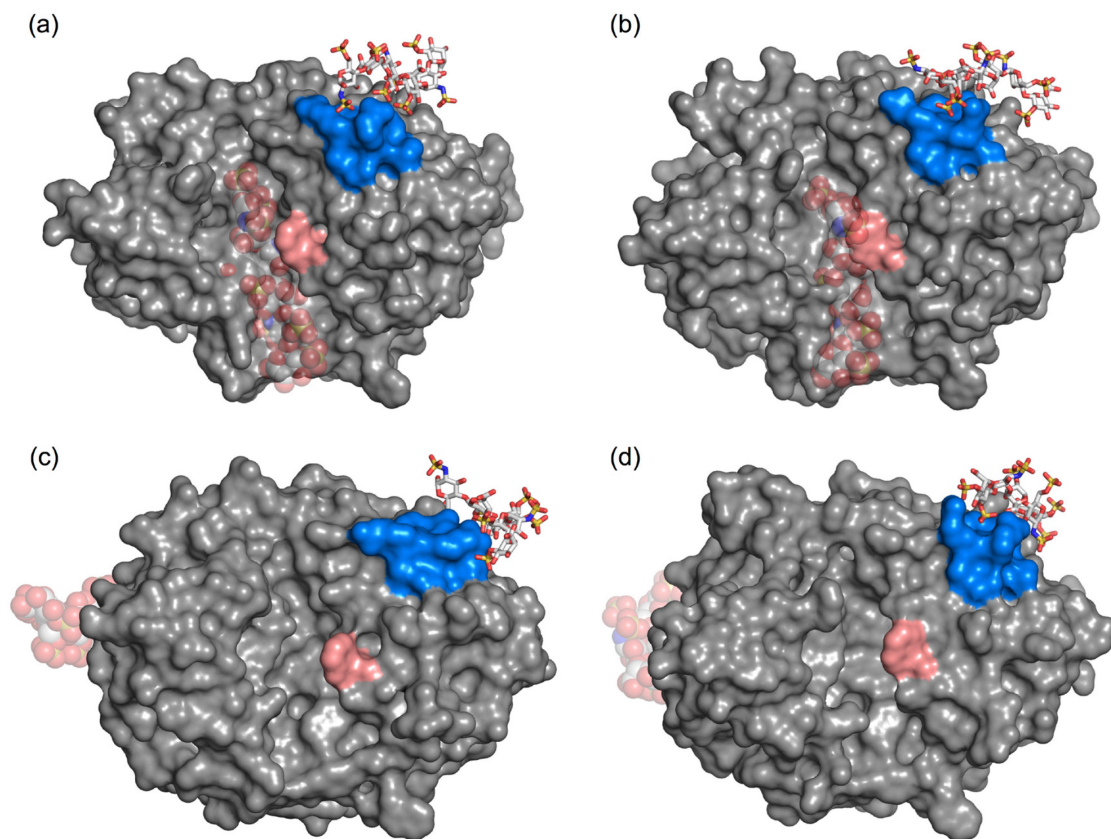


Figure 5.6 Snapshots from four independent MD simulations of heparin (white stick) binding to a putative heparin-binding site (blue surface) of YKL-40 (gray surface). The primary oligosaccharide binding site of YKL-40 is marked by an aromatic residue shown in salmon surface representation. Transparent spheres illustrate the initial simulation positions of heparin. In two cases, (a) and (b), the heparin ligand was initially bound in the primary YKL-40 binding site. In both cases, the ligand was expelled from the primary binding site into solution and located the heparin-binding site through electrostatic interactions. Two additional simulations, (c) and (d), were initialized with the heparin ligand free in solution.

5.4.3 Polysaccharide ligand binding affinity

Each of the three polysaccharides maintaining contact with the primary binding site of YKL-40, cellohexaose (or likely any glucose derivative), chitohexaose, and hyaluronan, are feasible ligands. However, free energy calculations suggest that hyaluronan may preferentially bind with YKL-40 when chitin is not indicated as a foreign entity. The absolute free energies of binding cellohexaose, chitohexaose, and hyaluronan to YKL-40 were -3.01 ± 0.88 , -15.14 ± 3.01 , and -25.54 ± 1.10 kcal/mol, respectively. Repulsive, dispersive, and electrostatic components of the free energy changes are tabulated in Table 5.2.

Table 5.2 Energetic components of the free energy of ligand binding to YKL-40. All values are in kcal/mol.

System	ΔG_{repu}	ΔG_{disp}	ΔG_{elec}	ΔG_{rstr}	ΔG_{Tot}	ΔG_{b}
YKL-40 + Cellohexaose	90.75 ± 0.91	-91.42 ± 0.5	-73.31 ± 0.53	-0.26	-74.25 ± 0.86	-3.01 ± 0.88
Cellohexaose	75.61 ± 0.43	-68.72 ± 0.31	-78.12 ± 0.38	0	-71.23 ± 0.67	
YKL-40 + Chitohexaose	128.14 ± 2.8	-128.86 ± 1.00	-97.49 ± 1.79	-1.32	-99.52 ± 2.68	-15.14 ± 3.01
Chitohexaose ^a	78.81 ± 1.08	-73.22 ± 0.72	-89.98 ± 0.81	0	-84.39 ± 1.41	
YKL-40 + Hyaluronan	104.76 ± 0.67	-105.09 ± 0.36	-333.79 ± 1.0	-0.31	-334.43 ± 1.03	-25.54 ± 1.10
Hyaluronan	79.65 ± 0.33	-73.17 ± 0.31	-315.38 ± 0.36	0	-308.9 ± 0.6	

^a Hamre, Jana [217]

The free energy of solvation for chitohexaose was previously calculated by our group as part of a study on family 18 chitinases [217]; this value has been used in our calculation of chitohexaose binding affinity to YKL-40 for computational efficiency. The methods used to calculate solvation free energy of chitohexaose were identical to those described here. Furthermore, the binding free energy of chitohexaose to YKL-40 is in good agreement with that of homologous family 18 chitinases, despite mutation of the catalytic motif in the lectin.

Chitohexaose and cellohexaose are both neutral ligands but display a significant difference in binding affinity to YKL-40. Electrostatic interactions appear to be one of the more significant contributors to the enhanced affinity of chitohexaose over cellohexaose (Table 5.2). For cellohexaose, the change in the electrostatic component of binding free energy was unfavorable (4.81 ± 0.65 kcal/mol), while the same component for chitohexaose was energetically favorable (-7.51 ± 1.96 kcal/mol). In the case of hyaluronan, electrostatic interactions play an even greater role in enhancing affinity of the ligand for YKL-40 (-18.41 ± 1.08 kcal/mol). This increasing electrostatic contribution is reflective of increasing number of electronegative atoms in the sidechains of carbohydrates as we go from cellopentaose to chitohexaose to hyaluronan. We observe no significant differences in cellohexaose, chitohexaose, or hyaluronan binding to YKL-40 arising from Weeks-Chandler-Anderson (WCA) dispersion and repulsion (Table 5.2). This is largely a function of the molecular similarity of the pyranose rings comprising the monomeric units of three oligosaccharides (Figure 5.1). The pyranose rings of carbohydrates bound in the active sites of glycoside hydrolases, and by extension, the binding clefts of lectins, form carbohydrate- π stacking interactions with surrounding

aromatic residues along the clefts [247]. In YKL-40, these stacking interactions are formed in the -3 and -1 binding sites with residues Trp31 and Trp352, respectively. Naturally, any polysaccharide ligand capable of binding in the YKL-40 binding cleft will likely exhibit a similarly favorable WCA binding free energy component. In the following section, we expand upon the molecular-level interactions that contribute to polysaccharide binding affinity in YKL-40.

Based on these results, it is unlikely that a cello-oligomer would bind in the cleft of YKL-40 over a chito-oligomer, and thus, while there is potential for YKL-40 to bind a cello-oligomer or glucose, it would not be inhibitory. Hyaluronan, on the other hand, likely competes with chito-oligomers in binding, which is due in large part to the electrostatic favorability of hyaluronan's charged side chains in the YKL-40 binding cleft. Clinical data supports hyaluronan as a biomarker for cancer prognosis and inflammation [236, 248], the same events in which YKL-40 appears at elevated serum levels [42]. To our knowledge, there are no studies evaluating the coexistence of YKL-40 and hyaluronan. The cell receptor protein CD44 has been implicated in hyaluronan binding interactions and is also involved in confounding scenarios, both aggravating and improving inflammation [249]. Sequence alignment of YKL-40 with the hyaluronan-binding domain of human CD44 [250], using BLASTP 2.3.0 [251], shows no homology, further suggesting that this YKL-40-hyaluronan binding is different from previously known hyaluronan-binding proteins [252].

5.4.4 Polysaccharide Binding Dynamics

YKL-40 is highly homologous with carbohydrate-active enzymes found in glycoside hydrolase family 18 [74, 253]. Despite lacking catalytic ability, the primary polysaccharide binding site of YKL-40 exhibits remarkable similarity to these family 18 chitinases. As such, one may reasonably expect that ligand binding within this family will demonstrate similar trends, regardless of evolutionary origin. Indeed, we observe that chitohexaose, cellohexaose, and hyaluronan binding in the primary binding site of YKL-40 follow a general pattern common to carbohydrate-active enzymes. Namely, that ligand binding interactions are mediated by carbohydrate- π stacking interactions with aromatic residues, and hydrogen bonding interactions are critical to overall ligand affinity and stability. We investigate these trends quantitatively through analysis of the MD simulation trajectories, including root-mean-square deviation (RMSD) of the protein, root-mean-square fluctuation (RMSF) of both the protein and the ligands over the course of the simulation, hydrogen bonding analysis, degree of solvation of the ligand, and interaction energy of the ligand with the protein.

Cellohexaose, chitohexaose, and hyaluronan binding in the primary YKL-40 binding site did not adversely affect protein dynamics. In each case, binding the polysaccharide ligand did not significantly disturb the protein backbone (i.e., protein fold), and the ligand remained relatively unperturbed over the course of the simulation. The RMSD of the protein (Figure 5.7a) is a measure of deviation over the course of the simulation from the initial configuration, which was the first frame of the simulation following *NPT* density equilibration. The relatively consistent RMSD of the protein backbones suggests the simulations reached a local equilibrium. The magnitude of the

RMSD change over 250 ns is small given the significant chemical differences in the three ligands examined, which indicates the primary YKL-40 binding site is forgiving of small charged side chains such as the carboxylate of hyaluronan. The RMSF fluctuation of the protein backbone similarly describes fluctuation of a given protein residue from the average position over the course of the entire simulation. As with the RMSD calculation, the RMSF of the protein backbone suggests the binding of chitohexaose and cellohexaose does little to disturb the overall protein conformation (Figure 5.7b). In the case of hyaluronan binding, we observe increased fluctuation in residues 178-189, 225-235, and 300-325 over that of cellohexaose and chitohexaose bound YKL-40. Both loops 225-235 and 300-325 are located away from the primary carbohydrate-binding site; the increase in flexibility in these loops appears to be related to solvent exposed polar residues sampling bulk solution and is likely unrelated to hyaluronan binding. Segment 178-189, comprising part of a β -sheet and a small α -helix just beneath the +1 and +2 binding sites, becomes increasingly mobile as its interaction with hyaluronan is lost in the formation of the sharp V-shape. Despite localized increases in backbone flexibility, the overall protein structure largely remains in the same initial conformation, as evidenced by the similarity in RMSD (Figure 5.7a).

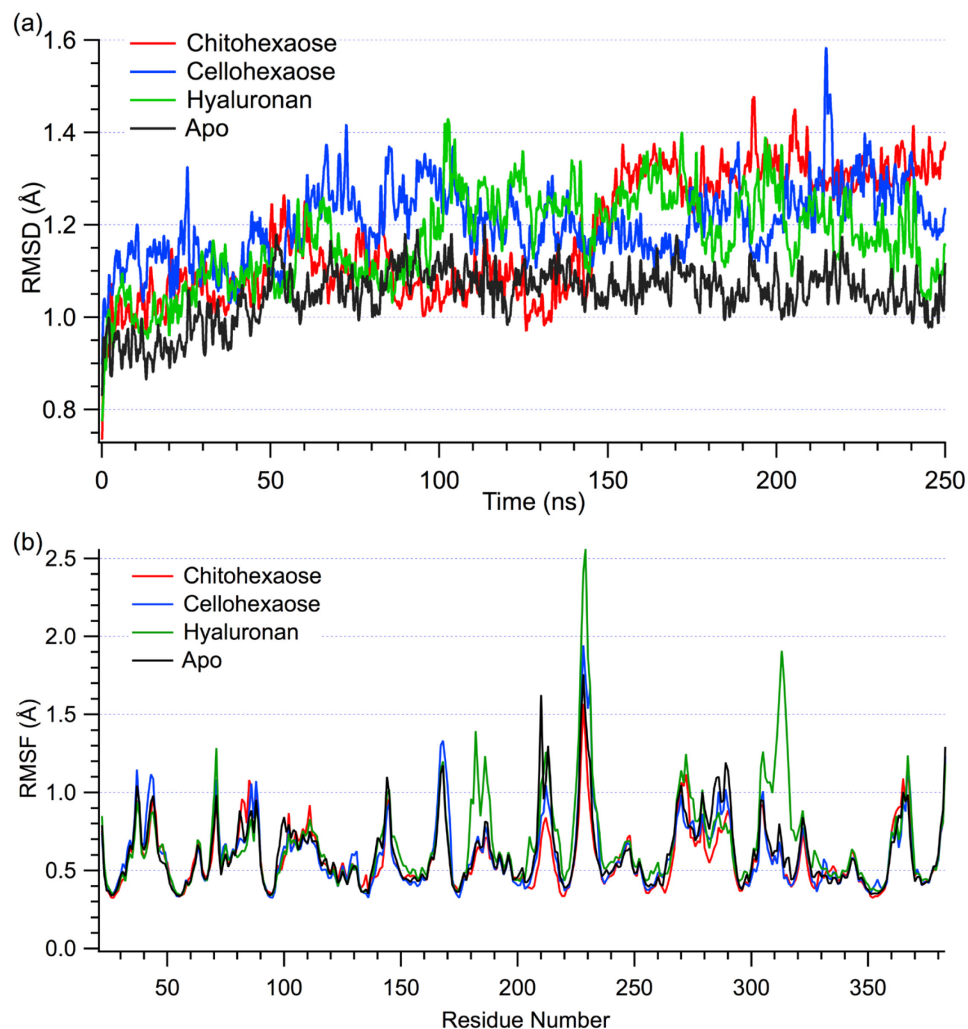


Figure 5.7 (a) Root-mean-square deviation over 250-ns MD simulations and (b) root-mean-square fluctuation of YKL-40 without a ligand (apo) and bound to chitohexaose, cellohexaose, and hyaluronan. Binding of chitohexaose, cellohexaose, and hyaluronan do not significantly alter the dynamics of YKL-40.

The RMSF of the ligand, averaged over 250 ns as a function of binding site, provides a measure of relative ligand stability. Error was estimated by block averaging over 2.5 ns blocks. Ligand stability over the course of the simulation suggests hyaluronan

is as stable, if not more so, as chitohexaose in the primary binding site (Figure 5.8a). Although, cellohexaose appears to be more stable relative to the two other ligands at the ends of the binding cleft. This latter finding is a function of the length of the side chains attached to the pyranose rings of each of the ligands. Of the three carbohydrates, the cello-oligomer has the shortest side chains, and thus, the ligand fluctuates less, as it does not need to rearrange as significantly to induce binding. As shown above, this does not necessarily correspond to the most thermodynamically preferential ligand and lower RMSF could also be interpreted hypothetically as loss of translational and conformational freedom, resulting in unfavorable entropic contribution.

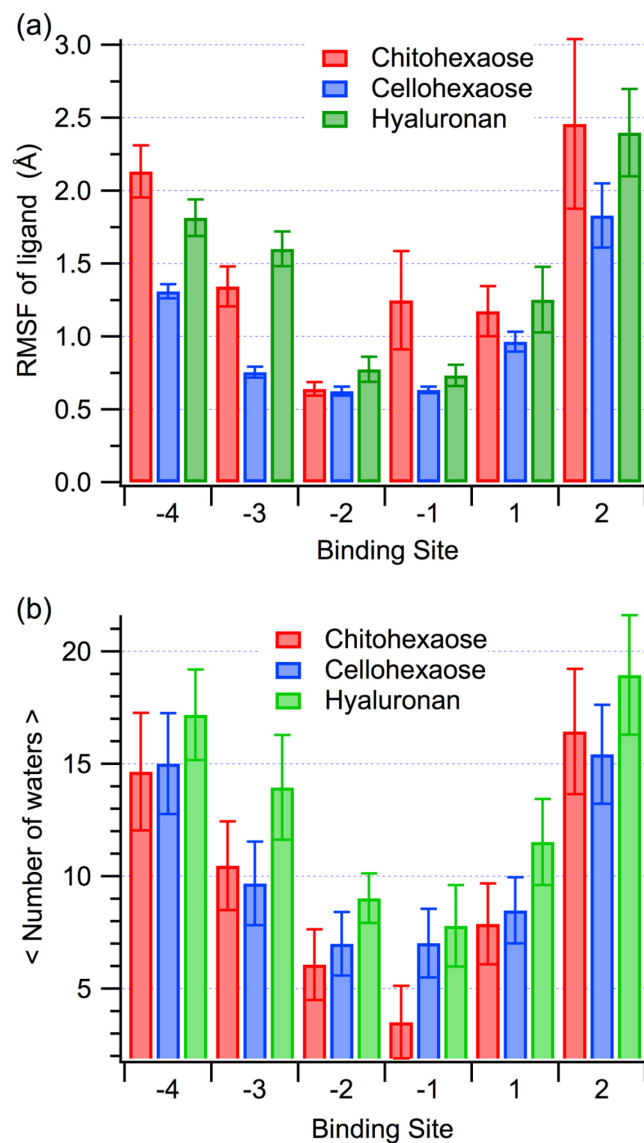


Figure 5.8 (a) RMSF of the polysaccharide ligands on a per-binding-subsite basis. Error bars were calculated using block averaging over 2.5 ns. (b) Average number of water molecules within 3.5 Å of each ligand monomer. Error bars represent one standard deviation.

The hydrogen-bonding partners of chitohexaose, cellohexaose, and hyaluronan are quite different, largely as a result of the conformational change of hyaluronan (Table 5.3). Defining a hydrogen bond as a polar atom within 3.4 Å and 60° of a donor, we

identified the formation of donor-acceptor pairs and percent occupancy of these hydrogen bonds between the protein and each carbohydrate moiety over the course of the 250-ns MD simulations (Table 5.3). As described above, hyaluronan formed a sharp “V-shape” in the polysaccharide binding cleft, which minimized steric hindrance and in turn, modified accessible hydrogen bonding partners relative to chito- and cellohexaose. Hydrogen bonds at the +1 site, between glucuronic acid and Asp207, Arg263, and Tyr141, stabilized the hyaluronan conformation (Table 5.3). In the -1 subsite, chitohexaose primarily hydrogen bonds with the side chain of Tyr206 and the main chain of Trp99. In the cases of both cellohexaose and hyaluronan, the interaction with Tyr206 was abolished, and instead, supplemented by Trp99 alone. In the -2 subsite, the oxygen of the chitohexaose acetyl forms a long-lived hydrogen bond with the indole nitrogen of the buried Trp352; neither hyaluronan nor cellohexaose interact with the -2 site through this tryptophan. Rather, Trp31, which stacks with the pyranose in the -3 subsite, acts as a hydrogen donor to the -2 subsite glucuronic acid side chain of hyaluronan. In the case of cellohexaose, the main chain of a solvent-exposed asparagine, Asn100, almost exclusively mediates hydrogen bonding in the -2 site. The +2, -3, and -4 binding subsites exhibit less frequent hydrogen bonding between the ligand and the protein, and there is little consistency in bonding partners across ligands. Certainly, these variations will manifest in enthalpic contributions to ligand binding, as even a single hydrogen bond may account for 1 – 7 kcal/mol of binding free energy in biological systems [219, 220]; such is likely the case for cellohexaose and chitohexaose binding to YKL-40, where the latter exhibits both greater hydrogen bonding capability and a more favorable binding free energy.

