BỘ GIÁO DỤC VÀ ĐÀO TẠO VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Thị Bình

NGHIÊN CỨU ĐA DẠNG KHU HỆ VI KHUẨN QUANH NẤM MỤC TRẮNG THỦY PHÂN LIGNOCELLULOSE VÀ KHAI THÁC GENE MÃ HÓA CELLULASE BẰNG KỸ THUẬT METAGENEOMICS

LUẬN ÁN TIẾN SĨ CÔNG NGHỆ SINH HỌC

Hà Nội - 2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Thị Bình

NGHIÊN CỨU ĐA DẠNG KHU HỆ VI KHUẨN QUANH NẤM MỤC TRẮNG THỦY PHÂN LIGNOCELLULOSE VÀ KHAI THÁC GENE MÃ HÓA CELLULASE BẰNG KỸ THUẬT METAGENEOMICS

LUẬN ÁN TIẾN SĨ CÔNG NGHỆ SINH HỌC

Mã số: 9.42.02.01

Xác nhận của Học viện Khoa học và Công nghệ

KT. GIÁM ĐỐC
PHÓ GIÁM ĐỐC

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Thị Trung

Thầy hướng dẫn 1

GS.TS. Trương Nam Hải

Thầy hướng dẫn 2

TS. Lê Thị Thu Hồng

Hà Nội - 2023

LỜI CAM ĐOAN

Tôi xin cam đoan:

Luận án là công trình nghiên cứu được thực hiện chủ yếu bởi cá nhân tôi và các cộng sự dưới sự hướng dẫn khoa học của GS.TS. Trương Nam Hải và TS. Lê Thị Thu Hồng tại Phòng Kỹ thuật di truyền, Viện Công nghệ Sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Các số liệu và kết quả trong luận án là hoàn toàn trung thực. Một phần lớn kết quả đã được công bố trên các tạp chí khoa học chuyên ngành với sự cho phép của đồng tác giả, một phần chưa được công bố.

Tôi xin hoàn toàn chịu trách nhiệm về lời cam đoan này!

Hà Nội, ngày 28 tháng 11 năm 2023

Nghiên cứu sinh

Nguyễn Thị Bình

LỜI CẢM ƠN

Lời đầu tiên, tôi xin được bày tỏ lòng biết ơn sâu sắc tới GS. TS. Trương Nam Hải và TS. Lê Thị Thu Hồng, Phòng Kỹ thuật di truyền, Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và công nghệ Việt Nam đã dành nhiều thời gian và tâm huyết để định hướng nghiên cứu, hướng dẫn, giúp đỡ và tạo mọi điều kiện cho tôi trong suốt quá trình thực hiện luận án này. Luận án được thực hiện bằng nguồn kinh phí của đề tài Nghị định thư Việt Nam – Cộng hòa Liên bang Đức giai đoạn 2018 – 2021 mã số NĐT.50.GER/18 do GS. TS. Trương Nam Hải làm chủ nhiệm.

Tôi xin chân thành cảm ơn các thầy cô giáo, các cán bộ đào tạo của Khoa Công nghệ sinh học, Ban Lãnh đạo Học viện Khoa học và công nghệ, Viện Hàn lâm Khoa học và công nghệ Việt Nam đã hướng dẫn, chỉ bảo cho tôi những kiến thức, kỹ năng cần thiết cũng như tạo mọi điều kiện thuận lợi cho tôi trong học tập và bảo vệ luận án.

Tôi xin bày tỏ lòng biết ơn sâu sắc đến các cán bộ phòng Kỹ thuật di truyền, Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam đã tận tình giúp đỡ, hướng dẫn nghiên cứu và tạo mọi điều kiện về cơ sở vật chất để tôi có thể hoàn thiện thực nghiệm nghiên cứu.

Tôi xin được cảm ơn các thầy cô bộ môn Công nghệ Sinh học, Ban chủ nhiệm Khoa Khoa học Tự nhiên và công nghệ, trường Đại học Thủ đô Hà Nội đã giúp đỡ, động viên và tạo điều kiện cho tôi trong thời gian học tập và nghiên cứu.

Cuối cùng, tôi xin được cảm ơn gia đình, bạn bè, đồng nghiệp đã động viên, khích lệ tôi trong suốt quá trình học tập và thực hiện luận án.

Tôi xin trân trọng cảm ơn!

Hà Nội, ngày 28 tháng 11 năm 2023

Nghiên cứu sinh

Nguyễn Thị Bình

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT TRONG LUẬN ÁN

Tên viết tắt	Tên tiếng Anh	Tên tiếng Việt
APS	Ammonium persulfate	Ammonium persulfate
ARDB	Antibiotic Resistance Genees Database	Cơ sở dữ liệu về gene kháng thuốc kháng sinh
BLAST	Basic Local Alignment Search Tools	Công cụ so sánh mức độ tương đồng về trình tự nucleotide/amino acid
BSA	Bovine Serum Albumin	Bovine Serum Albumin
bp	Base pair	Cặp base
CAZY	Carbohydrate-Active enzymes	Cơ sở dữ liệu về các enzyme tham gia chuyển hóa carbohydrate
CBD	Carbohydrate-binding domain	Vùng liên kết carbohydrate
CBH	Cellobiohydrolases	Cellobiohydrolases
CBM	Carbohydrate-binding vùng/cấu trúc	Vùng/cấu trúc liên kết carbohydrate
CD	Catalytic Domain	Vùng xúc tác
CMC	Carboxymethyl cellulose	Carboxymethyl cellulose
COG	Cluster of Orthologous groups	Cơ sở dữ liệu protein của sinh vật nhân sơ/nhân chuẩn đơn bào
CSDL		Cơ sở dữ liệu
EC	The Enzyme Commission	Hội đồng về enzyme
EDTA	Ethylene Diamine Tetracetic Acid	Ethylene Diamine Tetracetic Acid
eggNOG	Evolutionary genealogy of genees: Non-supervised Orthologous Groups	Cơ sở dữ liệu chứa các nhóm Orthologous
Expasy	Expert Protein Analysis	Hệ thống Phân tích protein

	System	chuyên sâu
Gb	Gigabyte	Gigabyte
GH	Glycoside hydrolase	Emzyme thủy phân liên kết glycosidic
GO	Genee Ontology	Bản thể gene học
His	Histidine	Amino acid histidine
HMM	Hidden Markov models	Mô hình đại diện Markov ẩn
HTS	High Throughput Sequencing	Giải trình tự thông lượng cao
IPTG	Isopropyl-β-D-thiogalactosidase	Isopropyl-β-D-thiogalactosidase
KEGG	Kyoto Encyclopedia of Genees and Geneomes	Cơ sở dữ liệu về hệ gene và hệ gene Kyoto
Km	The Michaelis constant	Hằng số Michaelis biểu thị nồng độ cơ chất cho phép enzyme đạt được một nửa Vmax.
KOG	Eukaryotic Orthologous groups	Cơ sở dữ liệu từ 7 hệ gene sinh vật nhân chuẩn (ba loài động vật, 1 loài thực vật, *Arabidosis thaliana,* 2 loài nấm và các ký sinh trùng nội bào)
LBA	Luria-Betani Ampicillin	Môi trường nuôi cấy LB có bổ sung ampicillin
LCA	Least Common Ancestor	Ít tổ tiên chung nhất
MCS	Multi-cloning site	Vùng đa nối
MEGAN	MEtaGeneomic Analyser	Phần mềm phân tích trình tự đa hệ gene
NCBI	National Center for Biotechnology Information	Trung tâm thông tin về Công nghệ Sinh học Quốc gia

NR	Non-redundant	Không dư thừa
NGS	Next Geneeration Sequencing	Giải trình tự gene thế hệ mới
OD	Optimal density	Mật độ quang học
ORF	Open Reading Frame	Khung đọc mở
PBS	Phosphate buffer	Đệm phosphate
PERISCOPE	Periplasmic expression classifier for soluble protein expression	Phần mềm ước đoán mức độ biểu hiện của protein dạng hòa tan trong khoang chu chất
PFAM	Protein Family	Cơ sở dữ liệu các họ protein
pNP	*p*-NitroPhenol	*p*-NitroPhenol
pNPG	*p*-NitroPhenol-β-Glucoside	*p*-NitroPhenol-β-Glucoside
PHYRE	Protein Homology/analogY Recognition Engine	Protein tương đồng/tương tự
SDS	Sodium dodecyl sulphate	Sodium dodecyl sulphate
SIB	Swiss Institute of Bioinformatics	Viện nghiên cứu Tin sinh học Thụy Sỹ
SVM	Support Vector Machine	Vector hỗ trợ phân tích tự động
SWISS-PROT	Swiss Protein	Dữ liệu các trình tự đã được xác định chức năng qua thực nghiệm
TBI	Taiwan Bioinformatic Institue	Viện nghiên cứu Tin sinh học Đài Loan
TEMED	Tetramethylethylenediamine	Tetramethylethylenediamine
Tm	Temperature melting	Nhiệt độ nóng chảy
Vmax	The maximum velocity	Tốc độ phản ứng tối đa đạt được khi enzyme bão hòa với cơ chất

DANH MỤC BẢNG TRONG LUẬN ÁN

DANH MỤC HÌNH VẼ, ĐỒ THỊ TRONG LUẬN ÁN

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Trong những năm gần đây, do nguồn nguyên liệu hóa thạch ngày càng cạn kiệt cùng với nhu cầu phát triển kinh tế bền vững, thân thiện với môi trường nên nhiều nguồn nguyên liệu sinh học đang được tìm kiếm. Trong đó lignocellulose là nguồn sinh khối tự nhiên có trữ lượng lớn, rẻ tiền và có khả năng tái tạo cao được cho là nguồn nguyên liệu sinh học có vai trò quan trọng trong nền kinh tế. Lignocellulose được sử dụng làm nguyên liệu thô để sản xuất các nhiên liệu sinh học và do đó được coi là nhiên liệu sinh học thứ hai. Nhiều công nghệ đã được phát triển để sản xuất nhiều sản phẩm khác nhau như rượu, axit hữu cơ từ sinh khối lignocellulose đồng thời các quy trình công nghệ sản xuất các chất có nguồn gốc từ sinh khối lignocellulose hiện có cũng ngày càng được mở rộng và phát triển nhằm nâng cao hiệu quả kinh tế thu được.

Sinh khối lignocellulose được chuyển hóa qua ba giai đoạn chính gồm: tiền xử lý bằng tác nhân khác nhau một cách hiệu quả; phân giải cellulose, hemicellulose bằng các enzyme cellulase, hemicellulase để tạo đường đơn C_5 và C_6; lên men đường đơn và các quá trình xử lý tiếp theo để tạo sản phẩm mong muốn như cồn sinh học, axit hữu cơ… Hiện nay, giá thành các sản phẩm sinh học sản xuất từ lignocellulose còn khá cao so với các sản phẩm sản xuất từ sinh khối hóa thạch. Một trong những nguyên nhân là do khó khăn trong quá trình phân giải cellulose và hemicellulose bằng các enzyme sinh học so với thủy phân bằng tác nhân lý, hóa. Để nâng cao hiệu quả của quá trình thủy phân cũng như giảm giá của các sản phẩm thu được và thúc đẩy phát triển kinh tế sinh học thì việc tìm ra enzyme có thể tham gia hiệu quả vào quá trình thủy phân cellulose, hemicellulose có vai trò quan trọng. Trong đó các enzyme cellulase có vai trò quan trọng trong việc nâng cao hiệu quả phân giải sinh khối lignocellulose do trong sinh khối này cellulose thường chiếm tỉ lệ lớn. Ở các hệ sinh thái có sự phân giải lignocellulose diễn ra mạnh mẽ như ruột mối, dạ cỏ trâu bò… sẽ là nguồn tiềm năng để tìm kiếm và khai thác các enzyme phân giải cellulose có giá trị cao trong công nghiệp. Đã có nhiều công trình khác nhau công bố về nghiên cứu sự đa dạng của vi sinh vật và khai thác các gene mã hóa enzyme thủy phân lignocellulose hiệu quả trong các hệ sinh thái này [1, 2].

Nấm mục trắng và đất xung quanh nấm mục trắng cũng là hệ sinh thái mà sự phân hủy lignocellulose diễn ra mạnh mẽ. Trong đó, nấm mục trắng có khả năng phân hủy tất cả các thành phần trong cấu tạo của gỗ và đặc biệt hiệu quả trong việc phân giải lignin không đặc hiệu. Bản chất không đặc hiệu của các hệ thống phân hủy lignin từ nấm mục trắng đã khiến các nhà nghiên cứu phát hiện ra việc sử dụng chúng trong phân hủy sinh học một số lượng lớn các chất gây ô nhiễm môi trường. Trong quá trình nấm phân giải gỗ đó đã xảy ra các phản ứng oxi hóa khử dẫn đến quá trình axit hóa nhanh và mạnh môi trường đất, quá trình chuyển hóa thứ cấp của nấm tạo ra chất độc trong đất. Vì vậy, các vi sinh vật tồn tại trong đất xung quanh khu nấm mục trắng phải có những đặc điểm đặc biệt về đa dạng loài và các enzyme tham gia vào quá trình chuyển hóa các chất. Các enzyme ở vi khuẩn có thể là các enzyme riêng rẽ hoặc các phức hợp enzyme giúp nấm mục trắng phân giải hiệu quả cellulose và hemicellulose [3]. Mặc dù có nhiều nghiên cứu về nấm mục trắng và vi khuẩn đất xung quanh khu nấm mục trắng nhưng cơ chế đằng sau sự tương tác này vẫn chưa được sáng tỏ, các đặc tính chức năng của chúng vẫn cần được xác định lại bằng thực nghiệm. Ở Việt Nam, cho đến nay vẫn chưa có nghiên cứu về đa dạng các loài vi khuẩn ở rừng Quốc Gia Cúc Phương nói chung và đa dạng loài các vi khuẩn đất xung quanh khu nấm mục trắng nói riêng cũng như khai thác các enzyme phân giải cellulose của vi khuẩn trong hệ sinh thái này.

Để khai thác các enzyme mong muốn từ các khu hệ vi sinh vật khác nhau như dạ cỏ dê, ruột mối, đất, nước thải…thì ngoài con đường truyền thống là phân lập từ ngân hàng gene, ngày nay kỹ thuật metageneomic đã được sử dụng rộng rãi. Đây là kỹ thuật hiện đại, có hiệu quả cao sử dụng kết quả giải trình tự gene thế hệ mới để có thể đánh giá đa dạng thành phần loài và tìm kiếm, khai thác các gene mới mã hóa enzyme đích từ vi sinh vật không thông qua nuôi cấy.

Nhằm phân tích và đánh giá mức độ đa dạng thành phần loài của vi sinh vật trong đất xung quanh khu nấm mục trắng nói chung và phân tích đa dạng loài vi sinh vật sinh cellulase nói riêng bằng kỹ thuật metageneomics, từ đó khai thác và lựa chọn được enzyme mã hóa cellulase có đặc tính mới không thông qua nuôi cấy, chúng tôi đã tiến hành nghiên cứu đề tài: "***Nghiên cứu đa dạng khu hệ vi khuẩn quanh nấm mục trắng thủy phân lignocellulose và khai thác gene mã hóa cellulase bằng kỹ***

thuật Metageneomics".

2. Mục tiêu của đề tài

Đánh giá được đa dạng của khu hệ vi sinh vật đất mùn xung quanh khu nấm mục trắng phân hủy lignocellulose và xác định được đa dạng enzyme tham gia vào quá trình phân giải lignocellulose, khai thác và lựa chọn được enzyme phân giải cellulose có tiềm năng ứng dụng trong thực tiễn sản xuất từ khu hệ vi khuẩn đất xung quanh khu nấm mục trắng ở rừng Quốc gia Cúc Phương bằng kỹ thuật Metageneomics.

3. Đối tượng nghiên cứu

- Các vi sinh vật trong đất mùn xung quanh khu nấm mục trắng có sự thủy phân lignocellulose trong rừng quốc gia Cúc Phương.

4. Nội dung nghiên cứu

- Phân tích và đánh giá mức độ đa dạng loài của khu hệ vi khuẩn trong đất xung quanh khu nấm mục trắng thủy phân lignocellulose bằng kỹ thuật Metageneomics;

- Phân tích và đánh giá mức độ đa dạng của các enzyme tham gia phân giải lignocellulose của khu hệ vi khuẩn đất xung quanh khu nấm mục trắng thủy phân lignocellulose bằng kỹ thuật Metageneomics;

- Tìm kiếm và lựa chọn các trình tự gene mới mã hóa cellulase có tiềm năng ứng dụng bằng các công cụ tin sinh học;

- Nghiên cứu biểu hiện tái tổ hợp của một gene đã lựa chọn, tinh chế và đánh giá tính chất của enzyme β-glucosidase.

5. Ý nghĩa khoa học và thực tiễn của đề tài

5.1. Ý nghĩa khoa học

- Đánh giá được sự đa dạng của vi khuẩn quanh khu nấm mục trắng, đặc biệt là sự đa dạng của các vi khuẩn sản sinh enzyme phân giải lignocellulose không thông qua nuôi cấy bằng phương pháp metageneomics.

- Cung cấp thêm các trình tự DNA mã hóa cellulase có khả năng phân hủy phế phụ phẩm nông nghiệp, công nghiệp chứa cellulose.

5.2. Ý nghĩa thực tiễn

Xác định được các enzyme mới tham gia phân giải nguyên liệu chứa cellulose từ vi khuẩn trong đất quanh khu nấm mục trắng. Các enzyme này có vai trò quan trọng trong sản xuất các nhiên liệu sinh học thế hệ thứ hai và phân giải sinh học các chất gây ô nhiễm môi trường

6. Đóng góp mới của đề tài

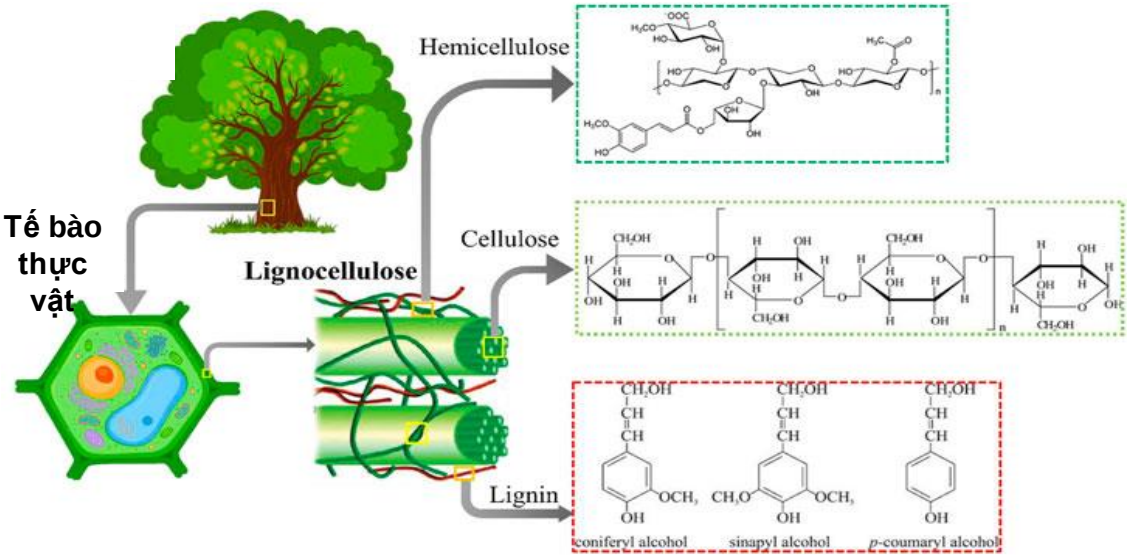
- Đây là nghiên cứu đầu tiên về đa dạng vi khuẩn xung quanh khu nấm mục trắng phân giải lignocellulose ở rừng Quốc gia Cúc Phương bằng kỹ thuật Metageneomics.

- Đã nghiên cứu được đa dạng enzyme tham gia phân giải lignocellulose ở khu hệ vi khuẩn xung quanh nấm mục trắng ở rừng Quốc gia Cúc Phương, lựa chọn và đánh giá được tính chất của enzyme β-glucosidase GH3S2 từ DNA đa hệ gene của vi khuẩn đất mùn xung quanh nấm mục trắng.

CHƯƠNG 1. TỔNG QUAN TÀI LIỆU

1.1. Khái quát chung về lignocellulose

Lignocellulose là tên gọi chung cho sinh khối thực vật được cấu tạo từ ba thành phần chính là cellulose, hemicellulose và lignin (Hình 1.1). Cellulose và hemicelluloses liên kết chặt chẽ với lignin. Cellulose là một polymer được cấu tạo từ các monomer là β-D-glucopyranose, đây là thành phần chính trong cấu trúc của thành tế bào thực vật thường chiếm tỷ lệ 38 – 50% [4]. Tiếp đến là hemicellulose chiếm tỷ 17 – 32% có cấu trúc không đồng nhất, có sự phân nhánh cao thường được cấu tạo từ các đường đơn pentose và hexose [5]. Các hemicellulose tạo ra các liên kết chéo giữa các cellulose. Lignin chiếm 15 – 30% bao gồm các polyphenol thơm, được sinh tổng hợp và tạo thành cấu trúc bao bọc xung quanh hai thành phần cellulose và hemicelluloses, cung cấp thêm độ bền cơ học cho thành tế bào, chống lại côn trùng hoặc điều kiện ẩm ướt.



Hình 1.1. Cấu tạo của lignocellulose [4]

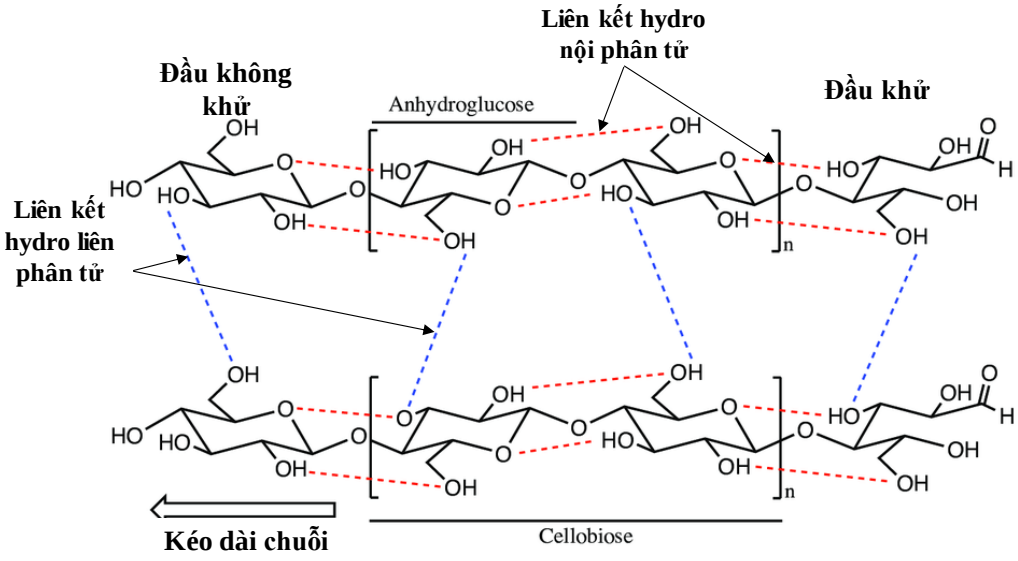
Nói chung, thành phần của lignocellulose phụ thuộc vào nguồn gốc của chúng như gỗ cứng hay gỗ mềm, cỏ hay cây công nghiệp, sản phẩm thải trong nông nghiệp hay sản xuất công nghiệp (Bảng 1.1) [5]. Ở một số nguyên liệu, cellulose thường chiếm tỷ lệ khá cao như sợi cotton 90%, cây gai dầu khô 57%, gỗ mềm 45- 50 %, cao lương ngọt 45%, bã mía 42% [6].

Bảng 1.1. Tỉ lệ thành phần của lignocellulose trong các nguyên liệu khác nhau [5]

Vật liệu	Cellulose (%)	Hemicellulose (%)	Lignin (%)
Bã mía	42	25	20
Cao lương ngọt	45	27	21
Cây phong	40-45	24-40	18-25
Gỗ mềm	45-50	25-35	25-35
Bắp ngô	45	35	15
Thân ngô	38	26	19
Rơm rạ	32,1	24	18
Vỏ quả hạch	25-30	25-30	30-40
Báo	40-55	25-40	18-30
Cỏ	25-40	25-40	10-30
Lúa mì	29-35	26-32	16-21
Chất thải chuối	13,2	14,8	14
Bã mía	54,87	16,52	23,33

1.1.1. Cellulose

Cellulose có công thức phân tử $(C_6H_{10}O_5)_n$ là polysaccharide mạch thẳng được cấu tạo từ các monosaccharide là β-D-glucopyranose. Các phân tử đường đơn này liên kết với nhau bởi liên kết β-(1-4) glucosidic [4], vì vậy cellulose có cấu trúc bền vững, khó bị thủy phân. Thông thường, trung bình mỗi vi sợi cellulose có khoảng 5000-7000 đơn phân glucose [7], số lượng đơn phân này thay đổi phụ thuộc vào nguồn gốc của cellulose như: bông 1000 – 3000 đơn phân, bột gỗ 500 – 1500 đơn phân [8]…

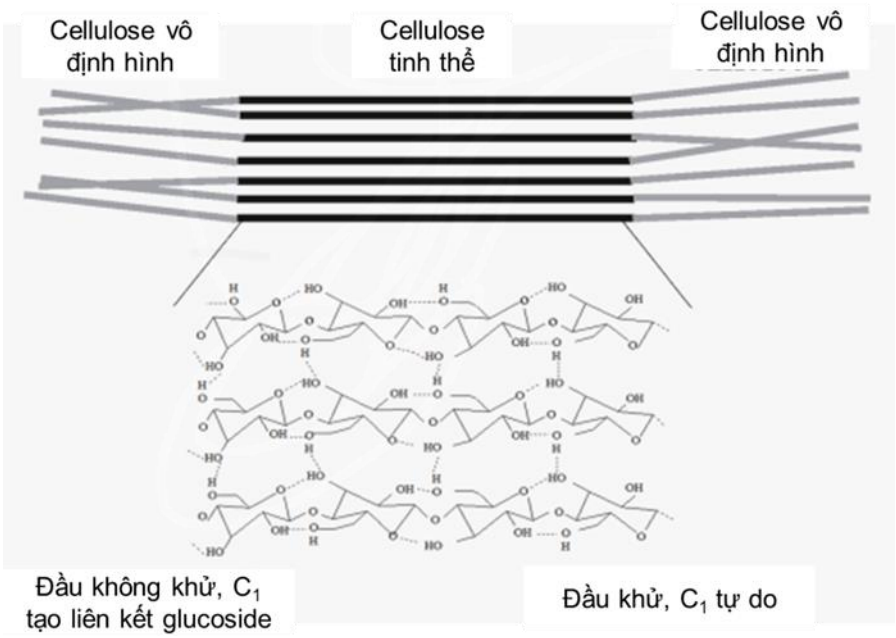


Hình 1.2. Cấu trúc của cellulose [8]

Các đơn phân D-glucose của cellulose có năm nhóm hydroxyl (OH) trong đó

có ba nhóm ở vị trí C2, C3, C6 tham gia hình thành liên kết hydro là liên kết đóng vai trò quan trọng trong cấu trúc của cellulose [6]. Các nhóm -OH này tham gia tạo ra các liên kết hydro nội phân tử, liên kết hydro liên phân tử và liên kết giữa các phân tử cellulose kề nhau làm cho cấu trúc chuỗi sợi của cellulose rất bền vững. Cellulose có cấu trúc 2 đầu, một đầu không khử có cấu trúc vòng khép kín, đầu còn lại có tính khử chứa nhóm carbonyl tự do (Hình 1.2). Vì vậy, cellulose là phân tử phân cực, trong đó các đơn phân glucose mới được gắn thêm vào đầu không khử để kéo dài chuỗi [9].

Nishikawa và Ono (1913) đã phát hiện ra các vi sợi cellulose đơn lẻ thường sắp xếp theo các trật tự khác nhau để hình thành trạng thái kết tinh của cellulose. Những vùng nào mà các vi sợi cellulose sắp xếp có trật tự cao, hình thành một số lượng lớn các liên kết hydro trong vi sợi và giữa các vi sợi, lực Van der Waals lớn (gọi là vùng tinh thể) thì cấu trúc cellulose rất bền vững. Còn những vùng mà các chuỗi cellulose sắp xếp không theo trật tự chặt chẽ, liên kết với nhau lỏng lẻo (gọi là vùng vô định hình) thì cấu trúc cellulose kém bền vững, dễ bị thủy phân. Vì vậy, chỉ số kết tinh (crystallinity index - CI) là một trong những chỉ số quan trọng nhất ảnh hưởng đến khả năng bị thủy phân của cellulose. Nhìn chung, trong tự nhiên, các chỉ số kết tinh dao động từ 40% đến 95%, phần còn lại là cellulose vô định hình [10].



Hình 1.3. Cấu trúc tinh thể và cấu trúc vô định hình của cellulose [9]

Thông thường, có khoảng 24 - 36 các vi sợi cellulose xếp xen kẽ với nhau theo hướng hình thành ít các liên kết hydro nội phân tử, tăng số lượng liên kết hydro liên phân tử, liên kết hydro hướng ra ngoài tạo nên một tổng thể cấu trúc cellulose rất vững chắc, ít bị hòa tan trong nước [11] (Hình 1.3). Nghiên cứu công hưởng từ hạt nhân (NMR) của cellulose cho thấy nhiều dữ liệu về các dạng cấu trúc tinh thể của cellulose đã được phân tích. Do các đơn phân D-glucose của cellulose có thể tạo ra nhiều loại liên kết hydro khác nhau, thêm vào đó là sự sắp xếp khác nhau của các vòng pyranose và sự chuyển đổi vị trí của các nhóm – OH so với mặt phẳng vòng carbon nên cellulose có thể tồn tại nhiều dạng cấu trúc tinh thể khác nhau [12]. Có 4 dạng cấu trúc tinh thể khác nhau của cellulose đã được xác định là celluloses I, II, III, IV trong đó dạng cấu trúc quan trọng nhất là cellulose I và II.

Ở thực vật và nhiều loài vi khuẩn Gram dương, Gram âm được báo cáo có khả năng tổng hợp cellulose như *Clostridium thermocellum, Streptomyces* sp., *Ruminococcus* sp., *Pseudomonas* sp., *Cellulomonas* sp., *Bacillus* sp., *Serratia, Proteus, Staphylococcus* sp., *Bacillus subtilis* [13]. Người và động vật không tổng hợp được cellulase nên không tiêu hóa được cellulose. Nhưng động vật nhai lại do có hệ vi khuẩn đường ruột nên có khả năng tổng hợp cellulase tiêu hóa cellulose trong cỏ thành chất dinh dưỡng cho cơ thể. Tuy cellulose không có giá trị dinh dưỡng với người và động vật nhưng có tác dụng hỗ trợ quá trình tiêu hóa, điều hòa hàm lượng đường trong máu, giảm mỡ máu, giảm cân và giảm ung thư đại tràng, điều hòa hệ vi sinh đường ruột, thải các sản phẩm thải ra khỏi cơ thể.

1.1.2. Hemicellulose

Hemicellulose là polysaccharide dị hợp, bao gồm các chuỗi phân tử đường có độ phân nhánh cao và gồm nhiều loại như glucuronoxylan, glucomannan và một số polysaccharite khác. Mức độ trùng hợp của hemicellulose vào khoảng dưới 200, có mạch bên có thể bị acetyl hóa. Hemicellulose được cấu tạo chủ yếu bởi các phân tử đường D-glucose, D-galactose, D-mannose, D-xylose, L-arabinose, axit D-glucuronic và axit 4-O-methyl-D-glucuroni [14]. Ở cỏ và vỏ trấu, hemicellulose chủ yếu là arabinan, galactan và xylan trong khi đó hemicellulose ở gỗ cứng và gỗ mềm chủ yếu là mannan [15]. Chúng được xếp vào nhóm hemicellulose có đường tham

gia cấu trúc mạch chính như xylan, mannan và glucan với xylan và mannan phổ biến nhất. Galactan, arabinan và arabinogalactan cũng thuộc nhóm hemicellulose nhưng không chứa liên kết β-1,4 trong cấu trúc. Trong gỗ cứng chứa chủ yếu hemicellulose loại glucuronoxylan (O-acetyl-4-O-methyl-glucurono-β-D-xylan) trong đó mạch chính được tạo bởi xylospyranose. Trong ligncellulose, hemicellulose chiếm khoảng 25 – 35% và trọng lượng phân tử trung bình là nhỏ hơn 30.000 đvC. Cellulose và hemicellulose liên kết chặt với nhau trên bề mặt của vi sợi cellulose. Hemicellulose ban đầu được cho là chất trung gian trong quá trình sinh tổng hợp cellulose.

1.1.3. Lignin

Lignin là thành phần không phải carbohydrate, chiếm tỉ lệ nhỏ nhất trong sinh khối lignocellulose (10-25%). Lignin góp phần tạo ra độ cứng chắc và tính kỵ nước cho thành tế bào thực vật đồng thời bảo vệ các polysaccharide khỏi sự phân hủy của các vi sinh vật. Lignin là polyme sinh học có chứa vòng thơm, trọng lượng phân tử cao và được tạo thành từ các tiểu đơn vị là phenylpropan (lignin syringyl (S), guaiacyl (G) và hydroxyphenyl (H)), nhóm methoxyl và các hợp chất poly phenol gắn các thành phần của thành tế bào với nhau [16]. Các phenylpropan này được ký hiệu là có 0, 1, 2 nhóm methoxyl gắn vào các vòng tạo ra cấu trúc đặc biệt I, II và III. Các cấu trúc này phụ thuộc vào nguồn thực vật mà chúng thu được. Cấu trúc I tồn tại ở cỏ, cấu trúc II có trong gỗ (cây lá kim) trong khi cấu trúc III tồn tại trong gỗ rụng lá. Lignin hoạt động giống như một chất keo, lấp đầy khoảng trống xung quanh phức hợp cellulose và hemicellulose. Lignin ngăn cản sự tiếp cận của cellulase với cellulose và làm giảm đáng kể hiệu quả của các enzyme trong quá trình chuyển hóa sinh khối lignocellulose.

1.2. Cellulase

1.2.1. Khái quát chung về cellulase

Enzyme cellulase cùng với hai nhóm enzyme là hemicellulase và enzyme tiền xử lý tham gia vào quá trình phân hủy sinh khối lignocellulose. Cellulase thuộc nhóm enzyme glycoside hydrolase (GH) (EC 3.2.1.-) có vai trò thủy phân liên kết β-1,4-glycoside trong phân tử cellulose tạo thành các sản phẩm cello-oligosaccharide, cellobiose và glucose [10] hoặc phân cắt các liên kết glycosidic giữa hai hay nhiều carbohydrate hoặc giữa một carbohydrate và một gốc không phải carbohydrate. Hiện

nay, GH đã được phân thành 130 họ khác nhau. Cellulase được xếp vào các họ GH khác nhau trên cơ sở tương đồng về trình tự amino acid. Cellulase có vùng xúc tác (catalytic domain CD) phân cắt liên kết glycosidic, vùng gắn cơ chất (carbohydrate-binding module CBM) và ở một số loại cellulase có vùng/cấu trúc phụ trợ giống FN3 [17], [18].

Cellulase có nguồn gốc từ nhiều sinh vật khác nhau vi khuẩn, nấm, thực vật, động vật [19]. Trong đó cellulase ở nấm và vi khuẩn có sự đa dạng lớn của các enzyme phân giải thành tế bào do sự phong phú về nguồn gene, khác biệt của các mRNA trưởng thành và các quá trình sửa đổi sau dịch mã [20]. Vì vậy, nấm và vi khuẩn trở thành đối tượng chủ yếu để nghiên cứu cellulase quy mô công nghiệp. Một số loại nấm và vi khuẩn phân giải cellulose được trình bày ở bảng 1.2.

Bảng 1.2. Một số loại nấm và vi khuẩn tham gia phân giải cellulose và nguồn gốc của chúng [19]

Vi khuẩn	**Nguồn gốc**	**Nấm**	**Nguồn gốc**
Cellulomonas fimi	Đất	*Geotrichum candidum*	Đất, phân trộn
Cellvibrio japonicas	Đất	*Penicillium chrysogeneum*	Đất, gỗ mục
Cytophaga hutchinsonii	Đất, Phân trộn	*Phanerochaete chrysosporium*	Phân trộn
Paenibacillus polymyxa	Phân trộn	*Rhizopus oryzae*	Đất, chất hữu cơ chết
Pseudomonas fluorescens	Đất, bùn	*Trichocladium canadense*	Đất
Pseudomonas fluorescens	Đất	*Trichoderma reesei*	Đất, vải mục nát
Bacillus brevis	Ruột mối	*Trichoderma longibrachiatum*	Đất
B. thuringiensis	Ruột sâu bướm	*Chaetomium thermophilum*	Đất
Bacillus cereus	Đất, dạ cỏ	*Corynascus*	Phân trộn
B. subtilis		*thermophilus*	nấm

Hiện nay, nấm *Chytridiomycetes* và *Basidiomycetes* là các nhóm sinh vật sinh cellulase được nghiên cứu nhiều nhất nhờ khả năng tiết một lượng lớn các enzyme phân giải cellulose có hoạt tính cao [21], [22]. Trong đó nấm *Basidiomycetes* có khả năng phân hủy gỗ một cách hiệu quả nhất. Những loại nấm hiếu khí này tiết ra các enzyme ngoại bào phân hủy lignocellulose. Không giống như nấm hiếu khí, một số

nấm kỵ khí *Chytridiomycetes* có phức hợp đa enzyme tương tự như cellulosome của vi khuẩn [23], [24], một số loài kỵ khí sống trong ống tiêu hóa của động vật nhai lại như *Anaeromyces, Caecomyces, Neocallimastix, Orpinomyces* và *Piromyces*. Trong gỗ mục nát và đất rừng, thành phần của nấm tiết enzyme phân giải cellulose có *Zygomycetes* đại diện là *Mucor, Ascomycetes* và *Basidiomycetes* được đại diện bởi các chi như *Trichoderma, Aspergillus, Penicillium*… Hai trong số các loại nấm được nghiên cứu nhiều nhất vì tính liên quan đến công nghiệp của chúng là *Trichoderma reesei* và *Phanerochaete chrysosporium* [25]. Hỗn hợp cellulase của *T. reesei* bao gồm các exoglucanase (80%), endoglucanases (15%) [18] và một lượng β-glucosidase nhỏ vì vậy cần phải bổ sung từ các nguồn khác như Aspergilli [26]. Các enzyme từ Aspergilli hầu hết có hoạt tính tổng số cellulase thấp [27], tuy nhiên β-glucosidase của chúng có hoạt tính cao. Chi Aspergillus là một trong những nhóm sinh vật sản xuất cellulase đa dạng tạo ra tác động nổi bật trong quá trình xử lý sinh học [28]–[33]. Ngày nay, hơn 14.000 loại nấm có khả năng phân giải cellulose và các hợp chất phức tạp đã được biết đến [34].

Việc phát hiện ra các đặc tính phân giải cellulose đặc biệt của vi khuẩn từ các chi *Clostridium* và *Thermotoga* đã góp phần vào việc chuyển dần nguồn enzyme phân giải cellulose từ các nguồn nấm sang các nguồn vi khuẩn. Đặc điểm của cellulase từ những loài này là chịu nhiệt và có khả năng hoạt động tốt ở các điều kiện nhiệt độ cao từ 60 - 125°C, vì vậy chúng là những ứng cử viên quan trọng để cải thiện kinh tế-công nghệ của quá trình đường hóa sinh khối. Vi khuẩn từ các chi *Clostridium* và *Thermotoga* cũng tạo ra hệ thống enzyme gọi là cellulosome để thủy phân hiệu quả cấu trúc phức tạp của cellulose [35]. Một số nhóm vi khuẩn tiết cellulase được biết đến là *Bacillus, Cellulomonas, Streptomyces, Cytophaga, Cellvibrio* và *Pseudomonas*. Vi khuẩn kỵ khí và hiếu khí có các cách khác nhau để phân hủy cellulose. Trong khi các vi khuẩn kị khí sử dụng cellulosome để phân hủy cellulose thì vi khuẩn hiếu khí tiết các enzyme riêng biệt hoạt động hiệp đồng để phân hủy cơ chất. Các vi khuẩn kỵ khí bộ *Clostridiales* (ngành Firmicutes) phân giải cellulose thường được phát hiện trong đất mùn, dạ cỏ của trâu bò, dê, nước thải, côn trùng. Vi khuẩn hiếu khí bộ *Actinomycetales* (ngành Actinobacteria) đã được tìm thấy trên đất, nước, mùn, phế thải nông nghiệp và lá cây mục nát tiết cellulase [36]. Do các vi khuẩn

sinh enzyme phân giải cellulose có sự đa dạng cao nên có thể xếp vi khuẩn thành ba nhóm: (1) vi khuẩn lên men kị khí điển hình là Gram dương (*Clostridium* và *Ruminococcus*) nhưng với một số loài Gram âm (*Butyvibrio* và *Acetivibrio*) có liên quan về mặt phát sinh loài với *Clostridium* (Fibrobacter); (2) vi khuẩn Gram dương hiếu khí (*Cellulomonas* và *Thermobifida*) và (3) vi khuẩn sợi hiếu khí (*Cytophaga* và *Sporocytophaga*) [37].

1.2.2. Phân loại cellulase

Trong tự nhiên, quá trình thủy phân cellulose được thực hiện nhờ sự hoạt động phối hợp của ít nhất ba loại cellulase chính là β-1,4-endoglucanase (EC3.2.1.4), exoglucanase hoặc cellobiohydrolase (EC 3.2.1.91) và β–glucosidase (EC 3.2.1.21). Ba loại enzyme này khác nhau về cấu trúc và cơ chế hoạt động, trong đó endoglucanase thủy phân các liên kết β-1,4-glucoside bên trong chuỗi cellulose để tạo ra các đầu chuỗi mới; exoglucanase thủy phân các liên kết glucoside ở hai đầu của chuỗi để giải phóng các phân tử cellobiose hoặc glucose hòa tan; β-glucosidase thủy phân các cellobiose thành glucose. Hoạt động phối hợp của ba enzyme này trên cellulose tinh thể có mức độ hoạt động và hiệu quả thủy phân cao hơn nhiều so với tổng hoạt động của các enzyme đơn lẻ cho thấy đây là một phức hệ enzyme thủy phân cellulose rất hiệu quả [38].

1.2.2.1. Endoglucanase

Endoglucanase là nhóm enzyme đầu tiên tham gia thủy phân cellulose. Các enzyme này phân cắt từ bên trong trong các sợi cellulose tại các vùng vô định hình, tạo ra các oligosaccharid với các kích thước khác nhau và tạo ra các đầu chuỗi mới có thể bị tấn công bởi các exoglucanase. Hoạt tính cao nhất của enzyme này thường xảy ra đối với các dạng cellulose hòa tan hoặc cellulose vô định hình được xử lý bằng axit. Endoglucanase ở các loại nấm *Sclerotium rolfsii* và *Gloeophyllum sepiarium* có trọng lượng 44 – 90 kD. Nói chung, endoglucanase không bị glycosyl hóa, pH tối ưu 4 – 5 (endoglucanase duy nhất được biết đến với độ pH trung tính là từ Basidiomycete (*Volvariella volvacea*), nhiệt độ tối ưu trong khoảng từ 50 đến 70°C [39]. Các endoglucanase khác nhau có các vùng/cấu trúc xúc tác thuộc các họ GH5-9, 12, 44, 45, 48, 51 và 74. Các endoglucanase của nấm thường có một vùng/cấu trúc xúc tác, có thể có hoặc không CBM, trong khi các endoglucanase của vi khuẩn có thể có nhiều

vùng/cấu trúc xúc tác, CBM và các vùng/cấu trúc khác chưa xác định chức năng [40]. Các vùng/cấu trúc xúc tác của hầu hết các endoglucanase có một vị trí hoạt động hình khe/rãnh cho phép endoglucanase liên kết và phân cắt cellulose để tạo glucose, các cellodextrin tan hoặc các đoạn cellulose không hòa tan. Tuy nhiên, một số endoglucanase có thể thủy phân các cellulose tinh thể và tạo ra các sản phẩm chính là cellobiose hoặc các cellodextrin dài hơn [41].

1.2.2.2. Exoglucanase

Enzyme exoglucanase (cellobiohydrolases) xúc tác quá trình thủy phân từ hai đầu của vi sợi cellulose tạo ra sản phẩm chính là các phân tử cellobiose, được thủy phân bởi các β-glucosidase. Chúng chiếm từ 40 đến 70% trong hệ thống cellulase và có khả năng phân cắt cellulose ở các vùng tinh thể [31]. Các exoglucanase phân cắt đặc hiệu trên các đầu của cellulose, chẳng hạn như ở *T. reesei* cellobiohydrolase (CBH) I và II lần lượt tác động lên đầu chuỗi cellulose có tính khử và không khử. Các enzyme này có kích thước nhỏ hơn endoglucanase, mức độ glycosyl hóa thấp (khoảng 0 - 12%), pH tối ưu của chúng là 4 đến 5, với nhiệt độ tối ưu từ 37 đến 60°C, tùy thuộc vào sự kết hợp enzyme-cơ chất cụ thể. Exoglucanase có mặt trong cellulase của nấm mục trắng, một số loài nấm mục nâu Basidiomycetes như *Fomitopsis palustris* [42]. Cellulose tinh thể (Avicel) là cơ chất tốt cho exoglucosidase, tuy nhiên một số endoglucanase có thể giải phóng đáng kể đường khử từ Avicel. Các exoglucanase khác nhau của vi khuẩn và nấm có các vùng/cấu trúc xúc tác thuộc các họ GH5, 6, 7, 9, 48 và 74. Exoglucanase của nấm hiếu khí chỉ có ở họ GH6 và 7, của nấm kỵ khí thuộc họ GH48; exoglucanase của vi khuẩn hiếu khí có trong họ GH6 và 48, của vi khuẩn kị khí thuộc họ GH9 và 48. Đặc điểm cấu trúc quan trọng nhất trong vùng/cấu trúc xúc tác của các enzyme exoglucanase là cấu trúc đường hầm được hình thành bởi hai vòng bề mặt. Đường hầm có thể bao phủ toàn bộ (ví dụ: họ GH7) hoặc một phần của vị trí đang hoạt động (ví dụ: họ GH48). Vị trí hoạt động dạng đường hầm của exoglucanase cho phép enzyme thủy phân cellulose theo cách độc đáo [43]. Họ GH48 exoglucanase được cho là đóng vai trò quan trọng trong quá trình thủy phân cellulose tinh thể của hệ thống cellulase của vi khuẩn. Vai trò của chúng được cho là tương tự như vai trò của *Trichoderma* CBHI (Cel7A).

1.2.2.3. β-Glucosidase

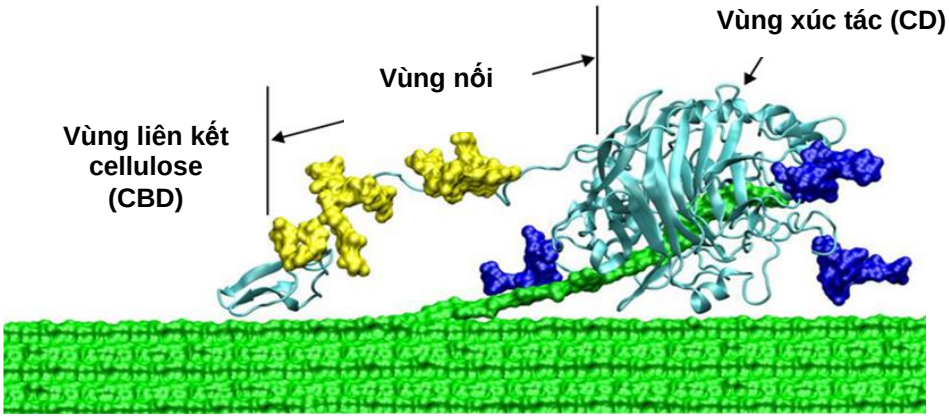
β-D-glucosidases thủy phân cellobiose hòa tan và các cellodextrin khác để tạo ra glucose nhằm loại bỏ sự ức chế cellobiose [8]. β-glucosidase có một vị trí hoạt động hình túi, cho phép chúng liên kết rồi tách glucose ra khỏi cellobiose hoặc cellodextrin. β-Glucosidase sử dụng cơ chế giữ nguyên cấu hình vòng hoặc cơ chế nghịch đảo cấu hình vòng của glucose sau khi thủy phân. β-glucosidase giữ nguyên cấu hình vòng phân cắt liên kết β-glucoside tạo thành glucose dạng β trong khi β-glucosidase nghịch đảo cấu hình vòng thì glucose tạo thành có cấu hình α. Các enzyme này có kích thước khoảng 35 - 640 kDa và chúng có thể là đơn phân hoặc tồn tại dưới dạng đồng phân lập thể. Hầu hết các β-glucosidase đều được glycosyl hóa, một số trường hợp như β-glucosidase 300 kDa từ *Trametes versicolor* glycosyl hóa có thể cao hơn 90%. Độ pH tối ưu của chúng nằm trong khoảng từ 3,5 đến 5,5 và nhiệt độ tối ưu của chúng nằm trong khoảng từ 45 đến 75°C.

β-Glucosidase là một loại enzyme có nguồn gốc từ các loài: vi khuẩn, nấm, thực vật và động vật. Trong đó, nấm được xem là nguồn sản xuất β-glucosidase chính như nấm sợi *Acremonium persicinum*, *Thermomyces lanugin*osus–SSBP, *Aspergillus niger* [44]. β-glucosidase cũng được tìm thấy ở vi khuẩn vì khả năng xúc tác mạnh mẽ và nhiều đặc tính giá trị của cellulase vi khuẩn [45]. β-glucosidase từ nhiều loài vi khuẩn cũng đã được tinh chế và xác định tính chất như *Flavobacterium johnsoniae*, *Lactobacillus brevis* [46], *Caldicellulosiruptor saccharolyticus* [47]. Các β-glucosidase được sản sinh ra dưới dạng các enzyme nội bào, ngoại bào hoặc liên kết bề mặt tế bào [32]. Trong khi phần lớn các β-glucosidase của nấm được tổng hợp ngoại bào và thuộc GH3 [48] thì hầu hết các β-glucosidase của vi khuẩn là nội bào và thuộc GH1 ví dụ như ở khuẩn *Baciulus circulans* subsp. *Alkalophilus*. β-Glucosidase phân cắt các liên kết β-D-glucoside từ nhiều hợp chất khác nhau giải phóng sản phẩm cuối cùng là glucose. Do có sự khác biệt rất nhiều về tính đặc hiệu cơ chất, đặc biệt là đối với gốc aglycone khiến việc phân loại β-glucosidase là một thách thức. Có hai cách phân loại: 1) phân loại theo cơ chất và 2) phân loại dựa trên nhận dạng trình tự nucleotide và phân tích nhóm kỵ nước. Tùy thuộc các cơ chất bị phân giải, β-glucosidase được phân loại thành ba nhóm: 1) aryl-β-glucosidases chỉ thủy phân liên kết aryl-β-glucoside, 2) cellobiases chỉ thủy phân cellobiose và 3) đặc hiệu cơ chất rộng, β-glucosidase thủy phân phạm vi rộng của cơ chất có các liên kết

khác nhau như liên kết β 1-4, β 1-3 glucoside (thường thấy ở các β-glucosidase có nguồn gốc từ vi sinh vật).

