The investigation of type-specific features of the copper coordinating AA9 proteins and their effect on the interaction with crystalline cellulose using molecular dynamics studies

A thesis submitted in fulfilment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

IN BIOINFORMATICS

of

RHODES UNIVERSITY, SOUTH AFRICA

In the Department of Biochemistry and Microbiology

Faculty of Science

by

Vuyani Moses

March 2017

Abstract

AA9 proteins are metallo-enzymes which are crucial for the early stages of cellulose degradation. AA9 proteins have been suggested to cleave glycosidic bonds linking cellulose through the use of their Cu^{2+} coordinating active site. AA9 proteins possess different regioselectivities depending on the resulting cleavage they form and as result, are grouped accordingly. Type 1 AA9 proteins cleave the C1 carbon of cellulose while Type 2 AA9 proteins cleave the C4 carbon and Type 3 AA9 proteins cleave either C1 or C4 carbons. The steric congestion of the AA9 active site has been proposed to be a contributor to the observed regioselectivity. As such, a bioinformatics characterisation of type-specific sequence and structural features was performed. Initially AA9 protein sequences were obtained from the Pfam database and multiple sequence alignment was performed. The sequences were phylogenetically characterised and sequences were grouped into their respective types and sub-groups were identified. A selection analysis was performed on AA9 LPMO types to determine the selective pressure acting on AA9 protein residues. Motif discovery was then performed to identify conserved sequence motifs in AA9 proteins. Once type-specific sequence features were identified structural mapping was performed to assess possible effects on substrate interaction. Physicochemical property analysis was also performed to assess biochemical differences between AA9 LPMO types. Molecular dynamics (MD) simulations were then employed to dynamically assess the consequences of the discovered type-specific features on AA9-cellulose interaction. Due to the absence of AA9 specific force field parameters MD simulations were not readily applicable. As a result, Potential Energy Surface (PES) scans were performed to evaluate the force field parameters for the AA9 active site using the PM6 semi empirical approach and least squares fitting. A Type 1 AA9 active site was constructed from the crystal structure 4B5Q, encompassing only the Cu^{2+} coordinating residues, the Cu^{2+} ion and two water residues. Due to the similarity in AA9 active sites, the Type 1force field parameters were validated on all three AA9 LPMO types. Two MD simulations for each AA9 LPMO types were conducted using two separate Lennard-Jones parameter sets. Once completed, the MD trajectories were analysed for various features including the RMSD, RMSF, radius of gyration, coordination during simulation, hydrogen bonding, secondary structure conservation and overall protein movement. Force field parameters were successfully evaluated and validated for AA9 proteins.

MD simulations of AA9 proteins were able to reveal the presence of unique type-specific binding modes of AA9 active sites to cellulose. These binding modes were characterised by the presence of unique type-specific loops which were present in Type 2 and 3 AA9 proteins but not in Type 1 AA9 proteins. The loops were found to result in steric congestion that affects how the Cu^{2+} ion interacts with cellulose. As a result, Cu^{2+} binding to cellulose was observed for Type 1 and not Type 2 and 3 AA9 proteins. In this study force field parameters have been evaluated for the Type 1 active site of AA9 proteins and this parameters were evaluated on all three types and binding. Future work will focus on identifying the nature of the reactive oxygen species and performing QM/MM calculations to elucidate the reactive mechanism of all three AA9 LPMO types.

Declaration

I declare that this thesis is my own, unaided work, unless otherwise stated. It is being submitted for the degree of Doctor of Philosophy at Rhodes University. It has not been submitted before for any degree or examination in any other university.

Signature....  ...

Date.....03-08-2017.....

Acknowledgements

I would like to thank the supervisor of my PhD study Prof. Özlem Taştan Bishop. Her encouragement and advice has been instrumental in the completion of this work. Without her patience and guidance, this work would not have reached its maturity. I truly appreciate all the opportunities that she has provided for me to be where I am today.

A very big thank to my co-supervisor Dr. Kevin Lobb for taking the time to guide me through various aspects of computational chemistry. I appreciate the opportunity you gave me to access a different perspective about my research. I have learnt a lot from working with you and for that I am truly grateful.

Thank you to Dr Rowan Andrew Hartheley for his graciously spending his time assisting me with the analysis and writing up my results. The time I worked with you has been very insightful.

I would like to thank Prof. Brett Pletschke for his initial inputs on AA9 proteins.

Thanks to all RUBi members who I have interacted with through the course of my studies. Thanks for the friendships which were much needed through this time of my life. Special thanks to David Lawrence-Penkler, Thommas Musyoka and Olivier Sheik Amamuddy for the general discussions we would have about my work.

Thank you to the CMCCD research group for hosting me at the Chemistry Department when I needed assistance with my research.

Thank you to Ngonidzashe Faya for kindly providing the scripts I used to generate the all vs all sequence alignments and the MEME heat maps.

Special thanks to Thobeka Ethel Shibe for her support and encouragement. You have been amazing and I appreciate everything you have done to help me through my studies.

Thank you to the Natural Research Foundation for awarding me the Scarce Skills scholarship for PhD study. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Thank you to the Center of High Performance Computing (CHPC) for providing the computational resources I required for performing this study.

Dedication

"This work is dedicated to my mother Noluthando Cynthia Moses."

Table of Contents

List of Equations

List of Figures

List of Tables

List of Webservers

DataMonkey	http://www.datamonkey.org/
Genbank	https://www.ncbi.nlm.nih.gov/genbank/
Mafft	http://mafft.cbrc.jp/alignment/software/
PAL2NAL	http://www.bork.embl.de/pal2nal/
PDB	http://www.rcsb.org/pdb/home/home.do
Promls3D	http://prodata.swmed.edu/promals3d/promals3d.php
Selecton	http://selecton.tau.ac.il/

Supplementary Information

Additional files

Additional file 1: Accession numbers of the AA9 protein sequences in the dataset

Additional file 2: Nucleotide sequence alignment of Type 1 AA9 proteins

Additional file 3: Nucleotide sequence alignment of Type 2 AA9 proteins

Additional file 4: Nucleotide sequence alignment of Type 3 AA9 proteins

Additional file 5: Full sequence alignment of AA9 protein sequences

Additional file 6: Full alignment of Type 1 AA9 proteins

Additional file 7: Full alignment of Type 2 AA9 proteins

Additional file 8: Full alignment of Type 3 AA9 proteins

Additional file 9: Summary of motifs identified by MEME

Additional file 10: Amino acid counts of AA9 proteins

Scripts

Acronyms

3D Three Dimensional

AA9 Auxiliary Activity 9

AA10 Auxiliary Activity 10

AA11 Auxiliary Activity 11

AA13 Auxiliary Activity 13

CBM Carbohydrate binding module

DOPE Discrete Optimized Protein Energy

GDT_TS Global Distance Test Total Score

GH61 glycoside hydrolase family 61

HMM Hidden Markov Models

JTT Jones Taylor Thornton

LPMO Lytic Polysaccharide Monooxygenases

MAFFT Multiple Sequence Alignment based on Fast Fourier Transform

MEGA Molecular Evolutionary Genetic Analysis

MK Merz-Kollman

MQAP Model Quality Assessment Program

MSA Multiple Sequence Alignment

MUSCLE MUltiple Sequence Comparison by Log – Expectation

NCBI National Centre for Biotechnology Information

NJ Neighbour Joining

NNI Nearest Neighbour Interchange

NMR Nuclear Magnetic Resonance

PDB Protein Data Bank

PES Potential Energy Scans

RMSD Root Mean Square Deviation

RMSD Root Mean Square Fluctuation

QM/MM Quantum Mechanics and Molecular Mechanics

PROMALS PROfile Multiple Alignment with Local Structure

Symbols used

Cα- alpha carbon
Cβ - beta carbon
α -alpha
β -beta
ϕ -phi
ψ -psi

List of Amino Acids

Amino acid	Three letter code	One letter code
Alanine	**ALA**	A
Arginine	**ARG**	R
Asparagine	**ASN**	N
Aspartic Acid	**ASP**	D
Cysteine	**CYS**	C
Glutamic Acid	**GLU**	E
Glutamine	**GLN**	Q
Glycine	**GLY**	G
Histidine	**HIS**	H
Isoleucine	**ILE**	I
Leucine	**LEU**	L
Lysine	**LYS**	K
Methionine	**MET**	M
Phenylalanine	**PHE**	F
Proline	**PRO**	P
Serine	**SER**	S
Threonine	**THR**	T
Tryptophan	**TRP**	W
Tyrosine	**TYR**	Y
Valine	**VAL**	V

List of Research Outputs

Research articles

Moses, V., Hatherley, R. & Tastan Bishop, O. 2016, "**Bioinformatic characterization of type-specific sequence and structural features in auxiliary activity family 9 proteins**", *Biotechnology for biofuels,* vol. 9, pp. 239.

Moses, V., Tastan Bishop, O. & Lobb, K. 2017, **"The Evaluation and Validation of Copper (II) Force Field Parameters of the Auxiliary Activity Family 9 Enzymes",** *Chemical physics letters* , pending review

Conference attendance

Oral presentation

Vuyani Moses, Özlem Tastan Bishop, Kevin Lobb. **The determination of force field parameter of the Type 1 Copper (II) active site for the Auxiliary Activity Family 9 enzymes**. CHPC national meeting, East London, South Africa, 5-9 December 2016.

Poster presentations

Vuyani Moses, Brett Pletschke, Özlem Tastan Bishop. **The sequence and structural analysis of the Auxiliary Activity Family 9 enzymes.** *Joint SASBi-SAGS Congress*, Kwalata Game Ranch, Tshwane, South Africa, 23-26 September 2014.

Vuyani Moses, Özlem Tastan Bishop. **The Sequence and Structural Analysis of LPMO types of the Auxiliary Activity Family 9 Enzymes.** *ISMB/ECCB conference*, Dublin, Ireland, 10-14 July 2015.

Contribution to publications

Moses *et al*. 2016: **Bioinformatic characterization of type-specific sequence and structural features in auxiliary activity family 9 proteins.**

For this paper I performed all experiments and data analysis. The writing of this article was done by myself, Dr. RA Hatherley and Prof. Ö Tastan Bishop.

Moses *et al*. 2016: **The Evaluation and Validation of Copper (II) Force Field Parameters of the Auxiliary Activity Family 9 Enzymes.**

All experiments for this article were performed by myself. The data analysis and writing of this article was done by myself, Prof. Ö Tastan Bishop and Dr K Lobb.

Thesis Structure

Overview

The aim of this study was to characterise the unique features of Auxiliary Activity family 9 (AA9) LPMO types and to assess what effect these features may have on AA9 substrate interaction. This thesis consists of five Chapters (Chapter 1,2,3,4 and 5) and three other sections which are the Supplementary information, Scripts and References. Chapter 1 entails the literature review of AA9 proteins and their substrate cellulose. Chapter 2 details the type-specific characterization of sequence and structural features of AA9 proteins. Chapter 3 describes the determination of the Cu^{2+} force field parameters of the AA9 active site. Chapter 4 describes the validation of the force field parameters using Molecular Dynamics (MD) simulations. Finally, Chapter 5 is a discussion of all the findings of this study and discussion of future prospects.

Chapter 1

Included in this chapter are AA9 features which contain the modularity of AA9 proteins and the presence of regioselectivity among AA9 proteins which consequently results into their grouping as Type 1, 2 and 3 LPMOs. The presence of regioselectivity among AA9 LPMO types suggest a previously uncharacterized structural configuration of AA9 active sites that results in the observed regioselectivities. As a result, a type-specific sequence and structural investigation of AA9 features was performed.

Chapter 2

Chapter 2 details the bionformatic investigation of unique sequence and structural features of AA9 LPMO types. This investigation was achieved by obtaining a large dataset of protein sequences from the Pfam database and grouping them with *Neurospora crassa* sequences. This allowed for a type-specific characterization of AA9 protein sequences. Separating AA9 protein sequences into their respective types was achieved through aligning the AA9 protein sequences and conducting phylogenetic analysis. However, it was revealed that the dataset consisted of redundant, diverse and short fragmented sequences. As a result, these sequences were removed from the dataset

resulting in a final dataset of 153 AA9 protein sequences. Once separated the AA9 sequences were then characterized using all vs all sequence alignments, physicochemical property analysis and motif analysis. Through structural mapping all unique type-specific features were mapped onto respective AA9 crystal structure and these crystal structures were aligned on to the cellulose substrate. Type-specific features were successfully identified on AA9 protein sequences. It was found that the identified type-specific features had the potential to interact with cellulose; however this observation provided a need for further validation through MD studies. The findings of this chapter were summarized in the form of a publication (Moses, Hatherley & Tastan Bishop 2016).

Chapter 3

To assess the structural effect on the AA9 cellulose interaction of our previous findings, MD simulations needed to be performed. However, because AA9 proteins are Copper coordinating metalo-enzymes, MD simulations could not be applied in a straightforward manner. As such, force field parameter determination of AA9 parameters was performed and the results of the analysis is shown in this Chapter. A subset of the Type 1 4B5Q crystal structure was selected to perform the parameter determination. The subset was used instead of the complete crystal structure to save the computational cost associated with using large structure for Quantum Mechanics (QM) studies. Potential energy surface (PES) scans and least squares fitting was performed to determine the ideal bond stretch, angle bend and torsional parameters required for an accurate description of the AA9 active site. Other considerations included the correct Lennard-Jones parameters for the Cu^{2+} atom of the active site and the charges of the active site. As such, the restrained electrostatic potential (RESP) of the active site was calculated. The Lennard-Jones parameters for the copper active site were initially estimated using the AA9 subset. Due to the study's inability to consolidate the bonded force-field component for the Cu^{2+} to Water Oxygen interaction from the calculations, an unrealistic Lennard-Jones parameter was determined for this particular interaction. As a result the Lennard-Jones parameters for the Cu^{2+} were used from literature. The literature search yielded two Lennard-Jones parameter sets. All the resulting force field parameters were then validated using MD simulations on all three AA9 LPMO types using both Lennard-Jones parameter sets.

Chapter 4

Chapter 4 details the interaction of all three AA9 LPMO types with cellulose. MD simulations were employed to validate the use of the AA9 Type 1 force field parameters and to characterize the AA9-cullulose interaction. Due to the similarity of AA9 active sites across LPMO types, the newly generated Type 1 force field parameters were compatible with Type 2 and 3 AA9 proteins. As a result, MD simulations were conducted on all three types. The MD simulations revealed that the force field parameters that have been generated are adequate to accurately describe the AA9 active site of all three AA9 LPMO types during the course of an MD simulations. Two Lennard-Jones parameter sets were used to perform MD simulations (termed biased and unbiased). The unbiased Lennard-Jones parameter set was evaluated for Cu^{2+} in acetonitrile and the biased Lennard-Jones parameter set was evaluated in water. As a result, for each AA9 LPMO type, two MD simulations were created. Due to limitation in computation resources only 10 ns MD runs were created for each MD run of each LPMO type. For each resulting trajectory, the RMSD, RMSF, radius of Gyration and potential energy was monitored to assess the stability of the proteins during the simulations. The overall movement of the proteins was also assessed visually using VMD. It was found the AA9 proteins can be regarded as stable through the course of a simulation. AA9 proteins were also found to have motion across the cellulose substrate and Type 1 proteins were found to have potential binding to cellulose. The presence of type-specific features was found to result in the steric congestion of the AA9 active site which is a likely cause of the observed regioselectivity of AA9 proteins.

Chapter 5

In this chapter the findings obtained in Chapters 2, 3 and 4 are discussed and potential future work is proposed.

Chapter 1

Literature Review

1. Background

The imminent depletion of fossil fuels has resulted in the need for the identification of new sustainable alternative sources of energy. There are many potential substitutes being proposed for fossil fuels (Capellán-Pérez et al. 2014). Biofuels have been shown to be promising alternatives to fossil fuels, however the degradation of plant biomass to produce biofuels remains a challenge. The degradation of plant biomass has potential applications in various industries due to the fact that plant biomass is the most abundant source of carbon on planet (Mäkelä, Donofrio & de Vries 2014). The cell walls of plant biomass consists of lignocellulosic material. Lignocellulose is primarily composed of cellulose, hemicellulose pectin, and lignin. Cellulose is the major component of lignocellulose, and is a major contributor to cell wall complexity of plants. In the biofuel industry hydrolysis of lignocellulose is the rate limiting step due to the complexity of plant biomass (Himmel et al. 2007). Fungal organisms have evolved to be efficient at degrading plant biomass. This efficiency is largely attributed to the fact that fungal organisms encode enzymes which synergistically break down lignocellulose resulting in the rerelease of glucose.

1.1. The structure of cellulose

Cellulose is found in two forms: Crystalline and amorphous. When in crystalline form, the cellulose crystal is composed of long chains of glucose residues that are held together by a network of hydrogen bonding; this hydrogen bonding results in structures called microfibrils. Within the crystalline microfibrils, cellulose contains non crystalline regions - the armophous regions (Ruel, Nishiyama & Joseleau 2012). In its simplest form, a single anhydro-glucose unit is represented by the following formula ($C_6H_{10}O_5$), this represents a two ring anhydro-glucose. A covalent bond is formed between the C1 and C4 carbon of respective glucose units in an alternating manner. This results in the formation of a β 1-4 glycosidic bond between repeat units. The linear chain of repeat units adopts a flat ribbon conformation. In a cellulose chain, the number of repeat units is usually

found in the range of 10000 and 15000. Enzymatic hydrolysis of cellulose is largely affected by the nature of the cellulosic substrate and its physical state (Arantes, Saddler 2010).

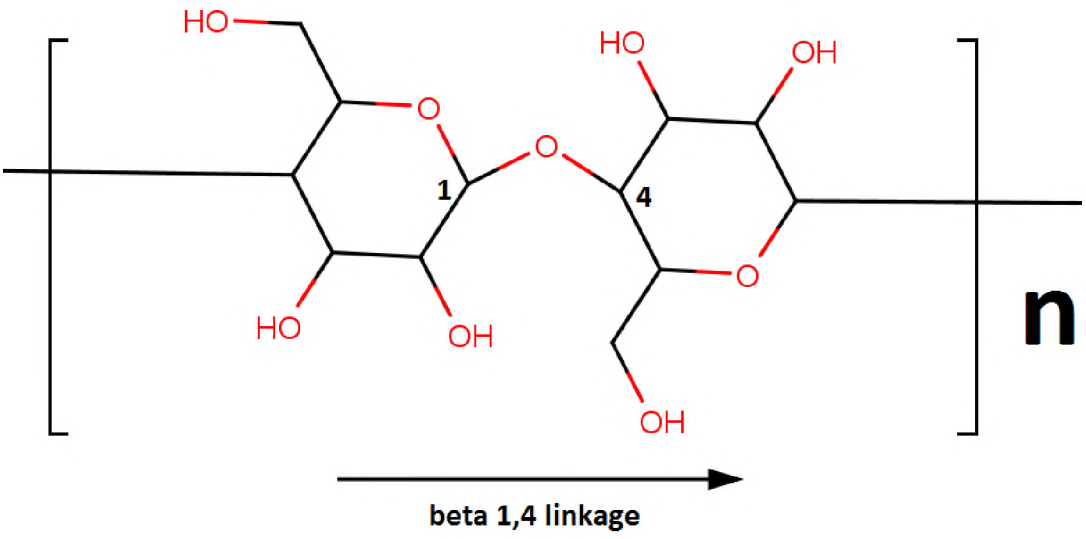

Figure 1.1: Cellulose repeat unit linked by a β 1,4 linkages. The repeat unit consists of two linked glucose residues.

The exact value of the repeat n is largely affected by the source of the cellulose. The structure of the cellulose repeat unit is shown in Figure 1.1. The hydrogen bonding network in cellulose causes cellulose chains to adopt a linear conformation (Moon et al. 2011). Cellulose is arranged in a homopolymer that is made up of two D-glucopyranose residues. The two ends of a cellulose chain consists of chemically distinct moeities. The first end possesses a free anomeric carbon on the D-glucopyranose. The other end of the D-glucopyronose has an anomeric carbon that forms the glycosidic bond between sugar residues. These ends are termed the reducing and non-reducing ends respectively (PÉrez, Samain 2010)

1.2. Cellulose degradation by fungal organisms

Fungal organisms have been demonstrated to exhibit great efficiency in degrading cellulose. This efficiency is largely attributed to the multiple enzymes encoded by fungi that are crucial for cellulose break down. This is due to the fact that fungi cannot readily uptake polysaccharides - as

a result, degradation of polysaccharides by a wide array of enzymes is required (van, de Vries 2011). The analysis of the genomes of cellulolytic fungi has revealed the presence of oxidative enzymes that are crucial for plant biomass breakdown. Currently, the method in which cellulose is broken down by these enzymes remains debated. However, there are three hypotheses that have been proposed to address the possible mechanism of degradation. These hypotheses involve the enzymatic hydrolysis of cellulose, fenton chemistry and the enzymatic oxidation of cellulose (Phillips et al. 2011). In the first hypothesis, it is believed that cellulases hydrolyze the β-1,4-glycosidic linkages that are on glucan chains. The hydrolysis of the β-1,4-glycosidic linkages is achieved by acid-base catalysis (Phillips et al. 2011). In the genomes of fungal organisms, genes that play a role in cellulose degradation are found in abundance. The expression of these genes as hydrolytic proteins results in the production of free glucose molecules through the degradation of cellulose (Merino, Cherry 2007, White, Brown 1981). Cellulose breakdown is mediated by three classes of enzymes which are the endo-β-glucanases, the exo-β-glucanase and the β-glucosidases. These enzymes have all been shown to synergistically degrade cellulose (Jeoh, Wilson & Walker 2002). Endo-glucanases are responsible for decreasing the length of the cellulose chain by hydrolyzing the β-1,4-glycosidic bonds randomly, exo-β-glucanases remove free glucose monomers from cellulose chains (Baker et al. 1998) and β-glucosidases, convert cellobiose to glucose through hydrolysis (Percival Zhang, Himmel & Mielenz 2006). Evidence has been presented that shows cellulose degradation can occur through non-enzymatic means. The enzymes required for cellulose breakdown are often too large to penetrate the plant cell wall. Due to this, pre-treatment by chemical means may be required for cellulose degradation (Goodell et al. 1997, Guerra et al. 2004). It has been proposed that oxidation through Fenton chemistry may modify the backbone of lignocellulose allowing enzymatic hydrolysis to occur. In this process metal ions are reduced resulting in hydroxyl radicals that randomly oxidize cellulose (Eastwood et al. 2011, Zhu et al. 2016). Fungal organisms are known to encode enzymes called cellobiose dehydrogenases (CDHs). CDHs have been shown to generate free hydroxyl radicals via Fenton chemistry (Mansfield, De Jong & Saddler 1997). This observation suggests that Fenton chemistry is important for cellulose breakdown. Cellulose degradation has also been shown to occur by enzymatic oxidation of cellulose by utilizing enzymes called oxidase and oxidoreductases. Through oxidative enzyme catalysis the glucan chain can be cleaved without having to remove the crystalline cellulose through the generation of oxygen radicals (Phillips et al. 2011).

1.3. The Auxiliary Activity Family 9

The Auxiliary Activity Family 9 (AA9) are enzymes of fungal origin that have been shown to increase the rate of cellulose degradation. For their enzymatic activity AA9 proteins require their coordinated Copper metal ion (Cu^{2+}), molecular oxygen and an electron donor. AA9 proteins are classified as Lytic Polysaccharide Monooxygenases (LPMOs) (Vaaje-Kolstad et al. 2010). Due to their mischaracterization as endoglucanases, AA9 proteins were referred to as the Glycoside hydrolase Family 61 (GH61). This mischaracterization was due to early studies that showed AA9 proteins (then GH61) having weak endo-1,4-beta-D-glucanase activity (Karkehabadi et al. 2008, Karlsson et al. 2001, Karlsson et al. 2001). More recent studies have shown that AA9 proteins are not traditional glycoside hydrolases due to the absence of the classical glycoside hydrolase binding cleft (Karkehabadi et al. 2008, Harris et al. 2010). As a LPMO, AA9 proteins degrade target crystalline cellulose allowing for further degradation by classical cellulases (Horn et al. 2012, Li et al. 2012). As a result, these enzymes have become important for the biofuel industry (Correa, dos Santos & Pereira 2016). The exact mode of AA9 cellulose interaction is unknown. However, it has been proposed that the Cu^{2+} coordinating AA9 proteins bind and cleave the glycosidic linkages of cellulose. AA9 proteins are believed to be crucial for the early stages of cellulose degradation. In order for traditional cellulases to have access to cellulose for degradation, it is important for the cellulose crystal to be destabilized. AA9 proteins are proposed to be involved in initial disruption of the cellulose structure (Vaaje-Kolstad et al. 2010, Quinlan et al. 2011). This initial cleavage of cellulose is believed to be the rate limiting step in cellulose degradation (Beeson et al. 2012, Beeson et al. 2015). There are currently many AA9 proteins - 2000 AA9 sequences are reported in the CAZy database (Levasseur et al. 2013) (www.cazy.org). However, only a small portion of these have been characterized experimentally. Due to sequence diversity displayed by AA9 proteins, it is possible that other AA9 proteins may cleave other substrates such as chitin (Horn et al. 2012, Li et al. 2012, Leggio, Welner & De Maria 2012). This notion is supported by the demonstration of hemicellulose degradation by AA9 proteins (Agger et al. 2014). This is further supported by the fact that the expression of AA9 genes is affected by different substrates (Hori et al. 2011, Yakovlev et al. 2012).

1.3.1. The AA9 Cu^{2+} active site

Even though the mechanism of AA9 proteins is unknown, there have been studies that suggest the involvement of copper-oxyl radical (Kim et al. 2014). AA9 proteins have been shown to introduce an oxygen atom (^{18}O) from molecular oxygen ($^{18}O_2$) into the resulting cleaved polysaccharide (Vaaje-Kolstad et al. 2010). In a recent study it was shown that molecular oxygen is introduced to polysaccharides at either the C1 or C4 carbon of glucose (Frandsen et al. 2016). The Cu^{2+} atoms of metalo-proteins are generally coordinated in one of three ways which are termed; Type I, Type II, and Type III (Adman 1991, Rubino, Franz 2012). The Type I Copper centers possess a distorted tetrahedral geometry that is coordinated by two His residues, as well as a Cys residue. The last residue required in this coordination can be a Met, Asp, Glu or a carbonyl atom from the protein backbone (Rubino, Franz 2012). The Type II Cu^{2+} centers are generally coordinated by one or more His residues and the other coordination positions can be occupied by any of the following residues: Met, Asp, Glu, Gln and Tyr. Type II copper centers coordinate water molecules. Type II copper centers adopt one of two geometries depending on coordinated ligands, either a distorted square planar or distorted square pyramidal geometry (Rubino, Franz 2012, MacPherson, Murphy 2007). In Type III Copper centers, three His residues coordinate two antiferromagnetically coupled Cu^{2+} (Adman 1991, Rubino, Franz 2012). Electron paramagnetic resonance (EPR) parameters have shown that the active Cu^{2+} of AA9 protein is a Type II Copper (Cu(II)) that is coordinated by a Histidine brace (Quinlan et al. 2011). The spectroscopic and computational analysis of the Cu^{2+} of AA9 proteins has a four coordinate tetragonal geometry in its oxidized state. Upon reduction by molecular oxygen the enzyme has a three coordinate T-shaped structure (Kjaergaard et al. 2014).

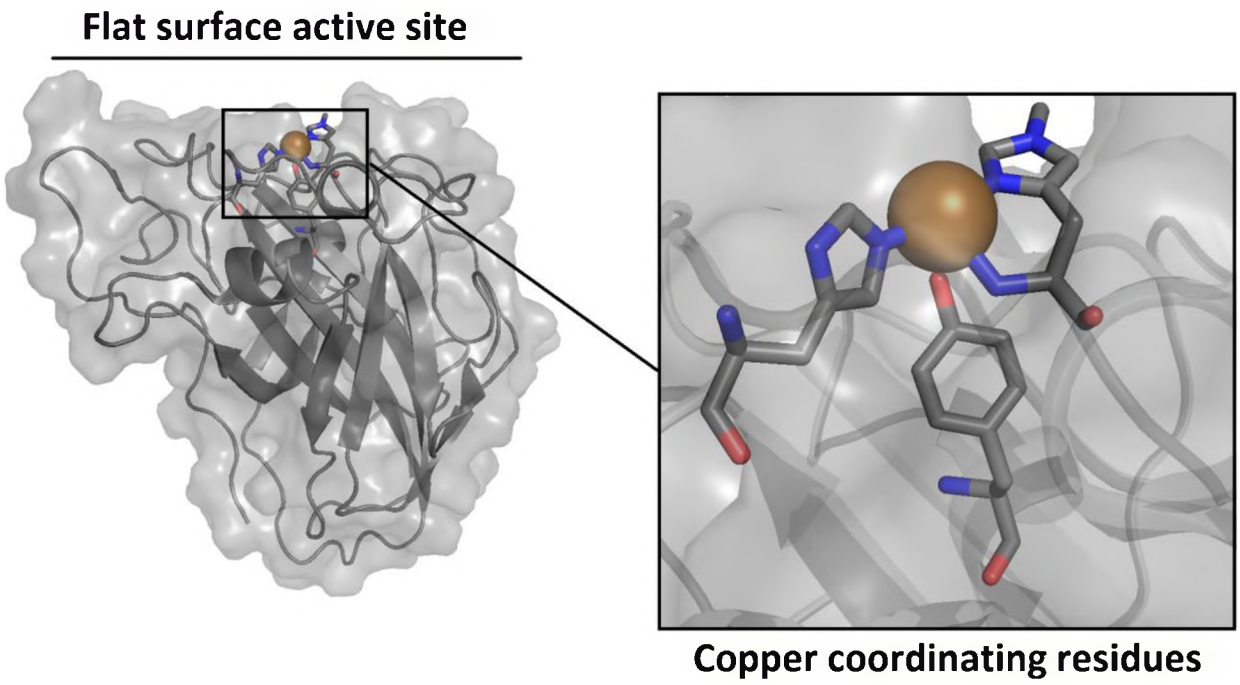

Figure 1.2: The copper coordinating flat surface active site of AA9 proteins. The Type 2 LPMO crystal structure 4EIR is shown in cartoon surface and the copper atom is shown as a sphere while the copper coordinating residues are shown as sticks.

The histidine brace of AA9 proteins consist of an N-terminal His residue which provides two coordination positions which are the N terminal Nitrogen and Nδ from imidazole ring and Nε atom from the second His imidazole ring (Hemsworth, Davies & Walton 2013) as shown in Figure 1.2. In some AA9 proteins the N-terminal His residue is methylated however, the function of the methylation remains unknown. The Cu^{2+} is also coordinated by a buried Tyr residue. The Cu^{2+} ion possesses octahedral geometry which undergoes Jahn–Teller distortion (Li et al. 2012).

1.3.2. AA9 regioselectivity

Depending on the cleavage product they form, AA9 LPMOs can be grouped into three types. All three AA9 LPMO types contain a Type II Cu^{2+} active site. The cleavage product formed is determined by the site of oxidation. AA9 proteins can either oxidize the C1 or the C4 carbon of a glucose ring structure. In rare instances the oxidation of the C6 carbon has also been reported (Li et al. 2012). When cleavage occurs at the C1 carbon, the resulting product is aldonolactone. When the cleavage position is the C4 carbon, the resulting cleavage products is 4-ketoaldose (Phillips et al. 2011, Beeson et al. 2012). Due to this observed regioselectivity, AA9 proteins are classified

into 3 types. Type 1 AA9 proteins cleave the C1 carbon of glucose and produce aldonolactone. Type 2 AA9 proteins cleave the C4 carbon of glucose producing 4-ketoaldose. Type 3 LPMOs cleave at both C1 and C4 carbons producing aldonolactone and 4-ketoaldose respectively (Phillips et al. 2011, Quinlan et al. 2011, Beeson et al. 2012). All three AA9 LPMO types contain a Type II Cu^{2+} active site.

1.3.3. Synergism of AA9 proteins

LPMO have been known to have a positive effect on cellulose degradation. As a result, numerous studies have been performed to determine the factors that play a role in polysaccharide degradation by LPMOs. It has been shown that AA9 proteins cooperate with a variety of cellulases to degrade cellulose. The degradation of cellulose was also shown to be affected by the different biomass components to achieve an increase in biomass conversion. The major contributor being the relative amounts of accessible crystalline to amorphous cellulose within the various components (Hu et al. 2014). Post translational modifications on AA9 proteins may have a possible effect of synergism. It was found that the expression of AA9 proteins without post translational modifications such as methylation and glycosylation may have decreased activity (Kim et al. 2015). Even though the decrease in activity was minimal, it was the authors' belief that studying the effect of post translational modifications on AA9 proteins may provide further insights into the activity of AA9 proteins.

1.3.4. Other AA families

Studies have demonstrated that the presence of LPMO families that have activity on cellulose (Phillips et al. 2011, Quinlan et al. 2011, Forsberg et al. 2014a), chitin (Forsberg et al. 2014b, Forsberg et al. 2014b, Hemsworth et al. 2014), cellodextrins (Isaksen et al. 2014a), hemicellulose (Agger et al. 2014), and starch (Vu et al. 2014, Lo Leggio et al. 2015b). To date, the known LPMO families are AA9, AA10, AA11 and AA13. AA10 enzymes are primarily of bacterial origin and activity studies have demonstrated that AA10 enzymes are active on chitin in an oxygen dependent reaction (Vaaje-Kolstad et al. 2010). However, such as the BlAA10A from *Bacillus licheniformis,* AA10 enzymes that have been determined to have activity on cellulose and chitin (Forsberg et al. 2014a). The AA11 enzymes are fungal in origin similar to AA9 proteins which resulted in their misclassification as AA9 proteins (Busk, Lange 2015). However, it was later shown that these enzyme were a distinct family hence their reclassification as AA11. AA11 proteins have been

shown to have enzymatic activity on chitin (Hemsworth et al. 2014). Though studies have almost always emphasized the role of LPMOs in the degradations of crystalline cellulose, recent studies have shown an LPMO family which oxidize non crystalline hemicelluloses and soluble cello-oligosaccharides (Agger et al. 2014, Isaksen et al. 2014b, Frommhagen et al. 2015, Bennati-Granier et al. 2015). The AA13 enzymes proposed to be involved in degradation non crystalline hemicelluloses (Lo Leggio et al. 2015a, Vu, Marletta 2016).

1.3.5. AA structural features

Even though AA9, AA10, AA11 and AA13 have different observed substrate interactions, they do share common features. The copper active site of AA9, AA10, AA11 and AA13 in their respective crystal structures is found to be surface exposed on the flat surface active site of LPMOs where it is exposed to the solvent (Lo Leggio et al. 2015b). This flat surface active site is often believed to be critical for cellulose oxidation as this is the region that most likely binds to polysaccharide substrate. However, unlike AA9, AA10 and AA11, the AA13 enzymes possess a small groove across the active site surface which could explain why AA13 may interact with soluble polysaccharides (PÉrez, Samain 2010). Polar residues and aromatic residues are important for polysaccharide binding. Mutagenesis studies were able to show the involvement of hydrophilic residues in polysaccharide binding for AA10 Proteins (Vaaje-Kolstad et al. 2005). Mutagenesis studies have also been used to confirm the importance of aromatic residues on the proposed AA9 binding surface (Harris et al. 2010). These residue features of LPMOs are believed to be crucial for the binding of LPMOs to polysaccharides as it is believed that aromatic residues or hydrophilic residues form stacking interactions with sugar residues on polysaccharides (Frandsen, Lo Leggio 2016). AA9 proteins like all LPMOs, have a conserved β-sandwich fold. This fold is generally found in proteins that play a role in molecular recognition such as the immunoglobulin heavy chain variable domain and fibronectin III. Similar to immunoglobulins and fibronectin III, LPMOs contain highly variable loop regions that are believed to be crucial for substrate binding (Li et al. 2012).

1.3.6. Modularity of AA9 proteins

A large proportion of AA9 protein sequences in the Pfam database were found to be associated with varying numbers of carbohydrate binding modules (CBMs) (Boraston et al. 2004). The glycosidic bonds of polysaccharides often present the challenge of accessibility to the active site

glycoside hydrolases. As a consequence, some AA9 proteins are associated CBMs. These CBMs are responsible for enhancing the accessibility of glycosidic bonds by associating the enzyme with the substrate. The structural analysis of 22 CBM families revealed that the CBMs from different families are similar in terms of structure (Shoseyov et al., 2006). CBMs use their hydrophobic surface which binds cellulose. Substrate disruption by CBMs was first observed in *Cellulomonas fimi* endoglucanase which showed the substrate disruption without any detectable hydrolytic activity (Din et al., 1991). Other proposed functions of CBMs include a proximity effect and a targeting function. Comparison of the structures of CBM1 and AA9 showed that these two are similar in terms the polar aromatic residues found on the substrate binding surface. The polar aromatic residues on CBM1 have a different spatial distribution to those on the AA9 protein because of the size differences between the two (Leggio, Welner & De Maria 2012, Kraulis et al. 1989). This observation suggests that AA9 bind cellulose in a similar manner to CBM1. The number of polar residues on AA9 tends to vary which is believed to affect substrate binding and product formation (Li et al., 2012).

1.3.7. Previous Cellulose AA9 interaction studies

QM and MD analysis that aim to elucidate the interaction between AA9 proteins and cellulose have been performed. In one such study, the spatial arrangement of the aromatic residues found on the AA9 active site was investigated. It was found that the planar flat surface active site residues of the AA9 active site share similar spatial arrangements to glucose rings of cellulose (Li et al. 2012). This finding suggests that the flat surface active site residues are possibly involved in substrate binding and orientation. In another study, MD simulations were performed to elucidate the interaction between AA9 proteins and their substrate cellulose (Wu et al. 2013). It was found that upon binding to cellulose the AA9 crystal structure undergoes conformational changes in many active site regions. These conformational changes were found to result in the alignment of three tyrosine residues that are found on the AA9 active to the cellulose substrate. This study was performed by constructing a three layered cellulose crystal only the C1 and C4 carbons of all layers were fixed, resulting in a rigid substrate. The AA9 crystal structure was placed on the cellulose substrate such that Tyr-28 and Tyr-198 align over the middle chain on the crystal surface, and Tyr-75 aligns over the edge chain. MD simulation were performed for 100 ns using the Chemistry at HARvard Macromolecular Mechanics (CHARMM) software package (Brooks et al. 2009). Due

to the fact that there are no AA9 force field parameters for the Cu^{2+} active site, the authors elected to use parameters from copper containing systems (Wang et al. 2011b, Ungar, Scherer & Voth 1997, Zhu et al. 2008, Babu, Lim 2006, Wang et al. 2011a). There were no dihedral parameters included in the study and the active site was harmonically restrained using a force constant of 10 kcal/mol/A. Another study sought to elucidate the enzymatic mechanism that results in cellulose cleavage by AA9 proteins (Kim et al. 2014). The investigation involved the use of the density functional theory (DFT) (Jones 2015) to investigate the reactive oxygen species responsible for cleavage, to compare the proposed methods of cleavage and to investigate the role of methylation on the N-terminus Histidine residue. In this study it was found that the AA9 proteins are likely to utilize a copper-oxyl mediated oxygen rebound mechanism for cellulose cleavage.

1.3.8. The knowledge gap

AA9 proteins are well known for being abundant in fungal genomes and are sequence diverse. As a result, it is important to understand the effect of sequence diversity on the structure of AA9 proteins and assess its possible effects on AA9 type-specific substrate interaction. This work, seeks to elucidate the type-specific features of AA9 proteins that contribute to their substrate regioselectivity. As a result, not much is known about the unique sequence features of respective AA9 LPMO types and the effect of these features on the structures and function of AA9 proteins. It is therefore important to analyze the sequences of AA9 proteins in a type-specific manner in order to elucidate type-specific features of these proteins. Once identified, these features will be investigated for their possible effect on substrate interaction using MD simulations. However, due to the absence of AA9 specific force field parameters of the Cu^{2+} AA9 active, it is important to first elucidate parameters required to perform MD simulations.

1.3.9. Overview of the work

The aim of this study was to determine the type-specific features of AA9 LPMO types that drives regioselectivity. To determine the type-specific features unique to AA9 LMPO bioinformatics characterization was performed. Once identified, the effect of the type-specific features was determined by employing MD simulations. It has been previously postulated that the configuration of the AA9 active site results in the observed regioselectivity of AA9 LPMO types (Hemsworth, Davies & Walton 2013). As result, the study of the sequence and structural features is likely to provide insights that will aid in furthering the understanding of AA9 cellulose interaction.

In this study the type-specific features of AA9 proteins were identified through the use of the methodology described below. AA9 protein and nucleotide sequences were obtained from literature and the Pfam database (Finn et al. 2016). Once the sequences were obtained, multiple sequence alignment and phylogenetic clustering were performed. All vs all sequence identify calculations (Faya, Penkler & Tastan Bishop 2015) were done in order to assess the extent of conservation in AA9 sequences in general as well as among AA9 LPMO types. Motif discovery (Jonassen et al. 2002, Redhead, Bailey 2007) of sequences was performed on AA9 sequences to identify regions that are common in all AA9 types and regions that are specific to individual AA9 LPMO types. A type-specific physicochemical property analysis (Faya, Penkler & Tastan Bishop 2015) was performed to identify characteristics unique to each type. A selection analysis (Massingham, Goldman 2005) was performed on AA9 nucleotide sequences to understand the selective pressures that are present on AA9 protein sequences. Once unique structural features were identified, they were mapped onto 4B5Q, EIR and 3ZUD AA9 crystal structures for Type 1, 2 and 3 AA9 proteins respectively to investigate the possible effects on type-specificity and substrate interaction. To further investigate the effect of type-specific features on substrate interaction, MD simulations were performed. However, prior to MD simulation it was important to elucidate the bond stretch, angle bend, dihedral parameters and Cu^{2+} Lennard-Jones parameters for the Cu^{2+} AA9 active site of AA9 proteins using Potential energy surface (PES) scans. Once evaluated, the force field parameters were validated with MD simulations on all three AA9 types by monitoring various properties throughout the course of the MD simulation.

Chapter 2

Sequence Analysis of AA9 proteins

2. Sequence analysis of AA9 proteins

AA9 proteins are known to be abundant in fungal genomes. The fact that AA9 proteins are divided into three LPMO types suggests that the may be sequence differences characterizing these different types, though this has been previously unexplored. It is therefore important to assess the effect of sequence diversity on AA9 LPMO types, its effect on structure of AA9 proteins and the possible effect on AA9 substrate interaction. In this chapter the main aim is to employ bioinformatic sequence analysis techniques to identify unique type-specific features on AA9 protein sequences. This will be achieved through obtaining the relevant AA9 protein and nucleotide sequences from the Pfam database (Finn et al. 2016), employing techniques such as multiple sequence alignment (MSA) and motif discovery to identify type-specific conserved features and phylogenetic analysis cluster the sequences into their respective LPMO types. A physicochemical property analysis will be performed to identify a chemical distinction between AA9 protein LPMO types. The evolutionary pressures that are at play on AA9 protein and nucleotide sequences will be investigated using phylogenetic analysis and a selective pressure analysis. With phylogenetic analysis, the phylogenetic relationships between AA9 protein LPMO types will be investigated and the AA9 sequences will be grouped into the respective types. Selection analysis will be used to detect the selective pressures that are present on specific residues on AA9 sequences. The selection pressure of that could be at play a role in AA9 proteins can either be positive, neutral and negative. All type-specific features will be mapped on AA9 crystal structures to better understand how these features affect AA9 structures and substrate interaction.

2.1. Introduction

The amino acid sequence of a protein governs the 3D structure of it. The 3D structure in turn determines the protein's function. In the case of the AA9 proteins, there is an evident sequence diversity in these proteins as reported in various databases. The implications of this sequence diversity has been suggested to result in AA9 proteins being able to metabolize more than one

substrate. To further understand the effect of sequence diversity on the regioselectivity of AA9 LPMO types the sequence will be investigated at DNA and protein levels.

2.1.1. DNA sequence analysis

The advancement of DNA sequencing approaches has resulted in a large abundance of nucleic data. This data has resulted in an increase in the amount of DNA sequence information available for testing evolutionary hypotheses (Pareek, Smoczynski & Tretyn 2011). This wealth of data includes collections of homologous gene sequences from different species as well as multiple sequences of particular genes sampled from different individuals within a single species. This data clearly indicates that there are abundant changes in DNA sequence between species as well as large amounts of DNA sequence polymorphism within species. The aim of this section is to assess the residue specific selective pressure on AA9 protein types. This was investigated by identifying residues that are under positive, neutral and negative selection.

2.1.1.1. Ratio between nonsynonymous and synonymous

The ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions (Ka or dN) per synonymous (Ks or dS) site is often used as a measure of selective constraint on a protein (Kryazhimskiy, Plotkin 2008). Synonymous sites are sites that undergo nucleotide changes that do not result in the change in amino acid residues. On the other hand, nonsynonymous sites are sites that undergo nucleotide changes that result in the change in amino acid residues. In certain instances this is referred to as Ka/Ks, or dN/dS, or simply as ω ('omega'). There are three possibilities: Ka/Ks< 1: This is the case for the vast majority of protein coding genes. It indicates selective constraint against amino acid replacements (known as 'negative' or 'purifying' selection). Ka/Ks= 1: This indicates that the amino acid sequence is under no selective constraint. This is expected for pseudogenes. Ka/Ks> 1: Indicates positive (or 'diversifying') selection for amino acid replacement (Dasmeh et al. 2014). This is sometimes observed in genes encoding antigenic proteins of pathogens, which are under strong selective pressure to change in order to avoid the immune response of the host for example, the gene encoding HIV envelope protein or the hemagglutinin A gene of human influenza virus. Searching for genes with Ka/Ks> 1 is a common method for identifying 'positively selected' genes from whole genome comparisons between species (Mayrose et al. 2013, Mugal, Wolf & Kaj 2014).

2.1.1.2. Tools for selection

The detection of positive and negative selection on AA9 proteins was carried out using two online servers namely DataMonkey (Delport et al. 2010) and Selecton (Stern et al. 2007). These two webservers were chosen because they allow the identification of selection pressure in protein sequences through the use of various methods. The DataMonkey webserver, which is based on the Hyphy package, has three methods for calculating selection at a specific codon site. These methods are SLAC, FEL and REL (Pond, Frost 2005). Selecton on the other hand uses the M8 model for evaluating selection. Both Selecton and DataMonkey take nucleotide sequences as input. For this analysis, nucleotide sequences were obtained from GenBank for each of the three AA9 LPMO types.

2.1.1.2.1. Models for heterogeneous selection pressure

Models used for selection all take into account the following parameters: the branch lengths, the transitions and tranversion ration and base frequencies at each of the three positions in the codon (Pond, Frost 2005). The models used to detect selection on AA9 proteins are displayed in Table 2.1. The models were used to detect positive, neutral and negative selection. The M8 model contains additional components which maker it suitable for performing positive selective analyses (Stern et al. 2007).

Table 2.1: Models used by Selecton for selection.

Model 0 (M0)	
Model 1 (M1)	Computes one dS and dN ratio for the entire sequence.
Model 2 (M2)	Similar to M1 but with an estimated dS/dN ratio as an extra class.
Model 3 (M3)	Takes into account the distribution shape of dS/dN.
Model 4 (M4)	Fixes dS/dN at specific values.
Model 5 (M5)	This model gives dS/dN a gamma distribution.
Model 6 (M6)	This model gives dS/dN two gamma distributions.
Model 7 (M7)	This model gives dS/dN a beta distribution.
Models 8-11 (M8-M11)	These models add an extra component to their computation to account for positive selection. In the M8 model, an extra class is added to the beta model this allows for the calculation of positive selection. Selecton implements this model for all positive selection analysis

2.1.1.2.2. Single likelihood ancestor counting (SLAC)

The SLAC approach is used to count the number of nonsynonymous (dN) and synonymous (dS) changes in sequences. Once the number of dN and dS sites has been obtained, the difference between dN and dS is computed. If the difference between dS and dN is significant, then inference can be made on the selective pressure on the codon site (Suzuki 2004). SLAC is an improved version of the Suzuki–Gojobori method (Suzuki, Gojobori 1999).

2.1.1.2.3. Fixed effect likelihood (FEL)

The FEL approach is a likelihood-based method that estimates the ratio of dN and dS for each site in a sequence alignment. The potentially large number of rate and other parameters (e.g. branch lengths) is optimized during the process. Without this optimization, the computation would be impossible (Pond, Frost 2005).

2.1.1.2.4. Random effects likelihood (REL)

REL is similar to FEL however, it also allows a variation in rate for dS and dN. REL generally considers a relatively smaller number of parameters as opposed to FEL as a result, REL optimization of all parameters is possible (Pond, Frost 2005).

2.1.2. Multiple Sequence Alignment

To identify conserved type-specific features of AA9 proteins that determine AA9 regioselectivity sequence alignments may be used. In order to identify these conserved features multiple sequence alignment (MSA) will be employed to compare linear AA9 protein sequences against each other to observe conservation. MSAs identify conservation through aligning homologous positions in DNA or protein sequences into columns (Do et al. 2005). By assessing the similarity in the column, the conservations of the protein or nucleotide sequences may be gauged. The conservation can information about the evolutionary history of the proteins being analyzed or even elucidate the function of a particular protein. One of the major contributors to the quality of resulting MSA is the use of scoring matrices. A protein scoring matrix contains information about amino acid substitution rates that occur overtime. The scoring matrix can be used to determine how favourable an amino acid substitution is (Pearson 2013). Another contributor to the quality of an MSA is the quality of the input sequences used in the calculation. If the sequences fall within the twilight zone, meaning the sequences have a sequence identity below 30%, a drop in alignment quality may be expected. Another contributor to the quality of a multiple sequence alignment is defining an objective function that will generate the best alignment. The objective function may also be used to gauge the quality of the generated sequence alignment. The alignment of sequences is achieved by adding a match or mismatch for each position aligned. For unaligned amino acids, a gap penalty is assigned to penalize a mismatch.

2.1.2.1. Alignment programs

Various alignment programs have been developed for sequence alignments. In this study two alignment programs were used which were MAFFT (Katoh et al. 2002) and PROMALS3D (Pei, Grishin 2014). MAFFT was used on large datasets in order to save on computational cost. PROMALS3D was applied to the filtered datasets in order to generate more accurate alignments due to the structural information incorporated by the program.

2.1.2.1.1. MAFFT

MAFFT is a MSA program that is based on the fast Fourier transform (FFT). FFT for the identification of homologous segments in proteins or nucleotide sequences. . The FFT is utilized because substitutions of amino acids with different ones can occur. However the substituted amino acid is usually exchanged with one that has similar physicochemical properties such as the size or polarity. For FFT calculation of the correlation ($c(k)$) between two amino sequences is required. This is then followed by identification of homologous regions. The two aligned sequences are searched for homologous regions. If homologous regions are detected then $c(k)$ peaks will correspond to the homologous regions. The FFT method is limited to the positional lag of k in a homologous region and is not affected by the position of the region. To locate the positions, a sliding window analysis is performed. A window size used is 30, with the degree of local homologies calculated for the first 20 highest peaks in c(k). To obtain an alignment of the sequences, a dividing homology matrix is constructed in order. The segments in the sequences are arranged consistently to ensure optimal arrangement of the homologous segments (Katoh et al. 2002, Katoh, Toh 2008).

2.1.2.1.2. PROMALS3D

In a PROfile Multiple Alignment with predicted Local Structures and 3D constraints (PROMALS3D) alignment, the first stage involves the alignment of similar sequences. This is achieved through the application of a scoring function. The scoring function consists of a weighted sum of pairs BLOSUM62 scores. The second stage is another alignment stage were a sequence is chosen as the target sequence. A search for additional homologues of the target is performed with PSI-BLAST (Altschul et al. 1997) from the UNIREF90 (Suzek et al. 2007, Suzek et al. 2014) database. PSIPRED (McGuffin, Bryson & Jones 2000) is then used to predict the secondary structural elements of the homologues. Both the representatives are then profile-profile aligned using a hidden Markov Model (HMM) to the predicted secondary structures. This is performed to generate the posterior probabilities of the residue matches. The posterior probabilities are then used sequence-based constraints through the derivation of a probabilistic consistency scoring function. Finally the representative target sequences are then progressively aligned using a consistency scoring function. The pre-aligned groups from the first stage are merged to form an

alignment of representatives resulting in the final multiple alignment of all sequences (Pei, Grishin 2014).

2.1.3. Motif analysis

DNA and protein sequences may also be investigated for the presence of sequence conserved motifs. In protein sequences a motif can have four definitions. Motifs can be regarded as series of conserved residues that may evolve separately from the remaining residues in a protein sequence. The conservation of motifs may suggest a specialized function for that sequence. A motif may also be regarded as a region on a protein that is critical for the structural integrity of a protein. These structural motifs can possess constraints that determine how secondary structural elements form. A conserved sequence motif can be thought of as region on a protein that is important signalling purposes such as signal peptides and trans-membrane regions. Lastly conserved sequence motifs may simply be used to distinguish a group of proteins from another by showing evolutionary relatedness between proteins (Bork, Koonin 1996).

2.1.3.1. Motif discovery with MEME

In this study, motif discovery was used to identify conserved motifs in AA9 proteins. To achieve this, the Multiple Expectation Maximisation for Motif Elicitation (MEME) (Bailey et al. 2006) was used. MEME is a web-server for detection of potentially biologically important motifs in related DNA or protein sequences. MEME program estimates the parameters that could have been used to generate the dataset using the MM algorithm. This best described by a two component finite mixture model. MEME takes a dataset of unaligned sequences as input. The first component of the algorithm identifies a set of small similar sequences with a fixed width within the dataset. The second component of the algorithm is the description of all the positions in the dataset which is termed the background. By estimating the frequency of motifs in the dataset, the MM model is fitted to the dataset (Bailey et al. 2006).

2.1.3.2. Motif comparison with MAST

In order to help assess the quality of the discovered motifs, the Motif Alignment and Search Tool (MAST) (Bailey, Gribskov 1998) is used. MAST is an integral part of the MEME-suite and is executed concurrently with MEME jobs to assess the likelihood that two motifs are similar or different from each other by calculating the pairwise correlation. By trying all possibilities

alignments between the two motifs creates the maximum. The maximum is sums all Pearson's correlation coefficients of the aligned columns this is then divided by the length of the shortest motif in the pair (Bailey, Gribskov 1998). If a pair of motifs has a high Pearson correlation (>0.60) then that pair of motifs are considered to be similar.