Table 5.3 Hydrogen bonding pairs from polysaccharide-bound molecular dynamics simulations. A hydrogen bond was defined as a polar atom having a donor-acceptor distance of 3.4 Å and a 60° cutoff angle. Occupancy refers to the percent of the simulation during which the hydrogen bond was formed. Occupancies less than 10% have not been reported unless relevant in comparison.

Binding Site	Cellohexaose			Chitohexaose			Hyaluronan		
	Donor	Acceptor	Occupancy	Donor	Acceptor	Occupancy	Donor	Acceptor	Occupancy
-4	BGLC1-SC	GLU70-SC	56.28%	NAG1-SC	GLU36-SC	9.32%	LYS289-SC	GCU1-SC	13.40%
				LYS289-SC	NAG1-MC	8.48%	GCU1-SC	TRP31-MC	12.55%
-3	TRP69-SC	BGLC2-SC	53.32%	NAG2-SC	GLU290-SC	69.52%	ASN100-SC	NAG1-SC	35.71%
	BGLC2-SC	GLU70-SC	34.76%	ASN100-SC	NAG2-MC	67.68%	TRP69-SC	NAG1-SC	9.76%
	ASN100-SC	BGLC2-SC	21.80%						
-2	ASN100-MC	BGLC3-SC	87.40%	TRP352-SC	NAG3-MC	93.24%	TRP31-SC	GCU2-SC	41.46%
				ASN100-MC	NAG3-SC	66.00%	ASN100-MC	GCU2-SC	22.27%
				NAG3-SC	GLU290-SC	30.32%	TRP99-MC	GCU2-SC	16.03%
				NAG3-SC	ASN100-SC	13.44%	ASN100-SC	GCU2-SC	13.60%

-1	TRP99-MC	BGLC4-SC	76.20%	TYR206-SC	NAG4-MC	75.16%	TRP99-MC	NAG2-MC	86.76%
				TRP99-MC	NAG4-SC	39.56%			
				TYR206-SC	NAG4-SC	16.52%			
				NAG4-SC	ASP207-SC	15.16%			
+1	BGLC5-SC	TYR141-SC	32.32%	NAG5-MC	ASP207-SC	74.08%	GCU3-SC	ASP207-SC	96.19%
	BGLC5-SC	ASP207-SC	18.08%	NAG5-SC	TYR141-SC	17.00%	ARG263-SC	GCU3-SC	76.88%
	TYR141-SC	BGLC5-SC	13.52%	TYR141-SC	NAG5-SC	15.04%	TYR141-SC	GCU3-SC	62.75%
				ARG263-SC	NAG5-MC	14.28%			
+2	TYR141-SC	BGLC6-SC	52.04%	NAG6-MC	TYR141-SC	45.68%	TRP99-SC	NAG3-SC	48.14%
				TYR141-SC	NAG6-SC	18.88%			
				TRP99-SC	NAG6-SC	10.28%			

SC – Side chain; MC – Main chain; BGLC – β -D-glucose; NAG – N-acetyl- α -D-glucosamine; GCU – β -D-glucuronic acid.

Key aromatic residues – Trp31, Trp99, and Trp352 – play a significant role in binding all three oligosaccharides. Notably, these tryptophans are conserved in other lectins, including mammary gland protein (MGP-40) and mammalian lectin Ym1 [131, 225]. According to previous structural studies, these aromatic residues form hydrophobic stacking interactions with pyranose moieties at the -3, +1, and -1 binding subsites, respectively [44]. As mentioned above, this carbohydrate- π stacking was observed across the three polysaccharide ligands as a result of the chemically similar carbohydrate “backbone” of pyranose rings. However, at the +1 site of hyaluronan, the stacking interaction with Trp99 was not maintained. Instead, prominent hydrogen bonding forces the +1 pyranose ring in an orientation that is perpendicular to aromatic Trp99 (Figure 5.5). Nevertheless, the similarity in WCA contributions to binding free energy for all three polysaccharides suggests this +1 site stacking interaction weakly contributes to the overall binding free energy.

The degree to which the binding cleft of YKL-40 was accessible to water molecules did not change significantly with the bound polysaccharide. The degree of ligand solvation was determined by calculating the average number of water molecules within 3.5 Å of a given pyranose ring over the course of the simulations (Figure 5.8b); error is given as one standard deviation. Chitohexaose and cellohexaose display similar degrees of solvation across the length of the cleft. Hyaluronan allows a moderate increase in degree of solvation of the cleft by comparison to chitohexaose, across the -3 and +1 subsites, where its sharp V-shape again contributes to variation in behavior. Given the similarity in solvent accessibility within the binding cleft regardless of ligand, it is

unlikely that entropic contributions from solvation play a role in the observed differences in ligand binding free energy.

5.4.5 Conformational changes in the YKL-40 binding site

Crystal structures of YKL-40 bound with chito-oligosaccharides suggest that YKL-40 undergoes a conformational change upon chitin ligand binding [43], contrary to suggestions that lectin binding sites, in general, are “pre-formed” to accommodate their natural substrates and undergo little change upon sugar binding [13]. Houston et al. reported that the residues forming a loop (residues 209 to 213) near the primary YKL-40 binding cleft occupy an unusual conformation in apo YKL-40 when compared to the ligand bound YKL-40 structure, where Trp 212 lines the +2 and +3 subsites [43]. However, a second structural investigation published concurrently did not observe a similar conformation change in either of two crystal structures (1NWR and 1NWS), where no ligand occupied either the +2 or +3 subsites [44]. Additionally, the positioning of Trp99 at the +1 site in both apo structures of human YKL-40 (1HJX and 1NWR) and the homologous MGP-40 (1LJY) differs from that of holo-YKL-40 and homologous mammalian lectin Ym1 (1E9L) [43, 44, 131, 225], with the tryptophan blocking the binding cleft in the apo form. This conformational variation as a function of binding site occupancy has been proposed as a tryptophan-mediated gating mechanism for ligand binding in chitolectins [224].

Based on MD simulations we did not observe data suggesting binding cleft rearrangement is important in polysaccharide binding to YKL-40. To investigate possible loop rearrangement upon ligand unbinding, the apo YKL-40 simulation was prepared by

undocking the bound chitin oligomer. One can reasonably expect that over the course of a 250-ns MD simulation, the 5-amino acid residue loop would, at a minimum, sample a variety of conformations indicating flexibility in this region. However, in examining the trajectory of this loop with respect to its initial position, we did not observe the peptide loop returning to the unusual conformation even for a single frame (Figure 5.9). This suggests that the crystallographic apo conformation may have resulted from serendipitous crystal packing interactions and may not represent a typical conformational behavior. Additionally, the phenomenon of tryptophan mediated gating, according to which one would expect the Trp99 to return to the “pinched” conformation of the apo state, was not observed. Though, we note the likelihood of observing that the latter behavior, i.e., returning to a “pinched” conformation, in an unbiased MD simulation is low and may require overcoming an energy barrier through enhanced sampling approaches.

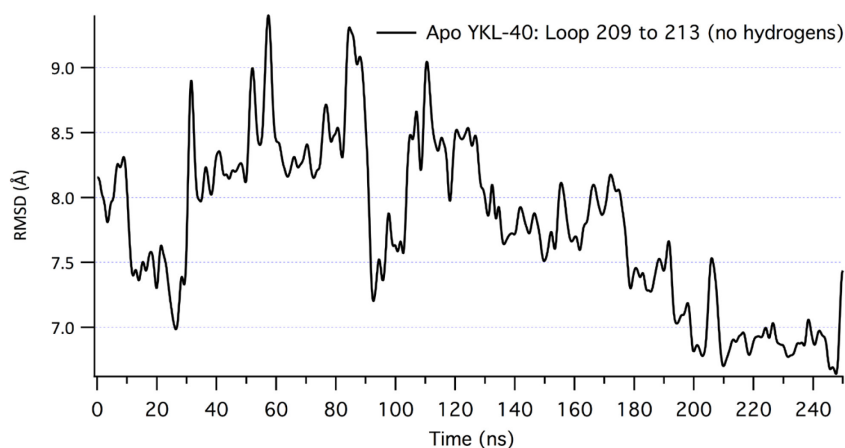


Figure 5.9. Root mean square deviation of loop of residues 209 to 213 from the unusual configuration in apo YKL-40 crystal structure during 250-ns MD simulation of apo YKL-40 prepared by removing the bound ligand from holo crystal structure.

5.5 Conclusions

We constructed polysaccharide-bound YKL-40 models to understand the molecular-level interactions of the protein with potential physiological ligands. MD simulations as well as free energy calculations overwhelmingly suggest polysaccharide ligands, in particular chito-oligomers and hyaluronan, are preferential physiological ligands of YKL-40. The ability of YKL-40 to bind the polysaccharide ligands is related to the ability of the carbohydrate ligands to adopt the primary binding cleft. These ligands are able to form longer-lived hydrogen bonds deeper in the hydrophobic interior of the protein. Additionally, electrostatic interactions play a key role in ligand recognition and affinity to YKL-40. Improper alignment of side chains of heparan sulfate with the residues lining the YKL-40 cleft, additional large and highly charged side chains of heparin and chondroitin sulfate prohibit these ligands from binding in the primary binding cleft. However, the smaller, negatively charged side chain of hyaluronan interacts favorably in the primary binding cleft and contributes significantly to the affinity of this molecule. Additionally, we confirmed the non-specific interaction of heparin with the putative heparin-binding domain suggested from previous structural studies. The charged side chains repeatedly and spontaneously interact with charged residues at this secondary surface-binding site. Based on this study, we suggest that hyaluronan is a preferential physiological ligand of YKL-40, which may explain the pervasive association of YKL-40 with the physical maladies in which hyaluronan has also been associated. These findings not only identify physiological ligands of YKL-40, they enable future efforts to rationally guide design of YKL-40 inhibitors, the design of which, is invaluable in the control of inflammation-based disorders and possibly several types of cancer.

Chapter 6 – Protein-protein interactions of YKL-40: Identification and characterization of collagen binding sites

This chapter focuses on how these carbohydrate-binding proteins can also be involved in surface interaction with other proteins as well. We worked on determining the unknown binding site for collagen on YKL-40, analysis of protein-protein binding abilities and calculation of relative binding affinities. This chapter has been adapted with permission from Kognole and Payne [223], Copyright © 2017, American Society for Biochemistry and Molecular Biology.

6.1 Introduction

Till now, the clinical use of YKL-40 as a biomarker in various mammalian diseases and structural properties of this chitinase-3-like-1 lectin have been described in the general introduction (Section 1.3.2). In the previous chapter we reported the carbohydrate-binding abilities of YKL-40 with hyaluronan being thermodynamically most preferred polysaccharide and heparin finding the surface binding site, however the story of binding partners for YKL-40 does not stop there. As YKL-40 protein has been observed to be up-regulated in various diseases that involve connective tissue remodeling, and based on YKL-40's noted affinity for various types of collagen and uncharted participation in collagen fibril formation [45], collagen has also been considered as a potential physiological ligand. Hence, in this chapter we specifically focus on identification and characterization of the collagen binding site of YKL-40 and exploring different surface properties that facilitate this protein-protein interaction. Collagen, unlike the other potential physiological ligands, is a macromolecular protein

triple helix structure (Figure 6.1); there are at least 27 distinct types of human collagen, forming a variety of biological networks, all of which are constructed of a basic Gly-Xxx-Yyy repeating amino acid sequence [254]. Generally, the unspecified amino acids, Xxx and Yyy, are proline (P) and hydroxyproline (O), respectively (Figure 6.1).

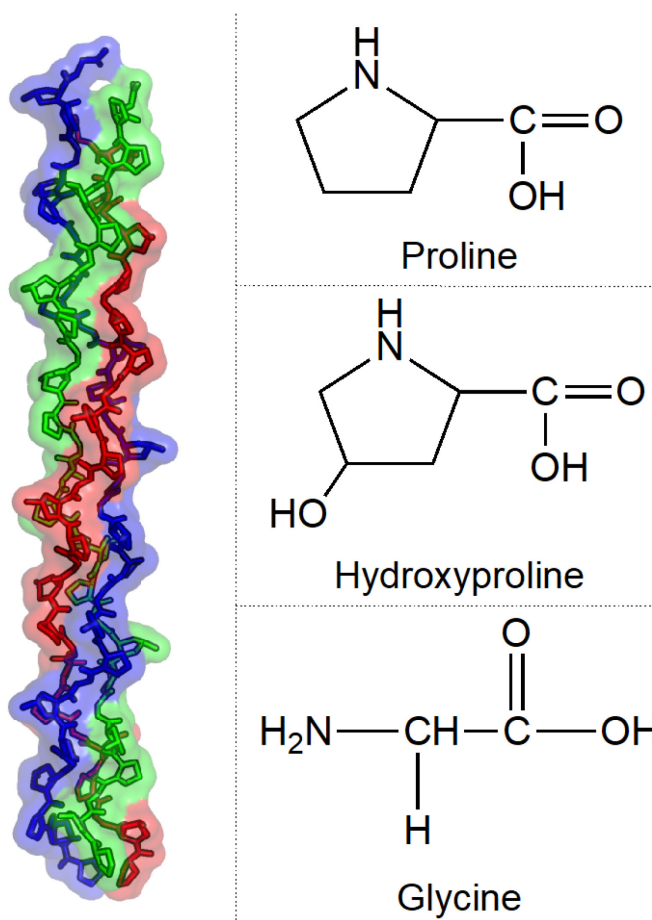


Figure 6.1 Triple helical structure of collagen from crystal structure (PDB ID – 1CAG) that exhibits strands with repeating amino acid sequence of Gly-Pro-Hyp. Three strands are shown in three colors. Structural formulas of glycine (G), proline (P) and hydroxyproline (O) are also shown.

Model collagen peptides have been observed in two different symmetries: the original Rich and Crick model with $10/3$ symmetry, 10 units in 3 turns, and the $7/2$ symmetry of a more tightly symmetrical triple helix [255-257]. On the molecular scale, collagen type will have relatively little impact on binding to YKL-40. However, symmetry may have an impact on hydrogen bonding to the binding site, and thus overall affinity, which will provide unique insight into physiological relevance. To date, model collagen peptides of a true $10/3$ symmetry have not been reported. Rather, the peptides either have a $7/2$ helical pitch or are somewhat “intermediate” in symmetry leading some to believe that the $7/2$ symmetry is representative of the true collagen helical structure [258]. However, it is not known how universally true this hypothesis is, as the structures of model peptides capture just a small subsection of the larger macromolecular structure [259].

With a broad range of possible collagen architectures, we have selected four representative model collagen peptides whose structures are both available from crystallographic evidence and span the $10/3$ and $7/2$ symmetries to the greatest possible extent. The first collagen peptide considered is that of the basic collagen peptide model, PDB ID 1CAG [260]. The reported 1.9 Å resolution structure exhibits a single Gly to Ala substitution and $7/2$ symmetry overall. Near the substitution site, the helix relaxes somewhat from $7/2$ symmetry, though not so much as to change overall symmetry. The second collagen model peptide we consider is a variation of the 1CAG peptide, where we reverted the alanine substitution to its native glycine. Minimization of this structure returns the helix to full $7/2$ symmetry; we refer to this peptide as “native 1CAG” here. The third model represents a segment from type III homotrimer collagen with

approximate 10/3 symmetry in the middle part of the helix (PDB ID 1BKV) [255]. This middle part of the 1BKV model peptide, also referred as the T3-785 peptide, has an imino acid-poor sequence of GITGARGLA. Our fourth model, 1Q7D, is a triple helical collagen-like peptide sequence including a hexapeptide Gly-Phe-Hyp-Gly-Glu-Arg (GFOGER) motif in the middle [261]; this motif is not sufficiently long to exhibit 10/3 symmetry, exhibiting, rather, an intermediate degree of 7/2 helical symmetry. This 1Q7D model is known to bind the integrin $\alpha 2 \beta 1$ -I domain protein [262], and the GFOGER motif is found in the $\alpha 1$ chain of type I collagen.

6.2 Methods

6.2.1 Docking of collagen triple helix on YKL-40

As stated in Chapter 6, all MD simulations were constructed based on the chitohexaose-bound YKL-40 crystal structure deposited by Houston et al. (PDB ID 1HJW) [43]. However, construction of the collagen-bound YKL-40 models required docking calculations to appropriately position the ligand. The collagen peptides are significantly larger than any of the carbohydrate ligands; thus, it is unlikely that a collagen molecule occupies the primary YKL-40 binding site in the same manner as chito-oligomer. Standard affinity-based docking calculations, such as the ones performed in AutoDock, are not feasible for determination of an initial collagen-binding domain given the size and flexibility of the triple helix structures. Rather, the collagen peptides were docked on the basis of molecular shape complementarity using the online web server PatchDock Beta v.1.3 [154, 156]. In the case of each of the four collagen-like model peptides, PatchDock predicted two potential occupancies along the surface of YKL-40, site A and site B. Binding site A corresponds to the primary carbohydrate-

binding domain of YKL-40, though the collagen ligand was not as deeply entrenched in the cleft as chitohexaose. Binding site B is located on the opposite side of YKL-40 from the primary binding cleft. Thus for each collagen-like peptide, two MD simulations were constructed representing the two potential binding sites. Figure 6.2 illustrates results of the docking with predicted collagen binding sites A and B for the 1Q7D collagen-like model peptide.

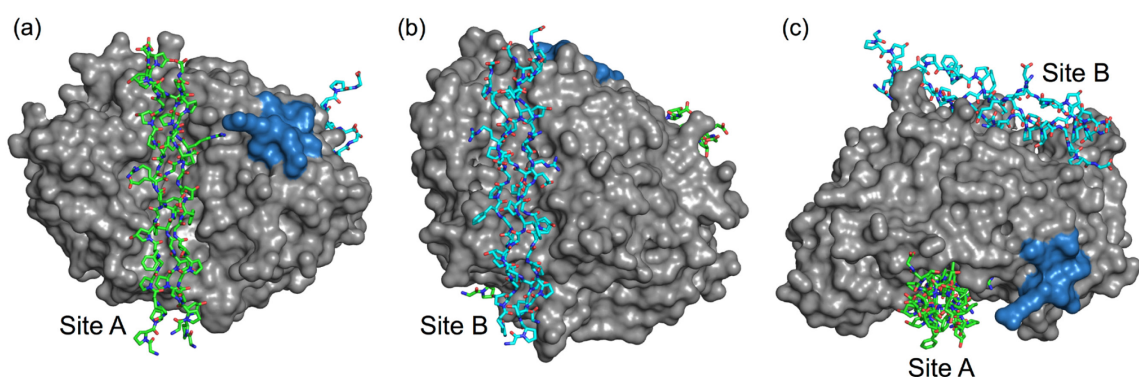


Figure 6.2 Molecular shape complementarity docking calculations predict collagen-like peptides will bind to YKL-40 in two possible orientations. (a) the front view of YKL-40 (gray surface) with collagen docked in site A (green stick), (b) the back view where collagen is docked in site B (cyan stick), and (c) top view of YKL-40 illustrating relative positions of binding sites. The putative heparin-binding subsite is shown in blue surface to aid in visualization of relative orientation of the protein-protein complexes. This particular figure shows the integrin-binding collagen peptide, 1Q7D [261], in the predicted binding sites along the surface of YKL-40; similar docking was carried out for other collagen models.

6.2.2 Molecular Dynamics Simulation

Most of the simulation setup procedure and parameters were exactly the same as described in earlier chapter for YKL-40 with carbohydrate ligands. The only difference was that the simulation box size was larger in this case with collagen being a larger binding partner. The YKL-40 complex with collagen-like peptides was solvated in $120 \text{ \AA} \times 120 \text{ \AA} \times 120 \text{ \AA}$ cubic boxes. The CHARMM36 force field with the CMAP correction [166, 191, 192] was used to describe YKL-40 and the collagen ligands. The parameters for hydroxyproline were determined using ParamChem, which determines force field parameters based on analogy with CHARMM General Force Field (CGenFF) program version 0.9.7 beta [263]. The CMAP corrections for hydroxyproline were adopted simply based on the analogy between proline and hydroxyproline residues. Water was modeled using the TIP3P force field [196, 197]. All simulations used explicit solvent.

