Manuscript submitted to: Volume 1, Issue 1, 59-80.

AIMS Molecular Science DOI: 10.3934/molsci.2014.1.59

Received date 27 November 2013, Accepted date 7 January 2014, Published date 30 January 2014

Research Article

Identifying the Enzymatic Mode of Action for Cellulase Enzymes by Means of Docking Calculations and a Machine Learning Algorithm

Somisetti V. Sambasivarao [1], David M. Granum [1], Hua Wang [2] and C. Mark Maupin [1,] *

[1] Department of Chemical and Biological Engineering, Colorado School of Mines, Golden, CO 80401, USA

[2] Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401, USA

* **Correspondence:** Email: cmmaupin@mines.edu; Tel: +1-303-273-3197;
Fax: +1-303-273-3730.

Abstract: Docking calculations have been conducted on 36 cellulase enzymes and the results were evaluated by a machine learning algorithm to determine the nature of the enzyme (i.e. endo- or exo-enzymatic activity). The docking calculations have also been used to identify crucial substrate-enzyme interactions, and establish structure-function relationships. The use of carboxymethyl cellulose as a docking substrate is found to correctly identify the endo- or exo-behavior of cellulase enzymes with 92% accuracy while cellobiose docking calculations resulted in an 86% predictive accuracy. The binding distributions for cellobiose have been classified into two distinct types; distributions with a single maximum or distributions with a bi-modal structure. It is found that the uni-modal distributions correspond to exo- type enzyme while a bi-modal substrate docking distribution corresponds to endo- type enzyme. These results indicate that the use of docking calculations and machine learning algorithms are a fast and computationally inexpensive method for predicting if a cellulase enzyme possesses primarily endo- or exo- type behavior, while also revealing critical enzyme-substrate interactions.

Keywords: Carboxymethyl cellulose; Cellobiohydrolase; Cellobiose; Cellulase; Docking; Endoglucanase; Machine learning; Product inhibition

1. Introduction

Cellulosic biomass, the most abundant natural polymer on Earth, can be used as a feedstock for chemicals and liquid transportation with a significantly reduced carbon footprint as compared to petroleum [1-3]. Cellulosic biomass is an inherently versatile feed stock consisting of glucose units linked through a β-1,4-glycosidic bond. This renewable, green feedstock can be used as a cheap and nontoxic raw material in a myriad of chemical and biological reactions for the production of transportation fuel and value added chemicals and materials [1-8]. However, present technologies for the biodegradation of crystalline cellulose are not sufficiently optimized for general commercial scale production and industrial applications [9,10]. This shortcoming is primarily due to the recalcitrant nature of lignocellulosic material [2,3].

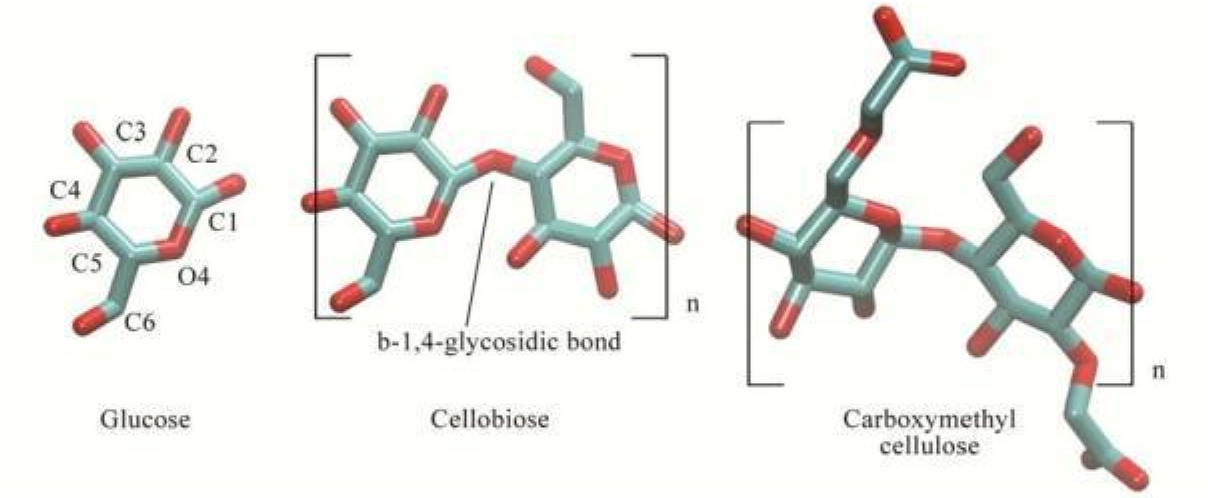

Figure 1 Licorice representation of glucose (left), the repeat cellobiose unit of a cellulose strand (middle), and the repeat carboxymethyl cellulose unit of amorphous cellulose (right). The carbon numbering system is depicted for the glucose molecule while the β-1,4-glycosidic bond is identified for the cellobiose repeat unit.

The difficulty in degrading cellulose is better understood through examining its molecular level structure and role in nature. Cellulose, which is produced biosynthetically and is vitally important to the global carbon flow, is composed of D-glucose units joined in a linear fashion by β-1,4-glycosidic bonds, Figure 1 [2]. The natural resistance of cellulose to degradation is in part due to the stability of the o-glycosidic bond, and to a larger extent the overall structure of cellulose, which is precisely arranged to maximize the strong hydrogen bonds between adjacent cellulose chains and the relatively weaker hydrophobic interaction between cellulose sheets [11,12]. This natural resistance to deconstruction (i.e. biomass recalcitrance) is primarily responsible for the high cost of biomass conversion to glucose [13].

In nature, biological organisms utilize several enzymes for the purpose of hydrolyzing crystalline cellulose. An efficient mixture of these enzymes (i.e. cellulases) work synergistically to hydrolyze cellulose as it is degraded from crystalline and/or amorphous cellulose to small, soluble cellulose fragments (e.g. cellobiose) and finally to glucose. In general, these enzyme mixtures consist of three different classifications of cellulase enzymes; endoglucanases (EG; EC 3.2.1.4), cellobiohydrolases (CBH; EC 3.2.1.91), and β-glucosidases (BG; EC 3.2.1.21), which are collectively referred to as glycosyl hydrolases (GH) [14,15,16]. The EGs are responsible for disrupting the crystalline structure of cellulose by randomly hydrolyzing accessible internal β-1,4-glycosidic bonds and thereby exposing individual cellulose polysaccharide chains. The CBHs act on the reducing (CBHI) and non-reducing (CBHII) ends of the exposed polysaccharide chains in a processive manner, releasing predominantly disaccharides (cellobiose) and to a lesser extent trisaccharides. The final component of the enzyme mixture is the BGs, which further hydrolyze the cellobiose to glucose. The slowest (i.e. rate limiting step) and most complicated aspect of hydrolyzing cellulose to glucose is believed to be the disruption of the crystalline cellulose substrate by EG and the CBHs [17].

Cellulases are classified into different families according to the primary amino acid sequence homology, secondary and tertiary structure, and their catalytic residues (CAZY database at http://www.cazy.org) [16-19]. All cellulases have multiple glucose binding sites (e.g. −7 to +2) with negative numbers representing non-reducing sugar binding sites, and positive numbers representing reducing sugar binding sites. In this numbering system, the −1 and +1 binding sites are responsible for catalysis of the cellulose fragments through hydrolysis of the o-glycosidic bond [20]. Across all cellulase families, a commonly accepted standard for classification of endo- or exo- behavior involves the ability of the enzyme to hydrolyze either carboxymethyl cellulose (CMC) or highly crystalline cellulose (Avicel) [17,21,22,23]. If the enzyme of interest shows measurable kinetics with CMC, a bulky derivative of cellulose (Figure 1), the enzyme is classified as an endo- cellulase, while significant activity on Avicel is given an exo-cellulase classification. The ability to discriminate between endo- and exo- activity based on the presence or absence of a bulky substrate (i.e. CMC) is due to the shape of the endo- or exo- active site groove [18,21,24]. Endo- and exo-cellulase active sites are formed by the closure of two distal loops flanking the active site region. In the case of exo-cellulases, these extensive loops remain in a *closed* conformation that results in a stable active site tunnel, while in endo-cellulases, these loops are shorter in length, resulting in a more *open* active site groove [21]. In exo-cellulases, the substrate is threaded through the active site tunnel, which contributes to the processive catalytic action observed in CBH enzymes [24,25]. In the case of the endo-cellulases, the distal loops have a reduced length resulting in an open active site resembling a groove or cleft, which allows the enzyme to adsorb onto a cellulose surface and bring its active site into close contact with a cellulose polymer. The critical importance of the distal loop length on the catalytic activity has been shown by experimental studies in which a deletion of 15 amino acids in one of the distal tunnel forming loops of a CBH (Cel6B from *Cellulomonas fimi)* resulted in a significant enhancement of endo- type activity for the enzyme [26].

