Genetics of Aquatic Organisms 7(1), GA527
https://doi.org/10.4194/GA527

RESEARCH PAPER

In Silico Structural Analysis, Classification, and Functional Annotation of an Uncharacterized Protein from an Aquatic Fungus *Lindgomyces ingoldianus*

Jayzon G. Bitacura[1,2,*], Mudjekeewis D. Santos[3]

[1]University of Santo Tomas, The Graduate School, España Boulevard, 1015 Manila, Philippines
[2]Visayas State University, Department of Biological Sciences, Visca, Baybay City 6521, Leyte, Philippines
[3]National Fisheries Research and Development Institute, Genetic Fingerprinting Laboratory, 101 Mother Ignacia Ave., South Triangle, Quezon City, 1101, Philippines

How to cite

Bitacura, J.G., Santos, M.D. (2023). *In Silico* Structural Analysis, Classification, and Functional Annotation of an Uncharacterized Protein from an Aquatic Fungus *Lindgomyces ingoldianus*. *Genetics of Aquatic Organisms, 7(1), GA527*. https://doi.org/10.4194/GA527

Article History
Received 16 June 2022
Accepted 17 November 2022
First Online 6 December 2022

Corresponding Author
Tel.: +639176075221
E-mail: jayzon.bitacura@vsu.edu.ph

Keywords
Protein characterization
Structure prediction
Molecular phylogeny
Protein function

Abstract

An uncharacterized protein from *Lindgomyces ingoldianus* was initially annotated to contain various domains with promising biotechnological applications. Thus, this study was conducted to determine the structural characteristics, classification, and potential function of this protein through *in silico* methods. Results revealed that this protein has a neutral charge and is unstable and non-polar. It is predicted to have a signal peptide, glycoside hydrolase family 114 (GH114) domain, low complexity region, and fungal type cellulose-binding domain (fCBD) or type 1 carbohydrate-binding module (CBM1) region. Structural characterization and phylogenetic analysis revealed that this protein is an endo-α-1,4-polygalactosaminidase enzyme. This protein was also predicted to contain 36 active sites and is extracellularly secreted. Molecular docking analysis showed that it could bind galactosaminogalactan (GAG), a key virulence factor for *Aspergillus fumigatus* chronic infections. The binding of this protein to GAG was much better than Ega3, which could be attributed to the presence of the fCBD region that is unique to this protein. It is hypothesized that the fCBD domain helps in carbohydrate recognition and holds them in place for maximum catalysis in the GH114 domain. Finally, this protein is found to be related to its orthologue from the plant pathogenic fungus *Zopfia rhizophila*.

Introduction

Fungi represent one of the most diverse groups of organisms on earth. Organisms under Kingdom Fungi are eukaryotic and bear spores. They are achlorophyllous organisms, meaning they do not have chlorophyll and are thus, not capable of photosynthesis. Fungi generally reproduce both sexually and asexually. They could either occur as single-celled organisms as yeasts or multicellular organisms by forming filaments. Fungal cells contain cell walls that are made up of chitin or cellulose or both, together with many other complex organic molecules (Alexopoulos & Wims, 1979)

Fungi are recognized for their various roles as pathogens of plants and animals, as one of the key drivers in the decomposition of many organic and inorganic materials, and as a source of compounds with enormous biotechnological potentials (Newbound *et al*., 2010). One group of fungi, the *Dothideomycetes*, is the largest class under this kingdom. This group comprises species with an incredible diversity of lifestyles that have evolved multiple times. The first large-scale whole-

genome comparison of 101 *Dothideomycetes* species was reported by Haridas *et al.* (2020). Their study produced a high-confidence phylogeny leading to the reclassification of 25 organisms, provided a clearer picture of the relationships among the various families, and indicated that pathogenicity evolved multiple times within this class. They also identified gene family expansions and contractions across the *Dothideomycetes* phylogeny linked to ecological niches providing insights into genome evolution and adaptation across this group.

One species included in the study mentioned above is *Lindgomyces ingoldianus* strain ATCC 200398 (Shearer & Hyde) Hirayama *et al*. This species is a freshwater fungus isolated from submerged decorticated wood. Part of the whole-genome shotgun sequence of this species is an unplaced genomic scaffold BDR25scaffold_49 (Accession No.: NW_022985210). Unplaced genomic scaffolds are those sequences found in an assembly with unknown chromosome placement. This genomic scaffold was predicted to express an mRNA (Accession No.: XM_033697727) which encodes an uncharacterized protein BDR25DRAFT_381691 (Accession No.: XP_033540912). This protein was initially annotated to contain glycoside hydrolase family 114 (GH114), UV excision repair protein Rad23, and fungal cellulose-binding domain (CBM1) regions.

GH114 has been shown to disrupt microbial biofilm produced by *Aspergillus fumigatus* (Bamford *et al*., 2019). Moreover, Rad23 plays a central role in proteasomal degradation of misfolded proteins and DNA repair (Dantuma *et al*., 2009), while CBM1 is found in many proteins that catalyze the recognition and degradation of carbohydrates (Gilkes *et al*., 1991). The predicted presence of these domains makes this uncharacterized protein an interesting subject for investigation. Thus, this study was undertaken to confirm the expression of the uncharacterized protein from the mRNA expressed by the unplaced genomic scaffold of *L. ingoldianus*, as well as determine its physicochemical characteristics, domain architecture, tertiary structure, classification, active site location, subcellular localization, cleavage site, and ligand interaction through *in silico* methods.

Materials and Methods

Confirmation of mRNA Translation

The fasta sequence of the mRNA (Accession No.: XM_033697727) expressed by unplaced genomic scaffold BDR25scaffold_49 (Accession No.: NW_022985210) from *L. ingoldianus* was submitted to Expasy's Translate Server (Galsteiger *et al*., 2005). Output format was set at Compact: M, -, no spaces on both forward and reverse strands. The selected open reading frame (ORF) was then aligned with uncharacterized protein BDR25DRAFT_381691 (Accession No.: XP_033540912) using Clustal Omega from the EMBL-EBI server (McWilliam *et al*., 2013).

Physicochemical Characterization

The physicochemical characteristics of the uncharacterized protein from *L. ingoldianus* were determined by submitting the amino acid sequence to Expasy's ProtParam Server (Galsteiger *et al*., 2005). This server is a tool that allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user-entered protein sequence.

Domain Architecture Analysis

To investigate the identification of the conserved domains found in the uncharacterized protein of *L. ingoldianus*, domain architecture analysis was performed using the Simple Molecular Architecture Research Tool (SMART) (Letunic *et al*., 2021). SMART allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signaling, extracellular, and chromatin-associated proteins are detectable. These domains are extensively annotated concerning phyletic distributions, functional class, tertiary structures, and functionally important residues. Each domain is found in a non-redundant protein database, and search parameters and taxonomic information are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa.

Structure Prediction

The online tool I-TASSER (Yang & Zhang, 2015) was used for homology modeling of the uncharacterized protein from *L. ingoldianus*. I-TASSER or Iterative Threading ASSEmbly Refinement is a hierarchical approach to protein structure prediction and structure-based function annotation. It first identifies structural templates from the PDB by multiple threading method LOMETS, with full-length atomic models constructed by iterative template-based fragment assembly simulations. Function insights of the target are then derived by re-threading the 3D models through the protein function database BioLiP. The PDB files of the models generated by I-TASSER were then visualized using EzMol v2.1 Molecular Display Wizard (Reynolds *et al*., 2018).

3D Structure Refinement and Quality Validation

The top predicted structure (Model 1) of I-TASSER was refined using local Protein structure REFinement via Molecular Dynamics (*loc*PREFMD) tool (Feig, 2016). This tool refines the predicted 3D structure of the protein by improving its local stereochemistry at the same time

preserving the overall fold presented. The method uses force field-based minimization and sampling via molecular dynamics simulations with a modified force field to bring bonds, angles, and torsion angles into an acceptable range for high-resolution protein structures.

Furthermore, an assessment of the predicted model quality was done by submitting the top model predicted by I-TASSER and the model refined by locPREFMD to Ramachandran Plot Server (Anderson et al., 2005) and ProSA-web server (Wiederstein & Sippl, 2007). The refined structure was finally validated using VERIFY3D (Eisenberg *et al*., 1997).

Protein Classification

To confirm the relationship of the uncharacterized protein from *L. ingoldianus* to other proteins with known functions, molecular phylogenetic analysis was done. Initial identification of the protein was done through homology analysis using BLASTp (Johnson *et al*., 2008) and homology modeling using I-TASSER (Yang & Zhang, 2015) as previously. FASTA sequences of representative proteins were retrieved from UniProt KB (2021) and were aligned using Clustal W. Phylogenetic tree was constructed through Maximum Parsimony analysis using all sites for gaps/missing data treatment, tree-bisection-reconnection (TBR) as MP search method, and with 1000 bootstrap replicates. The alignment and construction of the phylogenetic tree were done using MEGA 11 software (Tamura *et al*., 2021).

Active Site Detection

To determine the active site of the uncharacterized protein from *L. ingoldianus*, the refined predicted 3D structure was submitted to Computed Atlas of Surface Topography (CASTp) of Proteins (Dundas *et al*., 2006) server. This tool provides an online resource for locating, identifying and quantifying concave surface areas on three-dimensional protein structures.

