FEBS Open Bio 4 (2014) 936–946

journal homepage: www.elsevier.com/locate/febsopenbio

# Structure and stability of metagenome-derived glycoside hydrolase family 12 cellulase (LC-CelA) a homolog of Cel12A from *Rhodothermus marinus* ☆

Hiroyuki Okano [a], Masashi Ozaki [a], Eiko Kanaya [a], Joong-Jae Kim [a,1], Clement Angkawidjaja [a,b], Yuichi Koga [a], Shigenori Kanaya [a,*]

[a] Department of Material and Life Science, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
[b] International College, Osaka University, 1-30 Machikaneyama-cho, Toyonaka, Osaka 560-0043, Japan

ARTICLE INFO

*Article history:*
Received 17 October 2014
Revised 27 October 2014
Accepted 27 October 2014

Enzyme: Endoglucanase (EC 3.2.1.4)

*Keywords:*
Leaf–branch compost
Metagenome
Glycoside hydrolase family 12 cellulase
Flexible linker
Stability
Crystal structure

ABSTRACT

**Ten genes encoding novel cellulases with putative signal peptides at the N-terminus, termed pre-LC-CelA–J, were isolated from a fosmid library of a leaf–branch compost metagenome by functional screening using agar plates containing carboxymethyl cellulose and trypan blue. All the cellulases except pre-LC-CelG have a 14–29 residue long flexible linker (FL) between the signal peptide and the catalytic domain. LC-CelA without a signal peptide (residues 20–261), which shows 76% amino acid sequence identity to Cel12A from *Rhodothermus marinus* (*Rm*Cel12A), was overproduced in *Escherichia coli*, purified and characterized. LC-CelA exhibited its highest activity across a broad pH range (pH 5–9) and at 90 °C, indicating that LC-CelA is a highly thermostable cellulase, like *Rm*Cel12A. The crystal structure of LC-CelA was determined at 1.85 Å resolution and is nearly identical to that of *Rm*Cel12A determined in a form without the FL. Both proteins contain two disulfide bonds. LC-CelA has a 16-residue FL (residues 20–35), most of which is not visible in the electron density map, probably due to structural disorder. However, Glu34 and Pro35 form hydrogen bonds with the central region of the protein. ΔFL-LC-CelA (residues 36–261) and E34A-LC-CelA with a single Glu34 → Ala mutation were therefore constructed and characterized. ΔFL-LC-CelA and E34A-LC-CelA had lower melting temperatures ($T_m$) than LC-CelA by 14.7 and 12.0 °C respectively. The $T_m$ of LC-CelA was also decreased by 28.0 °C in the presence of dithiothreitol. These results suggest that Glu34-mediated hydrogen bonds and the two disulfide bonds contribute to the stabilization of LC-CelA.**



## 1. Introduction

A metagenomic approach is an efficient method to isolate novel enzymes useful for industrial purposes as well as for academic research [1–4]. In EXPO Park, Japan, leaves and branches cut from the trees are collected periodically, mixed with urea, and agitated for composting. The temperature increases up to ~70 °C inside this compost [leaf–branch compost (LC)] and then decreases to ~50 °C roughly one year later upon completion of composting. This compost is expected to be rich in various thermophilic microorganisms and is therefore a promising source of genes encoding novel thermostable enzymes. We have isolated a novel cutinase, LC-cutinase, with polyethylene terephthalate (PET) degrading activity [5], and 12 novel RNases H1, LC1 ~ LC12-RNases H1 [6], from this compost using a metagenomic approach. Structural and functional studies of these enzymes have shown that LC-cutinase is a kinetically robust protein with a slow unfolding rate [7]; LC9-RNase H1 is a bacterial RNase H1 with an atypical DEDN active site motif, evolutionarily distinct from those with a typical DEDD active site motif [8]; and LC11-RNase H1 is a prokaryotic RNase H1 with a unique substrate recognition mechanism, which interacts with four non-consecutive, instead of consecutive, 2′-OH groups of the RNA strand of RNA/DNA substrates [9,10]. However, no cellulases have been isolated from this compost.

*Abbreviations:* GH family, glycoside hydrolase family; LC-CelA, GH family 12 cellulase from leaf–branch compost; SP, signal peptide; FL, flexible linker; CM-cellulose, carboxymethyl cellulose; DTT, dithiothreitol; CD, circular dichroism; GdnHCl, guanidine hydrochloride

☆ The study represents a portion of the dissertation submitted by Mr. Hiroyuki Okano to Osaka University in partial fulfillment of the requirement for his PhD.

* Corresponding author. Tel./fax: +81 6 6879 7938.
*E-mail address:* kanaya@mls.eng.osaka-u.ac.jp (S. Kanaya).
[1] Present address: Department of Biological Sciences, University of Calgary, Calgary, Alberta T2N 1N4, Canada.

http://dx.doi.org/10.1016/j.fob.2014.10.013
2211-5463/

Cellulases hydrolyze β-1,4-glycoside bonds in cellulose and have received much attention because of their potential application in the production of bioethanol from cellulosic biomass [11–15]. Cellulases consist of three types of enzymes, endo-1,4-β-D-glucanase (endoglucanase; EC 3.2.1.4), exo-1,4-β-D-glucanase (cellobiohydrolase; EC 3.2.1.91 and 3.2.1.176), and β-glucosidase (EC 3.2.1.21). It has recently be shown that lytic polysaccharide monooxygenase enzyme AA9, formerly known as glycoside hydrolase (GH) family 61, represent another class of cellulose degrading enzymes [16]. These enzymes synergistically hydrolyze cellulose to produce glucose [16,17]. According to the CAZy database (http://www.cazy.org/), endoglucanase, cellobiohydrolase and β-glucosidase are classified into twelve, five and six families respectively. GH family 12 is one of the endoglucanase families. A GH family 12 cellulase (Cel12A) from *Rhodothermus marinus*, termed *Rm*Cel12A, is one of the most thermostable endoglucanases identified to date [18]. *Rm*Cel12A has a putative signal peptide (SP, residues 1–17) and a flexible linker (FL, residues 18–37) at the N-terminus. The crystal structure of *Rm*Cel12A without these SP and FL regions (residues 38–261) has been determined in substrate-free [19] and substrate-bound [20] forms. It has been proposed that dimerization, and a cluster of aromatic residues in the active site cleft, contribute to the adaptation of *Rm*Cel12A to a high temperature environment [20].

Thermostable enzymes are generally more useful than thermolabile ones in industry because of their higher temperature, chemical, and pH stability, lower production cost, and longer shelf life [21–23]. In addition, as the reaction temperature increases, the reaction rate increases, the solubility of the substrate and product increases, viscosity of the reaction mixture decreases, and the risk of microbial contamination decreases. Therefore, it would be informative to examine whether thermostable enzymes with cellulose-degrading activity can be isolated from a leaf–branch compost by a metagenomic approach.

In this report, we showed that ten novel cellulases were isolated from leaf–branch compost by a metagenomic approach. We also showed that one of them, LC-CelA, which has high amino acid sequence identity to *Rm*Cel12A, was highly thermostable. X-ray crystallographic studies and structure-based mutational studies indicate that hydrogen bonds formed between the C-terminal region of the FL and the central region of the protein, and the two disulfide bonds contribute to the stabilization of LC-CelA.

## 2. Materials and methods

### 2.1. Cells, plasmids, and enzymes

*Escherichia coli* BL21-CodonPlus(DE3)-RP was from Stratagene (La Jolla, CA, USA). Plasmid pET25b was from Novagen (Madison, WI, USA). *E. coli* BL21-CodonPlus(DE3)-RP transformants were grown in lysogeny broth (LB) medium (10 g Tryptone; 5 g Yeast extract; 10 g NaCl in 1 L $H_2O$) supplemented with 50 mg $L^{-1}$ ampicillin.

### 2.2. Construction of DNA library and screening

The 4-month-old leaf–branch compost made in EXPO Park, Japan was used to construct a metagenomic DNA library. The temperature and pH of this leaf–branch compost are 67 °C and pH 7.5. Extraction of DNA from this compost and construction of a DNA library for metagenomic study using CopyControl™ Fosmid Library Production kit (EPICENTRE Biotechnologies, Madison, WI, USA) were performed as described previously [5]. This DNA library was spread on LB-agar plates containing 12.5 μg $mL^{-1}$ chloramphenicol, 0.01% L-arabinose, 0.5% CM-cellulose, and 0.1 mg $mL^{-1}$ trypan blue. CM-cellulose and trypan blue have been used as a cellulose substrate and a chromogenic dye respectively for detection of cellulolytic activity [24]. The resultant plates were incubated at 37 °C for several days. Plasmids were extracted from colonies, which form halos around them due to hydrolysis of CM-cellulose. Genes encoding CM-cellulose degrading enzymes were identified by transposon mutagenesis using EZ-Tn5TM<T7/KAN-2> Promoter Insertion kit (EPICENTRE Biotechnologies), according to the procedures recommended by the supplier. Nucleotide sequence of the gene was determined by an ABI Prism 3100 DNA sequencer (Applied Biosystems, Tokyo, Japan). Oligonucleotides for sequencing were synthesized by Hokkaido System Science (Sapporo, Hokkaido, Japan).

### 2.3. Construction of plasmids

For construction of plasmid pET-LC-CelA used to overproduce LC-CelA (residues 20–261 of pre-LC-CelA) in a form with Met-Asp at the N-terminus, the gene encoding LC-CelA was amplified by PCR using the fosmid vector harboring the pre-LC-CelA gene as a template. The sequences of the PCR primers were 5′-GATCCATGGATCTGTTCCCAGAGAAAAATG-3′ for 5′-primer and 5′-CAAGAATTCACTTTAGCGTGCGGTC-3′ for 3′-primer, where underlines represent the *Nco*I site for 5′-primer and *Eco*RI site for 3′-primer. The resultant DNA fragment was digested with *Nco*I and *Eco*RI, and ligated into the *Nco*I- *Eco*RI sites of pET25b to generate plasmid pET-pelB-LC-CelA, which harbors the gene encoding the pelB-LC-CelA fusion protein. The small *Nde*I and *Nco*I fragment encompassing the pelB leader sequence was then deleted from this plasmid by digestion with *Nde*I and *Nco*I, the 5′-overhang filled in by Klenow DNA polymerase, and re-ligation to generate pET-LC-CelA.

