PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

PCCP

ARTICLE

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Structural and Dynamic Evolution of the Amphipathic N-terminus Diversify Enzyme Thermostability in the Glycoside Hydrolase Family 12

Xukai Jiang,[a] Guanjun Chen,[a] and Lushan Wang*[a,b]

Understanding the molecular mechanism underlying protein thermostability is central to the process of efficiently engineering thermostable cellulases, which can provide potential advantages in accelerating the conversion of biomass into clean biofuels. Here, we explored the general factors that diversify enzyme thermostability in glycoside hydrolase family 12 (GH12) using comparative molecular dynamics (MD) simulations coupled with a bioinformatics approach. The results indicated that protein stability is not equally distributed over the whole structure: the N-terminus is the most thermal-sensitive region of the enzymes with a β-sandwich architecture and it tends to lose its secondary structure during the course of protein unfolding. Furthermore, we found that the total interaction energy within the N-terminus is appreciably correlated with enzyme thermostability. Interestingly, the internal interactions within the N-terminus are organized in a special amphipathic pattern in which a hydrophobic packing cluster and a hydrogen bonding cluster lie at the two ends of the N-terminus. Finally, bioinformatics analysis demonstrated that the amphipathic pattern is highly conserved in GH12 and that, besides, the evolution of the amino acids in the N-terminal region is an inherent mechanism underlying the diversity of enzyme thermostability. Taken together, our results demonstrate that the N-terminus is generally the structure that determines enzyme thermostability in GH12, and it thereby is also an ideal engineering target. The dynameomics study of a protein family can give a general view of protein functions, which will offer wide applications in future protein engineering.

1. Introduction

Cellulose that consists of β-1,4-linked glucose units is the most abundant and sustainable resource on earth. In biorefinery processes, diverse cellulases synergistically hydrolyze cellulosic biomass into soluble sugars that are subsequently converted into environmentally friendly biofuels. Biological conversion of cellulosic biomass can reduce dependence on fossil fuels and offers a chance to resolve environmental problems such as global warming and air pollution.[1, 2] Consequently, accelerating the conversion rate of biomass is one of the most crucial steps in biorefinery processes. Rapid progress towards this goal and some impressive results have been achieved through optimization of the pretreatment process and protein engineering of cellulases.[3-6] Among these efforts, the use of thermostable cellulases in industrial production provides some unique advantages, such as enhanced specific activity, higher levels of stability and inhibition of microbial contamination.[7] Unfortunately, only a few cellulases present in nature can remain both stable and active at high temperatures (above 55 °C).[8] Thus, a discrepancy between the need and the reality makes it necessary to design thermostable cellulases by protein engineering.

Protein stability is determined by the ΔG (free energy change) of approximately 5–20 kcal/mol between the native and denatured states.[9] ΔH (enthalpy change) and ΔS (entropy change) are considered to be dependent upon intramolecular interactions and conformational flexibility, respectively.[10] An evident increase in temperature causes the thermal denaturation of the protein, along with increased conformational perturbation and even structural crash.[11] Interestingly, a previous study about the human growth hormone (hGH) revealed that the concomitant increase with temperature of the electrostatic interactions stabilized the interhelical salt bridges and gave rise to the decreased flexibility and increased stability.[12] Moreover, it has been widely accepted that the loss of internal hydrophobic interaction and the change of hydration behavior as the

[a] State Key Laboratory of Microbial Technology, Shandong University, Jinan 250100, China.Email: lswang@sdu.edu.cn Phone: +86-531-88366202.
[b] State Key Laboratory of Biochemical Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, 10090, China
Electronic Supplementary Information (ESI) available: additional Figures S1 and S2. See DOI: 10.1039/x0xx00000x

temperature increases are key factors of protein denaturation.[11, 13, 14] Importantly, the protein denaturation has been proved to be non-uniform, only some parts of the structures are sensitive to the environmental change on matter if it's caused by cold or heat.[11, 15] The above described nature of protein stability and the protein thermal response imply that introducing new interactions at some specific sites seems to be an operative approach to improving the protein stability. The introduced interactions could have an effect on the flexibility of local structures and consequently modify protein stability. On the basis of this theoretical foundation, certain experiments have improved protein thermostability by incorporating new disulfide bridges,[16, 17] salt bridges,[18] hydrogen bonds[19, 20] and hydrophobic packing interactions.[21] These experimental studies have described good examples of protein engineering; however, the problem of how to quickly search for the proper targets within the whole protein structure remains a restriction of engineering studies.

Fortunately, protein unfolding studies using molecular dynamics (MD) simulations have provided novel insights into protein stability.[14, 22, 23] In these studies, the initial unfolding sites in the unfolding pathway can be localized. Such sites with high conformational flexibility are considered to be weak points in the protein structure. Stabilizing the initial unfolding sites has been shown to be an effective technique for designing protein stability.[17, 24]. Notably, previous unfolding studies have revealed that the initial unfolding sites are inconsistent across different proteins, probably due to the diversity of protein sequences and structures, which impairs the search for engineering targets. Robins et al. proposed in their review article that generally reliable rules for predicting stabilizing changes among mutations are still enigmatic, thus the optimization and screening of protein variants still requires a large research team, as well as massive amounts of experimental work.[25]

Based on sequence and structural identity, proteins can be divided into diverse protein families. Proteins in the same family often possess similar topology. Thus, it becomes interesting and meaningful to investigate whether proteins in the same family adopt a similar unfolding process related to their thermostability. Glycoside hydrolases (GH) have been divided into many different families in the carbohydrate-active enzymes database (CAZY).[26] Of these GH families, the cellulases in GH12 have been widely used in the textile and pulp industries. They typically adopt a β-sandwich fold and cleave β-1,4-linked glycosidic bonds randomly at the exposed positions of crystal cellulose with a retaining mechanism.[27] Previous studies have revealed that GH12 enzymes display several unique features such as: being induced at the initial stage of cellulose degradation with the smallest cellulases (around 20 kDa) lacking a carbohydrate binding module (CBM); being found in archaea, bacteria and Eukaryota; and possessing a wide range of optimal temperatures for growth.[27, 28] These properties of GH12 make its members ideal candidates for studying protein stability in terms of the whole GH family. Compared to previous studies that mainly focused on a single protein, this approach may contribute to a consensus view of protein stability and thus provide more guidelines for protein engineering.

In this study, we selected GH12 as a model and investigated its unfolding processes by comparative MD simulations coupled with bioinformatics analysis. As a green technique, this structural dynameomics approach allows us to understand protein functions in atomic detail with relatively little environmental loading and reagent consumption.[29] Firstly, the overall stability and unfolding dynamics of the mesophilic and thermophilic enzymes of GH12 were characterized. Further, we investigated the changes in conformational flexibility and secondary structure among family members to identify the thermal-sensitive regions in the enzymes with β-sandwich architecture. Then, the interaction energy and interaction network were explored to illustrate how the thermal-sensitive regions diversify the thermostability of GH12 enzymes. Finally, we examined and generalized the predictions from MD simulations in other GH12 enzymes using the bioinformatics method. Our study provides potential applications in designing thermostable cellulases.

2. Methods

2.1 Protein preparation

For this study, a data set containing homologous mesophilic and thermophilic cellulases from GH12 was constructed. Five cellulases were selected including the mesophilic enzymes *Hs*Cel12A (PDB: 1OA3) from *Hypocrea schweinitzii* and *Tr*Cel12A (PDB: 1H8V) from *Trichoderma reesei*, and the thermophilic enzymes *Ss*Cel12A (PDB: 1OA4) from *Streptomyces* sp. 11AG8, *Hg*Cel12A (PDB: 1OLR) from *Humicola grisea* and *Rm*Cel12A (PDB: 1H0B) from *Rhodothermus marinus*. These enzymes have experimentally determined crystal structures that show folding into a similar β-sandwich architecture. Besides, their biochemical thermostabilities have been characterized by measurement of their melting temperatures, which range from 49 to 95 ˚C. The evolutionary locations in the phylogenetic tree of GH12 and the overall cellulase structure are shown in Figure 1. Details of selected cellulases are displayed in Table 1. All protein structures were obtained from the Protein Data Bank (www.rcsb.org). The melting temperatures were reported in previous studies[30, 31].

