**Protein Engineering, Design & Selection** vol. 23 no. 10 pp. 759–768, 2010
Published online August 3, 2010 doi:10.1093/protein/gzq049

# Probing the mechanism of cellulosome attachment to the *Clostridium thermocellum* cell surface: computer simulation of the Type II cohesin–dockerin complex and its variants

**Jiancong Xu[1,2,3,4] and Jeremy C.Smith[1,2]**

[1]Center for Molecular Biophysics, Oak Ridge National Laboratory, P.O. Box 2008 Oak Ridge TN 37831-6164; University of Tennessee, M407 Walters Life Sciences 1414 Cumberland Avenue, Knoxville, TN 37996, USA, [2]BioEnergy Science Center, Oak Ridge, TN, USA and [3]Present address: University of Washington, Seattle, WA 98105, USA

[4]To whom correspondence should be addressed.
E-mail: xujc@u.washington.edu

**Received February 1, 2010; revised June 8, 2010; accepted June 30, 2010**

**Edited by Peter Tieleman**

**The recalcitrance of lignocellulosic biomass to hydrolysis is the bottleneck in cellulosic ethanol production. Efficient degradation of biomass by the anaerobic bacterium *Clostridium thermocellum* is carried out by the multi-component cellulosome complex. The bacterial cell-surface attachment of the cellulosome is mediated by high-affinity protein–protein interactions between the Type II cohesin domain borne by the cell envelope protein and the Type II dockerin domain, together with neighboring X-module present at the C-terminus of the scaffolding protein (Type II coh–Xdoc). Here, the Type II coh–Xdoc interaction is probed using molecular dynamics simulations, free-energy calculations and essential dynamics analyses on both the wild type and various mutants of the *C. thermocellum* Type II coh–Xdoc in aqueous solution. The simulations identify the hot spots, i.e. the amino acid residues that may lead to a dramatic decrease in binding affinity upon mutation and also probe the effects of mutations on the mode of binding. The results suggest that bulky and hydrophobic residues at the protein interface, which make specific contacts with their counterparts, may play essential roles in retaining a rigid cohesin–dockerin interface. Moreover, dynamical cross-correlation analysis indicates that the X-module has a dramatic effect on the cohesin–dockerin interaction and is required for the dynamical integrity of the interface.**

*Keywords*: cellulosome/cohesin–dockerin/free-energy perturbation/molecular dynamics

## Introduction

Cellulosic biomass holds great potential as a sustainable and renewable energy source that can substitute fossil transportation fuel. To achieve a cost-effective conversion of biomass to biofuels, the problems must be overcome arising from biomass recalcitrance, i.e. the natural resistance of plant cell walls to microbial and enzymatic deconstruction (Himmel *et al*., 2007). One potential approach for improving current bioconversion efficiency is to understand and modify the properties of efficient multi-enzyme bacterial cellulosomes (Ding *et al*., 2008).

Cellulosomes are large, surface-attached systems that are primarily produced by anaerobic bacteria and can efficiently break down crystalline cellulose and other plant cell wall-associated polysaccharides (Bayer *et al*., 1998, 2004; Doi *et al*., 2003). The most extensively studied cellulosome system belongs to *Clostridium thermocellum* and comprises a variety of cellulases and hemicellulases organized around a non-catalytic integrating protein, the scaffoldin (CipA). It is widely believed that the intimate association of enzymes with complementary specificities causes an increase in synergy by bringing these enzymes into close proximity (Fierobe *et al*., 2002, 2005). The presentation of the enzyme complex on the surface of the anaerobic bacterium also ensures that the host organism is able to efficiently utilize the polysaccharides generated by the cellulolytic enzymes (Lu *et al*., 2006).

Catalytic components of cellulosomal enzymes are bound to CipA through a high-affinity non-covalent interaction of enzyme-derived Type I dockerin modules with the nine highly conserved Type I cohesin modules within the scaffoldin subunit. In addition to the cohesin domains, the CipA also harbors a cellulose-binding domain, which targets the cellulosome to its substrate, a hydrophilic X-module of unknown function and a C-terminal Type II dockerin. The Type II dockerin is responsible for anchoring the cellulosome to the proteoglycan layer of the *C. thermocellum* cell surface through high-affinity interactions with Type II cohesin domain borne by the cell-envelope proteins SdbA, OlpB and Orf2P (Lemaire *et al*., 1995; Leibovitz and Beguin, 1996; Leibovitz *et al*., 1997). Evidence for bacterial association of other mesophilic clostridia cellulosomes is, however, less clear, as they lack Type II dockerin within their scaffolding proteins: these cellulosomes may contain other domains that mediate attachment to the membrane of host organisms.

In efforts to obtain an understanding of the molecular binding mechanism responsible for the Type I and II cohesin–dockerin associations, the fundamental building blocks of the cellulosome, X-ray and NMR studies have been performed to solve the three-dimensional structures of individual domains in solution and their heterodimeric complexes. In general, both types of cohesin domain share the same overall jellyroll topology composed of nine β-strands connected by loops and arranged in two β-sheets (Carvalho *et al*., 2003, 2005). However, additional secondary structures were found in Type II cohesin, including a crowning α-helix between β-strands 6 and 7 and two β-flaps that interrupt

strands 4 and 8 (Noach *et al*., 2005). These Type-II-specific secondary structural elements border the presumed binding surface of the cohesin domain and are thought to contribute to the cohesin–dockerin interaction and specificity. The sequences of the dockerin domains comprise a 22-amino acid tandem repeat joined by a short linker region (Schaeffer *et al*., 2002), justifying a dual-binding mode of dockerin modules to cohesins as revealed by the crystal structure of the cohesin in complex with a mutated dockerin (Carvalho *et al*., 2007). Although divergent in sequence, in the complexed state, both dockerin domains possess two $Ca^{2+}$-binding loop-helix motifs resembling the classical EF-hand. In the non-complexed state, however, dockerins adopt a flexible conformation differing from that of EF-hand domains, suggesting that a conformational change occurs upon dockerin binding to the cohesin.

Although structurally related, sequence comparisons of *C. thermocellum* Type I and II cohesins and dockerins reveal only limited similarity: ~20% for cohesins and ~30% for dockerins (Mechaly *et al*., 2001), consistent with the lack of cross-type and cross-species specificity between Type I and II pairs (Leibovitz and Beguin, 1996). The overall cellulosome architecture is therefore determined by the specificities of different cohesin–dockerin interactions, and their number and distribution in different cellulosomal subunits. A thorough understanding of the binding mechanisms and specificities of the cohesin–dockerin interactions will lay a foundation enabling selective incorporation of an optimized set of enzymes into defined functional complexes in the design of improved cellulosome chimera for more efficient degradation of biomass.

The high-resolution crystal structures allow us to design appropriate computational experiments and to determine the driving forces behind the cohesin–dockerin interaction at the molecular level. In a previous work, molecular dynamics (MD) simulations and computationally demanding free-energy calculations were performed on both the wild type (WT) and D39N mutant of the *C. thermocellum* Type I cohesin–dockerin complex (Xu *et al*., 2009). The simulation results revealed the Type I cohesin–dockerin association process at a level beyond experimentally accessible detail, and also identified specific domains and amino acid residues that may be critically involved in the process. Simulations of the D39N mutant, which has been demonstrated experimentally to dramatically reduce the binding affinity (Handelsman *et al*., 2004), indicate that the mutation triggers significant global protein flexibility and causes a major change in the hydrogen-bonding network in the recognition strips.

