PROGRESS IN ENGINEERING APPLICATION AND TECHNOLOGY
e-ISSN: 2773-5303

PEAT

Vol. 5 No. 2 (2024) p. 598-610
https://publisher.uthm.edu.my/periodicals/index.php/peat

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

Alvin Owen Miharil Jukiri[1], Aisyah Mohamed Rehan[1*]

[1] *Department of Chemical Engineering Technology, Faculty of Engineering Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Pagoh Education Hub, KM1, Jalan Panchor, 86400, Muar Johor, MALAYSIA*

*Corresponding Author: aisyahr@uthm.edu.my
DOI: https://doi.org/10.30880/peat.2024.05.02.064

Article Info

Received: 27 June 2024
Accepted: 15 July 2024
Available online: 25 November 2024

Keywords

CAZymes, in-silico protein modeling, genomic data mining

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 that allow microbial conversion of plant biomass to industrially valuable products. There is a need for more research to characterise carbohydrate-active enzymes from Malaysia. This study aims to perform in silico genomic data mining for CAZymes elucidation from selected genome assemblies isolated in Malaysia. The National Center for Biotechnology Information (NCBI) database is utilized to investigate Malaysia's diversified genome assembly, discovering 1386 complete genomes. From 2019 to 2024, there are 34 bacterial genomes and one viral genome isolated by Malaysian researchers. CAZymes from six bacterial genomes were identified and categorised. Two bacteria; *Thermobifida fusca* and *Parageobacillus caldoxylosilyticus* were chosen for further investigation. Three CAZymes from *Thermobifida fusca* were chosen for subsequent protein modelling due to their acceptable homology range to available protein structures. To fully characterise the structural and functional properties of these CAZymes, online protein modelling tools were used. Ramachandran plot assessment, Verify3D, and QMEAN4 online tools were used to evaluate the quality of the protein models. Then, I-TASSER was used to predict the ligand-binding sites inside each CAZyme sequence, providing information on their functional capabilities. From this study, 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 the in-silico characterization, the proteins' structure and function have been elucidated. Future research should focus on refining tertiary models, conducting protein-ligand docking experiments and molecular dynamics (MD) simulations, to gain insights into their stability, flexibility, and functional mechanisms in a simulated physiological environment. This could lead to the discovery of new inhibitors or activators, laying the groundwork for experimental validation and future industrial applications.

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 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 [1][2]. Plants use photosynthesis to convert carbon dioxide and water into sugars, which are further converted into carbohydrates like starches and cellulose [4]. 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 [3].

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 [5]. CAZymes are also involved in biological processes underlying diseases like cancer, diabetes, Alzheimer's, and AIDS [6]. Due to Malaysia's biodiversity, bioprospecting and discovering new products are extensively performed by scientists and commercial companies. The CAZy database, dbCAN3 meta server [13], and PlantCAZyme have assisted in genome mining of CAZymes from bacteria, fungi, viruses, and plants [4].

Malaysia's unique genome assembly has led to the discovery of a novel beta-glucosidase from *Jeotgalibacillus Malaysiensis*, a marine bacterium. This is similar to a genomic mining study on *Glaciozyma antarctica,* a psychrophilic yeast from Antarctica [7]. The aim of this study is to perform in silico genomic data mining for CAZyme elucidation from selected Malaysian genome assemblies for industrial biotechnology and biomedical applications. The study aims to provide valuable insights into industrial biotechnology and biomedical applications.

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 [11]. 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 [4][12]. The results from the annotation tools on the dbCAN3 meta server were retrieved from the dbCAN3 results page.

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

For CAZymes protein without any structural data, the 3D protein modeling will be performed. 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 [14] and ExPASy SWISS-MODEL [15]. 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.

2.5 Ligand prediction & docking experiment

Ligand prediction in CAZymes involves predicting potential substrates with which enzymes may bind or act. Bioinformatics techniques and databases like the Carbohydrate-Active Enzymes (CAZy) database are used for this prediction. 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 AutoDock 4.2 and AutoDock Vina suite for docking experiments.

