MINISTRY OF EDUCATION
AND TRAINING

VIETNAM ACADEMY
OF SCIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

Nguyen Thi Binh

BACTERIAL DIVERSITY SURROUNDING WHITE-ROT FUNGI HYDROLYZES LIGNOCELLULOSE AND MINING CELLULASE-CODING GENES BY METAGENOMICS

SUMMARY OF DISSERTATION IN BIOTECHNOLOGY

Code: 9.42.02.01

Hanoi – 2023

The dissertation is completed at: Graduate University of Science and Technology, Vietnam Academy Science and Technology.

Supervisors:

Supervisor 1: Prof. Dr. Truong Nam Hai - Institute of Biotechnology
Supervisor 2: Dr. Le Thi Thu Hong - Institute of Biotechnology

Referee 1: Prof. Dr. Pham Xuan Hoi – Agricultural Genetic Institute
Referee 2: Assoc. Prof. Dr. Do Huu Nghi – Institute of Natural Products Chemistry
Referee 3: Assoc. Prof. Dr. Nguyen Quang Huy – VNU Hanoi - University of Science

The dissertation was examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at 9:00, 8th November 2023

The dissertation can be found at:

1. Graduate University of Science and Technology Library
2. National Library of Vietnam

INTRODUCTION

1. The urgency of the thesis

Lignocellulosic is a natural, inexpensive and abundant source of biomass that is continuously renewable and has been identified as a natural source of biomass to promote the bioindustry. However, converting this biomass by biological enzymes faces many difficulties, especially glycolysis. Therefore, the cost of biological products is still quite high compared to products produced with fossil fuels. In order to reduce costs as well as promote biological economic development, finding an enzyme that can effectively participate in the hydrolysis of cellulose plays an important role because cellulose contains a high percent in lignocellulosic biomass. The white-rot fungi and the bacteria in the soil surrounding the white-rot fungi are the ecosystems in which the decomposition of lignocellulose takes place strongly. This is a potential source for finding and exploiting enzymes of industrial value. In Vietnam, there has been no research on the diversity of bacteria in Cuc Phuong National Forest and the diversity of soil bacteria around the white-rot fungi, as well as the exploitation of fecal enzymes cellulose degrading bacteria in this ecosystem. For screening the desired enzyme from the microbiome without culture, metagenomics has many advantages. In order to evaluate the diversity of bacteria composition in the soil surrounding the white-rot fungi as well as to mine the enzyme encoding cellulase with many new properties without culture, we used metagenomic techniques to conduct the study: " ***Bacterial diversity surrounding white-rot fungi hydrolyzes lignocellulose and mining cellulase-coding genes by metagenomics*** ".

2. Research objectives of the thesis

To evaluate the diversity of the bacteria around the lignocellulose-degrading white-rot fungi and identify the diversity of enzymes involved in lignocellulose decomposition, exploiting and selecting cellulose-degrading enzymes with potential for practical application from the soil microbial flora around the white-rot fungi in Cuc Phuong national forest by metagenomics technique.

3. The main contents of the thesis

- Analyze and evaluate the diversity of bacteria in the humus surrounding the white-rot fungi using metagenomics techniques;
- Analyze and evaluate the diversity of enzymes involved in lignocellulosic decomposition of the soil microbiota surrounding the lignocellulosic white-rot fungi using metagenomics techniques;
- Mine new gene sequences encoding cellulase with potential applications by bioinformatics tools;
- Study the recombinant expression of a selected gene, purified and evaluated the properties of β-glucosidase enzyme.

4. Scientific and practical significance of the work

4.1. Scientific significance

- To evaluate the diversity of bacteria around the white-rot fungi, especially the diversity of bacteria producing lignocellulose-degrading enzymes without culture by metagenomics.
- Provide more DNA sequences encoding cellulase capable of degrading agricultural and industrial by-products containing cellulose.

4.2. Practical significance

New enzymes involved in decomposing cellulose-containing materials from bacteria in the soil surrounding white-rot fungi were identified. These enzymes play an important role in the production of second-generation biofuels and the biodegradation of environmental pollutants.

CHAPTER 1. DOCUMENTARY OVERVIEW

1.1. General overview of lignocellulose

Lignocellulose is composed of three main components: cellulose, hemicellulose and lignin. Cellulose and hemicelluloses are tightly bound to lignin. Cellulose is the main component in the structure of plant cell walls, usually accounting for 38-50%. Cellulose is composed of the D-glucose monomer and is responsible for the mechanical strength of plant cell walls. Next is hemicellulose accounting for 17-32% with heterogeneous structure, with high branching, usually composed of pentose and hexose single sugars. The hemicellulose creates cross-links between the cellulose. Lignin accounts for 15 - 30% including aromatic polyphenols, is biosynthesized and forms a wrap structure around two components cellulose and hemicelluloses, providing additional mechanical strength to the cell wall, resistance to insects or conditioning wet conditions.

1.2. Cellulases

Cellulase enzyme is one of three important groups of enzymes involved in the decomposition of lignocellulose biomass. Cellulase is classified as a glycoside hydrolase (GH) enzyme (EC 3.2.1.-), capable of cutting β-1,4-glycoside linkages in cellulose molecules to form cello-oligosaccharides, cellobiose and glucose products. . In nature, complete hydrolysis of cellulose is accomplished by the synergistic action of three main types of cellulase: (1) endoglucanase (EC 3.2.1.4) that hydrolyzes the intracellular β-1,4-glucoside bonds molecules of the cellulose chain randomly to generate new chain ends, (2) exoglucanase (EC 3.2.1.74, EC 3.2.1.91) cleaves the cellulose chain at the reduced and non-reducing ends to release cellobiose or glucose molecules soluble and (3) β -glucosidase (EC 3.2.1.21) hydrolyzes the cellobiose to glucose. In recent years, in the world, there have been many studies on exploiting cellulase genes using metagenomic techniques. In Vietnam, since 2012, Truong Nam Hai's research group has started using metagenomic techniques to exploit genes encoding lignocellulose hydrolysed enzymes from the Vietnamese termite gut microbiota, microorganisms in the rumen indigenous goats in Vietnam, microorganisms of Binh Chau hot spring…

1.3. White-rot fungi and the microflora around the lignocellulose hydrolyzed white-rot fungi

White-rot fungi are the only group of organisms capable of degrading all components of lignocellulose. When living on wood substrates, white-rot fungi are the group with the most effective ability to decompose lignin. This ability is due to the fact that white-rot fungihave a unique non-specific extracellular enzyme system as well as intracellular oxidative enzymes, from which white-rot fungi can completely mineralize lignin substrates into CO_2 and decompose a series of different substances. Therefore, white-rot fungi participate in the carbon cycle and play an important role in providing nutrients in tropical forests. To effectively decompose lignocellulose, not only fungi but also the microbial community in the soil surrounding the white-rot fungi are involved. The interactions between fungi and bacteria living in the same area can be supportive and/or competitive. During the process of wood decomposition by fungi, environmental conditions become selective for bacteria. Bacteria that exist in these conditions must have special and new properties. Therefore, microorganisms in the soil surrounding white-rot fungi are an important source for searching for lignocellulase degrading genes.