Subcellular Localization & Signal Peptide Analysis

To predict the subcellular localization of the uncharacterized protein from *L. ingoldianus*, its fasta sequence was submitted to the DeepLoc-1.0 server (Almagro Armenteros *et al*., 2017). This server predicts the subcellular localization of eukaryotic proteins using the Neural Networks algorithm trained on UniProt proteins with experimental subcellular localization evidence. It only uses the sequence information to perform the prediction. Furthermore, the amino acid sequence of the uncharacterized protein was submitted to the SignalP-6.0 server (Teufel *et al*., 2022) to determine the cleavage site and further analyze the secretion and translocation of the protein encoded by *L. ingoldianus* unplaced genomic scaffold BDR25scaffold_45. This online server predicts the presence and type of signal peptides and the location of their cleavage sites in proteins from Archaea, Gram-positive Bacteria, Gram-negative Bacteria, and Eukarya.

Protein-Protein Interaction Analysis

Protein-protein interactions were predicted using STRING v.11.5 software (Szklarczyk *et al*., 2021). On the web server, "GH114" was used in the query, and a fungal species was selected among the matches displayed.

Ligand/Substrate Interaction

HDOCK server (Yan *et al*., 2020) was used to determine the interaction of the uncharacterized protein from *L. ingoldianus* to its probable ligand or substrate. This docking server is based on a hybrid algorithm of template-based modeling and *ab initio-free* docking. Model 1 of the predicted structure was compared against its closest structural homolog, indicated by I-TASSER. Before submission for docking, the target protein sequence was cleaved first at a site determined by SignalP-6.0. Also, the structure of the carbohydrate substrate was constructed first using the Carbohydrate Builder tool on the Glycam website (Grant *et al*., 2016).

Protein Evolution Analysis

Orthology analysis was conducted to determine the relationship between this uncharacterized protein from *L. ingoldianus* and those same proteins from other fungal species. This was done by homology analysis followed by molecular phylogenetic analysis. BLASTp (Johnson *et al*., 2008) was used to search for protein orthologues. Only those above 90% identity with the query sequence were retrieved. The retrieved sequences were then aligned using Clustal W. Phylogenetic tree was constructed through Maximum Parsimony analysis using all sites for gaps/missing data treatment, tree-bisection-reconnection (TBR) as MP search method, and with 1000 bootstrap replicates. The alignment and construction of the phylogenetic tree were done using MEGA 11 software (Tamura *et al*., 2021).

Results and Discussion

Translated mRNA

Figure 1A shows several ORFs found in both forward and reverse strands of the translated mRNA expressed by unplaced genomic scaffold BDR25scaffold_49 from *L. ingoldianus*. Among these ORFs, the longest one was from the 5'3' Frame 2 (Figure 1A pink box). All the amino acid residues of this ORF were aligned with the amino acid residues of the uncharacterized protein BDR25DRAFT_381691 (Figure 1B). This confirms that the uncharacterized

protein BDR25DRAFT_381691 is encoded by the mRNA expressed by unplaced genomic scaffold BDR25scaffold_49.

Physicochemical Characteristics

The physicochemical characteristics of the uncharacterized protein from *L. ingoldianus* are summarized in Table 1. This protein has 370 amino acid residues and a molecular weight of 40, 442.44 Da. The theoretical isoelectric point of 6.71 suggests that this protein is neutral. The equal presence of positively charged and negatively charged residues further supports this protein's neutrality. Moreover, the computation of the instability index at more than 40 classifies this protein as unstable. The aliphatic index of the protein is 57, which regards it as a positive factor for increasing its thermostability. Lastly, the grand average of hydrophobicity (GRAVY) of this protein is -0.582. A negative GRAVY implies a non-polar characteristic of this protein.

Domain Architecture

The uncharacterized protein from *L. ingoldianus* was shown to contain a signal peptide, glycoside hydrolase family 114 (GH114) domain, low complexity region, and fungal type cellulose-binding domain (fCBD) also known as carbohydrate-binding module 1 (CBM1) (Figure 2). The signal peptide comprises the amino acid residues 1-17, the GH114 domain of the residues 27-259, the low complexity region of the residues 287-332,

Figure 1. Expasy's Translate tool shows the possible ORF encoded by the mRNA expressed by the unplaced genomic scaffold from *L. ingoldianus* in the pink box (A). Alignment of the ORF (VIRT-47012:5'3' to the uncharacterized protein (XP_033540912.1) shows all of the residues to be identical.

and fCBD of the residues 336-370. This result is not consistent with the previous annotation done by Haridas *et al.* (2020). Their study reported that Rad23 domain was included as part of the predicted regions found in this protein. However, the SMART analysis done in this study does not have this domain as one of those confidently predicted domains, repeats, motifs, and features. This inconsistency is understandable because the previous study used the JGI annotation pipeline to predict the function of all the genes found in 101 Dothideomycete genomes that they studied compared to this study which is only specific to this particular uncharacterized protein.

Signal peptides (SPs) are short peptides located in the N-terminal of proteins, carrying information for protein secretion, and are common to all prokaryotes and eukaryotes (Owji *et al*., 2018). The GH114 domain family is recognized as a glycosyl-hydrolase family, number 114. It is found in endo-α-1,4-polygalactosaminidase, a rare enzyme. It is proposed to be TIM-barrel, the most common structure amongst the catalytic domains of glycosyl-hydrolases (Naumov & Stepushchenko, 2011). Moreover, low-complexity regions (LCRs) are amino acid sequences that contain repeats of single amino acids or short amino acid motifs and are highly abundant in eukaryotic proteins (Toll-Riera *et al*., 2012). Finally, the fCBD or the CBM1 domain is fungi's small four-cysteine cellulose-binding domain. This domain is commonly found in carbohydrate degrading enzymes. These enzymes generally consist of a catalytic domain joined to a cellulose-binding domain (CBD) by a short linker sequence rich in proline and hydroxy-amino acids (Gilkes *et al*., 1991). In the case of this uncharacterized protein, the GH114 could serve as the catalytic domain that is joined to fCBD by a low complexity region (Figure 2)

Table 1. Physicochemical characteristics of the uncharacterized protein from *L. ingoldianus*

Physicochemcial Characteristics	Values
Number of amino acids	370
Molecular weight	40442.44
Theoretical pI	6.71
Amino acid composition (No.)	
Ala (A)	26
Arg (R)	11
Asn (N)	20
Asp (D)	28
Cys (C)	12
Gln (Q)	17
Glu (E)	9
Gly (G)	25
His (H)	2
Ile (I)	14
Leu (L)	23
Lys (K)	26
Met (M)	9
Phe (F)	12
Pro (P)	32
Ser (S)	31
Thr (T)	39
Trp (W)	9
Tyr (Y)	11
Val (V)	14
Total number of negatively charged residues (Asp+Glu)	37
Total number of positively charged residues (Arg+Lys)	37
Atomic composition	
Carbon (C)	1772
Hydrogen (H)	2747
Nitrogen (N)	479
Oxygen (O)	563
Sulfur (S)	21
Formula	$C_{1772}H_{2747}N_{479}O_{563}S_{21}$
Total number of atoms	5582
Extinction coefficients:	
Assuming all pairs of Cys residues form cystines	66640
Assuming all Cys residues are reduced	65890
Estimated half-life (hrs)	>20
Instability index (II)	48.71
Alipathic index	57.00
Grand average of hydrophobicity (GRAVY)	-0.582

Predicted Structure

Figure 3 shows the backbone (Figure 3A) and surface (Figure 3B) views of the top final model (Model 1) for the predicted tertiary structure of the uncharacterized protein from *L. ingoldianus.* Model 1 has the highest C-score of -1.43, with an estimated TM-score of 0.54±0.15 and an estimated RMSD of 9.9±4.6Å. According to Yang and Zhang (2015), C-score is a confidence score for assessing the quality of predicted models by I-TASSER. It is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-score is typically in the range of [-5,2], where a C-score of higher value signifies a model with increased confidence and vice-versa. Moreover, TM-score and RMSD are known standards for measuring structural similarity between two structures which are usually used to measure the accuracy of structure modeling when the native structure is known.

The Top 5 predicted models are based on the templates used by I-TASSER for threading. The top 10 threading templates include hydrolases, transcription, and toxin proteins. Of the ten templates, Ega3, an endo-α-1,4 polygalactosaminidase found in *A. fumigatus*, was ranked 1st, 3rd, 5th, 7th, and 9th (Table 2). Furthermore, the top 10 structural analogs of the uncharacterized protein in PDB were composed of endo-α-1,4-galactosaminidase from *A. fumigatus*, endo-α-N-acetylgalactosaminidase from *Bifidobacterium longum*, Lacto-N-biosidase from *Bifidobacterium bifidum*, N-acetyl-β-D-glucosaminidase from *Streptococcus gordonii,* Endo-α-N-acetylgalactosaminidase from *Streptococcus pneumoniae,* N-acetyl-α-glucosaminicase from *Homo sapiens*, Trimethylamine dehydrogenase from *Methylophilus methylotrophus*, Chitibiose from *Flavobacterium johnsoniae*, β-galactosidase from *Trichoderma reesei*, and Histamine dehydrogenase from *Pimelobacter simplex*, respectively (Table 3). Of these ten structural analogs, only the 1st ranked endo-α-1,4-galactosaminidase from *A. fumigatus* aligned perfectly with model 1 generated by I-TASSER (Figure 3C). The rest of the structural analogs have domains not present in the uncharacterized protein from *L. ingoldianus*.

