


RESEARCH

Open Access



# In silico characterization of the GH5-cellulase family from uncultured microorganisms: physicochemical and structural studies

Rahmat Eko Sanjaya<sup>1,2,3</sup>, Kartika Dwi Asni Putri<sup>2</sup>, Anita Kurniati<sup>1,2,4</sup>, Ali Rohman<sup>2,5</sup> and Ni Nyoman Tri Puspaningsih<sup>2,5\*</sup> 

## Abstract

**Background:** Hydrolysis of cellulose-based biomass by cellulases produce fermented sugar for making biofuels, such as bioethanol. Cellulases hydrolyze the  $\beta$ -1,4-glycosidic linkage of cellulose and can be obtained from cultured and uncultured microorganisms. Uncultured microorganisms are a source for exploring novel cellulase genes through the metagenomic approach. Metagenomics concerns the extraction, cloning, and analysis of the entire genetic complement of a habitat without cultivating microbes. The glycoside hydrolase 5 family (GH5) is a cellulase family, as the largest group of glycoside hydrolases. Numerous variants of GH5-cellulase family have been identified through the metagenomic approach, including CelGH5 in this study. University-CoE-Research Center for Biomolecule Engineering, Universitas Airlangga successfully isolated CelGH5 from waste decomposition of oil palm empty fruit bunches (OPEFB) soil by metagenomics approach. The properties and structural characteristics of GH5-cellulases from uncultured microorganisms can be studied using computational tools and software.

**Results:** The GH5-cellulase family from uncultured microorganisms was characterized using standard computational-based tools. The amino acid sequences and 3D-protein structures were retrieved from the GenBank Database and Protein Data Bank. The physicochemical analysis revealed the sequence length was roughly 332–751 amino acids, with the molecular weight range around 37–83 kDa, dominantly negative charges with pI values below 7. Alanine was the most abundant amino acid making up the GH5-cellulase family and the percentage of hydrophobic amino acids was more than hydrophilic. Interestingly, ten endopeptidases with the highest average number of cleavage sites were found. Another uniqueness demonstrated that there was also a difference in stability between in silico and wet lab. The *I* values indicated CelGH5 and ACA61162.1 as unstable enzymes, while the wet lab showed they were stable at broad pH range. The program of SOPMA, PDBsum, ProSA, and SAVES provided the secondary and tertiary structure analysis. The predominant secondary structure was the random coil, and tertiary structure has fulfilled the structure quality of QMEAN4, ERRAT, Ramachandran plot, and Z score.

\* Correspondence: [ni-nyoman-t-p@fst.unair.ac.id](mailto:ni-nyoman-t-p@fst.unair.ac.id)

<sup>2</sup>University-CoE-Research Centre for Bio-Molecule Engineering, 2nd Floor ITD Building, Kampus C Universitas Airlangga, Mulyorejo, Surabaya, East Java 60115, Indonesia

<sup>5</sup>Department of Chemistry, Faculty of Science and Technology, Kampus C Universitas Airlangga, Mulyorejo, Surabaya, East Java 60115, Indonesia  
Full list of author information is available at the end of the article

**Conclusion:** This study can afford the new insights about the physicochemical and structural properties of the GH5-cellulase family from uncultured microorganisms. Furthermore, *in silico* analysis could be valuable in selecting a highly efficient cellulases for enhanced enzyme production.

**Keywords:** Biofuel, Cellulase, Glycoside hydrolase 5 family, CelGH5, Uncultured microorganism, Computational tool, GenBank Database, Protein Data Bank

## Background

Cellulases are a group of enzymes that have the ability to hydrolyze cellulose polymers into glucose monomers by hydrolyzing the  $\beta$ -(1  $\rightarrow$  4) glycosidic bonds. Cellulases consist of three main enzymes: endo- $\beta$ -1,4-glucanase (EC 3.2.1.4),  $\beta$ -glucosidase (EC 3.2.1.21), and exoglucanase. Exoglucanase consists of cellobiohydrolase I (EC 3.2.1.176) and cellobiohydrolase II (EC 3.2.1.91). Cellulases are classified into the carbohydrate acting enzymes (CAZy) in the group of the glycoside hydrolase (GH) [1]. Glycoside hydrolase (EC 3.2.1.-) is a well-known enzyme that hydrolyzes the glycosidic bond between two or more carbohydrates or between a carbohydrate and a non-carbohydrate moiety (<http://www.cazy.org/Glycoside-Hydrolases.html>). The grouping of enzymes into GH was based on conserved amino acid sequences and classified into several families [2–4]. Enzymes that are in the same family have similar amino acid sequences and three-dimensional structures. The GH5 family is the cellulase family, has at least 56 subfamilies, the largest glycoside hydrolase family [5]. Most of the GH5 members are multi-modular, including a catalytic module, substrate-binding module, and unidentified.

Cellulose is the most abundant biopolymer on Earth and is found in plant cell walls. It is a linear polysaccharide of glucose linked by  $\beta$ -1,4-glycosidic bonds. Cellulose is the main load-bearing polysaccharide, consisting of long chains of glucose strongly packed together due to H-bonds. It is embedded in a matrix of lignin, hemicelluloses, and pectin [6]. In addition to being highly abundant in plants, cellulose is also synthesized by some bacterial strains, such as *Acetobacter*, *Rhizobium*, *Xanthococcus*, *Pseudomonas*, *Azotobacter*, *Aerobacter*, and *Alcaligenes* [7]. Cellulose produced by bacterial strains is known as bacterial cellulose (BC). Animals (tunicates), algae, and protists can also produce cellulose [8]. As such, cellulose is the main target for renewable fuel production, such as bioethanol. The production of biofuel from renewable materials can provide economic and environmental benefits [9, 10]. However, bioethanol production using cellulosic materials requires high temperatures and harsh conditions [11, 12]. Hydrolysis of cellulosic materials and the saccharification process for bioethanol production enzymatically requires cellulase as it can perform under harsh conditions, such as high temperatures, high salinity, broad pH ranges, and stable in the presence of organic compounds [13–16].

Cellulases can be obtained from cultured and uncultured microorganisms. Cellulases from cultured microorganisms are defined as cellulases isolated by the cultivation of microorganisms under laboratory conditions. Cellulases produced from cultured microorganisms known as microbial cellulases [17, 18]. Cellulases were produced by microorganisms, such as *Aspergillus flavus* [19], *Bacillus* sp. [20], and other species of bacteria, fungi, and actinomycetes [16]. In contrast, the cultivation-independent (uncultured) technique is constrained by the fact that the majority of microorganisms, particularly those found in soil, cannot be cultivated in the laboratory [21]. Notably, much information is held within the genomes of uncultured microorganisms, and metagenomic technologies can investigate this potential [22]. Metagenomics is a method of analyzing and collecting functional genes from uncultured microorganisms or without the cultivation of microorganisms. It is an emerging approach to studying microbial communities in the environment [23]. Uncultured microorganisms represent a significant part of natural biodiversity. Microorganisms that can be cultured by standardized laboratory techniques comprise only 0.1–1% of the natural ecosystem [24–26]. Genes constructed based on metagenomic approaches have shown to be effective in identifying novel genes with specific activities [27–30].

The metagenomics-derived cellulases exhibit various characteristics and have commercial applications. Several members of the GH5-cellulase family have been identified using metagenomic approaches [28, 31, 32]. For example, a novel cellulase with unusual catalytic properties was isolated and characterized from a sugarcane soil metagenome (CelE1) [29] and CelGH5 from waste decomposition of oil palm empty fruit bunches (OPEFB) soil. CelE1 showed optimal activity at pH 7.0 and 50 °C with remarkable activity at alkaline conditions. Interestingly, CelE1 has a relative activity of 60% after incubation at 70 °C and has a higher activity at low temperatures (10–50 °C). This indicates that CelE1 is a thermotolerant enzyme with relative catalytic activity (> 65%) in the 10–70 °C temperature range. CelGH5 catalytic activity increased twofold after 4.0 M NaCl addition at pH 7.0, 55 °C. This indicates that CelGH5 is a halophilic with relative catalytic activity > 200% (unpublished data). Other cellulases from the metagenomic approach have unique properties, including cellulases from soil [27] and enriched culture from a hot

spring [33], with hydrolytic activity increasing and stable in the presence of salt.

Understanding the properties and characteristics of cellulases can be achieved through their amino acid sequences and 3D structures. Therefore, predictions of cellulase properties can be considered an initial reference in developing the properties and characteristics of cellulases in the future. Most researchers' current focus has been on the large-scale production of industrial enzymes for industrial purposes using multiple functional genes cloned on expression hosts. However, numerous variations—molecular weight, stability, amino acid composition, family, and secondary and tertiary structures—have been observed between different recombinant proteins produced from functional genes [34]. The availability of software and internet tools can be used to understand the overall physicochemical characteristics (i.e., primary, secondary, and tertiary structures, functional analysis, domains and motifs, and phylogenetic analysis) of the GH5-cellulase family from uncultured microorganisms. To date, no research has been conducted relating to the *in silico* analysis of the GH5 cellulase family from uncultured microorganisms. Only cellulases from *Bacillus* [34] and *Ruminococcus albus* [35] have been reported; these were analyzed using the *in silico* approach. Moreover, this information about the GH5 cellulase family from uncultured microorganisms retrieved from various tools and databases could be valuable in selecting a highly efficient strain for enhanced commercial enzyme production. The present study aimed to utilize *in silico* tools for the physicochemical and structural characterization of the GH5-cellulase family from uncultured microorganisms.

## Methods

### Sequence retrieval

Cellulase amino acid sequences from uncultured microorganisms were taken from GenBank, NCBI (<http://www.ncbi.nlm.nih.gov/>) based on the CAZy database belonging to the glycoside hydrolase family 5 on 2 September 2020. The sequences were kept in FASTA format, and unspecified or truncated sequences were removed. After reducing the data using the CD-HIT program ([http://weizhong-lab.ucsd.edu/cdhit\\_suite/cgi-bin/index.cgi?cmd=cd-hit](http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit)), 26 cellulases of GH5 family sequences were discovered. In addition, a sequence with an identity of CelGH5 was retrieved from University-CoE-Research Centre for Bio-Molecule Engineering (BioME), Universitas Airlangga, Surabaya, Indonesia, on 2 September 2020. CelGH5 sequence was obtained using the metagenomic approach from compost soil of palm oil waste and was also used in this study. Thus, a total of 27 different cellulase sequences were used in this study.

### Physicochemical properties

The physicochemical properties: molecular weight, theoretical pI, instability index, aliphatic index, and GRAVY were analyzed using ExPASy-ProtParam tools (<https://web.expasy.org/protparam/>) [36].

### Stability analysis

Stability analysis was done to CelGH5 for supporting the *in silico* data on the physicochemical properties. The pH and an additive stability assay for CelGH5 was carried out using the ThermoFluor assay. Protein melting temperature ( $T_m$ ) was determined by monitoring protein unfolding with the fluoroprobe, which emits fluorescence that can be quantified as a function of temperature when bound to hydrophobic protein domains [37]. The ThermoFluor assay was performed on a real-time PCR (RT-PCR) instrument (IT-IS Life Science Ltd., Ireland). Solutions of 2.5  $\mu$ l of 80X SYPRO<sup>TM</sup> Orange (Thermo Fisher, USA), 2.5  $\mu$ l of 10 mg/ml CelGH5 enzyme, and 45  $\mu$ l of test compound (buffer and additives) were added to the real-time PCR tube (GenFollower, China). Buffer test using a buffer screen of Britton-Robinson (BR) buffers [1:1:1 acetic acid:H<sub>3</sub>PO<sub>4</sub>:boric acid] ranging from pH 2.0–12.0 and protein buffer [50 mM phosphate pH 8.0, 250 mM NaCl, 5 mM Imidazole] as control. Additive test using 13 additives with water as control (Fig. 2). Samples were heated in real-time PCR from 37 °C to 97 °C in increments of 0.025 °C/s with initial and final holds were 10 s. The changes of the fluorescence were recorded every 0.025 °C using a fluorescence detector.

### Primary structure analysis

Amino acid composition, hydrophilic, and hydrophobic residues were calculated from the primary structure using the CLC main workbench 8.1.2 software (QIAGEN) [38]. The motifs or sequence consensus were identified using Multiple EM for Motif Elicitation (MEME) server (<http://meme-suite.org/tools/meme>) [39]. The maximum number of motifs was set as 6. It used a maximum width of 50 amino acids and a minimum width of 6 amino acids set along with other factors as default values.

### Secondary structure analysis

The secondary structure was obtained using Self-Optimized Prediction Method with Alignment (SOPMA) tool. The results obtained were the percentage composition of  $\alpha$ -helix,  $\beta$ -sheet, turns, and random coil ([https://npsa-prabi.ibcp.fr/cgi-bin/npsa\\_automat.pl?page=/NPSA/npsa\\_sopma.html](https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html)) [40]. In order to confirm the predicted secondary structure, pictorial overviews of some experimental cellulase structures were retrieved from PDB RCSB, and the secondary structure was generated using PDBsum (<http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html>). Additionally, information of its Ramachandran

plots was generated by the PROCHECK tool on PDBsum.