1.4. Metagenomic and some bioinformatics tools, databases used in DNA metagenome mining

Metagenomic is a method of studying multiple genomes (metagenomes) of all microorganisms obtained directly from natural environmental samples without going through culture. Metagenomics approaches multiple genomes in two main methods: (1) Gene isolation based on establishing a multi-genome DNA library and (2) DNA sequence mining and gene isolation based on sequencing data Direct multi-genome DNA.

Currently, to research and exploit and search for potential genes from multi-genome DNA data, there are usually 3 steps: (1) extracting and sequencing multi-genome DNA samples; (2) assemble short read segments into long contig segments; (3) use specialized software to estimate gene function. The genes of multigenome DNA are predicted in terms of taxonomy and gene function. Based on the similarity of the multi-genome DNA sequence of the sample obtained with the sequences of the reference databases, the taxonomic unit and function of the genes can be estimated. Currently, the database on taxonomic units of genes commonly used is the NR database, reliable databases on gene functions such as: KEGG, eggNOG, COG, KOG, PFAM... However, there is no database that fully contains everything. information about taxonomic units and biological functions of genes in multi-genome DNA. So consolidating databases in a single program is necessary. In this study, the gene encoding cellulase was exploited using metagenomics techniques based on multigenome DNA direct sequencing data. There are many bioinformatics tools that have been used to mine genes from multi-genome DNA data such as: BLAST, Periscope, Phyre2, AcalPred...

CHAPTER 2. MATERIALS AND METHODS

2.1. Materials and chemicals

- Samples of humus around white-rot fungi in Cuc Phuong National Park were taken in the rainy season (May - June of the year).

- Research location: Department of Genetic Engineering, Institute of Biotechnology, Vietnam Academy of Science and Technology

- Microbial strains of Invitrogen (USA) and Biochemistry laboratory, University of Saarland (Germany) were used as clones and gene expression strains. The vector pET22b(+) of Novagene (USA) was used as the expression vector of gh3s2 gene in *E. coli* Rosetta 1.

- Pair of primers to amplify the bacterial *16S rDNA* gene:

27F: 5'-GAGTTTGATCCTGGCTCAG-3'

1527R: 5'-AGAAAGGAGGTGATCCAGCC-3'

- Chemicals and equipment used in the experiments are of clear origin.

2.2. Methods

There are 3 groups of research methods used in the thesis, including:

2.2.1. Microbiological and molecular methods

2.2.2. Protein biochemistry methods

2.2.3. Bioinformatics methods

The research process of the thesis is carried out as shown in Figure 2.2.

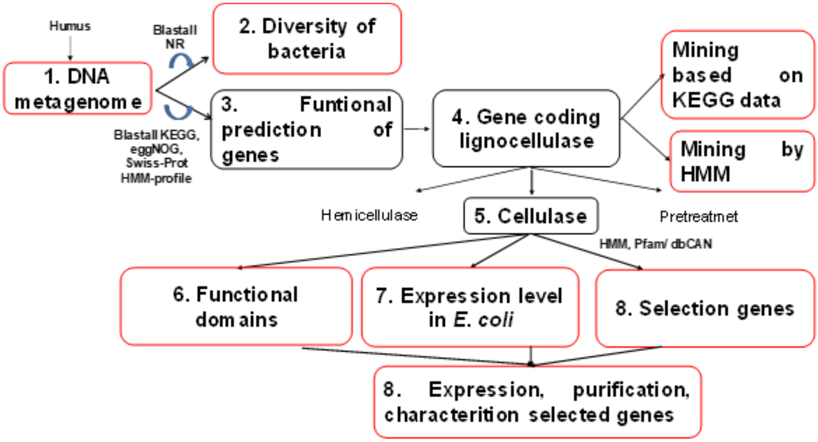
Humus
Blastall NR
1. DNA metagenome
2. Diversity of bacteria
3. Funtional prediction of genes
Blastall KEGG, eggNOG, Swiss-Prot HMM-profile
4. Gene coding lignocellulase
Mining based on KEGG data
Mining by HMM
Hemicellulase
5. Cellulase
Pretreatmet
HMM, Pfam/ dbCAN
6. Functional domains
7. Expression level in *E. coli*
8. Selection genes
8. Expression, purification, characterition selected genes

Figure 2.1. Flowchart of the research process in the thesis

CHAPTER 3: RESULTS AND DISCUSSION

3.1. Investigate the diversity of humus bacteria surrounding white-rot fungi

3.1.1. Extraction and purification of DNA metagenome

Humus samples were collected and mixed together to ensure microbial diversity for DNA metagenome extraction. The results of successful DNA metagenome extraction with DNA concentrations from 112 to 145 ng/µl, A260/A280 are in the range of 1.8 to 2, when PCR with 16S primer gives good results. Thus, DNA metagenome was extracted and purified, eligible for sequencing.

3.1.2. Results of DNA metagenome sequencing

From about 100 µg of DNA metagenome was sequenced, a total size of about 51.82 Gb was obtained (Table 3.2).