2.2 System preparation

The denaturation of proteins typically occurs on a microsecond time scale.[32] However, it's very difficult to perform MD simulations within the time period in which a protein naturally unfolds. Therefore, accelerating the unfolding process by increasing the simulation temperature is a necessary step. Herein, three parallel simulations were set up for each cellulase with the temperature of 300, 400 and 500 K. The elevated simulation temperatures can effectively accelerate

the unfolding process, which helps to identify the initially unfolding sites of the protein. A variety of studies have proved that high simulation temperatures do not influence the native unfolding pathway of the protein, especially in the early stage.[14, 22] Moreover, this technique has been applied into many engineering works and the thermostability of many enzymes has been successfully improved by rigidifying the initially unfolding sites,[17, 33] further verifying the feasibility of this technique.

2.3 Molecular dynamics simulations

In the present study, the protein was solvated using the SPC model.[34] This model has been successfully applied into a variety of thermal unfolding studies.[22, 35, 36] A cubic box was constructed to perform MD calculations. Water molecules that overlapped with the protein heavy atoms were removed. The total numbers of atoms in different solvent systems were greater than 30,000. To produce a neutral system with 0.1 mol/L ionic concentration, appropriate amounts of Na^+ and Cl^- were added by replacing water molecules with ions randomly. All MD simulations were performed under periodic boundary conditions using the GROMACS software package.[37] CHARMM27 force field[38] was used to describe the protein. The force field parameters are thought to be effective at room temperature, but a previous study provided qualitative validation that they are also valid at highly elevated temperatures.[39] In order to eliminate steric interference, steepest energy minimization was performed for every system to give the maximum force below 1000 $KJ{\cdot}mol^{-1}{\cdot}nm^{-2}$. We also performed 200-ps position restraint MD to further equilibrate the systems. Complete equilibration was assessed by the convergence of the potential energy and the temperatures of the systems. Lastly, 50-ns production MD simulations with three replicas were performed in the NPT ensembles. V-rescale and Parrinello-Rahman methods were used to control the system temperature and pressure, respectively.[40, 41] The LINCS algorithm was used to constrain all bonds to hydrogen atoms in the protein and the SETTLE algorithm was used for the water molecules.[42, 43] The Particle Mesh Ewald (PME) method was used to evaluate the long-range electrostatic interactions.[44] The non-bonded pair list cutoff was 10.0 Å and the pair list was updated every 10 fs.

2.4 Simulation analysis

All the biophysical properties were analyzed using the internal tools in Gromacs. Names of the tools were listed in the brackets after the corresponding parameters. The hydrophobic SASA (g_sasa) and Rg (g_gyrate) were calculated to investigate the change in global structure.[45] In present study, a hydrogen bond (g_hbond) was considered to be formed if the acceptor-donor distance was less than 0.35 nm and the acceptor-hydrogen-donor angle was less than 30°.[46, 47] The residue-level contact (g_mdmat) was calculated based on the smallest distance matrices between residue pairs. A contact was defined if the minimum distance was less than 0.5 nm. Through calculating the radial distribution functions (g_rdf), the hydration was analyzed by counting the number of water molecules in the hydration layer.[48] The entropy values (g_anaeig) were calculated based on the Quasiharmonic approach.[49] After removing the overall translational and rotational motions by superimposing the Cα atoms of each snapshot structure onto the starting structure using least-square fitting, the RMSD (g_rms) and RMSF (g_rmsf) were calculated to analyze structural stability and flexibility.[50] The interaction energy within the N-terminus was calculated based on the amber force field equation.[38] After constructing an autocephalous index for the residues in the N-terminus, the Coulomb and Lennard-Jones interactions were calculated.[51, 52] The total interaction energy was the sum of these two parts. The collective motions were also investigated by principal components analysis (PCA). The covariance matrix was generated using atomic coordinates of Cα atoms. Diagonalization of the covariance matrix generated the eigenvectors, each having a corresponding eigenvalues (g_covar).[47] The trajectory was projected onto a particular eigenvector to reveal concerted motions (g_anaeig).

2.5 Statistical analysis

More enzymes from GH12 were sampled to perform the bioinformatics analysis, which helps to examine and generalize the deductions from MD simulations. In total, eight mesophilic enzymes from *Streptomyces halstedii* (CelA), *Pectobacterium carotovorum* (CelB, CelS), *Aspergillus niger* (Xeg12A, EglA), *Fomitopsis palustris* (Cel12A), *Hypocrea schweinitzii* (Cel12A) and *Trichoderma reesei* (Cel12A), and seven thermophilic enzymes from *Streptomyces* sp. 11AG8 (Cel12A), *Rhodothermus marinus* (Cel12A), *Thermotoga maritime* (Cel12A, Cel12B), *Thermotoga neapolitana* (CelA, CelB) and *Humicola grisea* (Cel12A) were selected as the data set.

Multiple sequence alignments were performed using ClustalW.[53] Alignment results were used to create an N-terminal sequence profile using WebLogo,[54] giving a graphic description of sequence conservation and evolution. The amino acid composition was calculated using MEGA5.[55] Based on the amino acid composition of the N-terminus, the groups of percentages for each kind of amino acid can be calculated for every mesophilic and thermophilic enzyme. Through two-tailed, heteroscedastic t test, the difference between the two groups of percentages was defined as *P* value. $P < 0.1$ defines the significant difference based on the fundamental principle used in a previous study.[56] Detailed information about these bioinformatics packages and the flowchart of performing the bioinformatics analysis has been described in our previous paper.[57, 58]

3. Results and Discussion

In our work, the simulations were repeated three times. Although the results are not exactly same, the main items that will be

discussed in the present article are analogous in all simulations. The RMSD values for these systems are shown in Figure S1. Nevertheless, we will discuss only the first simulations of the three replicas in the next sections.

3.1 Protein unfolding and structural stability of GH12 cellulases

To investigate the structural stability of the mesophilic (*Tr*Cel12A and *Hs*Cel12A) and thermophilic cellulases (*Ss*Cel12A, *Hg*Cel12A and *Rm*Cel12A) in GH12, three parallel simulations at 300, 400 and 500 K were performed. The root mean square deviations (RMSDs) of the Cα atoms with respect to the crystal structures were calculated to illustrate the structural differences among the enzymes. A smaller RMSD represents a more stable protein structure.[24, 35] Figure 2 shows the RMSDs for *Tr*Cel12A and *Ss*Cel12A, which remained balanced at around 0.1 nm at 300 K. Then, the RMSD values slightly increased at 400 K, but the increase was larger in *Tr*Cel12A than in *Ss*Cel12A. This phenomenon was also appeared in the other two replicas of simulations, suggesting its reproducibility. The different RMSD patterns between *Tr*Cel12A and *Ss*Cel12A indicate that the simulation temperature of 400 K exerts different levels of influences on the protein unfolding of these two enzymes due to their different thermal stabilities. Finally, the RMSD values of both *Tr*Cel12A and *Ss*Cel12A sharply increased at 500 K and exhibited maximum deviations of 0.8 and 0.4 nm, respectively, which indicated that their structures may undergo a constant evolution at high temperatures. As expected, the thermophilic *Ss*Cel12A displayed lower RMSD values than *Tr*Cel12A, which is consistent with their different thermostabilities.[30] The RMSD values of the other three cellulases (*Hs*Cel12A, *Hg*Cel12A and *Rm*Cel12A) revealed similar results, shown in Figure S1.

