RESERVE THIS SPACE

Modeling the Cellulosome Using Multiscale Methods

Yannick J. Bomble, Michael F. Crowley, Qi Xu, Michael E. Himmel

BioEnergy Science Center
National Renewable Energy Laboratory
Golden CO 80401 USA

Deriving renewable liquid fuels from biomass using microbial conversion, which utilizes free enzymes or cellulosomes for degrading cell wall material to sugars, is an attractive solution for today's energy challenges. The study of the structure and mechanism of these large macromolecular complexes is an active and ongoing research topic worldwide, with the goal of finding ways to improve biomass conversion using cellulosomes. Here, we present methods for illuminating the structure and function of systems of this size and complexity using molecular modeling. We show examples of these methods as applied to a range of sizes and time scales from atomistic models of enzymatic modules to coarse-grained models of the entire cellulosomal complex of scaffold and enzymes. Normal mode analysis, fluctuations, hydrogen-bond analysis of enzymes, as well as sampling techniques for cellulosome assembly are described and the results presented. For example, the mechanism of the immunoglobulin-like module of GH9 proteins is shown to be

RESERVE THIS SPACE

determined largely by hydrogen bond networks, and the exact hydrogen bonds were identified. Finally, by using coarse-grained modeling and parameter scanning techniques, the assembly of cellulosomal complexes is shown to be dominated by their size and shape and not by their mass.

Introduction

The most common processes for producing fuels from biomass require fermentation by either yeast or bacteria after fermentable sugars are produced. A new thrust in the field of cellulosic ethanol production is the study of microorganisms capable of converting biomass directly to fermentable products using a process known as Consolidated Bioprocesssing (CBP). Several organisms are good candidates for such a task, including *Clostridium thermocellum*, which produces large enzyme complexes known as cellulosomes. Cellusomes differ from free cellulases in the sense that most of the catalytic enzymes are strongly bound to a scaffolding protein.

The cellulosome concept was first introduced by Bayer and coworkers as the cellulase system of *C. thermocellum* (1, 2, 3). In most cases, the cellulosome is composed of two subunits – a non-catalytic scaffolding and the enzymes that attach to it by a cohesin-dockerin mechanism. A strong interaction exists between the multiple cohesin modules on the scaffoldin and the enzyme-borne dockerin modules (4, 5). The primary scaffoldin of the cellulosome from *C. thermocellum*, cellulosome-integrating protein (CipA), contains a carbohydrate binding module (CBM), which binds strongly to plant cell wall polysaccharides and nine cohesins, and is thus able to accommodate nine different enzymes. The CBM modules are also present in some cellulosomal enzymes; for example, the processive endoglucanase CbhA, a family 9 glycosyl hydrolase (GH9) (6, 7).

It has been recognized that different types of cohesins and dockerins exist in different microbial species and that the recognition between cohesin and dockerin is both type- and species-specific. Several research groups have used these findings to try to understand and improve the action of cellulosomes using a so-called “designer cellulosome” by assembling different types of cohesins from different microbial species. Bayer and coworkers (1, 8, 9) used this idea to probe two different questions: (1) do the enzyme patterns on the scaffoldin provide a synergistic action on crystalline cellulose, and (2) is there the potential to assemble enzymes from different species with superior activities on different substrates? The first engineered cellulosome was composed of two cohesins able to accommodate two cellulases (10, 11). The resulting chimeras exhibited enhanced activity on crystalline cellulose over the same free cellulases. In 2005, Fierobe and coworkers created a new tri-functional engineered cellulosome by developing a third divergent cohesin-dockerin pair (12). The tri-functional engineered cellulosome was found to be superior in function when compared to the bi-functional one. When the tri-functional engineered cellulosome was decorated with one hemicellulase (GH10) and two cellulases, it performed with superior activity on both cellulose and hemicellulose in hatched straw.

Another aspect of great interest is the origin of the possible synergistic functions of the cellulosome. One of the main explanations for the cellulosome’s performance is the flexibility of its quaternary structure. It has been shown that restricting enzyme mobility negatively affect cellulase activity, thus implying that flexibility is a key ingredient in the function of the cellulosome (13, 14).

Molecular simulations are helpful for gaining a deeper understanding of the function and versatility of the CipA assembly. Knowing the enzymatic environment necessary to attain a particular enzyme configuration on the scaffold gives insight into the way a microbial cell regulates the cellulosome population and composition near a cell wall. Probing the role of the plasticity of the cellulosome on its dynamics and self-assembly process is also an important goal. Determining the function and mode of action of the primary cellulosomal enzymes and modules may help design an improved cellulosome with improved activity. Several numerical modeling techniques can be used to answer these questions, including the more detailed all-atom molecular dynamics simulations to the less computationally expensive coarse-grained models.

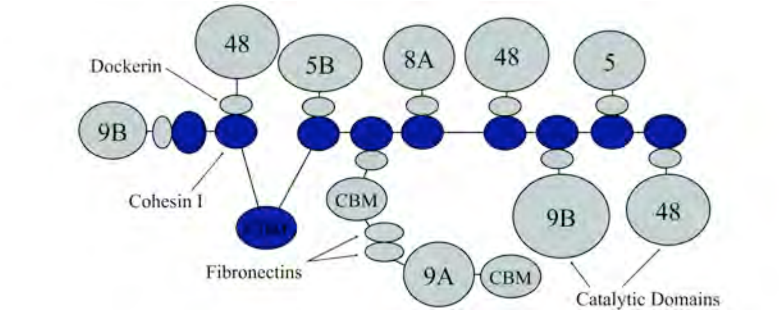


Figure 1. Concept of the first coarse-grained model for the CipA of C. thermocellum. The scaffoldin subunit (blue) contains nine cohesins and a carbohydrate binding module. The cellulolytic enzymes (grey) bind to cohesin partners with their dockerins.

Cellulosome concept and architecture

Cellulosomes from *C. thermocellum* can adopt different structures from the simplest three to nine-cohesin scaffoldin proteins to the more complex assemblies of multiple scaffoldin proteins organized on an additional scaffold, OlpB. In this chapter, our discussion will be solely based on the nine-cohesin stucture of CipA (Figure 1). A list of the CipA components and the cellulosomal enzymes considered in this chapter can be found in Table I.

Table I: Architecture of the cellulosomal protein complexes

Protein	Modules	Molecular Mass (kDa)
CipA	2COH-CBM3a-7COH	197
Cel5B	GH5-DOC	64
Cel48A	GH48-DOC	83
CbhA	CBM4-GH9-2FN3-CBM3b-DOC	138

The linkers between CipA modules vary greatly in length and are important contributors to the flexibility of the cellulosome. Cellulosomal enzymes can have simple structures, including two modules (a dockerin and a catalytic module) connected by a flexible linker, or be more complex with more than seven modules. The cellulosome is believed to bind to cellulose with the CipA-borne CBM3, but other complex enzymes whose architectures include CBMs are also believed to provide additional anchors. Moreover, many cellulosomal enzymes contain different types of carbohydrate binding modules, making them more appropriate to handle different types of substrates. Some CBMs seem to have an anchor function, whereas others have been hypothesized to be helper CBMs capable of holding a single cellulose chain and feeding it to its catalytic module partner (15). Several cellulosomal enzymes have protein modules with unknown function, such as immunoglobulin-like modules that are believed to stabilize the catalytic modules of family 9 enzymes. Fibronectin-like modules, also known as X-domains, are another case of a protein module whose function in the cellulosome is not understood. In general, fibronectins are believed to play the role of cellulose disruptors and facilitate the digestion of cellulose.

Despite the number of different modules present in the cellulosome, its quaternary structure is stable because of the high affinity between cohesins and dockerins. As mentioned earlier, in *C. thermocellum* this affinity is non-specific, and each dockerin can equally bind to any cohesin. The type I cohesin-dockerin complex is shown in Figure 2. The recognition strip, involving two helices on the dockerin and several beta strands on the cohesin, provides an almost planar binding surface. This interaction is mediated by Ca^{2+}, which is essential for the complex to maintain structure (4, 16).

