Endoglucanase Peripheral Loops Facilitate Complexation of Glucan Chains on Cellulose via Adaptive Coupling to the Emergent Substrate Structures

Yuchun Lin,[†,‡] Gregg T. Beckham,[§,⊥] Michael E. Himmel,[‖] Michael F. Crowley,[‖] and Jhih-Wei Chu*[,†,▽,○]

[†]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, California 94704, United States

[‡]State Key Laboratory of Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu, Sichuan 610041, People's Republic of China

[§]National Bioenergy Center, and [‖]Biosciences Center, National Renewable Energy Laboratory, Golden, Colorado 80401, United States

[⊥]Department of Chemical Engineering, Colorado School of Mines, Golden, Colorado 80401, United States

[▽]Department of Biological Science and Technology, and [○]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

Ⓢ *Supporting Information*

ABSTRACT: We examine how the catalytic domain of a glycoside hydrolase family 7 endoglucanase catalytic domain (Cel7B CD) facilitates complexation of cellulose chains from a crystal surface. With direct relevance to the science of biofuel production, this problem also represents a model system of biopolymer processing by proteins in Nature. Interactions of Cel7B CD with a cellulose microfibril along different paths of complexation are characterized by mapping the atomistic fluctuations recorded in free-energy simulations onto the parameters of a coarse-grain model. The resulting patterns of protein–biopolymer couplings also uncover the sequence signatures of the enzyme in peeling off glucan chains from the microfibril substrate. We show that the semiopen active site of Cel7B CD exhibits similar barriers and free energies of complexation over two distinct routes; namely, scooping of a chain into the active-site cleft and threading from the chain end into the channel. On the other hand, the complexation energetics strongly depends on the surface packing of the targeted chain and the resulting interaction sites with the enzyme. A revealed principle is that Cel7B CD facilitates cellulose deconstruction via adaptive coupling to the emergent substrate. The flexible, peripheral segments of the protein outside of the active-site cleft are able to accommodate the varying features of cellulose along the simulated paths of complexation. The general strategy of linking physics-based molecular interactions to protein sequence could also be helpful in elucidating how other protein machines process biopolymers.
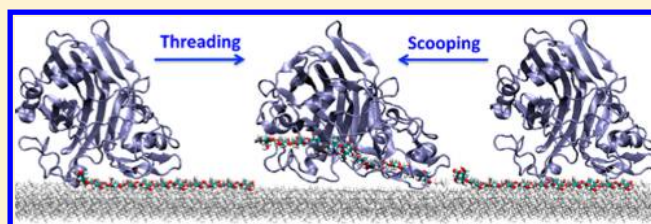
INTRODUCTION

In biological systems, synthesis,[1−3] repair,[4−6] decomposition,[7−9] and other intricate processing of biopolymers[10] require specialized proteins to accomplish. The structure−function relationships in handling macromolecules, though, are much less developed than those in active-site chemistry and small-molecule binding. An essential cause to this imbalance in knowledge is that the coupled conformational changes between an enzyme and its biomolecule substrate are difficult to resolve. In this regard, molecular simulation methods can potentially be used to supply the missing details. Such capability is illustrated here by applying a simulation framework we developed to an important example of enzyme-facilitated biomaterial processing: endoglucanase detaching cellodextrins at the solid−liquid interface of a cellulose microfibril and directing the glucan chains into the active-site cleft.[11−15] The cellulase activities in complexing and decomplexing glucan chains are also

determining factors of the rates of cellulose decomposition into soluble sugars.[11−15]

As an aggregate of linear glucose polymers, cellulose appears in plant cell walls mainly as microfibrils of 0.1−100 μm in length and 2−20 nm in diameter.[16] Within a flat sheet in the microfibril (Figure 1C), each glucan chain is hydrogen bonded to two neighbors via equatorial hydroxyl groups.[17−19] All of the axial CH moieties of glucose residues in the layer face out of the plane for different sheets of cellodextrin chains to form CH−O contacts and stack together. The crystalline network makes cellulose recalcitrant against deconstruction.[20] Detaching a glucan chain from cellulose in water requires a free energy of 2 kcal/mol/glucose-residue at room temperature.[21−23]
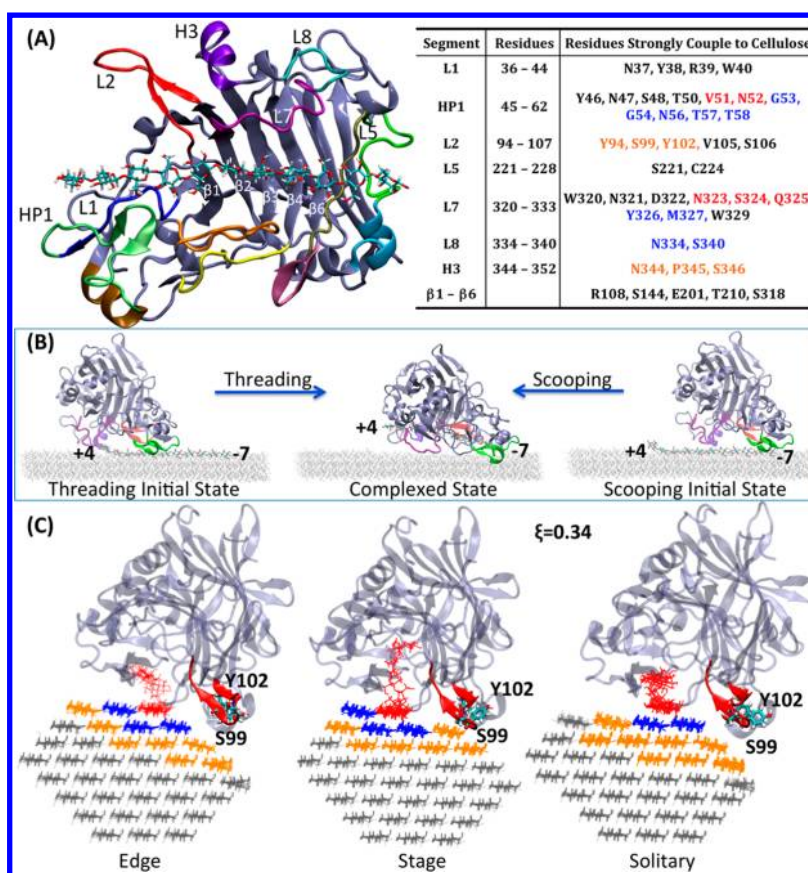
Figure 1. Simulation model of Cel7B CD complexation. (A) X-ray structure with indications of structure segments. The residues that have strongly coupled to the targeted glucan chain (TGC), the nearest-neighbor chains of the TGC (NNC), and the rest of the nearby chains on the microfibril (RNC) during the complexation pathways are listed in black, blue, and orange, respectively, in the right table. Deconstruction residues that exhibit strong couplings to both TGC and NNC are colored red. (B) Initial and final states of the two pathways resolved in this work: scooping and threading complexation. (C) Front view of the Cel7B CD and microfibril complex with the TGC partially in the active-site cleft for the edge, stage, and solitary configurations of the TGC considered in this work. Positions of the TGC (red), NNC (blue), and RNC (orange) chains are also illustrated.