A list of simulations and calculations performed to meet the objective of this study is given in Table 6.1. As described earlier, docking calculations of collagen-like peptides on YKL-40 indicated two potential binding surfaces; for these cases, the description in Table 6.1 lists both site and ligand. The free energy calculations performed are also indicated. Free energy calculations being highly expensive in terms of computational resources, only selective systems were included in this calculation. In addition to these protein-protein complexes, collagen-like peptide models were solvated in water separately, without YKL-40. However they were not used in any calculation or analysis as their only purpose was to preliminarily confirm whether the triple helical collagen-like peptides are eligible for dynamic studies and we have proper set of parameters to represent their chemical properties.

Table 6.1 Simulations and calculations performed in the investigation of the binding of collagen ligands to YKL-40.

Case No.	System	MD simulation	Free Energy Calculation
8 & 9	YKL-40 + collagen (1CAG) at site A & B	250 ns	--
10 & 11	YKL-40 + collagen (native 1CAG) at site A & B	250 ns	Umbrella Sampling
12 & 13	YKL-40 + collagen (1BKV) at site A & B	250 ns	--
14 & 15	YKL-40 + collagen (1Q7D) at site A & B	250 ns	Umbrella Sampling

The constructed protein-ligand systems were minimized in vacuum and subsequently solvated with water and sodium ions. Using CHARMM [166], the solvated systems were extensively minimized and heated to 300 K for 20 ps, which was followed by MD simulation for 100 ps in the *NPT* ensemble. The coordinates following density equilibration were used as a starting point for 250 ns of MD simulation in the *NVT* ensemble at 300 K using NAMD [169]. Explicit procedural details were similar to the Chapter 6.

6.2.3 Free Energy Calculations: Umbrella Sampling

Convergence challenges make FEP/ λ -REMD inappropriate for determining the binding free energy of the much larger and more flexible collagen-like model peptides. Thus, umbrella sampling was used to determine the work required to detach the collagen ligands from the shallow clefts of YKL-40. Over the entire reaction coordinate, this value

equates to binding affinity, enabling relative comparison of collagen peptide affinity. The MD umbrella sampling simulations used a native-contacts based reaction coordinate analogous to that defined by Sheinerman and Brooks and as implemented in recent cellulose decrystallization studies [209, 210, 264]. Here, a native contact was defined as YKL-40 protein residue within 12 Å of a collagen peptide residue; distance was defined by center of geometry of a given residue. The cutoff distance was selected to be larger than the non-bonded cutoff distance, ensuring that the collagen ligand was no longer interacting with YKL-40. Additionally, the water boxes of the collagen-YKL-40 systems were made bigger to accommodate the required separation distance.

The change in free energy was determined as a function of the reaction coordinate, ρ , formulated as the weighted sum of the states of the native contacts. The initial coordinates of the bound systems were selected from 250-ns equilibrated snapshots. The initial number of native contacts and their weights were calculated from these snapshots. An initial reaction coordinate of 0 (normalized) corresponds to this initial condition, and a final reaction coordinate of 1 corresponds to all of the native contacts being outside the 12-Å cutoff (i.e., the ligand is decoupled and freely sampling the bulk). The reaction coordinate was divided into 20 windows evenly spaced along the reaction coordinate, and each window was sampled for 5 ns, where the reaction coordinate was maintained at the specified value using a harmonic biasing force with the force constant of 10000 kcal/mol. The potential of mean force profiles were calculated using the weighted histogram analysis method (WHAM), and error analysis was performed using bootstrapping [181].

6.3 Results and Discussion

6.3.1 Protein-protein binding in YKL-40

Based on biochemical characterization, it is clear that YKL-40 functionally interacts with collagen. For example, Bigg et al. uncovered the ability of YKL-40 to specifically bind types I, II, and III collagen fibers [45], and Iwata et al. recently discovered that YKL-40 secreted by adipose tissue inhibits degradation of type I collagen by matrix metalloproteinase-1 and further stimulates the rate of type I collagen formation [265]. However, a lack of structural evidence has precluded development of an understanding of the molecular nature of these interactions. Using molecular docking, MD simulation, and free energy calculations, we describe interactions of four collagen-like peptides with two putative protein-binding sites along the surface of YKL-40. The selection of model peptides, as well as multiple binding sites, encompasses as many potential binding modes as feasible to describe protein-protein binding dynamics and relative affinity of YKL-40 for collagen.

6.3.2 Ligand Binding Dynamics and Comparison of Model Collagen-like Peptides.

Dynamics of the collagen-like peptide ligands varies with both binding site and the pitch of the triple helix. Root-mean-square deviation illustrates the relative stability of each collagen peptide in each of the two binding sites, A and B (Figure 6.3). Although the molecular docking results in very close contacts between collagen and YKL-40 (Figure 6.2), such that collagen appears to be almost buried in the primary carbohydrate-binding site of YKL-40, minimization and MD simulation results in slight rise and shift in the position of collagen for every model at binding site A. Each of the four ligands maintains association with the binding site A over the course of 250 ns, though with

slightly different protein-protein contacts with YKL-40 (Figure 6.4). Native 1CAG, 1BKV, and 1Q7D attained relative stability in a position not significantly different from the initial docked position, but the 1CAG peptide (Movie 7.1), with disrupted helical content resulting from the glycine to alanine mutation, required an adjustment in pitch before associating with YKL-40. This relative change in position is shown in the RMSD of the peptides during first 50 to 100 ns before stabilization (Figure 6.3a). Binding site B accommodates helical pitches of $7/2$ collagen peptides, as native 1CAG and 1Q7D associated with YKL-40 with very little change in orientation relative to the initial docked positions. The 1CAG ligand was expelled from binding site B as was the somewhat imperfect $10/3$ pitched-1BKV peptide. This suggests YKL-40 may avoid physiological interactions with certain collagen fibril domains, especially those having imperfect helical pitches. The integrin-binding collagen-like peptide 1Q7D demonstrated the greatest stability among collagen peptides in both binding sites (Figure 6.3) and formed more native contacts with YKL-40 than the other three collagen peptides at binding site A (Figure 6.4). We anticipate that the GFOGER motif plays substantial role in mediating the interaction of this collagen peptide with YKL-40.

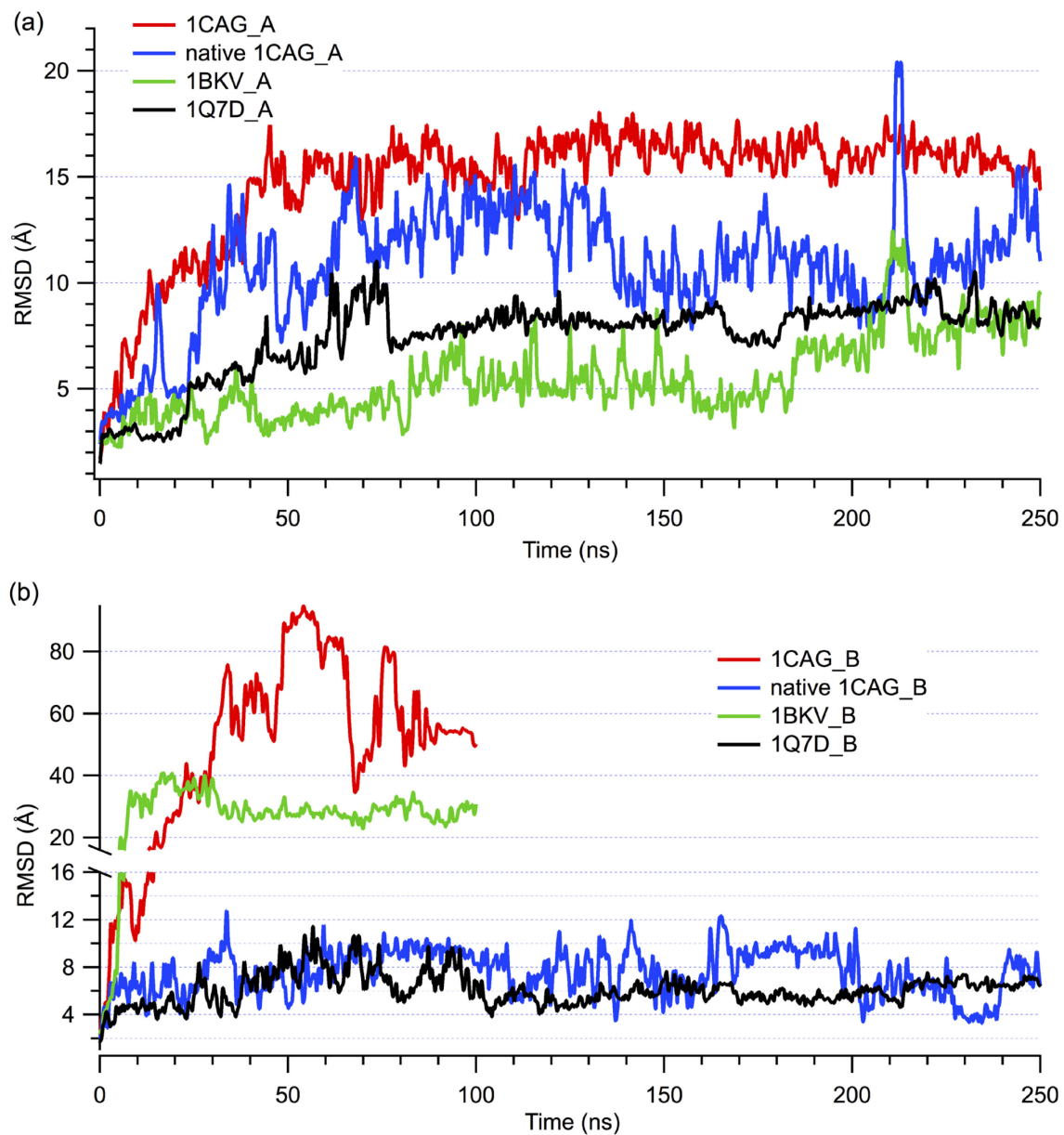


Figure 6.3 Root-mean-square deviation of collagen-like peptides over the course of 250-ns MD simulations at (a) collagen binding site A and (b) collagen binding site B. Each of the four collagen model peptides are shown.

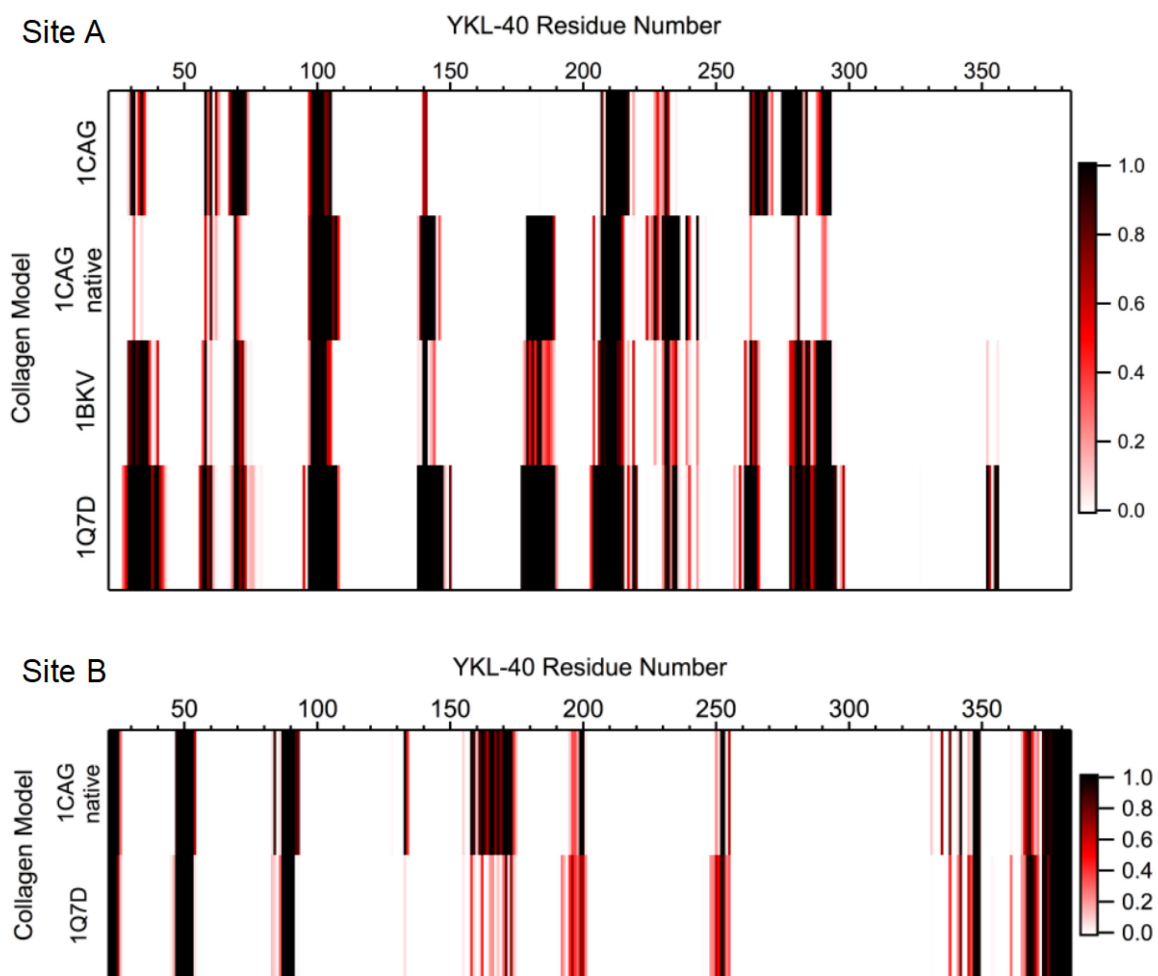


Figure 6.4 Native contact analysis of each collagen-like peptide model binding to YKL-40 at site A and at site B. The color scale represents the normalized frequency (i.e., fractional percentage of frames in which the contact was formed) of the respective YKL-40 residue as a native contact. A native contact was defined as anytime a collagen residue was within 12 Å of a YKL-40 residue where distance was defined by center of geometry of a given residue. Only frames from the last 100 ns simulation, following the period of equilibration, were considered in this analysis.

Examining the number of native contacts between each collagen peptide and binding site A of YKL-40 reveals several common interaction sites mediate collagen binding and helps narrow down key regions of interest (Figure 6.4). YKL-40 residues 69 to 71, 98 to 108, 205 to 215, and 230 to 235 interact with all four collagen peptides, and likely contribute to binding affinity, as we will discuss below. The region of YKL-40 between residues 179 and 189 associates with native 1CAG, 1BKV, and 1Q7D, but not with the original 1CAG as this peptide with relaxed symmetry needed to adjust its position from docked conformation to stabilize the interactions. The 1Q7D model formed the greatest number of interactions with YKL-40 residues relative to the other three models. Similar native contact analysis for binding site B shows that even N-terminal and C-terminal residues of YKL-40 are involved in collagen binding at binding site B (Figure 6.4). It shows that, unlike binding site A, there is little difference in number of interactions of model 1Q7D and native 1CAG collagen peptide with the binding site B of YKL-40.

To better understand the interactions collagen makes with YKL-40, identified through the native contact analysis, we calculated electrostatic and van der Waals interaction energies of each YKL-40 residue with each collagen peptide over the 250-ns MD simulations (Table 6.2). Visual inspection of the simulations reveals aromatic residues in the binding sites, such as Trp212 and Trp99 in binding site A and Phe49 in binding site B, were involved in aromatic-proline stacking interactions with the collagen triple helices. Such interactions are favorable, occurring due to both hydrophobic effects and interaction between the π aromatic face and the polarized C-H bonds [266]. This is illustrated in the van der Waals component of the interaction energy, where at binding

site A, Trp69, Trp71, Trp99, Trp212 and Phe234 show substantial favorable interaction with collagen peptides, though the contribution varies with each collagen peptide (Table 6.2). Additionally, acidic and basic residues of the integrin-binding GFOGER motif from collagen-like peptide 1Q7D form ionic interactions with the counter-ionic amino acids of YKL-40, also known as salt bridges. Specifically, 1Q7D forms three salt bridges at binding site A and one salt bridge at binding site B (Figure 6.5). At site A, Arg105, Asp207, and Arg263 of YKL-40 interact with Glu(a11), Arg(c12) and Glu(c11) of 1Q7D, respectively, where the a, b, and c in the parenthesis corresponds to one of the three strands of the collagen model. Notably, Glu(a11), Arg(c12) and Glu(c11) belong to the GFOGER integrin-binding motif. At site B, Lys23 of YKL-40 forms a salt bridge with Glu(a11) of 1Q7D. As anticipated, the GFOGER motif played a substantial role in the interaction of this collagen peptide with YKL-40, but its role was different from that of the integrin binding mechanism, which further involves coordination of a metal ion [262]. Nevertheless, salt-bridges and hydrophobic contacts are very important in both cases, significantly contributing to the electrostatic component of the binding affinity of this collagen peptide relative to collagen peptides lacking acidic or basic amino acids (e.g., native 1CAG) (Table 6.2). The hydroxyl oxygens of hydroxyprolines from 1CAG and native 1CAG appear to be involved in ionic interactions with acidic YKL-40 residues (favorable electrostatic interaction energies, Table 6.2), though as a result of hydrogen bonding rather than salt-bridge formation. We note that the interaction energies of YKL-40 residues with residues of each collagen model peptide are not conserved as a result of differences in the collagen sequences, particularly in the middle regions consisting of different imino-triplets.

Table 6.2 Interaction energies of YKL-40 residues with collagen peptides. The values are reported in terms of average interaction energy between major YKL-40 residues and collagen as a whole. van der Waals and electrostatic contributions are also provided separately. Residues with total average interaction energy greater than -4.18 kJ/mol have not been reported unless relevant to discussion. All the energies are in kJ/mol.