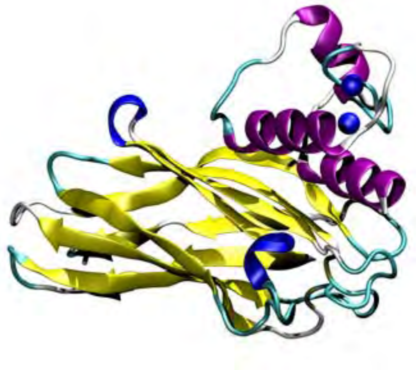


Figure 2. Structure of the cohesin-dockerin complex from the cellulosome of C. thermocellum (1OHZ) color coded by structures.

The cellulosome is an amazingly complex molecular assembly that can degrade cellulose using a wide variety of enzyme combinations, which are probably adjustable as the nature of the substrate changes. Any insight into the formation and action of the cellulosome would help us understand the roles of such complex systems in the natural degradation of cellulose and cell walls by bacteria.

Function of some cellulosomal modules

C. thermocellum produces a wide variety of enzyme families; among them, the family 9 enzymes are intriguing because they contain both endoglucanase and exoglucanases and can have rather complex architectures. They are divided into four groups based on their constructs (17), groups A through D. Group A includes enzymes containing only a catalytic module that can be linked to a dockerin. In the case of Cel9M in C. cellulyticum (18), group B includes an additional CBM3a located at the C-terminus (19). Group C includes enzymes with an immunoglobulin-like module at the N-terminus of the GH9 (20) catalytic module. Finally, group D includes enzymes that contain a CBM4 module and an Ig-like module at the N-terminus of the GH9 (7) catalytic domain.

The immunoglobulin-like module found in several of the family 9 cellulases from *C. thermocellum*, which belong to group C and D (Figure 3), is a protein module without a well-known mode of action. One of the main hypotheses for its mechanism is simply that it provides stability to the catalytic module. It has been shown that removing the Ig-like module will reduce the activity of several catalytic modules drastically (21). The mechanism by which the Ig provides this stability is still uncertain; and, while a possible mechanism has been proposed, there is no clear evidence supporting it. Several simulation techniques can be used to probe the hypothesized mechanism (see next section).

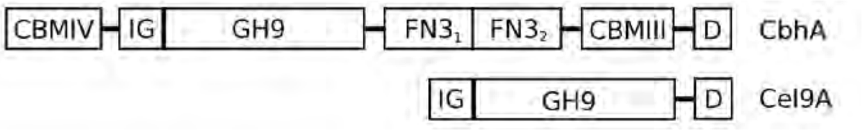


Figure 3. Constructs of CbhA and Cel9A.

Several family 9 enzymes from *C. thermocellum* exhibit an immunoglobulin-like module attached to a catalytic module. Specifically, CbhA, Cel9A, and CelK have been shown to lose most of their enzymatic activity upon removing the Ig-like module. This Ig-like module consists of about 99 amino acids directly attached to the catalytic module via a interface involving close to 40 amino acids from both modules. Several studies have investigated the possible causes of such a phenomenon in CbhA. One should note that only one x-ray structure each is available for Cel9A and CbhA. Both structures exhibit the same construct, with ten hydrogen bonds at the Ig-catalytic domain interface. However, only three of the ten hydrogen bonds are conserved between CbhA and Cel9A. These three hydrogen bonds are believed to contribute to the function of the Ig-like module by stabilizing the catalytic module as well as the catalytic cleft. In both enzymes, there exist hydrogen bond networks that appear to stabilize or at least mediate catalytic residues. Both CbhA and Cel9A have a catalytic cleft with several aromatic residues able to interact with, and thus guide, a cellulose chain (Figure 4).

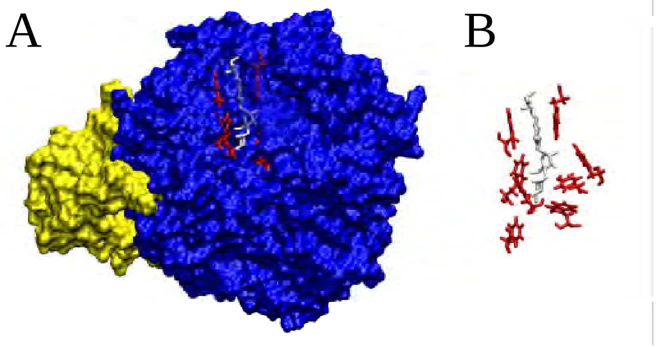


Figure 4. Surface plot of the Ig-like protein (yellow) and GH9 (Blue) with cellotetraose (white). (A) licorice representation is used to highlight the aromatic residues in the catalytic cleft. (B) Blow-up of the aromatic residues located in the catalytic cleft.

The hydrogen bond network described for Cel9A is shown in Figure 5. Both Thr-23 and Asp-51 form conserved hydrogen bonds with Gly-399; whereas Asp-53 forms a strong hydrogen bond with Tyr-408, which is located on a flexible loop connected to an important catalytic residue, Trp-410. Trp-410 is close to the substrate cleavage site. The experimental work of Kataeva and coworkers (in which Thr-23, Asp-51, and Asp-53 were mutated to alanyl residues) showed that several mutants could be created *in silico* to analyze the importance of each hydrogen bond on the dynamics and structure of the catalytic module. They also analyzed the configuration resulting from the removal of the Ig-like protein.

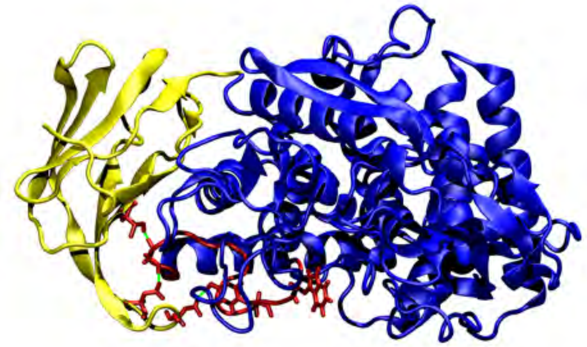


Figure 5. Structure of the catalytic (blue) and immunoglobulin-like (yellow) modules from Cel9A. The hydrogen bonding network, including an important catalytic residue, is shown in red and the three hydrogen bonds in green.

Computational methods

Several computational methods are well suited to study these systems and span the different length scales and complexity present in cellulosomal systems. Here we present some strategies for using these methods to provide insight into the potential improvement of CBP microorganisms.

Coarse-grained modeling of the cellulosome assembly

Advances in computer architectures and molecular mechanics packages have allowed larger and larger simulations; systems with more than 100,000 atoms can now be routinely simulated for hundreds of nanoseconds or microseconds (22). Also, coarse-grained modeling has been a critical addition to the computational techniques available when simulating larger macromolecular assemblies representing millions of atoms by utilizing a reduction in the number of particles by a factor of up to 10-20 (23, 24, 25, 26). While these techniques are useful, they are not well suited to the study of the formation of large macromolecular assemblies, such as cellulosomes. To understand how the cellulosome assembles close to the cell wall in a free-enzyme bath, we plan to conduct hundreds of simulations on the timescale of hundreds of nanoseconds with more than 1 million atoms. We will use the coarse-grained model proposed here to attempt to capture the most essential properties of the cellulosome and predict how these intrinsic properties will govern the enzyme configuration on the CipA scaffold. We also hope to gain insight into the dynamics of the cellulosome during and after its initial formation.

Functional form and parameters

The protein structure model consists of large spheres, called "beads," representing large regions of protein volume, up to 30 Angstroms radius, that are held together by a network of restraints to mimic the shape and flexibility of globular proteins, dockerins, cohesins, and linkers. These beads have no charge, and there is very little attractive potential between the beads. Each sphere, or bead, represents from three amino acids for linker regions to tens of amino acids in large globular protein regions. The restraints between beads are defined to be as simple as single bonds between beads in a linker, to networks of bonds between beads in globular-shaped protein modules. Special interactions are included to mimic the attraction of dockerins for cohesins. The model was developed within CHARMM (a molecular mechanics program package) (27). The CHARMM package offers considerable flexibility to the user for creating new pseudo atoms, has functionality for specific non-bond interactions between particular atom types, and allows additional parameters to be specified in the topology and parameters files.