In the present study, computational simulation studies are undertaken to gain insight into the structure–function relationship of the Type II cohesin–dockerin. The crystal structure of the Type II complex including the Type II dockerin, its neighboring X-module from the cellulosome scaffold of *C. thermocellum* and the Type II cohesin module (Type II coh–Xdoc) has been determined (Adams *et al*., 2006) and has provided direct structural insight into the mechanism of Type II cohesin–dockerin recognition and cellulosome attachment. The Type II cohesin–dockerin interface involves one face of the Type II cohesin β-barrel consisting of strands 8, 3, 6 and 5 and the planar surface formed by both $Ca^{2+}$-binding loop-helix motifs of Type II dockerin (Fig. 1). The interaction displays a very pronounced hydrophobic

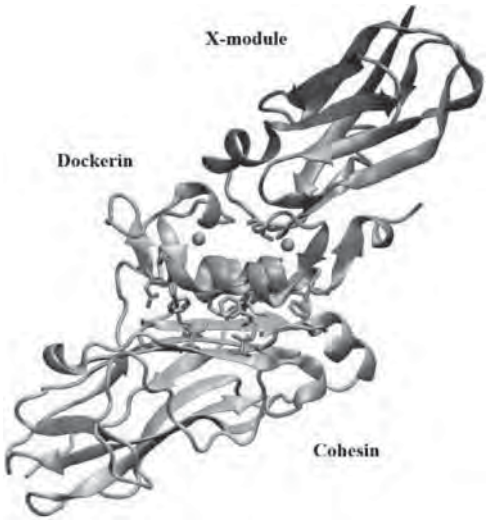


**Fig. 1.** Crystal structure of the Type II cohesin–dockerin complex in cartoon representation with cohesin in green, dockerin in orange and X-module in ice blue. Key residues involved in the interdomain interaction are highlighted in licorice mode, and colored by atom names. A colour version of Figure 1 is available as supplementary data in *PEDS* online.

feature involving a number of non-polar residues including Ile 93, Ile95, Ala110, Phe162 and Pro153 of the cohesin and Met144, Met148, Phe154, Phe121 and Val114 of the dockerin. Hydrophilic residues also exist at the interface and are likely to play an essential role in the recognition and formation of the complex: these involve Gln52, Asn54, Asn106, Ser112 and Ser151 of the cohesin domain and Asn122, Asn143 and Gln145 from α-helices of the dockerin domain. The crystal structure also revealed a hydrophobic interface between the Type II dockerin and the adjacent X-module, the latter adopting a highly stable immunoglobulin-like fold with two β-sheets. The X-module has been proposed to play a role in structural stability and enhanced solubility of cellulosomal components (Mosbah *et al*., 2000; Kataeva *et al*., 2004; Adams *et al*., 2005, 2006).

In addition to the structural determination, biophysical and mutagenesis studies have also begun to provide insight into the Type II cohesin–dockerin interaction, although fewer studies have been conducted than for Type I. An association constant value of $1.44 \times 10^{10}\ M^{-1}$ was reported for the Type II cohesin–dockerin with the X-module (Adams *et al*., 2006), which can be compared with $5.6 \times 10^{8}\ M^{-1}$ without the X-module (Jindou *et al*., 2004). Further, dockerin mutagenesis studies have revealed the importance of positions 10 and 11 in the second $Ca^{2+}$-binding loop (Met144 and Gln145, respectively) for recognition of the Type II cohesin (Schaeffer *et al*., 2002). No significant functional relevance was demonstrated for the residues at the equivalent positions in the first $Ca^{2+}$-binding loop.

In the present work, in order to identify the molecular determinants dictating the Type II cohesin–dockerin recognition and binding specificity, we computationally investigated the functional relevance of different structural elements of the Type II coh–Xdoc modules by mutational approaches. Using MD tools, a series of mutants of sites lining the protein contact surface on both the cohesin and dockerin is generated. The resulting complexes are then examined using the 'alchemical' free-energy perturbation (FEP) approach to determine whether the mutations destabilize the binding and

impair the cohesin–dockerin interaction. Further, to identify large concerted motions that could be critical to the recognition and binding functions exerted by these protein domains, an essential dynamics analysis of the WT and selected mutants is performed. Finally, the functional role that the X-module might play in tight binding between the protein partners in the Type II complex is investigated.

The paper is organized as follows: a description of the model and simulation protocols is given in the 'Methods' section. The results of the free-energy calculations and MD simulations of the WT complex and variants are presented in the 'Results and Discussion' section. Conclusions from the work are given in the final section.

## Methods

### *System preparation and MD simulations*

Starting coordinates of Type II coh–Xdoc complex were extracted from the Protein Data Bank (ID: 2B59; Adams *et al*., 2006). The model system comprises 166 residues from the SdbA Type II cohesin domain, 163 C-terminal residues of the C. *thermocellum* CipA scaffolding subunit, including the dockerin domain and the X-module, and 2 dockerin-bound $Ca^{2+}$ ions, solvated in a rectangular box of explicit TIP3P (Jorgensen *et al*., 1983) water molecules, with the total number of atoms being ~91 000. Two $Ca^{2+}$ ions were also added to neutralize the system. All the mutants referred to in this paper were constructed by replacing the mutation site residue with an Ala.

All MD calculations were performed with the NAMD software package (Phillips *et al*., 2005) with the CHARMM27 force field (MacKerell *et al*., 1998). The particle mesh Ewald approach was used for computation of electrostatic forces (Darden *et al*., 1993). Periodic boundary conditions were employed in all directions, with a shift cutoff at 12 Å for electrostatic interactions, and a switch cutoff at 10 Å for the van der Waals terms. The pair list regeneration distance of non-bonded atoms was 14 Å. The box size was adjusted to make sure that the periodic images of the protein do not overlap with the protein in the primary cell during the simulation. The starting structures were subjected to energy minimization using 1000 steps of the steepest descent and 2000 steps of the conjugate gradient method. After minimization, the structures were equilibrated by performing a 30-ps MD simulation with a weak harmonic restraint of 0.5 kcal $mol^{-1}$ $Å^{-2}$ on all $C_\alpha$ atoms. After releasing the constraints, constant number, pressure, and temperature ensemble simulations were subsequently conducted for 20 ns. The temperature and the normal pressure were maintained at 300 K and 1 bar, respectively, using Langevin dynamics and the Langevin piston method (Martyna *et al*., 1994; Feller *et al*., 1995). The first 10 ns of each simulation was considered as the equilibration phase and discarded from further analysis.

The simulation trajectories were analyzed with tools either from the GROMACS package (van der Spoel *et al*., 2005) or local code. Computer-aided structure analysis was performed using the VMD software (Humphrey *et al*., 1996).