These structural prediction and alignment results reveal that the uncharacterized protein encoded by the

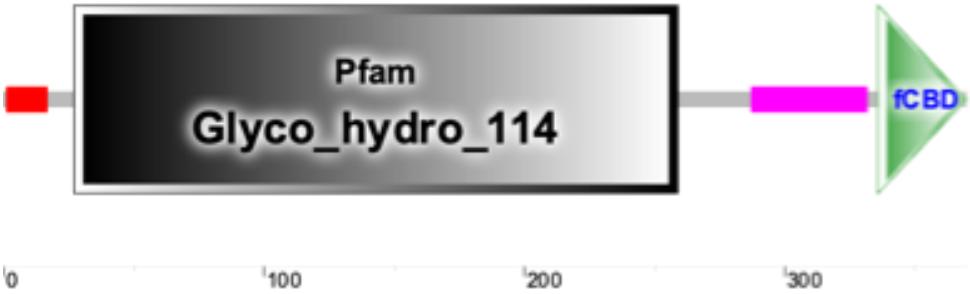

Figure 1. Domain architecture of the uncharacterized protein from L. ingoldianus. This protein consists of a signal peptide (red bar), GH114 domain (black box), low complexity region (pink bar), and fCBD region (green triangle).

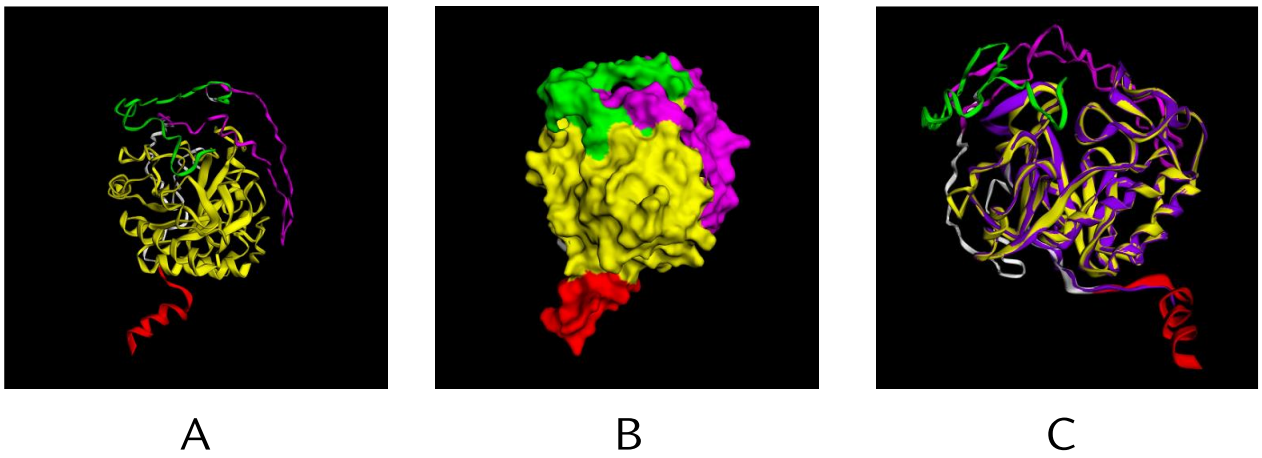

Figure 2. Top final model (Model 1) of the tertiary structure of the uncharacterized protein from L. ingoldianus predicted by I-TASSER shown in cartoon (A) and surface (B) views and in alignment with the top structural analog (C). The structural analog is colored purple in C while the model 1 structure from A-C shows the signal peptide (red), GH114 domain (yellow), low complexity region (pink, and fCBD region (green).

mRNA from the unplaced genomic scaffold of *L. ingoldianus* could probably be an endo-α-1,4-polygalactosaminidase with GH114 catalytic domain same as Ega3 from *A. fumigatus.* Ega3 is an active glycoside hydrolase that disrupts GAG-dependent *A. fumigatus* and Pel polysaccharide-dependent *Pseudomonas aeruginosa* biofilms at nanomolar concentrations (Bamford *et al*., 2019).

Refined and Validated 3D Structure

Figure 4 shows the top predicted model by I-TASSER that is refined by *loc*PREFMD. Although the refined model has not really reached the targeted values for various stereochemical parameters, this version has significantly enhanced values compared to the crude/submitted model (Table 4). In terms of the assessment of model quality, Ramachandran Plot analysis revealed that the refined model has increased observation values than the crude model. The crude model has highly preferred observations of 79.076%, preferred observations of 13.859%, and questionable observations of 7.065% (Figure 5A). After the refinement of the model, the values improved to 91.304%, 5.707%, and 2.989%, respectively (Figure 5B).

Furthermore, ProSA-web analysis showed that the overall model quality of the refined model is higher compared to the crude model. The Z-Score of the crude model was equivalent to -5.56 (Figure 5C) and increased to -5.61 in the refined model (Figure 5D). Finally,

Table 2. Top 10 threading templates used by I-TASSER to predict the structure of the uncharacterized protein from *L. ingoldianus*.

Rank	PDB Hit	Name	Class	Organism	Ident 1	Ident 2	Coverage	Norm Z-score
1	6oj1A	Ega3 (endo-α-1,4-polygalactosaminidase)	Hydrolase	*Aspergillus fumigatus*	0.46	0.32	0.66	2.79
2	7o0eG	GH30 (mutant E188A) complexed with aldotriuronic acid	Hydrolase	*Thermothelomyces thermophilus* ATCC 42464	0.10	0.20	0.94	1.15
3	6oj1	Ega3 (endo-α-1,4-polygalactosaminidase)	Hydrolase	*Aspergillus fumigatus* Af293	0.46	0.32	0.66	6.64
4	7ktrC	SAGA coactivator complex (TRRAP, core)	Transcription	*Homo sapiens*, unclassified *Rhodococcus*	0.09	0.18	0.93	1.08
5	6oj1	Ega3 (endo-α-1,4-polygalactosaminidase)	Hydrolase	*Aspergillus fumigatus* Af293	0.46	0.32	0.66	4.37
6	7wabA	prolyl endoprotease (PEP)	Hydrolase	*Aspergillus niger*	0.10	0.23	0.93	1.05
7	6oj1A	Ega3 (endo-α-1,4-polygalactosaminidase)	Hydrolase	*Aspergillus fumigatus* Af293	0.47	0.32	0.66	2.94
8	7qfpA	Botulinum neurotoxin serotype E	Toxin	*Clostridium botulinum*	0.05	0.22	0.97	1.05
9	6oj1	Ega3 (endo-α-1,4-polygalactosaminidase)	Hydrolase	*Aspergillus fumigatus* Af293	0.47	0.32	0.66	6.96
10	7m10A	RNA polymerase II pre-initiation complex (PIC1)	Transcription	*Saccharomyces cerevisiae*	0.06	0.19	0.82	0.72

Table 3. Proteins structurally close to the uncharacterized protein from *L. ingoldianus* in the PDB as identified by TM-align.

Rank	PDB Hit	Name	Class	Organism	TM-score	RMSD	IDEN	Coverage
1	6oj1A	endo-α-1,4-polygalactosaminidase (GH114)	Hydrolase	*Aspergillus fumigatus*	0.663	0.41	0.459	0.665
2	2zxqA	endo-α-N-acetylgalactosaminidase (GH101)	Hydrolase	*Bifidobacterium longum*	0.617	4.90	0.072	0.819
3	4h04A	Lacto-N-biosidase (GH20)	Hydrolase	*Bifidobacterium bifidum* JCM 1254	0.610	4.97	0.078	0.824
4	2epoA	N-acetyl-β-D-glucosaminidase (GCNA) (GH20)	Hydrolase	*Streptococcus gordonii*	0.610	4.82	0.043	0.803
5	3ecqB	Endo-α-N-acetylgalactosaminidase (GH101)	Hydrolase	*Streptococcus pneumoniae* R6	0.603	5.08	0.077	0.822
6	4xwhA	N-acetyl-α-glucosaminidase (GH89)	Hydrolase	*Homo sapiens*	0.595	4.84	0.054	0.797
7	2tmdA	Trimethylamine dehydrogenase	Oxidoreductase	*Methylophilus methylotrophus* W3A1	0.591	5.13	0.063	0.805
8	6yhhA	Chitobiose (GH20)	Hydrolase	*Flavobacterium johnsoniae* UW101	0.590	4.46	0.038	0.754
9	3og2A	β-galactosidase (GH35)	Hydrolase	*Trichoderma reesei*	0.588	5.06	0.064	0.811
10	3k30A	Histamine dehydrogenase (HADH)	Oxidoreductase	*Pimelobacter simplex*	0.588	5.13	0.068	0.800

VERIFY3D analysis for validating the quality of the refined model reported a "pass" result. The analysis revealed that 86.76% of the residues have averaged 3D-1D score >= 0.2 (Figure 5E). According to Eisenberg *et al*. (1997) if at least 80% of the amino acids have scored >= 0.2 in the 3D/1D profile, then the model passes the quality validation test.

