MALAYSIAN JOURNAL OF SCIENCE
MJS
ISSN 1394 - 3065

FAST-UTHM Global Nexus Conference 2024

Special issue

Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)

https://mjs.um.edu.my

In Silico Characterization of Selected Carbohydrate Active Enzymes (CAZymes) from Malaysia's Unique Genome Assemblies Through Genomic Data Mining

Alvin Owen Miharil@Jukiri[1a], Aisyah Mohamed Rehan[2a*], Azzmer Azzar Abdul Hamid[3b] and Kamarul Rahim Kamarudin[4c]

Abstract: In silico characterization of protein is usually performed to elucidate the structure and function of protein targets prior to experimental approach, which helps to narrow down target candidates, and reduce overall time, cost and effort. Carbohydrate-active enzymes (CAZymes) are enzymes involved in conversion of biomass to industrially valuable products. More research is needed to characterise CAZymes from Malaysia. This study aims to elucidate CAZymes from selected genome assemblies isolated in Malaysia by in silico genomic data mining. The National Center for Biotechnology Information database is utilized to investigate Malaysia's genome assemblies, discovering 1386 complete genomes. From 2019 to 2024, there are 34 bacterial genomes isolated by Malaysian researchers. CAZymes from six bacterial genomes were categorised. Three CAZymes from *Thermobifida fusca* were chosen for subsequent protein modelling due to their acceptable homology range to available protein structures. To characterise the structural and functional properties of these CAZymes, online protein modelling tools were used (dbCAN3 meta server, I-TASSER and ExPASy SWISS-MODEL), as well as to evaluate the quality of the protein models (Ramachandran plot assessment, Verify3D and QMEAN4 score). Then, the ligand-binding sites inside each CAZyme sequence were predicted using I-TASSER's COACH algorithm. Three target CAZymes annotated as endo-1,4-beta-xylanase, amylo-alpha-1,6-glucosidase and endoglucanase respectively have been modelled, their protein model assessed for its quality, and their active sites and potential ligand predicted. Through in-silico characterization, the proteins' structure and function have been elucidated. Future studies, employing refined tertiary models, protein-ligand docking, and molecular dynamics simulations, aim to identify novel inhibitors or activators. This will facilitate experimental validation and unlock diverse industrial biotechnology applications. Potential benefits include enhanced enzyme catalysis for biofuel production, development of novel biopharmaceuticals, and optimization of metabolic pathways for sustainable chemical synthesis. These findings could significantly impact industrial processes and contribute to bio-based economies.

Keywords: CAZymes, In-silico Protein Modeling, Genomic Data Mining

1. Introduction

Carbohydrates are a major class of large biopolymers found in all cells, along with nucleic acids, proteins, and lipids. Carbohydrate-active enzymes (CAZymes) are enzymes that can degrade, modify, or create glycosidic bonds. These enzymes are essential catalysts in numerous biological processes, driving the breakdown and synthesis of complex carbohydrates. CAZymes are divided into several classes and are secreted by microorganisms, allowing the microbial conversion of plant biomass into industrially valuable products like bioenergy and biofuel production (Wardman et al., 2022). This microbial activity complements the plant's own carbohydrate processing pathways. Plants use photosynthesis to convert carbon dioxide and water into sugars, which are further converted into carbohydrates like starches and cellulose (Sharkey, 2024). CAZyme genes are abundant in plant and plant-degrading microbes' genomes. Precursors produced by these enzymes can be used to generate bio-based products like food, paper, textiles, animal feed, and biofuels (Bains et al., 2024).

CAZymes are crucial in bioenergy, agriculture, human nutrition, and disease prevention. They break down complex carbohydrates in the gut microbes, allowing them to be absorbed by the intestinal epithelium (Lee et al., 2022). CAZymes are also involved in biological processes underlying diseases like cancer, diabetes, Alzheimer's, and AIDS (Ye et al., 2022; Nin-Hill et al., 2023). This broad range of functional roles underscores the importance of exploring CAZymes from diverse sources, with significant scientific and industrial interest.

Due to Malaysia's biodiversity, bioprospecting and discovering new products are extensively performed by scientists and commercial companies. Bioinformatic tools are utilized to enable these discoveries. The CAZy database, dbCAN3 meta server (Zheng et al., 2023), and PlantCAZyme database have

Authors information:

[a]Department of Chemical Engineering Technology, Faculty of Engineering Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Education Hub, KM 1, Jalan Panchor, 86400, Panchor, Johor, MALAYSIA. E-mail: an200042@student.uthm.edu.my[1], aisyahr@uthm.edu.my[2]

[b]Department of Biotechnology, Kulliyyah of Science, International Islamic University Malaysia, Kuantan Campus, Bandar Indera Mahkota, 25200 Kuantan, Pahang, MALAYSIA. E-mail: azzmer@iium.edu.my[3]