### Tertiary structure analysis

The tertiary structures of four cellulases from uncultured microorganisms belonging to the GH5 family were determined. Structures with PDB ID 4EE9, 4HTY, 4M1R, and 5I2U were retrieved from PDB RCSB (<https://www.rcsb.org/>) on 2 September 2020, and their tertiary structures were further analyzed. QMEAN scores (<https://swissmodel.expasy.org/qmean/>) and ERRAT values (<https://saves.mbi.ucla.edu/>) were used to validate and evaluate the 3D structures. QMEAN4 was used to fit cumulative QMEAN values on a global scale at a range of 0 to 1 [41, 42]. ERRAT values were related to the resolution of protein structure. An average overall quality factor from ERRAT values around or higher 95% represents the high resolution of the structures, and the lower resolutions (2.5 to 3 Å) were approximately 91% [43, 44]. ProSA-web was used to assess the Z score and energy plots (<https://prosa.services.came.sbg.ac.at/prosa.php>). The desirable Z score should be < 1 compared to a nonredundant set of PDB structures [42, 45].

### Functional analysis

In order to determine the functional linkage and protein stability, the presence and absence of cysteine bonds (disulfide bonds) and their bonding pattern were predicted by CYS\_REC (Softberry, Inc.) [46] ([http://www.softberry.com/berry.phtml?topic=cys\\_rec&group=programs&subgroup=propt](http://www.softberry.com/berry.phtml?topic=cys_rec&group=programs&subgroup=propt)). The protein sequences of cellulase were analyzed by a conserved domain database (CDD) to determine conserved domains (<https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>) [47]. Potential cleavage sites were identified by using The Peptide Cutter tool ([https://web.expasy.org/peptide\\_cutter/](https://web.expasy.org/peptide_cutter/)). The Peptide Cutter predicts potential cleavage sites cleaved by proteases or chemicals in a given protein sequence [48].

### Multiple sequence alignment and phylogenetic tree construction

The alignments of the amino acid sequences of cellulases were created using Clustal Omega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) [49–51] and generated by ESPript 3.0 program [52]. Cladograms of the GH5-cellulase family sequences from uncultured microorganisms were constructed through a maximum likelihood method based on the JTT matrix model [53] using the MEGA X software [54].

## Results

### Sequence retrieval

Twenty-six amino acid sequences were obtained from GenBank, and one sequence from our collection (Table

1) was added. Amino acid sequences were downloaded in FASTA format and used to analyze the physicochemical characteristics, primary and secondary structure, functional analyses, domains and motifs, and phylogenetic analyses.

### Physicochemical properties

Physicochemical properties of a protein, like molecular weight, pI, instability index, aliphatic index, and the average of hydrophobicity, are the preliminary properties to determine the uniqueness of proteins or enzymes [36]. The average molecular weight of GH5-cellulase family from uncultured microorganisms was 54862.07 Da or 54.86 kDa. The cellulase with the accession number of ACA61137.1 had a pI above 7 (pI > 7), 8.55, and another cellulase fell under 7 (pI < 7; Table 2). An isoelectric point (pI) below 7 (pI < 7) indicates the acidic nature of the protein. On the other hand, a pI of more than 7 depicts the alkaline nature. Negative charges (–R) of the sequences were computed based on numbers of aspartic acid and glutamic acid, while positive charges (+R) were based on numbers of arginine and lysine. Table 2 showed that the majority of cellulase had their pI lower than 7, indicating that the numbers of aspartic acid and glutamic acid for each cellulase sequence were more than arginine and lysine, except ACA61137.1 that had a pI > 7. Six sequences from the 27 selected sequences had *I*I values of more than 40. This means that these sequences (ACA61162.1, ACA61171.1, ACH67609.1, AOA60285.1, AOA60286.1, and CelGH5) were predicted unstable in the test tubes. GRAVY index of cellulases had negative values ranging from –0.562 to –0.207. This result revealed that all GH5-cellulases from uncultured microorganisms had good interactions with water. The increasing positive scores indicated a greater hydrophobicity. The aliphatic index of a protein was defined using the aliphatic side chains such as alanine, valine, isoleucine, and leucine. It was a positive factor that could increase the thermostability of globular proteins [55]. The aliphatic index of the GH5-cellulase family was ranging from 62.20 to 84.28. The high aliphatic index refers to the fact that protein may be stable for a wide range of temperatures.

### Stability analysis

CelGH5 gave  $T_m$  values at pH 2.5 to pH 11.0 and no apparent  $T_m$  values at pH 2.0, pH 11.5, and pH 12.0 (Fig. 1). These results indicate that CelGH5 has a wide pH range, from acidic to basic. At pH 4.0, the highest  $T_m$  value is obtained. This suggests that CelGH5 is more stable in acidic environments. Although it gives a  $T_m$  value in alkaline conditions (pH 7.5–11.0), the provided  $T_m$  is lower than the  $T_m$  of the control. Additives added to protein solutions could be stabilized or destabilized (Fig. 2). Thirteen additives were tested, and it was found that 7 additives gave a

**Table 1** Details of selected sequences with their protein accessions

No.	Protein accessions	Name	Source	Length (aa)
1	ACA61132.1	Cellulase	Uncultured microorganism	553
2	ACA61135.1	Cellulase	Uncultured microorganism	552
3	ACA61137.1	Cellulase	Uncultured microorganism	546
4	ACA61140.1	Cellulase	Uncultured microorganism	537
5	ACA61144.1	Cellulase	Uncultured microorganism	512
6	ACA61145.1	Cellulase	Uncultured microorganism	532
7	ACA61149.1	Cellulase	Uncultured microorganism	520
8	ACA61152.1	Cellulase	Uncultured microorganism	346
9	ACA61160.1	Cellulase	Uncultured microorganism	518
10	ACA61162.1	Cellodextrinase	Uncultured microorganism	332
11	ACA61171.1	Cellobiosidase	Uncultured microorganism	386
12	ACH67609.1	Cellulase	Uncultured microorganism	345
13	ADB80100.1	Endoglucanase	Uncultured microorganism	532
14	ADB80110.1	Endoglucanase	Uncultured microorganism	343
15	ADB80112.1	Cellodextrinase	Uncultured microorganism	370
16	ADK55024.1	CelA	Uncultured microorganism	551
17	ADR64667.1	Cellulase	Uncultured microorganism	592
18	ADR64668.1	Cellulase	Uncultured microorganism	719
19	AEX97595.1	Cellulase	Uncultured microorganism	751
20	AEX97596.1	Cellulase	Uncultured microorganism	473
21	AFQ39736.1	Cellulase	Uncultured microorganism	559
22	AHB33631.1	Endo-1,4- $\beta$ -D-glucanase	Uncultured microorganism	552
23	AHW46443.1	Cellulase	Uncultured microorganism	531
24	AOA60285.1	Cellulase	Uncultured microorganism	341
25	AOA60286.1	Cellulase	Uncultured microorganism	344
26	AOA60287.1	Cellulase	Uncultured microorganism	515
27	CelGH5 (this study)	Cellulase	Uncultured microorganism	333

lower  $T_m$  value than the control or destabilizing properties. Imidazole, EDTA, NaCl,  $(\text{NH}_4)_2\text{SO}_4$ ,  $\text{CaCl}_2$ ,  $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$ , and glucose were destabilized additives that should be avoided in CelGH5 storage. The other additives, glycerol, 2-mercaptoethanol, urea, KCl, arabinose, and galactose, had  $T_m$  values similar to the control.

#### Primary structure analysis

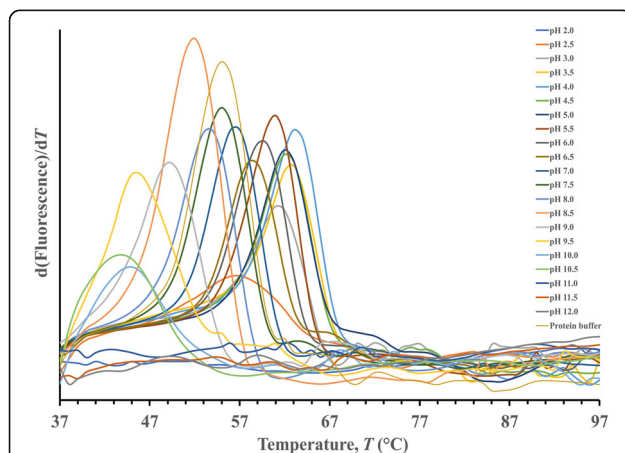
Proteins differ from one another by their primary structures. Primary structure studies reveal the characteristics of all proteins. The amino acid composition of the GH5-cellulase family from uncultured microorganisms was determined using the CLC Main Workbench 8.1.2 software (QIAGEN). Figure 3 showed that alanine (8.5%) was the most abundant amino acid in all these sequences, followed by glycine (7.2%), leucine (7.0%), threonine (6.8%), aspartic acid (6.6%), glutamic acid (6.4%), and valine (6.3%). The composition of cysteine

had the least quantity as compared to all amino acids. Figure 3 showed the comparative percentage average of amino acids in the GH5-cellulase family from uncultured microorganisms. Hydrophobicity was calculated by the number of hydrophobic residues (alanine, phenylalanine, glycine, isoleucine, leucine, methionine, proline, valine, tryptophan) and hydrophilic residues (cysteine, asparagine, glutamine, serine, threonine, tyrosine). All cellulase sequences analyzed were hydrophobic (Fig. 4). ADR64667.1 was a sequence with the highest hydrophobic residue percentage, whereas AEX97595.1 had the lowest.

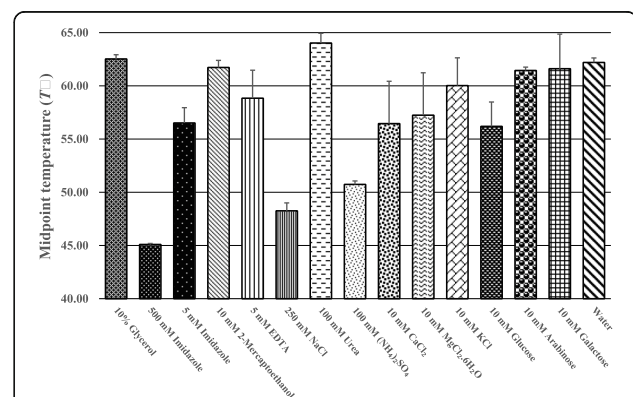
MEME software can determine the conserved motif of a full-length protein. Table 3 showed six conserved motifs of all 27 sequences of the GH5-cellulase family from uncultured microorganisms. Five out of six motifs were identified as GH 5 family motifs, and there was no information for one motif.

**Table 2** Physicochemical properties computed using ExPASy-ProtParam tool

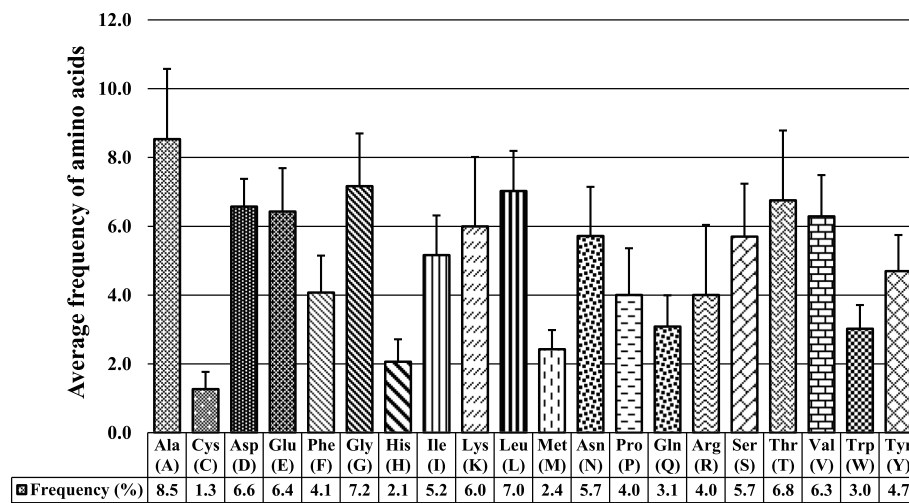
Accession	Mol. Wt. (Da)	pI	II	AI	GRAVY	-R	+R
ACA61132.1	62309.11	6.69	26.05	71.30	-0.538	66	65
ACA61135.1	61703.39	5.80	22.41	70.40	-0.497	66	60
ACA61137.1	60675.41	8.55	28.92	75.59	-0.415	56	60
ACA61140.1	60318.78	5.20	22.82	72.46	-0.467	70	53
ACA61144.1	56867.11	5.52	30.56	74.12	-0.353	61	51
ACA61145.1	58714.68	5.50	36.24	67.59	-0.451	60	49
ACA61149.1	57251.13	4.97	26.20	71.13	-0.312	67	48
ACA61152.1	39812.18	4.79	38.04	69.68	-0.273	46	29
ACA61160.1	56643.65	5.00	33.62	77.97	-0.233	68	49
ACA61162.1	38377.09	4.99	40.12	74.97	-0.401	53	36
ACA61171.1	45519.42	5.15	46.34	80.91	-0.549	60	43
ACH67609.1	39813.14	5.91	42.66	80.90	-0.344	44	36
ADB80100.1	59396.57	5.56	33.30	62.20	-0.498	60	49
ADB80110.1	39072.39	4.91	38.73	75.34	-0.239	46	29
ADB80112.1	43461.81	4.99	39.17	75.14	-0.562	60	38
ADK55024.1	62292.29	5.39	26.55	74.01	-0.501	71	60
ADR64667.1	64828.05	6.01	34.79	72.45	-0.329	74	65
ADR64668.1	79700.04	4.85	23.17	74.18	-0.372	91	61
AEX97595.1	83414.86	4.91	29.01	68.77	-0.480	91	65
AEX97596.1	51859.43	4.47	27.19	75.05	-0.207	56	29
AFQ39736.1	62735.57	6.30	23.73	72.61	-0.541	66	63
AHB33631.1	62552.79	5.60	26.59	75.85	-0.456	71	60
AHW46443.1	57361.16	5.16	26.56	66.57	-0.419	56	38
AOA60285.1	40643.19	5.59	44.78	84.28	-0.515	54	45
AOA60286.1	40450.78	5.25	44.09	82.21	-0.428	54	41
AOA60287.1	57844.96	4.68	33.97	82.37	-0.371	79	48
CelGH5 (this study)	37656.79	6.72	41.17	77.72	-0.303	39	38