Despite the significant recalcitrance of cellulose, many microorganisms secrete cellulase enzyme cocktails to decompose cellulose into soluble sugars as a food source.[11−15] In particular, the ascomycete fungus *Trichoderma reesei* (*Hypocrea jecorina*) has been well characterized in terms of its ability to decompose plant cell walls.[24] *T. reesei* secretes a mixture of processive and nonprocessive cellulases, and the most abundant cellobiohydrolase (Cel7A) and endoglucanase (Cel7B) in the cocktail are both from glycoside hydrolase family 7 (GH7).[25] The catalytic domain (CD) of Cel7B has peripheral loops partially enclosing the active site to form a semiopen cleft for endoglucanase activity.[26] Characterizing how the cellulase enzyme alleviates the aforementioned free-energy cost of detaching glucan chains from cellulose is key to resolving the overall energetics of enzymatic decomposition.[21]

To elucidate how Cel7B facilitates complexation, we jointly applied all-atom molecular dynamics (MD), path optimization, free-energy simulation, and coarse-grain (CG) analysis of the atomistic MD data obtained in these calculations. In particular, we examine whether the semiopen active-site cleft of Cel7B allows complexing a glucan chain via a scooping path wherein Cel7B begins binding from the middle of the chain, and a threading path wherein the chain is fed into the active-site cleft from the reducing end. Both routes can be resolved by altering the initial position of Cel7B CD on a microfibril relative to the

reducing end of a targeted glucan chain in path optimization. Common barriers and free-energy differences between end states for the two paths were observed in complexing different glucan chains on the cellulose surface. The potentials of mean force (PMFs) of complexation, though, are sensitive to the particular surface packing of the targeted chain. The calculated strengths of enzyme−cellulose couplings during the simulated paths elucidate how the sequence and structure of Cel7B CD modulate the free-energy landscape of complexation.

■ SIMULATION MODELS AND ANALYSIS

All-Atom Models of Cel7B CD Complexation on a Cellulose Microfibril. An X-ray structure of Cel7B CD (PDB ID 1EG1)[26] was employed for model construction of the adsorbed and complexed states of Cel7B CD as reported previously.[27] The simulation model of the microfibril[19] shown in Figure 1 contains 36 glucan chains in the I$_\alpha$ form.[18] This model was employed previously in studying the solvation structures of cellulose[19] and calculating the free-energy profile of deconstructing the surface glucan chains off the microfibril.[23] A snapshot of the microfibril structure in the production MD run after equilibration in the previous study was randomly picked for model construction of the Cel7B−microfibril complex. In the adsorbed state, the enzyme sits on an edge of the model microfibril initially with 100 ns MD simulation in

explicit water for equilibrating the protein orientation with respect to the microfibril, and their contact structures[19,27] were selected for model construction of the Cel7B−microfibril complex. The targeted glucan chain has the reducing end free of glycosidic bonding for detachment by the enzyme. In the complexed state, Cel7B CD has 11 residues of the targeted glucan chain from the reduced end in the active-site cleft on the microfibril according to the X-ray structure of Cel7A CD bound with an oligosaccharide.[27,28] An all-atom MD simulation of ~100 ns was performed to equilibrate the Cel7B CD−microfibril orientation and the contact structures of the cellulase and the microfibril in the complexed state. At the end of the simulations of the adsorbed and complexation states, a snapshot is taken to build the initial and final replicas for the reaction path optimization of Cel7B CD complexation. For studying the complexation of different glucan chains on the cellulose surface, the same protocol of using ~100 ns MD simulations to equilibrate the adsorbed and complexed states was employed for the model construction of reaction path optimization and free-energy simulations.

The entire microfibril was included in the previous simulations of the adsorbed and the complexed states,[27] but only the glucan chains within two layers of the protein are included in the complex model for the reaction path optimization and free-energy calculations of Cel7B CD complexation performed in this work, as shown in Figure 1. To retain the crystalline structures in the reduced model of the microfibril, heavy atoms of the glucan chains in the bottom layer and at the two ends of each glucan chain were restrained to their positions via harmonic potentials with a force constant of 100 kcal/mol/Å^2 in the restrained MD simulations for calculating the free-energy profile of complexation. Compared to the earlier model, the microfibril was also expanded longitudinally to 20 glucose residues to enable the simulation of the threading pathway conducted in this work.

In computing the free-energy profiles of complexation, the Cel7B CD−microfibril complex is solvated in explicit solvent of ~28 000 water molecules with 72 Na^+ and 70 Cl^- ions to mimic the ionic strength of the physiological condition and charge neutralization of the system. The entire system contains ~95 000 atoms. The CHARMM22 all-atom force field[29] with the CMAP correction and the TIP3P water model[30,31] were used in all of the MD simulations performed in this work. The particle mesh Ewald method[32] was employed for calculating long-range electrostatics. The short-range nonbonded interactions are calculated with a cutoff of 12 Å with a switch function turned on at 10 Å. Langevin dynamics with a damping coefficient of 0.5 ps^{-1} was used to maintain the system at the optimal temperature of Cel7B at 313 K, and the Langevin piston method was used to maintain pressure at 1 atm with the same damping coefficient. All-atom MD simulations were performed using the NAMD software[33] with a time step of 2 fs and all covalent bonds associated hydrogen atoms constrained at their equilibrium values.[29] System setup and other analyses were performed using the CHARMM software[29] and in-house codes.

Reaction Path and Free-Energy Simulations. Following a study reported previously,[23] minimum-energy path optimization using a chain-of-replicas method[34,35] with the "Fast Analytical Continuum Treatment of Solvation" implicit solvent model[36] was conducted first to obtain the pathways of complexation. All of the reaction path optimization calculations were performed with the CHARMM software.[29] Solvent and

ion atoms were not included in the calculations of path optimization. The energy-minimized structures of the adsorbed and complexed states are the initial and final replica, respectively. On the basis of the analysis of coupling strengths in the MD simulations of the Cel7B CD−microfibril complex,[27] the Cartesian coordinates of the three sets of atoms listed in Table 1 were employed to calculate the distances between

Table 1. Sets of Atoms Used To Define Rms Distances between Replicas in Path Optimization

A1	heavy atoms of the targeted glucan chain.
A2	heavy atoms of the glucan chains that make intersheet nearest-neighbor contacts with the targeted chain.
A3	heavy atoms of the residues of Cel7B CD with significant coupling strengths to the complexed glucan chain in an MD simulation.[27] For each glucose residue, the amino acid with the highest coupling strength was selected (Tyr38, Arg39, Trp40, Ser48, Val51, Val105, Ser144, Thr210, Ser221, Trp320).

replicas via the root-mean-square fit distance (rms)[35,36] in minimum-energy path optimization. Except enforcing equal distances between replicas, all other degrees of freedom are relaxed to minimize the sum of potential energies of all replicas.