Within our template, the interactions between coarse-grained beads can be expressed as a sum of traditional classical bonded and non-bonded terms as follows.

Non-Bonded terms

The non-bonded interactions are represented by a 6-12 Lennard-Jones (LJ) potential energy function,

$$(1)\quad E_{nb} = \sum_{i,j>i} \varepsilon_{ij} \left[\left(\frac{r_{\min}}{r} \right)^{12} - 2\left(\frac{r_{\min}}{r} \right)^{6} \right]$$

where $r_{\min}$ represents the closest distance of approach between two particles, ε_{ij} is the strength of their interaction, and r is the distance between two pseudo atoms. The vdW radii are defined to accurately reproduce the radii of the module represented by the pseudo atoms, and the interaction is defined to produce a shallow LJ potential well, so as to avoid unnatural attractions between pseudo atoms. The coarse-grained beads approximate hard spheres that have limited interactions with one another.

The electrostatic effects were neglected in our model because of the limited number of pseudo atoms or beads per protein (Figures 6-11). A specific interaction was added between the pseudo atoms in the binding site of the cohesin and dockerin proteins using an additional set of non-bonded parameters between specific atom pairs. The binding energy was set to 13 Kcal.mol^{-1}, a value between the experimental (12 kcal.mol^{-1}) (5) and theoretically determined value of 14.5 kcal.mol^{-1} (28).

Bonded terms

The bonded interactions are defined by the internal energy terms,

$$(2)\quad E_b = \sum k_r (r - r_0)^2 + \sum k_\theta (\theta - \theta_0)^2 + \sum k_\varphi (1 + \cos(\varphi - \varphi_0))$$

where r, θ, and φ are the distance, angle, and torsional angles between connected coarse-grained beads; r_0, θ_0, and φ_0 are the coarse-grained bond, angle, and torsional angle equilibrium values; and k_r, k_θ , and k_φ are the force constants. The force constants between beads of the same module are large, making the substructure rigid, while inter-modular linker regions have a wide range of flexibility. The distance, angles, and torsional angles were chosen to fit the original (all-atom) structure.

Scaffold subunit

The polymeric scaffold of *C. thermocellum* CipA, includes nine cohesin proteins connected by linker peptides of 10–30 amino acids in length and an additional carbohydrate binding module, CBM3 (Figure 6 and Figure 7). To provide the flexibility of the all-atom structure, each linker bead in the coarse-grained representation represents three amino acids (Figure 7). The all-atom and the coarse-grained representations of the full-length CipA are shown in Figure 7 and Figure 6, respectively. The linker regions offer the plasticity required by the cellulosome to assume the most appropriate configuration given a particular substrate. There is a clear need for a finer grained representation of these linkers than the coarser grained representation of the other components.

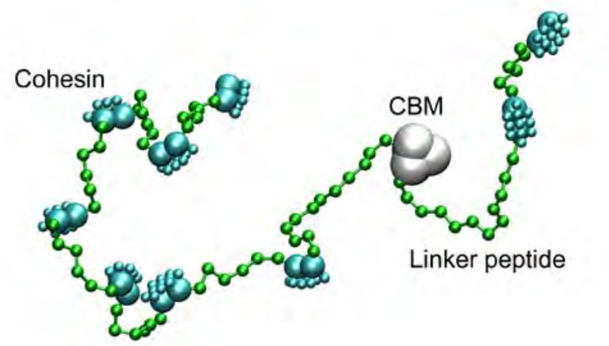


Figure 6. Coarse-grained representation of CipA from C. thermocellum.

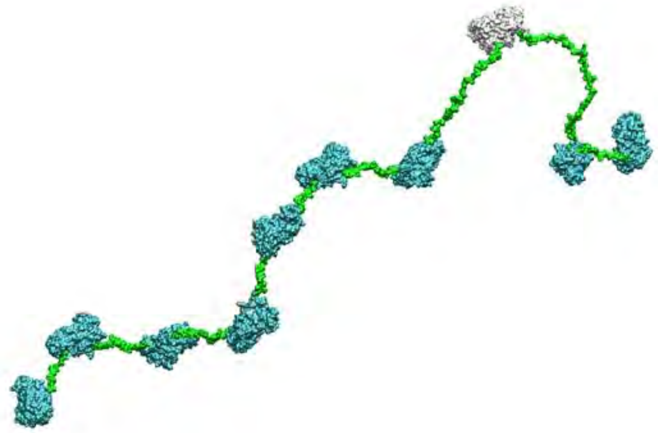


Figure 7. All-atom representation of CipA from C. thermocellum. The structure of one of the cohesins is known and reported in the literature. The other cohesins were obtained from homology modeling.

Cohesin and dockerin

The cohesins have a flat binding surface able to interact with the dockerin subunits of the cellulosomal enzymes. The architecture of the coarse-grained cohesin was conceived to accurately describe the binding interaction and create a flat binding surface while conserving the overall van der Waals volume of the protein module (Figure 8). The dockerin is constructed with a mating flat surface to match the cohesin. There are three special "attractor beads" in a row across the center of the mating surface of the cohesin and dockerin that are given special attracting properties for each other. The attractor beads are surrounded on the backside of the mating surface by beads that prevent multiple bindings to the same cohesin or dockerin simply by steric hindrance.

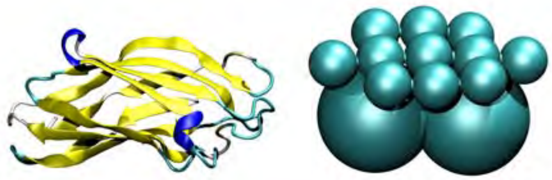


Figure 8. All-atom and coarse-grained representations of the cohesin from CipA.

Cellulosomal enzymes

As mentioned earlier, *C. thermocellum* is able to produce a wide variety of enzymes with different architecture and complexity. Three of these enzymes were selected in our study: the exocellulase Cel48A, the endoglucanase Cel5B, and the processive endoglucanase CbhA. They essentially encompass the complexity of the cellulosomal enzymes found in *C. thermocellum*. The construct details for these enzymes and the scaffoldin protein can found in Table I. The linkers between modules vary greatly in length, between 3–10 amino acids. Cel5B and Cel48A have a rather simple architecture including a catalytic module, a linker, and a dockerin. CbhA is a much more complex modular protein, including modules with mostly unknown functions, such as fibronectin-like (7, 29) and immunoglobulin-like modules (7, 21), as well as two types of carbohydrate binding modules, CBM3b and CBM4 (7). All of the enzymes studied here have a dockerin protein capable of binding to any cohesin on the scaffold without specificity; and the coarse-grained model dockerins, similar to cohesins, have an engineered flat binding platform. The coarse-grained representations of these enzymes are shown in Figures 9-11 along with their all atom counterparts. Note that the shape of the enzymes is accurately reproduced; and we should be able to model some important properties in our simulations, such as volume exclusion, mass effects, and flexible linkers.

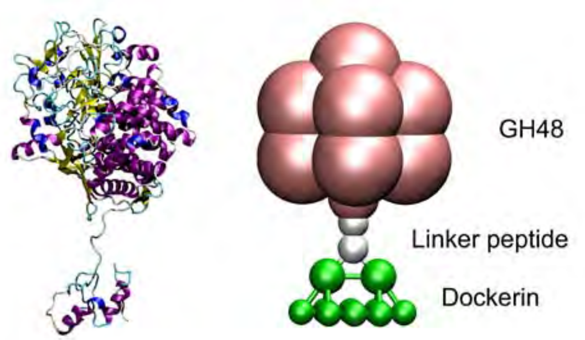


Figure 9. All-atom and coarse-grained representations of Cel48A.

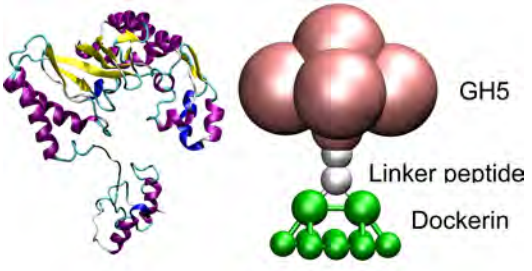


Figure 10. All-atom and coarse-grained representations of Cel5B.