**Fig. 1** Thermostability analysis of CelGH5 in BR buffer at various pH values. The melting temperature ( $T_m$ ) is defined as the midpoint temperature of the protein folding–unfolding transition [56].  $T_m$  is the first derivative of the fluorescence emission as a function of temperature ( $dF/dT$ ). Here,  $T_m$  is represented as the highest part of the curve



**Fig. 2** Midpoint temperatures of the protein-unfolding transition ( $T_m$ ) for CelGH5 in the presence of the additives. The control experiment is water, represented as a reference



**Fig. 3** Amino acid composition of GH5-cellulases family from uncultured microorganisms computed using the CLC workbench 8.1.2 software

### Secondary structure analysis

The secondary structure contained  $\alpha$ -helix,  $\beta$ -sheet or strand, and turns. However, one structure was not classified in the three usual groups; this was called a random coil. The SOPMA server analyzed the percentage or composition of  $\alpha$ -helix,  $\beta$ -turn, extended strand, and random coils. Secondary structure analyses showed the percentage of each conformation. SOPMA revealed that the random coil was much greater than other secondary structures, such as helix, sheet, and turn. The random coil is usually described as a more flexible and dynamic folded chain region than other secondary conformational structures [57]. Table 4 showed the comparative percentage of  $\alpha$ -helix, strands,  $\beta$ -turns, and random coil within all GH5-cellulase sequences. Sequences with accession numbers ACA61162.1, ACH67609.1, ADB80112.1, AOA60285.1, AOA60286.1, and CelGH5 had higher percentages of  $\alpha$ -helix than random coils. The high alanine content might be due to the six sequences with more  $\alpha$ -helix structures than other structures.

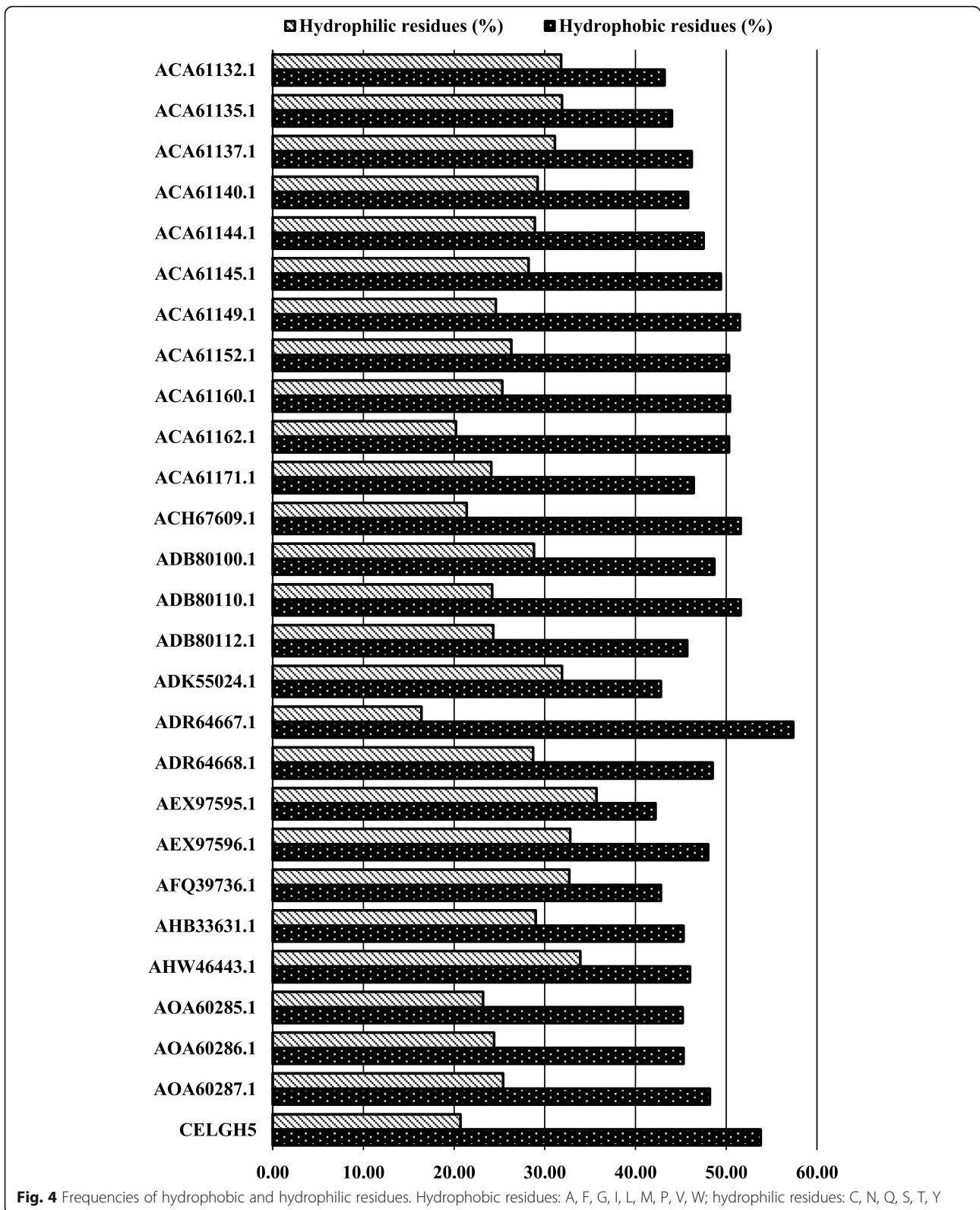
Cellulase structures with PDB ID 4EE9, 5I2U, 4M1R, and 4HTY were cellulases belonging to glycoside hydrolase family 5, recently identified via the metagenome approach. PDB ID 4EE9 was identified from the Antarctic soil [58], 5I2U was isolated from soil metagenome [27], 4M1R was from sugarcane soil [29], and 4HTY was from a metagenomic library. Commonly, cellulases from the GH5 family have a typical TIM-barrel fold consisting of  $\alpha$ -helices and stranded parallel  $\beta$ -sheet as a core, and another secondary structure, like  $\beta$ -turn and coil. Table 4 showed the average conformational structures of cellulase dominated by random coils (42.77%), followed by  $\alpha$ -helix (31.77%), strand (17.19%), and  $\beta$ -turn (8.28%). Figure 5 showed two schematic wiring diagrams of different cellulase structures. This figure confirmed from the

predicted secondary structure that random coil had the highest content, followed by  $\alpha$ -helix, strand, and  $\beta$ -turn. Disulfide bridges (Fig. 5a) connected cysteine residues 270 and 312. A small  $\beta$ -hairpin connected two strands in between residues 95 and 98. A small  $\beta$ -hairpin was also found in Fig. 5b that connected residues 22 and 25. There was no disulfide bridge found in Fig. 5b.

### Tertiary structure analysis

The tertiary structures of the selected GH5-cellulase family were evaluated and assessed using computational tools. QMEAN4, ERRAT, Z score, and Ramachandran plot were quality parameters to assess and evaluate the tertiary structures of the GH5-cellulase family. There were four GH5-cellulase families from uncultured microorganisms that were structured (PDB ID 5I2U, 4EE9, 4M1R, 4HTY). Table 5 showed QMEAN4, ERRAT, Z score, and Ramachandran plot of four cellulase structures. A larger QMEAN4 score indicated a better structure, whereas negative scores referred to an unstable structure [45]. QMEAN4 predicted the global model structure quality based on a linear combination of four descriptors: local geometry, distance-dependent interaction, agreement of predicted secondary structure and solvent accessibility, and solvation potential. Figure 6a showed that the QMEAN4 scored 0.09, which represented a reliable 3D structure. The results also showed that the QMEAN4 Z score was compared to the nonredundant set of PDB structures. The QMEAN4 Z score of the structure was included in the group of PDB structures with a QMEAN Z score of less than 1.

ProSA (Protein Structure Analysis) evaluated the accuracy of protein structure or model structure for prediction structure. The analysis was carried out based on statistical analyses of experimental protein structures,





**Table 3** The six motifs of the GH5-cellulases family from uncultured microorganisms found among the 27 sequences

Length	Sequence	Occurrence at different site	Conserved domain
37	EMD TDGKVBDAWMARVKEVDYVIDEGMYCIINVHHD	17	GH 5
41	TWRRTAQHETCWGQPVTKPELIKMMKEAGFGAIRVPVTWYQ	14	GH 5
33	YNTNKERYEKLWKQIAEEFKDYGQKLLFEAYNE	22	GH 5
41	YKAINSYAKSFVTTVRATGGNNATRNLIIVNTYAASSTPNAM	17	GH 5
50	ALYAMDYLIKKAKEAGIGTFYWMGLSDGDYRSLPAFNQPDLAETJJKAYY	13	No information
50	HIIFQLHSYPNWQSESNKSEIDNLIISNIKSNNLRAPVIIEYATFTTW	7	GH 5

either by X-ray crystallography or NMR spectroscopy. The validation result of the 3D structure was a Z score. The 3D structure would be accurate if it had a Z score within the Z score range of the experimental protein structure [59]. Figure 6c showed that cellulase's Z score was -9.36, which was included within the Z score range of the protein structure experimental with X-ray spectroscopy. ERRAT and Ramachandran plot were two other parameters to determine the quality of the tertiary structure. ERRAT values were related to structure resolutions. High resolution of 3D structures generally produces values around 95% or higher and lower resolutions would be present if the average overall quality factor is around 91%. Figure 6e reveals the overall quality factor of cellulase structure with the ERRAT value of 94.965%, a good enough structure resolution. A good quality model based on the Ramachandran plot would be expected to have over 90% in the most favored regions. The Ramachandran plot in Fig. 6d showed that residues in the favored region were less than 90%.

#### Functional analysis

In this study, the cysteine residues were determined using the CYS\_REC server. Table 6 reveals that among 27 protein sequences, 16 protein sequences contained cysteine residues connected by disulfide bonds. The presence of these disulfide bridges was regarded as a positive factor for stability at the molecular level. The amount of disulfide bonds was also calculated to determine the structure because of its role in protein folding. The CYS\_REC server also determined the specific residue number connected by disulfide bonds between cysteine residues. For example, the sequence with accession number AEX97595.1 had more than one sequence of disulfide bridges.

Table 6 showed the results of sequence analysis using CDD interactive web-based tools. It can be asserted that the sequence contained not only cellulase domains but also other domains. AEX97595.1 was the only cellulase with CBM among 27 sequences. AEX97595.1 had modular architecture, Cellulase - Dockerin\_I - CBM\_4\_9. The presence of CBM could increase the binding capacity of cellulase

to the substrate, indirectly helping the catalysis process of cellulose by cellulase. ACA61144.1, ACA61149.1, ACA61160.1, and AHW46443.1 had a Big 5 domain located before the cellulase domain. Meanwhile, ACA61145.1 and ADB80100.1 had the BACON domain. Another sequence had only cellulase domains without other domains. Information of conserved domains in cellulase sequence could be the engineering object to increase the ability or stability of cellulases.

Protease digestion is a valuable method for determining correct metabolism, enzymatic digestion, and high-order protein structure simplification. In addition to proteases, it is also important to identify chemicals that can cleave peptide chains. This study found ten endopeptidase/chemical that has the highest average number of cleavage sites in GH5-cellulase family sequences of the uncultured microorganisms. Those are Asp-N endopeptidase, chymotrypsin, formic acid, glutamyl endopeptidase, LysC, LysN, pepsin, proteinase K, thermolysin, and trypsin (Fig. 7).