	Residue #	VdW-Avg	Elec-Avg	Total Avg		Residue #	VdW-Avg	Elec-Avg	Total Avg
1Q7D_A	ARG263	0.079	-5.635	-5.557	Native ICAG_A	ASP232	0.000	-6.651	-6.651
	THR184	-0.270	-3.917	-4.188		TRP99	-3.250	-1.806	-5.056
	LYS182	-0.173	-3.992	-4.165		TRP212	-3.746	0.084	-3.662
	TRP212	-3.268	-0.468	-3.736		VAL183	-2.367	-0.387	-2.754
	ASP207	-0.175	-3.432	-3.607		PHE234	-2.445	-0.020	-2.465
	TYR141	-1.061	-2.314	-3.376		ASN100	-1.265	-0.606	-1.871
	GLU70	0.010	-3.197	-3.187		GLU290	-0.166	-1.182	-1.348
	GLU290	-0.809	-1.939	-2.748		THR184	-0.678	-0.588	-1.266
	ARG145	0.033	-2.461	-2.428		GLN104	-0.525	-0.544	-1.068
	TYR34	-1.861	-0.195	-2.056		TYR141	-0.909	-0.092	-1.001
	ASN100	-1.631	-0.393	-2.024		ASP207	-0.158	-0.770	-0.928
	TRP99	-1.365	-0.552	-1.917					
	PRO142	-0.091	-1.330	-1.421					
	VAL183	-1.500	0.143	-1.357					
	GLU36	-0.347	-0.937	-1.283					
	Residue #	VdW-Avg	Elec-Avg	Total Avg		Residue #	VdW-Avg	Elec-Avg	Total Avg
ICAG_A	GLU70	-0.211	-5.383	-5.594	1BKV_A	ASP207	-0.970	-14.302	-15.272
	TRP99	-3.128	-0.966	-4.094		PHE208	-1.266	-3.398	-4.663
	GLU290	-0.576	-2.309	-2.884		ALA180	-0.468	-4.172	-4.640
	ASN100	-1.971	-0.605	-2.576		TYR141	-1.302	-3.032	-4.334
	TRP69	-1.520	-0.763	-2.283		TRP99	-3.116	-1.091	-4.207
	TRP71	-2.146	-0.106	-2.252		HIS209	-1.763	-1.774	-3.537

ICAG_A	ALA211	-1.386	-0.670	-2.056	IBKV_A	TRP212	-1.972	-0.960	-2.931
	TRP212	-1.583	-0.300	-1.882		LYS182	-0.377	-2.365	-2.742
	ASP207	-0.074	-1.337	-1.411		SER179	-0.419	-2.309	-2.728
	TYR34	-1.309	0.178	-1.131		GLU290	-1.192	-1.388	-2.580
	TRP31	-0.803	-0.121	-0.924		ARG213	-0.433	-2.107	-2.540
IQ7D_B						TYR206	-0.247	-2.058	-2.305
						GLY210	-0.652	-1.647	-2.299
						ALA211	-0.195	-1.887	-2.082
						TYR34	-1.660	-0.298	-1.958
						GLU36	-0.443	-1.035	-1.478
						VAL183	-1.421	0.035	-1.386
						ASN100	-0.947	-0.348	-1.296
						TRP31	-1.096	-0.089	-1.184
	Residue #	VdW-Avg	Elec-Avg	Total Avg		Residue #	VdW-Avg	Elec-Avg	Total Avg
	LYS23	-0.190	-20.619	-20.809	Native ICAG_B	ASN89	-2.861	-6.322	-9.182
	TYR22	-1.145	-11.156	-12.301		LYS377	-1.703	-6.249	-7.952
	LYS91	0.011	-9.232	-9.221		ASP378	-1.224	-3.781	-5.004
	PHE49	-3.618	-0.449	-4.066		ALA381	-2.240	-0.846	-3.086
	ASP367	-0.062	-3.849	-3.912		GLN166	-1.533	-1.004	-2.536
	LYS377	-1.282	-2.186	-3.469		THR52	-1.952	-0.552	-2.504
	THR52	-0.760	-2.420	-3.179		GLN171	-1.526	-0.721	-2.247
	ASP47	-0.227	-1.987	-2.215		PHE49	-1.761	-0.218	-1.980
	LYS253	-0.280	-1.631	-1.911		TYR22	-1.055	-0.631	-1.686
	ASN89	-1.654	-0.172	-1.825		LYS91	-0.891	-0.689	-1.581
	ASP378	-0.225	-1.367	-1.592		LEU50	-0.677	-0.685	-1.362
	ALA381	-1.250	0.077	-1.174		HIS53	-0.669	-0.541	-1.210
						ASP199	-0.130	-0.880	-1.010

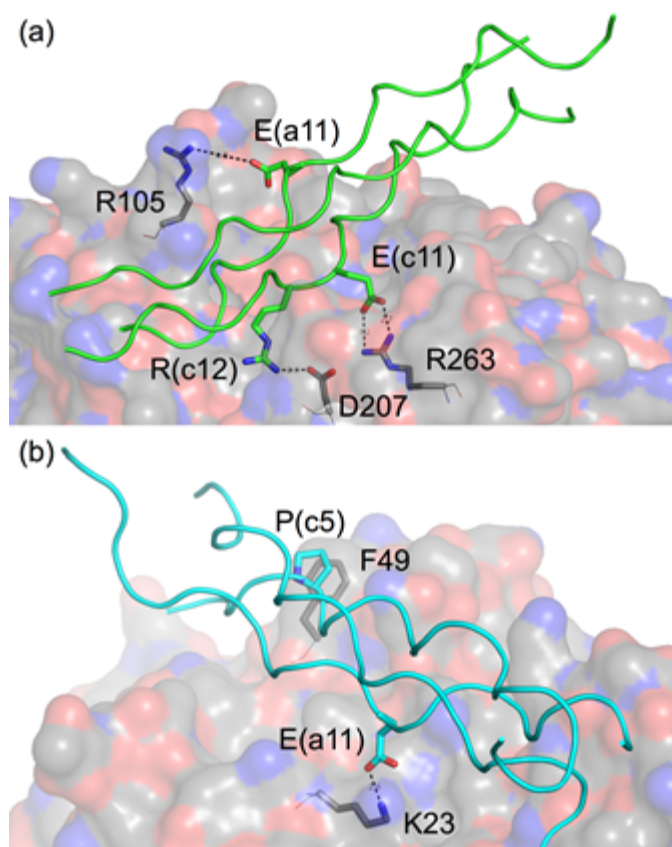


Figure 6.5. Collagen binding with YKL-40. (a) Salt bridges formed between the 1Q7D collagen peptide (green cartoon) and binding site A (gray surface) (b) Salt bridge interactions of the 1Q7D collagen peptide (cyan cartoon) with binding site B (gray surface).

From MD simulation, we observe substantial hydrogen bonding between the collagen peptides and YKL-40 across the length of each binding site, which contributes to overall stability and binding affinity. The hydrogen bonding analysis for the collagen-YKL-40 systems was performed as described above for the polysaccharide ligands; pairs exhibiting greater than 10% occupancy over the simulation are reported individually in Table 6.3. The YKL-40 residues responsible for hydrogen bonding are not consistent across each collagen model (Table 6.3). In general, Glu70, Trp99, Asn100, Tyr141,

Arg145, Ser179, Lys182, Thr184, Asp207, Arg213, Phe218, Asp232, Arg263 in binding site A form hydrogen bonds with the peptides. Similarly at site B, Tyr22, Lys23, Asn87, Asn89, Lys91, Lys377, Asp378 are typically involved in hydrogen bonding. The variation in hydrogen bonding pairs between YKL-40 and the collagen peptides is a natural extension of the varying amino acids along the repeating Gly-Xxx-Yyy sequence; there are many different potential donor and acceptor pairs in each case. Hydroxyproline residues play a crucial role both as donor and acceptor in most pairs, benefitting from the extra hydroxyl group relative to proline. Although all the collagen peptides maintain association with collagen-binding site A, the hydrogen-bonding characteristics are slightly different for each, which will in turn lead to affinity differences. The relaxed helical pitch of the 1CAG peptide effectively disrupts hydrogen bonding, and affinity for the ligand is lost at binding site B. The 1BKV peptide model with 10/3 symmetry was also unable to remain associated with binding site B, suggesting that binding site B may be more sensitive to helical pitch and prefers 7/2 symmetrical helices.

Table 6.3 Hydrogen bonding pairs between YKL-40 and collagen model peptides at binding site A, including percentage occupancy, over 250-ns MD simulations. A hydrogen bond was considered to be a polar atom having a donor-acceptor distance of 3.4 Å and a 60° cutoff angle. Occupancies above 100% mean that the same pair was involved in more than one type of hydrogen bond.

1Q7D_A			Native 1CAG_A		
Donor	Acceptor	Occupancy	Donor	Acceptor	Occupancy
ARG263-SC	GLU11-SC	164.84%	ARG213-SC	HYP8-SC	79.68%
ARG12-SC	ASP207-SC	126.36%	HYP8-SC	ASP232-SC	76.56%
ARG12-SC	THR184-SC	51.96%	GLN104-MC	HYP14-SC	19.60%
ARG12-SC	ALA291-MC	26.76%	SER103-MC	HYP14-SC	16.24%
HYP9-SC	GLU290-SC	25.56%	ARG233-SC	HYP2-SC	13.44%
HYP6-SC	GLU70-SC	18.80%	HYP17-SC	ASN100-SC	10.32%
TYR141-SC	GLU11-SC	17.28%	other pairs		123.68%
ASN100-SC	HYP9-SC	14.56%			
ARG12-SC	SER179-SC	13.44%			
HYP6-SC	TYR34-MC	13.24%			
ARG12-SC	ASP207-MC	12.04%			
other pairs		100.20%			
Total		585.04%	Total		339.52%

1CAG_A			1BKV_A		
Donor	Acceptor	Occupancy	Donor	Acceptor	Occupancy
HYP20-SC	GLU70-SC	75.72%	TRP99-SC	ALA17-MC	64.92%
ASN100-SC	HYP17-MC	35.84%	ARG14-MC	PHE218-MC	56.88%
GLY214-MC	HYP5-MC	29.24%	ARG11-SC	SER179-SC	54.08%
ARG213-SC	HYP5-SC	26.20%	LYS182-SC	THR8-SC	45.60%
HYP14-SC	ALA291-MC	17.80%	ARG11-SC	TYR141-SC	42.56%
GLY214-MC	GLY6-MC	11.00%	ARG11-SC	ALA180-MC	40.04%

other pairs	112.08%	ARG11-MC	THR184-SC	33.00%
		ARG11-SC	ASP207-MC	28.64%
		ARG11-SC	ASP207-SC	21.56%
		ARG213-SC	THR11-MC	20.24%
		ARG14-SC	GLY210-MC	15.92%
		ARG11-SC	TYR206-MC	13.72%
		TYR141-SC	GLY12-MC	13.04%
		TRP212-SC	GLY12-MC	12.76%
		ARG14-SC	ALA211-MC	11.04%
		other pairs		80.28%
Total	307.88%	Total		554.28%

1Q7D_B			Native 1CAG_B		
Donor	Acceptor	Occupancy	Donor	Acceptor	Occupancy
LYS23-SC	GLU11-SC	120.84%	ASN89-SC	GLY18-MC	94.68%
ASN87-SC	HYP6-SC	61.20%	HYP17-SC	ASN89-MC	84.72%
ASN89-SC	HYP6-MC	55.32%	LYS377-SC	HYP20-MC	59.92%
LYS91-SC	HYP9-SC	23.80%	ASN89-SC	HYP17-MC	47.84%
LYS377-SC	HYP3-SC	21.12%	HYP23-SC	ASP378-SC	32.80%
LYS91-SC	GLU11-SC	18.68%	LYS377-SC	GLY21-MC	30.20%
LYS377-SC	GLY1-MC	16.52%	GLN166-SC	HYP8-MC	18.40%
TYR22-MC	GLU11-SC	15.48%	HYP20-SC	ALA381-MC	16.92%
GLN171-SC	HYP15-SC	11.76%	ASN87-SC	HYP20-SC	15.68%
GLN171-SC	HYP18-SC	11.16%	HYP11-SC	LYS169-MC	15.28%
THR52-SC	GLU11-SC	10.00%	GLN171-SC	HYP11-MC	15.04%
other pairs		129.24%	LYS91-SC	HYP17-SC	12.80%
			other pairs		98.36%
Total		495.12%	Total		542.64%

6.3.3 Collagen-like peptide binding affinity

The relative binding affinity of collagen-like peptides to YKL-40 was determined from umbrella sampling MD simulations. Here, we report the binding affinity of the 1Q7D collagen-like peptide, which is the integrin binding peptide with an overall 7/2 helical pitch [261, 262], at both sites A and B. We have also calculated binding affinity of native 1CAG at binding site A for comparison of binding affinities of two different collagen peptides having different residue substitutions and helical pitches. Unfortunately, in case of umbrella sampling for the native 1CAG peptide at site B, we were unable to obtain statistically reliable results, and thus, we will not discuss findings relative to affinity of this model.

The umbrella sampling MD simulations of the 1Q7D collagen peptide at both sites A and B show that YKL-40 has similar affinity for 1Q7D at both sites; whereas, the native 1CAG collagen peptide appears to bind with a lower affinity than 1Q7D at binding site A (Figure 6.6). We note that the last umbrella sampling window in case 1Q7D at binding site A shows a sudden, sharp increase in the PMF, which is an artifact of the use of native contacts as an umbrella sampling reaction coordinate. As the standard C-terminals of three strands of collagen helix are negatively charged, they are attracted to the nearby, highly positively charged surface of heparin-binding site. As a result, the final window of the PMF overestimates the work to remove the 1Q7D peptide from binding site A exclusively (Figure 6.6). Removing this latter window from the calculation, the free energy of binding 1Q7D is -9.75 ± 1.12 kJ/mol in site A and -10.26 ± 1.19 kJ/mol in site B. The free energy of binding native 1CAG in site A is -5.26 ± 0.81 kJ/mol. The relatively low statistical uncertainty at each window along the potential of mean force

suggests sampling of the system was sufficient, providing a meaningful estimate of binding affinity.

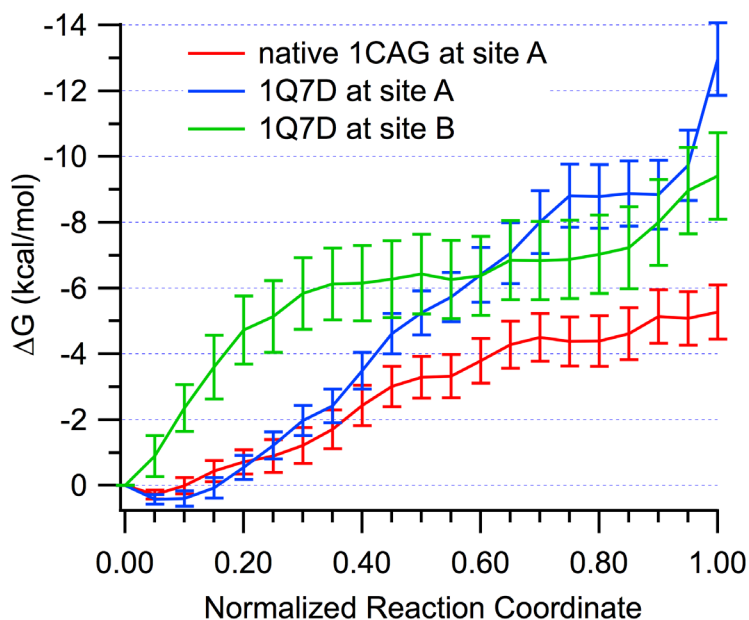


Figure 6.6. Binding free energy obtained from umbrella sampling MD simulations of the YKL-40-collagen peptide systems, interpreted as negative of potential of mean force (PMF) to decouple the partners. The collagen peptides in question are 1Q7D (at both sites) and 1CAG (at site A only). The free energy is shown as a function of normalized reaction coordinate, where the reaction coordinate is fraction of native contacts.

The potential of mean force determined from umbrella sampling MD simulations determines the amount of work required to pull the collagen ligand from the binding site. As free energy is a state function, the difference between the beginning and end state is the binding affinity, regardless of path taken, as reported above. The path can provide information as to barriers to unbinding; however, the collagen peptides are readily removed from the binding sites along a relatively smooth path. This suggests there is

little conformational rearrangement required of YKL-40 in the release of the collagen ligand. The difference between affinity for 1Q7D and native 1CAG at binding site A is reflected in the total hydrogen bonding occupancy in those two cases (Table 6.3). Notably, the binding free energies of collagen to YKL-40 are approximately half that of the tighter binding polysaccharide ligands. This suggests that YKL-40 will bind both hyaluronan and chito-oligomers over collagen in the presence of all three. This does not rule out collagen as a physiological ligand, but strongly supports hyaluronan as a preferred physiological ligand of YKL-40.

6.4 Conclusions

The docking of triple helical collagen-like peptide models on YKL-40 surface based on molecular shape complementarity at two surface binding sites again proves that exteriors of protein than primary binding site are equally significant in its functionality. Analysis of protein-protein binding dynamics, compared over various collagen models, provides detailed characteristics of surface binding residues at both the proposed binding sites. Binding site A showed more adaptability towards different helical symmetries and mutant disruptions. Binding site B only accommodated collagen peptides with 7/2 symmetry. Native contact analysis was very helpful in pointing out the most important residues/loops that were involved in protein-protein interaction. 1Q7D collagen-like peptide model, with the GFOGER-motif, appears to be the most favorable and stable binding partner at both the sites among the four cases studied. The affinity of surface binding of collagen was much less than the affinities for chitohexaose and hyaluronan in the primary binding cleft. The demonstrated ability of YKL-40 to bind collagen molecules with two surface binding sites adds another dimension to its functionality in

extracellular matrix, as it has already been experimentally shown to affect collagen fibril formation. However, further investigation of significance of this protein-protein interaction is needed and combined with experimental evaluation of hyaluronan binding, we could ultimately use this knowledge in order to utilize the mammalian glycoprotein in many roles than just biomarker.

Chapter 7 – Conclusions and future work

This dissertation has covered a crucial aspect of non-catalytic proteins having the unique abilities to bind not only various morphologies of substrates but also multiple sugar codes, sometimes with the aid of surface-binding sites. As the underlying mechanisms of such substrate recognition abilities are important to various applications, we set out to develop a molecular-level understanding of interesting protein-carbohydrate binding sites, their adaptability for ligand orientations, dynamical differences in binding to various morphologies, and thermodynamic preferences to certain protein-carbohydrate pairs; such an investigation provides critical insights that are beneficial towards the development of biotechnological applications, for example better use of CBM diversity in developing enzyme cocktails for efficient enzymatic hydrolysis. In this chapter, we discuss how our work can impact the future directions of this research area.

Through our CBM studies, we reached two main conclusions about Type B CBMs, based on MD simulations and free energy calculations. In the first part of the study, Chapter 3, we not only confirmed the original Johnson, et al. [61] hypothesis that *C. fimi* CBM4s are capable of binding cello-oligomers with the reducing end of the pyranose at either end of the binding cleft, but also showed that this bi-directional binding phenomenon also holds true for four other Type B CBMs. From this, we hypothesize that all cellulose-specific CBMs with a β -sandwich fold may exhibit this ability. This bi-directional binding ability of cellulose-specific CBMs suggests that multi-modular enzymes are evolved to recognize the free glycan chains with higher efficiency. It is another indication nature has evolutionarily selected for proteins with mechanisms

like bi-directional substrate recognition in response to the structural features of cellulose, with the approximate symmetry of sidechains, as a means to break it down in faster ways to maintain the balance of the carbon cycle. Nonetheless, we have more to learn from nature to capitalize on it for industrial and developmental purposes. Accordingly, CBMs with their basic attributes being independently folding proteins, abundant and inexpensive matrix attachment regions, and compliant binding specificities have already been employed in many applications, for example high-capacity purification tags for protein isolation [35]. The bi-directional binding ability of Type B CBMs could be harnessed for applications in which a directional preference is not critical, for example targeting of functional molecules to materials containing cellulose. CBMs are already being used to probe for polysaccharides in plant cell walls [94], and knowing that the CBMs can bind bi-directionally necessitates considering additional factor in designing probes for differentiation of the substrate morphologies. The molecular-level knowledge of these Type B CBM binding clefts can be utilized to engineer proteins with more flexibility towards binding orientations where ligands can offer symmetric interactions.

In Chapter 4, the architectural features of CBM binding sites in three families from the same type of CBMs were studied, extensively illustrating various characteristic differences that facilitate the tighter oligomeric binding within twisted platforms. The investigation further leads to the fact that this difference in binding site, within same family, is evolved to target distinct regions of non-crystalline cellulose and not only single glycan chains. The study highlights an overlooked element in structural characterizations of proteins that, although the active site or ligand-binding site is of prime importance towards proteins functionality, the rest of the protein surface and,

specifically, the neighboring loops of the binding site may play a critical role in defining the larger aspect of protein functionality. We anticipate our work will motivate more in-depth ‘active site’ characterizations that extend beyond the immediate protein-carbohydrate interface. Undoubtedly, MD simulations facilitate such investigations once structural data from either X-ray crystallography or NMR spectroscopy is available, but homology modeling is proving to be an increasingly reliable tool for structural biologists. We envision extension of this research would involve mutations in the exterior loops around the twisted binding platforms of family 17 and 28 CBMs, which would be tested across range of non-crystalline substrates; from this, we would be able to identify the true contributions to specificities. This information would also benefit development of refined CBM probes for probing plant cell wall architecture by enabling biotechnology with differential arrays of CBMs capable of recognizing specific morphologies on non-crystalline substrates. Exact identification the composition of pretreated biomass in terms of crystalline, non-crystalline, and soluble substrate, with better resolution of non-crystalline morphologies, allows design of complementary enzyme cocktails for efficient and faster hydrolysis. The different hydrogen bonding patterns across the two platforms pose a great example that even the small differences in binding site have evolved to provide the functional specificity.