### *Principal component analysis*

Principal component analysis (PCA) or essential dynamics analysis of an MD trajectory characterizes the collective motion in the system and separates the large-scale collective motions from random thermal fluctuations. PCA is based on the diagonalization of the covariance matrix of the interatomic fluctuations after the removal of the whole-molecule translational and rotational movement. The covariance matrix is defined as:

$$\mathrm{Cov}(i,j) = \langle (r_i(t) - \langle r_i \rangle_t) \bullet (r_j(t) - \langle r_j \rangle_t) \rangle_t$$

where $r_i$ and $r_j$ are the Cartesian coordinates of atom $i$ and $j$, respectively. The eigenvectors and eigenvalues of the covariance matrix yield the collective dynamic modes and their amplitudes.

The cosine content ($c_i$) of the principal component ($p_i$) was introduced (Hess, 2002) as a good indicator of whether a trajectory has sampled the free energy landscape sufficiently for convergence. $c_i$ can be extracted from the covariance analysis and is defined as:

$$c_i = \frac{2}{T}\left(\int \cos(i\pi t) p_i(t) \mathrm{d}t\right)^2 \left(\int p_i^2(t) \mathrm{d}t\right)^{-1}$$

The value of the cosine content varies between 0 (no cosine) and 1 (perfect cosine). It has been demonstrated that insufficient sampling could lead to high cosine content, representative of random diffusion (Hess, 2002). The evaluation of the cosine contribution to the first eigenvectors is sufficient to give a reliable idea of the convergence of the conformational sampling.

Both the average coordinates and the covariance matrix of the fluctuations about the average coordinates were calculated from the MD trajectories. In the present analysis, the trajectories determined from the last 5 ns of the equilibrated all-atom MD simulations in both the WT and mutants were used for construction of the covariance matrix. The g_anaeig and g_covar programs in GROMACS 3.3 (van der Spoel *et al*., 2005) were employed to calculate the covariance matrix elements, and porcupine plots (Tai *et al*., 2001) were used to visualize the collective dynamic modes. The first three and last three residues were removed before the PCA to avoid the incorporating excessive terminal motions.

### *Alchemical FEP calculations*

The change in free energy of binding due to point mutation of the cohesin and the dockerin domains was obtained from a thermodynamic cycle in which the free energy was computed between two distinct cohesin or dockerin domains both in the free and complexed states. For both the cohesin and the dockerin, the initial coordinates for the state free in solution were generated from the well-equilibrated WT configuration by removal of its counterpart, and then subjected to energy minimization and 10 ns equilibration using the same protocol as described in the 'System preparation and MD simulations' section.

Point mutations in both states were performed employing the FEP method (Straatsma and McCammon, 1992; Kollman, 1993; Gilson *et al*., 1997) implemented in NAMD. For each mutation, five independent FEP calculations were performed to obtain statistically reliable results. For each alchemical transformation, either in bulk water or in the bound complex, the reaction path was divided into 30 states of uneven widths, each corresponding to a different $\lambda$-value. Narrow

intermediate states were defined toward the end points of the transformation. For every $\lambda$ point 50 ps of equilibration was followed by 150 ps of data collection, corresponding to a total simulation length of 6 ns for each transformation.

### *Time-correlated atomic motions*

In order to investigate the correlated motion between different regions of a protein, such as the domain–domain communication, we calculated the cross-correlation coefficients for the $C_\alpha$ displacements using the MD trajectories with or without the presence of the X-module. The snapshots used were between 10 and 20 ns. The cross-correlation coefficient Corr($i$,$j$) is given by:

$$\mathrm{Corr}(i,j) = \frac{\langle \Delta r_i \bullet \Delta r_j \rangle}{\langle \Delta r_i^2 \rangle^{1/2} \bullet \langle \Delta r_j^2 \rangle^{1/2}}$$

where $\Delta r_i$ and $\Delta r_j$ are the vector displacements from the mean position of the $C_\alpha$ atoms in residues $i$ and $j$, respectively. These coefficients were collected in matrix form and displayed as a two-dimensional dynamical cross-correlation map (DCCM; Swaminathan *et al*., 1991). Positive values indicate that two atoms move in the same direction, whereas the negative values indicate anticorrelated motion. Prior to the DCCM analysis, translational and rotational components of the motion of the protein complex were removed by superimposing all MD snapshot structure on the X-ray crystal structure.

## Results and discussion

### *Probing hot spots on the Type II cohesin–dockerin interface*

Structural and mutagenesis studies have hitherto focused mainly on the Type I interaction, whereas our knowledge regarding the attachment of the cellulosome to the bacterial cell surface through the Type II cohesin–dockerin interaction is more limited. However, the crystal structure of the Type II coh–Xdoc complex has provided insight into structural elements that are likely to play a key role in the binding and specificity in the Type II interaction (Adams *et al*., 2006). Since the cohesin–dockerin complex is stabilized by key interaction sites on each protein surface, it is an important first step to identify these key hot spots, i.e. residues that may lead to a dramatic decrease in binding affinity upon mutation.

To identify key interaction residues in a high-throughput and systematic way, the computational alanine-scanning approach (Massova and Kollman, 1999; Kortemme and Baker, 2002; Kortemme *et al*., 2004) was used to scan the complete Type II coh–Xdoc interface. Both the experimental and the computational alanine-scanning mutagenesis have been proved to be powerful tools for analyzing important interactions in protein–protein interfaces (Cunningham and Wells, 1989; Massova and Kollman, 1999; Kortemme and Baker, 2002). Alanine scanning can reveal a cluster of large hydrophobic side chains at the Type II coh–Xdoc interface that, when mutated, show lower binding affinity to their partners. This approach is a systematic analysis that should be of general use to identify the side chains that most strongly modulate the interaction between two protein partners. It uses a simplified free-energy function to characterize the effect of the deletion of individual amino acid residue side chains beyond the β-carbon atom on the affinity. Hot spots can be operationally defined as residues showing a change in the binding free energy ($\Delta\Delta G_{\mathrm{bind}}$) by more than 1 kcal mol$^{-1}$ when replaced with alanine. By this criterion, 11 residues from the cohesin and the dockerin domains are predicted to be energetically significant (Table I), most of which possess bulky side chains. These residues are all located at the planar interface formed by the β-strands 8, 3, 6 and 5 of cohesin and the two loop-helix motifs of dockerin (Fig. 1).

The computational alanine-scanning approach is based on the assumption that the mutations do not introduce any changes in the overall binding modes of the cohesin and the dockerin, and only limited side-chain conformational changes are taken into account. This assumption may not work for all mutations. Therefore, the 11 residues identified as hot spots by alanine scanning were subjected to alchemical FEP, a more rigorous and accurate method, as outlined in the 'Methods' section. The non-conservative residues and the residues that were not identified to be significant by alanine scanning were not included in the subsequent FEP study. This selection procedure aims to identify all the mutations that lower the binding affinity and destabilize the binding.