1.2.3. Cấu trúc và cơ chế xúc tác của cellulase

Cấu trúc của cellulase thường gồm ba vùng là: vùng có vai trò xúc tác (Catalytic Domain - CD), một hoặc một số vùng có vai trò gắn kết với carbohydrate (Carbohydrate-Binding Domain - CBD hay còn gọi là CBM và đoạn trình tự peptide nối giữa hai vùng CD và và vùng CBD [49] (Hình 1.4).

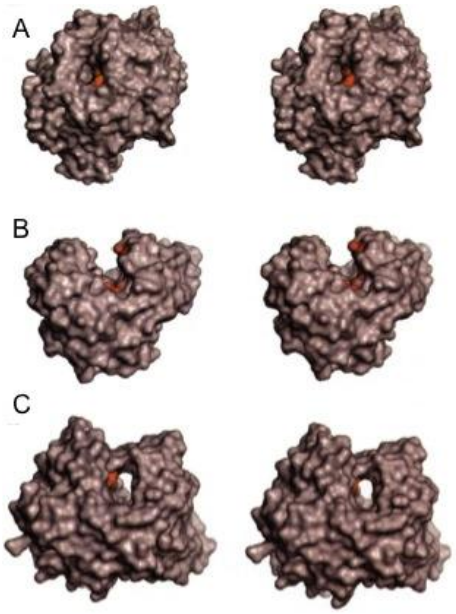


Hình 1.4. Mô hình cấu trúc chung của cellulase [49]

Vùng CD chiếm trên 70% trình tự protein. Phân tích trình tự vùng này ở các cellulase khác nhau cho thấy chúng rất đa dạng và vị trí xúc tác của enzyme có ba cách sắp xếp cấu trúc không gian: (1) Dạng túi (thủy phân các polymer hoặc dimer vô định hình như cellulose hoặc cellobiose); (2) Dạng khe hở (đối với endoglucanase thủy phân polymer tinh thể); (3) Dạng khe ngầm (đối với exoglucanase thủy phân polymer tinh thể) (Hình 1.5). Vùng CD được glycosyl hóa đầu N và thực hiện chức năng phân cắt liên kết β-glucoside thông qua cơ chế thủy phân axit sử dụng chất cho proton và nucleophyle/base như axit glutamic hoặc axit aspartic [23].

Vùng CBD tham gia vào quá trình thủy phân bằng cách giữ vùng CD gần cơ chất, do đó sự có mặt của vùng CBD rất quan trọng trong hoạt động của cellulase. Vùng CBD thường được O-glucosyl hóa để tránh sự phân cắt của protease, chứa từ 30 đến khoảng 200 amino acid và thường tồn tại thành 1, 2, hoặc 3 vùng trong protein. Vị trí của chúng trong protein có thể là cả hai, đầu C hoặc N và đôi khi nằm ở vị trí trung tâm. Các vùng CBD của các cellulase khác nhau có trình tự khác nhau đáng kể. Các CBD đưa enzyme tiến vào gần hơn với cơ chất, gắn kết với cơ chất làm tăng tốc

độ xúc tác của enzyme với cơ chất. Việc loại bỏ CBM khỏi enzyme hoặc khỏi protein khung trong cellulosome làm giảm đáng kể tính enzyme của nó. Sự có mặt của CBD góp phần cải thiện khả năng liên kết và hoạt động của cellulase trên các chất nền không hòa tan nhưng không ảnh hưởng đến hoạt động của chúng trên các chất nền hòa tan [50].



Hình 1.5. Cấu trúc không gian vùng xúc tác của cellulase (A): Dạng túi; (B): Dạng khe hở; (C): Dạng khe ngầm [23]

Đoạn peptide nối là một đoạn trình tự chứa từ 6 – 59 amino acid nối giữa hai vùng CD và vùng CBD. Đoạn peptide này rất linh hoạt cho phép các vùng trong cấu trúc của enzyme có thể hoạt động độc lập. Các enzyme khác nhau thì đoạn peptide nối này khác nhau nhưng chúng đều giàu proline, treonine và serine như trình tự PTPTPTPTT(PT)7 của enzyme endoglucanase ở *C. fimi* và trình tự NPSGGNPPGGNPPGTTTTRRPATTTGSSPG của cellobiohydrolase I ở *T. reesei.* Treonine và serine còn lại của đoạn peptide nối được O-glycosyl hóa cao để được bảo vệ khỏi sự phân giải của protease. Nếu đoạn peptide nối quá ngắn hoặc không tồn tại thì hoạt động của cả hai miền CBD và CD bị ảnh hưởng và giảm ái lực. Một số ví dụ về vùng/cấu trúc của cellulase ở các vi khuẩn khác nhau được thể hiện ở bảng 1.3.

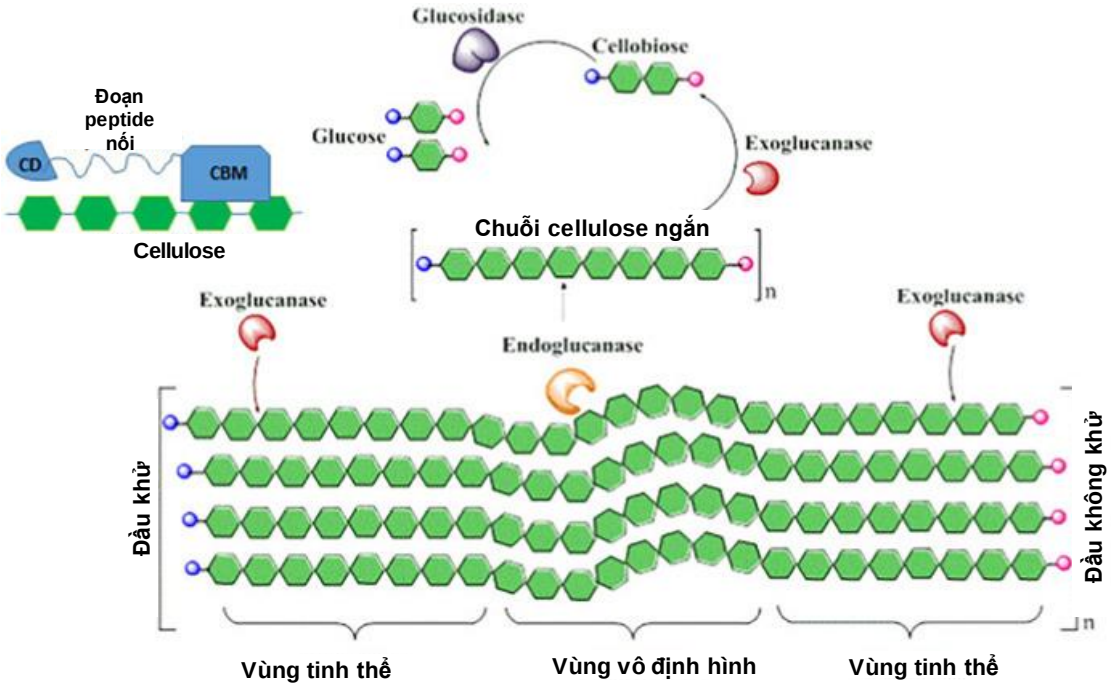
Do cellulose có cấu trúc chặt chẽ nên để phân hủy được cellulose, trước hết bề mặt của cellulose cần phải được nới lỏng để giúp các enzyme có thể xâm nhập và tiếp

xúc trực tiếp với các sợi cellulose ở bên trong [48]. Một khi các enzyme cellulase có thể xâm nhập vào được mạng lưới cellulose thì chúng tiến hành thủy phân cellulose từng bước để giải phóng glucose.

Bảng 1.3. Vùng/cấu trúc của cellulase ở một số loại vi khuẩn khác nhau [59]

Vi khuẩn	Cấu trúc vùng/cấu trúc	Genee Bank code
Anaerocellum thermophilum	GH9-(CBM3) 3 -GH48	ACM60955
Bacillus subtilis	GH5-CBM3 CAA82317	CAA82317
Clostridium phytofermentans	GH9-CBM3-(Ig)2-CBM3	ABX43720
Clostridium thermocellum	GH48-(Doc) 2	AAA23226
Clostridium thermocellum	GH26-GH5-CBM11-(Doc) 2	AAA23225
Cellulomonas fimi	GH48-Fn3-CBM2	AAB00822
Thermobifida fusca	CBM2-Fn3-GH48	AAD39947

Để thủy phân hoàn toàn cellulose, cần có sự hoạt động kết hợp của ít nhất 3 loại enzyme là endoglucanase, exoglucanase và β-D glucosidase trong đó mỗi enzyme có vai trò khác nhau (Hình 1.6) [51].

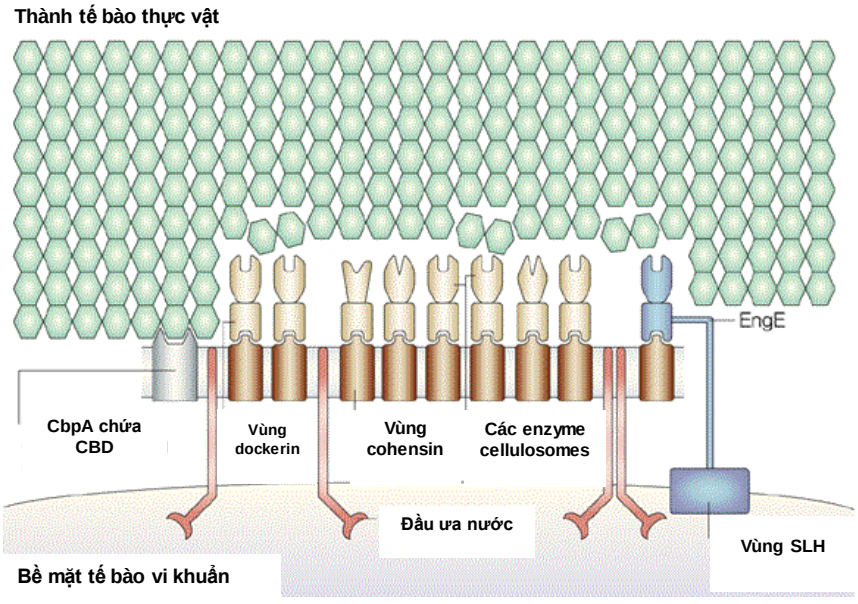


Hình 1.6. Cơ chế hoạt động của cellulase [51]

Đầu tiên, endoglucanase tấn công ngẫu nhiên và phân cắt các liên kết β 1,4 -glucoside bên trong các chuỗi cellulose, đặc biệt là ở các vùng vô định hình có cấu

trúc kém chặt chẽ, tạo ra các chuỗi oligosaccharid có các đầu khác nhau. Tiếp theo là exoglucanase (cellobiohydrolase) thủy phân các chuỗi này từ hai đầu của chúng tạo ra glucose, cellobiose và oligosaccharide ngắn. Hai enzyme này hoạt động hiệp đồng và thường bị ức chế bởi cellobiose [52]. Cuối cùng, β-glucosidase phân hủy cellobiose và các oligosaccharide ngắn thành các đơn vị glucose, do đó loại bỏ các ức chế cellobiose trên endoglucanase và cellobiohydrolases [53]. Đối với cellulose tinh thể, hoạt động của β-glucosidase hầu như không đáng kể. Hoạt động của ba loại enzyme trên xảy ra đồng thời, nâng cao hiệu quả phân giải cellulose.

Ở các vi sinh vật kị khí như *Clostridium*, *Acetivibrio*, *Bacteroides* và *Ruminococcus* thường sản xuất một lượng lớn phức hợp đa enzyme được gọi là cellulosome bám vào bề mặt tế bào vi sinh vật để phân giải cellulose. Các cellulosome không chỉ phân giải cellulose mà nó còn phân hủy thành tế bào thực vật. Cellulosome là phức hợp enzyme ngoại bào lớn có khả năng phân hủy cellulose, hemicelluloses và pectin. Chúng có thể là phức hợp enzyme ngoại bào lớn nhất được tìm thấy trong tự nhiên mặc dù kích thước cellulosome riêng lẻ nằm trong khoảng từ 0,65 MDa đến 2,5 MDa, một số polycellulosomes đã được báo cáo là lớn tới 100 MDa.



Hình 1.7. Cấu trúc cellulosome của vi khuẩn [36]

Gần đây, cấu trúc cellulosome phức tạp với nhiều protein khung cho phép liên kết nhiều enzyme hơn đã được tìm thấy [36]. Các kết nối cohensin-dockerin thành phần quy định cấu trúc tổng thể của cellulosome. Như vậy, hệ thống phức hợp enzyme

cellulosome hoạt động có hiệu quả trong việc phân giải cellulose do cấu trúc của nó, khả năng gắn kết hiệu quả với cơ chất và sự đa dạng của các enzyme thủy phân hoạt động hiệp đồng. Cellulosome chưa được xác định ở vi khuẩn phát triển trên 65 ºC và chưa được xác định trong vi khuẩn cổ [54].

1.2.4. Ứng dụng của cellulase

Cellulase là chất xúc tác sinh học có vai trò quan trọng, chúng có tiềm năng ứng dụng to lớn trong sản xuất. Trong công nghiệp dệt may, enzyme cellulase là nhóm enzyme lớn thứ ba được sử dụng trong ngành dệt may, đặc biệt đối với quá trình dệt ướt, phân hủy sinh học vải denim, đánh bóng sinh học sợi dệt, làm mềm hàng may mặc và loại bỏ thuốc nhuộm dư thừa khỏi vải. Cellulase còn được ứng dụng trong ngành công nghiệp giấy và bột giấy theo hướng tái chế và tái sử dụng giấy: nghiền thành bột, khử kim loại, xử lý sinh học chất thải công nghiệp, tẩy trắng và tăng cường chất xơ. Trong công nghiệp giặt và chất tẩy rửa thì các enzyme kiềm được sử dụng rộng rãi. Cellulase kiềm là chất phụ gia phù hợp nhất với chất tẩy rửa thông thường. Vì cellulase có khả năng loại bỏ đất và các hạt bụi bẩn từ các khoảng không của vải. Cellulase loại bỏ các cấu trúc thô ráp của sợi cellulose làm tăng độ bóng và mịn cho vải. Việc áp dụng cellulase trong nông nghiệp làm tăng năng suất cây trồng và hạn chế bệnh thực vật. Có nhiều loại cellulase vi khuẩn có khả năng thúc đẩy tăng trưởng thực vật, cải thiện năng suất cây trồng, bảo vệ cây trồng khỏi bệnh tật. Trong y tế, các cellulase được sản xuất bằng quá trình lên men tự nhiên của *Trichoderma reesei* và *Bacillus licheniformis* đã được đưa vào hỗn hợp enzyme nhằm tiêu hóa thực phẩm giàu chất xơ như trái cây và rau, ngũ cốc, các loại đậu, cám, các loại hạt và hạt. Cellulase từ nấm có thể áp dụng trong việc kiểm soát các mầm bệnh [55]. Ngày nay, cellulase được ứng dụng ngày càng nhiều trong công nghệ sinh học thực phẩm như: nước ép trái cây và rau quả, giảm độ nhớt của mật hoa, cô đặc chất tinh khiết, thay đổi các đặc điểm bề ngoài của quả [56], trong ngành sản xuất thức ăn chăn nuôi, cellulase được bổ sung để nâng cao khả năng sử dụng thức ăn có nguồn gốc từ ngũ cốc và để tăng giá trị dinh dưỡng cho thức ăn gia súc …

1.2.5. Tình hình nghiên cứu khai thác gene mã hóa cellulase ở thế giới và Việt Nam

1.2.5.1. Tình hình khai thác gene mã hóa cellulase trên thế giới

Trong sinh khối lignocellulose thì cellullose chiếm tỉ lệ lớn. Vì vậy, để nâng cao hiệu quả sử dụng sinh khối lignocellulose thì việc tìm kiếm và phân lập các chủng sinh enzyme phân giải cellulose và nghiên cứu đặc điểm, tính chất của enzyme này có vai trò quan trọng. Tuy nhiên, nhà khoa học đã nhận thấy số lượng vi sinh vật có thể phân lập được thông qua nuôi cấy là rất ít. Vì vậy, việc sử dụng kỹ thuật metageneomic nhằm nghiên cứu và khai thác các gene mã hóa cellulase trực tiếp từ môi trường không thông qua nuôi cấy có nhiều thuận lợi. Năm 2006, lần đầu tiên Xu và cộng sự phân tích DNA đa hệ gene của khu hệ vi sinh vật từ cặn bột giấy cho thấy đa dạng các vi khuẩn trong môi trường này gồm 4 ngành Spirochaetes, Proteobacteria, Bacteroidetes và Firmicutes. Việc sàng lọc chức năng gene thu được hai gene mã hóa endoglucanase, ba gene mã hóa exoglucanase và hai gene β-glucosidase [57]. Dữ liệu DNA đa hệ gene của vi sinh vật trong manh tràng thỏ cũng được dự đoán chức năng gene và sàng lọc cellulase [58]. Theo đó mười một gene mã hóa cellulase gồm bốn gene endo-β-1,4-glucanase và bảy gene β-glucosidase đã được phân lập. Theo dữ liệu của Guo và cộng sự (2008), từ dữ liệu $4,8x10^6$kb DNA đa hệ gene của vi sinh vật trong dạ cỏ trâu bò đã phân lập được 118 gene có hoạt tính β-glucosidase. Việc sàng lọc các gene này cho thấy tám gene có hoạt tính β-glucosidase cao ở pH 5,0 và 37°C và một trong số tám gene đó tiếp tục được khảo sát sâu hơn, thu được kết quả gene chọn lọc có độ tương đồng cao với gene mã hóa β-glucosidase từ *Bacillus* sp.. Năm 2008, Kim và cộng sự cũng đã tìm ra gene mới mã hóa endoglucanase từ mẫu đất rừng ở Hàn Quốc. Khai thác dữ liệu DNA đa hệ gene của vi sinh vật trong mẫu đất, nhóm nghiên cứu đã thấy một dòng pCM2 sử dụng carboxymethyl cellulose (CMC) làm nguồn carbon duy nhất. Các phân tích sâu hơn cho thấy hai gene celM2 và xynM2 chứa số amino acid lần lượt là 226 và 662 amino acid, trong đó trình tự amino acid suy diễn của celM2 tương đồng 36% với trình tự cellulase từ *Synechococcus* sp., trình tự amino acid của xynM2 tương đồng 59% với trình tự của endo-1,4-beta-xylanaseA từ *Cellulomonas pachnodae*. CelM2 tái tổ hợp thể hiện hoạt tính phân giải cơ chất CMC cao nhất ở pH 4,0 và 45°C. Mặc dù enzyme CelM2 có thủy phân cả cellulose tinh thể và xylan nhưng không thủy phân trên các cơ chất oligosaccharid như cellobiose, pNP-beta-cellobioside… Những kết quả này cho thấy CelM2 là một loại endoglucanase mới. Ngoài ra có rất nhiều gene mã hóa callulase mới được phát hiện

trên nhiều đối tượng khác nhau như ruột bào ngư [59], hệ vi sinh vật loài ruồi *Hermetia illucens* [60], phân trùn quế [61]…

1.2.5.2. Tình hình khai thác gene mã hóa cellulase ở Việt Nam

Ở Việt Nam, các phân tích và đánh giá về thành phần các loài vi sinh vật và phân lập, tìm kiếm các gene mã hóa cellulase đã được tiến hành từ những năm 2000. Trong đó các nghiên cứu đều tiến hành theo hướng phân lập các chủng sinh cellulase và tìm kiếm các enzyme mới của nấm mốc, xạ khuẩn [59]–[61]. Theo hướng này, Phan MTT và cộng sự (2012) đã phân lập các chủng vi khuẩn từ vùng ngập mặn tỉnh Nam Định và lựa chọn được chủng vi khuẩn *Bacillus* sp VLSH08 có khả năng sinh tổng hợp endo-1,4 β-glucanase ngoại bào. Kết quả kiểm tra cho thấy chủng *Bacillus* sp VLSH08 tương đồng 98% với chủng *Bacillus amyloliquefaciene* JN999857 và các enzyme thu được từ nhóm này đều thuộc cellulase [62]. Quyen và cộng sự (2018) đã tiến hành nghiên cứu đa dạng nấm mốc trong 6 mẫu đất ở rừng Mã Đà (Đồng Nai) trong đó có 19 chủng thuộc nhóm *Aspergillus niger*, 3 chủng thuộc C*urvularia* sp., 9 chủng thuộc *Penicilium lilacinum*, 2 chủng thuộc *Penicilium* sp.1, 3 chủng thuộc *Penicilium* sp.2, 3 chủng thuộc *Penicilium* sp.3, 2 chủng thuộc *Penicilium* sp.4, 1 chủng thuộc *Penicilium* sp.5, 3 chủng thuộc *Penicilium* sp.6, 3 chủng thuộc *Penicilium* sp.7 và 2 chủng thuộc *Trichoderma* sp. Nghiên cứu khả năng phân giải cellulose trên môi trường Czapek-Dox bổ sung 1% CMC cho thấy, tất cả các chủng nấm mốc này đều có khả năng phân giải cellulose, trong đó các chủng có hoạt tính cellulase cao thuộc chi *Penicilium* [63]. Tuy nhiên, việc nghiên cứu phụ thuộc môi trường nuôi cấy không thể đánh giá đầy đủ mức độ đa dạng loài các vi sinh vật và tìm kiếm được các gene mới mã hóa cellulase. Từ năm 2012, Trương Nam Hải và cộng sự đã bắt đầu sử dụng kỹ thuật metageneomic trong khai thác gene mã hóa enzyme thủy phân lignocellulose từ khu hệ vi sinh vật ruột mối Việt Nam bằng kỹ thuật metageneomic. Kết quả phân tích dữ liệu DNA đa hệ gene của vi sinh vật cho thấy khu hệ vi sinh vật rất phong phú khoảng 1460 loài, với 12 bộ phong phú nhất là Spirochaetales, Lactobacillales, Bacteroidales, Clostridiales, Enterobactees, Pseudomonades trong đó có 316 ORF có liên quan đến sự phân hủy cellulose bao gồm β-glucosidase, licheninases, endoglucanases, cellobiosidases, và phosphorylase cellobiose [64]. Cũng bằng kỹ thuật metageneomic, DNA đa hệ gene của vi sinh vật

trong dạ cỏ dê một số địa phương ở Việt Nam cũng đã được Do TH và cộng sự (2018) nghiên cứu khai thác, kết quả thu được 9 Gb DNA đa hệ gene trong đó có 816 ORF mã hóa 11 họ GH của cellulase [2]. Năm 2021, dữ liệu DNA đa hệ gene của vi sinh vật suối nước nóng Bình Châu đã được xác định có kích thước 9,4 GB. Qua phân tích đã xác định được phân loại học vi sinh vật gồm 41 ngành, 57 lớp, 128 bộ, 245 họ, 825 chi và 2.250 loài khác nhau; bộ dữ liệu về các gene mã hóa cho cellulase gồm 82 trình tự mã hóa cho endoglucanase, exoglucanase và β-glucosidase [65]. Như vậy, vi sinh vật từ các hệ sinh thái nhỏ có quá trình phân hủy cellulose mạnh như ruột mối, dạ cỏ dê hay suối nước nóng đã được nghiên cứu thành phần loài và nghiên cứu khai thác, tìm kiếm các mã hóa cellulase. Trong nghiên cứu này, chúng tôi tiếp tục sử dụng kỹ thuật metageneomic để phân tích, đánh giá đa dạng thành phần loài và tìm kiếm các gene mã hóa cellulase từ DNA đa hệ gene của khu hệ vi sinh vật ở hệ sinh thái có quá trình phân hủy lignocellulose diễn ra cũng rất mạnh mẽ đó là đất xung quanh khu nấm mục trắng ở vườn Quốc gia Cúc Phương.

1.3. Nấm mục trắng và khu hệ vi sinh vật xung quanh khu nấm mục trắng thủy phân lignocellulose

1.3.1. Nấm mục trắng

Lignocellulose là một nguồn sinh khối dồi dào cung cấp nguyên liệu cho ngành sản xuất nhiên liệu và hóa chất. Tuy nhiên, quá trình phân giải thành phần carbohydrate của lignocellulose bị cản trở bởi lignin. Đây là chất khó phân hủy hóa học và sinh học do lignin có đặc điểm cấu trúc hóa học phức tạp và các liên kết không thống nhất. Có ba nhóm nấm khác nhau với tác động và cơ chế phân giải lignocellulose khác nhau đã được xác định đó là nấm mục mềm, nấm mục nâu và nấm mục trắng. Trong đó, nấm mục trắng là nhóm duy nhất có khả năng phân hủy tất cả các thành phần của lignocellulose trong rơm rạ: lignin, cellulose và hemicellulose [66]. Khi sống trên giá thể gỗ, nấm mục trắng là nhóm có khả năng phân hủy lignin hiệu quả nhất [67]. Khả năng này có được là do nấm mục có hệ thống enzyme ngoại bào độc đáo không đặc hiệu cũng như các enzyme oxi hóa nội bào, từ đó nấm mục trắng có thể khoáng hóa hoàn toàn cơ chất lignin thành CO_2 [68] và phân hủy một loạt các chất khác nhau gồm các chất độc gây ô nhiễm có mùi thơm như hydrocarbon đa vòng thơm, polychlorinated biphenyls, thuốc nhuộm azo, thuốc trừ sâu và dược

phẩm. Vì vậy, nấm mục trắng tham gia vào chu trình carbon và đóng vai trò quan trọng trong việc cung cấp chất dinh dưỡng trong các rừng nhiệt đới [69] đồng thời nấm mục trắng và đất xung quanh khu nấm mục trắng cũng là một nguồn quan trọng để tìm kiếm các gene phân giải lignocellulase. Một số nấm mục trắng có khả năng phân giải các thành phần của gỗ là *Phanerochaete chrysosporium, Phanerochaete carnosa, Pleurotus ostreatus, Pyctisnoparinusa cin, Stropharia coronilla* và *Trametes versicolor* [69].

1.3.2. Tương tác giữa nấm mục trắng và khu hệ vi sinh vật xung quanh nấm mục trắng

Cùng hệ sinh thái với nấm mục trắng thì khu hệ vi sinh vật đất xung quanh nấm mục trắng cũng là đối tượng tiềm năng để nghiên cứu khai thác, tìm kiếm các gene mới mã hóa enzyme tham gia chuyển hóa cellulose [70]–[72]. Nấm mục trắng có khả năng phân giải các thành phần của lignocellulose trong đó khả năng phân giải lignin là hiệu quả nhất. Để phân giải hiệu quả lignocellulose thì không chỉ có sự tham gia của nấm mà còn có cả khu hệ vi sinh vật trong đất xung quanh khu nấm mục trắng. Hiện nay, có nhiều nghiên cứu chứng minh vai trò của vi khuẩn và nấm trong thủy phân lignocellulose nhưng các công trình nghiên cứu mối tương tác chặt chẽ giữa nấm và vi sinh vật thì còn khá ít. Haq và cộng sự (2022) đã nghiên cứu xác định quần xã vi sinh vật xung quanh nấm mục trắng *Fomes fomentarius* trên thân cây bạch dương. Kết quả cho thấy quần xã vi sinh vật xung quanh nấm này đều được thống trị bởi Protobacteria tiếp theo là Firmicutes, Actinobacteria, Acidobacteria và ở xung quanh khu nấm mục đều có độ đa dạng vi sinh vật kém hơn [3]. Trong nghiên cứu của Boer và cộng sự khi đánh giá ảnh hưởng của nấm gây bệnh thối trắng lên quần xã vi sinh vật trên các khối gỗ sồi vô trùng nhận thấy vi khuẩn kém đa dạng hơn ở môi trường tươi. Như vậy ở các môi trường chọn lọc như gỗ mục nát, giữa nấm và vi khuẩn có sự tương tác qua lại với nhau để cùng tồn tại. Vi khuẩn thích ứng được với điều kiện môi trường thường xuyên thay đổi và khắc nghiệt do nấm tạo ra, trong khi nấm là sinh vật nhân chuẩn có nhu cầu dinh dưỡng cao hơn, hệ enzyme có khả năng oxi hóa cao hơn nên phân hủy lignocellulose tốt hơn [73]. Các nhà khoa học khi tiến hành đồng nuôi cấy nấm mục trắng với vi sinh vật khác trong phòng thí nghiệm thì khả năng chuyển hóa lignocellulose của nấm tăng lên. Folman và cộng sự đã nghiên

cứu ảnh hưởng của 2 nấm mục *Hypholoma fasciculare* và *Resinicium bicolor* lên số lượng và thành phần vi khuẩn sinh sống trên các khối gỗ sồi từ đất rừng. Tổng số vi khuẩn xung quanh và số vi khuẩn sống trên khối gỗ có nấm mục trắng là rất ít so với tổng số vi khuẩn xung quanh khối gỗ đối chứng. Điều này cho thấy nấm mục trắng đã cạnh tranh với các vi khuẩn sống trong cùng khu hệ sinh thái. Sự có mặt của nấm mục trắng dẫn đến sự thay đổi tương đối số lượng các họ vi khuẩn xung quanh cây gỗ [74]. Wieschen và cộng sự cũng đã nghiên cứu tương tác giữa nấm mục trắng và vi khuẩn đất dựa trên đánh giá sự phân hủy các hợp chất gây ô nhiễm đất (là các hidrocacbon thơm đa vòng, gồm 3 dạng: ba, bốn và năm vòng. Hai loài nấm được thử nghiệm là *Dichomitus squalens* và *Pleurotus ostreatus* tiết ra lượng enzyme phân giải lignin là như nhau nhưng *P. ostreatus* có khả năng khoáng hóa các hydrocacbon thơm tốt hơn, đặc biệt là các hydrocacbon 5 vòng thơm. Trong khi đó, vi sinh vật xung quanh khu nấm mục trắng lại có khả năng phân hủy các hợp chất thơm 3 vòng và 4 vòng thơm một cách mạnh mẽ. Trong đồng nuôi cấy nấm mục trắng và vi khuẩn, khả năng khoáng hóa các hợp chất gây ô nhiễm của cả vi khuẩn đất và *P. ostreatus* bị hạn chế một phần do tương tác đối kháng nhưng cơ bản vẫn được duy trì. Do đó, với sự có mặt của *P. ostreatus* đã làm tăng đáng kể quá trình phân giải các hydrocacbon thơm khối lượng phân tử cao, đồng thời làm giảm sự khoáng hóa của các hợp chất thơm khối lượng phân tử thấp.

Như vậy, sự thủy phân lignocellulose của các enzyme từ nấm mục trắng để có hiệu quả cao thì thường được kết hợp cùng với enzyme của các vi sinh vật sống trong cùng khu hệ sinh thái. Sự tương tác giữa nấm và vi khuẩn sống trong cùng khu vực có thể là quan hệ hỗ trợ và/hoặc cạnh tranh [75]. Khi đồng nuôi cấy nấm và vi khuẩn, vi khuẩn không những tiết enzyme để nâng cao hiệu quả thủy phân lignocellulose cùng với nấm, mà vi khuẩn còn giúp tạo ra môi trường thuận lợi cho nấm chuyển hóa lignocellulose, không cạnh tranh nguồn dinh dưỡng với nấm, sử dụng các chất có phân tử lượng thấp từ nấm chuyển hóa để làm thức ăn, cung cấp nguồn nitơ cho nấm, giúp phòng tránh nhiễm vi sinh vật độc hại. Trong quá trình nấm phân giải gỗ, điều kiện môi trường trở lên rất chọn lọc đối với vi khuẩn do quá trình acid hóa nhanh và mạnh, là sản phẩm của các phản ứng oxi hóa khử và sự có mặt của các chất độc từ nấm là sản phẩm của quá trình chuyển hóa thứ cấp. Vi khuẩn tồn tại trong những điều

kiện này phải có những đặc tính đặc biệt và mới mẻ. Ở những hệ sinh thái trên cạn, nơi mà sự phân hủy các chất hữu cơ phức tạp là đáng kể thì nấm tồn tại nhiều. Sự xuất hiện nhiều của nấm ở hệ sinh thái trên cạn có tác động mạnh mẽ đến mức độ tiến hóa của cộng đồng vi khuẩn ở hệ sinh thái này. Một mặt, sự phân hủy các chất hữu cơ phức tạp như lignin đã làm mất các vi khuẩn cũ, mặt khác sự xuất hiện của nấm đã tạo ra các chủng vi khuẩn mới tương ứng với nó.

1.4. Metageneomic và một số công cụ tin sinh, cơ sở dữ liệu được sử dụng trong khai thác DNA đa hệ gene

Theo truyền thống, nghiên cứu về vi sinh vật thường dựa trên việc nuôi cấy, tuy nhiên việc này có một số nhược điểm. Metageneomics là một phương pháp mới để nghiên cứu về tổng số bộ gene của quần xã vi sinh vật trong một môi trường cụ thể bằng cách sử dụng sàng lọc chức năng gene hoặc sàng lọc trình tự. Trong metageneomics, việc nghiên cứu các hệ gene của các vi sinh vật trong một quần xã không chỉ cho biết về di truyền, sinh lý và hóa sinh của các vi sinh vật mà còn cung cấp thông tin chi tiết về vòng tuần hoàn dinh dưỡng và năng lượng trong quần xã, cấu trúc bộ gene, chức năng gene, di truyền quần thể và chuyển gene giữa các thành viên của một quần thể sinh vật không thể nuôi cấy. Nghiên cứu metageneomics đang phát triển nhanh chóng trong y học, nông nghiệp, bảo vệ môi trường và các lĩnh vực khác.

Metageneomics cung cấp thông tin về chức năng gene của các quần xã vi sinh vật và do đó đưa ra mô tả rộng hơn nhiều so với các khảo sát về nguồn gốc gene thường chỉ dựa trên sự đa dạng của một gene chẳng hạn như gene 16S rRNA. Bằng kỹ thuật metageneomics, các thông tin về chất xúc tác sinh học hoặc các enzyme mới, mối liên kết giữa chức năng và phát sinh loài đối với các sinh vật chưa được nuôi cấy có thể được phát hiện. Metageneomics cũng là một công cụ mạnh mẽ để tạo ra các giả thuyết mới về chức năng của vi sinh vật như quang dị dưỡng dựa trên sinh vật quang dị dưỡng hoặc vi khuẩn cổ oxi hóa amoniac.

Đây còn là công cụ được sử dụng để khai thác các gene mã hóa các enzyme mới có ý nghĩa trong công nghiệp và sản xuất ở các địa điểm khác nhau như: đất [70], nước, ruột mối [71], dạ cỏ của động vật nhai lại [72]…

1.4.1. Các phương pháp khai thác gene bằng metageneomics

Metageneomics khai thác đa hệ gene theo hai hướng chính là: (1) Thiết lập thư viện DNA đa hệ gene và từ đó phân lập gene và (2) dựa trên dữ liệu giải trình tự trực tiếp DNA đa hệ gene để từ đó khai thác, tìm kiếm và phân lập gene. Trong đó, cách nghiên cứu dựa trên dữ liệu giải trình tự trực tiếp DNA đa hệ gene tỏ ra có nhiều ưu thế [72].

1.4.1.1. Phân lập gene từ thư viện DNA đa hệ gene

Công nghệ metageneomics trong giai đoạn đầu của sự phát triển chủ yếu dựa vào thư viện DNA đa hệ gene để phân lập gene. Sử dụng phương pháp này đã phát hiện nhiều enzyme phân giải cellulose như: từ thư viện DNA đa hệ gene của vi sinh vật trong dạ cỏ trâu có 61 ORF khác nhau có hoạt tính cellulase đã được phân lập, trong đó 13 ORF có hoạt tính endoglucanase [76]; từ thư viện DNA đa hệ gene của vi sinh vật sống trong chất thải của nhà máy giấy, có 7 gene mã hóa cellulase đã được xác định gồm: 2 ORF có hoạt tính endoglucanase, 3 ORF có hoạt tính exoglucanase và 2 ORF có hoạt tính β-glucosidase [58]; từ DNA đa hệ gene của vi sinh vật trong dạ dày thỏ có 11 ORF mã hóa enzyme có hoạt tính cellulase đã được phân lập bao gồm: 4 ORF mã hóa endo-β-1,4-glucanase thuộc họ GH5 và GH3, 7 ORF mã hóa β-glucosidase. Từ 102.000 ORF của thư viện DNA đa hệ gene vi sinh vật ruột lợn Yorkshire, có 11 gene mã hóa cellulase, 4 gene mã hóa hemicellulase, 1 gene mã hóa polygalacturonase, 1 gene mã hóa enzyme thuộc họ mananase và 1 gene mã hóa cellobiose phosphorylase đã được phân lập [71]. Từ thư viện cosmid với các đoạn chèn là DNA đa hệ gene vi sinh vật đất bón phân hữu cơ kích thước trung bình khoảng 33 kb gồm khoảng 100.000 dòng, Pang (2009) đã tách dòng được 3 gene gồm *umcel9A* kích thước 1.852 bp, *umcel9B* kích thước 1.740 bp và *umcel9C* kích thước 1.761 bp đều mã hóa endoglucanase thuộc họ GH9 và 1 gene *umcel5A* kích thước 1.047 bp mã hóa endoglucanase thuộc họ GH5 [58]. Ngoài ra, nhiều gene khác cũng được sàng lọc từ thư viện DNA đa hệ gene, ví dụ, từ khoảng 930.000 dòng của thư viện DNA đa hệ gene của 3 mẫu đất khác nhau đã sàng lọc được 5 dòng thể hiện hoạt tính 4-hydroxybutyrate dehydrogeneease trên môi trường cơ chất 4-hydroxybutyrate.

Mặc dù, metageneomics khai thác đa hệ gene thông qua xây dựng và sàng lọc các thư viện đa hệ gene là hướng tiếp cận không quá phức tạp, ít tốn kém, tuy nhiên nó có 3 hạn chế: (1) Sự giới hạn của các hệ thống sàng lọc; (2) không phải tất cả các

gene có thể được biểu hiện một cách hiệu quả trong *E. coli*; (3) thư viện đa hệ gene có kích thước bị giới hạn. Ngoài ra, một gene hoàn chỉnh có thể thể hiện hoạt tính tốt trên cơ chất hay không còn phụ thuộc vào sự phù hợp và vị trí gắn kết của nó với promoter của vector dùng để tạo thư viện.

1.4.1.2. Nghiên cứu khai thác gene từ DNA đa hệ gene

Hiện nay, để nghiên cứu khai thác và tìm kiếm các gene tiềm năng từ dữ liệu DNA đa hệ gene, thường có 3 bước: (1) tách chiết và giải trình tự các mẫu DNA đa hệ gene; (2) tập hợp các đoạn read ngắn thành các đoạn contig dài; (3) sử dụng các phần mềm chuyên dụng để ước đoán chức năng gene.

Trong các năm qua, giải trình tự DNA đa hệ gene đã dần chuyển từ công nghệ giải trình tự Sanger cổ điển sang giải trình tự thế hệ mới. Giải trình tự Sanger được coi là phương pháp chuẩn để giải trình tự vì tỷ lệ lỗi thấp, chiều dài đọc lớn (> 700 bp) và kích thước chèn lớn (> 30 Kb đối với fosmid hoặc nhiễm sắc thể nhân tạo của vi khuẩn). Kỹ thuật này vẫn có thể áp dụng hiệu quả với mục tiêu tạo ra các bộ gene gần hoàn chỉnh với độ đa dạng thấp, tuy nhiên tốn nhiều công sức, chi phí. Kỹ thuật giải trình tự thông lượng cao (High Throughput Sequencing - HTS) được áp dụng ngày càng nhiều cho các mẫu DNA đa hệ gene và có nhiều đánh giá tuyệt vời. Kỹ thuật HTS (454/ Roche và Illumina/Solexa) tạo ra độ dài đọc trung bình 600-800 bp, đủ dài để chỉ gây ra các sai khác nhỏ trong mỗi lần chú thích [77]. Cho đến nay, Illumina đã sản xuất được máy giải trình tự HiSeq 2500 mới có khả năng cho 900 Gb-1Tb/mỗi lần chạy trong 6 ngày ở chế độ chạy công suất cao và 200 – 300 Gb/mỗi lần chạy trong 60 giờ ở chế độ chạy nhanh. Với chi phí thấp, độ dài đọc lớn và khả năng ứng dụng cao trong nghiên cứu metageneome thu nhận từ môi trường đã làm cho kỹ thuật giải trình tự HTS bằng công nghệ Illumina được sử dụng ngày càng phổ biến. DNA đa hệ gene sau khi được giải trình tự là các dữ liệu thô, các dữ liệu này sẽ được chọn lọc để thu được các dữ liệu tinh. Dữ liệu tinh là các đoạn trình tự ngắn riêng rẽ (gọi là các read) được tập hợp và lắp ráp lại thành các đoạn contig có kích thước dài hơn [78]. Quá trình sắp xếp các read có thể được thực hiện theo 2 cách: lắp ráp dựa trên tham chiếu (đồng lắp ráp) hoặc lắp ráp *de novo*. Các read sẽ được tập hợp dựa trên trình tự tham chiếu đã có bằng các phần mềm như Newbler (Roche), AMOS http://sourceforge.net/projects/amos/, hoặc MIRA. Phương pháp này có hiệu

quả khi bộ dữ liệu DNA đa hệ gene của mẫu với bộ gene tham chiếu sự nhiều tương đồng. Tuy nhiên, sự khác biệt trong DNA đa hệ gene của mẫu so với tham chiếu có thể có nghĩa là gene mẫu bị phân mảnh hoặc các vùng khác nhau không được che phủ. Một lượng lớn các read cũng có thể được lắp ráp *de novo* dựa trên sơ đồ Bruijn. Tuy nhiên, phương pháp này yêu cầu về bộ nhớ của máy khá lớn, thời gian vài ngày và đối với các quần xã vi sinh vật có thành phần chi và loài phức tạp như môi trường đất thì việc lắp ráp và sắp xếp các read khá khó khăn và có nhiều sai lệch [79]. Các read sau khi được tập hợp và chỉnh sửa sẽ được ước đoán gene bằng nhiều phần mềm chuyên dụng như FragGeneeScan (FGS), MetaGeneeMark (MGM), MetaGeneeAnnotator (MGA)/Metagenee...

Các gene của DNA đa hệ gene được dự đoán về đơn vị phân loại của gene và chức năng gene. Dựa vào mức độ tương đồng của trình tự DNA đa hệ gene của mẫu thu được với các trình tự của các CSDL tham khảo sẽ ước đoán được đơn vị phân loại và chức năng của các gene. Hiện nay CSDL về đơn vị phân loại của gene thường dùng là CSDL NR (là CSDL chứa các trình tự non – redundant từ ngân hàng gene cùng với các trình tự từ các dữ liệu ngân hàng khác như Refseq, PDB, SwissProt, PIR và PRF), các CSDL về chức năng gene đáng tin cậy như: Kyoto Encyclopedia of Genees and Geneeomes (KEGG) (https://www.kegg.jp) là CSDL phân loại chức năng gene theo con đường chuyển hóa [80], evolutionary geneealogy of genees: Non-supervised Orthologous Groups (eggNOG) là CSDL phân loại chức năng theo tiến hóa của gene [81], Clusters of Orthologous Group (COG) là một hệ thống của các họ gene từ các geneom hoàn chỉnh, KOG- eukaryotic orthologous groups là CSDL từ 7 hệ gene của sinh vật nhân chuẩn: 3 loài động vật, 1 loài thực vật *Arabidopsis thaliana*, 2 loài nấm và các ký sinh trùng nội bào, CSDL protein families (PFAM) là CSDL về các họ protein [82]…. Tuy nhiên, không có một CSDL nào chứa đầy đủ tất cả các thông tin về đơn vị phân loại và chức năng sinh học của gene trong DNA đa hệ gene. Nên việc hợp nhất các CSDL trong một chương trình duy nhất là cần thiết và đã được triển khai trong phiên bản mới nhất của MG-RAST và IMG/M [83].

1.4.2. Một số công cụ tin sinh để khai thác dữ liệu DNA đa hệ gene

1.4.2.1. Sử dụng BLAST để so sánh với CSDL của NCBI

BLAST (Basic Local Alignment Search Tool)

(https://blast.ncbi.nlm.nih.gov/Blast.cgi) là một trong các công cụ phổ biến dùng trong sinh học tính toán, dựa trên thuật toán tìm kiếm những trình tự axit nucleic/protein tương đồng lưu trữ trên nhiều CSDL. Khi CSDL phù hợp được đưa vào, BLAST sẽ tìm kiếm trong ngân hàng NCBI các chuỗi giống với chuỗi ban đầu. Có 5 loại BLAST cơ bản bao gồm BLASTn tìm kiếm các trình tự nucleotide tương đồng với trình tự nucletide của DNA đầu vào trong CSDL DNA, BLASTp tìm kiếm tất cả trình tự amino acid tương đồng với trình tự amino acid của protein đầu vào trong CSDL protein, BLASTx tìm kiếm các protein tiềm năng được mã hóa bởi các chuỗi nucleotide đưa vào, tBLASTn tìm kiếm các trình tự nucleotide mã hóa protein tương đồng với protein đưa vào, tBLASTx tìm kiếm các trình tự nucleotide tương tự như trình tự đưa vào dựa trên các protein mà chúng mã hóa. Trong nghiên cứu về DNA đa hệ gene, sau khi dự đoán được các gene mã hóa enzyme đích có thể sử dụng công cụ BLASTp để dự đoán đơn vị phân loại loài của gene, chú thích chức năng gene hay điều tra vùng bảo tồn của mỗi enzyme do ORF mã hóa [84].

Để dự đoán các mức độ phân loại của loài, các ORF trong dữ liệu DNA đa hệ gene của mẫu được so sánh các ORF với CSDL NR để tiến hành phân loài. Ngoài ra, đơn vị phân loại của loài còn được dự đoán bằng phần mềm MEGAN (MEtaGeneomic Analyser). Cấp độ phân loại loài của mỗi trình tự được xác định bằng thuật toán LCA (Least Common Ancestors). Thuật toán này sẽ căn cứ vào mức độ bảo thủ của trình tự gene để xếp gene đó vào các nhóm phân loại loài khác nhau. Để dự đoán chức năng gene, các trình tự amino acid tương ứng được so sánh với một số cơ sở dữ liệu như: KEGG, eggNOG, Swiss - Prot, COG, GO, CAZy, ARDB... Các CSDL này sẽ được phối hợp với nhau để đưa ra dự đoán chung nhất về chức năng của gene.

1.4.2.2. Phân tích các vùng chức năng của ORF bằng HMM profile

Các protein để thực hiện được chức năng xúc tác sinh học, chúng thường có một hoặc nhiều vùng chức năng, thường được gọi là vùng (domain). Việc xác định các vùng chức năng có trong protein có ý nghĩa quan trọng trong việc nâng cao khả năng xúc tác của protein cũng như ứng dụng trong sản xuất công nghiệp. Cơ sở dữ liệu Pfam (http://pfam-legacy.xfam.org/) là một tổ hợp các họ protein và vùng chức năng của protein được sử dụng rộng rãi để phân tích các hệ gene, đa hệ gene và để

định hướng thử nghiệm trên các protein và hệ thống cụ thể [85]. Pfam dựa trên mô hình Markov ẩn (cấu hình HMM) của các vùng chức năng của protein hoàn chỉnh. Việc xác định các vùng chức năng của protein, các thành viên trong họ protein và sự bắt cặp là dựa trên sự tương đồng về trình tự và các cấu hình HMM để xác định chính xác và sắp xếp các thành viên [86]. Mô hình đại diện HMM không chỉ tính toán trên một điểm bắt cặp mà còn tính toán tổng các xác suất trên toàn bộ tập hợp bắt cặp. Mô hình đại diện HMM cho biết thông tin cụ thể về các vị trí trong trình tự, loại gốc amino acid hay nucleotide nào xuất hiện nhiều nhất, khả năng xuất hiện các đột biến chèn hoặc mất, vì vậy cách tiếp cận này có nhiều thuận lợi. HMM đặc biệt có ý nghĩa khi nghiên cứu vùng chức năng của các họ, khi mà có thể sử dụng một mô hình đại diện cho một họ gồm hàng trăm trình tự riêng lẻ tương đồng [87]. Tuy nhiên, việc triển khai mô hình đại diện HMM trước đây chậm hơn BLAST khoảng 100 lần. Điều này làm giảm hiệu quả của chúng, vì tốc độ tính toán rất quan trọng với kích thước ngày càng tăng nhanh chóng của CSDL hiện đại. Hiện nay đã có phần mềm mới là HMMER3 giúp tìm kiếm nhanh như BLAST, trong khi vẫn giữ được sức mạnh của việc sử dụng công nghệ suy luận xác suất. Trong mô hình đại diện HMM, để tìm kiếm những trình tự tương đồng, kết quả ban đầu được lọc với giá trị E (e-value) nhỏ hơn e^{-10}, tỷ lệ chiều dài đoạn tương đồng dùng để tìm kiếm so với chiều dài mô hình đó lớn hơn 0,75 và tỷ lệ giá trị bias:score nhỏ hơn 0,1. Mô hình đại diện HMM chuyển kết quả so sánh đa trình tự thành một hệ thống điểm (score) đặc trưng cho từng vị trí, từ đó có thể sử dụng để so sánh trình tự, tìm kiếm trong CSDL các trình tự tương đồng. Hiện nay, Pfam khớp với 72% trình tự protein đã biết nhưng đối với các protein có cấu trúc đã biết thì Pfam khớp với 95% [88]. Để dự đoán các vùng chức năng của gene dựa trên CSDL Pfam, các trình tự protein quan tâm được tập hợp dưới dạng file fasta và gửi lên trang web của HMMer (https://www.ebi.ac.uk/interpro/search/sequence/) giá trị e – value được sử dụng là 1.0, kết quả sẽ trả về địa chỉ e-mail cá nhân sau 2 – 3 ngày tùy thuộc số lượng, chiều dài các trình tự và số lượng các vùng chức năng trên các protein đích.