Several biophysical properties were calculated to characterize the changes of the enzyme during the course of unfolding, including hydrophobic solvent accessible surface area (SASA), radius of gyration (Rg), entropy values and the number of hydration water. As shown in Table 2, the hydrophobic SASA and Rg values of both mesophilic and thermophilic enzymes increased as the temperature increased from 300 to 500 K, suggesting that exposure of the hydrophobic regions and expansion of the proteins occurred during the course of unfolding. Moreover, the increase of entropy values revealed that the enzymes become more disordered at a higher temperature. The number of hydration water also increased as the temperature went up, indicating that when the protein unfolded, the hydration around it was also enhanced.[11] Importantly, it should be noted that the mesophilic cellulases (*Tr*Cel12A and *Hs*Cel12A) always exhibited a larger increase in hydrophobic SASA and Rg values compared to the thermophilic cellulases (*Ss*Cel12A, *Hg*Cel12A and *Rm*Cel12A). Such differences indicated that the thermophilic cellulases might unfold more slowly than the mesophilic enzymes, which has subsequently been proved in the analysis of their secondary structures.

Additionally, the average numbers of intramolecular hydrogen bonds within the whole enzymes were calculated to characterize the transitions in intramolecular interactions at different temperatures. As presented in Table 2, the number of hydrogen bonds decreased in all cellulases as the temperature increased. However, it is worth noting that the reduction in hydrogen bonds was more dramatic in the mesophilic enzymes. Besides, the mesophilic enzymes always possessed fewer hydrogen bonds than the thermophilic enzymes at 300, 400 and 500 K. The smallest distance matrices between residue pairs were calculated and a contact was defined if the distance is below 0.5 nm, then we investigated the change of the internal contact, shown in Figure S2. Similarly, the number of the internal contacts gradually diminished with the temperature increasing. These results revealed the role of intramolecular interactions (including hydrogen bond and contact) in stabilizing the protein structure. In addition, we investigated the correlation between the T_m of enzymes and the data listed in Table 2. Only the number of intramolecular hydrogen bonds exhibits positive correlation with T_m. However, the variations (increase or decrease) of the properties in Table 2 with the temperature increase are always larger in mesophilic enzymes, which is consistent with their different thermostability. On the basis of the above analysis, our results verified the previous deduction that the initial unfolding of the proteins featured protein expansion and the disruption of intramolecular interactions[22, 23]. Moreover, the differences in thermostability between the mesophilic and thermophilic enzymes can be definitely observed in our simulations, which confirmed the reliability and reproducibility of the simulations and laid a foundation for the following analysis.

3.2 Identification of thermal-sensitive regions of GH12 cellulases

The variable root mean square fluctuation (RMSF) is often used to evaluate the conformational flexibility of proteins. Larger values of RMSF manifest as higher conformational flexibility.[58] So, increases in RMSF can be used to represent the response of the structural conformation to environmental changes. In order to identify the thermal-sensitive regions of GH12 cellulases, we calculated the RMSF changes of these enzymes for the simulations at 300 and 400 K. We selected these two systems because they defined the native state and the initially unfolding state, respectively, based on the analysis of structural stability (Figure 2 and Table 2). As shown in Figure 3, the gradient colors from blue to red depict the increasing extent of RMSF. Although the RMSF of most protein regions rose a little, the RMSF of several regions increased significantly with increasing temperature. In particular, two regions (loop 1 and loop 3 in the N-terminus as labeled in Figure 1) always had the largest RMSF increases in all five cellulases studied, implying that the N-terminus may be the general thermal-sensitive region. The RMSF increases for loop 3 were also found to be slightly smaller in thermophilic enzymes than in mesophilic enzymes, revealing the different thermostability of

their N-termini. Apart from the N-terminus, it should be noted that some other thermal-sensitive regions also existed in each cellulase, but these only displayed moderate RMSF increases. These regions tended to be individual sites because their locations in the protein structures were variable.

Next, we investigated the time evolution of the secondary structure in the simulations at 500 K, shown in Figure 4a. This gives more information about the structural flexibility of GH12 cellulases. As expected, the N-termini of the mesophilic enzymes (*Tr*Cel12A and *Hs*Cel12A) gradually lost their native secondary structure in the simulations, which can be attributed to violent fluctuations of the thermal-sensitive loops (L1 and L3) at high temperatures. In contrast, the N-termini of the thermophilic enzymes (*Ss*Cel12A, *Hg*Cel12A and *Rm*Cel12A) unfolded more slowly. The discrepancies in the N-terminal unfolding behavior between the mesophilic and thermophilic enzymes suggested that the N-termini of the thermophilic enzymes were more stable than those of the mesophilic enzymes. Furthermore, in order to investigate the relationship between protein dynamics and denaturation, we performed principal component analysis (the eigenvalues of corresponding eigenvectors were shown in Figure S3) and visualized the collective motions of enzymes along the first mode based on the simulations at 300 and 400 K. As shown in Figure 4b, the dynamics of the enzyme significantly changed in the N-terminus with the temperature increasing. At 400 K, the motions of the N-terminus tended to be disordered, while other regions of the enzyme still remained in their native state. These results further demonstrated the instability of the N-terminus and revealed the link between protein dynamics change and the N-terminus denaturation.

Protein stability is determined by the free energy difference between the native state and denatured state, which is influenced by intramolecular interactions and conformational flexibility.[9, 10] Thus, protein stability is thought to be linked to the whole protein structure. However, based on our results, it was interesting to find that the conformational stability seemed not to be equally distributed over the whole protein structure as some local structures were more sensitive to high temperature. These thermal-sensitive regions unfolded more easily at high temperature and thus should be considered the weak points of the protein structure. The N-termini were always representatives of these thermal-sensitive regions in the studied cellulases. Interestingly, a previous study performed by Sengupta et al. demonstrated that the structural unfolding of protein was correlated with surface hydration. The hydration and protein unfolding do not occur uniformly over the whole protein surface, but are sensitive to local structural propensity.[15] Based on this, the non-uniform unfolding in the N-terminus of GH12 enzymes might be inherently related to the non-uniform surface hydration, and further research is required to examine this hypothesis.

In a previous computational study of GH11 xylanases that fold into a β-sandwich architecture that is similar to the architecture of the GH12 cellulases, the N-terminus was also found to be a thermal-sensitive region.[59] Thus, the thermal sensitivity of the N-terminus might be a general characteristic of all members of the GH-C family (including GH11 and GH12 families) that display β-sandwich architecture. This conclusion can be supported by the fact that the β-sandwich architecture is mainly stabilized by the interactions between the adjacent β-strands; however, the β-strands A1, B1 and B2 in the N-terminus are significantly shorter than others in the enzymes. Previous studies focusing on the relationship between protein flexibility and stability have shown that the thermal-sensitive regions are usually the initial unfolding points of protein structure.[60] Therefore, the N-terminus seems to be a generally influential factor for enzyme thermostability in the GH-C family. Besides, in previous studies of GH-C enzymes, mutations in the N-terminal region could significantly enhance enzyme thermostability,[16, 61, 62] suggesting that the N-terminus may serve a determinant function in protein stability and is a generally available engineering target. On the other hand, it should be noted that the cellulases studied possess some individual thermal-sensitive sites in their structures, which are also potential engineering targets.[17] However, their locations in the protein structures are not coincident in different enzymes, and this is a possible reason why the engineering targets of rational design are so diverse in different proteins. In addition, the effects of the individual thermal-sensitive regions on enzyme thermostability were expected to be weaker than the effects of the N-terminus in GH-C family based on the fact that the RMSF increases in the individual sites were evidently lower than in the N-terminal region (Figure 3). Taken together, the above analysis demonstrates that identifying the general structure involved in enzyme thermostability is a valuable way for researchers to rapidly identify useful targets, which helps to improve engineering efficiency.