[c]Department of Technology and Natural Resources, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Education Hub, KM 1, Jalan Panchor, 84600 Panchor, Johor, MALAYSIA. E-mail: kamarulr@uthm.edu.my[4]

**Corresponding Author: aisyahr@uthm.edu.my*

Received: May 25, 2025
Accepted: July 30, 2025
Published: June 30, 2026

DOI: https://doi.org/10.22452/mjs.vol45sp1.8
Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)

assisted in the genome mining of CAZymes from bacteria, fungi, viruses, and plants (Zhang et al., 2018; Yang et al., 2024). Malaysia's unique genome assembly has led to the discovery of a novel beta-glucosidase from *Jeotgalibacillus Malaysiensis*, a marine bacterium (Liew et al., 2018; Godse et al., 2021). This is similar to a genomic mining study on *Glaciozyma antarctica*, a psychrophilic yeast from Antarctica (Nooraisyah et al., 2015). Over the years, advances have been made to unearth the potential of psychrophilic enzymes studied from this yeast and its cold adaptation strategies (Yusof et al., 2021).

While CAZymes are implicated in diverse fields, including bioenergy and disease, their potential within Malaysia's unique biodiversity remains largely unexplored. Existing databases like CAZy and dbCAN3 facilitate genome mining, yet a significant gap exists in the in-silico elucidation of novel CAZymes from Malaysian genome assemblies. This research aims to provide in-silico characterization of CAZymes from selected Malaysian genome assemblies, offering new insights for industrial biotechnology and biomedical applications, leveraging Malaysia's untapped genomic resources.

2. Materials and methods

2.1 Genome Sequence Assembly

In this research, the main approach used was reference-based genome assembly. The complete genome sequence assembly for bacteria, fungi, virus, and plant division submitted by Malaysian depositors in the NCBI database and NCBI GenBank between 2019 and until the end of 2024 were used (Sayers et al., 2022). Filters were activated to exclude anomalous and partial assembly. Animal genome assemblies were also excluded. Protein FASTA sequences for each assembly were downloaded.

2.2 CAZymes annotation

Protein FASTA sequences from each assembly were then submitted to the dbCAN3 meta server to annotate the proteins using HMMER and DIAMOND via CAZy, dbCAN, and PPR tools, respectively (Zhang et al., 2018; Sayers et al., 2022). The results from the annotation tools on the dbCAN3 meta server were retrieved from the dbCAN3 results page (Zheng et al., 2023).

2.3 Selection of target Novel CAZymes

Genome assemblies for potential novel CAZymes, which are crucial for understanding glucose metabolism, carbohydrate breakdown, and organisms' interactions with carbohydrates, were selected.

2.4 Protein modelling of target Novel CAZymes

3D protein modeling will be used to determine the structure of CAZymes that lack existing structural data. BLASTp will be used to check the availability of structural homologs in Protein Data Bank (PDB). The protein tertiary structure will be predicted using three different server which were I-TASSER (Zhou et al., 2022) and ExPASy SWISS-MODEL (Waterhouse et al., 2024). All three predicted structures will then be validated using Ramachandran plot assessment, Verify3D and QMEAN4 score. From the validation, the best predicted structure will be selected for structural refinement. Model structure refinement will be performed using ModRefiner algorithm tool from I-TASSER database (Zhou et al., 2022).

2.5 Ligand prediction & docking experiment

Ligand prediction in CAZymes involves predicting potential substrates with which enzymes may bind or act. Machine-learning-based ligand predictors use techniques like support vector machines (SVM) and decision trees to identify critical residues, geometries, and conformations. Molecular docking is used to visualize binding modes and determine binding affinities. Popular docking applications like AutoDock, GOLD, and Glide help researchers understand enzyme-ligand pairings, leading to the design or optimization of new CAZymes for specific applications. This research will use COACH algorithm from I-TASSER server suite that generates ligand binding-site predictions by matching the target models with proteins in the BioLiP database (Zhou et al., 2022).

3. Results and discussions

3.1 Genome assemblies deposited by Malaysian researcher

The NCBI database is a valuable tool for acquiring genome sequence assemblies focusing on Malaysia's unique genetic composition. Researchers can access a unique reservoir of genetic data, providing insights into the region's biodiversity. A search using the NCBI database server showed 1386 complete genome assemblies by Malaysian scientists, with filters used to narrow the search and a refined selection process focusing on bacterial genomes. The search resulted in 34 complete bacterial genome assemblies, one from viruses, and zero from plants (Table 1, Sayers et al., 2022).

3.2 Annotation of CAZymes families

The study focused on bacterial genome assemblies due to time constraints and their potential as future CAZyme producers. Six randomly chosen bacterial genomes were analyzed using the NCBI dataset to identify potential advantages (Table 2). Direct NCBI annotation was used to retrieve and calculate CAZyme numbers within the genome assemblies. The dbCAN3 database was used to confirm NCBI annotations (Zheng et al., 2023). The NCBI genome assembly search was filtered for complete genome assemblies, resulting in complete annotations for each genome sequence. However, compiling annotations was challenging due to different specific CAZyme names and families (Sayers et al., 2022).