1.4.2.3. Dự đoán mức độ biểu hiện của gene trong E. coli

Mức độ biểu hiện của protein tái tổ hợp trong vật chủ *E. coli* có ý nghĩa quan trọng. Mức độ biểu hiện này có vai trò quan trọng trong việc thu được các protein ở

dạng hòa tan và hoạt tính của các protein đích. Mức độ biểu hiện này trong khoang chu chất chịu ảnh hưởng của nhiều yếu tố khác nhau như: trình tự amino acid, các đoạn peptide tín hiệu [89], tốc độ gấp của protein [90]… Một số thuật toán và công cụ tính toán đã được phát triển để dự đoán khả năng hòa tan của protein và tốc độ gấp của protein dựa trên mối tương quan giữa trình tự amino acid và hai đặc tính quan trọng này của protein [91], [92]. Gần đây, Periscope (Periplasmic expression classifier for soluble protein expression) được xem như một công cụ dự đoán mức độ biểu hiện của gene ngoại lai trong tế bào vi khuẩn *E. coli*. Phần mềm dự đoán mức độ biểu hiện gene Periscope dựa trên mô hình SVM (Support Vector Mechine) gồm 4 bước: (1) Xây dựng hệ thống CSDL. Thay vì sử dụng các CSDL có sẵn như các phần mềm dự đoán khác, Periscope sử dụng các dữ liệu thực tế thông qua các công trình công bố trên thư viện NCBI. Protein có nồng độ 100 mg/l hoặc lớn hơn được định nghĩa là có mức độ biểu hiện cao, mức độ biểu hiện thấp có nồng độ 0,5 mg/l hoặc nhỏ hơn. Những giá trị ở giữa hai mốc này được coi là có mức độ biểu hiện trung bình. CSDL này được phân tách ngẫu nhiên thành hai bộ để xây dựng và kiểm định thử mô hình ước đoán với tỉ lệ lần lượt là 85%:15%. (2) Xây dựng bảng thuộc tính và sàng lọc thuộc tính. Có tất cả 7903 thuộc tính trích từ trình tự amino acid được xác định để xây dựng lên Periscope. (3) Thuật toán SVM cho mô hình phân biệt và hồi quy. (4). Kiểm định hiệu suất của mô hình.

Để cung cấp quyền truy cập vào công cụ dự đoán hai giai đoạn Pericsope này, một máy chủ web trực tuyến (http://lightning.med.monash.edu/periscope/index.jsp) đã được thiết kế tương đối dễ sử dụng. Khi người dùng gửi các trình tự amino acid, Periscope thực hiện dự đoán bằng cách sử dụng các mô hình đã xây dựng và sau đó trả về mức độ biểu hiện và lượng protein hòa tan được dự đoán trong khoang chu chất của *E. coli*. Nó cho phép gửi tối đa năm chuỗi truy vấn ở định dạng FASTA mỗi lần gửi và không có giới hạn về độ dài của chuỗi truy vấn. Periscope với cấu trúc hai giai đoạn còn có thể dự đoán định lượng các protein hòa tan trong *E. coli*. Dựa trên trình tự amino acid được cung cấp của peptit tín hiệu và protein đích, Periscope có thể phân loại biểu hiện của protein đích dạng tan thành ba mức độ: biểu hiện cao, trung bình hoặc thấp và dự đoán thêm lượng protein hòa tan trong *E. coli,* tính bằng đơn vị mg/l. Các kết quả này có thể truy xuất trực tiếp hoặc gửi về email của người dùng.

1.4.2.4. Ước đoán cấu trúc không gian và vị trí gắn cơ chất của enzyme

Để ước đoán cấu trúc không gian của các chuỗi protein, có thể sử dụng nhiều phần mềm. Trong đó, Phyre2 (http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index) là phần mềm dễ sử dụng, đưa lại kết quả trong thời gian ngắn. Phyre 2 dựa trên các nguyên tắc tương đồng ở các vùng bảo tồn cao của protein, cho phép dự đoán cấu trúc không gian của protein ở các cấp độ khác nhau, chức năng, nguồn gốc của protein…. Trình tự amino acid của các protein đích được quét với CSDL các trình tự protein và tìm kiếm sự tương đồng trong cấu trúc bậc hai, cấu trúc bậc ba của các protein đó và xuất ra kết quả. Để dự đoán cấu trúc không gian của protein, người dùng sẽ nhập các trình tự chuỗi amino acid định dạng fasta (thường tối đa là 100 trình tự trong 1 lần) và chờ khoảng thời gian nhất định (tùy thuộc vào chiều dài chuỗi, số lượng trình tự tương đồng, tần số lặp lại...) công cụ sẽ đưa ra một dự đoán về cấu trúc không gian của protein. Kết quả dự đoán sẽ được trả về e-mail đăng ký. Kết quả nhận được cho biết chức năng của protein đang nghiên cứu dựa trên chức năng của gene khuôn, độ bao phủ so với gene khuôn mẫu và độ tin cậy của kết quả thu được. Các gene khuôn được sắp xếp theo chiều giảm dần của độ tương đồng. Kết quả cũng cho thấy các cấu trúc bậc hai của protein đích so với cấu trúc bậc 2 của khuôn, cấu trúc của protein trong không gian ba chiều và một số các trình tự amino acid đặc biệt là vùng bảo thủ của protein đích. Các cấu trúc không gian và trình tự amino acid này có vai trò quan trọng trong việc thể hiện chức năng của protein [93].

1.4.2.5. Dự đoán khả năng chịu kiềm/axit của enzyme

Giá trị pH có vai trò quan trọng, ảnh hưởng đến khả năng xúc tác của enzyme. Hầu hết các enzyme hoạt động tốt trong phạm vi pH từ 6 đến 8, một số enzyme cụ thể chỉ hoạt động tốt trong điều kiện axit mạnh (nghĩa là pH < 5,0) hoặc kiềm mạnh (tức là pH > 9,0). Zhang và cộng sự đã trình bày một mô hình ngẫu nhiên để phân biệt các enzyme axit với các enzyme kiềm bằng cách sử dụng thông tin về trình tự và cấu trúc. Mô hình có thể đạt được độ chính xác tổng thể là 90,7% trong quá trình xác nhận chéo 10 lần. Tuy nhiên, độ chính xác vẫn chưa như mong đợi. Hơn nữa, họ không cung cấp máy chủ web để các nhà khoa học thực nghiệm có thể thu được kết quả mong muốn bằng cách áp dụng các phương pháp của họ [94]. Gần đây, Fan và

cộng sự đã thiết kế máy chủ web miễn phí gọi là Pred-enzyme để dự đoán các enzyme có tính axit và kiềm. Công cụ dự đoán có thể đạt được độ chính xác tổng thể là 94,01% trong quá trình xác nhận chéo 10 lần. Tuy nhiên, công cụ dự đoán của họ cần thông tin về bản đồ ngữ nghĩa gene (GO). Trong khi hầu hết các protein không có thông tin GO (<50%) [95]. Nếu một protein truy vấn chưa được chú thích trong CSDL GO và chưa có thông tin về bản đồ ngữ nghĩa gene thì dự đoán với mô hình sẽ không khả dụng. Khắc phục những nhược điểm này, nhằm dự đoán trước điều kiện pH tối ưu cho hoạt động của các enzyme, máy chủ trực tuyến AcalPred đã được sử dụng (http://lin.uestc.edu.cn/server/AcalPred). Khi ta cung cấp các trình tự chuỗi protein dạng fasta, phần mềm sẽ trả về kết quả là hai chỉ số thể hiện khả năng chịu kiềm và khả năng chịu axit của protein sau 1 – 3 phút [96]. Nếu protein có chỉ số chịu kiềm là từ 0,5 – 1 và chỉ số chịu axit từ 0 – 0,5 thì enzyme đó hoạt động tốt trong môi trường kiềm và giá trị chịu kiềm càng gần 1 thì enzyme hoạt động tối ưu trong môi trường càng kiềm cao và ngược lại. Các enzyme có chỉ số chịu kiềm từ 0 – 0,5, chỉ số chịu axit từ 0,5 – 1 thì enzyme đó hoạt động tốt trong môi trường axit. Phần mềm này cho phép người dùng ước đoán cùng lúc hàng trăm trình tự acid amin khác nhau.

1.4.2.6. Dự đoán khả năng chịu nhiệt của enzyme

Việc dự đoán khả năng chịu nhiệt của enzyme có vai trò quan trọng vì đây là tiền đề cho việc lựa chọn được các enzyme chịu nhiệt có tiềm năng ứng dụng cao trong các hoạt động sản xuất công nghiệp [97]. Để thực hiện việc này, phần mềm miễn phí TBI của Đài Loan, Trung Quốc đã được thiết kế dựa trên tổng hợp các đặc điểm chịu nhiệt của protein đã nghiên cứu như trình tự, thành phần amino acid của protein, liên kết hidro giữa các phân tử, tương tác kỵ nước, lực Van de waals… và các enzyme từ các vi sinh vật sống ở các suối nước nóng từ đó xây dựng vector hỗ trợ cho việc dự đoán đặc tính này [98]. Enzyme được dự đoán chịu nhiệt ở ba mức là dưới 55°C, 55 – 65°C và trên 65°C. Với phần mềm TBI, người dùng có thể ước đoán cùng lúc khả năng chịu nhiệt của nhiều trình tự amino acid khác nhau.

1.4.3. Một số cơ sở dữ liệu

1.4.3.1. The National Center for Biotechnology Information (NCBI)

Trung tâm Thông tin Công nghệ Sinh học Quốc gia (NCBI) http://www.ncbi.nlm.nih.gov tại Viện Y học Quốc gia Hoa Kỳ được thành lập để phát

triển hệ thống thông tin cho sinh học phân tử. Ngoài việc lưu trữ CSDL trình tự axit nucleic của GeneBank®, NCBI còn cung cấp các phân tích và truy xuất dữ liệu trong GeneBank và các dữ liệu sinh học khác được thực hiện thông qua trang web của NCBI. Các tài nguyên NCBI bao gồm Entrez, tiện ích lập trình Entrez, MyNCBI, PubMed, PubMed Central, Genee, trình duyệt phân loại NCBI, BLAST, Liên kết BLAST (BLink), Primer-BLAST, COBALT, Splign, RefSeq, UniGenee, HomoloGenee, ProtEST, dbMHC, dbSNP, dbVar, Epigeneomics, Cơ quan đăng ký kiểm tra di truyền, bộ gene và các công cụ liên quan, trình xem bản đồ, trình tạo mô hình, trình xem bằng chứng, lưu trữ theo dõi, lưu trữ đọc trình tự, dự án sinh học, mẫu sinh học, công cụ định kiểu gene retrovirus, cơ sở dữ liệu tương tác protein HIV-1/người, biểu hiện gene Omnibus, thăm dò, di truyền Menden trực tuyến ở động vật, cơ sở dữ liệu mô hình phân tử, cơ sở dữ liệu miền được bảo tồn, công cụ truy xuất cấu trúc miền bảo tồn, hệ thống sinh học, các cụm protein và cơ sở dữ liệu phân tử nhỏ. Nhiều ứng dụng web bổ sung cho các chương trình BLAST được tối ưu hóa để tìm kiếm các dữ liệu chuyên biệt. Tất cả các tài nguyên này có thể được truy cập thông qua trang chủ của NCBI.

1.4.3.2. KEGG (Kyoto Encyclopedia of Genees and Geneomes) [99]

KEGG là CSDL tích hợp gồm 16 CSDL được hiển thị bằng mã màu của các trang web và phân loại thành CSDL về chức năng của các cấu trúc sinh học như tế bào, sinh vật và hệ sinh thái, các thông tin từ cấp bộ gene và phân tử. Thông tin bộ gene được lưu trữ trong CSDL GENEES là tập hợp tất cả các bộ gene được giải trình tự hoàn toàn và một số bộ gene được giải trình tự một phần với chú giải cập nhật về các chức năng của gene. Thông tin chức năng bậc cao được lưu trữ trong CSDL PATHWAY, CSDL này chứa các biểu diễn đồ họa của các quá trình trong tế bào như trao đổi chất, vận chuyển màng, truyền tín hiệu và chu kỳ tế bào. Cơ sở dữ liệu PATHWAY được bổ sung bởi một tập hợp các bảng nhóm sinh vật nhân sơ/nhân chuẩn đơn bào cho thông tin về các con đường được bảo tồn, thường được mã hóa bởi các gene liên kết vị trí trên nhiễm sắc thể và đặc biệt hữu ích trong việc dự đoán các chức năng của gene. CSDL thứ ba trong KEGG là LIGAND cho thông tin về các chất hóa học, phân tử enzyme và phản ứng enzyme. KEGG cung cấp các công cụ đồ họa Java để duyệt bản đồ gene, so sánh hai bản đồ gene và thao tác trên bản đồ biểu

hiện, cũng như các công cụ tính toán để so sánh trình tự, so sánh đồ thị và tính toán đường dẫn. CSDL KEGG được cập nhật hàng ngày và được cung cấp miễn phí (http://www.geneome.ad.jp/kegg/). KEGG cũng chứa các thông tin về sức khỏe như các bệnh, thuốc cũng như những các sản phẩm sinh học khác. CSDL KEGG đã được Phòng thí nghiệm Kanehisa thuộc Đại học Kyoto phát triển từ năm 1995 và hiện là CSDL tham chiếu nổi bật để tích hợp và giải thích các dữ liệu phân tử quy mô lớn được tạo ra bằng giải trình tự bộ gene thông lượng cao.

1.4.3.3. Pfam (Protein families database)

Pfam là CSDL về các họ protein được sử dụng rộng rãi, chứa 14.831 họ được xếp theo cách thủ công trong phiên bản 27.0. Trong những năm gần đây, số lượng họ đã tăng lên 17.929 họ trong phiên bản 32.0 và CSDL này liên tục được cải thiện. Mỗi họ protein được xác định qua 2 trình tự và một mô hình đại diện HMM. Mô hình HMM là mô hình xác suất, được xây dựng từ một tập hợp các trình tự có các đoạn đặc trưng cho họ protein. Việc xây dựng mô hình đặc trưng này rất cần thiết vì nó cung cấp nền tảng cho các hiểu biết về các amino axit đặc biệt, khoảng trống và độ dài trong mô hình HMM. Trong Pfam, mô hình HMM được tìm kiếm dựa trên một tập hợp chuỗi lớn UniProt Knowledgebase (UniProtKB) [100] để tìm tất cả các đặc trưng cho họ protein. Các vùng trình tự đạt điểm cao hơn ngưỡng được cho là đặc trưng cho họ protein. Mô hình đại diện HMM được xây dựng và tìm kiếm bằng phần mềm HMMER (http://hmmer.janelia.org) [101]. Dữ liệu pfam có sẵn ở nhiều định dạng bao gồm tệp (lấy từ cơ sở dữ liệu MySQL) và bảng tương quan, cả hai đều có thể được tải xuống từ trang FTP (ftp://ftp.sanger.ac.uk/pub/cơ sở dữ liệu/Pfam). Trang web Pfam (có tại hệ thống máy chủ ở Anh http://pfam.sanger.ac.uk/, hệ thống máy chủ ở Mỹ http://pfam.janelia.org và Thụy Điển (http://pfam.sbc.su.se/)) cung cấp các cách khác nhau để truy cập nội dung CSDL, cung cấp biểu diễn đồ họa và quyền truy cập, tương tác vào dữ liệu.

CHƯƠNG 2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Vật liệu, hóa chất

2.1.1. Đối tượng nghiên cứu

Các mẫu đất mùn xung quanh nấm mục trắng thủy phân mạnh thân cây gỗ trong Vườn Quốc gia Cúc Phương có vị trí địa lý GPS 20.27776; 105.71137. Các mẫu đất này được lấy vào mùa mưa (tháng 5 – 6 trong năm 2017) trong bán kính 10 km. Nhiệt độ trung bình ở Cúc Phương là 20,6°C, độ ẩm và lượng mưa hàng năm lần lượt là 90% và 2138 mm. Đây là khu bảo tồn thiên nhiên lớn nhất và có độ đa dạng sinh học cao ở Việt Nam. 45 mẫu đất mùn (mỗi mẫu lấy khoảng 100 g) xung quanh khu vực có nấm mục trắng phân hủy gỗ đã được thu thập (Hình 2.1). Giá trị pH của các mẫu đất này dao động trong khoảng 6,9 – 7,3. Các mẫu đất mùn được bảo quản trong hộp đá ở 4°C và chuyển về phòng thí nghiệm.

Hình 2.1. Các vị trí mẫu đất mùn xung quanh các khu nấm mục trắng khác nhau được thu thập

2.1.2. Địa điểm nghiên cứu

Phòng Kỹ thuật di truyền, Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

2.1.3. Các chủng vi sinh vật, plasmid và cặp mồi sử dụng trong nghiên cứu

- Các chủng vi sinh vật: chủng *E. coli* DH10B (*F-mcrA* Δ(*mrr-hsd*RMS-*mer*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*Al *end*Al *ara*D139 Δ(*ara leu*) 7697 *gal*U *gal*K

*rps*L *nup*G λ-) của hãng Invitrogene (Mỹ) được sử dụng cho thí nghiệm tách dòng gene.

Các chủng *E. coli* BL21 (DE43) (F^- *omp*T *hsd* S_B (r_B^- m_B^-) *gal dcm* (DE3), Rosetta 1 (F^- *omp*T *hsd* S_B (r_B^- m_B^-) *gal dcm* (DE3) pRARE (Am^R), JM109 (DE3) (*end*A1 *rec*A1 *gyr*A96 *thi hsd*R17 (rk^- mk^+) *rel*A1 *sup*E44 λ− Δ(*lac-pro*AB) F′ *tra*D36 *pro*AB *lac*IqZΔM15 λDE3, *E. coli* C43 (DE3) (F- *omp*T *gal dcm hsd*SB (r_B^- m_B^-) (DE3), Soluble (DE3) (F^- *omp*T *hsd*S_B (r_B^- m_B^-) *gal dcm* (DE3)† (Lucigene), nhận từ phòng thí nghiệm Hóa sinh, Đại học Tổng hợp Saarland (CHLB Đức) được sử dụng làm chủng biểu hiện trong thí nghiệm biểu hiện gene.

- Plasmid: Vector pET22b(+) của hãng Novagenee (Mỹ) được sử dụng làm vector biểu hiện gene *gh3s2* trong các chủng biểu hiện *E. coli*. Vector pET22b(+) có chiều dài 5493 bp và có chứa đầy đủ các thành phần cần thiết phục vụ cho mục đích tách dòng và biểu hiện như: Trình tự khởi đầu sao chép (ori), gene chỉ thị chọn lọc (gene kháng kháng sinh Ampicillin - am^R), gene quy định chuỗi tín hiệu tiết pelB đầu N của protein có tác dụng hướng protein ngoại lai tổng hợp ngoài tế bào tế bào, vùng đa nối MCS (Multiple Cloning Site) có chứa điểm cắt của một số enzyme cắt hạn chế, phía sau vị trí này còn có một đoạn trình tự mã hóa cho 6 amino acid Histidine (đuôi His-tag) để thuận lợi cho quá trình tinh sạch protein bằng cột sắc kí ái lực, promoter T7 kiểm soát quá trình phiên mã, genee *lacI* mã hóa protein ức chế lac, *lacO* giúp điều hòa quá trình phiên mã.

- Cặp mồi khuếch đại gene *16S rDNA* của vi khuẩn:

27F: 5'-GAGTTTGATCCTGGCTCAG-3'

1527R: 5'-AGAAAGGAGGTGATCCAGCC-3'

2.1.4. Hóa chất và thiết bị

- Các hóa chất: Tris-HCl, sodium EDTA, sodium monohydrogene phosphate, sodium chloride, acrylamide/bis-acrylamide, APS, SDS, TEMED, ethidium bromide, methanol, ethanol, phenol, chloroform, isoamylalcohol, isopropanol, esculine, *p*NPG, isopropyl β- D- thiogalactopyranoside (IPTG), acetic acid, calcium chloride, sodium carbonate, potassium chloride, potassium acetate, 2-mercapto-ethanol, sodium hydroxide, bromophenol, glycine, imidazolee, disodium hydrogenee

phosphate dodecahydrate, sodium dihydrogenee phosphate dihydrate, nickel (II) chloride và một số hóa chất thông thường trong phòng thí nghiệm sinh học phân tử.

- Các enzyme được sử dụng: Enzyme cắt hạn chế *Nco*I và *Xho*I (Thermo Sciencetific, Mỹ), enzyme cellulase 0,05U (Sigma, Mỹ);

- Máy móc, thiết bị: Máy giải trình tự thế hệ mới HiSeq 2500 (Illumina HiSeq, San Diego, Mỹ), máy đọc ELISA ELx800 (BioTek, Mỹ), máy Nanophotometer P330 (Implen, Đức), máy Qubit™ 4 fluorometer (Thermo Fisher Scientific, Mỹ), máy PCR (Applied Biosystems, Mỹ), tủ nuôi cấy tế bào, máy lắc tế bào (Multitron, Đức), tủ lắc ổn nhiệt (New Jersey, Mỹ), máy li tâm nhỏ, máy li tâm lớn (Sorvall RC5B, Mỹ), bể ổn nhiệt (Mỹ), bể điện di, thiết bị biến tính protein, máy đo UV (Bio-Rad, Mỹ), cân điện tử, cân phân tích (Precisa, Thụy Sĩ), máy đo pH (Hana Instrument, Mỹ), máy hút chân không speed Vac Sc 110 (Savant, Mỹ), cột sắc kí ái lực Hitrap (Healthcare, Thụy Điển), tủ lạnh sâu -80ºC(Panasonic, Nhật), máy NanoDrop (Implen, Đức), Máy quang phổ UV-VIS 1650 (Shimadzu, Nhật Bản).

2.1.5. Môi trường nuôi cấy và một số dung dịch được sử dụng

2.1.5.1. Môi trường nuôi cấy

* Các thành phần môi trường nuôi cấy: Cao nấm men, bacto peptone, potassium monohydrogene phosphate, potassium dihydrogene phosphate (Merck, Đức), agar (Himedia, Ấn Độ), glucose, glycerol, sodium chloride (GH tech, Trung Quốc).

- Môi trường LBA lỏng: 0,5% cao nấm men; 1% bacto peptone; 1% NaCl hòa tan với nước cất một lần bổ sung ampicillin đến nồng độ cuối cùng là 100 µg/ml.

- Môi trường LBA đặc: Môi trường LB lỏng bổ sung thêm 1,5% agar và được bổ sung ampicillin đến nồng độ cuối cùng là 100 µg/ml.

- Môi trường TB: 1,2% bacto peptone; 2,4% cao nấm men; 72 mM K_2HPO_4; 17 mM KH_2PO_4; 0,4% glycerol; sau đó bổ sung ampicillin đến nồng độ cuối cùng là 100 µg/ml.

- Môi trường TB cải biến: 1,2% bacto peptone; 2,4% cao nấm men; 72 mM K_2HPO_4; 17 mM KH_2PO_4; 0,24% glucose; sau đó bổ sung ampicillin đến nồng độ cuối cùng là 100 µg/ml.

- Môi trường SB: 3,2% bacto peptone; 2% cao nấm men; 0,5% NaCl; sau đó bổ sung ampicillin đến nồng độ cuối cùng là 100 µg/ml.

- Môi trường PE: 1% cao nấm men; 2% bacto peptone; sau đó bổ sung ampicillin đến nồng độ cuối cùng là 100 µg/ml.

2.1.5.2. Một số dung dịch được sử dụng

- Các dung dịch tách chiết DNA plasmid từ *E. coli*: gồm dung dịch I (50 mM glucose; 25 mM Tris-HCI, pH 8,0; 10 mM EDTA, pH 8,0), dung dịch II (0,2 N sodium hydroxide; 1% SDS), dung dịch III (3 M potassium acetate; 11,5% acetic acid). Dung dịch phenol: chloroform: isoamylalcohol tỷ lệ theo thể tích tương ứng 25: 24: 1.

- Dung dịch sử dụng trong điện di DNA gồm dung dịch TAE 50 lần (24,2 g Tris-base; 5,71 ml acetic acid; 10 ml EDTA 0,5 M, pH 8,0: bổ sung nước đến 100 ml). Dung dịch nhuộm gel ethidium bromide (EtBr) 0,5 g/ml. Dung dịch sử dụng trong điện di protein gồm đệm xử lý mẫu protein 6 lần (7 ml Tris-HCl 1 M, pH 6,8; 3 ml glycerol 100%; 1 g SDS; 0,6 ml 2-mercapto-ethanol; 1,2 mg bromophenol). Đệm chạy điện di protein (0,05 M Tris; 0,192 M glycine; 0,1% SDS; pH 8,4). Dung dịch coomassie (coomassie brilliant blue 0,1% w/v; methanol 30% v/v; acetic acid 10% v/v). Dung dịch tẩy chất nhuộm coomassie (methanol 40% v/v: acetic acid 10% v/v).

- Dung dịch tinh chế protein gồm dung dịch cân bằng cột là đệm PBS 50 mM pH7 không chứa NaCl (gồm 0,45 mM KCl; 1,67 mM Na_2HPO_4; 0,3 mM KH_2PO_4), dung dịch rửa mẫu (gồm 0,45 mM KCl; 1,67 mM Na_2HPO_4; 0,3 mM KH_2PO_4; bổ sung 20 mM và 50 mM imidazolee) và dung dịch thu mẫu (gồm 0,45 mM KCl; 1,67 mM Na_2HPO_4; 0,3 mM KH_2PO_4; bổ sung 300 mM imidazolee) và một số dung môi hữu cơ khác.

2.2. Phương pháp nghiên cứu

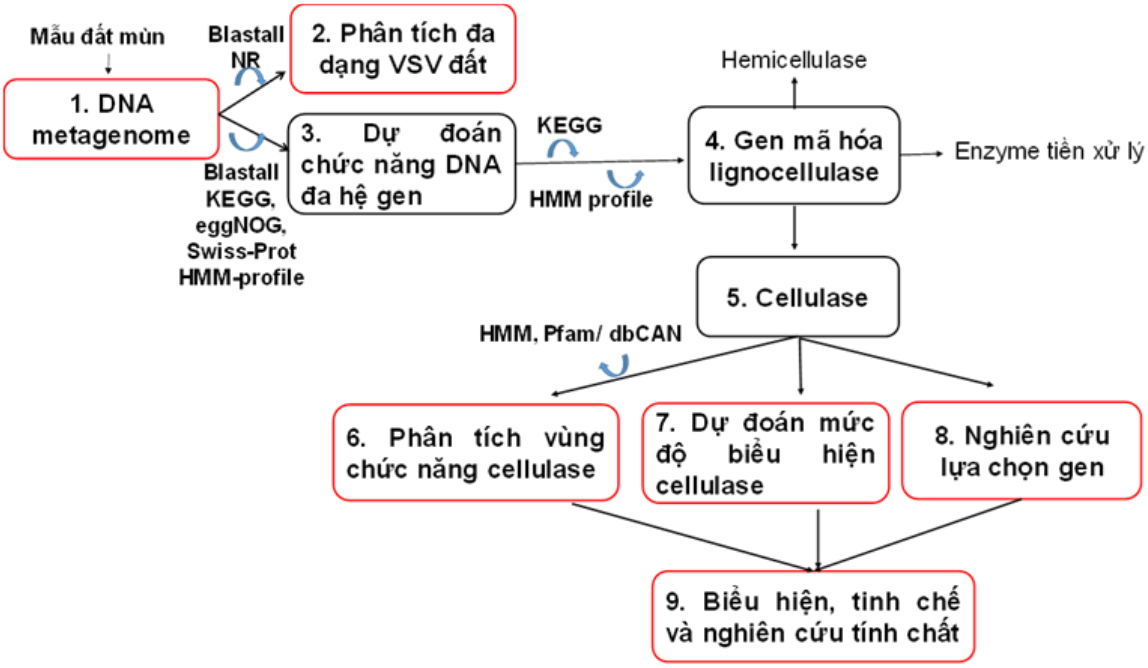
Đề tài được thực hiện theo các bước cơ bản như Hình 2.2.

2.2.1. Các phương pháp vi sinh và sinh học phân tử

2.2.1.1. Tách chiết DNA đa hệ gene của vi sinh vật đất

Các mẫu đất mùn đã thu thập được trộn đều, sau đó hòa vào đệm PBS 1X, pH 7,4 và tiến hành ly tâm phân pha với các tốc độ khác nhau để tách sơ bộ mẫu dùng

cho tách chiết DNA đa hệ gene. Mẫu được ly tâm 500 vòng/phút trong 10 phút để các tạp chất kích thước lớn có trong đất lắng xuống, thu dịch nổi và bỏ cặn. Sau đó ly tâm 2 lần, mỗi lần ly tâm 600 vòng/phút trong 10 phút để loại bỏ dần các tạp chất tồn tại trong đất. Tiếp tục ly tâm 5000 vòng/phút trong 1 phút để thu được mẫu dùng cho tách chiết DNA đa hệ gene. Hòa toàn bộ mẫu thu được với dung dịch PBS 1x pH 7,4 có bổ sung 20% glycerol. Mẫu được bảo quản tại -80^{o}C trước khi tiến hành tách chiết DNA đa hệ gene.



Hình 2.2. Sơ đồ quy trình nghiên cứu trong luận án

Mẫu dùng cho mỗi lần tách chiết DNA đa hệ gene là từ khoảng 10 g mẫu đất mùn ban đầu đã được xử lý theo các bước nêu trên, mẫu này được cho vào ống falcol 50 ml, bổ sung 20 ml đệm ly giải gồm 100 mM đệm Tris-HCl, 100 mM EDTA, 100 mM Na_2HPO_4, 1,5 M NaCl và 1% CTAB (tất cả các dung dịch này đều có pH 8,0) với 0,1 mg/ml protease K và được ủ ở 37°C trong 30 phút, có lắc nhẹ. Sau khi ủ, mẫu được xử lý với 3 ml 20% SDS và tiếp tục ủ ở 65°C trong 30 phút, thỉnh thoảng lắc nhẹ. Sau đó, phần dịch nổi phía trên được thu lại bằng cách ly tâm các mẫu ở tốc độ 7000 vòng/phút trong 5 phút, ở 4^{o}C và chuyển sang ống mới. Sau đó, phenol/chloroform/isoamyl alcohol (25:24:1 v/v) được thêm vào để tinh sạch các mẫu DNA. Lớp phía trên chứa DNA đa hệ gene được thu lại sau khi ly tâm ở tốc độ 6500 vòng/phút trong 10 phút ở 4°C. Mẫu DNA được kết tủa lại bằng cách bổ sung 6 ml

isopropanol, sau đó ly tâm ở tốc độ 13.000 vòng/phút trong 10 phút. DNA được rửa bằng ethanol lạnh 70%. Tủa DNA được làm khô trong máy speedvac và được hòa lại với 300 µl nước khử ion vô trùng. DNA đa hệ gene đã tách chiết được kiểm tra chất lượng dựa trên các tiêu chí: (1) mức độ đứt gãy của DNA đa hệ gene bằng điện di trên gel agarose 0,8%, (2) Nồng độ và độ tinh sạch của mẫu DNA được đo bằng máy đo Nanophotometer P330 (IMPLEN, Đức), (3) sự có mặt hay không của các chất ức chế polymerase trong mẫu dựa trên PCR khuếch đại gene 16S rDNA (vì DNA được giải trình tự bằng phương pháp tổng hợp). DNA đa hệ gene từ ba lần tách chiết được trộn vào với nhau và khoảng 100 µg mẫu DNA tổng số này đã được gửi đến BGI-Hong Kong Co. Ltd. để giải trình tự metageneome.

2.2.1.2. Giải trình tự DNA đa hệ gene bằng máy HiSeq2500 của Illumina

Kỹ thuật giải trình tự DNA đa hệ gene thông lượng cao được chia làm 3 giai đoạn: tạo thư viện NGS (Next Geneeration Sequencing), tạo nhóm DNA và giải trình tự DNA bằng phương pháp tổng hợp trên hệ thống Hiseq Illumina 2500 do công ty BGI, Trung Quốc thực hiện.

2.2.1.3. Biến nạp DNA plasmid vào tế bào chủ E. coli

- Quy trình tạo tế bào *E. coli* khả biến: Các chủng tế bào *E. coli* DH10B, BL21, Rosetta 1, JM109, Soluble, C43 khả biến được tạo ra theo phương pháp của Sambrook và cộng sự (2001) [102]. Theo đó, tế bào vi khuẩn được xử lý lạnh, ủ với 100 mM $CaCl_2$ ở các thể tích khác nhau. Cuối cùng, tế bào được hòa vào 100 mM $CaCl_2$ có bổ sung glycerol vô trùng để đạt nồng độ 15%, bảo quản ở -80°C trong các ống eppendorf.

- Quy trình biến nạp DNA plasmid vào vi khuẩn *E. coli*: Phương pháp sốc nhiệt đã được sử dụng để biến nạp DNA plasmid vào vi khuẩn *E. coli* [103]. Tế bào *E. coli* khả biến lấy ra từ -80°C được bảo quản trong đá 30 phút rồi bổ sung DNA plasmid và ủ mẫu trong đá khoảng 30 phút. Sau đó, mẫu được sốc nhiệt ở 42°C trong 1 phút 30 giây và ủ lại 4°C trong 2 phút. Mẫu được nuôi và cấy trải trên đĩa môi trường LB đặc có bổ sung ampicilin 100 µg/ml và ủ ở 37°C qua đêm.

2.2.1.4. Tách chiết DNA plasmid từ tế bào E. coli

DNA plasmid trong tế bào *E. coli* có kích thước và khối lượng nhỏ hơn nhiều so với DNA nhiễm sắc thể, vì vậy DNA plasmurid có thể được tách ra dưới dạng vòng

đóng. Việc tách chiết DNA plasmid này được thực hiện theo phương pháp của Sambrook và cộng sự [103]. Phương pháp này về cơ bản là các tế bào vi khuẩn nuôi cấy sẽ được hòa tan bằng các dung dịch Sol I, II, III trong điều kiện lạnh để làm tan các thành phần cấu tạo của tế bào, sau đó mẫu tế bào tiếp tục được hòa với dung dịch loại protein, ly tâm để thu pha lỏng ở phía trên chứa DNA plasmid. DNA plasmid được làm sạch, hòa tan trở lại và bảo quản ở -20°C, điện di kiểm tra trên gel agarose 0,8%.

2.2.1.5. Cắt kiểm tra DNA plasmid bằng enzyme cắt hạn chế

Để kiểm tra DNA plasmid tái tổ hợp, hai loại enzyme cắt hạn chế là *Xho*I và *Nco*I (2 U/µl) được sử dụng. Có hai phản ứng cắt DNA plasmid, mỗi phản ứng có tổng thể tích là 10 µl bao gồm 3 µl DNA plasmid (hàm lượng 10 µg/ml), 2 µl đệm tango 2X, phản ứng 1 bổ sung 0,3 µl enzyme hạn chế *Xho*I (2 U/µl) còn lại là nước cất, phản ứng 2 được bổ sung 0,3 µl *Xho*I và 0,3 µl *Nco*I (2 U/µl) còn lại là nước cất. Hỗn hợp các thành phần được trộn đều, ủ ở nhiệt độ 37°C qua đêm. Sản phẩm của phản ứng cắt sẽ được điện di trên gel agarose 0,8% để kiểm tra.

2.2.1.6. Điện di DNA trên gel agarose

Điện di DNA trên gene agarose 0,8% được sử dụng để kiểm tra kích thước đoạn DNA plasmid và các sản phẩm sau khi cắt bằng enzyme cắt hạn chế là là DNA đích và plasmid. Phương pháp điện di này được thực hiện theo Sambrook và cộng sự [103].

2.2.2. Các phương pháp hóa sinh protein

2.2.2.1. Phương pháp biểu hiện gene gh3s2

Các chủng tế bào *E. coli* mang plasmid pET22b(+)gh3s2 tái tổ hợp được cấy chuyển vào 5 ml môi trường LBA, nuôi lắc 200 vòng/phút ở 37°C qua đêm. Sau đó, các dịch tế bào đó được chuyển sang môi trường LB có bổ sung 100 µg/ml ampicillin mới sao cho OD_{600} đạt 0,1 và tiếp tục lắc 200 vòng/phút ở 37°C cho đến khi OD_{600} đạt đến giá trị phù hợp cho biểu hiện gene là 0,6. Lúc này, chất cảm ứng cho biểu hiện gene là isopropyl β- D- thiogalactopyranoside (IPTG) được bổ sung để đạt nồng độ cuối cùng là 0,5 mM IPTG, nhiệt độ lên men được sử dụng là 30°C, nuôi lắc 200 vòng/phút trong 4 giờ. Sau khi lên men, dịch nuôi cấy tế bào sẽ được li tâm tốc độ 5000 vòng/phút, trong 5 phút để thu các tế bào.

Sau khi lựa chọn được chủng tế bào để biểu hiện gene, nhằm thu được sản phẩm protein có hàm lượng cao và hoạt tính sinh học tốt, các điều kiện ảnh hưởng đến hiệu quả của quá trình lên men được tối ưu bao gồm: nhiệt độ biểu hiện được khảo sát là 18°C, 20°C, 25°C, 30°C, 37°C; thành phần của các môi trường nuôi cấy gồm có 5 môi trường: LB, TB, TB cải biến, SB và PE. Thêm vào đó, các nồng độ chất cảm ứng IPTG 0,05 mM, 0,1mM, 0,3 mM, 0,5 mM, 1mM, 1,2 mM, 1,5 mM; thời điểm cảm ứng khi mật độ tế bào đạt giá trị OD 0,4; 0,8; 1,0; 1,5; 3,0; 4,0 và thời điểm thu mẫu tối ưu sau khi cảm ứng 1, 2, 3, 4, 5, 6 và 22 giờ cũng đã được khảo sát.

2.2.2.2. Phương pháp tách chiết protein tái tổ hợp từ E. coli

Sau khi lên men, tế bào được hòa tan trở lại trong đệm 20 mM Tris HCl, pH=8 đến giá trị OD_{600} là 10. Để kiểm tra sự biểu hiện của protein, tế bào sau khi lên men được siêu âm để phá vỡ tế bào với cường độ 65% công lực, 3 giây "bật", 3 giây "tắt" trong 10 phút. Sau khi siêu âm, dịch protein tổng số được phân pha tan và pha không tan bằng li tâm lạnh ở 4^{o}C với tốc độ 12000 vòng/phút trong 10 phút. Phần dịch nổi phía trên được thu lại sang ống efpendorf khác, phần không tan lắng xuống sẽ được được hòa tan trở lại bằng đệm 20 mM Tris HCl, pH=8 với thể tích tương đương. Các mẫu protein tổng số, pha tan và pha không tan được kiểm tra lại bằng điện di biến tính trên gel polyacrylamide 12,5%.

2.2.2.3. Điện di biến tính protein trên gel polyacrylamide-SDS

- Chuẩn bị gel: Hai loại gel polyacrylamide được chuẩn bị với thành phần và nồng độ như mô tả trong Bảng 2.1. Các thành phần được bổ sung theo thứ tự và đảo đều trước khi cho vào giá đổ bản gel. Gel tách được chuẩn bị trước cho đến khi đông hoàn toàn mới chuẩn bị tiếp lớp gel cô. Bản gel được ổn định sau khi gel cô được chuẩn bị khoảng 30 phút.

Bảng 2.1. Thành phần gel polyacrylamide

Thành phần	**Gel tách (12,5%)**	**Gel cô (5%)**
dH_2O	0,55 ml	0,45 ml
Tris-HCl 6,8	-	0,2 ml
Tris-HCl 8,8	1,125 ml	-
Glycerol 50%	0,9 ml	-
Acrylamide 30%	1,89 ml	0,14 ml
SDS 10%	45 µl	4 µl
APS 10%	30 µl	8 µl
TEMED	3 µl	1 µl

Tổng	**4,543 ml**	**0,803 ml**

- Quy trình điện di:

Mẫu protein tổng số, protein ở pha tan và pha tủa được xử lý bằng đệm xử lý mẫu (sample buffer 6X) và ủ ở 95°C trong 10 phút. Sau đó, mẫu được cho vào giếng và chạy điện di với cường độ dòng điện 10 mA cho mỗi bản gel cho đến khi mẫu qua hết lớp gel cô. Sau đó, cường độ dòng điện được tăng lên 20 mA cho mỗi bản gel khi mẫu đến lớp gel tách. Kết thúc điện di, gel được nhuộm Coomassie Brilliant Blue R250. Sau khi ủ với thuốc nhuộm, gel được rửa sạch bằng dung dịch tẩy nhuộm cho đến khi quan sát được rõ ràng các băng protein.

2.2.2.4. Tinh sạch protein GH3S2 bằng sắc ký ái lực His-tag [104]

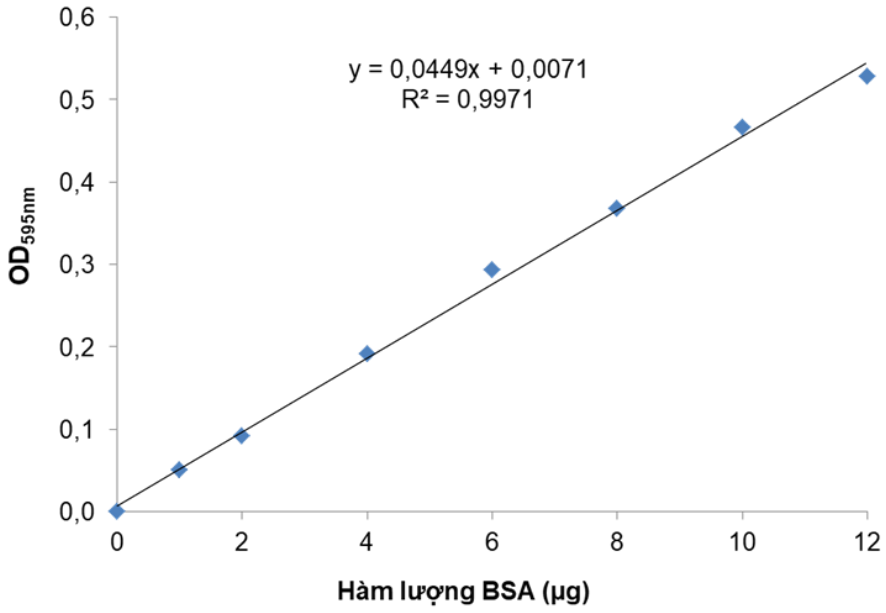
Mẫu tế bào sau khi biểu hiện được hòa lại trong nước để đưa về OD_{600} = 10, lấy 15 ml mẫu tế bào này vào ống falcon loại 50 ml. Mẫu tế bào này được phá vỡ bằng sóng siêu âm power 65%, chu kỳ 3 giây on, 3 giây off trong 10 phút sau đó bổ sung 300 mM PBS (không chứa NaCl) để đưa về nồng độ cuối cùng là 50 mM PBS (không chứa NaCl) pH 7,0. Dịch tế bào được ly tâm lạnh ở tốc độ 8000 vòng/ phút trong 10 phút, 4°C để thu pha tan chứa protein GH3S2, dịch này được sử dụng để tinh chế enzyme. Cột sắc kí ái lực His-tag được rửa và cân bằng cột bằng đệm 50 mM PBS (không chứa NaCl) pH 7,0. Sau đó, bơm 15 ml dịch protein pha tan đã chuẩn bị lên cột histrap 5 ml với tốc độ chậm khoảng 1 ml/phút. Dịch thu được trong lúc đưa protein lên cột (F) sẽ được thu lại để kiểm tra mức độ bám của protein GH3S2. Sau đó, rửa các protein tạp lần lượt bằng 5 thể tích cột (5 CV) đệm 50 mM PBS pH 7,0 có chứa imidazole các nồng độ là 20 mM và 50 mM. Dịch rửa được thu lại để kiểm tra các protein đi ra khỏi cột (W1, W2). Enzyme GH3S2 được thu lại với đệm 50 mM PBS pH 7,0 có chứa imidazole 300 mM vào các ống eppendorf (1 ml/ống). Cột được rửa hết các protein bằng đệm 50 mM PBS pH 7,0 chứa imidazole 500 mM và cân bằng cột trở lại bằng 25 ml đệm 50 mM PBS. Các phân đoạn chứa enzyme GH3S2 sẽ được gộp lại, bổ sung glycerol đến nồng độ cuối cùng 10%. Mẫu protein đã tinh sạch thu được tiếp tục được loại bỏ muối bằng thẩm tích lạnh qua đêm trong túi thẩm tích 10 kDa (Thermo Scientific, Mỹ) trong đệm 50 mM PBS, pH 7,0, glycerol 10%. Hàm lượng protein có trong mẫu được xác định bằng phương pháp Bradford [105]. Sau đó, protein được điện di trên gel polyacrylamide với các hàm lượng khác nhau trong các giếng điện di kết hợp với phần mềm Image Lab v6.1.0 build 7

(https://www.bio-rad.com/en-vn/product/image-lab-software) để kiểm tra độ tinh sạch của protein. Sản phẩm protein sau khi loại muối và đạt được độ sạch theo yêu cầu sẽ tiếp tục được sử dụng cho các thí nghiệm xác định hoạt tính và nghiên cứu tính chất, đặc điểm của enzyme.

2.2.2.5. Xác định độ sạch của protein GH3S2 sau tinh chế

Protein sau khi được biểu hiện và tinh chế có thể được đánh giá độ sạch tương đối bằng phần mềm Image Lab. Theo đó, mẫu protein đã tinh sạch và loại muối được điện di trên gel polyacrylamide với các hàm lượng protein ở các giếng khác nhau với thang chuẩn protein. Sau khi điện di, bản gel được nhuộm bằng thuốc nhuộm comassie, rửa sạch nhiều lần cho đến khi băng hiện rõ nét thì đưa lên máy scan để quét. Chế độ quét được lựa chọn sao cho ảnh đạt được chất lượng tốt nhất. Ảnh quét được chuyển về chế độ đen trắng và đưa vào phần mềm Image Lab version 6.1.0 build 7 (https://www.bio-rad.com/en-vn/product/image-lab-software) để phân tích làn và định lượng tương đối hàm lượng protein. Phần mềm sẽ nhận biết và quét để định lượng tương đối protein tổng số trên giếng dựa trên những vùng xác định thấy có băng protein. Các protein GH3S2 được xác định bằng mức độ đậm của băng tương ứng. Từ đó, độ sạch của protein đích được xác định chính là tỷ lệ giữa mức độ đậm của băng GH3S2 so với toàn bộ các băng protein ở mỗi đường chạy.

2.2.2.6. Xác định hàm lượng protein bằng phương pháp Bradford [105]



Hình 2.3. Đường chuẩn BSA được đo OD ở bước sóng 595 nm

- Xây dựng đường chuẩn: Trước khi xác định hàm lượng protein trong mẫu thu được cần xây dựng đường chuẩn BSA. Đường chuẩn sẽ được thiết lập bằng BSA với 8 giá trị từ 0 đến 12 µg BSA được pha với nước deion vô trùng (Hình 2.3). Mỗi ống thí nghiệm có tổng thể tích là 1000 µl bao gồm 800 µl BSA đã được pha loãng bằng nước deion vô trùng (có hàm lượng khác nhau từ 0 đến 12 µg) sau đó bổ sung 200 µl Bradford 5X, trộn đều mẫu bằng máy vortex và để ở nhiệt độ phòng trong 5 phút. Mẫu được đo OD ở bước sóng 595 nm, kết quả này sẽ được sử dụng để thiết lập đường chuẩn thể hiện mối tương quan giữa OD_{595} và nồng độ BSA.

- Xác định hàm lượng protein trong mẫu: 800 µl mẫu được pha loãng trong đệm 50 mM PBS, pH 7,0 ở các nồng độ khác nhau + 200 µl Bradford 5X, trộn đều bằng máy vortex và để ở nhiệt độ phòng trong 5 phút. Tiến hành đo OD_{595} tương tự như phần thiết lập đường chuẩn, dựa vào phương trình thể hiện mối tương quan giữa OD_{595} và hàm lượng BSA để tính hàm lượng protein trong mẫu.

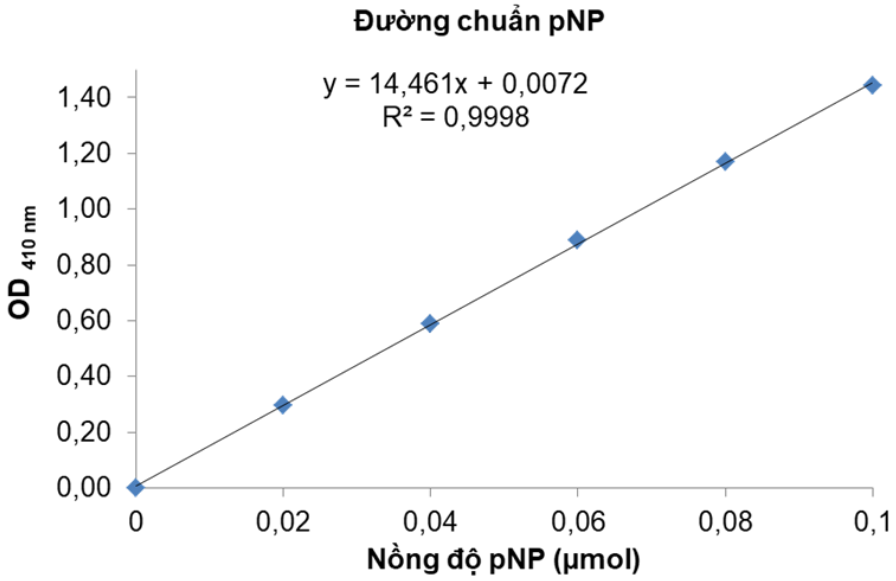
2.2.2.7. Xác định hoạt tính của enzyme β-glucosidase

* Khảo sát hoạt tính β-glucosidase: β-glucosidase có khả năng phân cắt esculin thành glucose và esculetin. Trong môi trường có sắt, ion sắt bị esculetin khử tạo màu nâu hoặc nâu thẫm. Thí nghiệm kiểm tra khả năng thủy phân esculin của protein tái tổ hợp được thực hiện dựa theo phương pháp của Veena và đồng tác giả [106]. Đĩa thạch môi trường LBA có bổ sung thêm esculin (3 g/l) và ferric ammonium citrate (0,2 g/l) được chuẩn bị sẵn. Sau đó, dùng dụng cụ tạo các giếng có đường kính khoảng 0,5 cm trên đĩa thạch. Hút 50 µl protein tổng số pha tan nhỏ vào một giếng. Sau đó, đĩa được ủ ở 37°C trong khoảng 16-20 giờ. Đối chứng âm là 50 µl đệm PBS 50 mM, pH 7,0 và đối chứng dương là 50 µl cellulase (Sigma, 0,05 U).

* Xác định hoạt độ β-glucosidase [107]

- Xây dựng đường chuẩn: Pha loãng chất chuẩn pNP (code 1048, Sigma) từ nồng độ 0,1 µmol/ml bằng đệm 50 mM PBS pH 7,0 về các nồng độ từ 0 đến 0,1 µmol trong tổng thể tích 200 µl (mỗi nồng độ lặp lại 3 lần) (Hình 2.4). Sau đó bổ sung 800 µl 0,2 M Na_2CO_3 vào mỗi ống, trộn đều. Các ống thí nghiệm được đo OD ở bước sóng 410 nm. Kết quả này được sử dụng để xây dựng đường chuẩn thể hiện mối tương quan giữa OD_{410} và nồng độ pNP. Phương trình đường chuẩn: $y = 14{,}461x + 0.0072$ ($R^2 = 0{,}9998$), trong đó x là µmol pNP, y là giá trị OD_{410}.