2.1.4. Phylogenetic analysis

Phylogenetic analyses to be employed to investigate the related of protein sequences (Yang, Rannala 2012). Phylogenetic analysis can have also be applied to predict the function of an unknown gene (Searls 2003). A phylogenetic tree is a series of branches that are connected by nodes. The branches of a phylogenetic tree represent the existence of a genetic lineage through evolutionary history. The node of a phylogenetic tree represents the formation of a new lineage. The phylogenetic tree represents the relatedness of species among each other. The nodes in a phylogenetic tree represent speciation events or the formation of new species (Yang, Rannala 2012).

2.1.5. Physicochemical property analysis

The physicochemical properties of proteins are important due to the fact that they give insights into the function of proteins. Physicochemical properties have also been shown to play a role in complex formation, folding and protein stability of proteins (Baker, Agard 1994). In this study the aromaticity, hydrophobicity, molecular weight, isoelectric point (pI), instability index and grand average of hydrophobicity (GRAVY) index were investigated for AA9 proteins. Boxplots were generated in R to graphically represent the difference in physicochemical properties among AA9 LPMO types. Statistical analysis was then performed on the observed physicochemical properties to observe any significant differences in the calculated physicochemical properties among AA9 LPMO types.

2.1.6. Statistical analysis

The statistical analysis on the observed physicochemical properties was performed using the R software package (Team 2013) in order to compare the distributions of physicochemical properties displayed in boxplots. The student t-test was used to determine the difference between the three AA9 proteins at a 5% level of significance ($p<=0.05$). The student t-test compares the means of the samples and decides if they are significantly different from each other. In this study the t-test

was used to determine if there were any type-specific physicochemical properties observed. The t–test assumes that the samples have a normal distribution (Witt, McGrain 2016).

2.2. Structural mapping of unique type-specific features

It has been demonstrated that the structure of a protein is a determinant of its function. The structure of a protein tends to be more conserved than its sequence (Wright, Dyson 1999). Due to this observation, AA9 proteins that generally have low sequence similarity may have similar functions due to structural conservation. As a result, to better understand the function of AA9 proteins it is important to gain a better understanding of the type-specific structural feature present on AA9 LPMO types. The common ways protein structures are determined is be utilizing X-ray crystallography and nuclear magnetic resonance (NMR). In cases were crystallographic or NMR structures are not present a viable model has to be created. As a result, in order to obtain a viable 3D models of AA9 protein sequences of interest, homology modelling was used to elucidate the unknown structures of interest (Wiltgen, Tilz 2009).

2.2.1. Homology modelling

In the absence of an experimentally elucidated protein structure, homology modelling may be employed. Homology modelling is a powerful technique that can be used to predict the structure of a protein. The principle of homology modelling is that the structure of proteins is more conserved as opposed to the primary amino acid sequence. It has been shown that the structure of proteins is ten times more conserved than the primary protein sequence (Illergard, Ardell & Elofsson 2009). With homology modelling the unknown query sequence is modelled based on already available crystal structures called templates. The accuracy of the generated homology model is highly dependent on the template used. When a suitable model template has been obtained, an alignment of the query sequence; target with a homologous structure template is performed. The alignment is performed in order to transfer the 3D coordinates of the template to the query sequence. After aligning the template and query sequence the model is built using the coordinates as constraints (Cavasotto, Phatak 2009).

2.2.1.1. Template identification

The first stage of homology modelling is the identification of a suitable template. A suitable template is one that has the highest sequence identity with the query sequence. However, it has been shown that structures may be modelled with templates having a sequence identity of at least 25% (Honarparvar et al. 2014). It is also important to consider the quality of the template by looking at features such as R-factor and the resolution (Fiser 2010). The alignment between query and template is important for accuracy due to the fact that features of the template used will be carried over to the resulting structure of the query sequence. As such, caution must be taken in selecting an appropriate template from the Protein Data Bank (PDB). To aid in the identification of suitable templates for query sequences, tools such as the HHpred webserver (Soding, Biegert & Lupas 2005) have been developed. The HHpred webserver identifies suitable templates by searching the PDB to identify a template based on the user specified query sequence. The input can either be a single query sequence or MSA. Using multiple iterations of PSI-BLAST (Altschul et al. 1997), the HHpred webserver first aligns homologous sequences from NCBI to the input query sequence. The user is allowed to specify the E-value threshold, PSI-BLAST iterations, the minimum sequence identity and the minimum number of PSI-BLAST matches to aid in the identification of a suitable template. Upon obtaining a match with PSI-BLAST, PSIPRED (Jones 1999) is used to annotate the predicted secondary structure to the alignment. Then from the alignment, a profile hidden Markov model (HMM) (Eddy 1998) is created. The HMM is used to represent the alignment statistically. For each column of the alignment the profile describes the probability that profile is one of the 20 amino standard acids. These include the quality of the template and the coverage the query and the target have (di Luccio, Koehl 2011). In order to generate a good model, a high quality template has to be used. In the absence of a suitable template, then multiple templates may be used to generate the homology model. The use of multiple template accounts missing information not represented in individual templates such as the presence of gaps, or low sequence similarity between target and template (Sali, Blundell 1993, Eswar et al. 2003).

2.2.1.2. Aligning the template and query sequence

Various approaches may be used to align the unknown query sequence with the structure template. The most straight forward approach would be to pair wise align the query and the template. If the

sequences are similar with sequences identities generally above 30 % pair wise alignments may be used with minimal concern. However in cases were the sequence identity is below 30%, this approach may not produce reliable results (Xiang 2006). The alignment stage is very important for model building because errors that are generated at this stage can be carried forward to the generated model (Sali, Blundell 1993, Sanchez, Sali 1997).

2.2.1.3. Homology modelling using spatial restraints

The program MODELLER (Webb, Sali 2016) is used for performing homology modelling. To do this MODELLER generates a number of spatial restraints on the query sequence based on the user supplied template and alignment that have to be satisfied. The restrains are created by assuming the distances of the aligned residues between the query and the template are similar. Features such as bond lengths, bond angles, dihedral angles, and non-bonded atom-atom contacts are then used to support the template derived restraints. Deviations from the specified restraints are minimized resulting in the homology model. Due to the high diversity in loop regions, loop modelling may have to be performed (Fiser, Do & Šali 2000, Jamroz, Kolinski 2010). As previously stated the loops of AA9 proteins are believed to be critical for substrate binding as a result, their correct modelling is of importance (Sali, Blundell 1993, Eswar et al. 2003).

2.2.1.4. The refinement of loops

When generating a homology model one of the most important things to consider is the modelling of loops. These loop regions present challenges due to the fact that they tend to show the greatest variability with respect to amino acid composition. The use of spatial restraints to loop regions is often difficult. As a result an energy function may be applied to relax the backbone of loops to match what would be observed in native structures (Sali, Blundell 1993). The MODELLER programs has a variety of loop refinement methods that use a scoring function specifically designed for loop modelling. MODELLER allows users to decide between automatic or manual loop refinements. The first stage in loop refinement involves creating the initial conformation of the loops. This is achieved by aligning the protein residues uniformly on the loop back bone. The initial loop conformation is then randomized by 5Å in each of the Cartesian directions. The model optimization is performed two times. First optimization is only applied to loop atoms and second

optimization is applied to the complete structure. A mean force that is atomistic distance-dependent for on bond interactions is used to optimize the structure (Melo, Feytmans 1997).

2.2.1.5. Model evaluation

The reliability of the information that is extracted from the homology model heavily reliant on how accurate the homology model is. It is therefore important to validate the models to determine the overall quality of the generated structure. Protein structures can be evaluated globally using scoring functions. The scoring functions used to assess protein quality are based on the properties of amino acids of found in known protein structures. Protein structures may also be validated through comparisons of structural features present on the homology model with those found on experimentally determined structures (Wiltgen, Tilz 2009, Sali, Blundell 1993, Sanchez, Sali 1997). The programs used for evaluating the quality of models generally use a qualitative or relative scoring function as opposed to the detection of correct and incorrect regions. The most native models generally have the lowest free energy (Zemla et al. 2001). As a result, the free energy can therefore be used to predict the quality of a model by sampling the potential energy surface using a molecular mechanics force field (Eramian et al. 2008). Multiple tools may be used to validate protein structures. The quality of the models can be ascertained from use of various validation programs and their consensus on the quality of the structure (Bishop, Beer, Tjaart A. P. de & Joubert 2008).

2.2.1.5.1. Normalized DOPE score

The Discrete Optimised Protein Energy (DOPE) score is a statistical potential that is dependent on atomic distance. The DOPE score is used to evaluate structures. Proteins in their native form possess their lowest free energy. As a result, the DOPE score is a measure to predict the stability of proteins. To measure protein stability the statistical potential i.e. DOPE score is computed. Generally the DOPE score is distance based potential that is obtained from sampling native structures. It is impossible to compare two different proteins with the DOPE score as the score is protein specific. Hence, the normalized DOPE (DOPE Z) score can be employed to permit comparison of two separate structures (Shen, Sali 2006).

2.2.1.5.2. Rosetta energy score

The Rosetta energy score unlike the DOPE score, is not based on native structures. The basis of the Rosetta energy score is the thermodynamics principles. The way large biomolecules structurally arrange themselves in equilibrium conditions will favor the state with the lowest free energy. For the Rosetta energy score, a mathematical function is used to calculate the most favored structural arrangement by taking in to account the hydrogen bonding and electrostatics of the system (Lazaridis, Karplus 1999).

2.2.1.5.3. MetaMQAPII

The MetaMQAPII webserver (Pawlowski et al. 2008) is a combination of various other model evaluation servers which include; VERIFY3D(Eisenberg, Lüthy & Bowie 1997), ProSA (Wiederstein, Sippl 2007), ANOLEA (Melo et al. 1997), BALA-SNAPP (Krishnamoorthy, Tropsha 2003), TUNE (Lin, May & Taylor 2002), REFINER (Boniecki et al. 2003) and PROQRES (Wallner, Elofsson 2006). MetaMQAPII uses these listed to detect local inaccuracies crystal structures and homology models. MetaMQAPII assigns each residue of a structure into one of 315 groups. For each of the groups a specific linear regression is created to predict the deviation of a residues based on a set of parameters evaluated by the server. The deviation is ranked on a scale of 1 to 100. The ranking can then be converted to distance (Å). The input for the MetaMQAPII server is a PDB structure file and the generated output consists of the three files. Of these files an output PDB file is generated that be visualized in PyMol (Schr\odinger 2015) to visual gauge the deviation observed.

2.2.1.5.4. Model evaluation using RAMPAGE

Often the most important locus regarded for the distortion of the covalent geometry in protein structures is the Cα. This is due to the fact that the Cα acts as a link that unites the protein residue sidechain to its relative position on the backbone of the protein. When the junction between the side chain and the back bone is incorrectly described, an incorrect local minimum is generated. The incorrect local minimum is also accompanied by a distortion in the Cα geometry to compromise on the energy change. Due to this, for structure validation, three major components are considered. These components include the conformation of the backbone, the conformation of the side chain and the geometry Cα. The validation program RAMPAGE utilizes Ramachandran

diagrams in order to assess the ϕ versus ψ angles of the backbone to determine if the conformation is favorable or not (Lovell et al. 2003).

2.3. Methodology

2.3.1. Data retrieval

The Pfam 27.0 database (Finn et al. 2016) was used to retrieve AA9 protein sequences (Pfam id: PF03443). At the time the Pfam databased contained 827 AA9 modular protein sequences with additional domains. All the retrieved proteins sequences were obtained from 87 fungal and one plant organism (*Zea mays*). The MAFFT alignment webserver (Katoh, Toh 2008) was used to identify duplicate sequences, highly variable sequences as well as short fragments of AA9 protein sequences. These sequences were then removed from the dataset. To remove these sequences, the AA9 domains was extracted from all aligned sequences. After all the AA9 domains were extracted, a Python script was used to remove sequences with a 100% sequence identity (duplicates). This resulted in a smaller dataset of 139 AA9 protein sequences. Sequences that had already been characterized as specific AA9 types and PDB files were then added to the dataset (Table 2.2 and Table 2.3) (Li et al. 2012).

Table 2.2: Available AA9 PDB structures.

PDB sequences				
PDB ID	**Uniprot accession**	**Organism**	**AA9 PMO type**	**Citation**
4B5Q	H1AE14	*Phanerochaete chrysosporium*	**Type 1**	(Wu et al. 2013)
3EII	D0VWZ9	*Thielavia terrestris*	**Type 1**	(Harris et al. 2010)
3EJA	D0VWZ9	*Thielavia terrestris*	**Type 1**	(Harris et al. 2010)
4EIS	Q7SA19	*Neurospora crassa*	**Type 1**	(Li et al. 2012)
4EIR	Q1K8B6	*Neurospora crassa*	**Type 2**	(Li et al. 2012)
2YET	G3XAP7	*Thermoascus aurantiacus*	**Type 3**	(Quinlan et al. 2011)
3ZUD	G3XAP7	*Thermoascus aurantiacus*	**Type 3**	(Quinlan et al. 2011)
2VTC	GUN7	*Trichoderma reesei*	**Type 3**	(Karkehabadi et al. 2008)

Table 2.3: ***Neurospora crassa*** **reference sequences.**

Neurospora crassa **reference sequences**				
Protein	**Uniprot**	**Organism**	**AA9 PMO type**	**Citation**
NCU00836	Q7SCJ5	*Neurospora crassa*	**Type 1**	(Li et al. 2012)
NCU03328	Q1K4Q1	*Neurospora crassa*	**Type 1**	(Li et al. 2012)
NCU02344	Q7S411	*Neurospora crassa*	**Type 1**	(Li et al. 2012)
NCU01050	Q1K8B6	*Neurospora crassa*	**Type 2**	(Li et al. 2012)
NCU02240	Q7S439	*Neurospora crassa*	**Type 2**	(Li et al. 2012)
NCU02916	Q7SHI8	*Neurospora crassa*	**Type 2**	(Li et al. 2012)
NCU07898	Q7SA19	*Neurospora crassa*	**Type 3**	(Li et al. 2012)
NCU05969	Q7S1V2	*Neurospora crassa*	**Type 3**	(Li et al. 2012)
NCU07760	Q7S111	*Neurospora crassa*	**Type 3**	(Li et al. 2012)

These sequences were from the *Neurospora crassa* fungal organism and in this study these sequences were used as reference for Type 1, Type 2 and Type 3 AA9 LPMO types. The accession numbers of the 139 sequences are available in Additional file 1. AA9 crystal structures (Table 2.2) added to the dataset resulting in the final dataset included of 153 AA9 domain sequences, some of which also contained a signal peptide region. The corresponding AA9 nucleotide sequences were retrieved from Genbank using a Python script. The nucleotide sequences were then grouped into their respective types.

2.3.2. Multiple Sequence Alignment

The PROMALS3D alignment tool was used to align the final dataset. PROMALS3D generally generates more alignments, due to the fact that it takes into account structural information when generating alignments. From the crystal structures in Table 2.2, 4B5Q, 4EIR and 3ZUD were used as input for Type 1, 2 and 3 LMPO types respectively. Phylogenetic clustering and sequence similarity with reference sequences used to cluster the sequences in the final dataset into their respective Types. All vs all sequence identity calculations were performed using Matlab to assess

the extent of sequence conservation in AA9 sequences as a whole and in individual AA9 Types. To obtain a codon alignment needed for both Selecton and DataMonkey, the nucleotide sequences were submitted to the Codon Alignment v2.1.0 tool to obtain the translated nucleotide sequences. Once the translated sequences were obtained, the MAFFT webserver was used to align the protein sequences. PAL2NAL (Suyama, Torrents & Bork 2006) was then used to generate a codon alignment of the nucleotide sequences using the MAFFT aligned protein fasta file.

2.3.3. Selection

To study the site specific selective pressure on AA9 protein sequences Selecton and DataMonkey were used. The input nucleotide AA9 alignment sequence files used for this analysis can be found in the electronic files Additional file 2, 3 and 4 for Type 1, 2 and 3 AA9 proteins respectively. The universal genetic code was used to analyse the data. For DataMonkey, two evolutionary models were employed to assess the evolutionary pressures that are present on AA9 proteins. These models were SLAC and FEL. The level of significance used was 0.1 to identify sites under positive, neutral or negative selection. The models used for selection in Selecton was the M8 model as this model is capable of detecting positive selection. Similar to DataMonkey, the universal genetic code was used. These methods were applied to AA9 sequences that were grouped into their respective LPMO types in order to observe if the was any type-specific selective on AA9 proteins.

2.3.4. Phylogenetic analysis

To cluster AA9 domains into their respective types and to observe type-specific variations in AA9 sequences phylogenetic analysis was applied. The Molecular Evolutionary Genetic Analysis (MEGA) v6.0 (Tamura et al. 2013) was used to generate all phylogenetic trees for the analysis. To select the best evolutionary model for computing the phylogenetic analysis the Bayesian information criterion (BIC) scores was used. The evolutionary models that displayed lowest BIC scores were selected for further analysis. Three gap deletions were used to select the best (90%, 95% and 100%). For the top three models Maximum Likelihood (ML) trees were generated considering each gap deletion. Phylogenetic trees were created using the Nearest-Neighbor-Interchange (NNI) and a maximum heuristics search was conducted. By default, MEGA uses Neighbor Join and BioNJ algorithms to generate the initial trees. For each constructed tree 1000 bootstrap replicates were performed using a very strong branch swap filter. The top three evolutionary models where determined to be Whelan And Goldman model (WAG) (Whelan,

Goldman 2001), WAG and Gama distribution (WAG+G) and WAG+G with Invariant sites (WAG+G+I) in that order. For all three models, three specified gap deletions were used to perform phylogenetic calculations resulting in nine phylogenetic trees. The generated phylogenetic trees were then compared to their respective bootstrap consensus tree to select the best tree. The best phylogenetic tree was obtained using the WAG+G+I evolutionary model with a 90% gap deletion.

2.3.5. Physicochemical property analysis

To observe type-specific physicochemical property differences in AA9 proteins physicochemical property analysis was done. This analysis was performed after phylogenetic clustering on individual AA9 types. The properties that were analyzed were the aromaticity (Lobry, Gautier 1994), GRAVY index, isoelectric points (Bjellqvist et al. 1994), instability index (Guruprasad, Reddy & Pandit 1990), molecular weights and the amino acid residue composition. The aromaticity of a protein is a relative measure of the proportion of aromatic residues of a protein sequence. The ProteinAnalysis class from the ProtParam module in BioPython was used to calculate the Aromaticity. The GRAVY index is an index that describes the hydropathicity/solubility of a protein. Proteins with a positive GRAVY index are regarded as hydrophobic and while a protein with a negative GRAVY index is considered hydrophilic (Kyte, Doolittle 1982). To determine if there were any significant differences in the observed physicochemical properties of different types, the student t-test was used. The R package was used to conduct the t-test at a 5% level of significance.

2.3.6. Homology modelling

For the identified Type 1 sequence variant termed *Aspergillus niger* (*A. niger*) AA9 homolog 9 (Uniprot accession: G3XUH5.1) a homology model was calculated. Due to characterization of the *A. niger* homolog 9 as Type 1, the template chosen for modelling was the 4B5Q crystal structure. The HHpred webserver (Soding, Biegert & Lupas 2005) was used to identify this template. The 4B5Q was regarded as the best template to model *A. niger* AA9 homolog 9 due to the observed phylogenetic clustering of the 4BQ5 structure with other Type 1 sequences. The 4B5Q crystal structure was also found to have the highest sequence identity (39%) with *A. niger* AA9 homolog 9 had the least gaps which with no disruption of the HHpred predicted secondary elements. MODELLERv9.12 (Webb, Sali 2014) was used to perform homology modelling of *A. niger*

homolog 9. Using very slow refinement 100 models were created. To select the best models, homology models were ranked using the DOPE Z-score (Shen, Sali 2006). The best 3 models were assessed with the MetaMQAPII (Pawlowski et al. 2008) and RAMPAGE was used to select the best model.

2.3.7. Motif analysis

Motif discovery was performed on the final dataset with the Multiple Em for Motif Elucidation (MEME) webserver (Bailey et al. 2006). MEME was used detect motifs that have a size range between 6 and 50 residues. The Motif Alignment Search Tool (MAST) (Bailey, Gribskov 1998) was find motifs that occur on the same region on the protein. 100 motifs were initially identified however the most optimal number of motifs was determined to be 30**.** Matlab was used to parse the MEME log file to visualize the conservation of the motifs using a heat map. Type-specific motifs were identified and these were mapped on respective crystal structures to observe unique structural features and potential interaction with cellulose. This analysis was performed on all AA9 proteins in order to determine motifs that are common in all AA9 proteins and also those that are type-specific.

2.3.8. Manual docking and structure mapping

To assess structurally important unique type-specific features and their potential interaction with cellulose, manual docking was performed. The protocol followed for manual docking was similar to that performed by previous studies (Li et al. 2012). The protocol involves aligning the aromatic residues on surface of AA9 proteins to the glucose residues of the cellulose face. The surface exposed aromatic residues described by Li *et al.* (Li et al. 2012) were identified in respective AA9 crystal structures and aligned to sugar residues on the cellulose surface. The cellulose substrate was constructed such the five cellulose chains were created. Each cellulose chain consists of 12 pyranose residues. The cellulose substrate was constructed using Iβ coordinates (Nishiyama, Langan & Chanzy 2002). Identified Type-specific features crystal structures 3EJA, 4EIR and 3ZUD to represent were Type 1, 2 and 3 AA9 proteins respectively. Type 1 proteins used Tyr-190, Tyr-191 and Tyr-67 on the 3EJA structure to align to the cellulose pyranose residues of cellulose. The aromatic residues His-1 and Tyr-206 were used to align the 4EIR Type 2 crystal structure to cellulose. Type 3 AA9 crystal structure was aligned using His-1, Tyr-24 and Tyr-212 to cellulose.

For the *A. niger* homology model 9 His-1, Trp-34 and Trp-207 were aligned to cellulose. The planar flat surface active site aromatic residues are used to align AA9 proteins due to their similar spatial organization to the small Carbohydrate Binding Modules Family I (CBMI).

2.4. Results

In this Chapter type-specific features were identified on respective AA9 proteins. The evolutionary selective pressure was also assessed. Type-specific AA9s were identified at a sequence, structural and physicochemical level. This analysis was performed on the 153 AA9 domains sequences obtained from the Pfam database including the reference sequences. AA9 protein sequences were found to be under negative to neutral selection as opposed to positive selection.

2.4.1. Multiple sequence alignment shows AA9 PMO type-specific inserts

The final dataset of 153 AA9 sequences were aligned using PROMALS3D. The full AA9 PROMALS3D alignment is found in Additional file 5. The alignment results showed that the N-terminus of AA9 proteins is highly variable as compared to their C-terminus. Inspection of the N-terminus revealed the presence of type-specific insertions and deletions as shown in previous studies (Vu et al. 2014). Finding suggests that the inserts of the N-terminus region is crucial for type-specificity. Three distinct sequence groups were found in Additional file 5, these sequence groups were found to be associated with type-specific inserts as shown in Figure 2.1.

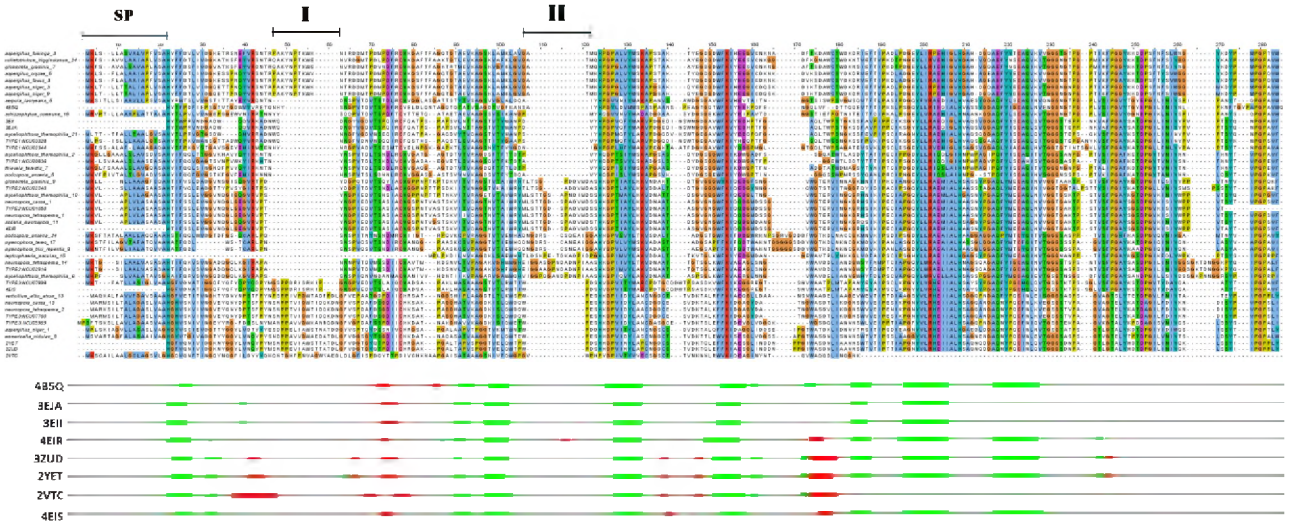

Figure 2.1: PROMALS3D alignment of AA9 domains. An alignment of 47 representative sequences to indicate the different AA9 sequences retrieved. The signal peptide (SP) is shown above the sequences, along with insert regions I and II. The PDB IDs of structures used as PROMALS3D input are shown at the bottom along with a representation of secondary structure along their sequences. Alpha helices are shown in red and beta-sheets are shown in green.

In Figure 2.1 two type-specific inserts were found. These inserts were called insert I and insert II. In Type 1 AA9 LPMO sequences both inserts are absent. The Type 2 AA9 LPMO protein sequences were associated with insert II. While the Type 3 AA9 LPMO sequences were associated with insert I. there was also a small group of Type 1 LPMO sequences possessed an 8 residue insert in the insert I region. Phylogenetic clustering and motif analysis both revealed that these sequences were type 1 LPMOs. As a result, because of the unique nature of these sequences a single representative sequence was selected for homology modelling. This was done in order to assess the effect of the 8 residue insert on the AA9 active site surface and its role in substrate interaction. In previous studies (Vu et al. 2014), it has been observed that protein structures with region I inserts (Figure 2.1) have a modification on the flat active site surface that is unique to Type 3 LPMO sequences. Similarly, Type 2 LPMOs were also found to have modification characteristic of the insert II. As a result, the relative position of the inserts on respective AA9 crystal structures and the homology model was determined by mapping these regions on their respective structures as shown in Figure 2.2. The structures were then aligned to cellulose in order to gauge the possible contribution of the type-specific regions on cellulose binding.

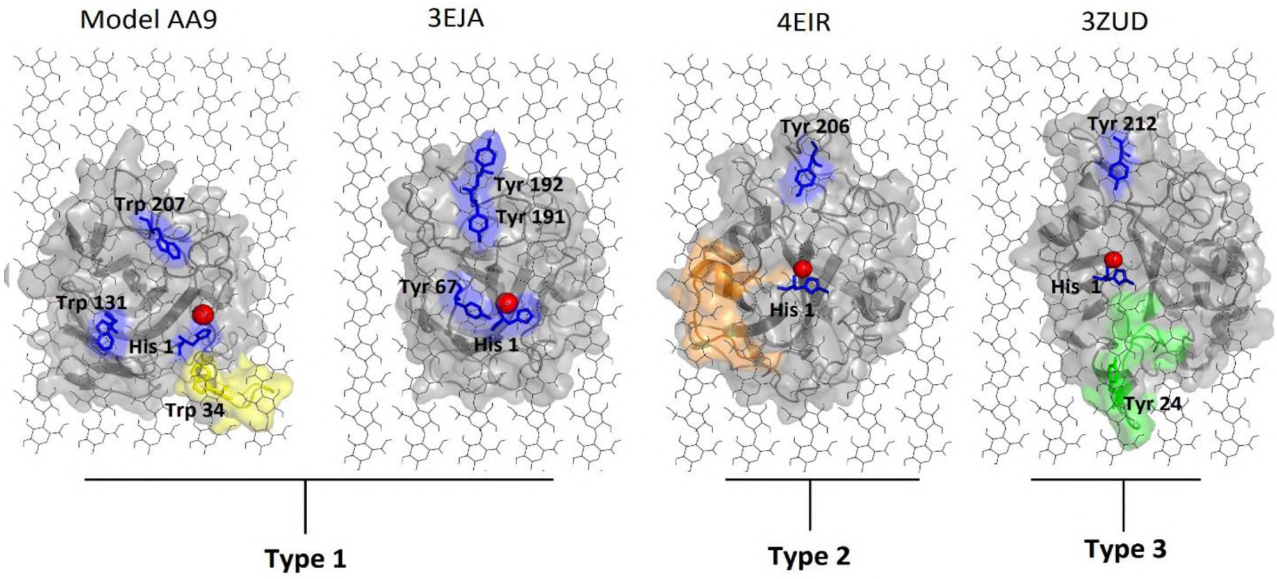

Figure 2.2: Structural representation of type-specific inserts of AA9 domains. Proteins are shown as cartoons and colored in gray, with the inserts colored in green and orange for region I and II respectively. The Type 1 insert II is colored in yellow. The cellulose is shown in black lines and the flat surface aromatic residues are represented in sticks and are colored in blue, with the type II copper ion shown as a red sphere. The planar residues Trp-34 on the model AA9 is colored in yellow as it forms part of the region I insert. On the 3ZUD crystal structure the Tyr 24 residue is colored in green as it forms part of the region I insert.

In Figure 2.2, the AA9 structures were manually aligned to the constructed cellulose Iβ substrate. The alignment was achieved using the planar aromatic residues on the AA9 active site. On the 3ZUD crystal structure the Tyr 24 residue is colored in green as it forms part of the region I insert. Due to an absence of a crystal structure of the *A. niger* homolog 9 Type 1 LPMO sequence, homology modelling was employed to study the presence of the 8 residue insert of the active site surface. Several studies have proposed that AA9 proteins use the planar aromatic residues on their flat surface active site with cellulose (Karkehabadi et al. 2008, Harris et al. 2010). It has been suggested that the steric congestion of the AA9 active surface may influence how the different AA9 types interact with cellulose (Hemsworth, Davies & Walton 2013). AA9 proteins have been shown to have active sites that have aromatic residues similar to those of the surface-binding (Type A) CBMs (Boraston et al. 2004). Due to the presence of planar aromatic residues on the AA9 active site, the AA9 structures were aligned to cellulose Iβ using the planar aromatic residues (Figure 2.2). Planar aromatic residues were found on the region I in the Type 1 variants and Type 3 AA9 proteins. Insert I is localized near the active site in both Type 1 and 3 AA9 proteins (4B5Q and 3ZUD). In particular, insert I of the modelled *A. niger* homolog 9 AA9 protein was possessed the planar aromatic residue Trp-34 (Figure 2.2). The Trp-34 residue is located towards the end of the eight residue insert. The position of this insert suggests that this motif can possibly interact with cellulose. This observation also occurs in the Type 3 crystal structure 3ZUD. The planar Tyr-24 residue is localized in the region I insert in the Type 3 crystal structure. The planar Tyr-24 residue was well conserved in Type 3 AA9 proteins (Figure 2.1). The planar aromatic have been suggested in substrate interaction studies (Leggio, Welner & De Maria 2012). The region I insert I both Type 1 and 3 was found to be associated with polar residues (Figure 2.1). In Type 1 AA9 protein sequences, relative to the *A. niger* homolog 9, were identified to be Thr-32 and Lys-33. Through manual docking this polar residues were found to have potential interaction with cellulose. In Type 3 AA9 protein sequences the region I insert was similarly constituted with polar residues. The polar residues were identified as Ser-26, Thr-37, Thr-39 and Glu-40 on the 3ZUD crystal structure. The localization of insert I showed that the insert has the potential to interact with cellulose. In the Type 2 sequences the specific to insert II was detected. The region II insert in the Type 2 4EIR crystal structure was found to be devoid of planar aromatic residues however, polar residues were detected. The polar residues on insert II relative to the 4EIR crystal structure were identified as Ser-70, Thr-71, Thr-72, Ser-75 and Asp-78. The orientation of these residues on the

A9 structure revealed that these residues were facing the cellulose substrate. In Figure 2.1 it can be seen that the Region I insert is more conserved relative to insert II. Studies have hypothesized about the effects of the active site configuration on their ability to metabolize C1, C4 or both cleavage positions (Hemsworth, Davies & Walton 2013). The localization of these type-specific inserts on the AA9 active site and their potential interaction with cellulose suggest a role in substrate binding, orientation and cleavage (Figure 2.2).

2.4.1.1. All vs all sequence alignment heat maps

All vs all sequence identity calculations were assessed to investigate the extent of sequence conservation in all AA9 proteins in the dataset and the individual sequence groups. The cluster of the sequences groups observed in Figure 2.1 and the reference sequences in Table 2.3 was monitored to help clustering the sequences in to their respective Types. Individual PROMALS3D alignments were created for each AA9 PMO type (Additional files 6, 7 and 8 for Type 1, 2 and 3 PMOs, respectively) using PROMALS3D. Through All vs all sequence identity calculations (Figure 2.3) it was found that individual AA9 types have higher sequence conservation as opposed the complete alignment of AA9 proteins.

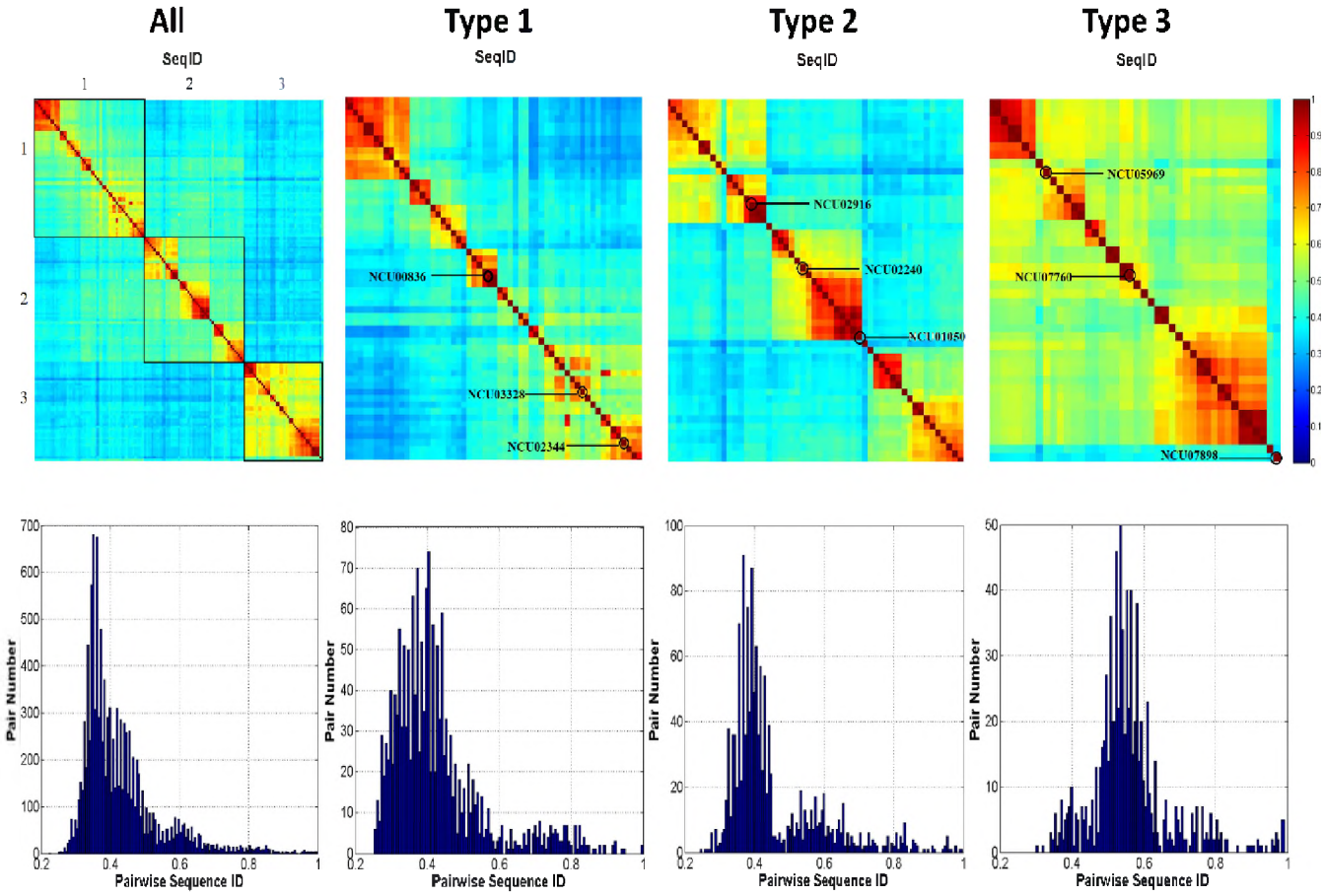

Figure 2.3: Sequence identity heat maps. The pairwise sequence identity values of all AA9 domains as well as sequences from each type of AA9 PMO are shown. The heat maps at the top show these scores as a color-coded matrix, of every sequence vs every sequence in their respective groupings shown. Heat maps are colored from blue to red with red show high conservation and blue showing low conservation. In the first panel (All), the blocks are numbered to indicate sequences from each AA9 PMO type. The positions of the Type 1, 2 and 3 reference sequences within each heat map are indicated. In the histogram at the bottom, the pairwise sequence identity values are shown and the number of sequence pairs with this value is shown on the y-axis. Low conservation across was observed for all AA9 LPMO Types in the alignment.

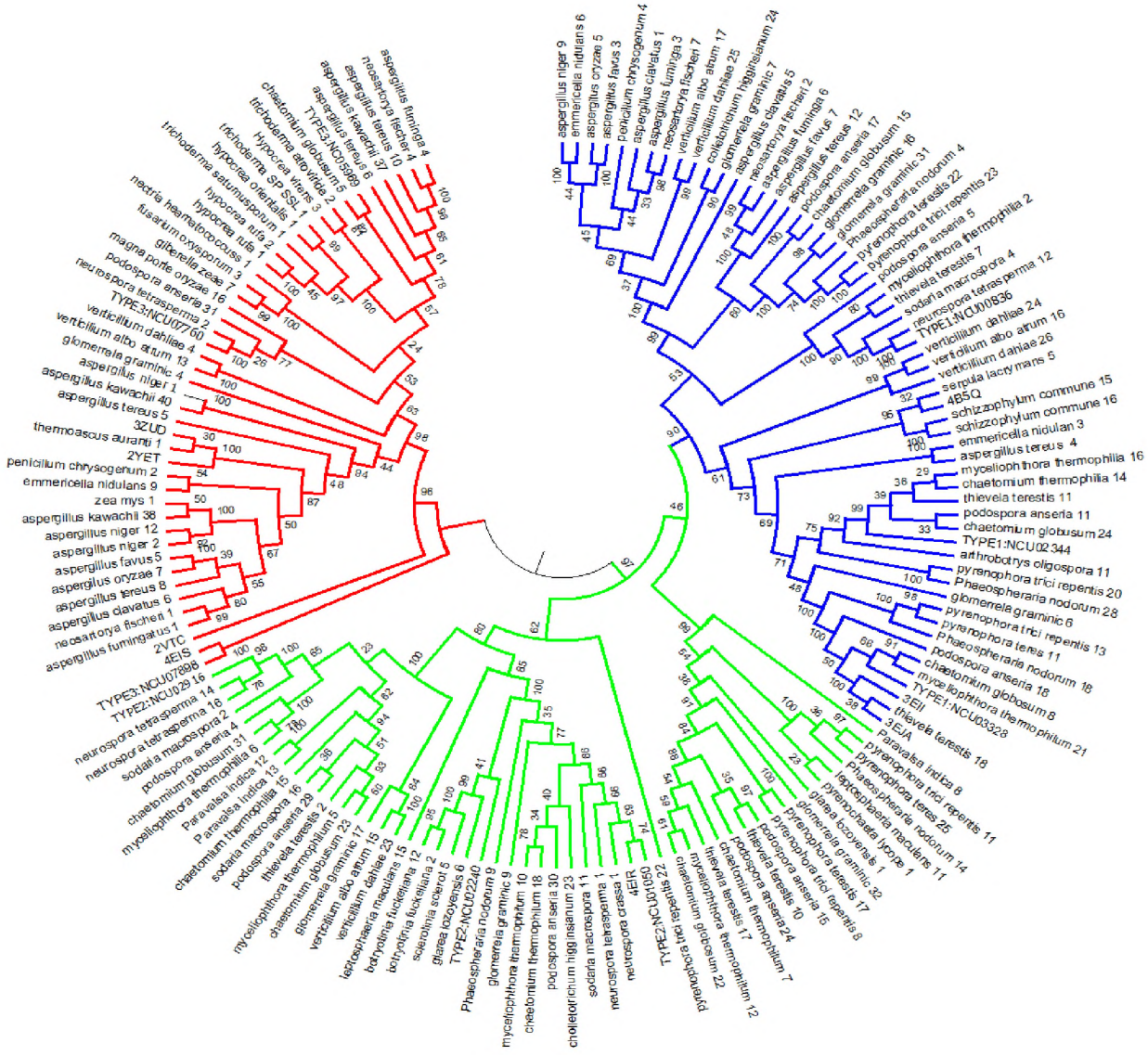

Figure 2.4: Molecular phylogenetic analysis by Maximum Likelihood method of AA9 proteins at 90% site coverage. Maximum likelihood tree constructed using MEGA6. Sequences are colored as follows: Type 1 – Blue; Type 2 – Green; Type 3 – Red.

Three distinct conservation regions were observed as indicated by dark lines. The observed conservation in the full alignment were found to correspond to the expected AA9 types in the dataset. As a result, despite the low sequence diversity among all AA9 proteins, a higher sequence identity among AA9 LPMO types was shown. The localization of reference sequences were mapped to the heat map (Figure 2.3) for Type 1, 2 and 3 proteins. This was done to show which types the reference sequences were related to. In the all heat map, sequence similarity between Type 1 and Type 2 sequences was observed as shown by the green blocks between the two. The

histogram of the distribution of sequence identity for the all vs all heat maps shows that the majority of AA9 domains have a sequence identities that fall between the range of 0.3 – 0.4 showing the high variation as characteristic of AA9 proteins. The Type 1 LPMO protein sequences were found to have high sequence variation. There were distinct conservation of sequences observed within Type 1 LPMOs as indicated by the distinct zones. This suggests that there could be possible Type 1 protein variants. The variation of Type 1 sequences was investigated phylogenetically (Figure 2.4). It was found that there were distinct phylogenetic clusters which were found to correspond to conservation observed previously in Figure 2.3. Mapping the conservation regions into phylogenetic tree is shown in Figure 2.5. This mapping resulted in the identification of sequence clusters that could indicate possible variation amongst AA9 proteins.

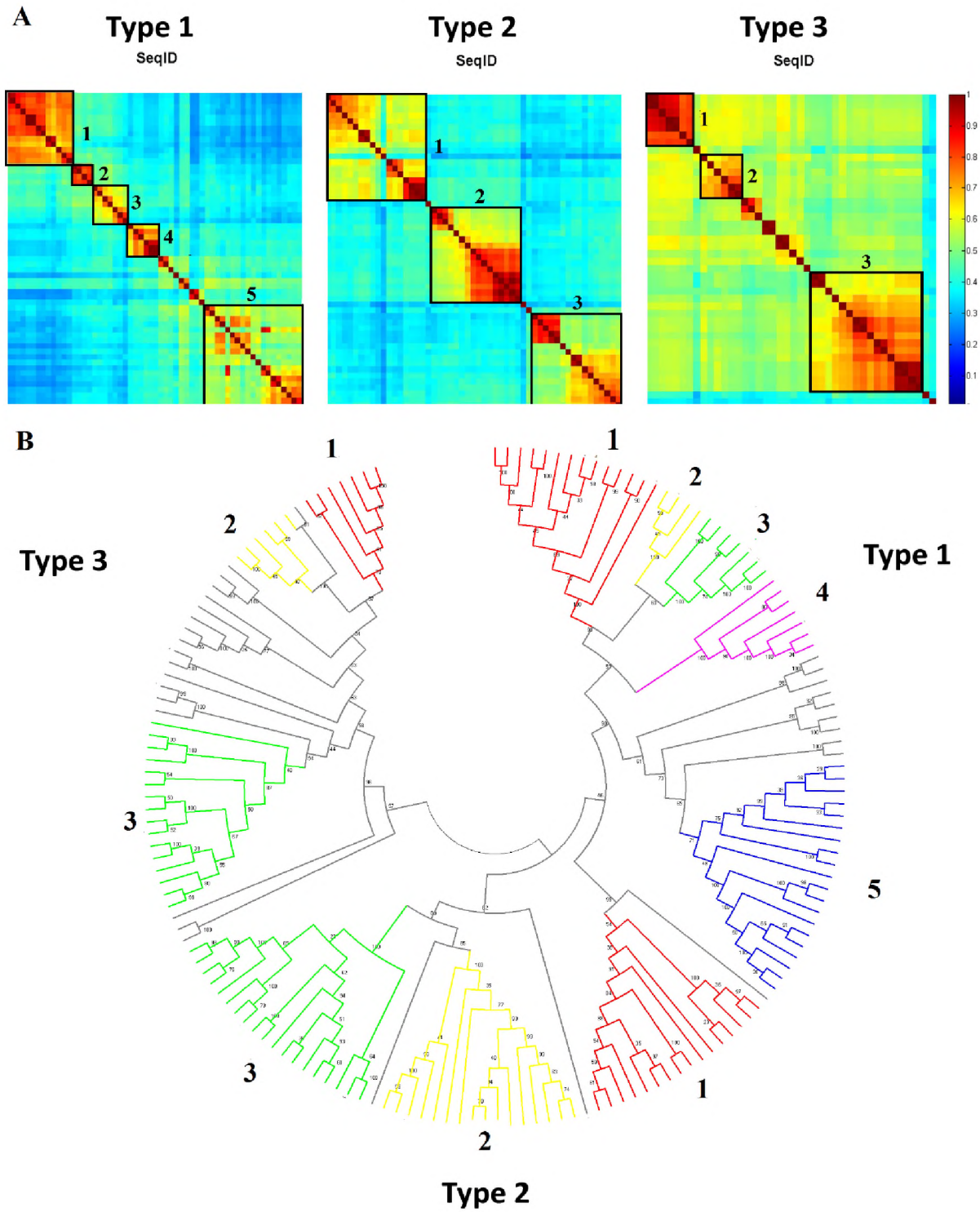

Figure 2.5: Mapping of sequence identity conservation boxes to the phylogenetic tree. A) Clusters within the each sequence identity heat map in Figure 3 are shown as boxes. B) The positions of these sequences are then highlighted on the phylogenetic tree from Figure 2.4.

For Type 2 AA9 protein sequences, three possible variations were identified as revealed by the distinct conservation boxes. Phylogenetic clusters were identified that correspond with the three identified Type 2 conservation (Figure 2.4). The distributions of sequences identities here as a series of histograms (Figure 2.3) for Type 2 protein sequences. It was revealed that the Type 1 LPMOs, generally have sequence identities below 0.4. The Type 3 AA9 protein sequences were found to be the most conserved. Generally the sequence identities of Type 3 protein sequences were above 0.5. However, these Type 3 sequences had very low sequence identities when comparing it to other types. It is possible that the presence of conservation region among AA9 LPMO types could reveal a previously undescribed phylogenetically distinct sub-groupings. AA9 proteins are renowned for being sequence diverse, also the modularity of proteins (Li et al. 2012) and the presence of distinct sequence sub-groups suggests AA9 proteins with more specialized. Histograms revealed the extent of conservation among AA9 types. The mean values were also calculated for all AA9 sequences and each types. For all sequences, Type 1, 2 and 3 groups the, mean sequence identities were 0.39, 0.44, 0.46 and 0.57. The corresponding standard deviations were 0.02, 0.02, 0.04 and 0.05 respectively. This reveals that Type 3 had higher conservation than the other AA9 groups. Figure 2.5 reveals that there were two Type 3 protein sequences with very low sequence identity as opposed to the other sequences. The variable sequences were identified as the reference sequence NCU07898 and the crystal structure sequence 2VTC. In the phylogenetic tree, the NCU07898 and 2VTC sequences shown to form out-groups on the Type 3 phylogenetic branch as shown in Figure 2.4.

2.4.2. Type-specific motifs identified, which also reveal sub-groups

Motifs analysis was performed on AA9 proteins in order to identify type-specific conserved regions in respective AA9 LPMO types. The most optimal number of motifs was determined to be 30. The MEME webserver was used to identify the maximum number of motifs was set at 30. When a number above 30 was used, insignificant motifs were identified (Figure S1). The 30 motifs identified are summarised in Additional file 9 were the regular expression are shown. The MEME results were summarized using a heat map showing the extant of conservation of the motifs on a particular sequence. The motifs are ordered by the order returned by MEME, are displayed in Figure 2.6A. It was revealed that the motifs 1, 2, 3, 4, 5, 7 and 8 were conserved in all of the AA9 sequences with a few exceptions.

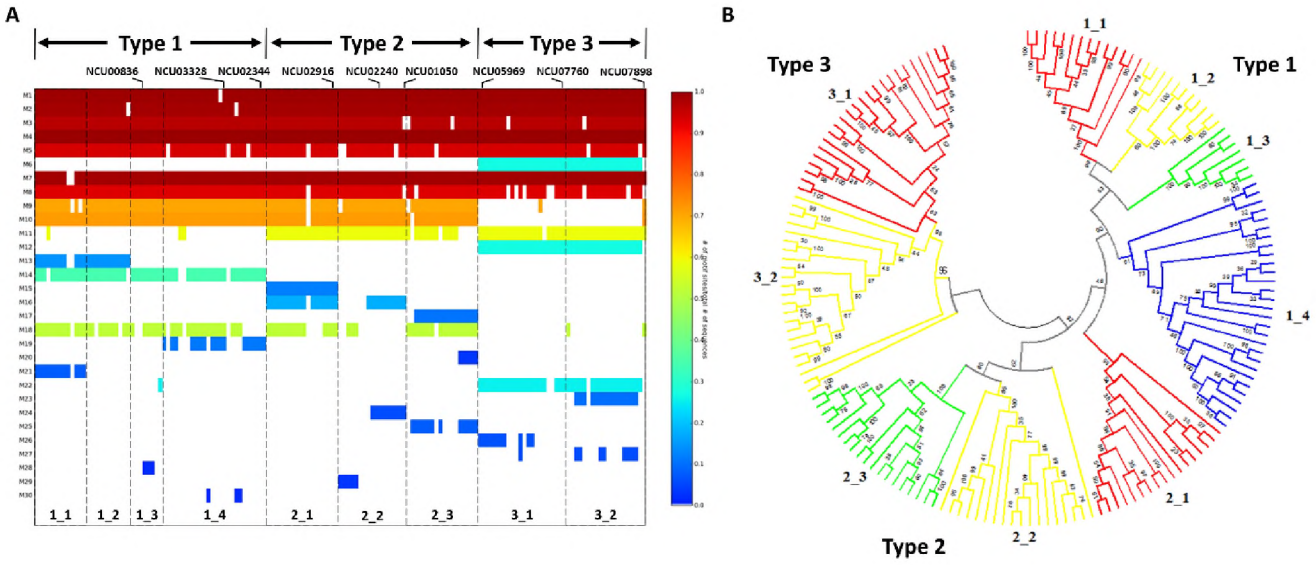

Figure 2.6: Motif analysis of AA9 domains. A) Heat map representing the extant of conservation of the identified 30 motifs on AA9 sequences (labelled on the left). Sequences are grouped according to type starting from Type 1-3. Motifs are colored based on conservation, as shown in the key. Divisions are indicated to show the sub-groups within each AA9 PMO type identified during motif analysis. The position of each reference sequences from *Neurospora crassa* is also shown on the motif heat maps. B) Mapping of AA9 sub-groups to the phylogenetic tree produced for AA9 domains. The sub-groups indicated in (A) are shown mapped to the phylogenetic tree from Figure 2.4.

Motif characterization of Type 1 AA9 LPMO protein sequences revealed the presence of four distinct sequence sub-groups. These sub-groups were characterized by the conservation of certain motifs as shown in highlighted in Figure 2.6A. In all Type 1 protein sequences, motifs 14 and 18 were present in all sub-groups. Type 1 protein sequences with motifs 13 and 21 were referred to as sub-group Type 1_1. Type 1 protein sequences with motifs Type 1_2 were shown to be similar to Type 1_1 but motif 21 was not present in these sequences. Type 1_3 sub-group sequences were associated with motifs 14 and 18 only, while the Type 1_4 subgroup had motif 19 together with motifs 14 and 18.

The Type 2 AA9 protein sequences were divided into three sub-groups. The Type 2_1 sub-group was associated with motifs 15, 16 and 18. The Type 2_2 sub-group was only associated with motif 16. The Type 2_3 sub-group possess the motifs 17 and 18. A few sequences of the Type 2_3 sub-group were found to be associated with motif 20. The Type 3 AA9 protein sequences were found to be least variable sequences with respect to motifs as shown in multiple sequence alignment (Figure 2.5).

All Type 3 protein sequences were found to possess motif 22. Two sub-groups were detected for Type 3 protein sequences. The Type 3_1 sub-group was associated with motif 26 and the Type 3_2 sub-group was associated with motif 23 and 27.