Only two genomes, *Thermobifida fusca* and *Parageobacillus caldoxylosilyticus*, were chosen for structure and function prediction after analysis and risk assessment of their CAZyme annotations. The exclusion criteria were to remove bacteria that is pathogenic to humans or can cause infection to human. This is critical for handling researchers' safety as well as for assessing research feasibility with low risk. These chosen genomes, recognised for their efficient breakdown abilities and resistance to high temperatures, respectively, provide potential avenues for industrial and environmental biotechnology advancements.

DOI: https://doi.org/10.22452/mjs.vol45sp1.8
Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)

Table 1. List of 34 genome sequence assemblies with complete genome

Genome Assembly ID:		Species name:
1.	**ASM1503458v1**	*Thermobifida fusca* (high G+C Gram-positive bacteria)
2.	**ASM1927293v1**	*Parageobacillus caldoxylosilyticus* (firmicutes)
3.	**ASM2520084v1**	*Pseudomonas aeruginosa* (g-proteobacteria)
4.	**ASM159764v2**	*Mycobacterium tuberculosis* (high G+C Gram-positive bacteria)
5.	**ASM1907668v1**	*Acinetobacter baumannii* (g-proteobacteria)
6.	**ASM199636v2**	*Vibrio parahaemolyticus* (g-proteobacteria)
7.	**ASM3070428v1**	*Vibrio parahaemolyticus* (g-proteobacteria)
8.	**ASM2542621v1**	*Stenotrophomonas maltophilia* (g-proteobacteria
9.	**ASM967662v1**	*Limosilactobacillus fermentum* (firmicutes)
10.	**ASM1735231v1**	*Priestia megaterium* (firmicutes)
11.	**ASM2954238v1**	*Streptococcus suis* (firmicutes)
12.	**UKMPMC2000**	*Burkholderia pseudomallei* (b-proteobacteria)
13.	**UKMD286**	*Burkholderia pseudomallei* (b-proteobacteria)
14.	**ASM1421793v1**	*Shigella sonnei* (enterobacteria)
15.	**ASM2429712v1**	*Streptomyces cavourensis* (high G+C Gram-positive bacteria)
16.	**ASM2458281v1**	*Acinetobacter colistiniresistens* (g-proteobacteria)
17.	**ASM3207508v1**	*Komagataeibacter nataicola* (a-proteobacteria)
18.	**ASM970730v1**	*Rhodococcus sp. AQ5-07* (high G+C Gram-positive bacteria)
19.	**ASM2580919v1**	*Paenibacillus sp. PSB04* (firmicutes)
20.	**ASM3132634v1**	*Aureispira sp. CCB-E* (CFB group bacteria)
21.	**ASM2241457v2**	*Streptomyces sp. MUM 178J* (high G+C Gram-positive bacteria)
22.	**ASM3277389v1**	*Teredinibacter sp. KSP-S5-2* (g-proteobacteria)
23.	**ASM2054666v1**	*Marinobacter sp. CA1* (g-proteobacteria)
24.	**ASM432894v1**	*Aquitalea sp. USM4* (b-proteobacteria)
25.	**ASM2615116v2**	*Cryobacterium sp. SO2* (high G+C Gram-positive bacteria)
26.	**ASM421021v2**	*Cryobacterium sp. SO1* (high G+C Gram-positive bacteria)
27.	**ASM970728v1**	*Arthrobacter sp. AQ5-05* (high G+C Gram-positive bacteria)
28.	**ASM2996405v1**	*Arthrobacter sp. EM1* (high G+C Gram-positive bacteria)
29.	**ASM2315155v2**	*Photobacterium sp. CCB-ST2H9* (g-proteobacteria)
30.	**ASM1340350v1**	*Acinetobacter baumannii* (g-proteobacteria)
31.	**UKMR15**	*Burkholderia pseudomallei* (b-proteobacteria)
32.	**ASM2995351v1**	*Bacillus altitudinis* (firmicutes)
33.	**ASM2124932v1**	*Mycolicibacterium fortuitum* (high G+C Gram-positive bacteria)
34.	**ASM452194v2**	*Streptomyces sp. S6* (high G+C Gram-positive bacteria)

3.3 CAZymes from Malaysia's unique genome assemblies

This study analyzed target genome assemblies for potential novel CAZymes, which are crucial for understanding glucose metabolism, carbohydrate breakdown, and organisms' interactions with carbohydrates. Protein modelling of these CAZymes was performed using in silico tools like I-TASSER (Zhou et al., 2022) and ExPASy SWISS-MODEL (Waterhouse et al., 2024) for protein structure prediction and analysis. The Ramachandran plot evaluation, Verify3D, and QMEAN4 score were used for quality validation of the structural models, ensuring a reliable protein model for active sites and ligand binding site prediction. This research could aid in developing new diagnostic tools for infectious illnesses and microbial-related ailments.