For constructions of plasmids pET-LC-CelA-His and pET-ΔFL-LC-CelA-His used to overproduce LC-CelA-His and ΔFL-LC-CelA-His (residues 36–261) in a form with Met at the N-terminus and a His-tag at the C-terminus, the genes encoding LC-CelA and ΔFL-LC-CelA were amplified by PCR using the fosmid vector harboring the pre-LC-CelA gene as a template. The sequences of the 5′-primers were 5′-GGATGCCATATGCTGTTCCCAGAG-3′ for LC-CelA and 5′-GAAGATCATATGACGGCTACAGTATGCG-3′ for ΔFL-LC-CelA. The sequence of the 3′-primer was 5′-TTCGTCGACGCGTGCGGTCAAAGTAAC-3′ for both LC-CelA and ΔFL-LC-CelA. In these sequences, underlines represent the *Nde*I site for 5′-primer and *Sal*I site for 3′-primer. The resultant DNA fragments were digested with *Nde*I and *Sal*I, and ligated into the *Nde*I- *Sal*I sites of pET25b.

Plasmid pET-E34A-LC-CelA-His used to overproduce E34A-LC-CelA-His in a form with Met at the N-terminus and a His-tag at the C-terminus was constructed by PCR with the KOD-plus mutagenesis kit (Toyobo Co., Ltd., Osaka, Japan), according to the procedures recommended by the supplier. Plasmid pET-LC-CelA-His was used as a template. The mutagenic primers were designed in such a way that the GAG codon for Glu34 was changed to GCT for Ala.

PCR was performed in 25 cycles with Gene Amp PCR system 2400 (Applied Biosystems) using KOD DNApolymerase (Toyobo). PCR primers and mutagenic primers were synthesized by Hokkaido System Science. The nucleotide sequence was confirmed by an ABI Prism 3100 DNA sequencer (Applied Biosystems).

### 2.4. Overproduction and purification

*E. coli* BL21-CodonPlus(DE3) transformants with the pET25b derivatives were cultivated at 37 °C. When the absorbance of the culture at 600 nm reached around 0.5, isopropyl-β-D-thio galactopyranoside (IPTG) was added to the culture medium and cultivation was continued for an additional 4 h. Cells were then harvested by centrifugation at 6000*g* for 10 min, suspended in

10 mM Tris–HCl (pH 8.0) containing 1 mM EDTA (TE buffer), disrupted by sonication lysis, and centrifuged at 30,000*g* for 30 min. The supernatant was collected, dialyzed against TE buffer, incubated at 70 °C for 30 min for heat treatment, and centrifuged at 30,000*g* for 30 min. The subsequent purification procedures were carried out at 4 °C. For purification of LC-CelA-His and its derivatives with a C-terminal His-tag, the supernatant obtained after heat treatment was dialyzed against 20 mM Tris–HCl (pH 7.0) containing 10 mM imidazole and 0.3 M NaCl, and applied to a Ni Sepharose 6 Fast Flow column (GE Healthcare) equilibrated with the same buffer. The protein was eluted from the column by linearly increasing the imidazole concentration from 10 to 300 mM. The fractions containing the protein were collected and dialyzed against 10 mM Tris–HCl (pH 7.0). For purification of LC-CelA without a His-tag, the supernatant obtained after heat treatment was loaded onto a HiTrap Q HP column (GE Healthcare, Tokyo, Japan) equilibrated with TE buffer containing 1 mM DTT. The protein was eluted from the column by linearly increasing the NaCl concentration from 0 to 1 M. The fractions containing the protein were collected and applied to a Hi-Load 16/60 Superdex 200 pg column (GE Healthcare) equilibrated with TE buffer containing 50 mM NaCl for gel filtration chromatography. The fractions containing the protein were pooled.

The production level of the protein in *E. coli* cells and the purity of the protein were analyzed by SDS–polyacrylamide gel electrophoresis [25] using a 12% polyacrylamide gel, followed by staining with Coomassie brilliant blue (CBB). The amount of the protein was estimated from the intensity of the band visualized by CBB staining using the Scion Image program. The N-terminal amino acid sequence of the protein was determined by a Procise automated sequencer model 491 (Applied Biosystems). The protein concentration was determined from the UV absorption on the basis that the absorbance of a 0.1% (1.0 mg $mL^{-1}$) solution at 280 nm is 3.37 for LC-CelA, 2.96 for LC-CelA-His, 3.15 for ΔFL-LC-CelA-His and 2.97 for E34A-LC-CelA-His. These values were calculated by using $\varepsilon$ = 1,526 $M^{-1}$ $cm^{-1}$ for tyrosine and 5225 $M^{-1}$ $cm^{-1}$ for tryptophan at 280 nm [26].

### 2.5. Sequence analysis

Blast searches of the amino acid sequences of metagenome-derived cellulases deduced from their nucleotide sequences were performed using the DDBJ blastp search tool (http://blast.ddbj.nig.ac.jp/blastn?lang = en). The flexible and hydrophilic regions of each amino acid sequence were predicted by using the PROTSCALE tool (http://web.expasy.org/protscale/). Calculation of p*I* for the corresponding amino acid region is performed by using the Compute p*I*/Mw tool (http://web.expasy.org/compute_pi/). Domain search was performed by using the SMART tool (http://smart.embl.de/) [34].

### 2.6. Determination of enzymatic activity

The enzymatic activity was determined at the temperatures indicated by the dinitrosalicylic acid (DNS) stopped method [27] using CM-cellulose as a substrate. The reaction mixture (100 μL) contained 100 mM sodium phosphate (pH 7.0) and 1% (w/v) CM-cellulose (low viscosity grade, Sigma–Aldrich Co., St Louis, MO, USA). The enzymatic reaction was initiated by adding an appropriate amount of the enzyme and terminated by adding 10 μL of 10% SDS and boiling for 5 min. The reaction time was 10 min. The resultant solution was mixed with 300 μL of the DNS solution prepared as described previously [27], boiled for 5 min, and cooled on ice. After centrifugation at 17,000*g* for 5 min, an aliquot of the supernatant (100 μL) was withdrawn, diluted twice by distilled water, and measured for absorption at 500 nm ($A_{500}$). The amount of reducing sugars released from the substrate was estimated from the $A_{500}$ value by using glucose as standard. One unit of enzymatic activity was defined as the amount of enzyme that produced 1 μmol of reducing sugars per min.

For analysis of the temperature dependence of activity, the activity was determined at pH 7.0 and various temperatures ranging from 40 to 100 °C. For the analysis of pH dependence of activity, the activity was determined at 90 °C and various pHs ranging from pH 4.0 to 10.5. The buffers used for this analysis were 100 mM sodium citrate (pH 4.0–6.0), 100 mM sodium phosphate (pH 6.0–8.0), and 100 mM glycine–NaOH (pH 8.0–10.5).

### 2.7. Measurement of circular dichroism (CD) spectra

The far-UV (200–260 nm) CD spectrum of the protein was measured at 25 °C on a J-725 spectropolarimeter (Japan Spectroscopic, Tokyo, Japan). The protein was dissolved in 10 mM Tris–HCl (pH 7.0). The protein concentration was 0.1 mg $mL^{-1}$ and a cell with an optical path length of 2 mm was used. The mean residual ellipticity ($\theta$, deg $cm^2$ $dmol^{-1}$) was calculated using an average amino acid molecular mass of 110 Da.

### 2.8. Thermal denaturation

The thermal denaturation curve of the protein was obtained by monitoring the change in CD values at 222 nm as the temperature was increased. The protein was dissolved in 10 mM Tris–HCl (pH 7.0) containing 3.0 M GdnHCl. The protein concentration and optical path length were 0.1 mg $mL^{-1}$ and 2 mm, respectively. The rate of temperature increase was 3.0 °C $min^{-1}$. The temperature of the midpoint of the transition, $T_m$, was calculated by curve fitting of the resultant CD values versus temperature data on the basis of a least-square analysis.

### 2.9. Crystallization

For crystallization, LC-CelA was dialyzed against 10 mM Tris–HCl (pH 8.0) and concentrated to 10 mg $mL^{-1}$ using an ultrafiltration system Centricon (Millipore, Billerica, MA, USA). The crystallization conditions were initially screened using crystallization kits from Hampton Research (Aliso Viejo, CA, USA) (Crystal Screens I and II) and Emerald BioStructures, Inc. & Emerald BioSystems (Bainbridge Island, WA, USA) (Wizard I, II, III and IV). The conditions were surveyed using sitting-drop vapor-diffusion method at 4 °C and 20 °C. Drops were prepared by mixing 1 μL each of the protein and reservoir solutions, and were vapor-equilibrated against a 100 μL reservoir solution. Crystals of LC-CelA appeared after a few weeks in Wizard I No. 28 [0.1 M 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) (pH 7.5), 0.2 M NaCl and 20% (w/v) polyethylene glycol (PEG) 3000]. To improve the crystal quality, the crystallization conditions were further optimized. Diffraction-quality crystals were obtained when the protein concentration was increased to 11.7 mg $mL^{-1}$ and the reservoir solution was changed to 0.1 M HEPES (pH 7.5) containing 0.2 M NaCl and 23% (w/v) PEG 3350.

### 2.10. X-ray diffraction data collection and structure determination

X-ray diffraction data set of LC-CelA was collected at −173 °C without cryoprotectant at a wavelength of 0.9 Å with the beam line BL44XU at SPring-8 (Hyogo, Japan). Diffraction data set was indexed, integrated and scaled using the HKL2000 program suite [28]. The structure was solved by the molecular replacement method using MOLREP [29] in the CCP4 program suite [30]. The 1.85 Å structure of *Rm*Cel12A (PDB: 1H0B) was used as a starting model. Automated model building was done by using ArpWarp [31]. Structural refinement was carried out by using REFMAC

[32] of the CCP4 program and the model was corrected using COOT [33]. The statistics for data collection and refinement are summarized in Table 3. The figures were prepared using PyMol (http://www. pymol.org).