3. Result and Discussion

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, 1 from viruses, and 0 from plants [11] [16].

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)

3.2 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. Direct NCBI annotation was used to retrieve and calculate CAZyme numbers within the genome assemblies. The dbCAN3 database was used to confirm NCBI annotations [13] [17]. 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 [12].

Table 2: List of six genome sequence assembly with the benefits and the annotation based on their CAZymes families

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

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

wereto 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.

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 [14] and ExPASy SWISS-MODEL [15] 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.

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), 9 carbohydrate esterases (CEs), 3 polysaccharide lyases (PLs) and 2 auxiliary activities (AAs) [8]. 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 [9].

Parageobacillus caldoxylosilyticus is a unique species from the thermophilic genus Parageobacillus [10]. Because of its high adaptability towards temperature, this bacterium is a promising candidate to producve industrially applicable thermostable enzymes, such as xylanase and maltogenic amylase [10]. 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 [10].

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 [14][15].

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.

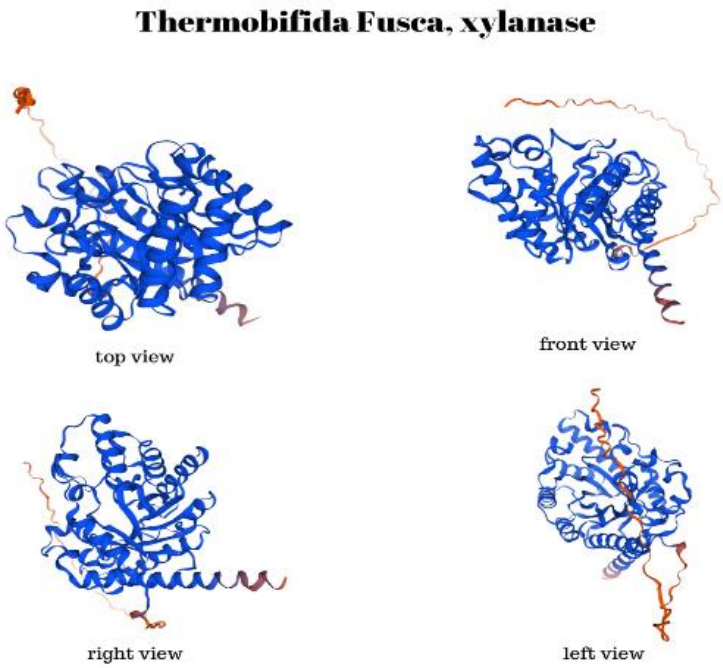
Progress in Engineering Application and Technology Vol. 5 No. 2 (2024) p. 598-610

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

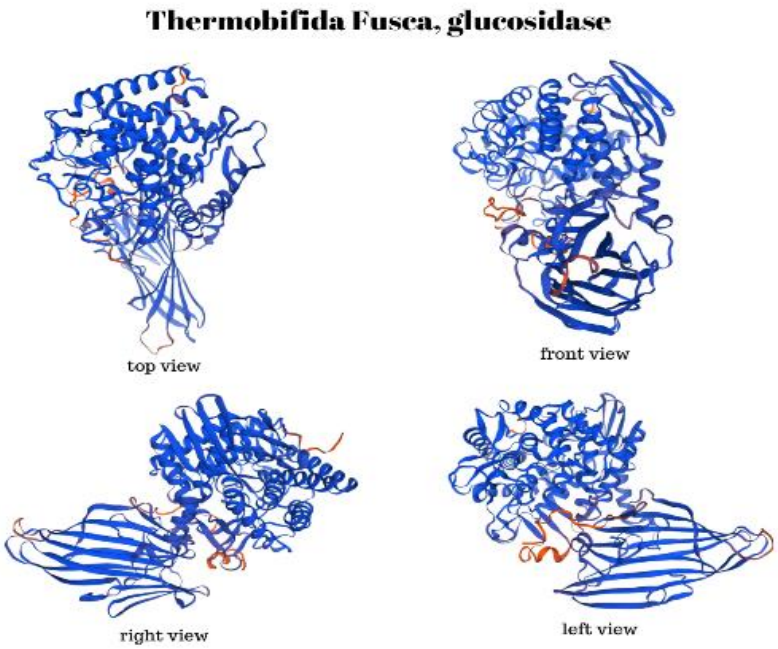
GH family	Product size (No of amino acid)	Protein/enzymes	Accession number
GH10	399	Endo-1,4-beta-xylanase	QOS59282.1
GH176	712	Amylo-alpha-1,6-glucosidase	QOS59445.1
GH6, CBM2	441	Endoglucanase	QOS57731.1

