**MINISTRY OF EDUCATION AND TRAINING**

**HA NOI NATIONAL UNIVERSITY OF EDUCATION**

**GIANG NGUYEN MINH**

# EXPLOIT DNA METAGENOME DATA, EXPRESS AND RESEACH CHARACTERS OF β-XYLOSIDASE FROM MICROORGANISM IN GUT OF *Coptotermes gestroi* IN VIET NAM

**Specialized: GENETICS**

**Code: 62. 42. 01. 21**

**SUMMARY OF BIOLOGICAL THESIS**

***Ha Noi, February, 2018***

The project was completed at: Genetics, Biology of department, Hanoi National University of Education and Genetics Engineering of department, Institute of Biotechnology, Vietnam Academy of Sciences

Science instructor: **Prof. Dr. Hai Truong Nam**

**Assoc. Prof. Dr. Lanh Dang Huu**

Critical Reviewer 1: **Prof. Dr**. Ton Phan Huu

Critical Reviewer 2: **Prof. Dr**. Mau Chu Hoang

Critical Reviewer 3: **Assoc. Prof. Dr** . Ham Le Huy

The thesis will be presented before Assessment Council in Hanoi National University of Education at 14h, October 25th 2017

The thesis can be found at the National Library, Hanoi
or Library of Hanoi National University of Education

# INTRODUCTION

## 1. Reason to choose the topic

Lignocellulose is a major component of plant biomass that is discharged mainly from the forestry, agriculture, paper industry, etc. This is a source of waste that causes environmental pollution if not thoroughly treated. However, this biomass can become a raw material for conversion into many valuable products. Methods that are treating plant biomass by physical and chemical factors have many major disadvantages such as: produces low monosacharide, demand for specialized equipment with high cost, solvent residue in product and environmental problems, ... The method of treating plant biomass by enzyme has many advantages such as: can actively control the reaction, the product of the eachother of stages is the material for many industries, environmental friendliness, etc. Therefore, the search and production of hydrolysis enzymes for plant biomass are still of interest in most countries in the world, including Vietnam.

Of the species, termites digest lignocellulose effectively. The studies showed that many microbial species in gut of termite secrete enzymes such as cellulase, hemicellulase which hydrolyze most of the lignocellulose into sugars. In the components of lignocellulose, hemicellulose has a heterogeneous structure and various types of biological molecules. Therefore if we want to decompose compeletely hemicellulose we need many group of enzymes such as xylanase, β-xylosidase, arabinanase, arabinofuranosidase, α-glucuronidase, α-galactosidase, acetyl xylan esterase, etc. In which β-xylosidase is capable of cutting at the non-reducing end position on the xylan-main chain of hemicellulose and xylobiose sugars into xylose sugars, it is essential in the enzymatic system to convert hemicellulose into monosacharide.

From the financial support of the Vietnam-Japan cooperation project: "Isolation of genes encoding lignocellulolytic enzymes in termites in Vietnam using Metagenomics technology" (2012-2015), DNA metagenome of symbiont in gut of *Coptotermes gestroi* was extracted and sequenced. The size of the metagenome DNA sequence was 5,431.60 Mb, of which 125,431 ORF and 587 ORF coded for lignocellulase was determined. The search and selection of genes from metagenome DNA sequences was based on the selection of each gene encoding the lignocellulase which was predicted by the BGI's company. After that each gene was built phylogenetic taxonomy by bioinformatic tool. However, this selection takes a lot of time because the DNA metagenome data is too large. On the other hand, there are only 10% of recombinant enzymes that are active in empirical expression. What is the way to search and exploit the gene coding lignocellulase from the DNA metagenome sequence data of the microorganism in *C. gestroi*. At the same time, the selection genes will be expressed successful, and have good activity.