Table 3.2. Results of DNA metagenome sequencing using the next generation sequencing system HiSeq Illuminar

Category		**Metric**	**Unit**
Read	Number	345.471.086	read
	Total base	51.820.662.900	base pair
Contig	Number	2.611.883	contig
	Avarage contig length	898	base pair
	Contig N50	1117	base pair
	Maximum contig length	611.845	base pair
Gene	Number	4.104.872	gene
	Avarage gene length	505	base pair
	Gene N50	615	base pair
	Maximum gene length	20.541	base pair

3.1.3. Analysis of soil microbial diversity

Table 3.3. Diversity of bacteria from DNA metagenom data using MEGAN software (version 6) based on NR database

	Number of genes	**Proportion (%)**	**Phyla**	**Class**	**Order**	**Family**	**Genera**	**Species**
Bacteria	3.884.879	99,69	111	83	170	406	1971	738
Archaea	293	0,01	9	12	18	23	50	8
Eukaryotae	1144	0,03	7	26	46	79	113	86
Viruses	10565	0,27	0	0	2	14	101	84
Toatal	3.896.881	100	131	118	237	523	2240	916

From data 51.82 Gb DNA metagenome of humus bacteria surrounding white-rot fungi in Cuc Phuong national park, 4,104,872 protein-coding genes have been identified. Of which, 3,923,046 genes (about 95.57%) are annotated in the NR database. By MEGAN software, these genes were identified and classified, with 3,896,881 genes classified into the kingdoms of bacteria, Eukaryotes, Archaea and viruses. In which, the number of genes classified as bacteria is absolutely dominant with 3,884,879 genes (accounting for 99.69% of total genes), the rest are archaea with 293 genes (0.01%), eukaryotes with 1144 genes (0.03%) and viruses with 10,565 genes (0.27%). Thus, bacteria have the largest number of genes and bacterial genes are classified into 111 phyla, 83 classes, 170 orders, 406 families, 1971 genera and only 738 species have been identified. In this bacterial kingdom, 93.26% of the total genes were identified at the taxonomic level. Of the 111 phyla identified, there are 5 common phyla, accounting for 92.59% of the total, and the rest are other phyla. Among them, Proteobacteria is the most common phylum with 3,106,400 genes accounting for about 75.68%. The following phyla are Bacteroidetes accounting for 13.11%, Actinobacteria 1.6%, Firmicutes 1.4%, Acidobacteria 0.8%. Thus, Proteobacteria is the dominant phylum with 5.77 times the number of genes than the second most common Bacteroidetes.

3.2. Mining genes encoding lignocellulase

3.2.1. Functional prediction of DNA metagenome

Genes were functionally annotated against the Swiss-Prot, KEGG, eggNOG, Nr and HMM-profile databases. There are 3,925,740 genes (95.64%) that are functionally annotated based on at least one of the above databases. Based on KEGG data, there are 2,809,791 genes (corresponding to about 68.45% of total genes) that have been identified as encoding proteins involved in the metabolism of substances in cells and the body. These proteins are involved in five metabolic groups: cellular processes, environmental information processing, genetic information processing, human disease, and metabolism. In the metabolism of various substances, carbohydrate metabolism has 297,103 protein-coding genes involved (accounting for about 13.56% of the total genes involved in metabolism).

3.2.2. Mining genes based on results annotated by KEGG

From 297,103 genes that are functionally identified as participating in carbohydrate metabolism on the KEGG database, there are 22,226 genes that are estimated to be genes encoding enzymes involved in lignocellulosic biomass degradation, including: (1) 907 genes encoding enzymes involved in biomass pretreatment are divided into 4 groups: pectinesterase 89.96%, feruloylesterase 8.27%, laccase 1.10% and expansin 0.67%; (2) 8301 genes encoding cellulase (including 5 groups arranged in descending order of β-glucosidase, endoglucanase, 6-phospho-beta- glucosidase, cellobiohydrolase, cellobiose phosphorylase); (3) 13,018 hemicellulase-encoding genes are classified into 20 groups, of which the xyloglucan-active β-D-galactosidase group is the most common 25.26% (3288 ORF).

3.2.3. Mining genes based on HMM representative model

Using the HMM model which is essentially based on motif similarity, 13 enzymes involved in lignocellulosic hydrolysis were exploited more efficiently than gene extraction based on sequence similarity in KEGG. These are CBM (1-84), arabinanase (GH43), galactanase, glucuronyl esterase, HPOXRE catalase, hydrogen peroxide oxidoreductase, LPMO, laccase, acetylxylanesterase, beta-glucuronidase, cellobiohydrolase, lichenase, beta-xylosidase. Of these, hydrogen peroxide oxidoreductase (belonging to the hemicellulase group) and LPMO (pretreatment enzyme) were not found based on the KEGG, CAZy data. This shows that when using the new tool HMM representative model, important groups of enzymes were found. This is the basis for a more complete understanding of the enzyme system involved in lignocellulose hydrolysis.

Table 3.6. Compare the results of identifying genes encoding lignocellulolytic enzymes using HMM and KEGG models

No	Enzyme	Number of genes based on HMM	Number of genes based on KEGG
1	*CBM (1-84)*	*3163*	*< 300*
2	*Arabinanase (GH43)*	*343*	-
3	*Galactanase*	*17*	-
4	*Glucuronyl esterase*	*22*	-
5	*HPOXRE catalase*	*224*	-
6	*Hydrogene peroxide oxidoreductase*	*224*	*0*
7	*LPMO*	*69*	*0*
8	*Laccase*	*1115*	*10*
9	*Axetylxylanesterase AXE1*	*79*	*1*
10	*β-glucuronidase*	*1044*	*277*
11	*Cellobiohydrolase*	*253*	*73*
12	*Lichenase*	*290*	*175*
13	*β-xylosidase*	*945*	*659*
14	β-mannosidase GH2	594	611
15	Xylanase (GH44)	599	659
16	Feruloylesterase	53	75
17	α-glucuronindase (GH76N)	102	161
18	α-L-arabinofuranosidase	431	1016
19	β-glucosidase	1118	4272
20	Endoglucanase	557	2216
21	Polygalacturonase	45	223
22	Mannanase	40	368
23	Xyloglucanase	14	3288
24	Expansin	0	7