Establishing if a cellulase enzyme has predominantly exo- or endo-cellulase behavior is important for future studies including efforts to bioengineer more efficient cellulase enzyme systems for industrial applications. Structural characteristics such as the loops enclosing the active site provide useful insights for specific enzymes, although it is difficult to quantitatively classify all cellulase enzymes using these structural features alone [27,28]. Kinetic measurements involving CMC and Avicel provide a better criterion for establishing endo- or exo-type activity, however these experiments are costly and time consuming, thus data is currenty unavailable for many cellulase enzymes. Furthermore, there will undoubtedly be many more cellulase structures available in the coming years from both experimental (i.e. x-ray) and computational (i.e. homology modeling), thus it is of interest to develop a quick, efficient method for classifying exo- or endo- type behavior. One possibility is to utilize available computational tools to develop models that can cheaply and easily predict the mode of action of cellulase enzymes. Furthermore, the use of computational methods allows for the investigation of other enzyme characteristics critical to functionality, which is difficult to study experimentally.

Here, we report the use of docking calculations between cellulase macromolecules, and the CMC and cellobiose ligands. The main objective of this study is to determine if docking calculations can be used as a quick and efficient means for accurately predicting endo- or exo-cellulase behavior. In total, 36 enzymes from 8 families were studied, all of which have been previously classified as possessing endo- or exo- activity by means of kinetic and/or X-ray studies. In addition to predicting the enzyme mode of action efficient computational studies such as docking allows for the identification of enzyme characteristics that may warrant further investigation, such as identification of cellulase mutants that may possess favorable characteristics (e.g. reduced product inhibition).

2. Materials and Method

2.1. Docking Calculations

The AutoDock 4.2 [29] program was used to conduct docking calculations on the cellulase enzymes, while the AutoDockTools [29] (ADT) software was used to prepare the input files. Two separate docking calculations were performed on each enzyme, one with cellobiose, and the other with CMC. The cellobiose ligand contained a total of 12 active torsions and the CMC ligand contained a total of 24 active torsions, with the torsion tree root of both ligands residing on the o-glycosidic bond oxygen. The ligands were then modified by adding polar hydrogens, merging all the non-polar hydrogens, and adding Gasteiger charges. The cellulase macromolecule structures were obtained from the protein data bank (PDB accession numbers are found in Table 1) and were modified by removing all crystallographic waters, metal ions, and associated ligands using the ADT program. The resulting macromolecule was then subjected to 1000 steps of steepest-descent and 1000 steps of conjugated gradient energy minimization using the UCSF Chimera package [30]. The resulting macromolecular structures were then used to prepare input files for performing docking simulations using the ADT software. In all docking calculations, the macromolecule was kept rigid while the ligand was allowed to sample the specified torsional parameters. The grid box created in ADT was sufficiently large to encompass the catalytic active tunnel/groove region. The dimensions for the grid boxes are given in Table S1 of the Supporting Information, and a representative grid box used for Cel7A [20] is shown in Figure 2. AutoGrid 4 [29] was used to produce grid maps with a default grid spacing of 0.375 Å. The docking calculations used the standard Autodock force field and the Lamarckian Genetic Algorithm (LGA) to search for the best docked ligand conformers. Each docking experiment consisted of 100 independent LGA runs with a population size of 150 and a random initial geometry for the ligand. The maximum number of energy evaluations for each LGA run was set at 25,000,000 with the maximum number of generations ranging from 1000 to 27000 depending on convergence. The maximum number of top individuals that automatically survived was set to 1, the mutation rate was set to 0.02, the crossover rate was set to 0.8, the translational step size was set to 2 Å, and the quaternions and torsion step size was set to 50°. For the analysis of the docking calculations, 100 conformers were considered for every complex, and the resulting docking clusters were calculated with a tolerance of 2.0 Å for the root mean squared deviation (rmsd) on the heavy atoms. Eight docking experiments were repeated four times to check the reproducibility of the docking patterns, and it was verified that consistent results are obtained. Furthermore, for cellulase enzymes that have a cellobiose bound in the crystal structure (14 total studied here), the lowest binding energy cellobiose pose was compared to the crystal structure, and the resulting RMSDs were all < 2.5 Å, indicating the docking results are accurately reproducing experimental data.

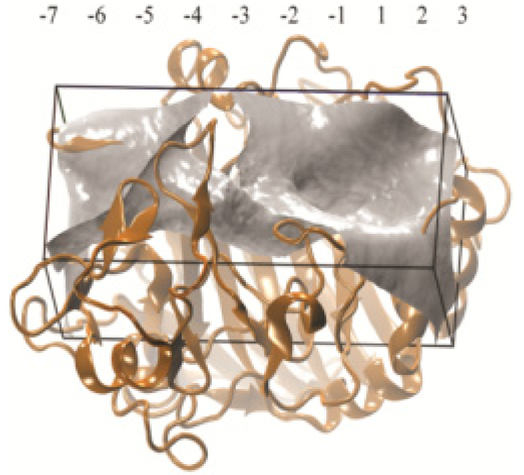

Figure 2 Representative docking grid box around the active site groove for Cel7A. The enzyme is represented in ribbon form while the surface of the active site groove is represented by the gray surface. The subsites of the catalytic tunnel/groove are indicated above the docking grid box.

2.2. Coarse Grained subsite docking analysis

To systematically analyze the CMC or cellobiose docking results, a protocol is required that will effectively work on a wide range of cellulase enzymes and substrates. Therefore, a distance dependent method was created that utilized the center of mass for each of the glucose rings to describe the spatial position and orientation for each of the docking solutions. The binding sites of the macromolecule, ranging from −7 to +2, were represented by spheres of equal diameter. The positions of the spheres were determined such that the spheres encompassed binding site residues identified by X-ray crystallography. The resulting active sub-site spheres smoothly and continuously tracked along the active site tunnel/groove much like beads on a string, as shown in Figure 3. The distances between an active sub-site sphere and the center of mass of the glucose ring were calculated for all of the identified ligand-macromolecule binding poses. The ligand for a particular binding pose was then assigned to the closest active sub-site sphere or pair of spheres. To eliminate assigning a ligand to an active site sphere when the ligand was not bound in the active site tunnel/groove, a distance cutoff of 6 Å was applied (i.e. if the closest active site sphere to a particular ligand is greater than 6 Å then that ligand is removed from the resulting histogram).

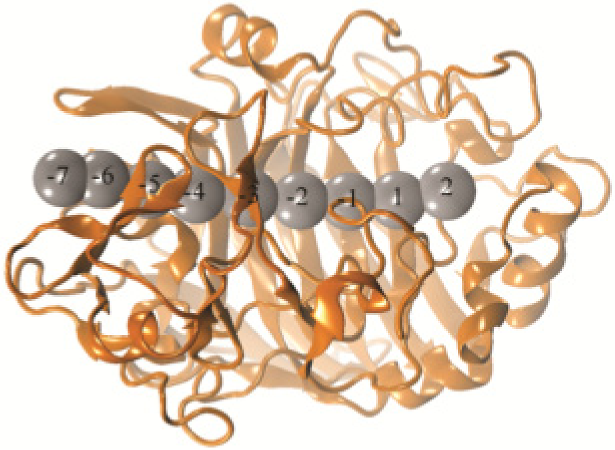

Figure 3 Coarse grained spherical representation of the macromolecules active sub-site regions. The macromolecule, Cel7A, is represented by the orange ribbon while the spherical active sub-site beads are represented by the gray spheres.

2.3. Carboxymethyl cellulose docking classification

The coarse grained subsite docking analysis was utilized on the carboxymethyl cellulose (CMC) docking poses, which can be used to indicate endo- or exo-cellulase behavior. Specifically, if the CMC docking study results in one or more conformations with a favorable binding energy residing inside the active site, the enzyme is classified as having endo-cellulase behavior, while the absence of a CMC docking conformation is indicative of exo-cellulase behavior. For all enzymes, a docked conformation is considered ‘inside’ the active tunnel/groove if it resides within the loops covering the active tunnel/groove sub-sites and has a favorable binding energy. As the length of the active groove may vary substantially between cellulase enzymes, the exact sub-site location that is considered part of the active tunnel/groove will vary between enzymes.