The search and selection of the lignocellulase gene from 587 ORF was performed by hand. The first we will select each gene encoding for the lignocellulase which estimated by the sequencing company. Then we investigate the protein's conservative region and initialize the phylogenetic tree for each enzyme. Finally, the ORF will be selected according to (1) the new sequence, (2) the specific and specific activity center, and (3) the simple sequence. However, this way takes a lot of time because of too large DNA metagenome data. Moreover, most of the genes selected in this way are difficult to express successfully. For example, after selecting eight genes for expression, six genes do not express or express insoluble and only two genes were expressed weakly. Limitations of the previous study required the development of an efficient method which help to find out exactly lignocellulase gene from the DNA metagenome data and must have the proper activity of the target enzyme after expression. In addition, bacterial diversity and lignocellulase in termites have not been studied in the past. So we conducted the topic " EXPLOIT DNA METAGENOME DATA, EXPRESS AND RESEACH CHARACTERS OF β-XYLOSIDASE FROM MICROORGANISM IN GUT OF *Coptetermes gestroi* IN VIET NAM ".

## 2. Objectives

### 2.1. General objectives

Studies on bacterial diversity and lignocellulase in gut of *C. gestroi* from DNA metagenome data. At the same time, building an effective method to search quickly target genes encoding of lignocellulase from DNA metagenome data of microorganisms in the *C. gestroi*. That genes have exactly target enzyme after expression.

### 2.2. Specific objectives

-Exploit available DNA metagenome data of microorganism in gut of *C. gestroi* to reseach diversity of bateria and lignocellulase;

- Build a new method to search for and select genes from DNA metagenome data of microorganisms in the *C. gestroi*;

- Express and reseach characters of lignocellulose hydrolysis enzyme.

**3. Research content**

- Study the diversity of bacteria in gut of *C. gestroi* and lignocellulase following the CAZY classification system from the available DNA metagenome data;
- Search for amino acid sequences of β-xylosidase which has been studied in detail actions and select reliable bioinformatics tools for constructing probes. Using a probe for rapid search of the β-xylosidase gene DNA metagenome data of microorganisms in the *C. gestroi*;
- Build probe to search for the gene encoding lignocellulase from the available DNA metagenome data;
- Express and study the characters of β-xylosidase.

**4. Objects**

DNA metagenome data of microorganisms in gut of *C. gestroi* of the Genetic Engineering Department, Institute of Biotechnology, Vietnam Academy of Sciences and technology, some databases and bioinformatic tools.

**5. The reseach scopes**

We study the diversity of bacteria in gut of *C. gestroi* and the method to build probe which use to exploit and select gene from DNA metagenome data. Gene is expressed in *E. coli Rosseta 1* and identified the characters of β-xylosidase.

**6. The scientific and practical significance of the topic**

**6.1. Scientific significance**

- Analyzed of bacteria in gut of *C. gestroi* in Vietnam and diversity of lignocellulase according to CAZY classification;

Provided a new and effective method for searching target genes from the DNA metagenome data by probes;

Expressed and identified the β-xylosidase activity of free-living microorganisms in gut of *C. gestroi*.

**6.2. Practical significance**

β-xylosidase is a new, well-functioning enzyme that works optimally in alkaline and high temperature environments. This enzyme promises to be very effective in the application of plant biomass hydrolysis, especially when combined with alkaline and thermal treatment of plant biomass.

**7. New contributions of the thesis**

This is the first study of the bateria in gut of *C. gestroi* in Vietnam and the diversity of CAZY's lignocellulose hydrolysis enzymes from DNA metagenome data;

Building probe from gene or amino acids sequences that have been studied to extract genes from DNA metagenome sequence data by Metagenomics is a novel approach;

β-xylosidase which is a new enzyme derived from microorganisms in gut of *C. gestroi* in Vietnam is expressed and refined successfully with good activity, optimized in alkaline and high temperature environments.