With the adsorbed and complexed states of Cel7B CD on the microfibril as the end replicas, we used the approach of increasing the resolution of the path stagewise[35,36] for resolving the pathways of complexation. Starting from a linear interpolation between the initial and final replicas, stage-1 used a three-replica path with a single intermediate. The stage-1 path optimization is thus finding a minimum-energy structure of the enzyme−microfibril complex on the equal-rms distance hyperplane in between the adsorbed and the complexed state. Without the couplings to other higher-energy intermediates through the equal-rms distance constraints as in the cases when using many replicas right to start with, molecular structures can readily relax on the sole hyperplane in the three-replica path optimization. In stage-2 path optimization, we doubled the path resolution by inserting a new replica in between two neighboring replicas in the optimized result of stage-1.[35,36] The rounds of resolution doubling and path optimization stopped when doubling the resolution results in insignificant changes (±1 kcal/mol) in the profile of the accumulated work along the optimized path.[35,36] We found that the smooth, energy-minimized 33-replica paths in stage-5 with the rms distance between neighboring replicas being ~0.70 Å were sufficient. This result is also consistent with the case of resolving the pathways of detaching an 11-residue glucan chain from cellulose without an enzyme.[23] The molecular structures along the optimized paths of 33 replicas were then used as the starting configurations for free-energy simulations. During the path optimization at different stages, we also added kinetic energy potentials[35,36] with the strength ranging from 0.001 to 0.01 kcal/mol/Å^2 to ensure stable optimization without forming a zigzag path.[35,36] For each path of complexation resolved in this work, we performed several rounds of minimum-energy path optimization and consistently observed the same patterns and structures of protein−substrate couplings.

Although the stagewise generation of reaction path was employed to avoid bias on the resulting mechanism that would be induced by assuming an arbitrary initial pathway, each path optimization only gives a single curve in the high-dimensionality energy landscape of the Cel7B CD−microfibril complex. To examine the reproducibility of the observed mechanism of

how the enzyme stabilizes cellulose deconstruction, we also resolve different complexation pathways by altering the adsorbed state while keeping the complexed state unchanged. In particular, by shifting the position of Cel7B CD with respect to the reducing end of the targeted glucan chain, we obtained the scooping and threading paths of complexation as shown in Figure 1B. Furthermore, we also analyzed how the two routes process different surface chains on the microfibril substrate. The results illustrate that the same mechanism was observed in this class of complexation pathway.

On the basis of the results of minimum-energy path optimization, we defined a general order parameter to calculate the free-energy profile of complexation. Since each optimized path represents a distinct way of breaking the contacts between A1−A2 and forming those of A1−A3 atoms in Table 1, the collective variable for computing the complexation PMF at 313 K and 1.0 atm in explicit water is defined as the difference in the contact number[37] between A1 and A3 and that between A1 and A2 normalized to vary from 0 to 1:

$$\xi = C1(\text{A1−A3contact number})$$
$$- C2(\text{A1 − A2contact number})$$

The correspondence between ξ and the optimized path was also verified by the nearly linear growing value of ξ with the index of replica along the paths. Therefore, each replica is assigned to a value of the order parameter, and the path is parametrized by 33 equally spaced ξ values. Furthermore, this order parameter can be applied to both the scooping and threading routes of complexation for their PMFs to be compared directly under the same footing.

The PMF along ξ for each path resolved in this work was calculated by integrating the mean forces calculated at 33 equally spaced values with 10 ns of production-stage MD data at each hyperplane. Before the production-stage MD, the molecular structure of each replica along the reaction path was solvated, minimized, heated, and equilibrated with a force constant of 100 kcal/mol/Å^2 restraining ξ to the targeted value. This strength of restraint potential is sufficient to retain ξ very close to the target value and gives consistent mean force values via $-k(\langle\xi\rangle_i - \xi_{i0})$ with local perturbations of the force constant. For the simulation on the ith hyperplane of ξ, $\langle\xi\rangle_i$ is the averaged value and ξ_{i0} is the targeted value. The error bars of mean forces were calculated via block averages of the 10 ns production-stage MD data. Furthermore, systematic off-path drifting was not observed in the MD simulations restrained on both scooping and threading routes, indicating that the sampling was around local minima in the path space that have high probabilities to be visited in dynamic trajectories.

Analysis of Enzyme−Cellulose Couplings along Complexation Paths. The molecular complex of Cel7B CD and the microfibril is modeled as an elastic network model with the force constants and bond lengths computed from an all-atom MD trajectory via fluctuation matching.[38] Each protein residue has a CG site at the C_α position and a site at the center of mass of side chain atoms. Each glucose residue has four CG sites to represent the interaction network in cellulose. The coupling strength between a protein residue and a cellulose segment is the sum of the force constants of their connecting bonds. Coupling strengths were calculated for every MD segment of 4 ns following the protocol developed in a previous work.[27] The average and standard deviation of coupling strength in a given trajectory were calculated by block averages.

The calculated force constants are the effective mechanical couplings that best reproduce the correlated structure fluctuations observed in the MD data. The coupling strengths between protein residues have been shown to capture the highly conserved and coevolved residues of several protein systems[38] including Cel7B CD.[27] With the aforementioned protocol, a coupling strength between a protein residue and a glucan chain exceeding 2.5 kcal/mol/Å^2 usually signals molecular contacts such as ring−ring stacking and hydrogen bonding. In such cases, the protein residue is considered to have a strong coupling strength to the glucan chain. Such specific molecular interactions observed in the optimized reaction paths mostly result in strong couplings in the restrained MD simulation, indicating that local minima in the path space were indeed identified. In addition to the targeted glucan chain (TGC), Cel7B CD also interacts with the nearest-neighbor chains (NNC) next to the TGC and the remaining nearby chains on the microfibril (RNC). Glucan chains in the TGC, NNC, and RNC parts of cellulose are defined in Figure 1C. The Cel7B CD residues that have strongly coupled to the TGC (black), NNC (blue), or RNC (orange) during complexation pathways are colored in Figure 1A. Residues that exhibit strong couplings to both the TGC and NNC during complexation are defined as deconstruction residues, which are listed in red in Figure 1A.