### 2.11. Protein data bank accession number

The coordinates and structure factors for LC-CelA have been deposited in the Protein Data Bank under code PDB: 3WX5.

## 3. Results and discussion

### 3.1. Cloning of cellulase genes from metagenomic DNA library

The size of the metagenomic DNA library prepared from the leaf–branch compost was approximately $2.1 \times 10^4$ colony forming units (CFU). The restriction fragment length polymorphisms of ten randomly selected clones using restriction enzyme *Bam*HI showed non-redundant patterns and an average insert size of 35 kb. Screening of the library for genes encoding cellulose degrading enzymes was performed using CT-agar plates (agar plates containing CM-cellulose and trypan blue). Of approximately 6000 clones screened, 24 clones gave a halo on CT-agar plates when they were incubated at 37 °C for 3 days. Ten of them, which gave a halo on CT-gellan gum plates (gellan gum plates containing CM-cellulose and trypan blue) at 50 °C, were chosen to determine the nucleotide sequences of the genes responsible for halo formation. Determination of the nucleotide sequences of these genes by transposon mutagenesis indicated that all ten clones harbored genes encoding cellulases, that were different from one another. Because these cellulases have a putative signal peptide (SP) at their N-terminus (as described below), they are termed pre-LC-CelA∼J. Clones harboring the genes encoding pre-LC-CelA, pre-LC-CelD, pre-LC-CelF, pre-LC-CelG, and pre-LC-CelI produced a halo on CT-gellan gum plates even at 80 °C, suggesting that these cellulases are highly thermostable.

### 3.2. Amino acid sequences of pre-LC-CelA∼J

Blast searches of the amino acid sequences of pre-LC-CelA∼J deduced from their nucleotide sequences indicated that all the cellulases are novel, as summarized in Table 1. The number of constituent amino acid residues in these cellulases varies from 261 to 782. The highest identity between the amino acid sequence of either one of these cellulases and that available from the database varies from 42% to 76%. This result indicates that a metagenomic approach is useful to increase our knowledge on sequence space of cellulase. The nucleotide sequences of the genes encoding pre-LC-CelA∼J are deposited in GenBank under accession numbers KF626648–KF626657.

The amino acid sequences of pre-LC-CelA∼J are schematically shown in Fig. 1. Domain searches using the SMART tool [34] allow us to predict a SP, a catalytic module of cellulase, and a cellulose binding domain/module. All ten cellulases have a 16–36 residue long SP at their N-terminus, suggesting that they are secretory proteins. They also have a catalytic module of cellulase. Three (pre-LC-CelA, pre-LC-CelD and pre-LC-CelE) have a catalytic module of GH family 12 cellulase, two (pre-LC-CelF and pre-LC-CelG) have a catalytic module of GH family 9 cellulase, and two (pre-LC-CelB and pre-LC-CelJ) have a catalytic module of GH family 6 cellulase. Pre-LC-CelI, pre-LC-CelH, and pre-LC-CelC have catalytic modules of GH family 3, 44, and 51 cellulases respectively. In addition, pre-LC-CelD and pre-LC-CelF have a cellulose binding domain II (CBDII) and pre-LC-CelF has a cellulose binding module 3 (CBM3). Analyses of the hydrophilic and flexible regions using the PROTSCALE tool allow us to predict a hydrophilic flexible linker (FL). All the cellulases, except for pre-LC-CelG, have a 14–29 residue long FL between the putative SP and the catalytic domain. Pre-LC-CelG has a 28 residue long FL within a catalytic module of cellulase. These FLs are acidic, because their p*I* values range from 3.28 to 5.52.

Pre-LC-CelA∼J are probably secreted in microorganisms as LC-CelA∼J, because SPs are removed by signal peptidase upon secretion. Of them, LC-CelA is expected to be the most stable cellulase, because it shows high amino acid sequence identity to GH family 12 cellulase from *R. marinus* (*Rm*Cel12A) (Table 1). *Rm*Cel12A has a 20 residue long FL (residues 18–37) [35]. Removal of this FL decreases the $T_m$ value of *Rm*Cel12A by 8.4 °C and its half-life at 90 °C from 5 h to 2 h [35], suggesting that the FL contributes to the stabilization of *Rm*Cel12A. However, this stabilization mechanism remains to be understood, because the crystal structure of *Rm*Cel12A has been determined only in a form without the FL [19,20]. Therefore, in order to examine whether LC-CelA is, like *Rm*Cel12A, a highly thermostable enzyme, and to analyze the role of the FL, we decided to overproduce LC-CelA in *E. coli*, purify and characterize it, and determine its crystal structure.

### 3.3. Comparison of amino acid sequences of pre-LC-CelA and pre-RmCel12A

The amino acid sequence of pre-LC-CelA is compared with that of pre-*Rm*Cel12A in Fig. 2. Both proteins consist of 261 residues. However, the regions predicted as SP (residues 1–19) and FL (residues 20–35) in pre-LC-CelA are slightly different from those predicted as SP (residues 1–17) and FL (residues 18–37) in pre-*Rm*Cel12A. According to the crystal structure of ΔFL-*Rm*Cel12A (*Rm*Cel12A without FL, residues 38–261) in complex with the substrate [20], Trp9, Trp26, Trp69, Trp159, Trp161, Tyr163, and Trp209

**Table 1**
List of cellulases isolated from leaf–branch compost and proteins with the highest amino acid sequence identities.

| Cellulases | No. of residues | Protein with the highest sequence identity | | | |
|---|---|---|---|---|---|
| | | Protein | Source organism[a] | Accession No. | Identity (%) |
| Pre-LC-CelA | 261 | Glycoside hydrolase family 12 | *Rhodothermus marinus* (65 °C) | G2SJ29 | 76 |
| Pre-LC-CelB | 596 | Cellulase | *Plesiocystis pacifica* SIR-1 (18 °C) | A6G3Z7 | 50 |
| Pre-LC-CelC | 515 | α-L-Arabinofuranosidase-like protein | *Opitutus terrae* (30 °C) | B1ZZJ5 | 45 |
| Pre-LC-CelD | 400 | Glycosyl hydrolase family 12 | *Thermobispora bispora* (50 °C) | D6Y3Q3 | 45 |
| Pre-LC-CelE | 286 | Cellulase 12A | *Streptomyces sp.* 11AG8 (35 °C) | Q9KIH1 | 43 |
| Pre-LC-CelF | 782 | Uncharacterized protein | *Streptosporangium roseum* (26 °C) | D2B808 | 62 |
| Pre-LC-CelG | 577 | Glycosyl hydrolase family 9 | *Microcoleus sp.* PCC 7113 (25 °C) | K9WM66 | 42 |
| Pre-LC-CelH | 742 | Cellulase | *Candidatus methylomirabilis oxyfera* (25 °C) | D5MF13 | 64 |
| Pre-LC-CelI | 592 | Glucosidase-like glycosyl hydrolase | *Xenococcus sp.* PCC 7305 (25 °C) | L8M232 | 57 |
| Pre-LC-CelJ | 326 | Glycoside hydrolase family 6 | *Frankia sp.* EUN1f (30 °C) | D3D8M1 | 53 |

[a] The optimum growth temperature of each organism is shown in parenthesis.

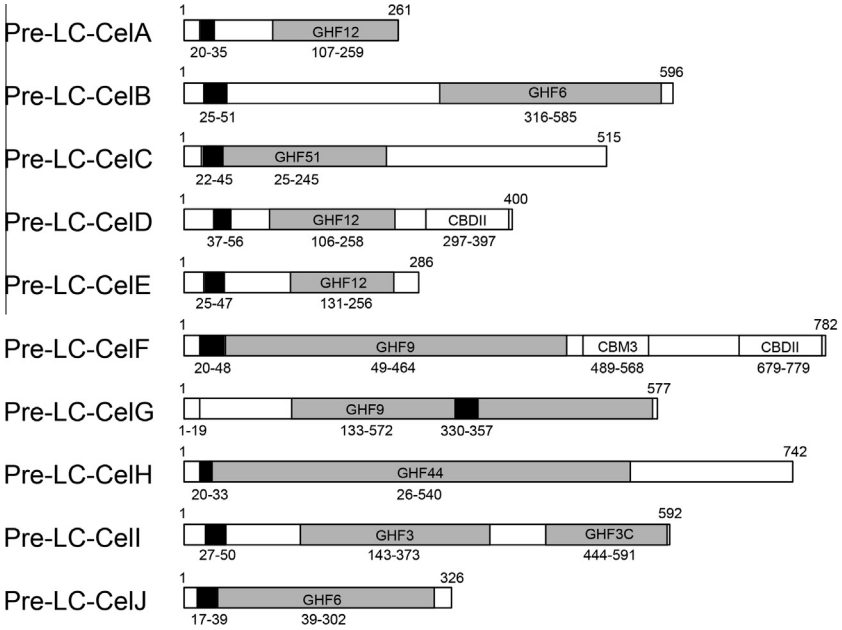


**Fig. 1.** Schematic representation of the primary structures of metagenome-derived cellulases from leaf–branch compost. Putative signal peptides (N-terminal open boxes), putative flexible linkers (black boxes), glycoside hydrolase family (GHF) domains (gray boxes), and cellulose binding domains/modules (C-terminal open boxes) are shown. The numbers above the sequence represent the positions of the N- and C-terminal residues of each protein. The numbers below the sequence represent the positions of the N- and C-terminal residues of each region or domain.