As stated earlier, Section 1.3.1.5, there is more to the story of Type B CBMs, pertaining to the fact that, in nature, four out of six CBMs studied here are found in tandem. The tandem Type B CBM construction further confounds fundamental understanding, as the two tandem CBM systems, *CfCBM4-1/CfCBM4-2* and *BspCBM17/BspCBM28*, differ in their binding abilities. There are two basic

dissimilarities in these tandems. First, both have either sandwich or twisted binding platforms in their individual CBMs, and second, there is significant difference in linker length. The *Bsp*CBM17/*Bsp*CBM28 tandem, having twisted platform CBMs and a longer peptide linker length, exhibits cooperative binding effects, with up to 100-fold higher rate of association over that of the individual CBMs [106]. The tandem *Cf*CBMs have a very short linker in combination with sandwich platforms of both individuals, and show only additive affinity improvement relative to separate domains [98]. Using these tandems *Cf*CBM4-1/*Cf*CBM4-2 and *Bsp*CBM17/*Bsp*CBM28 as representative models, we plan future studies to address the questions about cooperative and additive carbohydrate binding. The relative orientation of two partners in the wild-type tandems, length of the linker region between the two, and relative presence of the bound glycan chains are parameters that may contribute to cooperativity. We have obtained preliminary results examining the wild-type tandems, where the individual CBMs were connected by modeling the linker peptide between them based on the deposited sequences [105, 267]. After allowing a long equilibration before production MD, we obtained short 25-ns trajectories that we have analyzed through residue-residue cross-correlations (Figure 7.1) and principal component analysis (PCA) (Figure 7.2). In the case of tandem *Cf*CBM4s with a short linker, a large amount of strong negative correlations were observed between the residues from the two different CBMs; while in tandem *Bsp*CBMs, correlation across the CBMs was relatively very low. Additionally, we observed a large network of strong, positive correlations in tandem *Bsp*CBMs between residues of the same CBM, and these two CBMs only interact through the linker. In contrast, the tandem *Cf*CBMs show strong positive correlations that extend across the linker to connect residues of one CBM to that

of the partner CBM. This is a promising indication that short linkers induce highly interdependent motions of the CBMs, which may restrict co-cooperativity. The PCA analysis provides the essential dynamic information that allows us to describe the protein motions that contribute the most to conformational variation. Figure 7.2 shows that the tandem *Bsp*CBMs have highly scattered conformations in a trajectory along largest principal components (PC), while tandem *Cj*CBMs have closely related groups of conformers.

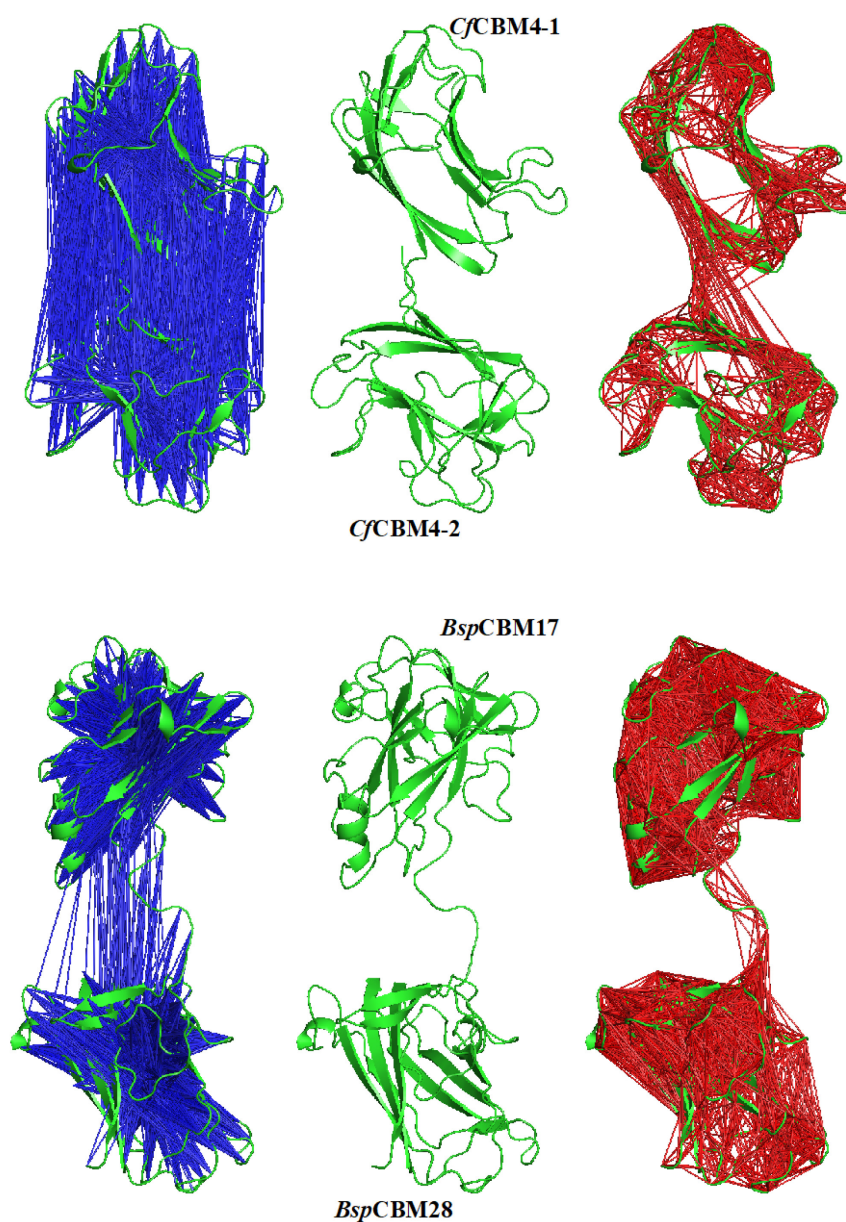


Figure 7.1 Visualization of residue-residue cross-correlation of tandem CBMs calculated based on RMSD of the protein backbone (α -carbon only). The linker length in *BspCBMs* is longer than that in *CfCBMs*. The blue lines represent strong negative correlation between a pair of residues, while the red lines represent strong positive correlation.

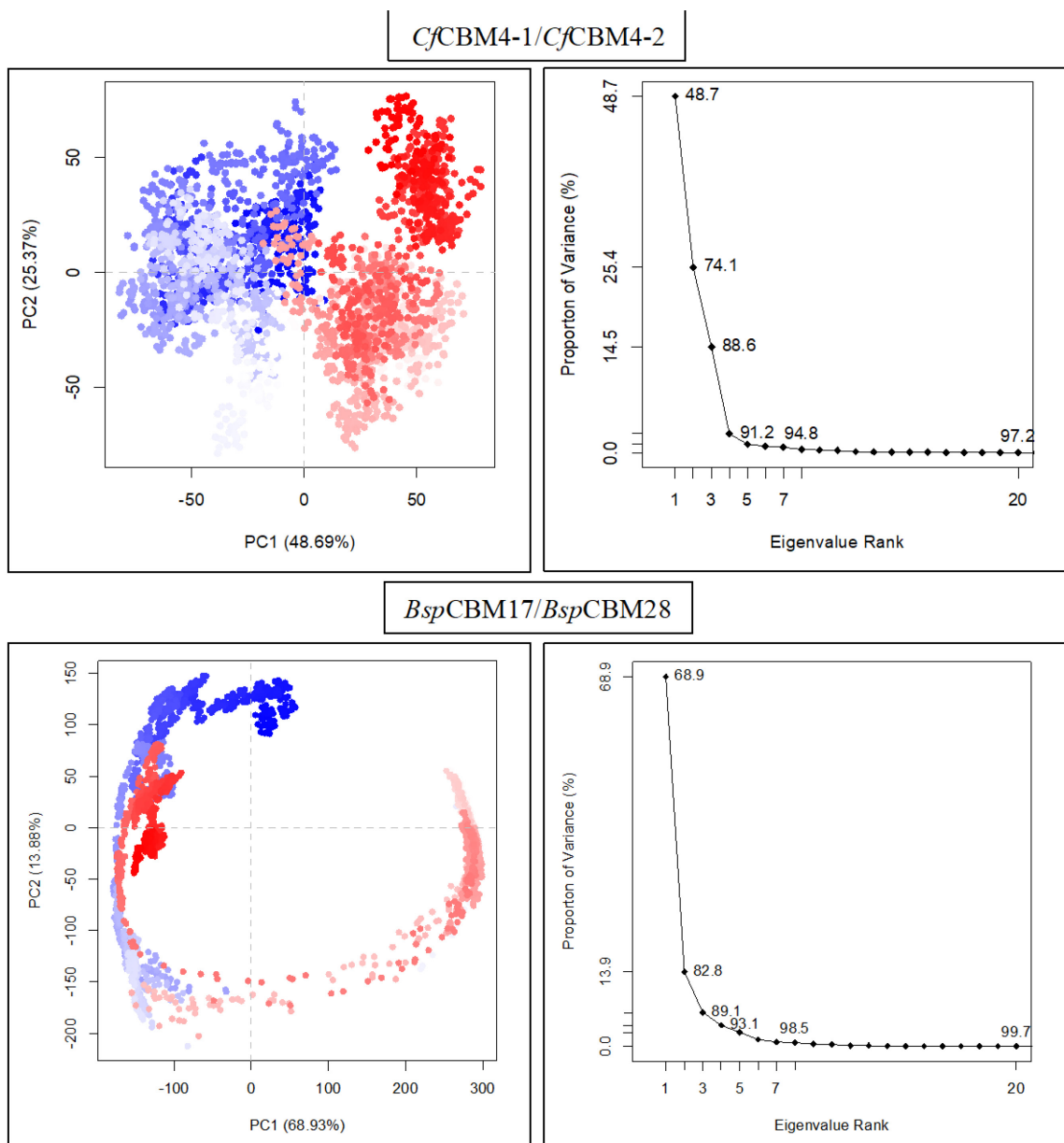


Figure 7.2 PCA analysis of preliminary MD simulation data. Left panels illustrate the clustering of conformers on a principle components 1 and 2 (PC1-PC2) space. A continuous color scale (from blue to red) is used to follow the sequence of conformers in the trajectory. The right panels show the percentage of the total mean square displacement (or variance) of atomic fluctuations captured in each dimension characterized by their corresponding eigenvalue.

Through long-timescale MD simulations and post-process analysis, such as these above, future studies will address protein-protein networking in these tandem CBMs, comparing individual CBM dynamics and binding mechanisms. With variations in linker lengths, through swapping of the linkers between tandems, adding cello-oligomeric ligands in the binding sites of CBMs to consider possible allosteric effects, and further introduction of the catalytic module in the modeling, I expect to characterize the overall behavioral attributes of these multi-modular enzymes.

For the other non-catalytic protein, YKL-40, studied in this dissertation apart from CBMs we also arrive at remarkable conclusions. We have covered most of the ECM components having the potential to be a physiological ligand of this multi-functional protein. In Chapter 5, we modeled and analyzed the binding mechanism of YKL-40 for its known binding partner, chitohexaose, along with 5 different oligosaccharides, where chitohexaose, hyaluronan and cellohexaose could remain in the primary binding site; heparin was able to bind to a surface binding site. Moreover, in Chapter 5, the same mammalian glycoprotein was able to bind four different model collagen peptides through two different surface binding sites. The absolute thermodynamic favorability of YKL-40 for hyaluronan over not only chitohexaose but also collagen model peptides calls for prompt experimental confirmation of this prediction. *In vitro* binding studies of YKL-40 with hyaluronan and structural characterization of their interactions would provide further resolution of this novel interaction, which along with *in vivo* studies for expression levels of both the partners in various malignant tissues, will reveal a clear picture about the physiological significance of this strong biomolecular binding. Nonetheless, the details of these non-catalytic protein-carbohydrate interactions discussed in this dissertation can be

adapted to develop new inhibitory molecules/pathways for this lectin. The affinity of the YKL-40 surface for heparin and other similar, highly negatively charged GAGs also needs further attention, as this surface-binding site may play a role in cell-cell and cell-matrix interactions. Heparin-Sepharose chromatography is used to purify YKL-40 [45]; a similar purification approach could also be developed for other proteins by modifying a surface patch with high positive charge density in the heparin-binding motif, GRRDKQH. The collagen binding site identification and characterization reveals that YKL-40 has very unique surface residues that can specifically recognize binding partners of many types and symmetries of collagen triple helices. The significance of having two binding sites needs to be further analyzed, as it may play a crucial role in bringing the two helices together in fibril formation; however, the experimental evidence of YKL-40's involvement with collagen fibril formation is ambiguous and reports both inhibitory and stimulatory effects based on different forms of protein [45]. Study of this protein-protein relationship will clarify the higher expression levels of YKL-40 in inflammatory joint diseases like osteoarthritis.

In summary, this dissertation applies state-of-art computational approaches to gain molecular-level understanding of inconspicuous mechanisms and unidentified interactions in two studies of carbohydrate recognition by non-catalytic proteins namely, Type B CBMs, recognizing oligomeric and non-crystalline cellulose, and YKL-40, binding to GAGs. The protein-protein interactions, such as tandem CBMs or collagen binding of YKL-40, inevitably come into the picture, as many biological events are highly interdependent and often must be studied simultaneously for effective analysis of results.

Appendix

A1 Supporting information related to CBMs

As the title suggests, Appendix A1 reports the additional details of simulation procedures. Also, some additional results about data analysis and free energy calculation methods that are not exactly related to actual theme of the dissertation are reported here. Information in this appendix has been adapted with permission from Kognole and Payne [184], Copyright © 2015, Oxford University Press.

A1.1 Additional methods for Chapter 3

The CBM-ligand complex systems illustrated in Figure 3.1 were constructed using CHARMM as described in the manuscript [166]. Protonation states of the titratable residues were determined by H++ and manual inspection [187-190]. GLU14, GLU55 and ASP120 from *Cf*CBM4-1 were protonated whereas no residues were protonated in *Cf*CBM4-2.

A1.1.1 Molecular Dynamics (MD) Simulation

The CBM-cellopentaose systems were minimized in vacuum in a stepwise fashion using the method of steepest descent. For 1000 steps, the ligand side-chains were minimized, holding all other atoms fixed. The entire ligand was minimized for an additional 1000 steps, with the protein held fixed. Finally, the protein and ligand were minimized for another 1000 steps with no constraints. An additional 1000 steps of minimization using Adopted Basis Newton-Raphson algorithm completed the vacuum minimization procedure. The minimized systems were then solvated; *Cf*CBM4-1 systems

were solvated in 65 Å square-box (~27,300 atoms), and *Cf*CBM4-2 systems were solvated in 60 Å square-box (~22,000 atoms). Sodium atoms, 13 for *Cf*CBM4-1 and 13 for *Cf*CBM4-2, were added to neutralize the system charge, as required for application of Particle Mesh Ewald electrostatic approximations. The solvated systems were re-minimized in a stepwise fashion. Holding the protein and the ligand fixed, the water molecules were minimized for 1000 steps of steepest descent. The restraints on the ligand side chains were then removed, and the water and ligand side chains were minimized for 1000 steps of steepest descent. The entire ligand and water molecules were then minimized for 1000 steps of steepest descent, and then, the entire solvated complex was minimized for 2000 steps of steepest descent and 2000 steps with adopted basis Newton-Raphson method.

Then the systems were heated from 100 to 300 K for 20 ps and equilibrated in the *NPT* ensemble for 100 ps. The Nosé-Hoover thermostat was used to control temperature in CHARMM [212, 213]. Shake was used to fix the distances to hydrogen atoms [215]. Non-bonded interactions were truncated with a 10-Å cutoff. The Particle Mesh Ewald method with a 6th order b-spline [216], a Gaussian distribution width of 0.320 Å, and a mesh size of $90 \times 90 \times 90$ was used to describe the electrostatics. All simulations used a 2-fs time step. The CHARMM36 force field with the CMAP correction was used to describe the protein [166, 191, 192], and the CHARMM36 carbohydrate force-field was used for the ligand [193-195]. Water was modeled using the modified TIP3P force field [196, 197]. Production MD simulations were performed using NAMD in the *NVT* ensemble at 300 K for 250 ns using a 2 fs time step [169]. Temperature was controlled

using Langevin thermostat in NAMD [214]. All other simulation parameters were the same as those described above for the CHARMM equilibration.

A1.1.2 Free Energy Simulation

To determine the absolute binding free energy of the *Cj*CBM4-1-RE and *Cj*CBM4-1-NRE systems to cellopentaose per the defined thermodynamic pathway (Figure 3.3), the solvation free energy of the cellopentaose ligand must be determined. Accordingly, a solvated ligand system was constructed in a fashion similar to the protein-ligand systems. The cellopentaose ligand was solvated in a 65-Å square-box (~9,000 water molecules), requiring no sodium ion additions. The solvated cellopentaose was minimized, heated, and equilibrated using the same protocol as outlined above. Free energy simulations were started from the 0.1-ns equilibration, which is sufficient for effective diffusion of the highly mobile water molecules.

One hundred twenty-eight Free Energy Perturbation (FEP) λ -windows, treated as replicas, were run concurrently, obtaining an acceptance ratio of > 80% along the alchemical path. These 128 replicas were distributed into 72 repulsive, 24 dispersive, and 32 electrostatic replicas. The MD simulations were performed using periodic boundary conditions in the *NVT* ensemble at 300K with a 1-fs time step. The systems were propagated with the multiple time step integration scheme and Langevin dynamics. Forty sequential 100 ps production runs were performed with a replica exchange frequency of 1/100 steps. The force field parameters were the same as described above for the MD simulations. Initial structures of the enzyme-ligand complexes were taken from the final coordinate set of the equilibration trajectory (0.1 ns). Translational and rotational

restraining potentials were applied in the enzyme-ligand free energy calculations. The distance from the initial center of mass of the ligand to the initial center of mass of the protein was maintained through an applied $10 \text{ kcal/mol/\AA}^2$ harmonic restraint.

Energies collected from the MD simulations were used to determine repulsive, electrostatic, and dispersive contributions to the free energy of binding. The Multistate Bennett Acceptance Ratio (MBAR) was applied to determine free energy and statistical uncertainty associated with each 0.1 ns interval [176]. From the 40 independent FEP/ λ -REMD calculations ($40 \times 100 \text{ ps}$), the last 30 (3 ns) were used to obtain the average free energy of binding. The contribution of this restraint to the free energy was determined via numerical integration with Simpson's rule as described by Deng and Roux [173]. The error associated with the binding free energy was obtained from determining the standard deviation over the last 3 ns for each of the two sets of calculations and combining standard deviation using error propagation rules.

A1.2 Additional results

A1.2.1 The effect of replica-exchange frequency on the sampling and convergence

The time progression of the 40 consecutive 0.1 ns FEP calculations is illustrated in Figure A.1a. The free energies reported in Chapter 3 (Table 3.1) were obtained from the final 3 ns. Statistical uncertainty of each point has been determined using MBAR. The effect of replica exchange frequency on statistical uncertainty and acceptance ratio was also considered. One generally expects that increasing the exchange frequency will improve the statistical sampling, perhaps leading to faster convergence. The reported free energy values were obtained using an exchange frequency of 1/100 steps, or every 0.1 ps,

as described in the Methods. We also examined exchange frequencies of every 1/1000 steps and 1/10 steps. All other parameters were identical. The binding free energies for CfCBM4-1-RE with exchange frequency of 1/1000 and 1/10 were -4.71 ± 1.17 kJ mol⁻¹ and -4.88 ± 1.00 kJ mol⁻¹, respectively (Figure A.1b). We found a moderate improvement in uncertainty with increased exchange frequency; however, progress toward convergence did not appear to be affected (Figure A.1b). Given the exponential increase in computational expense, using an exchange frequency of 1/10 for the CfCBM4-1-NRE free energy calculation was not warranted.