■ RESULTS AND DISCUSSION

The Paths and PMF Profiles of Complexation. Definitions of Cel7B CD segments are shown in Figure 1A. In adsorbing at the edge chain, Cel7B CD contacts the microfibril surface with HP1 and L7.[27] The two enzyme segments are separated by four glucose residues along the TGC and located at the opposite sides of the active-site cleft. Glucose residues in the TGC are numbered from −7 to +4 as indicated in Figure 1B, with +4 being the reducing end. The active site of Cel7B CD targets the glycosidic bond between the −1 and +1 units.

Along both the scooping and threading routes at different minima in the path space as discussed in the methods section, a conformational change occurs in the Asn52 and Ser324 regions of HP1 and L7, respectively, in which these deconstruction residues shift from the top of the cellulose surface to the next sheet to separate the TGC from NNC, Figure 2. According to the relative positions of Asn52 and Ser324 to the TGC, the complexation process can thus be divided into three stages: L7-wedging, HP1-wedging, and progression. The movies of both minimum-energy paths viewed from two orthogonal angles are recorded in Movies S1−S4 in the Supporting Information. The calculated PMF profiles of Cel7B CD complexing the edge chain are shown in Figure 2 with representative snapshots to exemplify the three stages.

Along the scooping route, complexation starts from HP1 and L7 flanking the TGC from above the −5 and −1 glucose, respectively, Figure 1B. In the first stage of L7-wedging, the reducing-end segment (+1 to +4) of the TGC shifts away from its crystalline packing for Ser324 in L7 to insert in between the TGC and NNC at the −1 site. In the next HP1-wedging stage, the structure change of the TGC is coupled with that of Cel7B CD until Asn52 in HP1 wedges in the microfibril at the −5 site of the TGC.

In the adsorbed state of the threading route, Asn52 contacts the +3 unit of the TGC and Ser324 is 2 glucose units away from the reducing end, Figure 1B, to model an exoinitiation
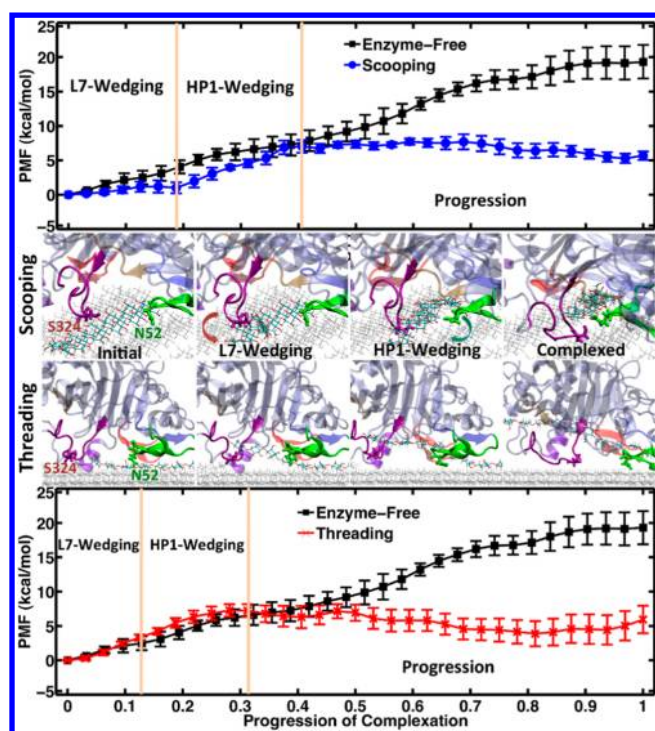
Figure 2. PMF profiles of complexing the edge chain in Figure 1C along scooping and threading paths with representative snapshots to exemplify the three stages of complexation.

process. As the Cel7B CD threads toward the nonreducing end in the L7-wedging stage, Ser324 wedges in the microfibril at replica 4 with its side chain hydrogen bonding to the +4 glucose of the TGC. The next HP1-wedging stage spans from replica 5 to 11 and finishes at Ser324 reaching the +3 and Asn52 wedging in at the −1 site, Figure 2. After the wedging stages, L7 and HP1 move closer to each other to continue to lift the TGC up in the progression stage, in which both paths have a mostly flat PMF as shown in Figure 2.

Along both paths, most of the free-energy increase occurred in the HP1-wedging stage, and the highest PMF (∼7.7 kcal/mol) and the free energy of complexation (∼5.7 kcal/mol) are similar. The agreement in PMF barrier illustrates that both routes can rely on biologically relevant energetics to accomplish and the two minima in the path space likely have similar probabilities. Agreement in the free energy of complexation, or the PMF difference between end states, indicates that thermodynamic consistency was achieved in our path and free-energy simulations.

The PMF of detaching a glucan chain of the same length from cellulose in water without an enzyme is also shown in Figure 2 for comparison. It is clear that Cel7B CD reduces the free energy of the detached state by ∼13.6 kcal/mol compared to the enzyme-free case. The enzyme-modulated PMF becomes significantly different in the progression stage as the disrupted interactions of the TGC with NNC start to be compensated by the couplings with Cel7B CD. Next, the enzyme–cellulose couplings along the paths are analyzed to elucidate the how the Cel7B CD facilitates complexation.

Evolution of Enzyme–Cellulose Couplings during Complexation Paths. The coupling of an enzyme residue to cellulose is decomposed into the TGC, NNC, and RNC parts defined in Figure 1C. Although Figure 2 provides a pictorial view of the molecular interactions occurring during the

scooping and threading routes of complexation, based the quantitative analysis of coupling strengths, Figure S1 of the Supporting Information records the strongly coupled Cel7B CD segments to the 11 glucose residues of the TGC as a function of ξ. For the glucose sites of NNC and RNC in parallel, the strongly coupled enzyme segments are shown in Figure S2 and Figure S3 of the Supporting Information. The patterns of the scooping and threading processes clearly reflect different sequences of coupling to the glucose residues of the TGC that give rise to similar complexation barriers and free energies. It is also apparent that HP1 and L7 play important roles in coupling to cellulose along both paths.