The FEP results (Table I) show a significant loss in the binding free energies when residues Gln52 and Phe162 from the cohesin and Met144, 148 and Phe121 from the dockerin are mutated to alanine. In particular, the Met144Ala mutation causes a dramatic loss in the binding free energy of ~5.7 kcal mol$^{-1}$. The data are in agreement with the experimental observation that double substitution of Met144 and Gln145 with Ser and Thr abolishes Type II cohesin recognition (Schaeffer *et al*., 2002). Met144 sits at position 10 of the second α-helix, and both a comparison of Type I and II dockerin sequences and site-directed mutagenesis studies have suggested that the two tandem pairs at positions 10 and 11 of the two $Ca^{2+}$-binding loops collectively play a role in the binding and specificity (Schaeffer *et al*., 2002; Adams *et al*., 2005). In the crystal structure, Met144 hydrogen bonds with Gln52 of the cohesin, the residue contributing the most to the cohesin–dockerin FEP-binding energy among all sampled residues, through a backbone atom interaction (Met144-N:Gln52-Oε), and by participating in hydrophobic interactions with the side chains of a number of non-polar residues including Phe162 and Ala110. Met144, whose physical properties are significantly different from its counterpart Ser45 in the Type I dockerin, may be the most critical residue in cohesin recognition and could by itself account for the observed binding specificity in the Type II interaction, i.e. the Type II specificity might be driven by the Gln52-Met144 pair, equivalent to the Asp39–Ser45 pair in Type I. The Gln52–Met144 pair is highly conserved and central to both the polar and non-polar interactions between the protein partners in Type II.

Gln145 at position 11 has also been proposed to be important for the binding (Schaeffer *et al*., 2002; Adams *et al*., 2006). Glu145 establishes indirect hydrogen-bond connections with Pro153 at the β-flap 8 and Gln52 via bridging water molecules, and non-polar contacts with Phe162 and Val168. Analysis of the dynamic trajectories shows that the Gln145Ala mutation does not disrupt the water-mediated

**Table I.** FEP derived free-energy changes between the WT complex and the alanine exchanged complexes ($\Delta\Delta G_{bind}$, kcal mol$^{-1}$) for the Type II coh–Xdoc complex

| $\Delta\Delta G_{bind}$ (kcal mol$^{-1}$) | Cohesin | | | | | |
|---|---|---|---|---|---|---|
| | Gln52Ala | Ile93Ala | Ile95Ala | Asn106Ala | Phe162Ala | |
| Mean | 7.85 | 2.04 | 1.24 | 2.40 | 4.32 | |
| SD | 1.5 | 0.55 | 0.68 | 1.35 | 1.08 | |
| | Dockerin | | | | | |
| | Met144Ala | Gln145Ala | Met148Ala | Phe154Ala | Phe121Ala | Asn122Ala |
| Mean | 5.68 | 1.29 | 4.74 | 1.50 | 3.12 | 1.61 |
| SD | 0.48 | 0.8 | 0.6 | 0.9 | 1.10 | 1.05 |

The binding free-energy calculations and the statistical analysis were carried out on five trajectories from separate runs (see the 'Methods' section). SD, standard deviation.

hydrogen-bond interactions with Gln57 and Pro153 of the cohesin, and also the non-polar contact with Phe162 still exists, consistent with the observation of a less dramatic loss in the FEP-binding free energy (~1.3 kcal mol$^{-1}$). Therefore, Gln145 may not play as an important role in the recognition code of cohesin as does Met144.

Leu111 and Leu112, at the analogous positions of the first α-helix, were not identified to be critical by the present alanine scan, although inspection of the crystal structure suggests that these two residues may nevertheless contribute to the binding through hydrophobic interactions with Ile93 and Ile95 of Type II cohesin. Further, the mutation of Ile93 and Ile95 is also not identified as being as energetically significant as others (~2.0 and 1.2 kcal mol$^{-1}$, respectively). Hence, the calculations suggest that a single mutation of these Leu or Ile residues will not have a dramatic effect on the affinity of the Type II interaction.

The substitution of Met148 by alanine is accompanied by a ~4.7 kcal mol$^{-1}$ loss in the calculated binding free energy. The Met148 side chain faces the short crowning α-helical segment that occurs between β-strands 6 and 7 of cohesin and forms a number of hydrophobic contacts with Phe162 and Ile114. Another dockerin residue that also contributes significantly to the cohesin binding is Phe121, which is located at the C-terminal of the first α-helix with its side chain pointing toward the inner part of the cohesin–dockerin interface. The Phe121 side chain is mainly engaged in the hydrophobic interaction with Met144, whereas the backbone atoms are exposed to the solvent. The largest input to the binding energy for these hydrophobic groups may come from the gain in van der Waals interaction that is unlikely to be cancelled by unfavorable changes in the solvation free energies (Massova and Kollman, 1999). According to a survey of crystallographic structures deposited in the PDB database both Phe and Met have a high propensity to occur in protein–protein interfaces (Jones and Thornton, 1997) and may generally play relatively important roles in protein binding. In the present calculations, Phe162, Phe121, Met144 and Met148 all make large contributions to the cohesin–dockerin binding.

The mutation results above provide useful information on the mechanism of the Type II cohesin–dockerin interaction. Even though the Type II interface has a very pronounced hydrophobic feature and multiple contacts are made with the cohesin by both dockerin helices, certain surface regions play more critical roles than others in forming and maintaining the complex structure. The cohesin residues that severely weaken the binding free energy, Gln52 and Phe162, both reside in the buried inner part of the interface, a hydrophobic groove region formed by the elevated/tilted β-strands 8, 3 and 6 and surrounded by the β-flap 8 and the short α-helix at the end of the β-barrel. These segments are the distinctive structural differences between the Type I and II cohesins and are very well conserved among all Type II cohesins, indicating that these secondary structural elements might be definitive of Type II cohesins (Noach *et al*., 2005). The information obtained from the present calculations serves to restrict the critical area of the putative cohesin–dockerin binding interface to the more occluded surface formed by β-strands 8, 3 and 6 of cohesin, where the specificity-determining residues may reside. Other hydrophobic residues that are not identified to be functionally critical by our calculations, such as the Ile93 and Ile95, are situated on the outskirts of this region.

Graphical inspection of Type I and II complex crystal structures reveals that the dockerin domains possess a remarkably different orientation on the Type I and Type II cohesin surface, whereas both Type II dockerin helices contact the cohesin surface over their entire length, in Type I only limited contact is established between the first α-helix and the corresponding part of cohesin (Adams *et al*., 2006; Xu *et al*., 2009). On the basis of the distinct orientation and lack of significant symmetry of Type II dockerin interface residues, involvement of both dockerin helices at the interface with Type II cohesin has been implied (Adams *et al*., 2006). However, the mutation results presented in this study suggest that the second α-helix, which accommodates Met144 and Met148 and is also in contact with the cohesin binding area made of β-strands 8, 3 and 6, plays a more critical role than the other helix in the cohesin recognition and high-affinity binding. Whether the dual-binding model exhibited by both α-helices in the Type I dockerin (Carvalho *et al*., 2007) also exists in the Type II cohesin–dockerin binding remains to be determined and merits more extensive experimental and computational investigation. Hall and Sansom (2009) performed coarse-grained MD on the Type I cohesin–dockerin, and the simulations were able to reveal both binding modes of the complex for the WT protein and a mutant of the dockerin. It will be of interest to apply similar multiscale computational approaches to study the association and binding mode of the Type II cohesin–dockerin and its variants.