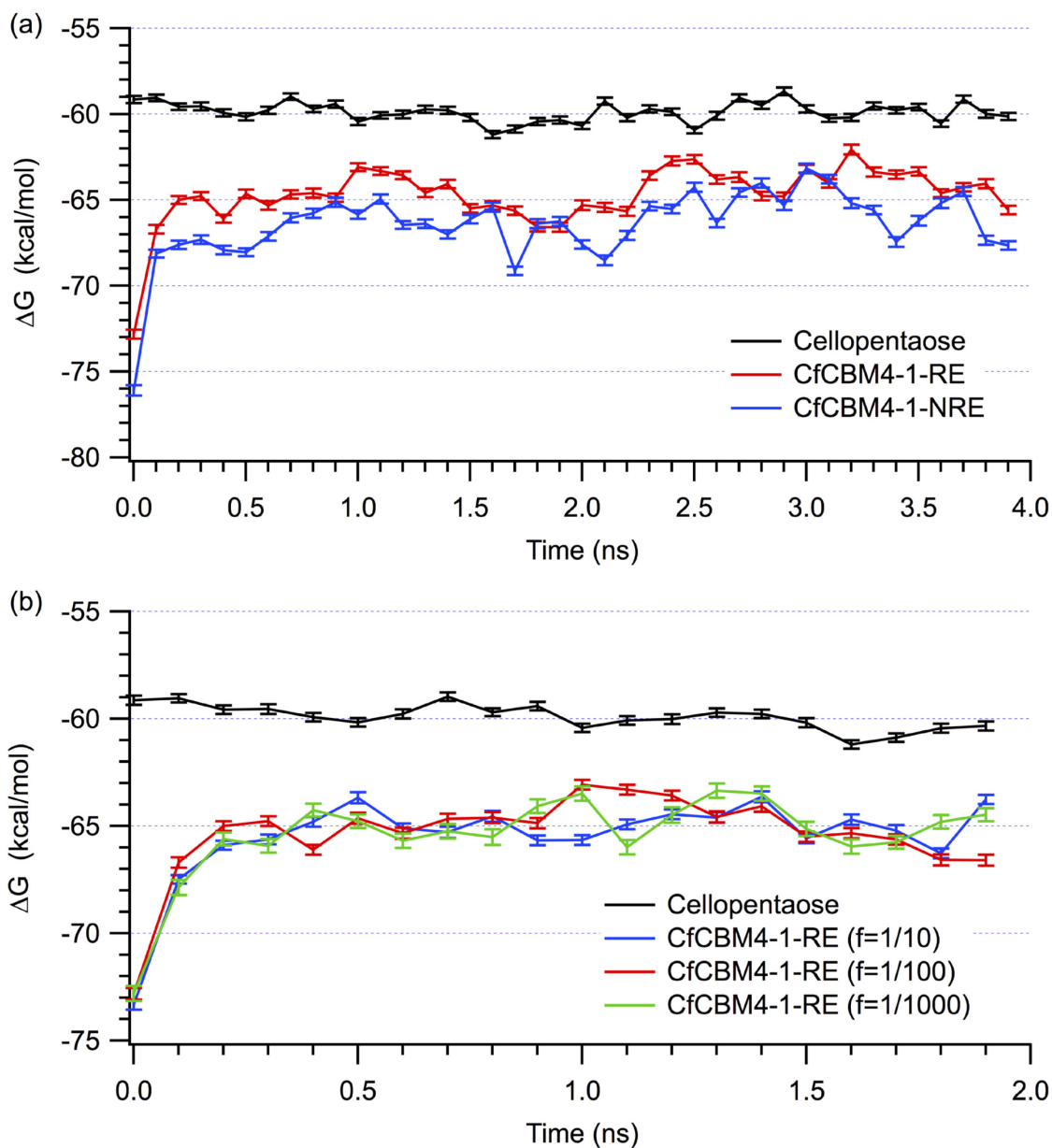


Figure A1.1 Calculated Gibbs free energy over 40 consecutive 0.1-ns calculations using FEP/ λ -REMD. (a) The difference between the average value for either *CfCBM4-1-RE* or *CfCBM4-1-NRE* and the cellopentaose solvation free energy represents the binding free energy. (b) The free energy calculation for *CfCBM4-1-RE* system using different replica exchange frequencies (in legends f = frequency) of 1/10 steps, 1/100 steps and 1/1000 steps.

A2 Supporting information related to YKL-40

Appendix A2 reports the details of force-field development required to setup the MD simulation. Information in this appendix has been adapted with permission from Kognole and Payne [223], Copyright © 2017, American Society for Biochemistry and Molecular Biology.

A2.1 Force-field parameterization for modeling heparin

Modeling of heparin in this study required development of new force-field parameters for a sugar, where the acetyl group in N-acetyl glucosamine was replaced by SO_3^{-1} (Figure A2.1). ParamChem was used to obtain an initial set of parameters based on analogy with available data [263, 268]. As the sulfamate anions were not explicitly supported, parameters obtained for $-\text{NHSO}_3$ group by analogy required optimization. The Force Field Toolkit (ffTK) Plugin Version 1.0 in VMD [238] was used to optimize the partial charges, bonds, angles, and dihedrals as described in the reference publication and provided examples. Parameters obtained using this approach are given in Table S2.

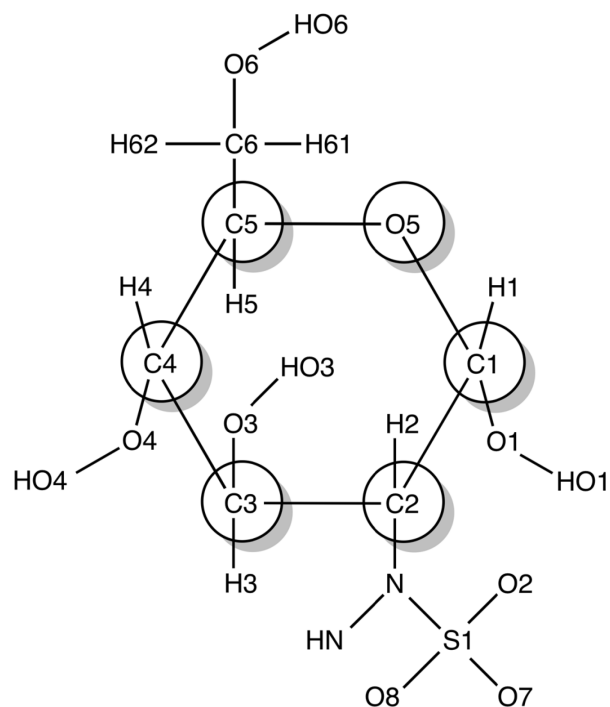


Figure A2.1 Atom labels of N-sulfo- α -D-glucosamine structure used for optimization of missing force-field parameters. The only missing parameters were the ones around N-S1 bond as documented in Table S2.

Table A2.1 CHARMM-additive parameters for GlcNS optimized using the fftK v.1.0 plugin in VMD. The atom labels are as illustrated in Figure A2.1.

Bonds	K_b		b₀
C2 – N	271.158		1.464
N – S1	332.175		1.823
N – HN	440.214		1.029
S1 – O2	540.346		1.452
Angles	K_{theta}		Theta₀
C1/C3 – C2 – N	91.721		112.507
N – C2 – H2	114.884		111.824
C2 – N – S1	124.591		117.44
C2 – N – HN	79.624		107.895
S1 – N – HN	74.629		129.979
N – S1 – O2	152.857		109.282
O2 – S1 – O7	103.66		105.957
Dihedrals	K_{chi}	n	Delta
N – C2 – C1 – O5	0.2	3	0
N – C2 – C3 – O3	0.2	3	0
N – C2 – C1 – O1	0.2	3	0
C4 – C3 – C2 – N	0.2	3	0
N – C2 – C3 – H3	0.2	3	0
N – C2 – C1 – H1	0.2	3	0
C1/C3 – C2 – N – S1	1.12	3	180
H2 – C2 – N – HN	0.527	3	180
H2 – C2 – N – S1	2.994	3	0
C2 – N – S1 – O2	1.048	3	180
NH – N – S1 – O2	0.831	3	0
C1/C3 – C2 – N – HN	1.575	1	0
O4* – C1 – C2 – N	0.2	3	0

*this O4 is from the glycosidic linkage this residue will be involved in.

References

1. Alberts, B., et al., *The Molecular Biology of the Cell*. Second Edition ed. 1989, New York: Garland Publishing, Inc.
2. Frietas, R.A., *Nanomedicine*. Vol. Volume I: Basic Capabilities. 1999, Georgetown, TX, USA: Landes Bioscience.
3. Nelson, D.L., M.M. Cox, and A.L. Lehninger, *Lehninger principles of biochemistry*. 2013, New York: W.H. Freeman.
4. Salmon, S. and S.M. Hudson, *Crystal Morphology, Biosynthesis, and Physical Assembly of Cellulose, Chitin, and Chitosan*. Journal of Macromolecular Science, Part C, 1997. **37**(2): p. 199-276.
5. Klemm, D., et al., *Cellulose: Fascinating Biopolymer and Sustainable Raw Material*. Angewandte Chemie International Edition, 2005. **44**(22): p. 3358-3393.
6. Lasky, L.A., *SELECTIN-CARBOHYDRATE INTERACTIONS AND THE INITIATION OF THE INFLAMMATORY RESPONSE*. Annual Review of Biochemistry, 1995. **64**: p. 113-139.
7. Karlsson, K.A., *Bacterium-host protein-carbohydrate interactions and pathogenicity*. Biochemical Society Transactions, 1999. **27**(4): p. 471.
8. Sacchettini, J.C., L.G. Baum, and C.F. Brewer, *Multivalent Protein–Carbohydrate Interactions. A New Paradigm for Supramolecular Assembly and Signal Transduction*. Biochemistry, 2001. **40**(10): p. 3009-3015.
9. Sharon, N. and H. Lis, *History of lectins: from hemagglutinins to biological recognition molecules*. Glycobiology, 2004. **14**(11): p. 53R-62R.
10. Davies, G.J., T.M. Gloster, and B. Henrissat, *Recent structural insights into the expanding world of carbohydrate-active enzymes*. Current Opinions in Structural Biology, 2005. **15**(6): p. 637-45.
11. Geijtenbeek, T.B.H., et al., *DC-SIGN, a Dendritic Cell-Specific HIV-1-Binding Protein that Enhances trans-Infection of T Cells*. Cell. **100**(5): p. 587-597.
12. Drickamer, K. and M.E. Taylor, *Biology of animal lectins*. Annual Review of Cell Biology, 1993. **9**: p. 237-64.
13. Weis, W.I. and K. Drickamer, *Structural basis of lectin-carbohydrate recognition*. Annual Review of Biochemistry, 1996. **65**: p. 441-473.

14. Lis, H. and N. Sharon, *Lectins: Carbohydrate-Specific Proteins That Mediate Cellular Recognition*. Chemical Reviews, 1998. **98**(2): p. 637-674.
15. Boraston, A.B., et al., *Structure and ligand binding of carbohydrate-binding module CsCBM6-3 reveals similarities with fucose-specific lectins and "galactose-binding" domains*. Journal of Molecular Biology, 2003. **327**(3): p. 659-69.
16. Boraston, A.B., et al., *Carbohydrate-binding modules: fine-tuning polysaccharide recognition*. Biochemical Journal, 2004. **382**: p. 769-781.
17. Lynd, L.R., et al., *Fuel Ethanol from Cellulosic Biomass*. Science, 1991. **251**(4999): p. 1318.
18. Faaij, A.P.C., *Bio-energy in Europe: changing technology choices*. Energy Policy, 2006. **34**(3): p. 322-342.
19. Shafiee, S. and E. Topal, *When will fossil fuel reserves be diminished?* Energy Policy, 2009. **37**(1): p. 181-189.
20. Brown, R.C. and T.R. Brown, *Why are We Producing Biofuels?: Shifting to the Ultimate Source of Energy*. 2012: Brownia LLC.
21. Meier, P.J., et al., *Potential for Electrified Vehicles to Contribute to U.S. Petroleum and Climate Goals and Implications for Advanced Biofuels*. Environmental Science & Technology, 2015. **49**(14): p. 8277-8286.
22. Robertson, G.P., et al., *Cellulosic biofuel contributions to a sustainable energy future: Choices and outcomes*. Science, 2017. **356**(6345).
23. Tilman, D., J. Hill, and C. Lehman, *Carbon-Negative Biofuels from Low-Input High-Diversity Grassland Biomass*. Science, 2006. **314**(5805): p. 1598.
24. Lynd, L.R., *OVERVIEW AND EVALUATION OF FUEL ETHANOL FROM CELLULOSIC BIOMASS: Technology, Economics, the Environment, and Policy*. Annual Review of Energy and the Environment, 1996. **21**(1): p. 403-465.
25. Hayes, D.J.M., *Second-generation biofuels: why they are taking so long*. Wiley Interdisciplinary Reviews: Energy and Environment, 2013. **2**(3): p. 304-334.
26. Brown, T.R. and R.C. Brown, *A review of cellulosic biofuel commercial-scale projects in the United States*. Biofuels, Bioproducts and Biorefining, 2013. **7**(3): p. 235-245.
27. Gelfand, I., et al., *Sustainable bioenergy production from marginal lands in the US Midwest*. Nature, 2013. **493**(7433): p. 514-7.

28. Sun, Y. and J.Y. Cheng, *Hydrolysis of lignocellulosic materials for ethanol production: a review*. Bioresource Technology, 2002. **83**(1): p. 1-11.
29. Himmel, M.E., et al., *Biomass recalcitrance: Engineering plants and enzymes for biofuels production*. Science, 2007. **315**(5813): p. 804-807.
30. Merino, S.T. and J. Cherry, *Progress and challenges in enzyme development for biomass utilization*. Advances in Biochemical Engineering/Biotechnology, 2007. **108**: p. 95-120.
31. Wyman, C.E., *What is (and is not) vital to advancing cellulosic ethanol*. Trends Biotechnol, 2007. **25**(4): p. 153-7.
32. Rouvinen, J., et al., *3-Dimensional Structure of Cellobiohydrolase-Ii from Trichoderma-Reesei*. Science, 1990. **249**(4967): p. 380-386.
33. Lynd, L.R., et al., *Microbial cellulose utilization: Fundamentals and biotechnology*. Microbiology and Molecular Biology Reviews, 2002. **66**(3): p. 506-577.
34. Levy, I. and O. Shoseyov, *Cellulose-binding domains: biotechnological applications*. Biotechnology Advances, 2002. **20**(3-4): p. 191-213.
35. Shoseyov, O., Z. Shani, and I. Levy, *Carbohydrate binding modules: Biochemical properties and novel applications*. Microbiology and Molecular Biology Reviews, 2006. **70**(2): p. 283-+.
36. Yamazaki, N., et al., *Endogenous lectins as targets for drug delivery*. Advanced Drug Delivery Reviews, 2000. **43**(2): p. 225-244.
37. Rudiger, H., et al., *Medicinal Chemistry Based on the Sugar Code: Fundamentals of Lectinology and Experimental Strategies with Lectins as Targets*. Current Medicinal Chemistry, 2000. **7**(4): p. 389-416.
38. Kang, B., et al., *Carbohydrate nanocarriers in biomedical applications: Functionalization and construction*. Chemical Society Reviews, 2015. **44**(22): p. 8301-8325.
39. Zhang, H., Y. Ma, and X.L. Sun, *Recent developments in carbohydrate-decorated targeted drug/gene delivery*. Medicinal Research Reviews, 2010. **30**(2): p. 270-289.
40. Gabius, H.-J. and J. Roth, *An introduction to the sugar code*. Histochemistry and Cell Biology, 2017. **147**(2): p. 111-117.

41. Bies, C., C.-M. Lehr, and J.F. Woodley, *Lectin-mediated drug targeting: history and applications*. Advanced Drug Delivery Reviews, 2004. **56**(4): p. 425-435.
42. Johansen, J.S., *Studies on serum YKL-40 as a biomarker in diseases with inflammation, tissue remodelling, fibroses and cancer*. Danish Medical Bulletin, 2006. **53**(2): p. 172-209.
43. Houston, D.R., et al., *Structure and ligand-induced conformational change of the 39-kDa glycoprotein from human articular chondrocytes*. Journal of Biological Chemistry, 2003. **278**(32): p. 30206-30212.
44. Fusetti, F., et al., *Crystal structure and carbohydrate-binding properties of the human cartilage glycoprotein-39*. Journal of Biological Chemistry, 2003. **278**(39): p. 37753-37760.
45. Bigg, H.F., et al., *The mammalian chitinase-like lectin, YKL-40, binds specifically to type I collagen and modulates the rate of type I collagen fibril formation*. Journal of Biological Chemistry, 2006. **281**(30): p. 21082-21095.
46. Bayer, E.A. and R. Lamed, *Ultrastructure of the cell-surface cellulosome of Clostridium thermocellum and its interaction with cellulose*. Journal of Bacteriology, 1986. **167**(3): p. 828-836.
47. Doi, R.H. and Y. Tamaru, *The Clostridium cellulovorans cellulosome: An enzyme complex with plant cell wall degrading activity*. Chemical Record, 2001. **1**(1): p. 24-32.
48. Teeri, T.T., et al., *Homologous domains in Trichoderma reesei cellulolytic enzymes - gene sequence and expression of Cellobiohydrolase-II*. Gene, 1987. **51**(1): p. 43-52.
49. Boraston, A.B., et al., *Carbohydrate-binding modules: diversity of structure and function.*, in *Recent Advances in Carbohydrate Bioengineering*, H.J. Gilbert, et al., Editors. 1999, Royal Society of Chemistry: Cambridge. p. 202-211.
50. Linder, M. and T.T. Teeri, *The roles and function of cellulose-binding domains*. Journal of Biotechnology, 1997. **57**(1-3): p. 15-28.
51. Bolam, D.N., et al., *Pseudomonas cellulose-binding domains mediate their effects by increasing enzyme substrate proximity*. Biochemical Journal, 1998. **331**: p. 775-781.
52. Srisodsuk, M., et al., *Trichoderma reesei cellobiohydrolase I with an endoglucanase cellulose-binding domain: action on bacterial microcrystalline cellulose*. Journal of Biotechnology, 1997. **57**(1-3): p. 49-57.

53. Teeri, T.T., et al., *Trichoderma reesei* cellobiohydrolases: why so efficient on crystalline cellulose? *Biochemical Society Transactions*, 1998. **26**(2): p. 173-178.
54. Tomme, P., et al., *Characterization and affinity applications of cellulose-binding domains*. *Journal of Chromatography B*, 1998. **715**(1): p. 283-296.
55. Reyes-Ortiz, V., et al., *Addition of a carbohydrate-binding module enhances cellulase penetration into cellulose substrates*. *Biotechnology for Biofuels*, 2013. **6**: p. 93-93.
56. Tomme, P., et al., *Comparison of a Fungal (Family-I) and Bacterial (Family-II) Cellulose-Binding Domain*. *Journal of Bacteriology*, 1995. **177**(15): p. 4356-4363.
57. Crouch, L.I., et al., *The Contribution of Non-catalytic Carbohydrate Binding Modules to the Activity of Lytic Polysaccharide Monooxygenases*. *The Journal of Biological Chemistry*, 2016. **291**(14): p. 7439-7449.
58. Kljun, A., et al., *Comparative Analysis of Crystallinity Changes in Cellulose I Polymers Using ATR-FTIR, X-ray Diffraction, and Carbohydrate-Binding Module Probes*. *Biomacromolecules*, 2011. **12**(11): p. 4121-4126.
59. Boraston, A.B., et al., *Binding specificity and thermodynamics of a family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A*. *Biochemistry*, 2001. **40**(21): p. 6240-6247.
60. Notenboom, V., et al., *Crystal structures of the family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A in native and ligand-bound forms*. *Biochemistry*, 2001. **40**(21): p. 6248-6256.
61. Johnson, P.E., et al., *The cellulose-binding domains from *Cellulomonas fimi* beta-1, 4-glucanase CenC bind nitroxide spin-labeled cellooligosaccharides in multiple orientations*. *Journal of Molecular Biology*, 1999. **287**(3): p. 609-25.
62. Boraston, A.B., et al., *Differential oligosaccharide recognition by evolutionarily-related β -1,4 and β -1,3 glucan-binding modules*. *Journal of Molecular Biology*, 2002. **319**(5): p. 1143-1156.
63. Carrard, G., et al., *Cellulose-binding domains promote hydrolysis of different sites on crystalline cellulose*. *Proceedings of the National Academy of Sciences of the United States of America*, 2000. **97**(19): p. 10342-10347.
64. Boraston, A.B., et al., *Recognition and hydrolysis of noncrystalline cellulose*. *Journal of Biological Chemistry*, 2003. **278**(8): p. 6120-6127.