DOI: https://doi.org/10.22452/mjs.vol45sp1.8
Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)

Table 2. List of six genome sequence assemblies annotated based on their CAZymes

ID no.	Species Name	CAZymes families* (number of gene)					Benefits
		GHs	GTs	PLs	CEs	AAs	
ASM1503458v1	*Thermobifida fusca*	14	30	0	0	1	• Decomposition of Organic Matter • Production of Cellulases • Biotechnological Applications
ASM1927293v1	*Parageobacillus caldoxylosilyticus*	13	17	0	0	0	• High-Temperature Applications • Enzyme Production Xylanolytic Activity
ASM2520084v1	*Pseudomonas aeruginosa*	4	24	0	0	0	• Biodegradation • Antimicrobial Properties • Research Value
ASM159764v2	*Mycobacterium tuberculosis*	6	22	2	4	0	• Improved Diagnosis and Treatment • Public Health Benefits
ASM1907668v1	*Acinetobacter baumannii*	1	14	6	2	0	• Understanding Antibiotic Resistance • Importance in Infection Control
ASM3070428v1	*Vibrio parahaemolyticus*	4	23	15	1	0	• Improved Food Safety • Importance of Public Health Surveillance

*CAZymes families abbreviations: Glycoside Hydrolases= GHs, GlycosylTransferases = GTs, Polysaccharide Lyases = PLs, Carbohydrate esterases = Ces, Auxiliary Activities = Aas

3.4 Selection of target novel CAZymes

Research on novel Carbohydrate-Active Enzymes (CAZymes) for organisms like *Thermobifida fusca* and *Parageobacillus caldoxylosilyticus* is crucial for industrial biotechnology. These enzymes exhibit specific capabilities for breaking down complex polysaccharides, leading to advancements in biofuel production, waste management, and environmentally friendly solutions, leveraging microorganisms' natural proficiency in biomass degradation.

Thermophilic cellulolytic bacteria, like *Thermobifida fusca*, have potential industrial and biotechnological uses. Malaysian scientists analyzed the species' genome sequence, revealing 112 carbohydrate-active enzyme genes, including 40 glycoside hydrolases (GHs), 36 glycosyltransferases (GTs), 22 carbohydrate-binding modules (CBMs), nine carbohydrate esterases (CEs), three polysaccharide lyases (PLs) and two auxiliary activities (AAs) (Yildiz et al., 2022). Two genes for lytic polysaccharide monooxygenase (LPMO) were also discovered. The ninth genome was used to investigate potential industrial and agricultural uses. Three novel CAZymes were selected from 109: xylanase, glucosidase, and endoglucanase. These enzymes are essential in lignocellulose-degrading microorganisms, breaking down complex plant polysaccharides into simpler sugars. Their synergistic activity is crucial for effective biomass deconstruction, a crucial step in bioeconomy's efforts to exploit renewable resources (Nargotra et al., 2022).

Parageobacillus caldoxylosilyticus is a unique species from the thermophilic genus Parageobacillus (Wang et al., 2022). Because of its high adaptability towards temperature, this bacterium is a promising candidate to produce industrially applicable thermostable enzymes, such as xylanase and maltogenic amylase (Wang et al., 2022). This makes it an interesting subject for researchers studying bioconversion processes and biofuel generation. It also has potential biotechnological uses beyond biofuel production, such as bioethanol and biochemical synthesis. However, due to a lack of genomic knowledge, the extent of its powers and underlying genetic pathways is still in its infancy. *Parageobacillus thermoglucosidasius*, with high sequence similarities to *P. caldoxylosilyticus*, possesses more research findings on its genome assembly, offering potential research and application opportunities (Wang et al., 2022).

DOI: https://doi.org/10.22452/mjs.vol45sp1.8
Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)

Table 3. The Carbohydrate hydrolysis enzymes (GH) of T. fusca

GH family	Product size (No of amino acid)	Protein/enzymes	Accession number
GH10	399	Endo-1,4-beta-xylanase	QOS59282.1
GH176	712	Amylo-alpha-1,6-glucosidase	QOS59445.1
GH6, CBM2	441	Endoglucanase	QOS57731.1

3.5 Protein modelling target novel CAZymes

Protein modelling is being used to study novel Carbohydrate-Active enzymes (CAZymes), providing a new approach to understanding and using these biological catalysts for biofuel production and medicine creation. Scientists use advanced bioinformatics methods to make 3D modelling predictions, using platforms like I-TASSER and ExPASy SWISS-MODEL. The models' correctness and dependability are assessed using computational tests like Ramachandran plot analysis, Verify3D, and QMEAN4 score (Zhou et al., 2022; Waterhouse et al., 2024).