- Xác định hoạt tính của enzyme GH3S2: Hoạt tính β-glucosidase của GH3S2 được xác định dựa vào khả năng thủy phân cơ chất được sử dụng phổ biến là *p*-nitrophenol-β-glucoside (*p*NPG), giải phóng p-nitrophenol (*p*NP) [8]. Một đơn vị hoạt tính của β-glucosidase là lượng enzyme cần thiết để xúc tác cho phản ứng giải phóng ra 1 µmol *p*NP trong thời gian 1 phút [8]. Hoạt tính của enzyme GH3S2 được xác định bằng cách: lấy 20 µl enzyme tổng số pha tan (được pha loãng 12-30 lần trong đệm 50mM PBS pH 7,0) trộn với 180 µl 5 mM *p*NPG. Ống đối chứng có thành phần tương tự ống phản ứng chỉ thay 20 µl enzyme bằng 20 µl đệm 50mM PBS pH 7,0. Phản ứng được ủ ở 37°C thời gian 15 phút, sau đó dừng phản ứng bằng cách bổ sung 800 µl 0,2 M Na_2CO_3 rồi trộn đều. Mẫu này được đo OD_{410}, dựa vào đường chuẩn thể hiện mối quan hệ giữa OD_{410} và nồng độ *p*NP để tính lượng pNP tạo ra.



Hình 2.4 Đường chuẩn pNP được đo OD ở bước sóng 410 nm

2.2.2.8. Xác định ảnh hưởng của nhiệt độ, pH, các ion kim loại và glucose lên hoạt tính của GH3S2

Để xác định ảnh hưởng của nhiệt độ đến hoạt tính của enzyme GH3S2, phản ứng xác định hoạt tính của enzyme GH3S2 và cơ chất *p*NPG được ủ ở các điều kiện nhiệt độ 30°C, 35°C, 37°C, 40°C, 50°C trong thời gian 15 phút. Enzyme được hòa trong đệm 50 mM PBS có pH thay đổi pH 5,0; pH 5,5; pH 6,0; pH 6,5, pH 7,0; pH 8,0 và thực hiện các phản ứng xác định hoạt tính. Các ion kim loại Na^+, K^+, Ca^{2+}, Mg^{2+}, Ni^{2+}, Mn^{2+}, Fe^{2+}, Cu^{2+} được thêm vào phản ứng xác định hoạt tính của GH3S2 để đạt nồng độ ion kim loại cuối cùng là 1 mM. Để xác định ảnh hưởng của glucose

đến khả năng xúc tác của GH3S2, glucose được thêm vào phản ứng xác định hoạt tính đến nồng độ cuối cùng 2, 4, 6, 8, 10, 15, 20, 25, 30, 50. 100. 150, 200, 250, 300 mM trước khi thực hiện phản ứng.

2.2.2.9. Xác định độ bền của enzyme với nhiệt độ, pH

Độ bền của enzyme với nhiệt độ được xác định bằng cách enzyme được ủ ở các nhiệt độ 37°C, 40°C, 45°C, 50°C trong khoảng thời gian 1, 2, 3, 4, 6, 12 (giờ). Enzyme cũng được xác định độ bền với pH bằng cách pha enzyme trong đệm 50 mM PBS có pH khác nhau: 5,0; 6,0; 7,0; 8,0 trong thời gian 1, 2, 3, 4, 6, 12 (giờ). Sau đó, mẫu enzyme sẽ được lấy ra để xác định hoạt tính ở điều kiện nhiệt độ và pH tối ưu là 37°C và pH 6,0.

2.2.2.10. Xác định thông số động học của GH3S2

Hoạt tính của protein GH3S2 được xác định ở các điều kiện nhiệt độ và pH tối ưu 37°C và pH 6,0, nồng độ cơ chất *p*NPG thay đổi từ 1-10 mM. Hỗn hợp phản ứng gồm: 1 µg enzyme trong 20 µl đệm 50 mM PBS pH 6,0 được bổ sung 180 µl *p*NPG có nồng độ từ 1-10 mM được ủ ở 37°C thời gian15 phút. Sau đó, tiếp tục bổ sung 800 µl 0,2 M Na_2CO_3 để dừng phản ứng, mẫu được đo OD ở bước sóng 410 nm. Các kết quả này được sử dụng để thiết lập đồ thị thể hiện mối tương quan giữa tốc độ phản ứng với nồng độ cơ chất theo Linewever – Burk, trong đó V là số µmol *p*NP được giải phóng ra trong 1 phút, 1/[S] được tính là 1/[*p*NPG] với nồng độ *p*NPG được tính là mM, từ phương trình đó xác định được các thông số động học của GH3S2 là K_m, V_{max}.

2.2.3. Các phương pháp tin sinh học

2.2.3.1. Phân tích trình tự DNA đa hệ gene vi sinh vật

DNA đa hệ gene tách chiết từ mẫu đất mùn được giải trình tự bằng hệ thống Illumina HiSeq 2500 (Illumina HiSeq, San Diego, Mỹ) để thu được các dữ liệu thô. Dữ liệu này bao gồm hàng triệu read ngắn (1 read là 1 đoạn DNA được đọc trình tự). Trước hết các trình tự có chất lượng kém của dữ liệu thô được loại bỏ để thu được dữ liệu tinh nhờ công cụ SOAPnuke. Các trình tự có chất lượng kém là các read chứa ≥ 5% các base không rõ ràng, các read chứa trình tự adapter (mặc định là 15 base bao phủ bởi các read và adapter), các read chứa 50% base có chất lượng thấp (Q<20) trở lên. Sau đó, các dữ liệu đã được lọc sẽ được tập hợp *de novo* bằng hai phần mềm

IDBA (version 1.1.0) [108] https://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/ (ngày khai thác 03/08/2019) và phần mềm MEGAHIT (version 1.0) [109] https://github.com/voutcn/megahit (ngày khai thác 03/08/2019) với một loạt các kích thước k-mer khác nhau. Và kích thước k-mer phù hợp nhất sẽ được lựa chọn để tập hợp các dữ liệu tinh thành các contig. Các kết quả sắp xếp, lắp ráp này sẽ được kiểm tra lại bằng cách so sánh các contig thu được với các read ngắn tham gia cấu thành nó bằng phần mềm Bowtie 2 [110] với tham số "-p8-very-sensitive-local-k 100-score-min L,0,1.2". Sau đó, sử dụng phần mềm dự đoán gene MetaGeneeMark (phiên bản 2.10, với các tham số có sẵn của phần mềm) để dự đoán gene từ các contig đã lắp ráp [111]. Các gene dự đoán được phân nhóm bằng cách sử dụng CD-HIT [112] với ngưỡng tương đồng trình tự là 95% và ngưỡng bao phủ liên kết là 90% [113]. Các trình tự DNA đa hệ gene đã được đăng ký trên ngân hàng dữ liệu SRA (SRA-sequence read archive) với mã đăng ký PRJNA715592.

2.2.3.2. Phân tích đa dạng vi sinh vật nói chung và đa dạng các vi sinh vật mang gene mã hóa lignocellulase từ dữ liệu DNA đa hệ gene

Đơn vị phân loại của các gene được thực hiện bằng cách BLASTp với trình tự protein trong cơ sở dữ liệu NR (chứa các trình tự non-redundant cùng với các trình tự từ các dữ liệu ngân hàng khác như Refseq, PDB, SwissProt, PIR, PRF). Sau đó, file kết quả thu được sau khi BLASTp với NR sẽ được tiếp tục phân tích bằng phần mềm MEGAN (MetaGeneomic Analyser version 4.6) [112]. Phần mềm này đọc kết quả BLASTp như là thông tin đầu vào và xếp các gene vào các node trong thang phân loại của NCBI sử dụng thuật toán LCA (Least Common Ancestors) [112]. Thuật toán LCA căn cứ vào mức độ bảo thủ của trình tự gene để xếp các gene vào nhóm phân loại tương ứng. Thang phân loại của NCBI được thể hiện hình cây và kích thước của các node thể hiện số lượng các gene được xếp vào nhóm phân loại tương ứng. Các gene được phân loại đến cùng một mức và thuộc cùng một nhóm phân loại được tính tổng và kết quả phân loại được vẽ bằng công cụ bổ trợ Krona trong Excel.

2.2.3.3. Dự đoán chức năng của DNA đa hệ gene

Tất cả các gene đã được dự đoán sẽ được so sánh một số CSDL đáng tin cậy bao gồm SwissProt, KEGG (Kyoto Encyclopedia of Genees and Geneomes – phân loại chức năng theo con đường chuyển hóa) [114], EggNOG (Evolutionary

geneealogy of genee: Non-supervised Orthologous Groups, Version: 3.0 – phân loại chức năng theo tiến hóa của gene) [115] và Nr (Non-redundant protein sequence database) với giá trị e-value nhỏ hơn 10^{-5} [112], HHM - profile. Tổng hợp các kết quả so sánh này, trình tự protein nào tương đồng với nhiều CSDL khác nhau sẽ được chú giải về chức năng.

2.2.3.4. Khai thác gene mã hóa enzyme lignocellulase

Trong khuôn khổ của luận án này, nhằm dự đoán chức năng của các gene theo con đường chuyển hóa carbohydrate nên các kết quả thu được khi so sánh với CSDL KEGG sẽ tiếp tục được phân tích sâu hơn. So với các CSDL khác thì trong kết quả so sánh các gene với CSDL KEGG, các gene mã hóa lignocellulase được xác định nhiều hơn. Vì vậy, trước hết các gene mã hóa lignocellulase được nghiên cứu dựa trên kết quả xác định chức năng của KEGG và được phân loại dựa trên số EC (Enzyme Commission) [114]. Các gene mã hóa lignocellulase cũng được khai thác dựa trên mô hình đại diện HMM từ Pfam nhằm khai thác hiệu quả các enzyme.

2.2.3.5. Phân tích vùng chức năng của các gene mã hóa cellulase bằng PFAM và HMMER

Các ORF được xác định chức năng mã hóa cellulase dựa trên dữ liệu KEGG sẽ được xác định các vùng chức năng sử dụng CSDL Pfam và phần mềm HMMer từ dữ liệu dbCAN (https://www.ebi.ac.uk/Tools/hmmer/search/phmmer) [116]. Để dự đoán các vùng chức năng của ORF mã hóa cellulase dựa trên CSDL Pfam, các trình tự protein đích dưới dạng file fasta được cung cấp, lựa chọn sử dụng tham số e – value là 1.0 và xác nhận gửi lên thông qua trang web của HMMer, sau 3 – 4 ngày kết quả sẽ được trả về e-mail cá nhân của người gửi. Các số liệu thu được sẽ tiếp tục được xử lý bằng phần mềm Microsoft Excel.

2.2.3.6. Dự đoán mức độ biểu hiện của các ORF mã hóa cellulase trong tế bào E. coli

Mức độ biểu hiện của các ORF mã hóa cellulase trong hệ biểu hiện *E. coli* được dự đoán bằng phần mền Periscope (**Peri** plasmic expre**s**sion **c**lassifier for s**o**luble **p**rotein **e**xpression) truy cập miễn phí tại http://lightning.med.monash.edu/periscope/._Sau khi người dùng gửi trình tự các amino acid của protein, Periscope sẽ thực hiện dự đoán và trả về kết quả định lượng

tương đối mức độ biểu hiện của protein và lượng protein pha tan (là protein có thể có hoạt tính) trong *E. coli*. Periscope cho phép sử dụng tối đa 5 chuỗi polypeptide định dạng fasta trong mỗi lần gửi và không giới hạn độ dài chuỗi polypeptide truy vấn. Sau khi người dùng gửi các trình tự amino acid, Periscope dựa trên sự kết hợp của đoạn peptide tín hiệu và trình tự protein đích đã phân loại mức độ biểu hiện của protein dạng tan thành ba mức: cao, trung bình và thấp, ngoài ra còn có cả chức năng dự đoán về lượng protein dạng tan tính bằng đơn vị mg/l.

2.2.3.7. So sánh trình tự protein với trình tự trên NCBI bằng công cụ Blast

Công cụ BLASTp đã được sử dụng nhằm tìm kiếm các trình tự tương đồng trong CSDL NCBI với trình tự protein đang quan tâm về hai thông số là độ bao phủ và mức độ tương đồng. Để thực hiện việc tìm kiếm này, chúng tôi cung cấp trình tự amino acid của protein quan tâm định dạng fasta và so sánh với CSDL chuẩn của NCBI. BLASTp sẽ tiến hành tìm các vùng trên protein đích giống với các vùng trong CSDL và trả về kết quả các trình tự và mức độ tương đồng với chuỗi protein đang quan tâm trong 2 – 3 phút (100 trình tự có mức độ tương đồng cao nhất sẽ được hiển thị ở kết quả chính).

2.2.3.8. Dự đoán cấu trúc không gian và vị trí gắn cơ chất của enzyme

Để dự đoán cấu trúc bậc hai và cấu trúc bậc ba của các protein, phần mềm Phyre2 (http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index) đã được sử dụng. Người dùng gửi trình tự protein định dạng fasta, cung cấp thông tin e-mail cá nhân, sử dụng modelling mode bình thường để nhận được một kết quả duy nhất không bao gồm tất cả các trình tự trong protein trong khoảng thời gian ngắn. Phyre2 cho phép gửi từ 5 - 6 trình tự cùng một lúc ở phần tìm kiếm chuyên sâu trong chế độ chuyên gia và kết quả dự đoán cấu trúc điển hình sẽ được trả về e-mail người gửi trong 30 phút đến vài giờ. Đối với các trình tự khó dự đoán cấu trúc, phyre2 có thể chạy tự động và trả về kết quả sau 4 – 7 ngày. Các kết quả trả về gồm cấu trúc bậc hai, mô hình cấu trúc bậc ba của protein, thành phần miền và chất lượng mô hình của protein so với mô hình tham chiếu.

2.2.3.9. Dự đoán khả năng chịu axit hay kiềm của enzyme

Phần mềm AcalPred tại địa chỉ http://lin-group.cn/server/AcalPred đã được sử dụng để dự đoán khả năng chịu axit hay kiềm của protein. Người dùng nhập trình tự

protein đích vào ô tìm kiếm, phần mềm sẽ trả kết quả về khả năng ưa axit hay kiềm của protein trong vài phút. Phần mềm cho phép tìm kiếm tối đa mỗi lần 100 trình tự. Mỗi trình tự đưa vào sẽ thu được chỉ số chịu axit và chỉ số chịu kiềm. Nếu chỉ số chịu axit từ 0,5 – 1, chỉ số chịu kiềm từ 0 – 0,5 thì đó là enzyme chịu axit và ngược lại là enzyme chịu kiềm.

2.2.3.10. Dự đoán khả năng chịu nhiệt của enzyme

Phần mềm của TBI (http://www.tbi.org.tw/tools/) đã được sử dụng để dự đoán khả năng chịu nhiệt của enzyme. Độ bền nhiệt của enzyme được dự đoán ở 3 mức là: nếu chỉ số Tm >1 thì Tm dự đoán là trên 65^{o}C, nếu 0<Tm<1 thì Tm dự đoán là 55^{o}C – 65^{o}C và nêu Tm<0 thì Tm là dưới 65^{o}C. Dữ liệu đầu vào của TBI là trình tự amino acid và sau vài phút sẽ có kết quả trả về.

2.2.3.11. Tối ưu mã và tổng hợp gene mã hóa β-glucosidase khai thác từ dữ liệu DNA đa hệ gene vi sinh vật xung quanh nấm mục trắng

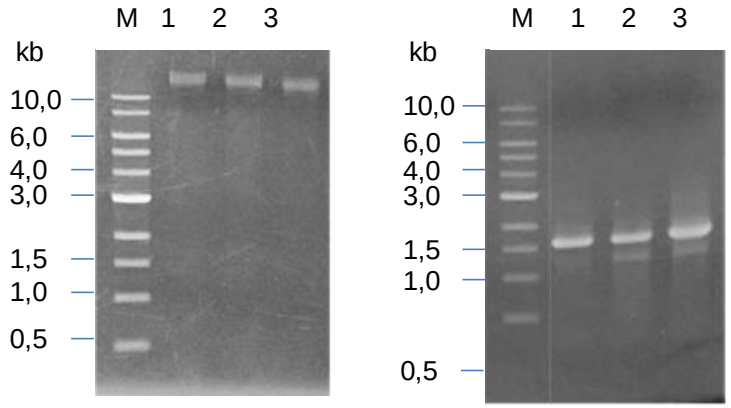
Để đảm bảo các gene β-glucosidase được biểu hiện là phù hợp với hệ biểu hiện *E. coli*, gene mã hóa enzyme β-glucosidase đã được kiểm tra sự phù hợp mã bộ ba bằng phần mềm trực tuyến (Rare Codon Analysis Tool) của hãng Genescript (https://www.genescript.com/tools/rare-codon-analysis). Để phân tích các mã bộ ba, người dùng chỉ cần cung cấp tên vật chủ biểu hiện, tên sinh vật mang gene và trình tự mã hóa của gene bắt đầy từ ATG. Phần mềm phân tích dựa trên chỉ số CAI (Codon Adaption Index – Chỉ số phù hợp mã bộ ba) cho biết trình tự ban đầu có cần tối ưu để biểu hiện trong vật chủ hay không. Sau đó, bằng phần mềm tối ưu mã bộ ba của Genescript, các gene này đã được tối ưu mã bằng phần mềm của Genescript để làm tăng chỉ số CAI lên 1 nhưng không làm thay đổi trình tự amino acid. Sau khi tối ưu, gene được đặt tổng hợp tại công ty Genescript (Mỹ).

CHƯƠNG 3: KẾT QUẢ VÀ THẢO LUẬN

3.1. Nghiên cứu đa dạng khu hệ vi khuẩn đất quanh khu nấm mục trắng

3.1.1. Tách chiết, tinh sạch DNA đa hệ gene của vi sinh vật đất

Chất lượng DNA đa hệ gene tách chiết ảnh hưởng rất lớn đến kết quả thu được khi giải trình tự gene và phân tích số liệu. Vì vậy, đối với mỗi loại mẫu khác nhau cần lựa chọn phương pháp tách chiết DNA đa hệ gene khác nhau để thu được kết quả mong muốn. Đất là một trong những nguồn tiềm năng để tìm kiếm các enzyme và vi sinh vật mới cho phân giải lignocellulose. Đặc biệt là mẫu đất ở xung quanh khu vực có sự phân giải mạnh lignocellulose. Trong thí nghiệm này, 45 mẫu đất mùn xung quanh nấm mục trắng phân giải mạnh thân cây gỗ ở vườn quốc gia Cúc Phương, Ninh Bình được thu thập, trộn lại với nhau để đảm bảo tính đa dạng của vi sinh vật trong mẫu thu được cho tách chiết DNA đa hệ gene. Kế thừa những kinh nghiệm của giai đoạn trước, phương pháp tách chiết DNA đa hệ gene của mẫu đất mùn này là phương pháp tách chiết bằng phenol đã được mô tả ở phần phương pháp. Để có đủ lượng DNA đa hệ gene tinh sạch cho giải trình tự, DNA đa hệ gene đã tiến hành tách chiết, tinh chế tất cả ba lần. Kết quả điện di kiểm tra được thể hiện trên Hình 3.1A.



A **B**

Hình 3.1. (A) Điện di đồ kiểm tra DNA đa hệ gene sau tách chiết, (B): Sản phẩm PCR gene 16S rDNA từ khuôn là DNA đa hệ gene tương ứng; 1-3: mẫu DNA đa hệ gene của 3 lần tách chiết lặp lại.

Kết quả trên điện di đồ cho thấy trên cả ba đường chạy, ba mẫu DNA đa hệ gene đều xuất hiện một băng đậm duy nhất có kích thước cao trên 10 kb, các băng này sáng, rõ nét và không thấy dải mờ ở vị trí thấp hơn. Điều này chứng tỏ DNA đa

hệ gene của vi sinh vật đã được tách thành công mà ít bị đứt gãy. Độ lớn của các mẫu điện di là tương đương nhau chứng tỏ quá trình thu mẫu và tách chiết là ổn định.

Sau khi điện di kiểm tra, mẫu DNA đa hệ gene của vi sinh vật được đo nồng độ và độ sạch bằng máy nanophotometer P330. Để đánh giá sơ bộ độ sạch của mẫu DNA đa hệ gene, giá trị tỉ lệ A_{260}/A_{280} được sử dụng. DNA hấp thụ ánh sáng mạnh nhật ở bước sóng 260 nm và protein hấp thụ ánh sáng mạnh ở bước sóng 280 nm. Do đó, giá trị cao của tỉ lệ A_{260}/A_{280} cho thấy sự hiện diện của DNA nhiều hơn và mẫu DNA tinh khiết hơn, nếu tỉ lệ này có giá trị thấp cho thấy sự hiện diện của protein chiếm ưu thế hơn. Thông thường khi tách chiết DNA có độ tinh khiết tốt thì tỉ lệ này thường có giá trị này trong khoảng 1,8-2. Kết quả đo nồng độ và độ sạch của mẫu DNA đa hệ gene được thể hiện ở Bảng 3.1

Bảng 3.1. Kết quả đo nồng độ và độ sạch của mẫu DNA đa hệ gene vi sinh vật xung quanh khu nấm mục trắng

Mẫu	**Nồng độ (ng/µl)**	**A_{260}/A_{280}**
1	145	1,921
2	112	1,922
3	126	1,931

Kết quả đo cho thấy nồng độ DNA thu được dao động trong khoảng từ 112 đến 145 ng/µl (tức là trên 100 ng/µl) và các kết quả A260/A280 đều xấp xỉ 1,90. Kết quả này tương tự với kết quả đo DNA đa hệ gene của mẫu nước suối nước nóng Bình Châu với nồng độ DNA là 139,3 ng/µl, A260/A280 đạt 1,84 [65]. Như vậy, mẫu DNA đa hệ gene từ vi sinh vật đất đã được tinh sạch, loại bỏ các tạp chất không mong muốn.

Mẫu DNA đa hệ gene sau khi được tách chiết, làm sạch được đánh giá sự tồn tại hay không của các chất ức chế hoạt động của enzyme. Chất ức chế này ảnh hưởng đến phản ứng PCR bằng cách ngăn chặn sự tương tác giữa DNA khuôn và enzyme *Taq* polymerase. Phản ứng PCR khuếch đại gene *16S rDNA* đã được sử dụng để đánh giá sự có mặt hay không của chất ức chế hoạt động của polymerase. Kết quả trên điện di đồ (Hình 3.1B) cho thấy khi PCR bằng mồi 16S sử dụng DNA đa hệ gene làm khuôn cho kết quả tốt. Trên cả 3 đường chạy đều xuất hiện băng PCR sáng duy nhất, có kích thước khoảng 1,5 kb, tương đương với kích thước lý thuyết của đoạn gene *16S rDNA* 16S ở vi khuẩn. Điều đó chứng tỏ phản ứng PCR vẫn diễn ra, trong mẫu DNA đa hệ gene không tồn tại chất ức chế polymerase. Như vậy, bằng việc kiểm tra

nồng độ, độ sạch bằng xác định giá trị tỷ lệ A_{260}/A_{280}, sự không tồn tại của các chất ức chế polymerase trong DNA đa hệ gene cho thấy, mẫu DNA đa hệ gene đã được tách chiết, tinh sạch thành công đảm bảo điều kiện cho giải trình tự. Để có sự đa dạng thành phần loài các vi sinh vật, lượng DNA đa hệ gene giống nhau từ ba lần tách chiết được trộn lại, sử dụng máy Qubit fluorometer đo được nồng độ DNA 113,25 µg/µl. Tổng lượng DNA đa hệ gene được sử dụng để chuyển cho BGI giải trình tự là 100 µg.

3.1.2. Kết quả giải trình tự DNA đa hệ gene vi khuẩn đất

DNA đa hệ gene của vi sinh vật đất quanh nấm mục trắng ở vườn Quốc gia Cúc Phương được giải trình tự bằng hệ thống HiSeq Illuminar. Kết quả này được sử dụng để phân tích, nghiên cứu đa dạng khu hệ vi khuẩn nói chung, nghiên cứu đa dạng vi khuẩn sinh cellulase nói riêng và tìm kiếm các gene ứng viên tiềm năng cho thủy phân sinh khối lignocellulose.

Bảng 3.2. Kết quả giải trình tự DNA đa hệ gene bằng hệ thống HiSeq Illuminar

	Loại phân tích	Kết quả thu được	Đơn vị
Read	Số lượng	345.471.086	read
	Tổng kích thước các read	51.820.662.900	cặp base
Contig	Số lượng	2.611.883	contig
	Kích thước trung bình	898	cặp base
	Kích thước N50	1117	cặp base
	Kích thước lớn nhất	611.845	cặp base
Gene	Số lượng	4.104.872	gene
	Kích thước trung bình	505	cặp base
	Kích thước N50	615	cặp base
	Kích thước lớn nhất	20.541	cặp base

Từ khoảng 100 µg DNA đa hệ gene được sử dụng để giải trình tự, kết quả giải trình tự ban đầu thu được ở dạng thô, có lẫn các trình tự bị lỗi và trình tự adapter. Những trình tự nhiễu này được loại bỏ khỏi dữ liệu thô nhờ công cụ SOAPnuke, thu được dữ liệu tinh là 345.471.086 read (1 read là một đoạn DNA được đọc trình tự, mỗi read thường có khoảng 100 – 200 bp [117]) với tổng dung lượng khoảng 51,82 Gb. Dữ liệu tinh được các phần mềm chuyên biệt là phần mềm IDBA (version 1.1.0) [108] và phần mềm MEGAHIT (version 1.0) [109] xử lý, lắp ráp và phân tích để thu được các contig có trình tự dài hơn. Có tổng số 2.611.883 contig được tạo ra với tổng chiều dài là 2.346 Mb. Trong đó, chiều dài trung bình của các contig là 898 bp, contig trung vị N50 có chiều dài là 1117 bp (50% các contig có kích thước ≥ 1117 bp), contig

có chiều dài lớn nhất là 611.845 bp. Sau đó, phần mềm MetaGenee Mark đã được sử dụng và dự đoán được 4.104.872 ORF (open reading frame) mã hóa protein tương đương khoảng 2.074 Mb. Trong đó, chiều dài trung bình của các gene là 505 bp, chiều dài của N50 là 615 bp (50% các gene có chiều dài ≥ 615 bp) và gene dài nhất có kích thước 20.541 bp (Bảng 3.2)

3.1.3. Phân tích đa dạng vi khuẩn đất quanh khu nấm mục trắng

Từ dữ liệu 51,82 Gb DNA đa hệ gene của vi sinh vật đất quanh khu nấm mục trắng ở vườn Quốc gia Cúc Phương, có 4.104.872 gene mã hóa protein đã được xác định bằng phần mềm MetaGenee Mark. Trong đó, có 3.923.046 gene (khoảng 95,57%) được chú giải trong cơ sở dữ liệu NR (là CSDL chứa các trình tự non-redundant và các trình tự khác như Refseq, PDB, Swiss-Prot, PIR và PRF). Bằng phần mềm MEGAN (MEtaGeneome ANalyzer) (version 4.6) các gene này đã được xác định phân loại, có 3.896.881 gene được xếp vào các giới vi khuẩn, sinh vật nhân chuẩn (Eukaryote), vi khuẩn cổ (Archaea) và virus. Trong đó, số gene được xếp vào giới vi khuẩn là chiếm ưu thế tuyệt đối với 3.884.879 gene (chiếm khoảng 99,69% tổng số gene), các giới còn lại là vi khuẩn cổ với 293 gene (0,01%), sinh vật nhân thực với 1144 gene (0,03%) và virus là 10.565 gene (0,27%). Như vậy, vi khuẩn có số lượng gene nhiều nhất và các gene của vi khuẩn được xếp vào 111 ngành, 83 lớp, 170 bộ, 406 họ, 1971 chi và chỉ có 738 loài được xác định (Bảng 3.3).

Bảng 3.3. Kết quả phân tích đa dạng từ dữ liệu DNA đa hệ gene vi sinh vật đất sử dụng phần mềm MEGAN (version 6) dựa trên CSDL NR

	Số gene	**Tỉ lệ (%)**	**Ngành**	**Lớp**	**Bộ**	**Họ**	**Chi**	**Loài**
Vi khuẩn	3.884.879	99,69	111	83	170	406	1971	738
Vi khuẩn cổ	293	0,01	9	12	18	23	50	8
Sinh vật nhân chuẩn	1144	0,03	7	26	46	79	113	86
Virus	10.565	0,27	0	0	2	14	101	84
Tổng	3.896.881	100	131	118	237	523	2240	916

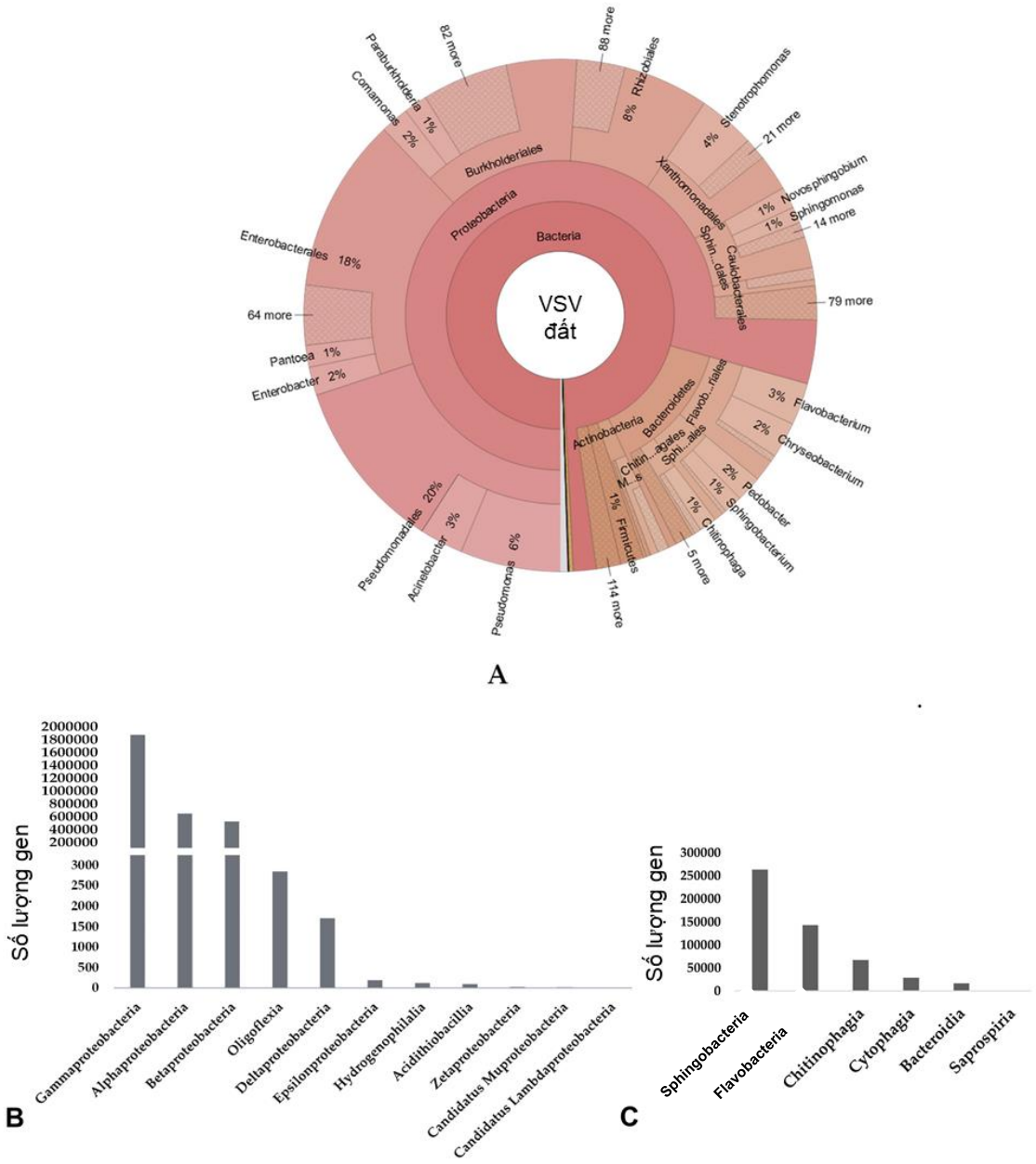
Trong khi đó, vi khuẩn cổ được xếp vào 9 ngành, sinh vật nhân chuẩn xếp vào 7 ngành và virus chưa được xếp vào mức phân loại ngành. Kết quả này lớn hơn nhiều so với công bố về thành phần loài trước đó của Praeg và cộng sự (2020) trong nghiên cứu về quần xã vi sinh vật xung quanh vùng rễ *Larix decidua*-một loài cây chiếm ưu

thế ở dãy Alps [118]. Theo đó, các vi sinh vật ở quần xã này gồm vi khuẩn được xếp vào 26 ngành, vi khuẩn cổ được xếp vào 4 ngành và nấm được xếp vào 6 ngành.

Như vậy, kết quả phân tích mẫu DNA đa hệ gene phần lớn là DNA của vi khuẩn. Trong giới vi khuẩn này, có 93,26% của tổng số gene được xác định ở bậc phân loại ngành. Trong 111 ngành vi khuẩn được xác định, có 5 ngành phổ biến chiếm 92,59% tổng số còn lại là các ngành khác. Trong số đó, Proteobacteria là ngành phổ biến nhất với 3.106.400 gene chiếm khoảng 75,68%. Các ngành tiếp theo là Bacteroidetes chiếm 13,11%, Actinobacteria 1,6%, Firmicutes 1,4%, Acidobacteria 0,8%. Như vậy, Proteobacteria là ngành chiếm ưu thế lớn có số lượng gene cao gấp 5,77 lần ngành Bacteroidetes phổ biến thứ hai. Kết quả này cũng tương tự với các kết quả công bố trước đó về đa dạng vi sinh vật đất [3]. Theo kết quả nghiên cứu của Rui Wang và cộng sự (2017) trên đất bị nhiễm vi khuẩn gây héo thực vật và đất thường không nhiễm vi khuẩn gây héo, có 26 ngành vi khuẩn được xác định. Trong đó, Proteobacteria cũng là ngành phổ biến nhất chiếm 27%, tiếp theo là các ngành Actinobacteria 14%, Acidobacteria 14%, Chloroflexi 8% and Firmicutes 6% [119]. Trong nghiên cứu thành phần các loài vi sinh vật trong mẫu đất bị nhiễm kim loại nặng Cadmium và đất không bị nhiễm Cadmium ở gần nhà máy sản xuất phân bón ở Shuangsheng, Tứ Xuyên, Trung Quốc, kết quả cho thấy Proteobacteria là ngành phổ biến nhất ở cả hai mẫu đất với tỉ lệ ở đất nhiễm kim loại là 57,85% và đất không bị nhiễm kim loại là 38,56% [120]. Khi khảo sát quần xã vi sinh vật trong rễ cây *Larix decidua* và khu hệ vi sinh vật đất xung quanh, Praeg và cộng sự (2020) cũng nhận thấy có 26 ngành vi khuẩn trong đó Proteobacteria là 36%, Acidobacteria 16%, Actinobacteria 11%, Bacteroidetes 7%, Candidatus Saccharibacteria 6%, Verrucomicrobia 5%, Planctomycetes 4% [118]. Điều này cho thấy Proteobacteria là ngành có ưu thế trong khu hệ vi sinh vật đất nói chung và ở các vùng đất có đặc điểm đặc biệt nói riêng.

Xét mức phân loại lớp, có 93,68% các gene được xác định ở mức phân loại này và được xếp vào 83 lớp. Lớp phổ biến nhất là Gammaproteobacteria 61,70%, tiếp theo là lớp Betaproteobacteria 11,35% và Alphaproteobacteria 6,85%, ba lớp này đều thuộc ngành Proteobacteria. Hai lớp tiếp theo là Sphingobacteria 6,39% và Flavobacteriia 5,45% thuộc ngành Bacteroidetes. Các lớp còn lại có tỉ lệ thấp, dưới

1%. Nhiều nghiên cứu cũng chỉ ra rằng trong các hệ sinh thái mà quá trình phân hủy diễn ra mạnh như mùn trong thảm thực vật rừng nhiệt đới, đất dưới các tử thi thì hệ vi sinh vật thay đổi theo hướng các ngành Proteobacteria, Actinobacteria, Firmicutes tăng lên, đặc biệt là họ Alphaproteobacteria và Gammaproteobacteria [121].



Hình 3.2. (A). Phân tích đa dạng của khu hệ vi sinh vật đất xung quanh nấm mục trắng ở mức phân loại: Giới, ngành, bộ, chi; (B). Đa dạng các lớp thuộc ngành Proteobacteria; (C). Đa dạng các lớp thuộc ngành Bacteroideres

Ở mức phân loại bộ, có 3 bộ chiếm tỉ lệ lớn là Pseudomonadales chiếm 29,16%, Enterobacterales chiếm 22,26%, Burkholderiales chiếm 11,19%. Tiếp theo

là các bộ Sphingomonadales chiếm 6,39%, Xanthomonadales chiếm 5,88%, Flavobacteriales chiếm 5,44%, Sphingomonadales chiếm 3,40%, Rhizobiales chiếm 2,66%, Alteromonadales chiếm 1,68%, còn lại là các bộ có tỉ lệ thấp dưới 1 %. Ba họ chiếm tỉ lệ cao nhất là Pseudomonadaceae với 16,3%, Enterobacteriaceae chiếm 14,44% và Moraxellaceae chiếm 11,02%. Ở mức phân loại chi, chỉ có 45,27% trong tổng số gene được phân loại ở mức này và tỉ lệ của tất cả các chi đều dưới 10%. Mức phân loại loài cũng được xác định, tuy nhiên chỉ có 0,55% tổng số gene được phân loại vào 738 loài. Điều này cho thấy vẫn còn một số lượng rất lớn các trình tự gene chưa được chú giải ở mức phân loại sâu như chi và loài. Mười loài trội điển hình trong đất xung quanh khu nấm mục trắng thủy phân gỗ là *Pseudomonas putida, Enterobacter cloacae, Acinetobacter johnsonii, Beauveria bassiana, Stenotrophomonas maltophilia, Enterobacter cancerogeneus, Cedecea davisae, Acinetobacter baumannii, Salmonella enterica, Shewanella decolorationis*.

Như vậy, vi khuẩn đất có độ đa dạng cao, thành phần và sự đa dạng của khu hệ vi khuẩn trong đất phụ thuộc vào nhiều yếu tố sinh học và các đặc điểm hóa lý [122], bao gồm: chất dinh dưỡng [123], sử dụng đất, ô nhiễm đất [124]…Trong đó, pH được xem là một trong những yếu tố quan trọng, có mối quan hệ chặt chẽ với thành phần và số lượng loài của cộng đồng vi khuẩn đất [125]. Vai trò quan trọng này của pH là do các vi khuẩn có khoảng pH hoạt động tối ưu hẹp [126]. Nhiều nghiên cứu cho rằng khu hệ vi khuẩn có độ đa dạng cao trong môi trường trung tính và ở môi trường axit độ đa dạng của cộng đồng vi khuẩn giảm xuống [122], [127]. Các ngành Proteobacteria, Actinobacteria và Acidobacteria sinh trưởng phát triển ưu thế trong môi trường đất trung tính hoặc hơi kiềm [128]. Độ đa dạng của các lớp trong ngành Proteobacteria tăng lên khi pH tăng, đặc biệt là lớp Gammaproteobacteria, trong khi đó hầu hết các ngành Actinobacteria và Bacteriodetes ít bị ảnh hưởng bởi độ pH của đất [126]. Có lẽ sự sinh trưởng của các ngành này chịu ảnh hưởng bởi tổ hợp các yếu tố khác như dinh dưỡng, kết cấu đất, sử dụng đất… hơn là pH. Trong mẫu nấm mục trắng của chúng tôi, khoảng pH thu được là 6,9 – 7,3. Đây là khoảng pH phù hợp cho ngành Proteobacteria phát triển, đặc biệt là lớp Gammaproteobacteria.

3.2. Nghiên cứu khai thác gene mã hóa enzyme tham gia thủy phân lignocellulose

3.2.1. Dự đoán chức năng của DNA đa hệ gene của hệ vi khuẩn đất

Nhằm nghiên cứu về chức năng của các DNA đa hệ gene vi khuẩn đất quanh khu nấm mục trắng, toàn bộ 4.104.872 gene trong dữ liệu thu được đã được xác định chức năng gene bằng BLASTp dựa trên các CSDL gồm Swiss-Prot (dữ liệu các protein đã được xác định chức năng qua thực nghiệm), KEGG (phân loại chức năng theo con đường chuyển hóa), eggNOG (phân loại chức năng theo tiến hóa của gene) và Nr (CSDL các trình tự non-redundant từ ngân hàng genee), HMM-profile của Pfam.

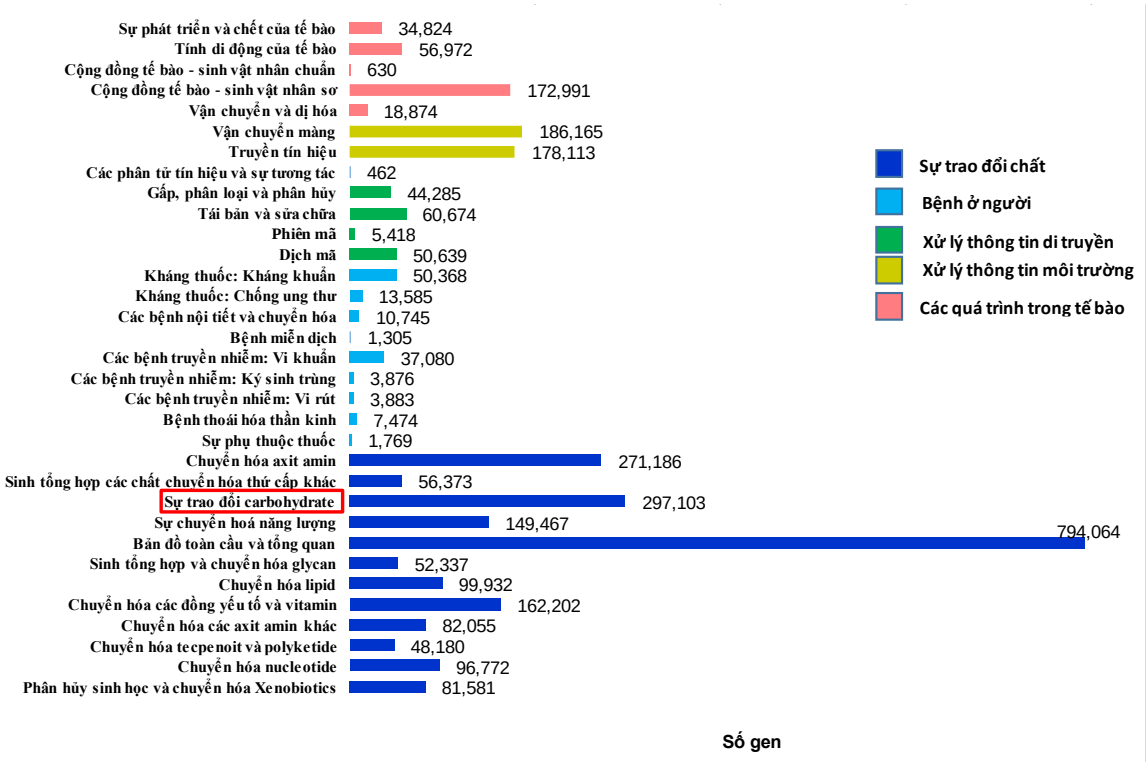
Bảng 3.4. Số lượng gene từ dữ liệu DNA đa hệ gene được chú giải chức năng dựa trên CSDL khác nhau

	Tổng số gene ban đầu	**NR**	**Swiss-Prot**	**KEGG**	**eggNOG**	**Tổng số gene chú giải được**
Số gene	4.104.872	3.923.046	2.382.630	2.809.791	3.279.853	3.925.740
%	100%	95,57%	58,04%	68,45%	79,90%	95,64%

Có một số lượng lớn các gene đã được chú giải chức năng. Cụ thể có 3.925.740 gene (tương ứng khoảng 95,64% tổng số gene) được chú giải chức năng dựa trên ít nhất một trong bốn CSDL. Dựa trên CSDL NR, số gene được chú giải là lớn nhất với 3.923.046 gene (chiếm khoảng 95,57% tổng số gene), tiếp sau đó là có 3.279.853 gene (tương ứng 79,90%) được xác định chức năng dựa trên CSDL eggNOG, trong khi đó dựa trên cơ sở dữ liệu Swiss-Prot chỉ có 58,04% gene được chú giải (2.382.630 gene) (Bảng 3.4).

Trong số các CSDL trên thì KEGG là CSDL bao gồm các gene được phân loại theo chức năng sinh học đối với tế bào và cơ thể sinh vật. Với mục đích ước đoán chức năng gene và khai thác các gene mã hóa enzyme tham gia phân giải lignocellulose thì dữ liệu KEGG cho kết quả có độ chính xác cao. Do đó, kết quả xác định chức năng gene dựa trên CSDL KEGG được sử dụng cho phân tích tiếp theo trong nghiên cứu DNA đa hệ gene. Dựa trên dữ liệu KEGG, có 2.809.791 gene (tương ứng khoảng 68,45% tổng số gene) được xác định chức năng mã hóa protein tham gia vào chuyển hóa các chất trong tế bào và cơ thể. Các protein này tham gia vào 5 nhóm chuyển hóa bao gồm: các quá trình trong tế bào, xử

lý thông tin môi trường, xử lý thông tin di truyền, bệnh ở người, sự trao đổi chất. Trong đó, trao đổi chất là quá trình có sự tham gia của nhiều gene nhất với 2.191.252 gene (tương ứng khoảng 69,98% tổng số gene được xác định), sau đó là đến các quá trình xử lý thông tin môi trường 11,63%, các quá trình trong tế bào 9,08%, xử lý thông tin di truyền 5,16% và bệnh ở người 4,15%. Trong quá trình trao đổi các chất khác nhau, trao đổi carbohydrate có 297.103 gene mã hóa protein tham gia (chiếm khoảng 13,56% trong tổng số các gene tham gia trao đổi chất) (Hình 3.3).



Hình 3.3. Sơ đồ chú giải chức năng gen từ dữ liệu DNA đa hệ gen vi sinh vật đất dựa trên CSDL KEGG

3.2.2. Khai thác gene mã hóa lignocellulase dựa trên kết quả chú giải chức năng bởi KEGG

Từ 297.103 gene được xác định chức năng là tham gia vào quá trình chuyển hóa carbohydrate trên CSDL KEGG, có 22.226 gene được ước đoán là các gene mã hóa các enzyme có tham gia vào quá trình phân giải sinh khối lignocellulose. Trong đó có 907 gene được chú giải mã hóa các enzyme tham gia vào tiền xử lý sinh khối, 8301 gene mã hóa cellulase và 13.018 gene mã hóa hemicellulase (Bảng 3.5).

Bảng 3.5. Các ORF mã hóa enzyme phân giải lignocellulose được khai thác từ DNA đa hệ gene của vi sinh vật quanh khu nấm mục trắng

Tên enzyme	Số ORF	Số ORF hoàn chỉnh	Số ORF hoàn chỉnh có domain	Số loại domain
1. Enzyme tiền xử lý	**907**	**216**	**198**	**19**
Pectinesterase (EC 3.1.1.11)	815	199	181	16
Feruloylesterase (EC 3.1.1.73)	75	12	12	2
Laccase (EC 1.10.3.2)	10	5	5	1
Expansin	7	0	0	0
2. Cellulase	**8301**	**1279**	**1058**	**81**
β-glucosidase (EC 3.2.1.21)	4272	503	475	26
Endoglucanase (EC 3.2.1.4)	2216	548	367	47
6-phospho-beta- glucosidase (EC 3.2.1.86)	1718	213	210	2
Cellobiohydrolase (EC 3.2.1.91)	73	15	6	6
Cellobiose phosphorylase (EC 2.4.1.20)	22	0	0	0
3. Hemicellulase	**13018**	**2087**	**1828**	**151**
Xyloglucan-active β-D-galactosidase (EC 3.2.1.23)	3288	330	298	36
α-L-fucosidase (EC 3.2.1.51)	2279	464	413	30
α-galactosidase (EC 3.2.1.22)	1033	163	134	15
α-L-arabinofuranosidase (EC 3.2.1.55)	1016	169	161	7
endo-β-1,4 xylanase (EC 3.2.1.8)	885	230	175	15
α-D- xylosidexylohydrolase (EC 3.2.1.177)	762	62	55	9
1,4-beta-xylosidase (EC 3.2.1.37)	659	146	134	4
β-mannosidase (EC 3.2.1.25)	611	46	37	4
oligosaccharide reducing-end xylanase (EC 3.2.1.156)	552	100	73	12
β-mannanase (3.2.1.78)	368	87	81	16

Endopolygalacturonaselyase, (EC 4.2.2.2)	341	60	52	7
β-fructofuranosidase (EC 3.2.1.26)	255	38	36	5
β-D-glucuronidase (EC 3.2.1.31)	227	33	28	3
Exopolygalacturonase (EC 3.2.1.67)	223	74	69	2
Licheninase (EC 3.2.1.73)	175	52	52	4
α-glucuronidase (EC 3.2.1.139)	161	17	16	1
Exopolygalacturonaselyase (EC 4.2.2.9)	142	9	9	1
Endopolygalacturonase (EC 3.2.1.15)	38	6	4	1
endo-transglycosylase/hydrolase (EC 2.4.1.207)	2	1	1	1
Acetylxylanesterase (EC 3.1.1.72)	1	0	0	0

Trong 907 ORF được chú giải mã hóa cho các enzyme tiền xử lý, các ORF này được xếp vào 4 nhóm là pectinesterase, feruloylesterase, laccase và expansin. Trong đó, pectinesterase là nhóm enzyme phổ biến nhất với 815 ORF (tương ứng 89,96%%), tiếp theo là các nhóm feruloylesterase 75 ORF (8,27%), laccase (1,10%) và còn lại expansin (0,67%). Các nhóm enzyme khác thường tham gia vào quá trình tiền xử lý như lignin peroxidase, lytic polysaccharide, monooxygenease, manganese peroxidase không được tìm thấy trong dữ liệu. Có 8301 ORF được chú giải mã hóa cho cellulase chia thành 5 nhóm sắp xếp theo thứ tự giảm dần là β-glucosidase (EC 3.2.1.21), endoglucanase (EC 3.2.1.4), 6-phospho-beta- glucosidase (EC 3.2.1.86), cellobiohydrolase (EC 3.2.1.91), cellobiose phosphorylase (EC 2.4.1.20); trong đó phần lớn là các ORF mã hóa cho β-glucosidase chiếm 51,46% (4272 ORF), tiếp theo là endoglucanase 26,70%, 6-phospho-beta- glucosidase 20,70%. Nhóm enzyme cellulase khác là cellobiose dehydrogenease không được tìm thấy trong dữ liệu. Xét trong nhóm enzyme hemicellulase, có 13.018 ORF được chú giải mã hóa cho hemicellulase được xếp vào 20 nhóm, trong đó nhóm xyloglucan-active β-D-galactosidase (EC 3.2.1.23) là nhóm phổ biến nhất 25,26% (3288 ORF), tiếp theo là các nhóm α-L-fucosidase (EC 3.2.1.51) chiếm 17,51% (2279 ORF), α-galactosidase

(EC 3.2.1.22) chiếm 7,94%, α-L-arabinofuranosidase (EC 3.2.1.55) chiếm 7,80%, các nhóm còn lại có số lượng ORF dưới 1000. Một số nhóm enzyme khác thuộc hemicellulase như acetyl xylan esterase, acetyl mannan esterase, α-D-xylosidase, α-L-fucosidase không được tìm thấy trong dữ liệu.