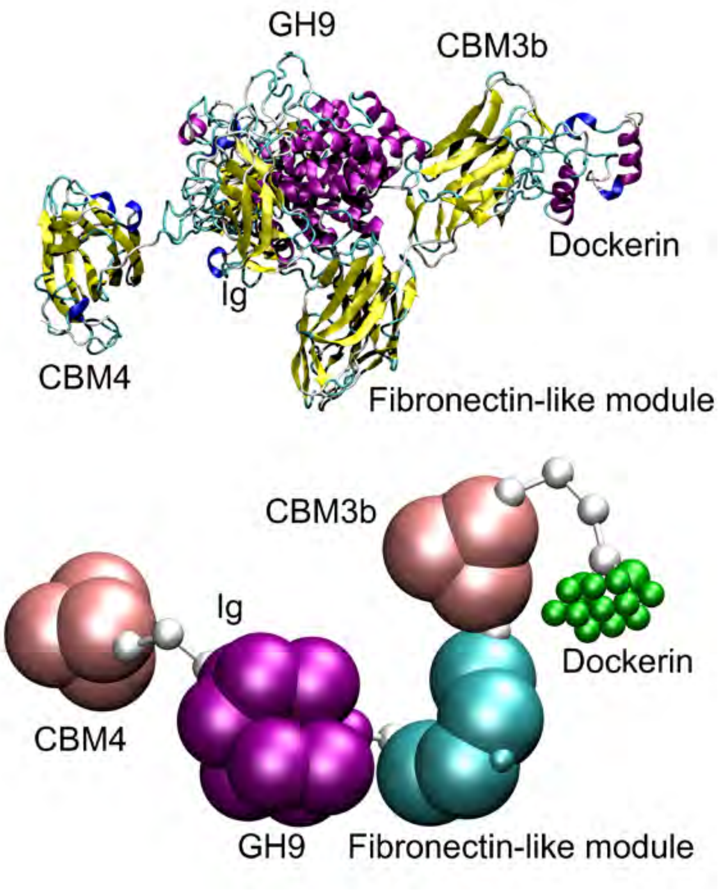


Figure 11. All-atom and coarse-grained representations of CbhA.

Experimental setup

All simulations were conducted with the CHARMM package. We tried to reproduce the enzymatic environment around the scaffoldin close to the cell wall. The simulation box has a volume of $1x10^9$ $Å^3$ (1000Å x 1000Å x 1000Å) (Figure 12). The total enzyme concentration varies from 30–120 total enzyme molecules per scaffoldin molecule and per box. The initial configurations were always randomly generated, and different random seeds for both the initial positions and the initial velocities were used to reproduce the random nature of the enzymatic environment and to eliminate the possibility of biases in our results. Initial simulations were performed with the full-length scaffold (9 cohesin). However, for clarity, the second part of this study used a 4-cohesin scaffold. Periodic boundary conditions employing a cubic box with sides measuring 1000 Å ensured a fixed concentration of enzymes in each simulation. Non-bonded interactions were cut at 99 Å, and the individual snapshots were registered every 1000 steps. Each trajectory was equilibrated for 100,000 steps with a time step of 2 fs, and trajectories were run for 30–100 ns. In our subsequent binding studies, we performed 30 simulations of 30-ns duration for each different configuration in which total concentration, ratio of enzymes, or binding constants were varied to achieve meaningful statistical analysis.

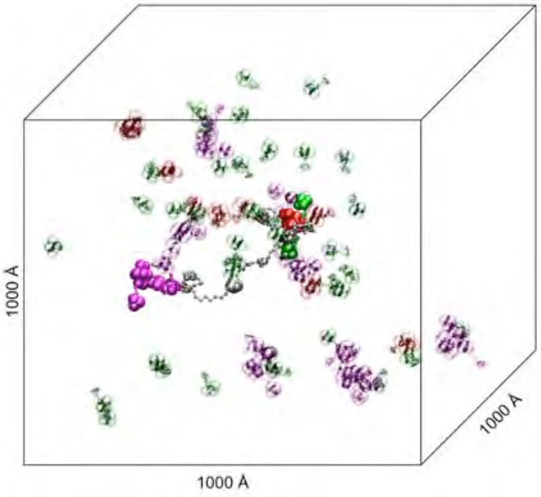


Figure 12. Simulation box with a scaffoldin molecule and 60 cellulosomal enzymes. The enzymes bound on the scaffold have solid colors.

Results and discussion

First observations were made using a 9-cohesin scaffold in the simulation box without any enzymes in solution. The scaffold adopts compact configurations reminiscent of the TEM images by Mayer and coworkers (30). Starting from an extended configuration, the scaffold tends to adopt a more compact form. In this configuration, the scaffold may be more shielded from the outside, which might explain results found by Bayer and coworkers (31). They showed that removing enzymes docked on the scaffold was easier when the cellulosome was bound on cellulose where it would adopt a more extended configuration, but much harder when free in solution.

The second observations were made when enzymes were added to the system. When an equal ratio of each enzyme is added for a total enzyme count of 60, the scaffold is fully populated with enzymes within less than 50 ns. The behavior of CipA is greatly modified whenever CbhA binds to a cohesin, which is caused by the large mass of that enzyme; but is not as affected by the binding of smaller enzymes. CbhA seems to lock the scaffold in a given location and prevents it from freely diffusing through the box the way it did before binding occurred. This behavior contributes to the nature of sequential binding of enzymes on the remaining binding sites, because the scaffold will not be able to diffuse freely. Also, the volume excluded by the first enzymes binding to the scaffold is a contributing factor in defining the probability of other enzymes binding.

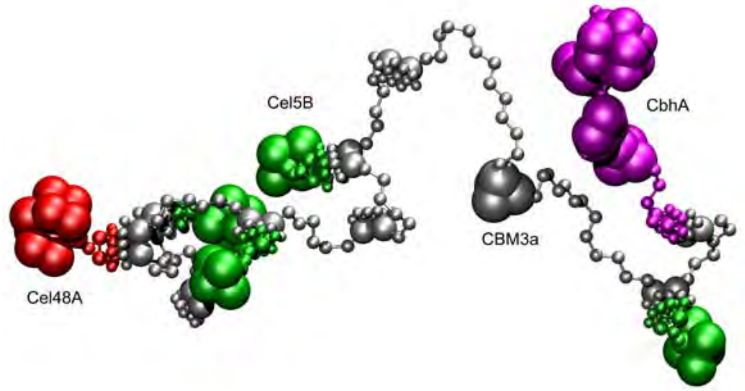


Figure 13. Coarse-grained representation of a partially populated scaffold.

The main focus of this study was to understand the driving forces behind the different cellulosome configurations. As mentioned above, we focused on a 4-cohesin scaffold without the CBM protein. Some of the results from competitive binding studies between three cellulosomal enzymes are summarized in Table II. CbhA tends to bind to the scaffoldin protein more significantly than Cel48A and Cel5B. The size or flexibility of CbhA could be responsible for this behavior; and subsequent studies varying size, mass, and radius of gyration of a given enzyme will help to understand this phenomenon.

Table II. Average cellulosome population arising from 30 replicated runs for a given ratio of enzymes in the simulation box.

Enzyme in solution Cel5B/Cel48A/CbhA (percentage)	33/33/33	41/41/18	50/50/0
Enzyme on the scaffold (percentage)	20/25/55	33/36/31	45/55/0

A detailed parameter scan of the total concentration of enzymes and enzyme ratio is being conducted and will shine more light on the binding dynamics of these enzymes that represent more than 1000 independent calculations. Response surface methodology will be used to define the environment necessary for a particular cellulosome configuration. Because of its modularity, it appears that the CbhA enzyme doesn't diffuse as quickly as the Cel5B and Cel48A because of its increased number of internal motions and therefore has more time to "feel" a cohesin partner. However, the results shown in Table II already indicate that this model could provide great insights into the cellulosome self-assembly and how the cell might regulate its scaffold configuration. There is even the possibility that the binding behavior of CbhA could be linked to the expensive nature of its construction, and that the cell doesn't need to secrete large amounts of this enzyme to be significantly present in cellulosomal assemblies.