The two β strands of hairpin HP1 form a three-face, twisted β sheet with the L1 strand defined as Asn37-Trp40.[26] This β sheet constitutes the part of the active-site cleft that binds the −7 to −4 portion of the TGC. The N-terminal strand of HP1 (Tyr46-Thr50) is next to L1 and forms one of the faces in the twisted β sheet. Along the scooping path, the coupling strength of a N-terminal residue, Ser48, to the TGC is shown in Figure 3 to illustrate the typical trend of a gradual increase after the wedging stages.
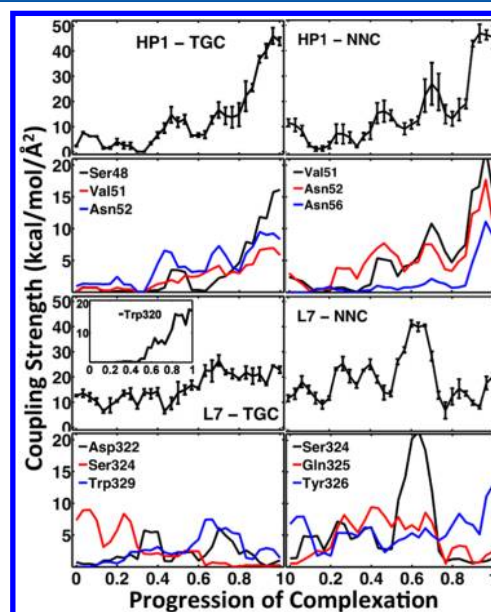

Figure 3. Coupling strengths of the HP1 and L7 segments of Cel7B CD to the TGC and NNC parts of cellulose during the scooping pathway of complexation. The profiles of individual protein residues with strong coupling strengths are also shown to illustrate the underlying details.

The hairpin turn of HP1 at Val51 and Asn52 extends toward the active-site cleft for wedging in the microfibril as discussed earlier. The coupling strengths of Val51 and Asn52 to TGC along the scooping path are presented in Figure 3. The enhanced strengths around ξ = 0.42 after HP1-wedging indicate that these residues block the TGC from back contacting with the NNC on the surface. Peaks in Figure 3 for the coupling strengths of Val51 and Asn52 to NNC after HP1-wedging are also clear. The C-terminal strand (Gly53-Tyr58) of HP1 forms the third face of the twisted β sheet. The residues there cannot reach the TGC but have strong couplings to the NNC chains after HP1-wedging. The profile of the Asn56−NNC coupling strength shown in Figure 3 is an evident example.

The sequence and structure of HP1 thus contribute three roles to facilitate complexation. First, the N-terminal strand forms part of the active-site cleft to bind the TGC. Second, the hairpin turn contains deconstruction residues to wedge in the microfibril. Third, the C-terminal strand couples to NNC to block the detached TGC from adhering back to cellulose. These behaviors can also be observed in the threading path for which the profiles of HP1−cellulose couplings are shown in Figure S4 of the Supporting Information.

L7 locates at the opposite side of the active-site cleft to HP1 with the N-terminal (Trp320-Glu322) and C-terminal (Trp329-Gly333) forming parts of the active-site cleft to cover the −2 to +2 portion in binding the TGC. The C-terminal of L7 attaches to L8 and is more flexible than the N-terminal region. The strengths of Trp320-TGC and Trp329-TGC couplings along the scooping pathway are shown in Figure 3 to illustrate the distinct profiles of the two terminals.

The centerpiece of L7 (Asn323-Gln325) extends out of the active-site cleft to wedge in the microfibril as discussed earlier. The profile of Ser324−TGC coupling along the scooping path shown in Figure 3 has high strengths in the L7-wedging stage followed by an overall trend of gradual decrease. The profile of Ser324−NNC coupling has higher strengths at intermediate ξ values with the maximum occurring at $\xi = \sim0.6$ that is also the instance at which the reducing end of the TGC back contacted the cellulose surface as mentioned earlier. The Ser324−NNC coupling thus prevents the TGC from adhering back further. The Gln325−NNC coupling also has higher strengths during the pathways than those in the end states.

Another special pattern of coupling to cellulose is for Tyr326 of L7 that is too far from the active-site cleft to reach the TGC during complexation. In the adsorbed state, the hydroxyl group of Tyr326 interacts with the equatorial moieties of the chain next to TGC in the same sheet as a result of the specific local conformation. Path and PMF simulations indicate that after L7-wedging, Tyr326 is brought to the NNC chains in the next sheet with direct ring−ring contacts. The strength of Tyr326−NNC coupling in Figure 3 thus becomes stronger toward the end of complexation. Asn334 and Ser340 in L8 are also brought to the defected site to strongly couple to NNC after L7-wedging and are listed in blue as Tyr326 in Figure 1A.

It is clear that L7 facilitates complexation also via the three roles stated earlier for HP1-forming part of the active-site cleft, wedging in the microfibril, and coupling to NNC to block TGC from adhering back to cellulose surface. The more extended structure of L7 than that of HP1 and L7 being closer to the β sandwich core than HP1 does reflect in the strengths of many L7−TGC and L7−NNC couplings to have higher values in the intermediate states than those in the end states (see Figure 3). The functional roles of such residues would thus be difficult to rationalize without resolving the pathways and PMF profiles. The threading path demonstrates similar behaviors in L7−cellulose couplings, Figure S4 of the Supporting Information.

Dependence of Complexation PMF on the Surface Configurations of Glucan Chains. The results presented thus far establish that Cel7B CD adapts to the exposed structural features on cellulose for modulating the complexation PMF. It is thus expected that the free-energy landscape of complexation would be sensitive to the surface packing of glucan chains. Figure 4 plots the PMF of Cel7B CD complexing the TGC of the edge, stage, or solitary scenario shown in Figure 1C and illustrates the strong dependence of PMF on the surface arrangement of TGC. For all three cases, the total
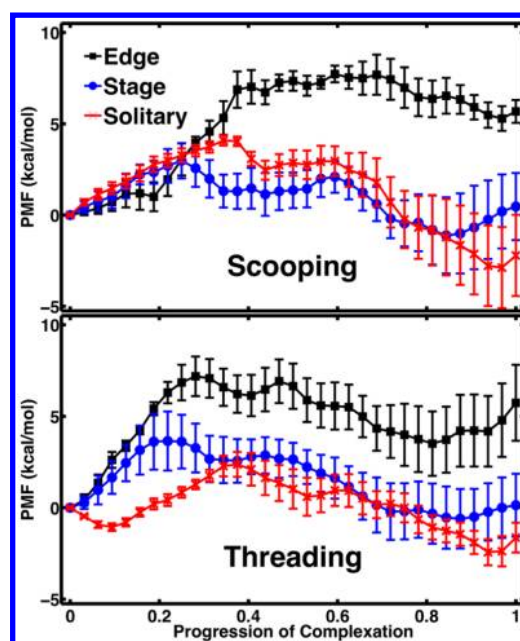

Figure 4. PMF profiles of Cel7B CD complexing the edge, stage, and solitary chains shown in Figure 1C along the scooping and threading pathways.

strength of Cel7B CD coupling to TGC plus that to NNC has almost identical curves with ξ as Supporting Information Figure S5 shows. The distinct PMF profiles observed in Figure 4 are thus not due to drastically different interaction strengths of Cel7B CD with the cellulose parts that are directly involved in deconstruction but as a result of the specific patterns of enzyme−substrate couplings. Furthermore, the aforementioned patterns of enzyme−substrate coupling along both routes also persist in the complexation of different surface chains, indicating reproducibility of the reported mechanistic details.