Protein Classification

BLASTp analysis and I-TASSER prediction initially identify the uncharacterized protein as GH114 domain-containing protein. Thus, molecular phylogenetic analysis of its sequence together with other glycoside hydrolases was done to confirm this identity. Phylogenetic analysis of the uncharacterized protein from *L. ingoldianus* confirms that this protein is an endo-α-1,4-polygalactosaminidase with a catalytic domain belonging to the glycoside hydrolase 114 family (GH114). The phylogenetic tree in Figure 6A shows that this uncharacterized protein was grouped with the two GH114 domain-containing endo-α-1,4-polygalactosaminidase sequences from *A. fumigatus* with a 100% bootstrap support value.

The GH114 group shares the same ancestor with endo-α-N-acetylgalactosaminidase or the GH101 family. Also shown in the tree are resolved clades of GH35 domain-containing β-galactosidases, GH20 domain-containing Chitooligosaccharide deacetylases, GH89 domain-containing N-acetyl-α-glucosaminidase, GH35

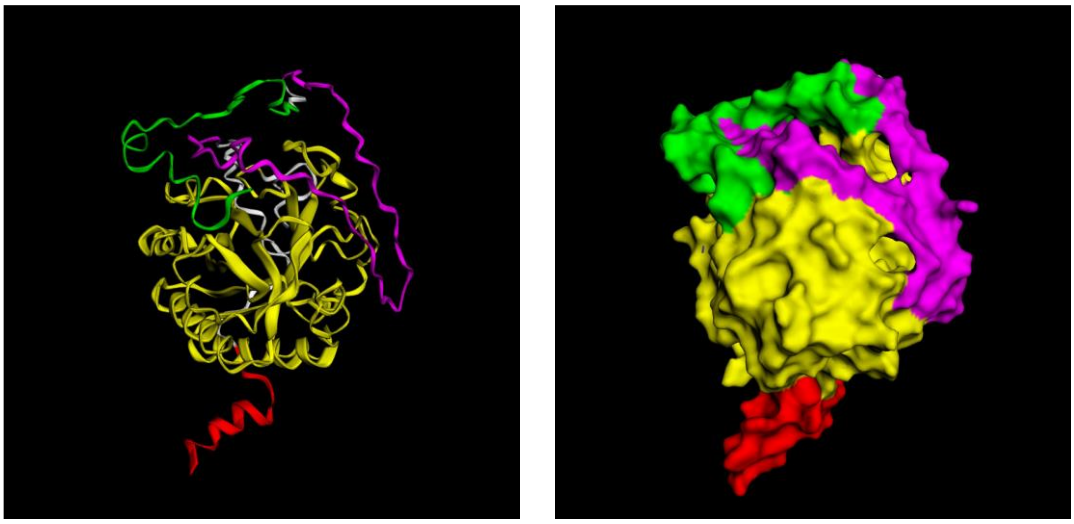

Figure 4. Cartoon (A) and surface (B) views of the top structural model (Model 1) refined by *loc*PREFMD. The structure shows the signal peptide (red), GH114 domain (yellow), low complexity region (pink, and fCBD region (green).

Table 4. Comparison of the stereochemical parameters of the top structural model predicted by I-TASSER and its refined version by locPREFMD.

Parameters	Submitted	Refined	Goal
Total score	2.924	1.869	< 1.0
Clash score	3.948	0.537	0
C-beta deviations	25	4	0
Sidechain rotamer outliers	15.990	5.330	< 0.3 %
Phi-psi backbone outliers	11.410	2.990	< 0.2 %
Phi-psi backbone favored region	71.470	86.410	> 98 %
21-window residue average	0.000	0.000	1.0
Phi-psi backbone favored region	64.300	79.100	> 90 %
Phi-psi backbone allowed region	28.900	17.700	
Phi-psi backbone general region	5.500	2.900	< 1 %
Phi-psi backbone disfavored region	1.300	0.300	< 0.2 %
Phi-psi backbone un-allowed region	0.179	0.087	< 0.2 %
Chi1-chi2 sidechain un-allowed region	0.076	0.021	< 0.2 %
G-factor dihedrals	-1.040	-0.600	> -0.5
G-factor covalent bonds	-0.240	0.020	> -0.5
G-factor overall interactions	-0.680	-0.330	> -0.5
Favorable main chain bond lengths	99.300	99.800	100 %
Favorable main chain angles	88.000	92.300	100 %
Sidechain ring planarity	84.000	94.100	100 %

domain-containing α-galactosidases, and the outgroup Oxidoreductases (EC 1). The maximum parsimony tree has a length of 7716, consistency index (CI) of 0.905780, retention index (RI) of 0.827766, and composite index of 0.749774.

A closer look at the aligned sequences of the uncharacterized protein from *L. ingoldianus* and the two sequences from *A. fumigatus* showed several conserved sequences found within their GH114 domains (Figure 6B). The GH114 domain of the first Ega3 sequence (pdb|6OJ1|A) is from 41 up to 273 amino acid residues. On the other hand, the GH114 domain of the second Ega3 sequence ((pdb|6OJB|A) is from 58 up to 290 amino acid residues. Lastly, in terms of their domain architecture, only the uncharacterized protein from *L. ingoldianus* has an fCBD linked to the GH114 domain via a low complexity region compared to the other two (Figure 6C), which makes it a unique form of an endo-α-1,4-polygalactosaminidase enzyme.

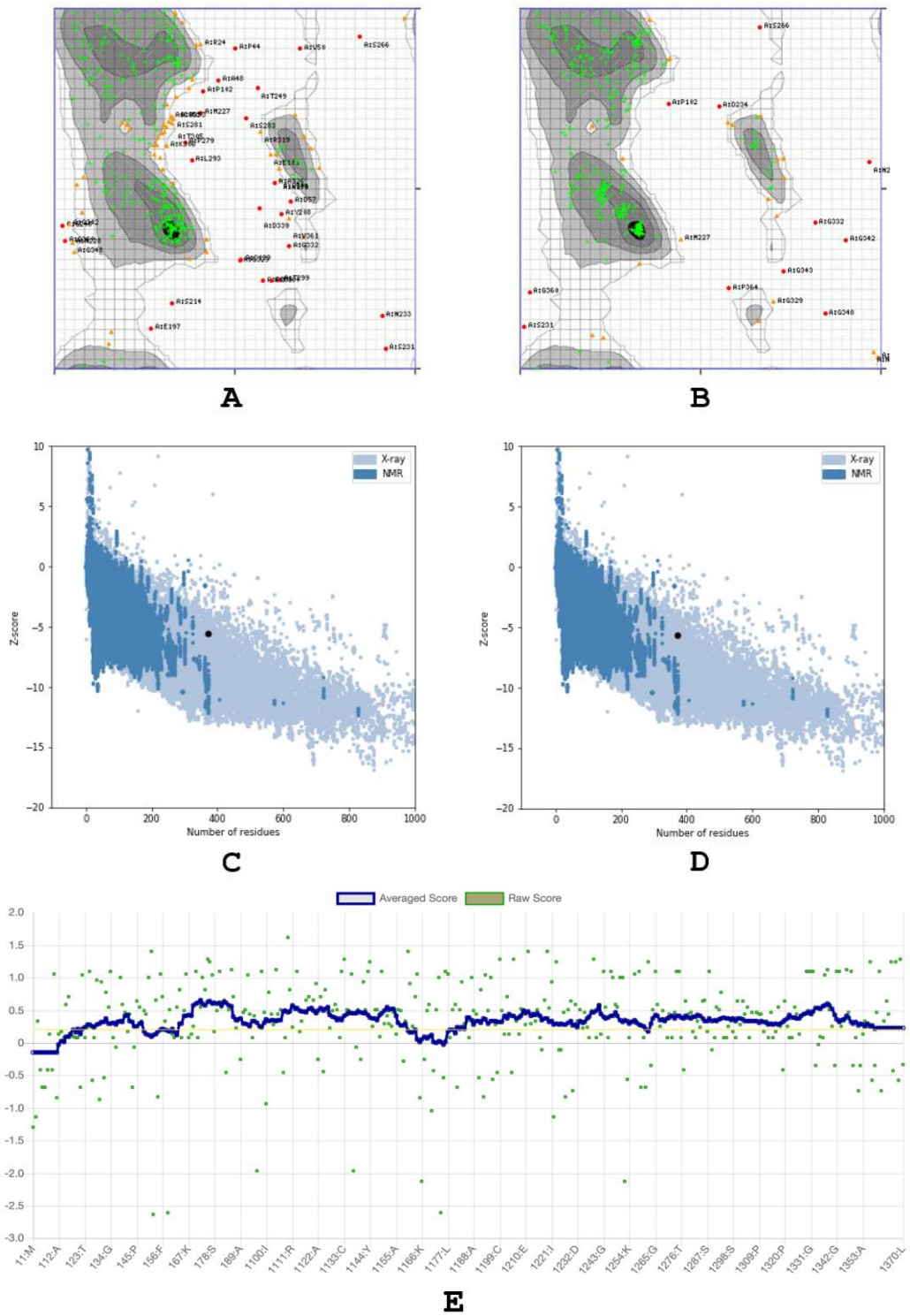

Figure 5. Ramachandran Plot Server and ProSA-web server analysis for the quality assessment of the top predicted model by I-TASSER (A&C) and the refined model by locPREFMD (B&D), as well as the refined model's validation through VERIFY3D (E).