Cellulosomes may attain their activity through their plasticity and special arrangements of the enzymes on the scaffold. Coarse-grained modeling proved to be an adequate tool to study these phenomena. However, more detailed simulations are needed to truly understand the interaction of the cellulosome with cellulose and the function of each individual protein involved in the hydrolysis process. These proteins include catalytic modules, carbohydrate binding modules, and modules such as the fibronectin-like or X domains. Several of these proteins, such as the fibronectins, have an unknown function, and others seem to have functions that differ from their fungal counterparts. In particular, several of the cellulosomal CBMs seem to have a unique function. In the next section, we study the family 9 enzyme of *C. thermocellum*, which contains many of these protein modules with different physical and chemical properties.

Normal mode analysis of Cel9A

Normal mode analysis (NMA) (32, 33, 34, 35) provides a computationally inexpensive way to study large-scale behaviors of molecular assemblies. NMA has several advantages over classical molecular dynamics (MD), even though it approximates the global potential by a harmonic function (34). First, it provides a clearer representation of the collective motions of biomolecules through a few of the lowest energy vibrational modes. Second, it makes evaluating entropy contributions and other thermodynamic properties straightforward. Finally, it is more affordable when long timescales are required for sampling times sufficient to display the low-frequency modes. While it is common practice to use elastic-network model or all-atom normal mode analysis in gas phase to approach this problem, some of the finer details may be lost in the process. Recently, NMA was extended to take advantage of the popular generalized born theory for implicit treatment of solvation effects. This new implementation (36, 37) was applied to long nucleic acid duplexes and was shown to accurately describe large-scale properties of these duplexes (37). The same method can be used as a first approach to gather information about the possible function of the Ig-like module as well as the mechanism by which GH9 endoglucanases may acquire a cellulose chain before hydrolysis of the 1,4-beta-D-glucosidic linkage.

The normal mode analyses were carried out with the molecular mechanics program package NAB (38, 39), now part of Amber10 (40, 41) ambertools using the parameter set parm99SB (42, 43); and we used the pairwise approach of Hawkins and coworkers for the Generalized Born (GB) model (44, 45). The structures were minimized using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno Truncated Newton Conjugate minimization technique to obtain a root mean square (RMS) gradient below 1 x 10^{-8} kcal/mol-Å. This level of convergence is necessary to avoid contamination from translational and rotational modes into true internal modes. The diagonalization of the Hessian matrix was done using the ARPACK (46) routines in combination with a Cholesky decomposition and inversion of the Hessian matrix, therefore providing a better separation of the eigenvalues to enhance convergence. The analysis of the normal modes was performed with a modified version of the program PTRAJ with additional functionalities. The first four normal modes of the Ig-GH9 module are shown in Figure 14 using a porcupine representation. It is commonly acknowledged that the first 10-20 normal modes are enough to describe the large-scale dynamics of a given molecule. Twenty normal modes were enough to converge the root-mean square fluctuations (RMSF) shown in Figure 16, and the first five modes dominate the fluctuations. The most dominant mode (mode1) shows a hinge motion opening the catalytic cleft around the substrate chain. The dominant motion could shine some light on the possible mechanism by which the enzyme acquires a cellulose chain before catalysis. The other normal modes are more localized, but still show a lot of motion at the bottom of the cleft as well as the flexible nature of the Ig module with respect to the CD module and within itself. Also shown is another hinge motion between the Ig and CD modules, with the hinge being the linker between the two modules. Figure 15 shows the flexible regions of the Ig-CD construct for Cel9A as determined from residue fluctuations. CbhA exhibits the same basic frequency modes and overall fluctuations as Cel9A. The high flexibility regions include loops and alpha helices at the bottom of the catalytic module close to the substrate. The atomic fluctuations calculated using the normal modes for Cel9A agree with with the atomic fluctuations calculated from crystallographic temperature factors β_i using Equation 3 and are compared in Figure 16.

$$(3) \quad \left\langle (r_i)^2 \right\rangle = \left(\frac{3\beta_i}{8\pi^2} \right)$$

$$(4) \quad Corr(i,j) = \frac{\left\langle \Delta r_i \bullet \Delta r_j \right\rangle}{\sqrt{\left\langle (\Delta r_i)^2 \right\rangle \bullet \left\langle (\Delta r_j)^2 \right\rangle}}$$

$$(5) \quad \left\langle \Delta r_i \bullet \Delta r_j \right\rangle = \sum_{k=7}^{3N} \frac{k_b T}{\lambda_k} \frac{d_{ik} d_{jk}}{\sqrt{m_i m_j}}$$

$$(6) \quad \left\langle (\Delta r_i)^2 \right\rangle = \sum_{k=7}^{3N} \frac{k_b T}{\lambda_k} \frac{{d_{ik}}^2}{m_i}$$

While the amplitudes of the fluctuations are not necessarily important, in contrast the relative fluctuations are a more relevant comparison to b-factors.

In this case, they describe the main features well. The relative fluctuations agree with experimental measurements of B-factors. This is reassuring and supports the accuracy of the NMA protocol used here.

The eigenvalues and eigenvectors can also be used to describe the correlation of motion of different protein modules. This is described by Equations 4-6 where d_{ik} and d_{jk} are the vector displacements for the kth mode and atom i or j, respectively. The cross-correlation maps of Ig-GH9 calculated for Cel9a are shown in Figure 17. The immunoglobulin-like module shows a strong correlation of motion within itself, probably due to the fact that it is composed of beta strands with strong interactions. One of the most interesting features of these maps is the fact that the Ig module, or at least several residues within the module, appear to have a strong correlation of motion with several residues of the catalytic module, including a strong positive correlation with residues 389 to 410 and also several other important loops within the vicinity of the catalytic cleft. This supports the hypothesis that these loops are closely coupled with the Ig module and that the removal of Ig or selected mutations in Ig may interfere with the dynamics of the catalytic residues, especially amino acid 410. However, a more careful investigation is required to unambiguously prove the function of the Ig module. NMA at least shows that the hypothesis mentioned earlier is relevant and deserves to be studied with a more time-consuming method such as MD simulations.

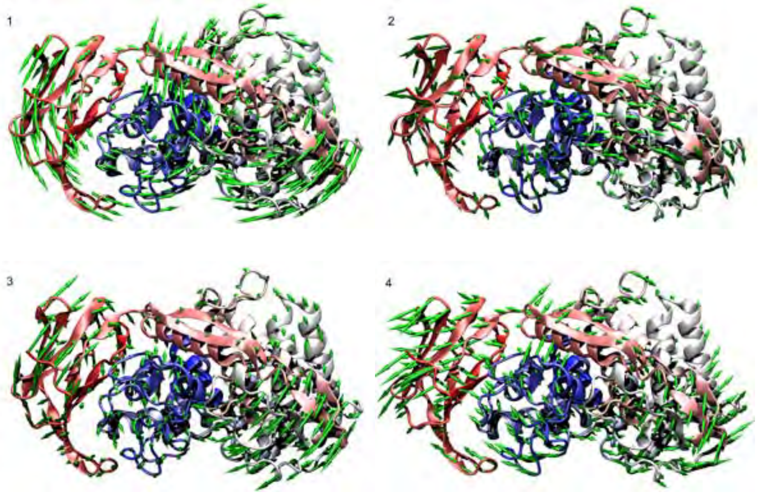


Figure 14. First four normal modes of Ig-Gh9 modules for Cel9A. The structure is color coded by amino acid sequence number.