As discussed earlier, in Cel7B CD adsorbing around the edge chain, L7 can only contact the sheet above that of the NNC. These preorganized enzyme−cellulose interactions have to be disrupted for L7 to shift from the upper layer to the next to wedge in (Figure 2). For the stage TGC resulting from removing the edge chain in Figure 1C, the originally covered glucan chains in the next sheet become accessible on one side. Therefore, in complexing the stage chain, the deconstruction residues of L7 can latch on at the sheet of NNC to begin the process. The calculated PMF profiles in Figure 4 indeed show that complexing the stage TGC has a barrier of ∼3.0 kcal/mol and free energy of complexation of about 0.5 kcal/mol; both are lower than those of targeting the edge chain. The solitary TGC in Figure 1C has no equatorial neighbors, and both HP1 and L7 can flank the TGC at the layer of the NNC. Detaching the solitary TGC along the two paths has a negative free energy of complexation of −2.2 kcal/mol. These results indicate that the apparent free energy of enzymatic cellulose decomposition[21] would depend on the spatial profile of the local conformation of individual glucan chains.

Although the solitary TGC has a favorable free energy of complexation, lower than that of targeting the stage chain, the barrier is slightly higher by ∼1.1 kcal/mol. Along the scooping path, the split between the stage and solitary PMFs shown in Figure 4 starts at the onset of the HP1-wedging stage where $\xi = \sim0.3$. The drop in PMF when complexing the stage chain at this position corresponds to the drastic increase in the strength

of enzyme−RNC coupling shown in Supporting Information Figure S5. Along the complexation paths of the other two TGC configurations, edge and solitary, however, the Cel7B CD does not exhibit strong couplings to RNC. This result indicates that even though the net strength of enzyme coupling to TGC and NNC is more or less insensitive to chain configurations (Supporting Information Figure S5), specific interactions with RNC can play a role in modulating the PMF. Dissecting the details indicates that the main contributor to the additional coupling to RNC in complexing the stage chain is L2 on the same side of L7. The strengths of L2−RNC couplings along the scooping path are shown in Figure 5. The complexation of
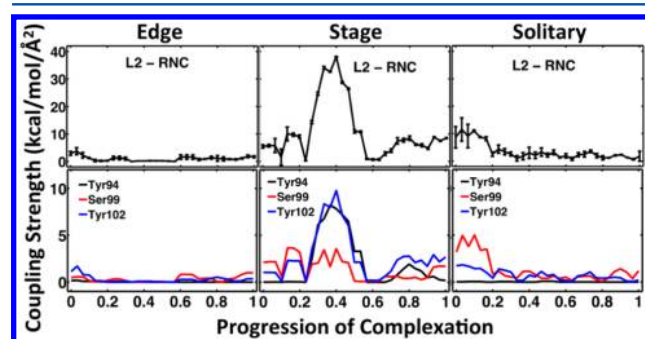

Figure 5. Coupling strength of L2 segment of Cel7B CD to the RNC part of cellulose during the scooping pathway of complexation. The profiles of individual protein residues with strong coupling strengths are also shown to illustrate the underlying details.

stage and solitary chains along the threading path also exhibits clear correspondences between the PMF profiles (Figure 4) and the strengths of enzyme−RNC couplings (Supporting Information Figure S5).

Sequence Signatures of Cel7B CD for Complexation. In binding glucan chains, amino acids with a planar side chain can exhibit shape-complemented hydrophobic interactions with glucose rings.[39,40] Nonplanar, nonpolar side chains can couple to glucose but tend to have weaker strengths. Polar atoms in peptide bonds and side chains also contribute to binding cellulose by hydrogen bonding to the OH groups. The ∼50 Å long active-site cleft of the concave β-sandwich fold of Cel7B CD contains highly conserved Tyr38, Trp40, Trp329, and Trp320 with the other types of amino acids mentioned above filled in between to hold a cellodextrin chain with 11 residues.[26]

Regarding other protein functionalities facilitating complexation, likely candidates include amino acids that can specifically couple to the hydroxyl groups, aliphatic sugar rings, and/or ether linkages in glucan chains because they may "wedge in" the microfibril and also block the TGC from adhering back to NNC or RNC. The key deconstruction residues of Cel7B CD have small, amphiphilic (Ser324 in L7) and polar-planar (Asn52 in HP1) side chains. An advantage of the small size of Ser and Asn lies in minimizing perturbation in the packing of glucan chains on the cellulose surface for the protein residues to wedge in. These properties (small, amphiphilic, polar-planar) also appear to be favored in the sequences of the peripheral segments of Cel7B CD for coupling to TGC, NNC, and RNC outside of the active-site cleft. Ser/Thr and Asn/Gln appear frequently as the red (to both TGC and NNC), blue (to NNC), and orange (to RNC) residues listed in Figure 1A. Gly, Met, and Pro can also be found occasionally in these positions. In the peripheral loops of Cel7B CD that fulfill the

aforementioned roles, Trp and Tyr that have shape complementarity to the hydrophobic surfaces of cellulose are relatively infrequent.

Although the alignment of 340 nonredundant, globally alignable sequences with high similarity to that of Cel7B CD[27] generally shows low conservation at the level of matching the specific amino acid type in the residue positions of wedging and blocking functionalities, it reveals the emergent behavior described above. Supporting Information Figure S6 records the occurrence frequencies of different amino acids at these positions and indicates that Tyr and the other smaller-sized amino acids mentioned earlier that are capable of interacting with glucan chains have high probabilities to take over the sequence of Cel7B CD. For example, Ser mostly likely replaces Asn52 in HP1. Ser324 in L7 has a high chance to be switched to Tyr, Ala, or Thr. His and Tyr mostly substitute for Asn323. Tyr326 that binds NNC has the highest chance to be replaced by Asn. The complexation-facilitating positions of Cel7B CD thus appear to afford alternative solutions in sequence to accommodate the diverse functional groups exposed by glucan chains. The observed pattern is that amino acids capable of coupling to glucose residues are generally observed in these positions and those of smaller sizes are preferred. Even in the class of shape-complemented surface binders, the smaller Tyr has much higher probabilities than those of Trp, Supporting Information Figure S6.