3.3 Interactions within the N-terminus of GH12 cellulases

Because of the importance of the N-terminus for enzyme thermostability in GH12 family, the following work mainly focused on this region. In order to investigate the relationship between the N-terminus and enzyme thermostability in GH12, we calculated the total interaction energies (residue–residue interaction energies) within the N-terminus for the cellulases and plotted the energy values as a function of their melting temperatures. It was easy to show (Figure 5) that there is a substantially positive correlation between the interaction energy values and the enzyme thermostabilities with a coefficient of determination (R^2) of 0.94. Considering that internal interactions typically contribute to the stability of the protein conformation and structure,[63, 64] these calculations indicated that the N-termini of thermophilic enzymes are more stable than those of mesophilic ones. A stable N-terminus could postpone the initial time of unfolding and thus improve the thermostability of the GH12 cellulases. Correlation analysis indicated that the stabilities of the N-terminus and the whole enzyme structure are coupled in the course of molecular evolution. In GH12, even in the GH-C family, a more stable N-terminus may be part of a more stable enzyme. Importantly,

this correlation information could not be derived by performing an MD study for a single protein.

In order to clarify why the N-termini of thermophilic enzymes are more stable, we investigated the interaction network within the N-terminus (including A1, B1, B2, A2 β-strands, and their linking loops) using Protein Interactions Calculator.[65] As shown in Figure 6a, it was interesting to find that the internal interactions within the N-terminus are organized in a special amphipathic pattern: a hydrophobic packing cluster consisting of hydrophobic residues padding the space between the A2 and B2 β-strands and a hydrogen bonding cluster composed of polar residues connecting the L1 and L3 loops together. Intricate interactions, just like adhesion-agent, integrate the structural elements in the N-terminus into a whole. Our previous experimental study of *Tr*Cel12A demonstrated that several residues (Trp-7, Asn-20 and Trp-22) located in the hydrogen bonding area participate in substrate binding.[66] Thus, we speculated that the hydrogen bonding network stabilizing the L1 and L3 loops may produce a beneficial effect on not only enzyme thermostability, but also the substrate binding process. Importantly, this special amphipathic pattern of internal interactions appears in all studied cellulases (shown in Figure S4), but was not found in GH11 xylanases, suggesting that this amphipathic pattern may be a unique stabilizing foundation of the GH12 family.

In order to further characterize the roles of the amphipathic pattern in the stability of the N-terminus, the number of contacts in the hydrophobic packing cluster and the number of hydrogen bonds in the hydrogen bonding cluster were determined. As shown in Figures 6b and 6c, the numbers of contacts and hydrogen bonds in the thermophilic enzymes were greater than those in the mesophilic enzymes. These data suggest that the internal interaction network plays an important role in enzyme thermostability than the hydrogen bonding cluster. Notably, the number of hydrophobic contacts changed a little in five cellulases as the temperature increased from 300 to 400 K, but the number of hydrogen bonds significantly decreased at 400 K. This indicates that at 400 K, the hydrophobic packing network of the N-terminus still remained relatively integral, while the hydrogen bonding network has been broken up.

Based on these results, it became clear that the presence of amphipathic interactions within the N-terminus was necessary for the enzyme thermostability of GH12 cellulases. The enhanced stability of thermophilic enzymes can be attributed to their more intricate interaction network within the N-terminus. Thus, mutagenesis with the aim of improving the thermostability of the GH12 enzymes should be focused on the generally thermal-sensitive N-terminus. Interestingly, an experimental study of *Tr*Cel12A showed that, in the A35V variant, stability had increased by 7.7 ˚C but, in the A35S variant, stability had decreased by 4.0 ˚C.[30] One interpretation of such a discrepancy is that Ala-35 is precisely located in the hydrophobic packing cluster, and the introduction of more hydrophobic residues (V) is more stabilizing than polar residues (S). Hence, the amphipathic pattern of the interactions needs to be taken into consideration before designing mutagenesis in the N-terminus.

3.4 Analysis of the amino acid sequence of the N-terminus

MD simulations could provide more details to enhance our understanding of how the N-terminus dynamics and interactions potentiate enzyme thermostability in the GH 12 family. However, it seems infeasible to investigate all proteins in this family with the limitations of computing resources. In order to compensate for this drawback and to examine the findings from MD simulations, we analyzed the amino acid sequences of the N-termini of other GH12 enzymes using a bioinformatics method. The data set consisted of eight mesophilic and seven thermophilic homologous enzymes (listed in section 2.4). The sequence profiles of the N-termini are displayed in Figure 7a. According to the amphipathic patterns in the N-termini, the sequences were divided into hydrogen bonding areas and hydrophobic packing areas. Sequence positions 10, 15, 17, 33 and 35 (with red boxes) are mainly occupied by hydrophobic residues. In particular, it is these sites that participate in the hydrophobic packing cluster. Similarly, the sequence sites (with blue boxes) engaged in the hydrogen bonding cluster primarily consist of polar residues. Our previous study have revealed that the conservation of amino acids is connected with their molecular function;[57] however, in the present study, surprisingly we found that the amino acids in the key positions of the amphipathic pattern remained relatively conserved in property but diverse in species. This interesting phenomenon suggests that the conservation of the amphipathic pattern in the GH12 family might be more crucial for N-terminus stability.

On the other hand, sequence variability in the key positions, in turn, might reflect the special adaptions of the enzymes to different temperature conditions. In order to investigate this conjecture, the amino acid compositions of the N-terminus were investigated. As shown in Figure 7b, the amino acid compositions exhibited different characteristics in mesophilic enzymes in comparison to thermophilic ones. The percentages of some amino acids were significantly different in mesophilic and thermophilic enzymes, according to the criterion of a *P* value below 0.1.[56] For clarity, the relative percentages of amino acids are displayed in Figure 7c to demonstrate the preferences for amino acids in the mesophilic and thermophilic enzymes. It can be seen that there were more charged (E and R) and highly hydrophobic amino acids (L, I, M and V) in the thermophilic enzymes. However, based on the hydrophobicity coefficients of amino acids,[67] the amino acids preferred by mesophilic enzymes featured weak polarity and hydrophobicity. The higher numbers of charged and highly hydrophobic residues in the thermophilic enzymes may contribute to the formation of stronger electrostatic and hydrophobic interactions in their N-termini. A previous study also emphasized the importance of charged amino acids, but neglected the roles of hydrophobic amino acids in protein stability.[56] Taken together, these results demonstrate that the evolution of the amino acids in the N-termini was an inherent

mechanism of diversifying enzyme thermostability across the GH12 family.

5. Conclusions

In this study, we performed MD and bioinformatics studies to investigate the protein thermostability in GH12 family. The differences in overall stability and unfolding dynamics between the mesophilic and thermophilic enzymes in the GH12 family were characterized. Furthermore, the N-terminus was identified as the general thermal-sensitive region in the GH12 family that tends to lose its native secondary structure during the simulations. Together with the results of a previous study, this conclusion may also be valid for all enzymes with β-sandwich architecture in the GH-C family. The interaction analysis revealed that the stability of the N-terminus and that of the whole enzyme are appreciably correlated and, besides, the amphipathic interaction pattern is the structural foundation of N-terminal stability. Finally, bioinformatics analysis revealed that the amphipathic patterns are highly conserved in GH12 enzymes. The amino acid preference in the N-terminus is inherent mechanism underlying enzyme thermostability. In short, our results demonstrate that the structural and dynamics evolution of the amphipathic N-terminus is the general mechanism that diversifies enzyme thermostability in the GH12 family. The protocol used in this study would also be applicable to other protein families to determine their special stabilizing mechanisms.