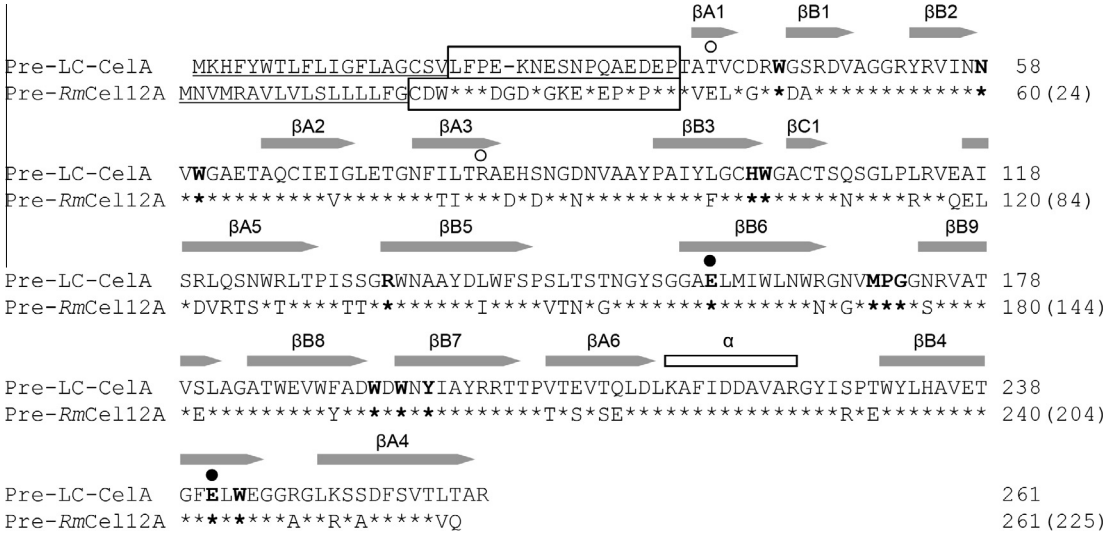


**Fig. 2.** Alignment of the amino acid sequences of pre-LC-CelA and pre-Cel12A from *Rhodothermus marinus* (pre-*Rm*Cel12A). The accession numbers are KF626648 for pre-LC-CelA and G2SJ29 for pre-*Rm*Cel12A. For the pre-*Rm*Cel12A sequence, the identical residues with those in the pre-LC-CelA sequence are indicated by asterisks (*). Gaps are shown by dashes. A putative signal peptide (SP) is underlined. The region predicted as a flexible linker (FL) between a putative SP and a catalytic domain is boxed. The amino acid residues responsible for dimerization and catalytic activity of *Rm*Cel12A are denoted above the sequences by open and closed circles respectively. The aromatic and non-aromatic amino acid residues that form the active site cleft of *Rm*Cel12A and corresponding residues of LC-CelA are shown in boldface. Numbers represent the positions of the amino acid residues that start from the initiator methionine residue for each protein. Numbers in parentheses represent the positions of the amino acid residues that start from the initiator methionine for the *Rm*Cel12A derivative without SP and FL used to determine the crystal structure [Met-(Thr38-Gln261)]. The ranges of the secondary structures of LC-CelA, which are identical to those of *Rm*Cel12A, are shown above the sequences.

(Trp45, Trp62, Trp104, Trp195, Trp197, Tyr199, and Trp245 respectively in pre-*Rm*Cel12A) form an extensive aromatic network in the active site cleft, to which the substrate binds. Asn24, His67, Arg100, Met136, Pro137, and Gly138 (Asn60, His103, Arg136, Met172, Pro173, and Gly174 respectively in pre-*Rm*Cel12A) also form this cleft. Two catalytic residues, Glu124 and Glu207 (Glu160 and Glu243 respectively in pre-*Rm*Cel12A) are located at this cleft. All of these residues are fully conserved in pre-LC-CelA.

### 3.4. Overproduction and purification of LC-CelA and LC-CelA-His

Because the SP of pre-*Rm*Cel12A is cytotoxic to *E. coli* cells, and the production level of pre-*Rm*Cel12A greatly increases on removal of the SP [35], LC-CelA (residues 20–261) was overproduced in *E. coli* without the SP, either in a non-His-tagged or a His-tagged form. LC-CelA in a non-His-tagged form with Met-Asp at its N-terminus is simply designated as LC-CelA, whereas LC-CelA in a His-

tagged form with Met at the N-terminus and a His-tag at the C-terminus is designated as LC-CelA-His. LC-CelA and LC-CelA-His were used for structural analysis and biochemical characterization respectively. Upon induction for overproduction, LC-CelA and LC-CelA-His equally accumulated in *E. coli* cells, in both soluble and insoluble forms. The fraction of the protein in soluble form was 20–30%. This protein was purified to give a single band on SDS–PAGE by heat treatment, followed by nickel affinity chromatography (LC-CelA-His) or ion-exchange and gel filtration chromatographies (LC-CelA) (data not shown). The amount of protein purified from 1 L culture was typically 3 mg for both LC-CelA and LC-CelA-His. The N-terminal amino acid sequence of LC-CelA was determined to be Met-Asp-Leu-Phe-, indicating that LC-CelA contains the entire FL region.

The molecular mass of LC-CelA was estimated to be 29 kDa by gel filtration chromatography. This value is comparable to the calculated mass (26.9 kDa), suggesting that LC-CelA exists as a monomer. ΔFL-*Rm*Cel12A has been reported to exist as a dimer [20]. The interface of the two monomers is stabilized by two salt bridges formed between Glu4 (Glu40 in pre-*Rm*Cel12A) of one monomer and Arg47 (Arg83 in pre-*Rm*Cel12A) of the other. Arg47 is conserved as Arg81 in LC-CelA, whereas Glu4 is replaced by Thr38 in LC-CelA. LC-CelA does not form a dimer, probably because the salt bridges that stabilize the interface of two monomers are not formed.

### 3.5. Activity of LC-CelA-His

The temperature dependence of the enzymatic activity of LC-CelA-His was analyzed by measuring the activity at various temperatures ranging from 40 to 100 °C using CM-cellulose as a substrate. As shown in Fig. 3A, LC-CelA-His exhibited the highest activity at 90 °C. This temperature was slightly lower than, but comparable to, that of *Rm*Cel12A in a form without the SP, which has been reported to be ⩾100 °C [35]. However, the specific activity of LC-CelA-His determined at 60 °C, 4.2 units $mg^{-1}$, was slightly higher than that of *Rm*Cel12A without the SP determined at 65 °C (3.1 units $mg^{-1}$) [35]. These results indicate that LC-CelA-His is a highly thermostable enzyme, like *Rm*Cel12A, and is slightly more active than *Rm*Cel12A.

The pH dependence of the enzymatic activity was analyzed by measuring the activity at various pHs ranging from pH 4.0 to 10.5. As shown in Fig. 3B, LC-CelA-His exhibited nearly maximal activity across a broad pH range (pH 5.0–9.0). *Rm*Cel12A also exhibits nearly maximal activity over a broad pH range (pH 5.0–8.0), but shows only ∼50% of the maximal activity at pH 9.0 [18]. This result indicates that the pH range suitable for activity of LC-CelA-His is larger than that of *Rm*Cel12A.

### 3.6. Stability of LC-CelA-His against heat inactivation

To analyze the stability of LC-CelA-His against irreversible heat inactivation, the enzyme (0.05 mg $mL^{-1}$) was incubated in 100 mM sodium phosphate (pH 7.0) at various temperatures ranging from 60 to 100 °C for 30 min, and the residual activity was determined at pH 7.0 and 60 °C. As shown in Fig. 4, LC-CelA-His almost fully retained activity after incubation at 90 °C for 30 min, whereas it almost totally lost activity after incubation at 95 °C for 30 min. This result is consistent with the observation that the optimum temperature for LC-CelA-His activity is 90 °C (Fig. 3A). LC-CelA-His almost completely lost activity after incubation at 100 °C for 30 min (Fig. 4), but it exhibited approximately 80% of the maximal activity in activity assays at 100 °C (Fig. 3A), probably because the enzyme hydrolyzed the substrate before it was fully denatured in those assay conditions.

### 3.7. Thermal denaturation of LC-CelA-His

To analyze the stability of LC-CelA-His more quantitatively, thermal denaturation of this protein was analyzed at pH 7.0 in the presence of 3 M guanidine hydrochloride (GdnHCl) by monitoring the change in CD values at 222 nm. Thermal denaturation of the protein was reversible in this condition. It was necessary to use 3 M GdnHCl because the protein was not fully denatured even at 100 °C in the presence of ⩽2 M GdnHCl. Fig. 5 shows the thermal denaturation curve of LC-CelA-His measured in the presence of 3 M GdnHCl. The midpoint of the transition of this curve, $T_m$, is 86.8 °C (Table 2). The $T_m$ value of *Rm*Cel12A has been reported to be 102.9 °C in the absence of denaturant [35]. Because LC-CelA-His is not fully denatured even at 100 °C in the absence of denaturant, LC-CelA-His may be nearly as stable as *Rm*Cel12A.

### 3.8. Activity and stability of LC-CelA

To examine the effects of attachment of a C-terminal His-tag, the activity and stability of LC-CelA were determined and compared with those of LC-CelA-His. The activity of LC-CelA was determined at 60 °C using CM-cellulose as a substrate. The specific activity of LC-CelA determined at this temperature was 4.3 units $mg^{-1}$, which was comparable to that of LC-CelA-His (4.2 units $mg^{-1}$). The stability of LC-CelA was analyzed at pH 7.0 in the presence of 3 M GdnHCl by CD spectroscopy. Thermal dena-

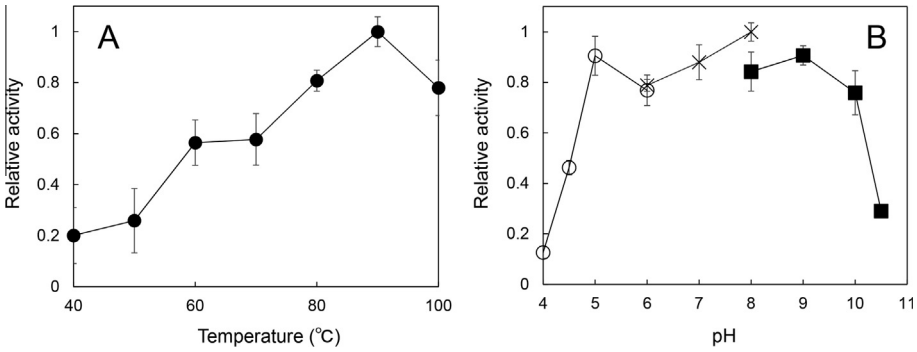


**Fig. 3.** Optimum temperature and pH for activity of LC-CelA-His. The temperature (A) and pH (B) dependencies of the enzymatic activity of LC-CelA-His are shown. The activity was determined at pH 7.0 and the temperatures indicated (A) or at 90 °C and the pHs indicated (B) using 1% (w/v) carboxymethyl-cellulose (CM-cellulose) as a substrate, as described in Experimental procedures. The buffers used to analyze the pH dependence of the activity were 100 mM sodium citrate (pH 4.0–6.0), 100 mM sodium phosphate (pH 6.0–8.0) and 100 mM Glycine–NaOH (pH 8.0–10.5). The experiment was carried out at least twice, and errors from the average values are indicated by *vertical lines*.