## Chapter 1. OVERVIEW OF RESEARCH

### 1.1. LIGNOCELLULOSE VÀ DECOMPOSE

**1.1.1. Lignocellulose**: The role and composition of lignocellulose

**1.1.2. Lignocellulose metabolism**: Describe the enzyme system involved in lignocellulose metabolism

### 1.2. METAGENOMICS, BIOLOGICAL TOOLS AND DATA RESOURCES FOR DNA ANALYSIS METAGENOME

**1.2.1. Metagenomics:** Describe the general concept of Metagenomics, its approach in research and application.

**1.2.2. Some tools are used to analyze data**

Descriptive basis of data building, the way to use and reliability of some bioinformatic tools used to analyse data including: BLAST, EXPASY, Phyre 2, TBI, Alcapred and Phylogeny.

**1.2.3. Data sources**

Description about the sources, amounts of data, data updates and reliability of the two data sources used in the study including CAZY and NCBI.

**1.2.4. DNA probes and applications**

A brief overview of DNA probes and their applications, particularly probes has been used in study about the DNA metagenome.

**1.3. β–XYLOSIDASE**

An overview of the β-xylosidase includes the names, operating patterns, spatial structure, activity, application and supply of β-xylosidase.

**1.4. MICROBIOLOGY COMMUNITIES AND LIGNOCELLULASE**

**1.4.1. Some communities attend to decompose lignocellulose:** General overview of lignocellulose microorganisms in the nature.

**1.4.2 Microorganism and lignocellulase in gut of termite**

Describe the biological characteristics of termites and the diversity of symbiotic microorganisms in the termite gut - the source of the lignocellulase.

**1.4.3. Research overview on microbial diversity and lignocellulase in gut of *C. gestroi* in Vietnam**

A summary of the reseach on microorganism and lignocellulase of C.gestroi in Vietnam and the issues need to be solved in this study.

## Chapter 2. OBJECTS, MATERIALS AND RESEARCH METHODS

**2.1. OBJECTS AND MATERIALS**

**2.1.1. Objects**

Using 125,432 ORFs from the DNA metagenome data of the microorganisms in gut of *C. gestroi* was annotated by BGI. Of these, there are 587 ORFs encoding lignocellulases as a source of data to exploit the Termite 2 _ GL0112518 gene which encodes β-xylosidase.

**2.1.2. Chemicals and machinery**

Chemicals: The chemicals were ordered from international standard companies such as Bio-Lab (USA), Fermentas (USA), Sigma (USA), Merck (Germany). Using machine of key technology department, Institute of Biotechnology, Vietnam Academy of Sciences and technology.