### *Collective modes of motion in the WT and mutants*

In proteins, sets of secondary structure, such as α-helices and β-sheets, may move collectively, and changes in these collective motions on complexation are indicative of mechanical cooperativity. Here, to characterize the collective modes of motion in the native and mutant proteins PCA ('essential dynamics'; Amadei *et al*., 1993) is applied. This analysis is based on the fluctuations of $C_\alpha$ atomic coordinates obtained from the trajectories generated by the MD simulations (see the 'Methods' section).

To examine whether reasonable convergence was achieved the essential dynamics analysis was applied by sampling time windows of different lengths ranging from 10 to 20 ns. The motion along any desired eigenvector can be visualized by projecting all trajectory frames onto a specific eigenvector; the trajectories thus generated reveal the motion in the direction defined by the eigenvector. In the present work, the first three largest-eigenvalue principal modes describe more than 80% of the total variance, and their amplitudes are remarkably different in the WT complex and mutants (Supplementary Fig. S1). The projections along each of the first principal components obtained by PCA are shown in Supplementary Fig. S2, along with the cosine content of the principal components. The cosine contents are generally lower than 0.5, with the smallest (0.0107) in the WT and largest (0.4520) in I93A. It should be noted that there is no conventional threshold of the cosine content value that separates sufficient and insufficient sampling. However, previous studies by Maisuradze *et al*. (Maisuradze and Leitner, 2007; Maisuradze *et al*., 2009) show that for proteins such a crossover may lie somewhere around 0.5. The length of a minimum MD simulation required for the convergence of dynamics may vary from system to system. For the Type II coh–Xdoc and mutants, the present 20-ns MD simulations appear to be enough time to provide a sufficient conformational sampling and ensure the correlation between the MD simulation data and protein dynamics characteristics.

Porcupine plots of the three largest PCA modes of the WT, Gln52Ala, Phe162Ala and Ile93Ala mutants, are shown in Supplementary data (Fig. S3). Animations of the largest PCA modes are also available as Supplementary data (Movie S1). The motions along the first three eigenvectors and their relevance to cohesin–dockerin complexation are examined here.

The projection of the MD motions along the largest-amplitude PCA mode in the WT Type II coh–Xdoc complex is illustrated in Fig. 2a. The width of the ribbon indicates the amplitude of the backbone motion and the direction given as going from the red to the blue colors. This mode (Supplementary Movie S1a) involves a concerted motion involving the solvent-exposed surface of the X-module and the loop/turn regions in the cohesin. The cohesin–dockerin interaction interface containing the conserved hydrophobic

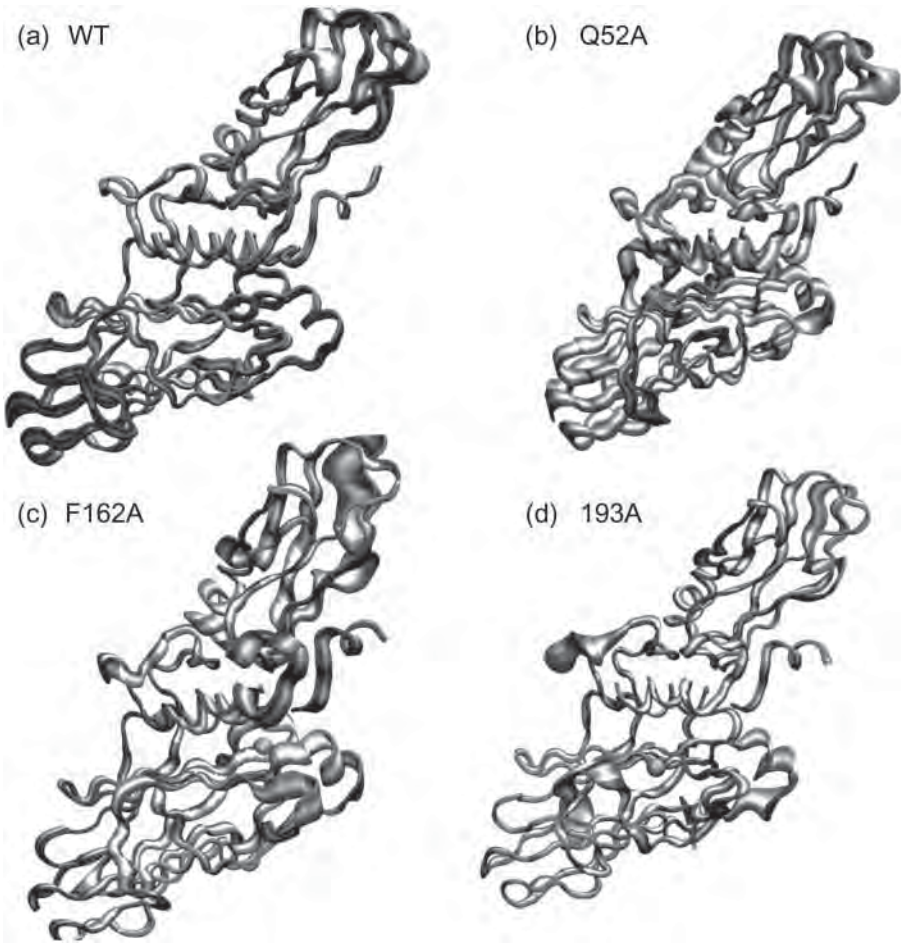


**Fig. 2.** Tube representation of the projections along the first principal component eigenvector for the (**a**) WT, (**b**) Gln52Ala, (**c**) Phe162Ala, and (**d**) Ile93Ala. $C_\alpha$ atoms of the mutation sites are indicated by the van der Waals spheres in cyan. A colour version of Figure 2 is available as supplementary data in *PEDS* online.

patch is found to move as a rigid unit. The second and the third modes (Supplementary Fig. S3a) exhibit complex mixed collective motions including a twisting of the β-flap 4 and the loop region connecting the first and the second α-helices, coupled with stretching of the X-module along the lateral direction. These modes may contribute to the domain binding by mediating both the lateral and the longitudinal contacts.

The largest-amplitude mode in the Gln52Ala mutant (Supplementary Movie S1b) is a domain-opening motion in which the cohesin and the dockerin domains move concertedly in the opposite longitudinal directions. This motion dissociates the two intimately connected domains, and as a result, the interaction surface is dramatically disturbed and the entire complex exhibits substantial mobility (Fig. 2b). Similar deformation behavior is also observed in the Phe162Ala mutant, in which the most significant mode (Movie S1c) involves the two domains rolling so as to give rise to a large shift in the positions of interfacial contacts. This behavior may arise from accessible space at the interface created by the deletion of the bulky hydrophobic side chain of Phe162. Consequently, the mode induces large-scale configurational fluctuations and structural rearrangement throughout the entire complex (Fig. 2c) and again dramatically affects the interdomain interaction, relative to the WT, within the Type II coh–Xdoc complex. In the final example, the Ile93Ala mutant (Fig. 2d), the amplitudes of the low-frequency modes are not as large as for Gln52Ala and Phe162Ala. The regions with a relatively high flexibility include the loops and terminal tails of the cohesin and the dockerin modules together with the solvent-exposed region of the X-module; the rigid cohesin–dockerin interface is, however, retained.

In summary, whereas in the erratic motion of the WT the interface is preserved, in the mutants the interface is significantly dynamically disrupted.