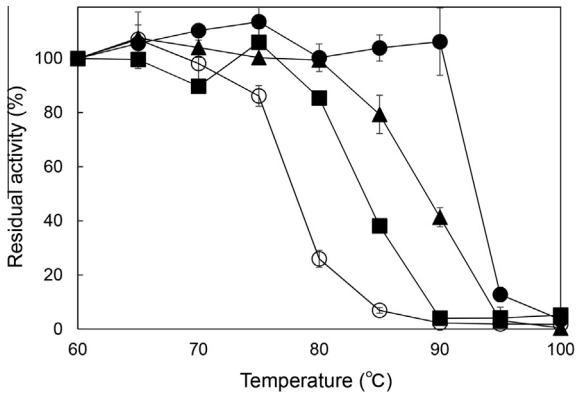


**Fig. 4.** Stability of LC-CelA-His and its derivatives against heat inactivation. LC-CelA-His was incubated at pH 7.0 and the temperatures indicated for 30 min in the absence of DTT (filled circle) or in the presence of 1 mM DTT (open circle). ΔFL-LC-CelA-His (filled square) and E34A-LC-CelA-His (filled triangle) were also incubated at the same condition in the absence of DTT. The residual activities were determined at pH 7.0 and 60 °C using CM-cellulose as a substrate, as described in Experimental procedures. The experiment was carried out at least twice, and errors from the average values are indicated by *vertical lines*.

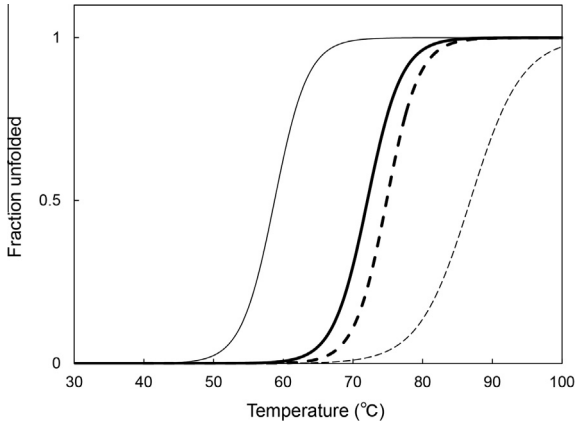


**Fig. 5.** Thermal denaturation curves of LC-CelA-His and its derivatives. Thermal denaturation curves of LC-CelA-His (thin dashed line), ΔFL-LC-CelA-His (thick solid line) and E34A-LC-CelA-His (thick dashed line) obtained in the absence of DTT and that of LC-CelA-His obtained in the presence of 1 mM DTT (thin solid line) are shown. These curves were obtained at pH 7.0 in the presence of 3M GdnHCl by monitoring the change in CD values at 222 nm as described in Experimental procedures.

turation of LC-CelA was reversible in this condition. The $T_m$ value of this protein was determined to be 86.0 °C, which was comparable to that of LC-CelA-His (86.8 °C). These results indicate that attachment of a C-terminal His-tag does not significantly affect the activity or stability of LC-CelA.

### 3.9. Crystal structure of LC-CelA

To examine whether the FL region of LC-CelA is folded into a defined structure, the crystal structure of LC-CelA was determined at 1.85 Å resolution. The asymmetric unit of the crystal structure consists of two LC-CelA molecules (A and B). Both molecules consist of residues 33–261. The N-terminal Met-Asp and most of the FL region (residues 20–32) are not visible in the electron density map probably due to structural disorder. The structures of these two molecules are nearly identical with a root-mean-square deviation (RMSD) value of 0.08 Å for 229 Cα atoms. We used the structure of molecule A in this study. Details of the data-collection statistics and refinement are summarized in Table 3.

The overall structure of LC-CelA (molecule A) is shown in Fig. 6A. This structure belongs to the typical β-jelly roll fold and consists of two β-sheets that pack with each other and one α-helix. The structure of LC-CelA excluding the FL region is nearly identical to that of ΔFL-*Rm*Cel12A (PDB: 1H0B) with a RMSD value of 0.33 Å for 226 Cα atoms. Superimposition of the structure of LC-CelA around the active site onto that of the ΔFL-*Rm*Cel12A-substrate (cellopentose) complex determined by the soaking method [20] shows that the steric configurations of all aromatic and non-aromatic residues forming the active site cleft of LC-CelA, including two acidic catalytic residues (Glu158 and Glu241), are nearly identical to those of *Rm*Cel12A (Fig. 6B). In the active site cleft of LC-CelA, a band of aromatic residues, mostly tryptophan residues, probably provides a series of glucose-binding subsites through stacking interactions with glucose residues, and charged and polar residues provide specificity through hydrogen bonding interactions with sugar molecules, as reported for *Rm*Cel12A [20]. In the crystal structure of ΔFL-*Rm*Cel12A, a HEPES molecule binds to the active site [19]. However, electron density of this molecule was not clearly detected in the LC-CelA structure, suggesting that no HEPES molecule binds strongly to the active site of LC-CelA. This may be due to slight differences in the active site geometry. The LC-CelA structure shows that two disulfide bonds are formed between Cys40 and Cys67 and between Cys100 and Cys105. The Cys40–Cys67 bond apparently anchors the N-terminal region of the catalytic domain (loop between βA1 and βB1 strands) to the central region (βA2 strand). The Cys100–Cys105 bond forms a short loop (residues 101–104). The Cys40–Cys67 bond is well conserved in GH family 12 cellulases from various organisms, whereas the Cys100–Cys105 bond is conserved only in GH family 12 cellulases from *R. marinus* and *Streptomyces* species.

As noted above, most of the FL region is structurally disordered. However, the C-terminal region of the FL (Asp33-Pro35) is visible in the electron density map (Fig. 6C), indicating that this region is not included in the FL. Interestingly, this region interacts with the central region (loop between βA2 and βA3 strands) via three hydrogen bonds, formed between Glu34 $O^{\delta1}$ and Thr74 N, Glu34 $O^{\delta1}$ and Glu73 N, and Pro35 O and Leu72 N. Because all of these residues are conserved in *Rm*Cel12A, these hydrogen bonds are

**Table 2**
Activities and stabilities of LC-CelA-His and its derivatives.

| Protein | [DTT] (mM) | Specific activity[a] (U/mg) | Relative activity (%) | $T_m$[b] (°C) | $\Delta T_m$[c] (°C) |
|---|---|---|---|---|---|
| LC-CelA-His | 0 | 4.2 | 100 | 86.8 | – |
| | 1 | 4.2 | 100 | 58.8 | −28.0 |
| ΔFL-LC-CelA-His | 0 | 3.9 | 93 | 72.1 | −14.7 |
| E34A-LC-CelA-His | 0 | 4.0 | 96 | 74.8 | −12.0 |

[a] The specific activity was determined at 60 °C either in the presence or absence of DTT using CM-cellulose as a substrate, as described in Experimental procedures. Each experiment was carried out at least twice and the average value is shown. Errors are within 10% of the values reported.
[b] The melting temperature ($T_m$), which is the temperature of the midpoint of the transition, was determined from the thermal denaturation curves shown in Fig. 5.
[c] $\Delta T_m = T_m$ determined − 86.8 °C.

**Table 3**
Data collection and refinement statistics of LC-CelA.

| Crystal | LC-CelA |
|---|---|
| Wavelength (Å) | 0.900 |
| Space group | C121 |
| *Cell parameters* | |
| a, b, c (Å) | 130.78, 59.58, 74.95 |
| α = γ, β (°) | 90.00, 122.89 |
| Molecules/asymmetric unit | 2 |
| Resolution range (Å) | 50.00–1.85 (1.88–1.85)[a] |
| Reflections measured | 297,834 |
| Unique reflections | 41,281 |
| Redundancy | 7.2 (7.4) |
| Completeness (%) | 99.4 (99.6)[a] |
| $R_{merge}$ (%)[b] | 12.0 (37.1)[a] |
| Average $I/\sigma$ ($I$) | 21.8 (5.6)[a] |
| *Refinement statistics* | |
| Resolution limits (Å) | 62.9–1.9 |
| No. of atoms | |
| Protein/water | 1798/507 |
| $R_{work}$ (%)/$R_{free}$ (%)[c] | 16.2/20.5 |
| *Rms deviations from ideal values* | |
| Bond lengths (Å) | 0.012 |
| Bond angles (°) | 1.200 |
| *Average B factors* (Å2) | |
| Protein | 15.2 |
| Water | 27.2 |
| *Ramachandran plot statistics* | |
| Preferred regions (%) | 96.9 |
| Allowed regions (%) | 3.1 |
| Molprobity score[d] | 1.35 |

[a] Values in parentheses are for the highest-resolution shell.
[b] $R_{merge} = \sum|I_{hkl} - \langle I_{hkl}\rangle|/\sum I_{hkl}$, where $I_{hkl}$ is an intensity measurement for reflection with indices *hkl* and $\langle I_{hkl}\rangle$ is the mean intensity for multiply recorded reflections.
[c] Free *R*-value was calculated using 5% of the total reflections chosen randomly and omitted from refinement.
[d] MolProbity score combines the clashscore, rotamer, and Ramachandran evaluations into a single score, normalized to be on the same scale as X-ray resolution [52].

probably also formed in that enzyme. We used the terminology of FL for the region predicted as FL (Leu20-Pro35), instead of the actual one (Leu20-Glu32), in this study.