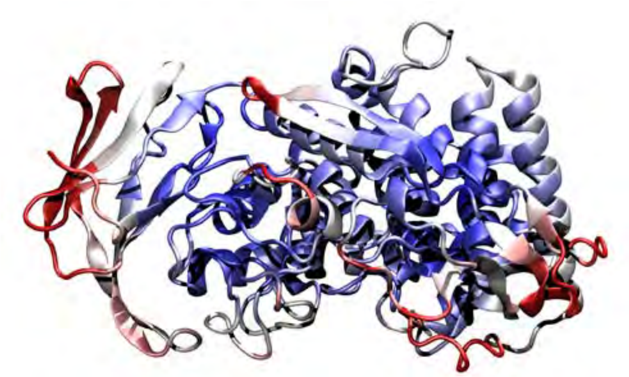
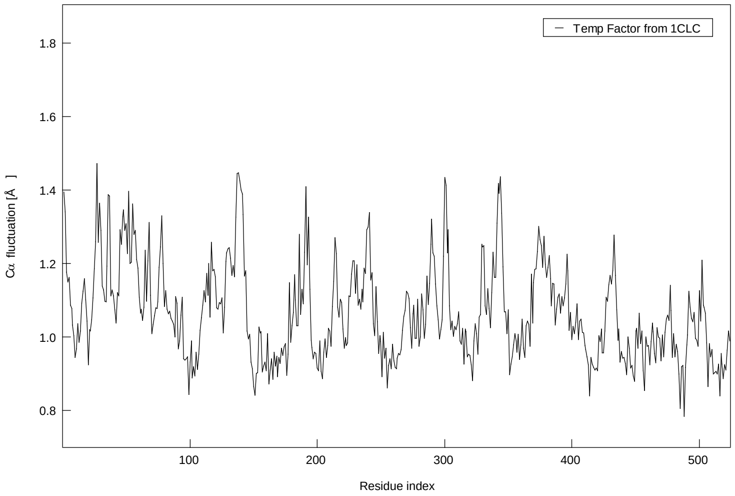


Figure 15. Stucture of Ig-Gh9 from Cel9A color coded by fluctuations (increasing from blue to red) using the first 300 normal modes.



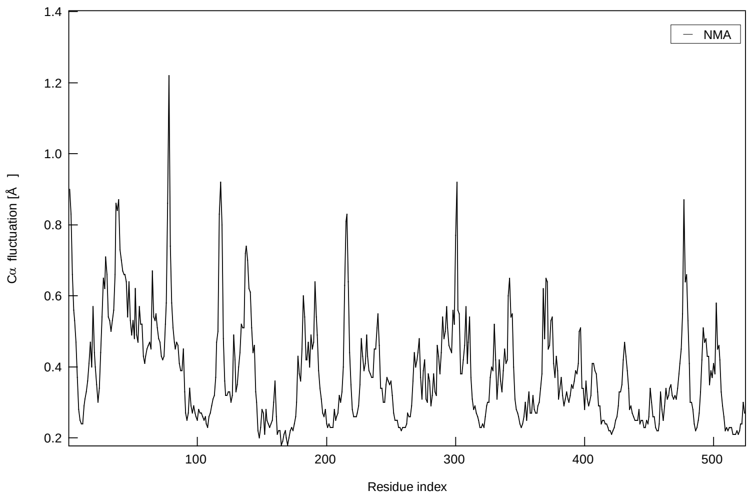


Figure16. Atomic fluctuations for Ig-Gh9 (Cel9A-1CLC) from crystallographic temperature factor and from normal mode analysis using the first 300 normal modes. The amplitutes are in Angström.

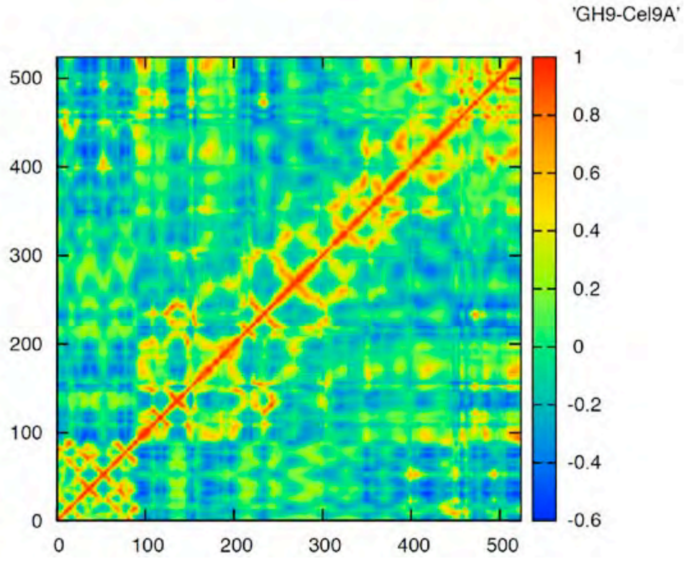


Figure 17. Residue cross-correlation map of Ig-Gh9 for Cel9A. A value of 1 shows a correlation of motion, while -1 is indicative of anti-correlation of motion, and zero represents a total lack of motion correlation. This map was calculated using the first 300 normal modes.

Molecular dynamics simulations

MD simulations were used to address the aforementioned problem – the function of the Ig module in several family 9 enzymes – in more detail using a set of analysis tools demonstrated in similar studies (47). All simulations in this section were carried out using the program, PMEMD, from Amber 10 and the parameter set parm99SB (42, 43). The proteins were solvated in a truncated octahedral box of TIP3P water molecules extending to 12 Å from the surface of the protein. A simulation time step of 2 fs was used along with SHAKE (48) to constrain covalent bonds between heavy and hydrogen atoms. The particle mesh Ewald method was used along with a non-bonded cutoff of 12 Å. The calcium ions were kept in their original positions from the pdb files, and the parameters usec for the calcium ions were taken from Aqvist (49). After equilibration, 15 ns of unconstrained MD were performed for dynamic sampling of states. Three replicates of the same starting configuration were run with different initial velocities to check the convergence of the fluctuations and other properties extracted from the trajectory and to insure proper statistical sampling. Removing the rotations and translations from the trajectories was done by rmsd, fitting the trajectory to the backbone of the entire protein in its initial post-equilibration configuration. Using a selected area of low mobility of the protein as inferred by NMA for rmsd fitting resulted in comparable findings. However, closer inspection of the cross-correlation map in Figure 19 shows that the rms fitting procedure is of crucial importance – as Ichiye and Karplus pointed out (50) – where a poor choice of rms fitting parameters can result in a loss of details in such map. It was clear from the map computed in this work that even the best set of parameters does not offer as much constrast as provided by normal mode analysis.

The RMS fluctuations of the Cα atom of the wild-type Ig-GH9 are in as good agreement with those calculated from crystallographic temperature factors (Figure 18) as the fluctuations calculated from the normal mode analysis. The fluctuations from the three replicas are almost indiscernible, except for a few flexible loops where the results are not as consistent. Given the overall consistency of the results, any difference in fluctuations caused by mutation can be linked to the effect of the mutation. It is worth noticing that the fluctuations calculated from MD are overestimated, as is always the case in the literature.