65. Araki, Y., et al., *Family 17 and 28 Carbohydrate-Binding Modules Discriminated Different Cell-Wall Sites in Sweet Potato Roots*. Bioscience, Biotechnology, and Biochemistry, 2010. **74**(4): p. 802-805.
66. Blake, A.W., et al., *Understanding the biological rationale for the diversity of cellulose-directed carbohydrate-binding modules in prokaryotic enzymes*. Journal of Biological Chemistry, 2006. **281**(39): p. 29321-29329.
67. Din, N., et al., *C-I-C-X Revisited - Intramolecular Synergism in a Cellulase*. Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(24): p. 11383-11387.
68. Din, N., et al., *Non-Hydrolytic Disruption of Cellulose Fibers by the Binding Domain of a Bacterial Cellulase*. Bio-Technology, 1991. **9**(11): p. 1096-1099.
69. Wang, L., Y. Zhang, and P. Gao, *A novel function for the cellulose binding module of cellobiohydrolase I*. Science in China Series C: Life Sciences, 2008. **51**(7): p. 620-629.
70. Giardina, T., et al., *Both binding sites of the starch-binding domain of Aspergillus niger glucoamylase are essential for inducing a conformational change in amylose* Edited by R. Huber. Journal of Molecular Biology, 2001. **313**(5): p. 1149-1159.
71. Gao, P.J., et al., *Non-hydrolytic disruption of crystalline structure of cellulose by cellulose binding domain and linker sequence of cellobiohydrolase I from Penicillium janthinellum*. Acta Biochimica Et Biophysica Sinica, 2001. **33**(1): p. 13-18.
72. Stahlberg, J., G. Johansson, and G. Pettersson, *A New Model for Enzymatic-Hydrolysis of Cellulose Based on the 2-Domain Structure of Cellobiohydrolase-I*. Bio-Technology, 1991. **9**(3): p. 286-290.
73. McLean, B.W., et al., *Carbohydrate-binding modules recognize fine substructures of cellulose*. Journal of Biological Chemistry, 2002. **277**(52): p. 50245-50254.
74. Lombard, V., et al., *The carbohydrate-active enzymes database (CAZy) in 2013*. Nucleic Acids Research, 2014. **42**(Database issue): p. D490-5.
75. Gilbert, H.J., J.P. Knox, and A.B. Boraston, *Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules*. Current Opinion in Structural Biology, 2013. **23**(5): p. 669-677.
76. Simpson, P.J., et al., *The structural basis for the ligand specificity of family 2 carbohydrate-binding modules*. Journal of Biological Chemistry, 2000. **275**(52): p. 41137-41142.

77. Sakon, J., et al., *Structure and mechanism of endo/exocellulase E4 from Thermomonospora fusca*. Nature Structural Biology, 1997. **4**(10): p. 810-818.
78. McNeil, M., et al., *Structure and Function of the Primary Cell Walls of Plants*. Annual Review of Biochemistry, 1984. **53**(1): p. 625-663.
79. Carpita, N.C. and D.M. Gibeaut, *Structural Models of Primary-Cell Walls in Flowering Plants - Consistency of Molecular-Structure with the Physical-Properties of the Walls during Growth*. Plant Journal, 1993. **3**(1): p. 1-30.
80. Payne, C.M., et al., *Fungal cellulases*. Chemical Reviews, 2015. **115**(3): p. 1308–1448.
81. Nishiyama, Y., P. Langan, and H. Chanzy, *Crystal structure and hydrogen-bonding system in cellulose I beta from synchrotron X-ray and neutron fiber diffraction*. Journal of the American Chemical Society, 2002. **124**(31): p. 9074-9082.
82. Nishiyama, Y., et al., *Crystal structure and hydrogen bonding system in cellulose I(alpha), from synchrotron X-ray and neutron fiber diffraction*. Journal of the American Chemical Society, 2003. **125**(47): p. 14300-14306.
83. Langan, P., Y. Nishiyama, and H. Chanzy, *X-ray Structure of Mercerized Cellulose II at 1 Å Resolution*. Biomacromolecules, 2001. **2**(2): p. 410-416.
84. Wada, M., et al., *Cellulose III Crystal Structure and Hydrogen Bonding by Synchrotron X-ray and Neutron Fiber Diffraction*. Macromolecules, 2004. **37**(23): p. 8548-8555.
85. Brown, R.M., *Cellulose structure and biosynthesis: What is in store for the 21st century?* Journal of Polymer Science Part A: Polymer Chemistry, 2004. **42**(3): p. 487-495.
86. Kim, I.J., et al., *Synergistic proteins for the enhanced enzymatic hydrolysis of cellulose by cellulase*. Applied Microbiology and Biotechnology, 2014. **98**(20): p. 8469-8480.
87. Taherzadeh, M.J. and K. Karimi, *Enzyme-Based Hydrolysis Processes for Ethanol from Lignocellulosic Materials: A Review*. Bioresources, 2007. **2**(4): p. 707-738.
88. Atalla, R.H. and D.L. Vanderhart, *NATIVE CELLULOSE - A COMPOSITE OF 2 DISTINCT CRYSTALLINE FORMS*. Science, 1984. **223**(4633): p. 283-285.
89. Baker, A.A., et al., *New insight into cellulose structure by atomic force microscopy shows the I-alpha crystal phase at near-atomic resolution*. Biophysical Journal, 2000. **79**(2): p. 1139-1145.

90. Kondo, T., E. Togawa, and R.M. Brown, Jr., *"Nematic ordered cellulose": a concept of glucan chain association*. *Biomacromolecules*, 2001. **2**(4): p. 1324-30.
91. Kulasinski, K., et al., *A comparative molecular dynamics study of crystalline, paracrystalline and amorphous states of cellulose*. *Cellulose*, 2014. **21**(3): p. 1103-1116.
92. Ciolacu, D., F. Ciolacu, and V.I. Popa, *Amorphous Cellulose - Structure and Characterization*. *Cellulose Chemistry and Technology*, 2011. **45**(1-2): p. 13-21.
93. Brown, R.M., *Cellulose and other natural polymer systems: biogenesis, structure, and degradation*. 2013: Springer Science & Business Media.
94. McCartney, L., et al., *Glycoside hydrolase carbohydrate-binding modules as molecular probes for the analysis of plant cell wall polymers*. *Analytical Biochemistry*, 2004. **326**(1): p. 49-54.
95. Pell, G., et al., *Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from *Cellvibrio japonicus* Xyn10C*. *Biochemistry*, 2003. **42**(31): p. 9316-9323.
96. Boraston, A.B., et al., *Identification and glucan-binding properties of a new carbohydrate-binding module family*. *Biochemical Journal*, 2002. **361**: p. 35-40.
97. Boraston, A.B., et al., *Specificity and affinity of substrate binding by a family 17 carbohydrate-binding module from *Clostridium cellulovorans* cellulase 5A*. *Biochemistry*, 2000. **39**(36): p. 11129-11136.
98. Tomme, P., et al., *Interaction of polysaccharides with the N-terminal cellulose-binding domain of *Cellulomonas fimi* CenC .1. Binding specificity and calorimetric analysis*. *Biochemistry*, 1996. **35**(44): p. 13885-13894.
99. Kormos, J., et al., *Binding site analysis of cellulose binding domain CBDN1 from endoglucanase C of *Cellulomonas fimi* by site-directed mutagenesis*. *Biochemistry*, 2000. **39**(30): p. 8844-8852.
100. Notenboom, V., et al., *Recognition of cello-oligosaccharides by a family 17 carbohydrate-binding module: an X-ray crystallographic, thermodynamic and mutagenic study*. *Journal of Molecular Biology*, 2001. **314**(4): p. 797-806.
101. Tsukimoto, K., et al., *Recognition of cellooligosaccharides by a family 28 carbohydrate-binding module*. *Febs Letters*, 2010. **584**(6): p. 1205-1211.
102. Araki, Y., et al., *Characterization of family 17 and family 28 carbohydrate-binding modules from *Clostridium josui* Cel5A*. *Bioscience, Biotechnology, and Biochemistry*, 2009. **73**(5): p. 1028-32.

103. Bolam, D.N., et al., *Evidence for synergy between family 2b carbohydrate binding modules in Cellulomonas fimi xylanase IIA*. Biochemistry, 2001. **40**(8): p. 2468-2477.
104. Boraston, A.B., et al., *Co-operative binding of triplicate carbohydrate-binding modules from a thermophilic xylanase*. Molecular Microbiology, 2002. **43**(1): p. 187-194.
105. Coutinho, J.B., et al., *Nucleotide-sequence of the endoglucanase-C gene (Cenc) of Cellulomonas fimi, its high-level expression in Escherichia coli, and characterization of its products*. Molecular Microbiology, 1991. **5**(5): p. 1221-1233.
106. Jamal, S., et al., *X-ray crystal structure of a non-crystalline cellulose-specific carbohydrate-binding module: CBM28*. Journal of Molecular Biology, 2004. **339**(2): p. 253-258.
107. Johnson, P.E., et al., *Interaction of soluble cellooligosaccharides with the N-terminal cellulose-binding domain of Cellulomonas fimi CenC .2. NMR and ultraviolet absorption spectroscopy*. Biochemistry, 1996. **35**(44): p. 13895-13906.
108. Johnson, P.E., et al., *Structure of the N-terminal cellulose-binding domain of Cellulomonas fimi CenC determined by nuclear magnetic resonance spectroscopy*. Biochemistry, 1996. **35**(45): p. 14381-14394.
109. Lemieux, R.U., *The Origin of the Specificity in the Recognition of Oligosaccharides by Proteins*. Chemical Society Reviews, 1989. **18**(3): p. 347-374.
110. Brun, E., et al., *Structure and binding specificity of the second N-terminal cellulose-binding domain from Cellulomonas fimi endoglucanase C*. Biochemistry, 2000. **39**(10): p. 2445-2458.
111. Cao, R.Y., Y.D. Jin, and D.G. Xu, *Recognition of Cello-Oligosaccharides by CBM17 from Clostridium cellulovorans: Molecular Dynamics Simulation*. Journal of Physical Chemistry B, 2012. **116**(21): p. 6087-6096.
112. Hou, T.J., et al., *Assessing the Performance of the MM/PBSA and MM/GBSA Methods. I. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations*. Journal of Chemical Information and Modeling, 2011. **51**(1): p. 69-82.
113. Payne, C.M., et al., *Glycoside hydrolase processivity Is directly related to oligosaccharide binding free energy*. Journal of the American Chemical Society, 2013. **135**(50): p. 18831-18839.

114. Coutinho, J.B., et al., *The binding of Cellulomonas fimi endoglucanase-C (Cenc) to cellulose and sephadex Is mediated by the N-terminal repeats*. Molecular Microbiology, 1992. **6**(9): p. 1243-1252.
115. Boraston, A.B., *The interaction of carbohydrate-binding modules with insoluble non-crystalline cellulose is enthalpically driven*. Biochemical Journal, 2005. **385**: p. 479-484.
116. Linder, M., et al., *Characterization of a double cellulose-binding domain - Synergistic high affinity binding to crystalline cellulose*. Journal of Biological Chemistry, 1996. **271**(35): p. 21268-21272.
117. Faibish, M., et al., *A YKL-40-neutralizing antibody blocks tumor angiogenesis and progression: A potential therapeutic agent in cancers*. Molecular Cancer Therapeutics, 2011. **10**(5): p. 742-751.
118. Francescone, R.A., et al., *Role of YKL-40 in the angiogenesis, radioresistance, and progression of glioblastoma*. Journal of Biological Chemistry, 2011. **286**(17): p. 15332-15343.
119. Ku, B.M., et al., *CHI3L1 (YKL-40) is expressed in human gliomas and regulates the invasion, growth and survival of glioma cells*. International Journal of Cancer, 2011. **128**(6): p. 1316-1326.
120. Park, J.-A., J.M. Drazen, and D.J. Tschumperlin, *The chitinase-like protein YKL-40 is secreted by airway epithelial cells at base line and in response to compressive mechanical stress*. Journal of Biological Chemistry, 2010. **285**(39): p. 29817-29825.
121. Sharif, M., et al., *Serum cartilage oligomeric matrix protein and other biomarker profiles in tibiofemoral and patellofemoral osteoarthritis of the knee*. Rheumatology, 2006. **45**(5): p. 522-526.
122. Johansen, J.S., et al., *Serum YKL-40 is increased in patients with hepatic fibrosis*. Journal of Hepatology, 2000. **32**(6): p. 911-920.
123. Volck, B., et al., *Studies on YKL-40 in knee joints of patients with rheumatoid arthritis and osteoarthritis. Involvement of YKL-40 in the joint pathology*. Osteoarthritis and Cartilage, 2001. **9**(3): p. 203-214.
124. Kirkpatrick, R.B., et al., *Induction and Expression of Human Cartilage Glycoprotein 39 in Rheumatoid Inflammatory and Peripheral Blood Monocyte-Derived Macrophages*. Experimental Cell Research, 1997. **237**(1): p. 46-54.

125. Létuvé, S., et al., *YKL-40 Is Elevated in Patients with Chronic Obstructive Pulmonary Disease and Activates Alveolar Macrophages*. The Journal of Immunology, 2008. **181**(7): p. 5167.
126. Johansen, J.S., et al., *Identification of proteins secreted by human osteoblastic cells in culture*. Journal of Bone and Mineral Research, 1992. **7**(5): p. 501-512.
127. Mizoguchi, E., *Chitinase 3-like-1 exacerbates intestinal inflammation by enhancing bacterial adhesion and invasion in colonic epithelial cells*. Gastroenterology, 2006. **130**(2): p. 398-411.
128. Nyirkos, P. and E.E. Golds, *Human synovial-cells secrete a 39-Kda protein similar to a bovine mammary protein expressed during the nonlactating period*. Biochemical Journal, 1990. **269**(1): p. 265-268.
129. Renkema, G.H., et al., *Synthesis, sorting, and processing into distinct isoforms of human macrophage chitotriosidase*. European Journal of Biochemistry, 1997. **244**(2): p. 279-285.
130. Fusetti, F., et al., *Structure of human chitotriosidase. Implications for specific inhibitor design and function of mammalian chitinase-like lectins*. Journal of Biological Chemistry, 2002. **277**(28): p. 25537-44.
131. Sun, Y.J., et al., *The crystal structure of a novel mammalian lectin, Ym1, suggests a saccharide binding site*. Journal of Biological Chemistry, 2001. **276**(20): p. 17507-14.
132. Varela, P.F., et al., *Crystal structure of imaginal disc growth factor-2. A member of a new family of growth-promoting glycoproteins from Drosophila melanogaster*. Journal of Biological Chemistry, 2002. **277**(15): p. 13229-36.
133. Henrissat, B. and G. Davies, *Structural and sequence-based classification of glycoside hydrolases*. Current Opinion in Structural Biology, 1997. **7**(5): p. 637-44.
134. Malinda, K.M., et al., *Gp38k, a Protein Synthesized by Vascular Smooth Muscle Cells, Stimulates Directional Migration of Human Umbilical Vein Endothelial Cells*. Experimental Cell Research, 1999. **250**(1): p. 168-173.
135. Lis, H. and N. Sharon, *Lectins: Carbohydrate-specific proteins that mediate cellular recognition*. Chemical Reviews, 1998. **98**(2): p. 637-674.
136. Siaens, R., et al., *(123)I-Labeled chitinase as specific radioligand for in vivo detection of fungal infections in mice*. Journal of Nuclear Medicine, 2004. **45**(7): p. 1209-16.

137. Lal, A., et al., *A public database for gene expression in human cancers*. Cancer Research, 1999. **59**(21): p. 5403-5407.
138. Lau, S.H., et al., *Clusterin plays an important role in hepatocellular carcinoma metastasis*. Oncogene, 2006. **25**(8): p. 1242-1250.
139. Wong, E.V., *Cells: Molecules and Mechanisms*. 2009, Louisville, KY, USA: Axolotl Academic Publishing Company.
140. Connor, J.R., et al., *Human cartilage glycoprotein 39 (HC gp-39) mRNA expression in adult and fetal chondrocytes, osteoblasts and osteocytes by in-situ hybridization*. Osteoarthritis and Cartilage, 2000. **8**(2): p. 87-95.
141. Hakala, B.E., C. White, and A.D. Recklies, *Human cartilage gp-39, a major secretory product of articular chondrocytes and synovial-cells, is a mammalian member of a chitinase protein family*. Journal of Biological Chemistry, 1993. **268**(34): p. 25803-25810.
142. Harvey, S., et al., *Chondrex: new marker of joint disease*. Clinical Chemistry, 1998. **44**(3): p. 509-516.
143. Johansen, J.S., et al., *Serum YKL-40 levels in health children and adults. Comparison with serum and synovial fluid levels of YKL-40 in patients with osteoarthritis or trauma of the knee joint*. British Journal of Rheumatology, 1996. **35**(6): p. 553-559.
144. Johansen, J.S., et al., *Serum YKL-40, a new prognostic biomarker in cancer patients?* Cancer Epidemiology Biomarkers & Prevention, 2006. **15**(2): p. 194-202.
145. Martí-Renom, M.A., et al., *Comparative Protein Structure Modeling of Genes and Genomes*. Annual Review of Biophysics and Biomolecular Structure, 2000. **29**(1): p. 291-325.
146. Cavasotto, C.N. and S.S. Phatak, *Homology modeling in drug discovery: current trends and applications*. Drug Discovery Today, 2009. **14**(13): p. 676-683.
147. Yarnitzky, T., A. Levit, and M.Y. Niv, *Homology modeling of G-protein-coupled receptors with X-ray structures on the rise*. Current Opinion in Drug Discovery & Development, 2010. **13**(3): p. 317-325.
148. Arnold, K., et al., *The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling*. Bioinformatics, 2006. **22**(2): p. 195-201.
149. Kiefer, F., et al., *The SWISS-MODEL Repository and associated resources*. Nucleic Acids Research, 2009. **37**(suppl_1): p. D387-D392.

150. Biasini, M., et al., *SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information*. Nucleic Acids Research, 2014. **42**(W1): p. W252-W258.
151. Benkert, P., M. Künzli, and T. Schwede, *QMEAN server for protein model quality estimation*. Nucleic Acids Research, 2009. **37**(suppl_2): p. W510-W514.
152. Hasegawa, H. and L. Holm, *Advances and pitfalls of protein structural alignment*. Current Opinion in Structural Biology, 2009. **19**(3): p. 341-348.
153. Fischer, D., et al., *A geometry-based suite of molecular docking processes*. Journal of Molecular Biology, 1995. **248**(2): p. 459-477.
154. Duhovny, D., R. Nussinov, and H.J. Wolfson, *Efficient unbound docking of rigid molecules*. Algorithms in Bioinformatics, Proceedings, 2002. **2452**: p. 185-200.
155. Halperin, I., et al., *Principles of docking: An overview of search algorithms and a guide to scoring functions*. Proteins: Structure, Function, and Bioinformatics, 2002. **47**(4): p. 409-443.
156. Schneidman-Duhovny, D., et al., *PatchDock and SymmDock: servers for rigid and symmetric docking*. Nucleic Acids Research, 2005. **33**: p. W363-W367.
157. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins*. Nature, 1977. **267**(5612): p. 585-590.
158. Karplus, M., *Molecular Dynamics Simulations of Biomolecules*. Accounts of Chemical Research, 2002. **35**(6): p. 321-323.
159. Borhani, D.W. and D.E. Shaw, *The future of molecular dynamics simulations in drug discovery*. Journal of Computer-Aided Molecular Design, 2012. **26**(1): p. 15-26.
160. Dror, R.O., et al., *Biomolecular Simulation: A Computational Microscope for Molecular Biology*. Annual Review of Biophysics, 2012. **41**(1): p. 429-452.
161. Verlet, L., *Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules*. Physical Review, 1967. **159**(1): p. 98-103.
162. Verlet, L., *Computer "Experiments" on Classical Fluids. II. Equilibrium Correlation Functions*. Physical Review, 1968. **165**(1): p. 201-214.
163. Swope, W.C., et al., *A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules*:

- Application to small water clusters*. The Journal of Chemical Physics, 1982. **76**(1): p. 637-649.
164. Allen, M.P. and D.J. Tildesley, *Computer simulation of liquids*. 2017: Oxford university press.
 165. Hockney, R.W. and J.W. Eastwood, *Computer Simulation using Particles*. 1988, Bristol: Adam Hilger.
 166. Brooks, B.R., et al., *CHARMM: The biomolecular simulation program*. Journal of Computational Chemistry, 2009. **30**(10): p. 1545-1614.
 167. Allen, M.P., *Introduction to molecular dynamics simulation*. Computational Soft Matter, 2004. **23**: p. 1-28.
 168. Adcock, S.A. and J.A. McCammon, *Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins*. Chemical Reviews, 2006. **106**(5): p. 1589-1615.
 169. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD*. Journal of Computational Chemistry, 2005. **26**(16): p. 1781-1802.
 170. Humphrey, W., A. Dalke, and K. Schulten, *VMD - Visual Molecular Dynamics*. Journal of Molecular Graphics, 1996. **14**: p. 33-38.
 171. Schrodinger, L., *The PyMOL Molecular Graphics System, Version 1.1r1*, 2010.
 172. Deng, Y.Q. and B. Roux, *Hydration of amino acid side chains: Nonpolar and electrostatic contributions calculated from staged molecular dynamics free energy simulations with explicit water molecules*. Journal of Physical Chemistry B, 2004. **108**(42): p. 16567-16576.
 173. Deng, Y.Q. and B. Roux, *Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant*. Journal of Chemical Theory and Computation, 2006. **2**(5): p. 1255-1273.
 174. Jiang, W., M. Hodoscek, and B. Roux, *Computation of absolute hydration and binding free energy with free energy perturbation distributed replica-exchange molecular dynamics*. Journal of Chemical Theory and Computation, 2009. **5**(10): p. 2583-2588.
 175. Jiang, W. and B. Roux, *Free energy perturbation hamiltonian replica-exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations*. Journal of Chemical Theory and Computation, 2010. **6**(9): p. 2559-2565.