3.6 Identification of target CAZymes genes on Thermobifida fusca

Researchers have identified Carbohydrate-Active Enzymes (CAZymes) in *Thermobifida fusca*, a bacterium known for its cellulolytic capabilities. Three CAZymes, xylanase, glucosidase, and endoglucanase, were chosen based on their high homologous fraction to well-characterized enzymes (Table 3). This discovery opens up new possibilities for biofuel generation and plant biomass biodegradation, enhancing environmental sustainability and renewable energy sources.

3.7 Determination of protein model quality

Protein models created using computational techniques are crucial for predicting active sites and simulating ligand binding. Verify3D, a tool on ExPASy SWISS-MODEL, helps scientists examine the structural integrity and functional usefulness of modelled CAZymes by comparing their 3D atomic models' suitability and correctness to their amino acid sequences. The correctness is determined by the environments of amino acid residues, such as the area of residue buried within the model, the fraction of side chain area covered by polar atoms, and the local secondary structure. This method allows scientists to verify structural predictions and ensure enzymes function like biological counterparts. By confirming the structural validity of these models, strengthens the basis for subsequent analyses, such as ligand docking and molecular dynamics simulations, ultimately facilitating the development of biotechnological and industrial applications. Figure 1 and Figure 2 shows the 3D modelling output for each side of the CAZymes.

3.8 Ramanchandran plot assessment

The Ramachandran plot is crucial for understanding the conformational dynamics of carbohydrate-active enzymes (CAZymes), which perform various biological tasks. By studying phi and psi torsion angles within CAZymes, researchers can determine structural flexibility and rigidity for substrate

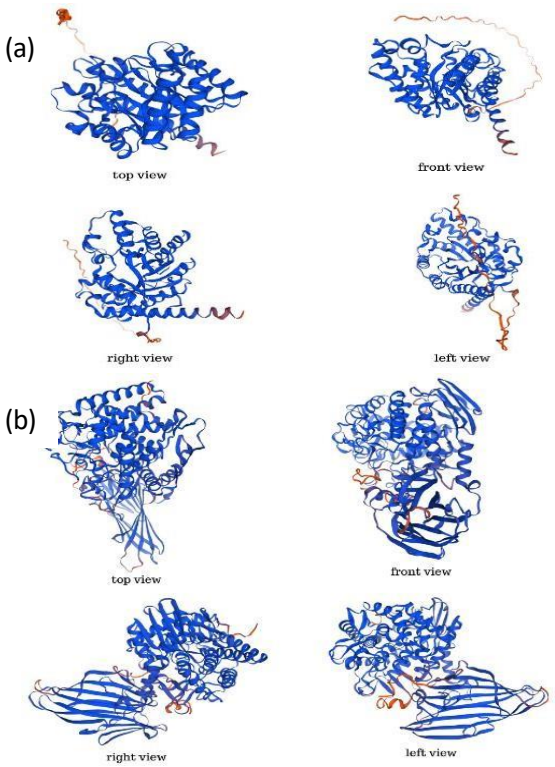

Figure 1. 3D modelling of endo-1,4-beta-xylanase (a) and 3D modelling of amylo-alpha-1,6-glucosidase (b)

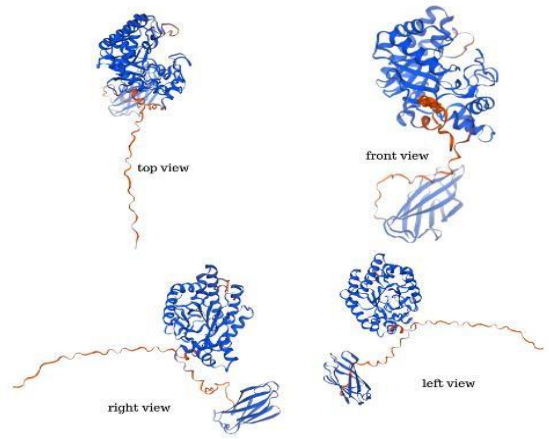

Figure 2. 3D modelling of Endoglucanase

identification and catalytic activity. This analytical technique has enabled the discovery of structure-function correlations within

DOI: https://doi.org/10.22452/mjs.vol45sp1.8
Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)

CAZyme families, thereby optimizing their applications in industrial and biotechnological processes. Figure 3 presents the Ramachandran plots for each modeled CAZyme, with detailed statistical summaries provided in Table 4. Notably, the xylanase model, depicted in Figure 3(a), exhibits the most favorable Ramachandran plot characteristics, demonstrating a lower number of outlier residues compared to the other CAZymes.