To make the detection of unique motif patterns easier, the input sequences were ordered based on phylogenetic clustering of AA9 sequences in (Figure 2.4). This was done to aid the identification of phylogenetic clusters that correspond to the observed motif sub-groups. Mapping of the subgroups to the phylogenetic tree shown was performed and the results are shown in Figure 2.6B. AA9 sequences of these sub-groups were found to form distinct phylogenetic branches. However, outgroups sequences in sub-groups 2_2 and 3_2 did not follow this observation. An overlap between the motif sub-groups and the sequence variants observed in Figure 2.5B. The variants were not detected in all AA9 sequences in Figure 2.5B however, a good overlap with motif sub-groups was observed (Figure 2.6B). The Type 1_1 specific motif 21, was determined to be the 8 residue insert found in insert I of Type 1 AA9 proteins shown in Figure 2.1 and displayed structurally in Figure 2.2. With respect to motif conservation across AA9 types, there were similar between AA9 Types. Between Type 1 and 2 AA9 proteins, motifs 9, 10 and 18 were shown to be common to both Type 1 and 2 sequences. However, in Type 3 sequences these motifs are absent. Type 3 and 2 sequences were both found to be associated with motif 11. Motif 18 was not part of the AA9 domain but was found to be associated with the signal peptide. Motif 14 was exclusively found Type 1 protein sequences. Type 2 sequences did not have any exclusively conserved motifs. The conservation of Type 3 sequences was demonstrated by sequence alignments. This was also reflected by motif analysis. Motifs 6, 12 and 22 were conserved throughout the sequences with motifs 6 and 12 being specific to Type 3 sequences. However, three proteins (crystal structures 4EIS, 2VTC and the reference sequence NCU07898) were found to lack these motifs. These three protein sequences were shown to form an out-group in the phylogenetic analysis. Motif analysis revealed the presence of type-specific motifs. The type-specific motifs were identified as Motifs 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19. On their representative Type 1, 2 and 3 crystal structures (3EJA, 4EIR and 3ZUD respectively), the type-specific motifs were mapped (Figure 2.7A). In Figure 2.7, the motifs 13, 17 and 18 were not added because the regions that these motifs occur were absent in the crystal structures. Sub-groups specific were identified which were Motifs 13 and 17. As stated previously, motif 18 was in the signal peptide region of AA9 proteins. The signal peptide region is removed from fully matured protein. This analysis identified type-specific

regions situated on the active site surface of AA9 protein sequences. Previous studies (Hemsworth, Davies & Walton 2013) have proposed that the arrangement of the active site surface. Configuration of the active site is a likely contributor to AA9 LPMO types. This finding is consistent with the findings of this study. Common motifs among AA9 PMO types were also demonstrated in Figure 2.7.

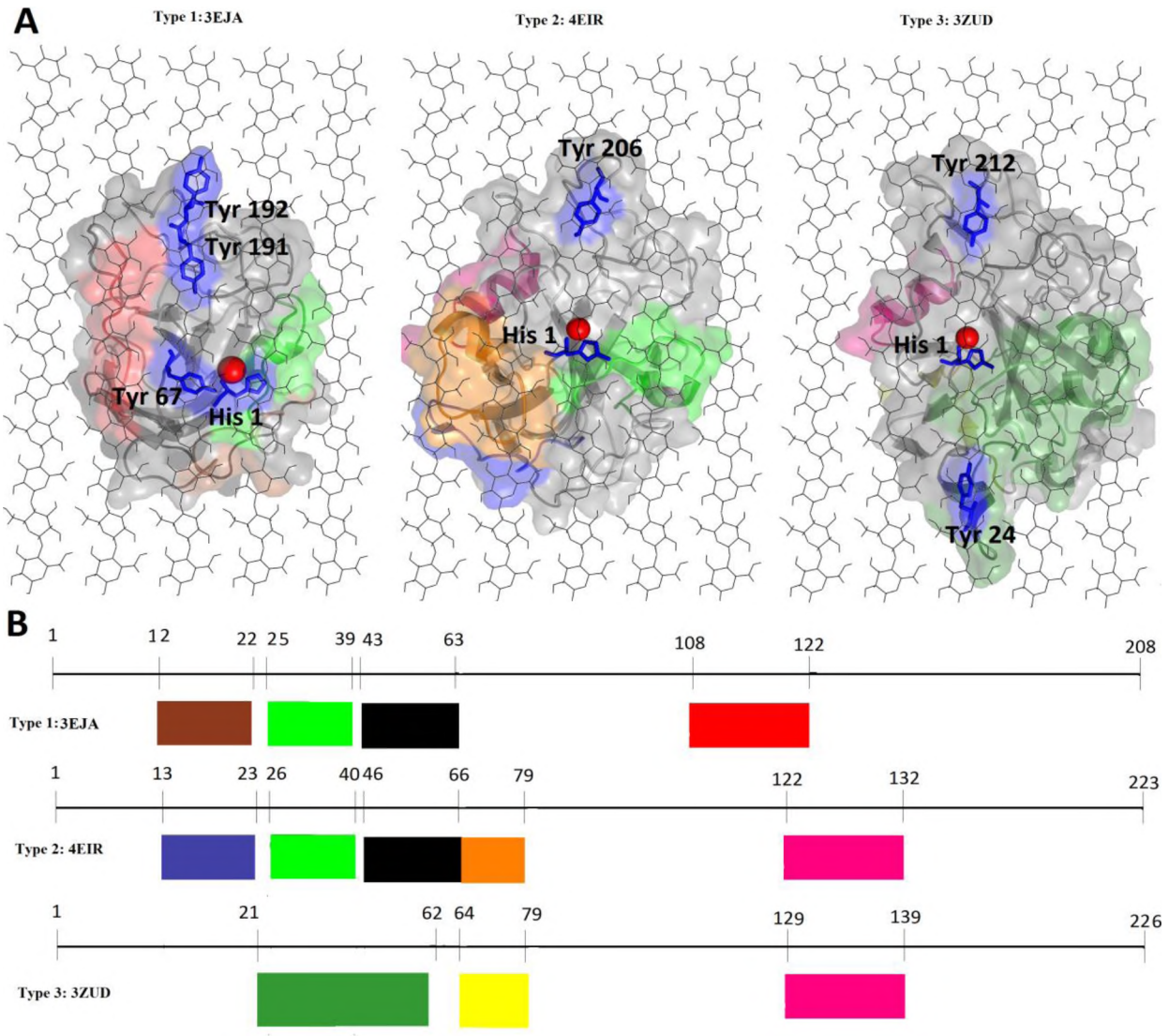

Figure 2.7: Visualization of type-specific motifs on crystal structures and linear sequences. A) Structural visualization of type-specific motifs**.** Crystal structures 3EJA, 4EIR and 3ZUD were used to represent the AA9 PMO types respectively. The AA9 structures were aligned onto cellulose Iβ using the planar aromatic residues on the AA9 active site. The flat surface aromatic residues are represented as sticks and are colored in blue, with the type II copper ion shown as a red sphere B) Linear representation of N-terminal type-specific motifs**.** The motifs are visualized on structures 3EJA, 4EIR and 3ZUD linear sequences to show type-specific motifs. The motifs shown are motifs 6, 9,

10, 11, 12, 14, 15, 16 and 19, colored based on where they are mapped to the structure (A). The motifs are colored as follows: Motif 6 dark green, motif 9 green, motif 10 black, motif 11 pink, motif 12 yellow, motif 14 brown, motif 15 orange, motif 16 dark blue and motif 19 red.

For both Type 1 and Type2 LPMOs example, motif 9 was conserved. Motif 19 was determined to be which helical structure present in both Type 1 and 2 sequences. Similarly, Type 2 and Type 3 LPMOs also had conserved motif 11. The N-terminus of AA9 domain sequences (Figure 2.7B) was found to be variable as shown in the sequence alignment. Motif 9 found to have potential interaction with cellulose chains 4 and 5 (Figure 2.7) for both Type 1 and 2 AA9 proteins. The motif 9 is replaced by the large motif 6 in Type 3 AA9 protein sequences. The motif 19 of Type 1 AA9 proteins had close proximity with cellulose chain 2 suggesting potential interaction. Motif 11 replaces motif 19 in Type 2 sequences (Figure 2.7). In general the manual docking findings suggests that the type-specific motifs are potential contributors to regioselectivity. The AA9 sub-groups identified are based on the different motif compositions of different sequences within each AA9 type (Figure 2.6). The molecular dynamics studies will be performed in subsequent chapters to further understand the effect of different active site configuration on AA9 substrate interaction.

2.4.3. Selective pressure found on AA9 sequences

To understand the different selection pressures that are present on AA9 proteins Selecton and DataMonkey were used. In both webservers, different models were used to investigate the Ka/Ks ratio specific to a specific amino acid position relative to the input alignment. The result of the analysis is mapped onto respective crystal structures for the results of Selecton. For Type 1 AA9 proteins the representative crystal structure 4B5Q was selected, for Type 2 AA9 protein sequences the 4EIR crystal structure and for the Type 3 AAA9 proteins the 3ZUD crystal structure were used. The results of the selection analysis is shown in Figure 2.8 below. For Selecton the results are mapped onto respective AA9 structures in Figure 2.8 A. For DataMonkey Ka/Ks ratio is plotted relative to the codon positions. The results are shown in Figure 2.8 B, C and D for Type 1, 2 and 3 AA9 proteins respectively.

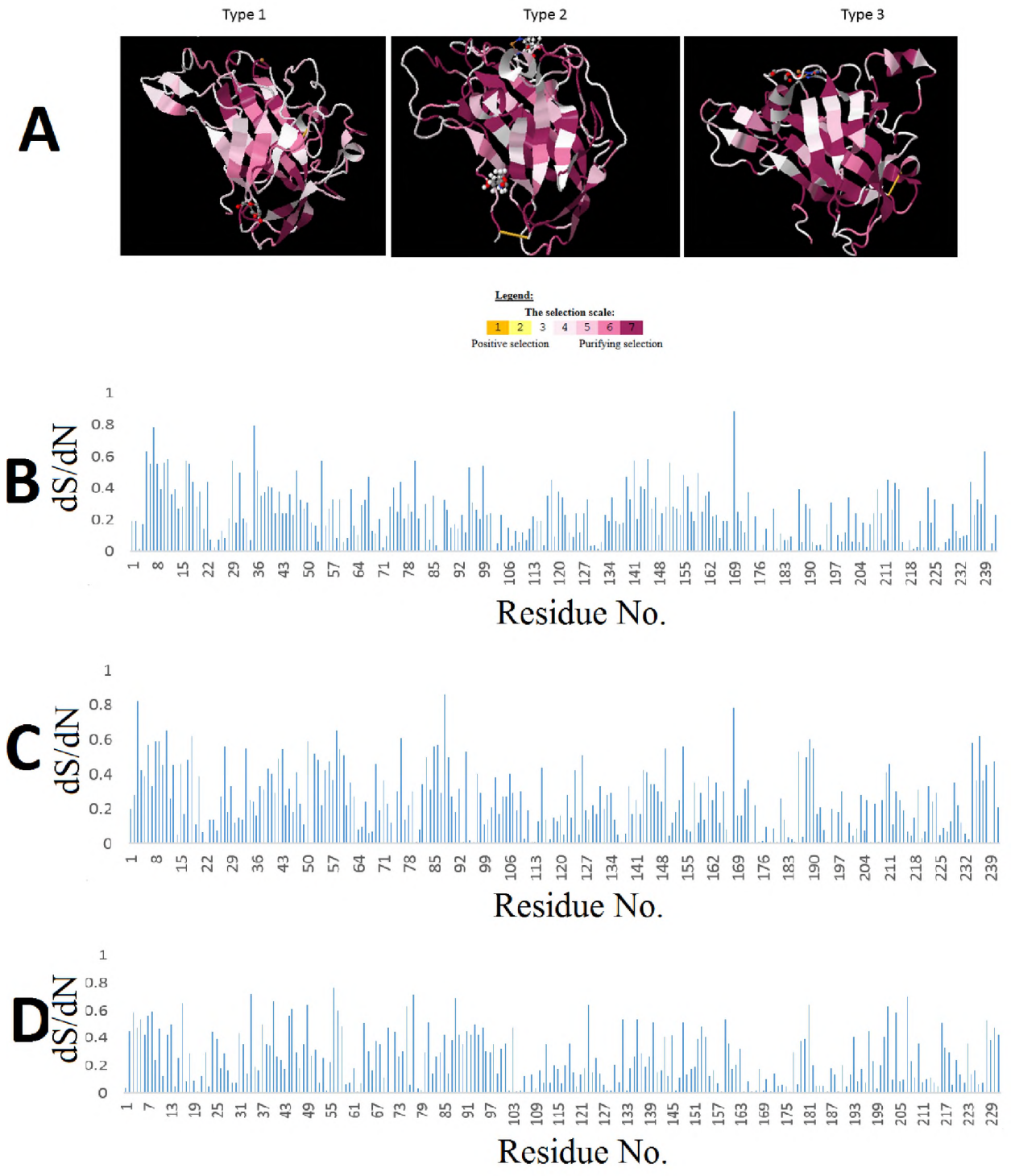

Figure 2.8: Selection on AA9 structures. A) Selection analysis of AA9 proteins using Selecton on Type 1, 2 and 3 AA9 proteins. The purple colors indicate sites that undergo negative selection while the white color indicates site that are undergoing neutral selection and yellow indicates sites that undergo positive selection. B, C, D are results of the DataMonkey analysis of selection for Type 1, 2 and 3AA9 proteins respectively. Sites with values greater than 1 are regarded as being positively selected.

For the selection analysis sites that were found to have dN/dS ratio less than 1 were considered as negatively selected, sites with a dN/dS ratio equal to 0 were considered as being under neutral selection while sites that had a dN/dS ratio above 1 were considered as being under diversifying or positive selection. The analysis showed that the AA9 proteins generally do not have any positively selected residues. However, patches indicative of both neutral and negative selection were found on AA9 structures. These findings were consistent in both the webservers used. The AA9 active site copper coordinating residues were found to generally undergo negative selection. Neutral selection was observed on the beta sandwich fold and the surface exposed active loop regions. These findings suggests that the selection pressure on the AA9 active site residues results in the conservation of these residues. The conservation of the AA9 Cu^{2+} coordinating residues is important for Cu^{2+} coordinating and therefore the negative selection of these residues is to be expected. Sites under positive selection are regarded as regions that undergo adaptive change in order to deal with certain changes (Yang et al. 2000). The fact that AA9 protein deal with a consistent substrate, a positive selection on residues of these enzymes is not to be expected.

2.4.4. Physicochemical differences is observed between different AA9 PMO types

To identify non-structural features that are specific to AA9 LPMO type's physicochemical property analysis was performed (Figure 2.9). The properties that were analyzed were Aromaticity, GRAVY, Isoelectric point, Instability index, and Molecular weight. Three boxplots for each evaluated physicochemical property for the respective AA9 types. The first feature analyzed was the Aromaticity of AA9 domains. Due to the fact that Aromaticity is a relative measure of aromatic residues in proteins. It is therefore a potentially important feature of AA9 proteins because of the proposed function of aromatic residues (Li et al. 2012).Figure 2.9 revealed the distribution of aromatic residues. It was found that there were significant differences between AA9 types with respect to Aromaticity.

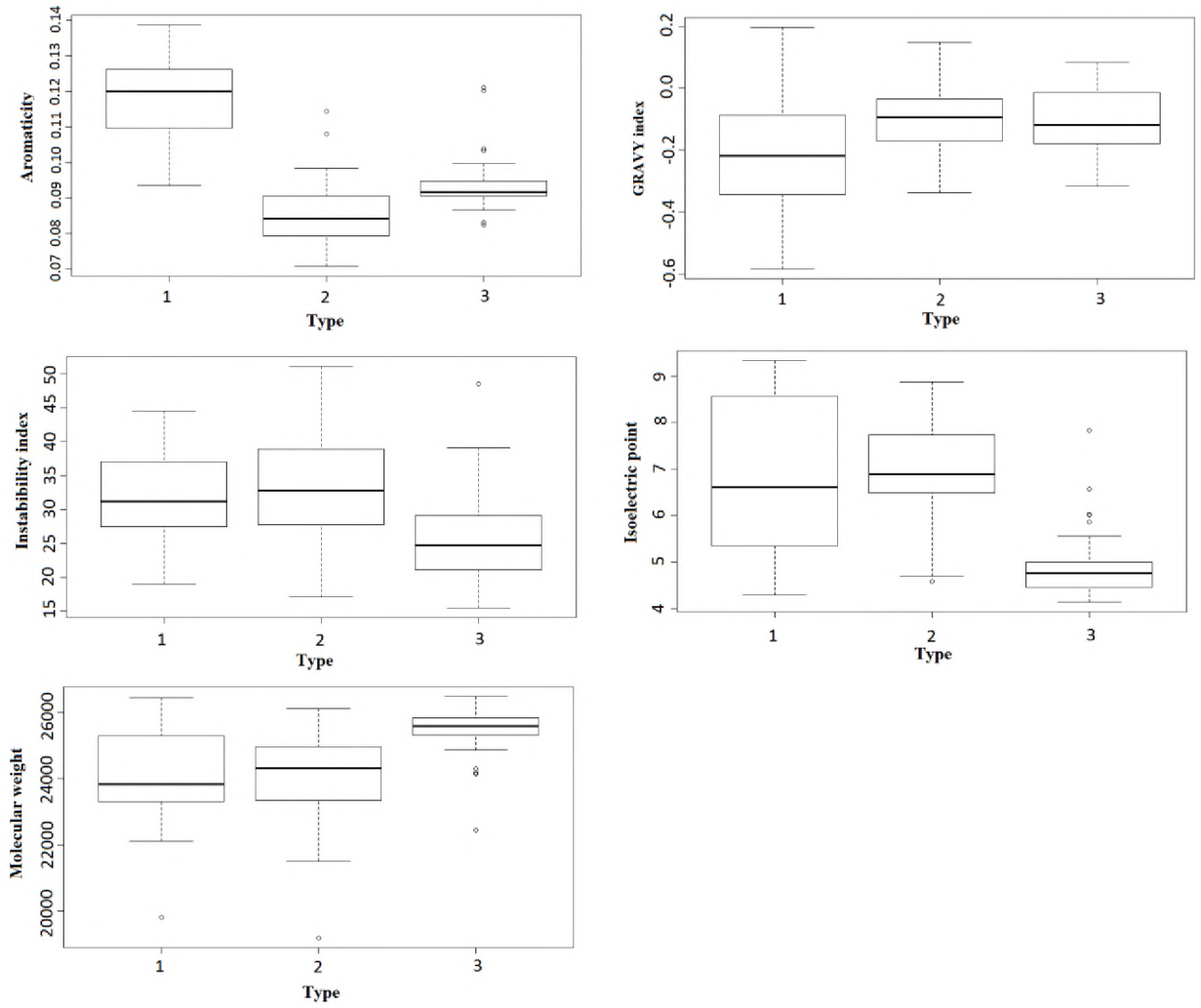

Figure 2.9: Boxplot representation of the distribution of the different physicochemical properties analyzed for Type 1, 2 and 3 in AA9 protein sequences. In their respective order, the properties analyzed were: Aromaticity, GRAVY index, Instability index, isoelectric point and Molecular weight. The numbering corresponds the respective types**.** The dark central line shows the median of the data. The lines on either side of the median show the upper and lower quartile which represent data that fall above 75% and data that falls below 25% respectively. The lines at the extremes represent the lower and upper fence which allows for the detection of outliers. 1.5 IQR (interquartile range) of the upper or lower quartile results in the fences. Outliers are indicated as dots that occur outside the whiskers of each plot.

The highest distribution of Aromaticity was observed for Type 1 LPMOs. Type 2 LPMOs were show to possess the lowest. It is well established that the active site AA9 domains has surface exposed residues. Of these the planar aromatic residues are considered to be crucial for function (Li et al. 2012). Due to this, the observed differences in Aromaticity in AA9 LPMO types may be important. To assess the differences between different AA9 different in terms of a specific

physicochemical property, the t-tests was used in R. The findings of the analysis are summarized in Table 2.3. The p-value between Type 1 sequences and both Type 2 and Type 3 was below 0.05 suggesting that significant differences among AA9 types in terms of Aromaticity.

Table 2.3: Results of the t-test, performed to compare the means of the different physicochemical properties at a 5% level of significance. In each block, a p-value is given, with a value < 0.05 indicating that the means of the two groups are significantly different.

Aromaticity		
	Type 1	Type 2
Type 2	2.20E-16	
Type 3	2.72E-12	1

GRAVY Index		
	Type 1	Type 2
Type 2	1	
Type 3	1	0.55

Instability index		
	Type 1	Type 2
Type 2	0.6848	
Type 3	5.47E-06	1.92E-05

Isoelectric point		
	Type 1	Type 2
Type 2	0.5705	
Type 3	3.54E-11	2.52E-16

Molecular weight		
	Type 1	Type 2
Type 2	0.5214	
Type 3	1	1

Figure 2.9 revealed quite a number of outliers in the aromaticity datasets. The outliers were identified for the respective LPMO types. For Type 2 sequences two outliers which were detected. These were: *Verticillium albo-atrum (V. albo-atrum)* AA9 homolog 15 and *Leptosphaeria maculans* (*L. maculans*) homolog 15. In in Figure 2.4 it shown these two sequences form two distinct sister groups within the Type 2 phylogenetic branch. Both *V. albo-atrum* AA9 homolog 15 and *L. maculans* homolog 15 were ultimately found to form outer groups in their respective sister groups. Their positions on the phylogenetic trees suggests that these sequences are very diverse. Within in Type 3, *V. albo-atrum* homolog 13, *Magnaporthe oryzae* (*M. oryzae*) homolog 16, NCU07898 and the crystal structure 2VTC were shown as to be upper quartile outliers with respect to aromaticity. The lower quartile outliers were determined to be *Aspergillus fumigatus* (*A. fumigatus*) homolog 4 and *Penicillium chrysogenum* (*P. chrysogenum*) homolog 2. On the phylogenetic tree it was found that the sequences NCU07898 and 2VTC sequences are outer

groups relative to Type 3 sequences. The other outlier did not appear to have any phylogenetic relationship relative to the other Type 3 sequences as shown by their clustering on the phylogenetic tree. In Figure 2.3 (bottom right corner) it is apparent that these two sequences have low sequence identity when compared to other Type 3 sequences.

In addition to aromaticity, the relative hydrophobicity was also of interest, so the GRAVY index was compared. Generally the LPMO types were hydrophilic. This is attributed to the negative GRAVY indices that were observed for most AA9 sequences. However, it was found that AA9 sequences in the upper quartile of all 3 AA9 type box plots are hydrophobic. Statistically, the distributions of the GRAVY indices AA9 LPMO types were not found to be significantly different (Table 2.3).

To assess the stability of AA9 proteins, the instability index was calculated for AA9 LPMO types. Proteins that have an instability index that fall below 40, have been shown as stable, while values above 40 are unstable (Guruprasad, Reddy & Pandit 1990). Most of AA9 sequences of the LPMO types were found to be stable as the instability index values mostly fall below 40. The Type 3 AA9 LPMOs were found to be the most stable group of proteins. With respect to all three types there were few sequences with instability indices above 40. One outlier was found for Type 3 protein sequences. The Type 3 *Fusarium oxysporum* (*F. oxysporum*) homolog 3 was found to have an instability of 48.477 suggesting this sequence is highly unstable. There was phylogenetic relationship observed for the detected instability index.

To aid in the elucidation of the functional importance, the isoelectric points (pI) of the three AA9 PMO types was calculated. The AA9 proteins are extracellular protein which may be expressed in various environments, as result it expected that AA9 proteins have a diversity with respect to physicochemical properties. A wide span of observed pI was seen for both Type 1 and 2 PMOs. For both these types, all these sequences possess acidic, neutral and basic pI values. Type 3 PMOs had more consistent pI values. The Type 3 sequences were found to span the neutral and highly acidic range with one acidic sequences being an exception. Four outliers for Type 3 AA9 proteins were identified. These were *M. oryzae* homolog 16, *Aspergillus clavatus* (*A. clavatus*) homolog 6, NCU07898 and the crystal structure 2VTC. The broad pH range observed Type 1 and 2 LPMO types may be functional importance given the extracellular nature of AA9 proteins. The acidic

nature of Type 3 protein sequences may indicate a specialized function for this group. The size of the AA9 LPMO types was passed in terms of molecular weight. A similar distribution of molecular weight was observed between Type 1 and 2 LPMOs ranging between 19-26.5 kDa. Type 3 AA9 LPMO sequences were found to have a size range between 23-27.6 kDa suggesting that these proteins are significantly larger than Type 1 and 2 LPMOs (Table 2.3). For all three AA9 PMO types, outliers were detected. Both Type 1 and 2 had one outlier each which were *Arthrobotrys oligospora* (*A. oligospora*) homolog 11 and *L. maculans* homolog 15 for Type 1 and 2. The outliers were for Type 3 the outliers identified as *Chaetomium globosum* (*C. globosum*) homolog 5, *Aspergillus terreus* (*A. terreus*) homolog 5, NCU07898, 2VTC and 3ZUD.

For AA9 LPMO types the composition amino acid was evaluated (Additional file 10). To visualize the results, 3D plots were generated (Figure 2.10). In the 3D plots the amino acids are ranked by their occurrence frequency from the lowest to the highest.

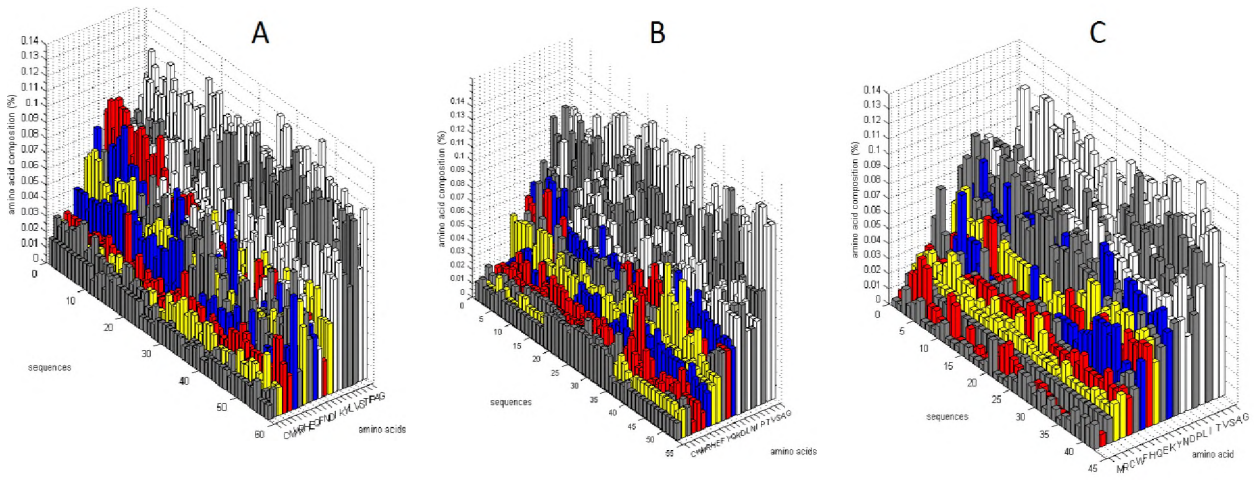

Figure 2.10: 3D plots showing the amino acid composition of the three AA9 PMO types. Amino acids counts are shown as a percentage for each sequence. The occurrence frequency of each amino acid is ranked from lowest to highest for clarity. Residues are colored as follows: Aromatic - Yellow; Positively charged – Red; Negatively charged – Blue; Hydrophobic – White. A) Type 1, B) Type 2 and C) Type 3 PMO sequences.

Across all types there was high variability in the amino acid composition of individual AA9 sequences. In all sequences it was shown that the hydrophobic residues (Ala, Gly, Ile, Val and Leu) have the highest frequency of occurrence as opposed to other residues. The polar residues Ser, Thr and Asp also had high occurrence frequency. However, Glu, Cys and Met had very low occurrence frequency in AA9 sequences. The residues that had the lowest occurrence frequencies

were charged and aromatic residues. Similarly the aromaticity previously observed (Figure 2.9) revealed relatively low content of aromatic residues in all AA9 LPMO types. Aromatic residues and polar residues are considered to be crucial for AA9 function. Even distribution of these residues in the AA9 domains and not just concentrated on the AA9 the active site. High amino acid residue variation in sequence amino acid composition was shown for AA9 sequences. The variation was not found to be type-specific variation could be observed.

2.5. Discussion

The goal of this chapter was to detect and characterize unique sequence and structural features specific to AA9 LPMO types. Initially, the intention was to apply the analysis to all AA9 sequences in the Pfam database. Short fragments and highly divergent sequences redundant sequences were subsequently removed. This resulted in a decrease of 827 sequences to 153 sequences of the dataset. The addition of reference sequences from *N. crassa* sequences to the final dataset was done to aid in the identification of type-specific sequence and structural features. Reference sequences were included in all experiment performed in this chapter to help associate of discovered features with specific AA9 LPMO types. Analysis of AA9 protein sequences was performed to determine unique AA9 LPMO type features. The AA9 sequences were aligned, motifs were discovered, and phylogenetically clustered and physicochemical property analysis was performed. Inserts on AA proteins were found to be important for type-specificity. Multiple sequence alignment of AA9 sequences allowed for the identification of these type-specific inserts. Two type-specific motifs were identified on AA9 protein sequences which were termed insert I and II. As shown in previous studies (Vu et al. 2014), it has been demonstrated that the absence or presence these inserts is crucial for type-specificity. Due to the diverse nature of AA9 proteins, the inserts may not be sufficient to draw conclusions about AA9 proteins. The N-terminus was found to have the highest degree variability as shown by motif analysis and multiple sequence analysis. The visualization of AA9 cellulose interaction was achieved through manually docking the AA9 domains and crystalline cellulose. This structural analysis, showed that the insert I may interact with cellulose aided by the conserved Tryptophan residue. The insert II of Type 2 sequences did not have as apparent interaction with cellulose due to the absence of aromatic residues on this region. However, the presence of polar residues on this insert is a likely contributor to cellulose interaction. At a 90% site coverage phylogenetic analysis was used to assess the evolutionary nature of AA9 proteins. AA9 proteins have been shown to have a degree of sequence diversity. However, all vs all sequence alignments showed that the AA9 proteins are more conserved within their respective types as opposed to AA9 proteins in general. The N-terminus of AA9 proteins was shown to have motifs which are crucial for distinguishing between AA9 LPMO types. Heat maps were used to identify sequence groups within AA9 LPMO types. The sequence groups were also found to form phylogenetic clusters. These sequence sub-groups were characterized slightly different motif composition of AA9 LPMO types. Similar to sequence alignment, the motif

analysis showed that the N-terminus is important for type-specificity. The presence of these sequence sub-groups may require further study for their potential for AA9-cellulose interaction and the involvement of the motifs identified. The preliminary analysis the AA9-cellulose interface showed that type-specific motifs had the potential to interact with cellulose. All AA9 proteins were subjected to a selection analysis using Selecton and DataMonkey. In all AA9 LPMO types, no positive selection was detected. However, the AA9 protein sequences were found to undergo neutral and negative selection suggesting that the proteins have a conserved and specialized function that is retained through purification selection. This selection is supported by the fact that AA9 proteins are known to have activity on cellulose. Due to this observation, we can speculate that AA9 proteins are not under any selective pressure to adapt to other substrates. Various physicochemical property analysis were evaluated to identify unique chemical features of AA9 types. AA9 proteins were determined to be a fairly stable group of enzymes with a few exceptions. Type 3 AA9 proteins are mostly acid while both Type 1 and 2 AA9 sequences do not have a particular preferred pI distribution. This suggest that fungal organisms have a repertoire of AA9 enzymes that are required in various environmental conditions. The importance of the aromatic residues has always been highlighted for AA9 function. These aromatic residues are important for coordination of the copper atom. These residues have also been proposed to interact with cellulose through stacking interactions. (Leggio, Welner & De Maria 2012). The global distribution of aromatic residues on AA9 and its effect on type specificity is an aspect of AA9 proteins that is not well understood. In AA9 proteins aromatic residues are distributed throughout the structure. The aromatic residues are found to protrude on the surface exposed active site of AA9 proteins. Residues located away from the active site surface are generally buried. This offers no clear difference between AA9 LPMO types. The localization of the protruding aromatic residues may suggest a possible role in substrate interaction (Li et al. 2012). Aromaticities across different AA9 LPMO types was calculated. The results of the analysis showed different compositions of aromaticities for the three AA9 LPMO types.

In this Chapter, it was possible to identify a wide selection of AA9 proteins. The selected AA9 LPMO types were shown to be diverse however unique type-specific features were observed. This was achieved through sequence and structural analysis of the observed features. Manual dockings were then used to investigate the effect of type-specific features on AA9-cellulose interaction. Manual docking was used to identify regions with the potential to interact with cellulose. However,

to validate these findings further validation is needed through the use of molecular dynamics (MD) simulations. MD studies for all three AA9 type were performed in Chapter 4.

Chapter 3

Force Field Parameter Determination

3. Copper (II)(Cu^{2+}) force field parameters

In various biological processes (Adman 1991) Copper (Cu)-bound proteins are crucial (Wise, Coskuner 2014). AA9 proteins are copper coordinating enzymes that interact with cellulose. The copper atom in AA9 protein is believed to result in the formation of a copper-oxyl radical that cleaves cellulose (Gudmundsson et al. 2014). Due to the critical nature of the Cu^{2+} atom in the function of AA9 proteins, it is important to study the interaction between cellulose and Cu^{2+}. There have been numerous studies describing the elucidation coordination geometries of coordination sites (Ando 2010, Bruschi et al. 2012, Hewitt, Rauk 2009, Hodak et al. 2009, Inoue, Shiota & Yoshizawa 2008). However, currently available force field parameters are not expansive enough to adequately describe the active site of AA9 proteins. As a result, in this study a subset of Cu^{2+} coordinating active site residues was created from a Type 1 crystal structure (4B5Q). The subset was then used to perform quantum mechanical (QM) calculations to generate Cu^{2+} parameters. Once evaluated, the Cu^{2+} force field parameters were validated using MD simulations in the subsequent Chapter 4.

3.1. Force field parameter determination

Various approaches have been put forth to aid in the elucidation of missing force field parameters of metal containing proteins for MD simulations (Wise, Coskuner 2014). A popular approach would be to use non-bonded approaches such as the weak van der Waals parameters and electrostatic forces to keep the metal coordinated by the protein residues (Cannistraro 1997). The non-bonded approach is relatively easy to implement into potential functions. However, this method may not be ideal due to the fact that weak van der Waals forces may not be sufficient to keep the metal bound. In cases were the metal does remain bound, nonspecific bonding to various other ligands may occur and the loss of correct coordination geometry may be observed (Wise, Coskuner 2014). Another important consideration is the charge of the system. The metal ion can

be treated with its formal charge or a partial charge in order to predict potential electrostatic interactions. The electrostatic interactions would then relate to the overall energy and structure of the system. Geometric constraints may be applied between the Cu^{2+} atom and the coordinating atoms. With this approach the Cu^{2+} atom will remain in the active site however, the coordination geometry may fluctuate unrealistically even affecting the overall structure of the protein (Wang et al. 2011a, Hodak et al. 2009, Inoue, Shiota & Yoshizawa 2008, Deng et al. 2006). Another approach is the use of dummy atoms on the metal ions. These dummy atoms can be attached to the Cu^{2+} ion and the charge is then distributed evenly throughout the atom (Furlan et al. 2010). To use these dummy atoms, it may be necessary to add additional parameters to the force field to ensure coordination occurs. This approach results in a better coordination geometry around the copper atom. However, nonspecific binding of the Cu^{2+} to other ligands may be observed and the Cu^{2+} atom may be released during the course of a simulation (Wise, Coskuner 2014). The final method that can be used is the bonded approach. This method treats the coordination around the copper center as covalent bonds. The bonded approach is far superior to the non-bonded approach because the Cu^{2+} active center is represented more accurately with respect to geometry. To fully utilize the bonded approach, the force field parameters describing the coordinating atoms around the Cu^{2+} have to be included in the force field for simulation. As such, the accuracy of the conducted MD experiments will depend on how well the force field parameters were elucidated. To obtain these parameters a series of single point energy calculations may be performed on a set of structures with a single variable feature being changed while everything else remains constant (a potential energy surface scan, PES). This feature can be bond stretch, angle bend or dihedral angle parameter (Zhu et al. 2008, Sabolovic et al. 2003, Mentler et al. 2005). These single point energy calculations are termed potential energy surface scans.

A more recent example of determination of coordinated Cu^{2+} force field parameters with respect to the AMBER force field is described in ref (Zhu et al. 2008) where bonded terms (for bonds, angles and dihedrals) were directly determined from the potential energy surfaces for Cu^{2+} coordinated by pyridylmethyl-amine and benimidazolylmethyl-amine ligands. In this study a similar protocol was followed in that a fit for each parameter required for three bonding terms was determined and included, while literature values for the one non-bonding term was included in the CHARMM 36 force field (Guvench et al. 2011), (Equation 1), for the Cu^{2+} containing metalo-protein AA9 active site.

3.1.1. Energy function

The energy function is a crucial component of any force field as it describes the relationship between a structures relative to its energy. A potential energy function is employed molecular dynamics simulations is composed of summations that describe the energy of stretching bonds, angle bending, rotations, non-bonded interactions such as van der Waals, and electrostatic interactions (Wise, Coskuner 2014, Mackerell 2004). In this study the CHARMM potential energy function described in Equation 1 is used (Hornak et al. 2006). The potential energy function by its self is not sufficient as a force field as a result, the potential energy function has to be used in conjunction with the parameters of the energy function. With the presence of these parameters a force field is generated which describes particular system.

$$E_{total} = \sum_{bonds} k_b\left(r - r_{0,b}\right)^2 + \sum_{angles} k_a\left(\theta - \theta_{0,a}\right)^2 + \sum_{dihedrals} k_{d,n}\left[1 + \cos\left(n\chi - \delta_{d,n}\right)\right] + \sum_{non-bonded} \epsilon\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right]$$

Equation 1: CHARMM energy potential

In Equation 1 four summations are illustrated. These summations describe the bond stretch, angle bending, torsional angle and van der Waals parameters which were considered. The first and second summations describe the bond stretch and bending terms, with r, $r_{0,b}$ and k_b, denoting the bond distance, equilibrium bond distance and bond stretch force constant, and $\theta, \theta_{0,a}$ and k_b which denote the bond angle, the equilibrium bond angle and the angle bending force constant. The third summation describes bond dihedral rotation with terms $k_{d,n}$, n and $\delta_{d,n}$ for the dihedral angle, force constant, periodicity and the phase angle, respectively. The last summation is for the non-bonded van der Waals interactions where $r_{ij}$the distance between two non-bonded atoms is. The evaluated terms are shown Figure 3.1.

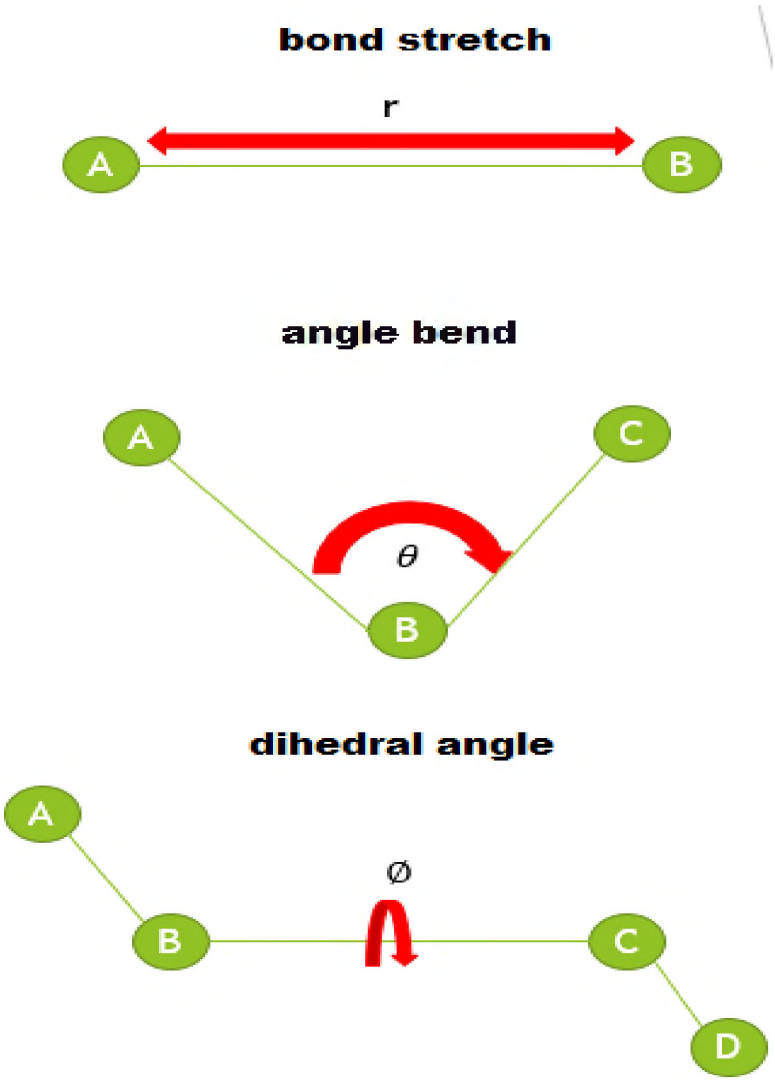

Figure 3.1: Force field parameters to be evaluated. The parameters evaluated were bond stretch angle bend and dihedral angles.

3.1.2. The AA9 active site

Due to the large nature of the AA9 protein it proved difficult to perform the elucidation of force field parameters on the complete structure, although attempts were made to optimize using various levels of theory within an ONIOM calculation (Chung et al. 2015). In the end, a subset of the AA9 Type 1 crystal structure 4B5Q (Wu et al. 2013) which encompasses the Cu-containing active site was used for PES scans. The use of a subset of the full structure was done in order to reduce the

computational cost, while retaining bonds critical to the active site. Residues occurring within 5 Å of the Cu^{2+} atom center were selected using (Script_1.py). Residues not selected were removed, and where this removal resulted in cleavage of amide bonds, the appropriate amine or carboxylic groups were restored. This resulted in an arrangement where the His 1 His 76 and Tyr 160 residues were present. Two water molecules were inserted to complete the octahedral geometry of the AA9 active site (Li et al. 2012), and this structure was optimized at the PM6 level of theory (Stewart 2013) resulting in the structure shown in Figure 3.2. The semi empirical PM6 level of theory has been shown to be applicable in the design of transition metal complexes (Fredin, Allison 2016). As a result, this method was used to describe the AA9 active site.

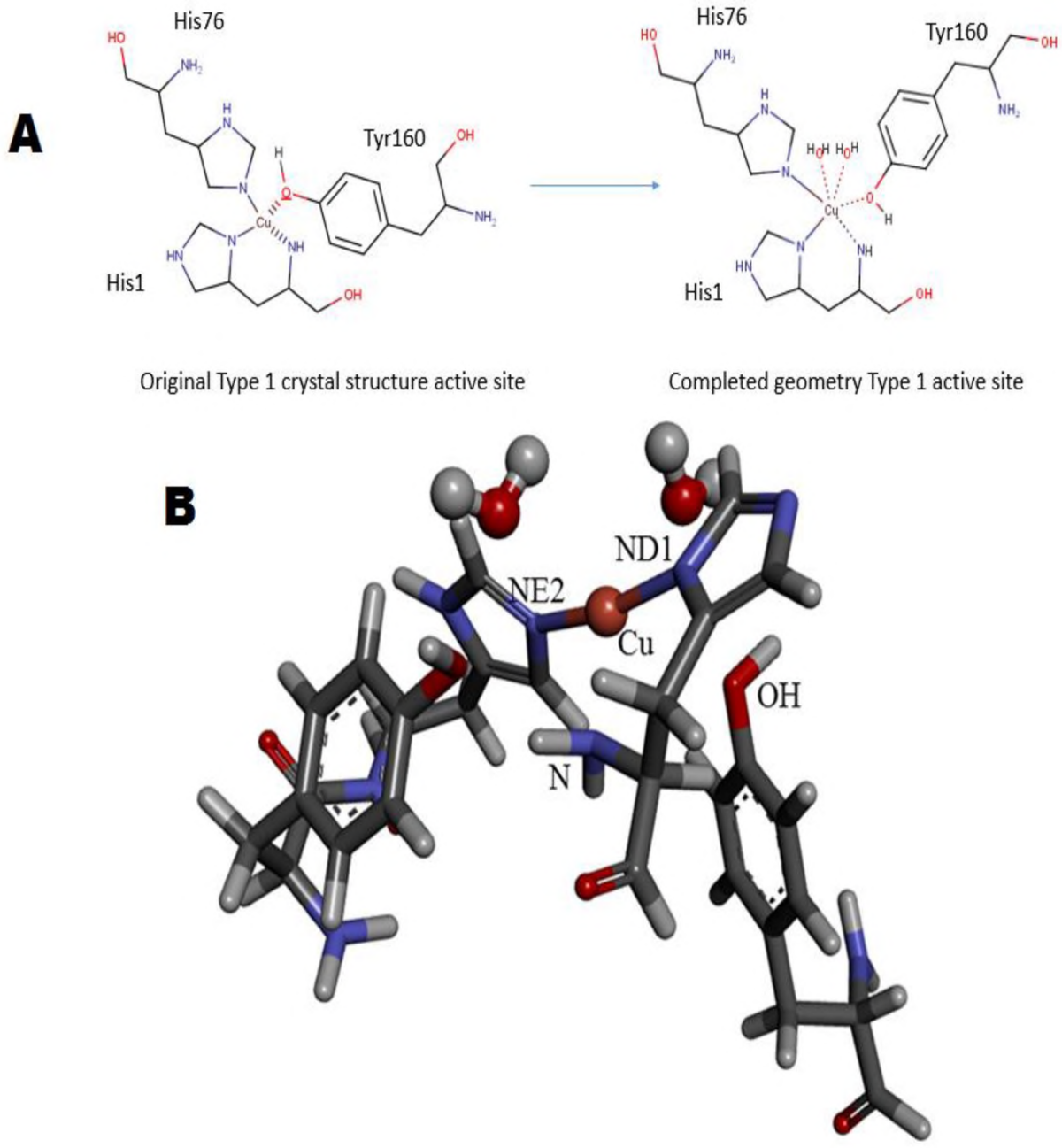

Figure 3.2: Completion of the correct Cu^{2+} geometry by addition of water. A) The addition of water to the subset, B) Final Subset of Type 1 active residues. His-1, His-76 and Tyr-160 were chosen as the subset residues

The geometry of the starting AA9 active site was optimized at the semi-empirical PM6 optimized AA9 active site structure is summarized in Table 3.1. Hessian matrix analyses were employed to unequivocally characterize the conformation thus obtained as a true minima on the potential energy surface. This subset describing the AA9 active site was then used to perform PES scans.

3.1.3. Potential energy surface scans

In order to determine the AA9 force field parameters, PES scans were performed with Gaussian09 (Frisch et al.). A series of potential energy searches in redundant coordinates were performed. The searches were done on coordinates that define the bond stretch, angle bend and torsional movement of atoms that coordinate the Cu^{2+} atom were performed. The PM6 semi-empirical approach was used on this subset to provide this series of potential energy surfaces. Computation of the force field parameters was achieved through least squares fitting of the generated PM6 energy profiles to terms in Equation 1. The force field parameters for the water molecules were not included in subsequent MD simulations. This was done in order to allow for exchange of water and other oxygen species at the active site. Instead of using these during dynamics simulations, a Lennard-Jones term (Table 3.3) was used.

3.1.4. Restrained electrostatic potential charges

To effectively and accurately represent the charge of the Cu^{2+} atom in the MD simulation the electrostatic potential (ESP) was evaluated. There are various approaches that are used to elucidate ESP charges. To calculate the ESP, a grid can be placed around a particular molecule and effective charges are then fitted on to selected sites of that molecule. This is based on the Merz-Kollman (MK) scheme (Singh, Kollman 1984). A second way of obtaining ESP charges uses grid methods (CHELPG) (Breneman, Wiberg 1990) and the restrained electrostatic potential (RESP) (Bayly et al. 1993) methods used in charge fitting. Alternatively the partial form of the electron density can be estimated from an atom within that molecule. The most popular method for charge determination is the Mulliken population analysis (Mulliken 1955). A disadvantage that comes with the Mulliken population analysis is its heavy reliance on the chosen basis set. The findings of the Mulliken population analysis are only meaningful if the basis set used has basis functions that fully encompass a particular atomic site. Mulliken charges generally become large when complete basis sets are used. To overcome this, Natural population analysis (Reed, Weinstock & Weinhold 1985) can be used. In this approach the basis set related problems of the Mulliken population analysis are resolved by taking into account the orthonormal natural atomic orbitals of the atoms in a molecule.

3.2. Methodology

This section details the methodology used (and previously summarized) to elucidated the force field parameters of the Type 1 LPMO AA9 active site of the crystal structure 4B5Q. As previously mentioned, due to the large nature of the AA9 proteins, a subset of residues was selected to perform analysis. The selected subset was then optimized and PES scans were performed. Using the least squares method the newly generated energy profiles were then fitted to the CHARMM energy function to extract the force field parameters. The new force field parameters were then evaluated through MD simulations using Type 1, 2 and 3 crystal structures (4B5Q, 4EIR and 3ZUD respectively). The results of this analysis is detailed in the MD section in Chapter 4.

3.2.1. Constructing the AA9 active site

A subset of the Type 1 AA9 active site was created from the 4B5Q crystal structure. The subset was selected such that that all the important coordinating positions are present and the geometry of the Type 1 AA9 active site was maintained. The two water molecules added to the subset were included to ensure completion of the AA9 active site geometry. The parameters that would be calculated are summarized in Table 3.1 below.

PES scans were then performed using this geometry optimized subset of the AA9 as a starting point. A total of six coordination positions were considered for calculation. Two coordination positions were found present on the His-1 ND1 nitrogen of the imidazole ring and N terminal nitrogen. The His-76 residue coordinated the copper atom with its NE2 nitrogen from imidazole ring. The Tyr-160 residue was found to coordinating the Cu^{2+} atom at the axial position with a relatively long bond length of 3.06 Å indicating Jahn-Teller distortions. Twelve angle and two dihedral angles were considered for calculation. The features that were analyzed are summarized in Table 3.1.

Table 3.1: Initial Type I AA9 x-ray crystal parameters.

Parameter	Crystal structure (Å)
Bonds	
Cu – OH (Tyr)	3.056
Cu – O (Water)	-
Cu – NE1	2.020
Cu – ND1	1.999
Cu – N	2.081
Angles	
ND1-Cu-NE2	174.577
ND1-Cu-N	91.784
Cu-ND1-CE1	127.270
Cu-ND1-CG	125.486
Cu-NE2-CD2	127.956
Cu-NE2-CE1	124.875
Cu-NH2-HT1	103.803
Cu-NH2-HT2	103.586
Cu-OH-HH	93.311
O(water)-Cu-O(water)	-
OH(Tyr)-Cu-O(water)	-
OH(Tyr)-Cu-N	83.372
Dihedral	
Cu-NH2-C-H	0.407
ND1-Cu-NH2-C	-58.437

3.2.2. Force field determination and parameter fitting

PES scans were performed using Gaussian09 on all the features listed in Table 3.1. The scans were done such that each parameter to be evaluated was changed at 50 steps at 0.1 Å increments (in two directions – increase bond length and decrease bond length) for bond stretches. As a result, for the bond stretch parameter a distance of 10 Å was sampled. For angle bend parameters and for dihedral angles a range of 10° was sampled. In cases were PES scans were found to generate energy profiles that did not follow a symmetric polynomial shape, the profiles were trimmed to extract the regions in the profile that could be fitted using polynomial functions. Once the energy profiles were

generated, a least squares fitting was applied to the energy profiles with the CHARMM energy profile and the force field parameters were generated.

3.2.3. RESP Charge evaluation

Partial atomic charges of the AA9 active site was calculated using the RESP (restrained electrostatic potential) charge-fitting method (Cieplak et al. 1995). RESP charges were determined at the B3LYP level of theory (Becke 1993, Lee, Yang & Parr 1988, Vosko, Wilk & Nusair 1980) using 6-31G(d) (Ditchfield, Hehre & Pople 1971, Hehre, Ditchfield & Pople 1972, Hariharan, Pople 1973, Hariharan, Pople 1974, Gordon 1980, Rassolov et al. 2001) basis for all non-metal atoms and the LANL2DZ (Dunning, Hay 1977) basis (with pseudopotential) for the Cu^{2+} center. The lowest energy conformation were submitted to this single-point ab initio calculations and the Löwdin (Mayhall, Raghavachari & Hratchian 2010), Mulliken, and electrostatic potential (ESP) derived charges were obtained.

3.3. Results

Force field parameters were calculated for the AA9 active site and the generated energy profiles for the bond stretch, angle bend and torsions are shown in Figure 3.3, 3.4 and 3.5 respectively. The fitted force field parameters to be used in the MD simulation were extracted, and are summarized in Table 3.4.

3.3.1. Type 1 AA9 Force field parameters

The AA9 force field parameters were determined for the Type 1 active site of the 4B5Q crystal structure were determined at the PM6 level of theory. The parameters evaluated were for the bond stretch, angle bend and dihedrals. To determine correct Lennard-Jones parameter for the Cu^{2+}, QM calculations and literature searches were performed.

3.3.1.1. Bond stretch parameters

The parameters of the bond stretch of the Cu^{2+} coordinating positions are represented in Figure 3.2. Out of the six coordination position only five were considered for the analysis for the proteins. These were His-1 imidazole nitrogen (N7-Cu), His-1 N-terminal nitrogen (N1-Cu), His-76 imidazole nitrogen (N49-Cu), Tyr-160 OH (O86-Cu) and an oxygen from one of the water molecules. Even though two water molecules were added to complete the geometry of the active site, only one calculation was performed because these parameters were anticipated identical.

A good least squares fit of force field parameters to the PM6 data was observed for the six calculated positions. For the Cu – OH (Tyr), Cu – O (Water), Cu – NE1, Cu – ND1 and Cu – N (His) coordinating positions, the corresponding force constants were 3.0 kcal mol^{-1} $Å^{-2}$, 61.3 kcal mol^{-1} $Å^{-2}$, 167.1 kcal mol^{-1} $Å^{-2}$, 250.9 kcal mol^{-1} $Å^{-2}$ and 94.8 kcal mol^{-1} $Å^{-2}$.