### 3.10. Stability of ΔFL-LC-CelA-His and E34A-LC-CelA-His

To examine whether hydrogen bonding interactions between the FL and central regions contribute to the stabilization of LC-CelA, ΔFL-LC-CelA-His and E34A-LC-CelA-His were constructed. ΔFL-LC-CelA-His and E34A-LC-CelA-His are the LC-CelA-His derivatives without the FL and with a single Glu34→Ala mutation, respectively. These proteins were purified to give a single band on SDS–PAGE as was LC-CelA-His by heat treatment, followed by the nickel affinity chromatography. The amounts of these proteins purified from 1 L culture were similar to that of LC-CelA or LC-CelA-His (3 mg). The far-UV CD spectra of ΔFL-LC-CelA-His and E34A-LC-CelA-His were similar to that of LC-CelA-His (Fig. 7), suggesting that neither removal of the FL region nor a single mutation in the FL significantly affects the gross structure of LC-CelA.

The stability of ΔFL-LC-CelA-His and E34A-LC-CelA-His with respect to heat inactivation was analyzed at pH 7.0 by incubating these proteins at various temperatures for 30 min and determining their residual activities at 60 °C. The results are shown in Fig. 4. Unlike LC-CelA-His, which retains maximal activity after incubation at 90 °C, ΔFL-LC-CelA-His lost approximately 60% and 100% of its activity after incubation at 85 °C and 90 °C respectively. Similarly, E34A-LC-CelA-His lost approximately 20 and 60% of its activity after incubation at 85 °C and 90 °C respectively. These results indicate that the thermal stability of the proteins decreases in the order LC-CelA-His > E34A-LC-CelA-His > ΔFL-LC-CelA-His.

The stability of ΔFL-LC-CelA-His and E34A-LC-CelA-His was also analyzed at pH 7.0 in the presence of 3 M GdnHCl using CD spectroscopy as described above for LC-CelA-His. Thermal denaturation of the proteins was reversible in this condition. The thermal denaturation curves of ΔFL-LC-CelA-His and E34A-LC-CelA-His are shown in Fig. 5. The midpoints of the transitions of these curves, $T_m$, are summarized in Table 2. The $T_m$ values of ΔFL-LC-CelA-His and E34A-LC-CelA-His are lower than that of LC-CelA-His by 14.7 °C and 12.0 °C, respectively, indicating that the FL region of LC-CelA contributes to protein stability (as does the FL of *Rm*Cel12A). The difference in $T_m$ values between LC-CelA-His and E34A-LC-CelA-His accounts for 82% of that between LC-CelA-His and ΔFL-LC-CelA-His. This result suggests that Glu34-mediated hydrogen bonding interactions are the major forces responsible for the FL-dependent stabilization of LC-CelA. Hydrogen bonds have been reported to make large contributions to protein stability [36]. The actual FL region (Leu20-Glu32) may not significantly contribute to the stabilization of LC-CelA.

### 3.11. Stability of LC-CelA-His in the presence of DTT

Proteins are usually [7,37–40], but not always [41,42], destabilized by removal of a native disulfide bond. To examine whether the two disulfide bonds of LC-CelA contribute to protein stability, thermal denaturation of LC-CelA-His was analyzed as described above, at pH 7.0 in the presence of 1 mM dithiothreitol (DTT), which reduces disulfide bonds to thiol groups, and 3 M GdnHCl. Thermal denaturation of LC-CelA-His was reversible in this condition. The thermal denaturation curve of LC-CelA-His measured in the presence of 1 mM DTT and 3 M GdnHCl is shown in Fig. 5. The $T_m$ value of LC-CelA-His determined in the presence of 1 mM DTT is lower than that determined in the absence of DTT by 28.0 °C (Table 2). This result indicates that disulfide bonds contribute significantly to the stabilization of LC-CelA-His. The $T_m$ value of LC-CelA-His (59.8 °C) determined in the presence of 10 mM DTT and 3 M GdnHCl was comparable to that (58.8 °C) determined in the presence of 1 mM DTT and 3 M GdnHCl, suggesting that the disulfide bonds of LC-CelA-His are fully reduced by 1 mM DTT.

It remains to be determined which disulfide bond contributes more to the stabilization of LC-CelA. However, it has been reported that the contribution of disulfide bonds to protein stability increases as the number of the residues in the loop formed by the disulfide bond increases [37]. That number is 26 for Cys40–Cys67 and four for Cys100–Cys105. Therefore, the Cys40–Cys67 bond may contribute significantly more to the stabilization of LC-CelA than the Cys100–Cys105 bond.

Disulfide bonds do not account for the differences in stability between LC-CelA or *Rm*Cel12A and a GH family 12 cellulase from *Streptomyces* species, because the two disulfide bonds of LC-CelA are conserved in these proteins. The $T_m$ values of *Rm*Cel12A and Cel12A from *Streptomyces* sp. 11AG8 (*St*Cel12A), which represents GH family 12 cellulases from *Streptomyces* species, have been reported to be 102.9 °C [35] and 65.7 °C [43] respectively in the absence of denaturant, indicating that *Rm*Cel12A is more stable than *St*Cel12A by 37.2 °C. Because LC-CelA is nearly as stable as *Rm*Cel12A, LC-CelA is also more stable than *St*Cel12A. Proteins are stabilized relative to each other by a combination of factors, including increased number of ion pairs, increased number of hydrogen bonds, increased number of disulfide bonds, increased number of proline residues in loop regions, and increased interior hydrophobicity [36]. Therefore, we compared these structural factors in LC-CelA, *Rm*Cel12A and *St*Cel12A (PDB: 1OA4). The ratio of interior apolar residues to total interior residues in LC-CelA (66.7%) is comparable to that in *Rm*Cel12A (65.2%), but significantly higher