**2.2. RESEARCH METHODS**

**2.2.1. Method for building probe:** We built the probe of β-xylosidase following in the figure 2.2.

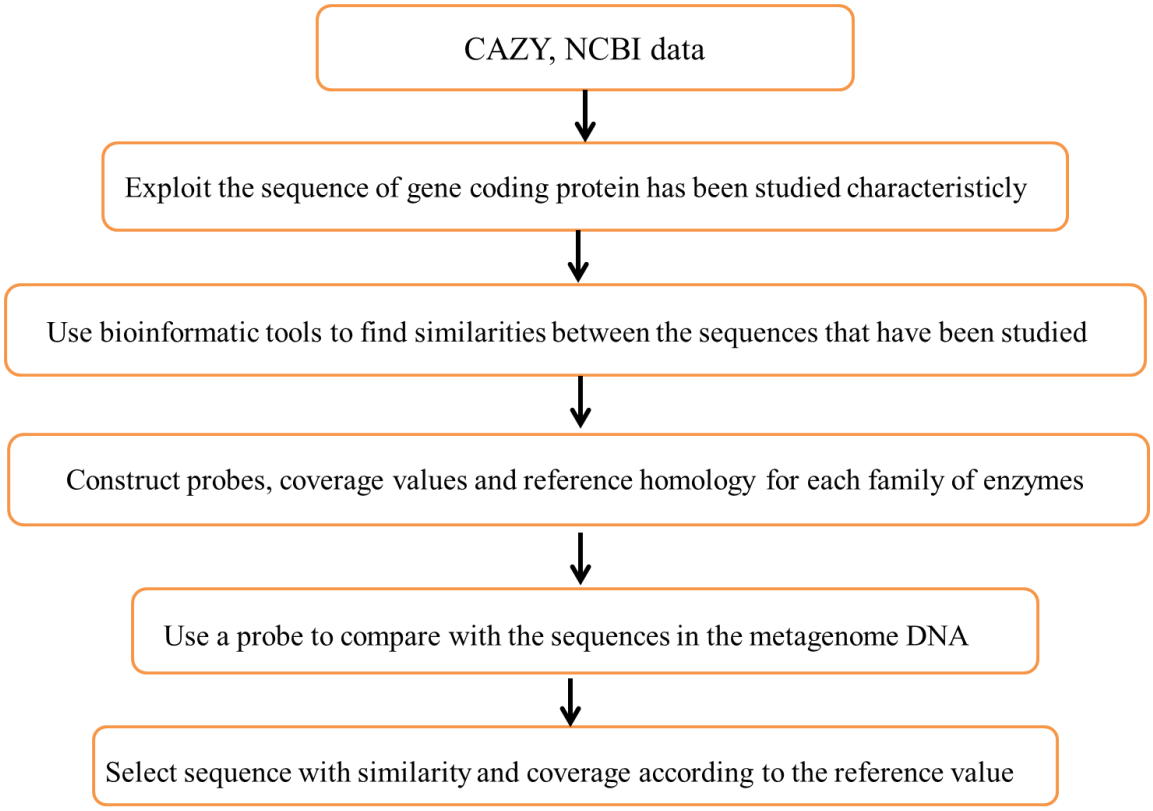


***Figure 2.2. Diagram of basic steps in building a probe***

**2.2.2. Methods to analyse data by bioinformatic softwares**

Analyze the diversity of typical microorganisms in *C. gestroi;* Analysis of ORF diversity of lignocellulase according to CAZY classification; Compare the ORF sequences of the DNA metagenome data with the NCBI database by BLAST; Design primers by FastPCR software; Check the position of the restriction enzyme on the gene by RestrictionMapper software; Transfer DNA sequence code to amino acid sequence using EXPASY translation program; Predict the structure of *Xbx14* by Phyre2 and Swiss model software; Predict the origin of the gene by the BLAST-Explorer tool; Determine the purity of *Xbx14* protein after refining by Quantity One software.

**2.2.3. Microbiological methods:** Keep the strain of *E.coli*; cultivation produces single colonies; bacterial culture.

**2.2.4. Molecular biological methods:** Sequencing DNA using new Illumina's HiSeq2000, the method of transforming plasmid DNA into *E. coli* by heat shock, the method of extracting plasmid DNA from *E.coli;*, *g*enetic screening; DNA/protein electrophoresis on agarose/PAGE-SDS gel, co-expression method, PAGE – SDS.

**2.2.5. Methods of protein biochemistry:** Protein purification method by fractional precipitation, quantify nucleic acid/protein by density measurement, determination of β-xylosidase activity, determine the effect of temperature and pH, determination of heat resistance of enzyme, determine the kinetic parameters of the enzyme.

# Chapter 3. RESULTS AND DISCUSSION

## 3.1. RESEACH DIVERSITY OF BACTERIA AND LIGNOCELLULASE ACCORDING TO CAZY SYSTEM FROM DNA METAGENOME DATA OF MICROORGANISM IN GUT OF *C. gestroi* IN VIET NAM