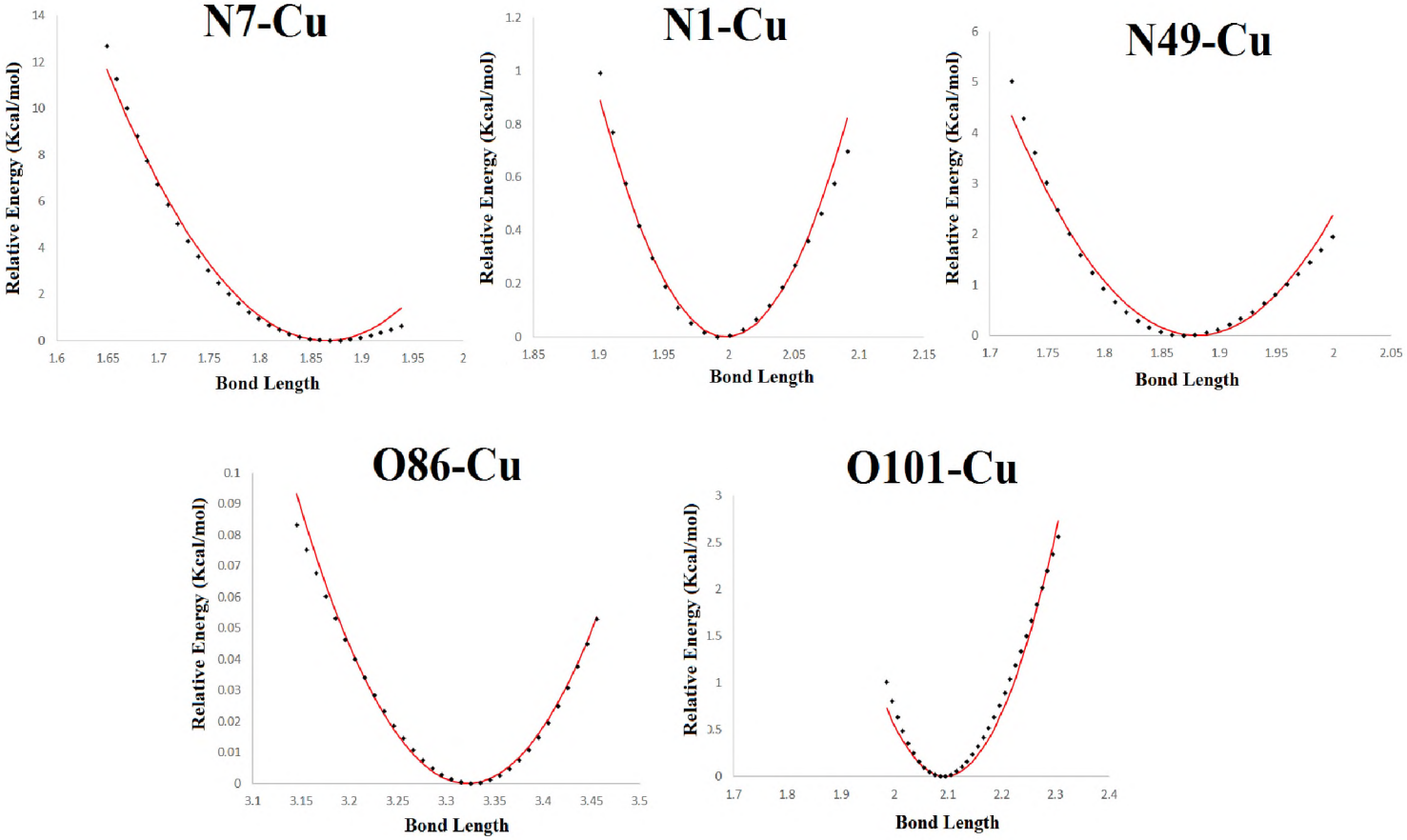

Figure 3.3: Energy profiles for the generated bond parameters. The MM fitting curve is shown with the red line and the bond stretch curves are shown with the black dots.

A weak force constant was found for the Tyr-160 OH to Cu^{2+} interaction indicating that this coordinating position is loosely bound, and this may translate to a bond length with a large degree of fluctuation at temperature. Associated with the weak binding is its relatively long equilibrium bond length of 3.322 Å. The remaining equilibrium distances were 2.095 Å, 1.880 Å, 1.865 Å and 1.998 Å. These values were found to be similar to those reported in literature for Cu^{2+} centers (Zhu et al. 2008). These values also match the crystal structure values as indicated in Table 3.1 and 3.2.

3.3.1.2. Angle bend parameters

A total of 11 angle bend parameters were chosen for modelling the AA9 active site (Table 3.1). These angles were selected to ensure that the correct octahedral geometry of the AA9 active is properly define and will be accurately described in subsequent MD simulations. The findings of the angle bend parameters are displayed in Figure 3.4.

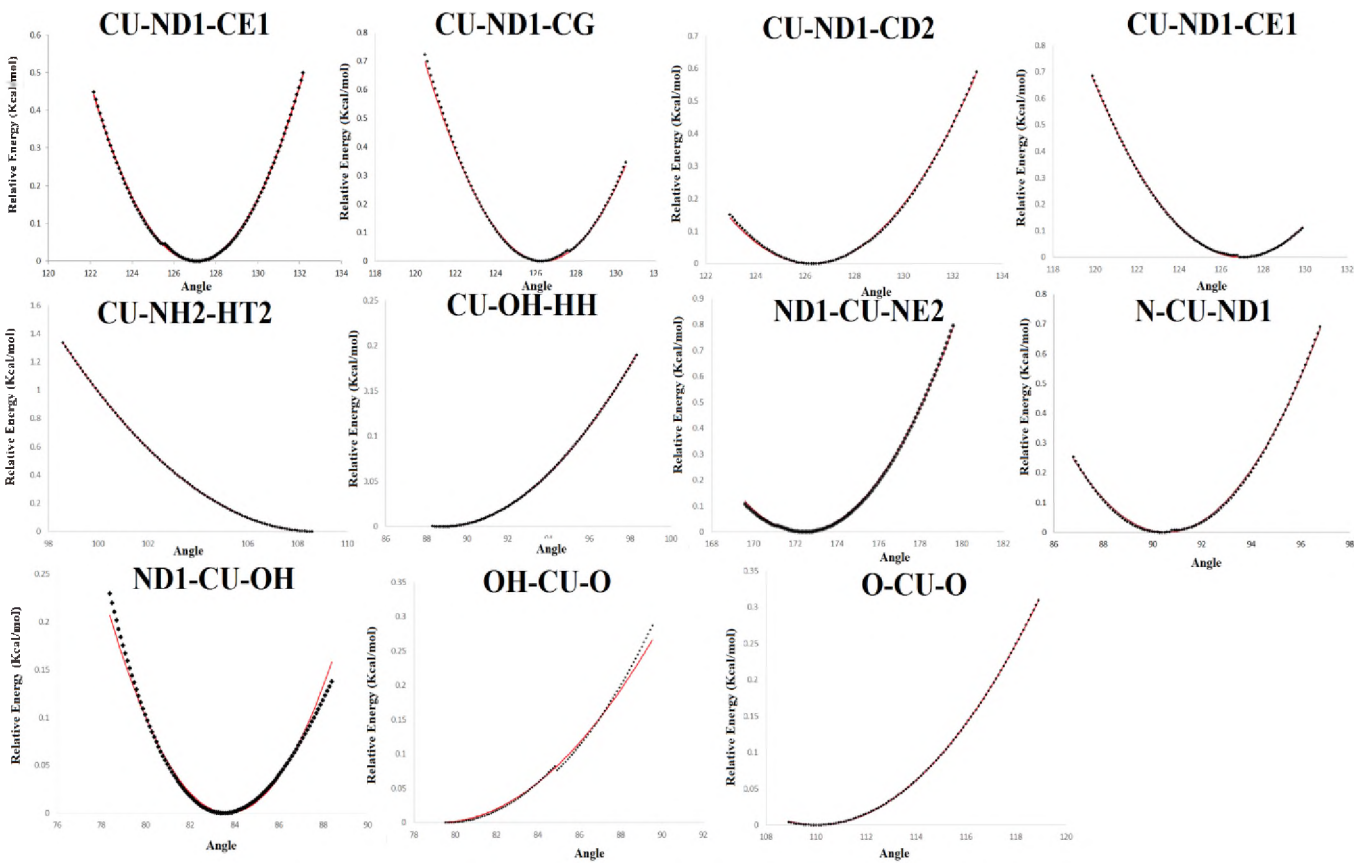

Figure 3.4: Energy profiles for the generated angle parameters. The MM fitting curve is shown with the red line and the angle bend curves are shown with the black dots.

These values are similar to the crystal structure values as shown in Table 3.2. The calculated angle bend parameters are shown in Figure 3.4 and displayed in Table 3.2, while the angles considered for calculation are also presented in Table 3.2. All angles of interest were defined by three atom centers, of which one is Cu^{2+}. The corresponding angle force constants (k_a) were found to be reasonable – of magnitudes corresponding to other similar centers in the CHARMM36 force field.

3.3.1.3. Rotations (Dihedral / Torsions)

Two dihedral angles were considered for the Type 1 AA9 active site. The results of the potential energy surface scans are shown in Figure 3.5. These two dihedral angles were chosen to ensure the planar arrangement of the His 1 aromatic imidazole relative to the copper atom and cellulose.

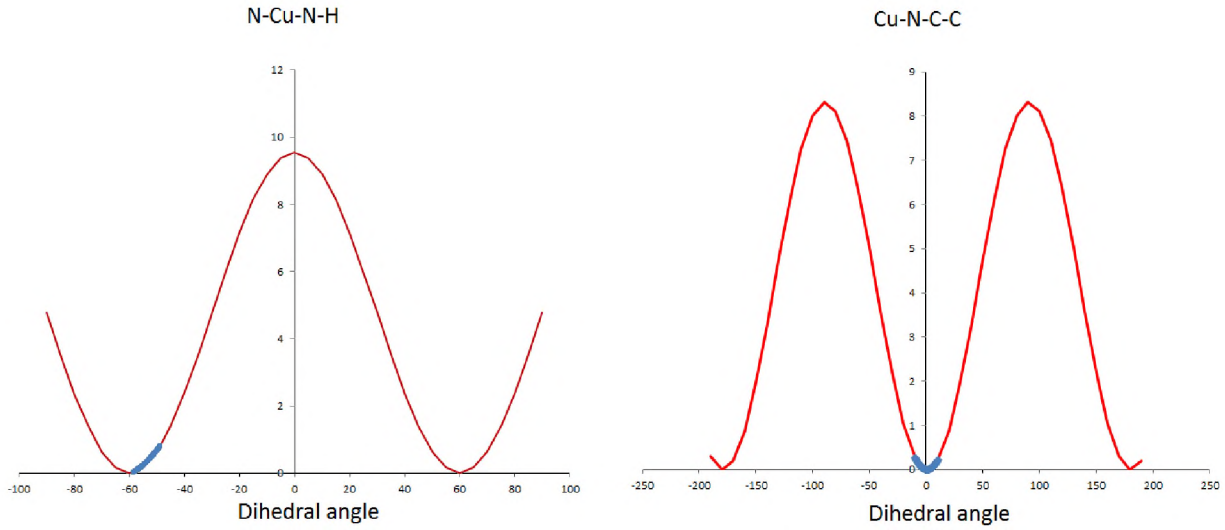

Figure 3.5: Energy profiles for the generated torsion parameters. The MM fitting curve is shown with the red line and the torsions curves are shown with the blue dots.

The following two dihedral angles were considered Cu-ND-C-C and NE-Cu-N-H, to account for the correct geometry of the histidine-copper interaction. The least squares fit to this PES data is shown in Figure 3.5. As seen in Figure 3.5, the range sampled by the PES scans was not large. This was done purposely to avoid energy fluctuations that result from disruption of the active site subset, by excessive rotation of the dihedral angle. All calculated parameters are summarized in Table 3.2.

3.3.2. Results summary

The force field parameters for the Type 1 active site of AA9 proteins were successfully evaluated at the semi empirical PM6 level of theory. The elucidated force field parameters are summarized in Table 3.2.

Table 3.2. Type 1 Force field parameters.

Parameter			
Bonds	$K_r(kcal(mol^2 A)^{-1})$	$r_{eq}(A)$	
Cu – O (Tyr)	3.000	3.322	
Cu – O (Water)	61.341	2.095	
Cu – NE1	167.147	1.880	
Cu – ND1	250.940	1.865	
Cu – N	94.784	1.998	
Angles	$K_\theta(kcal(mol\ radian^2 A^{-1})$	θ_{eq}**(degrees)**	
ND1-Cu-NE2	49.688	172.371	
ND1-Cu-N	57.557	90.536	
Cu-ND1-CE1	61.428	127.025	
Cu-ND1-CG	65.494	126.400	
Cu-NE2-CD2	42.676	126.250	
Cu-NE2-CE1	44.358	126.967	
Cu-NH2-HT1	-	-	
Cu-NH2-HT2	42.339	108.760	
Cu-OH-HH	6.700	88.643	
O(water)-Cu-O(water)	12.958	110.053	
O(Tyr)-Cu-O(water)	8.084	79.133	
O(Tyr)-Cu-O(NE)	23.830	83.711	
Dihedral	$V_n(kcal\ mol^{-1}$	**n**	**Γ**
Cu-ND-C-C	8.312	2	182
NE-Cu-N-H	9.536	3	0

The calculated angle bend parameters are shown in Figure 3.3, while the angles considered for calculation are presented in Table 3.2. All angles of interest were defined by three atom centers, of which one is Cu^{2+}. The force constants corresponding with angles (k_a) were found to be acceptable when compared to those of the CHARMM force field. The following two dihedral angles were considered Cu-ND-C-C and NE-Cu-N-H, to account for the correct geometry of the histidine-copper interaction. The least squares fit to this PES data is shown in Figure 3.4. All

calculated parameters are summarized in Table 3.2. The elucidated parameters where then validated by MD simulations however prior to this validation Lennard-Jones parameters for the Cu^{2+} had to be defined. The following section details the steps that were taken in selected the appropriate Lennard-Jones parameters for the MD simulation.

3.3.3. Lennard-Jones parameters for Cu

The prediction of bulk properties of molecules has been studied extensively in various mediums. In gases, molecular properties such as viscosity can be evaluated using information based on molecular pair interactions. To model the pair interaction, a potential energy function is required to model for each of the molecules be considered. One, two or more parameters can be used to distinguish between each molecule. A technique that that can n be used to describe this pair interaction is the Lennard-Jones function shown in Equation 2.

$$V(LJ) = \varepsilon_{CuO}\left[\left(\frac{Rmin}{r}\right)^{12} - 2\left(\frac{Rmin}{r};\right)^{6}\right]$$

Equation 2: Lennard-Jones

$$E_r = K(r - r_0)^2$$

Equation 3: Bonded component

$$\varepsilon_{CuO} = \sqrt{\varepsilon_{Cu} \times \varepsilon_O}$$

Equation 4: Energy well

$$Rmin = Rmin_{Cu} + Rmin_o$$

Equation 5: Van der Waals radius

Equation 2 describes the Lennard-Jones potential which is a mathematical approximation of the interaction between two neutral atoms. The Lennard-Jones parameters for Cu^{2+} were determined from the interaction between the Cu^{2+} and the water oxygen in the simplified AA9 active site. Water copper bond interaction parameters were evaluated to be 61.34 kcal/mol with a bond length of 2.0916 Å. The parameters were evaluated at the B3LYP/Mixed level, with 6-31G(d) basis for all atoms except copper, and the Los Alamos LANL2DZ basis set (with pseudopotential) for

copper. The water Lennard-Jones parameters were obtained from the CHARMM36 force field and used for fitting (Figure 3.6).

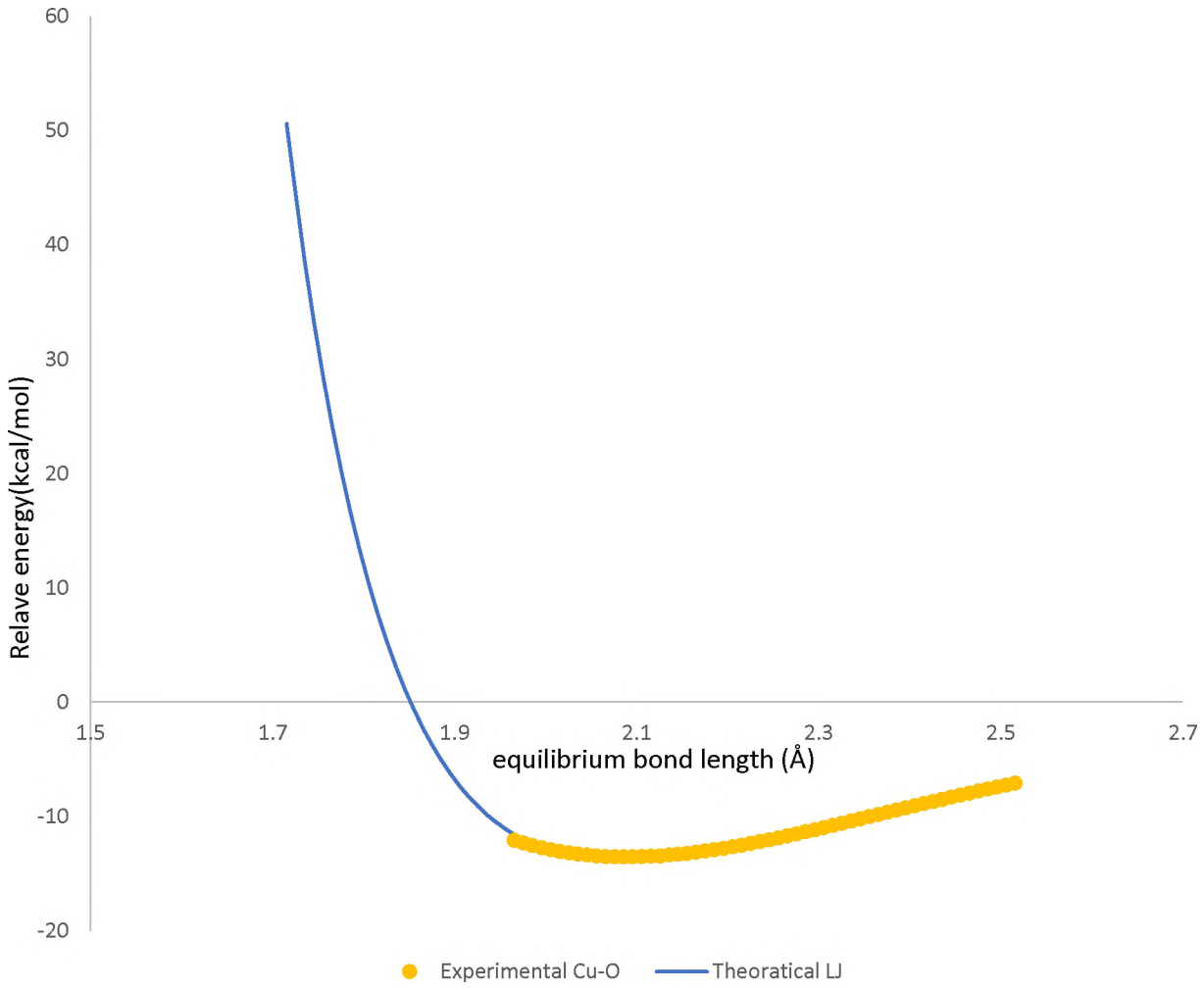

Figure 3.6: Data fitting analysis of experimental DFT Cu-O calculation and theoretical Lennard-Jones parameters for the Cu-O bond.

There was concern that the Lennard-Jones parameters for Cu^{2+} that we obtained from literature would bias the results of the study. The Literature Cu^{2+} Lennard-Jones parameters seemed to possess strong Van der Waals interactions with oxygen atoms of water, cellulose protein itself. The performed ab-initio results for the interaction between these elements suggests an even

stronger interaction to occur (Table 3.3). In order to mediate this interaction the more moderate parameters from literature were used.

Table 3.3: Lennard-Jones parameter estimation for Cu^{2+} based on Cu-O bond.

Rmin O	1.7682	**εO**	0.1521
Rmin Cu	0.310302	**εCu**	1245.922
Rmin	2.078502	E	13.76607

As such, the literature values for the Cu^{2+} Lennard-Jones parameters were used to simulate as shown in Table 3.4. The presence of two literature Lennard-Jones parameters permitted the design of two MD experiments: which were termed the biased and unbiased experiments. The biased dynamics would allow for the rapid interaction between the Cu^{2+} and oxygen atoms of the cellulose crystal. While the unbiased experiment with a relatedly low epsilon (ε) for Cu^{2+} atom, would provide a less interaction between Cu^{2+} and cellulose. As a result, simulations were performed on both Lennard-Jones parameter sets. The parameter sets used for MD simulations are shown in Table 3.4.

Table 3.4: Lennard-Jones (LJ) parameters for Cu^{2+} obtained from literature

Lennard-Jones set	Rmin	εCu	Reference
1(Biased)	0.413	410	(Torras, Aleman 2013)
2(Unbiased)	1.476	0.03198620	(Li, Merz 2014)

3.3.4. Determining how the Cu^{2+} charge is handled in the system

Evaluating the realistic charge on the AA9 active during the MD important is important in ensuring an accurate AA9-cellulose interaction. As a result, the ESP charges were determined at the B3LYP level of theory using 6-31G(d) basis for all non-metal atoms and the LANL2DZ basis (with pseudopotential) for the Cu^{2+} center. This analysis was performed in order to determine the realistic charge on the Cu^{2+} center. The findings of this analysis is shown in Figure 3.6 below.

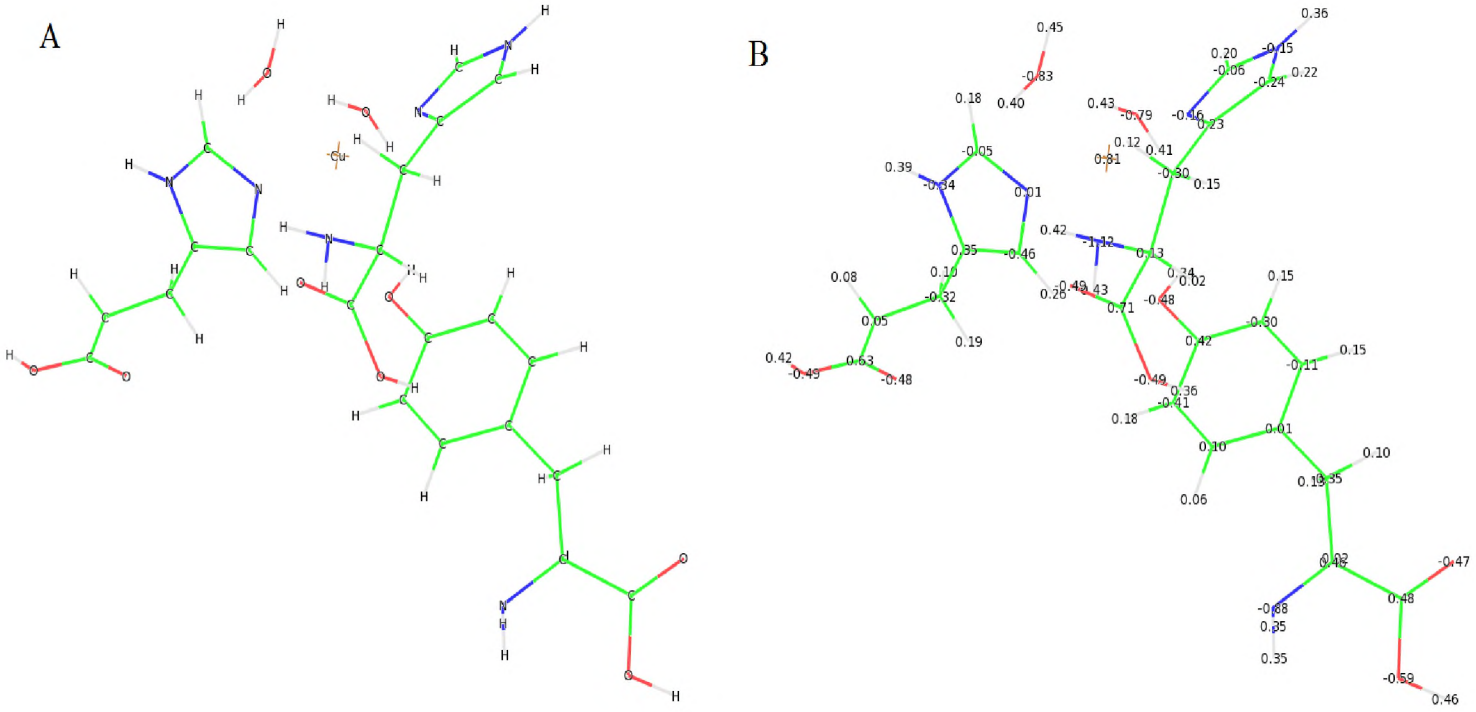

Figure 3.7: Atoms and RESP for the Cu^{2+} AA9 active site. Only residues His-1, His-76 and Tyr-160 are shown for visual clarity.

It was observed that from the RESP charge analysis (Figure 3.7) that the realistic charge of the Cu^{2+} atom is +0.81 as opposed to +2 total charge. To generate Figure 3.7, Script2.py was used to extract the charges generated by Gaussian09 and the charges were the mapped on to the B factor column of the PDB file. The resulting PDB file is then visualised in PyMol to show the charges.

3.3.5. Discussion

The semi-empirical PM6 approach and least squared fitting was used to determine the force field parameters of a subset of the Type 1 Cu^{2+} AA9 active site. The parameters that were considered for calculation were the bond stretch, angle bend and torsions/rotational bonds. Since the AA9 protein is relatively large, the elucidation of force field parameters was performed on a subset of active site residues from the 4B5Q crystal structure. The force field parameters were successfully evaluated and were validated using MD simulations (Chapter 4). The force field parameters were evaluated on a Type 1 AA9 active site subset, however due to the similarity of the AA9 active site across other AA9 LPMO types, the force field parameters may be readily applied to other AA9 types. The validation of the force field parameters will be discussed further in Chapter 4. MD simulations showed that the newly evaluated parameters were able to maintain the correct geometry of the AA9 active site for all three types during the course of simulation. In addition binding to cellulose was observed for the Type 1 AA9 MD experiments. The binding of the Cu^{2+} to cellulose is hypothesized to be the first step in the AA9-cellulose reaction mechanism. Future studies may involve Quantum Mechanical studies that take into account the reactive dioxygen species and observe what affect this might have on cellulose cleavage. These studies may involve placing the dioxygen species into proximity with the Cu^{2+} and the nascent glycosidic bond to observe the cleavage of glyosidic bonds through oxidation, similar to the studies previously performed on cellulose cleavage by AA9 proteins (Kim et al. 2014). Since the main goal of this study was to investigate the role of type-specific features in AA9-cellulose interaction, adding a reactive oxygen species to the QM studies or the MD simulation studies was beyond the scope of this study.

Chapter 4

Force Field Parameter validation

4. Comparative MD analysis of AA9 types

4.1. MD simulations

In order to fully understand how biological molecules functions it is important to understand their structure as well as dynamics (Karplus, Kuriyan 2005). To predict the dynamic features of proteins it is essential to use a correct model to describe the system of interest. The model used consists of the information about the structure and intermolecular interactions of the system. To test the accuracy of the model used, the results of the simulation are then compared to experimental data. As a result, computer simulations are a good way for testing theory. In cases where experimental data is not available to describe a system, computational approaches can be used to predict the properties of the system. Molecular dynamics (MD) is a combination of computer simulation and statistical mechanics with the aim of computing the equilibrium and movement of a many-body system (McCallum 1999). The equilibrium properties are energy, temperature and pressure. The movement of a system can involve the diffusion coefficient, shear viscosity and thermal conductivity. MD simulations are broken down into three steps. The first step involves specification of the initial positions and momenta of the particles in the system. To describe the interaction of the particles in the system with MD, a potential is required. The potential used to describe the system will determine how accurate the results obtained from the MD simulation will be. The second step of MD simulation involves evolving the system using Newton's second law of motion as shown in Equation 7.

$$f=ma$$

Equation 7: Newton's second law of motion

Where f is the force, m is the mass and a is the acceleration. In the MD simulation the particles move around in 3D space in the simulation in specific trajectories (Meller 2001). The third step of MD simulations involves the measurement of physical quantities as mathematical functions

describing the positions of the particles and their momenta. Then to interpret the measurements with respect to equilibrium properties, statistical mechanics is used. MD simulations can be used to study various features about various systems. In this study MD simulations are used to validate the Type 1 AA9 force field parameters determined in Chapter 3 as well as to perform a type-specific MD characterization of AA9 proteins. To achieve this the Type 1 force field parameters were inserted to the CHARMM 36 force field for both proteins and carbohydrates (Guvench et al. 2011). All simulations were performed with the CHARMM software package (Brooks et al. 2009).

4.2. Force fields

In molecular mechanics, molecular systems are described as group of nuclei that obey the classical, Newtonian laws of motion. The drawback of this approach is that it neglects the effects of electron movement of the system being analyzed. However, the negligence of the electron motions allows for a faster calculations of the force and energies active within the system. The potential energy of the starting state of the system i.e. the initial coordinates, is calculated using a force field equation. A force is then calculated to propagate the system to different coordinates. The force is derived from the energy of the initial coordinates. The derived force from the energy is generated resulting in the displacement of the coordinates (Mackerell 2004). For this study the CHARMM force field is used as illustrated in Equation 1 in Chapter 3. Generally force fields contain sets of parameters such that describe the force constants and their respective equilibrium bond lengths and angles, and dihedral angles and charges. The parameters are generated by experiment or quantum chemical calculations for specific atom sets. Force field parameterization is the process of generating force field parameters. This process involves selecting the group of atoms to be paramatized. Once selected the atom types are assigned depending on hybridization and bond types. Using ab initio methods, the parameters are calculated. A penalty function can be used to measure the difference between the force field and their respective reference values.

4.2.1. Protein force fields

When simulating biological macromolecules using MD, it is common to either use a united force field or an all atom force field (McCammon, Gelin & Karplus 1977). It is common for force fields to use both the united-atom and all-atom force fields, however in many cases the force field used for protein MD simulation studies, is the all-atom protein models. Examples of all-atom force

fields are the OPLS/AA (Jorgensen, Tirado-Rives 1988), the CHARMM FF (MacKerell et al. 1998), and AMBER (Cornell et al. 1996) force fields. In all three of these force fields the parameters were optimized taking into account how they treat proteins. HF/6-31G* supramolecular data is used to calculate partial atomic charges for both the OPLS and CHARMM FF. For the standard AMBER force field (parm99) the charges are based on the restrained electrostatic potential (RESP) charges which are computed using HF/6-31G*. A set of model compounds is utilized to evaluate the Lennard-Jones parameters using condensed phase simulations for all three force fields. The OPLS and CHARMM FF force fields were optimized using the TIP3P water model. The standard AMBER force field was optimized for the TIP3P, TIP4P, and SPC models. Due to the similar water dimer interaction energy for both TIP4P and SPC models, both these models can be used with the CHARMM FF and AMBER force fields. It has been observed that for the OPLS/AA, CHARMM FF and AMBER force fields, the intermolecular interactions are modelled well. However, differences in charge distributions can be observed locally (Ponder, Case 2003). The differences observed in charge distributions may possibly affect the local interactions such as the affecting the peptide bond length (Chen, Yin & MacKerell 2002). The intramolecular parameters for CHARMM FF and AMBER force fields are obtained from Quantum Mechanical calculations using small model compounds while the OPLS,/AA Force field is taken from the AMBER PARM94 (Cornell et al. 1996).

4.2.2. Carbohydrate force fields

The study of carbohydrates with MD simulations presents challenges for current empirical force fields (Mackerell 2004). The structures of carbohydrates such as monosaccharides are abundant with intermolecular hydrogen bonding that occurs between hydroxyl groups of the carbohydrate and water in the environment. On top of this, there are also different types of monosaccharides with different functions. Small molecule geometric and vibrational data and on -D-glucopyranose have been used to model hexo pyronose sugars (Ha et al. 1988). This model was found to generate incorrect conformations the exocyclic hydroxyl (Kouwijzer et al. 1993). The issue is that the model does not treat glycosidic bonds and only takes into account hydroxyl substituents. To improve the conformational properties of the exocyclic hydroxyl, the model has been re-optimized to make it consistent with the CHARMM22 and the CHARMM27 nucleic acid/lipid force fields (Kuttel, Brady & Naidoo 2002). There are also other CHARMM-compatible force fields. One such

example is the one developed from multiple QM calculations that were done on carbohydrate analogues (Reiling, Schlenkrich & Brickmann 1996). The D-glucose substituents use sulfates and sulfamates (Huige, Altona 1995). The GROMOS force field has been popularly used carbohydrate simulations (Lins, Hünenberger 2005, Scott et al. 1999).

The GROMOS force field is a united atom model that has been developed in aqueous environments. There have been variations on this force field that incorporate corrections that more accurately model the exo-anomeric effect (Ott, Meyer 1996). There is also an all atom alternative which provides missing parameters from CHARMM and also adjusts the Lennard-Jones parameters according to condensed phase simulations (Kouwijzer et al. 1993).

4.2.3. Solvation

The aqueous environment of the system to be analyzed is an important aspect to be considered to ensure accurate representation of the system in MD simulations. This may require the use of explicit or implicit water models. The explicit water models are regarded as being a more microscopically complete method. However, though less complete, the implicit water models have the advantage of less computational cost associated with them and direct yield of free energies of solvation (Mäkelä, Donofrio & de Vries 2014)

. The explicit water models commonly used to study biomolecular systems are TIP3P, TIP4P (Jorgensen et al. 1983) SPC, extended SPC/E (Berendsen, Grigera & Straatsma 1987), and F3C (Levitt et al. 1997) models. The specified models have been shown to perform similarly to bulk water at ambient temperatures. The most commonly used water model is TIP3P. The TIP3P model does have limitations which include a diffusion constant that is larger than what is observed experimentally and the overestimation of the height of the tetrahedral peak in the OOO radial distribution (Feller et al. 1996). However, the TIP3P water model does provide advantages that include satisfactory treatment of the energetics because most of the interactions between water and biomolecules involve either first or second shell hydration. The absence of the long-range structure is found not to cause problems. Similar to the TIP3P models is the SPC models using a tetrahedral geometry result in an increased structure. This increased structure is shown by the tetrahedral peak of the OOO radial distribution function (Mackerell 2004).

4.2.4. Boundary Conditions

In order to study the properties of a bulk system, MD simulations are used. However running very large systems results in a large computational cost. To overcome this, a smaller simulation box may be used. The use of a small simulation box would result in most of the molecules being in the edge of the box. As a result periodic boundary conditions (PBC) are required to handle this problem. PBC is employed in MD simulations to remove the surface of a simulation box thus removing the need for a large box. PBCs allow for an infinite lattice to be created by repeating the simulation box in space. This means that if a molecule were to leave the simulation box, it will enter in the same manner on the opposite face of the box in the same direction. For this study, AA9 proteins wee simulated in a cubic box (Tyrus, Gosz & DeSantiago 2007, Nguyen et al. 2012).

4.2.5. Trajectory analysis

The resulting trajectories were analyzed with respect to the Root Mean Square Deviation (RMSD, radius of gyration (Rg), residue wise Root Mean Square Fluctuation (RMSF) in this study. The RMSD was measured to assess the stability of the protein throughout the course of the simulation. The radius of gyration was monitored to observe any extreme changes in the protein during the course of the experiment. The RMSF was measured to assess which regions on AA9 protein structures contribute the most to protein movement as well as the general stability of the three AA9 types. Additionally the overall movement of AA9 proteins was observed and monitored with VMD (Humphrey, Dalke & Schulten 1996). The regions responsible for the overall interaction between AA9 protein structures and cellulose was analyzed using contact maps. This was achieved by performing pairwise distance calculations between the protein and the cellulose.

4.2.5.1. Root Mean Square Deviation (RMSD):

To assess the stability of the protein the simulation the RMSD is measured. RMSD is measured by assessing the deviation of a structure from a specific conformation. The RMSD of protein during an MD simulation is described in the following equation:

$$RMSD = \left(\frac{\sum_{N} (R_i - R_i^0)^2}{N} \right)^{1/2}$$

Equation 8: RMSD

N represents the total number of residues that are considered within the calculation. R_i represents the vector position of particle i in the selected frame. R_i^0 represents the vector position of particle i in the reference conformation. This is calculated after alignment of the selected frame to the reference structure. The alignment is performed using least square fitting. All RMSD calculations were performed on the Cα atoms of the backbone of AA9 proteins. The first frame of the MD simulations were use as the reference structure (Kufareva, Abagyan 2012, Carugo, Pongor 2001, Stone et al. 1995).

4.2.5.2. Radius of Gyration

The radius of gyration was calculated to measure the compactness of the protein throughout the course of the simulation. The formula used to compute radius of gyration is shown in shown Equation below:

$$r^2 gyr = \frac{\left(\sum_{i=1}^{n} w_i \left(r_i - \bar{r} \right)^2 \right)}{\sum_{i=1}^{n} w_i}$$

Equation 9: Radius of gyration

The r_i describes the position of the ith atom. The r represents the center which is weighted. The weight is assigned to atom i. The radius of gyration as computed for all three AA9 LPMO types to assess the contribution of the flat surface active site loop regions to the overall compactness of AA9 protein types (Stone et al. 1995, Lobanov, Bogatyreva & Galzitskaya 2008, Hong, Lei 2009).

4.2.5.3. Root Mean Square Fluctuation

The Root Mean Square Fluctuation (RMSF) is used to describe the observed local fluctuation on a structures. The equation below describes the calculation of RMSF:

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_i(t) - r_i^{ref})^2}$$

Equation 10: RMSF

T denotes the time of the simulation in which the average is extracted and r_i is the selected frame and the r_i^{ref} represents the reference position of particle i. the RMSF was computed observe the contribution of the AA9 loop regions to the overall movement of the AA9 protein structures on cellulose (Kufareva, Abagyan 2012, Kuzmanic, Zagrovic 2010).

4.3. Methodology

4.3.1. Structure preparation

The Discovery Studio Visualizer (Dassault Systèmes BIOVIA 2016) was used to construct the native model of the β-cellulose crystal. The cellulose crystal was constructed from coordinates and crystal parameters obtained from literature (Nishiyama, Langan & Chanzy 2002). The cellulose crystal was constructed such that it was a three layered beta cellulose crystal which consisted of 168 glucose residues. The cellulose crystal was composed of 14 chains with 5 chains on the top layer, 4 chains on the middle layer and 5 chains in the bottom layer. The staring orientation of the AA9-cellulose complex was created such that the planar aromatic residues (as described in previous studies (Li et al. 2012)) of the AA9 4B5Q crystal structure were aligned to the pyronose residue of the top layer of the β-cellulose crystal. The orientation of the AA9-cetllulse complex was used due to the observed binding of the small CBMI to cellulose using MD simulations (Harris et al. 2010). During the MD simulations, the small CBMI domains were found to orient to the cellulose substrate glucose using their planar aromatic residues (Harris et al. 2010). To approximate an infinite cellulose structure for MD simulations, the bottom layer of the three layered cellulose crystal was restrained using the C1 and C4 carbons. The other two layers were allowed to freely move during MD simulations.

4.3.2. Modified residues

As powerful as MD simulations can be, they are limited by the coverage of the force field used in the simulation (Mackerell 2004). In some cases, such as in AA9 proteins, the force field being used may not be adequate enough to completely describe the protein structures. In the case of AA9 proteins, there are three considerations which had to be addressed prior to simulating AA9 proteins with cellulose. The first being the absence of AA9 copper force field parameters. The second consideration is the presence of modified histidine residues in the active sites of AA9 proteins. The AA9 active His-1 residue has a methylation (Hemsworth et al. 2014, Bennati-Granier et al. 2015). This methylation is only present on Type 2 and 3 AA9 crystal structures (4EIR and 3ZUD respectively) (Figure 4.1). The third consideration was determining the correct Lennard-Jones parameters for the Cu^{2+} atom. Since the Lennard-Jones parameters describe the non-bonded interaction of the AA9 active site, the choice of a suitable parameter set is of importance.

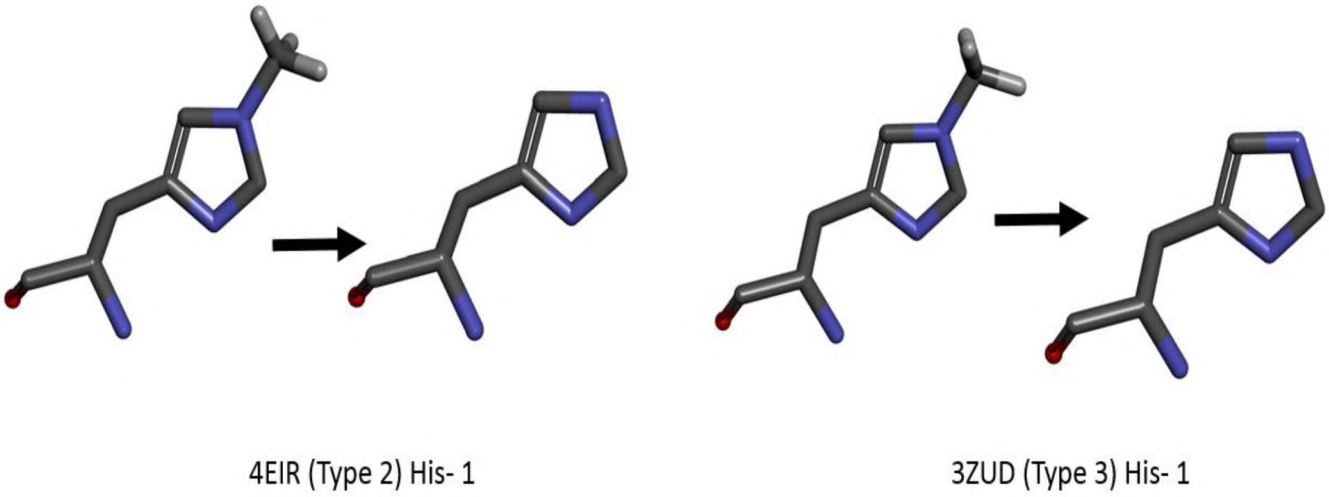

Figure 4.1. Methylation of terminal His 1 residue in crystal structures 4EIR and 3ZUD. The Type 1 crystal structure used in this study does not have this modification on the His 1 residue. The biological function of the methylation on the His 1 residue is currently unknown. Studies have found that, in lab experiments, the removal of the methylation in AA9 proteins not affect activity. As a result for both the 4EIR and 3ZUD crystal structure was removed from the His 1 residues prior to simulations.

4.3.2.1. Disulphide linkages

Submission of all three AA9 crystal structures (4B5Q, 4EIR and 3ZUD) to the PDBsum webserver (Laskowski 2001) revealed the presence of disulphide linkages. The disulphide bonds are usually a component of native proteins. Disulphide linkages are crucial for the stability and the function of proteins. The disulphide bonds have also been found to be important for maintaining the structure of the protein. Unlike the other bonds on protein structures, the cleavage of disulphide linkages may result in extreme conformational changes (Chinchio et al. 2007). The disulphide linkages found on AA9 structures are listed in Table 4.1 and are graphically represented in Figure 4.2.

For the Type 1 crystal structure the only 1 disulphide linkage was observed. The linkage occurred between residues Cys-43 and Cys-163. The Type 2 crystal structure had two disulphide linkages which were found in residues Cys-39 and Cys-171 for the first disulphide linkage and Cys-141 and Cys-223.

Table 4.1. Disulphide linkages found in AA9 structures

4B5Q(Type 1)	
Cys-43	Cys-163
4EIR (Type 2)	
Cys-39	Cys-171
Cys-141	Cys-223
3ZUD (Type 3)	
Cys-56	Cys-101
Cys-98	Cys-178

The Type 3 crystal structure was found to have two disulphide linkages. These linkages were found to occur in Cys-56 and Cys-101 for the first linkage and Cys-98 and Cys-178 for the second linkage.

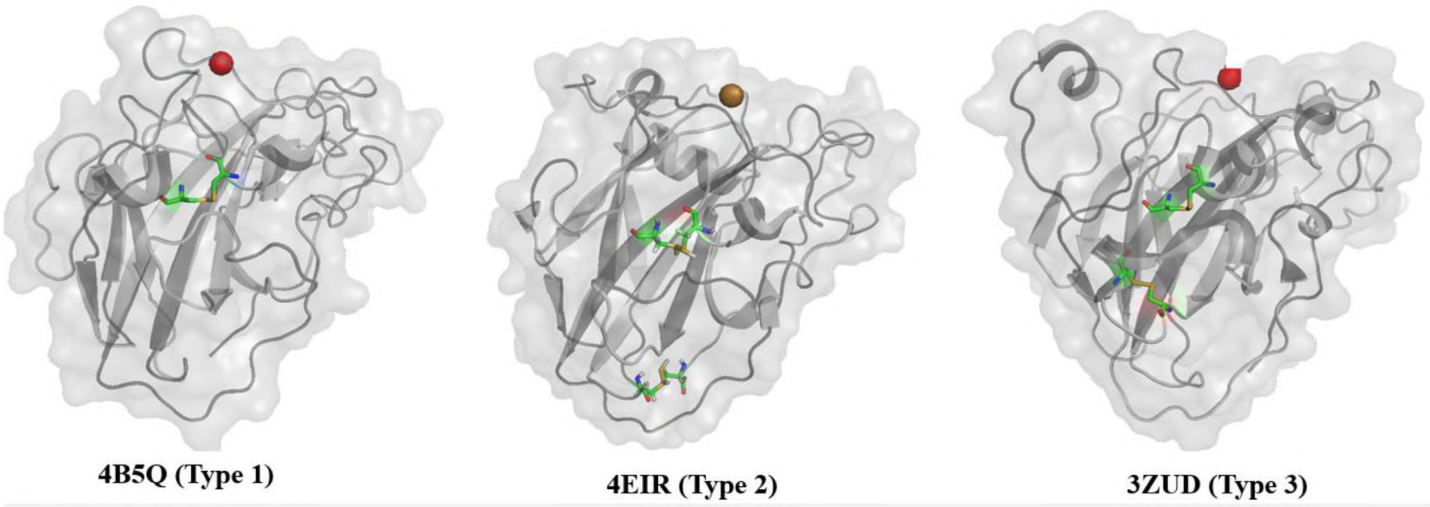

Figure 4.2. Disulphide linkages found on respective Type AA9 structures. The residues that form disulphide linkages are shown on stick representation. The three AA9 LPMO types are respectively represented by the crystal structures 4B5Q, 4EIR and 3ZUD.

Prior to MD simulations, the identified disulphide linkages (Table 4.1 and Figure 4.2) were patched into the CHARMM input files to ensure accurate representation of the AA9 type in the

simulations. SSBOND was used to create the disulphide bridges in all simulations (Hazes, Dijkstra 1988).

4.3.2.2. AA9 active sites

AA9 protein are Type II copper coordinating enzymes. They use a histidine brace to coordinate the metal center in all three AA9 types as shown in Figure 4.3. The AA9 active sites are well conserved throughout all AA9 LPMPO types. As a result, the AA9 Type 1 copper active site force field parameters that were determined were readily transferable to other AA9 types. The only issue to be considered was the methylation of the N-terminal Histidine residue 1 of both Type 2 and 3 crystal structures.

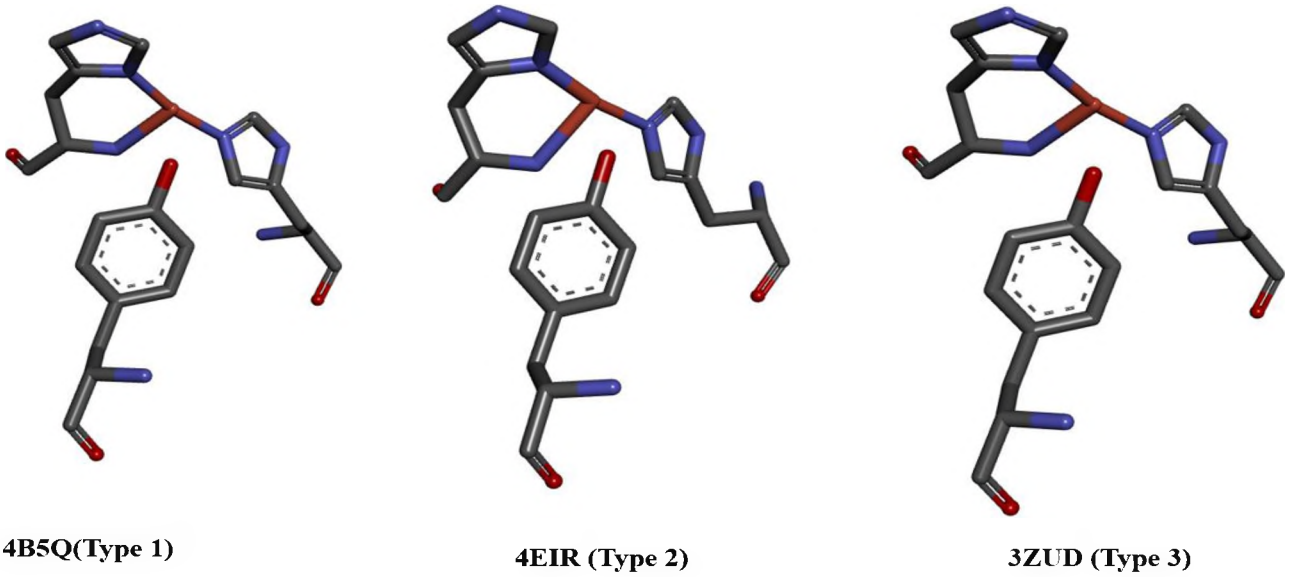

Figure 4.3. Copper coordinating active site residues for Type 1, 2 and 3 AA9 proteins. The conserved histidine brace of the AA9 active site are visualised for all three AA9 LPMO types.

As stated earlier, the AA9 active site is a Type II Copper (Cu(II)) that is coordinated by a Histidine brace (Quinlan et al. 2011). That means that there are two Histidine residues that coordinate the copper at three positions and a Tyrosine residue that occupies the fourth position (Kjaergaard et al. 2014).

4.3.2.3. Parameter translation across AA9 types

The elucidated Type 1 force field parameters were inserted into the CHARMM 36 force field to validate the parameters using the Type 1 crystal 4B5Q as input. The same parameters were inserted to the CHARMM 36 force field for both the Type 2 and 3 AA9 crystal structures (4EIR and 3ZUD). To perform the MD analysis for Type 2 and 3 AA9 protein structures, the Type 1 force field parameters from Chapter 3 were translated across to other AA9 types to make sure they are compatible.

Table 4.3. Translation of Type 1 force field parameters to Type 2 and 3 AA9 proteins.

Parameter	**Residues**		
Bonds	**4B5Q(Type 1)**	**4EIR (Type 2)**	**3ZUD (Type 3)**
Cu – O (Tyr)	Tyr-160-Cu-221	Tyr-168-Cu-224	Tyr-175-Cu-229
Cu – O (Water)	O-Cu-221	-	-
Cu – ND1	His-1-Cu-221	His-1-Cu-224	His-1-Cu-229
Cu – NE2	His-76-Cu-221	His-84-Cu-224	His-86-Cu-229
Cu – N	His-1-Cu-221	His-1-Cu-224	His-1-Cu-229
Angles			
ND1-Cu-NE2	His-1-Cu-221-His-76	His-1-Cu-224-His-84	His-1-Cu-229-His-86
ND1-Cu-N	His-1-Cu-221-His-1	His-1-Cu-224-His-1	His-1-Cu-229-His-1
Cu-ND1-CE1	His-1-His-1-Cu-221	His-1-His-1-Cu-224	His1-86-His-1-Cu-229
Cu-ND1-CG	His-1-His-1-Cu-221	His-1-His-1-Cu-224	His-86-His-1-Cu-229
Cu-NE2-CD2	His-76-His-76-Cu-221	His-84-His-1-Cu-224	His-86-His-1-Cu-229
Cu-NE2-CE1	His-1-His-1-Cu-221	His-84-His-1-Cu-224	His-86-His-1-Cu-229
Cu-NH2-HT1	His-1-His-1-Cu-221	His-1-His-1-Cu-224	His-86-His-1-Cu-229
Cu-NH2-HT2	His-1-His-1-Cu-221	His-1-His-1-Cu-224	His-86-His-1-Cu-229
Cu-OH-HH	Tyr-160-Cu-221	Tyr-168-Cu-224	Tyr-175-Cu-229
O(water)-Cu-O(water)	-	-	-
O(Tyr)-Cu-O(water)	-	-	-
O(Tyr)-Cu-O(NE2)	His-76-Cu-221-Tyr-160	His-84-Cu-224-Tyr-168	His-86-Cu-229-Tyr-175

Dihedral			
Cu-ND-C-C	Cu-221-His-1-His-1-His-1	Cu-224-His-1-His-1-His-1	Cu-229-His-1-His-1-His-1
NE-Cu-N-H	Cu-221-His-1-His-1-His-1	Cu-224-His-1-His-1-His-1	Cu-229-His-1-His-1-His-1

As a result, the translated position of the Type 1 4B5Q force field parameters for the Type 2 4EIR and Type 3 3ZUD AA9 crystal structures is shown in Table 4.3. Script_1.py revealed similar coordination pattern for both Type 2 and 3 AA9 proteins.

4.3.3. Force field parameter validation – MD simulations

The validation of the force field parameters was done using MD simulations. The protocol for MD simulation was implemented in CHARMM software package (Brooks et al. 2009). The MD protocol involved structure preparation, vacuum minimization, solvation, neutralization, minimization, equilibration and the final production run. A summary of the MD simulation approach followed is shown in Figure 4.4.