3.2.3. Khai thác gene mã hóa lignocellulase dựa trên mô hình HMM

Trong nghiên cứu chú giải chức năng gene, trình tự protein suy diễn đôi khi chỉ một phần được lắp ráp từ dữ liệu giải trình tự DNA đa hệ gene và như vậy có thể ảnh hưởng đến chú giải dựa trên sự tương đồng do không hoàn chỉnh và lỗi của các khung được lắp ráp. Trong trường hợp đó, mặc dù độ tương đồng kém nhưng các protein được dự đoán có xu hướng thực hiện các chức năng tương tự với những protein có cùng trình tự. Như vậy, chúng rất có thể có cùng kiểu motif. Mô hình HMM được xây dựng từ trình tự amino acid của các họ protein hoặc các domain đã biết sau đó chúng được sử dụng để tìm kiếm các trình tự chưa biết và phân loại chúng. Khai thác gene sử dụng mô hình đại diện HMM mà bản chất là dựa trên sự tương đồng về motif có thể chú giải được chức năng của những gene mà không có sự tương đồng cao về trình tự. Trong nghiên cứu này, khi khai thác gene mã hóa lignocellulase dựa trên mô hình HMM có 13 họ enzyme tham gia thủy phân lignocellulose đã được khai thác hiệu quả hơn so với việc khai thác gene dựa trên sự tương đồng về trình tự trong KEGG. Đó là CBM (1-84), arabinanase (GH43), galactanase, glucuronyl esterase, HPOXRE catalase, hydrogene peroxide oxidoreductase, LPMO, laccase, axetylxylanesterase, beta- glucuronidase, cellobiohydrolase, lichenase, beta-xylosidase. Trong đó hydrogene peroxide oxidoreductase (thuộc nhóm hemicellulase) và LPMO (enzyme tiền xử lý) là chưa được tìm thấy dựa trên dữ liệu KEGG, CAZy. Điều này cho thấy khi sử dụng công cụ mới là mô hình đại diện HMM, các nhóm enzyme quan trọng đã được tìm thấy. Đây là cơ sở để hiểu biết đầy đủ hơn về hệ enzyme tham gia thủy phân lignocellulose.

Bảng 3.6. So sánh kết quả xác định gen mã hóa enzyme phân giải lignocellulose bằng mô hình HMM và KEGG

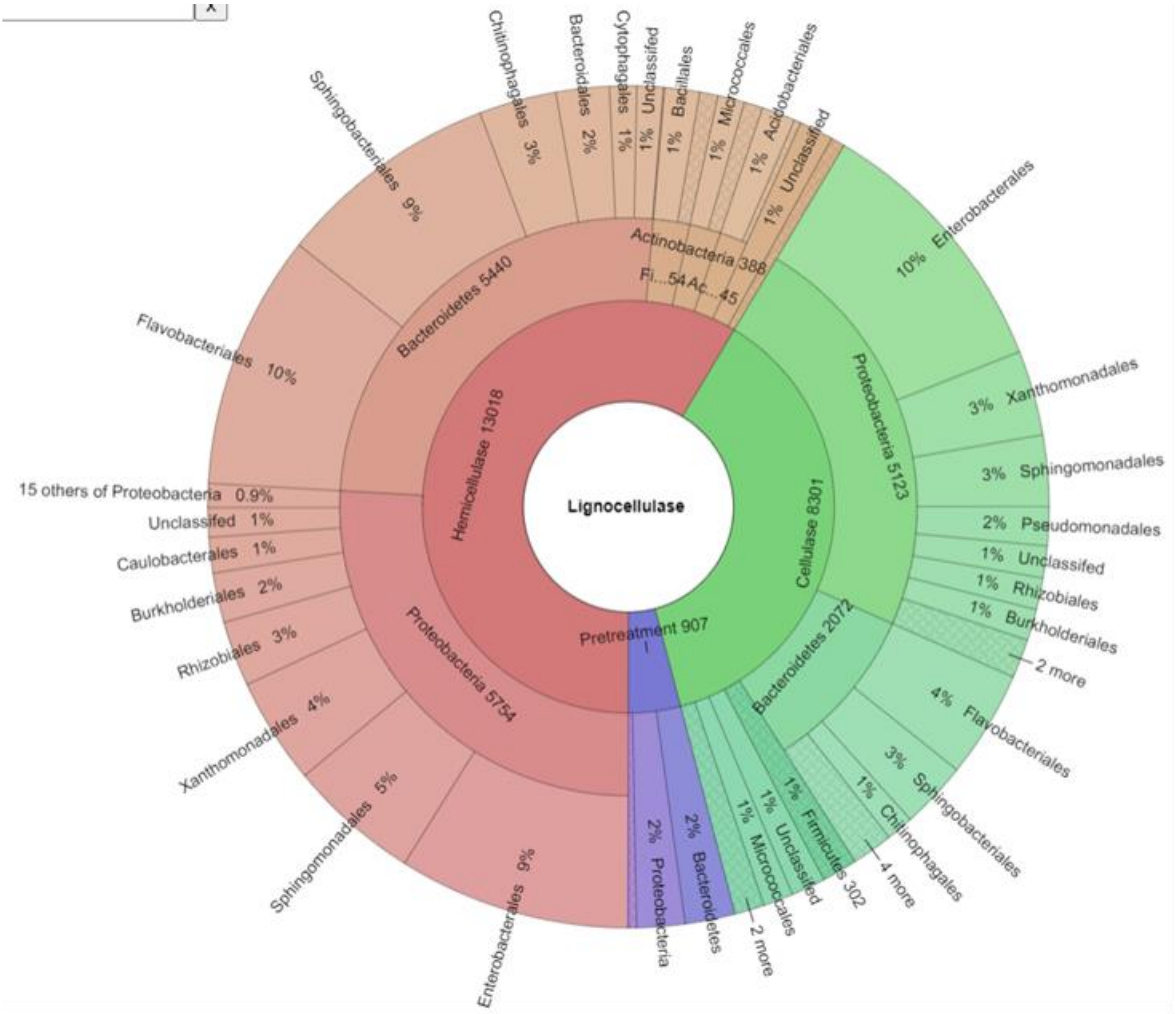
STT	Tên enzyme	Số lượng gene dựa trên HMM	Số lượng gene dựa trên KEGG
1	*CBM (1-84)*	*3163*	*< 300*
2	*Arabinanase (GH43)*	*343*	-

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. Nghiên cứu đa dạng các vi sinh vật mang gene mã hóa lignocellulase

Trong số 22.226 gene mã hóa enzyme tham gia phân hủy lignocellulose có phần lớn các gene đã xác định được nguồn gốc vi khuẩn ở các cấp độ phân loại, và chỉ 107 (chiếm 0,49%) không xác định được đơn vị phân loại. Trong số này, có 22.092 gene (chiếm 99,39%) là thuộc về vi khuẩn được xếp vào 28 ngành, trội nhất là ngành Proteobacteria (11.288 gene, chiếm 50,79%), tiếp theo là Bacteroidetes (8.164 gene, 36,73%), Firmicutes 3,43%, Actinobacteria 3,30%, Acidobacteria 1,99%, Verucomicrobia 0,53%, Cyanobacteria 0,11% Planctomycetes 0,11% và tổng số 20 ngành khác chiếm 0,22% (Hình 3.4). Tỷ lệ Bacteroidetes/Proteobacteria (0,72; 1) trong gene mã hóa enzyme tham gia phân hủy lignocellulose cao hơn nhiều so với tỷ lệ này trong tổng số cấu trúc vi khuẩn của mùn xung quanh khu nấm mục trắng (0,17: 1). Điều này cho thấy Bacteroidetes đóng vai trò quan trọng trong quá trình thủy phân lignocellulose. Ở cấp độ bộ, phân tích cũng cho thấy Enterobacterales là bộ nổi bật nhất chiếm 20,06%, tiếp theo là Flavobacters 15,14%, Sphingobacteria 11,62%.

Phân tích sâu hơn với nhóm enzyme tiền xử lý chúng tôi thấy rằng ngành Bacteroidetes là ngành phong phú nhất (427 gene, chiếm 47,08%), cao hơn một chút so với Proteobacteia (45,31%). Trong khi đó, đối với nhóm hemicellulase, Proteobacteria (44,20%) cao hơn so với Bacteroidetes (43,52%). Đối với cellulase, tỷ lệ giữa Proteobacteria và Bacteroidetes khác biệt đáng kể, đạt 2,4 lần tương ứng với Proteobacteria 61,72% và Bacteroidetes 24,96%. Như vậy, tỷ lệ Proteobacteia/Bacteroidetes trong DNA đa hệ gene vi sinh vật xung quanh nấm mục trắng là 5,77, trong khi đối với cellulase thì tỷ lệ Proteobacteria/Bacteroidetes là 2,4. Do đó, Bacteroidetes dường như đóng một vai trò quan trọng hơn trong quá trình thủy phân lignocellulose.



Hình 3.4. Đa dạng vi sinh vật mang gen mã hóa lignocellulase ở ngành và bộ

So sánh ở mức độ phân loại bộ thể hiện sự khác biệt to lớn giữa các nhóm enzyme. Các bộ Flavobacteriales, Sphingobacteriales, Enterobacterales lần lượt chiếm 29,88%, 19,63%, 17,42% là ba bộ có vai trò quan trọng trong nhóm enzyme

tiền xử lý; bộ Enterobacterales 27,90% và Flavobacterales 11,02% là hai bộ phổ biến trong cellulase; và đối với hemicellulase thì các bộ Flavobacteriales, Enterobacterales và Sphingobacteriales là bộ chiếm ưu thế lần lượt là 16,72%, 15,26%, 14.65%. Trong khi đó, trong tổng số hệ vi sinh vật đất mùn xung quanh nấm mục trắng, các bộ Sphingobacteriales 6,39%, Xanthomonadales 5,88%, Flavobacteriales 5,44% thuộc bộ phổ biến thứ hai dưới 10%, các bộ chiếm ưu thế nhất là Pseudomonadales 29,16%, tiếp theo là Enterobacteriales 22,26% và Burkholderiales 11,19%. Ngược lại, Pseudomonadales chỉ chiếm lần lượt là 3,75%; 4,04%; 0,45% trong nhóm enzyme tiền xử lý, cellulase, hemicelulase. Do đó, Pseudomonadales là bộ điển hình có trong hệ vi sinh vật đất mùn nhưng không phải là bộ chứa gene tham gia mã hóa lignocellulase. Enterobacteriales là bộ chiếm ưu thế trong cả mẫu mùn và enzyme phân giải lignocellulose. Bộ Flavobacteriales chiếm ưu thế trong tất cả các vi sinh vật chứa enzyme lignocellulase. Do đó, các bộ Flavobacteriales và Enterobacterales lần lượt thuộc ngành Bacteroideles, Proteobacteria đóng một vai trò quan trọng trong quá trình phân giải lignocellulose của mùn. Có nhiều nghiên cứu cũng chỉ ra rằng Bacteroidetes tổng hợp được cellulase ở nhiều hệ sinh thái khác nhau [129]. Trong nghiên cứu của Vries và cộng sự (2015), khi nghiên cứu thành phần loài các vi sinh vật ở đất nông nghiệp được xử lý theo các cách khác nhau, thấy phần lớn các enzyme cellulase được chú giải cho Proteobacteria, Actinobacteria và Bacteroidetes [130]. Vi khuẩn thuộc Bacteroidetes có vai trò quan trọng trong phân giải polysaccharide và được tìm thấy ở hầu hết các hệ sinh thái [131]. Có nhiều nghiên cứu cho thấy, Bacteroidetes thường chiếm khoảng 10% trong thành phần các vi sinh vật đất [132]. Các vi khuẩn thuộc ngành này cũng được biết có chứa nhiều gene mã hóa cho các enzyme phân giải polysaccharide, các gene được sắp xếp trong cụm gene gọi là PULs (polysaccharide utilization loci). Trong nghiên cứu về đất than bùn ở Bắc Cực, phần lớn các gene nghiên cứu tham gia mã hóa cho enzyme phân giải sinh khối lignocellulose được xác định thuộc các ngành Bacteroidetes, Actinobacteria, Verrucomicrobia (chiếm khoảng 70% các gene). Trong kết quả nghiên cứu này cho thấy, cả hai ngành Proteobacteria và Bacteroidetes đều có mặt trong khu hệ vi sinh vật trong đất xung quanh khu nấm mục trắng phân hủy lignocellulose trong đó Proteobacteria chiếm tỉ lệ 75,68% và Bacteroides chiếm tỉ lệ 13,11%, trong khi đó

trong các gene mã hóa lignocellulase thì Proteobacteria chiếm tỉ lệ giảm xuống 61,72% và Bacteroidetes tăng lên 1,90 lần với tỉ lệ 24,96%. Điều đó cho thấy ngành Bacteroidetes, bộ Flavobacteriales có sự phát triển ưu thế trong số các vi sinh vật chứa gene mã hóa enzyme phân giải lignocellulose. Kết quả này cũng phù hợp với kết quả được công bố trước đó của Soares và cộng sự (2012). Việc giải trình tự gene 16S rRNA ở các chủng được phân lập từ đất Nam Cực đã cho thấy bộ Flavobacteriia là nhóm chính của vi khuẩn tham gia phân giải cellulose [133]. Đặc biệt, Edwards và cộng sự (2010) khi nghiên cứu thành phần loài trong các môi trường có sự phân hủy polysaccharide mạnh cũng thấy chỉ ra rằng các bộ Gammaproteobacteria thuộc ngành Proteobacteria và bộ Flavobacteriia thuộc ngành Bacteroidetes [134] là các bộ chiếm ưu thế tuyệt đối ở các môi trường này.

Trong số các gene mã hóa enzyme phân giải lignocellulose, ở luận án này chúng tôi tiếp tục tiến hành khai thác và lựa chọn gene tiềm năng mã hóa enzyme thủy phân cellulose để tổng hợp/phân lập gene.

3.3. Nghiên cứu khai thác và lựa chọn gene tiềm năng mã hóa cellulase

3.3.1. Phân tích các vùng chức năng của cellulase

Dựa trên việc tham chiếu với CSDL KEGG, có 8301 ORF được xác định mã hóa enzyme cellulase. Các ORF này gồm 5 loại enzyme gồm: (1) endoglucanase có 2216 ORF mã hóa endoglucanase EC 3.2.1.4 – thủy phân các liên kết 1,4-β-D-glucoside bên trong mạch của các chuỗi cellulose để tạo ra các chuỗi ngắn hơn, lichenin và cereal β-D-glucan; (2) exoglucanase có 73 ORF mã hóa cellobiohydrolase EC 3.2.1.91 - thủy phân liên kết (1,4)-β-D-glucoside ở hai đầu của các chuỗi ngắn cellotetraose, giải phóng cellobiose; (3) 4272 ORF mã hóa β-glucosidase EC 3.2.1.21- xúc tác phản ứng phân cắt liên kết glycoside để giải phóng phân tử β-D-glucose từ hợp chất glycoside hoặc oligosaccharide; (4) 1718 ORF mã hóa 6-phospho-β-glucosidase EC 3.2.1.86 – xúc tác phản ứng 6-phospho-β-D-glucosyl-(1,4)-D-glucose + H_2O -> D-glucose + D-glucose 6-phosphate; (5) 22 ORF mã hóa cellobiose phosphorylase EC 2.4.1.20 – xúc tác phản ứng cellobiose + phosphate -> α-D-glucose 1-phosphate + D-glucose). Trong số 8301 gene được chú giải mã hóa cho enzyme cellulase có 1279 gene (15,41%) là chứa gene đầy đủ có cả hai đầu 5’ và 3’ bao gồm: 548 gene endoglucanase, 15 gene exoglucanase loại cellobiohydrolase,

503 gene β-glucosidase, 213 gene 6-phospho-β-glucosidase còn lại là 7022 gene không đầy đủ (thiếu đầu 5', đầu 3' hoặc thiếu cả hai đầu). Trong nghiên cứu khai thác, phân tích cấu trúc vùng chức năng cellulase, chúng tôi đã ưu tiên lựa chọn 1279 ORF đầy đủ đã được dự đoán có cả đầu 3', đầu 5' để phân tích. (Bảng 3.7).

Bảng 3.7. Các ORF mã hóa cellulase trong DNA đa hệ gene vi sinh vật đất

Enzyme	ORF đầy đủ	ORF mất đầu 5'	ORF mất đầu 3'	ORF mất 2 đầu	Tổng
Endoglucanase (EC 3.2.1.4)	**548**	447	503	718	**2216**
Cellobiohydrolase (EC 3.2.1.91)	**15**	5	11	42	**73**
β-glucosidase (EC 3.2.1.21)	**503**	765	1065	1939	**4272**
6-phospho-β-glucosidase (EC 3.2.1.86)	**213**	397	454	654	**1718**
Cellobiose phosphorylase (EC 2.4.1.20)	**0**	3	7	12	**22**
Tổng	**1279**	**1617**	**2040**	**3365**	**8301**

Các enzyme thủy phân cellulose thường có cấu trúc gồm nhiều vùng chức khác nhau. Các vùng này có vai trò quan trọng ảnh hưởng đến cơ chế hoạt động và hoạt tính của enzyme cellulase. Chúng tôi tiến hành nghiên cứu các vùng chức năng (domain) của các gene mã hóa cellulase sử dụng CSDL Pfam và mô hình đại diện HMM. Kết quả thu được trong số 1279 gene đầy đủ mã hóa các nhóm enzyme cellulase được sử dụng cho phân tích thì có 1058 gene có domain bao gồm: 367 gene mã hóa endoglucanase, 6 gene mã hóa exoglucanase loại cellobiohydrolase, 475 gene mã hóa β-glucosidase, 210 gene mã hóa enzyme 6-phospho β-glucosidase. Kết quả được thể hiện trên Bảng 3.8.

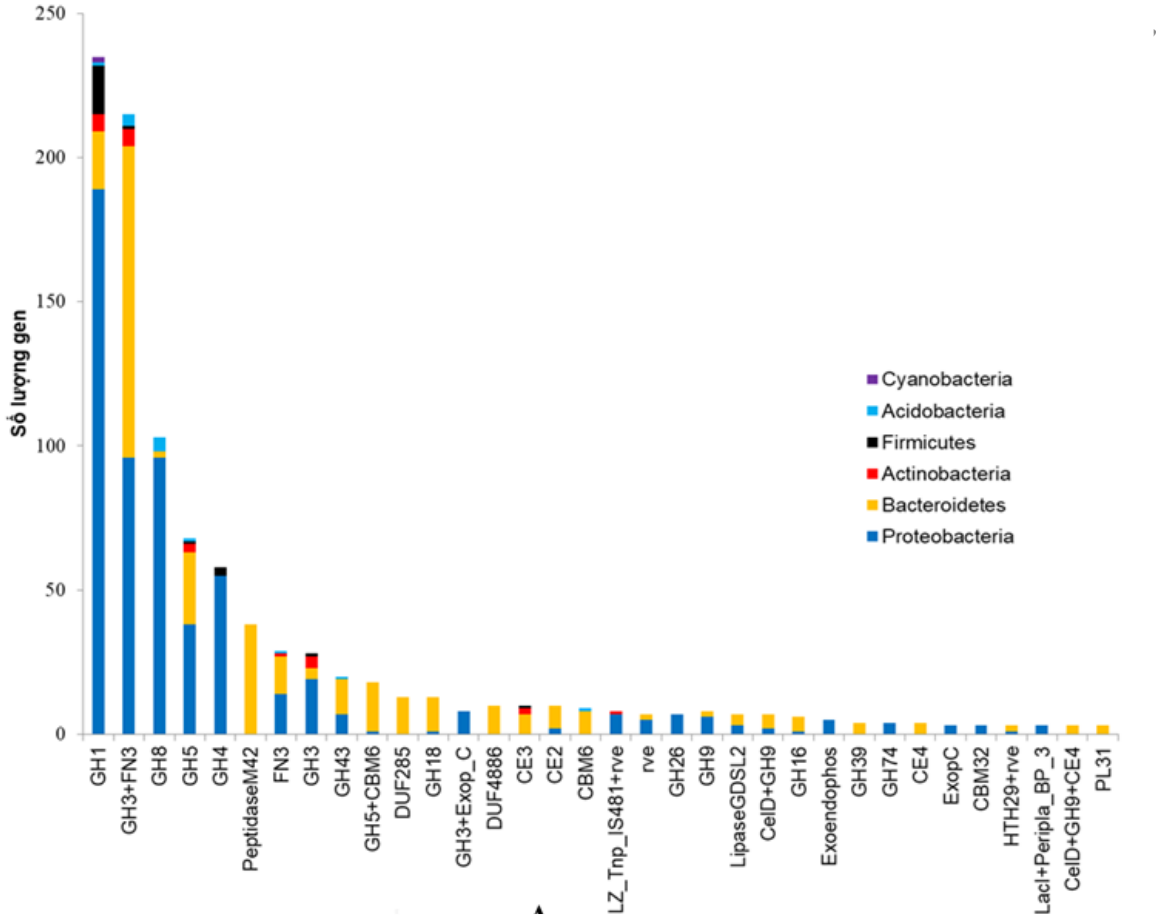
Bảng 3.8. Kết quả phân tích vùng chức năng của các ORF hoàn chỉnh mã hóa cellulase

Enzyme	Số ORF hoàn chỉnh có domain		Loại domain
	Tổng số	Số ORF theo loại domain	
Endoglucanase (EC 3.2.1.4)	367	105	GH8
		72	GH5
		38	PeptidaseM42

		18	GH5-CBM6
		14	DUF285
		13	GH18
		10	CE2
		97	40 loại khác
Cellobiohydrolase (EC 3.2.1.91)	6	1	Alginate_lyase
		1	Amidase 3
		1	CBM2
		1	CBP_BcsO
		1	GH128+Laminin G3
		1	Znribbon8
β-glucosidase (EC 3.2.1.21)	475	220	GH3+FN3
		93	GH1
		29	FN3
		29	GH3
		20	GH43
		11	GH3+Exop_C
		10	DUF4886
		10	CE3
		8	LZ_Tnp_IS481+rve
		8	rve
		5	Exoendophos
		5	GH16
		4	LipaseGDSL2
		3	ExopC
		3	HTH29+rve
		3	LacI+Peripla_BP_3
		2	CBM32
		2	GH89
		2	rve3
		1	AP_endonuc2
		1	CBM32+GH55
		1	GH87
		1	GH16+CBM32
		1	GH43+CBM32+LamininG3
		1	GxDLY+Lipase_GDSL_3+ CE3
		1	HTH1+GH1
		1	SpoIIM
6-phospho-β-glucosidase	210	152	GH1
		58	GH4
Tổng	**1058**		**81 loại domain**

Kết quả nhận được cho thấy trong 1058 ORF hoàn chỉnh mã hóa cellulase chứa 81 loại domain. Trong đó, domain phổ biến nhất thuộc họ GH (chiếm trên 80% ORF hoàn chỉnh có domain). Đại diện là GH1 có 245 ORF trong đó 189 ORF (tương

ứng 77,14%) thuộc ngành Proteobacteria, 20 ORF (8,16%) thuộc ngành Bacteroidetes còn lại là thuộc các ngành khác. Tiếp theo là domain GH3+FN3 (220 ORF) trong đó 96 ORF (43,67%) thuộc ngành Proteobacteria, 108 ORF (49,09%) thuộc ngành Bacteroidetes. Sau đó là các họ GH khác như họ GH8 (105 ORF), GH5 (72 ORF), GH4 (58 ORF) trong đó tỉ lệ các ORF thuộc ngành Proteobacteria lần lượt



Hình 3.5. Các ngành vi khuẩn mang ORF hoàn chỉnh có domain mã hóa cellulase

Phân tích theo từng nhóm enzyme cho thấy, thuộc nhóm endoglucase có 367 ORF với 47 loại domain. Trong đó, domain GH8 là phổ biến nhất với 105 ORF, 96 ORF trong số này (91,43%) thuộc ngành Proteobacteria, chỉ có 2 ORF (1,90%) thuộc ngành Bacteroidetes, 5 ORF (4,77%) là ngành Acidobacteria và 2 ORF thuộc các ngành khác. Loại domain phổ biến thứ hai trong nhóm enzyme này là GH5 với 72 ORF. Các ORF chứa domain GH5 hầu hết thuộc hai ngành Proteobacteria (52.78%) và Bacteroidetes (34.72%). Tiếp theo là các loại domain Peptidase M42 (38 ORF), GH5-CBM6 (18 ORF), DUF285 (14 ORF), GH18 (13 ORF), CE2 (10 ORF) trong

đó phần lớn các domain thuộc ngành Bacteroidetes với tỉ lệ lần lượt là 100%, 94,44%,92,86%, 92,31%, 80,00%. Còn lại 43 loại domain (97 ORF) thuộc nhiều nhóm phân loại khác nhau.

Trong nhóm enzyme exoglucanase chỉ có 6 ORF với 6 loại domain khác nhau. Trong số các domain thuộc ORF mã hóa exoglucanase có 3 domain Alginate_lyase, Amidase 3, GH128+Laminin G3 thuộc ngành Bacteroidetes, 2 domain CBM2 và Znribbon8 thuộc ngành Acidobacteria, domain CBP_BcsO thuộc ngành Proteobacteria.

Nhóm enzyme β-glucosidase là nhóm enzyme có số lượng ORF nhiều nhất 475 ORF (44,90%) với 27 loại domain. Trong nhóm enzyme này, domain có số lượng nhiều nhất là domain GH3+FN3 với 220 ORF, trong đó 96 ORF (43,64%) thuộc ngành Proteobacteria và 108 ORF (49,09%) thuộc ngành Bacteroidetes còn lại 16 ORF thuộc các ngành khác. Tiếp đó là domain GH1 (93 ORF) trong đó 60 ORF (64,52%) thuộc ngành Proteobacteria, 20 ORF (21,51%) thuộc ngành Bacteroidetes và 13 ORF còn lại thuộc một số ngành khác. Ngoài ra còn nhiều domain khác được tìm thấy như GH4, FN3, GH3, GH43, GH3+Exop_C, DUF4886, CE3, LZ_Tnp_IS481+rve, rve, Exoendophos, GH16, LipaseGDSL2, ExopC, HTH29+rve, LacI+Peripla_BP_3, CBM32, GH89, rve3, AP_endonuc2. Các domain này đều thuộc nhóm Proteobacteria, trong khi đó một số domain chỉ tìm thấy ở nhóm Baceroidetes là: DUF4886, CE3, GH89, GH16+CBM32, GH43+CBM32+LaminiG3, GxDLY+Lipase_GDSL_3+CE3. Trong nhóm 6-phospho-β-glucosidase, domain phổ biến nhất là GH1 với 152 ORF trong đó có 129 ORF (84,87%) thuộc ngành Proteobacteria, 15 ORF thuộc ngành Firmicutes còn lại chưa được phân loại ngành. Domain còn lại trong nhóm enzyme này là GH4 với 58 ORF, trong đó 55 ORF (94,83%) thuộc ngành Proteobacteria và 3 ORF thuộc ngành Firmicutes.

Từ các kết quả phân tích domain của các ORF mã hóa cellulase có thể thấy các enzyme mã hóa cellulase có cấu trúc domain khá đơn giản, chúng không chứa các domain hoạt tính khác hoặc các domain không có chức năng xúc tác mà chỉ có một domain xúc tác. Có lẽ đây là đặc điểm đặc trưng của các enzyme cellulase của khu hệ vi khuẩn đất quanh khu nấm mục trắng. Như vậy trong phân tích của chúng tôi, các ORF thuộc nhóm endoglucase chứa domain GH8 nhiều nhất, nhóm β-

glucosidase chứa domain GH3+FN3, GH1 là nhiều nhất, nhóm 6-phospho-β-glucosidase chứa domain GH1 nhiều nhất với sự ưu thế của ngành Proteobacteria. Ở các nghiên cứu khác, khi phân tích DNA metageneeome của các vi sinh vật trong dạ cỏ dê thu được GH5 và GH9 thể hiện hoạt động endoglucanase, trong khi đó GH3 được dự đoán có hoạt tính β-glucosidase. Các enzyme này hầu hết đều được phân loại vào ngành Bacteroidetes, một số enzyme được phân loại vào ngành Firmiticute như GH5, GH6, GH9 sẽ đi kèm với CBM như CBM2, CBM3, CBM4, CBM63 [2]. Trong thí nghiệm của Inoue và cộng sự (2014) khi tinh chế cellulase từ nấm *Talaromyces cellulolyticus* thu được các họ GH3 hoạt tính β-glucosidase, GH5 hoạt tính endoglucanase, GH6 và GH7 hoạt tính cellobiohydrolase. Trong CSDL CAZy mô tả các β-glucosidase chịu trách nhiệm xử lý các oligosaccharide nhỏ chủ yếu được tìm thấy trong GH1 và GH3, trong khi các endo- và exocellulase chủ yếu có trong GH5, GH6, GH8, GH9, GH12, GH44, GH45, GH48 [135]. Đáng chú ý, trong kết quả phân tích domain của các gene mã hóa β-glucosidase chỉ ra rằng riêng trong nhóm β-glucosidase có khoảng 90% của các domain GH3 có liên kết vùng/cấu trúc FN3 và Exop_C. Đây là các vùng/cấu trúc độc lập và ít được nghiên cứu. Cấu trúc GH3+FN3 xuất hiện cả ở ngành Proteobacteria và Bacteroidetes, cấu trúc GH3+Exop_C chỉ xuất hiện ở ngành Proteobacteria. FN3 là loại vùng/cấu trúc liên kết phổ biến nhất chịu trách nhiệm nới lỏng bề mặt cellulose, làm bong tróc sợi cellulose và hướng chuỗi cellulose vào lõi xúc tác để dễ dàng chuyển đổi cơ chất. Ngoài ra sự có mặt của vùng/cấu trúc FN3 còn giúp enzyme được hình thành và hoạt động. Kết quả này cũng thông nhất với các công bố trước đó của Nguyen và cộng sự (2021) khi khai thác DNA đa hệ gene ở dạ cỏ dê thu được 90,9% cellulase GH3 chứa vùng/cấu trúc FN3 [136]. Vùng/cấu trúc Exop_C thường ít gặp, vai trò chính của vùng/cấu trúc này không chỉ là liên kết với cơ chất mà còn có vai trò ổn định cấu trúc cần thiết cho hoạt động của enzyme.

3.3.2. Dự đoán mức độ biểu hiện của các gene mã hóa cellulase

Số lượng các gene thu được sau khi phân tích DNA đa hệ gene thường rất lớn. Vì vậy, để đạt được hiệu quả cao trong nghiên cứu thực nghiệm thì mức độ biểu hiện ngoại bào của các gene nói chung và mức độ biểu hiện dạng tan nói riêng cần được dự đoán. Sự biểu hiện của các protein dạng tan trong vật chủ giúp cho các protein giữ

nguyên được cấu trúc không gian và có hoạt tính sinh học. Mức độ biểu hiện này phụ thuộc vào độ tương thích của gene cần được biểu hiện và vật chủ. Trong số các vật chủ biểu hiện hiện nay thì *E. coli* là hệ biểu hiệu phổ biến và đơn giản nhất. Mức độ biểu hiện dạng tan của 1058 gene hoàn chỉnh có đầy đủ cấu trúc domain mã hóa cho enzyme cellulase đã được xác định bằng phần mềm Periscope. Các gene đại diện cho mỗi nhóm cấu trúc domain và có mức độ biểu hiện cao nhất so với các gene còn lại trong nhóm được trình bày ở Bảng 3.9.

Bảng 3.9. Dự đoán mức độ biểu hiện của gene mã hóa cellulase trong E. coli

Enzyme	**Loại domain**	**Mã gene đại diện có mức độ biểu hiện cao nhất trong nhóm**	**Mức độ biểu hiện (mg/l)**
Endoglucanase	GH8	GL0183420	3739
		GL1155166	3726
		GL0051672	3622
		GL0127466	3201
		GL0176868	3196
		GL0791089	2806
		GL0946225	2752
		GL0565361	2497
		GL0613574	2367
		GL0699893	2020
	GH5	GL0285761	3199
		GL0361483	2785
		GL0599940	2613
		GL0472979	2366
		GL0309031	2246
	PeptidaseM42	GL2894807	1067
	GH5-CBM6	GL0472979	2366
	DUF285	GL0168545	5382
		GL0614297	5243
		GL0239003	5012
		GL0188991	4881
		GL0652637	4856
		GL0001438	4375
		GL0042321	3391
	GH18	GL0560255	36
	CE2	GL0144694	15
	43 loại khác		
Exoglucanase	Alginate_lyase	GL0212614	743
	Amidase 3	GL0221923	9
	CBM2	GL2034110	9
	CBP_BcsO	GL0879211	15
	GH128-Laminin G3	GL0058533	14

β-glucosidase	GH3+FN3	GL0554917	4268
	GH1	GL0186901	2849
	FN3	GL2121620	2320
	GH3	GL0173907	3608
		GL0336364	2302
		GL0524609	1912
		GL0168583	1911
		GL0168583	1809
	GH43	GL1276531	2478
		GL0801723	2430
		GL1531450	1203
	GH3+Exop_C	GL0050362	4626
	DUF4886	GL0245593	22
	CE3	GL0464911	24
	LZ_Tnp_IS481+rve	GL0280494	25
	rve	GL1394039	22
		GL0888773	20
	Exoendophos	GL0437370	22
	GH16	GL0003443	2100
	LipaseGDSL2	GL0143432	27
	ExopC	GL1796064	1827
	HTH29+rve	GL1261227	22
	LacI+Peripla_BP_3	GL0732032	20
	CBM32	GL0418067	25
	GH89	GL0037389	2329
	rve3	GL1983913	19
	AP_endonuc2	GL0596682	40
	CBM32+GH55	GL0415923	15
	GH87	GL1311891	16
	GH16+CBM32	GL0278102	21
	GH43+CBM32+LaminiG3	GL0130082	39
	GxDLY+Lipase_GDSL_3+CE3	GL0475588	11
	HTH1+GH1	GL0975522	18
	SpoIIM	GL1042070	14
6-phospho β-glucosidase	GH1	GL0494307	4714
	GH4	GL0335762	1549
		GL0413390	1093
		GL0436665	1078

Kết quả xác định mức độ biểu hiện ở *E. coli* cho thấy trong 1058 gene hoàn chỉnh có domain mã hóa cellulase, các gene thuộc nhóm endoglucanase và β-glucosidase được xác định có khả năng biểu hiện cao hơn nhóm exoglucanase. Trong nhóm endoglucase, các gene chứa domain GH8 được dự đoán có mức biểu hiện cao

nhất với các mã GL0183420, GL1155166, GL0051672, GL0127466, GL0176868 đều có khả năng biểu hiện trên 3000 mg/l. Các gene này thuộc ngành Proteobacteria và Acidobacteria. Tiếp theo là các gene chứa domain GH5 thuộc ngành Proteobacteria và Bacteroidetes có mức biểu hiện trên 2000 mg/l. Ngoài ra một số gene chứa domain PeptidaseM42, GH5-CBM6, DUF285 thuộc ngành Bacteroidetes cũng biểu hiện tốt trong hệ biểu hiện *E. coli*. Trong nhóm β-glucosidase, các gene có domain GH3 thuộc ngành Proteobacteria đều mức độ biểu hiện tốt: nhiều đại diện cấu trúc domain GH3+FN3 có mức độ biểu hiện cao trên 4000 mg/l, gene có domain GH3 có mức biểu hiện trên 1800 mg/l, mã gene GL0050362 với cấu trúc domain GH3+Exop_C có mức độ biểu hiện cao nhất 4626 mg/l. Bên cạnh đó, các gene β-glucosidase chứa domain GH4, GH43, GH1 thuộc ngành Proteobacteria, gene β-glucosidase chứa domain GH16 thuộc ngành Bacteroidetes cũng biểu hiện tốt. Nhóm 6-phospho β-glucosidase, gene có domain GH1 có mức độ biểu hiện cao nhất 4714 mg/l, một số gene chứa domain GH4 có mức độ biểu hiện trên 1000 mg/l.

3.3.3. Nghiên cứu lựa chọn gene mã hóa cellulase

Ở quanh nấm mục trắng và khu đất xung quanh diễn ra sự phân hủy cellulose mạnh có sự tham gia của nhiều nhóm vi sinh vật. Trong đó, các vi sinh vật phân hủy cellulose mạnh thường có nhiều loại enzyme với lượng β-glucosidase cao và các vi sinh vật cơ hội thường chỉ chứa β-glucosidase [137]. Trong quá trình phân hủy đó, β-glucosidase tham gia vào quá trình thủy phân các liên kết glucoside trong các đường đôi tạo sản phẩm là đường đơn, để các đường đôi này không ức chế ngược hai enzyme endoglucanase và exoglucanase, đảm bảo quá trình phân hủy cellulose được diễn ra thuận lợi.

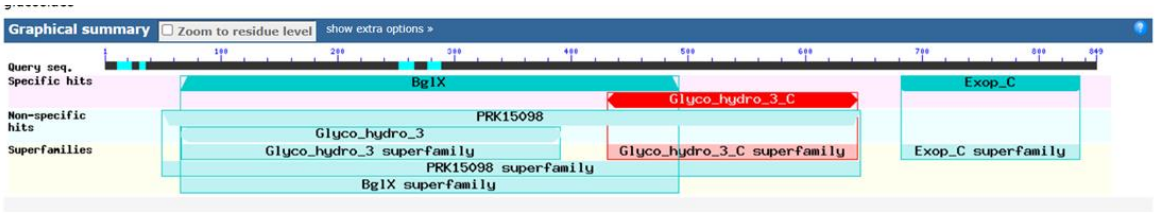
Dựa trên CSDL KEGG để tham chiếu, có 8301 gene được xác định mã hóa cellulase. Xét về phân loại ngành, Proteobacteria là ngành được xác định nhiều nhất 5123 gene (61,72%) trong các gene mã hóa cellulase. Đây có thể là đặc trưng về mặt loài của khu hệ vi sinh vật quanh khu nấm mục trắng ở vườn Quốc gia Cúc Phương. Xem xét về khía cạnh cấu trúc gene, trong 8301 gene được chú giải mã hóa cellulase có 1058 gene đầy đủ có domain thuộc 3 nhóm enzyme chính cần thiết để thủy phân hoàn toàn cellulose thành glucose là endoglucanase (367 ORF đầy đủ), cellobiohydrolase (6 ORF đầy đủ) và β-glucosidase (475 ORF đầy đủ). Trong khuôn

khổ của luận án, chúng tôi nghiên cứu lựa chọn một gene mã hóa β-glucosidase để tiến hành phân lập gene. Trong nhóm β-glucosidase, GH3 là cấu trúc domain phổ biến nhất (260 ORF) trong đó có 220 ORF có cấu trúc GH3+FN3, 11 ORF có cấu trúc GH3+Exop_C và 29 ORF chỉ chứa GH3. Các cấu trúc phụ trợ như FN3, Ig, CMC, Exop_C là độc lập và ít được nghiên cứu. Trong các nghiên cứu trước, vùng phụ trợ FN3 được coi là vùng đặc trưng của cellulase GH3 trong dạ cỏ dê với tỉ lệ 90,3% [136]. Nhằm tìm kiếm các gene mới, có đặc điểm khác biệt thì các gene mã hóa β-glucosidase có cấu trúc phụ trợ như Exop_C, FN3 sẽ được ưu tiên lựa chọn. Kết hợp với kết quả dự đoán mức độ biểu hiện, mã gene GL0050362 có cấu trúc GH3+Exop_C có mức độ biểu hiện cao nhất 4626 mg/l được lựa chọn. Vì vậy, trong nghiên cứu ở vi sinh vật xunh quanh nấm mục trắng này, mã gene có cấu trúc phụ trợ mới Exop_C được dự đoán tính chất bằng một số công cụ tin sinh.

3.3.3.1. Dự đoán vùng bảo thủ của gene bằng BLASTp

Việc tìm kiếm các vùng tương đồng giữa gene ứng viên với các trình tự khác đã được xác định chứa năng bằng thực nghiệm có vai trò quan trọng trong việc dự đoán chức năng gene. Thêm vào đó, khi sự tương đồng giữa gene ứng viên và các gene khác càng thấp thì khả năng chúng là gene mới, có nhiều tiềm năng trong nghiên cứu và ứng dụng càng cao. Tuy nhiên, nếu sự tương đồng là quá thấp thì có thể sẽ gặp khó khăn khi thực nghiệm biểu hiện. Vì vậy, các gene được lựa chọn thường là các gene có độ tương đồng khoảng trên 85% với CSDL khi so sánh. Kết quả khi so sánh gene GL0050362 với CSDL trên ngân hàng gene cho thấy gene GL0050362 có độ bao phủ 99-100% và độ tương đồng từ 96% trở lên với β-glucosidase của nhiều loại vi khuẩn như: *Stenotrophomonas maltophilia* (ID: VUR03699.1), *Stenotrophomonas sepilia* (ID: PZT38871.1), *Pseudomonas aeruginosa* (ID: CRP60742.1), *Pseudomonas hibiscicola* (ID: WP_019659734.1), *Stenotrophomonas* sp. HMSC10F07 (ID: oFU99884.1). Điều này cũng cho thấy tính khả thi khi tiến hành biểu hiện gene và nghiên cứu tính chất enzyme sau này. Dựa trên cơ sở dữ liệu NR, gene GL0050362 được dự đoán thuộc ngành Proteobacteria, lớp Gammaproteobacteria, Bộ Xanthomonadales, Họ Xanthomonadaceae, Chi *Stenotrophomonas*.

Kết quả xác định cấu trúc protein do gene GL0050362 mã hóa bằng BLASTp cho thấy protein này có ba vùng đặc hiệu (specific hit) gồm: (1) là vùng BglX (thuộc siêu họ BglX) tương ứng với hai vùng không đặc hiệu (non-specific hit) PRK15098 [138] và GH3-N (theo số liệu được liên kết với pfam00933) mã hóa β-glucosidase và các glycosidase tham gia vào quá trình trao đổi carbohydrate (theo cơ sở dữ liệu COG1472); (2) là vùng GH3-C (thuộc siêu họ GH3-C) tham gia vào quá trình xúc tác và có thể liên kết beta-glucan (theo số liệu được liên kết với pfam01915) [139]; (3) Exop_C (thuộc siêu họ Exop_C) giống vùng liên kết với Galactose, đây là vùng đầu C được tìm thấy trong ExoP (exo-1,3/1,4-beta-glucanase) từ *Pseudoalteromonas*. Vùng này chứa một nếp gấp β thường gặp trong glycosyl hydrolase (GH7, 11, 12 và 16) và trong một số vùng/cấu trúc liên kết carbohydrate. Vùng này được cho rằng không chỉ có vai trò định hướng liên kết với cơ chất mà còn giúp làm ổn định cấu trúc cần thiết cho hoạt động của ExoP [140] (Hình 3.6).



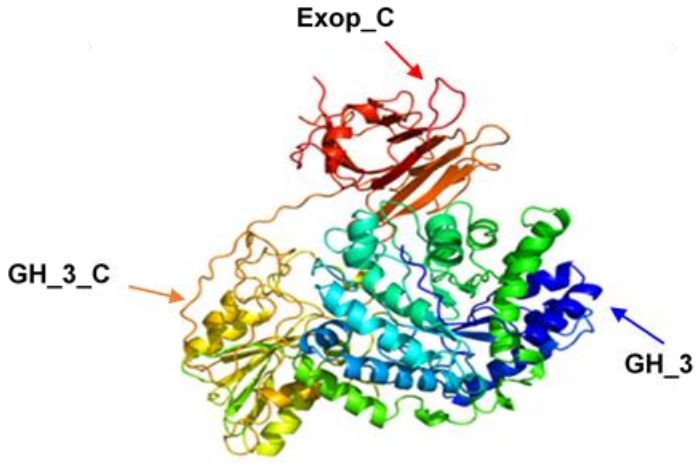
Hình 3.6. Kết quả dự đoán chức năng gen GL0050362 bằng BLASTp. Các vùng đặc hiệu (specific hit): BglX (COG1472), GH_3_C (pfam01915), Exop_C (pfam18559).

3.3.3.2. Dự đoán cấu trúc không gian của protein

Vì các cấp độ cấu trúc bậc cao của protein có xu hướng bảo thủ hơn nhiều so với trình tự amino acid của chúng trong quá trình tiến hóa nên các cấu trúc không gian của enzyme do gene GL0050362 mã hóa được khảo sát bằng phần mềm Phyre2. Kết quả thu được trong cấu trúc bậc 2 của enzyme có 29% xoắn α, 21% xoắn β và 16% không xác định dạng cấu trúc. Phyer2 còn dự đoán cấu trúc không gian ba chiều của enzyme, từ đó có thể dự đoán sâu hơn về trung tâm hoạt động, vùng bảo tồn, vị trí xúc tác của enzyme.

Trong mô hình cấu trúc không gian ba chiều, protein được xác định dựa trên khuôn enzyme β-glucosidase từ *Pseudoalteromonas* sp. bb1 (c3f93D) có độ bao phủ

93% và độ tin cậy 100%. Cấu trúc không gian bậc 3 của gene này có 47% tương đồng với β-glucosidase của khuôn c3f93D với độ tin cậy 100%, có ba vùng đặc hiệu GH-3, GH-3-C và Exop_C, ngoài ra gene này có vùng bảo tồn cao [HIS]249 giống nhau giữa protein ứng viên và khuôn c3f93D, mặt khác enzyme ứng viên còn có vị trí xúc tác [GLY]848 liên quan đến hoạt tính β-glucosidase theo ước đoán của Phyer2 (Hình 3.7).



Hình 3.7. Mô hình cấu trúc không gian của gen ứng viên sử dụng Phyre2 dựa trên khuôn c3f93D

3.3.3.3. Dự đoán một số tính chất của enzyme ứng viên

Ngoài cấu trúc không gian thì một số đặc điểm ảnh hưởng đến khả năng xúc tác của enzyme như pH, nhiệt độ tối ưu cho hoạt động của enzyme cũng được dự đoán. Mỗi loại enzyme hoạt động tốt ở một điều kiện pH cụ thể, các enzyme ưa axit hoặc kiềm có nguồn gốc từ vi sinh vật ưa axit hoặc ưa kiềm có thể có nhiều ứng dụng trong công nghiệp sản xuất. Khi đưa trình tự amino acid của enzyme ứng viên dạng FASTA vào phần mềm xác định khả năng chịu axit/kiềm AcalPred thu được kết quả xác xuất enzyme chịu axit và enzyme chịu kiềm lần lượt là 0,507957 và 0,492043. Hai xác xuất này là gần giống nhau và enzyme ứng viên là enzyme có pH trung tính, hơi ngả axit. Kết quả này cùng phù hợp với các nghiên cứu trước đó cho rằng điều kiện pH 6,0-7,5 là pH tối ưu cho hoạt động của các enzyme β-glucosidase [141].

Nhiệt độ là một trong những yếu tố quan trọng ảnh hưởng đến khả năng xúc tác của enzyme. Việc ước đoán khả năng chịu nhiệt của enzyme không chỉ thuận lợi trong lựa chọn điều kiện cho thực nghiệm mà còn là cơ sở để lựa chọn các gene có thể ứng dụng trong thực tiễn sản xuất. Công cụ TBI xác định khả năng chịu nhiệt của

enzyme theo 3 mức: nếu Tm (temperature melting) > 1, enzyme chịu được nhiệt độ trên 65°C, nếu Tm từ 0-1 thì enzyme chịu được nhiệt độ là 55°C-60°C và Tm<0 thì nhiệt độ tối ưu cho hoạt động của enzyme là dưới 55°C. Kết quả xác định khả năng chịu nhiệt của protein ứng viên có Tm là 0,6606, như vậy nhiệt độ tối ưu cho hoạt động của enzyme từ 55°C-65°C.

Dựa trên các kết quả xác định vùng hoạt tính, mức độ biểu hiện trong hệ biểu hiện *E. coli*, khảo sát vùng bảo thủ và cấu trúc không gian cũng như một số tính chất của protein suy diễn, mã gene GL0050362 đã được lựa chọn để biểu hiện và nghiên cứu tính chất của enzyme. Mã gene GL0050362 được kí hiệu là gene *gh3s2*.