### *The function of the X-module*

The C-terminal region of the *C. thermocellum* scaffolding subunit adopts a rigid, elongated structure due to the intimate interface between the Type II dockerin and the X-module. The Ig-like X-modules are found not only in the scaffoldings of thermophilic and mesophilic bacteria but also in cellulolytic enzymes (Kataeva *et al*., 2004). Biophysical studies have suggested that within the scaffoldings, the X-module may act as a conformational linker and solubility enhancer and could be involved in the cellulosome attachment to the bacterial cell surface directly or indirectly (Mosbah *et al*., 2000; Adams *et al*., 2006). Free cellulases deletion of the Ig-like X-module results in complete inactivation of the catalytic module (Kataeva *et al*., 2004). In the Type II coh–Xdoc complex, the X-module directly contributes to the cohesin–dockerin binding only through hydrogen-bond connections between Ser20 of the X-module and Glu167 of Type II cohesin, but nevertheless increases the cohesin–dockerin binding affinity by 2 orders of magnitude (Jindou *et al*., 2004; Adams *et al*., 2006).

To further understand the functional role of the X-module in structural stability and cell-surface attachment, MD simulations of the cohesin–dockerin complex were carried out with and without the X-module. From the resulting trajectories, the root mean square deviation (RMSD) of the dockerin domain (Fig. 3a) with respect to the initial crystal structure was first compared. Only the cohesin was chosen for this least-squares fitting, because comparison of the Type II cohesin structure in the absence and the presence of the Xdoc modular pair indicates that the cohesin undergoes little conformational change upon binding to the Xdoc module (backbone RMSD of 0.99 Å; Adams *et al*., 2006). As shown in Fig. 3a, the mean RMSD of the dockerin $C_\alpha$ atoms in the simulation with the X-module is ∼2 Å, but increases to 4 Å

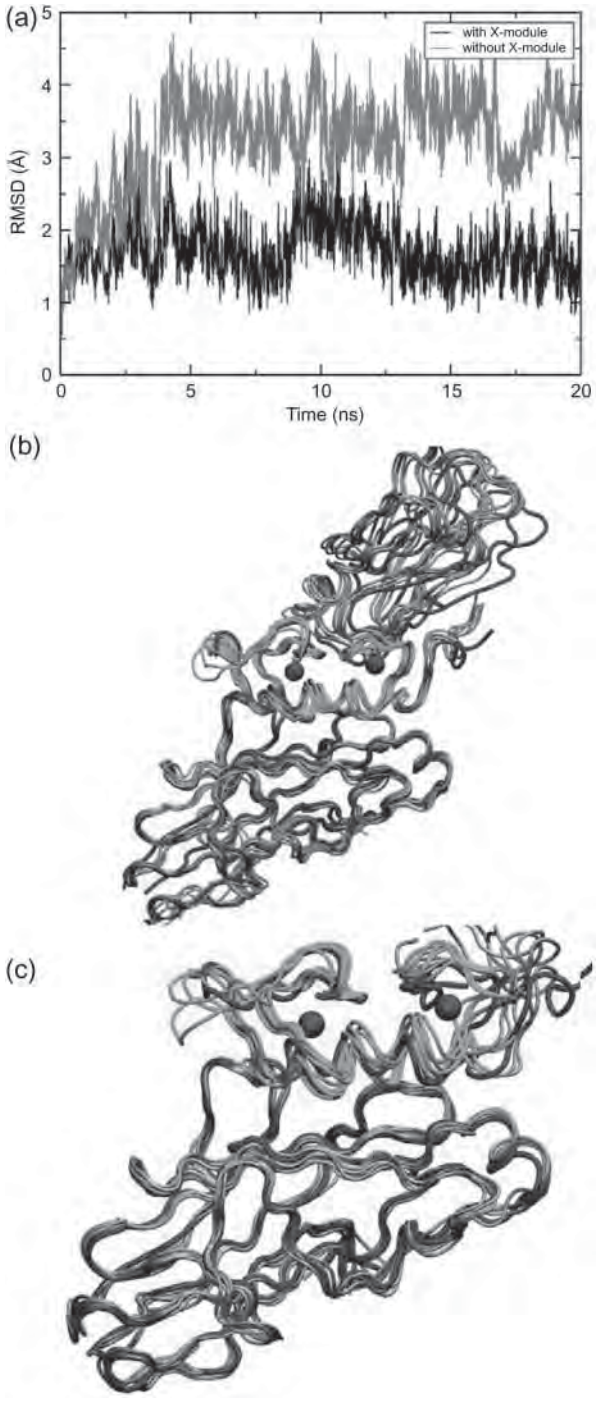


**Fig. 3.** (**a**) $C_\alpha$ RMSDs of the dockerin from MD of the systems with and without the X-module. (**b**) Displacement of the dockerin domain relative to the cohesin domain in the presence of the X-module and (**c**) in the absence of the X-module. The dockerin domain that experiences large fluctuation when the X-module is absent, including the two α-helices and $Ca^{2+}$-binding loops, is highlighted in green. A colour version of Figure 3 is available as supplementary data in *PEDS* online.

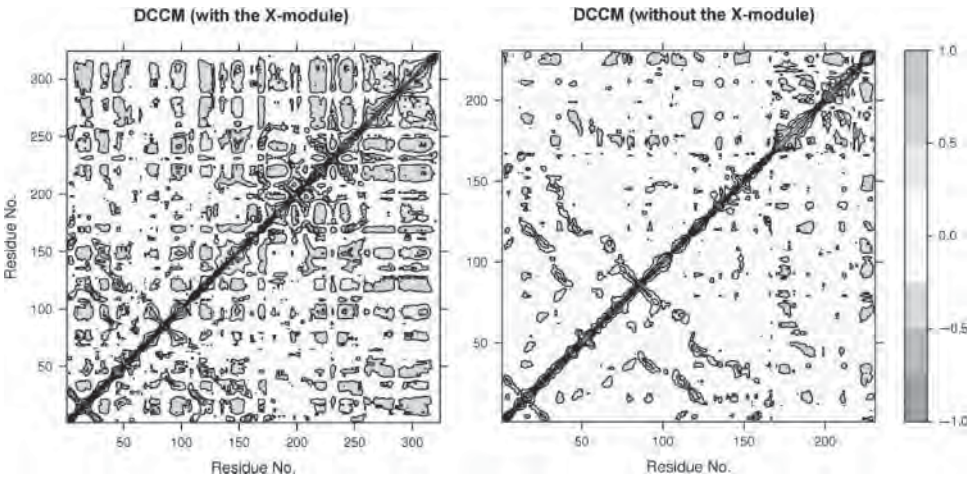


**Fig. 4.** Residue–residue-based correlated motions (DCCM) within the Type II coh–doc complex with or without the presence of the X-module. The first 166 residues are from the cohesin domain, and the rest from the dockerin (65 amino acids) and the X-module if present. A colour version of Figure 4 is available as supplementary data in *PEDS* online.

when the X-module is removed. This indicates that the dockerin, when not connected to the X-module, becomes unstable and deviates largely from the crystal structure.