176. Shirts, M.R. and J.D. Chodera, *Statistically optimal analysis of samples from multiple equilibrium states*. Journal of Chemical Physics, 2008. **129**(12): p. 1-10.
177. Torrie, G.M. and J.P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling*. Journal of Computational Physics, 1977. **23**(2): p. 187-199.
178. Kästner, J., *Umbrella sampling*. Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**(6): p. 932-942.
179. Virnau, P. and M. Müller, *Calculation of free energy through successive umbrella sampling*. The Journal of Chemical Physics, 2004. **120**(23): p. 10925-10930.
180. Kumar, S., et al., *THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method*. Journal of Computational Chemistry, 1992. **13**(8): p. 1011-1021.
181. Grossfield, A., *WHAM: the weighted histogram analysis method , version 2.0.9*, in <http://membrane.urmc.rochester.edu/content/wham>.
182. Mills, M. and I. Andricioaei, *An experimentally guided umbrella sampling protocol for biomolecules*. The Journal of Chemical Physics, 2008. **129**(11): p. 114101.
183. Efron, B. and R.J. Tibshirani, *An introduction to the bootstrap*. 1994: CRC press.
184. Kognole, A.A. and C.M. Payne, *Cello-oligomer-binding dynamics and directionality in family 4 carbohydrate-binding modules*. Glycobiology, 2015. **25**(10): p. 1100-1111.
185. Frka-Petesic, B., B. Jean, and L. Heux, *First experimental evidence of a giant permanent electric-dipole moment in cellulose nanocrystals*. Epl, 2014. **107**(2).
186. Sugiyama, J., H. Chanzy, and G. Maret, *Orientation of cellulose microcrystals by strong magnetic-fields*. Macromolecules, 1992. **25**(16): p. 4232-4234.
187. Anandakrishnan, R., B. Aguilar, and A.V. Onufriev, *H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations*. Nucleic Acids Research, 2012. **40**(Web Server issue): p. W537-41.
188. Myers, J., et al., *A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules*. Proteins-Structure Function and Bioinformatics, 2006. **63**(4): p. 928-938.

189. Gordon, J.C., et al., *H++: a server for estimating pK(a)s and adding missing hydrogens to macromolecules*. Nucleic Acids Research, 2005. **33**: p. W368-W371.
190. Onufriev, A., et al. *H++*, V. 3.1. Available from: <http://biophysics.cs.vt.edu/H++>.
191. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins*. Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.
192. MacKerell, A.D., M. Feig, and C.L. Brooks, *Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations*. Journal of Computational Chemistry, 2004. **25**(11): p. 1400-1415.
193. Guvench, O., et al., *Additive empirical force field for hexopyranose monosaccharides*. Journal of Computational Chemistry, 2008. **29**(15): p. 2543-2564.
194. Guvench, O., et al., *CHARMM additive all-atom force field for glycosidic linkages between hexopyranoses*. Journal of Chemical Theory and Computation, 2009. **5**(9): p. 2353-2370.
195. Guvench, O., et al., *CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate-protein modeling*. Journal of Chemical Theory and Computation, 2011. **7**(10): p. 3162-3180.
196. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water*. Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
197. Durell, S.R., B.R. Brooks, and A. Bennaïm, *Solvent-induced forces between 2 hydrophilic groups*. Journal of Physical Chemistry, 1994. **98**(8): p. 2198-2202.
198. Wang, J., Y. Deng, and B. Roux, *Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials*. Biophysical Journal, 2006. **91**(8): p. 2798-814.
199. Baranauskienė, L., et al., *Titration calorimetry standards and the precision of isothermal titration calorimetry data*. International Journal of Molecular Sciences, 2009. **10**(6): p. 2752-2762.
200. Johnson, P.E., et al., *Calcium binding by the N-terminal cellulose-binding domain from *Cellulomonas fimi* beta-1,4-glucanase CenC*. Biochemistry, 1998. **37**(37): p. 12772-12781.

201. Johnson, P.E., et al., *Interaction of soluble cellooligosaccharides with the N-terminal cellulose-binding domain of Cellulomonas fimi CenC .2. NMR and ultraviolet absorption spectroscopy*. Biochemistry, 1996a. **35**(44): p. 13895-13906.
202. Alahuhta, M., et al., *The unique binding mode of cellulosomal CBM4 from Clostridium thermocellum cellobiohydrolase A*. Journal of Molecular Biology, 2010. **402**(2): p. 374-87.
203. Szabo, L., et al., *Structure of a family 15 carbohydrate-binding module in complex with xylopentaose. Evidence that xylan binds in an approximate 3-fold helical conformation*. Journal of Biological Chemistry, 2001. **276**(52): p. 49061-5.
204. Luis Asensio, J., et al., *Carbohydrate-aromatic interactions*. Accounts of Chemical Research, 2013. **46**(4): p. 946-954.
205. Wimmerova, M., et al., *Stacking interactions between carbohydrate and protein quantified by combination of theoretical and experimental methods*. Plos One, 2012. **7**(10): p. doi:10.1371/journal.pone.0046032.
206. Richardson, J.S., *The anatomy and taxonomy of protein structure*. Advances in Protein Chemistry, 1981. **34**: p. 167-339.
207. Tormo, J., et al., *Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose*. The EMBO Journal, 1996. **15**(21): p. 5739-51.
208. McLean, B.W., et al., *Analysis of binding of the family 2a carbohydrate-binding module from Cellulomonas fimi xylanase 10A to cellulose: specificity and identification of functionally important amino acid residues*. Protein Engineering, 2000. **13**(11): p. 801-809.
209. Beckham, G.T., et al., *Molecular-level origins of biomass recalcitrance: decrystallization free energies for four common cellulose polymorphs*. Journal of Physical Chemistry B, 2011. **115**(14): p. 4118-4127.
210. Payne, C.M., et al., *Decrystallization of oligosaccharides from the cellulose I beta surface with molecular simulation*. Journal of Physical Chemistry Letters, 2011. **2**(13): p. 1546-1550.
211. Beglov, D. and B. Roux, *Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations*. The Journal of Chemical Physics, 1994. **100**(12): p. 9050-9063.
212. Hoover, W.G., *Canonical dynamics - equilibrium phase-space distributions*. Physical Review A, 1985. **31**(3): p. 1695-1697.

213. Nose, S. and M.L. Klein, *Constant pressure molecular-dynamics for molecular-systems*. Molecular Physics, 1983. **50**(5): p. 1055-1076.
214. Schneider, T. and E. Stoll, *Molecular-dynamics study of a 3-dimensional one-component model for distortive phase-transitions*. Physical Review B, 1978. **17**(3): p. 1302-1322.
215. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical-integration of cartesian equations of motion of a system with constraints - molecular-dynamics of N-alkanes*. Journal of Computational Physics, 1977. **23**(3): p. 327-341.
216. Essmann, U., et al., *A smooth particle mesh Ewald method*. Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.
217. Hamre, A.G., et al., *Thermodynamic relationships with processivity in Serratia marcescens family 18 chitinases*. Journal of Physical Chemistry B, 2015. **119**(30): p. 9601-9613.
218. Creagh, A.L., et al., *Stability and oligosaccharide binding of the N1 cellulose-binding domain of Cellulomonas fimi endoglucanase CenC*. Biochemistry, 1998. **37**(10): p. 3529-3537.
219. Bronowska, A.K., *Thermodynamics of ligand protein interactions implications for molecular design*. Thermodynamics-Interaction Studies - Solids, Liquids and Gases. 2011: InTech. 1-48.
220. Emsley, J., *Very strong hydrogen bonding*. Chemical Society Reviews, 1980. **9**(1): p. 91-124.
221. Cockburn, D. and B. Svensson, *Surface binding sites in carbohydrate active enzymes: an emerging picture of structural and functional diversity*, in *Carbohydrate Chemistry*. 2013, The Royal Society of Chemistry. p. 204-221.
222. Payne, C.M., et al., *Glycosylated linkers in multimodular lignocellulose-degrading enzymes dynamically bind to cellulose*. Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(36): p. 14646-14651.
223. Kognole, A.A. and C.M. Payne, *Inhibition of Mammalian Glycoprotein YKL-40: IDENTIFICATION OF THE PHYSIOLOGICAL LIGAND*. Journal of Biological Chemistry, 2017. **292**(7): p. 2624-2636.
224. Zaheer-ul-Haq, et al., *Family 18 chitolectins: comparison of MGP40 and HUMGP39*. Biochemical and Biophysical Research Communications, 2007. **359**(2): p. 221-226.

225. Mohanty, A.K., et al., *Crystal structure of a novel regulatory 40-kDa mammary gland protein (MGP-40) secreted during involution*. Journal of Biological Chemistry, 2003. **278**(16): p. 14451-14460.
226. De Ceuninck, F., et al., *Purification of guinea pig YKL40 and modulation of its secretion by cultured articular chondrocytes*. Journal of Cellular Biochemistry, 1998. **69**(4): p. 414-424.
227. Iozzo, R.V., *Matrix proteoglycans: From molecular design to cellular function*. Annual Review of Biochemistry, 1998. **67**(1): p. 609-652.
228. Rabenstein, D.L., *Heparin and heparan sulfate: structure and function*. Natural Product Reports, 2002. **19**(3): p. 312-331.
229. Mikami, T. and H. Kitagawa, *Biosynthesis and function of chondroitin sulfate*. Biochimica et biophysica acta, 2013. **1830**(10): p. 4719-33.
230. Malavaki, C., et al., *Recent advances in the structural study of functional chondroitin sulfate and dermatan sulfate in health and disease*. Connective Tissue Research, 2008. **49**(3-4): p. 133-139.
231. Plaas, A.H.K., et al., *Chemical and Immunological Assay of the Nonreducing Terminal Residues of Chondroitin Sulfate from Human Aggrecan*. Journal of Biological Chemistry, 1997. **272**(33): p. 20603-20610.
232. Semino, C.E., et al., *Homologs of the Xenopus developmental gene DG42 are present in zebrafish and mouse and are involved in the synthesis of Nod-like chitin oligosaccharides during early embryogenesis*. Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(10): p. 4548-4553.
233. Meyer, M.F. and G. Kreil, *Cells expressing the DG42 gene from early Xenopus embryos synthesize hyaluronan*. Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(10): p. 4543-4547.
234. Varki, A., *Biological roles of oligosaccharides - All of the theories are correct*. Glycobiology, 1993. **3**(2): p. 97-130.
235. Hardingham, T.E. and H. Muir, *The specific interaction of hyaluronic acid with cartilage proteoglycans*. Biochimica et Biophysica Acta (BBA) - General Subjects, 1972. **279**(2): p. 401-405.
236. Fraser, J.R.E., T.C. Laurent, and U.B.G. Laurent, *Hyaluronan: Its nature, distribution, functions and turnover*. Journal of Internal Medicine, 1997. **242**(1): p. 27-33.

237. Comper, W.D. and T.C. Laurent, *Physiological function of connective tissue polysaccharides*. Physiological Reviews, 1978. **58**(1): p. 255.
238. Mayne, C.G., et al., *Rapid parameterization of small molecules using the force field toolkit*. Journal of Computational Chemistry, 2013.
239. Mulloy, B., et al., *N.m.r. and molecular-modelling studies of the solution conformation of heparin*. Biochemical Journal, 1993. **293** (Pt 3): p. 849-58.
240. Angyal, S.J., *The composition and conformation of sugars in solution*. Angewandte Chemie International Edition in English, 1969. **8**(3): p. 157-166.
241. Davies, G. and B. Henrissat, *Structures and mechanisms of glycosyl hydrolases*. Structure, 1995. **3**(9): p. 853-859.
242. Davies, G.J., A. Planas, and C. Rovira, *Conformational analyses of the reaction coordinate of glycosidases*. Accounts of Chemical Research, 2012. **45**(2): p. 308-316.
243. Sinnott, M.L., *Catalytic mechanisms of enzymatic glycosyl transfer*. Chemical Reviews, 1990. **90**(7): p. 1171-1202.
244. Vocadlo, D.J. and G.J. Davies, *Mechanistic insights into glycosidase chemistry*. Current Opinion in Chemical Biology, 2008. **12**(5): p. 539-555.
245. Hamre, A.G., et al., *Processivity, substrate positioning, and binding: The role of polar residues in a family 18 glycoside hydrolase*. Biochemistry, 2015. **54**(49): p. 7292-7306.
246. Cardin, A.D. and H.J.R. Weintraub, *Molecular modeling of protein-glycosaminoglycan interactions*. Arteriosclerosis, 1989. **9**(1): p. 21-32.
247. Asensio, J.L., et al., *Carbohydrate–aromatic interactions*. Accounts of Chemical Research, 2013. **46**(4): p. 946-954.
248. Josefsson, A., et al., *Prostate cancer increases hyaluronan in surrounding nonmalignant stroma, and this response is associated with tumor growth and an unfavorable outcome*. The American Journal of Pathology, 2011. **179**(4): p. 1961-1968.
249. Aruffo, A., et al., *CD44 is the principal cell surface receptor for hyaluronate*. Cell, 1990. **61**(7): p. 1303-1313.
250. Liu, L.-K. and B. Finzel, *High-resolution crystal structures of alternate forms of the human CD44 hyaluronan-binding domain reveal a site for protein interaction*.

Acta Crystallographica. Section F, Structural Biology Communications, 2014. **70**(Pt 9): p. 1155-1161.

- 251. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.
- 252. Toole, B.P., *Hyaluronan and its binding proteins, the hyaladherins*. Current Opinion in Cell Biology, 1990. **2**(5): p. 839-844.
- 253. Bleau, G., et al., *Mammalian chitinase-like proteins*. EXS, 1999. **87**: p. 211-21.
- 254. Brodsky, B. and A.V. Persikov, *Molecular structure of the collagen triple helix*. Fibrous Proteins: Coiled-Coils, Collagen and Elastomers, 2005. **70**: p. 301-339.
- 255. Kramer, R.Z., et al., *Sequence dependent conformational variations of collagen triple-helical structure*. Nature Structural Biology, 1999. **6**(5): p. 454-457.
- 256. Okuyama, K., et al., *Crystal structure of (Gly-Pro-Hyp)₉: Implications for the collagen molecular model*. Biopolymers, 2012. **97**(8): p. 607-616.
- 257. Rich, A. and F.H. Crick, *The molecular structure of collagen*. Journal of Molecular Biology, 1961. **3**: p. 483-506.
- 258. Okuyama, K., et al., *Revision of collagen molecular structure*. Biopolymers, 2006. **84**(2): p. 181-91.
- 259. Shoulders, M.D. and R.T. Raines, *Collagen structure and stability*. Annual Review of Biochemistry, 2009. **78**(1): p. 929-958.
- 260. Bella, J., et al., *Crystal-structure and molecular-structure of a collagen-like peptide at 1.9-angstrom resolution*. Science, 1994. **266**(5182): p. 75-81.
- 261. Emsley, J., et al., *Structure of the integrin alpha 2 beta 1-binding collagen peptide*. Journal of Molecular Biology, 2004. **335**(4): p. 1019-1028.
- 262. Emsley, J., et al., *Structural basis of collagen recognition by integrin alpha 2 beta 1*. Cell, 2000. **101**(1): p. 47-56.
- 263. Vanommeslaeghe, K., et al., *CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields*. Journal of Computational Chemistry, 2010. **31**(4): p. 671-690.
- 264. Sheinerman, F.B. and C.L. Brooks, *Calculations on folding of segment B1 of streptococcal protein G*. Journal of Molecular Biology, 1998. **278**(2): p. 439-456.

- 265. Iwata, T., et al., *YKL-40 secreted from adipose tissue inhibits degradation of type I collagen*. Biochemical and Biophysical Research Communications, 2009. **388**(3): p. 511-516.
- 266. Zondlo, N.J., *Aromatic–proline interactions: electronically tunable CH/ π interactions*. Accounts of Chemical Research, 2013. **46**(4): p. 1039-1049.
- 267. *UniProt: the universal protein knowledgebase*. Nucleic Acids Research, 2017. **45**(D1): p. D158-D169.
- 268. Yu, W.B., et al., *Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations*. Journal of Computational Chemistry, 2012. **33**(31): p. 2451-2468.

Vita

Personal Information

Place of birth Jayasingpur, Maharashtra, India.

Education

2012 – Present Ph.D. Chemical Engineering, University of Kentucky, USA.

2008 – 2012 B.E. Chemical, Institute of Chemical Technology, Mumbai, India.

Honors and Awards

2017 - Outstanding Graduate Student award in Chemical Engineering by College of Engineering, University of Kentucky.

2015 - Dissertation Enhancement Award (DEA) 2015-2016 awarded by Graduate School, University of Kentucky. \$1600 of financial aid to attend workshop on Macromolecular Simulation Software organized by CECAM at Jülich, Germany.

2015 - National Science Foundation Travel Award (\$1500) to attend workshop on Macromolecular Simulation Software organized by CECAM at Jülich, Germany.

2013 - Graduate School Academic Year Fellowship for the 2013-2014 academic year awarded by College of Engineering, University of Kentucky.

Publications

1. **Kognole AA**, Payne CM (2017), “Inhibition of the Mammalian Glycoprotein YKL-40: Identification of the Physiological Ligand”. *Journal of Biological Chemistry* 292: 2624-2636.
2. **Kognole AA**, Payne CM (2015), “Cello-oligomer-binding dynamics and directionality in family 4 carbohydrate-binding modules”. *Glycobiology* 25: 1100-1111.
3. Borisova AS, Isaksen T, Dimarogona M, **Kognole AA**, Mathiesen G, et al. (2015), “Structural and Functional Characterization of a Lytic Polysaccharide Monooxygenase with Broad Substrate Specificity”. *Journal of Biological Chemistry* 290: 22955-22969.
4. Dimarogona M, **Kognole AA**, Liu B, Wu M, Westereng B, Crowley MF, Kim S, Payne CM, Sandgren M, “Crystal structure and molecular dynamics studies of an AA9 LPMO from the tree-killing fungus *Heterobasidion irregulare*”. (To be submitted to *FEBS Journal*)
5. **Kognole AA**, Payne CM, “Cellulose-specific Type B Carbohydrate Binding Modules: Understanding Substrate Recognition Mechanisms Through Molecular Simulation”. (Under preparation).