3.4. Biểu hiện, tinh chế và nghiên cứu tính chất protein GH3S2

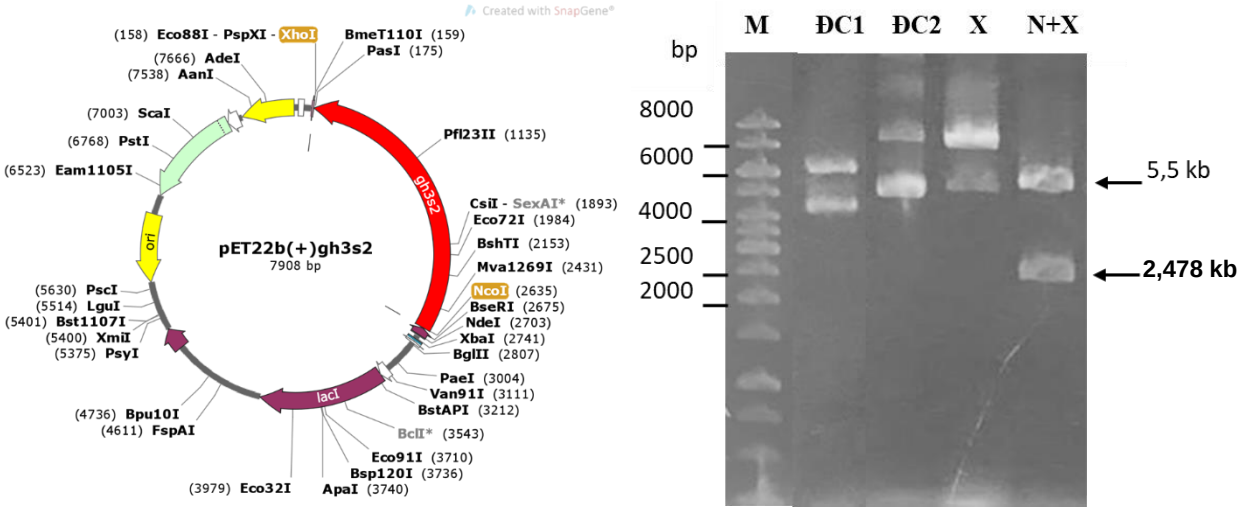
3.4.1. Nghiên cứu biểu hiện gene gh3s2

3.4.1.1. Thiết kế vector tái tổ hợp pET22b(+) mang gene gh3s2

Gene *gh3s2* ban đầu được khai thác từ DNA đa hệ gene của vi sinh vật đất có chiều dài 2547 bp mã hóa protein có 849 amino acid. Protein này chứa đoạn peptide tín hiệu tiết dài 26 amino acid (từ amino acid 1 đến amino acid 26) theo phần mềm trực tuyến Phobius dự đoán. Đoạn tín hiệu tiết này có vai trò quan trọng trong việc chuyển vị các protein đã được tổng hợp ra ngoài màng sinh chất ở các nhóm sinh vật [142]. Ở sinh vật nhân sơ như *E. coli*, phần lớn sự chuyển vị của protein chưa cuộn xoắn, gấp khúc là qua kênh Sec. Các protein này thường chứa tín hiệu kị nước tại đầu N của chúng [143]. Một con đường khác là con đường chuyển vị arginine đôi (twin-arginine translocation – Tat) trong đó đầu N của tín hiệu tiết có chứa motif đặc trưng Arg-Arg. Đây là con đường chuyển các protein đã gấp cuộn sau dịch mã. Trong nghiên cứu này, gene *gh3s2* được chúng tôi thiết kế không bao gồm trình tự mã hóa cho tín hiệu tiết. Như vậy, sau khi loại bỏ trình tự này thì gene *gh3s2* có kích thước là 2483 bp tương ứng sẽ tổng hợp protein có kích thước là 91,04 kDa. Trình tự của gene *gh3s2* sau khi được xác định các mã hiếm và tối ưu mã bộ ba được trình bày trong phụ lục 2 (Phụ lục 2: Trình tự nucleotide của gene *gh3s2* sau khi tối ưu mã bộ ba và trình tự amino acid tương ứng).

Gene *gh3s2* sau khi được lựa chọn và tối ưu mã sẽ được đặt tổng hợp và được chèn vào vectơ biểu hiện pET22b(+). DNA plasmid tái tổ hợp pET22b(+)gh3s2 đã được tách dòng trong chủng tách dòng *E.coli* DH10b. Nhằm khẳng định gene *gh3s2*

trong DNA plasmid tái tổ hợp pET22b(+)gh3s2, vectơ này sẽ được cắt kiểm tra bằng enzyme cắt hạn chế. Các vị trí enzyme cắt hạn chế trên vector được kiểm tra bằng phần mềm trực tuyến http://www.restrictionmapper.org/ cho thấy *Nco*I chỉ có 1 vị trí cắt và *Xho*I chỉ có 1 vị trí cắt. Các vị trí cắt của enzyme cắt hạn chế trên pET22b(+)gh3s2 được thể hiện trên hình 3.8A. Điện di đồ kiểm tra sản phẩm cắt vectơ tái tổ hợp pET22b(+)gh3s2 được thể hiện trong hình 3.8B.



Hình 3.8. (A). Các vị trí cắt của enzyme cắt hạn chế trên pET22b(+)gh3s2. (B). Điện di đồ sản phẩm cắt vectơ tái tổ hợp pET22b(+)gh3s2. ĐC1: vector không mang gen pET22b(+); ĐC2: vector tái tổ hợp pET22b(+)gh3s2; X: sản phẩm cắt vector tái tổ hợp bằng XhoI; N+X: sản phẩm cắt bằng tổ hợp NcoI và XhoI

Trên điện di đồ thấy đường chạy ĐC1 là vectơ không mang gene có kích thước nhỏ hơn so với đường chạy ĐC2 là plasmid tái tổ hợp có mang gene *gh3s2*. Điều đó chứng tỏ ở plasmid tái tổ hợp đã tách chiết đã được chèn thêm gene *gh3s2* nên có kích thước lớn hơn. Khi cắt vector tái tổ hợp với một enzyme cắt hạn chế *Xho*I thì thu được plasmid mở vòng có kích 7,978 kb (đường chạy X). Khi cắt vector tái tổ hợp bằng tổ hợp hai enzyme cắt hạn chế *Nco*I và *Xho*I (đường chạy N+X) sẽ thu được hai đoạn gồm một đoạn có kích thước 2,478 kb chính là gene *gh3s2* và một đoạn pET22b(+) có kích thước 5,5 kb. Trên điện di đồ thấy sản phẩm cắt vector tái tổ hợp bằng một và hai enzyme cắt hạn chế đều thu được các băng DNA có kích thước đúng. Như vậy, vector tái tổ hợp pET22b(+)gh3s2 đã được tổng hợp thành công.

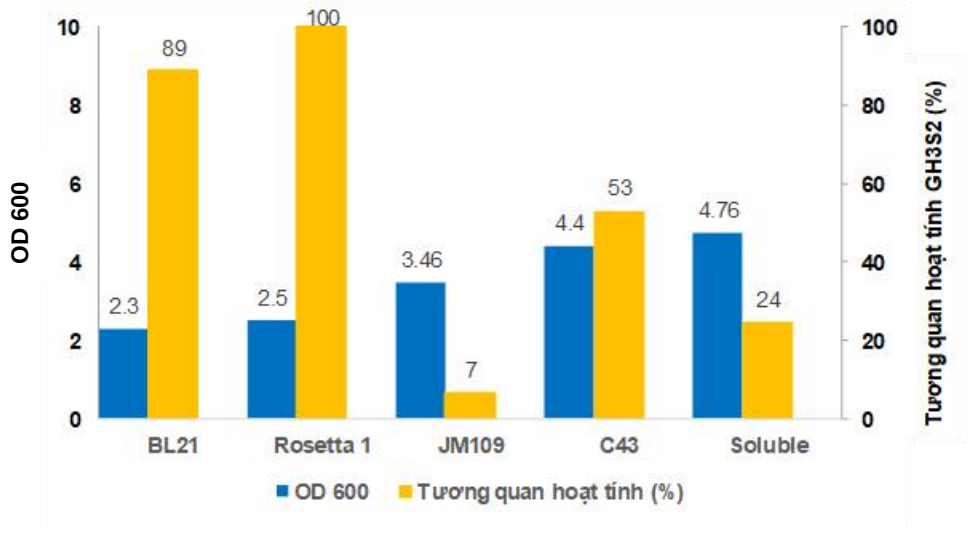
3.4.1.2. Nghiên cứu lựa chọn chủng biểu hiện protein GH3S2

Để biểu hiện được các protein tái tổ hợp cần có các vật chủ phù hợp. *E. coli* là một trong những vật chủ được thường lựa chọn để biểu hiện protein ngoại lai. Việc sử dụng *E. coli* làm vật chủ để biểu hiện protein đã được thực hiện từ lâu và nó đã trở thành nền tảng biểu hiện phổ biến nhất vì các lý do sau: (i) *E.coli* có tốc độ sinh trưởng nhanh, vòng đời ngắn, (ii) mật độ tế bào khi nuôi cấy có thể đạt được cao dẫn đến hàm lượng protein được biểu hiện lớn (iii) DNA ngoại lai được biểu hiện dễ dàng và hiệu quả [144]. Tuy nhiên, tùy thuộc vào gene ngoại lai, vào các chủng biểu hiện khác nhau và mức độ phù hợp của gene với vật chủ mà mức độ biểu hiện của gene và trạng thái hoạt động của gene là khác nhau.

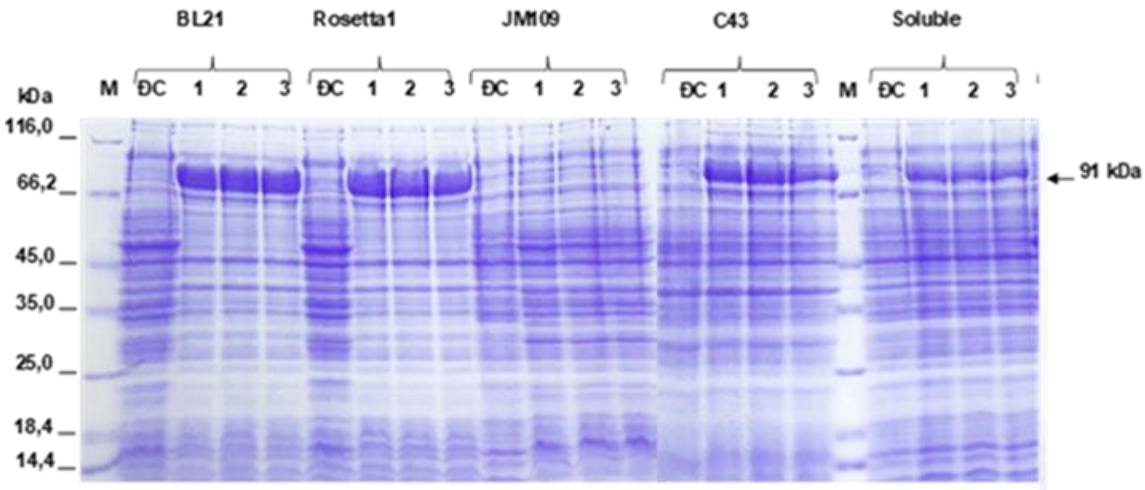
Protein tái tổ hợp GH3S2 được nghiên cứu biểu hiện trong 5 chủng *E. coli* khác nhau gồm: BL21, Rosetta 1, JM109, C43, Soluble. Ở chủng *E. coli* BL21, Rosetta 1, C43 và Soluble (DE3) có một số gene đã bị gây đột biến để quá trình biểu hiện gene được diễn ra thuận lợi: loại bỏ gene *omp*T để cho protein ngoại lai không bị phân hủy trong tế bào vật chủ và hàm lượng các protein tái tổ hợp vẫn giữ cấu trúc không gian được tăng cường từ đó giữ được hoạt tính sinh học, thích hợp biểu hiện protein ngoại lai ở mức độ cao [145], thiếu gene *hsd*S có chức năng mã hóa protease phân giải plasmid ngoại bào xâm nhập vào tế bào chủ, làm tăng hiệu quả biến nạp các DNA plasmid tái tổ hợp vào tế bào vật chủ, thiếu gene *gal* có chức năng kích thích tế bào sử dụng nguồn carbon là galactose cho các hoạt động sinh trưởng phát triển [146], thiếu gene *dcm* giúp DNA ngoại lai không bị methyl hóa cytosine thứ hai trong trình tự 5'-CC (A/T) GG-3', do đó protein ngoại lai được biểu hiện chính xác trong vật chủ. Ngoài ra, chủng *E. coli* Rosetta 1 còn có mang plamid pRARE mã hóa tRNAs cho các protein của sinh vật nhân chuẩn có chứa các bộ ba hiếm được sử dụng trong *E. coli* như AUA, AGG, AGA, CUA, CCC, GGA trên một plasmid kháng chloramphenicol tương ứng [145]. Chủng *E. coli* JM109 sinh trưởng tốt, có thể được biến nạp hiệu quả bằng nhiều phương pháp khác nhau và có đột biến endonuclease A^- dẫn đến tăng hiệu quả biểu hiện của DNA plasmid tái tổ hợp trong tế bào vật chủ. Chủng *E. coli* Soluble làm tăng khả năng biểu hiện ở dạng hòa tan của protein đích đặc biệt là các protein có nguồn gốc từ động vật có vú.

Gene *gh3s2* trong plasmid tái tổ hợp pET22b(+)gh3s2 được điều khiển phiên mã bởi promoter T7 bacteriaphage. Sự biểu hiện của gene đích *gh3s2* được cảm ứng

bởi nguồn T7-RNA polymerase từ vật chủ *E. coli*. Khi được cảm ứng đầy đủ, gần như tất cả các thành phần của tế bào vật chủ tập trung cho việc biểu hiện protein ngoại lai, sản phẩm protein mong muốn có thể đạt được 50% protein tổng số của tế bào vật chủ một thời gian ngắn sau cảm ứng. Promoter T7 cũng được cảm ứng bởi hợp chất isopropyl β-D-1-thiogalactopyranoside (IPTG) với hàm lượng thích hợp khi bổ sung vào môi trường nuôi cấy. Đây là một chất chuyển hóa lactose, kích hoạt promoter T7 và hoạt động phiên mã của operon lac và do đó nó được sử dụng để cảm ứng tạo sự biểu hiện protein.



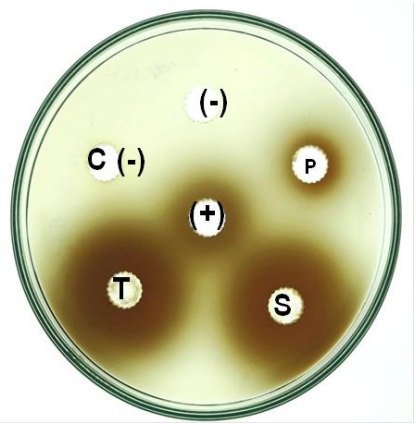
A



B

Hình 3.9. (A). Mật độ tế bào và hoạt tính của enzyme thu được khi biểu hiện trong các chủng biểu hiện; (B). Điện di đồ GH3S2 tổng số, ĐC: protein tổng số của đối chứng vector không mang gen; 1, 2, 3: protein tổng số các dòng khác nhau 1, 2, 3 mang gen gh3s2 cảm ứng IPTG; M: protein chuẩn (Thermo Sciencetific)

Kết quả kiểm tra sự biểu hiện của protein GH3S2 cho thấy trong các chủng BL21, Rosetta 1 protein GH3S2 được biểu hiện hiệu quả cao, lượng protein thu được là nhiều nhất được thể hiện bằng việc xuất hiện băng protein đậm tương ứng với kích thước của GH3S2 là 91 kDa. Khả năng biểu hiện của GH3S2 trong hai chủng C43 và Soluble thấp hơn rõ rệt, chủng JM109 gene không được biểu hiện (Hình 3.9B). Trong 2 chủng có mức độ biểu hiện gene tốt là BL21, Rosetta 1 thì ở chủng Rosetta có mật độ tế bào khi thu mẫu cao hơn, hoạt tính tương đối của enzyme tổng số khi biểu hiện ở chủng Rosetta cao hơn (Hình 3.9A). Dựa trên OD khi thu mẫu, hàm lượng tương đối của protein GH3S2 trên điện di đồ và hoạt tính của enzyme thu được, chủng Rosetta 1 được lựa chọn làm chủng biểu hiện enzyme GH3S2.



Hình 3.10. Kiểm tra hoạt tính của GH3S2 trên đĩa thạch LB sử dụng cơ chất esculin. T: protein tổng số, S: protein pha tan, P: protein pha không tan, C-: protein tổng số của pET22b(+), -: đệm, +: cellulase 0,05U

Hoạt tính β-glucosidase của enzyme GH3S2 sau khi biểu hiện trong môi trường LB có bổ sung 100 µg/ml ampicillin, cảm ứng 0,5 mM IPTG, nuôi lắc 200 vòng/phút ở 30°C trong 4 giờ với chủng biểu hiện *E. coli* Rosetta 1 cũng được kiểm tra với cơ chất esculin theo phương pháp của Vena và cộng sự (2011) [106]. β-glucosidase phân cắt esculin để tạo ra esculetin và glucose. Sau đó, sản phẩm esculetin khử các ion sắt trong môi trường để tạo ra sắt dẫn đến màu nâu và các vòng sáng sẽ được quan sát thấy trên đĩa cơ chất sau khi ủ qua đêm.

Kết quả xác định hoạt tính protein GH3S2 trên đĩa thạch cho thấy kích thước vòng màu nâu đối với các mẫu khác nhau có mức độ hoạt động β-glucosidase khác nhau. Mẫu protein tổng số và mẫu dịch pha tan có hoạt tính mạnh hơn so với pha

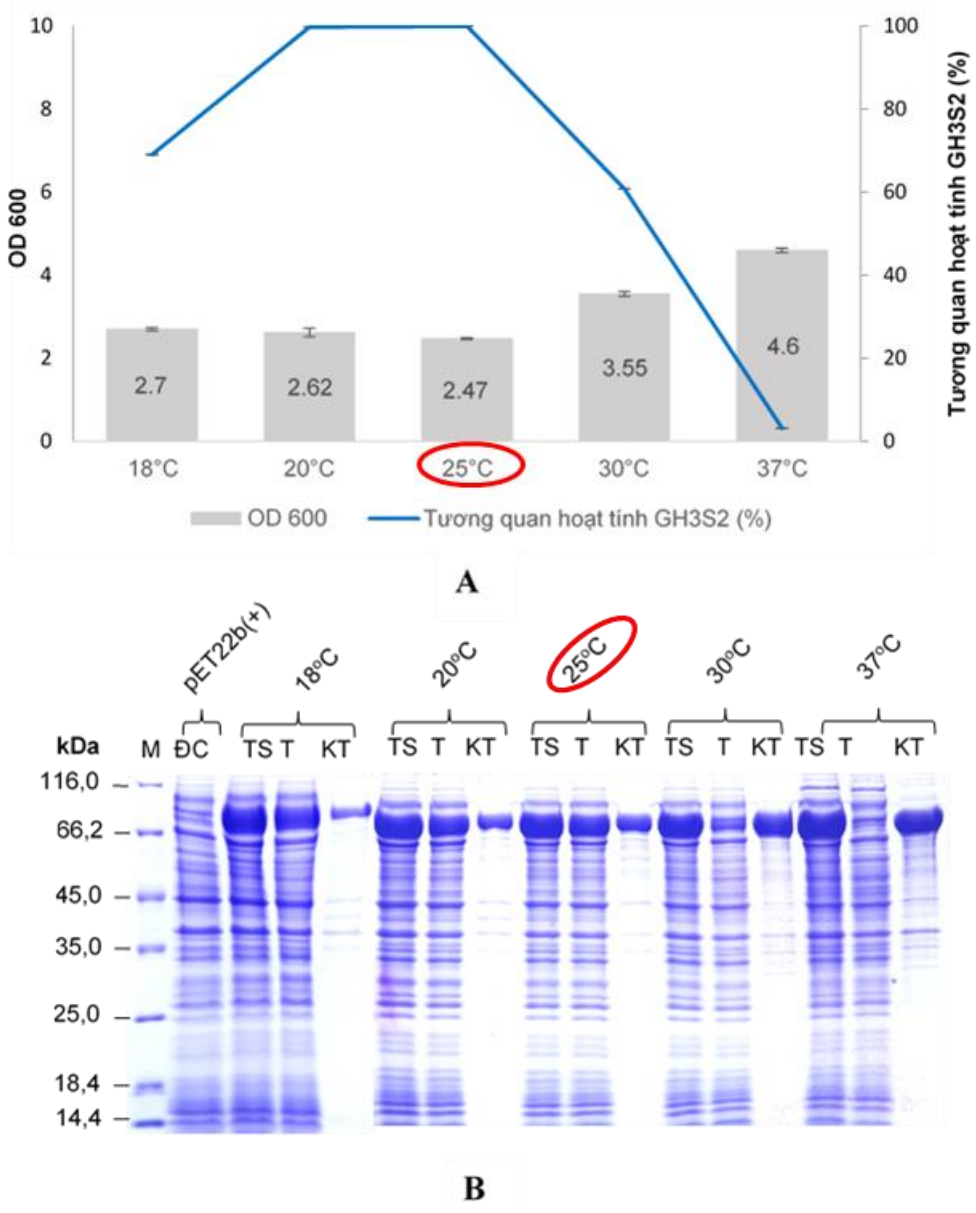
không hòa tan và các mẫu đối chứng âm. Đường kính của vòng màu nâu được tạo ra bởi hoạt tính của β-glucosidase trong mẫu protein tổng số và mẫu protein pha tan là khoảng 4 cm. Điều này chứng tỏ rằng protein GH3S2 đã được biểu hiện thành công ở dạng hòa tan và thể hiện hoạt tính β-glucosidase khá tốt (Hình 3.10).

3.4.1.3. Nghiên cứu ảnh hưởng của nhiệt độ nuôi cấy đến sự biểu hiện của GH3S2 trong E.coli Rosetta 1

Để thu được hàm lượng protein GH3S2 cao và enzyme có hoạt tính tốt, các điều kiện ảnh hưởng đến sự biểu hiện của gene được khảo sát để xác định các điều kiện tối ưu. Trong đó, nhiệt độ khi nuôi cấy là một trong những yếu tố ảnh hưởng lớn đến tốc độ sinh trưởng của *E. coli* từ đó ảnh hưởng đến hàm lượng protein tái tổ hợp biểu hiện được. Nhiệt độ thuận lợi cho sự sinh trưởng của *E. coli* là 37°C và protein GH3S2 cũng sẽ được biểu hiện tốt nhiều ở nhiệt độ này. Tuy nhiên, thông thường sự biểu hiện protein ở nhiệt độ cao thường dẫn tới việc hình thành các protein không tan hoặc các protein có hoạt tính kém [147]. Điều này là do ở nhiệt độ cao, protein được tổng hợp nhanh dẫn đến hàm lượng trong tế bào cao, protein chưa kịp cuộn xoắn đúng cấu trúc trong khi đó các liên kết kị nước được hình thành mạnh mẽ và chính các liên kết đó đã dẫn đến sự kết tủa của protein (inclusion body) chiếm ưu thế hơn so với sự gấp cuộn đúng cấu trúc của protein. Một số protein phức tạp, protein ở sinh vật nhân chuẩn yêu cầu thời gian lâu hơn để có thể gấp cuộn về đúng cấu trúc [148]. Một trong những biện pháp hiệu quả để giữ cho protein đích hình thành và cuộn xoắn đúng cấu trúc không gian, giảm sự kết tủa của protein và để thu được protein có hoạt tính là giảm nhiệt độ lên men [149]–[151].

Để nghiên cứu ảnh hưởng của nhiệt độ nuôi cấy đến sự biểu hiện của protein GH3S2, chủng *E. coli* Rosetta 1 mang DNA tái tổ hợp được cảm ứng 0,5mM IPTG, nuôi ở các điều kiện nhiệt độ là: 18°C, 20°C, 25°C, 30°C và 37°C. Kết quả mật độ tế bào cho thấy, khi nhiệt độ tăng thì mật độ tế bào cũng tăng và mật độ tế bào thu được lớn nhất ở 37°C (Hình 3.11 A). Điều này là phù hợp với quy luật sinh trưởng của vi khuẩn *E. coli*. Kết quả điện di kiểm tra protein pha tan khi biểu hiện ở các điều kiện nhiệt độ khác nhau cho thấy, ở cả năm điều kiện nhiệt độ đều xuất hiện băng đậm nét tương ứng với kích thước 91 kDa chứng tỏ protein GH3S2 đã biểu hiện tốt ở các nhiệt độ thí nghiệm. Ở 25°C lượng protein sau khi biểu hiện tồn tại pha tan là nhiều nhất,

ở nhiệt độ 18°C và 20°C lượng protein pha tan ít hơn. Khi nhiệt độ tăng lên đến 30°C thì lượng protein ở pha tan giảm dần và đến 37°C thì không thu được protein ở pha tan (Hình 3.11 B).

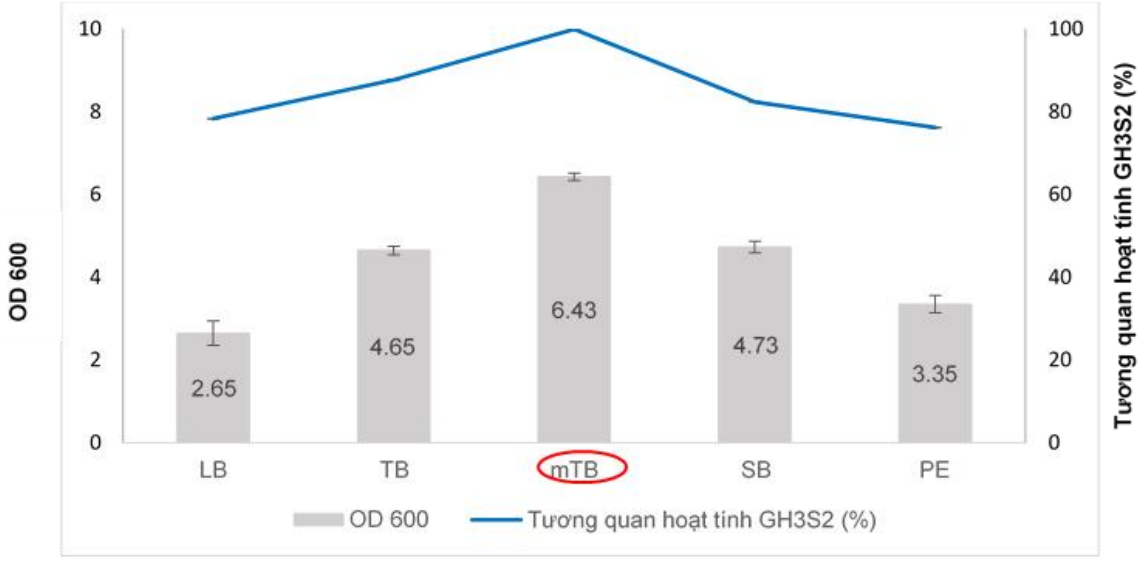


Hình 3.11. Ảnh hưởng của nhiệt độ đến mật độ tế bào, sự biểu hiện và hoạt tính của GH3S2. (A). Mật độ tế bào, hoạt tính GH3S2 ở các nhiệt độ khác nhau; (B). Điện di đồ GH3S2 biểu hiện. T, T, KT: protein tổng số, pha tan, pha không tan, C: mẫu đối chứng pET22b(+) không mang gen gh3s2, M: protein chuẩn (Thermo

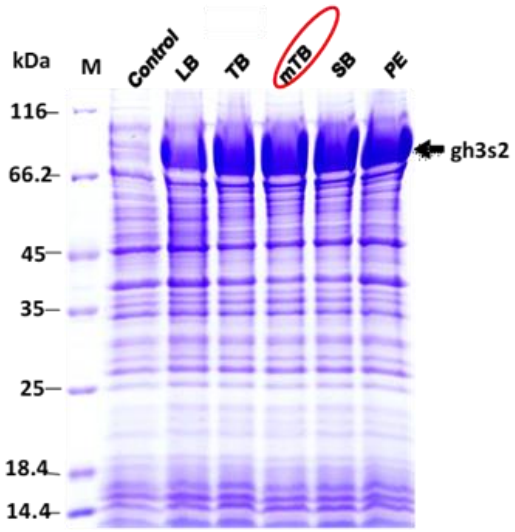
Protein sau khi được biểu hiện ở các điều kiện nhiệt độ khác nhau, pha protein tan sẽ được kiểm tra hoạt tính enzyme theo phương pháp của Dashtban và cộng sự (2010). Kết quả cho thấy ở 20°C và 25°C enzyme có hoạt tính cao nhất và tương đồng nhau. Kết hợp mật độ tế bào thu được, lượng protein ở pha tan, hoạt tính enzyme khi

biểu hiện (hình 3.11 A) và chi phí năng lượng, nhiệt độ biểu hiện cho protein GH3S2 được lựa chọn trong các nghiên cứu tiếp theo là 25°C. Các kết quả nghiên cứu trước đó cũng cho rằng khi giảm nhiệt độ biểu hiện protein có thể tăng khả năng thu được protein ở pha tan và các enzyme có hoạt tính sinh học cao [151], [152].

3.4.1.4. Nghiên cứu ảnh hưởng của thành phần môi trường nuôi cấy đến sự biểu hiện của GH3S2 trong E.coli Rosetta 1



A



B

Hình 3.12. Ảnh hưởng của môi trường nuôi cấy đến mật độ tế bào E. coli, sự biểu hiện và hoạt tính của GH3S2. (A). Mật độ tế bào và hoạt tính của GH3S2 khi nuôi cấy trong 5 môi trường khác nhau (B). Điện di đồ sản phẩm protein thu được.

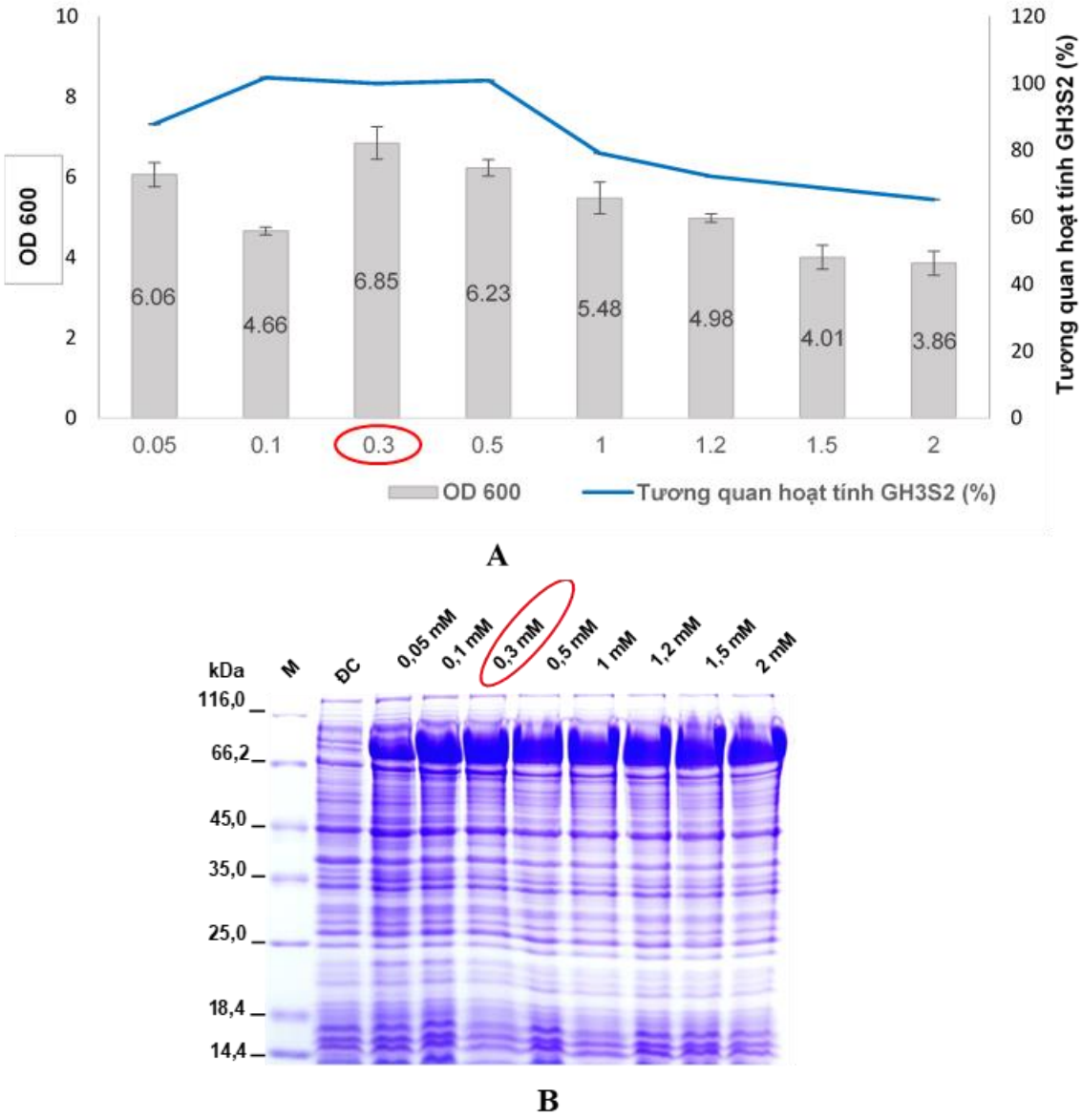
Môi trường nuôi cấy cung cấp chất dinh dưỡng cho sự sinh trưởng của các tế bào chủ vì vậy chúng có ảnh hưởng lớn đến tốc độ sinh trưởng của tế bào và hàm lượng protein ngoại lai thu được. Năm môi trường nuôi cấy đã được kiểm tra bao gồm LB, TB, TB cải biến, SB và PE. So với môi trường LB tiêu chuẩn, các môi trường khác đều làm tăng sự biểu hiện của protein GH3S2. Đặc biệt, khi glycerol trong môi trường TB chuẩn được thay thế bằng glucose trong môi trường TB cải biến thì mật độ tế bào thu được trong môi trường TB cải biến gấp 2,4 lần mật độ tế bào thu được từ môi trường LB chuẩn (Hình 3.12 A).

Kết quả kiểm tra lượng protein biểu hiện được cho thấy lượng protein GH3S2 được tạo ra từ các tế bào nuôi cấy trong môi trường TB cải biến cũng cao tương tự như khi nuôi cấy tế bào trong các môi trường TB, SB, PE (Hình 3.12 B). Điều này chỉ ra rằng nguồn carbon trong môi trường nuôi cấy có vai trò quan trọng, ảnh hưởng đến lượng protein GH3S2 thu được do nguồn carbon ảnh hưởng đến mật độ tế bào đạt được khi nuôi cấy [152]. Kiểm tra hoạt tính của protein thu được cho thấy trong môi trường TB cải biến, hoạt tính của GH3S2 là cao nhất (Hình 3.12 A). Kết hợp mật độ tế bào nuôi cấy, hàm lượng protein thu được và hoạt tính của protein (Hình 3.12A) thì môi trường TB cải biến là môi trường được lựa chọn sử dụng trong các thí nghiệm tiếp theo.

3.4.1.5. Nghiên cứu ảnh hưởng của nồng độ chất cảm ứng IPTG đến sự biểu hiện của GH3S2

Gene mã hóa protein GH3S2 được gắn vào plasmid pET22b(+) tạo ra DNA plasmid tái tổ hợp pET22b(+)gh3s2 hoạt động dưới sự kiểm soát bởi T7 promoter trên vector. Chất cảm ứng quá trình phiên mã và dịch mã tổng hợp protein GH3S2 là IPTG. Khi môi trường có đầy đủ chất cảm ứng, vật chủ *E. coli* tổng hợp T7 RNA polymerase và enzyme này bám vào vị trí T7 promoter khởi đầu sự phiên mã tổng hợp protein ngoại lai. Khi đó, hầu hết các thành phần của tế bào đều tập trung cho biểu hiện của protein. Vì vậy, lượng IPTG có vai trò quan trọng quyết định hiệu quả của quá trình biểu hiện. Tuy nhiên, IPTG là hóa chất độc hại cho tế bào vật chủ ở nồng độ cao và làm giảm hiệu quả biểu hiện protein đích [153], [154]. Hơn nữa, giá thành của IPTG khá cao. Vì vậy, để GH3S2 có thể biểu hiện hiệu quả với hàm lượng

cao, hoạt tính tốt thì cần khảo sát nồng độ chất cảm ứng IPTG tối ưu cho sự biểu hiện này.

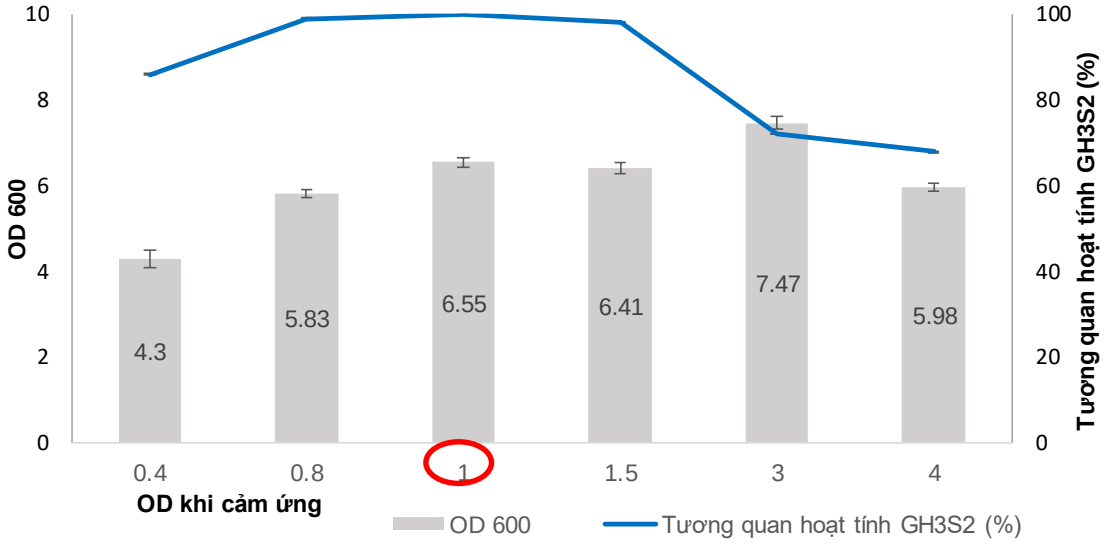


Hình 3.13. Ảnh hưởng của nồng độ IPTG đến mật độ tế bào, sự biểu hiện và hoạt tính của GH3S2. (A). Mật độ tế bào và hoạt tính GH3S2 khi cảm ứng IPTG có nồng độ khác nhau (B). Điện di đồ sản phẩm GH3S2 thu được.

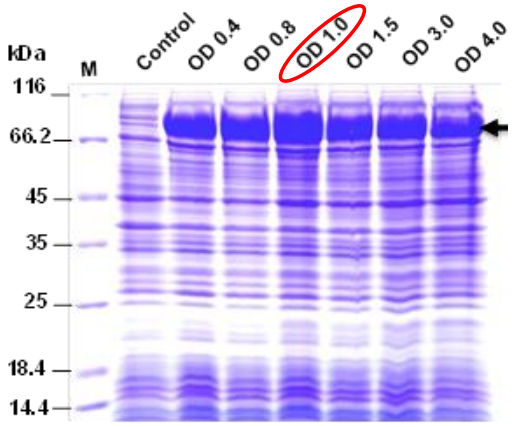
Để kiểm tra ảnh hưởng của IPTG đến hiệu quả khi biểu hiện GH3S2 thì các nồng độ IPTG từ 0,05 đến 2 mM được bổ sung vào dịch tế bào khi nuôi cấy. Mật độ tế bào thu được khi cảm ứng 4 giờ là tăng dần khi nồng độ IPTG tăng dần từ 0,05 đến 0,3 mM và sau đó mật độ tế bào giảm dần từ 0,5 mM IPTG (Hình 3.13 A). Lượng protein GH3S2 thu được cũng tăng dần từ 0,05 đến 0,3 mM và đạt cân bằng ở các nồng độ tiếp theo (Hình 3.13 B). Hoạt tính của GH3S2 ở nồng độ 0,1 và 0,3 mM là

tương đương nhau, sau đó giảm dần ở các nồng độ tiếp theo (Hình 3.13 A). Phối hợp các điều kiện về mật độ tế bào khi thu mẫu, mức độ biểu hiện và hoạt tính của protein đích, để đạt được hiệu quả biểu hiện tốt nhất thì IPTG có nồng độ 0,3 mM thích hợp cho biểu hiện protein GH3S2 và được sử dụng trong các nghiên cứu tiếp theo.

3.4.1.6. Nghiên cứu ảnh hưởng của mật độ tế bào khi cảm ứng đến sự biểu hiện của GH3S2



A



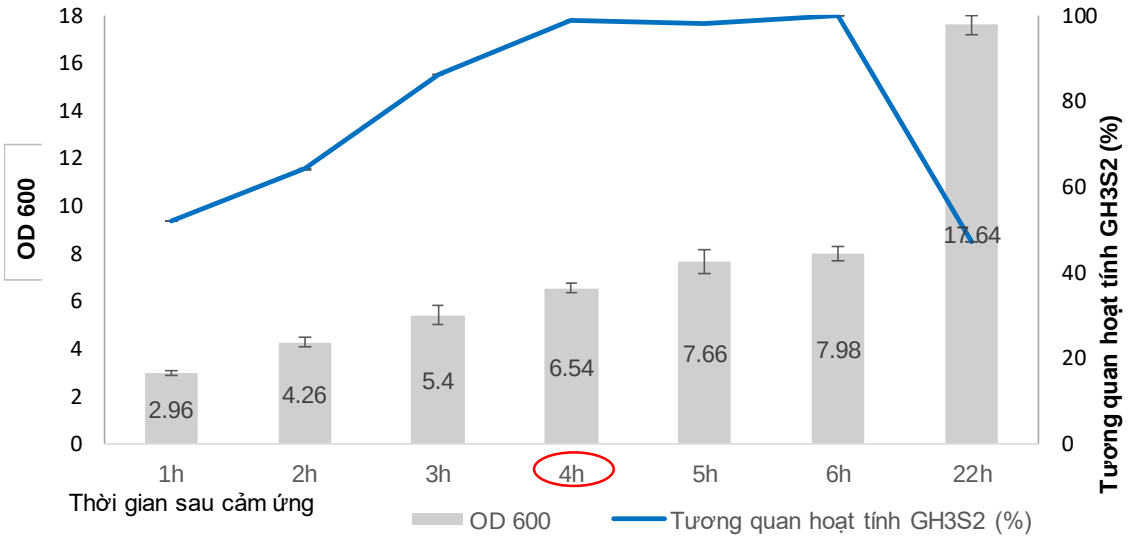
B

Hình 3.14. Ảnh hưởng của OD cảm ứng đến mật độ tế bào, sự biểu hiện và hoạt tính của GH3S2. (A). Mật độ tế bào và hoạt tính GH3S2 thu được khi cảm ứng ở các thời điểm khác nhau (B). Điện di đồ sản phẩm protein thu được

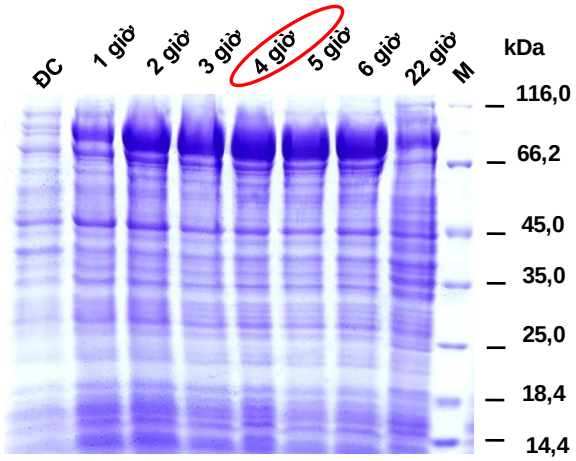
Kết quả thu các tế bào sau nuôi cấy cho thấy, khi mật độ tế bào lúc cảm ứng tăng lên thì mật độ tế bào khi thu mẫu cũng tăng nhanh, khi cảm ứng lúc mật độ tế

bào là 1 thì mật độ tế bào thu mẫu đạt được giá trị cao nhất, sau đó duy trì ổn định (Hình 3.14 A). Kết quả điện di và xác định hoạt tính sơ bộ cũng cho thấy cảm ứng lúc mật độ tế bào là 1 thì protein GH3S2 có mức độ biểu hiện tốt nhất và hoạt tính cao nhất (Hình 3.14 B). Vì vậy, để thu được sản phẩm biểu hiện GH3S2 cao nhất và hoạt tính tốt nhất thì mật độ tế bào của mẫu lúc cảm ứng là 1. Kết quả này cũng chỉ ra rằng cảm ứng ở giữa pha lũy thừa cho hiệu quả sản xuất protein tái tổ hợp cao nhất.

3.4.1.7. Nghiên cứu xác định thời gian thu mẫu GH3S2 tối ưu sau cảm ứng



A



B

Hình 3.15. Ảnh hưởng của thời gian sau cảm ứng đến mật độ tế bào, sự biểu hiện và hoạt tính GH3S2. (A). OD thu mẫu và hoạt tính của protein thu được khi cảm ứng các khoảng thời gian khác nhau. (B). Điện di đồ sản phẩm protein thu được.

Trong điều kiện nuôi cấy thích hợp, tế bào vật chủ có thể sinh tổng hợp protein đích ngay sau khi được cảm ứng. Theo thời gian, lượng protein này sẽ tăng lên và đạt tối đa vào một thời điểm nhất định. Sau đó, do chất dinh dưỡng giảm dần và các sản phẩm chuyển hóa tăng lên nên hiệu quả sinh tổng hợp protein ngoại lai có thể giảm. Như vậy, thời gian sau cảm ứng có ảnh hưởng đến mật độ tế bào khi thu mẫu, hoạt tính của protein GH3S2 thu được cũng như hiệu quả của quá trình biểu hiện.

Để xác định khoảng thời gian nuôi cấy sau cảm ứng thích hợp cho sự biểu hiện của protein GH3S2, mẫu lên men được tiến hành thu sau mỗi một giờ cảm ứng liên tục cho đến 6 giờ và 22 giờ sau khi cảm ứng. Kết quả xác định mật độ tế bào khi thu mẫu cho thấy, thời gian tăng lên từ 1 giờ đến 6 giờ thì mật độ tế bào thu mẫu cũng tăng lên sau đó chậm dần (Hình 3.15 A). Lượng protein ngoại lai GH3S2 được biểu hiện cũng tăng dần từ 1 giờ đến 4 giờ và duy trì ổn định đến 6 giờ (Hình 3.15 B). Hoạt tính của enzyme thu được cũng đạt cao nhất trong khoảng thời gian từ 4-6 giờ. Khi thời gian sau cảm ứng là 22 giờ thì lượng protein đích giảm mạnh và hoạt tính thấp (Hình 3.15 A). Tổng hợp kết quả xác định mật độ tế bào khi thu mẫu, hàm lượng protein thu được và hoạt tính protein cho thấy nuôi cấy tế bào sau khi cảm ứng 4 giờ sẽ cho hiệu quả biểu hiện GH3S2 tốt nhất.

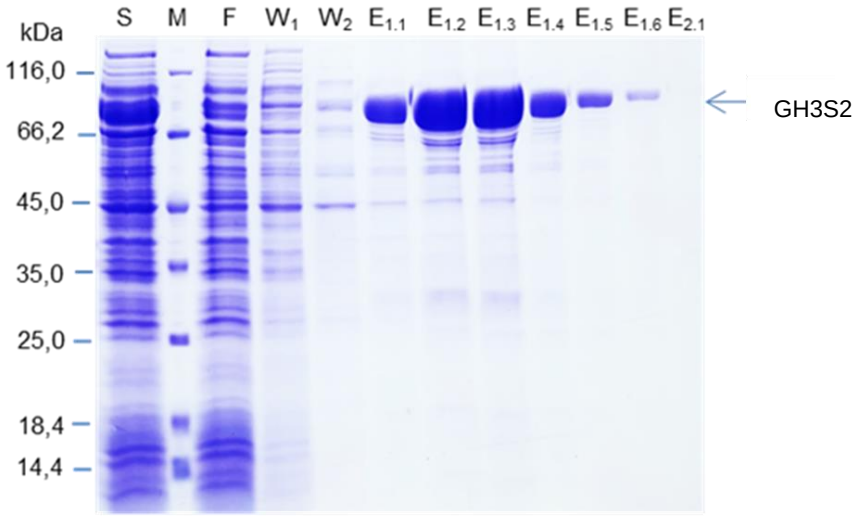
Như vậy, các điều kiện tối ưu để biểu hiện protein GH3S2 đã được xác định là biểu hiện trong chủng biểu hiện *E. coli* Rosetta 1 ở 25°C, môi trường TB cải biến, nồng độ chất cảm ứng là 0,3 mM IPTG, thời điểm cảm ứng khi mật độ tế bào là 1 và thu mẫu sau khi cảm ứng 4 giờ.

3.4.2. Tinh chế protein tái tổ hợp GH3S2 bằng cột sắc ký ái lực

Tinh chế enzyme là quá trình làm tinh khiết enzyme đích từ hỗn hợp enzyme ban đầu của tế bào. Thông thường, quá trình tinh chế sẽ dựa trên các đặc điểm sai khác của protein đích với các protein khác trong hỗn hợp như trọng lượng của protein, các đặc điểm hóa lý hay tương tác của protein với các chất khác. Theo thiết kế plasmid tái tổ hợp, gene *gh3s2* được ghép nối vào plasmid pET22b(+) là vector có thêm trình tự mã hóa 6 amino acid Histindin (His-tag) ở đầu 3’của gene. Vì vậy, protein GH3S2 có thể được tinh chế từ hỗn hợp protein bằng cách sử dụng cột sắc ký ái lực. Đó là do protein GH3S2 có đuôi his-tag có thể liên kết nhanh và mạnh với ion Ni^{2+} của cột sắc ký ái lực và được giữ lại trên giá thể. Các protein khác có các his nằm riêng rẽ và rải

rác trong protein nên ái lực kém với Ni^{2+}, khi hỗn hợp protein được bơm qua cột thì các protein này sẽ trôi qua giá thể.

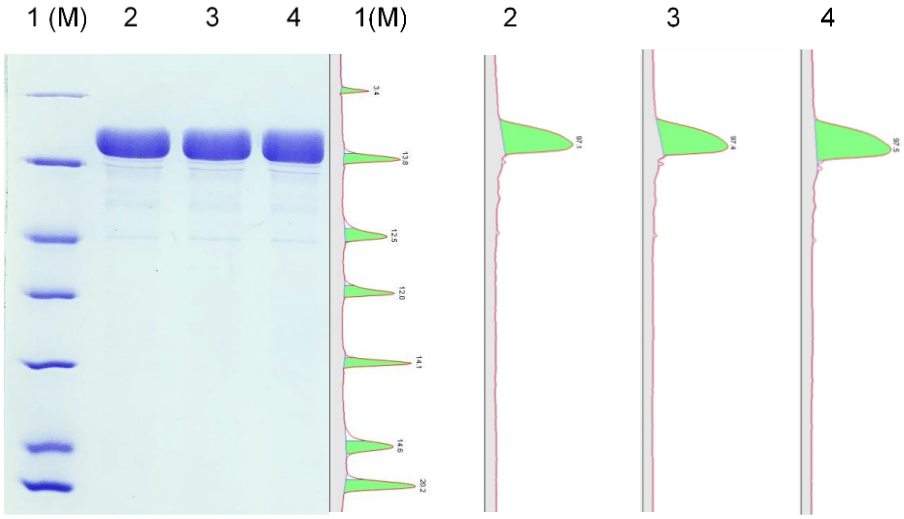
Kết quả điện kiểm tra sau tinh chế cho thấy protein GH3S2 bám cột rất tốt và được thôi ra khỏi giá thể khá tập trung khi sử dụng đệm có chứa nồng độ 300 mM imidazole. Protein GH3S2 được thu thành 6 phân đoạn, mỗi phân đoạn là 1 ml/1 eppendorf trong đó protein được thôi ra chủ yếu ở các phân đoạn 1, 2, 3, 4, đặc biệt là các phân đoạn 2 và 3. Ở các phân đoạn này, lượng lớn GH3S2 được thôi ra thể hiện bằng băng protein to và đậm nét (Hình 3.16). Các phân đoạn 2, 3, 4 được thu lại, trộn với nhau và được thẩm tích loại muối để không ảnh hưởng đến hoạt tính của enzyme. Protein GH3S2 sau tinh chế được sử dụng để xác định hàm lượng protein thu được trong 1 lít dịch lên men. Kết quả thu được hàm lượng protein GH3S2 trong mẫu tinh sạch là 1,54 mg/ml. Như vậy, trong 1 lít dịch khi lên men lượng GH3S2 tinh sạch thu được là 41,80 mg. Mẫu protein được đánh giá độ sạch cũng như sử dụng cho các thí nghiệm xác định đặc điểm của protein GH3S2.