A potential reason for Ser/Thr, Asp/Gln, and the other amino acids mentioned above to be opted over Trp and Tyr to facilitate complexation is that throughout the paths, the enzyme has to adapt to the evolving biopolymer substrate at a solid−liquid interface to modulate the PMF profile. Complementing the structure set encountered in multiple paths likely requires a smoother free-energy landscape via Cel7B CD not having overly specific couplings to cellulose surfaces. Another potential reason is ensuring adequate rates of decomplexation for the endoglucanase to reset after each turnover. Aligning the sequence of Cel7A CD of *T. reesei*, the processive cellobiohydrolase analogue,[28] with that of Cel7B CD in Supporting Information Figure S7 shows that for the segments corresponding to the HP1 and L7 in Cel7B, Cel7A has more occurrences of Tyr and Trp, likely due to the higher surface specificity necessary to achieve processivity.

■ **CONCLUSION**

Processing biopolymers is an essential functionality of enzymes in Nature, for which cellulases must employ specific structure−property relationships to handle the mechanical rigidity of cellulose. This work identified complexation of the most abundant *T. reesei* endoglucanase on cellulose as a model system in this class of problems. The computational framework adopted here involves all-atom path optimization and free-energy simulations coupled with a trans-scale mapping scheme to convert the results of atomistic simulations into the parameters of a coarse-grain model. The objective was to capture the emergent properties of enzyme−biopolymer couplings along the processing paths.

A key principle established by the molecular simulations is that Cel7B CD adapts to couple to the evolving surface structures of cellulose during complexation. The PMF profiles calculated in separate simulations are thus a strong function of the local arrangements of glucan chains on the cellulose surface. The peripheral segments HP1 and L7 of Cel7B CD are shown to play key roles in complementing the diverse set of

biopolymer structures emergent during complexation. These peptide segments are composed of ~10 amino acid residues to accomplish three consecutive tasks: wedging in the microfibril, binding the TGC via the forming part of the active-site cleft, and affixing to NNC to block the TGC from adhering back to cellulose. Coordination of these functionalities significantly reduces the PMF profiles of perturbing the surface packing of glucan chains compared to the enzyme-free case. The versatility of the peripheral segments of Cel7B was also shown to allow the flexibility for adapting multiple paths (the scooping and threading routes) with similar free-energy costs to accomplish complexation.

The presented studies of Cel7B complexation show that the enzyme facilitates the cellulose processing via adaptive couplings to the emergent substrate structures. The principles deduced from molecular simulations also establish a guideline for dissecting the functional signature of complexing glucan chains in a multiple-sequence alignment. To accommodate the tight packing and recalcitrance of the macromolecular substrate, Cel7B exhibits peripheral segments with extended lengths to fulfill distinct roles in different stages of complexation. An important implication for the engineering of improved cellulases, based on the mechanism of complexation resolved in this work, is that the optimal protein sequence depends on the structural features of the substrate. Therefore, an integrated strategy is needed to jointly consider the impact of pretreatment on cellulose structure and the resulting kinetics of cellulase action. The approach of linking physics-based molecular properties to bioinformatics is also expected to be useful in studying other biopolymer-processing proteins.

ASSOCIATED CONTENT

ⓢ Supporting Information

Supporting Figures S1−S7; four movies S1−S4 to show the two different complexation mechanisms. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author
*Tel. 886-3-5712121, ext. 56996; e-mail jwchu@nctu.edu.tw.

Author Contributions
The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

REFERENCES

(1) Yonath, A. Polar Bears, Antibiotics, and the Evolving Ribosome (Nobel Lecture). *Angew. Chem., Int. Ed. Engl.* **2010**, *49*, 4341−4354.

(2) Tanovic, A.; Samel, S. A.; Essen, L.-O.; Marahiel, M. A. Crystal Structure of the Termination Module of a Nonribosomal Peptide Synthetase. *Science* **2008**, *321*, 659−663.

(3) Morgan, J. L. W.; Strumillo, J.; Zimmer, J. Crystallographic Snapshot of Cellulose Synthesis and Membrane Translocation. *Nature* **2012**, *493*, 181−186.

(4) Ali, M. M. U.; Roe, S. M.; Vaughan, C. K.; Meyer, P.; Panaretou, B.; Piper, P. W.; Prodromou, C.; Pearl, L. H. Crystal Structure of an Hsp90-Nucleotide-p23/Sba1 Closed Chaperone Complex. *Nature* **2006**, *440*, 1013−1017.

(5) Silva, P.; Thompson, E.; Bailey, S.; Kruse, O.; Mullineaux, C. W.; Robinson, C.; Mann, N. H.; Nixon, P. J. FtsH Is Involved in the Early Stages of Repair of Photosystem II in Synechocystis sp PCC 6803. *Plant Cell* **2003**, *15*, 2152−2164.

(6) Redinbo, M. R.; Stewart, L.; Kuhn, P.; Champoux, J. J.; Hol, W. G. Crystal Structures of Human Topoisomerase I in Covalent and Noncovalent Complexes with DNA. *Science* **1998**, *279*, 1504−1513.

(7) Erez, E.; Fass, D.; Bibi, E. How Intramembrane Proteases Bury Hydrolytic Reactions in the Membrane. *Nature* **2009**, *459*, 371−378.

(8) Huber, E. M.; Basler, M.; Schwab, R.; Heinemeyer, W.; Kirk, C. J.; Groettrup, M.; Groll, M. Immuno- and Constitutive Proteasome Crystal Structures Reveal Differences in Substrate and Inhibitor Specificity. *Cell* **2012**, *148*, 727−738.

(9) Hammond, S. M.; Bernstein, E.; Beach, D.; Hannon, G. J. An RNA-Directed Nuclease Mediates Post-Transcriptional Gene Silencing in Drosophila Cells. *Nature* **2000**, *404*, 293−296.

(10) Lodish, H.; Scott, M. P.; Matsudaira, P.; Darnell, J.; Zipursky, L.; Kaiser, C. A.; Berk, A.; Krieger, M. *Molecular Cell Biology*, 5th ed.; W. H. Freeman: New York, 2003.

(11) Zhang, Y.-H. P.; Lynd, L. R. A Functionally Based Model for Hydrolysis of Cellulose by Fungal Cellulase. *Biotechnol. Bioeng.* **2006**, *94*, 888−898.

(12) Bansal, P.; Hall, M.; Realff, M. J.; Lee, J. H.; Bommarius, A. S. Modeling Cellulase Kinetics on Lignocellulosic Substrates. *Biotechnol. Adv.* **2009**, *27*, 833−848.

(13) Jalak, J.; Väljamäe, P. Mechanism of Initial Rapid Rate Retardation in Cellobiohydrolase Catalyzed Cellulose Hydrolysis. *Biotechnol. Bioeng.* **2010**, *106*, 871−883.