3.7 Determination of protein model quality

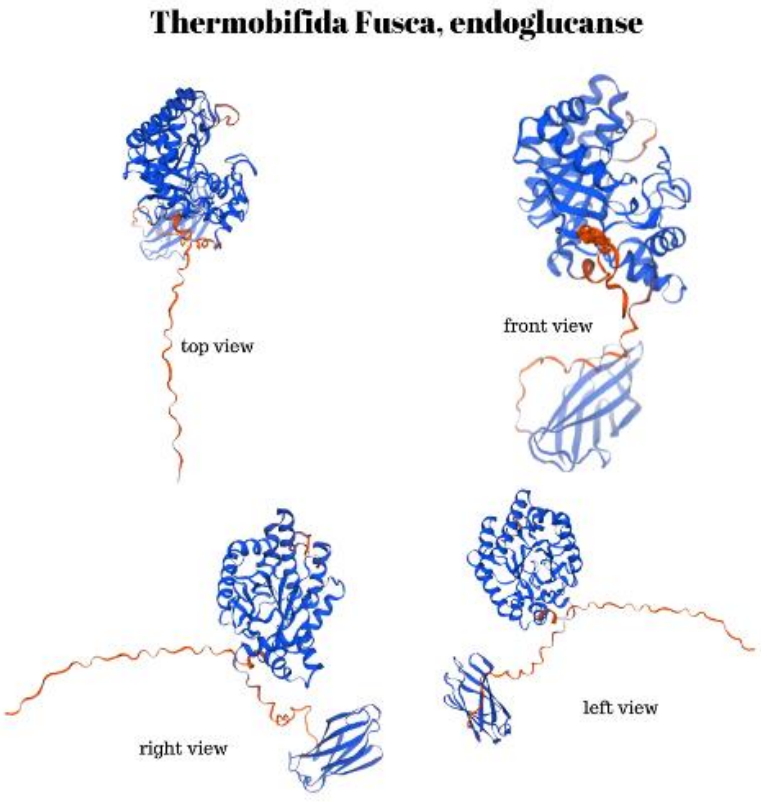
Protein models created using computational techniques are crucial for predicting active sites and simulating ligand binding. Verify3D, a tool on ExPASy SWISS-MODEL, helps scientists examine the structural integrity and functional usefulness of modelled CAZymes by comparing their 3D atomic models' suitability and correctness to their amino acid sequences [18] [19]. 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. Fig 1show the 3D modelling output for each side of the CAZymes.



(a)



(b)



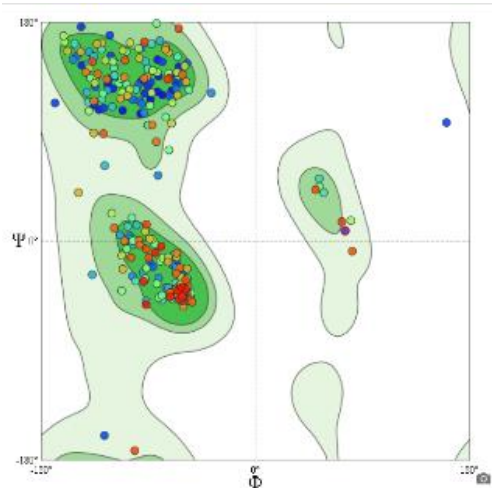
(c)