Hình 3.16. Điện di đồ kiểm tra các phân đoạn trong tinh chế GH3S2 bằng cột sắc ký ái lực. S: Protein tổng số pha tan; M: thang protein chuẩn (Thermo Sciencetific, SM0431); F: dịch thu được khi bơm mẫu lên cột; W1, W2 lần lượt là dịch rửa cột với đệm PBS 50 mM, pH 7 có chứa 20, 50 mM imidazol; E1.1 – 1.6: các phân đoạn thu mẫu chứa 300 mM imidazol; E2.1: phân đoạn rửa cột

Để xác định độ sạch của protein GH3S2, chúng tôi sử dụng điện di SDS-PAGE với lượng mẫu xác định và phân tích kết quả bằng phần mềm Image Lab để đánh giá

độ sạch tương đối của protein GH3S2. Mẫu trong một đường chạy là 2 µg, thí nghiệm được lặp lại 3 lần (lane 2,3,4 Hình 3.17A). Mỗi băng trên bản điện di được phần mềm thể hiện bởi một đường cong tương ứng trên sơ đồ (Hình 3.17B, C). Mỗi đỉnh trên đường cong được phần mềm tự động tính lượng protein tương ứng. Phần mềm sẽ nhận biết và quét để định lượng tương đối protein tổng số, protein GH3S2 cũng được xác định bằng mức độ đậm của băng tương ứng. Tỉ lệ giữa mức độ đậm của băng GH3S2 so với toàn bộ các băng protein ở mỗi đường chạy được xác định là độ sạch của protein GH3S2. Kết quả thu được độ sạch của GH3S2 sau khi tinh chế là 97,3%, đạt tiêu chuẩn cho việc sử dụng để tiến hành xác định đặc điểm enzyme. Kết quả này cao hơn kết quả thu được khi tinh chế protein bglA từ *Bacillus polymyxa* (cho độ sạch là 92,7%) [155].



A **B**

Hình 3.17. Kết quả kiểm tra độ sạch GH3S2 sau tinh chế (A). Điện di đồ GH3S2 sau tinh chế (2 µg); (B, (C). Kết quả đo độ sạch bằng phần mềm Image Lab

Sau khi tinh sạch protein GH3S2, hoạt tính β-glucosidase của mẫu protein tổng số và mẫu GH3S2 đã tinh sạch sẽ được xác định. Kết quả thu được hoạt tính của protein tổng số là 0,156 ± 0.01 U/mg, hoạt tính của GH3S2 tinh chế là 1,10 ± 0.02 U/mg. Như vậy protein GH3S2 đã được tính chế 7,05 lần và hiệu suất tinh chế là 40,06% (Bảng 3.10).

Bảng 3.10. Bảng tổng kết hiệu suất tinh chế protein GH3S2 tái tổ hợp

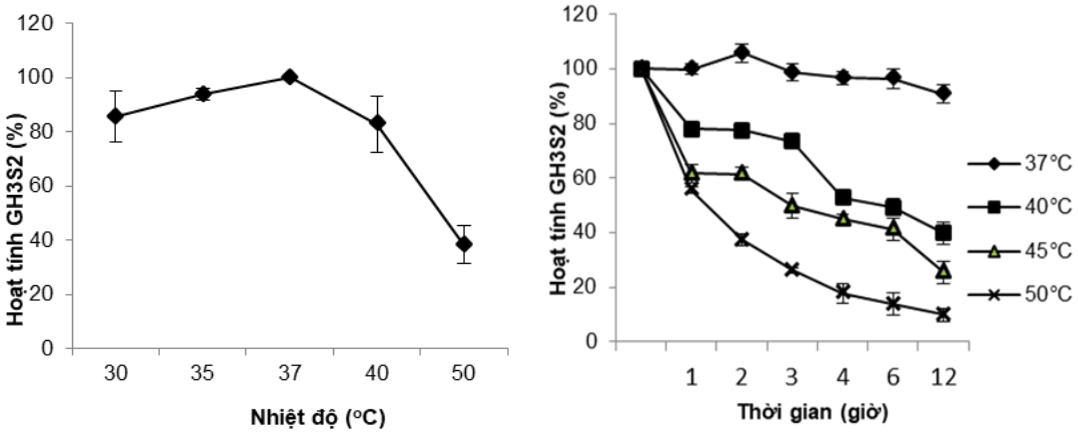
(*: tính trên 1 lit dịch nuôi cấy)

	Tổng lượng protein (mg)*	Tổng hoạt tính (U)*	Hoạt tính riêng (U/mg)	Độ sạch (lần)	Hiệu suất thu hồi (%)
Protein tổng số	735,68 ±0,6	114,77 ± 1,2	0,156 ± 0,01	1	100
GH3S2 tinh chế	41,80 ± 0,3	45,98 ± 0,8	1,10 ± 0,02	7,05	40,06

3.4.3. Nghiên cứu tính chất của protein tái tổ hợp GH3S2

3.4.3.1. Ảnh hưởng của nhiệt độ đến hoạt tính và độ bền nhiệt của GH3S2

Các vi sinh vật trong đất ở các khu rừng nhiệt đới như rừng quốc gia Cúc Phương thường ưa ấm, vì vậy mà enzyme GH3S2 của chúng nếu có cũng hoạt động tốt ở các điều kiện nhiệt độ từ 20°C đến khoảng 40°C. Do đó, để nghiên cứu ảnh hưởng của các điều kiện nhiệt độ khác nhau đến hoạt tính của enzyme GH3S2, thì các tác động của các nhiệt độ 30°C, 35°C, 37°C, 40°C, 50°C đến GH3S2 đã được khảo sát.



A **B**

Hình 3.18. Ảnh hưởng của nhiệt độ đến hoạt tính và độ bền nhiệt của enzyme GH3S2 theo thời gian

Kết quả thu được hoạt tính GH3S2 tăng lên khi nhiệt độ tăng từ 30°C đến 37°C sau đó giảm dần ở 40°C và giảm mạnh ở 50°C. Nếu coi hoạt tính của enzyme GH3S2 ở nhiệt độ tối ưu 37°C là 100% thì ở các nhiệt độ 30°C, 35°C, 40°C, 50°C hoạt tính của enzyme lần lượt là 82,19%, 93,66%, 80,97%, 37,26% (Hình 3.18 A). Như vậy, ở 37°C thì enzyme GH3S2 thể hiện hoạt tính cao nhất. So với kết quả dự đoán bằng

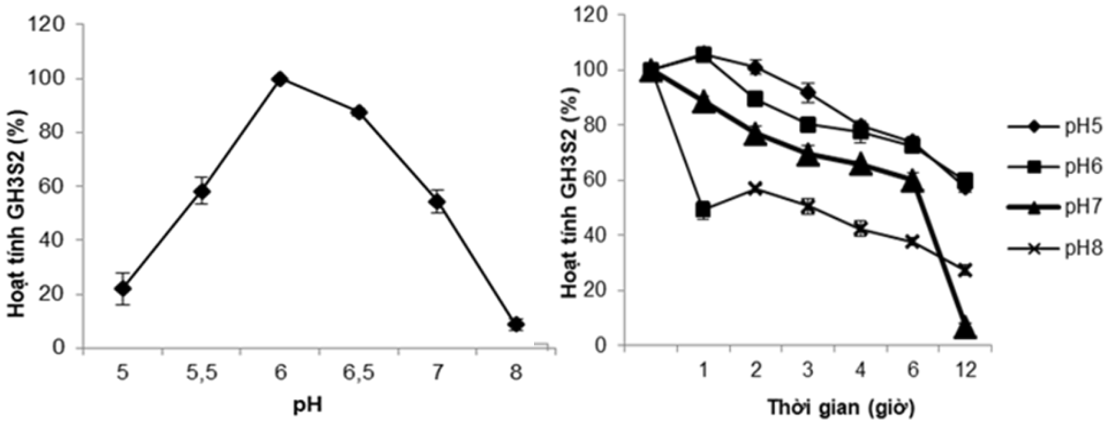
các phần mềm trực tuyến TBI thì nhiệt độ tối ưu cho hoạt động của enzyme là 55°C - 65° thì kết quả thu được ở mức nhiệt thấp hơn. Tuy nhiên, nhiệt độ này nằm trong khoảng nhiệt độ tối ưu của đa số cellulase của vi khuẩn (từ 35°C đến 50°C). Theo công bố của Gomes-Pepe và cộng sự, 37°C cũng là nhiệt độ tối ưu của enzyme β-glucosidase từ metageneome của vi khuẩn đất [156]. Nhiệt độ tối ưu này tương tự như GH3S2 ở dạ cỏ gia súc [157], *Proteus mirabilis* VIT117 nuôi trên vỏ tôm [158]. Tuy nhiên, nhiệt độ tối ưu cao hơn của β-glucosidase cũng được công bố: từ dịch tiêu hóa của ấu trùng đuông 55°C [159]; *A.fumigatus* Z5 là 60°C [160], β-glucosidase tái tổ hợp *Bgl.bli1* từ *Bacillus licheniformis* CGMCC 2876 hoạt động tối ưu ở nhiệt độ 60°C [161].

Để kiểm tra độ bền của enzyme GH3S2 với nhiệt độ, enzyme được xử lý ở các điều kiện nhiệt độ 37°C, 40°C, 45°C, 50°C trong các thời gian khác nhau 1, 2, 3, 4, 6 giờ và 12 giờ trước khi tiến hành các phản ứng xác định hoạt tính. Kết quả thu được cho thấy, ở điều kiện 37°C hoạt tính sinh học của enzyme GH3S2 khá ổn định, sau 12 giờ xử lý nhiệt độ này, hoạt tính của enzyme vẫn đạt 90,78%. Ở nhiệt độ 40°C, enzyme vẫn duy trì được khoảng trên 70% hoạt tính trong 3 giờ xử lý, từ giờ thứ 4 hoạt tính enzyme giảm nhanh. Ở các điều kiện 45°C, 50°C hoạt tính enzyme giảm gần như một nửa trong 1 giờ đầu, sau đó hoạt tính của enzyme giảm liên tục (Hình 3.18 B). Trong kết quả nghiên cứu của Lin Zhang và cộng sự (2017), cen 502 là một gene β-glucosidase từ vi khuẩn *Bursaphelenchus xylophylus* cũng chỉ bền ở nhiệt độ dưới 40°C [162].

3.4.3.2. Ảnh hưởng của pH đến hoạt tính và độ bền pH của enzyme GH3S2

pH là yếu tố quan trọng ảnh hưởng đến hoạt tính enzyme β-glucosidase. Để nghiên cứu pH ảnh hưởng đến hoạt tính của enzyme GH3S2, đệm với 6 giá trị pH khác nhau từ 5,0 đến 8,0 được sử dụng để pha loãng enzyme GH3S2 trước khi thực hiện phản ứng xác định hoạt tính. Kết quả thu được ở pH 6,0 enzyme GH3S2 thể hiện hoạt tính cao nhất. Kết quả thực nghiệm này đúng với dự đoán bằng công cụ tin sinh ban đầu là enzyme GH3S2 hoạt động tốt trong môi trường trung tính hơi ngả axit. Đây là giá trị pH tối ưu thường được công bố với những β-glucosidase vi khuẩn và GH3S2 cũng là enzyme vi khuẩn. Ở các điều kiện pH kiềm (8,0) hoặc axit (5,0) enzyme GH3S2 có hoạt tính bị giảm mạnh, chỉ đạt khoảng 20% hoạt tính ở điều kiện

tối ưu (Hình 3.19 A). Kết quả này cũng phù hợp với các công bố trước đó như enzyme β-glucosidase ở *F.oxysporum*, *Cellulomonas flavigenea*, *Clostridium thermocellum* có pH tối ưu cho hoạt động là 6,0 [11], [163]; β-glucosidase của *Bacillus licheniformis* có hoạt tính thấp ở môi trường axit và hoạt tính giảm mạnh ở pH 8,0 [161]; β-glucosidase của *Caulobacter crescentus* [164], enzyme β-glucosidase được phân lập từ vi khuẩn đất có hoạt tính thấp ở cả môi trường axit và kiềm [165]–[167]. Trong khi đó, các β-glucosidase có nguồn gốc từ nấm như *Aspergillus niger* thể hiện hoạt tính tối ưu ở pH 5,0 [168], *Sporiobolus pararoseus* hoạt tính cao ở pH 5,0 [169].



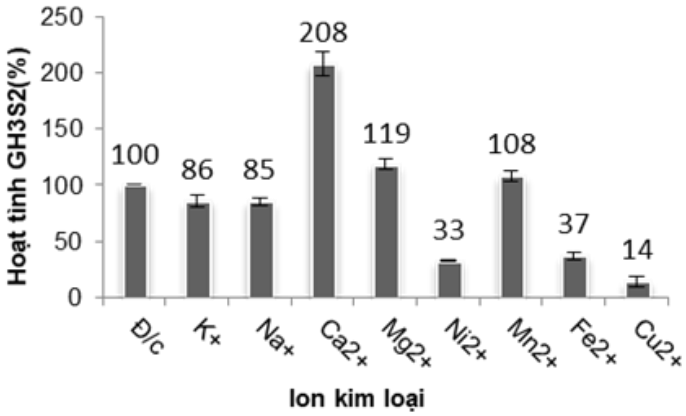
Hình 3.19. Ảnh hưởng của pH đến hoạt tính và độ bền pH của enzyme GH3S2 theo thời gian

Khi kiểm tra độ bền của enzyme với các điều kiện pH khác nhau, ở pH tối ưu 6,0 enzyme được duy trì sự ổn định khoảng 70% hoạt tính sau 6 giờ, ở pH trung tính 7,0 hoạt tính enzyme được duy trì 70% sau 4 giờ sau đó giảm xuống và pH càng cao hoạt tính của enzyme càng giảm (Hình 3.19 B). Trong thí nghiệm trước đó của Yin và cộng sự cũng cho thấy enzyme β-glucosidase có nguồn gốc từ vi sinh vật đất ở vùng cận nhiệt đới duy trì được 70% hoạt tính ở pH 6,0 – 7,5 [141].

3.4.3.3. Ảnh hưởng của một số ion kim loại đến hoạt tính của enzyme GH3S2

Các enzyme nói chung và enzyme thủy phân cellulose nói riêng rất nhạy cảm với các ion kim loại nặng. Các ion kim loại này tương tác và có thể liên kết với enzyme, từ đó làm thay đổi cấu trúc và hoạt tính của enzyme. Các ion kim loại có thể làm tăng hoặc giảm hoạt tính của enzyme tùy thuộc từng loại ion. Để nghiên cứu vai trò của ion kim loại trong việc thể hiện hoạt tính của enzyme GH3S2, trong nghiên cứu này 8 ion kim loại gồm 2 ion hóa trị I: K^+, Na^+ và 6 ion hóa trị II là Ca^{2+}, Mg^{2+},

Ni^{2+}, Mn^{2+}, Fe^{2+}, Cu^{2+} được sử dụng để xử lý enzyme với nồng độ cuối cùng là 1 mM. Kết quả xác định hoạt tính được thực hiện ở điều kiện tối ưu là 37°C, pH 6,0 trong thời gian 15 phút cho thấy, ion Ca^{2+}, Mg^{2+}, Mn^{2+} làm tăng hoạt tính của enzyme, trong đó có ion Ca^{2+} làm tăng mạnh mẽ hoạt tính enzyme của GH3S2 lên 2,08 lần và Mg^{2+} làm tăng thêm 1,19 lần so với mẫu không được bổ sung ion kim loại. Như vậy, nếu coi hoạt tính của mẫu không thêm ion kim loại là 100%, thì khi thêm Ca^{2+}, Mg^{2+}, Mn^{2+} hoạt tính của enzyme lần lượt là 208%, 119%, 108%. Trong khi đó, các ion Fe^{2+}, Ni^{2+}, Cu^{2+} làm giảm mạnh hoạt tính của enzyme xuống còn lần lượt là 37%, 33%, 14%. Các ion K^+, Na^+ ảnh hưởng không đáng kể đến hoạt tính của protein GH3S2, hoạt tính của enzyme sau khi xử lý ion K^+, Na^+ còn lại lần lượt là 86% và 85% (Hình 3.20). Điều này có thể do trung tâm xúc tác của GH3S2 chứa vị trí liên kết với các ion hóa trị II. Ảnh hưởng làm tăng hoạt tính GH3S2 của Ca^{2+} cũng đã được đề cập trong các nghiên cứu trước như: β-glucosidase của vi sinh vật đất cận nhiệt đới tăng hoạt tính lên 131% [141], β-glucosidase của *Streptomyces griseus* tăng hoạt tính lên 118% [170] khi bổ sung ion Ca^{2+}.

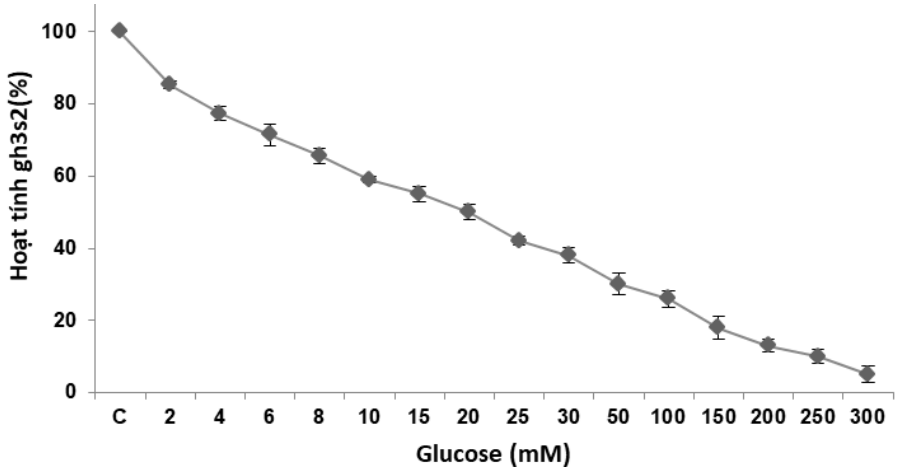


Hình 3.20. Ảnh hưởng của một số ion kim loại đến hoạt tính của GH3S2

3.4.3.4. Nghiên cứu ảnh hưởng của glucose đến hoạt tính của enzyme GH3S2

GH3S2 là một enzyme nhạy cảm với sự có mặt của glucose và thường bị giảm hoạt tính khi nồng độ của glucose tăng lên [171]. Việc tìm được GH3S2 có thể chịu được sự có mặt của glucose có ý nghĩa quan trọng khi phân giải cellulose trong công nghiệp sản xuất giấy, rượu, bia. Để khảo sát ảnh hưởng của glucose đến khả năng xúc tác của enzyme GH3S2, các nồng độ glucose từ 2-300 mM được bổ sung vào phản ứng xác định hoạt tính của GH3S2. Kết quả cho thấy khi bổ sung glucose đến nồng

độ 6 mM thì hoạt tính của enzyme GH3S2 chỉ bị ảnh hưởng ít và duy trì được khoảng 70% hoạt tính sau đó giảm xuống chỉ còn 6% khi nồng độ glucose tăng lên 300 mM (Hình 3.21). Điều này cho thấy, glucose và GH3S2 đã xảy ra tương tác cạnh tranh và glucose ức chế GH3S2 trong quá trình phân giải cơ chất pNPG [172]. Trong nghiên cứu trước đó của Chen và cộng sự (2017) cũng cho thấy khi nồng độ glucose khoảng 34 mM thì enzyme β-glucosidase từ *B. licheniformis* hoàn toàn mất hoạt tính [161], hay hoạt tính xúc tác của β-glucosidase từ nấm *Gongronella butleri* bị ức chế 50% khi hàm lượng glucose trong môi trường là 10 mM [70]…



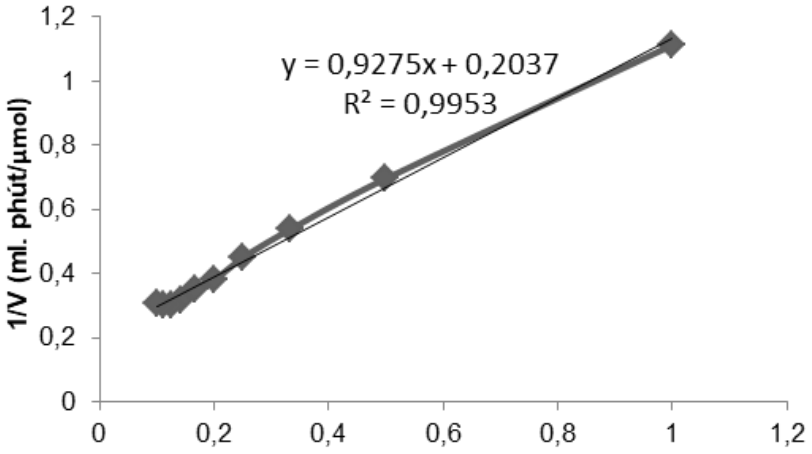
Hình 3.21. Ảnh hưởng của glucose đến hoạt tính của enzyme GH3S2

3.4.3.5. Đặc điểm động học của enzyme GH3S2

Để tìm hiểu sâu hơn về đặc điểm của enzyme GH3S2, các giá trị động học K_m, V_{max} của enzyme GH3S2 này đã được tính toán. K_m chính là nồng độ cơ chất cho phép enzyme đạt được một nửa vận tốc cực đại. Do đó, một enzyme có K_m cao cho thấy enzyme đó cần nồng độ cơ chất lớn để đạt được vận tốc cực đại và có ái lực thấp với cơ chất. V_{max} là tốc độ hoặc vận tốc tối đa của một phản ứng được xúc tác bằng enzyme GH3S2 dựa trên sự bão hòa của enzyme với cơ chất của nó. Giá trị này càng lớn cho thấy enzyme có hoạt tính càng mạnh. Để xác định các hằng số động học của GH3S2 (K_m, V_{max}), trong phản ứng xác định hoạt tính của enzyme nồng độ cơ chất *p*NPG được sử dụng là từ 1-10 mM pha trong đệm PBS 50 mM, pH 6,0 với lượng enzyme là 1 µg trong một phản ứng. Để đảm bảo độ chính xác, các thí nghiệm đều được thực hiện lặp lại 3 lần. Kết quả cho thấy mối tương quan giữa tốc độ phản ứng

với nồng độ *p*NPG tuân theo phương trình y = 0,9275x + 0,2037 với độ tin cậy R^2 = 0,9953 (Hình 3.22).

Mặt khác, phương trình tổng quát thể hiện mối quan hệ giữa tốc độ phản ứng và nồng độ cơ chất là$1/v = K_m.1/V_{max}.1/[S] + 1/V_{max}$ nên các giá trị K_m, V_{max} của GH3S2 được tính tương ứng là 4,55 mM và 4,91 U/mg. Ở điều kiện này, enzyme GH3S2 có hoạt tính riêng là 2,23 U/mg với cơ chất *p*NPG. Như vậy, ái lực với cơ chất *p*NPG của GH3S2 là không cao trong tương quan với một số β-glucosidase từ các đối tượng khác như: β-glucosidase từ vi khuẩn đất có K_m và V_{max} là 0,16 mM và 19,10 μmol/phút [165]; β-glucosidase của vi sinh vật đất *Exiguobacterium* sp. GXG2 có K_m và V_{max} là 1,1 mM và 12,14 U/mg [141]; β-glucosidase của vi sinh vật đất quanh gốc cây bạch đàn có K_m và V_{max} là 0,49 mM và 10,81 U/mg [156]… Tuy nhiên, có nhiều nghiên cứu cũng chỉ ra hoạt tính của β-glucosidase là thấp như β-glucosidase ở *Caulobacter crescentus* có K_m và V_{max} là 0,24 mM và 0,04 U/mg [164], β-glucosidase_32768 ở vi sinh vật sống trong suối nước nóng Bình Châu (Việt Nam) có K_m 0,66 mM và V_{max} đạt 81,81 μmol/min/mg [65].



Hình 3.22. Mối tương quan giữa tốc độ phản ứng của GH3S2 với nồng độ cơ chất theo Linewever – Burk

Tóm lại, kết quả nghiên cứu các đặc điểm của enzyme GH3S2 cho thấy đây là enzyme có hoạt tính β-glucosidase với hoạt tính riêng là 2,23 U/mg, nhiệt độ và pH tối ưu cho hoạt động là 37°C và 6,0, hoạt tính duy trì được 70% ở nồng độ glucose trong môi trường là 6 mM, hoạt tính của enzyme GH3S2 được tăng lên 2,08 lần khi bổ sung 1 mM ion Ca^{2+}. Như vậy, hoạt tính của GH3S2 được phân lập trực tiếp từ DNA đa hệ gene của vi sinh vật đất quanh khu nấm mục trắng có hoạt tính không

cao. Điều này có thể do trong quá trình lựa chọn gene mã hóa cellulase dựa trên vùng/cấu trúc chức năng, chúng tôi ưu tiên lựa chọn gene có vùng/cấu trúc mới (GH3+Exop_C) với cấu trúc phụ trợ Exop_

C nên hoạt tính của protein chưa được chú trọng. Wilson và cộng sự (2008) cũng cho rằng, ở nấm và vi khuẩn có hai cơ chế khác nhau để thủy phần cellulose, đó là (1) tiết các cellulase riêng rẽ, các cellulase này đều có CBM riêng và có ý nghĩa quan trọng trong việc phối hợp cùng phân giải cellulose tự nhiên, (2) là sử dụng phức hệ enzyme gọi là cellulosome trong đó hầu hết các enzyme trong cellulosome không có CBM và có vùng xúc tác bảo phủ nhau. Trong cả hai cơ chế này, endoglucanase và exoglucanase đều là các enzyme phổ biến nhất [173], hoạt động của β-glucosidase hầu như không đáng kể đối với cellulose tinh thể. Mặt khác, trong mối tương tác giữa vi sinh vật và nấm mục trắng, hoạt động phân giải cellulose có thể hầu hết do các nấm chuyên biệt thực hiện như *Basidiomycota*và *Ascomycota* [175], [42], [176]. Các nấm hiếu khí này có thể tiết cellulase ngoại bào để phân giải cellulose còn các hoạt động phân giải của vi khuẩn hầu như không đáng kể. Điều này cũng được Folman và cộng sự đề cập. Như vậy, có thể cho rằng nấm mục trắng đã có ảnh hưởng đến thành phần các vi sinh vật sống quanh đó và các vi sinh vật sống trong khu vực này cũng có những đặc điểm để thích nghi và có thể tồn tại ở đây [74].

KẾT LUẬN VÀ KIẾN NGHỊ

KẾT LUẬN

Từ những kết quả thu được của luận án, chúng tôi rút ra được các kết luận sau:

1. Đã xây dựng bộ dữ liệu DNA đa hệ gene của quần xã vi sinh vật đất xung quanh khu nấm mục trắng ở vườn Quốc gia Cúc Phương với dung lượng 51,82 Gb và phân tích được 3.896.881 ORF thuộc 131 ngành, 118 lớp, 237 bộ, 523 họ, 2240 chi và 916 loài, trong đó có 3.884.879 ORF thuộc giới vi khuẩn được thuộc 111 ngành, 83 lớp, 170 bộ, 406 họ, 1971 chi và 738 loài. Proteobacteria là ngành phổ biến nhất với 3.106.400 gene (75,68%) và ngành Bacteroidetes lớn thứ hai (13,11%) trong số các ngành được phân tích;

2. Dựa trên CSDL KEGG đã chú giải được chức năng của 22.226 gene mã hóa enzyme tham gia thủy phân lignocellulose trong đó 907 gene mã hóa enzyme và protein tham gia tiền xử lý, 8301 gene mã hóa enzyme cellulase và 13.018 gene mã hóa enzyme hemicellulase. Có 22.092 gene được phân loại thuộc 28 ngành của vi khuẩn, trong đó, trội nhất là ngành Proteobacteria (50,79%) và ngành Bacteroidetes (36,73%). Đã khai thác được 13 họ enzyme tham gia thủy phân lignocellulose bằng mô hình đại diện HMM;

3. Trong số 8301 gene mã hóa cellulase được chú giải bằng CSDL KEGG, đã xác định được 1058 gene hoàn chỉnh được phân tích các vùng/cấu trúc chức năng bao gồm các nhóm: (1) endoglucanase với 47 loại domain (367 gene), trong đó domain GH8 là phổ biến nhất; (2) exoglucanase với 6 domain (6 gene); (3) β-glucosidase với 27 loại domain (475 gene), trong đó domain phổ biến là GH3 với vùng/cấu trúc FN3, Exop_C), GH1; (4) 6-phospho-β-glucosidase với 2 domain GH1 và GH4 (210 gene); đã xác định được 1 gen mã hóa enzyme cellulase tiềm năng;

4. Đã biểu hiện thành công protein GH3S2 có kích thước khoảng 91 kDa trong chủng *E. coli* Rosetta1 ở 25°C, môi trường TB cải biến, 0,3 mM IPTG, cảm ứng ở OD_{600} là 1 và thu mẫu 4 giờ sau khi cảm ứng. Đã tinh chế được enzyme tái tổ hợp GH3S2 từ vi khuẩn *E. coli* Rosetta có độ sạch là 97,3 %, hàm lượng đạt 41,8 mg/lít dịch lên men; enzyme GH3S2 có K_m = 4,55 mM và V_{max} = 4,91 U/mg. Ion Ca^{2+} và Mg^{2+} làm tăng hoạt tính enzyme, trong khi đó, ion Ni^{2+} và Cu^{2+} làm giảm hoạt tính. Glucose ở nồng độ 6 mM ảnh hưởng nhẹ đến hoạt tính của GH3S2.

KIẾN NGHỊ

Nghiên cứu khả năng phối hợp các enzyme β-glucosidase GH3S2 với enzyme endoglucanase và exoglucanase để đánh giá hiệu quả thủy phân nguồn cơ chất cellulose.

DANH MỤC CÔNG TRÌNH CỦA TÁC GIẢ

1. **Nguyễn Thị Bình**, Đào Trọng Khoa, Lê Thị Thu Hồng, Trương Nam Hải, *Nghiên cứu khai thác các gene mã hóa enzyme oxi hóa đa đồng từ dữ liệu metageneome của khu hệ vi khuẩn quanh nấm mục trắng (Trametes versicolor) trong rừng Quốc gia Cúc Phương,* Hội nghị Công nghệ sinh học toàn quốc, 2020, tr 187-192.

2. **Nguyễn Thị Bình**, Nguyễn Hồng Dương, Nguyễn Thị Quý, Lê Thị Thu Hồng, Trương Nam Hải, *Nghiên cứu khai thác và biểu hiện gene mã hóa enzyme β-glucosidase từ dữ liệu metageneome của khu hệ vi khuẩn quanh nấm mục trắng (Trametes versicolor),* Hội nghị Công nghệ sinh học toàn quốc, 2021, tr 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 metageneomic analysis of microbial community contributing in lignocellulose degradation in humus samples harvested from Cuc Phuong tropical forest in Vietnam,* Diversity, 2022, 14(3), 220; https://doi.org/10.3390/d14030220

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*, Tạp chí Công nghệ sinh học, 2022, 20(3): 425-433.

5. **Nguyen Thi Binh**, Le Thi Thu Hong, Truong Nam Hai, *Using some bioinformatic tools to mining genees coding cellobiohydrolase from metageneome data of the bacteria surrounding white-rot fungi (Trametes versicolor) in Cuc Phuong National Park*, Tạp chí Khoa học Đại học Thủ đô Hà Nội tập 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*, Tạp chí Công nghệ sinh học, 2022, 20(4): 599-607.

TÀI LIỆU THAM KHẢO

[1] T. H. Do, T. T. Nguyen, T. N. Nguyen, N. G. Le, C. Nguyen, K. Kimura, N. H. Truong, "*Mining biomass-degrading genees through Illumina-based de novo sequencing and metageneomic analysis of free-living bacteria in the gut of the lower termite Coptotermes gestroi harvested in Vietnam,"* J. Biosci. Bioeng., vol. 118, no. 6, pp. 665–671, Dec. 2014, doi: 10.1016/j.jbiosc.2014.05.010.

[2] T. H. Do, N. G. Le, T. K. Dao, T. M. P. Nguyen, T. L. Le, H. L. Luu, K. H. V. Nguyen, V. L. Nguyen, L. A. Le, T. N. Phung, N. M. van Straalen, D. Roelofs, N. H. Truong, "*Metageneomic insights into lignocellulose-degrading genees through Illumina-based de novo sequencing of the microbiome in Vietnamese native goats' rumen,*" J. Gene. Appl. Microbiol., vol. 64, no. 3, pp. 108–116, 2018, doi: 10.2323/jgam.2017.08.004.

[3] I. U. Haq, B. Hillmann, M. Moran, S. Willard, D. Knights, K. R. Fixen, J.S. Schilling, "*Bacterial communities associated with wood rot fungi that use distinct decomposition mechanisms*," ISME Commun. 2022 21, vol. 2, no. 1, pp. 1–9, Mar. 2022, doi: 10.1038/s43705-022-00108-5.

[4] R. Sankaran, K, Markandan, K. S. Khoo, C. K. Cheng, V. Ashokkumar, B. Deepanraj, P. L. Show, "*The expansion of lignocellulose biomass conversion into bioenergy via nanobiotechnology*," Front. Nanotechnol., vol. 3, p. 96, Dec. 2021, doi: 10.3389/fnano.2021.793528/bibtex.

[5] Z. Anwar, M. Gulfraz, and M. Irshad, "*Agro-industrial lignocellulosic biomass a key to unlock the future bio-energy: A brief review*," J. Radiat. Res. Appl. Sci., vol. 7, no. 2, pp. 163–173, Apr. 2014, doi: 10.1016/j.jrras.2014.02.003.

[6] H. Chen, "*Biotechnology of lignocellulose: Theory and practice,"* Biotechnol. Lignocellul. Theory Pract., pp. 1–511, Jan. 2014, doi: 10.1007/978-94-007-6898-7.

[7] M. Lauria, F. Molinari, and M. Motto, "*Geneetic strategies to enhance plant biomass yield and quality- related traits for bio-renewable fuel and chemical productions*," Plants Futur., Oct. 2015, doi: 10.5772/61005.

[8] Y. H. P. Zhang and L. R. Lynd, *"Toward an aggregated understanding of enzymeatic hydrolysis of cellulose: noncomplexed cellulase systems,"*

Biotechnol. Bioeng., vol. 88, no. 7, pp. 797–824, Dec. 2004, doi: 10.1002/bit.20282.

[9] A. Zafar, M. N. Aftab, A. Asif, A. Karadag, L. Peng, H. U. Celebioglu, M. S. Afzal, A. Hamid, I. Iqbal, *"Efficient biomass saccharification using a novel cellobiohydrolase from Clostridium clariflavum for utilization in biofuel industry*," RSC Adv., vol. 11, no. 16, pp. 9246–9261, Mar. 2021, doi: 10.1039/d1ra00545f.

[10] Y. Yang, X. Zhang, Q. Yin, W. Fang, Z. Fang, X. Wang, X. Zhang, Y. Xiao, *"A mechanism of glucose tolerance and stimulation of GH1 β-glucosidases,*" Sci. Reports 2015 51, vol. 5, no. 1, pp. 1–12, Nov. 2015, doi: 10.1038/srep17296.

[11] A.V. Morant, K. Jørgenesen, C. Jørgenesen, S.M. Paquette, R. Sánchez-Pérez, B.L. Møller, S. Bak, *"Beta-glucosidases as detonators of plant chemical defense,"*. Phytochem. , vol. 69, no. 9, pp. 1795-813, June 2008 doi: 10.1016/j.phytochem.2008.03.006. epub 2008 may 9. pmid: 18472115.

[12] V. I. Kovalenko, "*Crystalline cellulose: structure and hydrogene bonds*," Russ. Chem. Rev., vol. 79, no. 3, pp. 231–241, May 2010, doi: 10.1070/rc2010v079n03abeh004065/xml.

[13] S. P. Gautam, P. S. Bundela, A. K. Pandey, Jamaluddin, M. K. Awasthi, and S. Sarsaiya, "*Diversity of cellulolytic microbes and the biodegradation of municipal solid waste by a potential strain,"* Int. J. Microbiol., vol. 2012, 2012, doi: 10.1155/2012/325907.

[14] H. Jørgenesen, J. B. Kristensen, and C. Felby, "*Enzymeatic conversion of lignocellulose into fermentable sugars: Challenges and opportunities,*" Biofuels, Bioprod. Biorefining, vol. 1, no. 2, pp. 119–134, Oct. 2007, doi: 10.1002/bbb.4.

[15] J. S. Brigham, W. S. Adney, and M. E. Himmel, "*Hemicellulases: Diversity and applications,*" Handb. Bioethanol, pp. 119–141, May 2018, doi: 10.1201/9780203752456-7.

[16] C. N. Hamelinck, G. Van Hooijdonk, and A. P. C. Faaij, "*Ethanol from lignocellulosic biomass: techno-economic performance in short-, middle- and*

long-term," Biomass and Bioenergy, vol. 28, no. 4, pp. 384–410, Apr. 2005, doi: 10.1016/j.biombioe.2004.09.002.

[17] S. Moraïs, Y. Barak, R. Lamed, D. B. Wilson, Q. Xu, M. E. Himmel, E. A. Bayer, *"Paradigmatic status of an endo- and exoglucanase and its effect on crystalline cellulose degradation*," Biotechnol. Biofuels, vol. 5, no. 1, pp. 1–9, Oct. 2012, doi: 10.1186/1754-6834-5-78/figures/4.

[18] M. Garvey, H. Klose, R. Fischer, C. Lambertz, and U. Commandeur, "*Cellulases for biomass degradation: comparing recombinant cellulase expression platforms*," Trends Biotechnol., vol. 31, no. 10, pp. 581–593, Oct. 2013, doi: 10.1016/j.tibtech.2013.06.006.

[19] S. Kim and C. H. Kim, "*Production of cellulase enzymes during the solid-state fermentation of empty palm fruit bunch fiber*," Bioprocess Biosyst. Eng. 2011 351, vol. 35, no. 1, pp. 61–67, Nov. 2011, doi: 10.1007/S00449-011-0595-Y.

[20] A. K. Badhan, B. S. Chadha, J. Kaur, H. S. Saini, and M. K. Bhat, *"Production of multiple xylanolytic and cellulolytic enzymes by thermophilic fungus Myceliophthora sp. IMI 387099*," Bioresour. Technol., vol. 98, no. 3, pp. 504–510, Feb. 2007, doi: 10.1016/j.biortech.2006.02.009.

[21] A. Ulrich, G. Klimke, and S. Wirth, "*Diversity and activity of cellulose-decomposing bacteria, isolated from a sandy and a loamy soil after long-term manure application*," Microb. Ecol. 2007 553, vol. 55, no. 3, pp. 512–522, Jul. 2007, doi: 10.1007/s00248-007-9296-0.

[22] V. Juturu and J. C. Wu, "*Microbial cellulases: Engineering, production and applications*," Renew. Sustain. Energy Rev., vol. 33, pp. 188–203, May 2014, doi: 10.1016/j.rser.2014.01.077.

[23] L. R. Lynd, P. J. Weimer, W. H. van Zyl, and I. S. Pretorius, "*Microbial cellulose utilization: fundamentals and biotechnology*," Microbiol. Mol. Biol. Rev., vol. 66, no. 3, pp. 506–577, Sep. 2002, doi: 10.1128/mmbr.66.3.506-577.2002.

[24] M. Dashtban, H. Schraft, and W. Qin, "*Fungal bioconversion of lignocellulosic residues; opportunities & perspectives*," Int. J. Biol. Sci., vol. 5, no. 6, pp. 578–595, 2009, doi: 10.7150/ijbs.5.578.

[25] R. Wahlström, J. Rahikainen, K. Kruus, and A. Suurnäkki, "*Cellulose hydrolysis and binding with Trichoderma reesei Cel5A and Cel7A and their core domains in ionic liquid solutions*," Biotechnol. Bioeng., vol. 111, no. 4, pp. 726–733, Apr. 2014, doi: 10.1002/bit.25144.

[26] V. Parisutham, T. H. Kim, and S. K. Lee, "*Feasibilities of consolidated bioprocessing microbes: From pretreatment to biofuel production*," Bioresour. Technol., vol. 161, pp. 431–440, Jun. 2014, doi: 10.1016/j.biortech.2014.03.114.

[27] D. L. Falkoski, V. M. Guimarães, M. N. de Almeida, A. C. Alfenas, J. L. Colodette, and S. T. de Rezende, "*Chrysoporthe cubensis: A new source of cellulases and hemicellulases to application in biomass saccharification processes*," Bioresour. Technol., vol. 130, pp. 296–305, Feb. 2013, doi: 10.1016/j.biortech.2012.11.140.

[28] P. Gangwar, S. I. Alam, S. Bansod, and L. Singh, "*Bacterial diversity of soil samples from the western Himalayas, India*," Can. J. Microbiol., vol. 55, no. 5, pp. 564–577, May 2009, doi: 10.1139/w09-011.

[29] M. Dashtban, M. Maki, K. T. Leung, C. Mao, and W. Qin, "*Cellulase activities in biomass conversion: Measurement methods and comparison*," Crit. Rev. Biotechnol., vol. 30, no. 4, pp. 302–309, Dec. 2010, doi: 10.3109/07388551.2010.490938.

[30] J. Zhou, L. Bao, L. Chang, Z. Liu, C. You, and H. Lu, "*Beta-xylosidase activity of a GH3 glucosidase/xylosidase from yak rumen metageneome promotes the enzymeatic degradation of hemicellulosic xylans*," Lett. Appl. Microbiol., vol. 54, no. 2, pp. 79–87, Feb. 2012, doi: 10.1111/j.1472-765x.2011.03175.x.

[31] R. E. Quiroz-Castañeda, J. L. Folch-Mallol, R. E. Quiroz-Castañeda, and J. L. Folch-Mallol, "*Hydrolysis of biomass mediated by cellulases for the production of sugars*," Sustain. Degrad. Lignocellul. Biomass - Tech. Appl. Commer., May 2013, doi: 10.5772/53719.

[32] K. P. Rajasree, G. M. Mathew, A. Pandey, and R. K. Sukumaran, "*Highly glucose tolerant β-glucosidase from Aspergillus unguis: NII 08123 for enhanced hydrolysis of biomass*," J. Ind. Microbiol. Biotechnol., vol. 40, no. 9,

pp. 967–975, Sep. 2013, doi: 10.1007/s10295-013-1291-5.

[33] E. M. Obeng, S. N. N. Adam, C. Budiman, C. M. Ongkudon, R. Maas, and J. Jose, *"Lignocellulases: a review of emerging and developing enzymes, systems, and practices*," Bioresour. Bioprocess. 2017 41, vol. 4, no. 1, pp. 1–22, Apr. 2017, doi: 10.1186/s40643-017-0146-8.

[34] D. B. Wilson, "*Microbial diversity of cellulose hydrolysis,"* Curr. Opin. Microbiol., vol. 14, no. 3, pp. 259–263, Jun. 2011, doi: 10.1016/j.mib.2011.04.004.

[35] R. Brunecky, M. Alahuhta, Y. J. Bomble, Q. Xu, J. O. Baker, S.Y. Ding, M. E. Himmel and V. V. Lunin, *"Structure and function of the Clostridium thermocellum cellobiohydrolase A X1-module repeat: enhancement through stabilization of the CbhA complex,"* urn:issn:0907-4449, vol. 68, no. 3, pp. 292–299, Feb. 2012, doi: 10.1107/S0907444912001680.

[36] E. Ransom-Jones, D. L. Jones, A. J. McCarthy, and J. E. McDonald, "*The Fibrobacteres: an important phylum of cellulose-degrading bacteria,"* Microb. Ecol., vol. 63, no. 2, pp. 267–281, Feb. 2012, doi: 10.1007/S00248-011-9998-1.

[37] R. H. Doi, "*Cellulases of mesophilic microorganisms: cellulosome and noncellulosome producers,*" Ann. N. Y. Acad. Sci., vol. 1125, pp. 267–279, 2008, doi: 10.1196/annals.1419.002.

[38] R. E. Quiroz-Castañeda and J. L. Folch-Mallol, "*Hydrolysis of biomass mediated by cellulases for the production of sugars,*" Sustain. Degrad. Lignocellul. Biomass - Tech. Appl. Commer., May 2013, doi: 10.5772/53719.

[39] V. Valášková and P. Baldrian, "*Degradation of cellulose and hemicelluloses by the brown rot fungus Piptoporus betulinus--production of extracellular enzymes and characterization of the major cellulases*," Microbiology, vol. 152, no. Pt 12, pp. 3613–3622, Dec. 2006, doi: 10.1099/mic.0.29149-0.

[40] X.-Z. Zhang and Y.-H. P. Zhang, "*Cellulases: Characteristics, Sources, Production, and Applications,*" Bioprocess. Technol. Biorefinery Sustain. Prod. Fuels, Chem. Polym., pp. 131–146, Jul. 2013, doi: 10.1002/9781118642047.ch8.

[41] J. J. Yoon, C. J. Cha, Y. S. Kim, and W. Kim, "*Degradation of cellulose by the major endoglucanase produced from the brown-rot fungus Fomitopsis pinicola,*" Biotechnol. Lett., vol. 30, no. 8, pp. 1373–1378, Aug. 2008, doi: 10.1007/s10529-008-9715-4.

[42] B.C. Song, K.Y. Kim, J.J. Yoon, S.H. Sim, K Lee, Y.S. Kim, Y.K. Kim, C.J. Cha, *"Functional analysis of a genee encoding endoglucanase that belongs to glycosyl hydrolase family 12 from the brown-rot basidiomycete Fomitopsis palustris,*" J Microbiol Biotechnol. 2008 Mar;18(3):404-9. PMID: 18388455.

[43] D. J. Vocadlo and G. J. Davies, "*Mechanistic insights into glycosidase chemistry,*" Curr. Opin. Chem. Biol., vol. 12, no. 5, pp. 539–555, Oct. 2008, doi: 10.1016/j.cbpa.2008.05.010.

[44] A. Sørensen, M. Lübeck, P. S. Lübeck, and B. K. Ahring, *"Fungal beta-glucosidases: a bottleneck in industrial use of lignocellulosic materials,*" Biomolecules, vol. 3, no. 3, p. 612, 2013, doi: 10.3390/biom3030612.

[45] S. Sethi, A. Datta, B. L. Gupta, and S. Gupta, *"Optimization of cellulase production from bacteria isolated from soil*," ISRN Biotechnol., vol. 2013, pp. 1–7, Feb. 2013, doi: 10.5402/2013/985685.

[46] H. Michlmayr, C. Schümann, N. M. Barreira Braz Da Silva, K. D. Kulbe, and A. M. Del Hierro, "*Isolation and basic characterization of a β-glucosidase from a strain of Lactobacillus brevis isolated from a malolactic starter culture,*" J. Appl. Microbiol., vol. 108, no. 2, pp. 550–559, Feb. 2010, doi: 10.1111/j.1365-2672.2009.04461.x.

[47] M. R. Hong, Y. S. Kim, C. S. Park, J. K. Lee, Y. S. Kim, and D. K. Oh, *"Characterization of a recombinant beta-glucosidase from the thermophilic bacterium Caldicellulosiruptor saccharolyticus,*" J. Biosci. Bioeng., vol. 108, no. 1, pp. 36–40, Jul. 2009, doi: 10.1016/j.jbiosc.2009.02.014.

[48] K. Gourlay, J. Hu, V. Arantes, M. Andberg, M. Saloheimo, M. Penttilä, J. Saddler., "*Swollenin aids in the amorphogeneesis step during the enzymeatic hydrolysis of pretreated biomass,*" Bioresour. Technol., vol. 142, pp. 498–503, 2013, doi: 10.1016/j.biortech.2013.05.053.

[49] O. Shoseyov, Z. Shani, and I. Levy, "*Carbohydrate binding modules:*

biochemical properties and novel applications," Microbiol. Mol. Biol. Rev., vol. 70, no. 2, pp. 283–295, Jun. 2006, doi: 10.1128/mmbr.00028-05.

[50] D. Guillén, S. Sánchez, and R. Rodríguez-Sanoja, "*Carbohydrate-binding domains: Multiplicity of biological roles*," Appl. Microbiol. Biotechnol., vol. 85, no. 5, pp. 1241–1249, Feb. 2010, doi: 10.1007/s00253-009-2331-y.

[51] A. Singh, S. Bajar, A. Devi, and D. Pant, "*An overview on the recent developments in fungal cellulase production and their industrial applications*," Bioresour. Technol. Reports, vol. 14, p. 100652, Jun. 2021, doi: 10.1016/j.biteb.2021.100652.

[52] A. Ahmed, F. ul-H. Nasim, K. Batool, and A. Bibi, "*Microbial β-glucosidase: sources, production and applications*," J. Appl. Environ. Microbiol. Vol. 5, 2017, Pages 31-46, vol. 5, no. 1, pp. 31–46, Mar. 2017, doi: 10.12691/jaem-5-1-4.

[53] S. H. Toushik, K.-T. Lee, J.-S. Lee, and K.-S. Kim, *"Functional applications of lignocellulolytic enzymes in the fruit and vegetable processing industries*," J. Food Sci., vol. 82, no. 3, pp. 585–593, Mar. 2017, doi: 10.1111/1750-3841.13636.

[54] S. E. Blumer-Schuette, I. Kataeva, J. Westpheling, M. W. Adams, and R. M. Kelly, *"Extremely thermophilic microorganisms for biomass conversion: status and prospects*," Curr. Opin. Biotechnol., vol. 19, no. 3, pp. 210–217, Jun. 2008, doi: 10.1016/j.copbio.2008.04.007.

[55] C.K.S. Pillai, W. Paul, and C.P. Sharma, *"Chitin and chitosan polymers: Chemistry, solubility and fiber formation*," Progress in Polymer Science. 2009;34:641-678

[56] J. Zhang, W. Xia, P. Liu, Q. Cheng, T. Tahirou, W. Gu, B. Li, "*Chitosan modification and pharmaceutical/biomedical applications*," Mar. Drugs, vol. 8, no. 7, pp. 1962–1987, 2010, doi: 10.3390/md8071962.

[57] Y. Q. Xu, C. J. Duan, Q. N. Zhou, J. L. Tang, and J. X. Feng, "C*loning and identification of cellulase genees from uncultured microorganisms in pulp sediments from paper mill effluent*," Wei Sheng Wu Xue Bao, vol. 46, no. 5, pp. 783–788, 2006.