2.4. Cellobiose docking pattern classification

Using the coarse graining technique, two general patterns were identified for the docking of cellobiose in the catalytic tunnel/groove of cellulases. The two patterns contained either a single maximum in the distribution, or a bi-modal distribution, which are indicative of exo- and endo-activity, respectively. To be considered a uni-modal pattern, the distribution must contain a single maximum, with the number of conformations decreasing or remaining approximately constant on either side of the maximum conformation sub-site. For a bi-modal pattern, the defining characteristic is that two relatively high occupancy sub-sites must be separated by one or more sub-site(s) with a lower occupancy. One of the high occupancy sub-sites must be at least 20% greater than the intervening low occupancy sub-site(s) and the other high occupancy sub-site must be at least 2% greater than the intervening low occupancy sub-site(s). It is important to note that because the length of the active groove can vary substantially between cellulases (e.g. Cel7 has 9 sub-sites while Cel9 has 6 sub-sites), the location of the distinguishing features of the binding distributions will vary between enzymes. Representative histogram plots of the binding motifs indicative of exo-/endo-activity are shown in Figure 4.

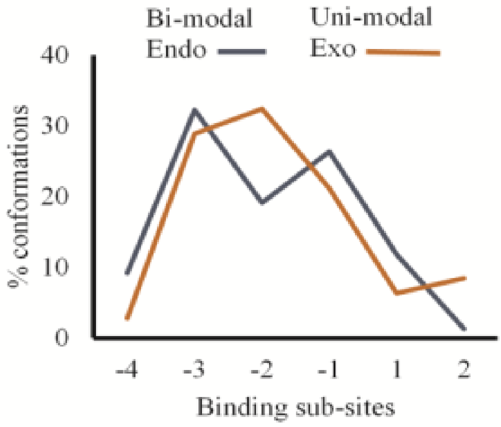

Figure 4 Representative docking modes of cellobiose with Cel5A (PDB ID: 3AZR) in orange and Cel12A (PDB ID: 1UU4) in dark blue.

2.5. Machine learning algorithm

In addition to the cellobiose docking pattern classification process it is beneficial to have an automated algorithm that can evaluate large amounts of docking data and predict the enzyme mode of action without the need for time consuming graphing and manual analysis. To this end a sparse representation based learning model was utilized to explore the association between the docking results and the experimentally determined enzymatic mode of action. The machine learning algorithm (MLA), based on our previous work [31,32], analyzed the binding patterns for cellobiose in terms of the number of binding events at each binding sub-site and/or the presence of CMC in the binding tunnel/groove for the various cellulase enzymes. This method addresses the group structure of binding events such that the learned regression model has better predictive performance for the enzymatic mode of action for cellulase enzymes (*i.e.* endo- *vs.* exo-cellulase activity).

$$\min_{\mathbf{W}} \sum_{i=1}^{n} \left\| \mathbf{W}^T \mathbf{X} - \mathbf{Y} \right\|_F^2 + \gamma_1 \left\| \mathbf{W} \right\|_{G_{2,1}} + \gamma_2 \left\| \mathbf{W} \right\|_{2,1} \qquad (1)$$

The benefit of this MLA is its ability to select features across multiple tasks by enforcing additional structure sparsity for jointly selected features across multiple tasks via a $\ell_{2,1}$-norm regularization[33,34]. In equation 1, the first term measures the regression loss, the second term couples all the regression coefficients of a group of features, and the third term penalizes the regression coefficient of each individual feature to select features across multiple learning tasks. **W** is the weight matrix that measures the relative importance of a particular piece of docking data in predicting the enzyme mode of action, γ is a trade-off parameter, and $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_n\right]$, $\mathbf{Y} = \left[\mathbf{y}_1, \ldots, \mathbf{y}_n\right]$, and $\|.\|_{G_{2,1}}$ are the proposed group $\ell_{2,1}$-norm($G_{2,1}$-norm) where matrices are boldface uppercase letters and vectors are boldface lowercase letters. 5-fold cross-validation was used for model training, validation, and predictive purposes to avoid over-fitting the model. For a thorough description of the machine learning algorithm the reader should consult the work of Wang et. al.[31,32]. It is noted that this new method is very close and motivated by our earlier work [32]. However, the MLA utilized here is redesigned to specifically solve the classification between endo- and exo- mode of action. This type of learning model naturally leads to feature (e.g. binding sites) selection, feature type selection, as well as the explicit feature relevance (e.g. how relevant a binding site is with respect to predicting the mode of action).

3. Results and Discussion

3.1. Exo-/Endo- Structural Features

Generally, cellulases from the same family have high primary sequence identity, structural similarity, and conserved catalytic residues. However, the length of the loops and their flexibility over the catalytic tunnel differ between cellulases in the same family, and it is believed that the nature of these loops plays a significant role in the different catalytic activities observed in the enzymes (*i.e.* endo- *vs.* exo-) [25,26]. To illustrate this point, the amino acid sequence and overall structures are compared for the family 7 cellulases Cel7A [20] and Cel7B, [35] which have exo- and endo- activities, respectively. On the basis of amino acid sequences, it is evident that some of the loops in the endo- Cel7B are significantly shorter than in the exo-Cel7A. Specifically, residue segments 191-204, 244-249, and 381-392 in Cel7A are missing in Cel7B, which results in structural changes that can be seen by comparing the superimposed structures of the two enzymes, Figure 5 (regions I, II and III). It is evident from Figure 5 that the loops marked I, II, and III are reduced or absent in Cel7B. The reduced or absent loop regions results in a more open groove structure around the catalytic site of Cel7B as compared to Cel7A, which is thought to allow endo- type mode of action on the cellulose surface. The presence of these loops in Cel7A leads to the formation of an active site tunnel, favoring an exo-processive mode of action. However, such structural characteristics are difficult to quantify due to the large variability in the size and position of the loop regions between enzymes.

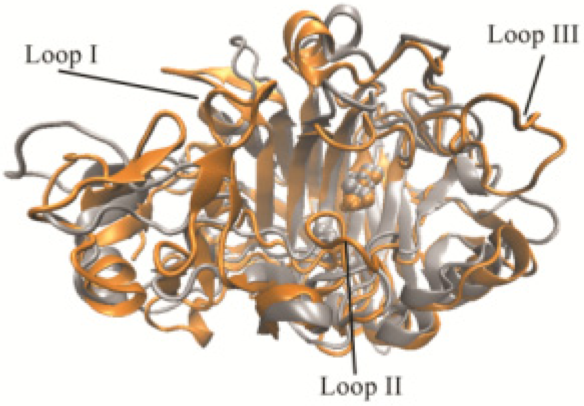

Figure 5. Superimposed structures of Cel7A (Orange) and Cel7B (Silver). PDB IDs: Ce7A: 7CEL, Cel7B: 2A39. Region I, II and III represent loops for residues 191-204, 244-249, and 381-392 respectively that are present in Cel7A and reduced or absent in Cel7B. Active site residues shown in CPK representation.

Utilizing the primary amino acid sequence and characteristic secondary structure for the 36 enzymes evaluated in this study the machine learning algorithm (MLA) predict the enzymatic mode of action with 81% accuracy. Additionally, grouping the amino acids (e.g. positively, negatively, and uncharged side chain) also resulted in an 81% predictive accuracy. The MLA analysis of the primary amino acid sequence identified loop residues at position 192 (loop I), 244 and 245 (loop II), and 391 (loop III) as possessing positive relative significance in correctly predict the enzyme mode of action. The analysis of the primary and secondary structure aspects represents a rapid method for classifying the mode of action with only the primary amino acid sequence and without the need for docking studies, kinetic experiments, or crystal structure data. Although for a more accurate predictive model additional calculations are found to be necessary.

3.2. CMC Docking

A commonly accepted experimental method for determining endo-cellulase activity is to either measure the cellulase enzymatic ability to hydrolyze CMC and/or obtain a crystal structure of the enzyme with CMC bound in the active groove [21,23]. Therefore, docking calculations were performed with CMC as a substrate to determine if such calculations could be used to indicate endo- or exo-cellulase behavior *in silico*. It was found that a favorable (i.e. negative binding energy) docked conformation of CMC within the active tunnel/groove of the enzyme can be used as a criterion for predicting exo- or endo-type cellulose behavior, which is represented in Figure 6. As the figure shows, the docked CMC is found inside the groove of Cel6B, while the docked CMC is outside the active groove of Cel6A, indicating these enzymes have endo- and exo- type behavior, respectively [21,36]. The results of the CMC docking calculations with all 36 enzymes utilized in this study are shown in Table 1. It is clear from Table 1 that the predicted mode of action is in good agreement with the experimental classification for all but three of the cellulase enzymes investigated; 2 exo-cellulases (PDB IDs: 2RFW and 2RFY) [37] and 1 endo-cellulase (PDB ID: 1QI0) [38]. The results of the CMC docking and MLA code reveal a 100% agreement with the 3 kinetically verified modes of action, 87% agreement with the 15 X-ray verified modes of action, and a 92% agreement with all 36 structures, which includes modes of actions determined by structural analysis. The results of the MLA code are found in Figure 7 where larger relative significance values correspond to larger favorable contributions to the predictive accuracy. The results from the MLA code indicate that the endo-enzymes have a larger favorable impact on the predictive model, most likely due to their larger presence in the MLA code (22 endo- *vs.* 14 exo-). This good agreement with experimental classification over a broad range of cellulase families is encouraging given the computationally inexpensive nature of these calculations.