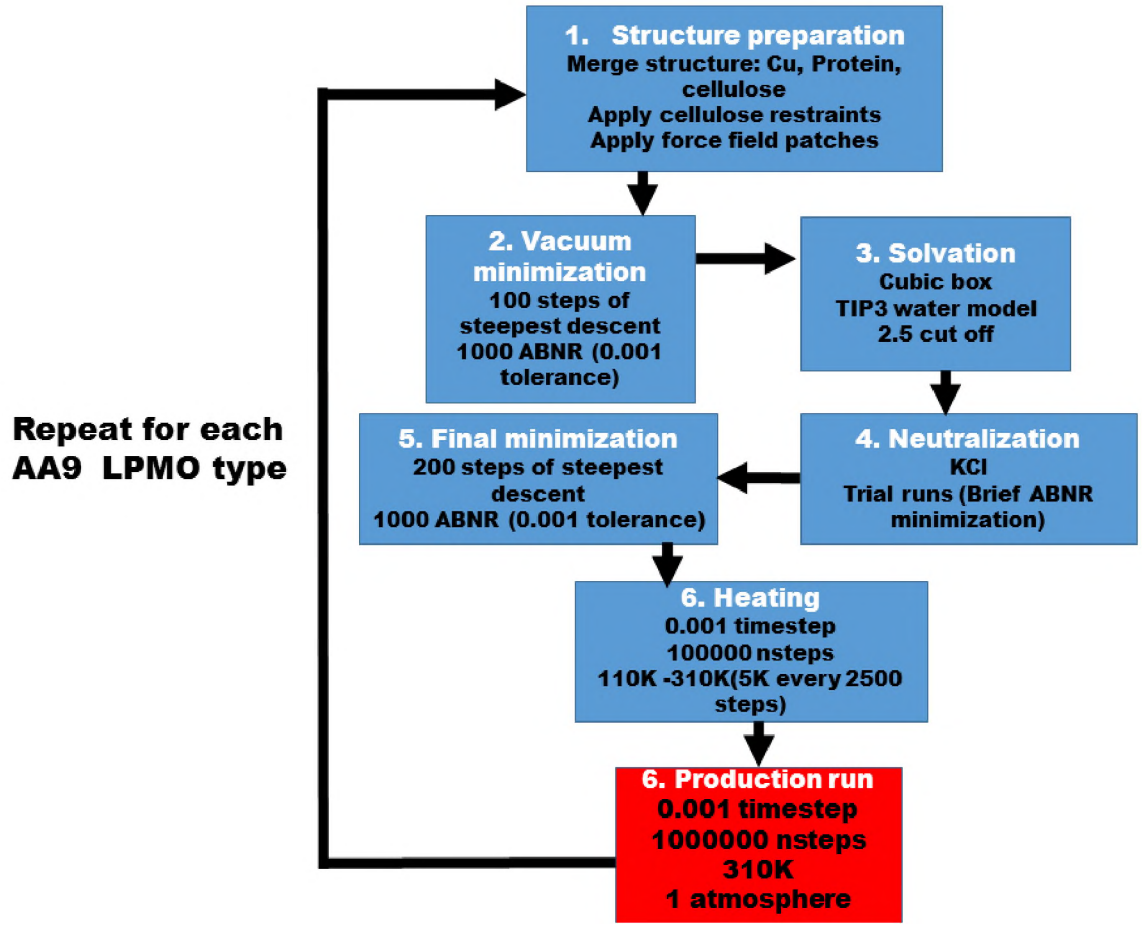

Figure 4.4: Molecular Dynamics flow diagram. This figure details the steps taken to validate force field parameters through MD simulations.

To simulate the AA9-cellulose a cubic box with the box dimensions of 90.0, 90.0 and 90.0 was used. The box used had periodic boundary conditions. The force field used for the simulation of all three AA9 types was the CHARMM 36 force field. This force field was chosen because it had parameters that describe the protein residues excluding the copper atom and the cellulose. The calculated copper parameters were then included into the force field. The system was then minimized *in vacuo* with a 100 steps of the steepest descent method. After this 1000 steps of the Adopted Basis Newton-Raphson (ABNR) were used to minimize using a 0.001 tolerance. After completion of the minimization step, the cellulose AA9 complex was solvated with TIP3P water (Mark, Nilsson 2001) within a cubic box. The cut off used for solvation was 2.5 Å. The solvated system was then neutralized using an excess of potassium cations (K^+) with 0.15M KCl. A second minimization was performed with 200 steps of steepest descent method and 1000 steps of ABNR

method. The gradient tolerance used was with a 0.001. Heating of the complete system was done for 100000 steps. The time step used was 0.001 (1fs). The system was heated from 110K to 310K using 5K increments every 2500 steps. This was done to negate the need to equilibrate the system. The final MD runs were performed 10^7 steps to generate a 10 ns MD run. The simulation was performed under constant pressure and temperature (CPT) conditions at 310K and 1 atmosphere pressure. All of these steps of dynamics were performed twice, using both literature available Cu^{2+} Lennard-Jones parameters.

4.3.4. CHARMM scaling test

To test the most optimal resources to perform MD simulations using the CHARMM software package, short MD runs were performed utilizing various combinations of resources. The tests were performed using 10000 steps of simulation using a 0.001 (1 fs) timestep. The test were performed on 2, 4, 8, 16, 32, 64 and 128 cores. To test the higher number of cores, starting from 32 cores, multiple nodes were used. Number of nodes used and the findings of the performance tests are shown in Figure 4.5.

	Number of nodes					
	1 node	2 nodes	3 nodes	4 nodes	5 nodes	6 nodes
2 cores	120.25	120.25	120.25	120.24	120.23	120.24
4 cores	60.15	60.15	60.15	60.15	60.15	60.15
8 cores	44.18	39.58	39.64	39.62	39.65	39.64
16 cores	23	25.61	23.07	23.05	22.99	22.99
32 cores	-	14.97	15.01	14.98	15.01	15.1
64 cores	-	-	12.72	13.38	12.82	12.82
128 cores	-	-	-	-	-	13.34

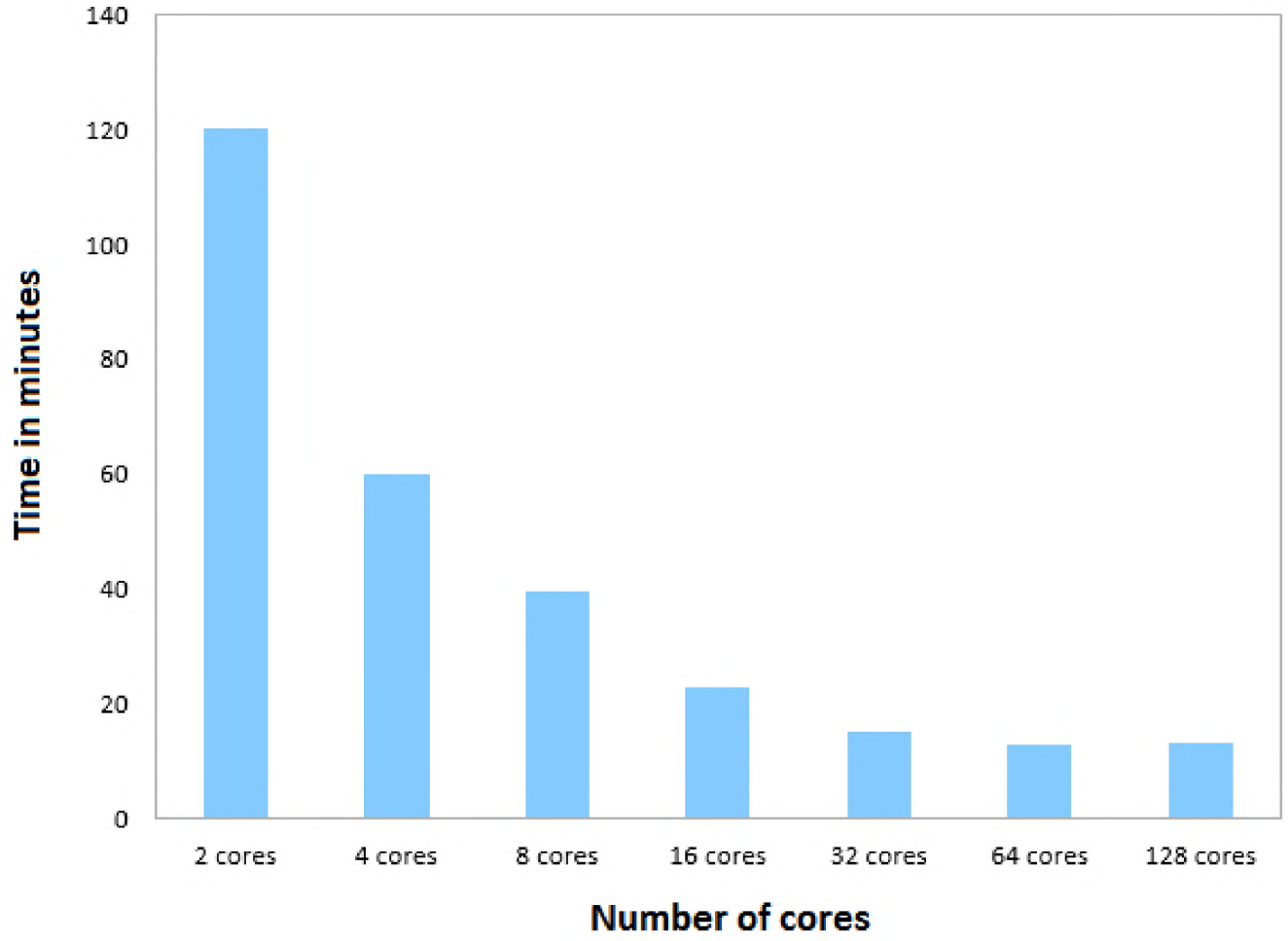

Figure 4.5: CHARMM scaling tests. The most optimal resources to perform MD simulations were tested on 2, 4, 8, 16, 32, 64 and 128 cores.

From Figure 4.5 it can be seen that there is no effect on performance from using multiple nodes. It was found the most optimal resources to use to perform MD simulations was determined to be 64 cores since it performed the simulation in the lowest amount of time generally below 13

minutes. As a result the resources that were selected to perform all subsequent simulations was 64 cores and to minimize internode communication time wastage, the lowest number of nodes (3 nodes) was used. All calculations for MD were performed using the computing resources available at the Centre for High Performance Computing (CHPC, Cape Town, South Africa). MD simulations were performed for all three AA9 LPMO types using the 4B5Q, 4EIR and 3ZUD crystal structures to represent Type 1, 2 and 3 AA9 LPMO types to study their interaction with cellulose.

4.3.5. Contact maps

To understand how the identified type–specific features of AA9 proteins affect substrate binding, contact maps were created. The contact maps were created for both biased and unbiased MD experiments for all three AA9 LPMO types. The contact maps were created at five time intervals for the 10ns trajectories. The time intervals chosen were; 0ns, 2ns, 4ns, 6ns, 8ns and 10ns. Script_3.py was used to measure the regions in respective AA9 proteins that make contact with the Top celluloses layers. The top cellulose layer consisted of five chains termed; M0, M3, M6, M9 and M12. Once the minimum contacting regions for both AA9 proteins were identified VMD was used to map out the findings.

4.3.6. Hydrogen bonding analysis

When type-specific contacts were identified, it was important assess how these regions were interacting with cellulose. Because the cellulose structure is infamous for its ability to form hydrogen bonding, the hydrogen bonding between AA9 LPMO types and cellulose was investigated. This was done to determine which area of the AA9 protein that interact with the top layer of cellulose. The criteria to describe a hydrogen bond was a maximum donor-acceptor distance of 3Å and an angle cutoff was 20◦. The hydrogen bonding analysis was performed using VMD.

4.3.7. DSSP analysis

To determine the stability of secondary structural elements during the MD simulations of all three AA9 LPMOs, DSSP analysis was performed using VMD. The structural elements that were investigated were beta-Turns (T), extended confirmations or Beta sheets (E), isolated bridges (B), 3-10 helices (G), Pi helices (I) and coils (C). Each 10 ns trajectory, for both biased and unbiased MD experiments was uploaded to VMD and the analysis was carried out. The DSSP algorithm assigns secondary structural elements to amino acids of a particular protein (Kabsch, Sander 1983). For the proteins in motion the continuous DSSP assignment is implemented by VMD (Andersen et al. 2002).

4.4. Results

For each of the AA9 LPMO types, two MD simulations were performed for the biased and unbiased Lennard-Jones parameter sets. The unbiased parameter set used is based on the Cu^{2+} ion that was established in TIP3P water (Li, Merz 2014). The second biased parameter set used was obtained from Cu^{2+} Lennard-Jones parameters that were established in acetonitrile (Torras, Aleman 2013). MD simulations were then performed for both parameter sets. To aid in the validation of the force field parameters various features were analyzed in the trajectories. After the force field and Lennard-Jones parameters were validated the AA9-cellulose interaction was analyzed to assess the type-specific interactions with cellulose. The features used to evaluate the AA9-cellulose interaction were the potential energy, root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration, and the bond lengths of the coordinating position of the Cu^{2+} cation were monitored throughout the simulation. To assess the overall movement of the AA9 proteins on the cellulose substrate the proteins were visualized using VMD and the RMSF was measured to assess the regions that contribute to the observed movement. The findings of this analysis are shown in for Figure 4.6, 4.7 and 4.8 for Type 1, 2 and 3 AA9 LPMO respectively for both biased and unbiased experiments. This was done to monitor regions of the AA9 protein that undergoing large displacements during the MD simulations and their contribution to movement relative to cellulose. The findings suggest that the major contributors to protein movement are the loop regions due to the fact that these regions displayed the highest degree displacement compared to the β-sandwich fold. RMSD, radius of gyration and the potential energy was used to evaluate the stability of the protein-cellulose complex during the MD simulation (Figure 4.12, 4.13 and 4.14 for Type 1, 2 and 3 respectively).

4.4.1. AA9 movement on cellulose

AA9 proteins were found to have movement across the cellulose substrate. The movement for Type 1, 2 and 3 AA9 proteins is summarized in Figures 4.6, 4.7 and 4.8 respectively. The movement of AA9 crystal structures on the cellulose substrate was quantified by visualizing the protein structures during the course of the simulation using VMD and measuring the RMSF.

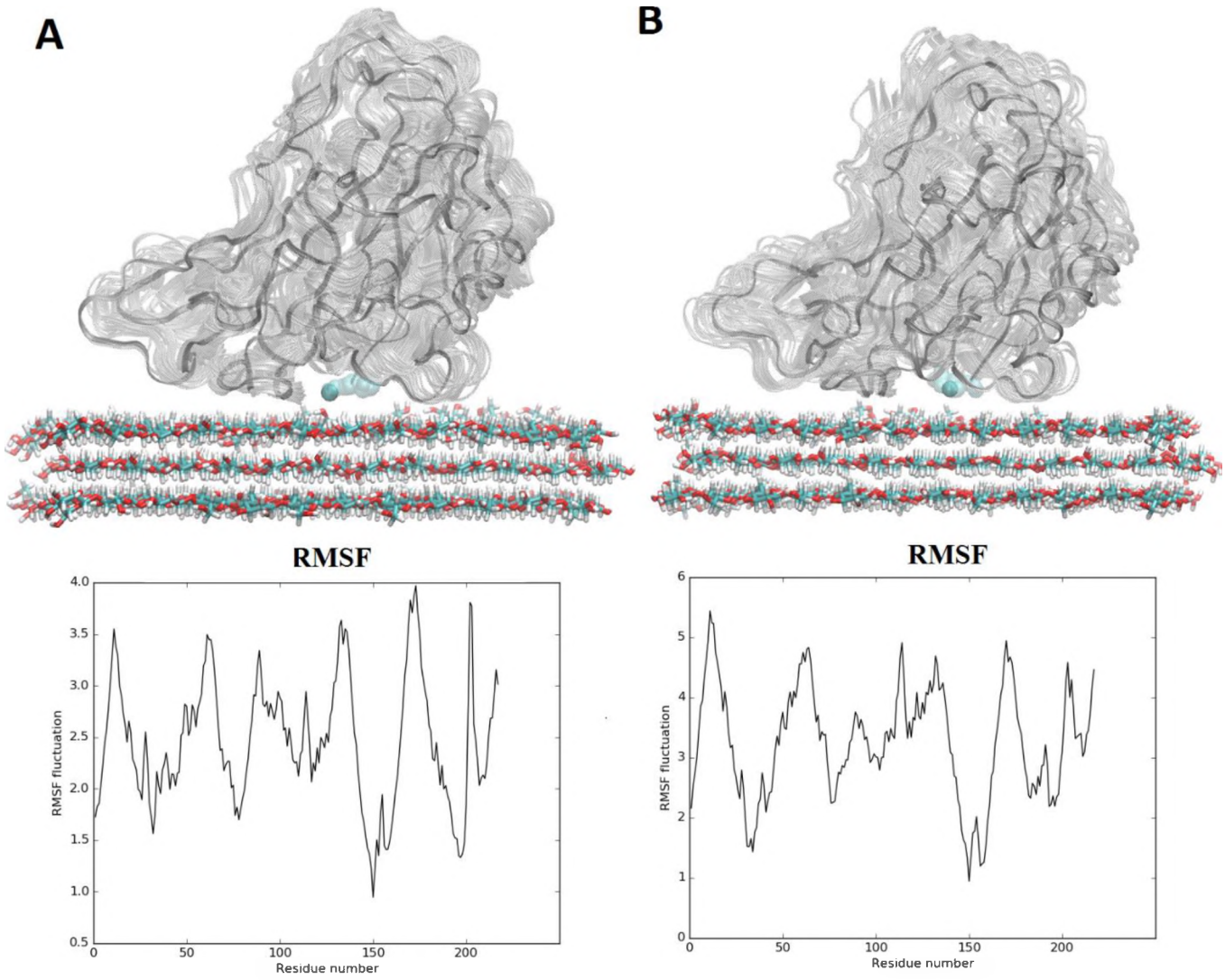

Figure 4.6: Relative motion on cellulose surface and root mean square fluctuation of biased and unbiased MD experiment for Type 1 AA9 proteins. The movement of the Type 1 AA9 protein (4B5Q) on the cellulose substrate is represented by the gray transparent representation relative to the starting structure. Snapshots were taken every 0.5 ns to show protein movement. RMSF is plotted per residue to show regions contributing to the observed movement. A) Biased MD experiment B) Unbiased MD experiment.

Through the course of the 10 ns simulation snapshots were taken to show the overall displacement of the protein from time 0 ns to 10ns. It can be seen that for the Type 1 MD experiment for both the biased and biased experiment there is movement across the cellulose substrate (Figure 4.6). During the displacement the active site of the Type 1 4B5Q crystal structure remains orientated towards the exposed surface layer of the cellulose. There were extreme fluctuations of the loop regions observed in both the biased and unbiased experiments. To a lesser degree the secondary structure regions also showed fluctuation. However, the observed fluctuation may be attributed to

the three AA9 proteins overall displacement across the celluloses substrate. Initial binding to cellulose was observed at time 0ns for the biased experiment and through the course of the simulation an exchange occurs between Cu^{2+} at and the hydroxyl atoms of the cellulose. The Type 1 4B5Q protein is found to initially bind one cellulose chain and through the course of the simulation the chain is replaced by another. This observation is illustrated in Figure 4.10 and 4.12 and will be further discussed in Section 4.4.3 in this Chapter. The movement of the metal ion from one cellulose chain to another is illustrated clearly in Figure 4.6 A for the Type 1 biased experiment. The blue sphere indicating the Cu^{2+} atom is shown moving in straight part across the cellulose substrate. This illustrates the cellulose chain exchange that occurs during the simulation. In the Type 1 unbiased experiment (Figure 4.6 B) the movement of the Cu^{2+} atom is not as well defined as the biased experiment. This is largely attributed to the more moderate Lennard-Jones parameter set used in this experiment. For the unbiased experiment binding to cellulose did not occur instantaneously as it did for the biased experiment due to the weak Lennard-Jones parameter used. Even with the weaker Lennard-Jones parameter set used, binding to cellulose was observed for the Type 1 unbiased MD experiment, however binding occurred much later in the simulation. As a result, there was no cellulose chain exchange observed for the unbiased Type 1 MD experiment. The lack of cellulose chain exchange is attributed to short nature of the MD run. Due to the weak Lennard-Jones parameters used for the unbiased experiment, 10 ns would be a sufficient time to observe exchange. As a result this, study may benefit from an extension of the MD runs.

The Type 2 MD simulations for the biased and unbiased experiments was performed and the results are shown in Figure 4.7. It was found that similar to Type 1 MD experiments, the Type 2 MD experiments also displayed movement of the AA9 protein relative to the cellulose substrate for both biased and unbiased experiments.

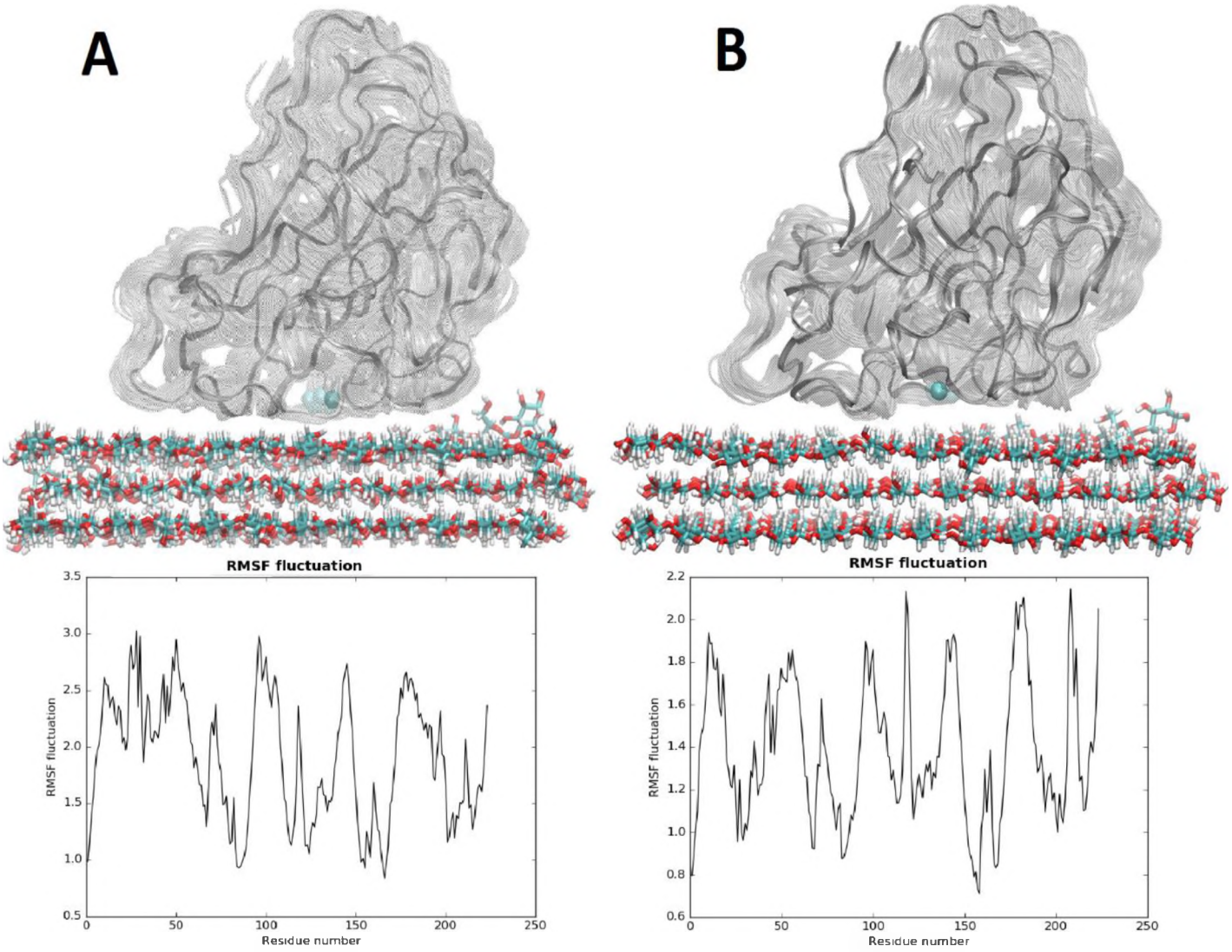

Figure 4.7: Relative motion on cellulose surface and root mean square fluctuation of biased and unbiased MD experiment for Type 2 AA9 proteins. The movement of the Type 2 AA9 protein (4EIR) on the cellulose substrate is represented by the gray transparent representation relative to the starting structure. Snapshots were taken every 0.5 ns to show protein movement. RMSF is plotted per residue to show regions contributing to the observed movement. A) Biased MD experiment B) Unbiased MD experiment.

Like the Type 1 MD experiment the displacement of the AA9 protein was attributed to fluctuation of the loop regions since the secondary element regions of both biased and unbiased experiments were found to fluctuate to a lesser extent. Based on the RMSF measurements for Type 1 and 2 MD experiments it is apparent that the Type 1 MD experiment has a higher degree of fluctuation as opposed to the Type 2 experiment. This suggests that during the course of the 10 ns MD simulation,

the displacement of the Type 2 4EIR crystal structure is less than that of the Type 1 crystal structure. This decrease in movement may be attributed to inserts that are present in the 4EIR crystal structure but are absent in the 4B5Q structure as shown previously in Chapter 2. Insert II on the 4EIR crystal structure was shown to be localized on the flat surface active site spanning Leu-69 to Met-80. Similarly, insert I on the 3ZUD crystal structure was found localized on the active site surface of the protein. Unlike the Type 1 MD experiments, binding to cellulose was not observed for the both the biased and unbiased experiments of the Type 2 MD experiments. The absence of binding to cellulose can be attributed to the insert II of Type 2 sequences that results in steric congestion that may prevent binding to cellulose. As previously hypothesized by an earlier study (Hemsworth, Davies & Walton 2013), the steric congestion of different AA9 Types is a likely contributor to type-specific regioselectivity. Due to the fact that binding to cellulose was not observed for Type 2 AA9 proteins, the use of a biased and unbiased Lennard-Jones parameter set had no affect the Type 2 AA9-cellulose binding. Since Biasing the Type 2 MD experiments did not have an effect on binding to cellulose, increasing the length of the MD experiments may provide further insights to this study. This is due to the fact that the 10 ns simulations performed may have not been long enough to observe cellulose binding.

The MD simulations of the biased and unbiased parameter sets are shown in Figure 4.8 for the Type 3 crystal structure 3ZUD. It was found that for both the biased and unbiased experiments the AA9 flat surface active site of the 3ZUD crystal structure remained oriented to the surface exposed top layer of cellulose. These findings are consistent with what was observed for both Type 1 and 2 MD experiments in the sense that movement across the cellulose substrate was observed. Like the Type 2 MD simulations, the Type 3 MD simulations showed a relatively less displacement across the cellulose substrate as opposed to the Type 1 MD simulations. In Chapter 2 it was revealed that the Type 3 AA9 sequences possess insert I. The presence of this insert is likely to result in a similar steric congestion that was observe for Type 2 AA9 proteins. As result of this steric congestion, the Type 3 MD experiments did not reveal any binding to cellulose. Both the biased and unbiased experiments yielded comparable results.

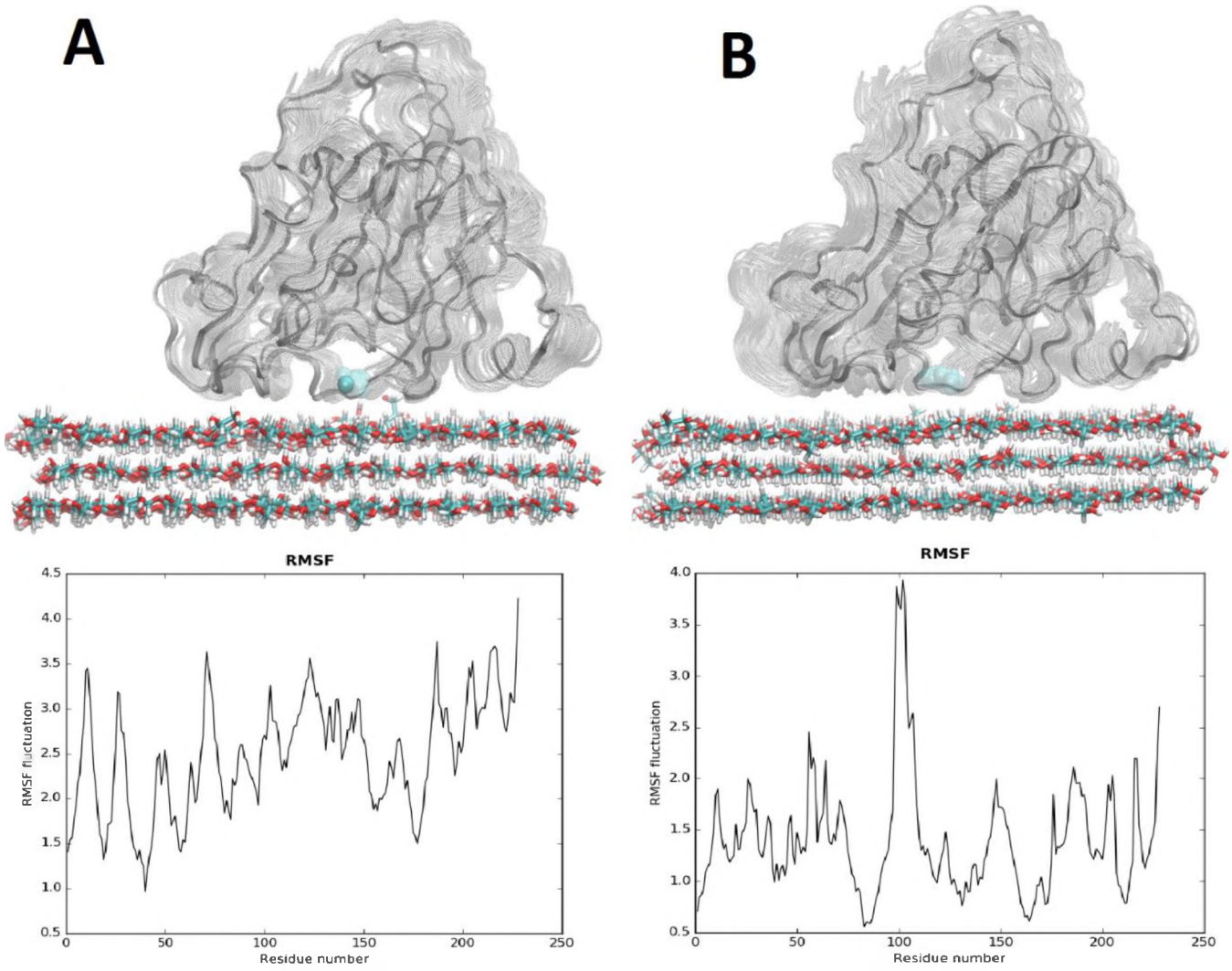

Figure 4.8: Relative motion on cellulose surface and Root mean square fluctuation of biased and unbiased MD experiment for Type 3 AA9 proteins. The movement of the Type 3 AA9 protein (3ZUD) on the cellulose substrate is represented by the gray transparent representation relative to the starting structure. Snapshots were taken every 0.5 ns to show protein movement. RMSF is plotted per residue to show regions contributing to the observed movement. A) Biased MD experiment B) Unbiased MD experiment.

Overall, for both biased and unbiased MD experiments binding to cellulose was observed for Type 1 proteins. Due to the different Lennard-Jones parameters used binding to cellulose was observed much sooner for the biased experiment as opposed to the unbiased experiments. For Type 2 and 3 AA9 proteins no binding to cellulose was observed. This is likely a consequence of the presence of Insert I and II on the active site surfaces of Type 2 and 3 AA9 proteins. These inserts possibly cause steric congestion of the active site that prevents binding to the cellulose substrate. Another

possibility is that 10ns runs where not long enough to result in cellulose binding as a result lengthening of MD runs may be beneficial. The Type 1 MD showed the highest displacement relative to the cellulose substrate. This is evidenced by the higher RMSF values observed. Type 2 and 3 MD though showing less displacement, where found to have more residues undergoing fluctuations. As previously shown in Figure 2.9 in Chapter 2, Type 2 and 3 AA9 proteins are generally larger than Type 1 proteins. As a results there are more regions to undergo fluctuations.

4.4.2. Secondary structure evolution in both biased and unbiased MD experiments

VMD was used to measure the stability of secondary structural elements during the simulation. This was achieved through the implementation of the DSSP algorithm using the Timeline extension (Stone et al. 1995) of the VMD program. The results of this analysis is shown in Figure 4.9 below.

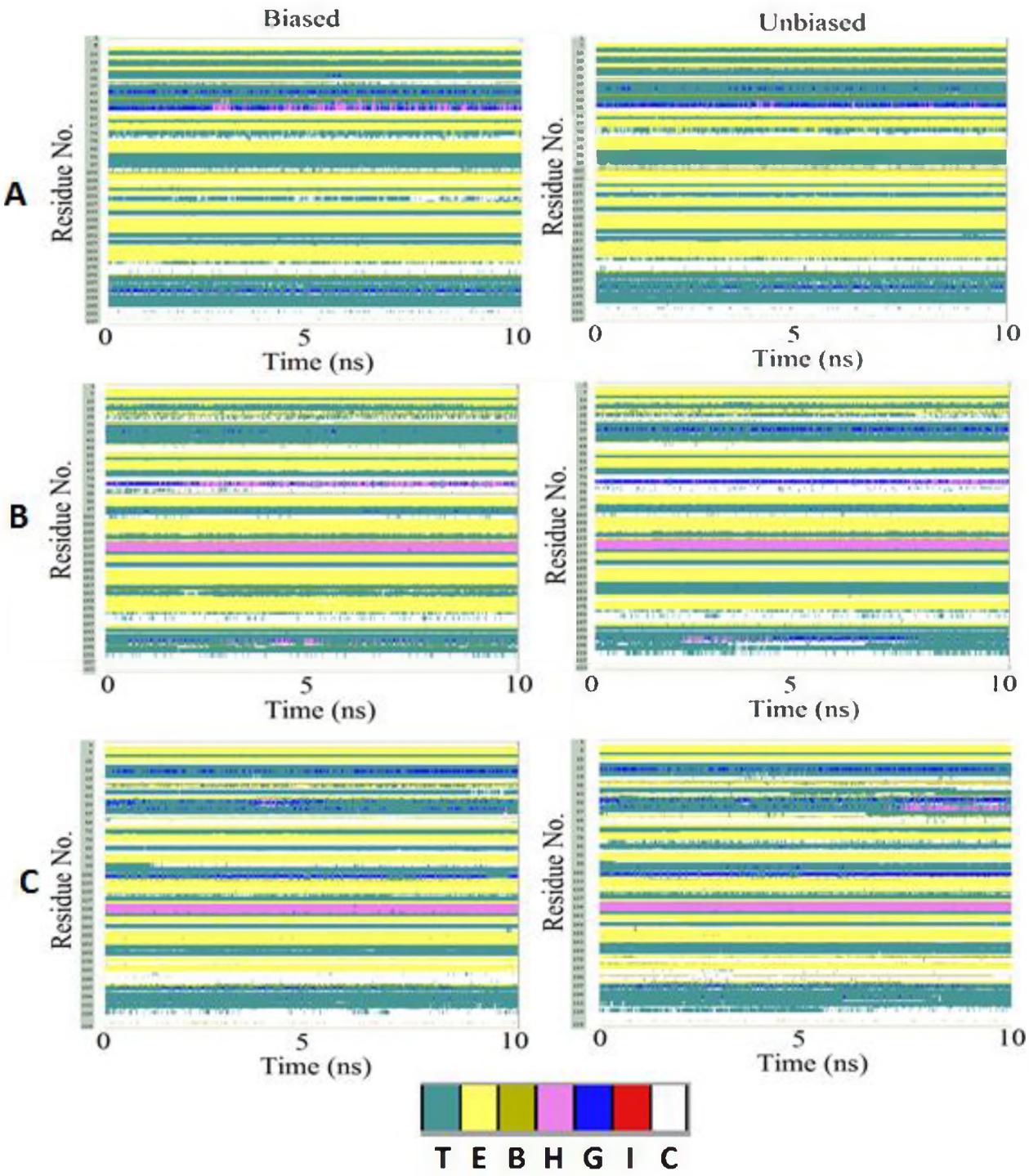

Figure 4.9: The conservation of secondary structural elements during MD. The secondary elements were monitored through the course of the 10 ns MD experiments. T represents the Turn, E is the extended confirmation or Beta sheet, B is an isolated bridge, G are 3-10 helices, I represents Pi helices and C represents coils. This analysis was performed for Type 1 (A), Type 2 (B) and Type 3 (C) biased and unbiased MD experiments.

The DSSP algorithm was used to assign secondary structure of the amino acids of AA9 proteins, in both biased and unbiased MD experiments. This analysis was carried out to assess the stability of secondary elements of AA9 proteins due to the high fluctuation observed in Figure 4.6, 4.7 and 4.8 with respect to RMSF. It was observed that AA9 proteins possess the rare structural element called 3-10 helices (G) (Figure 4.9). 3-10 helices are protein structural elements that are characterized by having two or more hydrogen bonds that occur between the main chain carbonyl of residue *i* and the amide hydrogen of the main chain *i*+3. There are 10 atoms in the ring that forms the hydrogen of 3-10 helices hence their name (Enkhbayar et al. 2006). The 3-10 helices are often believed to be unstable structures in proteins due to their low occurrence in proteins and their distorted hydrogen bonding network. This also appear to be the case with respect to AA9 proteins (Vieira-Pires, Morais-Cabral 2010). In Figure 4.9 it was revealed that for all 3 AA9 LPMO types MD simulations were found to form 3-10 helices at some stage in the simulation. The localization of 3-10 helices is shown in Figure 4.10.

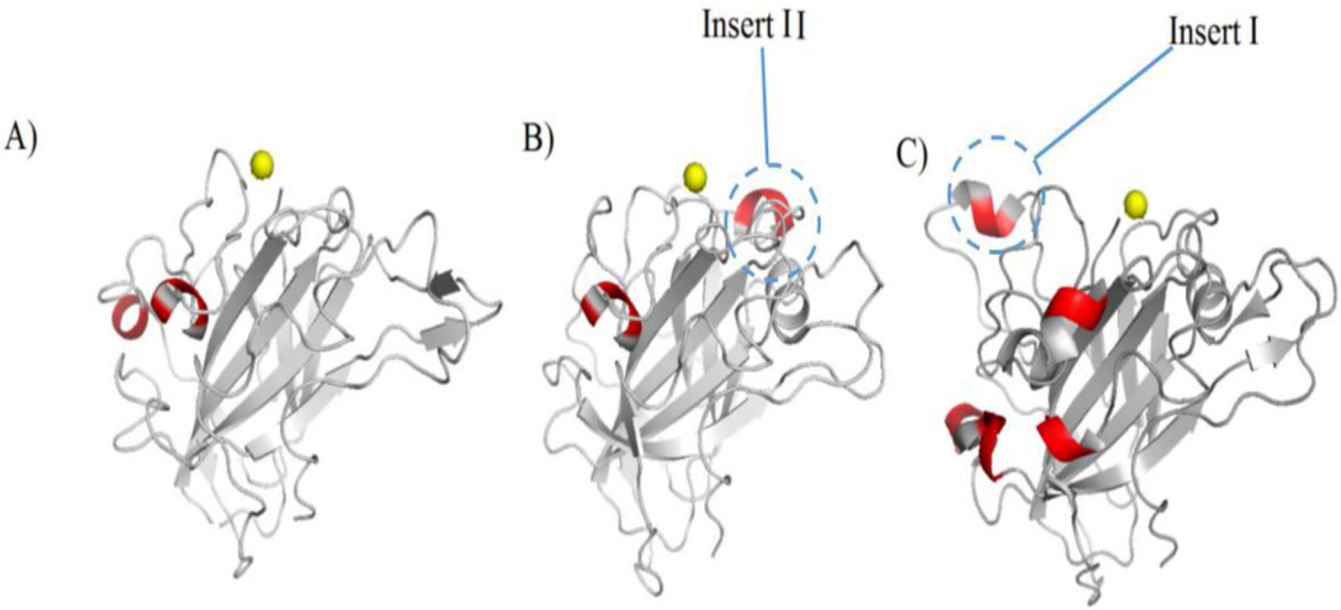

Figure 4.10: Localization of 3-10 helices on AA9 LPMO types. The crystal structure Type 1 (A), Type 2 (B) and Type 3 (C) are represented in cartoon and colored in gray. The 3-10 helices are shown in red.

It was found that the majority 3-10 helices is not found on the flat face active site of AA9 proteins. However, for Type 3 and 3 AA9 proteins (Figure 4.10 B and C respectively), 3-10 helices were found associated with their type-specific inserts. For Type 2 AA9 proteins (4EIR) the 3-10 helix

was found associated with Insert II and For Type 3 AA9 proteins (3ZUD) the 3-10 helix was associated with insert I. There was no 3-10 helix associated with the active site of Type 1 AA9 proteins (4B5Q) (Figure 4.10). The most prominent structural element in all AA9 proteins was found to be the extended conformation or beta sheet. This is to be expected as the beta-sandwich fold is a major constituent of AA9 proteins. The extended conformation was found to be conserved throughout all simulation showing that the beta-sandwich fold of AA9 LPMO types is rigid structure. Beta turns or tight loops (T) were also found to be prominent in AA9 structures. Like beta sheets, the beta turns were found to be well conserved during simulations. A less common structural element determined to be the alpha-helices. Type 1 LPMO types were found to show the least occurrence and low conservation of 3-10 helices. In Type 2 and 3 LPMO types there is a conserved alpha helical structures that is present throughout all simulations. Overall it can be seen that the α-helices, β-sheets of AA9 proteins are well conserved the simulation. However, 3-10 helices, beta turns and coils were found to have the least conservation as they were found to be constantly interchanging during the simulation.

4.4.3. Binding to cellulose

Binding of the Cu^{2+} to cellulose was detected for both the Type 1 MD experiments. There was no binding detected for Type 2 and 3 MD experiments. This was attributed to the presence of inserts in Zone I and II of AA9 proteins. The possible effects of inserts I and II had previously been investigated in Chapter 2 Figure 2.4. The inserts possibly form steric congestion around the Cu^{2+} of AA9 proteins and this results in the different regioselectivity displayed by these enzymes. It was show in Figure 4.8, 4.19 and 4.20 that all three AA9 LPMO types have type-specific interaction with the cellulose substrates. The inserts I and II were found to be contributors to these type-specific interactions. The bonds around the Cu^{2+} center were monitored through the course of the simulation this included all coordinating atoms from both protein and the cellulose substrate. The binding observed for the Type 1 MD simulations is illustrated in Figure 4.11.

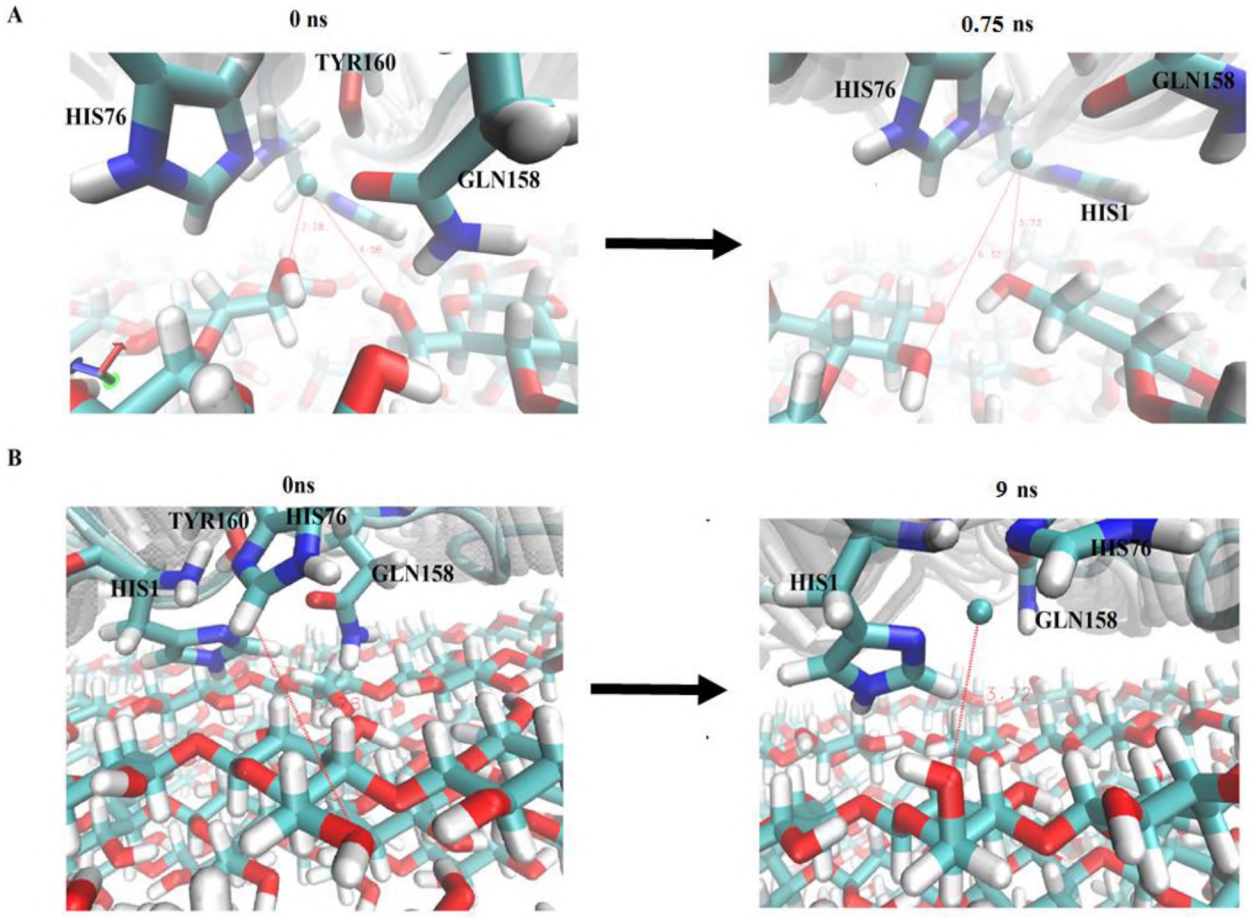

Figure 4.11: Cellulose binding during both Biased and Unbiased MD runs. Cellulose binding for A) Biased and B) Unbiased MD simulation.

Binding of the biased experiment (Figure 4.11 A) was observed during time 0 ns. This is was due to the fact that binding to cellulose had already occurred during the heating step as illustrated in Figure 4.12 A. Binding to cellulose occur during heating for the biased experiment however, binding does not happen in the unbiased experiment. This is attributed to the high epsilon value of the biased Lennard-Jones parameter used which results in relatively stronger attraction between the cellulose and Cu^{2+} atom.

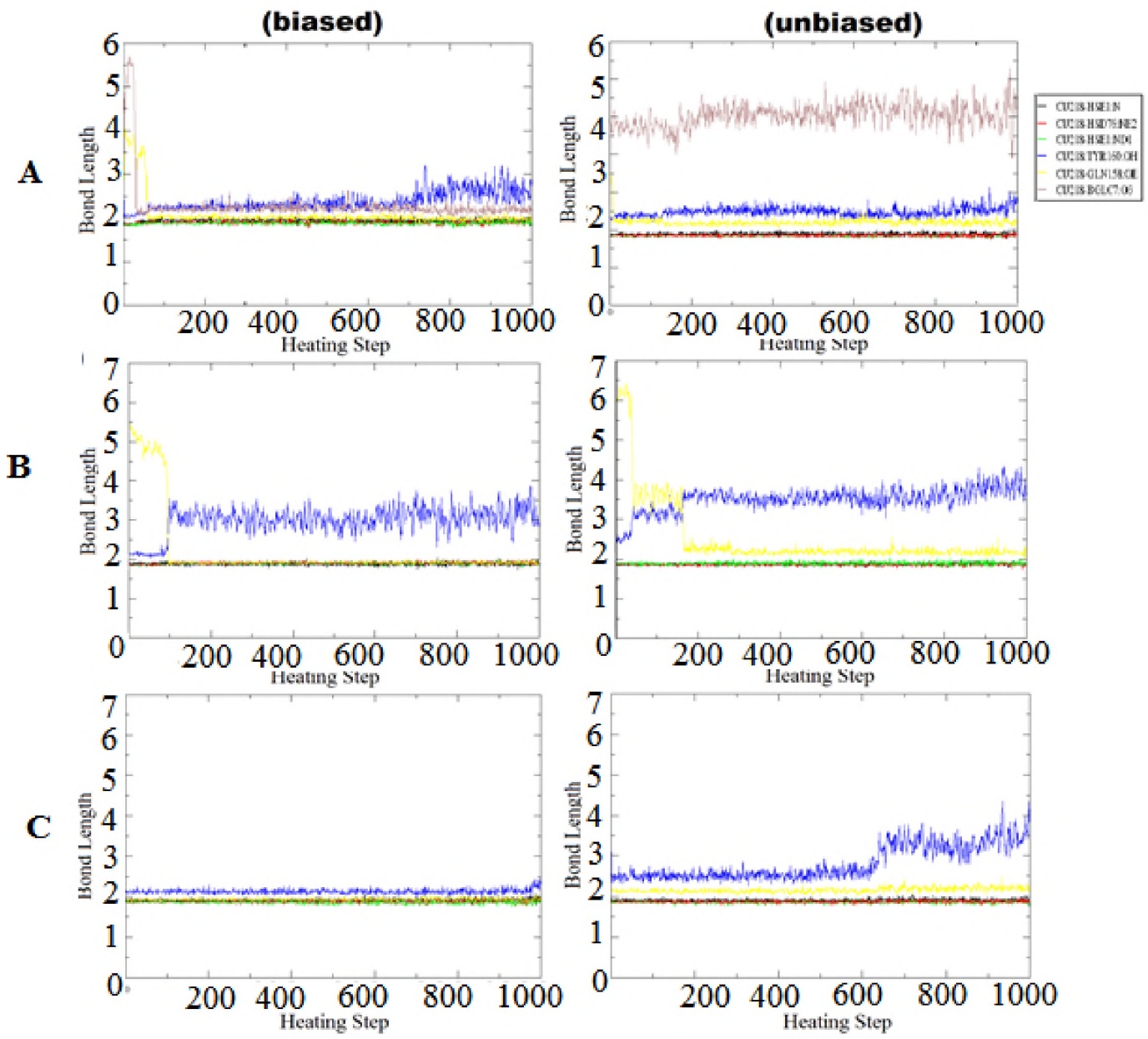

Figure 4.12: Coordination of the AA9 copper atom during heating for both biased and unbiased experiments. The distance between the AA9 active site Cu^{2+} and coordinating atoms are monitored during the heating steps. Bond length denotes the distance of the Cu^{2+} coordinating positions.

This observation is illustrated by the brown line in Figure 4.12 A which represents the relative distance of the cellulose hydroxyl group from the Cu^{2+}. It can be seen that during the unbiased experiment the cellulose is coordinated during the early stages of the heating step. On the other hand, in the unbiased experiment coordination of the cellulose substrate is not observed during the

heating step. For both the Type 2 and 3 MD experiments, there is no observable binding to cellulose (Figure 4.12 B and C). Interestingly the Gln-158 residue of the Type 1 4B5Q crystal structure was not considered for calculation or included in the subset for computation of force field parameters. However, it was found that during the heating step Gln-158 binds Cu^{2+} atom in both the biased and unbiased experiments of the Type 1 MD experiments. This is indicated by the yellow line in Figure 4.12 A. Equivalent residues were identified for the Type 2 and 3 MD experiment as indicated by the yellow lines in Figure 4.11 B and 4.11 C. All other parameters remained stable with minor fluctuations observed for both biased and unbiased MD experiments for all three AA9 LPMO types.

The coordination of the Cu^{2+} was monitored through the course of the MD simulation. The findings are summarized in Figure 4.13. For the Type 1 MD experiments, it was found that the Gln158 residue remains coordinated to Cu^{2+} during the biased experiment. However, in the unbiased experiment, the Gln-158 residue gets released at the start of the simulation. The biased Lennard-Jones parameters keeps the Gln-158 residue bound during the simulation while the more moderate unbiased Lennard-Jones parameter frees up the coordination position to interact with the aqueous environment. While the Gln-158 residue is released, the residue begins to fluctuate in the unbiased experiment, however it later stabilizes around 4 Å around the Cu^{2+} center. For the Type 2 MD experiments, the equivalent Gln-158 residue remains bound during both the biased and unbiased experiments. In the Type 3 MD experiments the Gln-158 residue equivalent remains bound in the biased experiment. However, in the unbiased experiment the residue is bound initially but is later released and fluctuates through the course of the simulation.

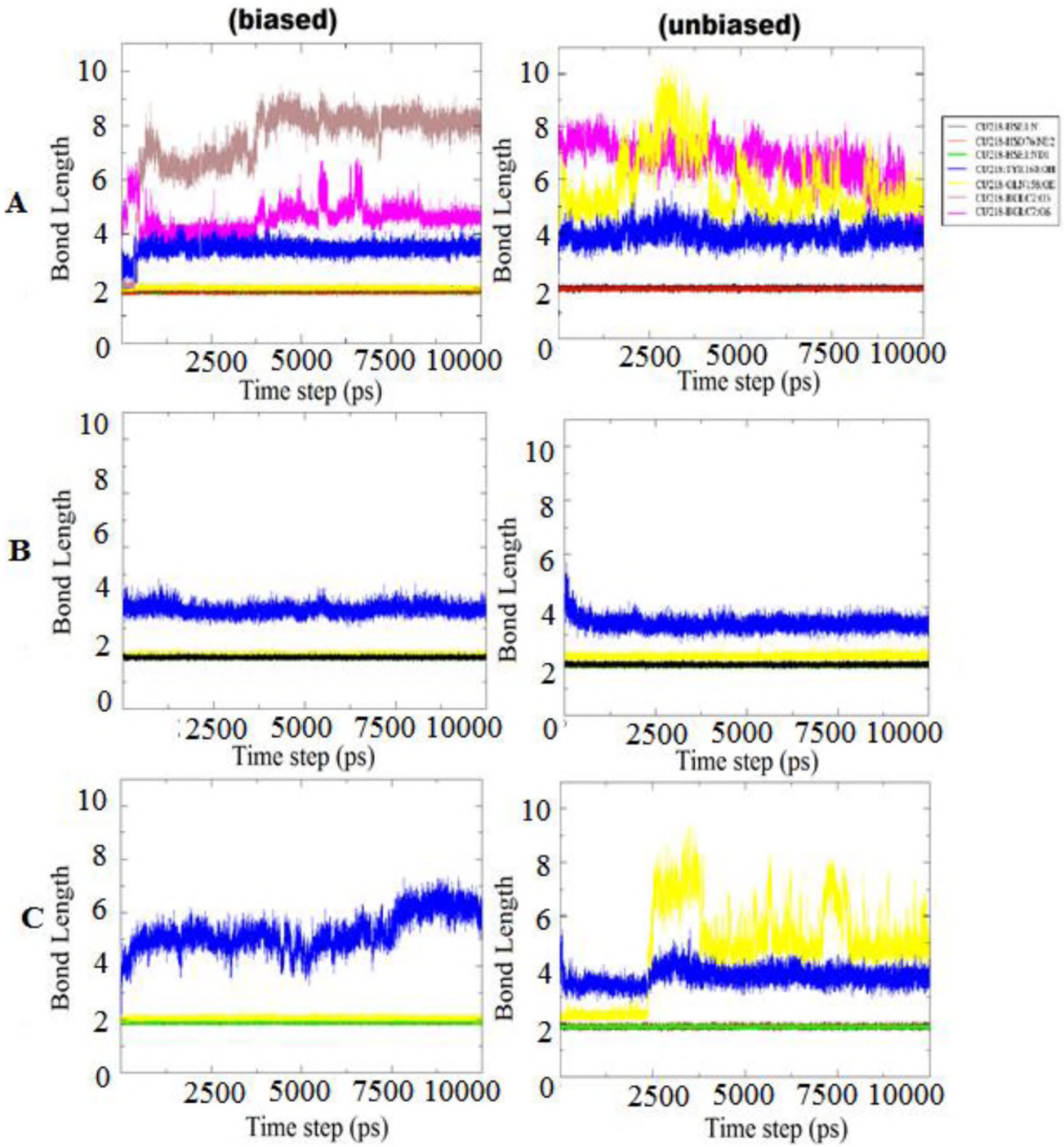

Figure 4.13: Coordination of the AA9 copper atom during the MD simulation for both biased and unbiased experiments of AA9 LPMO types. A) The coordination during the Type 1 biased and unbiased MD experiments. B) The coordination during the Type 2 biased and unbiased MD experiments. C) The coordination during the Type 3 biased and unbiased MD experiments. Bond length denotes the distance of the Cu^{2+} coordinating positions.