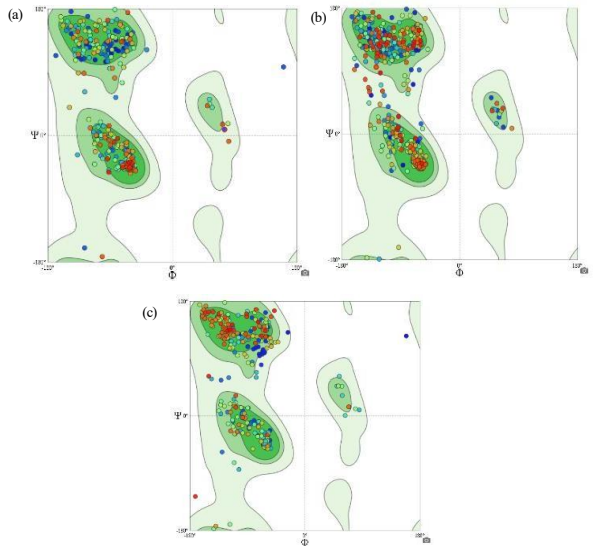

Figure 3. Ramachandran plot for Xylanase (endo-1,4-beta-xylanase) (a), Glucosidase (amylo-alpha-1,6-glucosidase) (b), and Endoglucanase (c)

This observation indicates a more energetically favorable arrangement of amino acid residues in the xylanase structure, suggesting higher model quality and potentially greater reliability in predicting its functional behavior. The reduced number of outliers signifies a well-defined secondary structure, which is crucial for accurate ligand binding and catalytic activity predictions. This superior conformation in xylanase reinforces its potential for targeted industrial applications, where precise structural understanding is paramount for enzyme engineering and optimization.

3.9 QMEAN4 score

The QMEAN4 score is a composite measure of protein model quality and structural reliability, indicating their accuracy. It is crucial in understanding the molecular mechanism of carbohydrate-activated enzymes (CAZymes), which are essential for carbohydrate breakdown, biosynthesis, and modification. An accurate QMEAN4 score indicates a well-structured and reliable model, impacting bioengineering experiments, biotechnological applications, and drug discovery. Figure 4 shows the output of QMEAN4 score for each of the CAZymes. Endoglucanase and xylanase show that output of -043 and -0.92 almost reached the ideal score of QMEAN4 score which is 1 and 0. Therefore, the structure of endoglucanase and xylanase are of relatively high structural accuracy, albeit below the ideal score of 1. These scores suggest that the generated models, particularly for endoglucanase and xylanase, provide a reliable foundation for subsequent analyses. The relatively close scores to the ideal indicate that the models are well constructed and likely to represent the true structures.

The implications of these QMEAN4 scores are significant. Firstly, the confidence in the model accuracy strengthens the validity of our ligand prediction findings. Secondly, these reliable models can be used for virtual screening of potential inhibitors or activators, which could lead to novel drug candidates or improved industrial enzyme performance. Furthermore, these models can guide site-directed mutagenesis experiments, enabling the fine-tuning of enzyme activity for specific biotechnological applications. While the QMEAN4 scores indicate good model quality, it's essential to acknowledge that they are in-silico predictions. Therefore, experimental validation, such as X-ray crystallography or NMR spectroscopy, is necessary to confirm the structural accuracy and refine our understanding of these CAZymes. This combined approach will maximize the potential of these enzymes for industrial and biomedical applications.

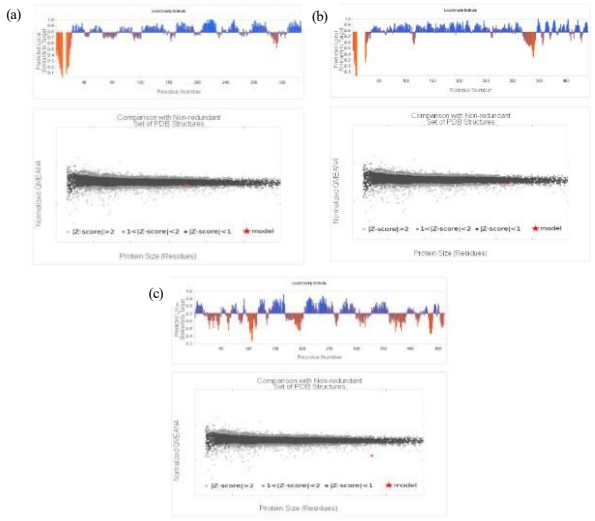

Figure 4. QMEAN4 score for Xylanase (endo-1,4-beta-xylanase) (a), Glucosidase (amylo-alpha-1,6-glucosidase) (b), and Endoglucanase (c)

3.10 Ligand site prediction

I-TASSER is a technique that predicts ligand site interactions with carbohydrate-active enzymes (CAZymes), which play a crucial role in carbohydrate breakdown, biosynthesis, and modification. By searching a large library of protein structures and improving models with ab initio folding simulations, I-TASSER helps understand how CAZymes interact with their substrates, revealing information about their catalytic processes (Zhou et al., 2022). This knowledge is essential for creating efficient enzymes for industrial uses like biofuel production and producing inhibitors for medicinal applications.