3.2.4. Bacterial diversity of carrying genes encoding lignocellulase

Among 22,226 genes encoding enzymes involved in lignocellulose degradation, 22,092 genes (accounting for 99.39%) belong to bacteria, classified into 28 phyla, the most dominant being Proteobacteria (11,288 genes, accounting for 50.79%), followed by phylum Proteobacteria (11,288 genes, accounting for 50.79%). followed by Bacteroidetes (8,164 genes, 36.73%). The ratio of Bacteroidetes/Proteobacteria (0.72:1) in the gene encoding the enzyme involved in lignocellulosic degradation was much higher than this ratio in the total bacterial structure of the humus surrounding the white-rot fungi (0.17: 1). This suggests that Bacteroidetes play an important role in the hydrolysis of lignocellulose. At the order level, the analysis also showed that Enterobacterales was the most prominent order accounting for 20.06%, followed by Flavobacters 15.14%, Sphingobacteria 11.62%. Further analysis with the group of pretreatment enzymes showed that the phylum Bacteroidetes was the most abundant (427 genes, accounting for 47.08%), slightly higher than that of Proteobacteia (45.31%). Meanwhile, for the hemicellulase group, Proteobacteria (44.20%) were higher than Bacteroidetes (43.52%). For cellulase, the ratio between Proteobacteria and Bacteroidetes was 2.4 times with Proteobacteria 61.72% and Bacteroidetes 24.96% respectively. Thus, the ratio of Proteobacteia/Bacteroidetes in microbial polygenomic DNA around white-rot fungi is 5.77, while for cellulase, the ratio of Proteobacteria/Bacteroidetes is 2.4. Therefore, Bacteroidetes seem to play a more important role in the hydrolysis of lignocellulose.

3.3. Mining, selecting potential genes encoding cellulase

3.3.1. The functional regions of cellulase

Based on the KEGG database, there are 8301 ORFs encoding cellulase enzymes, including 1279 complete ORFs, 1058 complete ORFs with 81 domains including: 367 ORF encoding endoglucanase, 6 ORF encoding cellobiohydrolase, 475 ORF encoding endoglucanase. encoding β-glucosidase, 210 ORFs encode 6-phospho β-glucosidase. In which, the most popular domain belongs to the GH family (accounting for over 80% of complete ORFs with domains). Representative of GH1 has 245 ORFs, of which 189 ORFs (77.14%) belong to Proteobacteria phylum, 20 ORFs

(8.16%) belong to Bacteroidetes phylum and the rest belong to other phyla. Next is the domain GH3+FN3 (220 ORF) in which 96 ORFs (43.67%) belong to the phylum Proteobacteria, 108 ORFs (49.09%) belong to the phylum Bacteroidetes. Then there are other GH families such as GH8 (105 ORF), GH5 (72 ORF), GH4 (58 ORF) in which the proportion of ORFs belonging to the Proteobacteria phylum is 91.43%, respectively; 52.78%, 94.83%.

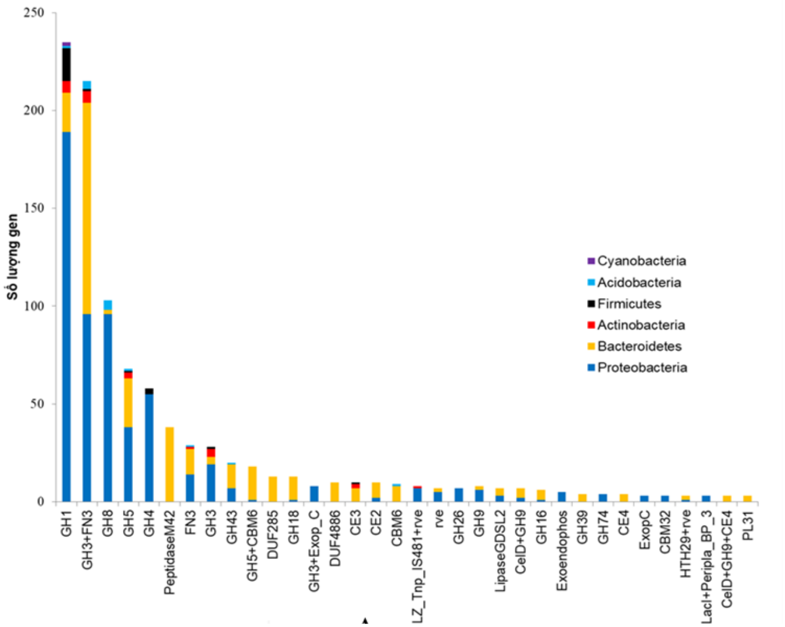

Figure 3.5. Bacterial phylums carrying complete ORFs have cellulase coding

Analysis by enzyme group showed that there were 367 ORFs in the endoglucase group with 47 domains. In which, domain GH8 is the most common with 105 ORFs, 96 of these (91.43%) belong to the Proteobacteria phylum, only 2 ORFs (1.90%) belong to the Bacteroidetes phylum. The second most common domain in this group of enzymes is GH5 with 72 ORFs. The ORFs containing the GH5 domain mostly belonged to two phyla Proteobacteria (52.78%) and Bacteroidetes (34.72%). In the group of exoglucanase enzymes, there are only 6 ORFs with 6 different domains. Among the ORF domains encoding exoglucanase, there are 3 domains Alginate_lyase, Amidase 3, GH128+Laminin G3 belonging to phylum Bacteroidetes, 2 domains CBM2 and Znribbon8 belonging to phylum Acidobacteria, domain CBP_BcsO belonging to phylum Proteobacteria. The group of β-glucosidase enzymes is the group of enzymes with the highest number of ORFs with 475 ORFs (44.90%) with 27 domains. In this group of enzymes, the domain with the most number is the GH3+FN3 domain with 220 ORFs, of which 96 ORFs (43.64%) belong to the Proteobacteria phylum and 108 ORFs (49.09%) belong to the Bacteroidetes phylum, the remaining 16 ORFs belong to the phylum Proteobacteria. other industries. Next is the GH1 domain (93 ORFs), of which 60 ORFs (64.52%) belong to the phylum Proteobacteria, 20 ORFs (21.51%) belong to the phylum Bacteroidetes and the remaining 13 ORFs belong to some other phyla. In the group of 6-phospho-β-glucosidase, the most common domain is GH1 with 152 ORFs, of which 129 ORFs (84.87%) belong to the Proteobacteria phylum, 15 ORFs belong to the Firmicutes phylum, the rest are unclassified. The remaining domain in this group of enzymes is GH4 with 58 ORFs, of which 55 ORFs (94.83%) belong to the Proteobacteria phylum and 3 ORFs belong to the Firmicutes phylum.