Detected Active Sites

CASTp analysis predicted 36 active sites within the refined model of the uncharacterized protein from *L. ingoldianus* (Figure 7). The top active sites of the modeled protein were identified between the area of 2335.104 and the volume of 1817.838.

Subcellular Localization and Signal Peptide Profile

Figure 8A shows that the uncharacterized protein from *L. ingoldianus* is an extracellular enzyme. It is synthesized inside the fungal cell and is secreted outside for its catalytic function. Analysis of the signal peptide sequence of the uncharacterized protein further revealed that it contains a Sec/SPI signal peptide. Sec/SPI is the "standard" secretory signal peptide transported by the Sec translocon and is cleaved by Signal Peptidase I (*Lep*). Moreover, the analysis revealed that this protein has the signal peptide cleavage site between amino acid residues 17 and 18 (Figure 8B).

The protein endo-α-1,4-polygalactosaminidase is a type of carbohydrate-active enzyme (CAZyme). CAZymes are secreted by filamentous fungi, especially those that could efficiently degrade lignocellulose (Gabriel *et al*., 2021). Production of various CAZymes is an inherent characteristic of the heterotrophic fungal lifestyle to efficiently degrade the available biomass in their habitat (Barrett *et al*., 2020). Their biomass substrates are often composed of different plant cell wall polysaccharides, primarily cellulose, hemicellulose, pectin, and lignin (Benoit *et al*., 2015). In the case of endo-α-1,4-polygalactosaminidase, a critical action of this enzyme is its ability to disrupt galactosaminogalactan (GAG). GAG is an integral component of the *A. fumigatus* biofilm matrix and a key

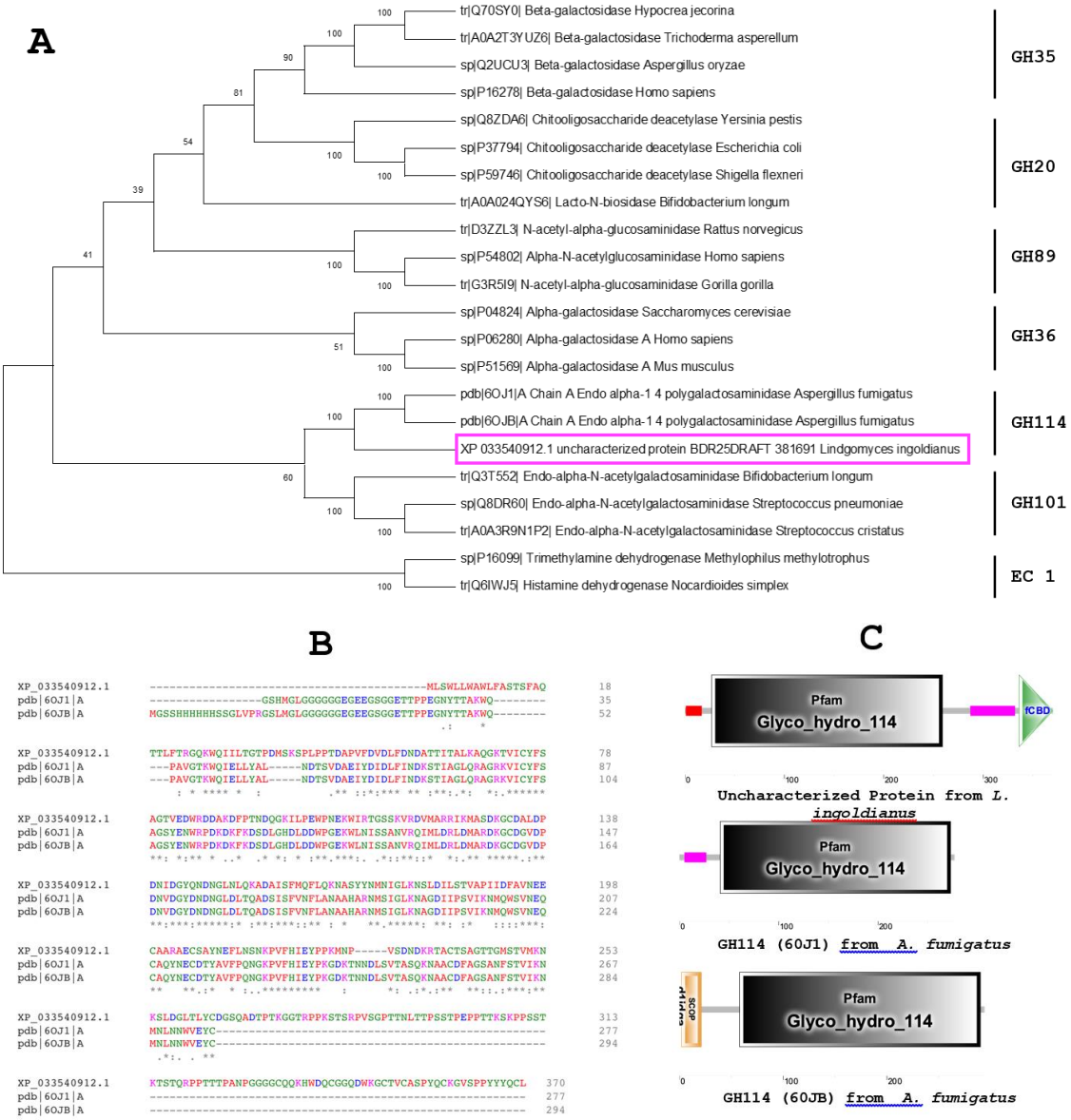

Figure 6. Molecular phylogenetic analysis of the uncharacterized protein from *L. ingoldianus* with other glycoside hydrolase proteins (A). Multiple sequence alignment of the uncharacterized protein to the two sequences of endo-α-1,4-polygalactosaminidase from A. fumigatus (B) and the comparison of their domain architectures (C).

virulence factor for causing chronic infections in patients with pre-existing lung conditions such as chronic obstructive pulmonary disease or cystic fibrosis (Bamford *et al*., 2019).

Protein-Protein Interaction

STRING query resulted in a match for endo-α-1,4-polygalactosaminidase protein interaction from *Aspergillus niger*. Figure 9 shows that endo-α-1,4-polygalactosaminidase (An08g04330) interacts with the following proteins: Udp-glucose 4-epimerase (An02g11320) which belongs to the NAD-dependent epimerase/dehydratase family, cell surface spherulin 4-like protein (Am02g11330 & An03g05560) which belongs to the Spherulation-specific family 4, plasma membrane ATPases (An01g05670, An02g12510, An16g05840), large subunit ribosomal protein lp2 (An16g04930) which belongs to the eukaryotic ribosomal protein P1/P2 family, endo-arabinase (An07g04930) which belongs to the Glycosyl hydrolases family 43, and glycosyl transferases group 1 family protein (An02g11400). All of these predicted functional partners are based on textmining.

Substrate Interaction

Molecular docking analysis conducted to compare catalytic activities of the enzyme endo-α-1,4-polygalactosaminidase from *A. fumigatus* and *L. ingoldianus* revealed almost the same interactions with the substrate galactosaminogalactan (GAG). GAG is a heterogeneous linear α-1,4–linked exopolysaccharide of galactose and GalNAc that is partially deacetylated after secretion.

Docking of GAG to Ega3 revealed only eight out of ten models showing substrate binding to the active sites of the protein. The other two models were bound to the N-terminal sequences of the enzyme (Figure 10A). On the other hand, all ten models showed GAG binding to the active sites of the endo-α-1,4-polygalactosaminidase from *L. ingoldianus* (Figure 10B). This optimized binding for the studied protein could be due to cleaving the signal peptide before subjecting it to docking analysis.

Comparing the structures of the enzymes with the bound substrates, the uncharacterized protein appears to be more globular than Ega3. Using MS and functional assays, Bamford *et al.* (2019) demonstrated that Ega3 is

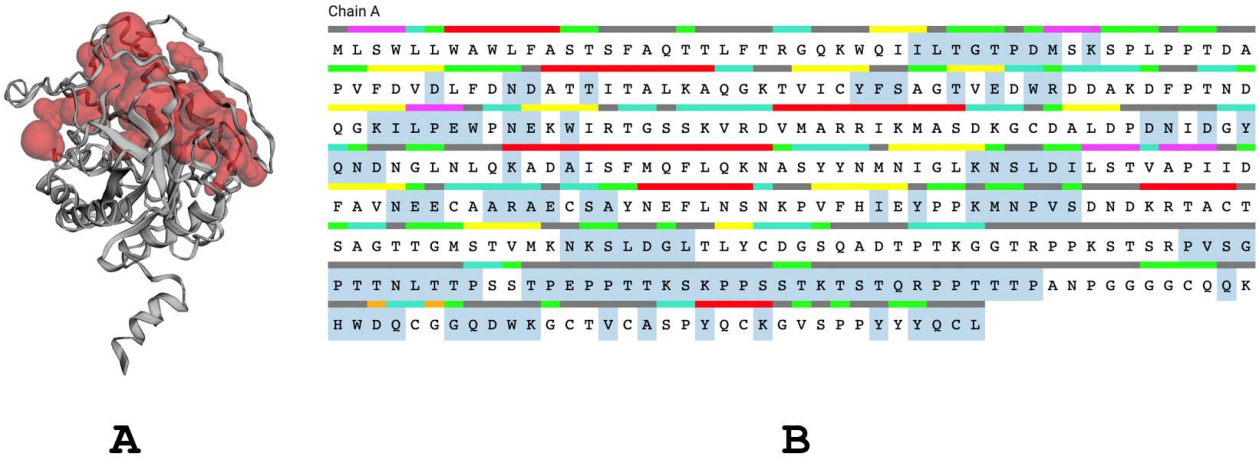

Figure 7. The active sites in the uncharacterized protein predicted by CASTp. The active sites are shown in red spheres found in the cartoon view of the protein (A) and in the specific amino acid residues highlighted in gray (B).