(14) Fox, J. M.; Levine, S. E.; Clark, D. S.; Blanch, H. W. Initial- and Processive-Cut Products from Cellobiohydrolase-Catalyzed Hydrolysis of Cellulose Reveal Rate-Limiting Steps and Role of Companion Enzymes. *Biochemistry* **2012**, *51*, 442−452.

(15) Cruys-Bagger, N.; Elmerdahl, J.; Praestgaard, E.; Tatsumi, H.; Spodsberg, N.; Borch, K.; Westh, P. Pre-Steady-State Kinetics for Hydrolysis of Insoluble Cellulose by Cellobiohydrolase Cel7A. *J. Biol. Chem.* **2012**, *287*, 18451−18458.

(16) Perez, S.; Mazeau, K. Conformations, Structures, and Morphologies of Celluloses. In *Polysaccharides: Structural Diversity and Functional Versatility*, 2nd ed.; Dumitriu, S., Ed.; Marcel Dekker: New York, 2004; pp 41−68.

(17) Nishiyama, Y.; Langan, P.; Chanzy, H. Crystal Structure and Hydrogen-Bonding System in Cellulose 1 Beta from Synchrotron X-ray and Neutron Fiber Diffraction. *J. Am. Chem. Soc.* **2002**, *124*, 9074−9082.

(18) Nishiyama, Y.; Sugiyama, J.; Chanzy, H.; Langan, P. Crystal Structure and Hydrogen Bonding System in Cellulose 1(Alpha), from Synchrotron X-ray and Neutron Fiber Diffraction. *J. Am. Chem. Soc.* **2003**, *125*, 14300−14306.

(19) Gross, A. S.; Chu, J.-W. On the Molecular Origins of Biomass Recalcitrance: The Interaction Network and Solvation Structures of Cellulose Microfibrils. *J. Phys. Chem. B* **2010**, *114*, 13333−13341.

(20) Himmel, M. E.; Ding, S.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. Biomass Recalcitrance: Engineering Plants and Enzymes for Biofuels Production. *Science* **2007**, *315*, 804−807.

10757

dx.doi.org/10.1021/jp405897q | *J. Phys. Chem. B* 2013, 117, 10750−10758

(21) Beckham, G. T.; Matthews, J. F.; Peters, B.; Bomble, Y. J.; Himmel, M. E.; Crowley, M. F. Molecular-Level Origins of Biomass Recalcitrance: Decrystallization Free Energies for Four Common Cellulose Polymorphs. *J. Phys. Chem. B* **2011**, *115*, 4118–4127.

(22) Gross, S. A.; Bell, A. T.; Chu, J.-W. The Thermodynamics of Cellulose Solvation in Water and the Ionic Liquid 1-Butyl-3-Methylimidazolim Chloride. *J. Phys. Chem. B* **2011**, *115*, 13433–13440.

(23) Cho, H. M.; Gross, A. S.; Chu, J.-W. Dissecting Force Interactions in Cellulose Deconstruction Reveals the Required Solvent Versatility for Overcoming Biomass Recalcitrance. *J. Am. Chem. Soc.* **2011**, *133*, 14033–14041.

(24) Martinez, D.; Berka, R. M.; Henrissat, B.; Saloheimo, M.; Arvas, M.; Baker, S. E.; Chapman, J.; Chertkov, O.; Coutinho, P. M.; Cullen, D.; et al. Genome Sequencing and Analysis of the Biomass-Degrading Fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat. Biotechnol.* **2008**, *26*, 553–560.

(25) Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B. The Carbohydrate-Active Enzymes Database (CAZy): An Expert Resource for Glycogenomics. *Nucleic Acids Res.* **2009**, *37*, D233–D238.

(26) Kleywegt, G.; Zou, J.; Divne, C.; Davies, G.; Sinning, I.; Stahlberg, J.; Reinikainen, T.; Srisodsuk, M.; Teeri, T.; Jones, T. 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.* **1997**, *272*, 383–397.

(27) Lin, Y.; Silvestre-Ryan, J.; Himmel, M. E.; Crowley, M. F.; Beckham, G. T.; Chu, J.-W. Protein Allostery at the Solid–Liquid Interface: Endoglucanase Attachment to Cellulose Affects Glucan Clenching in the Binding Cleft. *J. Am. Chem. Soc.* **2011**, *133*, 16617–16624.

(28) Divne, C.; Ståhlberg, J.; Teeri, T. T.; Jones, T. A. High-Resolution Crystal Structures Reveal How a Cellulose Chain is Bound in the 50 Å Long Tunnel of Cellobiohydrolase I from Trichoderma reesei. *J. Mol. Biol.* **1998**, *275*, 309–325.

(29) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(30) Mackerell, A. D. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

(31) Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(32) Darden, T.; York, D.; Pederson, L. Particle Mesh Ewald: An *N* log(*N*) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089.

(33) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(34) Brokaw, J. B.; Haas, K. R.; Chu, J.-W. Reaction Path Optimization with Holonomic Constraints and Kinetic-Energy Potentials. *J. Chem. Theory Comput.* **2009**, *5*, 2050–2061.

(35) Haas, R. K.; Chu, J.-W. Decomposition of Energy and Free Energy Changes by Following the Flow of Work along Reaction Path. *J. Chem. Phys.* **2009**, *131*, 144105.

(36) Haberthür, U.; Caflisch, A. FACTS: Fast Analytical Continuum Treatment of Solvation. *J. Comput. Chem.* **2008**, *29*, 701–715.

(37) Iannuzzi, M.; Laio, A.; Parrinello, M. Efficient Exploration of Reactive Potential Energy Surfaces Using Car–Parrinello Molecular Dynamics. *Phys. Rev. Lett.* **2003**, *90*, 238302.

(38) Silvestre-Ryan, J.; Lin, Y.; Chu, J.-W. "Fluctuograms" Reveal the Intermittent Intra-Protein Communication in Subtilisin Carlsberg and Correlate Mechanical Coupling with Co-Evolution. *PLoS Comput. Biol.* **2011**, *7*, e1002023.

(39) Boraston, A. B.; Bolam, D. N.; Gilbert, H. J.; Davies, G. J. Carbohydrate-Binding Modules: Fine-Tuning Polysaccharide Recognition. *Biochem. J.* **2004**, *382*, 769–781.

(40) Payne, C. M.; Bomble, Y. J.; Taylor, C. B.; McCabe, C.; Himmel, M. E.; Crowley, M. F.; Beckham, G. T. Multiple Functions of Aromatic–Carbohydrate Interactions in a Processive Cellulase Examined with Molecular Simulation. *J. Biol. Chem.* **2011**, *286*, 41028–41035.