#### Multiple sequence alignment and phylogenetic tree construction

A multiple sequence alignment of retrieved cellulase sequences was performed by the Clustal Omega software and shown in Fig. 8. The sequence alignment identified several conserved amino acid residues (red column), like glycine (G), arginine (R), histidine (H), glutamic acid (E), asparagine (N), tyrosine (Y), and tryptophan (W). The most important residues in the GH5-cellulase sequence were two glutamic acids (E). The two glutamic acid residues had an important role in catalytic activity. Glutamic acid acted as a proton donor, and the other acted as a nucleophile [5, 60, 61]. Other residues had a role in stabilizing the structure and were also found in the cavity of active sites. Changes in amino acid residues in a conserved area could cause changes in the structure and function of these proteins. The phylogenetic tree of the GH5-cellulase family from uncultured microorganisms has been constructed with MEGA X using a maximum likelihood method based on the JTT matrix model with bootstrap replications

**Table 4** Secondary structure among different sequences of GH5-cellulase family from uncultured microorganisms

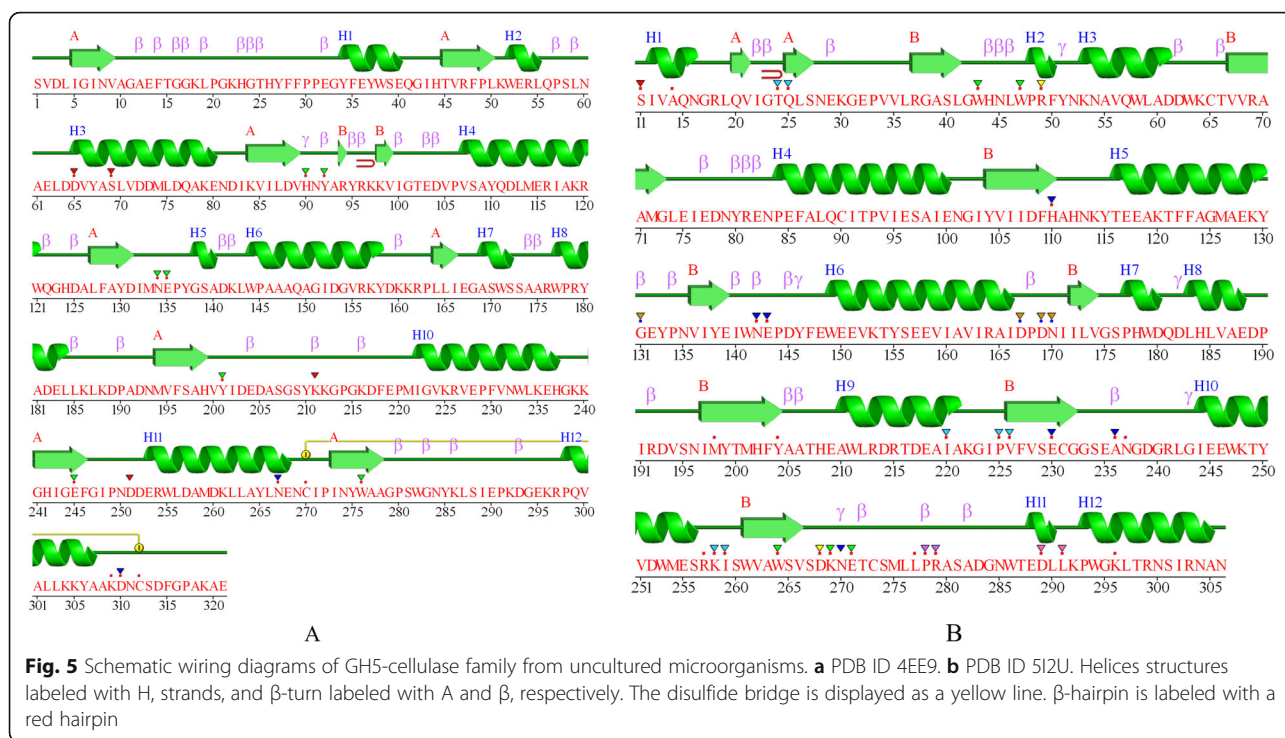
Identity	Contents of principal secondary structure			
	$\alpha$ -helix (%)	Extended strand (%)	$\beta$ -turn (%)	Random coil (%)
<b>Protein accession<sup>a</sup></b>				
ACA61132.1	33.45	20.61	3.62	42.31
ACA61135.1	35.69	20.11	5.07	39.13
ACA61137.1	32.05	20.70	4.40	42.86
ACA61140.1	29.42	23.46	4.28	42.83
ACA61144.1	29.10	19.53	4.10	47.27
ACA61145.1	28.95	18.05	4.14	48.87
ACA61149.1	30.19	18.27	4.62	46.92
ACA61152.1	31.50	17.92	7.23	43.35
ACA61160.1	28.76	18.53	3.28	49.42
ACA61162.1	43.07	14.16	8.13	34.64
ACA61171.1	39.90	13.73	5.44	40.93
ACH67609.1	43.77	13.62	5.80	36.81
ADB80100.1	29.70	18.05	3.76	48.50
ADB80110.1	30.32	17.20	7.29	45.19
ADB80112.1	42.43	14.05	4.86	38.65
ADK55024.1	29.76	22.87	4.54	42.83
ADR64667.1	32.60	16.72	6.93	43.75
ADR64668.1	29.39	21.59	8.36	40.67
AEX97595.1	29.16	23.83	6.66	40.35
AEX97596.1	39.11	13.95	4.44	42.49
AFQ39736.1	30.41	21.47	5.01	43.11
AHB33631.1	27.72	20.11	4.53	47.64
AHW46443.1	31.83	16.95	3.58	47.65
AOA60285.1	42.23	15.25	6.74	35.78
AOA60286.1	43.02	14.24	6.40	36.34
AOA60287.1	27.57	20.39	8.93	43.11
CelGH5	42.34	14.71	5.71	37.24
Average frequency	33.83	18.15	5.48	42.54
<b>PDB ID<sup>b</sup></b>				
4EE9	33.33	16.51	7.48	42.68
4M1R	36.49	18.58	8.78	36.15
5I2U	31.33	18.07	9.34	41.27
4HTY	25.91	15.60	7.52	50.97
Average frequency	31.77	17.19	8.28	42.77

<sup>a</sup>Predicted structure<sup>b</sup>Experimental structure

are 1000 replicates. Figure 9 showed a cladogram of cellulase and distributed into three nodes. The dominant node consisted of 14 nodes and was marked in red lines. The second group consisted of 10 nodes and was represented by a brown line, including our sequence, CelGH5. The last group consisted of 3 nodes and was marked by blue lines.

## Discussion

An isoelectric point is a condition in which the protein surface is covered with no charge or the net charge, and thus the protein charge, is zero. At an isoelectric point, proteins or enzymes are compact and stable. The isoelectric point calculation is important for determining purification buffer systems focusing on an isoelectric



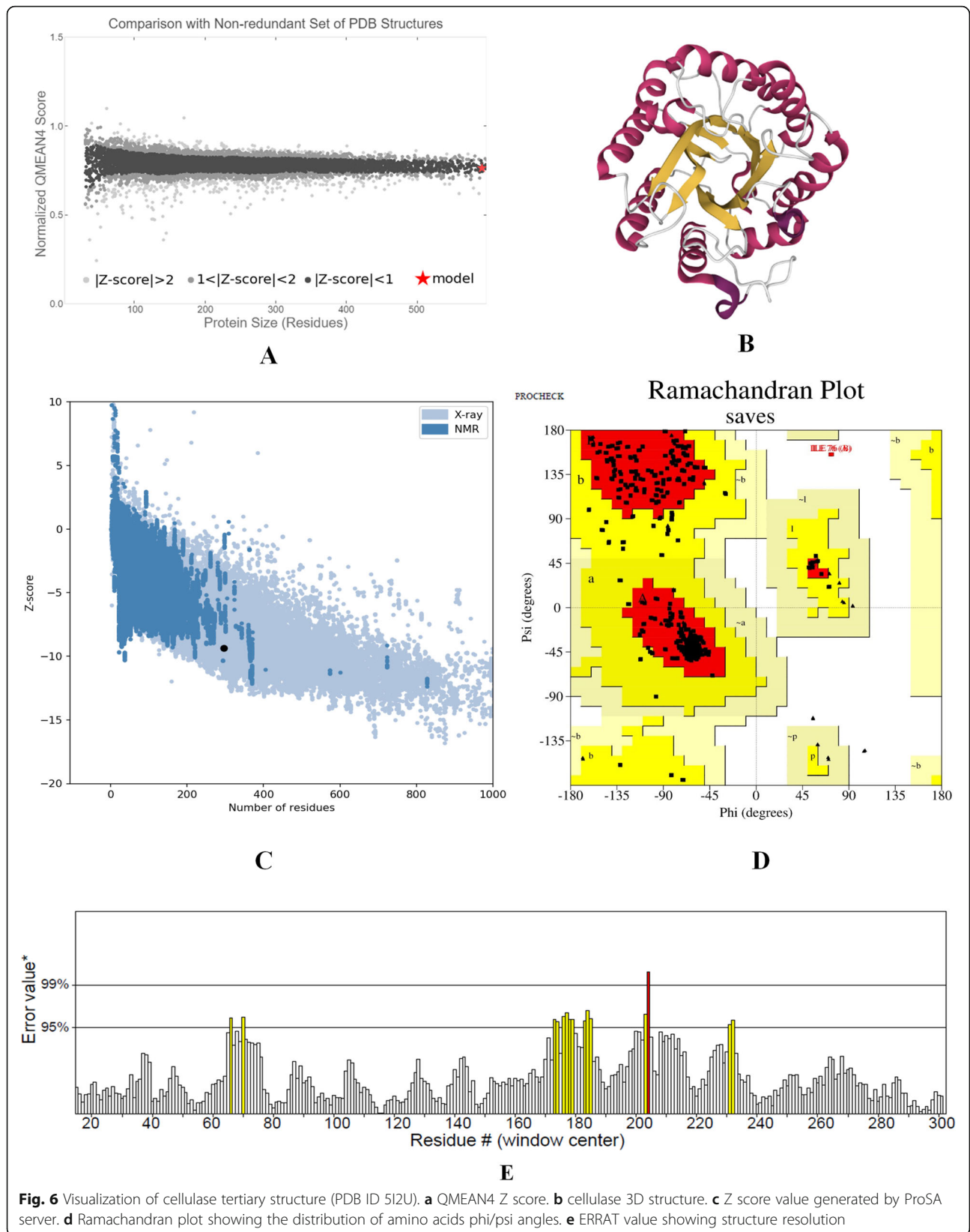
and buffer systems for crystallization. The high efficiency and promising nature of protein crystallization can be improved by determining the pI of the protein, followed by screening for a buffer range at or near that pI value (within 2–3 pH units of the pI) [62]. In the current study, 27 cellulase sequences retrieved from GenBank had an isoelectric point (pI) values of less than 7, except the sequence with accession number ACA61137.1, which had a pI of 8.55. This result indicates that the GH5-cellulase family from uncultured microorganisms had acidic properties. Hoda et al. [35] found that GH5 cellulase from *Ruminococcus albus* had pI values ranging between 4.39 and 4.53, suggesting moderately acidic properties.

An analysis of halophilic/halotolerant enzymes revealed a consensus in which these enzymes tended to have more acidic or negative residues than their non-

halophilic homologs [63]. The amount of glutamate and aspartate residues (–R) as acidic residues in the primary structure could not be used as references to determine the enzymes' acidity or halophilic properties. The acidity or halophilic properties of enzymes could be determined from the glutamate and aspartate residues on the enzymes' surfaces [27, 63, 64]; this would be known after determining the enzyme structure. The cellulase sequence analysis (PDB ID 5I2U) showed that 52 (16.7%) residues were acidic [27]. This result was relatively greater than that of other halophilic cellulases. The endoglucanase from *Bacillus subtilis* 168 (PDB ID: 3PZT) had 38 (11.6%) acidic residues [64], and the GH5 cellulases from *Thermoanaerobacterium*, which possessed halostable characteristics, only had 43 (11.3%) acidic residues [33]. CelGH5 possessed a slight difference between the negatively and positively charged

**Table 5** Comparison of QMEAN4, ERRAT, Ramachandran plot, and Z score for the quality assessment of three-dimensional structures

PDB ID	QMEA N 4 score	ERRAT quality factor (%)	Ramachandran plot				Z score
			Residues in favored region (%)	Residues in additional allowed region (%)	Residues in generously allowed region (%)	Residues in disallowed region (%)	
4EE9	0.54	96.154	89.7	10.3	0.0	0.0	–9.91
4M1R	0.40	95.606	88.1	11.9	0.0	0.0	–7.56
5I2U	0.09	94.965	87.7	11.9	0.0	0.4	–9.36
4HTY	0.11	96.530	90.0	9.7	0.3	0.0	–9.64



**Fig. 6** Visualization of cellulase tertiary structure (PDB ID 5I2U). **a** QMEAN4 Z score. **b** cellulase 3D structure. **c** Z score value generated by ProSA server. **d** Ramachandran plot showing the distribution of amino acids phi/psi angles. **e** ERRAT value showing structure resolution