Acknowledgements

We acknowledge the Supercomputer Centre in Shandong University at Weihai, for providing computational facilities. We acknowledge the suggestions from Dr. James Tipper and the anonymous reviewers, the help from Shijia Liu for revising the manuscript. This work was financially supported by a grant from The National Natural Science Foundation of China (31370111/31170071) and The Key Technologies R&D Program of Shandong Province (2015GSF121019).

Notes and references

1. R. F. Service, *Science*, 2013, **339**, 1374-1379.
2. R. F. Service, *Science*, 2014, **345**, 1111-1111.
3. R. M. Wahlstroem and A. Suurnaekki, *Green Chem.*, 2015, **46**, 694-714.
4. J. A. Rollin, Z. Zhu, N. Sathitsuksanoh and Y. H. Zhang, *Biotechnol. Bioeng.*, 2011, **108**, 22-30.
5. Y. S. Cheng, T. P. Ko, J. W. Huang, T. H. Wu, C. Y. Lin, W. Luo, Q. Li, Y. Ma, C. H. Huang, A. H. Wang, J. R. Liu and R. T. Guo, *Appl. Microbio. Biotechnol.*, 2012, **95**, 661-669.
6. D. L. Trudeau, T. M. Lee and F. H. Arnold, *Biotechnol. Bioeng.*, 2014, **111**, 2390-2397.
7. A. S. Bommarius, M. Sohn, Y. Kang, J. H. Lee and M. J. Realff, *Curr. Opin. Biotechnol.*, 2014, **29**, 139-145.
8. I. Wu and F. H. Arnold, *Biotechnol. Bioeng.*, 2013, **110**, 1874-1883.
9. T. J. Kamerzell and M. C. Russell, *J. Pharm. Sci.*, 2008, **97**, 3494-3517.
10. K. Teilum, J. G. Olsen and B. B. Kragelund, *Biochim. Biophys. Acta*, 2011, **1814**, 969-976.
11. P. Chatterjee and N. Sengupta, *Mol. BioSyst.*, 2016, **12**, 1139-1150.
12. J. M. Vinther, S. M. Kristensen and J. J. Led, *J. Am. Chem. Soc.*, 2010, **133**, 271-278.
13. Y. Li, B. Shan and D. P. Raleigh, *J. Mol. Biol.*, 2007, **368**, 256-262.
14. V. Daggett and M. Levitt, *J. Mol. Biol.*, 1993, **232**, 600-619.
15. P. Chatterjee, S. Bagchi and N. Sengupta, *J. Chem. Phys.*, 2014, **141**, 205103.
16. Y. Wang, Z. Fu, H. Huang, H. Zhang, B. Yao, H. Xiong and O. Turunen, *Bioresour. Technol.*, 2012, **112**, 275-279.
17. S. Zhang, Y. Wang, X. Song, J. Hong, Y. Zhang and L. Yao, *J. Chem. Inf. Model.*, 2014, **54**, 2826-2833.
18. G. I. Makhatadze, V. V. Loladze, D. N. Ermolenko, X. Chen and S. T. Thomas, *J. Mol. Biol.*, 2003, **327**, 1135-1148.
19. D. Z. Ayadi, A. H. Sayari, H. B. Hlima, S. B. Mabrouk, M. Mezghani and S. Bejar, *Int. J. Biol. Macromol.*, 2015, **72**, 163-170.
20. X. Duan, J. Chen and J. Wu, *Appl. Environ. Microbiol.*, 2013, **79**, 4072-4077.
21. J. Georis, F. D. L. Esteves, J. L. Brasseur, V. Bougnet, B. Devreese, F. Giannotta, B. Granier and J. M. Frere, *Protein Sci.*, 2000, **9**, 466-475.
22. L. Chen, X. Li, R. Wang, F. Fang, W. Yang and W. Kan, *J. Biomol. Struct. Dyn.*, 2015, **34**, 1-40.
23. K. Kumar, K. Patel, D. C. Agrawal and J. M. Khire, *J. Mol. Model.*, 2015, **21**, 163.
24. H. Zhang, J. Li, J. Wang, Y. Yang and M. Wu, *Biotechnol. Biofuels*, 2014, **7**, 112-120.
25. U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore and K. Robins, *Nature*, 2012, **485**, 185-194.
26. V. Lombard, H. G. Ramulu, E. Drula, P. M. Coutinho and B. Henrissat, *Nucleic Acids Res.*, 2014, **42**, D490-D495.
27. M. Sandgren, J. Stahlberg and C. Mitchinson, *Prog. Biophys. Mol. Biol.*, 2005, **89**, 246-291.
28. Karlsson J, Siika-aho M and T. M, *J. Biotechnol.*, 2002, **99**, 63-78.
29. A. L. Jonsson, K. A. Scott and V. Daggett, *Biophys. J.*, 2009, **97**, 2958-2966.
30. M. Sandgren, P. J. Gualfetti, A. Shaw, L. S. Gross, M. Saldajeno, A. G. Day, T. A. Jones and C. Mitchinson, *Protein Sci.*, 2003, **12**, 848-860.
31. D. Kapoor, V. Kumar, S. Chandrayan, S. Ahmed, S. Sharma, M. Datt, B. Singh, S. Karthikeyan and P. Guptasarma, *Biochim. Biophys. Acta*, 2008, **1784**, 1771-1776.
32. Y. Duan, L. Wang and P. A. Kollman, *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 9897-9902.
33. S. Badieyan, D. R. Bevan and C. Zhang, *Biotechnol. Bioeng.*, 2012, **109**, 31-44.
34. J. Zielkiewicz, *J. Chem. Phys.*, 2005, **123**, 104501.
35. M. Paul, M. Hazra, A. Barman and S. Hazra, *J. Biomol. Struct. Dyn.*, 2014, **32**, 928-949.