Fig 1 3D modelling of endo-1,4-beta-xylanase (a), 3D modelling of amylo-alpha-1,6-glucosidase (b), 3D modelling of Endoglucanase (c)

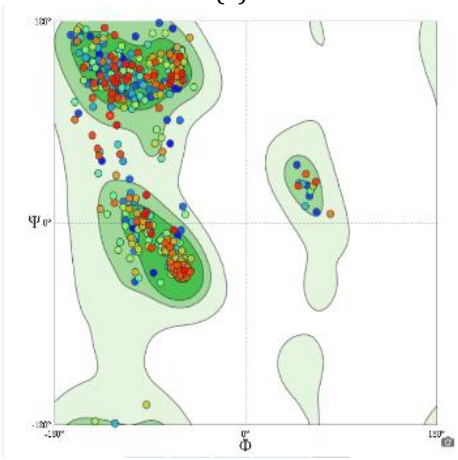
Progress in Engineering Application and Technology Vol. 5 No. 2 (2024) p. 598-610

3.8 Ramanchandran plot assessment

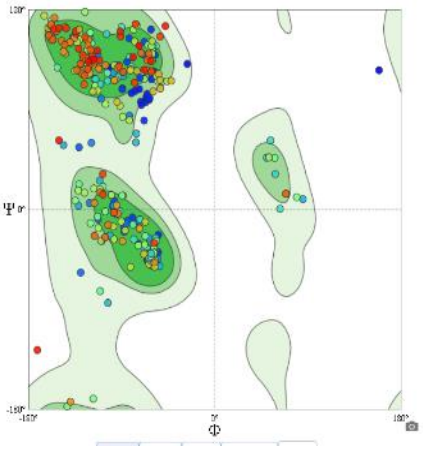
The Ramachandran plot is crucial for understanding the conformational dynamics of carbohydrate-active enzymes (CAZymes), which perform various biological tasks. By studying phi and psi torsion angles within CAZymes, researchers can determine structural flexibility and rigidity for substrate identification and catalytic activity. This analytical technique has enabled the discovery of structure-function correlations within CAZyme families, improving industrial and biotechnological uses. Fig 2 shows the Ramachandra plot for each of the CAZymes and table 4 show that the detail for the Ramachandran plot and xylanase at fig 2 (a) have the ideal model compared to other because of the outliers was less shows that the arrangement for amino acid better.



(a)



(b)



(c)

Fig 2 Xylanase (*endo-1,4-beta-xylanase*) Ramachandran plot (a), Glucosidase (*amylo-alpha-1,6-glucosidase*) Ramachandran plot (b), and *Endoglucanase* Ramachandran plot (c)

Table 4: Xylanase (*endo-1,4-beta-xylanase*) Ramachandran plot detail (a), Glucosidase (*amylo-alpha-1,6-glucosidase*) Ramachandran plot detail (b), *Endoglucanase* Ramachandran plot detail (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)

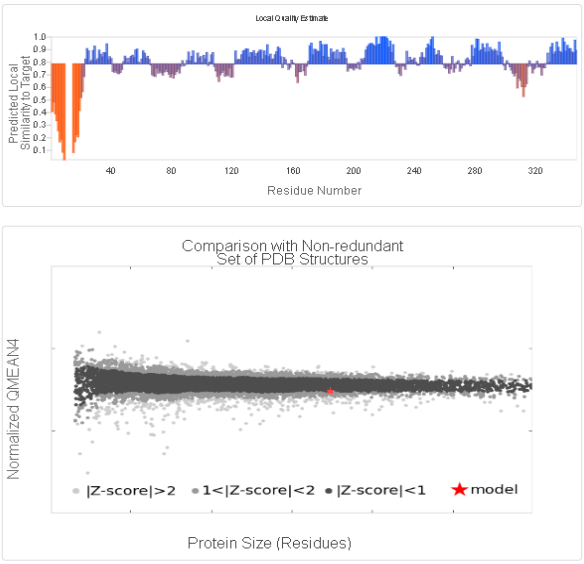
Positions	A402	A314
torsion angles phi (ϕ) and psi (ψ)	-118.65 and 137.63	-65.91 and -40.07
confidence	0.99	0.97
MolProbity score	1.11	
clash score	0.32	
Ramachandran-favoured regions	91.80%	
Ramachandran outliers	1.82%	
rotamer outliers	0.60%	