3.3. Cellobiose Docking

Although the CMC docking results can predict endo- and exo- behavior with high accuracy, CMC is not a natural substrate of cellulase enzymes, thus it is difficult to extrapolate docking results with CMC to the identification of substrate specific enzyme characteristics. Therefore, docking calculations with cellobiose were performed, which is a better representation of the natural cellulose substrate and is the main product of cellobiohydrolase enzymes. It was identified, as shown in Figure 4, that the docking modes for cellobiose in the catalytic tunnel/groove have either a uni-modal or bi-modal shaped docking distribution. Analysis of the docking results show that 12 of 14 (86%) enzymes that have been experimentally classified as exo-cellulase gave a uni-modal shaped docking distribution, while 17 of 22 (77%) enzymes experimentally classified as an endo-cellulase gave bi-modal shaped docking distributions for an overall 81% accuracy in identifying the enzymatic mode of action (100% kinetic and 87% X-ray).

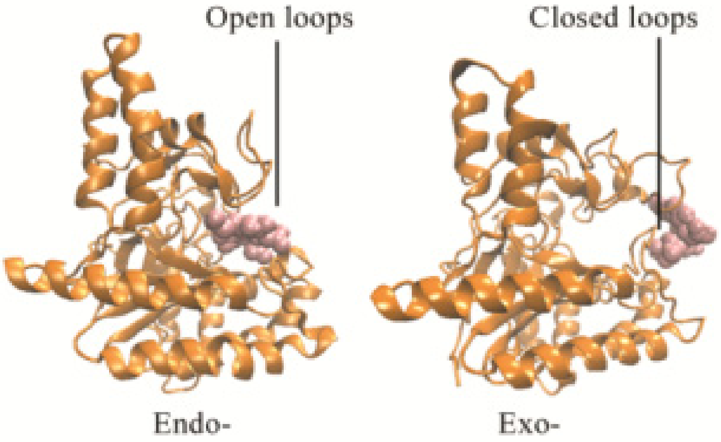

Figure 6. Typical docking poses for CMC (in pink color) with an endo-cellulase (left, Cel6B) and an exo-cellulase (right, Cel6A). PDB IDs: Cel6B: 1DYS, Cel6A: 1QK2.

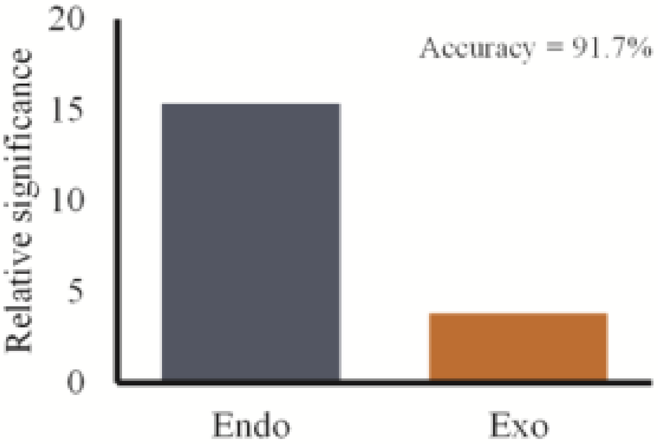

Figure 7. Machine learning algorithm results for the carboxymethyl cellulose docking calculation. A larger relative significance indicates a larger favorable contribution to the predictive model.

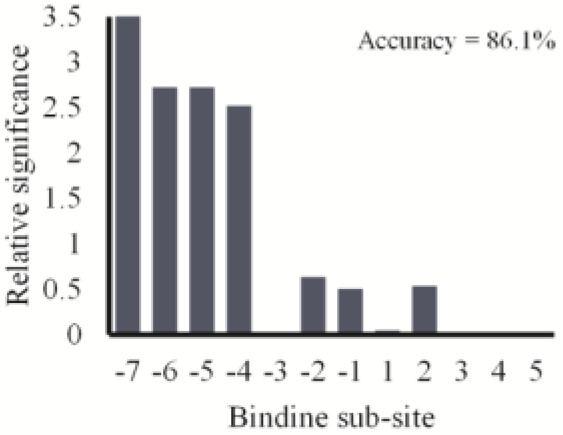

Figure 8. Machine learning algorithm results for the carboxymethyl cellulose docking calculation. A larger relative significance indicates a larger favorable contribution to the predictive model.

Along with the CMC results, the docking results with cellobiose as a substrate are also given in Table 1. For the endo-enzymes, five systems gave uni-modal/exo-distributions, which are accounted for by the closed nature of the loop region in the X-ray structure. The ability of cellobiose docking to predict the enzymatic mode of action (29/36, 81%) is found to be 11% lower than the CMC docking results. In general, the average binding energies at the various sub-sites in the exo-cellulase enzymes is almost equal or contain one minima at the sub-site with an increased number of conformations in 12 of 14 exo-enzymes, but this trend is not seen in the endo-cellulase enzymes. In 12 of 22 endo-cellulases, the two average binding energies at sub-sites with increased number of conformations are separated by a relatively more unfavorable binding energy with a low number of binding conformations.

Utilizing the MLA code and the cellobiose docking results for all available sub-binding sites yields an overall predictive accuracy of 86%. In addition, the MLA code is capable of indicating the relative importance for each sub-binding site as seen in Figure 8. The MLA code results indicate that the most significant contributors to the predictive accuracy are the sub-binding sites -7 to -4. These particular binding site, while found in both endo- and exo- type enzymes, are more prevalent in the exo-cellulases. The increased occurrence of these substrate binding sites in the exo-cellulases is commonly associated with their threading of the cellulose substrate into the active site tunnel and processive action. These sites are not as critical in the endo-cellulase enzymes due to their role in randomly clipping cellulose strands and the absence of processive ability. It is noted that some endo-cellulase enzymes have been reported to have exo-processive activity, although this is significantly reduced when compared to exo-celluase enzymes. Evaluation of Figure 8 also indicates that sub-binding sites -2, -1, and +2 are important in classifying endo- or exo-activity. These binding sites, while close to the catalytic site (*i.e.* sub-binding site -1 and +1), also reside under loop II, which has been implicated in defining the enzyme mode of action.

Table 1 The predicted exo- and endo- cellulose behavior using docking analysis and comparison to experimental findings.

			Experimental	Predicted	
	PDB ID	Source (Bacterial/Fungal)		CMC	Cellobiose
Cel7A	1CEL	*Trichoderma Reesei*	Exo[39]	Exo	Exo
	1Q9H	*Talaromyces Emersonii*	Exo[40]	Exo	Exo
	2V3R	*Trichoderma Reesei*	Exo	Exo	Exo
	2XSP	*Heterobasidion Annosum*	Exo[41]	Exo	Exo
	7CEL	*Trichoderma Reesei*	Exo[20]	Exo	Exo
Cel7B	1EG1	*Trichoderma Reesei*	Endo[42]	Endo	Endo
	2RFW	*Melanocarpus Albomyces*	Exo[37]	**Endo**	**Endo**
	2RFY	*Melanocarpus Albomyces*	Exo[37]	**Endo**	**Endo**
	2A39	*Humicola Iinsolens*	Endo[35]	Endo	Endo
Cel7D	1GPI	*Phanerochaete Chrysosporium*	Exo[43]	Exo	Exo
	1Z3T	*Phanerochaete Chrysosporium*	Exo[44]	Exo	Exo
Cel6A	1BVW	*Humicola Insolens*	Exo[45]	Exo	Exo
	1QK2	*Trichoderma Reesei*	Exo[36]	Exo	Exo
	1OCB	*Humicola Insolens*	Exo[46]	Exo	Exo
	2BOD	*Thermobifida Fusca*	Endo[47]	Endo	Endo
	2BVW	*Humicola Insolens*	Exo[48]	Exo	Exo
Cel6B	1DYS	*Humicola Insolens*	Endo[21]	Endo	Endo
Cel6C	3A9B	*Coprinopsis Cinerea*	Exo[49]	Exo	Exo
Cel5A	1QIO	*Bacillus Agaradhaerens*	Endo[38]	**Exo**	**Exo**
	2A3H	*Bacillus Agaradhaerens*	Endo[50]	Endo	Endo
	3AZR	*Thermotoga Maritima*	Endo[51]	Endo	Endo
	3QR3	*Trichoderma Reesei*	Endo[52]	Endo	**Exo**
Cel9A	3EZ8	*Alicyclobacillus Acidocaldarius*	Endo[53]	Endo	Endo
	3H2W	*Alicyclobacillus Acidocaldarius*	Endo[54]	Endo	Endo
Cle9G	1GA2	*Clostridium Cellulolyticum*	Endo[55]	Endo	Endo
	1K72	*Clostridium Cellulolyticum*	Endo[55]	Endo	Endo
Cel9M	1IA6	*Clostridium Cellulolyticum*	Endo[56]	Endo	Endo
	1IA7	*Clostridium Cellulolyticum*	Endo[56]	Endo	Endo
Cel12A	1UU4	*Humicola Grisea*	Endo[57]	Endo	Endo
	1H8V	*Trichoderma Reesei*	Endo[58]	Endo	**Exo**
	1OLR	*Humicola Grisea*	Endo[59]	Endo	**Exo**
	1W2U	*Humicola Grisea*	Endo[57]	Endo	**Exo**
Cel44A	2EJ1	*Clostridium Thermocellum*	Endo[60]	Endo	Endo
Cel45A	1L8F	*Melanocarpus Albomyces*	Endo[61]	Endo	Endo
	1OA7	*Melanocarpus Albomyces*	Endo[62]	Endo	Endo
Ce448F	1FAE	*Clostridium Cellulolyticum*	Endo[63]	Endo	Endo