36. S. Kundu and D. Roy, *J. Mol. Graph. Model.*, 2008, **27**, 88-94.
37. S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845-854.
38. W. D. Cornell, P. Cieplak and C. I. Bayly, *J. Am. Chem. Soc.*, 1995, **117**, 5179-5197.
39. D. E. Shaw, P. Maragakis and K. Lindorff-Larsen, *Science*, 2010, **330**, 341-346.
40. G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.
41. M. Parrinello and A. Rahman, *J. Appl. Phys*, 1981, **52**, 7182-7190.
42. B. Hess, H. Bekker and H. J. C. Berendsen, *J. Comput. Chem.*, 1997, **18**, 1463-1472.
43. S. Miyamoto and P. A. Kollman, *J. Comput. Chem.*, 1992, **13**, 952-962.
44. M. J. Abraham and J. E. Gready, *J. Comput. Chem.*, 2011, **32**, 2031-2040.
45. F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander and M. Scharf, *J. Comput. Chem.*, 1995, **16**, 273-284.
46.T. Fu, X. Wu, Z. Xiu, J. Wang, L. Yin and G. Li, *J. Theor. Comput. Chem.*, 2013, **12**, 1491-1499.
47. M. Liu, L. Wang, X. Sun and X. Zhao, *Sci. Rep.*, 2014, **4**, 5095.
48. P. Rani and P. Biswas, *J. Phys.: Condens. Matter*, 2014, **26**, 335102.
49. O. L. Rojas, R. M. Levy and A. Szabo, *J. Chem. Phys.*, 1986, **85**, 1037-1043.
50.V. N. Maiorov and G. M. Crippen, *Proteins*, 1995, **22**, 273-283.
51. B. A. Luty, I. G. Tironi and W. F. V. Gunsteren, *J. Chem. Phys.*, 1995, **103**, 3014-3021.
52. M. Zacharias, T. P. Straatsma and J. A. Mccammon, *J. Chem. Phys.*, 1994, **100**, 9025-9031.
53. M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins, *Bioinformatics*, 2007, **23**, 2947-2948.
54. G. E. Crooks, G. Hon, J. M. Chandonia and S. E. Brenner, *Genome Res.*, 2004, **14**, 1188-1190.
55. K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar, *Mol. Boil. Evol.*, 2011, **28**, 2731-2739.
56. A. Szilágyi and P. Závodszky, *Structure*, 2000, **8**, 493-504.
57. L. Tian, S. Liu, S. Wang and L. Wang, *Sci. Rep.*, 2016, **6**, 23605.
58. S. Liu, S. Shao, L, Li, Z. Cheng, L. Tian, P. Gao and L. Wang, *Carbohydr. Res.*, 2015, **418**, 50-56
59. L. Bu, M. F. Crowley, M. E. Himmel and G. T. Beckham, *J. Biol. Chem.*, 2013, **288**, 12175-12186.
60. M. Purmonen, J. Valjakka, K. Takkinen, T. Laitinen and J. Rouvinen, *Protein Eng. Des. Sel.*, 2007, **20**, 551-559.
61. A. Karshikoff, L. Nilsson and R. Ladenstein, *FEBS J.*, 2015, **282**, 3899-3917.
62. S. Zhang, K. Zhang, X. Chen, X. Chu, F. Sun and Z. Dong, *Biochem. Biophys. Res. Commun.*, 2010, **395**, 200-206.
63. C. Dumon, A. Varvak, M. A. Wall, J. E. Flint, R. J. Lewis, J. H. Lakey, C. Morland, P. Luginbuhl, S. Healey, T. Todaro, G. DeSantis, M. Sun, L. Parra-Gessert, X. Tan, D. P. Weiner and H. J. Gilbert, *J. Biol. Chem.*, 2008, **283**, 22557-22564.
64. V. Potapov, M. Cohen and G. Schreiber, *Protein Eng. Des. Sel.*, 2009, **22**, 553-560.
65. Q. A. Le, J. C. Joo, Y. J. Yoo and Y. H. Kim, *Biotechnol. Bioeng.*, 2012, **109**, 867-876.
66. K. G. Tina, R. Bhadra and N. Srinivasan, *Nucleic Acids Res.*, 2007, **35**, W473-476.
67. X. Zhang, S. Wang, X. Wu, S. Liu, D. Li, H. Xu, P. Gao, G. Chen and L. Wang, *Sci. Rep.*, 2015, **5**, 18357.
68. M. C. J. Wilce, M. I. Aguilar and M. T. W. Hearn, *Anal. Chem.*, 1995, **67**, 1210-1219.

Tables and Figures

Table 1. Characteristics of the studied GH12 cellulases.

Enzyme	Organism	PDB ID	GenBank	T_m (℃)
Mesophilic				
*Tr*Cel12A	*Trichoderma reesei*	1H8V	AAE59774.1	54.4
*Hs*Cel12A	*Hypocrea schweinitzii*	1OA3	AAM77711.1	49.2
Thermophilic				
*Ss*Cel12A	*Streptomyces sp. 11AG8*	1OA4	AAF91283.1	65.7
*Hg*Cel12A	*Humicola grisea*	1OLR	AAM77714.2	68.7
*Rm*Cel12A	*Rhodothermus marinus*	1H0B	AAB65594.1	95

Table 2. Average values of Rg, SASA and intramolecular H-bond numbers of GH12 cellulases in different temperature simulations.

Factors[a]	300 K	400 K	500 K	Increase/Decrease[b]
***Tr*Cel12A**				
SASA	37.35	39.46	59.53	22.18
Rg	1.64	1.67	1.91	0.27
H-bond	155	141	95	60
Entropy	15.5	16.2	21.9	6.4
Hydration water	1654	1706	1764	110
***Hs*Cel12A**				
SASA	36.62	39.70	60.72	24.1
Rg	1.65	1.68	2.04	0.39
H-bond	153	141	92	61
Entropy	15.4	16.7	21.9	6.5
Hydration water	1650	1708	1775	125
***Ss*Cel12A**				
SASA	36.05	38.16	51.43	15.38
Rg	1.62	1.64	1.78	0.16
H-bond	160	150	115	45
Entropy	15.1	17.4	20.0	4.9
Hydration water	1622	1659	1699	77
***Hg*Cel12A**				
SASA	42.13	42.68	59.16	17.03
Rg	1.71	1.73	1.95	0.24
H-bond	164	167	120	44
Entropy	15.9	16.8	19.5	3.6
Hydration water	1670	1697	1740	70
***Rm*Cel12A**				
SASA	34.27	37.13	50.05	15.78
Rg	1.67	1.68	1.81	0.14
H-bond	171	156	127	44
Entropy	16.2	17.6	20.0	3.8
Hydration water	1647	1665	1693	46

[a]Values of SASA, Rg and Entropy are given in nm^2, nm and $KJ/mol \cdot K$, respectively.
[b]The increases/decreases in values are defined by the data difference between 300K and 500K.

ARTICLE

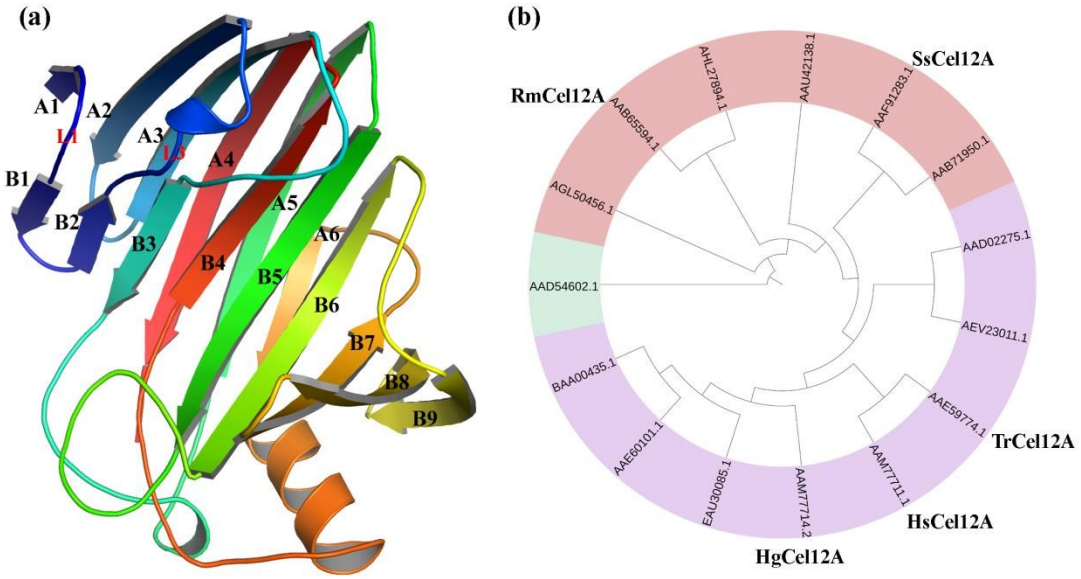

Figure 1. (a) Schematic β-sandwich architecture of the GH12 cellulases, color-ramped based on the residue number, starting with blue at the N-terminus and ending with red at the C-terminus. (b) The phylogenetic tree of the GH12 family, including all enzymes with known structures. Archaea, bacteria and Eukaryota are highlighted by *green*, *orange* and *purple*, respectively. The evolutionary locations of five selected cellulases are marked.