3.3.2. Prediction of expression levels of cellulase genes

The results of determining the expression level in E. coli showed that in 1058 complete genes with cellulase coding domains, genes belonging to the endoglucanase and β-glucosidase groups were determined to have higher expression than the exoglucanase group. In the endoglucase group, the genes containing the GH8 domain were predicted to have the highest expression levels with the codes GL0183420, GL1155166, GL0051672, GL0127466, GL0176868 all capable of expressing over 3000 mg/l. These genes belong to the phylum Proteobacteria and Acidobacteria. Next are genes containing the GH5 domain belonging to the phylum Proteobacteria and Bacteroidetes with expression levels above 2000 mg/l. In addition, some genes containing the domains PeptidaseM42, GH5-CBM6, DUF285 belonging to the Bacteroidetes phylum were also well expressed in the E. coli expression system. In the β-glucosidase group, genes with GH3 domain belonging to the Proteobacteria phylum have good expression: many representatives of the GH3+FN3 domain structure have high expression levels above 4000 mg/l, genes with GH3 domain have high expression levels. above 1800 mg/l, gene code GL0050362 with domain structure GH3+Exop_C had the highest expression level at 4626 mg/l. Besides, the β-glucosidase genes containing domains GH4, GH43, GH1 belonging to the Proteobacteria phylum, β-glucosidase genes containing the GH16 domain belonging to the Bacteroidetes phylum were also well expressed. Group 6-phospho β-glucosidase, gene with GH1 domain has the highest expression level of 4714 mg/l, some genes containing GH4 domain have expression level over 1000 mg/l.

3.3.3. Selection of genes encoding cellulase

3.3.3.1. Prediction domain regions of genes using BLASTp

The results of determining the protein structure of the GL0050362 gene encoded by BLASTp showed that this protein has three specific hit regions, including: (1) the BglX region (belonging to the BglX superfamily) corresponding to two nonspecific regions (non-specific hit) PRK15098 and GH3-N (data linked to pfam00933) encode β-glucosidases and glycosidases involved in carbohydrate metabolism (database COG1472); (2) is the GH3-C region (belonging to the GH3-C superfamily) involved in catalysis and can bind beta-glucan (as linked to pfam01915); (3) Exop_C (belonging to the Exop_C superfamily) resembles the Galactose-binding region, which is the C-terminal region found in ExoP (exo-1,3/1.4-beta-glucanase) from Pseudoalteromonas. This region contains a β-fold commonly found in glycosyl hydrolases (GH7, 11, 12 and 16) and in several carbohydrate regions/topologies. This region is thought to not only play a role in directing substrate binding, but also to help stabilize the structure required for ExoP activity.

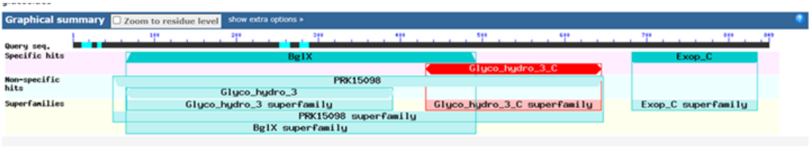

Figure 3.6. Prediction of GL0050362 gene function by BLASTp.

3.3.3.2. Spatial structure prediction and substrate binding of gh3s2 protein

In the three-dimensional structural model, the protein was identified based on the β-glucosidase enzyme template from Pseudoalteromonas sp. bb1 (c3f93D) has 93% coverage and 100% confidence. The tertiary spatial structure of this gene has 47% similarity with β-glucosidase of the c3f93D template with 100% confidence, has three specific regions GH-3, GH-3-C and Exop_C, in addition, this gene has The highly conserved region [HIS]249 is similar between the candidate protein and the c3f93D template, on the other hand the candidate enzyme also has the catalytic site [GLY]848 involved in β-glucosidase activity as estimated by Phyer2.

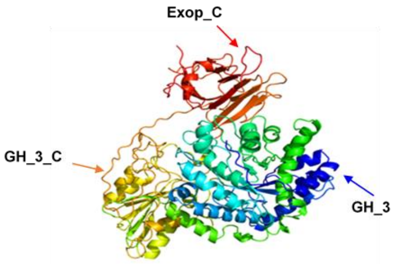

Figure 3.7. Structural modeling of candidate genes using Phyre2 based on c3f93D template

3.3.3.3. Prediction of some properties of candidate enzymes

AcalPred software determined the acid/alkaline tolerance of GH3S2 to be 0.507957 and 0.49204, respectively, this enzyme has a neutral pH, slightly acidic. The results of the investigation of the heat tolerance of the GH3S2 protein with a Tm of 0.6606, thus, the optimal temperature for enzyme activity is predicted from 55°C-65°C.

3.4. Expression, purification and characterization of GH3S2

3.4.1. Gh3s2 gene expression

3.4.1.1. Design a recombinant vector carrying gh3s2

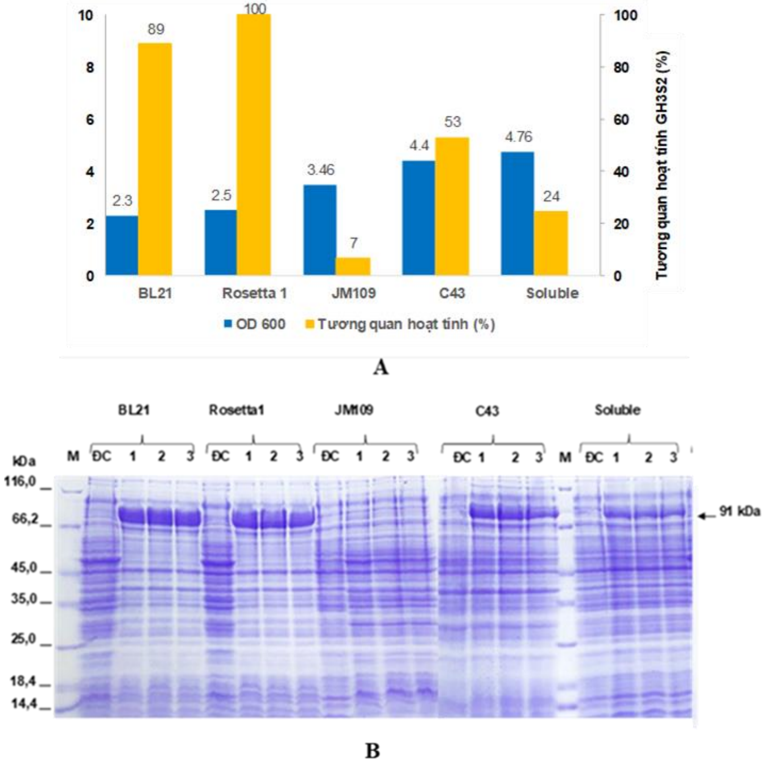

Figure 3.9. (A). Cell density and enzyme activity were obtained when expressed in expressed strains; (B). Electrophoresis of total GH3S2, DC: total protein of non-carrying vector control; 1, 2, 3: total protein of different lines 1, 2, 3 carrying the IPTG-inducible gh3s2 gene; M: standard protein (Fermentas)