Figure 5 shows the ligand prediction for each of the CAZymes. The I-TASSER's COACH algorithm output on Xylanase protein model predicted a 137-cluster ligand (Figure 5(a)), designated 4-O-beta-D-xylopyranosyl-beta-D-xylopranose (BXP), with critical residues i.e. E48, N49, K52, H85, W89, Q92, N133, E134, Q209,

DOI: https://doi.org/10.22452/mjs.vol45sp1.8
Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)

Table 4. Ramachandran plot detail for Xylanase (endo-1,4-beta-xylanase) (a), Glucosidase (amylo-alpha-1,6-glucosidase) (b), and Endoglucanase (c)

Positions	A257	A323
Torsion angles phi (φ) and psi (ψ)	-105.76 and 124.56	-97.85 and 0.56
Confidence	0.99	0.98
MolProbity score	1.11	
Clash score	0.82	
Ramachandran-favoured regions	94.83%	
Ramachandran outliers	0.52%	
Rotamer outliers	0.92%	

(a)

Positions	A428	A630
Torsion angles phi (φ) and psi (ψ)	-115.07 and 127.86	-95.84 and -4.04
Confidence	0.28	0.71
MolProbity score	2.34	
Clash score	3.21	
Ramachandran-favoured regions	86.55%	
Ramachandran outliers	6.29%	
Rotamer outliers	6.15%	

(b)

Positions	A402	A314
Torsion angles phi (φ) and psi (ψ)	-118.65 and 137.63	-65.91 and -40.07
Confidence	0.99	0.97
MolProbity score	1.11	
Clash score	0.32	
Ramachandran-favoured regions	91.80%	
Ramachandran outliers	1.82%	
Rotamer outliers	0.60%	

(c)

E240, W281, and W288, indicating their potential interactions and relevance in ligand binding.

Meanwhile, for the glucosidase, the predicted ligand site with a C-score of 0.10, indicated a 14-cluster ligand as alpha-D-glucopyranose (GCL) (Figure 5(b)). The binding region revealed residues i.e. R288, D289, W419, D421, H559, and E604, suggesting potential interactions with the glucoside enzyme. For endoglucanase ligand site prediction, the identified ligand is beta-D-glucopyranose (BGC), with a C-score of 0.28 and a 30-cluster size (Figure 5(c). The binding region revealed residues i.e. W72, Y104, P147, D148, S220, W262, K290, and D296 to be potentially involved in binding interactions.

These findings are significant for several reasons. Firstly, understanding the precise ligand-binding interactions of these CAZymes can facilitate the design of more efficient enzymes for industrial applications. For instance, optimized xylanases, glucosidases, and endoglucanases can enhance the breakdown of plant biomass in biofuel production, contributing to sustainable energy solutions. Secondly, the identified binding residues serve as potential targets for inhibitor development. This is particularly relevant in medicinal applications, where CAZyme inhibitors can be used to control carbohydrate metabolism in diseases like diabetes. Furthermore, the varying C-scores obtained for ligand predictions highlight the complexity of these interactions and the need for further experimental validation. The identified ligands and interacting residues provide a framework for future

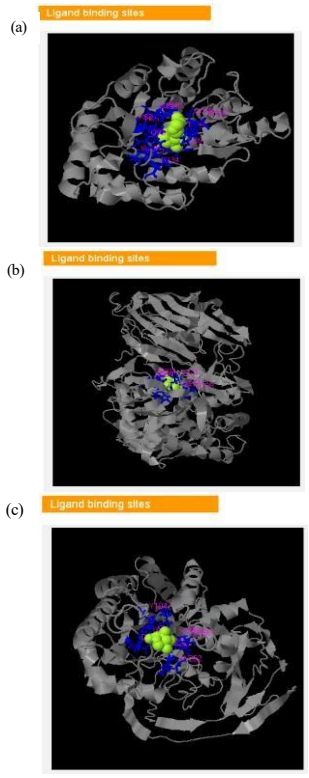

Figure 5. Ligand site prediction for Xylanase (endo-1,4-beta-xylanase) (a), Glucosidase (amylo-alpha-1,6-glucosidase) (b), and Endoglucanase (c)

DOI: https://doi.org/10.22452/mjs.vol45sp1.8
Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)

molecular dynamics simulations and in vitro assays, which will refine our understanding of these enzymes and their potential applications. The detailed ligand predictions from this study underscore the power of in-silico methods in elucidating enzyme function and paving the way for targeted biotechnological interventions.

4. Conclusion

This study successfully mined the NCBI database for genome sequence assemblies, prioritizing bacterial candidates and ultimately focusing on *Thermobifida fusca* and *Parageobacillus caldoxylosilyticus* due to their demonstrated industrial relevance and biosafety profiles. The study identified three new carbohydrate-active enzymes from *Thermobifida fusca*, including endo-1,4-beta-xylanase, amylo-alpha-1,6-glucosidase, and endoglucanase. In-silico methodologies, including protein modeling and active site prediction, were employed to elucidate the structural and functional properties of these enzymes, with quality assessment confirming model reliability. Future research should prioritize refined tertiary structure modeling, detailed protein-ligand docking experiments, and comprehensive molecular dynamics simulations. These investigations will facilitate the identification of specific inhibitors or activators, crucial for optimizing enzyme activity. This refined understanding will directly support experimental validation and pave the way for targeted industrial applications, particularly in biomass conversion and bioprocessing.