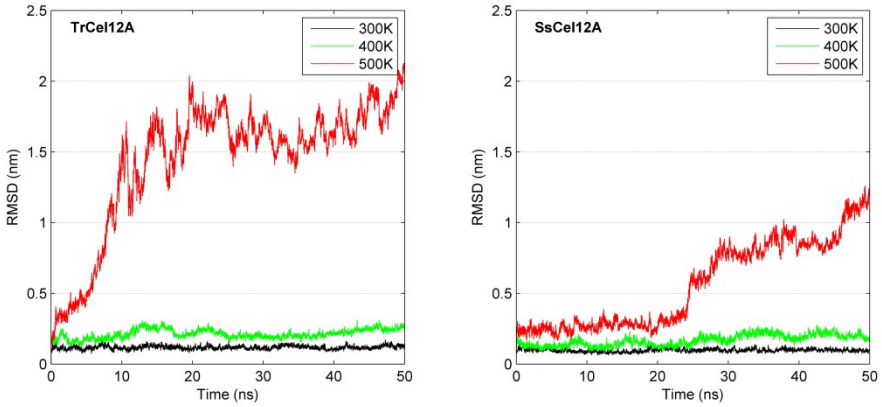

Figure 2. Time evolution of the backbone RMSD for *Tr*Cel12A and *Ss*Cel12A. The calculated data obtained from simulations at 300, 400 and 500 K are shown in *black*, *green* and *red*, respectively.

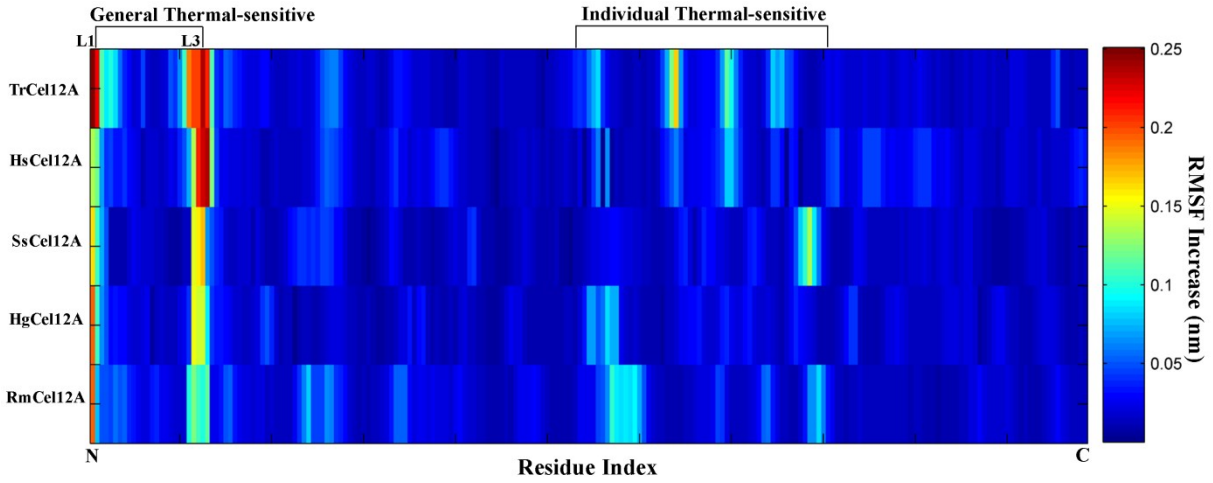

Figure 3. Increases in RMSF between the simulations at 300 and 400 K, showing the conformational flexibilities of the GH12 cellulases. Color-coding corresponds to increasing values, with *blue* being the lowest conformational flexibility and *red* being the highest conformational flexibility. The map was constructed using MATLAB software.

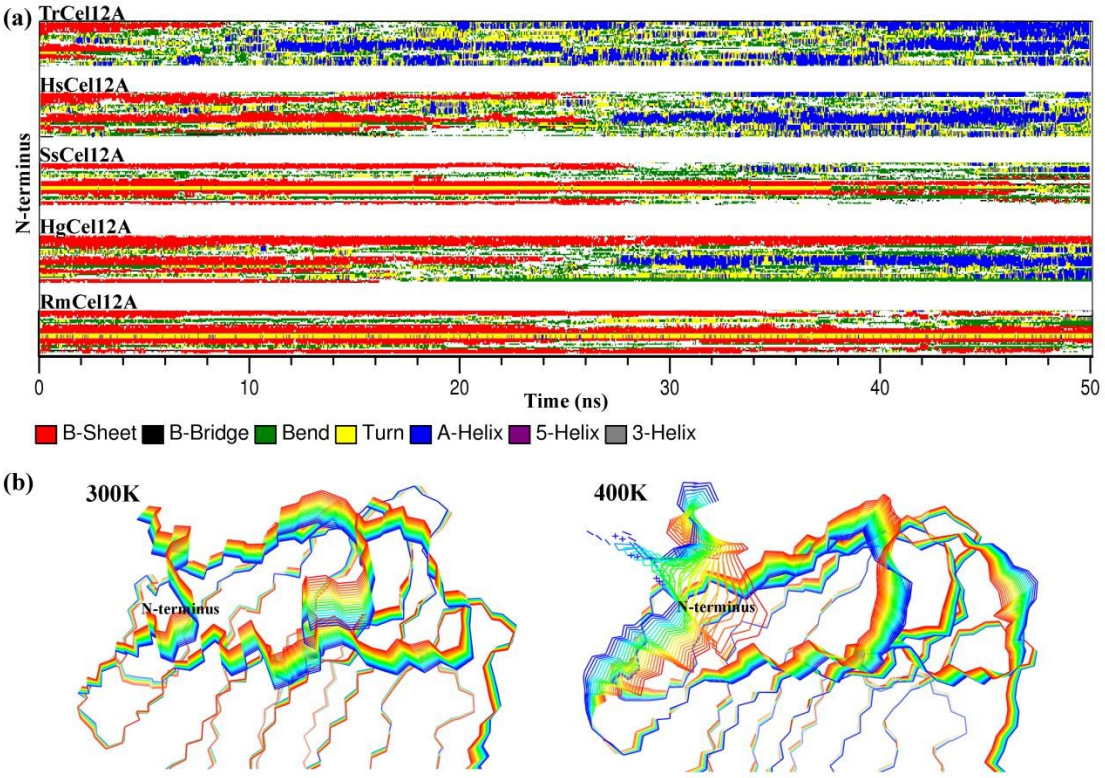

Figure 4. (a) Time evolution of the secondary structure elements in N-terminus of the GH12 cellulases in simulations at 500 K. (b) The collective motions of *Tr*Cel12A along the first mode based on the simulations at 300 and 400 K.

ARTICLE

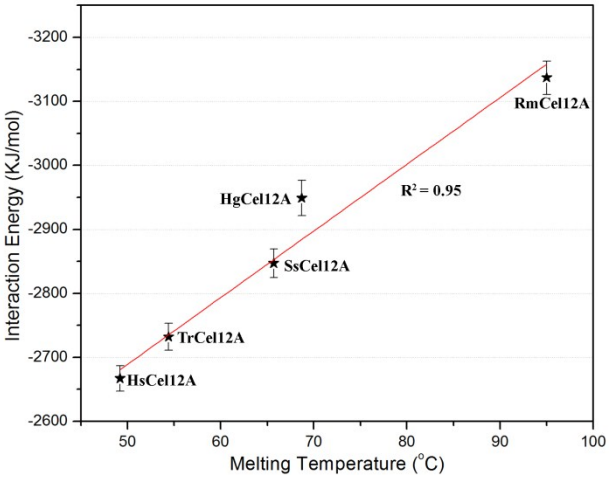

Figure 5. Averaged internal interaction energies within the N-termini were plotted against the melting temperatures of the GH12 cellulases. The coefficient of determination is shown beside the trend line (R^2 = 0.95).