Similar to the CMC dockings, the cellobiose docking method described here seems to provide a quick and accurate method for determining if an enzyme has predominately endo- or exo- characteristics. Furthermore, the cellobiose docking results can provide insights into unique enzyme structure-function relationships leading to the observed endo-/exo- behavior, as well as serving to predict other useful enzyme characteristics which may warrant further study, such as inhibition patterns, and residue interactions critical to substrate binding. Also, further investigation of the cellobiose docking results lends insight into why the predicted mode of action was incorrect for a select few enzymes. This is examined in greater detail in the following Cel7 and Cel6 family specific sections.

3.4. Family Cel7

Examining the overall cellobiose docking distributions provides useful insights into the different binding mechanisms utilized by endo- and exo-cellulase enzymes. The docking modes for cellobiose at the binding sub-sites of two Cel7 enzymes are depicted in Figure 9. It is evident that Cel7A possess a uni-modal shaped docking histogram that is in agreement with the experimentally classified exo-cellulase mode of action[20]. Inspection of the docking results indicate a gradual increase in the number of docked conformations at each sub-site, starting from the active site tunnel entrance (-7 and -6 sites) and ending at sub-site -4. The docking distribution reaches a maximum at sub-site -4 and then starts a gradual decrease until the catalytic active site is reached (-1 and +1).

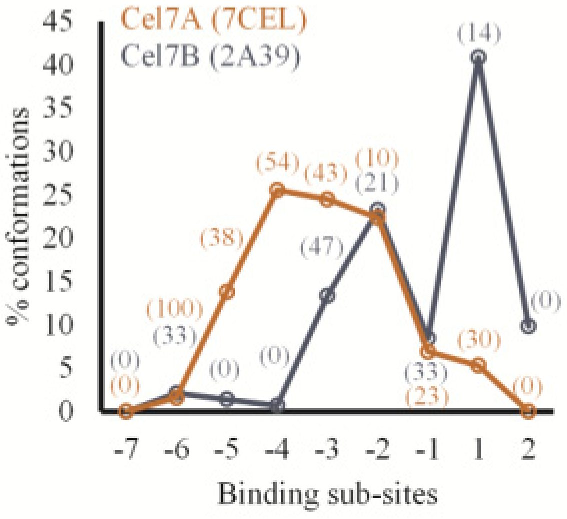

Figure 9 Number of docking conformations for cellobiose at sub-sites in the active site tunnel of Cel7A (orange), and Cel7B (blue). The percentages of conformations that are docked over neighboring subsites are given in brackets. PDB IDs: Cel7A: 7CEL, Cel7B: 2A39.

The uni-modal shape of the docking distribution is due to the increasingly favorable interactions from one site to the next as the substrate is threaded from the active site tunnel entrance (-7 sub-site) to sub-sites deep in the catalytic tunnel (-4 site). The uni-modal binding pattern suggests that in order to stop the cellulose strand from exiting the active site tunnel and re-annealing to the micro fibril substrate, the enzyme has evolved a ratchet-like mechanism to coax the substrate strand into the active site tunnel. It is hypothesized that this feature is necessary for the efficient threading of the substrate into the active site tunnel and the subsequent retention of the substrate for a long enough time to facilitate the observed processive nature of the enzyme. The overall shape of the docking distribution is clearly uni-modal and reveals a general structure-function relationship that illuminates how these particular enzymes thread the substrate into the active site tunnel and move in a processive fashion on the surface.

The decrease in the number of docking conformations between sub-sites -3 to -1 indicate the substrate in this region has a reduced amount of enzyme stabilization or enhanced conformational strain. The decreased enzyme stabilization near the catalytic sub-sites (-1 and +1) may allow the substrate to change from the standard chair configuration to the skew boat configuration, which is found to be necessary for catalysis [36,64]. A distortion of glucose from chair to half-chair, boat, or skew boat has also been observed in quantum mechanical/molecular mechanical (QM/MM) and QM calculations carried out on the glycosidic bond hydrolysis of a cellulose chain in Cel7A [65]. The increased flexibility would also be in line with the hypothesis that flexibility (usually enzyme flexibility) is correlated with increased enzyme activity [66].

While the histogram for Cel7A has a uni-modal shape, the docking histogram shape for the endo-cellulase Cel7B is significantly different. The most notable difference is that the maximum number of conformations is almost equally split between sub-sites -2 and +1, creating a bimodal distribution. This distribution indicates the enzyme may initially absorbs onto the substrate through strong interactions on the reducing side of the active site, which could temporarily ‘lock’ the enzyme onto the substrate, followed by the attachment of the +1 sub-site on the non-reducing side to the surface or vice versa. Such a binding mechanism would allow the entire active groove to absorb onto a cellulose sheet and bring the active site into close proximity with the cellulose polymer. The lower average binding energy for cellobiose conformations at product sub-sites in Cel7B, as compared to Cel7A, suggests that more cellobiose conformations docked at product sub-site.

Beyond providing insights into general substrate binding mechanisms, further examination of the docking distributions can provide details on key residue interactions. A complete list of amino acids interacting with cellobiose, and the number of conformations at each sub-sites are given in Table S2 and Figure S5A & 5B of the Supporting Information. Examination of the tables reveals four critical Trp residues at the -7, -4, -2 and the +1 sub-sites of the exo-cellulase Cel7A. These particular Trp residues were previously identified and shown to play an important role in the binding of cellulose/cellobiose units [25]. The importance of these Trp residues was also seen in all atomistic molecular dynamics simulations on aromatic-carbohydrate interactions in a processive cellulose [67]. The ability of the current docking calculations to identify these residues as critical stabilizing interactions is encouraging given the docking calculations are orders of magnitude less computationally intensive than all atom molecular dynamics simulations.

Given the polymeric nature of cellulose, an important characteristic of cellulose binding is the ability of a cellulase enzyme to bind across multiple sub-sites. In the current study, this idea translates to a cellobiose having the ability to bind across two sub-sites, thereby maximizing the binding energy. The values shown in the parentheses in Figure 9 indicate the percentage of cellobiose conformations at the sub-site that span (i.e. bridge or bind) across two sub-sites, with the first cellobiose glucose unit at the site shown in Figure 9 (e.g. sub-site -5) and the second glucose unit binding to the sub-site to the right (e.g. sub-site -4). As the values indicate, with the exception of the -7 entrance site, all sub-sites contain conformations that span across two consecutive sites. It is thought that this continuous nature of Cel7A aids in the exo-processive action of the enzyme by maximizing the enzyme contact with the substrate.