### 3.1.1. Study on the diversity of bacteria in gut of *C. gestroi*

#### *3.1.1.1. The composition of the bacterial species has a large amount of ORF in gut of C. gestroi*

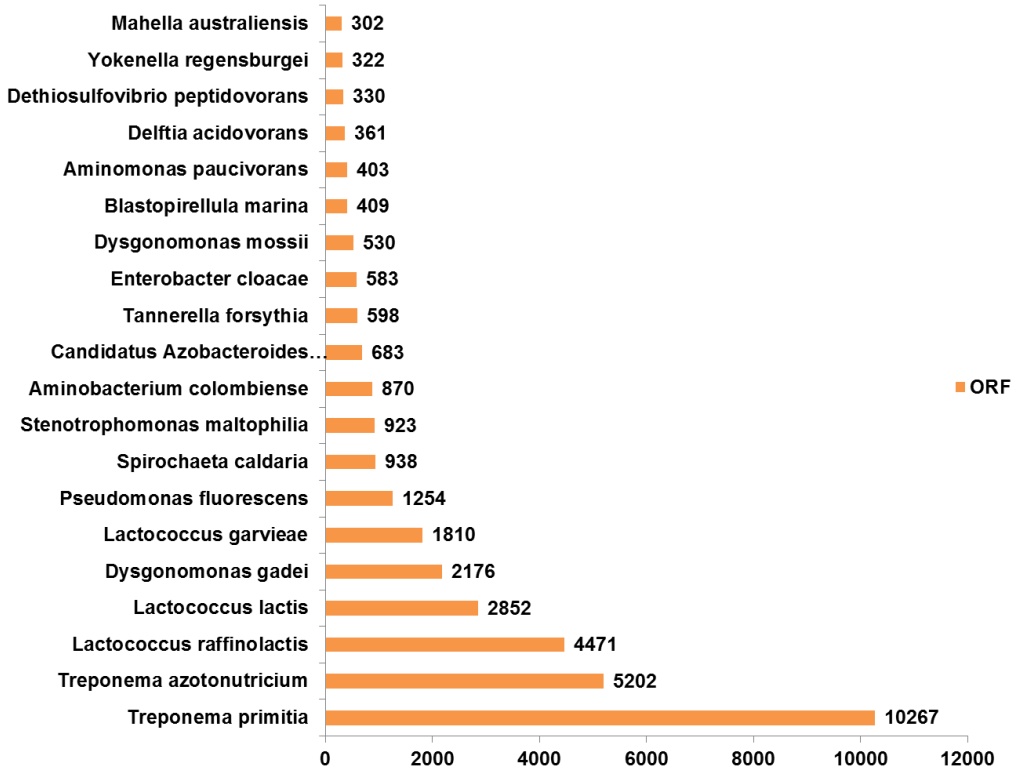


***Figure 3.1. A statistical map of the 20 largest ORF species in C. gestroi***

Previous research has used the Metagenomics to sequence the DNA metagenome of microorganisms in the gut of *C. gestroi* in Vietnam. According to the initial statistics we will focus on analyzing the largest number of microorganisms that are bacteria. This is a group of lignocellulase-secreting microorganisms that

decomposes lignocellulose completely in both low and high termites. Based on predicted species data, 20 bacterial counts of more than 300 ORFs were identified and filtered (Fig 3.1).

***3.1.1.2. Characters of bacteria species help C. gestroi to metabolise nitrogen***

Participation in the nitrogen metabolism have six species, including *Treponema primitia, Treponema azotaurricium, Candidatus Azobacteroides pseudotrichonymphae, Aminobacterium colombiense, Aminomonas paucivorans, Stenotrophomonas maltophilia*. In which there are three species *(Aminobacterium colombiense, Aminomonas paucivorans* và *Stenotrophomonas maltophilia)* who are only *C. gestroi*.

***3.1.1.3. Characters of bacteria species help C. gestroi to metabolise lignocellulose***

There are 16 species with large amounts of ORF in gut of *C. gestroi* involved in lignocellulose metabolism, including: *Treponema primitia, Treponema azotaurricium, Lactococcus raffinolactis, Lactococcus lactis, Lactococcus garvieae, Spirochaeta caldaria, Enterobacter cloacae, Mahella australiensis, Tannerella forsythia, Delftia acidovorans, Dethiosulfovibrio peptidovoran, Blastopirellula marina, Pseudomonas fluorescens, Yokenella regensburgei, Dysgonomonas gadei* và *D. mossii.* In which there are five species (*Mahella australiensis, Blastopirellula marina, Yokenella regensburgei, Delftia acidovorans, Dethiosulfovibrio peptidovorans)* who are only in *C. gestroi.*