[58] Y. Feng, C. J. Duan, H. Pang, X. C. Mo, C. F. Wu, Y. Yu, Y. L. Hu, J. Wei, J. L. Tang, J. X. Feng JX, *"Cloning and identification of novel cellulase genees from uncultured microorganisms in rabbit cecum and characterization of the expressed cellulases*," Appl. Microbiol. Biotechnol., vol. 75, no. 2, pp. 319–328, May 2007, doi: 10.1007/S00253-006-0820-9.

[59] D. Kim, S. N. Kim, K. S. Baik, S. C. Park, C. H. Lim, J. O. Kim, T. S. Shin, M. J. Oh, C. N. Seong, "*Screening and characterization of a cellulase genee from the gut microflora of abalone using metageneomic library*," J. Microbiol., vol. 49, no. 1, pp. 141–145, Feb. 2011, doi: 10.1007/S12275-011-0205-3.

[60] C. M. Lee, Y. S. Lee, S. H. Seo, S. H. Yoon, S. J. Kim, B. S. Hahn, J. S. Sim, B. S. Koo, "*Screening and characterization of a novel cellulase genee from the gut microflora of Hermetia illucens using metageneomic library*," J. Microbiol. Biotechnol., vol. 24, no. 9, pp. 1196–1206, Jul. 2014, doi: 10.4014/jmb.1405.05001.

[61] M. Yasir, H. Khan, S. S. Azam, A. Telke, S. W. Kim, and Y. R. Chung, *"Cloning and functional characterization of endo-β-1,4-glucanase genee from metageneomic library of vermicompost*," J. Microbiol., vol. 51, no. 3, pp. 329–335, Jun. 2013, doi: 10.1007/s12275-013-2697-5.

[62] M. T. T. Phan, V. Q. Nguyen, H. G. Le, T. K. Nguyen, and M. D. Tran, "*Molecular cloning genee and nucleotide sequence of the genee encoding an endo-1,4-beta-glucanase from Bacillus sp VLSH08 strain applying to biomass hydrolysis*," J. Vietnamese Environ., vol. 3, no. 2, pp. 80–86, Nov. 2012, doi: 10.13141/jve.vol3.no2.pp80-86.

[63] H. B. T. Quyên, P. N. P. Thảo, and N. M. P. Long, *"Khảo sát nấm mốc có khả năng phân giải cellulose thu nhận từ rừng Mã Đà, Đồng Nai*," Tạp chí khoa học Đại học mở Thành phố Hồ Chí Minh - kỹ thuật và công nghệ, vol. 13, no. 1, pp. 170–180, Oct. 2018, doi: 10.46223/hcmcoujs.tech.vi.13.1.454.2018.

[64] T. H. Do, T. K. Dao, K. H. V. Nguyen, N. G. Le, T. M. P. Nguyen, T. L. Le, T. N. Phung, N. M. van Straalen, D. Roelofs, N. H. Truong, "*Metageneomic analysis of bacterial community structure and diversity of lignocellulolytic bacteria in Vietnamese native goat rumen*," Asian-Australasian J. Anim. Sci.,

vol. 31, no. 5, pp. 738–747, Sep. 2017, doi: 10.5713/ajas.17.0174.

[65] T. T. Thủy, "*Nghiên cứu đánh giá đa dạng vi sinh vật, sàng lọc, thu nhận và xác định tính chất của cellulase suối nước nóng Binh Châu bằng kỹ thuật metageneomics,."* Luận án Tiến sĩ sinh học, Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và công nghệ Việt Nam, 2021

[66] A. I. Hatakka, O. K. Mohammadi, and T. K. Lundell, "*The potential of white-rot fungi and their enzymes in the treatment of lignocellulosic feed*," http://dx.doi.org/10.1080/08905438909549697, vol. 3, no. 1, pp. 45–58, Jan. 2009, doi: 10.1080/08905438909549697.

[67] Y. Hadar, "*Biodegradation of aromatic toxic pollutants by white rot fungi,*" Encycl. Mycol., pp. 197–204, Jan. 2021, doi: 10.1016/B978-0-12-819990-9.00066-4.

[68] M. Couturier and J. G. Berrin, "*The saccharification step: The main enzymeatic components*," Lignocellul. Convers. Enzyme. Microb. Tools Bioethanol Prod., pp. 93–110, Mar. 2013, doi: 10.1007/978-3-642-37861-4_5/cover/.

[69] M. Andlar, T. Rezić, N. Marđetko, D. Kracher, R. Ludwig, and B. Šantek, "*Lignocellulose degradation: An overview of fungi and fungal enzymes involved in lignocellulose degradation,*" Eng. Life Sci., vol. 18, no. 11, p. 768, Nov. 2018, doi: 10.1002/elsc.201800039.

[70] R. da S. S. Fl aacute via, F. L. G. Nayara, F. da P. Marcelo, G. F. Gustavo, and S. otilde es R. L. Rodrigo, *"Production and characterization of -glucosidase from Gongronella butleri by solid-state fermentation*," African J. Biotechnol. , vol. 15, no. 16, pp. 633–641, Apr. 2016, doi: 10.5897/ajb2015.15025.

[71] F. Warnecke, P. Luginbühl, N. Ivanova *et al.*, "*Metageneomic and functional analysis of hindgut microbiota of a wood-feeding higher termite,*" Nature, vol. 450, no. 7169, pp. 560–565, Nov. 2007, doi: 10.1038/nature06269.

[72] D. H. Huson, D. C. Richter, S. Mitra, A. F. Auch, and S. C. Schuster, "*Methods for comparative metageneomics,*" BMC Bioinforma. 2009 101, vol. 10, no. 1, pp. 1–10, Jan. 2009, doi: 10.1186/1471-2105-10-s1-s12. [90] W. De Boer, P. Verheggene, P. J. A. Klein Gunnewiek, G. A. Kowalchuk, and J. A. Van Veen, "*Microbial community composition affects soil fungistasis*," Appl. Environ.

Microbiol., vol. 69, no. 2, p. 835, Feb. 2003, doi: 10.1128/aem.69.2.835-844.2003.

[73] B. Liu, J. Liu, M. Ju, X. Li, and P. Wang, "*Bacteria-white-rot fungi joint remediation of petroleum-contaminated soil based on sustained-release of laccase*," RSC Adv., vol. 7, no. 62, pp. 39075–39081, Aug. 2017, doi: 10.1039/c7ra06962f.

[74] L.B. Folman, P.J. Klein Gunnewiek, L. Boddy, W.F. de Boer, *"Impact of white-rot fungi on numbers and community composition of bacteria colonizing beech wood from forest soil*," FEMS Microbiol. Ecol., vol. 63, no. 2, pp. 181–191, Feb. 2008, doi: 10.1111/j.1574-6941.2007.00425.x.

[75] G. Janusz, A. Pawlik, J. Sulej, U. Świderska-Burek, A. Jarosz-Wilkołazka, and A. Paszczyński, "*Lignin degradation: microorganisms, enzymes involved, geneomes analysis and evolution*," FEMS Microbiol. Rev., vol. 41, no. 6, pp. 941–962, Nov. 2017, doi: 10.1093/femsre/fux049.

[76] C. J. Duan, L. Xian, G. C. Zhao, Y. Feng, H. Pang, X. L. Bai, J. L. Tang, Q. S. Ma, J. X. Feng, "*Isolation and partial characterization of novel genees encoding acidic cellulases from metageneomes of buffalo rumens*," J. Appl. Microbiol., vol. 107, no. 1, pp. 245–256, Jul. 2009, doi: 10.1111/J.1365-2672.2009.04202.x.

[77] T.C. Glenn, "*Field guide to next-geneeration DNA sequencers*," Mol. Ecol. Resour., vol. 11, no. 5, pp. 759–769, Sep. 2011, doi: 10.1111/j.1755-0998.2011.03024.x.

[78] T. Thomas, J. Gilbert, and F. Meyer, "*Metageneomics - a guide from sampling to data analysis,"* Microb. Informatics Exp. 2012 21, vol. 2, no. 1, pp. 1–12, Feb. 2012, doi: 10.1186/2042-5783-2-3.

[79] A. Mikheenko, V. Saveliev, and A. Gurevich, "*MetaQUAST: evaluation of metageneome assemblies*," Bioinformatics, vol. 32, no. 7, pp. 1088–1090, Apr. 2016, doi: 10.1093/bioinformatics/btv697.

[80] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, *"The KEGG resource for deciphering the geneome*," Nucleic Acids Res., vol. 32, no. Database issue, Jan. 2004, doi: 10.1093/nar/gkh063.

[81] J. Muller, D. Szklarczyk , P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. von Mering, T. Doerks , L. J. Jensen, P. Bork, "*eggNOG v2.0: extending the evolutionary geneealogy of genees with enhanced non-supervised orthologous groups, species and functional annotations,*" Nucleic Acids Res., vol. 38, no. Database issue, Nov. 2010, doi: 10.1093/nar/gkp951.

[82] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, M. Punta, *"The Pfam protein families database,*" Nucleic Acids Res., vol. 38, no. Database issue, Nov. 2010, doi: 10.1093/nar/gkp985.

[83] V. M. Markowitz, N.N. Ivanova, E. Szeto, K. Palaniappan, K. Chu, D. Dalevi, I.M. Chen, Y. Grechkin, I. Dubchak, I. Anderson, A. Lykidis, K. Mavromatis, P. Hugeneholtz, N. Kyrpides, *"IMG/M: a data management and analysis system for metageneomes*," Nucleic Acids Res., vol. 36, no. Database issue, Jan. 2008, doi: 10.1093/nar/gkm869.

[84] Z. A. Dyson, R. J. Seviour, J. Tucci, and S. Petrovski, "*Geneome sequences of Pseudomonas oryzihabitans phage POR1 and Pseudomonas aeruginosa phage PAE1,*" Geneome Announc., vol. 4, no. 3, pp. 1515–1530, 2016, doi: 10.1128/geneomea.01515-15.

[85] J. Mistry, S. Chuguransky. L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer , S. C. E. Tosatto, L. Paladin, L. J. Raj S, Richardson, R. D. Finn, A. Bateman, *"Pfam: The protein families database in 2021,"* Nucleic Acids Res., vol. 49, no. D1, pp. D412–D419, Jan. 2021, doi: 10.1093/nar/gkaa913.

[86] E.L. Sonnhammer, S.R. Eddy, E. Birney, A. Bateman, and R. Durbin, *"Pfam: multiple sequence alignments and HMM-profiles of protein domains,*" Nucleic Acids Res. 1998 Jan 1;26(1):320-2. doi: 10.1093/nar/26.1.320. PMID: 9399864;

[87] S. R. Eddy, *"Accelerated Profile HMM Searches,*" PLOS Comput. Biol., vol. 7, no. 10, p. e1002195, Oct. 2011, doi: 10.1371/journal.pcbi.1002195.

[88] S.J. Sammut, R.D. Finn, and A. Bateman, "*Pfam 10 years on: 10,000 families and still growing,"* Brief. Bioinform., vol. 9, no. 3, pp. 210–219, May 2008, doi: 10.1093/bib/bbn010.

[89] D. Steiner, P. Forrer, M. T. Stumpp, and A. Plückthun, "*Signal sequences directing cotranslational translocation expand the range of proteins amenable to phage display*," Nat. Biotechnol., vol. 24, no. 7, pp. 823–831, Jul. 2006, doi: 10.1038/nbt1218.

[90] J. Song, K. Takemoto, H. Shen, H. Tan, M. M. Gromiha, and T. Akutsu, "*Prediction of protein folding rates from structural topology and complex network properties*," IPSJ Trans. Bioinforma., vol. 3, pp. 40–53, 2010, doi: 10.2197/ipsjtbio.3.40.

[91] E. Capriotti and R. Casadio, *"K-Fold: a tool for the prediction of the protein folding kinetic order and rate*," Bioinformatics, vol. 23, no. 3, pp. 385–386, Feb. 2007, doi: 10.1093/bioinformatics/btl610.

[92] P. Chaudhary, A. N. Naganathan, and M. M. Gromiha, *"Folding RaCe: a robust method for predicting changes in protein folding rates upon point mutations,"* Bioinformatics, vol. 31, no. 13, pp. 2091–2097, Jul. 2015, doi: 10.1093/bioinformatics/btv091.

[93] H.A. Ariyarathna, M.G. Francki, *"Phylogeneetic relationships and protein modelling revealed two distinct subfamilies of group II HKT genees between crop and model grasses*," Geneome, vol. 59, no. 7, pp. 509–517, May 2016, doi: 10.1139/gene-2016-0035.

[94] G. Zhang, H. Li, and B. Fang, *"Discriminating acidic and alkaline enzymes using a random forest model with secondary structure amino acid composition,"* Process Biochem., vol. 44, no. 6, pp. 654–660

[95] G. L. Fan, Q. Z. Li, and Y. C. Zuo, "*Predicting acidic and alkaline enzymes by incorporating the average chemical shift and genee ontology informations into the geneeral form of Chou's PseAAC*," Process Biochem., vol. 48, no. 7, pp. 1048–1053, Jul. 2013, doi: 10.1016/j.procbio.2013.05.012.

[96] H. Lin, W. Chen, and H. Ding, "*Acalpred: a sequence-based tool for discriminating between acidic and alkaline enzymes*," PLoS One, vol. 8, no. 10, Oct. 2013, doi: 10.1371/journal.pone.0075726.

[97] F. Pucci, R. Bourgeas, and M. Rooman, "*Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing*

HoTMuSiC," Sci. Reports 2016 61, vol. 6, no. 1, pp. 1–9, Mar. 2016, doi: 10.1038/srep23257.

[98] M. Ebrahimi, A. Lakizadeh, P. Agha-Golzadeh, E. Ebrahimie, and M. Ebrahimi, "*Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes*," PLoS One, vol. 6, no. 8, p. 23146, 2011, doi: 10.1371/journal.pone.0023146.

[99] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, Y. Yamanishi, "*KEGG for linking geneomes to life and the environment,"* Nucleic Acids Res., vol. 36, no. suppl_1, pp. D480–D484, Jan. 2008, doi: 10.1093/nar/gkm882.

[100] UniProt Consortium*, "Reorganizing the protein space at the Universal Protein Resource (UniProt),"* Nucleic Acids Res., vol. 40, no. Database issue, Jan. 2012, doi: 10.1093/nar/gkr981.

[101] S. R. Eddy, "*Accelerated Profile HMM Searches,*" PLoS Comput. Biol., vol. 7, no. 10, 2011, doi: 10.1371/journal.pcbi.1002195.

[102] T. Sambrook, J.; Fritsch, E. F.; Maniatis, "*In vitro amplication of dna by the polymerase,*" Mol. Cloning, pp. 494–500, 2001.

[103] J. Sambrook, E. F. Fritsch, and T. Maniatis, Molecular cloning: *A laboratory manual.*, Second edi., vol. 1. Cold Spring Harbor Laboratory Press, 1989.

[104] K. L. Franken, H. S. Hiemstra, K. E. van Meijgaarden, Y. Subronto, J. den Hartigh, T. H. Ottenhoff, J. W. Drijfhout, *"Purification of his-tagged proteins by immobilized chelate affinity chromatography: the benefits from the use of organic solvent,*" Protein Expr. Purif., vol. 18, no. 1, pp. 95–99, 2000, doi: 10.1006/prep.1999.1162.

[105] M. Bradford, "*A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding*," Anal. Biochem., vol. 72, no. 1–2, pp. 248–254, May 1976, doi: 10.1006/abio.1976.9999.

[106] V. Veena, P. Poornima, R. Parvatham, Sivapriyadharsini, and K. Kalaiselvi, "*Isolation and characterization of β-glucosidase producing bacteria from*

different sources," African J. Biotechnol., vol. 10, no. 66, pp. 14907–14912, 2011, doi: 10.5897/ajb09.314.

[107] Z. Fang, W. Fang, J. Liu J, Y. Hong , H. Peng, X. Zhang, B. Sun, Y. Xiao, *"Cloning and characterization of a ${\beta}$-glucosidase from marine microbial metageneome with excellent glucose tolerance*," J. Microbiol. Biotechnol., vol. 20, no. 9, pp. 1351–1358, Sep. 2010, doi: 10.4014/jmb.1003.03011.

[108] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "*IDBA-UD: a de novo assembler for single-cell and metageneomic sequencing data with highly uneven depth,*" Bioinformatics, vol. 28, no. 11, pp. 1420–1428, Jun. 2012, doi: 10.1093/bioinformatics/bts174.

[109] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "*MEGAHIT: an ultra-fast single-node solution for large and complex metageneomics assembly via succinct de Bruijn graph*", doi: 10.1093/bioinformatics/btv033.

[110] B. Langmead and S. L. Salzberg, "*Fast gapped-read alignment with Bowtie 2,*" Nat. Methods, vol. 9, no. 4, pp. 357–359, Apr. 2012, doi: 10.1038/nmeth.1923.

[111] W. Zhu, A. Lomsadze, and M. Borodovsky, "*Ab initio genee identification in metageneomic sequences,"* Nucleic Acids Res., vol. 38, no. 12, pp. e132–e132, Jul. 2010, doi: 10.1093/nar/gkq275.

[112] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "*MEGAN analysis of metageneomic data*," Geneome Res., vol. 17, no. 3, pp. 377–386, Mar. 2007, doi: 10.1101/GR.5969107.

[113] W. Li Ã and A. Godzik, *"Bioinformatics applications note Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,"* vol. 22, no. 13, pp. 1658–1659, 2006, doi: 10.1093/bioinformatics/btl158.

[114] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, Y. Yamanishi*, "KEGG for linking geneomes to life and the environment,"* Nucleic Acids Res., vol. 36, no. suppl_1, pp. D480–D484, Jan. 2008, doi: 10.1093/nar/gkm882.

[115] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold

, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, P. Bork, "*eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges,*" Nucleic Acids Res., vol. 40, no. D1, pp. D284–D289, Jan. 2012, doi: 10.1093/nar/gkr1060.

[116] J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, and M. Punta, "*Challenges in homology search: HMMER3 and convergenet evolution of coiled-coil regions,*" Nucleic Acids Res., vol. 41, no. 12, pp. e121–e121, Jul. 2013, doi: 10.1093/nar/gkt263.

[117] K. E. Wommack, J. Bhavsar, and J. Ravel, "*Metageneomics: read length matters,*" Appl. Environ. Microbiol., vol. 74, no. 5, pp. 1453–1463, Mar. 2008, doi: 10.1128/AEM.02181-07.

[118] N. Praeg and P. Illmer, "*Microbial community composition in the rhizosphere of Larix decidua under different light regimes with additional focus on methane cycling microorganisms*," Sci. Reports 2020 101, vol. 10, no. 1, pp. 1–16, Dec. 2020, doi: 10.1038/s41598-020-79143-y.

[119] R. Wang, H. Zhang, L. Sun, G. Qi, S. Chen, and X. Zhao, *"Microbial community composition is related to soil biological and chemical properties and bacterial wilt outbreak*," Sci. Reports 2017 71, vol. 7, no. 1, pp. 1–10, Mar. 2017, doi: 10.1038/s41598-017-00472-6.

[120] G. Feng, T. Xie, X. Wang, J. Bai, L. Tang, H. Zhao, W. Wei , M. Wang, Y. Zhao*, "Metageneomic analysis of microbial community and function involved in cd-contaminated soil,*" BMC Microbiol., vol. 18, no. 1, pp. 1–13, Feb. 2018, doi: 10.1186/S12866-018-1152-5/figures/7.

[121] K. L. Cobaugh, S. M. Schaeffer, and J. M. DeBruyn, "*Functional and structural succession of soil microbial communities below decomposing human cadavers*," PLoS One, vol. 10, no. 6, p. e0130201, Jun. 2015, doi: 10.1371/journal.pone.0130201.

[122] N. Fierer and R. B. Jackson, *"The diversity and biogeography of soil bacterial communities,*" Proc. Natl. Acad. Sci. U. S. A., vol. 103, no. 3, pp. 626–631, Jan. 2006, doi: 10.1073/pnas.0507535103.

[123] W. R. Cookson, M. Osman, P. Marschner, D. A. Abaye, I. M. Clark, D.V.

Murphy, E.A. Stockdale, C. A. Watson, *"Controls on soil nitrogene cycling and microbial community composition across land use and incubation temperature*," Soil Biol. Biochem., vol. 39, no. 3, pp. 744–756, Mar. 2007, doi: 10.1016/j.soilbio.2006.09.022.

[124] Y. Liang, J. D. V. Nostrand, Y. Deng, Z. He, L. Wu, X. Zhang, G. Li, J. Zhou, *"Functional genee diversity of soil microbial communities from five oil-contaminated fields in China*," ISME J. 2011 53, vol. 5, no. 3, pp. 403–413, Sep. 2010, doi: 10.1038/ismej.2010.142.

[125] S. J. Cho, M. H. Kim, and Y. O. Lee, "*Effect of pH on soil bacterial diversity*," J. Ecol. Environ., vol. 40, no. 1, pp. 1–9, Oct. 2016, doi: 10.1186/S41610-016-0004-1/figures/4.

[126] J. Rousk, , E. Bååth, P. C. Brookes, C. L. Lauber, C. Lozupone, J. G. Caporaso, R. Knight, N. Fierer, "*Soil bacterial and fungal communities across a pH gradient in an arable soil*," ISME J. 2010 410, vol. 4, no. 10, pp. 1340–1351, May 2010, doi: 10.1038/ismej.2010.58.

[127] Y. Wu, J. Zeng, Q. Zhu, Z. Zhang, and X. Lin, "*pH is the primary determinant of the bacterial community structure in agricultural soils impacted by polycyclic aromatic hydrocarbon pollution,"* Sci. Reports 2017 71, vol. 7, no. 1, pp. 1–7, Jan. 2017, doi: 10.1038/srep40093.

[128] Y. Yun, H. Wang, B. Man, X. Xiang, J. Zhou, X. Qiu , Y. Duan, A. S. Engel, *"The relationship between ph and bacterial communities in a single karst ecosystem and its implication for soil acidification,"* Front. Microbiol., vol. 7, no. DEC, p. 1955, 2016, doi: 10.3389/fmicb.2016.01955.

[129] R. V. Augimeri, A. J. Varley, and J. L. Strap, "*Establishing a role for bacterial cellulose in environmental interactions: Lessons learned from diverse biofilm-producing Proteobacteria*," Front. Microbiol., vol. 6, no. NOV, p. 1282, 2015, doi: 10.3389/fmicb.2015.01282/bibtex.

[130] M. De Vries, A. Schöler, S. Schöler, J. Ertl, Z. Xu, and M. Schloter, "*Metageneomic analyses reveal no differences in genees involved in cellulose degradation under different tillage treatments*," FEMS Microbiol. Ecol., vol. 91, p. 69, 2015, doi: 10.1093/femsec/fiv069.

[131] P. Lapébie, V. Lombard, E. Drula, N. Terrapon, and B. Henrissat, "*Bacteroidetes use thousands of enzyme combinations to break down glycans*," Nat. Commun. 2019 101, vol. 10, no. 1, pp. 1–7, May 2019, doi: 10.1038/s41467-019-10068-5.

[132] N. Fierer, *"Embracing the unknown: disentangling the complexities of the soil microbiome*," Nat. Rev. Microbiol., vol. 15, no. 10, pp. 579–590, Oct. 2017, doi: 10.1038/nrmicro.2017.87.

[133] F. L. Soares, I. S. Melo, A. C. F. Dias, and F. D. Andreote, *"Cellulolytic bacteria from soils in harsh environments*," World J. Microbiol. Biotechnol., vol. 28, no. 5, pp. 2195–2203, May 2012, doi: 10.1007/s11274-012-1025-2.

[134] J. L. Edwards, D. L. Smith, J. Connolly, J. E. McDonald, M. J. Cox, I. Joint, C. Edwards, A. J. McCarthy*, "Identification of carbohydrate metabolism genees in the metageneome of a marine biofilm community shown to be dominated by gammaproteobacteria, bacteroidetes*," Genees (Basel)., vol. 1, no. 3, p. 371, Dec. 2010, doi: 10.3390/genees1030371.

[135] H. Inoue, S. R. Decker, L. E. Taylor II, S. Yano, and S. Sawayama, "*Identification and characterization of core cellulolytic enzymes from Talaromyces cellulolyticus (formerly Acremonium cellulolyticus) critical for hydrolysis of lignocellulosic biomass*," vol. 7, pp. 1–13, 2014, doi: 10.1186/s13068-014-0151-5.

[136] K. H. V. Nguyen, T. K. Dao, H. D. Nguyen, K. H. Nguyen, T. Q. Nguyen, T. T. Nguyen, T. M. P. Nguyen, N. H. Truong, T. H. Do*, "Some characters of bacterial cellulases in goats' rumen elucidated by metageneomic DNA analysis and the role of fibronectin 3 module for endoglucanase function*," Anim. Biosci., vol. 34, no. 5, p. 867, May 2021, doi: 10.5713/ajas.20.0115.

[137] R. Berlemont and A. C. Martiny, "*Phylogeneetic distribution of potential cellulases in bacteria*," Appl. Environ. Microbiol., vol. 79, no. 5, pp. 1545–1554, Mar. 2013, doi: 10.1128/aem.03305-12/suppl_file/zam999104146so4.pdf.

[138] S. Lu, J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales , M. Gwadz, D. I. Hurwitz, G. H. Marchler , J. S. Song, N. Thanki, R. A. Yamashita

, M. Yang, D. Zhang, C. Zheng, C. J. Lanczycki, A. Marchler-Bauer, *"CDD/SPARCLE: the conserved domain database in 2020,*" Nucleic Acids Res., vol. 48, no. D1, pp. D265–D268, Jan. 2020, doi: 10.1093/nar/gkz991.

[139] J. N. Varghese, M. Hrmova, and G. B. Fincher, *"Three-dimensional structure of a barley beta-D-glucan exohydrolase, a family 3 glycosyl hydrolase,*" Structure, vol. 7, no. 2, pp. 179–190, 1999, doi: 10.1016/S0969-2126(99)80024-0.

[140] Y. Nakatani, S. M. Cutfield, N. P. Cowieson, and J. F. Cutfield, "*Structure and activity of exo-1,3/1,4-β-glucanase from marine bacterium Pseudoalteromonas sp. BB1 showing a novel C-terminal domain*," FEBS J., vol. 279, no. 3, pp. 464–478, Feb. 2012, doi: 10.1111/j.1742-4658.2011.08439.x.

[141] B. Yin, H. Gu, X. Mo, Y. Xu, B. Yan , Q. Li, Q. Ou, B. Wu, C. Guo, C. Jiang, *"Identification and molecular characterization of a psychrophilic GH1 β-glucosidase from the subtropical soil microorganism Exiguobacterium sp. GXG2,*" AMB Express, vol. 9, no. 1, Dec. 2019, doi: 10.1186/S13568-019-0873-7.

[142] L. Käll, A. Krogh, and E. L. L. Sonnhammer, *"A combined transmembrane topology and signal peptide prediction method,"* J. Mol. Biol., vol. 338, no. 5, pp. 1027–1036, May 2004, doi: 10.1016/j.jmb.2004.03.016.

[143] P. Singh, L. Sharma, S. R. Kulothungan, B. V. Adkar, R. S. Prajapati, P. S. Ali , B. Krishnan, R. Varadarajan, *"Effect of signal peptide on stability and folding of escherichia coli thioredoxin,*" PLoS One, vol. 8, no. 5, p. e63442, May 2013, doi: 10.1371/journal.pone.0063442.

[144] G. L. Rosano and E. A. Ceccarelli, *"Recombinant protein expression in Escherichia coli: Advances and challenges,*" Front. Microbiol., vol. 5, no. APR, p. 172, 2014, doi: 10.3389/fmicb.2014.00172/bibtex.

[145] M. Fathi-Roudsari, A. Akhavian-Tehrani, and N. Maghsoudi, *"Comparison of three escherichia coli strains in recombinant production of reteplase*," Avicenna J. Med. Biotechnol., vol. 8, no. 1, p. 16, Jan. 2016, Accessed: Jun. 08, 2022. [Online]. Available: /pmc/articles/pmc4717461/

[146] M. Sim, H. S. Seok, and J. Kim, "*A next-geneeration sequence clustering*

method for e. coli through proteomics-geneomics data mapping," Procedia Comput. Sci., vol. 23, pp. 96–101, Jan. 2013, doi: 10.1016/j.procs.2013.10.013.

[147] R. Vincentelli, C. Bignon, A. Gruez, S. Canaan, G. Sulzenbacher, M. Tegoni, V. Campanacci, C. Cambillau, *"Medium-scale structural geneomics: strategies for protein expression and crystallization,"* Acc. Chem. Res., vol. 36, no. 3, pp. 165–172, Mar. 2003, doi: 10.1021/ar010130s.

[148] D. M. Francis and R. Page, "*Strategies to optimize protein expression in e. coli*," Curr. Protoc. Protein Sci., vol. 61, no. 1, pp. 5.24.1-5.24.29, Aug. 2010, doi: 10.1002/0471140864.ps0524s61.

[149] T. San-Miguel, P. Pérez-Bermúdez, and I. Gavidia, "*Production of soluble eukaryotic recombinant proteins in E. coli is favoured in early log-phase cultures induced at low temperature*," Springerplus, vol. 2, no. 1, pp. 1–4, 2013, doi: 10.1186/2193-1801-2-89.

[150] A. Vera, N. González-Montalbán, A. Arís, and A. Villaverde, *"The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures*," Biotechnol. Bioeng., vol. 96, no. 6, pp. 1101–1106, Apr. 2007, doi: 10.1002/bit.21218.

[151] M. Dragosits, G. Frascotti, L. Bernard-Granger, F. Vázquez, M. Giuliani, K. Baumann, E. Rodríguez-Carmona, J. Tokkanen, E. Parrilli, M. G. Wiebe, R. Kunert, M. Maurer, B. Gasser, M. Sauer, P. Branduardi, T. Pakula, M. Saloheimo, M. Penttilä, P. Ferrer, M. Luisa Tutino, A. Villaverde, D. Porro, D. Mattanovich, *"Influence of growth temperature on the production of antibody Fab fragments in different microbes: A host comparative analysis*," Biotechnol. Prog., vol. 27, no. 1, pp. 38–46, Jan. 2011, doi: 10.1002/BTPR.524.

[152] R. Seyfi, V. Babaeipour, M. R. Mofid, and F. A. Kahaki, "*Expression and production of recombinant scorpine as a potassium channel blocker protein in Escherichia coli*," Biotechnol. Appl. Biochem., vol. 66, no. 1, pp. 119–129, Jan. 2019, doi: 10.1002/bab.1704.

[153] M. Gutiérrez-González, C. Farías, S. Tello, D. Pérez-Etcheverry, A. Romero, R. Zúñiga, C. H. Ribeiro, C. Lorenzo-Ferreiro, M. C. Molina, *"Optimization of culture conditions for the expression of three different insoluble proteins in*

Escherichia coli," Sci. Rep., vol. 9, no. 1, Dec. 2019, doi: 10.1038/S41598-019-53200-7.

[154] P. Dvorak, L. Chrast, P. I. Nikel, R. Fedr, K. Soucek, M. Sedlackova, R. Chaloupkova, V. de Lorenzo, Z. Prokop, J. Damborsky*, "Exacerbation of substrate toxicity by IPTG in Escherichia coli BL21(DE3) carrying a synthetic metabolic pathway*," Microb. Cell Fact., vol. 14, no. 1, pp. 1–15, Dec. 2015, doi: 10.1186/S12934-015-0393-3/figures/5.

[155] Y. Zhao, W.F. Liu, A.J. Mao, N Jiang, and Z.Y. Dong, *"Expression, purification and enzymeatic characterization of Bacillus polymyxa beta-glucosidase genee (bglA) in Escherichia coli.,"* Sheng Wu Gong Cheng Xue Bao, vol. 20, no. 5, pp. 741–744, Sep. 2004, Accessed: Aug. 24, 2021. [Online]. Available: https://europepmc.org/article/med/15974001

[156] E. S. Gomes-Pepe, E. G. M. Sierra, M. R. Pereira, T. C. L. Castellane, and E. G. M. De Lemos, "*Bg10: A novel metageneomics alcohol-tolerant and glucose-stimulated gh1 β-glucosidase suitable for lactose-free milk preparation,"* PLoS One, vol. 11, no. 12, Dec. 2016, doi: 10.1371/journal.pone.0167932.

[157] Y. Li, N. Liu, H. Yang, F. Zhao, Y. Yu, Y. Tian, X. Lu*, "Cloning and characterization of a new β-Glucosidase from a metageneomic library of Rumen of cattle feeding with Miscanthus sinensis*," BMC Biotechnol., vol. 14, no. 1, pp. 1–9, Oct. 2014, doi: 10.1186/1472-6750-14-85.

[158] S. Mahapatra, A. S. Vickram, T. B. Sridharan, R. Parameswari, and M. R. Pathy, *"Screening, production, optimization and characterization of β-glucosidase using microbes from shellfish waste*," 3 Biotech, vol. 6, no. 2, Dec. 2016, doi: 10.1007/s13205-016-0530-7.

[159] D. Y. A. Yapi, D. Gnakri, S. L. Niamke, and L. P. Kouame, *"Purification and biochemical characterization of a specific β"- glucosidase from the digestive fluid of larvae of the palm weevil, Rhynchophorus palmarum*," J. Insect Sci., vol. 9, Feb. 2009, doi: 10.1673/031.009.0401.

[160] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, M. Law, *"Comparison of next-geneeration sequencing systems*," J. Biomed.

Biotechnol., vol. 2012, 2012, doi: 10.1155/2012/251364.

[161] Z. Chen, T. Meng, Z. Li, P. Liu, Y. Wang, N. He, D. Liang*, "Characterization of a beta-glucosidase from Bacillus licheniformis and its effect on bioflocculant degradation,"* AMB Express, vol. 7, no. 1, Dec. 2017, doi: 10.1186/S13568-017-0501-3.

[162] L. Zhang, Q. Fu, W. Li, B. Wang, X. Yin, S. Liu, Z. Xu, Q. Niu., *"Identification and characterization of a novel β-glucosidase via metageneomic analysis of Bursaphelenchus xylophilus and its microbial flora.*," Sci. Rep., vol. 7, no. 1, pp. 14850–14850, Nov. 2017, doi: 10.1038/S41598-017-14073-W.

[163]F. D. Otajevwo and H. S. A. Aluyi, "*Cultural conditions necessary for optimal cellulase yield by cellulolytic bacterial organisms as they relate to residual sugars released in broth medium*," Mod. Appl. Sci., vol. 5, no. 3, p. p141, Jun. 2011, doi: 10.5539/mas.v5n3p141.

[164]P.I. Justo, j. M. Corrêa, A. Maller, M. K. Kadowaki,J. L. da Conceição-Silva R. F. Gandra, C. Simão Rde, "*Analysis of the xynB5 genee encoding a multifunctional GH3-BglX β-glucosidase-β-xylosidase-α-arabinosidase member in Caulobacter crescentus*," Antonie Van Leeuwenhoek, vol. 108, no. 4, pp. 993–1007, Oct. 2015, doi: 10.1007/s10482-015-0552-x.

[165] S.J. Kim, C. M. Lee CM, M.Y. Kim, Y. S. Yeo, S. H. Yoon, H. C. Kang, B.S. Koo*, "Screening and characterization of an enzyme with beta-glucosidase activity from environmental DNA*," J. Microbiol. Biotechnol., vol. 17, no. 6, pp. 905–912, Jun. 2007, Accessed: Jul. 20, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18050907/

[166] J. Kaur, B. S. Chadha, B. A. Kumar, G. S. Kaur, and H. S. Saini, *"Purification and characterization of ß-glucosidase from Melanocarpus sp. MTCC 3922*," Electron. J. Biotechnol., vol. 10, no. 2, pp. 260–270, 2007, doi: 10.4067/S0717-34582007000200010.

[167] H. Li, X. Xu, H. Chen, Y. Zhang, J. Xu, J. Wang, X. Lu, *"Molecular analyses of the functional microbial community in composting by PCR-DGGE targeting the genees of the β-glucosidase,"* Bioresour. Technol., vol. 134, pp. 51–58, Apr. 2013, doi: 10.1016/j.biortech.2013.01.077.

[168] S. Wei, Y. Semel, B. A. Bravdo, H. Czosnek, and O. Shoseyov, "*Expression and subcellular compartmentation of Aspergillus niger β-glucosidase in transgeneic tobacco result in an increased insecticidal activity on whiteflies (Bemisia tabaci),*" Plant Sci., vol. 172, no. 6, pp. 1175–1181, Jun. 2007, doi: 10.1016/j.plantsci.2007.02.018.

[169] M. A. Baffi, T. Tobal, J. Henrique, G. Lago, R.S. Leite, M. Boscolo, E. Gomes , R. Da-Silva*, "A novel β-glucosidase from Sporidiobolus pararoseus: characterization and application in winemaking,*" J. Food Sci., vol. 76, no. 7, Sep. 2011, doi: 10.1111/J.1750-3841.2011.02293.x.

[170] A. Uhoraningoga, G. K. Kinsella, J. M. Frias, B. J. Ryan, and G. T. Henehan, "*The statistical optimisation of recombinant β-glucosidase production through a two-stage, multi-model, design of experiments approach,"* Bioeng. 2019, Vol. 6, Page 61, vol. 6, no. 3, p. 61, Jul. 2019, doi: 10.3390/bioengineering6030061.

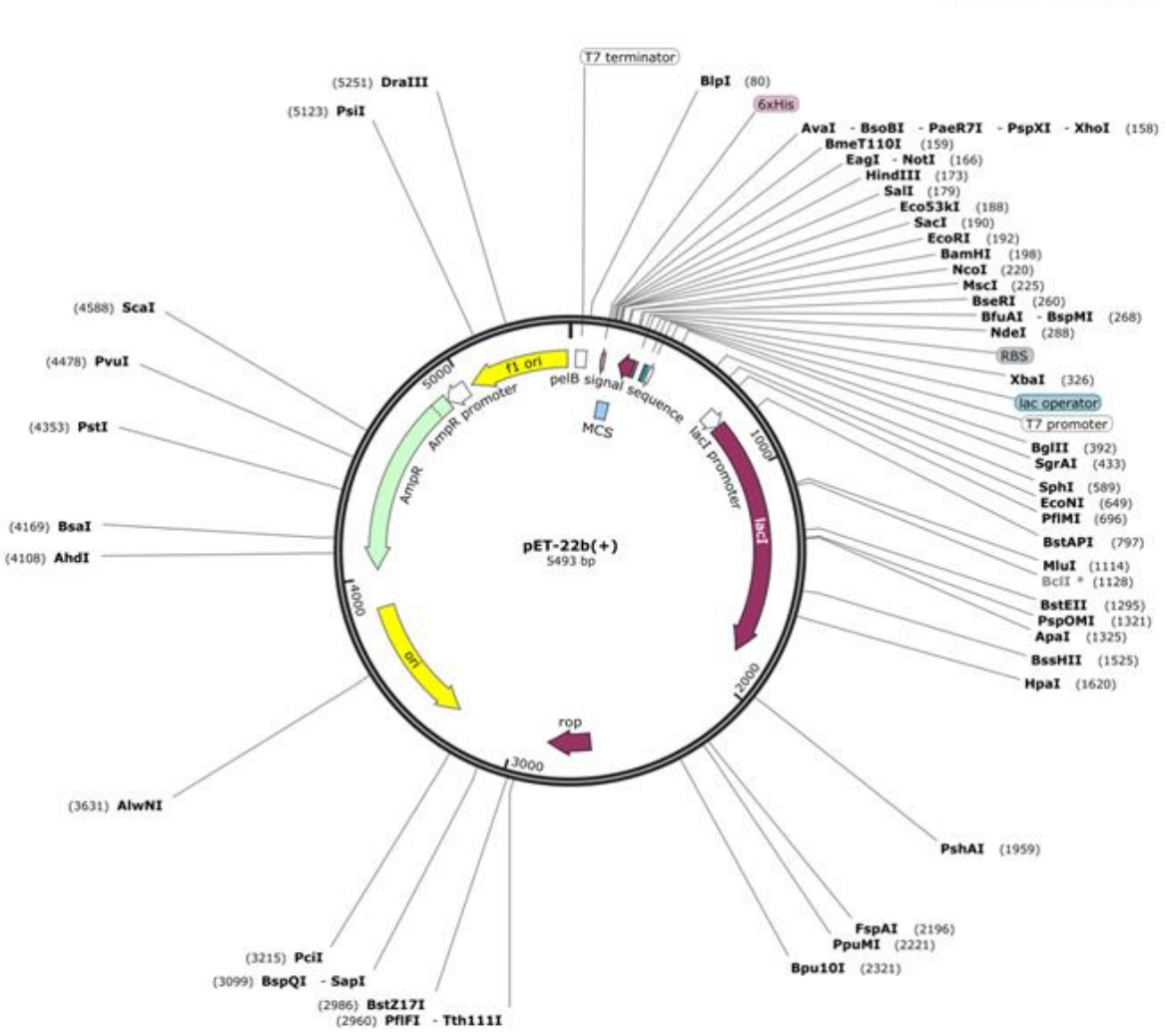
[171] R. R. Singhania, A. K. Patel, R. K. Sukumaran, C. Larroche, and A. Pandey, "*Role and significance of beta-glucosidases in the hydrolysis of cellulose for bioethanol production,"* Bioresour. Technol., vol. 127, pp. 500–507, 2013, doi: 10.1016/j.biortech.2012.09.012.

[172] Y. Feng, C. J. Duan, H. Pang, X. C. Mo, C. F. Wu, Y. Yu, Y. L. Hu, J. Wei, J. L. Tang, J. X. Feng, *"Cloning and identification of novel cellulase genees from uncultured microorganisms in rabbit cecum and characterization of the expressed cellulases,"* Appl. Microbiol. Biotechnol., vol. 75, no. 2, pp. 319–328, May 2007, doi: 10.1007/s00253-006-0820-9.

[173] D. B. Wilson, *"Three microbial strategies for plant cell wall degradation,"* Ann. N. Y. Acad. Sci., vol. 1125, no. 1, pp. 289–297, Mar. 2008, doi: 10.1196/annals.1419.026.

PHỤ LỤC

Phụ lục 1. Bản đồ gene của plasmid pET22b(+)



Phụ lục 2: Trình tự nucleotide của gene GH3S2 sau khi tối ưu mã bộ ba và trình tự amino acid tương ứng

atg aca tcg cag gcc ttc gtc atc cgc agc ggc gcg ctg gtc gcc gca ctg atg ctg gga
M T S Q A F V I R S G A L V A A L M L G
ttg ctc ggc tgc cgc ggc cag gac cgg gct gcc gcc gcc gca gcc acc gac aag gat ccc
L L G C R G Q D R A A A A A A T D K D P
tgg ccg gag gtc atc tgg ccc ctg gct gcg gac ccg gcg ctg gag aag cgc atc acc gac
W P E V I W P L A A D P A L E K R I T D
ctg atg gcc ggc atg acg gtg gag gaa aag gtc ggc cag ctg gtg cag ggt gac atc gcc
L M A G M T V E E K V G Q L V Q G D I A
agc gtc acc cca gat gat gtg cgc cgc tac cgg ctt ggc tcg atc ctg gcc ggt ggc aac
S V T P D D V R R Y R L G S I L A G G N
tcc gat ccc ggt ggc cgc tat gac gcg tcg ccg gcc gaa tgg ctg gcg ctg gcc gac gcc
S D P G G R Y D A S P A E W L A L A D A
ttc tac gac gcg tcc atg gac acg tcg aaa ggc ggc aag gcc atc ccg ctg ctg ttc ggc
F Y D A S M D T S K G G K A I P L L F G
atc gat gcc gtg cac ggg cag agc aac atc att ggc gcc acg ttg ttc ccg cac aac atc
I D A V H G Q S N I I G A T L F P H N I
ggg ctg ggc gcc acg cgc aat ccg gag ctg ctt cgg cag atc ggt ggc atc acc gcg ctg
G L G A T R N P E L L R Q I G G I T A L
gag acc cgc gtt acc ggc atg gaa tgg acg ttc gcg ccg acc gtt gcc gta ccc cag gat
E T R V T G M E W T F A P T V A V P Q D
gat cgc tgg gga cgc acc tac gaa ggc tac tcc gaa tcg ccg gac gtg gtg gcc agc tat
D R W G R T Y E G Y S E S P D V V A S Y
gcc gcc gcc atg gtg gag gga ttg cag ggc agg gtg gga acc ccg gag ttc ctc gat ggc
A A A M V E G L Q G R V G T P E F L D G
cgc cat gtg atc gcc tcg gtg aag cat ttc ctc ggc gac ggt ggc acc act gac ggc aag
R H V I A S V K H F L G D G G T T D G K
gac cag ggc gac acc cgc atc agc gag tca gat ctg gtg cgc atc cac gcc gcc gga tat
D Q G D T R I S E S D L V R I H A A G Y
ccg ccg gca atc gcc gcc ggc gcg cag acc gcg atg gcg tcg ttc aac agc gtc aac ggt
P P A I A A G A Q T A M A S F N S V N G
gaa aag atg cat ggg cac cgg cac tac ctt acc gat gta ctc aag ggc cgc atg aac ttc
E K M H G H R H Y L T D V L K G R M N F
ggt ggc ttc gtg gtg ggt gac tgg aat ggt cat gga cag gtc aag ggt tgc acc act aca
G G F V V G D W N G H G Q V K G C T T T
gac tgc ccg gcc acg atc aac gcg ggc ctg gac atg gcg atg gcc tcg gac agc tgg aag
D C P A T I N A G L D M A M A S D S W K
ggt ttc tac gag acg acg ctg gct gcg gtg aag gat ggg cgg atc acg ccg caa cgc ctg
G F Y E T T L A A V K D G R I T P Q R L
gac gat gcg gtg cgc cgg atc ctg cgg gtc aag ttc cgc ctt ggg ctg ttc gag gcc ggc
D D A V R R I L R V K F R L G L F E A G
agg cca tcc acg cgg gcc gtc ggc ggg cag ttc gca ctg atc ggc gcg ccg gca cat cgc
R P S T R A V G G Q F A L I G A P A H R

```
gcg gtt gcc cgg cag gcc gtg cgc gaa tcg ctg gtc ctg ctg aag aac cag aac ggc ctc
 A   V   A   R   Q   A   V   R   E   S   L   V   L   L   K   N   Q   N   G   L
ctg ccg ctg tcg ccg aag cag cgg atc ctc gtg gcc ggc gac ggt gcc gac gat gtc ggc
 L   P   L   S   P   K   Q   R   I   L   V   A   G   D   G   A   D   D   V   G
aag cag gcc ggc ggc tgg acg ctc aac tgg cag ggc acc ggc acc acc cgc aag gac ttc
 K   Q   A   G   G   W   T   L   N   W   Q   G   T   G   T   T   R   K   D   F
ccc aat gcg gac acg atc tac gag ggc atc gcg cgc cag gcc agg gcg gcc ggt ggt gaa
 P   N   A   D   T   I   Y   E   G   I   A   R   Q   A   R   A   A   G   G   E
gcc atg ctt tcc gtc gac ggt cgc tat gca gtg aag ccc gat gtg gcg gtg gtg gtg ttt
 A   M   L   S   V   D   G   R   Y   A   V   K   P   D   V   A   V   V   V   F
ggc gag gac ccc tat gcc gag ttc cag gga gac cgg ccg acg ctg gcc tac aag ccc ggc
 G   E   D   P   Y   A   E   F   Q   G   D   R   P   T   L   A   Y   K   P   G
aac gaa acg gac ctg gcg ctg ctc aag cgg ctc aag gcc gat ggc ata ccg gtt gtt gcg
 N   E   T   D   L   A   L   L   K   R   L   K   A   D   G   I   P   V   V   A
atc ttc ctg agc ggg cgg ccg ctc tgg gtg aac cgg gaa atc aat gcc gcc gat gcc ttc
 I   F   L   S   G   R   P   L   W   V   N   R   E   I   N   A   A   D   A   F
gtg gct gcg tgg ctg ccg ggt tcg gaa ggc gcc ggg att gcc gat gtg ctg ctg cgc gga
 V   A   A   W   L   P   G   S   E   G   A   G   I   A   D   V   L   L   R   G
agc gat ggc cgc gtg cag cac gat ttc aag ggc aag ctc agt ttc agc tgg ccg cgc act
 S   D   G   R   V   Q   H   D   F   K   G   K   L   S   F   S   W   P   R   T
gcc acc cag tac gcc aac aac gtg ggc cag aag gac tac gat cca ttg ttt gcg ttc ggc
 A   T   Q   Y   A   N   N   V   G   Q   K   D   Y   D   P   L   F   A   F   G
ttc ggc ctt acc tac gcc gac aac ggc ggc ctg gcc gcg cta ccg gag gca tcg ggc gta
 F   G   L   T   Y   A   D   N   G   G   L   A   A   L   P   E   A   S   G   V
acc ggc aac gaa ggc gcg acc ggc gtg ttc ttt gcg cgc ggt ggc gca ggc cct ggc atg
 T   G   N   E   G   A   T   G   V   F   F   A   R   G   G   A   G   P   G   M
gcg ctg cgg ctc gag gat gcc gct ggc cag ggc ctg agc gtg acc cgc gta ccg gac gca
 A   L   R   L   E   D   A   A   G   Q   G   L   S   V   T   R   V   P   D   A
ttg ccc gat gat cgg ctg aag atc acc ggc gtg gat cat ctg gcg cag gag gat ggg cga
 L   P   D   D   R   L   K   I   T   G   V   D   H   L   A   Q   E   D   G   R
cgc ctg gcc tgg tcg ggc aat ggc gaa gcc gtc gct gca ctg cag tcg cac acg gcg ctg
 R   L   A   W   S   G   N   G   E   A   V   A   A   L   Q   S   H   T   A   L
gac ctg cag cgc gaa tcc aac ggc gac ctg atg ctg ctg acc acg ctg cgg gtg gac gca
 D   L   Q   R   E   S   N   G   D   L   M   L   L   T   T   L   R   V   D   A
gcc ccg aag ggt gag gcg tgg ctg tcg gtc ggt tgc ggc gcg ggc tgc tcg gca cgc atc
 A   P   K   G   E   A   W   L   S   V   G   C   G   A   G   C   S   A   R   I
gcc atc ggg tcg tcg ctg gcg gcg ctt cca cag ggc cag tgg aag cgt gtc ggc gtg ccg
 A   I   G   S   S   L   A   A   L   P   Q   G   Q   W   K   R   V   G   V   P
ctg aag tgc ctg gcc agg gcg ggc gcc aag ctg gac gcg atc gac cga ccg tgg tcg gtg
 L   K   C   L   A   R   A   G   A   K   L   D   A   I   D   R   P   W   S   V
gtg acg ggc gat gcg atg acg atc tcc gtg tca cgc gtc gcg ctg ggt gcg ctg aac gaa
 V   T   G   D   A   M   T   I   S   V   S   R   V   A   L   G   A   L   N   E
gcc gag gtc acc ctc gga tgc gga gca tga
 A   E   V   T   L   G   C   G   A   -
```