The distances between the AA9 active site Cu^{2+} and coordinating atoms was monitored through the course of the MD run. This Figure denotes the bond length of the bound cellulose oxygen

(OH6)-copper bond during the biased dynamics. It is seen that this bond length increases and then an exchange with another cellulose oxygen atom occurs. The original bound OH atom is released and the protein binds another OH in an opposite chain. This exchange is shown is illustrated Figure 4.11 A. In Figure 4.14 A it can be seen that even in the absence of binding to cellulose for the unbiased experiment, there seems to be a trend towards cellulose binding. This is due to the observation of proximity of between the Cu^{2+} and this cellulose OH. Close to the end of the simulation the distance between them is reduced to 3.72 Å. This is indicative of potential binding to cellulose (Figure 4.14 B). It has been reported previously that cellulose binding may occur at 4.7 Å (Wu et al. 2013) as a result, the obtained value of 3.72 Å was deemed acceptable. Therefore for both biased and unbiased experiments, potential to bind cellulose was observed for both Type 1 experiments even though the unbiased conditions took longer to yield binding. Table 4.3 displays crystal structure for the evaluated parameters compared to the average equilibrium bond distance angle and dihedral parameters obtained from Type 1 MD experiments. The parameters were generally found to keep their respective equilibrium positions relative to the Cu^{2+} center. The residue Tyr-160 and its equivalents in the other AA9 Type was found to show the greatest degree of fluctuation. This is attributed largely to the low force constant that that was obtained for this parameter.

4.4.4. Force field parameter validation

To assess the stability of the force field parameters with respect to the Type 1 MD experiment, the average values for the parameters were collected during the course of the MD simulation. This was done to ensure that there were no extreme fluctuations in the evaluated parameters. The summary of these findings is shown in Table 4.4. With the exception of the Tyr-160 residue, it was found that the force field parameters fluctuate within acceptable limits of their crystal structure values. In the case of the Tyr-160 residue, a low force constant was observed for this parameter. As a result, extreme fluctuations from the crystal structure and equilibrium bond length was observed. However, the average MD values for this parameters were found to be acceptable as shown in Table 4.3.

Table 4.3. Type 1 AA9 crystal structure parameters compared to the average parameters obtained from MD experiments.

Parameter	Crystal structure	Biased experiment	Unbiased experiment
Bonds			
Cu – OH (Tyr)	3.056	2.898	3.22
Cu – O (Water)	-	-	-
Cu – NE2	2.020	1.882	1.909
Cu – ND1	2.000	1.881	1.962
Cu – N	2.081	1.878	1.974
Angles			
ND1-Cu-NE2	174.577	151.788	149.435
ND1-Cu-N	91.784	93.656	89.958
Cu-ND1-CE1	127.270	131.482	126.931
Cu-ND1-CG	125.486	123.788	125.962
Cu-NE2-CD2	127.956	122.300	124.739
Cu-NE2-CE1	124.8750	107.911	114.125
Cu-NH2-HT1	103.803	94.864	97.545
Cu-NH2-HT2	103.586		
Cu-OH-HH	93.311	91.017	95.356
O(water)-Cu-O(water)	-		
OH(Tyr)-Cu-O(water)	-		
OH(Tyr)-Cu-N	83.372	73.457	63.499
Dihedral			
Cu-NH2-C-H	0.407	-	-
ND1-Cu-NH2-C	-58.437	-	-

4.4.5. Protein stability

The stability of the proteins through the course of the simulation was monitored in order to assess the stability of all the input parameters of both the biased and unbiased MD experiments. The features that were monitored through the course of the simulation were the potential energy, root mean square fluctuation (RMSF), root mean square deviation (RMSD), radius of gyration and

bond lengths between cellulose and the Cu^{2+} cation. The movement and RMSF was monitored during the simulation and is shown in Figures 4.6, 4.7 and 4.8.

The Type 1 MD experiments showed a stabilization of the potential energy of the system for both biased and unbiased MD experiments as shown in Figure 4.14 A. For the RMSD of the Type 1 AA9 protein (4B5Q) (Figure 4.14 B), it was found not to reach convergence for both the biased and unbiased experiments. This was believed to be due to the overall movement of the protein across the cellulose substrate as demonstrated in Figure 4.6. Due to this movement, the convergence of the RMSD of the system for both the biased and unbiased experiments is unlikely to occur. As a result, a similar observation was made for radius of gyration (Figure 4.14 C). The gyration of both the biased and unbiased experiments is found to fluctuate through the simulation. Indicating that the Type 1 AA9 protein undergoes conformational changes during the MD simulations. However due to the stability observed with respect to the potential energy of the system both the biased and unbiased experiments were regarded as stable.

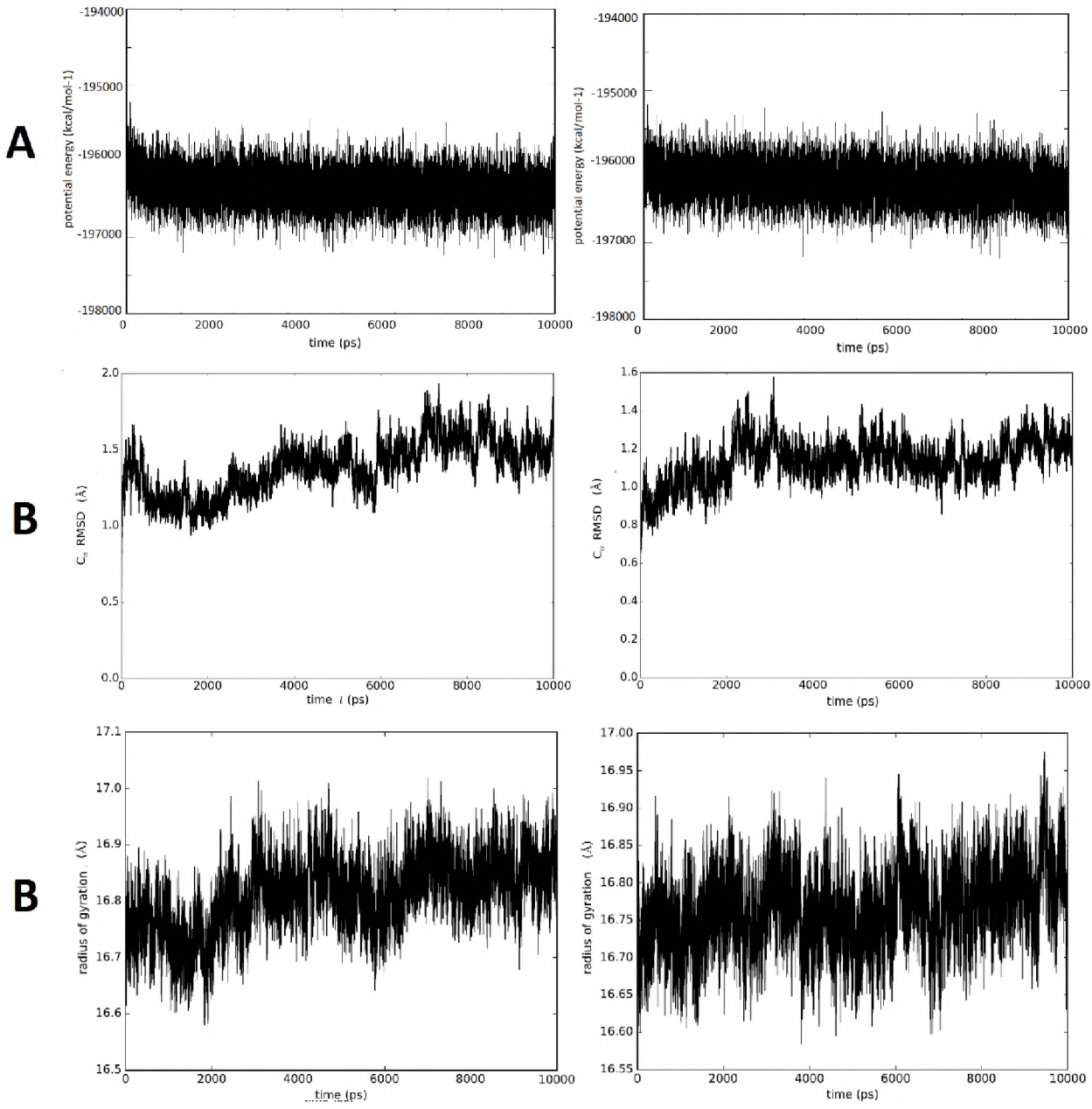

Figure 4.14: Protein stability measured by root mean square deviation and radius of gyration for Type 1 AA9 proteins (4B5Q). A) Potential energy of the Biased and Unbiased MD experiment, B) RMSD of the Biased and Unbiased MD experiment, C) Radius of gyration of the Biased MD experiment.

Similar to what was observed for Type 1 MD simulations, the Type 2 (4EIR) biased and unbiased MD experiments showed a similar trend as shown in Figure 4.15. The potential energy of the biased and unbiased MD experiments were shown to stabilize through the course of both simulations (Figure 4.15 A). However, similar to the Type 1 MD simulations, the Type 2 MD simulations failed to reach convergence with respect to RMSD (Figure 4.15 B). The radius of

gyration was found to show fluctuations for both biased and unbiased experiments indicating that the Type 2 4EIR crystal structure like the Type 1 4B5Q crystal structure also undergoes some form of conformational change during the 10 ns MD runs Figure 4.15 C.

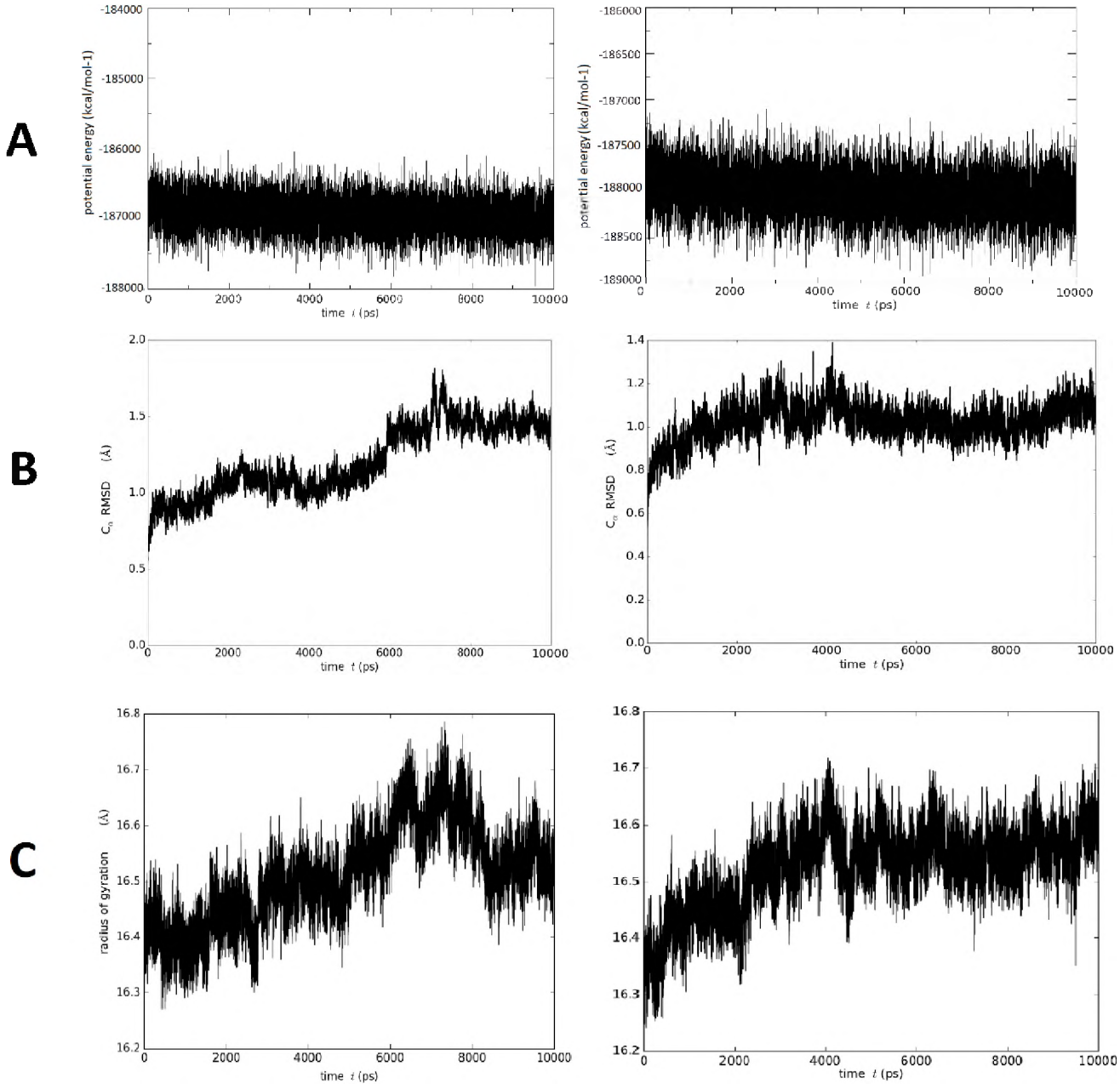

Figure 4.15: Protein stability measured by root mean square deviation and radius of gyration for Type 2 AA9 protein (4EIR). A) Potential energy of the Biased and Unbiased MD experiment, B) RMSD of the Biased and Unbiased MD experiment, C) Radius of gyration of the Biased MD experiment.

The stability of the Type 3 MD experiments was also monitored with respect to potential energy, RMSD and the radius of gyration. The result of this analysis can be found in Figure 4.16. For

Type 3 MD simulations (both the biased and unbiased experiments) the proteins were found to stabilize with respect to the potential energy of the system as shown in Figure 4.16 A.

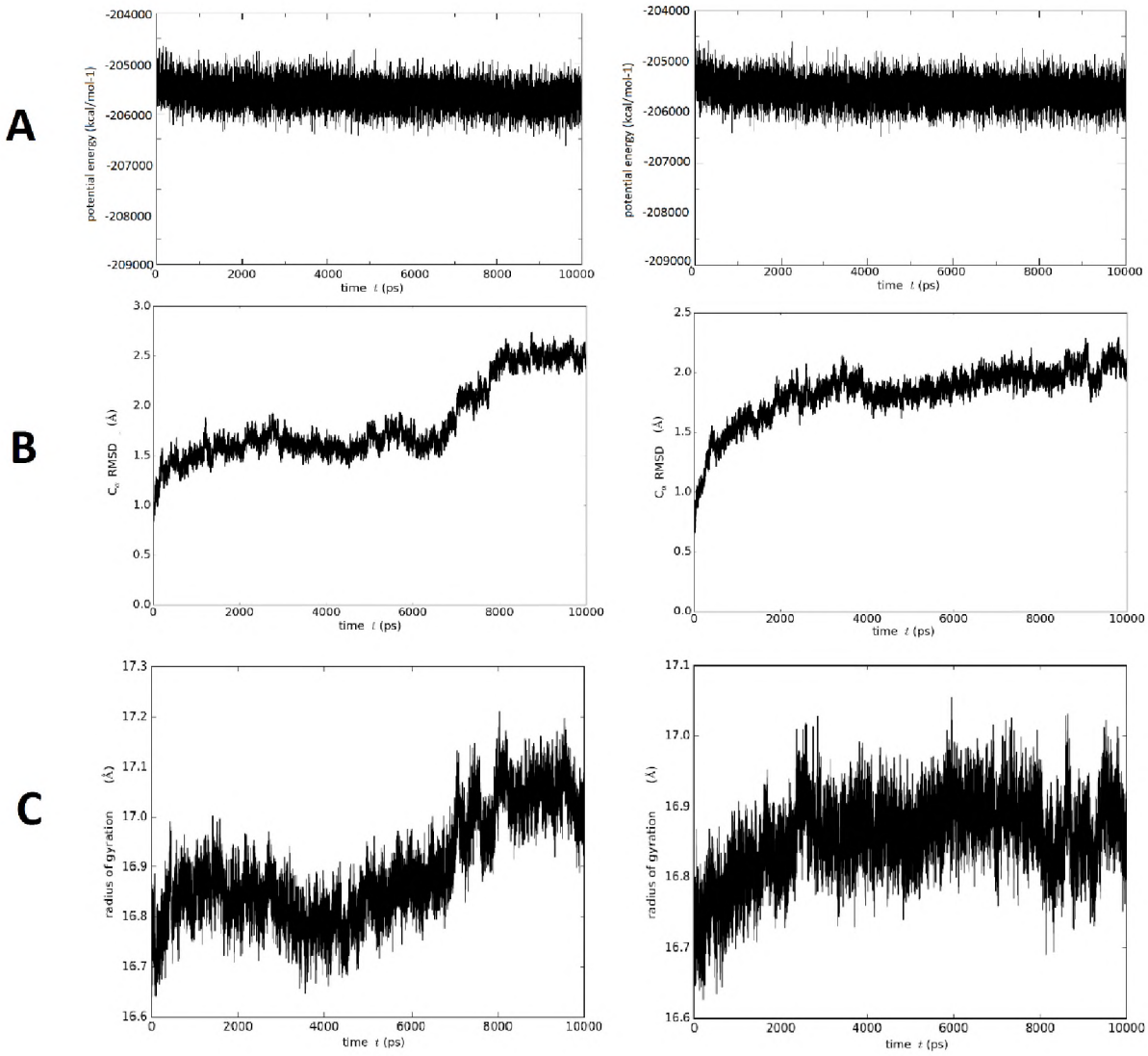

Figure 4.16: Protein stability measured by root mean square deviation and radius of gyration for Type 3 AA9 protein (3ZUD). A) Potential energy of the Biased and Unbiased MD experiment, B) RMSD of the Biased and Unbiased MD experiment, C) Radius of gyration of the Biased MD experiment.

Due to the short nature of the Type 3 (3ZUD) MD run stabilization of the RMSD was not yet fully achieved (Figure 4.16 B). The biased experiment showed failure to convergence while the unbiased experiment was nearing stability. Like the other AA9 types this failure to converge was attributed to the overall movement of the protein across the cellulose substrate. However increasing the length of the MD simulations may provide further insights into the stability of these proteins. Due to this movement, the convergence of the RMSD of the Type 3 MD simulations for

both the biased and unbiased experiments is unlikely to occur. As a result, a similar observation was made for radius of gyration (Figure 4.16 C). The radius gyration of both the biased and unbiased experiments is was also found to fluctuate in Type 3 MD experiments indicating possible conformational changes. The type 3 MD experiments were found to have stable potential energies of both biased and unbiased systems.

4.4.6. Type-specific contacts

To assess the unique Type-specific interactions that occur during the MD simulations for all three AA9 type, contact maps were generated. To create these contact maps, Script3.py was using. The Script3.py calculates the inter-atomic distances between the respective AA9 proteins and the top layer chains cellulose substrate. A distance cut-off is used to only capture the minimum distances which would be the contact points between the cellulose and the AA9 protein. The cellulose chains considered for calculation were the five top layer chains named M0, M3, M6, M9 and M12 as illustrated in Figure 4.16. The calculation was performed at five time intervals within the 10ns MD runs which were: 0ns, 2ns, 4ns, 6ns, 8ns and 10ns. These calculations were performed for both the biased and unbiased experiments for all three types. The findings of this analysis are shown in supplementary data Figure S3-8. For the Type 1 biased and unbiased experiment the contact maps shown in Figures S3 and S4 respectively. For the Type 2 biased and unbiased experiment the contact maps shown in Figures S5 and S6 respectively. For the Type 3 biased and unbiased experiment the contact maps shown in Figures S7 and S8 respectively.

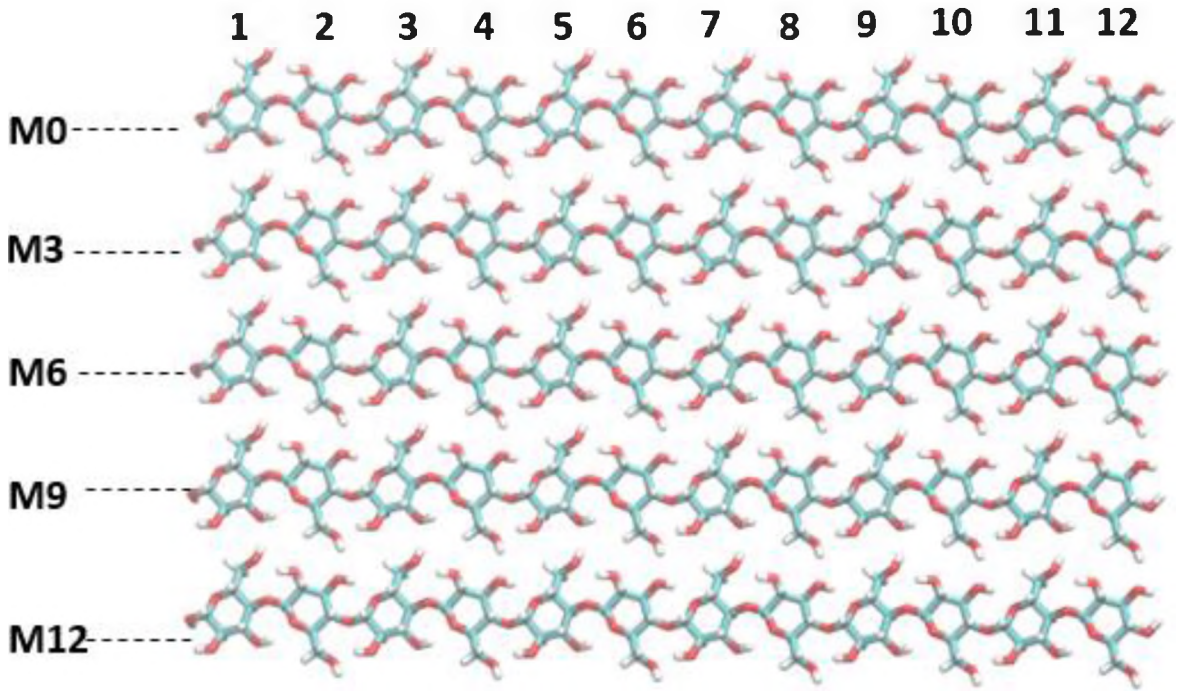

Figure 4.17: cellulose chains of the top layer of the cellulose substrate. Figure shows the five top layers of the cellulose substrate and the number sugar residues.

It was found that the regions of the protein that make contact with cellulose were similar for both unbiased and biased experiments were similar for all three types as shown. As a result, the structural mapping of the findings was only performed for unbiased experiment for all three types. The result of this analysis is shown in Figure 4.18, 4.19 and 4.20 for Type 1, 2 and 3 AA9 proteins respectively. A total of 8 contact regions was observed for AA9 proteins. Of these 8 contact regions, 4 were conserved in all AA9 proteins. The remaining contact regions were found to be type-specific. The result of this analysis is discussed in detail below.

4.4.6.1. Type 1 unique contacts

For Type 1 AA9 proteins 4 contact regions were identified these regions were named region 1, 2, 3 and 4 (Figure 4.18 A). Regions 1-4 were also found present on other AA9 types (Figure 4.19 A and 4.19 A). As expected the regions 1-4 are loops of the proposed flat binding face AA9 proteins. Regions 1-4 are colored in red in Figure 4.19. To assess the overall contributions of this regions to cellulose binding, the movement of this regions of the cellulose surface was monitored. This was achieved by taking 0.5ns snapshots of regions 1-4 from time 0ns to 10ns. The start positions of the

inserts is indicated by the solid lines in Figure 4.18 B while the overall movement is indicated by the transparent lines of the respective regions. It was found that the region 1 loop spans the both the M0 and M3 cellulose chain of the top layer of the cellulose substrate.

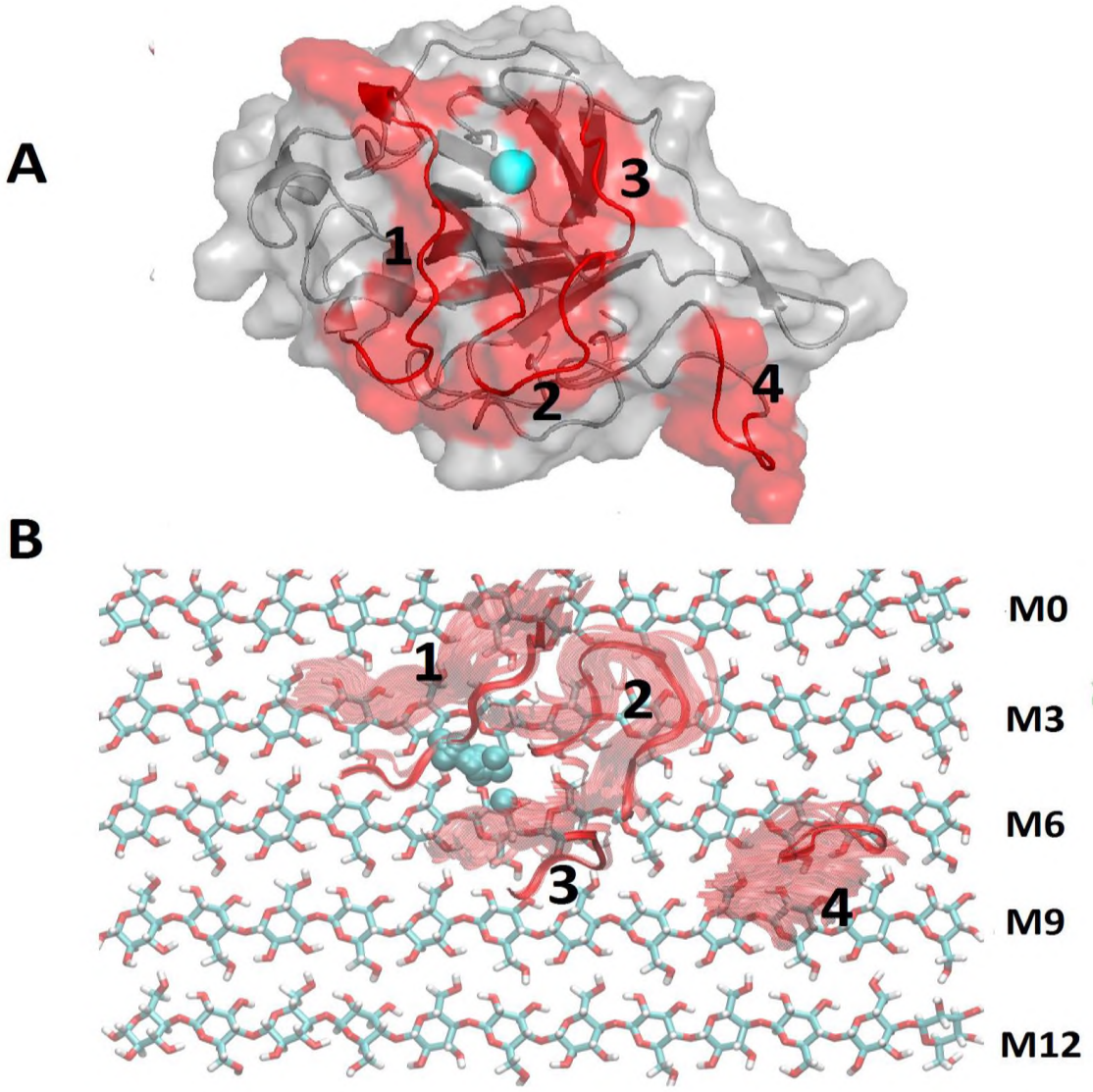

Figure 4.18: Type 1 AA9 – cellulose contact regions observed during MD simulation. Red denotes contact regions that are common in all AA9 types (1-4). A shows the localization of the unique contacts on the Type 1 4B5Q crystal structure. B shows how the contact regions interact with the top layer of cellulose.

Initially the region 1 loop is anchored on to the 7th glucose residue of the M0 chain and the 6th glucose residue of the M3 chain. As the simulation progresses the region 1 loop moves to the 6th glucose residue of the M0 chain and spans the 4th and 5th glucose residues of the M3 chain. The region 2 loop was initially found bound to the M3 cellulose chain on the 7th and 8th glucose

residues. Through the course of the simulation the region 2 loop remains bound to the original positions however movement to 6th glucose residue of the M3 chain is observed. The region 3 loop is original found anchored between the M6 and M9 chains. This region the moves to bind exclusively to the 6th and 7th glucose residues of the M6 cellulose chains. The region 4 loop is found originally anchored to the 10th and 11th glucose residues of the M6 chain and is later moved to the 9th and 10th glucose residue of the M9 chain. The overall movement of the regions appears to result in the overall displacement of the Cu^{2+} from the 6th glucose of the M6 chain to the 5th glucose of the M3 chain. This displacement has previously been demonstrated in Figure 4.11 and 4.12.

4.4.6.2. Type 2 unique contacts

On top of the region 1-4 loops, Type 2 AA9 proteins had 2 additional regions (Figure 4.19 A). These additional regions named region 5 and 6. The regions 1-4 are colored in red and the regions 5 and 6 are colored in yellow. Similar to what was observed for the type 1 MD simulation, the regions 1-4 are loops are part of the proposed flat binding face AA9 proteins. It was seen that in the Type 2 protein, the region 1 loop is initially anchored on the M3 cellulose chain on the 6th glucose residue. As the simulation progresses, the region 1 loop moves to the 6th glucose residue of the M3 chain to the 5th glucose residues of the M3 chain. The region 2 loop was initially found bound to the M3 and M6 cellulose chains. Initially the region 2 loop was bound to the on the 8th and 9th glucose residues of the M3 chain however, is later displaced by one glucose residue. On the M6 chain the region 2 insert is fluctuates between the 6th and 7th glucose residue. Through the course of the simulation the region 2 loop remains bound to the original positions similar to what was observed for the Type 1 MD experiments. The region 3 loop is found originally bound between the M6 and M9 cellulose chains and through the course of the simulation there is no significant displacement of the region 3 loop that is observed. The region 4 loop is original anchored between the on the 6th glucose residue of the M6 chain and is anchored between the M6 and M9 chains. The region 4 loop then is displaced across to the 9th and 10th glucose residues of the M6 cellulose chains. Two type-specific regions were identified for Type 2 MD experiment (4EIR). The Type 2 specific region 5 loop was initially bound to the between the M0 and M3 cellulose chains. Region 5 is later displaced in the simulation and binds to the 3rd glucose residue of the M3 chain. The other Type 2 specific Region 6 was initially found spanning the M6, M9 and M12 cellulose chains.

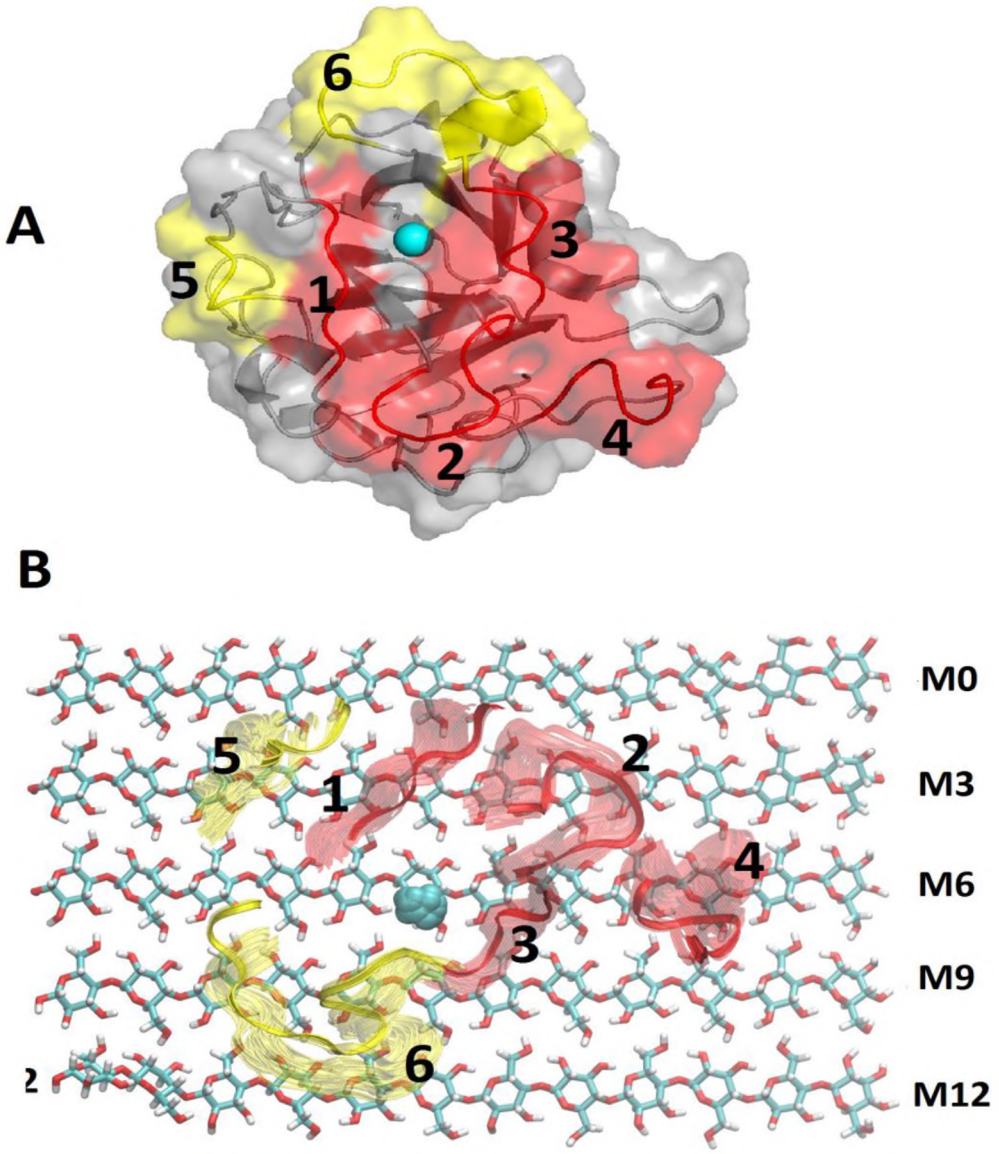

Figure 4.19: Type 2 AA9 – cellulose contact regions observed during MD simulation. Red denotes contact regions that are common in all AA9 types (1-4) and yellow denotes contact regions that are specific to Type 2 AA9 proteins (4EIR) (5 and 6). A shows the localization of the unique contacts on the Type 2 4EIR crystal structure. B shows how the contact regions interact with the top layer of cellulose.

The region 6 is then displaced through the course of the MD simulation to M9 and M12 chains were it binds on the 3, 4, 5 and 6th glucose residues of the M9 chain and the 4, 5 and 6th of the M12 chain. The overall movement of the regions also result in displacement of the Cu^{2+} (Blue sphere Figure 4.19 B), however, the displacement was not as profound as what was observed for the Type 1 MD experiment.

4.4.6.3. Type 3 unique contact regions

For the Type 3 MD experiments 6 contact regions were identified. The regions were found to consist of the common regions 1, 2, 3 and 4 (Figure 4.20 A) and additional Type 3 specific regions 7 and 8. Regions 1-4 are shown in red while the regions 7 and 8 are shown in green (Figure 4.20). Similar to what was observed in the other AA9 LPMO types the regions 1-4 loops were found to constitute the proposed flat binding face Type 3 AA9 proteins. It was observed that in the Type 3 MD experiments the region 1 loop is initially found spanning both the M0 and M3 cellulose chain. The region 1 loop is original bound to the 7^{th} residue of the M0 chain and on the 4^{th} of the M3 chain. Through the course of the simulation, the region 1 loop is displaced to the M0 chain and spans the 4, 5 and 6^{th} glucose residues of this chain. The region 2 loop was originally found bound to the 6, 7 and 8^{th} glucose residues of the M3 cellulose chain. As the simulation progresses, the region 2 loop remains was found to be displaced by one glucose residue relative to the M3 chain and additional binding to the 8^{th} glucose residue of the M0 chain was observed. The region 3 loop is found bound between the M6 and M9 chains. The region 2 loop is then displaced to the 6th and 7^{th} glucose residue of the M9 cellulose chain. The region 4 loop is found initially bound to the 8^{th} glucose residue of the M6 chain and is displaced to between the M6 and M9 cellulose chains. Like the Type 2 MD experiments, two type-specific regions were identified for Type 3 MD experiment (3ZUD). The Type 3 specific region 7 loop was originally bound to 8 and 9^{th} glucose residue of the M9 cellulose chain. Region 7 then displaced during the simulation and binds to the 8^{th} glucose residue of the M12 chain. The second Type 3 specific region was the region 8 loop which was initially bound between the M0 and M3 cellulose chains.

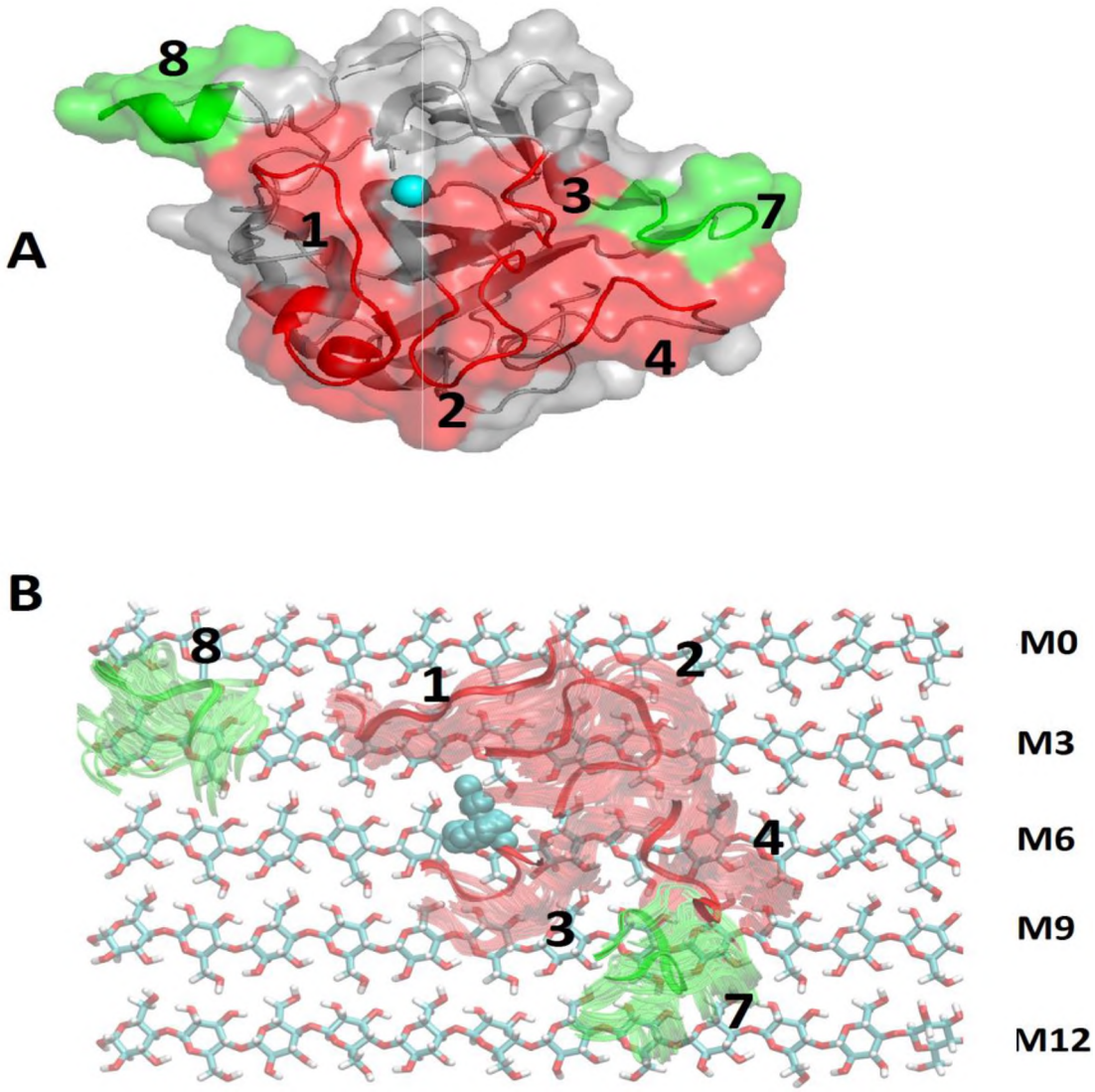

Figure 4.20: Type 3 AA9 – cellulose contact regions observed during MD simulation. Red denotes contact regions that are common in all AA9 types (1-4), and green denotes contact regions that are specific to Type 3 AA9 proteins (3ZUD) (7 and 8). A shows the localization of the unique contacts on the Type 3 3ZUD crystal structure. B shows how the contact regions interact with the top layer of cellulose.

The region 8 is later displaced through the course of the MD simulation to the 1st and 2nd glucose residues of the M3 chain. The overall movement of the regions also result in displacement of the Cu^{2+} (Blue sphere Figure 4.19 B) from between the M3 and M6 chain to the 6th glucose reside of the M6 cellulose chain. Unlike Type 1 AA9 proteins, no binding to cellulose was observed the Cu^{2+} of Type AA9 LPMOs.

Type-specific contact regions were identified using MD simulations. These contact regions were then correlated with the motif analysis results from Chapter 2. This analysis revealed an overlap

between the contact regions and the conserved motifs. It was found that the region 1 contact corresponds to the Motif 6 of Type 3 proteins and Motif 9 of both Type 2 and 3 proteins. The region 2 contact was found to correspond with Motif 2 which conserved in all AA9 types. The region 3 contact was found to correspond with the conserved Motif 4. Region 4 was found to correspond with Motif 3 that was also well conserved in all AA9 proteins. The Type 2 specific region 6 was found to correspond with insert II of Type 2 proteins. The region 5 contact was not associated with any motifs. The type-specific region 7 contact was found to correspond to the insert I that is characteristic of Type 3 AA9 proteins. The region 8 contact, similar to what was observed for region 5, was not associated with any motifs. This was attributed to the presence of regions in AA9 protein segments with low conservation.

4.4.7. Hydrogen bond analysis

The effect of hydrogen bonding on AA9 cellulose interaction during MD simulations was assessed. The analysis was performed on both biased and unbiased MD simulations of Type 1, 2 and 3 LPMO types of AA9 proteins. The findings of the analysis are shown in Figure 4.21.

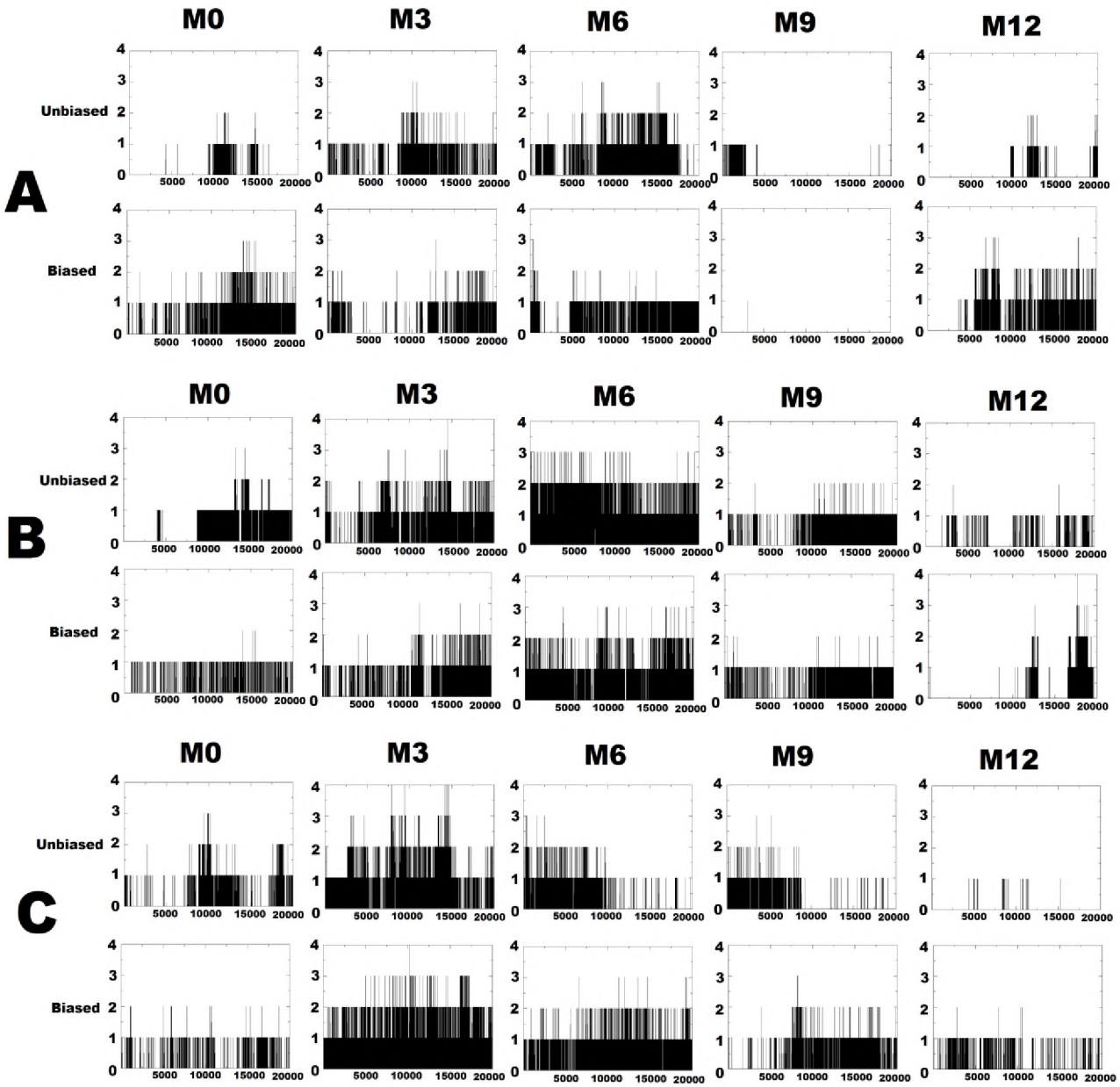

Figure 4.21: Hydrogen bonding analysis of both biased and unbiased experiments of AA9 LPMO types. The X-axis of the subplots denotes the number of hydrogen bonds on a particular cellulose chain and the Y-axis denotes the particular frame of the simulation. A) Hydrogen analysis performed the biased and unbiased Type 1 MD experiments, B) Hydrogen analysis performed the biased and unbiased Type 2 MD experiments and C) Hydrogen analysis performed the biased and unbiased Type 3 MD experiments.

VMD was used to identify hydrogen bonds that occur between the top layer of the cellulose substrate and the AA9 proteins. The criteria to define a hydrogen bond was defined as; a Donor-Acceptor distance of 3 Å, an angle cut-off of 20 degrees. The search was applied to all frames of the simulations. Hydrogen bonding analysis was performed the contribution of hydrogen bonding

to substrate binding and orientation. The hydrogen bonding analysis revealed that AA9 protein show the highest binding affinity to binding the central M3 and M6 cellulose chains. This is indicated by the large number of hydrogen bonds that were observed in these chains for all LPMO types as shown in Figure 4.21. The hydrogen bonding observed is a result of the regions 1-4 loops which were found to be common in all AA9 types (as shown in Figures 4.18, 4.19 and 4.20). Regions 1-4 were found to consistently interact with the M3 and M6 cellulose chain. As a result, there is prominent hydrogen bonding observed on both these chains. It is also important to consider that the 10 ns MD runs performed may have not been long enough to observe the overall movement of the AA9 protein on cellulose. Due to this, it is possible that the AA9 protein remains anchored to these chains because there was insufficient simulation time to observe binding to other cellulose chains. It therefore important for future studies to perform longer simulations to better understand the dynamics between cellulose and AA9 proteins. The hydrogen bonding analysis of Type 1 proteins (Figure 4.21 A) showed that there is minimal hydrogen bonding between the 4B5Q the M9 cellulose due to the minimal hydrogen bonding observed. This was attributed to the lack of interacting regions with this chain as seen in Figure 4.18 B. Unlike the Type 1 MD experiments, both the Type 2 and 3 MD experiments revealed the presence of hydrogen bonding to the M9 chain (Figure 4.21 A and 4.21 B). The presence of hydrogen bonding is attributed to the presence of type-specific regions in both Type 2 and 3 AA9 proteins. In Figure 4.18 B Type 2 proteins have region 6 loop which interacts with the M9 chain. In Figure 4.19 B Type 3 proteins have region 7 loop which interacts with the M9 chain. These type-specific contacts result in a more prominent hydrogen bonding with the M9 chain of cellulose substrate for Type 2 and 3 AA9 LPMO types. The region 1 loop of all AA9 LPMO types was also found to interact with the M0 chain in all the MD experiments as evidenced by the hydrogen bonding in Figure 4.21. Hydrogen bonding was also observed on the M12 chain however this was not type-specific as this was attributed to the dissociation of the cellulose substrate during the MD simulation. Due to the fact that the top layer of the cellulose substrate was unrestrained, dissociation of the M12 cellulose chain was often observed in the MD experiments. This dissociation resulted in the observed hydrogen bonding. To avoid this disassociation of the M12 cellulose chain the cellulose substrate may have been constructed such that the restraints were also applied to the top layer of cellulose. However, the application of restraints to the top layer of cellulose would have prevented the ability of the AA9 protein to freely interact with the top layer of cellulose.

4.5. Discussion

In order to validate the newly determined force field parameters as well as assess the overall type-specific interaction between AA9 LPMO types and their substrate cellulose MD simulations were performed. The bond stretch, angle bend, and torsions which were obtained from PES scans and least squares fitting were inserted into the CHARMM 36 force field for validation using MD simulations. However, in order to perform MD simulations a Lennard-Jones parameter set had to be selected in order to describe the behaviour of the Cu^{2+} in the TIP3P water environment. Estimation of the Lennard-Jones proved to be an issue as it is beyond the scope of this study to separate the bonded component from the QM calculation. This resulted in an unusable Lennard Jones parameter set as shown in Table 3.3. To overcome this, a literature search was performed in order to identify a suitable Lennard-Jones parameters for the Cu^{2+} metal center. The literature search yielded two Lennard-Jones parameter sets to be used in the MD experiments. Both the Lennard-Jones parameter sets were used for MD this resulted in biased and unbiased MD experiments. As a result for all three AA9 LPMO types two MD experiments were performed. One experiment would use the Lennard-Jones parameter set evaluated in water (unbiased) and the second experiment would use the Lennard-Jones parameter set which was evaluated in acetonitrile (biased). Due to the relatively high ε (energy well) of the biased Lennard-Jones parameter set, binding to cellulose was expected to occur sooner as opposed to the more moderate unbiased Lennard-Jones parameter set.

After insertion of force field parameters to the CHARMM 36 force field, MD simulations were performed. No additional parameterisation was necessary for the cellulose substrate because the CHARMM 36 force field also covers carbohydrates. As a result, all MD simulations were performed on a three layered cellulose substrate using coordinates supplied in ref(Nishiyama, Langan & Chanzy 2002). To ensure accurate representation of the cellulose within the simulation, the bottom layer of the cellulose substrate was fixed on the C1 and C4 carbons of the cellulose chains. This was done to create infinite cellulose models where the top two cellulose layers were held in position by hydrogen bonding. The force applied on the C1 and C4 carbons was 5 $kcal/mol^{-1}$. Other considerations prior to performing MD simulations were the presence of modified residues on AA9 crystal structures and the presence of disulphide linkages. The 4EIR and the 3ZUD crystal structures (Type 2 and 3 respectively) are methylated on the N-terminal His-1 residue. According

to literature, the methylation serves no known function due to the fact that removing the methylation has no effect on the activity of AA9 proteins. Instead of parametising a new histidine residue for the CHARMM 36 force field, the methylation was removed from the His-1 residues of the 4EIR and 3ZUD crystal structures. Removing the methylation resulted in His1 residues that were compatible with the CHARMM 36 force field. The presence of disulphide linkages in all three AA9 crystal structures (4B5Q, 4EIR, 3ZUD) was handled by specifying the location of the respective disulphide linkages using the SSBOND.

To speed up calculations, MD simulations were performed at the Center of High Performance Computing (CHPC). In order to determine the most optimal resources to perform MD simulations short MD runs were prepared. The tests were performed on 2, 4, 8, 16, 32, 64 and 128 cores using 10000 steps of simulation using a 0.001 (1 fs) timestep. It was found that the best resources to be used were 64 cores. There was no effect observed for changing the number of cores on performance. In cases were there slight differences observed internode communication or latent processes were attributed to that.

The Type 1 force field parameters were successfully determined and evaluated in all three AA9 LPMO types. The comparison of the observed average parameters and the original crystal structure values is shown in Table 4.4. It was determined that the parameters are generally able to maintain their equilibrium positions with slight deviations. The only exception to this was the Tyr-160 residue which showed the greatest degree of fluctuation in all LPMO type. The cause of the fluctuation was the low force constant of 3.0 kcal/mol-1 that was obtained from QM calculations. Interestingly, residues which were not previously considered in the calculation were found to coordinate the Cu^{2+} in all three AA9 types. This is best demonstrated by the Glu-157 residue of the Type 1 crystal structure 4B5Q. The Glu-157 residue was found to bind in both biased and unbiased MD experiments. However, it is important to note that in the biased experiments the Glu-157 residue is more tightly bound. Due to the fact that the Cu^{2+} active site of AA9 proteins has two free coordination positions, the Glu-157 residue, may be important for stabilizing the coordination geometry of the active site prior to substrate binding.