(c)

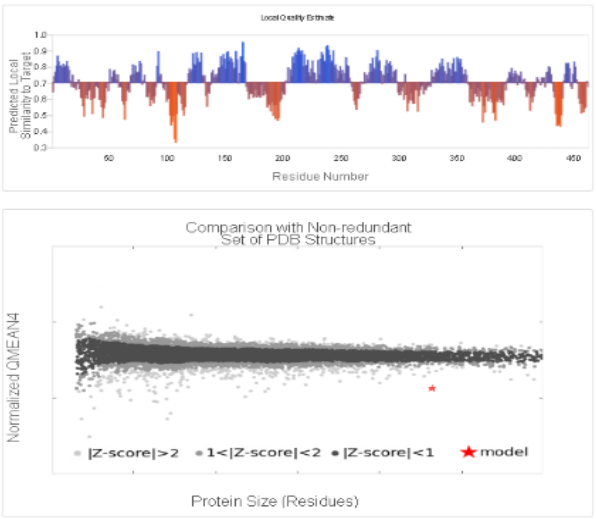
3.9 QMEAN4 score

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. Fig 3 shows the output of QMEAN4 score for each of the CAZymes. Endoglucanase and xylanase show that output of -043 and -0.92 where it's almost reached the ideal score of QMEAN4 score which is 1 and 0. Therefore, the structure of endoglucanase and xylanase are more accurate.

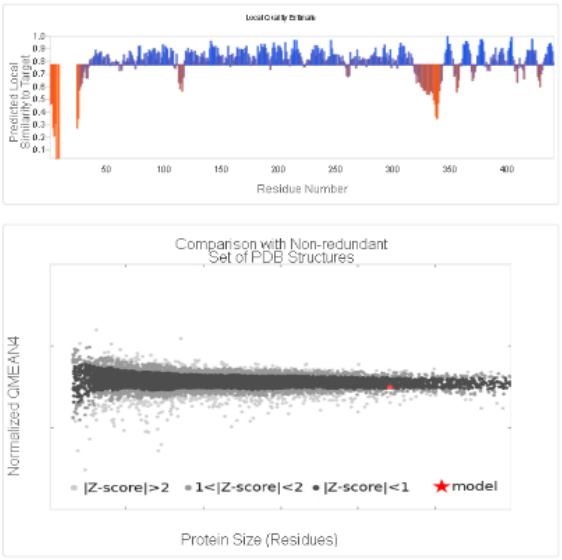
Progress in Engineering Application and Technology Vol. 5 No. 2 (2024) p. 598-610



(a)



(b)



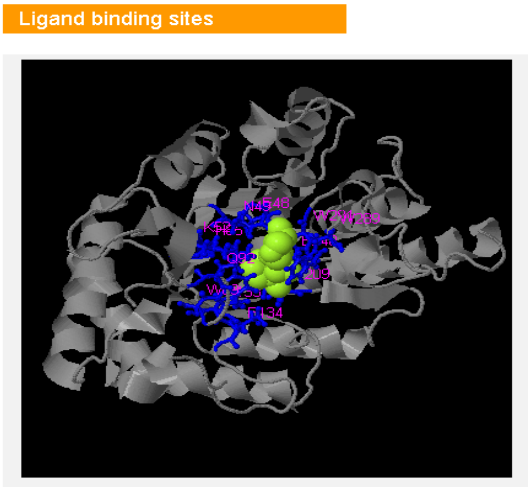
(c)