After being selected and optimizing the code, the *gh3s2* gene will be synthesized and attached to the pET22b(+) expression vector. Recombinant plasmid DNA pET22b(+)gh3s2 will be transformed into *E.coli* DH10b clone cells by heat shock method and perform plasmid extraction.

3.4.1.2 Selection of strains expressing GH3S2

We studied GH3S2 expression in 5 strains of *E. coli* including: BL21, Rosetta 1, JM109, C43, Soluble. The results showed that GH3S2 was expressed in strains BL21, Rosetta 1 (Figure 3.9B). Of these two strains, in the Rosetta strain with higher cell density, the relative activity of the total enzyme when expressed in the Rosetta strain was higher (Figure 3.9A). Combining the expression level, cell density and enzyme activity obtained, we selected Rosetta 1 as the GH3S2 expression strain. After expression in strain *E. coli* Rosetta 1, the GH3S2 enzyme's β-glucosidase activity was tested with the esculin substrate. The results showed that the brown ring size for different samples had different levels of β-glucosidase activity. This demonstrates that the GH3S2 protein was successfully expressed in soluble form and showed obvious activity (Figure 3.10).

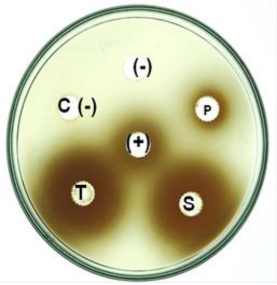

Figure 3.10. Check the activity of GH3S2 on LB agar plate using esculin substrate.

3.4.1.3. Study on the effect of culture temperature on the expression of GH3S2 in E. coli Rosetta 1

The *E. coli* Rosetta 1 strain carrying recombinant DNA was cultured at the following temperature conditions: 18°C, 20°C, 25°C, 30°C and 37°C. To improve expression efficiency, reduce energy costs, we chose 25°C as the temperature for GH3S2 expression in subsequent studies.

3.4.1.4. Study on the influence of culture medium composition on the expression of GH3S2 in E. coli Rosetta 1

Five cultures were examined including: LB, TB, modified TB, SB and PE. The medium of choice for expression of the recombinant protein GH3S2 is the modified cell medium.

3.4.1.5. The effect of IPTG inducer concentration on the expression of GH3S2

Different IPTG concentrations from 0.05 mM, 0.1mM, 0.3 mM, 0.5 mM, 1mM, 1.2 mM, 1.5 mM were tested for recombinant GH3S2 expression. The results showed that the appropriate IPTG concentration was selected as 0.3 mM for further studies.

3.4.1.6. The effect of optical density at induction on the expression of GH3S2 in E. coli Rosetta 1

To know the optimal cell density for the induction of GH3S2 protein synthesis, we investigated the OD at induction of 0.4; 0.8; 1.0; 1.5; 3.0; 4.0. The results showed that when the induction cell density was 1, the GH3S2 protein had the best expression level and high activity.

3.4.1.7. The optimal GH3S2 collection time after induction

To determine the appropriate post-induction time period for the expression of recombinant GH3S2 protein, fermentation samples were collected at intervals of 1, 2, 3, 4, 5, 6 and 22 hours after induction. . The results showed that the best amount of recombinant GH3S2 protein could be achieved after 4 hours after induction.

3.4.2. Purification GH3S2 by affinity chromatography column

When purified, GH3S2 adheres well and is removed from the substrate in the target protein fractions at a concentration of 300 mM imidazole. The results obtained that the GH3S2 content in the purified sample was 1.54 mg/ml. Thus, in 1 liter of culture solution, total protein amount of 735.68 mg was obtained, of which the amount of purified GH3S2 obtained was 41.80 mg.

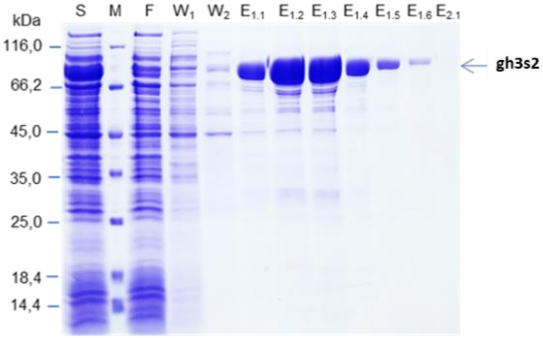

Figure 3.1. Electrophoresis of product in GH3S2 enzyme purification fractions on 12.5% polyacrylamide gel

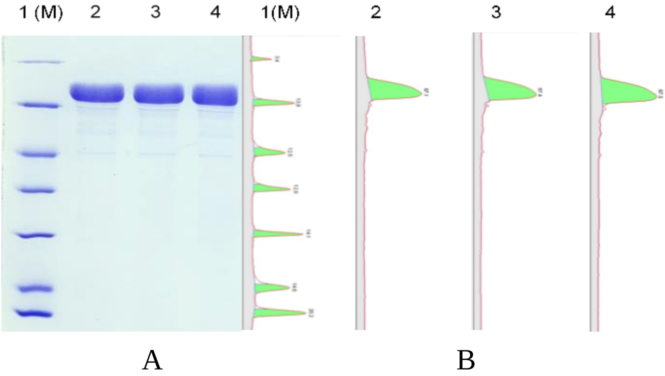

Figure 3.17. GH3S2 purity test results after purification (A). Electrophoresis of GH3S2 after purification (2 µg); (B). Cleanness measurement results using Image Lab software

To determine the purity of the GH3S2 protein, we used SDS-PAGE electrophoresis with a defined amount of sample and analyzed the results using Image Lab software to evaluate the relative purity of the GH3S2. The resulting purity of GH3S2 is 97.3%. The β-glucosidase activity of purified samples and total protein samples was determined in which the activity of total protein was 0.156 ± 0.01 U/mg, the activity of purified GH3S2 was 1.10 ± 0.02 U/mg. Thus, the GH3S2 protein was purified 7.05 times, the purification efficiency was 40.06% (Table 3.10).