In summary, based on the number of ORFs estimated from the DNA metagenome data of microorganisms in gut of *C. gestroi* in Vietnam, we found out 20 species of bacteria with very large ORF numbers. In which there are 03 species involved in nitrogen metabolism and 05 species involved in lignocellulose hydrolysis that are not present in other termites. As a result, There are not only protozoan but also bacteria to produce lignocellulase in gut of *C. gestroi* in Viet Nam

**3.1.2. The reseach diversity of lignocellulase in *C. gestroi***

***3.1.2.1. The ORF diversity of GH families hydrolyzing lignocellulose is classified according to CAZY***

Following BGI's anotation there are 587 ORF encoding the lignocellulase from the DNA metagenome data of the microorganism in gut of *C. gestroi*. To evaluate the diversity of lignocellulase, we classified them into the GHs in the CAZY system. Based on the results of CAZY's analysis, it will be easy to compare with similar research from other termites. Detailed results are presented in table 3.1.

***Table 3.1. Comparison of GH families from microorganisms in gut of C. gestroi with other termites***

| **GH No** | ***C. gestroi*** (symbiont)[1] | ***Nasutitermes*** (symbiont) [2] | ***R. speratus*** (symbiont)[3] | ***R. flavipes*** ((symbiont)[4] | ***R. flavipes*** (symbiont)[5] |
|---|---|---|---|---|---|
| GH1 | 187 | 22 | – | – | 3 |
| GH2 | 36 | 23 | – | 1 | 1 |
| GH3 | 205 | 69 | 1 | 10 | 2 |
| GH4 | 116 | 14 | – | – | – |
| GH5 | 136 | 56 | 3 | 11 | 2 |
| GH6 | 21 | – | – | – | – |
| GH7 | 20 | – | . | 35 | 10 |
| GH8 | 13 | 5 | 9 | 1 | 1 |
| GH9 | 16 | 9 | – | – | 2 |
| GH10 | 39 | 46 | . | 1 | 1 |
| GH11 | 52 | 14 | . | 10 | 5 |
| GH12 | 7 | – | – | – | – |
| GH13 | 10 | 48 | – | – | 10 |
| GH16 | 3 | 1 | – | 3 | 3 |
| GH17 | 2 | – | – | – | – |
| GH18 | 1 | 17 | – | 5 | 8 |
| GH20 | – | 15 | – | 3 | 4 |
| GH23 | – | 52 | – | – | – |
| GH25 | – | 1 | – | – | – |
| GH26 | 6 | 15 | . | 5 | 1 |
| GH27 | 24 | 4 | – | – | 1 |
| GH28 | 2 | 6 | – | – | – |
| GH30 | 2 | – | – | 1 | 3 |
| GH31 | 6 | 26 | – | – | – |