Fig 3 Xylanase (endo-1,4-beta-xylanase) QMEAN4 output (a), Glucosidase (amylo-alpha-1,6-glucosidase) QMEAN4 score (b), Endoglucanase QMEAN4 score (c)

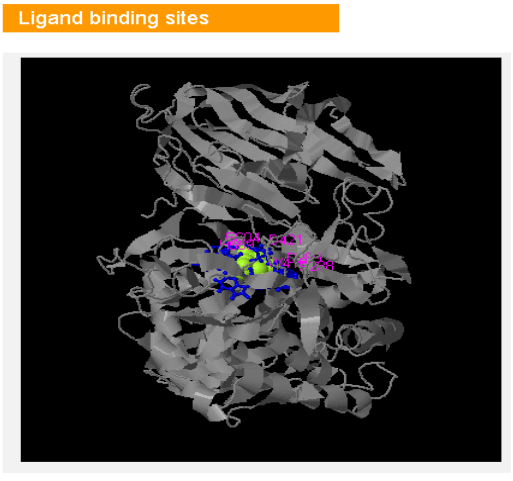
3.10 Ligand site prediction

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. This knowledge is essential for creating efficient enzymes for industrial uses like biofuel production and producing inhibitors for medicinal applications.

Equations and formulae should be typed in Math type and numbered consecutively with Arabic numerals in parentheses on the right-hand side of the page (if referred to explicitly in the text). They should also be separated from the surrounding text by one space. Fig 4 shows the ligand prediction for each of the CAZymes. (a) The I-TASSER study on Xylanase ligand predicted a 137-cluster ligand, designated 4-O-beta-D-xylopyranosyl-beta-D-xylopranose (BXP), with critical residues like E48, N49, K52, H85, W89, Q92, N133, E134, Q209, E240, W281, and W288, indicating their potential interactions and relevance in ligand binding. (b) The study predicted the glucosidase ligand site with a C-score of 0.10, identifying a 14-cluster ligand as alpha-D-glucopyranose (GCL). The binding region revealed residues like R288, D289, W419, D421, H559, and E604, suggesting potential interactions with the glucoside ligand. (c) The study achieved a C-score of 0.28 for endoglucanase ligand site prediction, with a 30-cluster size and the identified ligand being beta-D-glucopyranose (BGC). The binding region revealed residues like W72, Y104, P147, D148, S220, W262, K290, and D296 potentially involved in binding interactions.

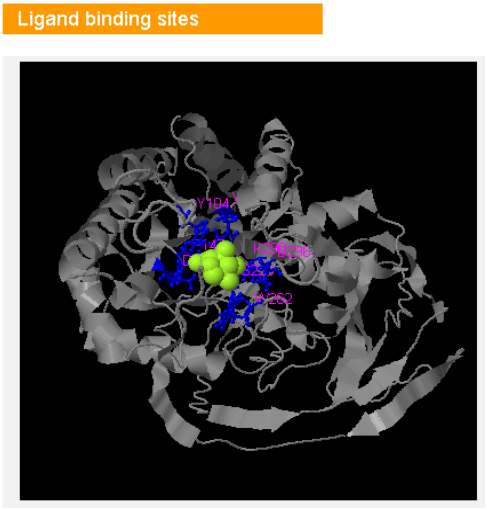


(a)



(b)

Progress in Engineering Application and Technology Vol. 5 No. 2 (2024) p. 598-610



(c)

Fig 4 Xylanase (endo-1,4-beta-xylanase) ligand site prediction (a), Glucosidase (amylo-alpha-1,6-glucosidase) ligand site prediction (b), Endoglucanase ligand site prediction (c)

4. Conclusion

The study gathered information on genome sequence assembly for bacteria, fungi, viruses, and plants from the NCBI database, focusing on bacteria. *Thermobifida fusca* and *Parageobacillus caldoxylosilyticus* were chosen for further investigation due to their potential relevance and safety. The study identified three new carbohydrate-active enzymes from *Thermobifida fusca*, including endo-1,4-beta-xylanase, amylo-alpha-1,6-glucosidase, and endoglucanase. I-TASSER and ExPASy SWISS-MODEL were used to evaluate the results. However, Objective 3 was not completed due to time restrictions and processing time for the docking experiment. Progress in predicting ligand binding sites is being made for future research.