5. Acknowledgement

The authors would like to thank the technical staff from the Chemical Engineering Technology Laboratory Cluster, Pagoh Campus Laboratory Management Office for their help and assistance. This research was supported by Universiti Tun Hussein Onn Malaysia through Tier 1 (Vote No.: H851).

6. References

Bains, R. K., Nasseri, S. A., Wardman, J. F., & Withers, S. G. (2024). Advances in the understanding and exploitation of carbohydrate-active enzymes, Current Opinion in Chemical Biology, 80, 102457.

Godse, R., Bawane, H., Tripathi, J., & Kulkarni, R. (2021). Unconventional β-glucosidases: a promising biocatalyst for industrial biotechnology, Applied Biochemistry and Biotechnology, 193, 2993-3016.

Lee, J. Y., Tsolis, R. M., & Bäumler, A. J. (2022). The microbiome and gut homeostasis. Science, 377(6601), eabp9960.

Liew, K. J., Lim, L., Woo, H. Y., Chan, K. G., Shamsir, M. S., & Goh, K. M. (2018). Purification and characterization of a novel GH1 beta-glucosidase from Jeotgalibacillus malaysiensis. International Journal of Biological Macromolecules, 115, 1094-1102.

Nargotra, P., Sharma, V., Lee, Y. C., Tsai, Y. H., Liu, Y. C., Shieh, C. J., ... & Kuo, C. H. (2022). Microbial lignocellulolytic enzymes for the effective valorization of lignocellulosic biomass: a review, Catalysts, 13(1), 83.

Nooraisyah, M. N., Siti Nur Hasanah, M. Y., Mahadi, N. M., Abu Bakar, F. D., & Murad, A. M. A. (2015). Genome Mining For Glycoside Hydrolases From The Psychrophilic Yeast Glaciozyma Antarctica PI12, Malaysian Applied Biology, 44(1).

Nin-Hill, A., Piniello, B., & Rovira, C. (2023). In silico modelling of the function of disease-related CAZymes, Essays in Biochemistry, 67(3), 355-372.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information, Nucleic acids research, 50(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Sharkey, T. D. (2024). The end game (s) of photosynthetic carbon metabolism, Plant Physiology, 195(1), 67-78.

Wang, Y., Wang, C., Chen, Y., Cui, M., Wang, Q., & Guo, P. (2022). Heterologous expression of a thermostable α-galactosidase from Parageobacillus thermoglucosidasius isolated from the lignocellulolytic microbial consortium TMC7, Journal of Microbiology and Biotechnology, 32(6), 749.

Waterhouse, A. M., Studer, G., Robin, X., Bienert, S., Tauriello, G., & Schwede, T. (2024). The structure assessment web server: for proteins, complexes and more, Nucleic Acids Research, gkae270.

Wardman, J. F., Bains, R. K., Rahfeld, P., & Withers, S. G. (2022). Carbohydrate-active enzymes (CAZymes) in the gut microbiome, Nature Reviews Microbiology, 20(9), 542-556.

Yang, Q., Chang, S., Zhang, X., Luo, F., Li, W., & Ren, J. (2024). The fate of dietary polysaccharides in the digestive tract, Trends in Food Science & Technology, 104606.

Ye, S., Shah, B. R., Li, J., Liang, H., Zhan, F., Geng, F., & Li, B. (2022). A critical review on interplay between dietary fibers and gut microbiota, Trends in Food Science & Technology, 124, 237-249.

Yildiz, S. Y., Finore, I., Leone, L., Romano, I., Lama, L., Kasavi, C., Nicolaus, B., Oner, E. T., & Poli, A. (2022). Genomic Analysis Provides New Insights Into Biotechnological and Industrial Potential of Parageobacillus thermantarcticus M1, Frontiers in Microbiology, 13. https://doi.org/10.3389/fmicb.2022.923038

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Peter Kamp Busk, Xu, Y., & Yin, Y. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation, Nucleic Acids Research, 46(W1), W95–W101.

DOI: https://doi.org/10.22452/mjs.vol45sp1.8
Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)

Zheng, J., Ge, Q., Yan, Y., Zhang, X., Huang, L., & Yin, Y. (2023). dbCAN3: automated carbohydrate-active enzyme and substrate annotation, Nucleic Acids Research, 51(W1), W115-W121.

Zhou, X., Zheng, W., Li, Y., Pearce, R., Zhang, C., Bell, E. W., ... & Zhang, Y. (2022). I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction, Nature Protocols, 17(10), 2326-2353.

DOI: https://doi.org/10.22452/mjs.vol45sp1.8
Malaysian Journal of Science 45 (Special Issue 1): 66-74 (June 2026)