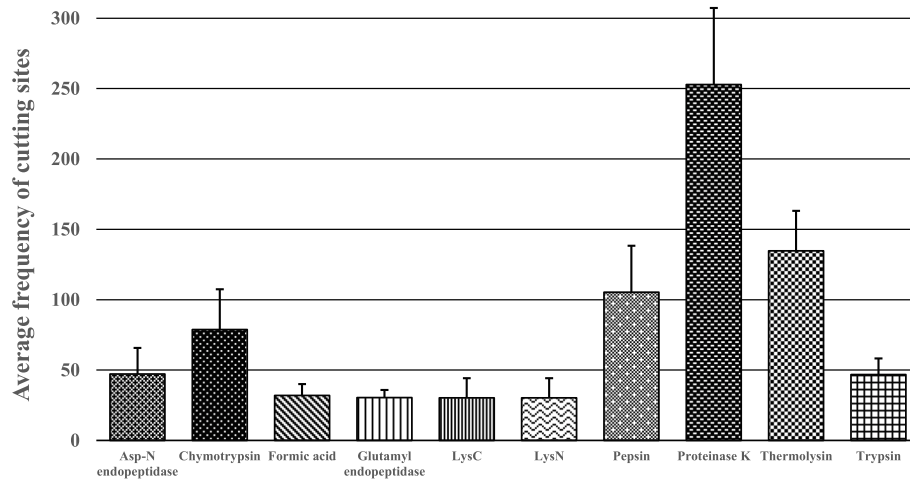
**Table 6** Disulfide bond prediction and conserved domain identification

Protein accession	Cys rec	Conserved domain	
		Domain	Position (aa)
ACA61132.1	Not SS-bounded	Cellulase	58-330
ACA61135.1	Not SS-bounded	Cellulase	58-326
ACA61137.1	Not SS-bounded	Cellulase	56-324
ACA61140.1	Not SS-bounded	Cellulase	56-324
ACA61144.1	61-477	Big_5	27-128
		Cellulase	187-480
ACA61145.1	22-174, 87-441, 325-473	BACON	36-120
		Cellulase	185-480
ACA61149.1	Not SS-bounded	Big_5	31-132
		Cellulase	185-492
ACA61152.1	33-313, 188-271	Cellulase	78-313
ACA61160.1	64-304	Big_5	32-130
		Cellulase	192-485
ACA61162.1	42-262	GH superfamily	55-311
ACA61171.1	21-255, 354-368	GH superfamily	65-357
ACH67609.1	Not SS-bounded	GH superfamily	25-320
ADB80100.1	161-325	BACON	34-117
		BACON	67-118
		Cellulase	185-480
ADB80110.1	26-206, 181-264	Cellulase	71-306
ADB80112.1	35-350	GH superfamily	57-353
ADK55024.1	11-62, 490-498	Cellulase	60-333
ADR64667.1	204-267, 241-502, 427-537	Cellulase	47-351
ADR64668.1	70-361	Cellulase	68-371
AEX97595.1	206-234, 238-333, 339-450, 562-731	Cellulase	64-347
		Dockerin_1	385-437
		CBM_4_9	478-569
		CBM_4_9	620-725
AEX97596.1	70-147, 262-344, 273-288	Cellulase	60-350
AFQ39736.1	Not SS-bounded	Cellulase	58-331
AHB33631.1	126-490	Cellulase	62-334
AHW46443.1	Not SS-bounded	Big_5	44-139
		Cellulase	186-480
AOA60285.1	77-81	Cellulase	31-315
AOA60286.1	Not SS-bounded	Cellulase	9-317
AOA60287.1	Not SS-bounded	Cellulase	91-333
CelGH5	Not SS-bounded	Cellulase	57-304

residues. Despite this result, CelGH5 had halophile properties with a relative activity of more than 200% in the presence of 3M NaCl (data not shown).

The instability index (*II*) showed an estimation of the protein stability in a test tube. The instability index portrayed a stable protein when the index value was less

than 40, and an unstable condition was shown when the index value was greater than 40. Six sequences from the 27 selected sequences had *II* values greater than 40. This means that these sequences (ACA61162.1, ACA61171.1, ACH67609.1, AOA60285.1, AOA60286.1, and CelGH5) were predicted unstable in test tubes. This result was in



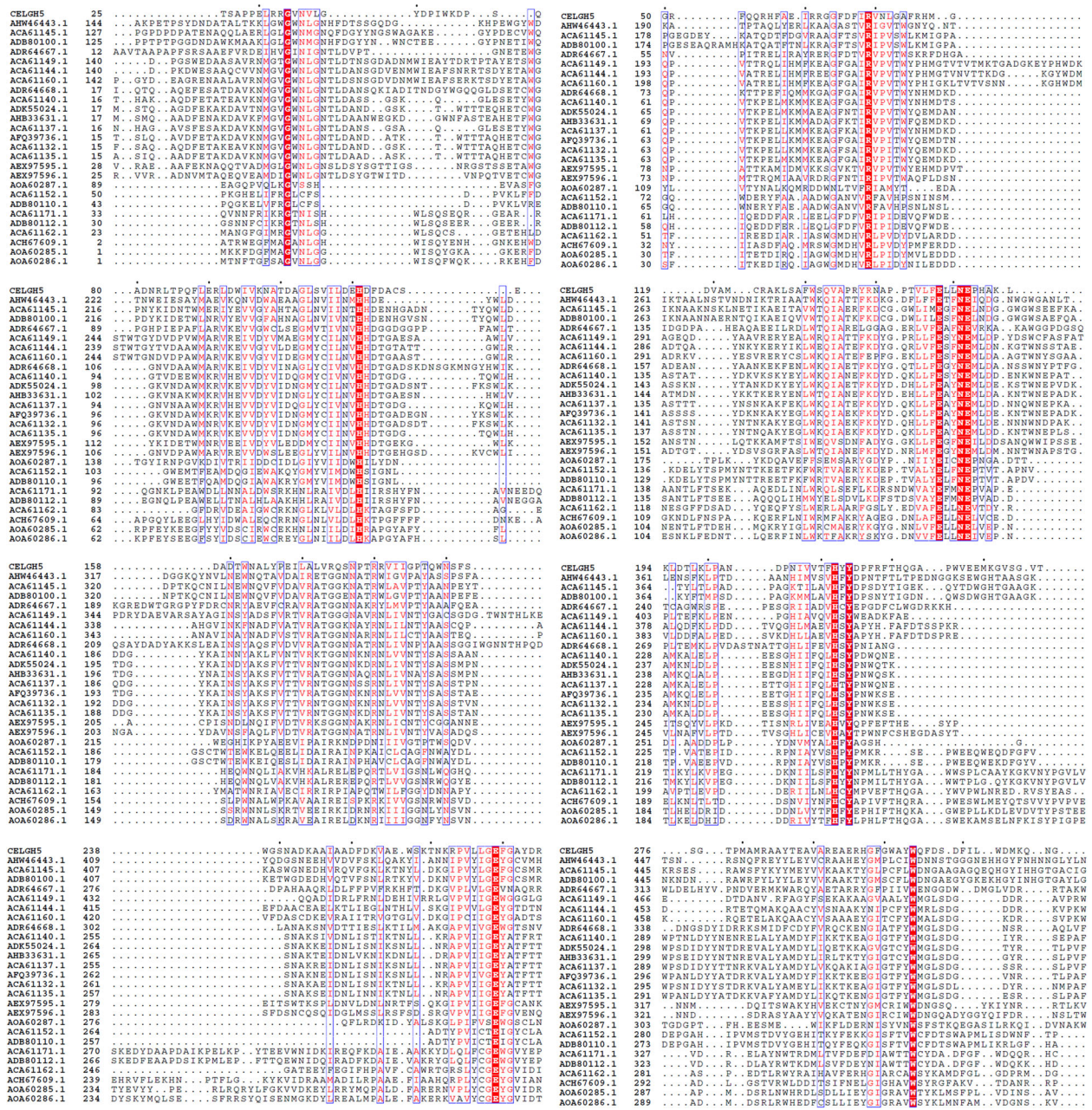
**Fig. 7** Average number of cleavage sites for the GH5-cellulase family from uncultured microorganisms as identified through the peptide cutter tool

contrast with that of Duan et al. [65], whose paper showed that ACA61162.1 had an optimal condition at pH 4.5 and was stable a pH range of 3.5 to 10.5 based on experimental data [65]; in contrast, the *II* showed different results. The analysis of this condition shows that environmental aspects, such as the autolysis of an enzyme, do not encapsulate the instability index calculation. Furthermore, the *II* model was only based on the primary sequence, and the secondary or tertiary structure contributions were not incorporated into the model [66]. Gamage et al. [66] calculated *II* values of three proteins; the results were consistent, similar to the degradation pattern observed by SDS-PAGE analyses. However, the unstable properties of  $\alpha$ -S1-casein displayed in the *II* value were not related to the natural degradation visualized on SDS-PAGE analyses.

Based on the *II* value, CelGH5 was categorized as an unstable protein. Nevertheless, the ThermoFluor assay revealed that CelGH5 has a wide pH range of 2.5 to 11.0. In this pH range, CelGH5 gave an emission signal recorded by RT-PCR and converted to a melting point ( $T_m$ ). The increase in melting temperature under different buffers or additives gave rise to a thermal shift that quantified the stabilization of the protein [67]. At pH values of 2.0, 11.5, and 12, no apparent  $T_m$  was observed, indicating that the CelGH5 structure is destabilized at these pHs. CelGH5 had the highest  $T_m$  value at pH 4.0 or acidic conditions. Therefore, CelGH5 is suggested to be stored in a pH 4.0 buffer. Apart from the pH 4.0 buffer, an additive, such as glycerol, can be added to the CelGH5 solution because it does not affect the CelGH5 stability (Fig. 2). Glycerol is a cryoprotectant that helps stabilize proteins by preventing the formation of ice crystals at  $-20$  °C, and thus the destruction of the protein structure. Other properties of CelGH5 include high

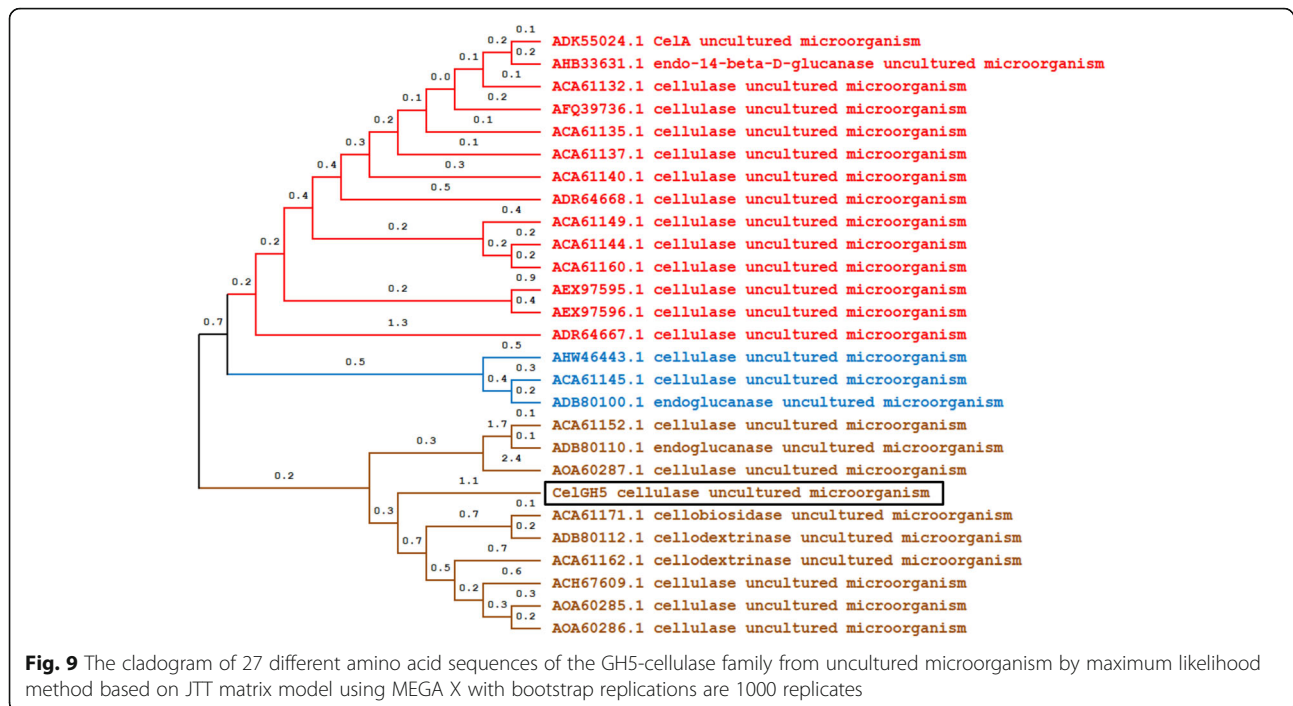
stability with residual activity of 52% after 240 h incubation at 55 °C (data not shown). Thus, the results showed that the most important experimental condition is the careful use of the *II* to predict in vitro protein stability. This condition tells us that the *II* prediction does not accommodate all relevant information in the determination of protein stability under in vitro conditions. The application of *II* prediction toward protein stability still depends on the intrinsic nature of the protein and conditions of the protein milieu.