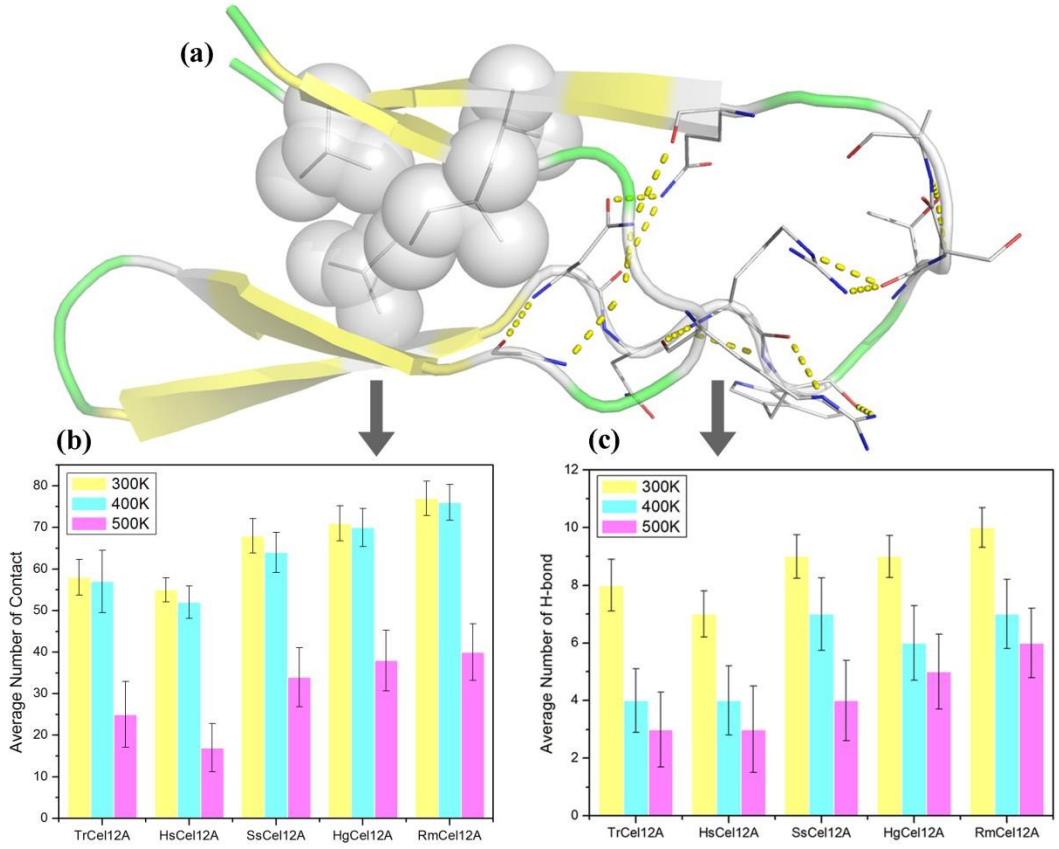

Figure 6. (a) Amphipathic pattern of internal interactions within the N-terminus. The residues that participate in the hydrophobic packing core and the hydrogen bonding network are shown as spheres and lines models, respectively. (b) The average number of hydrophobic contacts in the GH12 cellulases at 300, 400 and 500 K. (c) The average number of hydrogen bonds in the GH12 cellulases at 300, 400 and 500 K.

ARTICLE

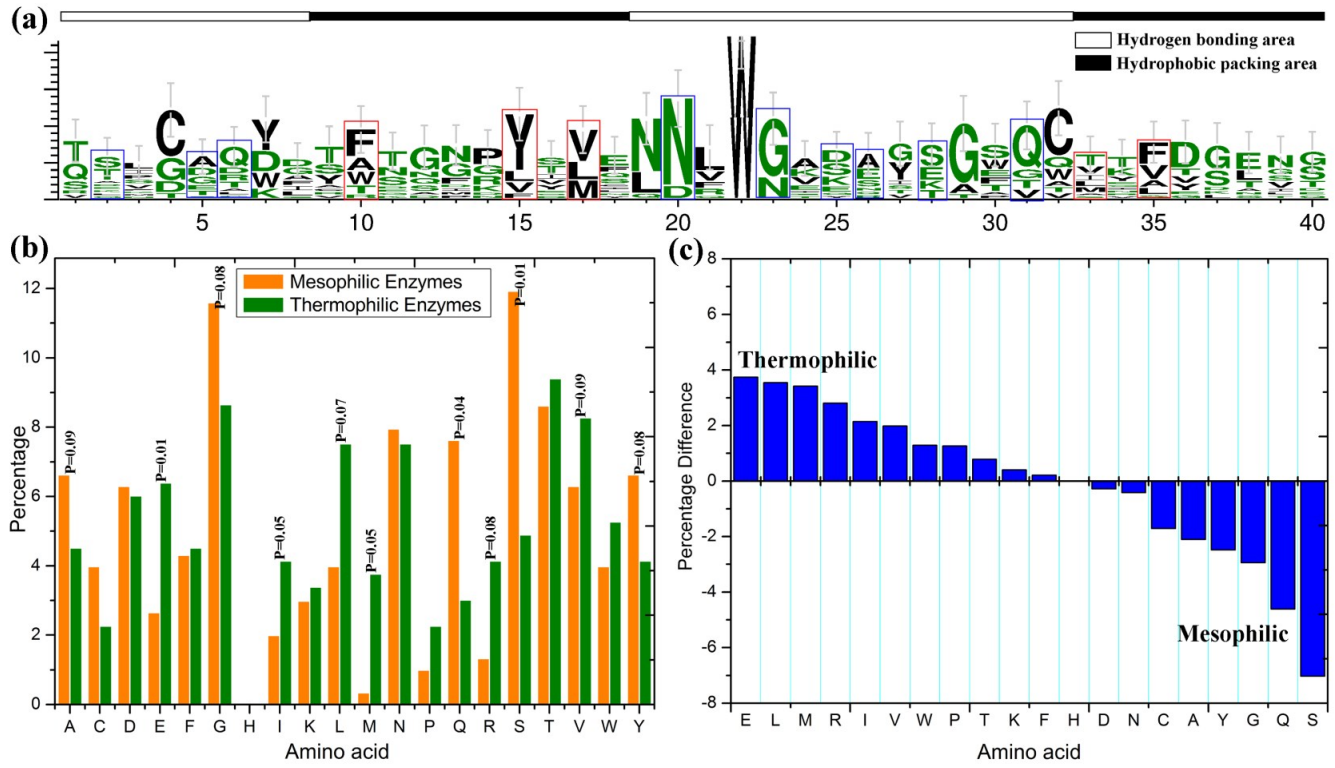

Figure 7. (a) Sequence profiles of the N-termini in the GH12 family. Hydrophilic and hydrophobic residues are shown in *green* and *black*, respectively. The residues sites participating in hydrophobic packing core and hydrogen bonding network were highlighted by *red* and *blue* boxes, respectively. (b) Comparison of the amino acid compositions in the N-termini of mesophilic and thermophilic enzymes in our dataset. P values were calculated using a two-tailed, heteroscedastic *t* test for each amino acid. Significant differences were defined as $P < 0.1$. (c) The preferences of amino acids: positive values indicate a thermophilic preference, while negative values indicate a mesophilic preference.