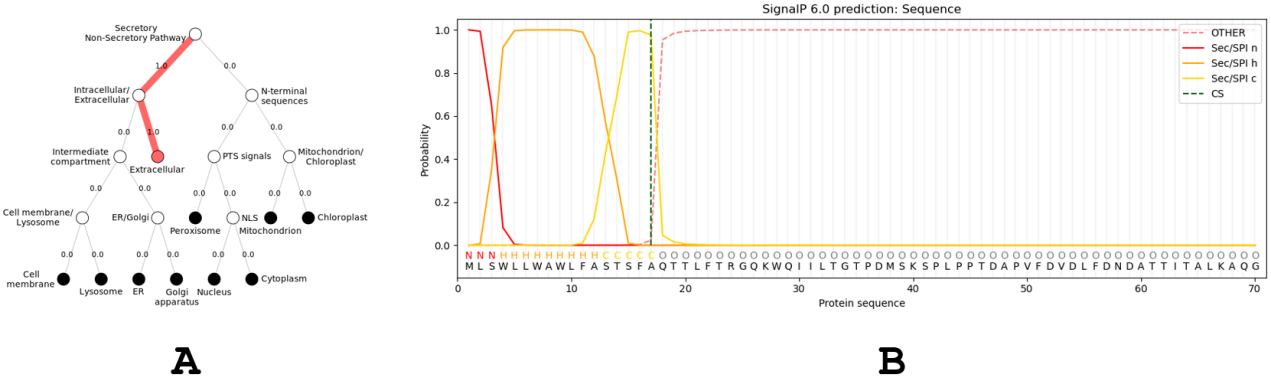

Figure 8. Results of the subcellular localization (A) and signal peptide analysis (B) of the uncharacterized protein from *L. ingoldianus*.

an endo-acting α-1,4-galactosaminidase whose activity depends on the conserved acidic residues, Asp-189 and Glu-247. X-ray crystallographic structural analysis of the apo Ega3 and an Ega3-galactosamine complex, at 1.76 and 2.09 Å resolutions, revealed a modified $(\beta/\alpha)_8$-fold with a deep electronegative cleft, which upon ligand binding is capped to form a tunnel. However, the cleft on the uncharacterized protein is deeper and more pronounced than the cleft found on the Ega3 (Figure 7). This difference could be due to the enzyme's fCBD and low complexity regions from *L. ingoldianus* (Figure 3). Thus, it is hypothesized that the fCBD part found in this protein does not just function for recognizing galactosaminogalactan but also for holding it in place for maximum contact with the active site of the catalytic domain GH114.

Finally, the mode of action of the endo α-1,4 polygalactosaminidase from *Pseudomonas* sp. 881 on galactosaminooligosaccharides (GOSs) was studied by Tamura *et al*. (1992). This enzyme could hydrolyze α-1,4

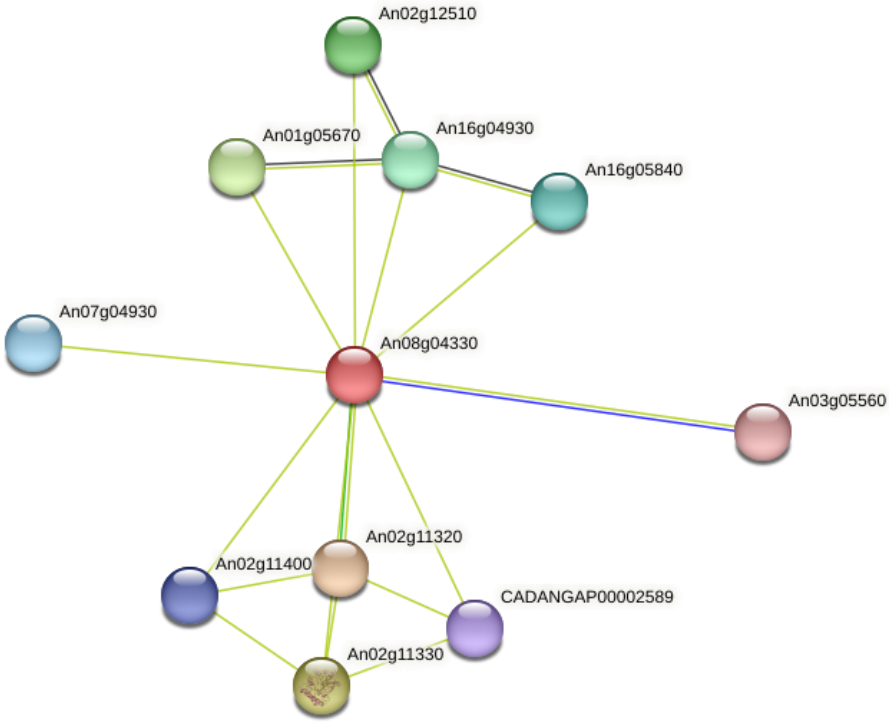

Figure 9. Interactions of endo-α-1,4-polygalactosaminidase from A. niger with other proteins as predicted by STRING v.11.5.

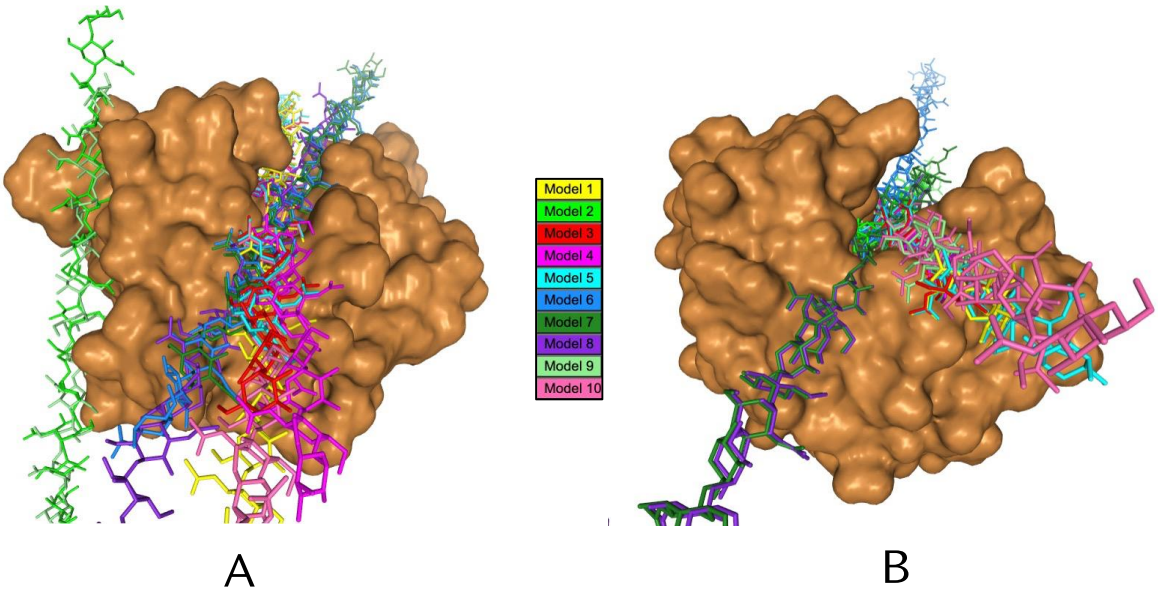

Figure 10. Molecular docking analysis results of GAG and Ega3 (A) and GAG and the uncharacterized protein from L. ingoldianus (B). All docking models of the substrate GAG were overlayed (sticks) on the enzyme endo-α-1,4-polygalactosaminidase (brown; surface view).

polygalactosamine to GOSs in an endo-split manner. They also found that tetraose and longer GOSs were hydrolyzed to galactosaminobiose and galactosaminotriose as the final products and galactosaminomonomer (galactosamine) could not be produced as an enzymatic product.

Protein Evolution

A total of 26 orthologous protein sequences were retrieved from NCBI after filtering only those accessions with above 90% identity in the BLASTp result. Molecular phylogenetic analysis of these sequences revealed that the endo α-1,4 polygalactosaminidase from *L. ingoldianus* is more closely related to CBM1 from *Zopfia rhizophila* (pink box; Figure 11A) than its other orthologues with a 74% bootstrap support value. The most parsimonious tree has a length of 1887, consistency index of 0.602014, retention index of 0.533250, and a composite index of 0.321024. *Z. rhizophila* is a plant pathogenic fungus that causes root rot in Asparagus (Sadowski, 1989).