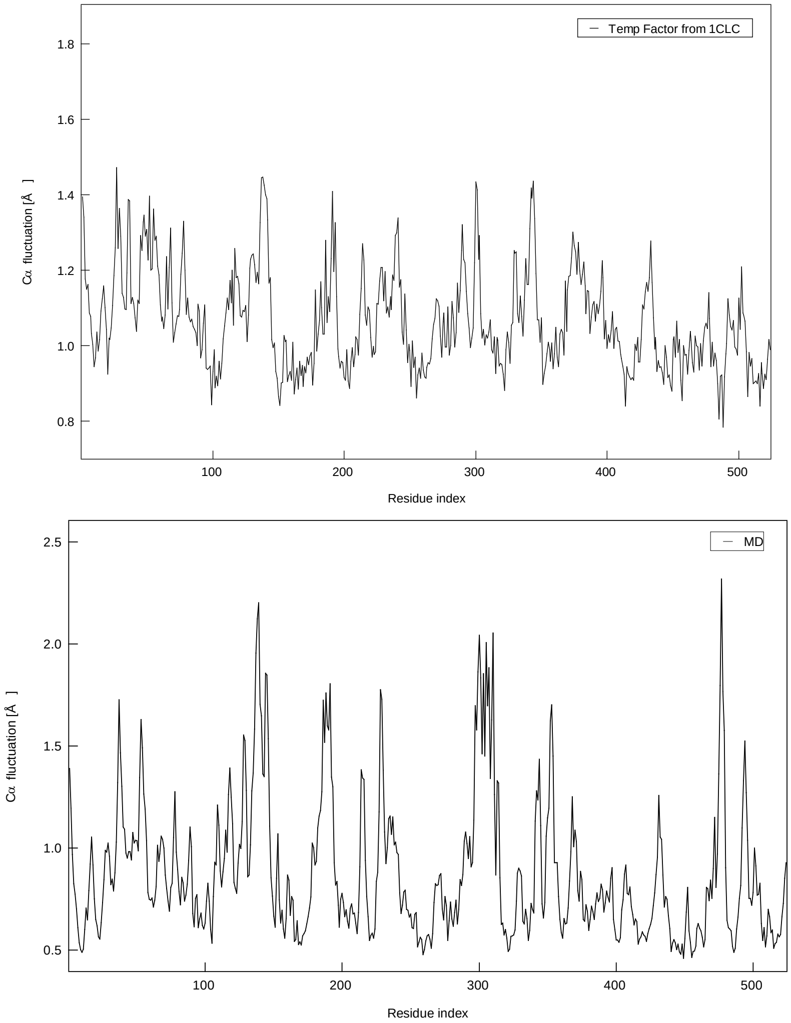


Figure 18. Atomic fluctuations for Ig-Gh9 (Cel9A-1CLC) from crystallographic temperature factor and from 15ns of molecular dynamics simulation.

Before starting experimental mutational studies, it is important to know which amino-acid residues are most likely to impact the structure or dynamics of the catalytic module. As mentioned earlier, three hydrogen bonds in Cel9A at the Ig-GH9 interface are conserved but their respective stability is unknown. The analysis program, PTRAJ, was used to follow the stability of those hydrogen bonds during 15 ns of MD simulations (Figure 20). It appears that only Asp-53 is able to create strong hydrogen bonds between the Ig and CD module in Cel9A. Thr-23 is also able to create a stable hydrogen bond in Cel9A. This analysis shows that Asp-51 is unable to strongly interact with the catalytic module as previously thought. It appears that only one or two of these conserved hydrogen bonds are good candidates for mutagenesis. A similar investigation for the remaining hydrogen bonds is being conducted; and even though these hydrogen bonds are not evolutionarily conserved, they most likely contribute to the interaction between Ig and the catalytic cleft.

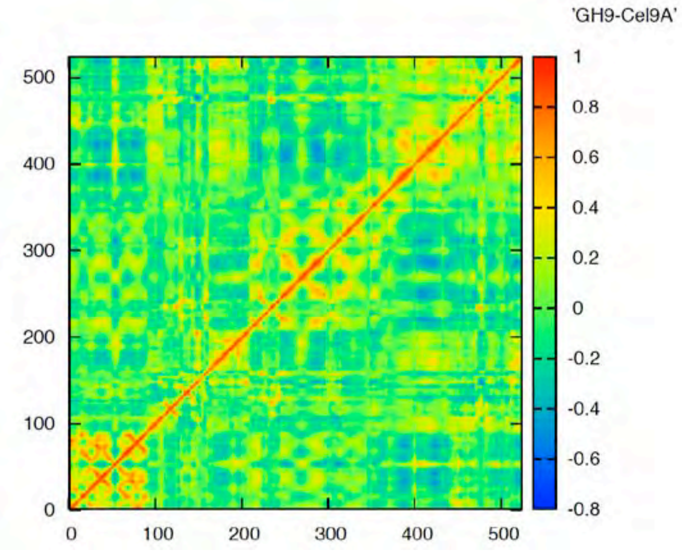


Figure 19. Residue cross-correlation maps of Ig-Gh9 Cel9A from 30 ns of molecular dynamics simulations. A value of 1 shows a correlation of motion, while -1 is indicative of an anti-correlation of motion, and zero represents a total lack of motion correlation.

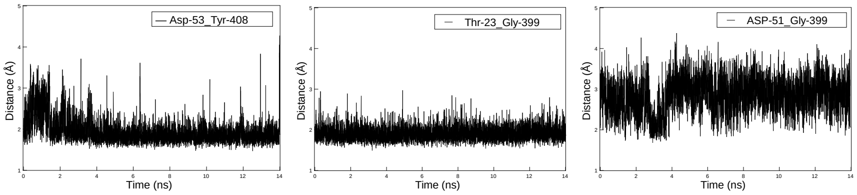


Figure 20. Distance between atoms involved in several hydrogen bonds between the Ig and catalytic modules over 15ns of simulation for Cel9A.

The effect of the extreme case of the Ig module's total removal is shown in Figure 21, where the fluctuations of the Cα atoms for Ig-GH9 and GH9 in Cel9A seem to present interesting differences in the vicinity of residues 390 to 425 as well as other less relevant loops. The features of the fluctuations appear to be substantially different and are not only restricted to a difference in the amplitude of a single peak. It would be encouraging to see the same behavior in some of the mutational studies for conserved or not conserved hydrogen bonds, as it would validate the hypothesis presented here. It is clear that dynamics of some of the residues inside the catalytic cleft are being perturbed, although it is not yet clear how this could affect the correct functioning of the enzyme. Substantial conformational changes have not been observed in these rather short simulations. Longer simulations with a generalized born model are being conducted as well as clustering analysis of the trajectory to better understand the difference in states visited for the wild type and mutated enzyme.

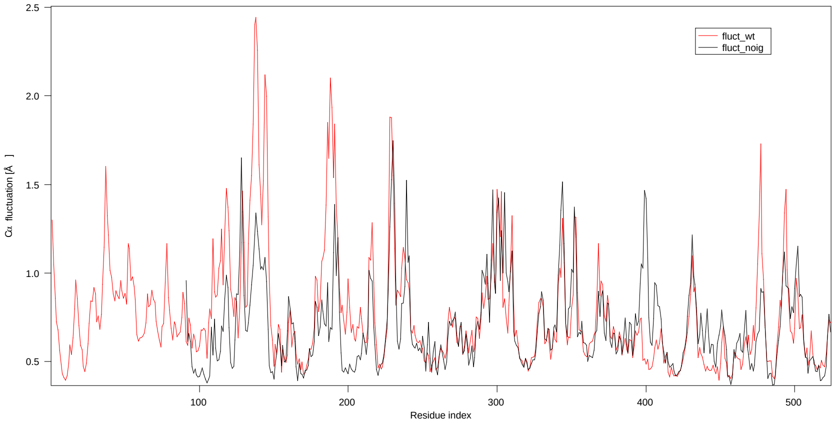


Figure 21 Atomic fluctuations for Ig-Gh9 (Cel9A) for the wild type and after ablation of the Ig module from 15ns of molecular dynamics simulation.

Conclusions

Whereas the results from normal mode analysis and molecular dynamics simulations to date are not enough to provide a definite answer about the function of the immunoglobulin-like module or the mode of action of the GH9 endoglucanases, they do seem to show the close relationship between the catalytic cleft and the Ig module. These computational tools demonstrate that the hypothesis presented several years ago is viable and that more careful analysis of this problem is not only needed, but worthwhile. Understanding the function of each individual protein (modules) of the *C. thermocellum* cellulosome is essential for improving the microorganism's performance in terms of biofuels production. Such understanding would impact both the improvement of the enzymes as well as cellulosomes.

Acknowledgments

This work was supported by the DOE Office of Science, Office of Biological and Environmental Research, through the BioEnergy Science Center (BESC), a DOE Bioenergy Research Center. Computational time for this research was supported in part by the Golden Energy Computing Organization at the Colorado School of Mines using resources acquired with financial assistance from the National Science Foundation and the National Renewable Energy Laboratory. Simulations were also performed in part using the Texas Advanced Computing Center Ranger cluster under the National Science Foundation Teragrid grant number TG-MCB090159. The authors would also like to thank Professor Ed Bayer for helpful discussions.