Comparing the binding modes of exo- and endo-cellulase Cel7 enzymes can also serve to predict inhibition patterns. For example, the spatial distribution for the total number of cellobiose conformations docked at each sub-site for Cel7A (PDB ID: 7CEL [20]) and Cel7B (PDB ID: 2A39[35]) are shown in Figure 10. The increased number of docking conformations found at the Cel7B sub-sites +1 and +2 (product) as compared to the Cel7A sub-site +1 and +2 (product) indicates Cel7B has increased product inhibition characteristics as compared to Cel7A. This is also seen in the docking histogram in Figure 9. This general trend holds for all Cel7B enzymes, which can be seen by examining Figure S5A and S5B in the Supporting Information. As the figures indicate, in comparison to Cel7A enzymes, all Cel7B enzymes have an increased amount of docking poses at sub-sites +1 and 2, indicating Cel7B endo-cellulases have elevated product inhibition patterns when compared to Cel7A exo-cellulases. This docking observation is in agreement with Voutilainen *et. al.* kinetic studies of cellobiose inhibition on cellulase enzymes that found the inhibition constant of Cel7B (K_i = 6 μM) to be three times less than Cel7A (K_i = 19 μM)[68]. In addition, experimental measurements of Cel7A and Cel7D indicate K_d values of 20 mM and 115 mM for cellobiose respectively, indicating the exo-cellulase enzymes do not suffer from severe product inhibition, again in agreement with the docking studies [44].

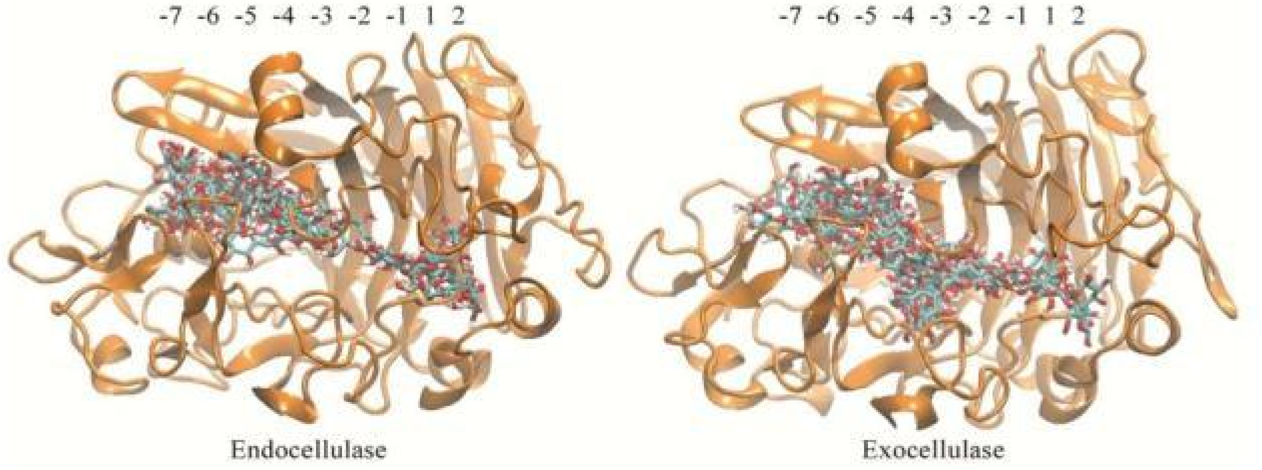

Figure 10. Comparison between exo- and endo-cellulase cellobiose docking poses. (Right) Cellobiose in licorice representation docked to exo-cellulase Cel7A (PDB ID: 7CEL), and (Left) an endo-cellulase, Cel7B (PDB ID: 2A39).

In addition to inhibition patterns, our studies can also reveal binding affinity trends. Specifically, Cel7B (PDB: 2A39 [35]) has a significantly increased number of docking poses at the reducing end sub-sites as compared to Cel7A (PDB: 7CEL) [20], Cel7D (PDB: 1GPI) [43], Cel7B (PDB: 1EG1)[42], and Cel7B (PDB: 2RFW) (see Figure 9 and Figure S5A &5B) [37]. Furthermore, the average binding energies for cellobiose conformations at reducing end sub-sites are lower in Cel7B (PDB: 1EG1) compared to the non-reducing sub-sites in other Cel7B enzymes (PDB: 1EG1 and 2A39). Specifically, the numbers of cellobiose conformations for 1EG1 are lower and have a higher binding energy as compared to 2A39. In general, these results indicate that Cel7B (2A39) has an increased affinity for the microfibril substrate. This can be experimentally observed as an increase in the equilibrium adsorption constant, K_{ads}, an enhanced amount of time residing on the microfibril, and/or an enhanced product inhibition pattern. Langmuir adsorption measurements by Ding et. al. on *T. reesei* Cel7A and Cel7B have shown that Cel7B has 80% higher adsorption rates on amorphous phosphoric acid-swollen cellulose (PACS) than Cel7A, confirming that the docking experiments correctly identify enzymes with increased substrate affinity [69].

3.5. Family Cel6 and Other Families

The cellobiose docking distributions of Cel6 cellulases are given in Figure 11. As opposed to the Cel7 family, the Cel6 family cleaves cellulose strands from the non-reducing ends, thus the cellulose enters the active tunnel from the +4 sub-site instead of the -3 subsite. The binding pattern for Cel6A[36] shows a uni-modal like structure similar to Cel7A, with the maximum number of conformations residing at the +2 or +1 subsites for Cel6A. Cel6B[21] shows a bi-modal distribution similar to that of Cel7B with the maximum distributions located at -2, +2, and +3 sub-sites.

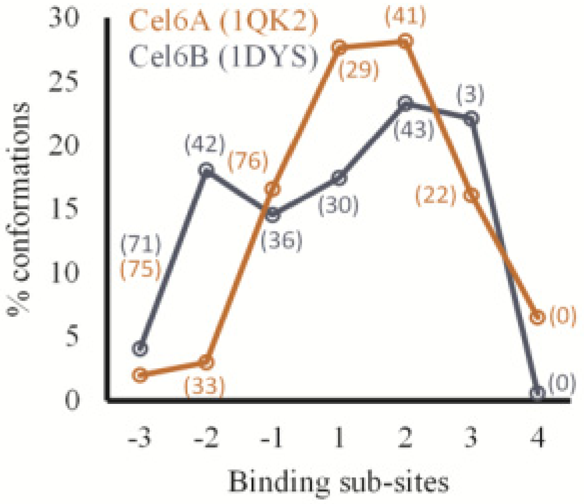

Figure 11. Number of docking conformations for cellobiose at sub-binding sites of Cel6A (orange), and Cel6B (blue). The percentage of conformations is docked over neighboring subsites given in brackets. PDB IDs: Cel6A: 1QK2, Cel6B: 1DYS.

Similar to the CMC docking results, the predicted mode of binding for Cel6C (in Figure S5D of the Supporting Information) is exo-, while experimental evidence shows this enzyme possesses both endo- and exo-type activity. A common trend seen in both docking studies is that certain enzymes previously classified as endo- or 'hybrid' endo-/exo- results in an exo-type docking classification. The reasoning for this anomaly appears to involve the loops flanking the active groove, which are believed to be highly flexible in endo-cellulases. This flexibility may be necessary so that the endo-cellulase can adopt an open conformation to adsorb its entire active groove onto a cellulose sheet. Once adsorbed onto the cellulose sheet, the loops then adopt a closed conformation, clamping the active groove onto the cellulose sheet such that hydrolysis can occur. A consequence of this flexibility is that endo-cellulase enzymes may possess a varying degree of exo-type activity if the loops are capable of closing. That is to say, an endo-cellulase may bind to a cellulose polymer, adopt the closed loop conformation, and then subsequently spend longer in this conformation than is necessary for a single catalytic event. This would allow the enzyme to proceed with a certain amount of exo-processive action. It is hypothesized that the degree of flexibility and the stability of the closed conformation is related to the degree of exo-cellulase activity seen in predominantly endo-cellulase enzymes.

From the above discussion, it is apparent that the static conformation seen in the crystal structure may be in two very distinct conformations for an endo-cellulase enzyme. One is a more open structure that allows the enzyme to absorb onto a microfibril, and is unique to endo-cellulases. The other is a closed conformation for performing hydrolysis on a polymer chain, which is common to both endo- and exo-cellulases. Thus, it is hypothesized that the reason Cel6C and certain other endo- or endo-/exo-cellulases studied here resulted in an exo-type docking classification is due to the static crystal structure being in a more closed conformation. That is to say, the enzyme is in a conformation that is indicative of a catalytic event that is the same in both endo- and exo-cellulases enzymes, instead of a conformation representative of the unique substrate binding performed by endo-cellulases. Given that Cel6C has been found to possess substantial endo- and exo-activity, it is expected that the enzyme when bound to substrate would adopt the closed conformation, resulting in docking and MLA classification as an exo-. Furthermore, the crystal structure of Cel6C used here was solved with a cellobiose bound in the active site, which again will induce or increase the probability of the enzyme adopting a closed conformation for the subsequent catalytic event. Overall, given that Cel6C has been experimentally shown to have substantial endo- and exo- activity coupled to the fact that the crystal structure had a cellobiose substrate bound in the active site, it is reasonable to assume that the enzyme is in a closed conformation, leading to the observed exo-type docking motif. The list of amino acids interacting with cellobiose, and the number of conformations at each sub-sites are given in Table S3 and Figure S5C & 5D of the Supporting Information.