Furthermore, although the orthologue from *Z. rhizophila* is named CBM1 in NCBI, comparison of their domain architectures showed both proteins are endo α-1,4 polygalactosaminidase (Figure11B). Moreover, a

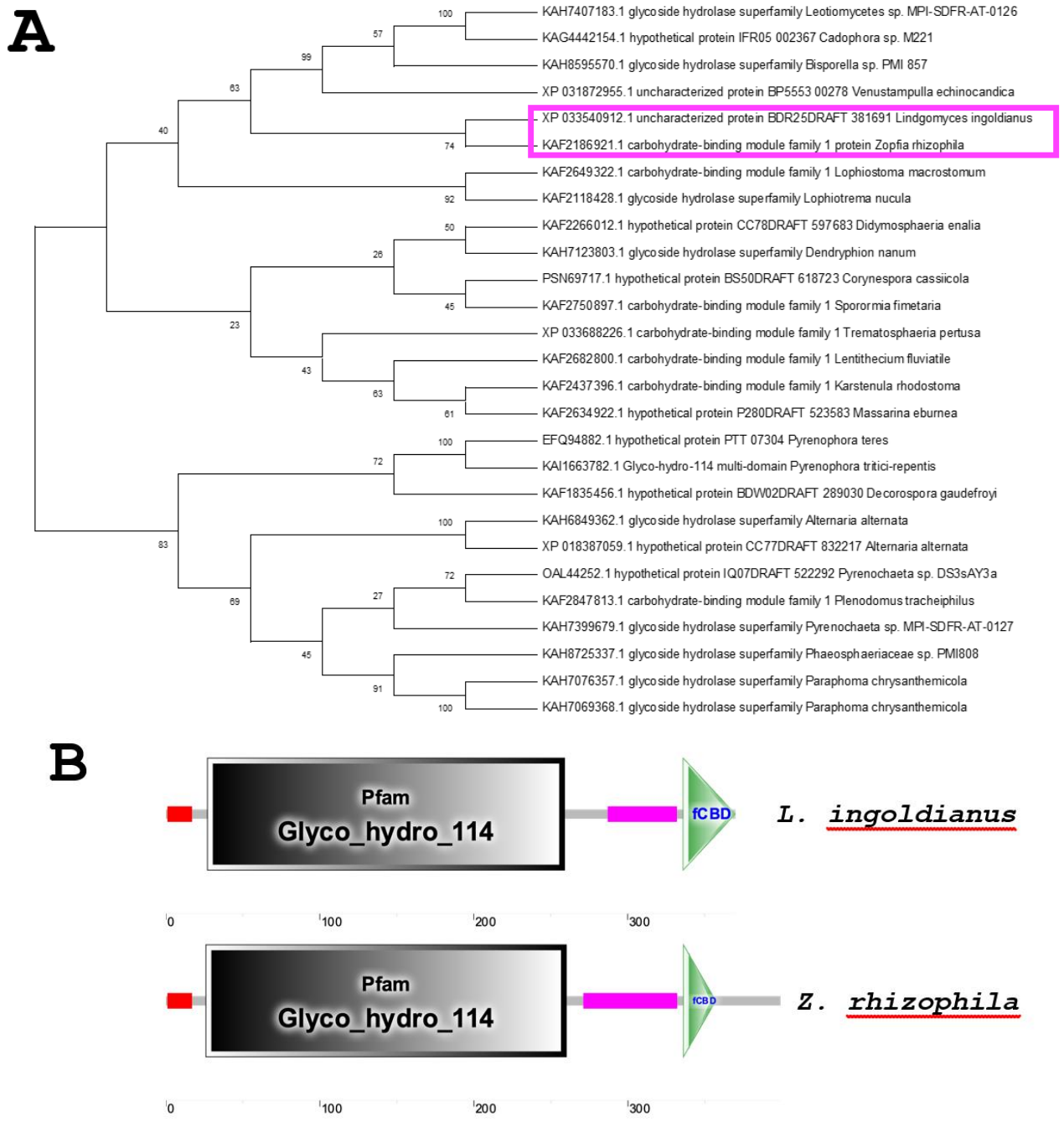

Figure 11. (A) Phylogenetic tree showing the relationship of endo-α-1,4-polygalactosaminidase from *L. ingoldianus* with its orthologues, (B) comparison of the domain architectures of the proteins from *L. ingoldianus* and *Z. rhizophila*.

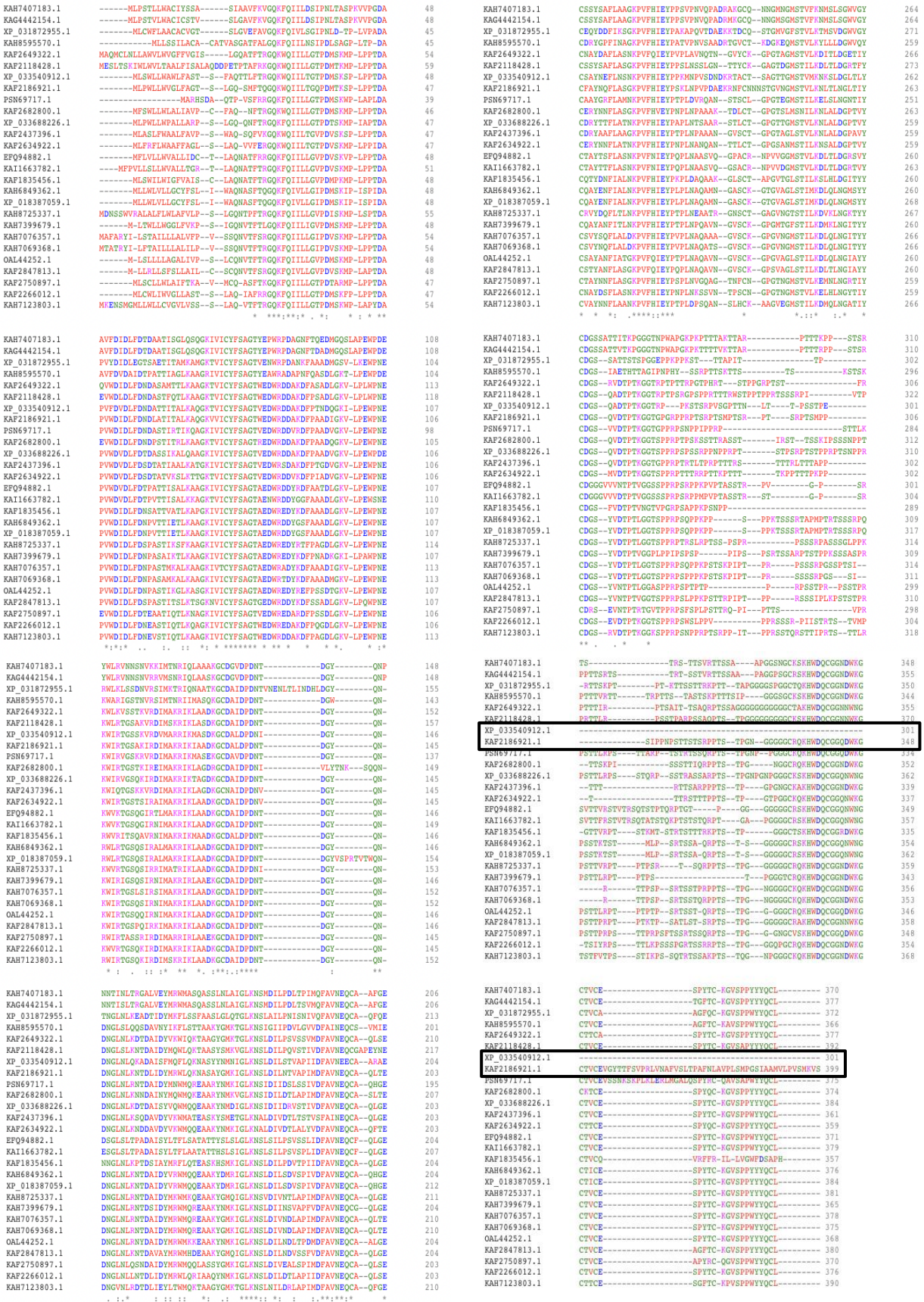

Figure 12. Multiple sequence alignment of endo-α-1,4-polygalactosaminidase from *L. ingoldianus* with its orthologues.

closer look at the alignment of the sequences of the proteins showed that both proteins differ from the rest of their orthologues in terms of the lacking and/or varying amino acid residues at their terminal ends (black box; Figure 12). Although, all these orthologous proteins obviously contain several conserved sequences making up the same domain architectures.

Finally, the BLASTp analysis did not reveal any similar protein sequences that are found on the same strain of species studied. Thus, analysis of paralogous genes embedded within the same genome of *L. ingoldianus* or any fungal species that produce endo α-1,4 polygalactosaminidase would be an interesting topic for future research.