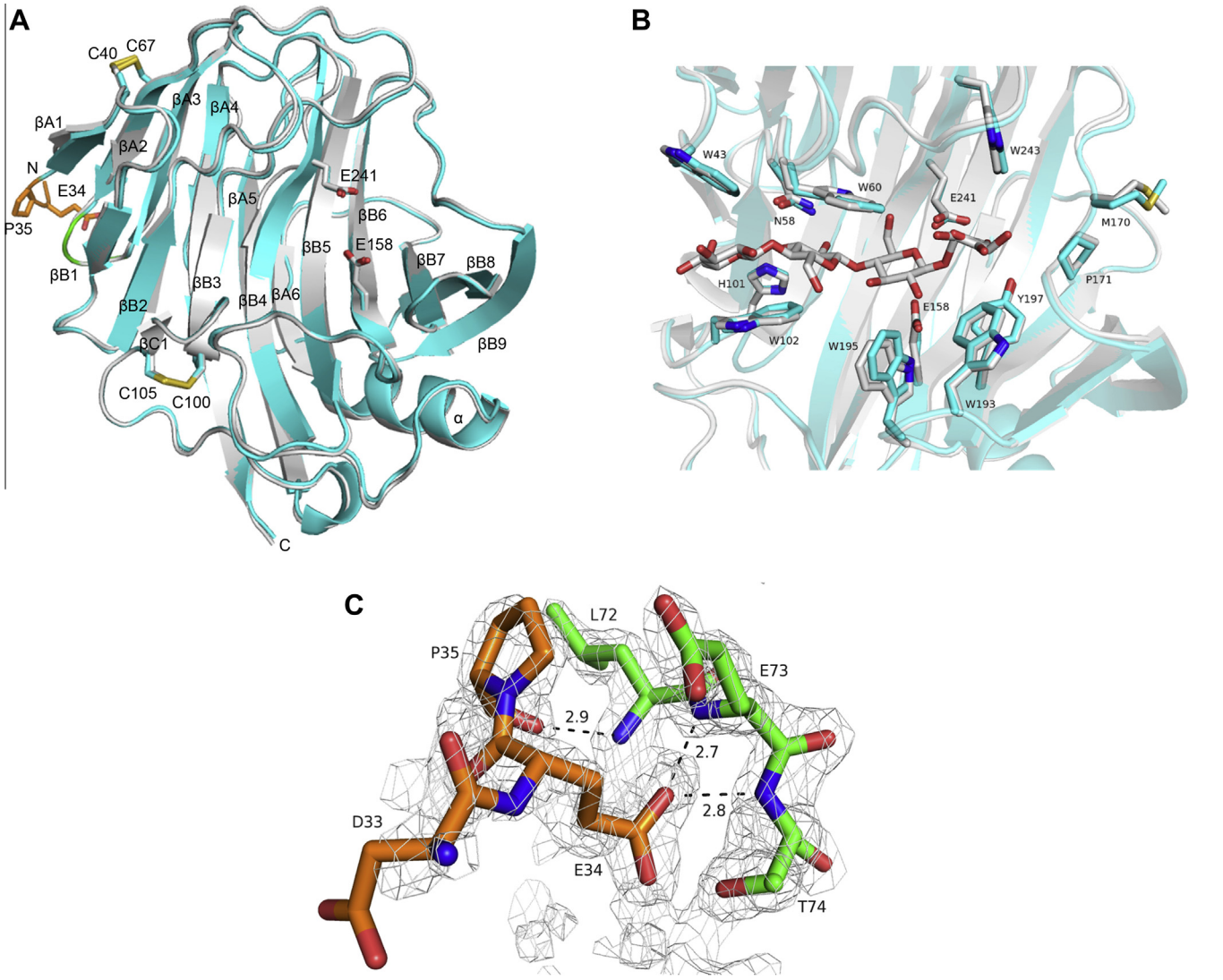


**Fig. 6.** Crystal structure of LC-CelA. (A and B) The entire structure (A) and the structure around the active site (B) of LC-CelA are superimposed on those of *Rm*Cel12A in complex with the substrate ($_D$(+)-cellopentose) (PDB: 2BWC). The structures of LC-CelA and *Rm*Cel12A are colored cyan and gray respectively. The region predicted as a flexible linker (FL) of LC-CelA is colored orange. The central region of LC-CelA that contacts FL is colored green. Two acidic active site residues (Glu158 and Glu241), two disulfide bonds (Cys40–Cys67 and Cys100–Cys105), twelve residues forming the substrate binding site of LC-CelA, and C-terminal two residues of FL (Glu34 and Pro35) are shown as stick models with labels. The corresponding residues of *Rm*Cel12A, except for those corresponding to Glu34 and Pro35, and the substrate bound to *Rm*Cel12A are also shown as stick models. In these stick models, the oxygen, nitrogen and sulfur atoms are colored red, blue and yellow respectively. The substrate and the residues forming the substrate binding site are not shown in panel A for simplicity. Six β-strands forming sheet A (βA1–βA6), nine β-strands forming sheet B (βB1–βB9), βC1-strand, and one α-helix are shown in panel A. N and C represent the N- and C-termini. (C) Electron density around the region, where hydrogen bonds are formed between the FL and central regions of LC-CelA. The $2F_o$–$F_c$ map contoured at the 1.5σ level is shown. The side chain and main chain atoms are shown as stick models, where the oxygen and nitrogen atoms are colored red and blue respectively, and the carbon atoms are colored orange and green for the FL and central regions respectively as in panel A. Three hydrogen bonds formed between Glu34 $O^{\delta1}$ and Thr74 N, between Glu34 $O^{\delta1}$ and Glu73 N, and between Pro35 O and Leu72 N are shown by broken lines together with their distances. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

than that in *St*Cel12A (60.8%). Similarly, the number of ion pairs in LC-CelA (10) is comparable to that in *Rm*Cel12A (8), but higher than that in *St*Cel12A (5). However, the number of hydrogen bonds and proline residues in loop regions of LC-CelA and *Rm*Cel12A are not consistently higher than in *St*Cel12A. The number of hydrogen bonds is 202 in LC-CelA, 177 in *Rm*Cel12A, and 197 in *St*Cel12A. The number of proline residues in loop regions is 8 in LC-CelA, 7 in *Rm*Cel12A, and 13 in *St*Cel12A. These results suggest that differences in interior hydrophobicity and in the number of ion pairs at least partially account for the difference in stability between LC-CelA or *Rm*Cel12A and *St*Cel12A.

### 3.12. Activities of ΔFL-LC-CelA-His and E34A-LC-CelA-His

The enzymatic activities of ΔFL-LC-CelA-His and E34A-LC-CelA-His were determined at 60 °C in the absence of DTT using CM-cellulose as a substrate. The activities of these proteins were not determined at the optimum temperature for activity of LC-CelA-His (90 °C), because they are thermally denatured at this temperature (Fig. 4). The specific activities of ΔFL-LC-CelA-His and E34A-LC-CelA-His determined in the absence of DTT were slightly lower than, but comparable to, that of LC-CelA-His (Table 2). These data indicate that the enzymatic activity of LC-CelA-His is not significantly affected by removal of, or mutation in, the FL. The specific

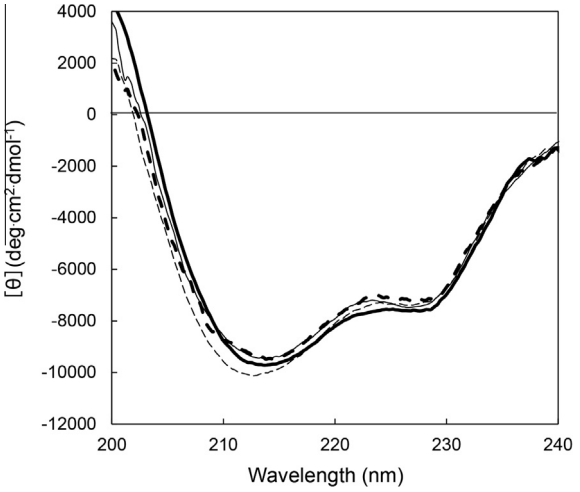


**Fig. 7.** CD spectra of LC-CelA-His and its derivatives. The far-UV CD spectrum of LC-CelA-His was measured at pH 7.0 and 25 °C either in the absence of DTT (thin dashed line) or in the presence of 1 mM DTT (thin solid line), as described in Experimental procedures. The far-UV CD spectra of ΔFL-LC-CelA-His (thick solid line) and E34A-LC-CelA-His (thick dashed line) were also measured at pH 7.0 and 25 °C in the absence of DTT.

activity of LC-CelA determined in the presence of 1 mM DTT was identical to that determined in the absence of DTT, indicating that the enzymatic activity of LC-CelA-His is not significantly affected by reduction of the disulfide bonds.

## 4. Conclusion

LC-CelA, a homolog of the highly thermostable GH family 12 cellulase from *R. marinus* (*Rm*Cel12A), was isolated from leaf–branch compost by a metagenomic approach and its crystal structure was determined. This structure is the first of a GH family 12 cellulase with an acidic flexible linker (FL) at the N-terminus. The structure shows that most of the region predicted as FL is disordered. This result supports the hypothesis that the FL is necessary to separate a hydrophobic signal peptide that anchors the enzyme to the cell from the catalytic core [35]. However, Glu34 and Pro35 located at the C-terminus of the FL assume a defined structure and form hydrogen bonds with the central region of the protein. The structure also shows that LC-CelA contains two disulfide bonds (Cys40–Cys67 and Cys100–Cys105). LC-CelA is a highly thermostable enzyme with an optimum temperature for activity of 90 °C. The stability of LC-CelA decreases by 14.9 °C on removal of the FL, mainly due to loss of hydrogen bonds between the FL and central regions of the protein. Removal of the FL destabilizes *Rm*Cel12A by 8.4 °C [35], probably due to loss of the residues equivalent to Glu34 and Pro35 of LC-CelA, which are conserved. Database searches indicate that these residues are conserved only in LC-CelA and *Rm*Cel12A, suggesting that stabilization by hydrogen bonds formed between the FL and central regions is unique to these proteins. The stability of LC-CelA also decreases by reducing disulfide bonds by 28 °C. The Cys40–Cys67 bond located near the N-terminal region may more significantly contribute to the stabilization of LC-CelA than Cys100–Cys105, because the number of the residues in the loop formed by Cys40–Cys67 is higher than that formed by Cys100–Cys105. It has been reported for GH family 11 xylanase that a unique N-terminal 11-residue extension contributes to protein stability by forming hydrogen bonds with the core region [44]. It has also been reported for this enzyme that mutations in the N-terminal region [45] and engineering of an N-terminal disulfide bond [46–49] stabilize the protein. GH family 12 cellulases with improved activity and stability have been engineered [50,51]. However, studies of protein engineering of GH family 12 cellulases are still limited. The results presented in this work provide valuable information that serves as a structural toolbox for the engineering of GH family 12 cellulases with improved function.

## Acknowledgments

The synchrotron radiation experiments were performed at Osaka University beam line BL44XU at SPring-8 with the approval of the Japan Synchrotron Radiation Research Institute (JASRI) (Proposal number 2013B6813). We thank Commemorative Organization for the Japan World Exposition '70 for donation of leaf–branch compost. This work was supported in part by a Grant (24380055) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