GRAVY analyses were calculated by adding the hydrophathy values [68] of each amino acid residue and dividing by the length of full sequences. The GRAVY index represented the solubility of proteins and positive interactions with water [69]. The increasing positive scores indicated greater hydrophobicity. A low GRAVY value represented good interaction between water and protein. The GRAVY index of cellulases had negative values ranging from  $-0.562$  to  $-0.207$ . This result revealed that all members of the GH5-cellulase family from uncultured microorganisms had good interactions with water. Although it was known that all analyzed cellulases had hydrophobic properties, it did not necessarily mean that they had a poor interaction with water. The GRAVY values and hydrophobic components of the amino acid sequence residues are a different matter. The hydrophobic residues in the formation of the three-dimensional structure are located inside or buried within the structure; thus, all surfaces interacting with water contain hydrophilic residues. Asparagine, cysteine, glutamine, serine, threonine, and tyrosine are hydrophilic amino acids that have a propensity to interact in the aqueous environment due to polarity properties; these residues are found on protein surfaces.



The high aliphatic index refers to protein stability under a wide range of temperatures. For example, the aliphatic index of the GH5 cellulase family ranged from 62.20 to 84.28. The higher the AI value, the greater the thermal stability of an enzyme. For example, the sequence with accession number AOA60285.1 was more stable than ADB80100.1. Interestingly, based on the *I*I value, AOA60285.1 was an unstable enzyme. This result reinforces the notion that the use of the *I*I as a reference in determining the stability of proteins or enzymes may also need to consider other influencing factors.

Primary structure analysis showed that alanine, glycine, leucine, threonine, aspartic acid, glutamic acid, and valine were the most abundant amino acids in the cellulase sequences analyzed. The number of cysteines was lower than other amino acids. Together with glycine, leucine, and glutamic acid, alanine had a greater tendency to build  $\alpha$ -helix secondary structures in the protein conformation. This was in contrast with threonine and valine, which usually built  $\beta$ -sheet secondary structures. The aspartic acid had the role of connecting with the solvent, supported by



hydrogen bonds. All analyzed cellulase sequences had hydrophobic properties because the majority of amino acid side chains had hydrophobic properties. Alanine, glycine, leucine, valine, proline, isoleucine, tryptophan, phenylalanine, and methionine had hydrophobic properties, and these amino acids were much more abundant than other amino acids.

MEME software revealed sequence motifs in all 27 sequences of the GH5-cellulase family, and a consensus of these sequences functioned as a signature sequence identifying the enzymes. Five out of six motifs were found, and one motif had no information. In order to confirm the conserved motif, an internet tool (<https://www.genome.jp/tools/motif/>) was used. With this tool, five motifs were confirmed as belonging to the GH5 family domain. The motifs also explained the diversity of the structures and functions of enzymes [70].

The SOPMA server analyzed the percentage of  $\alpha$ -helix,  $\beta$ -turn, extended strand, and random-coil compositions. Secondary structure analyses displayed the percentage of each conformation. The coil structure had a higher percentage than other conformations. These results align with Hoda et al. [30], who found in cellulase from *Ruminococcus albus* that random coils were the most dominant secondary structure, followed by  $\alpha$ -helix. The high percentage of coil might be caused by the high number of glycines and the presence of prolines [71]. A good glycine percentage in the sequence granted high flexibility to the polypeptide chain and

provided structural rigidity. The properties of proline were created in a coiling structure because of the crinkly polypeptide chains that interfered with the secondary structures. Sequences with accession numbers, ACA61162.1, ACH67609.1, ADB80112.1, AOA60285.1, and AOA60286.1, had lower random coil percentages than  $\alpha$ -helix structures. This was a result of the high number of alanines. These five sequences are likely to be present in cellulases from *Bacillus thuringiensis* and *Bacillus pumilus* [34], which have a higher  $\alpha$ -helix structure percentage than other secondary structures.

The different amino acid sequences influenced the properties and formed different structures. Alanine, glutamic acid, and leucine were uncharged amino acids that played a significant role in the high helix-forming propensities. In contrast, glycine and proline had only a few helix-forming propensities [72]. Proline did not have any amide hydrogens; thus, it could not donate any amide hydrogens. However, it could break or bend the helix structure; additionally, the side chains could be disrupted because of the steric position of the backbone of the preceding turn inside a helix [73]. Proline was also found in the edge strands of  $\beta$ -sheets and existed presumably to avoid an "edge-to-edge" protein association that might have led to aggregation and amyloid formation. Proline was seen as the first residue of the helix due to the rigidity of the structure. However, glycine also disturbed the flexibility conformation of  $\alpha$ -helical structures. Tyrosine, phenylalanine, tryptophan (a large



aromatic group residue), threonine, valine, and isoleucine ( $\beta$ -branched amino acids) were mostly found in the middle of  $\beta$ -sheets [74].

The secondary structure  $\beta$ -turns had the lowest percentage.  $\beta$ -turns or reverse turns usually connected different antiparallel  $\beta$ -strands. The  $\beta$ -turn was stabilized by hydrogen bonds connecting the carbonyl oxygen and amide hydrogen. The  $\beta$ -turn was arranged into the four amino acids with the carbonyl oxygen as the first residue and the amide hydrogen as the fourth residue. Glycine and proline tended to have arrangements of  $\beta$ -turns. Proline had a crucial role in building the cis conformation that supported the  $\beta$ -turn formation. Contrastingly, glycine just had a small R group that allowed for high flexibility. There are some theories concerning the role of  $\beta$ -turns in globular proteins. First of all,  $\beta$ -turns had weak bonds that could not support the secondary structures. Second,  $\beta$ -turns played a role in the folding process. However, both of these perspectives were still inaccurate and required further supporting experiments.

There were four GH5 cellulases from uncultured microorganisms that had been structured (i.e., PDB ID 5I2U, 4EE9, 4M1R, 4HTY). Cellulases from the GH5 family had a typical TIM-barrel fold consisting of  $\alpha$ -helices and  $\beta$ -sheets as a core structure, combined with other secondary structures, such as  $\beta$ -turns and coils. Figure 6b displays the tertiary structure of cellulases obtained using the metagenome approach (PDB ID 5I2U), with halophile properties. The evaluation and quality assessment of structures were performed with the QMEAN4, ERRAT, Z score, and Ramachandran plot. The QMEAN score revealed geometric aspects of the protein structures and the global arrangement of variable residues. A larger QMEAN4 score indicated a better structure, whereas negative scores referred to an unstable structure [45]. QMEAN4 predicted the global quality of model structure based on a linear combination of four descriptors: local geometry, distance-dependent interaction, agreement of predicted secondary structure and solvent accessibility, and solvation potential. The QMEAN4 of cellulase's 3D structures are represented in Fig. 6a. They depicted that the proteins were properly folded into a compact three-dimensional field. QMEAN4 scores of all cellulase structures varied from 0.09 to 0.54 (Table 5). Desirable QMEAN scores were 0–1 [42, 75]. The results also show that the QMEAN4 Z score was compared to the nonredundant set of PDB structures. The QMEAN4 Z score of the structure was included in the group of PDB structures, with a QMEAN Z score of less than 1. The verifications of the 3D structures were determined through crystallography represented by ERRAT values. ERRAT values were related to structure resolutions. ERRAT was also useful for analyzing protein structures from the numbers of non-bounded residues

with a cutoff of 3.5 Å between different pairs of atoms. The high 3D structure resolution generally produces values of approximately 95% or higher. Lower resolutions would be present if the average overall quality factor were roughly 91%. Figure 6e displays the overall quality factor of the cellulase structure, with an ERRAT value of 94.96%, a good enough structural resolution. ERRAT values under 91% indicated that the structure had a lower resolution of approximately 2.5 to 3.0 Å. The Ramachandran plot was constructed to show the positions of each amino acid residue (Fig. 6d). Analysis of the Ramachandran plot (PDB ID 5I2U) showed that 87.7% of residues were present in the most favored region (Table 5). Residues in the favored region of the Ramachandran plot equaling more than 90% represented a good quality structure [76, 77].

The cysteine was an amino acid that played an important role in determining the thermostability of proteins. Cysteine-cysteine residues, creating a disulfide bridge, could influence the stability and folding of proteins. This was caused by an oxidative folding process occurring in the thiol groups of cysteine. Some studies showed strategies to increase protein stability by mutating cysteine. When the native disulfide bond was removed, the stability decreased. Besides, adding disulfide bonds also improved the rigidity and stability of protein structure [78]. The presence of disulfide bridges was regarded as a positive factor for stability at the molecular level [79]. The successful disulfide-bonding analysis supported the accuracy of 3D enzyme structure prediction [80]. The cleavage of disulfide bonds affected the native conformation and biological function. Thus, failed folding of the formation caused by disulfide bonds may have been due to protein aggregates [81].

The peptide cutter tool found 27 proteases and chemicals that can cleave GH5-cellulase sequences from uncultured microorganisms. From the 27 proteases and chemicals, there are 10 that possess the highest average number of cleavage sites, including Asp-N endopeptidase, chymotrypsin, formic acid, glutamyl endopeptidase, LysC, LysN, pepsin, proteinase K, thermolysin, and trypsin. Meanwhile, caspase 1, caspase 2, caspase 4, caspase 6, and enterokinase are proteases with the lowest cleaving ability. The results of the peptide cutter tool cleavage sites could be useful when conducting studies on a portion of a protein, separating domains in a protein, and removing a tagged protein while expressing a fusion protein [57].

The conserved domain position had an important role in determining the catalytic site of the observed sequences. Through this process, other functional domains in the sequence could be identified. CDD is a protein database that lists all proteins that have been registered or deposited using multiple sequence

alignment models and full-length proteins. This database can also be used for the fast identification of proteins by looking at conserved domains in the protein sequence and classifying them into their respective families [47]. Based on the results, it was found that the sequence did contain not only cellulase domains but also other domains. The selected amino acid sequences had Big 5, BACON, Dockerin, and the carbohydrate-binding module (CBM). The presence of CBM could increase the binding capacity of cellulases to the cellulase substrate, indirectly helping the catalysis process [82]. The BACON domain was found in varied domain architectures and associated with various domains, including proteases and carbohydrate-active enzymes. The function of the BACON domain had an unclear relationship with carbohydrate metabolism but a strong connection to protease domains [83]. Dockerin is a domain that belongs to the cellulosome complex. Cellulosomes are multienzyme complexes with cellulosic activity and are usually found in anaerobic bacteria [84–87]. The sequence with accession number AEX97595.1 had a dockerin domain and was predicted as a bacterial cellulase-typical sequence. The bacterial immunoglobulin-like (Big) domain can be widely found in bacterial proteins with diverse biological functions such as adhesion and biofilm development [88].

Glycine, arginine, histidine, glutamic acid, asparagine, tyrosine, and tryptophan were conserved residues identified by the Clustal Omega software. These conserved residues played a pivotal role in the catalytic mechanism and were reported as cellulases from uncultured microorganisms or metagenomic approaches [28, 33, 58, 89]. Glutamic acid played an essential role in the GH5 family as a catalytic residue. Glutamic acid acted as a base or a catalytic nucleophile and a catalytic proton donor [90]. Three glutamic acid residues were found from multiple sequence alignments as conserved residues. Residues E148, E152, E269 were conserved glutamate from the CelGH5 sequence. It was predicted that E148 was the CelGH5 catalytic residue that acted as a proton donor, with E269 acting as a nucleophile. This prediction could be confirmed after determining the CelGH5 structure or aligning its sequence with other sequences whose structures had been determined. Histidine, asparagine, and tyrosine were conserved residues located between two catalytic residues. It was assumed that these residues were located in the CelGH5 cavity site that participated in substrate binding, stability, and hydrogen bond formation between catalytic residues and substrates [33, 58]. Histidine and tyrosine were conserved residues in the catalytic cavity site of cellulases from the soil metagenome library from Antarctica [58]. Glycine, arginine, and tryptophan played a role in the binding of the substrate and influenced hydrolysis activities [91].

The phylogenetic tree of GH5-cellulase was distributed into three nodes, with the dominant node consisting of 14 nodes and the minor nodes consisting of 3 nodes (Fig. 9). Every branch represented evolutionary lineages changing over time, and each lineage had a unique history [44]. CelGH5 clustered in the second group formed a new root and was a direct branch approaching the point of its ancestor. This indicates that CelGH5 is a metagenome GH5-cellulase sequence with a different typical sequence compared to other GH5-cellulase metagenome sequences. The cladogram branches further diverged into small branches, with every branch representing an evolution by the cellulases and each lineage having a unique history [44]. The vertical lines connecting horizontal lines revealed their irrelevance. The GH5-cellulase sequences from uncultured microorganisms diverged into three main daughter lineages; small branches resulted from the daughter branches. Branch length represented genetic changes among the sequences.