Table 3.10. Summary table of recombinant GH3S2 purification performance (*: calculated per 1 liter of culture solution)

Step	**Total protein (mg)**	**Total activity (U)**	**Specific activity (U/mg)**	**Purification fold**	**Yield (%)**
Crude extract	735.68±0,6	114.77±1.2	0.156±0,01	1	100
Purified GH3S2	41.80±0,3	45.98±0,8	1.10±0,02	7.05	40.06

3.4.3. Properties of recombinant protein GH3S2

3.4.3.1. Effect of temperature on the GH3S2 activity and thermal stability of GH3S2 enzyme

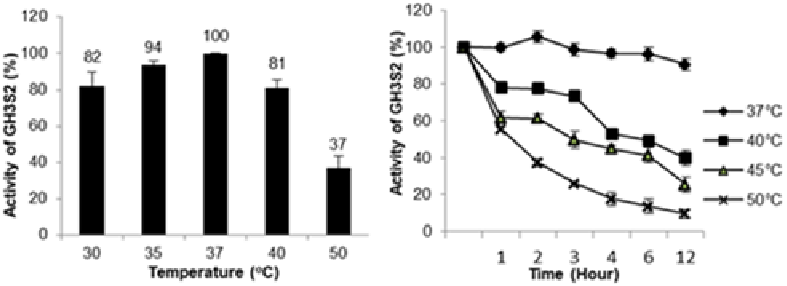

Figure 3.18. Effect of temperature on the activity and thermal stability of GH3S2 enzyme over time

The effect of temperature on the GH3S2 activity were tested. The results showed that at 37°C the enzyme had the highest activity. For thermal stability testing, the results showed that at a temperature of 37°C the activity of GH3S2 was almost stable, after 12 hours the enzyme activity still reached 90.78%.

3.4.3.2. Effect of pH on the activity and pH stability of GH3S2

Effect of pH on the activity showed that at pH 6.0 the GH3S2 had the highest activity. When testing the stability of the enzyme, at pH 6.0 the enzyme was maintained at about 70% of the activity after 6 hours, at pH 7.0, the enzyme activity was maintained at 70% after 6 hours. After 4 hours, it decreased and the higher the pH, the lower the enzyme activity.

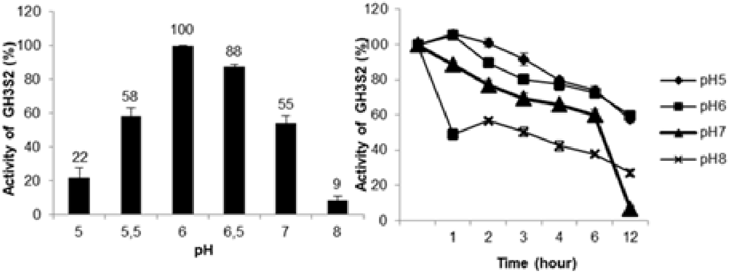

Figure 3.19. Effect of pH on the activity and pH stability of enzyme GH3S2

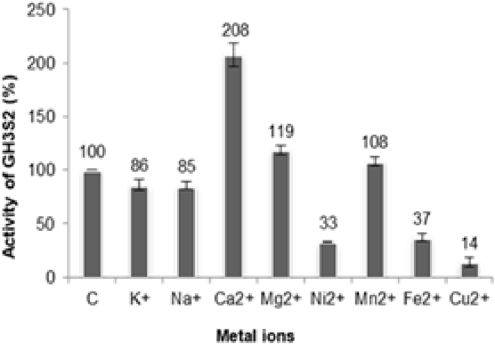

Figure 3.20. Effect of metal ions on GH3S2 activity

3.4.3.3. Effect of some metal ions on GH3S2 activity

Effect of metal ions on GH3S2 activity showed that when adding Ca^{2+}, Mg^{2+}, and Mn^{2+}, the enzyme activity was 208%, 119%, and 108%, respectively. Meanwhile, Fe^{2+}, Ni^{2+}, Cu^{2+} ions strongly reduced enzyme activity to 37%, 33%, 14%, respectively. K^{+}, Na^{+} ions have negligible influence on GH3S2 activity.

3.4.3.4. The effect of glucose on GH3S2 activity

β-glucosidase is an enzyme sensitive to the presence of glucose. Finding an enzyme that can tolerate the presence of glucose is important when degrading cellulose in industry. The effect of glucose on GH3S2 activity was carried out at different glucose concentrations from 2, 4, 6, 8, 10, 15, 20, 25, 30, 50, 100, 150, 200, 250, 300 mM. The results showed that, at a glucose concentration of 6 mM, the enzyme activity was only slightly affected and maintained about 70% then decreased to only 6% when the glucose concentration increased to 300 mM. The GH3S2 enzyme is sensitive to the presence of glucose.

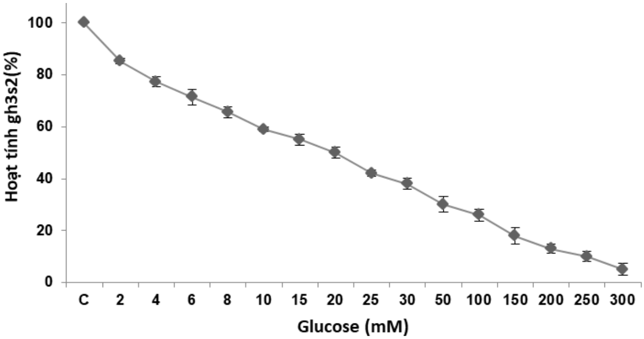

Figure 3.21. Effect of glucose on GH3S2 activity

3.4.3.5. Kinetic characteristics of the enzyme GH3S2

The kinetic constants of GH3S2 (Km, Vmax) were determined. Tthe reaction rate dependence on pNPG concentration follows the equation y = 0.9275x + 0.2037 with confidence R2 = 0.9953 (Figure 3.22). Based on the equation 1/v = Km.1/Vmax.1/[S] + 1/Vmax, the Km and Vmax values of

GH3S2 are calculated as 4.55 mM and 4.91 μmol/min, respectively. Under this condition, the enzyme GH3S2 has a specific activity of 2.23 U/mg with the pNPG substrate.