The docking modes for cellobiose at each sub-site for the other cellulase families studied here; Cel5, Cel9, Cel12, Cel44, Cel45, and Cel48, are discussed in specific family sections and Figure S5 of the Supporting Information, and the predicted modes of action from CMC and cellobiose docking are given in Table 1. To verify the reproducibility of the docking experiments, eight dockings were repeated four times. All of the repeating docking calculations resulted in similar docking patterns as seen in the original docking calculations with CMC and cellobiose ligands. These results indicate that docking calculations are a fast, reproducible tool to identify the mode of action for cellulase enzymes.

4. Conclusion

Both the CMC and cellobiose docking methods reported here can be used to identify an enzyme as an endo- or exo-cellulase. CMC as a docking substrate is found to identify the endo- or exo-behavior for all enzymes in this study except three. A second set of calculations were conducted using cellobiose as a docking ligand to directly probe the substrate interactions. The binding distributions have been classified into two families; distributions with a single maximum and bi-modal distribution that are indicative of exo- and endo-cellulase enzymes, respectively. The cellobiose binding motifs and MLA code indicates a comparable accuracy as the CMC docking with 86% and 92% in identifying endo- or exo- behavior, respectively. These results are an improvement over analyzing the primary and secondary structure components, which resulted in the MLA code predicting the enzymatic mode of action with 81% accuracy. These results indicate a fast and computationally inexpensive method for identifying the nature of the cellulase enzyme. Furthermore, the docking results were able to identify important enzyme characteristics in agreement with previous studies, indicating autodock calculations and the MLA code can serve as a cheap, efficient means of identifying cellulase characteristics that may warrant further study in future cellulase bioengineering efforts.

Acknowledgments

This research was supported by the National Science Foundation (NSF grant CBET-1337044). The authors would like to acknowledge the support from Colorado School of Mines Campus Computing, Communications and Information Technologies and helpful communications with Dr. Shubham Vyas.

Conflict of interest

The authors declare that there are no conflicts of interest related to this study.

References

1. Bhat MK, Bhat S (1997) Cellulose degrading enzymes and their potential industrial applications. *Biotechnol Advances* 15: 583-620.
2. Himmel ME, Ding SY, Johnson DK, et al. (2007) Biomass recalcitrance: Engineering plants and enzymes for biofuels production. *Science* 315: 804-807.
3. Himmel ME (2008) Biomass Recalcitrance: Deconstructing the Plant Cell Wall for Bioenergy. Wiley Blackwell.
4. Updegraff DM (1969) Semimicro determination of cellulose in biological materials. *Anal Biochem* 420-424.
5. Wang LS, Zhang YZ, Gao PJ (2008) A novel function for the cellulose binding module of cellobiohydrolase I. *Sci Chin Series C-Life Sci* 51: 620-629.
6. Edgar KJ, Buchanan CM, Debenham JS, et al. (2001) Advances in cellulose ester performance and application. *Prog Polym Sci* 26: 1605-1688.
7. Ragauskas AJ, Williams CK, Davison BH, et al. (2006) The path forward for biofuels and biomaterials. *Science* 311: 484-489.
8. Lynd LR, Laser MS, Brandsby D, et al. (2008) How biotech can transform biofuels. *Nat Biotechnol* 26: 169-172.
9. Schubert C (2006) Can biofuels finally take center stage? *Nat Biotechnol* 24: 777-784.
10. Andre G, Kanchanawong P, Palma R, et al. (2003) Computational and experimental studies of the catalytic mechanism of Thermobifida fusca cellulase Cel6A (E2). *Protein Eng* 16: 125-134.

11. Wolfenden R, Yuan Y (2008) Rates of spontaneous cleavage of glucose, fructose, sucrose, and trehalose in water, and the catalytic proficiencies of invertase and trehalas. *J Am Chem Soc* 130: 7548.
12. Nishiyama Y, Sugiyama J, Chanzy H, et al. (2003) Crystal structure and hydrogen bonding system in cellulose 1(alpha), from synchrotron X-ray and neutron fiber diffraction. *J Am Chem Soc* 125: 14300-14306.
13. CarleUrioste JC, EscobarVera J, ElGogary S, et al. (1997) Cellulase induction in Trichoderma reesei by cellulose requires its own basal expression. *J Biol Chem* 272: 10169-10174.
14. Bayer EA, Chanzy H, Lamed R, et al. (1998) Cellulose, cellulases and cellulosomes. *Curr Opin Structl Biol* 8: 548-557.
15. Boisset C, Fraschini C, Schulein M, et al. (2000) Imaging the enzymatic digestion of bacterial cellulose ribbons reveals the endo character of the cellobiohydrolase Cel6A from Humicola insolens and its mode of synergy with cellobiohydrolase Cel7A. *Appl Environ Microbiol* 66: 1444-1452.
16. Cantarel BL, Coutinho PM, Rancurel C, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* 37: D233-D238.
17. Wilson DB (2009) Cellulases and biofuels. *Curr Opin Biotechnol* 20: 295-299.
18. Davies G, Henrissat B (1995) Structures and mechanisms of glycosyl hydrolases. *Structure* 3: 853-859.
19. Henrissat B, Davies G (1997) Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol* 7: 637-644.
20. Divne C, Stahlberg J, Teeri TT, et al. (1998) High-resolution crystal structures reveal how a cellulose chain is bound in the 50 angstrom long tunnel of cellobiohydrolase I from Trichoderma reesei. *J Mol Biol* 275: 309-325.
21. Davies GJ, Brzozowski AM, Dauter M, et al. (2000) Structure and function of Humicola insolens family 6 cellulases: structure of the endoglucanase, Cel6B, at 1.6 angstrom resolution. *Biochem J* 348: 201-207.
22. Yang B, Willies DM, Wyman CE (2006) Changes in the enzymatic hydrolysis rate of avicel cellulose with conversion. *Biotechnol Bioeng* 94: 1122-1128.
23. Dashtban M, Maki M, Leung KT, et al. (2010) Cellulase activities in biomass conversion: measurement methods and comparison. *Crit Rev Biotechnol* 30: 302-309.
24. Teeri TT (1997) Crystalline cellulose degradation: New insight into the function of cellobiohydrolases. *Trends Biotechnol* 15: 160-167.
25. Rouvinen J, Bergfors T, Teeri T, et al. (1990) 3-dimensional structure of cellobiohydrolase-II from trichoderma-reesei. *Science* 249: 380-386.
26. Meinke A, Damude HG, Tomme P, et al. (1995) Enhancement of the endo-beta-1,4-glucanase activity of an exocellobiohydrolase by deletion of a surface loop. *J Biol Chem* 270: 4383-4386.
27. Kurasin M, Valjamae P (2011) Processivity of Cellobiohydrolases Is Limited by the Substrate. *J Biol Chem* 286: 169-177.
28. Breyer WA, Matthews BW (2001) A structural basis for processivity. *Protein Sci* 10: 1699-1711.
29. Morris GM, Huey R, Lindstrom W, et al. (2009) AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J Comput Chem* 30: 2785-2791.
30. Pettersen EF, Goddard TD, Huang CC, et al. (2004) UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* 25: 1605-1612.