## Conclusions

The present study provided new insight on in silico study to determine the characteristics of cellulases from uncultured microorganisms belonging to the GH5 family of the CAZy classification in terms of their physicochemical and structural properties. The sequence length was roughly 332–751 amino acids and had a molecular weight range around 37–83 kDa. Based on the amino acid charge, the dominant-selected cellulase sequences had negative charges and pI values below 7 (acidic). Alanine was the most abundant amino acid making up the GH5-cellulase family, and the percentage of hydrophobic amino acids was more than hydrophilic. Interestingly, ten endopeptidases with the highest average number of cleavage sites were found. Another uniqueness demonstrated that there was also a difference in stability between in silico and wet lab. The *I* values indicated CelGH5 and ACA61162.1 as unstable enzymes, while the wet lab showed they were stable at broad pH range. The predominant secondary structure was the random coil, with an average percentage of 42.54%. The tertiary structure of four cellulase structures from the metagenomic GH5 family has fulfilled the 3D-protein structure quality based on QMEAN4, ERRAT, Z score, and residues in the favored region on the Ramachandran plot. Glycine, arginine, histidine, glutamic acid, asparagine, tyrosine, and tryptophan were conserved residues found from multiple sequence alignments. This study is significant as a consideration in terms of further isolation, characterization, and selection of a highly efficient cellulases for enhancing enzyme production.

## Abbreviations

GH5: Glycoside hydrolase family 5; CAZY: Carbohydrate-Active Enzymes database; OPEFB: Oil palm empty fruit bunches; GRAVY: Grand average of hydropathicity; pI: Isoelectric point; Ii: Instability index; Al: Aliphatic index; -R: Number of negative residues (aspartic acid and glutamic acid); +R: Number of positive residues (arginine and lysine); aa: Amino acid; MEME: Multiple EM for motif elicitation; SOPMA: Self-Optimized Prediction Method with Alignment; CDD: Conserved Domain Database; CBM: Carbohydrate-binding module; BACON: Bacteroidetes-associated Carbohydrate-binding Often N-terminal; Big: Bacterial immunoglobulin-like domain; QMEAN: Qualitative model energy analysis; MEGA: Molecular Evolutionary Genetics Analysis; NCBI: National Center for Biotechnology Information; PDB: Protein Data Bank; RCSB: Research Collaboratory for Structural Bioinformatics

## Acknowledgements

The authors would like to appreciate the University-CoE-Research Center for Bio-Molecule Engineering, Universitas Airlangga, for supporting the research activity and facilities.

## Authors' contributions

RES, KDAP, AK, AR, and NNTP carried out literature search, survey, performed the computational analysis, and analyzed the data. RES writing original draft preparation. AR validated data. NNTP edited, validated data, finalized the manuscript, and supervised the overall study. All the authors read and approved the final manuscript for publication.

## Funding

The research was funded by UNAIR Research University (Research Group Grant, Universitas Airlangga, Rector Decree 347/UN3.14/PT/2020).

## Availability of data and materials

The authors declare that all generated and analyzed data have been included in the article.

## Declarations

### Ethics approval and consent to participate

Not applicable

### Consent for publication

Not applicable

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Mathematics and Natural Science Study Program, Faculty of Science and Technology, Kampus C Universitas Airlangga, Mulyorejo, Surabaya, East Java 60115, Indonesia. <sup>2</sup>University-CoE-Research Centre for Bio-Molecule Engineering, 2nd Floor ITD Building, Kampus C Universitas Airlangga, Mulyorejo, Surabaya, East Java 60115, Indonesia. <sup>3</sup>Chemistry Education Study Program, Faculty of Teacher Training and Education, Universitas Lambung Mangkurat, Jl. Brigjend. H. Hasan Basry, Banjarmasin, Kalimantan 70123, Indonesia. <sup>4</sup>Department of Health, Faculty of Vocational Studies, Kampus B Universitas Airlangga, Surabaya, East Java 60286, Indonesia. <sup>5</sup>Department of Chemistry, Faculty of Science and Technology, Kampus C Universitas Airlangga, Mulyorejo, Surabaya, East Java 60115, Indonesia.

Received: 25 April 2021 Accepted: 29 August 2021

Published online: 30 September 2021

## References

- Henrissat B, Claeysens M, Tomme P, Lemesle L, Mornon JP (1989) Cellulase families revealed by hydrophobic cluster analysis. *Gene* 81:83–95. [https://doi.org/10.1016/0378-1119\(89\)90339-9](https://doi.org/10.1016/0378-1119(89)90339-9)
- Henrissat B (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* 280:309–316. <https://doi.org/10.1042/bj2800309>
- Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* 293:781–788. <https://doi.org/10.1042/bj2930781>
- Henrissat B, Bairoch A (1996) Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* 316:695–696. <https://doi.org/10.1042/bj3160695>
- Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol* 12:186. <https://doi.org/10.1186/1471-2148-12-186>
- Houfani AA, Anders N, Spiess AC, Baldrian P, Benallaoua S (2020) Insights from enzymatic degradation of cellulose and hemicellulose to fermentable sugars – a review. *Biomass Bioenergy* 134:105481. <https://doi.org/10.1016/j.biombioe.2020.105481>
- Gao M, Li J, Bao Z, Hu M, Nian R, Feng D, An D, Li X, Xian M, Zhang H (2019) A natural in situ fabrication method of functional bacterial cellulose using a microorganism. *Nat Commun* 10:1–10. <https://doi.org/10.1038/s41467-018-07879-3>
- McNamara JT, Morgan JLW, Zimmer J (2015) A molecular description of cellulose biosynthesis. *Annu Rev Biochem* 84:895–921. <https://doi.org/10.1146/annurev-biochem-060614-033930>
- Darsono SM (2014) Pembuatan bioetanol dari lignoselulosa tandan kosong kelapa sawit menggunakan perlakuan awal iradiasi berkas elektron dan NaOH. *J Kim dan Kemasan* 36:245–252. <https://doi.org/10.24817/jkk.v36i2.1891>
- Saini R, Osorio-Gonzalez CS, Hegde K, Brar SK, Magdoui S, Vezina P, Avalos-Ramirez A (2020) Lignocellulosic biomass-based biorefinery: an insight into commercialization and economic standpoint. *Curr Sustain Energy Rep* 7:122–136. <https://doi.org/10.1007/s40518-020-00157-1>
- Maurya DP, Singla A, Negi S (2015) An overview of key pretreatment processes for biological conversion of lignocellulosic biomass to bioethanol. *3 Biotech* 5:597–609. <https://doi.org/10.1007/s13205-015-0279-4>
- Kumari D, Singh R (2018) Pretreatment of lignocellulosic wastes for biofuel production: a critical review. *Renew Sustain Energy Rev* 90:877–891. <https://doi.org/10.1016/j.rser.2018.03.111>
- Ingram T, Wörmeyer K, Lima JCI, Bockemühl V, Antranikian G, Brunner G, Smirnova I (2011) Comparison of different pretreatment methods for lignocellulosic materials. Part I: Conversion of rye straw to valuable products. *Bioresour Technol* 102:5221–5228. <https://doi.org/10.1016/j.biortech.2011.02.005>
- Li S, Yang X, Yang S, Zhu M, Wang X (2012) Technology prospecting on enzymes: application, marketing and engineering. *Comput Struct Biotechnol J* 2:1–11. <https://doi.org/10.5936/CSBJ.201209017>
- Bušić A, Marđetko N, Kundas S, Morzak G, Belskaya H, Ivančić Šantek M, Komes D, Novak S, Šantek B (2018) Bioethanol production from renewable raw materials and its separation and purification: a Review. *Food Technol Biotechnol* 56:289–311. <https://doi.org/10.17113/ftb.56.03.18.5546>
- Kuhad RC, Gupta R, Singh A (2011) Microbial cellulases and their industrial applications. *Enzyme Res* 2011:1–10. <https://doi.org/10.4061/2011/280696>
- Nigam P (2013) Microbial enzymes with special characteristics for biotechnological applications. *Biomolecules* 3:597–611. <https://doi.org/10.3390/biom3030597>
- Thapa S, Li H, O'Hair J, Bhatti S, Chen F-C, Nasr KA, Johnson T, Zhou S (2019) Biochemical characteristics of microbial enzymes and their significance from industrial perspectives. *Mol Biotechnol* 61:579–601. <https://doi.org/10.1007/s12033-019-00187-1>
- Bano A, Chen X, Prasongsuk S, Akbar A, Lotrakul P, Punnapayak H, Anwar M, Sajid S, Ali I (2019) Purification and characterization of cellulase from obligate halophilic *Aspergillus flavus* (TISTR 3637) and its prospects for bioethanol production. *Appl Biochem Biotechnol* 189:1327–1337. <https://doi.org/10.1007/s12010-019-03086-y>
- Sadhu S, Ghosh PK, Aditya G, Maiti TK (2014) Optimization and strain improvement by mutation for enhanced cellulase production by *Bacillus sp.* (MTCC10046) isolated from cow dung. *J King Saud Univ Sci* 26:323–332. <https://doi.org/10.1016/j.jksus.2014.06.001>
- Schmeisser C, Steele H, Streit WR (2007) Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol* 75:955–962. <https://doi.org/10.1007/s00253-007-0945-5>
- Guazzaroni ME, Beloqui A, Golyshin PN, Ferrer M (2009) Metagenomics as a new technological tool to gain scientific knowledge. *World J Microbiol Biotechnol* 25:945–954. <https://doi.org/10.1007/s11274-009-9971-z>