Figure 3b and c shows six instantaneous structures generated along the trajectory and superimposed onto the initial X-ray structure of the cohesin domain. When the X-module is connected to the dockerin, the two α-helices that are intimately involved in cohesin-binding fluctuate very little in position relative to the cohesin. However, when the X-module is released, the displacement of the first α-helix is seen to increase substantially, being elevated by ∼3 Å, and the helix is moved away from the cohesin.

The two $Ca^{2+}$-binding loops also fluctuate more highly on X-module dissociation. It has been demonstrated experimentally that in the Type I and II complexes, both calcium-binding segments of the dockerin are required for cohesin recognition (Pages *et al*., 1997; Lytle and Wu, 1998). The present results suggest that the increased affinity may be due to the X-module mediated stabilization of the dockerin structure and that the X-module is able to restrict the flexibility and orientation of the dockerin domain, keeping the binding sites locked in place.

To further investigate how the structural dynamics induced by the removal of the X-module may affect the domain–domain communication, the DCCM representing the correlated motions was averaged over the last 5 ns and is shown in Fig. 4. The cross-correlation map averaged over the last 5 ns was found to agree with results obtained over the time interval of 10–20 ns, indicating that a converged picture of the correlated motions is obtained. Positive correlation (blue regions in the figure) indicates that the two residues generally move in the same direction, whereas a negative value (pink regions) indicates that they move in opposite directions.

In general, when the X-module is present a stronger degree of correlation is found throughout the map both within the cohesin domain and between the cohesin and the dockerin domains. When the X-module is removed the most of the correlations disappear, and domain–domain communication between the cohesin and the dockerin is completely absent. This marked qualitative change in the correlation pattern further confirms that the X-module facilitates the tight mechanical interdomain packing, enabling the cooperation of the two domains during complexation.

Closer examination of the data in Fig. 4 also reveals that the DCCM map generally exhibits strong positive cross-correlation values between interdomain regions containing those residues that are in strong contact and therefore may play roles in the binding mechanism. The regions involved include the residues previously defined as members of the β-strands 8 and 3 of the cohesin, where Gln52 and Phe162, residues with strong $\Delta\Delta G_{bind}$ (Table I), are situated, and the entire second α-helix of the dockerin, whereas weak-to-moderate positive correlation was observed for residues 279–283 of the first α-helix [i.e. residues 294–298, according to the numbering of Adams *et al*. (2006)]. Several residues from the α-helices are also positively correlated with residues 135–138 that belong to the Type-II-specific β-flap 8 (Noach *et al*., 2005). As stated above, β-flap 8 is absent from the Type I cohesin structures but has been observed in both the solution structures of the Type II cohesin (Carvalho *et al*., 2005; Noach *et al*., 2005) and the crystal structure of the Type II cohesin–dockerin complex (Adams *et al*., 2006). The flap disrupts and alters the route of β-strand 8 by forming a 12-residue raised loop that closely borders the posterior face of the complex and could potentially interfere with the dockerin binding. The functional role of β-flap 8 is not very well understood and requires further experimental and computational investigation.

Other groups of residues that exhibit positive correlations with the second α-helix of the dockerin include residues 147–151 of the β-strand 8 and residues 97–102 of the Type-II-specific short α-helical segment that occurs between β-strands 6 and 7. Anti-correlation peaks between the interdomain regions of the protein complex are also evident and are found mostly between spatially distant regions. For example, the intense blocks of negative cross-correlation seen around residues ∼295–310, 210–220 and 235–245 correspond to motions between the extended dockerin contact surface and the peripheral areas of the cohesin and the X-module. Negligible negative anti-correlation was found between the cohesin and the rest of the X-module.

## Conclusion

Characterization of the mechanisms by which anaerobic bacteria degrade plant cell walls will provide insight of potential

use in designing strategies for cost-effective conversion of biomass to biofuels. As the Type II cohesin–dockerin interaction is the basis for cellulosome-microbe attachment, it is of fundamental interest to understand the structure–dynamics–function relationships of this interaction. In this paper, the molecular determinants dictating binding and specificity in the cohesin–dockerin interaction have been examined using *in silico* simulation methods. Mutations of both the cohesin and the dockerin counterparts have identified residues that are critical to the binding and may determine specificity. Examination of essential collective modes generated by PCA of MD trajectories in the WT and selected mutants demonstrate that elimination of bulky hydrophobic side chains located at the interdomain interface, such as those of Gln52 or Phe162, induces a major dynamical perturbation and a global conformational rearrangement. Residues with the favorable combination of bulky and hydrophobic attributes can sterically hinder unproductive conformational change that may otherwise occur at the cohesin–dockerin interface, therefore conferring rigidity to the interface regions surrounding it. These key residues are present not only in *C. thermocellum* cellulosomes but also in other members of the cohesin–dockerin family from different species, as they have been identified from sequence alignment as being either invariant or highly conserved across species (Carvalho *et al*., 2005). The presently explored mechanism of Type II cohesin–dockerin recognition may therefore also be highly conserved across different prokaryotic species.

Of particular interest are the results on the communication between the cohesin and the dockerin domains, obtained here from the examination of the correlated motion between different regions of the proteins as revealed by the DCCM. The cross-correlation analysis sheds light on the mechanical basis for the interaction and reveals that the X-module acts to mechanically strengthen the Type II binding communication in the *C. thermocellum* cellulosome. Its interaction with the dockerin affects the fold and stability of the cohesin–dockerin complex and facilitates the cooperation of the two domains during complexation.

Computational work such as present analysis sets the stage for the rational development of effective chimeric molecules comprised of dockerins with divergent specificity derived from different cellulosomal modules. The specificity and high affinity of the Type I cohesin–dockerin complex have been employed using synthetic biology to design and construct artificial designer cellulosomes (Fierobe *et al*., 2002, 2005; Mingardon *et al*., 2007). The results presented here suggest 'hot' segments of the Type II cohesin–dockerin complex, and some of these energetically important residues at the interface display the cross-type specificity that might be exploited in the design of specific interaction modulators. Work in this direction is in progress.

## Supplementary data

Supplementary data are available at *PEDS* online.

## Funding

The work was supported by a grant from the BioEnergy Science Center. The BioEnergy Science Center is a US Department of Energy BioEnergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. The research was supported in part by the National Science Foundation through TeraGrid resources provided by the National Institute for Computational Sciences.