References

1. Bayer, E. A.; Belaich, J. P.; Shoham, Y.; Lamed, R. Annual Review of Microbiology **2004,** 58, 521-554.
2. Demain, A. L.; Newcomb, M.; Wu, J. H. D. Microbiology and Molecular Biology Reviews **2005,** 69(1), 124-+.
3. Doi, R. H.; Kosugi, A. Nature Reviews Microbiology **2004,** 2(7), 541-551.
4. Chauvaux, S.; Beguin, P.; Aubert, J. P.; Bhat, K. M.; Gow, L. A.; Wood, T. M.; Bairoch, A. Biochem. J **1990,** 265(1), 261-265.

5. Carvalho, A. L.; Dias, F. M. V.; Nagy, T.; Prates, J. A. M.; Proctor, M. R.; Smith, N.; Bayer, E. A.; Davies, G. J.; Ferreira, L. M. A.; Romao, M. J.; Fontes, C. M. G. A.; Gilbert, H. J. Proceedings of the National Academy of Sciences of the United States of America **2007,** 104(9), 3089-3094.
6. Schubot, F. D.; Kataeva, I. A.; Chang, J.; Shah, A. K.; Ljungdahl, L. G.; Rose, J. P.; Wang, B. C. Biochemistry **2004,** 43(5), 1163-1170.
7. Zverlov, V. V.; Velikodvorskaya, G. V.; Schwarz, W. H.; Bronnenmeier, K.; Kellermann, J.; Staudenbauer, W. L. J. Bacteriol. **1998,** 180(12), 3091-3099.
8. Bayer, E. A.; Shimon, L. J. W.; Shoham, Y.; Lamed, R. Journal of Structural Biology **1998,** 124(2-3), 221-234.
9. Bayer, E. A.; Lamed, R.; Himmel, M. E. Curr. Opin. Biotechnol. **2007,** 18(3), 237-245.
10. Fierobe, H. P.; Mechaly, A.; Tardif, C.; Belaich, A.; Lamed, R.; Shoham, Y.; Belaich, J. P.; Bayer, E. A. J. Biol. Chem. **2001,** 276(24), 21257-21261.
11. Fierobe, H. P.; Bayer, E. A.; Tardif, C.; Czjzek, M.; Mechaly, A.; Belaich, A.; Lamed, R.; Shoham, Y.; Belaich, J. P. J. Biol. Chem. **2002,** 277(51), 49621-49630.
12. Ding, S. Y.; Bayer, E. A.; Steiner, D.; Shoham, Y.; Lamed, R. J. Bacteriol. **1999,** 181(21), 6720-6729.
13. Gilbert, H. J. Mol. Microbiol. **2007,** 63(6), 1568-1576.
14. Hammel, M.; Fierober, H. P.; Czjzek, M.; Kurkal, V.; Smith, J. C.; Bayer, E. A.; Finet, S.; Receveur-Brechot, V. J. Biol. Chem. **2005,** 280(46), 38562-38568.
15. Jindou, S.; Xu, Q.; Kenig, R.; Shulman, M.; Shoham, Y.; Bayer, E. A.; Lamed, R. FEMS Microbiol. Lett. **2006,** 254(2), 308-316.
16. Lytle, B. L.; Volkman, B. F.; Westler, W. M.; Wu, J. H. D. Arch. Biochem. Biophys. **2000,** 379(2), 237-244.
17. Bayer, E. A.; Shoham, Y.; Lamed, R. In *The Prokaryotes, an Evolving Electronic Resource for the Microbiological Community* Dvorkin, M.; Falkow, S.; Rosenberg, E.; Schleifer, K.-H.; Stackebrandt, E., Eds.; Springer: New York, NY,2000; 3 rd, pp 1–41.
18. Belaich, A.; Parsiegla, G.; Gal, L.; Villard, C.; Haser, R.; Belaich, J. P. J. Bacteriol. **2002,** 184(5), 1378-1384.
19. Sakon, J.; Irwin, D.; Wilson, D. B.; Karplus, P. A. Nature Structural Biology **1997,** 4(10), 810-818.
20. Chauvaux, S.; Souchon, H.; Alzari, P. M.; Chariot, P.; Beguin, P. J. Biol. Chem. **1995,** 270(17), 9757-9762.

21. Kataeva, I. A.; Uversky, V. N.; Brewer, J. M.; Schubot, F.; Rose, J. P.; Wang, B. C.; Ljungdahl, L. G. Protein Engineering Design & Selection **2004,** 17(11), 759-769.
22. Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. Biophys. J. **2008,** 94(10), L75-L77.
23. Noid, W. G.; Liu, P.; Wang, Y.; Chu, J. W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. J. Chem. Phys. **2008,** 128(24), -.
24. Noid, W. G.; Chu, J. W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. J. Chem. Phys. **2008,** 128(24), -.
25. Liu, P.; Izvekov, S.; Voth, G. A. J. Phys. Chem. B **2007,** 111(39), 11566-11575.
26. Villa, E.; Balaeff, A.; Mahadevan, L.; Schulten, K. Multiscale Modeling & Simulation **2004,** 2(4), 527-553.
27. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J. Comput. Chem. **1983,** 4(2), 187-217.
28. Xu, J. C.; Crowley, M. F.; Smith, J. C. Protein Sci. **2009,** 18(5), 949-959.
29. Kataeva, I. A.; Seidel, R. D.; Shah, A.; West, L. T.; Li, X. L.; Ljungdahl, L. G. Applied and Environmental Microbiology **2002,** 68(9), 4292-4300.
30. Mayer, F.; Coughlan, M. P.; Mori, Y.; Ljungdahl, L. G. Applied and Environmental Microbiology **1987,** 53(12), 2785-2792.
31. Morag, E.; Yaron, S.; Lamed, R.; Kenig, R.; Shoham, Y.; Bayer, E. A. J. Biotechnol. **1996,** 51(3), 235-242.
32. Tama, F. Protein and Peptide Letters **2003,** 10(2), 119-132.
33. Janezic, D.; Brooks, B. R. J. Comput. Chem. **1995,** 16(12), 1543-1553.
34. Case, D. A. Current Opinion in Structural Biology **1994,** 4(2), 285-290.
35. Brooks, B. R.; Janezic, D.; Karplus, M. J. Comput. Chem. **1995,** 16(12), 1522-1542.
36. Brown, R. A.; Case, D. A. J. Comput. Chem. **2006,** 27(14), 1662-1675.
37. Bomble, Y. J.; Case, D. A. Biopolymers **2008,** 89(9), 722-731.
38. Macke, T. A.; Case, D. A. In *Molecular Modeling of Nucleic Acids.*; Leontes, N. B.; SantaLucia, J., Jr., Eds.; American Chemical Society: Washington, DC,1998; 1, pp 379–393.
39. Macke, T.; Svrcek Seiler, W. A.; Brown, R. A.; Kolossvary, I.; Bomble, Y. J.; Case, D. A., *NAB Version 6*,
40. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; B., W.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A., *AMBER 10,* University of California, San Francisco

41. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. J. Comput. Chem. **2005,** 26(16), 1668-1688.
42. Wang, J. M.; Cieplak, P.; Kollman, P. A. J. Comput. Chem. **2000,** 21(12), 1049-1074.
43. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Proteins-Structure Function and Bioinformatics **2006,** 65(3), 712-725.
44. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Chem. Phys. Lett. **1995,** 246(1-2), 122-129.
45. Hawkins, C. J.; Cramer, G. D.; Truhlar, D. G. J. Phys. Chem. **1996,** (100), 19824–19839.
46. Lehoucq, R. B.; Sorensen, D. C.; Yang, C., SIAM: Philadelphia, PA, 1999
47. Gohlke, H.; Kuhn, L. A.; Case, D. A. Proteins-Structure Function and Bioinformatics **2004,** 56(2), 322-337.
48. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Journal of Computational Physics **1977,** 23(3), 327-341.
49. Aqvist, J. J. Phys. Chem. **1990,** 94(21), 8021-8024.
50. Ichiye, T.; Karplus, M. Proteins-Structure Function and Genetics **1991,** 11(3), 205-217.