## References

[1] Steele, H.L., Jaeger, K.E., Daniel, R. and Streit, W.R. (2009) Advances in recovery of novel biocatalysts from metagenomes. J. Mol. Microbiol. Biotechnol. 16, 25–37.
[2] Uchiyama, T. and Miyazaki, K. (2009) Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr. Opin. Biotechnol. 20, 616–622.
[3] Duan, C.J. and Feng, J.X. (2010) Mining metagenomes for novel cellulase genes. Biotechnol. Lett. 32, 1765–1775.
[4] Iqbal, H.A., Feng, Z. and Brady, S.F. (2012) Biocatalysts and small molecule products from metagenomic studies. Curr. Opin. Chem. Biol. 16, 109–116.
[5] Sulaiman, S., Yamato, S., Kanaya, E., Kim, J.J., Koga, Y., Takano, K. and Kanaya, S. (2011) Isolation of a novel cutinase homolog with polyethylene terephthalate degrading activity from leaf–branch compost using a metagenomic approach. Appl. Environ. Microbiol. 78, 1556–1562.
[6] Kanaya, E., Sakabe, T., Nguyen, N.-T., Koikeda, S., Koga, Y., Takano, K. and Kanaya, S. (2010) Cloning of the RNase H genes from a metagenomic DNA library: identification of a new type 1 RNase H without a typical active-site motif. J. Appl. Microbiol. 109, 974–983.
[7] Sulaiman, S., You, D.-J., Kanaya, E., Koga, Y. and Kanaya, S. (2014) Crystal structure and thermodynamic and kinetic stability of metagenome-derived LC-cutinase. Biochemistry 53, 1858–1869.
[8] Nguyen, T.N., You, D.-J., Kanaya, E., Koga, Y. and Kanaya, S. (2013) Crystal structure of metagenome-derived LC9-RNase H1 with atypical DEDN active site motif. FEBS Lett. 587, 1418–1423.
[9] Nguyen, T.N., Angkawidjaja, C., Kanaya, E., Koga, Y., Takano, K. and Kanaya, S. (2012) Activity, stability, and structure of metagenome-derived LC11-RNase H1, a homolog of *Sulfolobus tokodaii* RNase H1. Protein Sci. 21, 553–561.
[10] Nguyen, T.N., You, D.-J., Matsumoto, H., Kanaya, E., Koga, Y. and Kanaya, S. (2013) Crystal structure of metagenome-derived LC11-RNase H1 in complex with RNA/DNA hybrid. J. Struct. Biol. 182, 144–154.
[11] Wilson, D.B. (2009) Cellulases and biofuels. Curr. Opin. Biotechnol. 20, 295–299.
[12] Lynd, L.R., van Zyl, W.H., McBride, J.E. and Laser, M. (2005) Consolidated bioprocessing of cellulosic biomass: an update. Curr. Opin. Biotechnol. 16, 577–583.
[13] la Grange, D.C., den Haan, R. and van Zyl, W.H. (2010) Engineering cellulolytic ability into bioprocessing organisms. Appl. Microbiol. Biotechnol. 87, 1195–1208.
[14] Yanase, S., Yamada, R., Kaneko, S., Noda, H., Hasunuma, T., Tanaka, T., Ogino, C., Fukuda, H. and Kondo, A. (2010) Ethanol production from cellulosic materials using cellulase-expressing yeast. Biotechnol. J. 5, 449–455.
[15] Yamada, R., Nakatani, Y., Ogino, C. and Kondo, A. (2013) Efficient direct ethanol production from cellulose by cellulase- and cellodextrin transporter-co-expressing *Saccharomyces cerevisiae*. AMB Express 3, 34.
[16] Hu, J., Arantes, V., Pribowo, A., Gourlay, K. and Saddler, J.N. (2014) Substrate factors that influence the synergistic interaction of AA9 and cellulases during the enzymatic hydrolysis of biomass. Energy Environ. Sci. 7, 2308–2315.
[17] Lynd, L.R., Weimer, P.J., van Zyl, W.H. and Pretorius, I.S. (2002) Microbial cellulose utilization: fundamentals and biotechnology. Microbiol. Mol. Biol. Rev. 66, 506–577.
[18] Halldorsdottir, S., Thorolfsdottir, E.T., Spilliaert, R., Johansson, M., Thorbijarnardottir, S.H., Palsdottir, A., Hreggvidsson, G.O., Kristjansson, J.K., Holst, O. and Eggertsson, G. (1998) Cloning, sequencing and overexpression of a *Rhodothermus marinus* gene encoding a thermostable cellulase of glycosyl hydrolase Family 12. Appl. Microbiol. Biotechnol. 49, 277–284.
[19] Crennell, S.J., Hreggvidsson, G.O. and Nordberg Karlsson, E. (2002) The structure of *Rhodothermus marinus* Cel12A, a highly thermostable family 12 endoglucanase, at 1.8 Å resolution. J. Mol. Biol. 320, 883–897.
[20] Crennell, S.J., Cook, D., Minns, A., Svergun, D., Andersen, R.L. and Nordberg Karlsson, E. (2006) Dimerisation and an increase in active site aromatic groups as adaptations to high temperatures: X-ray solution scattering and substrate-

bound crystal structures of *Rhodothermus marinus* endoglucanase Cel12A. J. Mol. Biol. 356, 57–71.
[21] Bruins, M.E., Janssen, A.E. and Boom, R.M. (2001) Thermozymes and their applications: a review of recent literature and patents. Appl. Biochem. Biotechnol. 90, 155–186.
[22] Haki, G.D. and Rakshit, S.K. (2003) Developments in industrially important thermostable enzymes: a review. Bioresour. Technol. 89, 17–34.
[23] Unsworth, L.D., van der Oost, J. and Koutsopoulos, S. (2007) Hyperthermophilic enzymes – stability, activity and implementation strategies for high temperature applications. FEBS J. 274, 4044–4056.
[24] Jo, W.S., Bae, S.H., Choi, S.Y., Park, S.D., Yoo, Y.B. and Park, S.C. (2010) Development of detection methods for cellulolytic activity of *Auricularia auricula-judae*. Mycobiology 38, 74–77.
[25] Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680–685.
[26] Goodwin, T.W. and Morton, R.A. (1946) The spectrophotometric determination of tyrosine and tryptophan in proteins. Biochem. J. 40, 628–632.
[27] Miller, G.L., Blum, R., Glennon, W.E. and Burton, A.L. (1960) Measurement of carboxymethylcellulase activity. Anal. Biochem. 2, 127–132.
[28] Otwinowski, Z. and Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. 276, 307–326.
[29] Vagin, A. and Teplyakov, A. (2010) Molecular replacement with MOLREP. Acta Crystallogr. D Biol. Crystallogr. 66, 22–25.
[30] Collaborative Computational Project, Number 4 (1994) The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D Biol. Crystallogr. 50, 760–763.
[31] Langer, G., Cohen, S.X., Lamzin, V.S. and Perrakis, A. (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. Nat. Protoc. 3, 1171–1179.
[32] Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D Biol. Crystallogr. 53, 240–255.
[33] Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126–2132.
[34] Gilkes, N.R., Henrlssat, B., Kilburn, D.G., Miller Jr., R.C. and Warren, R.A.J. (1991) Domains in microbial β-l,4-glycanases: sequence conservation, function, and enzyme families. Microbiol. Rev. 55, 303–315.
[35] Wicher, K.B., Abou-Hachem, M., Halldórsdóttir, S., Thorbjarnadóttir, S.H., Eggertsson, G., Hreggvidsson, G.O., Nordberg Karlsson, E. and Holst, O. (2001) Deletion of a cytotoxic, N-terminal putative signal peptide results in a significant increase in production yields in *Escherichia coli* and improved specific activity of Cel12A from *Rhodothermus marinus*. Appl. Microbiol. Biotechnol. 55, 578–584.
[36] Pace, C.N., Scholtz, J.M. and Grimsley, G.R. (2014) Forces stabilizing proteins. FEBS Lett. 588, 2177–2184.
[37] Pace, C.N., Grimsley, G.R., Thomson, J.A. and Barnett, B.J. (1988) Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. J. Biol. Chem. 263, 11820–11825.
[38] Piatek, R., Bruździak, P., Wojciechowski, M., Zalewska-Piatek, B.M. and Kur, J. (2010) The noncanonical disulfide bond as the important stabilizing element of the immunoglobulin fold of the Dr fimbrial DraE subunit. Biochemistry 49, 1460–1468.
[39] Schulenburg, C., Weininger, U., Neumann, P., Meiselbach, H., Stubbs, M.T., Sticht, H., Balbach, J., Ulbrich-Hofmann, R. and Arnold, U. (2010) Impact of the C-terminal disulfide bond on the folding and stability of onconase. ChemBioChem 11, 978–986.
[40] Mason, J.M., Bendall, D.S., Howe, C.J. and Worrall, J.A.R. (2012) The role of a disulfide bridge in the stability and folding kinetics of *Arabidopsis thaliana* cytochrome c6A. Biochim. Biophys. Acta 1824, 311–318.
[41] Fernandes, A.T., Pereira, M.M., Silva, C.S., Lindley, P.F., Bento, I., Melo, E.P. and Martins, L.O. (2011) The removal of a disulfide bridge in CotA-laccase changes the slower motion dynamics involved in copper binding but has no effect on the thermodynamic stability. J. Biol. Inorg. Chem. 16, 641–651.
[42] Schmidt, B., Ho, L. and Hogg, P.J. (2006) Allosteric disulfide bonds. Biochemistry 45, 7429–7433.
[43] Sandgren, M., Gualfetti, P.J., Shaw, A., Gross, L.S., Saldajeno, M., Day, A.G., Jones, T.A. and Mitchinson, C. (2003) Comparison of family 12 glycoside hydrolases and recruited substitutions important for thermal stability. Protein Sci. 12, 848–860.
[44] Cheng, Y.S., Chen, C.C., Huang, C.H., Ko, T.P., Luo, W., Huang, J.W., Liu, J.R. and Guo, R.T. (2014) Structural analysis of a glycoside hydrolase family 11 xylanase from *Neocallimastix patriciarum*: insights into the molecular basis of a thermophilic enzyme. J. Biol. Chem. 289, 11020–11028.
[45] Dumon, C., Varvak, A., Wall, M.A., Flint, J.E., Lewis, R.J., Lakey, J.H., Morland, C., Luginbühl, P., Healey, S., Todaro, T., DeSantis, G., Sun, M., Parra-Gessert, L., Tan, X., Weiner, D.P. and Gilbert, H.J. (2008) Engineering hyperthermostability into a GH11 xylanase is mediated by subtle changes to protein structure. J. Biol. Chem. 283, 22557–22564.
[46] Fenel, F., Leisola, M., Jänis, J. and Turunen, O. (2004) A de novo designed N-terminal disulphide bridge stabilizes the *Trichoderma reesei* endo-1,4-β-xylanase II. J. Biotechnol. 108, 137–143.
[47] Jänis, J., Turunen, O., Leisola, M., Derrick, P.J., Rouvinen, J. and Vainiotalo, P. (2004) Characterization of mutant xylanases using fourier transform ion cyclotron resonance mass spectrometry: stabilizing contributions of disulfide bridges and N-terminal extensions. Biochemistry 43, 9556–9566.
[48] Wang, Y., Fu, Z., Huang, H., Zhang, H., Yao, B., Xiong, H. and Turunen, O. (2012) Improved thermal performance of *Thermomyces lanuginosus* GH11 xylanase by engineering of an N-terminal disulfide bridge. Bioresour. Technol. 112, 275–279.
[49] Li, H., Kankaanpää, A., Xiong, H., Hummel, M., Sixta, H., Ojamo, H. and Turunen, O. (2013) Thermostabilization of extremophilic *Dictyoglomus thermophilum* GH11 xylanase by an N-terminal disulfide bridge and the effect of ionic liquid [emim] OAc on the enzymatic performance. Enzyme Microb. Technol. 53, 414–419.
[50] Tishkov, V.I., Gusakov, A.V., Cherkashina, A.S. and Sinitsyn, A.P. (2013) Engineering the pH-optimum of activity of the GH12 family endoglucanase by site-directed mutagenesis. Biochimie 95, 1704–1710.
[51] Sandgren, M., Ståhlberg, J. and Mitchinson, C. (2005) Structural and biochemical studies of GH family 12 cellulases: improved thermal stability, and ligand complexes. Prog. Biophys. Mol. Biol. 89, 246–291.
[52] Chen, V.B., Arendall 3rd, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. 66, 12–21.