A total of six MD runs was generated for the three AA9 LPMO types (Two MD runs for each type). Upon completion of the MD runs the trajectories were analysed for various features which include; the movement of the AA9 protein relative to the cellulose substrate, local fluctuations of

the protein, protein stability as measured by RMSD, radius of Gyration and Potential energy. The coordination of the Cu^{2+} during the simulation was also monitored to ensure stability of the force field parameters and observe any binding that may occur. Contact maps between respective AA9 proteins were created to motor regions that are crucial for type-specific interactions during MD simulations. To assess the overall movement of AA9 structures relative to the cellulose substrate VMD was used. For each LPMO type (both biased and unbiased MD experiments), the trajectories were visualised and snaps shots of the protein including the Cu^{2+} were taken at every 0.5 ns interval to observe the overall movement of AA9 proteins relative to the cellulose substrate. The RMSF was measured to identify regions on the protein structure that are likely contributors to the observed motions of AA9 proteins. It was found that AA9 proteins generally have a directional motion on the surface of cellulose. This was demonstrated for all three AA9 types. Analysis of the RMSF, revealed the major contributors to protein movement were the loop regions. The loop regions of all AA9 LPMO types were found to possess relatively higher RMSF values as opposed to secondary structural elements. In general the RMSF values for AA9 types were high, especially for the Type 1 proteins. This was attributed to the displacement of the AA9 proteins across the cellulose substrate resulting in the relatively large RMSF values observed. Through the DSSP analysis the AA9 LPMO types were found to be structurally stable through the course of the simulation. This observation was based on the fact that AA9 proteins in all three LPMO types appeared to retain their β-sandwich fold through the course of the simulation. The elements that showed the highest instability were the 3-10 helices of AA9 proteins. Even though the function of 3-10 helices on AA9 structure is unknown, their presence in all three AA9 protein structures requires further investigation. The destabilization of 3-10 helices through the course of the simulation may be attributed to their overall motion relative to the cellulose substrate.

The stability of AA9 proteins through the course of the simulation was assessed by measuring by RMSD, radius of gyration and the Potential energy of the resulting trajectories. For all three types no convergence with respect to RMSD observed. For all three types changes greater than 0.5 Å for RMSD were observed. This was to be expected as the proteins were determined to have movement relative to the cellulose. As result the RMSD was regarded as not a suitable gauge for protein stability. The radius of gyration was also measured to assess any observed changes in compactness during the simulation. It was found that for all three LPMO types the radius of gyration showed a fluctuations. The fluctuations observed were indicative of conformational changes that occur as

the protein moves across the cellulose substrate. Due to these observations the Potential energy was calculated. The potential energies for all AA9 LPMO MD experiments was shown to be stable throughout. As a result, the potential energy was regarded as the best measure for assessing protein stability.

It was observed that through the course of the simulation, the flat surface active site of all AA9 LPMO types remains oriented towards the surface exposed layer of cellulose. This orientation provide an opportunity for the Cu^{2+} ion to potentially bind cellulose. As a result, binding to cellulose was observed for the Cu^{2+} of the Type 1 4B5Q crystal structure to cellulose. The binding to cellulose was shown to occur in both biased and unbiased Type 1 MD experiments. Even though binding was observed for both the biased and unbiased experiments of the Type 1 MD study, binding was shown to occur much sooner for the biased experiment. This was attributed to the different Lennard-Jones parameter sets used in the experiments. The biased Lennard-Jones parameter set also allowed for the detection of possible bound cellulose chain exchange between. Within the biased experiment experiments it was found the initially bound hydroxyl (OH) from one cellulose chain may be exchanged for another in a different cellulose chain. This observation suggests possible processivity possessed by AA9 proteins in the other LPMO types (Type 2 and 3), binding to cellulose was not observed. The absence of binding in Type 2 and Type 3 AA9 proteins was attributed to the presence of type-specific inserts of the flat surface active site of AA9 proteins. The presence of the Zone I insert in Type 3 AA9 proteins and the Zone II insert in Type 2 AA9 proteins (as discussed in in Chapter 2) is likely to result in steric hindrances that result in the inability to bind cellulose. As postulated in previous studies (Hemsworth, Davies & Walton 2013), the steric congestion of the AA9 active site may be the reason why AA9 proteins display such different regioselectivities. To further understand the effect of the steric congestions on the AA9 cellulose interaction, contact maps were generated using the biased MD trajectories for all three types. A total of eight contact regions were identified these regions were termed regions 1-8. Regions 1-4 were found in all AA9 proteins while regions 5 and 6 were specific to Type 2 AA9 proteins and regions 7 and 8 were found to be specific to Type 3 AA9 proteins. It was seen that the type-specific regions have potentially crucial role in substrate interaction due to the fact that in both Type 2 and 3 AA9 proteins they provide unique contacts there are absent in the other LPMO types. As a result, it can be said that all three AA9 LPMO types have unique binding modes as determined by MD studies. Interestingly these contact regions were also found to have strong

overlap with the previous sequence analysis and manual docking findings that are detailed in Chapter 2. The common regions (1-4), were found to correspond to the conserved motifs of AA9 proteins. The conservation of these regions is indicative of conserved interaction in AA9 proteins. However, the presence of type-specific contacts on the AA9 active site surface implicates these regions in AA9 regioselectivity. In particular the type-specific inserts I and II, which were shown to correspond to the region 7 (Type 3) and region 6 (Type 2) respectively. Hydrogen bonding analysis showed that AA9 LPMO types show the highest binding affinity to the central M3 and M6 cellulose chains. The high affinity to the M3 and M6 cellulose chains is attributed to the regions 1-4 loops. As previously discussed, the regions 1-4 loops are common in all AA9 LPMO types. These regions were found to make contact with the cellulose substrate suggesting that there is a common interaction that occurs among AA9 LPMO types. Hydrogen bonding analysis was also able to reveal interaction between the type-specific regions and cellulose chains.

Chapter 5

Conclusions and Future Work

5. Conclusions

5.1. Sequence and structural analysis

The characterizations of LPMO type-specific sequence and structural features of AA9 protein was achieved by identifying a dataset of AA9 proteins and performing sequence based bioinformatics analysis. In order to perform a type-specific analysis *N. crassa* protein sequences were added to the dataset. Sequences were aligned and redundant and short fragment sequences were removed resulting in the final dataset of 153 protein sequences. The resulting dataset was phylogenetically clustered into their respective LPMO types in order to perform type-specific analyses. Once sequences were grouped into their respective types, all vs all sequence analysis were performed which revealed the sequence identity inherent in these proteins. Even with the diversity displayed by AA9 LPMO types, unique sequence type-specific features were observed. The unique sequence features of AA9 LPMO types was found to have structural consequences to potential cellulose binding. It was shown through manual dockings the type-specific features cause a unique LPMO specific AA9-cellulose interaction. This resulted in the identification of regions that may have the potential to interact with cellulose. This prompted the need to develop an MD protocol to understand the dynamics of the AA9-cellulose interaction.

5.2. Force field parameter determination

The semi empirical PM6 method and least squares fitting was used to determine the force field parameters of the Type 1 Cu^{2+} AA9 active site. The parameters evaluated were for bond stretch, angle bend and torsions. Initially the Lennard-Jones parameters were estimated, however due to the complexity of subtracting the bonded component from the QM calculation this was abandoned and literature values were used. To save on the computational cost associated with performing these kind of calculations, a subset of active site residues was used. The representative Type 1 crystal structure 4B5Q was used for all subsequent calculations. The calculations were performed

on the Type 1 AA9 active site however, because of the similarity in active site residues of the LPMO types, the force field parameters may be readily applied to other AA9 types.

5.3. Force field parameter validation and MD studies

The newly determined force field parameters were then validated using MD simulations. The MD simulations were performed on all three AA9 types using the 4B5Q, 4EIR and 3ZUD crystal structures to represent Type 1, 2 and 3 AA9 proteins respectively. The parameters were found to be adequate for maintaining the geometry of the Cu^{2+} active site for all three LPMO types during MD simulation. The two Lennard-Jones parameter sets allowed for the use of a biasing parameter as well as a more moderate parameter for the MD simulations. The force field parameters were sufficient to describe potential for bonding between the Type 1 AA9 Cu^{2+} and the cellulose substrate. The binding observed during simulations could be indicative of the first stages of cellulose binding and subsequent degradation. The use of a biased Lennard-Jones parameter allowed for the more rapid cellulose binding and chain exchange. The ability of the AA9 proteins to exchange is indicative of proccessivity of AA9 proteins which is in line with their proposed method of enzymatic cleavage. The lack of binding for Type 2 and 3 AA9 proteins is a further confirmation of the hypothesis put forth by one study (Hemsworth, Davies & Walton 2013). Is suggested in the study that the likely contributor to regioselectivity of AA9 proteins is arrangement of the active site. In this study, the bioinformatics characterization of sequence and structural features specific to LPMO types was performed. Through molecular dynamics simulations it was shown that the different arrangements of the AA9 LPMO active sites results in different binding. These binding modes were found to be characterised by the presence of type-specific contact position between AA9 types and cellulose. The effect of the different binding regions was quantified by hydrogen bonding and a type-specific hydrogen bonding network was observed between cellulose and AA9 LPMO types. Displacement across the cellulose substrate was seen for the AA9 types. The displacement observed in AA9 proteins was attributed to the overall fluctuation of the loop regions. As a result, the loops of AA9 proteins may serve a dual function of substrate binding as well displacement across the substrate.

5.4. Future work

There are many interesting features of AA9 proteins that were not covered by this work. Future work may include taking into account reactive dioxygen species proposed to be responsible for cleavage. This can be investigated through design of MD simulations that include a paramatized oxygen molecule in the system to observe the resulting geometry in a dynamic environment. Alternatively, a combinatorial Quantum Mechanical and Molecular Mechanics (QM/MM) ONIOM approach may be used to investigate the cleavage mechanism of AA9 proteins. As previous studies (Kim et al. 2014) have attempted to show, oxidative cleavage of cellulose is likely facilitated by AA9 proteins. As such QM/MM calculations will have to be designed to show the type-specific cleavage of the C1 and C4 carbon of glucose residue of the cellulose substrate.

Supplementary Data

Chapter 2 Motif analysis

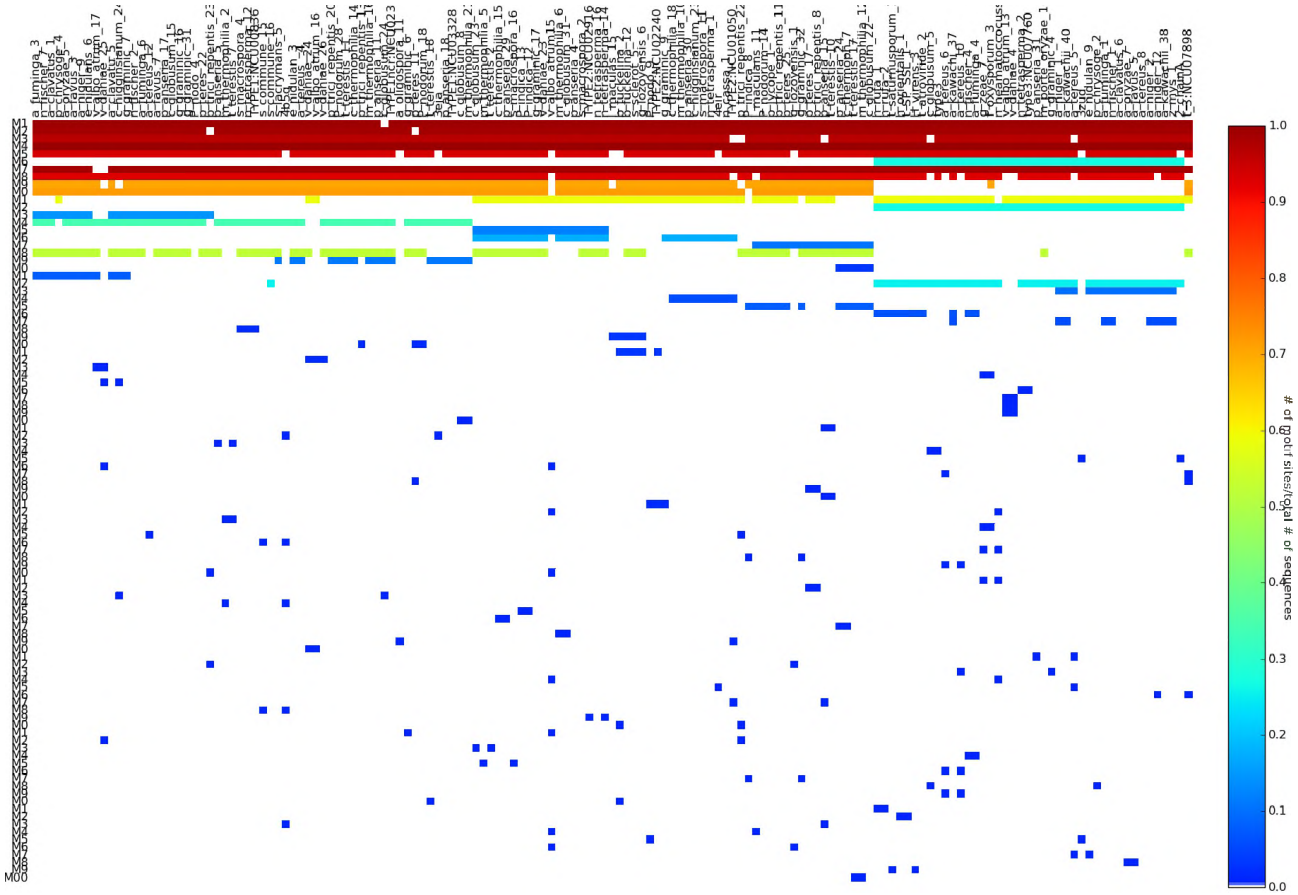

Figure S1: Motif analysis of AA9 domains. Heat representing the extant of conservation of the identified 30 motifs on AA9 sequences. Sequences are grouped according to Type starting from type 1-3. Red color indicates highly conserved motifs while lighter to blue colors represent low conservation.

Chapter 2 Model evaluation

Models for the *aspergillus niger* homolog 9 were evaluated with RAMPAGE and MetaMQAPII (Figure S2). The best model was the *aspergillus niger* homolog 9 19 homology model Figure S2 B.

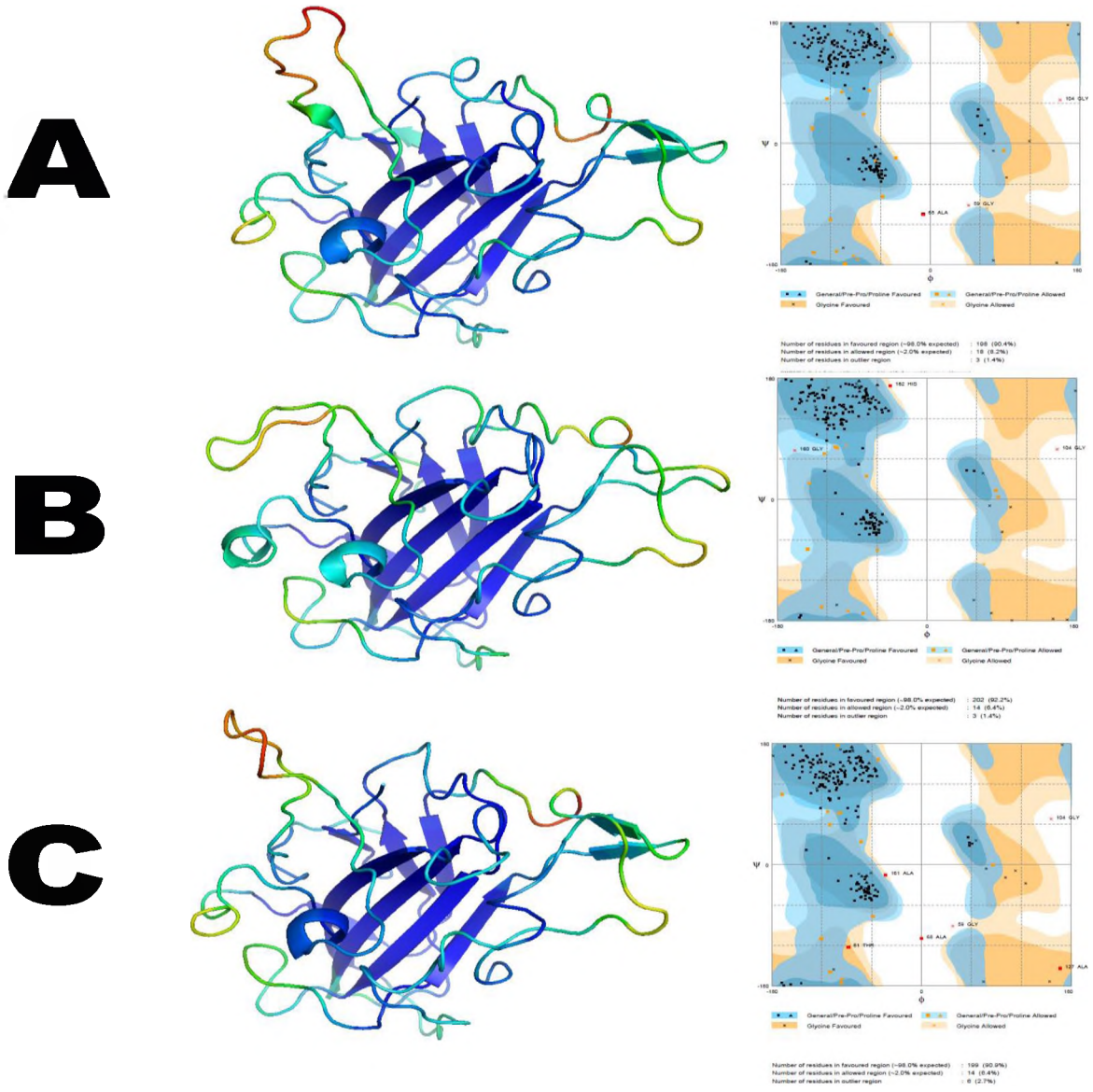

Figure S2: Model validation of the *aspergillus Niger* homolog 9 using MetaMQAPII and RAMPAGE. The top 3 of the 100 models generated models are shown A) is the *aspergillus niger* 08 structure, B), the *aspergillus niger* 19 structure and C) the *aspergillus niger* 23 structure.

Chapter 4 contact maps and unique regions Figure S3-11

Contact maps to identify type-specific regions responsible for cellulose interaction were generated using Script3.py. The contact maps were generated for the biased and unbiased trajectories for all three AA9 LPMO types. For the biased and unbiased Type 1 MD experiments the results are shown in Figure S3 and S4 respectively. For the biased and unbiased Type 2 MD experiments the results are shown in Figure S5and S6 respectively. For the biased and unbiased Type 3 MD experiments the results are shown in Figure S7 and S8 respectively. The unique contacts that were identified were mapped onto structures. It was seen that the contacts for both biased and unbiased experiments were similar. As result, only the unbiased contacts were shown in chapter 4. The results for unbiased experiments are summarised Figure S9, S10 and S11 for Type 1, 2 and LPMO types.

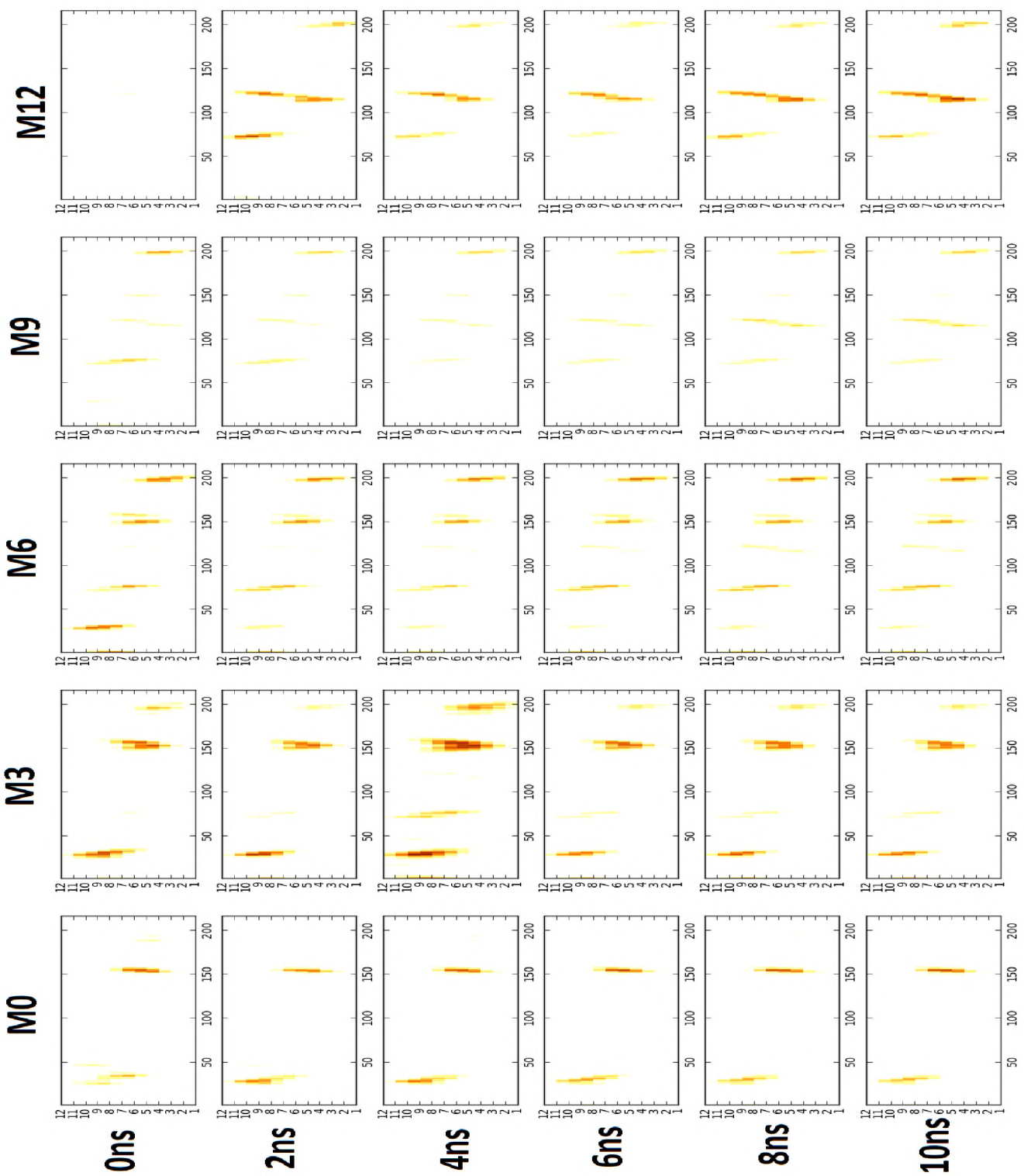

Figure S3. Contact map for the biased Type 1 10 ns MD simulation.

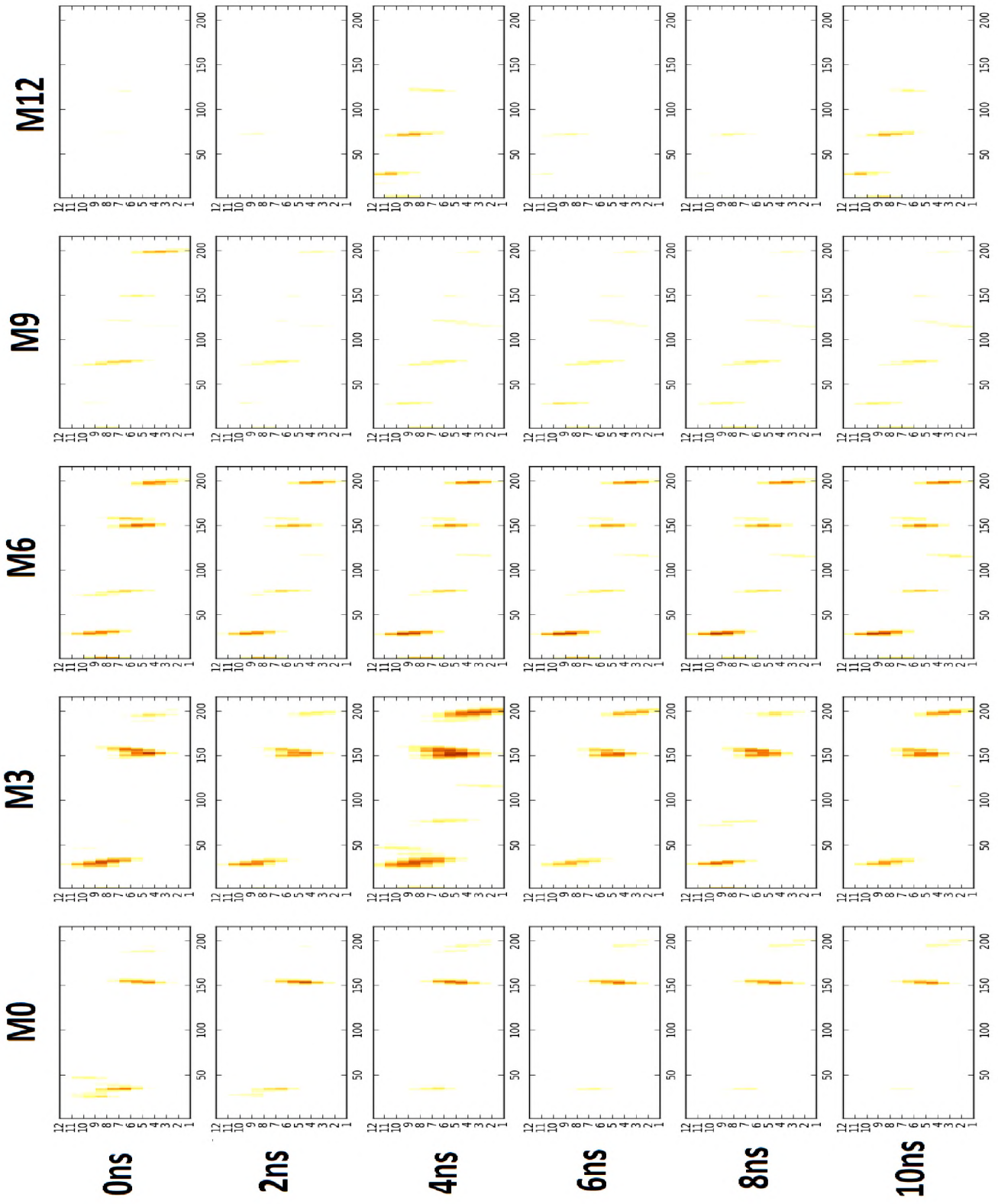

Figure S4: Contact map for the unbiased Type 1 10 ns MD simulation.

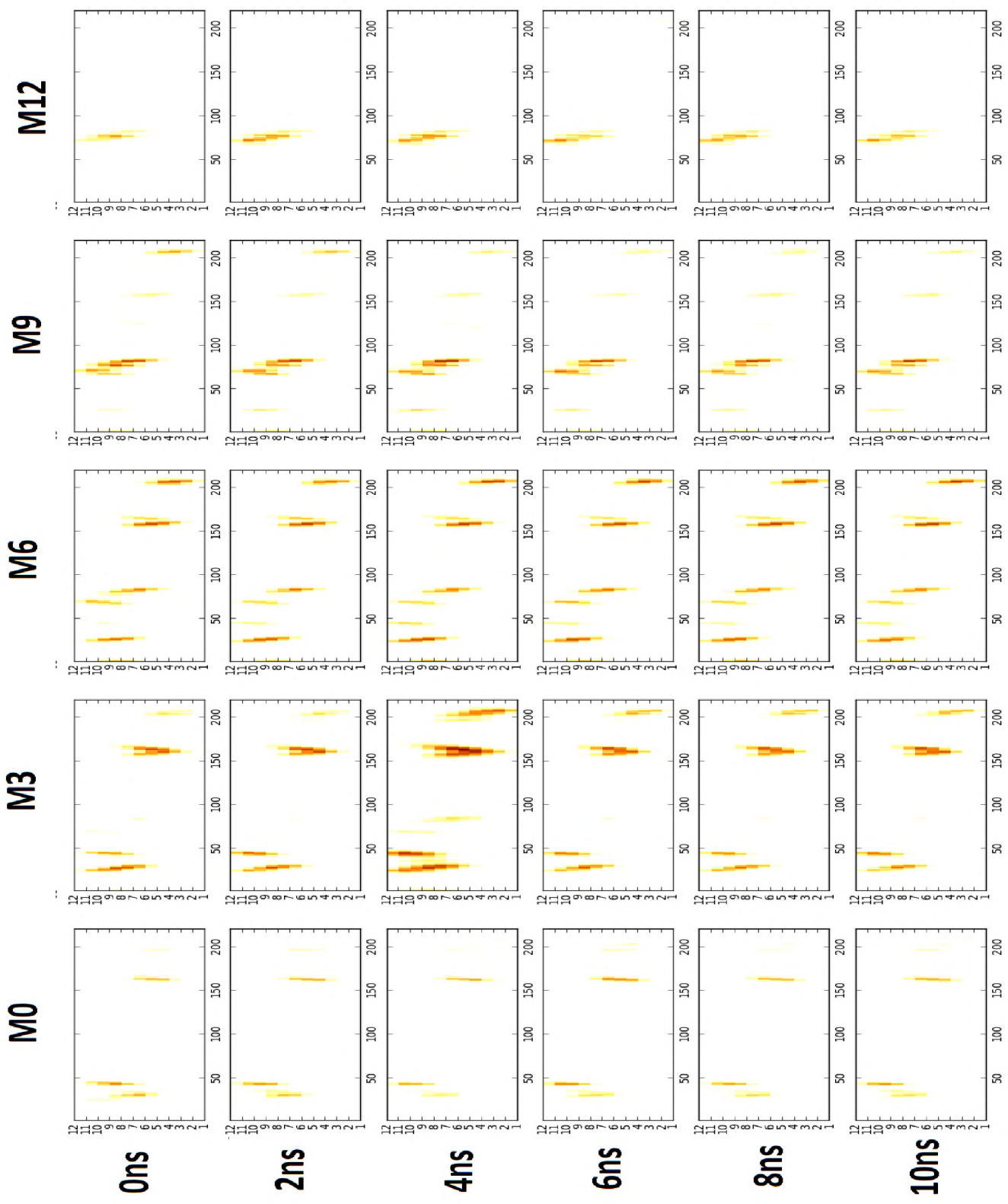

Figure S5: Contact map for the biased Type 2 10 ns MD simulation.

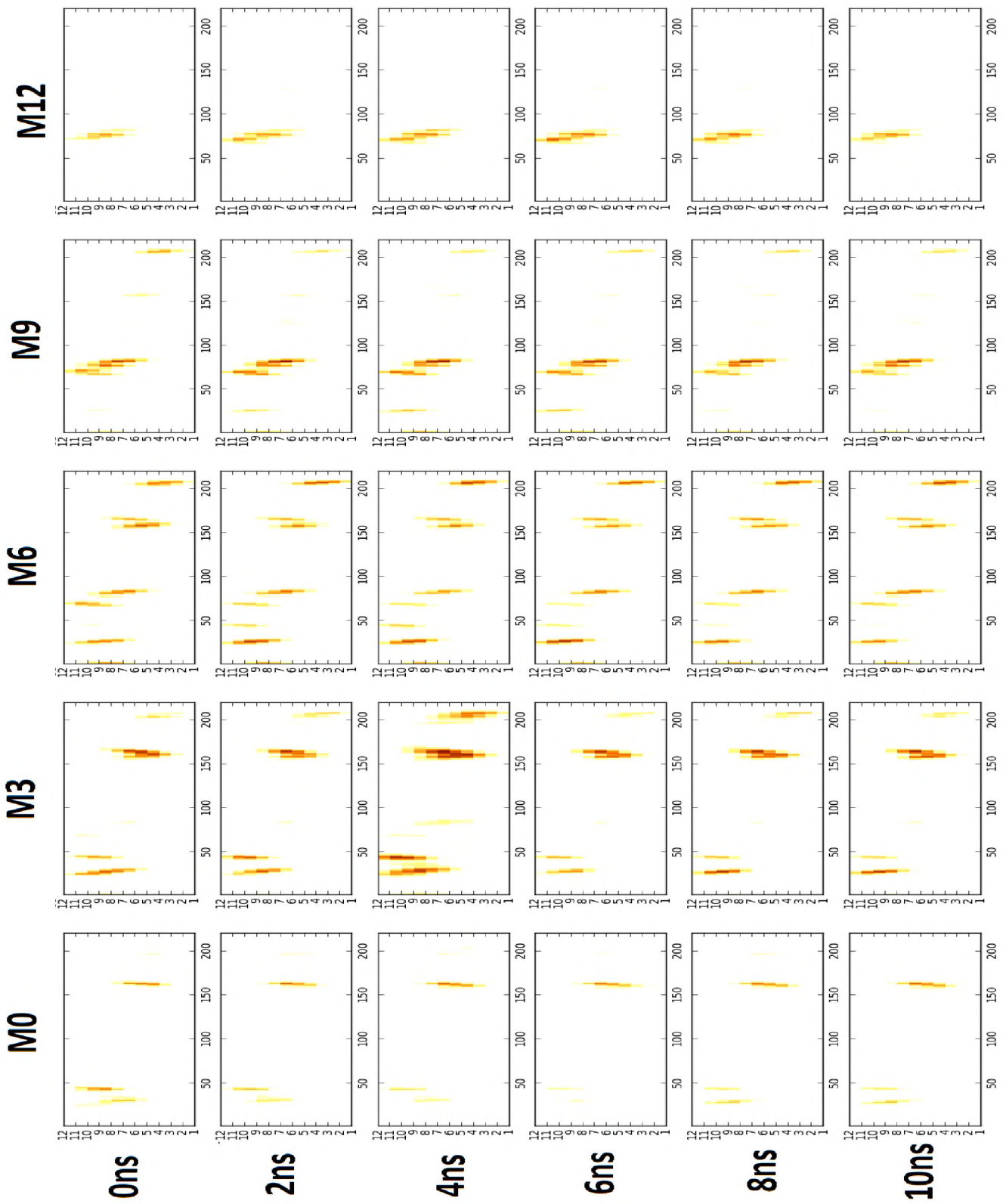

Figure S6: Contact map for the unbiased Type 2 10 ns MD simulation.

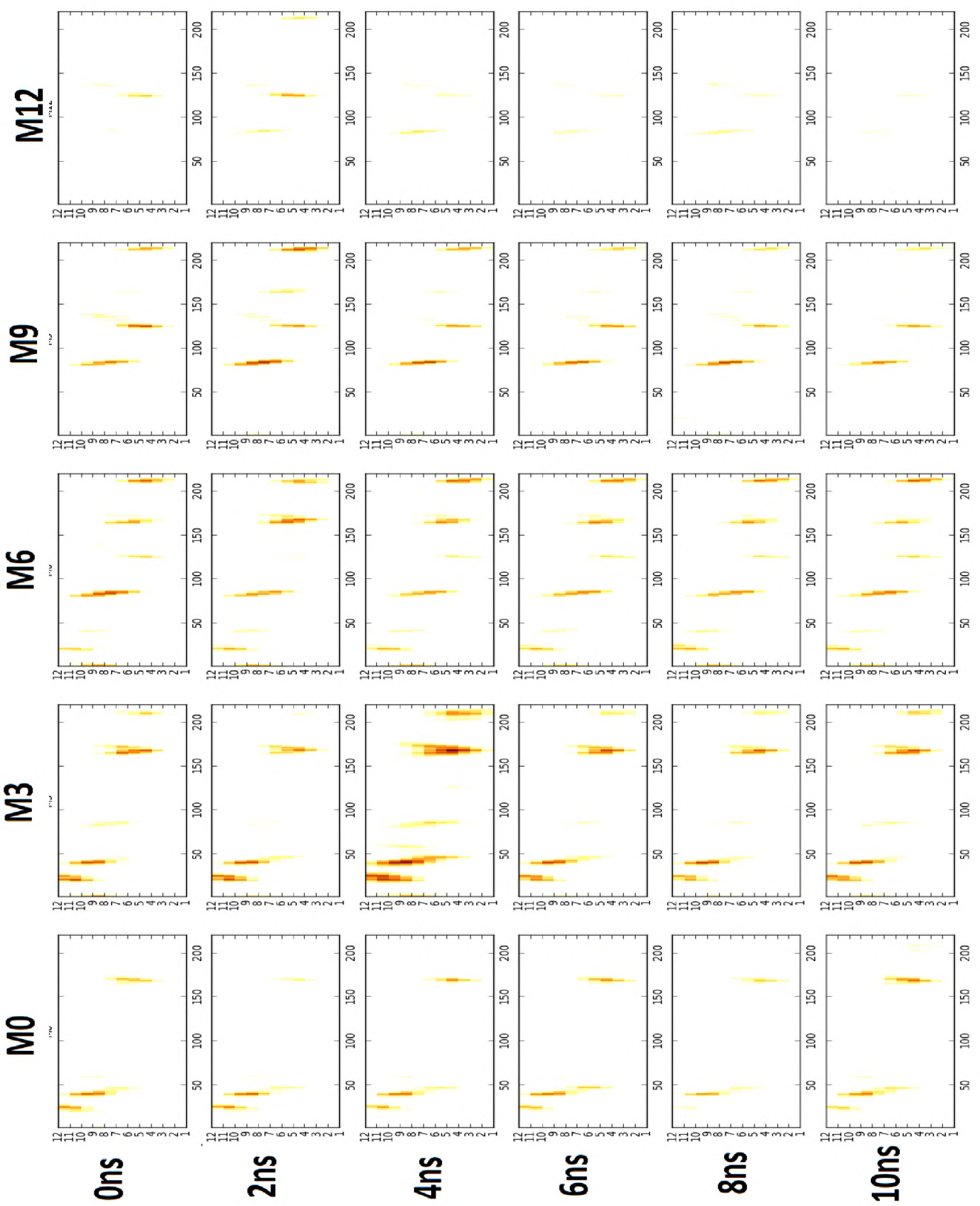

Figure S7: Contact map for the biased Type 3 10 ns MD simulation.

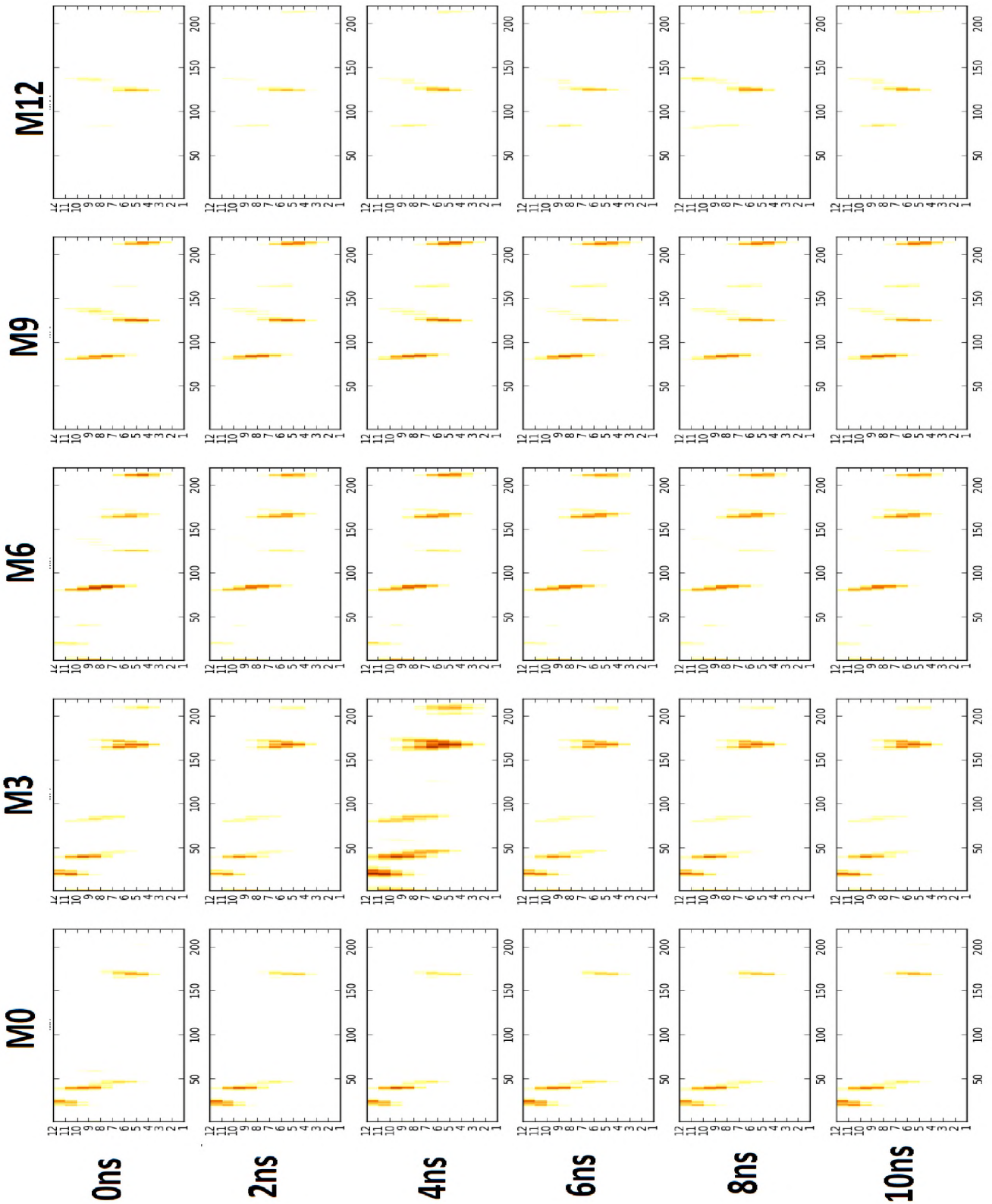

Figure S8: Contact map for the unbiased Type 3 10 ns MD simulation.

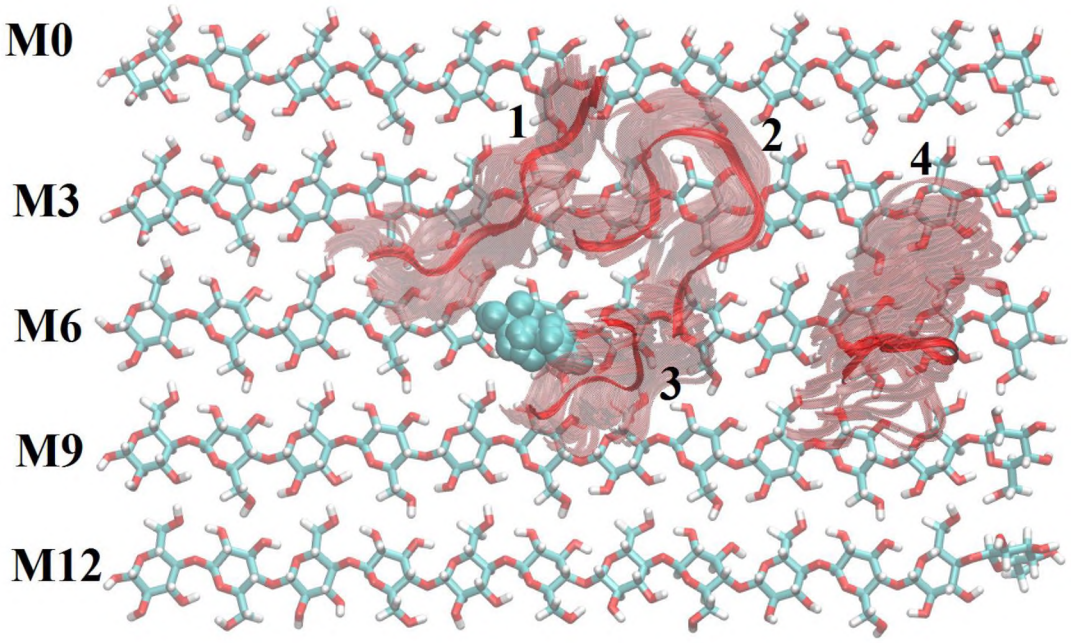

Figure S9: Contact regions from the Type 1 unbiased experiment. Red denotes contact regions that are common in all AA9 types (1-4) and were they are localized on the top layer of cellulose.

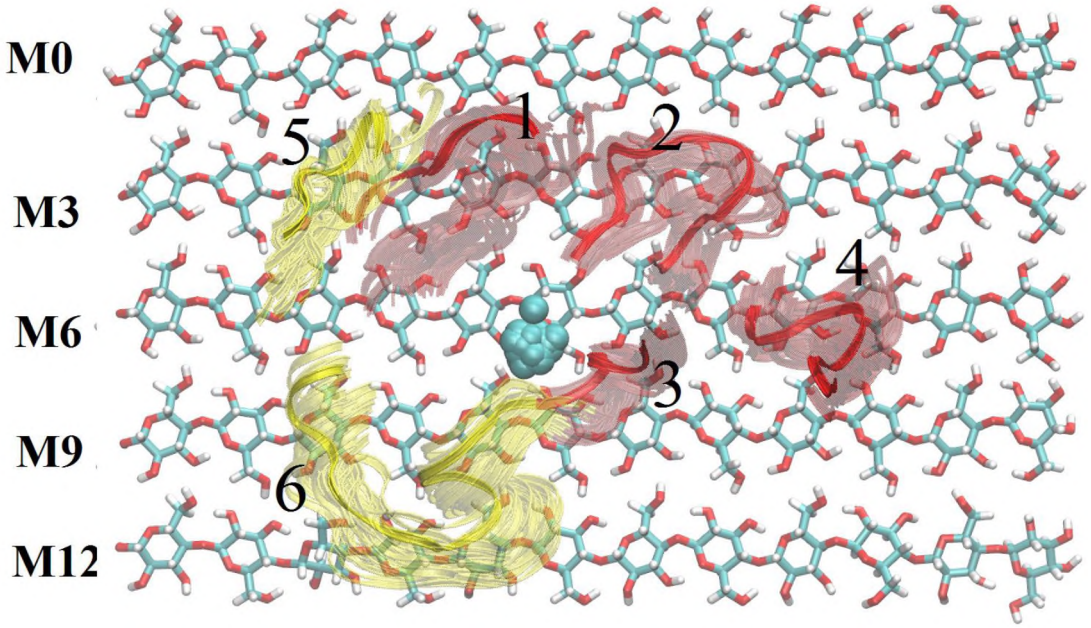

Figure S10: Contact regions from the Type 2 unbiased experiment. Red denotes contact regions that are common in all AA9 types (1-4) while yellow shows the type 2 specific contact regions (5 and 6) and were they are localized on the top layer of cellulose.

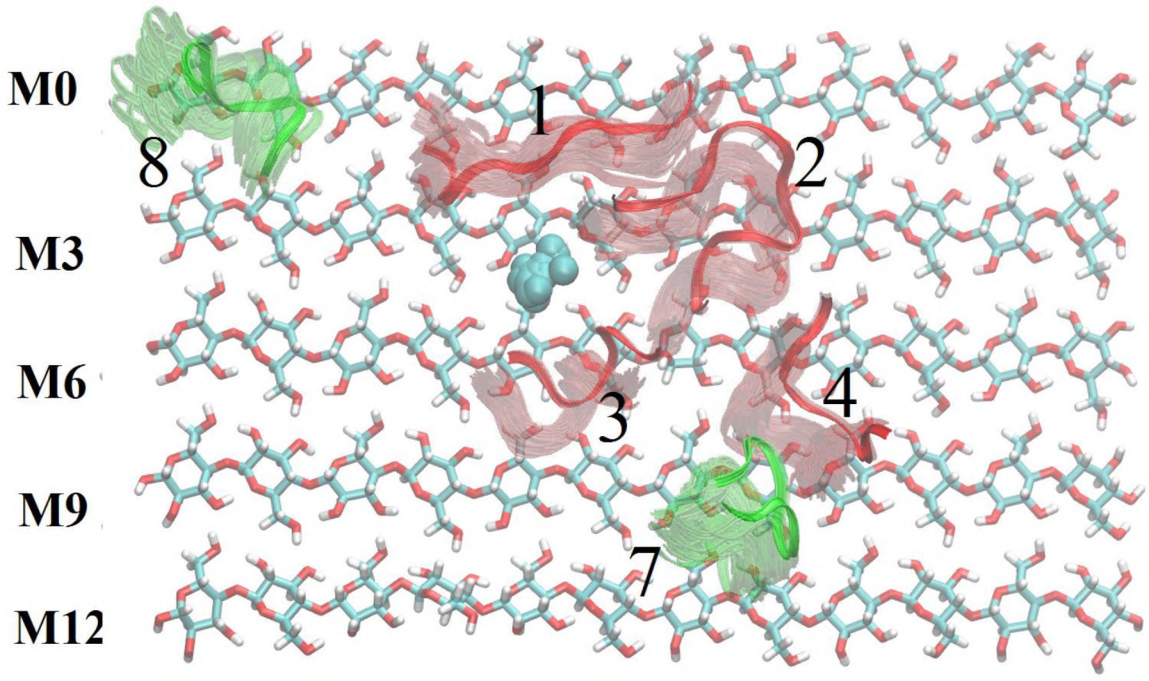

Figure S11: Contact regions from the Type 3 unbiased MD experiment. Red denotes contact regions that are common in all AA9 types (1-4) while green shows the Type 2 specific contact regions (7 and 8) and were they are localized on the top layer of cellulose.

Scripts

Script_1.py: script used to detect metal ions in crystal structures and measure the distances of their coordinating protein residues.

```
#script by Vuyani Moses
#script for detecting metal ions in pdb files and reporting the coordinating atoms based on user
supplied proximity
import math
filename=raw_input("Enter .pdb file   :  ")
f1=open(filename,"r")
pdblist=f1.readlines()
metallist=["Cu","CU","Mg","MG","Co","CO","Ni","NI","Zn","ZN"]#may be adjusted for mor metals
metalpdb=[] # metals found in the pdb file
allmetals=[]
allatoms=[] # all atoms in the pdb file
allatomsdist=[]# all atoms within specified distance
for line in pdblist:
        if line.startswith("HETATM"):
                a=line.split()
                for metal in metallist:
                        if metal==a[2]:
                                metalpdb.append([metal,a[1],a[2],a[3],a[4],a[5],a[6],a[7],a[8]])
        if line.startswith("ATOM"):
                b=line.split()
                allatoms.append([b[1],b[2],b[3],b[4],b[5],b[6],b[7],b[8]])

# count all residues in specified proximity
cutoff=raw_input("Enter minimun distance cut off :  ")
atomsinprox=[]
for met in metalpdb:
        x=met[6]
        y=met[7]
        z=met[8]
        for otheratom in allatoms:
                x1=otheratom[5]
                y1=otheratom[6]
                z1=otheratom[7]
                x_dist = (float(x) - float(x1))**2
                y_dist = (float(y) - float(y1))**2
                z_dist = (float(z) - float(z1))**2
                final_dist=math.sqrt(x_dist + y_dist + z_dist)
                if final_dist < int(cutoff):
                        atomsinprox.append([otheratom[1],otheratom[2],otheratom[3],otheratom[4],met
[1],met[2],met[3],met[4],met[5],final_dist])

print "The script detected "+str(len(metalpdb))+" metals in the provided .pdb file.\n"
print "ATOM\tAA\tID\tRes _No.\t\tMetal\tChain\tRes_No\tProximity\n"

f2=open("metal_distances.txt","a")
f2.write("The script detected "+str(len(metalpdb))+" metals in the provided .pdb file.\n")
f2.write("ATOM\tAA\tID\tRes _No.\t\tMetal\tChain\tRes_No\tProximity\n")
for i in atomsinprox:
        f2.write(i[0]+"\t"+i[1]+"\t"+i[2]+"\t"+i[3]+"\t"+"\t"+"\t"+i[6]+"\t"+i[7]+"\t"+i[8]+"\t"+str(i
[9])+"\n")
        print i[0]+"\t"+i[1]+"\t"+i[2]+"\t"+i[3]+"\t"+"\t"+"\t"+i[6]+"\t"+i[7]+"\t"+i[8]+"\t"+str(i[9])
```

Script_2.py: Python script used to map the evaluated RESP charges onto their respective atomic positons.

```
#Vuyani Moses
# script for mapping RESP charges on your pdb file

a=raw_input("enter your .pdb file  :  ")#input file
b=raw_input("enter gausian file  :  ")#gausian file
c=raw_input("enter output .pdb file :  ")# out file file

f1=open(a,'r')
f2=open(b,"r")
f3=open(c,"a")

pdb=f1.readlines()# reading pdbfile
charges=f2.readlines()# reading gausian output file
chargelist=[]
# reading charges
for line in charges:
         chargelist.append(line[12:].rstrip("\n"))
counter=0
newfilelist=[]
# writing charges to beta factor field in pdb file
for i in chargelist:
        a=pdb[counter]
        newline= a[:55]+" 2   "+chargelist[counter][:6]+"              "+a[-3:]
        counter=counter+1
        newfilelist.append(newline)

f3.writelines(newfilelist)
```

Script_3.py: Python script used to calculate the contact points between AA9 structures and their substrate cellulose during the course of an MD simulation.

```
#script by Vuyani Moses and Dr Kevin Lobb
#script for detecting metal ions in pdb files and reporting the coordinating atoms based on user
supplied proximity
import math
filename="06_4b5q_heated.pdb"
f1=open(filename,"r")
pdblist=f1.readlines()
pdblines=[]
pdblinenumber=0

oldresiduenumber=""
oldprotresidue=""
oldresid=""
MINDISTANCE=20
distance=0
for line in pdblist:

        if line.startswith("ATOM"):
                k=line.split()
                if k[10]=="A" or k[10]=="M0" or k[10]==("M1") or k[10]==("M2") or k[10]==("M3")or k[10]==
("M4") or k[10]==("M5") or k[10]==("M6") or k[10]==("M7") or k[10]==("M8") or k[10]==("M9")or k[10]==
("M10") or k[10]==("M11")or k[10]==("M12")or k[10]==("M14"):
                        pdblinenumber=pdblinenumber+1
                        pdblines.append(line)
for element in pdblines:
        c=element.split()

        if c[3]=="BGLC" and c[2]=="C1":
                c=element.split()
                atomnumber=int(c[1])
                bx=float(c[5])
                by=float(c[6])
                bz=float(c[7])
                residuenumber=c[4]
                resid=c[10]
                oldprotresidue=""
                mindistance =MINDISTANCE
                for element2 in pdblines:

                        b=element2.split()
                        if b[0]=="ATOM" and b[3]!="BGLC" and b[3]!="CU":

                                patomnumber=int(b[1])
                                px=float(b[5])
                                py=float(b[6])
                                pz=float(b[7])
                                proteinresidue=b[4]
                                if proteinresidue != oldprotresidue:
```

```
                                        if oldprotresidue !="":
                                                print "Cellulose "+residuenumber+" "+resid+"  Protein
"+oldprotresidue+" ", mindistance
                                                mindistance=MINDISTANCE
                                        oldprotresidue=proteinresidue

                                        for element3 in pdblines:
                                                kk=element3.split()
                                                catomnumber=int(kk[1])
                                                cx=float(kk[5])
                                                cy=float(kk[6])
                                                cz=float(kk[7])
                                                cresidue=kk[4]
                                                cresid=kk[10]
                                                if cresidue==residuenumber and cresid==resid:
                                                        distance = math.sqrt((bx-px)*(bx-px)+(by-py)*(by-
py)+(bz-pz)*(bz-pz))
                                                        if distance < mindistance:
                                                                mindistance=distance
                                        #else:
                                    if b[10]=="CU":
                                                print "Cellulose "+residuenumber+" "+resid+"  Protein
"+oldprotresidue+" ", mindistance
```

References

Adman, E.T. 1991, "Copper Protein Structures", *Advances in Protein Chemistry,* vol. 42, pp. 145-197.