31. Wang H, Nie F, Huang H, et al. (2012) Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics* 28: 229-237.
32. Wang H, Nie F, Huang H, et al. (2012) Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics* 28: i127-136.
33. Lee S, Zhu J, Xing EP (2010) Adaptive Multi-Task Lasso: with application to eQTL detection. *Adv Neural Informat Process Syst*: 1306-1314.
34. Puniyani K, Kim S, Xing EP (2010) Multi-population GWA mapping via multi-task regularized regression. *Bioinformatics* 26: i208-216.
35. Mackenzie LF, Sulzenbacher G, Divne C, et al. (1998) Crystal structure of the family 7 endoglucanase I (Cel7B) from Humicola insolens at 2.2 angstrom resolution and identification of the catalytic nucleophile by trapping of the covalent glycosyl-enzyme intermediate. *Biochem J* 335: 409-416.
36. Zou JY, Kleywegt GJ, Stahlberg J, et al. (1999) Crystallographic evidence for substrate ring distortion and protein conformational changes during catalysis in cellobiohydrolase Cel6A from Trichoderma reesei. *Struct Fold Des* 7: 1035-1045.
37. Parkkinen T, Koivula A, Vehmaanpera J, et al. (2008) Crystal structures of Melanocarpus albomyces cellobiohydrolase Ce17B in complex with cello-oligomers show high flexibility in the substrate binding. *Protein Sci* 17: 1383-1394.
38. Varrot A, Schulein M, Davies GJ (2000) Insights into ligand-induced conformational change in Cel5A from Bacillus agaradhaerens revealed by a catalytically active crystal form. *J Mol Biol* 297: 819-828.
39. Divne C, Stahlberg J, Reinikainen T, et al. (1994) The 3-dimensional crystal-structure of the catalytic core of cellobiohydroase-I from trichoderma-reesei. *Science* 265: 524-528.
40. Grassick A, Murray PG, Thompson R, et al. (2004) Three-dimensional structure of a thermostable native cellobiohydrolase, CBHIB, and molecular characterization of the cel7 gene from the filamentous fungus, Talaromyces emersonii. *Eur J Biochem* 271: 4495-4506.
41. Momeni MH, Payne CM, Hansson H, et al. (2013) Structural, Biochemical, and Computational Characterization of the Glycoside Hydrolase Family 7 Cellobiohydrolase of the Tree-killing Fungus Heterobasidion irregulare. *J Biol Chem* 288: 5861-5872.
42. Kleywegt GJ, Zou JY, Divne C, et al. (1997) The crystal structure of the catalytic core domain of endoglucanase I from Trichoderma reesei at 3.6 angstrom resolution, and a comparison with related enzymes. *J Mol Biol* 272: 383-397.
43. Munoz IG, Ubhayasekera W, Henriksson H, et al. (2001) Family 7 cellobiohydrolases from Phanerochaete chrysosporium: Crystal structure of the catalytic module of Cel7D (CBH58) at 1.32 angstrom resolution and homology models of the isozymes. *J Mol Biol* 314: 1097-1111.
44. Ubhayasekera W, Munoz IG, Vasella A, et al. (2005) Structures of Phanerochaete chrysosporium Cel7D in complex with product and inhibitors. *Febs J* 272: 1952-1964.
45. Varrot A, Hastrup S, Schulein M, et al. (1999) Crystal structure of the catalytic core domain of the family 6 cellobiohydrolase II, Cel6A, from Humicola insolens, at 1.92 angstrom resolution. *Biochem J* 337: 297-304.
46. Varrot A, Frandsen TP, von Ossowski I, et al. (2003) Structural basis for ligand binding and processivity in cellobiohydrolase Cel6A from Humicola insolens. *Structure* 11: 855-864.

47. Larsson AM, Bergfors T, Dultz E, et al. (2005) Crystal structure of Thermobifida fusca endoglucanase Cel6A in complex with substrate and inhibitor: The role of tyrosine Y73 in substrate ring distortion. *Biochemistry* 44: 12915-12922.

48. Varrot A, Schulein M, Davies GJ (1999) Structural changes of the active site tunnel of Humicola insolens cellobiohydrolase, Cel6A, upon oligosaccharide binding. *Biochemistry* 38: 8884-8891.

49. Liu Y, Yoshida M, Kurakata Y, et al. (2010) Crystal structure of a glycoside hydrolase family 6 enzyme, CcCel6C, a cellulase constitutively produced by Coprinopsis cinerea. *Febs J* 277: 1532-1542.

50. Davies GJ, Dauter M, Brzozowski AM, et al. (1998) Structure of the Bacillus agaradherans family 5 endoglucanase at 1.6 angstrom and its cellobiose complex at 2.0 angstrom resolution. *Biochemistry* 37: 1926-1932.

51. Wu TH, Huang CH, Ko TP, et al. (2011) Diverse substrate recognition mechanism revealed by Thermotoga maritima Cel5A structures in complex with cellotetraose, cellobiose and mannotriose. *Biochim Et Biophys Acta-Proteins and Proteomics* 1814: 1832-1840.

52. Lee TM, Farrow MF, Arnold FH, et al. (2011) A structural study of Hypocrea jecorina Cel5A. *Protein Sci* 20: 1935-1940.

53. Pereira JH, Sapra R, Volponi JV, et al. (2009) Structure of endoglucanase Cel9A from the thermoacidophilic Alicyclobacillus acidocaldarius. *Acta Crystallogr Sect D-Biological Crystallogr* 65: 744-750.

54. Eckert K, Vigouroux A, Lo Leggio L, et al. (2009) Crystal Structures of A. acidocaldarius Endoglucanase Cel9A in Complex with Cello-Oligosaccharides: Strong-1 and-2 Subsites Mimic Cellobiohydrolase Activity. *J Mol Biol* 394: 61-70.

55. Mandelman D, Belaich A, Belaich JP, et al. (2003) X-ray crystal structure of the multidomain endoglucanase Cel9G from Clostridium cellulolyticum complexed with natural and synthetic cello-oligosaccharides. *J Bacteriol* 185: 4127-4135.

56. Parsiegla G, Belaich A, Belaich JP, et al. (2002) Crystal structure of the cellulase Ce19M enlightens structure/function relationships of the variable catalytic modules in glycoside hydrolases. *Biochemistry* 41: 11134-11142.

57. Sandgren M, Berglund GI, Shaw A, et al. (2004) Crystal complex structures reveal how substrate is bound in the -4 to the +2 binding sites of Humicola grisea cel12A. *J Mol Biol* 342: 1505-1517.

58. Sandgren M, Shaw A, Ropp TH, et al. (2001) The X-ray crystal structure of the Trichoderma reesei family 12 endoglucanase 3, Cel12A, at 1.9 angstrom resolution. *J Mol Biol* 308: 295-310.

59. Sandgren M, Gualfetti PJ, Paech C, et al. (2003) The Humicola grisea Cell2A enzyme structure at 1.2 angstrom resolution and the impact of its free cysteine residues on thermal stability. *Protein Sci* 12: 2782-2793.

60. Kitago Y, Karita S, Watanabe N, et al. (2007) Crystal structure of Cel44A, a glycoside hydrolase family 44 endoglucanase from Clostridium thermocellum. *J Biol Chem* 282: 35703-35711.

61. Valjakka J, Rouvinen J (2003) Structure of 20K endoglucanase from Melanocarpus albomyces at 1.8 angstrom resolution. *Acta Crystallogr Sect D-Biol Crystallogr* 59: 765-768.

62. Hirvonen M, Papageorgiou AC (2003) Crystal structure of a family 45 endoglucanase from Melanocarpus albomyces: Mechanistic implications based on the free and cellobiose-bound forms. *J Mol Biol* 329: 403-410.

63. Parsiegla G, Reverbel-Leroy C, Tardif C, et al. (2000) Crystal structures of the cellulase Ce148F in complex with inhibitors and substrates give insights into its processive action. *Biochemistry* 39: 11238-11246.
64. Sulzenbacher G, Driguez H, Henrissat B, et al. (1996) Structure of the Fusarium oxysporum endoglucanase I with a nonhydrolyzable substrate analogue: Substrate distortion gives rise to the preferred axial orientation for the leaving group. *Biochemistry* 35: 15280-15287.
65. Li JH, Du LK, Wang LS (2010) Glycosidic-Bond Hydrolysis Mechanism Catalyzed by Cellulase Cel7A from Trichoderma reesei: A Comprehensive Theoretical Study by Performing MD, QM, and QM/MM Calculations. *J Phys Chem B* 114: 15261-15268.
66. Mine Y, Fukunaga K, Itoh K, et al. (2003) Enhanced enzyme activity and enantioselectivity of lipases in organic solvents by crown ethers and cyclodextrins. *J Biosci Bioeng* 95: 441-447.
67. Payne CM, Bomble Y, Taylor CB, et al. (2011) Multiple Functions of Aromatic-Carbohydrate Interactions in a Processive Cellulase Examined with Molecular Simulation. *J Biol Chem* 286: 41028-41035.
68. Voutilainen SP, Boer H, Alapuranen M, et al. (2009) Improving the thermostability and activity of Melanocarpus albomyces cellobiohydrolase Cel7B. *Appl Microbiol Biotechnol* 83: 261-272.
69. Ding H, Xu F (2004) Lignocellulolse Biodegradation, Chapter 9, *ACS* 154-169.

Supplementary

Supporting Information is available, which contains **Table S1** Dimensions of the grid boxes, **Table S2** List of amino acids interacting with cellobiose at each sub-site of Cel7A and Cel7B, **Table S3** List of amino acids interacting with cellobiose at each sub-site of Cel6A and Cel6B, **Figure S1** Family Cel5, **Figure S2** Family Cel9, **Figure S3** Family Cel45, **Figure S4** Family Cel44A and Cel48F, and **Figure S5**. Percentage of cellobiose conformations docked at sub-binding sites of cellulase enzymes.