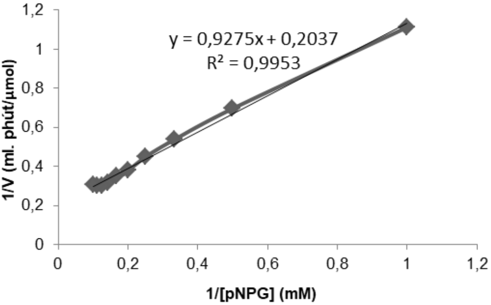

Figure 3.22. Reaction rate dependence of GH3S2 on pNPG substrate concentration according to Linewever – Burk

CONCLUSIONS AND RECOMMENDATIONS

CONCLUDE

1. Built a 51.82 Gb DNA metagenome data of the humus bacteria around the white-rot fungi in Cuc Phuong National Park and analyzed 3,896,881 ORFs belonging to 131 branches, 118 classes, 237 orders, 523 families, 2240 genera and 916 species, of which 3,884,879 ORFs belong to the bacterial kingdom and belong to 111 phyla, 83 classes, 170 orders, 406 families, 1971 genera and 738 species. Proteobacteria were the most common phylum with 3,106,400 genes (75.68%) and the phylum Bacteroidetes was the second largest (13.11%);

2. Based on the KEGG database, the functions of 22,226 genes encoding enzymes involved in lignocellulose hydrolysis have been annotated, in which 907 genes encoding enzymes involved in pretreatment, 8301 genes encoding cellulase enzymes and 13,018 genes encoding hemicellulases. There are 22,092 genes classified into 28 phyla of bacteria, in which, the most dominant are Proteobacteria (50.79%) and Bacteroidetes (36.73%). 13 families of enzymes involved in lignocellulosic hydrolysis

have been exploited using a representative model of HMM;

3. Among 8301 cellulase-encoding genes annotated by KEGG database, there are 1058 complete genes analyzed for functional regions: (1) endoglucanases with 47 domains (367 genes), in which domain GH8 is the most popular; (2) exoglucanase with 6 domains (6 genes); (3) β-glucosidase with 27 domains (475 genes), in which the common domains is GH3, GH1; (4) 6-phospho-β-glucosidase with 2 domains GH1 and GH4 (210 genes); a gene encoding a potential cellulase enzyme has been identified;

4. Successfully expressed GH3S2 in *E. coli* Rosetta1 at 25°C, modified medium, 0.3 mM IPTG, induction at OD_{600} =1 and sampled 4 hours later. The recombinant GH3S2 was purified with a purity of 97.3%, the content reached 41.8 mg/liter of fermentation liquid; GH3S2 enzyme has Km = 4.55 mM and Vmax = 4.91 U/mg. Ca^{2+} and Mg^{2+} ions increase enzyme activity, while Ni^{2+} and Cu^{2+} ions decrease activity. 6 mM glucose had a slight effect on the activity of GH3S2.

RECOMMENDATIONS

Studying the ability to combine β-glucosidase GH3S2 with endoglucanase and exoglucanase enzymes to evaluate the hydrolysis efficiency of cellulose.

NEW CONTRIBUTIONS OF THE THESIS

- This is the first study on bacterial diversity around lignocellulosic white-rot fungi in Cuc Phuong National Forest using Metagenomics technique.

- Investigate a variety of enzymes involved in lignocellulose degradation in the humus bacteria surrounding white-rot fungi in Cuc Phuong national forest, selected and evaluated the properties of the enzyme β-glucosidase GH3S2 from the DNA metagenome of humus bacteria around white-rot fungi.

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

1. **Nguyen Thi Binh**, Dao Trong Khoa, Le Thi Thu Hong, Truong Nam Hai, Research and exploitation of genes encoding multicoronoxidase enzymes from metagenome data of bacterial flora around white-rot fungi (*Trametes versicolor*) in Cuc Phuong national forest, National Biotechnology Conference, 2020, pp. 187-192.

2. **Nguyen Thi Binh**, Nguyen Hong Duong, Nguyen Thi Quy, Le Thi Thu Hong, Truong Nam Hai, Research on exploitation and expression of genes encoding β-glucosidase enzyme from metagenome data of bacterial flora around white-rot fungi (*Trametes versicolor*), National Biotechnology Conference, 2021, pp. 16-22.

3.Thi-Thu-Hong Le, **Thi-Binh Nguyen**, Hong-Duong Nguyen, Hai-Dang Nguyen, Ngoc-Giang Le, Trong-Khoa Dao, Thi-Quy Nguyen, Thi-Huyen Do, Nam-Hai Truong, De Novo metagenomic analysis of microbial community contributing in lignocellulose degradation in humus samples harvested from Cuc Phuong tropical forest in Vietnam, Diversity, 2022, 14(3), 220.

4. **Nguyen Thi Binh**, Nguyen Thi Quy, Do Thi Huyen, Le Thi Thu Hong, Truong Nam Hai, Selection of optimal culture conditions for expression of recombinant beta-glucosidase in *Escherichia Coli*, Journal of Biotechnology, 2022, 20(3): 425-433.

5. **Nguyen Thi Binh**, Le Thi Thu Hong, Truong Nam Hai, Using some bioinformatic tools to mining genes coding cellobiohydrolase from metagenome data of the bacteria surrounding white-rot fungi (*Trametes versicolor*) in Cuc Phuong National Park, Journal of Science, Hanoi Metropolitant University vol.62/2022: 119-126.

6. **Nguyen Thi Binh**, Nguyen Thi Quy, Le Thi Thu Hong, Truong Nam Hai, Purification and characterization of a recombinant beta-glucosidase in *Escherichia Coli*, Journal of Biotechnology, 2022, 20(4): 1-9.