23. Vieites JM, Guazzaroni M-E, Belouqui A, Golyshin PN, Ferrer M (2009) Metagenomics approaches in systems microbiology. *FEMS Microbiol Rev* 33: 236–255. <https://doi.org/10.1111/j.1574-6976.2008.00152.x>
24. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 5:R245–R249. [https://doi.org/10.1016/S1074-5521\(98\)90108-9](https://doi.org/10.1016/S1074-5521(98)90108-9)
25. Schloss PD, Handelsman J (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol* 6:229. <https://doi.org/10.1186/gb-2005-6-8-229>
26. Ravin NV, Mardanov AV, Skryabin KG (2015) Metagenomics as a tool for the investigation of uncultured microorganisms. *Russ J Genet* 51:431–439. <https://doi.org/10.1134/S1022795415050063>
27. Garg R, Srivastava R, Brahma V, Verma L, Karthikeyan S, Sahni G (2016) Biochemical and structural characterization of a novel halotolerant cellulase from soil metagenome. *Sci Rep* 6:1–15. <https://doi.org/10.1038/srep39634>
28. Yang C, Xia Y, Qu H, Li AD, Liu R, Wang Y, Zhang T (2016) Discovery of new cellulases from the metagenome by a metagenomics-guided strategy. *Biotechnol Biofuels* 9:138. <https://doi.org/10.1186/s13068-016-0557-3>
29. Alvarez TM, Paiva JH, Ruiz DM, Cairo JPL, Pereira IO, Paixão DAA, De Almeida RF, Tonoli CCC, Ruller R, Santos CR, Squina FM, Murakami MT (2013) Structure and function of a novel cellulase 5 from sugarcane soil metagenome. *PLoS One* 8:1–9. <https://doi.org/10.1371/journal.pone.0083635>
30. Alves LDF, Westmann CA, Lovate GL, De Siqueira GMV, Borelli TC, Guazzaroni ME (2018) Metagenomic approaches for understanding new concepts in microbial science. *Int J Genomics* 2018:1–15. <https://doi.org/10.1155/2018/2312987>
31. Missa H, Susilowati A, Setyaningsih R (2016) Diversity and phylogenetic relationship of cellulolytic bacteria from the feces of Bali Cattle in South Central Timor, East Nusa Tenggara, Indonesia. *Biodiversitas* 17:614–619. <https://doi.org/10.13057/biodiv/d170232>
32. Pimentel AC, Ematsu GCG, Liberato MV, Paixão DAA, Franco Cairo JPL, Mandelli F, Tramontina R, Gandin CA, de Oliveira NM, Squina FM, Alvarez TM (2017) Biochemical and biophysical properties of a metagenome-derived GH5 endoglucanase displaying an unconventional domain architecture. *Int J Biol Macromol* 99:384–393. <https://doi.org/10.1016/j.ijbiomac.2017.02.075>
33. Zarafeta D, Kassis D, Sayer C, Gudbergsdottir SR, Ladoukakis E, Isupov MN, Chatziannou A, Peng X, Littlechild JA, Skretas G, Kolis FN (2016) Discovery and characterization of a thermostable and highly halotolerant GH5 cellulase from an Icelandic hot spring isolate. *PLoS One* 11:1–18. <https://doi.org/10.1371/journal.pone.0146454>
34. Lugani Y, Sooch BS (2017) *In silico* characterization of cellulases from genus *Bacillus*. *Int J Curr Res Rev* 9:3–10. <https://doi.org/10.7324/IJCRR.2017.9136>
35. Hoda A, Tafaj M, Sallaku E (2021) *In silico* structural, functional and phylogenetic analyses of cellulase from *Ruminococcus albus*. *J Genet Eng Biotechnol* 19:58. <https://doi.org/10.1186/s43141-021-00162-x>
36. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein analysis tools on the ExPASy server. In: Walker JM (ed) *The Proteomics Protocols Handbook*. Humana Press, New Jersey, pp 571–607
37. Santos SP, Banderas TM, Pinto AF, Teixeira M, Carrondo MA, Romão CV (2012) Thermofluor-based optimization strategy for the stabilization and crystallization of *Campylobacter jejuni* desulfurubryethrin. *Protein Expr Purif* 81:193–200. <https://doi.org/10.1016/j.pep.2011.10.001>
38. Shoemaker KM, Moisaner PH (2015) Microbial diversity associated with copepods in the North Atlantic subtropical gyre. *FEMS Microbiol Ecol* 91:1–11. <https://doi.org/10.1093/femsec/fiv064>
39. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME suite: tools for motif discovery and searching. *Nucleic Acids Res* 37:202–208. <https://doi.org/10.1093/nar/gkp335>
40. Combet C, Blanchet C, Geourjon C, Deléage G (2000) NPS@: network protein sequence analysis. *Trends Biochem Sci* 25:147–150. [https://doi.org/10.1016/S0968-0004\(99\)01540-6](https://doi.org/10.1016/S0968-0004(99)01540-6)
41. Pramanik K, Kundu S, Banerjee S, Ghosh PK, Maiti TK (2018) Computational-based structural, functional and phylogenetic analysis of *Enterobacter* phytases. *3 Biotech* 8:1–12. <https://doi.org/10.1007/s13205-018-1287-y>
42. Benkert P, Künzli M, Schwede T (2009) QMEAN server for protein model quality estimation. *Nucleic Acids Res* 37:W510–W514. <https://doi.org/10.1093/nar/gkp322>
43. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* 2:1511–1519. <https://doi.org/10.1002/pro.5560020916>
44. Dutta B, Banerjee A, Chakraborty P, Bandopadhyay R (2018) *In silico* studies on bacterial xylanase enzyme: structural and functional insight. *J Genet Eng Biotechnol* 16:749–756. <https://doi.org/10.1016/j.jgeb.2018.05.003>
45. Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27: 343–350. <https://doi.org/10.1093/bioinformatics/btq662>
46. Bhattacharya M, Hota A, Kar A, Sankar Chini D, Chandra Malick R, Chandra Patra B, Kumar Das B (2018) *In silico* structural and functional modelling of antifreeze protein (AFP) sequences of ocean pout (*Zoarces americanus*, Bloch & Schneider 1801). *J Genet Eng Biotechnol* 16:721–730. <https://doi.org/10.1016/j.jgeb.2018.08.004>
47. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res* 48:D265–D268. <https://doi.org/10.1093/nar/gkz991>
48. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF (2005) Protein identification and analysis tools in the ExPASy server. In: Walker JM (ed) *The Proteomics Protocols Handbook*. Humana Press, New Jersey, pp 531–552
49. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539. <https://doi.org/10.1038/msb.2011.75>
50. Sievers F, Higgins DG (2014) Clustal omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol* 1079:105–116. [https://doi.org/10.1007/978-1-62703-646-7\\_6](https://doi.org/10.1007/978-1-62703-646-7_6)
51. Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* 27:135–145. <https://doi.org/10.1002/pro.3290>
52. Robert X, Gouet P (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* 42:W320–W324. <https://doi.org/10.1093/nar/gku316>
53. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* 8:275–282. <https://doi.org/10.1093/bioinformatics/8.3.275>
54. Kumar S, Stecher G, Li M, Nkryaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35: 1547–1549. <https://doi.org/10.1093/molbev/msy096>
55. Yakimov AP, Afanaseva AS, Khodorkovskiy MA, Petukhov MG (2016) Design of stable alpha-helical peptides and thermostable proteins in biotechnology and biomedicine. *Acta Naturae* 8:70–81. <https://doi.org/10.32607/20758251-2016-8-4-70-81>
56. Ericsson UB, Hallberg BM, DeTitta GT, Dekker N, Nordlund P (2006) Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal Biochem* 357:289–298. <https://doi.org/10.1016/j.jab.2006.07.027>
57. Dutta B, Deska J, Bandopadhyay R, Shamekh S (2021) *In silico* characterization of bacterial chitinase: illuminating its relationship with archaeal and eukaryotic cousins. *J Genet Eng Biotechnol* 19:19. <https://doi.org/10.1186/s43141-021-00121-6>
58. Delsaute M, Berlemont R, Dehareng D, Van Elder D, Galleni M, Bauvois C (2013) Three-dimensional structure of RBcel1, a metagenome-derived psychrotolerant family GH5 endoglucanase. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 69:828–833. <https://doi.org/10.1107/S1744309113014565>
59. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* 35:W407–W410. <https://doi.org/10.1093/nar/gkm290>
60. Davies G, Henrissat B (1995) Structures and mechanisms of glycosyl hydrolases. *Structure* 3:853–859. [https://doi.org/10.1016/S0969-2126\(01\)00220-9](https://doi.org/10.1016/S0969-2126(01)00220-9)
61. Davies GJ, Henrissat B (2002) Structural enzymology of carbohydrate-active enzymes: implications for the post-genomic era. *Biochem Soc Trans* 30:291–297. <https://doi.org/10.1042/bst0300291>
62. Kantardjiev KA, Rupp B (2004) Protein isoelectric point as a predictor for increased crystallization screening efficiency. *Bioinformatics* 20:2162–2168. <https://doi.org/10.1093/bioinformatics/bth066>
63. Graziano G, Merlino A (2014) Molecular bases of protein halotolerance. *Biochim Biophys Acta - Proteins Proteomics* 1844:850–858. <https://doi.org/10.1016/j.bbapap.2014.02.018>
64. Santos CR, Paiva JH, Sforça ML, Neves JL, Navarro RZ, Cota J, Akao PK, Hoffmam ZB, Meza AN, Smetana JH, Nogueira ML, Polikarpov I, Xavier-Neto

- J, Squina FM, Ward RJ, Ruller R, Zeri AC, Murakami MT (2012) Dissecting structure-function-stability relationships of a thermostable GH5-CBM3 cellulase from *Bacillus subtilis* 168. *Biochem J* 441:95–104. <https://doi.org/10.1042/BJ20110869>
65. Duan CJ, Xian L, Zhao GC, Feng Y, Pang H, Bai XL, Tang JL, Ma QS, Feng JX (2009) Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. *J Appl Microbiol* 107:245–256. <https://doi.org/10.1111/j.1365-2672.2009.04202.x>
  66. Gamage DG, Gunaratne A, Periyannan GR, Russell TG (2019) Applicability of instability index for in vitro protein stability prediction. *Protein Pept Lett* 26: 339–347. <https://doi.org/10.2174/0929866526666190228144219>
  67. Huynh K, Partch CL (2015) Analysis of protein stability and ligand interactions by thermal shift assay. *Curr Protoc Protein Sci* 79:28.9.1–28.9.14. <https://doi.org/10.1002/0471140864.ps2809s79>
  68. Biro JC (2006) Amino acid size, charge, hydrophobicity indices and matrices for protein structure analysis. *Theor Biol Med Model* 3:15. <https://doi.org/10.1186/1742-4682-3-15>
  69. Kyte J, Doolittle RF (1982) A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 157:105–132. [https://doi.org/10.1016/0022-2836\(82\)90515-0](https://doi.org/10.1016/0022-2836(82)90515-0)
  70. Mohammed A, Guda C (2011) Computational approaches for automated classification of enzyme sequences. *J Proteomics Bioinforma* 4:147–152. <https://doi.org/10.4172/jpb.1000183>
  71. Vidhya VG, Swarnalatha Y, Bhaskar A (2018) An *in silico* analysis of physicochemical characterization and protein-protein interaction network analysis of human anti-apoptotic proteins. *Asian J Pharm* 12:51397–51407
  72. Fujiwara K, Toda H, Ikeguchi M (2012) Dependence of  $\alpha$ -helical and  $\beta$ -sheet amino acid propensities on the overall protein fold type. *BMC Struct Biol* 12: 18. <https://doi.org/10.1186/1472-6807-12-18>
  73. Melnikov S, Mailliot J, Rigger L, Neuner S, Shin B, Yusupova G, Dever TE, Micura R, Yusupov M (2016) Molecular insights into protein synthesis with proline residues. *EMBO Rep* 17:1776–1784. <https://doi.org/10.15252/embr.201642943>
  74. Richardson JS, Richardson DC (2002) Natural  $\beta$ -sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A* 99: 2754–2759. <https://doi.org/10.1073/pnas.052706099>
  75. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242. <https://doi.org/10.1093/nar/28.1.235>
  76. Okano H, Kanaya E, Ozaki M, Angkawidjaja C, Kanaya S (2015) Structure, activity, and stability of metagenome-derived glycoside hydrolase family 9 endoglucanase with an N-terminal Ig-like domain. *Protein Sci* 24:408–419. <https://doi.org/10.1002/pro.2632>
  77. Saleem A, Rajput S (2020) Insights from the *in silico* structural, functional and phylogenetic characterization of canine lysyl oxidase protein. *J Genet Eng Biotechnol* 18:20. <https://doi.org/10.1186/s43141-020-00034-w>
  78. Gao X, Dong X, Li X, Liu Z, Liu H (2020) Prediction of disulfide bond engineering sites using a machine learning method. *Sci Rep* 10:10330. <https://doi.org/10.1038/s41598-020-67230-z>
  79. Savojardo C, Fariselli P, Alhamdoosh M, Luigi Martelli P, Pierleoni A, Casadio R (2011) Structural bioinformatics improving the prediction of disulfide bonds in eukaryotes with machine learning methods and protein subcellular localization. *Bioinformatics* 27:2224–2230. <https://doi.org/10.1093/bioinformatics/btr387>
  80. Yang J, He BJ, Jang R, Zhang Y, Shen H (2015) Accurate disulfide-bonding network predictions improve ab initio structure prediction of cysteine-rich proteins. *Bioinformatics* 31:3773–3781. <https://doi.org/10.1093/bioinformatics/btv459>
  81. Wiedemann C, Kumar A, Lang A, Ohlenschläger O (2020) Cysteines and disulfide bonds as structure-forming units: insights from different domains of life and the potential for characterization by NMR. *Front Chem* 8:280. <https://doi.org/10.3389/fchem.2020.00280>
  82. Griffo A, Rooijackers BJM, Hähl H, Jacobs K, Linder MB, Laaksonen P (2019) Binding forces of cellulose binding modules on cellulosic nanomaterials. *Biomacromolecules* 20:769–777. <https://doi.org/10.1021/acs.biomac.8b01346>
  83. Mello LV, Chen X, Rigden DJ (2010) Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *FEBS Lett* 584:2421–2426. <https://doi.org/10.1016/j.febslet.2010.04.045>
  84. Wojciechowski M, Różycki B, Huy PDQ, Li MS, Bayer EA, Cieplak M (2018) Dual binding in cohesin-dockerin complexes: the energy landscape and the role of short, terminal segments of the dockerin module. *Sci Rep* 8:5051. <https://doi.org/10.1038/s41598-018-23380-9>
  85. Pinheiro BA, Proctor MR, Martinez-Fleites C, Prates JAM, Mon VA, Davies GJ, Bayer EA, Fontes CMGA, Fierobe HP, Gilbert HJ (2008) The *Clostridium cellulolyticum* dockerin displays a dual binding mode for its cohesin partner. *J Biol Chem* 283:18422–18430. <https://doi.org/10.1074/jbc.M801533200>
  86. Juturu V, Wu JC (2014) Microbial cellulases: engineering, production and applications. *Renew Sustain Energy Rev* 33:188–203. <https://doi.org/10.1016/j.rser.2014.01.077>
  87. Bayer EA, Belaich J-P, Shoham Y, Lamed R (2004) The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* 58:521–554. <https://doi.org/10.1146/annurev.micro.57.03.0502.091022>
  88. Hüttner M, Prieto A, Aznar S, Bernabeu M, Glaría E, Valledor AF, Paytubi S, Merino S, Tomás J, Juárez A (2019) Expression of a novel class of bacterial Ig-like proteins is required for InChI plasmid conjugation. *PLoS Genet* 15: e1008399. <https://doi.org/10.1371/journal.pgen.1008399>
  89. Song Y, Lee K, Baek J, Kim M, Kwon M, Kim Y, Park M, Ko H, Lee J, Kim K (2017) Isolation and characterization of a novel endo- $\beta$ -1,4-glucanase from a metagenomic library of the black-goat rumen. *Brazilian J Microbiol* 48:801–808. <https://doi.org/10.1016/j.bjbm.2017.03.006>
  90. Rohman A, Dijkstra BW, Puspangsih NNT (2019)  $\beta$ -xylosidases: structural diversity, catalytic mechanism, and inhibition by monosaccharides. *Int J Mol Sci* 20:7–11. <https://doi.org/10.3390/ijms20225524>
  91. Yan J, Liu W, Li Y, Lai H-L, Zheng Y, Huang J-W, Chen C-C, Chen Y, Jin J, Li H, Guo R-T (2016) Functional and structural analysis of *Pichia pastoris*-expressed *Aspergillus niger* 1,4- $\beta$ -endoglucanase. *Biochem Biophys Res Commun* 475:8–12. <https://doi.org/10.1016/j.bbrc.2016.05.012>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen® journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)