Agger, J.W., Isaksen, T., Varnai, A., Vidal-Melgosa, S., Willats, W.G., Ludwig, R., Horn, S.J., Eijsink, V.G. & Westereng, B. 2014, "Discovery of LPMO activity on hemicelluloses shows the importance of oxidative processes in plant cell wall degradation", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 111, no. 17, pp. 6287-6292.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. 1997, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic acids research,* vol. 25, no. 17, pp. 3389-3402.

Andersen, C.A., Palmer, A.G., Brunak, S. & Rost, B. 2002, "Continuum secondary structure captures protein flexibility", *Structure (London, England : 1993),* vol. 10, no. 2, pp. 175-184.

Ando, K. 2010, "The axial methionine ligand may control the redox reorganizations in the active site of blue copper proteins", *The Journal of chemical physics,* vol. 133, no. 17, pp. 175101.

Arantes, V. & Saddler, J.N. 2010, "Access to cellulose limits the efficiency of enzymatic hydrolysis: the role of amorphogenesis", *Biotechnology for Biofuels,* vol. 3, no. 1, pp. 4.

Babu, C.S. & Lim, C. 2006, "Empirical Force Fields for Biologically Active Divalent Metal Cations in Water", *The Journal of Physical Chemistry A,* vol. 110, no. 2, pp. 691-699.

Bailey, T.L. & Gribskov, M. 1998, "Combining evidence using p-values: application to sequence homology searches", *Bioinformatics (Oxford, England),* vol. 14, no. 1, pp. 48-54.

Bailey, T.L., Williams, N., Misleh, C. & Li, W.W. 2006, "MEME: discovering and analyzing DNA and protein sequence motifs", *Nucleic acids research,* vol. 34, no. Web Server issue, pp. W369-73.

Baker, D. & Agard, D.A. 1994, "Kinetics versus Thermodynamics in Protein Folding", *Biochemistry,* vol. 33, no. 24, pp. 7505-7509.

Baker, J.O., Ehrman, C.I., Adney, W.S., Thomas, S.R. & Himmel, M.E. 1998, "Hydrolysis of cellulose using ternary mixtures of purified celluloses", *Applied Biochemistry and Biotechnology,* vol. 70, no. 1, pp. 395-403.

Bayly, C.I., Cieplak, P., Cornell, W. & Kollman, P.A. 1993, "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model", *The Journal of physical chemistry,* vol. 97, no. 40, pp. 10269-10280.

Becke, A.D. 1993, "Density-functional thermochemistry. III. The role of exact exchange", *The Journal of chemical physics,* vol. 98, no. 7, pp. 5648-5652.

Beeson, W.T., Phillips, C.M., Cate, J.H. & Marletta, M.A. 2012, "Oxidative cleavage of cellulose by fungal copper-dependent polysaccharide monooxygenases", *Journal of the American Chemical Society,* vol. 134, no. 2, pp. 890-892.

Beeson, W.T., Vu, V.V., Span, E.A., Phillips, C.M. & Marletta, M.A. 2015, "Cellulose degradation by polysaccharide monooxygenases", *Annual Review of Biochemistry,* vol. 84, pp. 923-946.

Bennati-Granier, C., Garajova, S., Champion, C., Grisel, S., Haon, M., Zhou, S., Fanuel, M., Ropartz, D., Rogniaux, H., Gimbert, I., Record, E. & Berrin, J.G. 2015, "Substrate specificity and regioselectivity of fungal AA9 lytic polysaccharide monooxygenases secreted by Podospora anserina", *Biotechnology for biofuels,* vol. 8, pp. 90-015-0274-3. eCollection 2015.

Berendsen, H.J.C., Grigera, J.R. & Straatsma, T.P. 1987, "The missing term in effective pair potentials", *The Journal of physical chemistry,* vol. 91, no. 24, pp. 6269-6271.

Bishop, A.O.T., Beer, Tjaart A. P. de & Joubert, F. 2008, "Protein homology modelling and its use in South Africa", *South African Journal of Science,* vol. 104, no. 1-2, pp. 2-6.

Bjellqvist, B., Basse, B., Olsen, E. & Celis, J.E. 1994, "Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions", *Electrophoresis,* vol. 15, no. 3-4, pp. 529-539.

Boniecki, M., Rotkiewicz, P., Skolnick, J. & Kolinski, A. 2003, "Protein fragment reconstruction using various modeling techniques", *Journal of computer-aided molecular design,* vol. 17, no. 11, pp. 725-738.

Boraston, A.B., Bolam, D.N., Gilbert, H.J. & Davies, G.J. 2004, "Carbohydrate-binding modules: fine-tuning polysaccharide recognition", *The Biochemical journal,* vol. 382, no. Pt 3, pp. 769-781.

Bork, P. & Koonin, E.V. 1996, "Protein sequence motifs", *Current opinion in structural biology,* vol. 6, no. 3, pp. 366-376.

Breneman, C.M. & Wiberg, K.B. 1990, "Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis", *Journal of Computational Chemistry,* vol. 11, no. 3, pp. 361-373.

Brooks, B.R., Brooks, C.L., MacKerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M. & Karplus, M. 2009, "CHARMM: The Biomolecular Simulation Program", *J.Comput.Chem.,* vol. 30, no. 10, pp. 1545-1614.

Bruschi, M., Bertini, L., Bonacic-Koutecky, V., De Gioia, L., Mitric, R., Zampella, G. & Fantucci, P. 2012, "Speciation of copper-peptide complexes in water solution using DFTB and DFT approaches: case of the [Cu(HGGG)(Py)] complex", *The journal of physical chemistry.B,* vol. 116, no. 22, pp. 6250-6260.

Busk, P.K. & Lange, L. 2015, "Classification of fungal and bacterial lytic polysaccharide monooxygenases", *BMC genomics,* vol. 16, pp. 368-015-1601-6.

Cannistraro, A.R.B.a.S. 1997, "Anomalous and anisotropic diffusion of plastocyanin hydration water", *EPL (Europhysics Letters),* vol. 37, no. 3, pp. 201.

Capellán-Pérez, I., Mediavilla, M., de Castro, C., Carpintero, Ó. & Miguel, L.J. 2014, "Fossil fuel depletion and socio-economic scenarios: An integrated approach", *Energy,* vol. 77, pp. 641-666.

Carugo, O. & Pongor, S. 2001, "A normalized root-mean-spuare distance for comparing protein three-dimensional structures", *Protein Science : A Publication of the Protein Society,* vol. 10, no. 7, pp. 1470-1473.

Cavasotto, C.N. & Phatak, S.S. 2009, "Homology modeling in drug discovery: current trends and applications", *Drug discovery today,* vol. 14, no. 13–14, pp. 676-683.

Chen, I.J., Yin, D. & MacKerell, A.D. 2002, "Combined ab initio/empirical approach for optimization of Lennard-Jones parameters for polar-neutral compounds", *Journal of Computational Chemistry,* vol. 23, no. 2, pp. 199-213.

Chinchio, M., Czaplewski, C., Liwo, A., Oldziej, S. & Scheraga, H.A. 2007, "Dynamic Formation and Breaking of Disulfide Bonds in Molecular Dynamics Simulations with the UNRES Force Field", *Journal of chemical theory and computation,* vol. 3, no. 4, pp. 1236-1248.

Chung, L.W., Sameera, W.M., Ramozzi, R., Page, A.J., Hatanaka, M., Petrova, G.P., Harris, T.V., Li, X., Ke, Z., Liu, F., Li, H.B., Ding, L. & Morokuma, K. 2015, "The ONIOM Method and Its Applications", *Chemical reviews,* vol. 115, no. 12, pp. 5678-5796.

Cieplak, P., Cornell, W.D., Bayly, C. & Kollman, P.A. 1995, "Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins", *Journal of Computational Chemistry,* vol. 16, no. 11, pp. 1357-1377.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. & Kollman, P.A. 1996, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules J. Am. Chem. Soc. 1995, 117, 5179â^'5197", *Journal of the American Chemical Society,* vol. 118, no. 9, pp. 2309-2309.

Correa, T.L., dos Santos, L.V. & Pereira, G.A. 2016, "AA9 and AA10: from enigmatic to essential enzymes", *Applied Microbiology and Biotechnology,* vol. 100, no. 1, pp. 9-16.

Dasmeh, P., Serohijos, A.W., Kepp, K.P. & Shakhnovich, E.I. 2014, "The influence of selection for protein stability on dN/dS estimations", *Genome biology and evolution,* vol. 6, no. 10, pp. 2956-2967.

Dassault Systèmes BIOVIA 2016, "*Discovery Studio Modeling Environment*", *Dassault Systèmes an Diego, .*

Delport, W., Poon, A.F., Frost, S.D. & Kosakovsky Pond, S.L. 2010, "Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology", *Bioinformatics (Oxford, England),* vol. 26, no. 19, pp. 2455-2457.

Deng, N.J., Yan, L., Singh, D. & Cieplak, P. 2006, "Molecular basis for the Cu2+ binding-induced destabilization of beta2-microglobulin revealed by molecular dynamics simulation", *Biophysical journal,* vol. 90, no. 11, pp. 3865-3879.

di Luccio, E. & Koehl, P. 2011, "A quality metric for homology modeling: the H-factor", *BMC Bioinformatics,* vol. 12, no. 1, pp. 48.

Ditchfield, R., Hehre, W.J. & Pople, J.A. 1971, "Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules", *The Journal of chemical physics,* vol. 54, no. 2, pp. 724-728.

Do, C.B., Mahabhashyam, M.S., Brudno, M. & Batzoglou, S. 2005, "ProbCons: Probabilistic consistency-based multiple sequence alignment", *Genome research,* vol. 15, no. 2, pp. 330-340.

Dunning, T.H. & Hay, P.J. 1977, "Gaussian Basis Sets for Molecular Calculations" in *Methods of Electronic Structure Theory*, ed. H.F. Schaefer, Springer US, Boston, MA, pp. 1-27.

Eastwood, D.C., Floudas, D., Binder, M., Majcherczyk, A., Schneider, P., Aerts, A., Asiegbu, F.O., Baker, S.E., Barry, K., Bendiksby, M., Blumentritt, M., Coutinho, P.M., Cullen, D., de Vries, R.P., Gathman, A., Goodell, B., Henrissat, B., Ihrmark, K., Kauserud, H., Kohler, A., LaButti, K., Lapidus, A., Lavin, J.L., Lee, Y., Lindquist, E., Lilly, W., Lucas, S., Morin, E., Murat, C., Oguiza, J.A., Park, J., Pisabarro, A.G., Riley, R., Rosling, A., Salamov, A., Schmidt, O., Schmutz, J., Skrede, I., Stenlid, J., Wiebenga, A., Xie, X., KÃ¼es, U., Hibbett, D.S., Hoffmeister, D., HÃ¶gberg, N., Martin, F., Grigoriev, I.V. & Watkinson, S.C. 2011, "The Plant Cell Wallâ€"Decomposing Machinery Underlies the Functional Diversity of Forest Fungi", *Science,* vol. 333, no. 6043, pp. 762.

Eddy, S.R. 1998, "Profile hidden Markov models", *Bioinformatics,* vol. 14.

Eisenberg, D., Lüthy, R. & Bowie, J.U. 1997, "[20] VERIFY3D: Assessment of protein models with three-dimensional profiles", *Methods in enzymology,* vol. 277, pp. 396-404.

Enkhbayar, P., Hikichi, K., Osaki, M., Kretsinger, R.H. & Matsushima, N. 2006, "310-helices in proteins are parahelices", *Proteins: Structure, Function, and Bioinformatics,* vol. 64, no. 3, pp. 691-699.

Eramian, D., Eswar, N., Shen, M.Y. & Sali, A. 2008, "How well can the accuracy of comparative protein structure models be predicted?", *Protein science : a publication of the Protein Society,* vol. 17, no. 11, pp. 1881-1893.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V.A., Pieper, U., Stuart, A.C., Marti-Renom, M., Madhusudhan, M.S., Yerkovich, B. & Sali, A. 2003, "Tools for comparative protein structure modeling and analysis", *Nucleic Acids Res,* vol. 31.

Faya, N., Penkler, D.L. & Tastan Bishop, O. 2015, "Human, vector and parasite Hsp90 proteins: A comparative bioinformatics analysis", *FEBS open bio,* vol. 5, pp. 916-927.

Feller, S.E., Pastor, R.W., Rojnuckarin, A., Bogusz, S. & Brooks, B.R. 1996, "Effect of Electrostatic Force Truncation on Interfacial and Transport Properties of Water", *The Journal of physical chemistry,* vol. 100, no. 42, pp. 17011-17020.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J. & Bateman, A. 2016, "The Pfam protein families database: towards a more sustainable future", *Nucleic acids research,* vol. 44, no. D1, pp. D279-85.

Fiser, A. 2010, "Template-Based Protein Structure Modeling" in *Computational Biology*, ed. D. FenyÃ¶, Humana Press, Totowa, NJ, pp. 73-94.

Fiser, A., Do, R.K.G. & Šali, A. 2000, "Modeling of loops in protein structures", *Protein Science,* vol. 9, no. 9, pp. 1753-1773.

Forsberg, Z., Mackenzie, A.K., Sorlie, M., Rohr, A.K., Helland, R., Arvai, A.S., Vaaje-Kolstad, G. & Eijsink, V.G. 2014a, "Structural and functional characterization of a conserved pair of bacterial cellulose-oxidizing lytic polysaccharide monooxygenases", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 111, no. 23, pp. 8446-8451.

Forsberg, Z., Rohr, A.K., Mekasha, S., Andersson, K.K., Eijsink, V.G., Vaaje-Kolstad, G. & Sorlie, M. 2014b, "Comparative study of two chitin-active and two cellulose-active AA10-type lytic polysaccharide monooxygenases", *Biochemistry,* vol. 53, no. 10, pp. 1647-1656.

Frandsen, K.E. & Lo Leggio, L. 2016, "Lytic polysaccharide monooxygenases: a crystallographer's view on a new class of biomass-degrading enzymes", *IUCrJ,* vol. 3, no. Pt 6, pp. 448-467.

Frandsen, K.E., Simmons, T.J., Dupree, P., Poulsen, J.C., Hemsworth, G.R., Ciano, L., Johnston, E.M., Tovborg, M., Johansen, K.S., von Freiesleben, P., Marmuse, L., Fort, S., Cottaz, S., Driguez, H., Henrissat, B., Lenfant, N., Tuna, F., Baldansuren, A., Davies, G.J., Lo Leggio, L. & Walton, P.H. 2016, "The molecular basis of polysaccharide cleavage by lytic polysaccharide monooxygenases", *Nature chemical biology,* vol. 12, no. 4, pp. 298-303.

Fredin, L.A. & Allison, T.C. 2016, "Predicting Structures of Ru-Centered Dyes: A Computational Screening Tool", *The journal of physical chemistry.A,* vol. 120, no. 13, pp. 2135-2143.

Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G.A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H.P., Izmaylov, A.F., Bloino, J., Zheng, G., Sonnenberg, J.L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., A., M., Jr.J., Peralta, J.E., Ogliaro, F., Bearpark, M., Heyd, J.J., Brothers, E., Kudin, K.N., Staroverov, V.N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J.C., Iyengar, S.S., Tomasi, J., Cossi, M., Rega, N., Millam, J.M., Klene, M., Knox, J.E., Cross, J.B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R.E., Yazyev, O., Austin, A.J., Cammi, R., Pomelli, C., Ochterski, J.W., Martin, R.L., Morokuma, K., Zakrzewski, V.G., Voth, G.A., Salvador, P., Dannenberg, J.J., Dapprich, S., Daniels, A.D., Farkas, Ã., Foresman, J.B., Ortiz, J.V., Cioslowski, J. & Fox, D.J. "Gaussian09 Revision E.01", *Gaussian09 Revision E.01,* .

Frommhagen, M., Sforza, S., Westphal, A.H., Visser, J., Hinz, S.W., Koetsier, M.J., van Berkel, W.J., Gruppen, H. & Kabel, M.A. 2015, "Discovery of the combined oxidative cleavage of plant xylan and cellulose by a new fungal polysaccharide monooxygenase", *Biotechnology for biofuels,* vol. 8, pp. 101-015-0284-1. eCollection 2015.

Furlan, S., Hureau, C., Faller, P. & La Penna, G. 2010, "Modeling the Cu+ binding in the 1-16 region of the amyloid-beta peptide involved in Alzheimer's disease", *The journal of physical chemistry.B,* vol. 114, no. 46, pp. 15119-15133.

Goodell, B., Jellison, J., Liu, J., Daniel, G., Paszczynski, A., Fekete, F., Krishnamurthy, S., Jun, L. & Xu, G. 1997, "Low molecular weight chelators and phenolic compounds isolated from wood decay fungi and their role in the fungal biodegradation of wood1", *Journal of Biotechnology,* vol. 53, no. 2–3, pp. 133-162.

Gordon, M.S. 1980, "The isomers of silacyclopropane", *Chemical Physics Letters,* vol. 76, no. 1, pp. 163-168.

Gudmundsson, M., Kim, S., Wu, M., Ishida, T., Momeni, M.H., Vaaje-Kolstad, G., Lundberg, D., Royant, A., Stahlberg, J., Eijsink, V.G., Beckham, G.T. & Sandgren, M. 2014, "Structural and electronic snapshots during the transition from a Cu(II) to Cu(I) metal center of a lytic polysaccharide monooxygenase by X-ray photoreduction", *The Journal of biological chemistry,* vol. 289, no. 27, pp. 18782-18792.

Guerra, A., MendonÃ§a, R., Ferraz, A., Lu, F. & Ralph, J. 2004, "Structural Characterization of Lignin during Pinus taeda Wood Treatment with Ceriporiopsis subvermispora", *Applied and Environmental Microbiology,* vol. 70, no. 7, pp. 4073-4078.

Guruprasad, K., Reddy, B.V. & Pandit, M.W. 1990, "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence", *Protein engineering,* vol. 4, no. 2, pp. 155-161.

Guvench, O., Mallajosyula, S.S., Raman, E.P., Hatcher, E., Vanommeslaeghe, K., Foster, T.J., Jamison, F.W. & MacKerell, A.D. 2011, "CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate-protein modeling", *J. Chem. Theory Comput.,* vol. 7, no. 10, pp. 3162-3180.

Ha, S.N., Giammona, A., Field, M. & Brady, J.W. 1988, "A revised potential-energy surface for molecular mechanics studies of carbohydrates", *Carbohydrate research,* vol. 180, no. 2, pp. 207-221.

Hariharan, P.C. & Pople, J.A. 1974, "Accuracy of AH n equilibrium geometries by single determinant molecular orbital theory", *Molecular Physics,* vol. 27, no. 1, pp. 209-214.

Hariharan, P.C. & Pople, J.A. 1973, "The influence of polarization functions on molecular orbital hydrogenation energies", *Theoretica chimica acta,* vol. 28, no. 3, pp. 213-222.

Harris, P.V., Welner, D., McFarland, K.C., Re, E., Navarro Poulsen, J., Brown, K., Salbo, R., Ding, H., Vlasenko, E., Merino, S., Xu, F., Cherry, J., Larsen, S. & Lo Leggio, L. 2010, "Stimulation of Lignocellulosic Biomass Hydrolysis by Proteins of Glycoside Hydrolase Family 61: Structure and Function of a Large, Enigmatic Family", *Biochemistry,* vol. 49, no. 15, pp. 3305-3316.

Hazes, B. & Dijkstra, B.W. 1988, "Model building of disulfide bonds in proteins with known three-dimensional structure", *Protein engineering,* vol. 2, no. 2, pp. 119-125.

Hehre, W.J., Ditchfield, R. & Pople, J.A. 1972, "Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules", *The Journal of chemical physics,* vol. 56, no. 5, pp. 2257-2261.

Hemsworth, G.R., Davies, G.J. & Walton, P.H. 2013, "Recent insights into copper-containing lytic polysaccharide mono-oxygenases", *Current opinion in structural biology,* vol. 23, no. 5, pp. 660-668.

Hemsworth, G.R., Henrissat, B., Davies, G.J. & Walton, P.H. 2014, "Discovery and characterization of a new family of lytic polysaccharide monooxygenases", *Nature chemical biology,* vol. 10, no. 2, pp. 122-126.

Hewitt, N. & Rauk, A. 2009, "Mechanism of hydrogen peroxide production by copper-bound amyloid beta peptide: a theoretical study", *The journal of physical chemistry.B,* vol. 113, no. 4, pp. 1202-1209.

Himmel, M.E., Ding, S.Y., Johnson, D.K., Adney, W.S., Nimlos, M.R., Brady, J.W. & Foust, T.D. 2007, "Biomass recalcitrance: engineering plants and enzymes for biofuels production", *Science (New York, N.Y.),* vol. 315, no. 5813, pp. 804-807.

Hodak, M., Chisnell, R., Lu, W. & Bernholc, J. 2009, "Functional implications of multistage copper binding to the prion protein", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 106, no. 28, pp. 11576-11581.

Honarparvar, B., Govender, T., Maguire, G.E., Soliman, M.E. & Kruger, H.G. 2014, "Integrated approach to structure-based enzymatic drug design: molecular modeling, spectroscopy, and experimental bioactivity", *Chemical reviews,* vol. 114, no. 1, pp. 493-537.

Hong, L. & Lei, J. 2009, "Scaling law for the radius of gyration of proteins and its dependence on hydrophobicity", *Journal of Polymer Science Part B: Polymer Physics,* vol. 47, no. 2, pp. 207-214.

Hori, C., Igarashi, K., Katayama, A. & Samejima, M. 2011, "Effects of xylan and starch on secretome of the basidiomycete Phanerochaete chrysosporium grown on cellulose", *FEMS microbiology letters,* vol. 321, no. 1, pp. 14-23.

Horn, S.J., Vaaje-Kolstad, G., Westereng, B. & Eijsink, V.G. 2012, "Novel enzymes for the degradation of cellulose", *Biotechnology for biofuels,* vol. 5, no. 1, pp. 45-6834-5-45.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. 2006, "Comparison of multiple Amber force fields and development of improved protein backbone parameters", *Proteins,* vol. 65, no. 3, pp. 712-725.

Hu, J., Arantes, V., Pribowo, A., Gourlay, K. & Saddler, J.N. 2014, "Substrate factors that influence the synergistic interaction of AA9 and cellulases during the enzymatic hydrolysis of biomass", *Energy & Environmental Science,* vol. 7, no. 7, pp. 2308-2315.

Huige, C.J.M. & Altona, C. 1995, "Force field parameters for sulfates and sulfamates based on ab initio calculations: Extensions of AMBER and CHARMm fields", *Journal of Computational Chemistry,* vol. 16, no. 1, pp. 56-79.

Humphrey, W., Dalke, A. & Schulten, K. 1996, "VMD: visual molecular dynamics", *Journal of Molecular Graphics,* vol. 14, no. 1, pp. 33-8, 27-8.

Illergard, K., Ardell, D.H. & Elofsson, A. 2009, "Structure is three to ten times more conserved than sequence--a study of structural response in protein cores", *Proteins,* vol. 77, no. 3, pp. 499-508.

Inoue, T., Shiota, Y. & Yoshizawa, K. 2008, "Quantum chemical approach to the mechanism for the biological conversion of tyrosine to dopaquinone", *Journal of the American Chemical Society,* vol. 130, no. 50, pp. 16890-16897.

Isaksen, T., Westereng, B., Aachmann, F.L., Agger, J.W., Kracher, D., Kittl, R., Ludwig, R., Haltrich, D., Eijsink, V.G. & Horn, S.J. 2014a, "A C4-oxidizing lytic polysaccharide monooxygenase cleaving both cellulose and cello-oligosaccharides", *The Journal of biological chemistry,* vol. 289, no. 5, pp. 2632-2642.

Isaksen, T., Westereng, B., Aachmann, F.L., Agger, J.W., Kracher, D., Kittl, R., Ludwig, R., Haltrich, D., Eijsink, V.G. & Horn, S.J. 2014b, "A C4-oxidizing lytic polysaccharide monooxygenase cleaving both cellulose and cello-oligosaccharides", *The Journal of biological chemistry,* vol. 289, no. 5, pp. 2632-2642.

Jamroz, M. & Kolinski, A. 2010, "Modeling of loops in proteins: a multi-method approach", *BMC Structural Biology,* vol. 10, no. 1, pp. 5.

Jeoh, T., Wilson, D.B. & Walker, L.P. 2002, "Cooperative and competitive binding in synergistic mixtures of Thermobifida fusca cellulases Cel5A, Cel6B, and Cel9A", *Biotechnology progress,* vol. 18, no. 4, pp. 760-769.

Jonassen, I., Eidhammer, I., Conklin, D. & Taylor, W.R. 2002, "Structure motif discovery and mining the PDB", *Bioinformatics (Oxford, England),* vol. 18, no. 2, pp. 362-367.

Jones, D.T. 1999, "Protein secondary structure prediction based on position-specific scoring matrices", *J Mol Biol,* vol. 292.

Jones, R.O. 2015, "Density functional theory: Its origins, rise to prominence, and future", *Rev.Mod.Phys.,* vol. 87, no. 3, pp. 897-923.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. 1983, "Comparison of simple potential functions for simulating liquid water", *The Journal of chemical physics,* vol. 79, no. 2, pp. 926-935.

Jorgensen, W.L. & Tirado-Rives, J. 1988, "The OPLS optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin", *Journal of the American Chemical Society,* vol. 110, no. 6, pp. 1657-1666.

Kabsch, W. & Sander, C. 1983, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features", *Biopolymers,* vol. 22, no. 12, pp. 2577-2637.

Karkehabadi, S., Hansson, H., Kim, S., Piens, K., Mitchinson, C. & Sandgren, M. 2008, "The first structure of a glycoside hydrolase family 61 member, Cel61B from Hypocrea jecorina, at 1.6 A resolution", *Journal of Molecular Biology,* vol. 383, no. 1, pp. 144-154.

Karlsson, J., Saloheimo, M., Siika-Aho, M., Tenkanen, M., Penttila, M. & Tjerneld, F. 2001, "Homologous expression and characterization of Cel61A (EG IV) of Trichoderma reesei", *European journal of biochemistry / FEBS,* vol. 268, no. 24, pp. 6498-6507.

Karplus, M. & Kuriyan, J. 2005, "Molecular dynamics and protein function", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 19, pp. 6679-6685.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. 2002, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform", *Nucleic acids research,* vol. 30, no. 14, pp. 3059-3066.

Katoh, K. & Toh, H. 2008, "Recent developments in the MAFFT multiple sequence alignment program", *Briefings in bioinformatics,* vol. 9, no. 4, pp. 286-298.

Kim, I.J., Nam, K.H., Yun, E.J., Kim, S., Youn, H.J., Lee, H.J., Choi, I. & Kim, K.H. 2015, "Optimization of synergism of a recombinant auxiliary activity 9 from Chaetomium globosum with cellulase in cellulose hydrolysis", *Applied Microbiology and Biotechnology,* vol. 99, no. 20, pp. 8537-8547.

Kim, S., Stahlberg, J., Sandgren, M., Paton, R.S. & Beckham, G.T. 2014, "Quantum mechanical calculations suggest that lytic polysaccharide monooxygenases use a copper-oxyl, oxygen-rebound mechanism", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 111, no. 1, pp. 149-154.

Kjaergaard, C.H., Qayyum, M.F., Wong, S.D., Xu, F., Hemsworth, G.R., Walton, D.J., Young, N.A., Davies, G.J., Walton, P.H., Johansen, K.S., Hodgson, K.O., Hedman, B. & Solomon, E.I. 2014, "Spectroscopic and computational insight into the activation of O2 by the mononuclear Cu center in polysaccharide monooxygenases", *Proceedings of the National Academy of Sciences,* vol. 111, no. 24, pp. 8797-8802.

Kouwijzer, M.L.C.E., Van Eijck, B.P., Kroes, S.J. & Kroon, J. 1993, "Comparison of two force fields by molecular dynamics simulations of glucose crystals: Effect of using ewald sums", *Journal of Computational Chemistry,* vol. 14, no. 11, pp. 1281-1289.

Kraulis, J., Clore, G.M., Nilges, M., Jones, T.A., Pettersson, G., Knowles, J. & Gronenborn, A.M. 1989, "Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from Trichoderma reesei. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing", *Biochemistry,* vol. 28, no. 18, pp. 7241-7257.

Krishnamoorthy, B. & Tropsha, A. 2003, "Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations", *Bioinformatics (Oxford, England),* vol. 19, no. 12, pp. 1540-1548.

Kryazhimskiy, S. & Plotkin, J.B. 2008, "The Population Genetics of dN/dS", *PLoS Genetics,* vol. 4, no. 12, pp. e1000304.

Kufareva, I. & Abagyan, R. 2012, "Methods of protein structure comparison", *Methods in molecular biology (Clifton, N.J.),* vol. 857, pp. 231-257.

Kuttel, M., Brady, J.W. & Naidoo, K.J. 2002, "Carbohydrate solution simulations: Producing a force field with experimentally consistent primary alcohol rotational frequencies and populations", *Journal of Computational Chemistry,* vol. 23, no. 13, pp. 1236-1243.

Kuzmanic, A. & Zagrovic, B. 2010, "Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors", *Biophysical journal,* vol. 98, no. 5, pp. 861-871.

Kyte, J. & Doolittle, R.F. 1982, "A simple method for displaying the hydropathic character of a protein", *Journal of Molecular Biology,* vol. 157, no. 1, pp. 105-132.

Laskowski, R.A. 2001, "PDBsum: summaries and analyses of PDB structures", *Nucleic acids research,* vol. 29, no. 1, pp. 221-222.

Lazaridis, T. & Karplus, M. 1999, "Effective energy function for proteins in solution", *Proteins: Structure, Function, and Bioinformatics,* vol. 35, no. 2, pp. 133-152.

Lee, C., Yang, W. & Parr, R.G. 1988, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density", *Phys.Rev.B,* vol. 37, no. 2, pp. 785-789.

Leggio, L.L., Welner, D. & De Maria, L. 2012, "A structural overview of GH61 proteins - fungal cellulose degrading polysaccharide monooxygenases", *Computational and structural biotechnology journal,* vol. 2, pp. e201209019.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P.M. & Henrissat, B. 2013, "Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes", *Biotechnology for biofuels,* vol. 6, no. 1, pp. 41-6834-6-41.

Levitt, M., Hirshberg, M., Sharon, R., Laidig, K.E. & Daggett, V. 1997, "Calibration and Testing of a Water Model for Simulation of the Molecular Dynamics of Proteins and Nucleic Acids in Solution", *The Journal of Physical Chemistry B,* vol. 101, no. 25, pp. 5051-5061.

Li, P. & Merz, K.M.,Jr 2014, "Taking into Account the Ion-induced Dipole Interaction in the Nonbonded Model of Ions", *Journal of chemical theory and computation,* vol. 10, no. 1, pp. 289-297.

Li, X., Beeson, W.T.,4th, Phillips, C.M., Marletta, M.A. & Cate, J.H. 2012, "Structural basis for substrate targeting and catalysis by fungal polysaccharide monooxygenases", *Structure (London, England : 1993),* vol. 20, no. 6, pp. 1051-1061.

Lin, K., May, A.C. & Taylor, W.R. 2002, "Threading using neural nEtwork (TUNE): the measure of protein sequence-structure compatibility", *Bioinformatics (Oxford, England),* vol. 18, no. 10, pp. 1350-1357.

Lins, R.D. & Hünenberger, P.H. 2005, "A new GROMOS force field for hexopyranose-based carbohydrates", *Journal of Computational Chemistry,* vol. 26, no. 13, pp. 1400-1412.

Lo Leggio, L., Simmons, T.J., Poulsen, J.C., Frandsen, K.E., Hemsworth, G.R., Stringer, M.A., von Freiesleben, P., Tovborg, M., Johansen, K.S., De Maria, L., Harris, P.V., Soong, C.L., Dupree, P., Tryfona, T., Lenfant, N., Henrissat, B., Davies, G.J. & Walton, P.H. 2015a, "Structure and boosting activity of a starch-degrading lytic polysaccharide monooxygenase", *Nature communications,* vol. 6, pp. 5961.

Lo Leggio, L., Simmons, T.J., Poulsen, J.C., Frandsen, K.E., Hemsworth, G.R., Stringer, M.A., von Freiesleben, P., Tovborg, M., Johansen, K.S., De Maria, L., Harris, P.V., Soong, C.L., Dupree, P., Tryfona, T., Lenfant, N., Henrissat, B., Davies, G.J. & Walton, P.H. 2015b, "Structure and boosting activity of a starch-degrading lytic polysaccharide monooxygenase", *Nature communications,* vol. 6, pp. 5961.

Lobanov, M.Y., Bogatyreva, N.S. & Galzitskaya, O.V. 2008, "Radius of gyration as an indicator of protein structure compactness", *Molecular biology,* vol. 42, no. 4, pp. 623-628.

Lobry, J.R. & Gautier, C. 1994, "Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes", *Nucleic acids research,* vol. 22, no. 15, pp. 3174-3180.

Lovell, S.C., Davis, I.W., Arendall, W.B.,3rd, de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S. & Richardson, D.C. 2003, "Structure validation by Calpha geometry: phi,psi and Cbeta deviation", *Proteins,* vol. 50, no. 3, pp. 437-450.

Mackerell, A.D.,Jr 2004, "Empirical force fields for biological macromolecules: overview and issues", *Journal of computational chemistry,* vol. 25, no. 13, pp. 1584-1604.

MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. 1998, "All-atom empirical potential for molecular modeling and dynamics studies of proteins", *The journal of physical chemistry.B,* vol. 102, no. 18, pp. 3586-3616.

MacPherson, I.S. & Murphy, M.E.P. 2007, "Type-2 copper-containing enzymes", *Cellular and Molecular Life Sciences,* vol. 64, no. 22, pp. 2887-2899.

Mäkelä, M.R., Donofrio, N. & de Vries, R.P. 2014, "Plant biomass degradation by fungi", *Fungal Genetics and Biology,* vol. 72, pp. 2-9.

Mansfield, S.D., De Jong, E. & Saddler, J.N. 1997, "Cellobiose dehydrogenase, an active agent in cellulose depolymerization", *Applied and Environmental Microbiology,* vol. 63, no. 10, pp. 3804-3809.

Mark, P. & Nilsson, L. 2001, "Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K", *J. Phys. Chem. A,* vol. 105, no. 43, pp. 9954-9960.

Massingham, T. & Goldman, N. 2005, "Detecting amino acid sites under positive selection and purifying selection", *Genetics,* vol. 169, no. 3, pp. 1753-1762.

Mayhall, N.J., Raghavachari, K. & Hratchian, H.P. 2010, "ONIOM-based QM:QM electronic embedding method using Löwdin atomic charges: Energies and analytic gradients", *The Journal of chemical physics,* vol. 132, no. 11, pp. 114107.

Mayrose, I., Stern, A., Burdelova, E.O., Sabo, Y., Laham-Karam, N., Zamostiano, R., Bacharach, E. & Pupko, T. 2013, "Synonymous site conservation in the HIV-1 genome", *BMC evolutionary biology,* vol. 13, pp. 164-2148-13-164.

McCallum, C.M. 1999, "Statistical Mechanics: Fundamentals and Modern Applications (Wilde, Richard E.; Singh, Surjit)", *Journal of chemical education,* vol. 76, no. 6, pp. 761.

McCammon, J.A., Gelin, B.R. & Karplus, M. 1977, "Dynamics of folded proteins", *Nature,* vol. 267, no. 5612, pp. 585-590.

McGuffin, L.J., Bryson, K. & Jones, D.T. 2000, "The PSIPRED protein structure prediction server", *Bioinformatics,* vol. 16.

Meller, J. 2001, "Molecular Dynamics" in *eLS* John Wiley & Sons, Ltd, .

Melo, F., Devos, D., Depiereux, E. & Feytmans, E. 1997, "ANOLEA: a www server to assess protein structures", *Proceedings.International Conference on Intelligent Systems for Molecular Biology,* vol. 5, pp. 187-190.

Melo, F. & Feytmans, E. 1997, "Novel knowledge-based mean force potential at atomic level", *Journal of Molecular Biology,* vol. 267, no. 1, pp. 207-222.

Mentler, M., Weiss, A., Grantner, K., del Pino, P., Deluca, D., Fiori, S., Renner, C., Klaucke, W.M., Moroder, L., Bertsch, U., Kretzschmar, H.A., Tavan, P. & Parak, F.G. 2005, "A new method to determine the structure of the metal environment in metalloproteins: investigation of the prion protein octapeptide repeat Cu(2+) complex", *European biophysics journal : EBJ,* vol. 34, no. 2, pp. 97-112.

Merino, S.T. & Cherry, J. 2007, "Progress and Challenges in Enzyme Development for Biomass Utilization" in *Biofuels*, ed. L. Olsson, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 95-120.

Moon, R.J., Martini, A., Nairn, J., Simonsen, J. & Youngblood, J. 2011, "Cellulose nanomaterials review: structure, properties and nanocomposites", *Chemical Society Reviews,* vol. 40, no. 7, pp. 3941-3994.

Moses, V., Hatherley, R. & Tastan Bishop, O. 2016, "Bioinformatic characterization of type-specific sequence and structural features in auxiliary activity family 9 proteins", *Biotechnology for biofuels,* vol. 9, pp. 239.

Mugal, C.F., Wolf, J.B. & Kaj, I. 2014, "Why time matters: codon evolution and the temporal dynamics of dN/dS", *Molecular biology and evolution,* vol. 31, no. 1, pp. 212-231.

Mulliken, R.S. 1955, "Electronic Population Analysis on LCAOâ€• MO Molecular Wave Functions. IV. Bonding and Antibonding in LCAO and Valenceâ€• Bond Theories", *The Journal of chemical physics,* vol. 23, no. 12, pp. 2343-2346.

Nguyen, V.-., Béchet, E., Geuzaine, C. & Noels, L. 2012, "Imposing periodic boundary condition on arbitrary meshes by polynomial interpolation", *Computational Materials Science,* vol. 55, pp. 390-406.

Nishiyama, Y., Langan, P. & Chanzy, H. 2002, "Crystal structure and hydrogen-bonding system in cellulose Ibeta from synchrotron X-ray and neutron fiber diffraction", *Journal of the American Chemical Society,* vol. 124, no. 31, pp. 9074-9082.

Ott, K. & Meyer, B. 1996, "Parametrization of GROMOS force field for oligosaccharides and assessment of efficiency of molecular dynamics simulations", *Journal of Computational Chemistry,* vol. 17, no. 8, pp. 1068-1084.

Pareek, C.S., Smoczynski, R. & Tretyn, A. 2011, "Sequencing technologies and genome sequencing", *Journal of Applied Genetics,* vol. 52, no. 4, pp. 413-435.

Pawlowski, M., Gajda, M.J., Matlak, R. & Bujnicki, J.M. 2008, "MetaMQAP: A meta-server for the quality assessment of protein models", *BMC Bioinformatics,* vol. 9, pp. 403-403.

Pearson, W.R. 2013, "Selecting the Right Similarity-Scoring Matrix", *Current protocols in bioinformatics,* vol. 43, pp. 3.5.1-9.

Pei, J. & Grishin, N.V. 2014, "PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information", *Methods in molecular biology (Clifton, N.J.),* vol. 1079, pp. 263-271.

Percival Zhang, Y.-., Himmel, M.E. & Mielenz, J.R. 2006, "Outlook for cellulase improvement: Screening and selection strategies", *Biotechnology Advances,* vol. 24, no. 5, pp. 452-481.

PÉrez, S. & Samain, D. 2010, "Structure and Engineering of Celluloses", *Advances in Carbohydrate Chemistry and Biochemistry,* vol. 64, pp. 25-116.

Phillips, C.M., Beeson, W.T., Cate, J.H. & Marletta, M.A. 2011, "Cellobiose dehydrogenase and a copper-dependent polysaccharide monooxygenase potentiate cellulose degradation by Neurospora crassa", *ACS chemical biology,* vol. 6, no. 12, pp. 1399-1406.

Pond, S.L. & Frost, S.D. 2005, "Datamonkey: rapid detection of selective pressure on individual sites of codon alignments", *Bioinformatics (Oxford, England),* vol. 21, no. 10, pp. 2531-2533.

Ponder, J.W. & Case, D.A. 2003, "Force fields for protein simulations", *Advances in Protein Chemistry,* vol. 66, pp. 27-85.

Quinlan, R.J., Sweeney, M.D., Lo Leggio, L., Otten, H., Poulsen, J.C., Johansen, K.S., Krogh, K.B., Jorgensen, C.I., Tovborg, M., Anthonsen, A., Tryfona, T., Walter, C.P., Dupree, P., Xu, F., Davies, G.J. & Walton, P.H. 2011, "Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 108, no. 37, pp. 15079-15084.

Rassolov, V.A., Ratner, M.A., Pople, J.A., Redfern, P.C. & Curtiss, L.A. 2001, "6-31G* basis set for third-row atoms", *Journal of Computational Chemistry,* vol. 22, no. 9, pp. 976-984.

Redhead, E. & Bailey, T.L. 2007, "Discriminative motif discovery in DNA and protein sequences using the DEME algorithm", *BMC Bioinformatics,* vol. 8, no. 1, pp. 385.

Reed, A.E., Weinstock, R.B. & Weinhold, F. 1985, "Natural population analysis", *The Journal of chemical physics,* vol. 83, no. 2, pp. 735-746.

Reiling, S., Schlenkrich, M. & Brickmann, J. 1996, "Force field parameters for carbohydrates", *Journal of Computational Chemistry,* vol. 17, no. 4, pp. 450-468.

Rubino, J.T. & Franz, K.J. 2012, "Coordination chemistry of copper proteins: How nature handles a toxic cargo for essential function", *Journal of inorganic biochemistry,* vol. 107, no. 1, pp. 129-143.

Ruel, K., Nishiyama, Y. & Joseleau, J. 2012, "Crystalline and amorphous cellulose in the secondary walls of Arabidopsis", *Plant Science,* vol. 193–194, pp. 48-61.

Sabolovic, J., Tautermann, C.S., Loerting, T. & Liedl, K.R. 2003, "Modeling anhydrous and aqua copper(II) amino acid complexes: a new molecular mechanics force field parametrization based on quantum chemical studies and experimental crystal data", *Inorganic chemistry,* vol. 42, no. 7, pp. 2268-2279.

Sali, A. & Blundell, T.L. 1993, "Comparative protein modelling by satisfaction of spatial restraints", *J Mol Biol,* vol. 234.

Sanchez, R. & Sali, A. 1997, "Advances in comparative protein-structure modelling", *Current opinion in structural biology,* vol. 7, no. 2, pp. 206-214.

Schr\odinger, L. 2015, *The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint Version~1.8.*

Scott, W.R.P., HÃ¼nenberger, P.H., Tironi, I.G., Mark, A.E., Billeter, S.R., Fennen, J., Torda, A.E., Huber, T., KrÃ¼ger, P. & van Gunsteren, W.F. 1999, "The GROMOS Biomolecular Simulation Program Package", *The Journal of Physical Chemistry A,* vol. 103, no. 19, pp. 3596-3607.

Searls, D.B. 2003, "Pharmacophylogenomics: genes, evolution and drug targets", *Nature reviews.Drug discovery,* vol. 2, no. 8, pp. 613-623.

Shen, M.Y. & Sali, A. 2006, "Statistical potential for assessment and prediction of protein structures", *Protein science : a publication of the Protein Society,* vol. 15, no. 11, pp. 2507-2524.

Singh, U.C. & Kollman, P.A. 1984, "An approach to computing electrostatic charges for molecules", *Journal of Computational Chemistry,* vol. 5, no. 2, pp. 129-145.

Soding, J., Biegert, A. & Lupas, A.N. 2005, "The HHpred interactive server for protein homology detection and structure prediction", *Nucleic acids research,* vol. 33, no. Web Server issue, pp. W244-8.

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E. & Pupko, T. 2007, "Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach", *Nucleic acids research,* vol. 35, no. Web Server issue, pp. W506-11.

Stewart, J.J. 2013, "Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters", *Journal of molecular modeling,* vol. 19, no. 1, pp. 1-32.

Stone, D.B., Schneider, D.K., Huang, Z. & Mendelson, R.A. 1995, "The radius of gyration of native and reductively methylated myosin subfragment-1 from neutron scattering", *Biophysical journal,* vol. 69, no. 3, pp. 767-776.

Suyama, M., Torrents, D. & Bork, P. 2006, "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments", *Nucleic acids research,* vol. 34, no. Web Server issue, pp. W609-12.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C.H. 2007, "UniRef: comprehensive and non-redundant UniProt reference clusters", *Bioinformatics,* vol. 23, no. 10, pp. 1282-1288.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H. & the, U.C. 2014, "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches", *Bioinformatics,* vol. 31, no. 6, pp. 926-932.

Suzuki, Y. 2004, "New methods for detecting positive selection at single amino acid sites", *Journal of Molecular Evolution,* vol. 59, no. 1, pp. 11-19.

Suzuki, Y. & Gojobori, T. 1999, "A method for detecting positive selection at single amino acid sites.", *Molecular biology and evolution,* vol. 16, no. 10, pp. 1315-1328.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013, "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0", *Molecular biology and evolution,* vol. 30, no. 12, pp. 2725-2729.

Team, R.C. 2013, "R: A Language and Environment for Statistical Computing", .

Torras, J. & Aleman, C. 2013, "Determination of new Cu+, Cu2+, and Zn2+ Lennard-Jones ion parameters in acetonitrile", *The journal of physical chemistry.B,* vol. 117, no. 36, pp. 10513-10522.

Tyrus, J.M., Gosz, M. & DeSantiago, E. 2007, "A local finite element implementation for imposing periodic boundary conditions on composite micromechanical models", *International Journal of Solids and Structures,* vol. 44, no. 9, pp. 2972-2989.

Ungar, L.W., Scherer, N.F. & Voth, G.A. 1997, "Classical molecular dynamics simulation of the photoinduced electron transfer dynamics of plastocyanin", *Biophysical journal,* vol. 72, no. 1, pp. 5-17.

Vaaje-Kolstad, G., Houston, D.R., Riemen, A.H., Eijsink, V.G. & van Aalten, D.M. 2005, "Crystal structure and binding properties of the Serratia marcescens chitin-binding protein CBP21", *The Journal of biological chemistry,* vol. 280, no. 12, pp. 11313-11319.

Vaaje-Kolstad, G., Westereng, B., Horn, S.J., Liu, Z., Zhai, H., Sorlie, M. & Eijsink, V.G. 2010, "An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides", *Science (New York, N.Y.),* vol. 330, no. 6001, pp. 219-222.

van, d.B. & de Vries, R.,P. 2011, "Fungal enzyme sets for plant polysaccharide degradation", *Applied Microbiology and Biotechnology,* vol. 91, no. 6, pp. 1477-1492.

Vieira-Pires, R. & Morais-Cabral, J. 2010, "3_{10} helices in channels and other membrane proteins", *J Gen Physiol,* vol. 136, no. 6, pp. 585.

Vosko, S.H., Wilk, L. & Nusair, M. 1980, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis", *Canadian Journal of Physics,* vol. 58, no. 8, pp. 1200-1211.

Vu, V.V., Beeson, W.T., Phillips, C.M., Cate, J.H. & Marletta, M.A. 2014, "Determinants of regioselective hydroxylation in the fungal polysaccharide monooxygenases", *Journal of the American Chemical Society,* vol. 136, no. 2, pp. 562-565.

Vu, V.V. & Marletta, M.A. 2016, "Starch-degrading polysaccharide monooxygenases", *Cellular and molecular life sciences : CMLS,* vol. 73, no. 14, pp. 2809-2819.

Vu, V.V., Beeson, W.T., Span, E.A., Farquhar, E.R. & Marletta, M.A. 2014, "A family of starch-active polysaccharide monooxygenases", *Proceedings of the National Academy of Sciences,* vol. 111, no. 38, pp. 13822-13827.

Wallner, B. & Elofsson, A. 2006, "Identification of correct regions in protein models using structural, alignment, and consensus information", *Protein science : a publication of the Protein Society,* vol. 15, no. 4, pp. 900-913.

Wang, Q., Werstiuk, N.H., Kramer, J.R. & Bell, R.A. 2011a, "Effects of Cu ions and explicit water molecules on the copper binding domain of amyloid precursor protein APP(131-189): a molecular dynamics study", *The journal of physical chemistry.B,* vol. 115, no. 29, pp. 9224-9235.

Wang, T., Andreazza, H.J., Pukala, T.L., Sherman, P.J., Calabrese, A.N. & Bowie, J.H. 2011b, "Histidine-containing host-defence skin peptides of anurans bind Cu2+. An electrospray ionisation mass spectrometry and computational modelling study", *Rapid communications in mass spectrometry : RCM,* vol. 25, no. 9, pp. 1209-1221.

Webb, B. & Sali, A. 2016, "Comparative Protein Structure Modeling Using MODELLER", *Current protocols in protein science,* vol. 86, pp. 2.9.1-2.9.37.

Webb, B. & Sali, A. 2014, "Comparative Protein Structure Modeling Using MODELLER", *Current protocols in bioinformatics / editoral board, Andreas D.Baxevanis ...[et al.],* vol. 47, pp. 5.6.1-5.6.32.

Whelan, S. & Goldman, N. 2001, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach", *Molecular biology and evolution,* vol. 18, no. 5, pp. 691-699.

White, A.R. & Brown, R.M. 1981, "Enzymatic hydrolysis of cellulose: Visual characterization of the process", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 78, no. 2, pp. 1047-1051.

Wiederstein, M. & Sippl, M.J. 2007, "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins", *Nucleic Acids Res,* vol. 35.

Wiltgen, M. & Tilz, G.P. 2009, "Homology modelling: a review about the method on hand of the diabetic antigen GAD 65 structure prediction", *Wiener medizinische Wochenschrift (1946),* vol. 159, no. 5-6, pp. 112-125.

Wise, O. & Coskuner, O. 2014, "New force field parameters for metalloproteins I: Divalent copper ion centers including three histidine residues and an oxygen-ligated amino acid residue", *J.Comput.Chem.,* vol. 35, no. 17, pp. 1278-1289.

Witt, P.L. & McGrain, P. 2016, "Comparing Two Sample Means t Tests", *Physical Therapy,* vol. 65, no. 11, pp. 1730-1733.

Wright, P.E. & Dyson, H.J. 1999, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm", *Journal of Molecular Biology,* vol. 293, no. 2, pp. 321-331.

Wu, M., Beckham, G.T., Larsson, A.M., Ishida, T., Kim, S., Payne, C.M., Himmel, M.E., Crowley, M.F., Horn, S.J., Westereng, B., Igarashi, K., Samejima, M., Stahlberg, J., Eijsink, V.G. & Sandgren, M. 2013, "Crystal structure and computational characterization of the lytic polysaccharide monooxygenase GH61D from the Basidiomycota fungus Phanerochaete chrysosporium", *The Journal of biological chemistry,* vol. 288, no. 18, pp. 12828-12839.

Xiang, Z. 2006, "Advances in homology protein structure modeling", *Current Protein & Peptide Science,* vol. 7, no. 3, pp. 217-227.

Yakovlev, I., Vaaje-Kolstad, G., Hietala, A.M., Stefanczyk, E., Solheim, H. & Fossdal, C.G. 2012, "Substrate-specific transcription of the enigmatic GH61 family of the pathogenic white-rot fungus Heterobasidion irregulare during growth on lignocellulose", *Applied Microbiology and Biotechnology,* vol. 95, no. 4, pp. 979-990.

Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A.M. 2000, "Codon-substitution models for heterogeneous selection pressure at amino acid sites", *Genetics,* vol. 155, no. 1, pp. 431-449.

Yang, Z. & Rannala, B. 2012, "Molecular phylogenetics: principles and practice", *Nature reviews.Genetics,* vol. 13, no. 5, pp. 303-314.

Zemla, A., Venclovas, Moult, J. & Fidelis, K. 2001, "Processing and evaluation of predictions in CASP4", *Proteins,* vol. Suppl 5, pp. 13-21.

Zhu, N., Liu, J., Yang, J., Lin, Y., Yang, Y., Ji, L., Li, M. & Yuan, H. 2016, "Comparative analysis of the secretomes of Schizophyllum commune and other wood-decay basidiomycetes during solid-state fermentation reveals its unique lignocellulose-degrading enzyme system", *Biotechnology for Biofuels,* vol. 9, pp. 42.

Zhu, Y., Su, Y., Li, X., Wang, Y. & Chen, G. 2008, "Evaluation of Amber force field parameters for copper(II) with pyridylmethyl-amine and benzimidazolylmethyl-amine ligands: A quantum chemical study", *Chem. Phys. Lett.,* vol. 455, no. 4–6, pp. 354-360.