## References

Adams,J.J., Webb,B.A., Spencer,H.L. and Smith,S.P. (2005) *Biochemistry*, **44**, 2173–2182.
Adams,J.J., Pal,G., Jia,Z.C. and Smith,S.P. (2006) *Proc. Natl Acad. Sci. USA*, **103**, 305–310.
Amadei,A., Linssen,A.B. and Berendsen,H.J. (1993) *Proteins*, **17**, 412–425.
Bayer,E.A., Shimon,L.J., Shoham,Y. and Lamed,R. (1998) *J. Struct. Biol.*, **124**, 221–234.
Bayer,E.A., Belaich,J.P., Shoham,Y. and Lamed,R. (2004) *Annu. Rev. Microbiol.*, **58**, 521–554.
Carvalho,A.L., Dias,F.M., Prates,J.A., Nagy,T., Gilbert,H.J., Davies,G.J., Ferreira,L.M., Romao,M.J. and Fontes,C.M. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 13809–13814.
Carvalho,A.L., Pires,V.M.R., Gloster,T.M., Turkenburg,J.P., Prates,J.A.M., Ferreira,L.M.A., Romao,M.J., Davies,G.J., Fontes,C.M.G.A. and Gilbert,H.J. (2005) *J. Mol. Biol.*, **349**, 909–915.
Carvalho,A.L., *et al*. (2007) *Proc. Natl Acad. Sci. USA*, **104**, 3089–3094.
Cunningham,B.C. and Wells,J.A. (1989) *Science*, **244**, 1081–1085.
Darden,T., York,D. and Pedersen,L. (1993) *J. Chem. Phys.*, **98**, 10089–10092.
Ding,S.Y., Xu,Q., Crowley,M., Zeng,Y., Nimlos,M., Lamed,R., Bayer,E.A. and Himmel,M.E. (2008) *Curr. Opin. Biotechnol.*, **19**, 218–227.
Doi,R.H., Kosugi,A., Murashima,K., Tamaru,Y. and Han,S.O. (2003) *J. Bacteriol.*, **185**, 5907–5914.
Feller,S.E., Zhang,Y., Pastor,R.W. and Brooks,B.R. (1995) *J. Chem. Phys.*, **103**, 4613–4621.
Fierobe,H.P., Mechaly,A., Tardif,C., Belaich,A., Lamed,R., Shoham,Y., Belaich,J.P. and Bayer,E.A. (2002) *Carbohydrate Bioengineering: Interdisciplinary Approaches*, The Royal Society of Chemistry, Cambridge, UK, pp. 113–123.
Fierobe,H.P., Mingardon,F., Mechaly,A., Belaich,A., Rincon,M.T., Pages,S., Lamed,R., Tardif,C., Belaich,J.P. and Bayer,E.A. (2005) *J. Biol. Chem.*, **280**, 16325–16334.
Gilson,M.K., Given,J.A., Bush,B.L. and McCammon,J.A. (1997) *Biophys. J.*, **72**, 1047–1069.
Hall,B.A. and Sansom,M.S.P. (2009) *J. Chem. Theory Comput.*, **5**, 2465–2471.
Handelsman,T., Barak,Y., Nakar,D., Mechaly,A., Lamed,R., Shoham,Y. and Bayer,E.A. (2004) *FEBS Lett.*, **572**, 195–200.
Hess,B. (2002)*Phys. Rev. E*, **65**, 0319101–03191010.
Himmel,M.E., Ding,S.Y., Johnson,D.K., Adney,W.S., Nimlos,M.R., Brady,J.W. and Foust,T.D. (2007) *Science*, **315**, 804–807.
Humphrey,W., Dalke,A. and Schulten,K. (1996) *J. Mol. Graph.*, **14**, 33–38.
Jindou,S., Kajino,T., Inagaki,M., Karita,S., Beguin,P., Kimura,T., Sakka,K. and Ohmiya,K. (2004)*Biosci. Biotechnol. Biochem.*, **68**, 924–926.
Jones,S. and Thornton,J.M. (1997) *J. Mol. Biol.*, **272**, 121–132.
Jorgensen,W.L., Chandrasekhar,J., Madura,J.D., Impey,R.W. and Klein,M.L. (1983) *J. Chem. Phys.*, **79**, 926–935.
Kataeva,I.A., Uversky,V.N., Brewer,J.M., Schubot,F., Rose,J.P., Wang,B.C. and Ljungdahl,L.G. (2004) *Protein. Eng. Des. Sel.*, **17**, 759–769.
Kollman,P.A. (1993) *Chem. Rev.*, **93**, 2395–2417.
Kortemme,T. and Baker,D. (2002) *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
Kortemme,T., Kim,D.E. and Baker,D. (2004) *Sci. STKE*, **2004**, pl2.
Leibovitz,E. and Beguin,P. (1996) *J. Bacteriol.*, **178**, 5335–5335.
Leibovitz,E., Ohayon,H., Gounon,P. and Beguin,P. (1997) *J. Bacteriol.*, **179**, 2519–2523.
Lemaire,M., Ohayon,H., Gounon,P., Fujino,T. and Beguin,P. (1995) *J. Bacteriol.*, **177**, 2451–2459.
Lu,Y.P., Zhang,Y.H.P. and Lynd,L.R. (2006) *Proc. Natl Acad. Sci. USA*, **103**, 19605–19605.
Lytle,B. and Wu,J.H.D. (1998) *J. Bacteriol.*, **180**, 6581–6585.
MacKerell,A.D., *et al*. (1998) *J. Phys. Chem. B*, **102**, 3586–3616.
Maisuradze,G.G. and Leitner,D.M. (2007) *Proteins*, **67**, 569–578.
Maisuradze,G.G., Liwo,A. and Scheraga,H.A. (2009) *J. Mol. Biol.*, **385**, 312–329.

Martyna,G.J., Tobias,D.J. and Klein,M.L. (1994) *J. Chem. Phys.*, **101**, 4177–4189.
Massova,I. and Kollman,P.A. (1999) *J. Am. Chem. Soc.*, **121**, 8133–8143.
Mechaly,A., Fierobe,H.P., Belaich,A., Belaich,J.P., Lamed,R., Shoham,Y. and Bayer,E.A. (2001) *J. Biol. Chem.*, **276**, 19678.
Mingardon,F., Chanal,A., Lopez-Contreras,A.M., Dray,C., Bayer,E.A. and Fierobe,H.P. (2007) *Appl. Environ. Microbiol.*, **73**, 3822–3832.
Mosbah,A., Belaich,A., Bornet,O., Belaich,J.P., Henrissat,B. and Darbon,H. (2000) *J. Mol. Biol.*, **304**, 201–217.
Noach,I., Frolow,F., Jakoby,H., Rosenheck,S., Shimon,L.J.W., Lamed,R. and Bayer,E.A. (2005) *J. Mol. Biol.*, **348**, 1–12.
Pages,S., Belaich,A., Belaich,J.P., Morag,E., Lamed,R., Shoham,Y. and Bayer,E.A. (1997) *Proteins*, **29**, 517–527.
Phillips,J.C., Braun,R., Wang,W., Gumbart,J., Tajkhorshid,E., Villa,E., Chipot,C., Skeel,R.D., Kale,L. and Schulten,K. (2005) *J. Comput. Chem.*, **26**, 1781–1802.
Schaeffer,F., Matuschek,M., Guglielmi,G., Miras,I., Alzari,P.M. and Beguin,P. (2002) *Biochemistry*, **41**, 2106–2114.
Straatsma,T.P. and McCammon,J.A. (1992) *Annu. Rev. Phys. Chem.*, **43**, 407–435.
Swaminathan,S., Harte,W.E. and Beveridge,D.L. (1991) *J. Am. Chem. Soc.*, **113**, 2717–2721.
Tai,K., Shen,T., Borjesson,U., Philippopoulos,M. and McCammon,J.A. (2001) *Biophys. J.*, **81**, 715–724.
van der Spoel,D., Lindahl,E., Hess,B., Groenhof,G., Mark,A.E. and Berendsen,H.J.C. (2005) *J. Comput. Chem.*, **26**, 1701–1718.
Xu,J., Crowley,M.F. and Smith,J.C. (2009) *Protein Sci.*, **18**, 949–959.