Acknowledgement

The author gratefully acknowledges financial assistance from the Faculty of Engineering Technology (FTK) at University Tun Hussein Onn Malaysia (UTHM) for this study.

References

[1] *5.2: Genome Assembly I- Overlap-Layout-Consensus Approach. (2020, October 5). Biology LibreTexts. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/05%3A_Genome_Assembly_and_Whole-Genome_Alignment/5.02%3A_Genome_Assembly_I-_Overlap-Layout-Consensus_Approach*

[2] *Lombard, V., Hemalatha Golaconda Ramulu, Élodie Drula, Coutinho, P. M., & Henrissat, B. (2013). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research, 42(D1), D490–D495. https://doi.org/10.1093/nar/gkt1178*

[3] *Origin of Stereoselectivity and Substrate/Ligand Recognition in an FAD-Dependent R-Selective Amine Oxidase. (2016). ACS Publications. https://pubs.acs.org/doi/abs/10.1021/acs.jpcb.6b09328*

[4] *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.*

[5] *Cockburn, D. W., & Koropatkin, N. M. (2016). Polysaccharide Degradation by the Intestinal Microbiota and Its Influence on Human Health and Disease. Journal of Molecular Biology/Journal of Molecular Biology, 428(16), 3230–3252. https://doi.org/10.1016/j.jmb.2016.06.021*

[6] Montgomery, S., Hiyoshi, A., Burkill, S., Lars Alfredsson, Shahram Bahmanyar, & Olsson, T. (2017). Concussion in adolescence and risk of multiple sclerosis. *Annals of Neurology*, *82*(4), 554–561. https://doi.org/10.1002/ana.25036

[7] Nooraisyah Mohamad Nor, Nur, S., N.M. Mahadi, & Munir, A. (2015, April). *Genome mining for glycoside hydrolases from the psychrophilic yeast glaciozyma antarctica PI12*. ResearchGate; Persatuan Biologi

Gunaan Malaysia. https://www.researchgate.net/publication/284916561_Genome_mining_for_glycoside_hydrolases_from_the_psychrophilic_yeast_glaciozyma_antarctica_PI12

[8] *Songul Yasar Yildiz, Ilaria Finore, Leone, L., Romano, I., Lama, L., Ceyda Kasavi, Nicolaus, B., Ebru Toksoy Oner, & 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*

[9] 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.

[10] 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.

[11] *National Center for Biotechnology Information*. (2024). Nih.gov. https://www.ncbi.nlm.nih.gov/

[12] *GenBank Overview*. (2022). Nih.gov. https://www.ncbi.nlm.nih.gov/genbank/

[13] *dbCAN3 server*. (2023). Unl.edu. https://bcb.unl.edu/dbCAN2/

[14] *I-TASSER server for protein structure and function prediction. (2024). Zhanggroup.org. https://zhanggroup.org/I-TASSER/*

[15] *SWISS-MODEL. (2024). Expasy.org. https://swissmodel.expasy.org/*

[16] *Pruitt, K. D., Tatusova, T., Brown, G., & Maglott, D. R. (2011, November 24). NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. ResearchGate; Oxford University Press.https://www.researchgate.net/publication/51837933_NCBI_Reference_Sequences_RefSeq_Current_status_new_features_and_genome_annotation_policy*

[17] *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. https://doi.org/10.1093/nar/gkad328*

[18] *Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2014). The I-TASSER Suite: protein structure and function prediction. Nature Methods, 12(1), 7–8. https://doi.org/10.1038/nmeth.3213*

[19] *Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., Tjaart A P de Beer, Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research, 46(W1), W296–W303. https://doi.org/10.1093/nar/gky427*