Conclusion

This study confirms the translation of an uncharacterized protein BDR25DRAFT_381691 (Accession No.: XP_033540912) from an mRNA (Accession No.: XM_033697727) expressed by an unplaced genomic scaffold BDR25scaffold_49 (Accession No.: NW_022985210). This protein has 370 amino acid residues and is predicted to have a neutral charge and unstable and non-polar characteristics. It consists of a signal peptide, GH114, low complexity, and fCBD regions. Structural characterization and phylogenetic analysis revealed that this protein is an endo-α-1,4-polygalactosaminidase enzyme with unique domain architecture. This protein was also found to contain 36 active sites and was predicted to be secreted extracellularly. Lastly, molecular docking analysis showed that it could bind galactosaminogalactan in its active site much better than Ega3 from *A. fumigatus*. Suggesting that this enzyme produced by *L. ingoldianus* could be potentially used for more efficient degradation of biofilms or other polymers composed of galactosaminooligosaccharides. Finally, his endo-α-1,4-polygalactosaminidase from *L. ingoldianus* was found to be closely related to its orthologue from a plant pathogenic fungus *Zopfia rhizophila*.

Ethical Statement

Not Applicable.

Funding Information

The authors received no specific funding for this work.

Author Contribution

JGB: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, writing – Original Draft Preparation; MDS: Conceptualization, Methodology, Project Administration, Supervision, Validation, Writing – Review & Editing

Conflict of Interest

The authors declare that they have no known competing financial or non-financial, professional, or personal conflicts that could have appeared to influence the work reported in this paper.

Acknowledgements

The first author is grateful to the Visayas State University for its extensive faculty development program and to the Department of Science and Technology (DOST) of the Philippines for awarding the Ph.D. scholarship at the University of Santo Tomas through the Accelerated Science and Technology Human Resource Development Program (ASTHRDP).

References

Alexopoulos, C.J. and Mims, C.W. (1979) Introductory Mycology. 3rd Edition. Wiley, New York.

Almagro Armenteros, J.J., Sønderby, C.K., Sønderby, S.K., Nielsen, H., & Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, *33*(21), 3387-3395. https://doi.org/10.1093/bioinformatics/btx431

Anderson, R.J., Weng, Z., Campbell, R.K., & Jiang, X. (2005). Main-chain conformational tendencies of amino acids. *Proteins: Structure, Function, and Bioinformatics*, *60*(4), 679-689. https://doi.org/10.1002/prot.20530

Barrett, K., Jensen, K., Meyer, A.S., Frisvad, J.C., & Lange, L. (2020). Fungal secretome profile categorization of CAZymes by function and family corresponds to fungal phylogeny and taxonomy: Example Aspergillus and Penicillium. *Scientific reports*, *10*(1), 5158. https://doi.org/10.1038/s41598-020-61907-1

Bamford, N.C., Le Mauff, F., Subramanian, A.S., Yip, P., Millán, C., Zhang, Y., ... & Howell, P.L. (2019). Ega3 from the fungal pathogen Aspergillus fumigatus is an endo-α-1, 4-galactosaminidase that disrupts microbial biofilms. *Journal of Biological Chemistry*, *294*(37), 13833-13849. https://doi.org/10.1074/jbc.RA119.009910

Benoit, I., Culleton, H., Zhou, M., DiFalco, M., Aguilar-Osorio, G., Battaglia, E., Bouzid, O., Brouwer, C., El-Bushari, H., Coutinho, P.M., Gruben, B.S., Hildén, K.S., Houbraken, J., Barboza, L., Levasseur, A., Majoor, E., Mäkelä, M.R., Narang, H.M., Trejo-Aguilar, B., van den Brink, J., … de Vries, R.P. (2015). Closely related fungi employ diverse enzymatic strategies to degrade plant biomass. *Biotechnology for biofuels*, *8*, 107. https://doi.org/10.1186/s13068-015-0285-0

Dantuma, N.P., Heinen, C., & Hoogstraten, D. (2009). The ubiquitin receptor Rad23: at the crossroads of nucleotide excision repair and proteasomal degradation. *DNA repair*, *8*(4), 449-460. https://doi.org/10.1016/j.dnarep.2009.01.005

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., & Liang, J. (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Research*, 34(suppl_2), W116-W118. https://doi.org/10.1093/nar/gkl282

Eisenberg, D., Lüthy, R., & Bowie, J.U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol*, 277, 396-404. https://doi.org/10.1016/s0076-6879(97)77022-8

Feig, M. (2016). Local protein structure refinement via molecular dynamics simulations with locPREFMD. *Journal of chemical information and modeling*, *56*(7), 1304-1312. https://doi.org/10.1021/acs.jcim.6b00222

Gabriel, R., Mueller, R., Floerl, L., Hopson, C., Harth, S., Schuerg, T., ... & Singer, S.W. (2021). CAZymes from the thermophilic fungus *Thermoascus aurantiacus* are induced by C5 and C6 sugars. *Biotechnology for biofuels*, *14*(1), 1-13. https://doi.org/10.1186/s13068-021-02018-5

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M.R., Appel, R.D., & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*, 571-607.

Gilkes, N.R., Henrissat, B., Kilburn, D.G., Miller Jr, R.C., & Warren, R. (1991). Domains in microbial beta-1, 4-glycanases: sequence conservation, function, and enzyme families. *Microbiological reviews*, *55*(2), 303-315. https://doi.org/10.1128/mr.55.2.303-315.1991

Grant, O.C., Tessier, M.B., Meche, L., Mahal, L.K., Foley, B.L., & Woods, R.J. (2016). Combining 3D structure with glycan array data provides insight into the origin of glycan specificity. *Glycobiology*, *26*(7), 772-783. https://doi.org/10.1093/glycob/cww020

Haridas, S., Albert, R., Binder, M., Bloem, J., LaButti, K., Salamov, A., ... & Grigoriev, I.V. (2020). 101 Dothideomycetes genomes: a test case for predicting lifestyles and emergence of pathogens. *Studies in mycology*, *95*(1), 5-169. https://doi.org/10.1016/j.simyco.2020.01.003

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T.L. (2008). NCBI BLAST: a better web interface. *Nucleic acids research*, *36*(suppl_2), W5-W9. https://doi.org/10.1093/nar/gkn201

Letunic, I., Khedkar, S., & Bork, P. (2021). SMART: recent updates, new developments, and status in 2020. *Nucleic acids research*, *49*(D1), D458-D460. https://doi.org/10.1093/nar/gkaa937

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., ... & Lopez, R. (2013). Analysis tool web services from the EMBL-EBI. *Nucleic acids research*, *41*(W1), W597-W600. https://doi.org/10.1093/nar/gkt376

Naumov, D.G., & Stepushchenko, O.O. (2011). Endo-alpha-1-4-polygalactosaminidases and their homologs: structure and evolution. *Molekuliarnaia biologiia*, *45*(4), 703-714.

Newbound, M., Mccarthy, M.A., & Lebel, T. (2010). Fungi and the urban environment: A review. *Landscape and urban planning*, *96*(3), 138-145. https://doi.org/10.1016/j.landurbplan.2010.04.005

Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A., & Ghasemi, Y. (2018). A comprehensive review of signal peptides: Structure, roles, and applications. *European journal of cell biology*, *97*(6), 422-441. https://doi.org/10.1016/j.ejcb.2018.06.003

Reynolds, C.R., Islam, S.A., & Sternberg, M.J. (2018). EzMol: a web server wizard for the rapid visualization and image production of protein and nucleic acid structures. *Journal of molecular biology*, *430*(15), 2244-2248. https://doi.org/10.1016/j.jmb.2018.01.013

Sadowski, C. (1989). The occurrence of Zopfia rhizophila Rabenh. on asparagus roots in Poland. In VII International Asparagus Symposium 271 (pp. 377-382). 10.17660/ActaHortic.1990.271.53

Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., ... & von Mering, C. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, *49*(D1), D605-D612. https://doi.org/10.1093/nar/gkaa1074

Tamura, J.I., Abe, T., Hasegawa, K., & Kadowaki, K. (1992). The mode of action of endo α-1, 4 polygalactosaminidase from Pseudomonas sp. 881 on galactosaminooligosaccharides. *Bioscience, biotechnology, and biochemistry*, *56*(3), 380-383. https://doi.org/10.1271/bbb.56.380

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Molecular biology and evolution*, *38*(7), 3022-3027. https://doi.org/10.1093/molbev/msab120

Teufel, F., Almagro Armenteros, J.J., Johansen, A.R., Gíslason, M.H., Pihl, S.I., Tsirigos, K.D., ... & Nielsen, H. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, 1-3. https://doi.org/10.1038/s41587-021-01156-3

Toll-Riera, M., Radó-Trilla, N., Martys, F., & Alba, M.M. (2012). Role of low-complexity sequences in the formation of novel protein-coding sequences. *Molecular biology and evolution*, *29*(3), 883-886. https://doi.org/10.1093/molbev/msr263

"UniProt: the universal protein knowledgebase in 2021." *Nucleic acids research* 49, no. D1 (2021): D480-D489. https://doi.org/10.1093/nar/gkaa1100

Wiederstein, M. & Sippl, M.J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* 35, W407-W410. https://doi.org/10.1093/nar/gkm290

Yan, Y., Tao, H., He, J., & Huang, S.Y. (2020). The HDOCK server for integrated protein-protein docking. *Nature protocols*, *15*(5), 1829-1852. https://doi.org/10.1038/s41596-020-0312-x

Yang, J., & Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. *Nucleic acids research*, *43*(W1), W174-W181. https://doi.org/10.1093/nar/gkv342