| | | | | | |
|---|---|---|---|---|---|
| GH32 | 16 | – | – | – | – |
| GH35 | 2 | 3 | – | – | – |
| GH36 | 44 | 5 | – | – | – |
| GH37 | 1 | – | – | – | 1 |
| GH38 | – | 11 | – | – | 2 |
| GH39 | 50 | 3 | – | – | – |
| GH42 | 30 | 24 | – | 1 | – |
| GH43 | 60 | 16 | . | – | 1 |
| GH44 | 6 | 6 | – | – | – |
| GH45 | – | 4 | 2 | 4 | 1 |
| GH47 | – | – | – | 1 | – |
| GH51 | 50 | 18 | – | – | – |
| GH52 | – | 3 | – | – | – |
| GH53 | 3 | 12 | – | 1 | 1 |
| GH55 | 1 | – | – | – | – |
| GH57 | – | 17 | – | – | – |
| GH58 | – | 1 | – | – | – |
| GH62 | 1 | – | . | – | – |
| GH65 | – | 6 | – | – | – |
| GH67 | 20 | 10 | – | – | – |
| GH68 | 1 | – | – | – | – |
| GH70 | – | – | – | – | 1 |
| GH74 | – | 7 | – | – | – |
| GH76 | – | – | – | – | 1 |
| GH77 | 1 | 14 | – | 1 | – |
| GH78 | 1 | – | – | – | – |
| GH79 | 18 | – | – | – | – |
| GH82 | 1 | – | – | – | – |
| GH85 | – | – | – | – | 1 |
| GH86 | 2 | – | – | – | – |
| GH87 | 4 | – | – | – | – |
| GH88 | – | 9 | – | – | – |
| GH91 | – | 1 | – | – | – |
| GH92 | – | 2 | – | 1 | – |
| GH93 | 1 | – | – | – | – |
| GH94 | 4 | 68 | – | – | – |
| GH95 | – | 12 | – | – | – |
| GH97 | 7 | – | – | – | – |
| GH98 | – | 1 | – | – | – |
| GH103 | – | 3 | – | – | – |
| GH106 | – | 2 | – | – | – |
| GH109 | – | 3 | – | – | – |
| GH113 | 3 | – | – | – | – |
| GH115 | 3 | – | – | – | – |
| GH116 | 1 | – | – | – | – |
| GH117 | 40 | – | – | – | – |
| GH119 | 2 | – | – | – | – |
| GH126 | 3 | – | – | – | – |
| GH128 | 3 | – | – | – | – |
| GH130 | 10 | – | – | – | – |

[1]Do et al., 2014 (*C. gestroi* symbiont metagenomic sequencing)

[2]Warnecke *et al.*(Scharf and Tartar, 2008) (*Nasutitermes* symbiont metagenomic sequencing).

[3]Todaka *et al.*(Scharf and Tartar, 2008) (*R. speratus* symbiont cDNA library).

[4]Tartar *et al.*(Scharf and Tartar, 2008) (*R. flavipes* symbiont cDNA library).

[5]Tartar *et al.*(Scharf and Tartar, 2008) (*R. flavipes* normalized gut cDNA library).

| |
|---|
| * GH identities were assigned using the CAZy (carbohydrate-active enzymes) database and nomenclature system (http://www.cazy.org/). Numbers indicate the number of genes identified within each GH. Solid circles (.) denote undeterminable numbers. Dashes (–) indicate missing GH members |

The results show that the greater number of GH families and ORFs belonging to each GH in the intestine of *C. gestroi* and higher other termites. We are interested in GH families that only appear in the gut of *C. gestroi* to determine the ability to decompose the components of lignocellulose, thereby we continue to select new genes from the available DNA metagenome data. Of the GH families classified by CAZY, the GH117 and GH130 families hydrolyse the branched chain of hemicellulose. These two enzyme families appear only in the *C. gestroi* with large numbers of ORFs of 40 (GH117) and 10 ORF (GH130), respectively. This demonstrates that *C. gestroi* has the potential for hydrolysing of hemicellulose. The abundance of hemicellulase from the DNA metagenome data of *C. gestroi* is a favorable basis for the selection genes coding hemicellulase in experiment.

#### *3.1.2.2. The ORF diversity of the β-xylosidase belongs to the GH according to the classification of CAZY*

In this study we are interested in the hemicellulase and one of the major enzymes involved in the hydrolysis of hemicellulose is β-xylosidase. The intestine of *C. gestroi*, there are 48 ORF encoding β-xylosidase, of which 23 ORFs are classified into 13 species, while 25 ORFs are not categorized into species. ORFs are classified into 8 different GHs and are dominated by GH43, 51, 116. Among the GH families containing β-xylosidase, the GH43 family appears in seven different bacterial species. Specially *Treponema primitia* contains ORF of 8 GH (Table 3.2).

***Table 3.2. Diversity of ORF and GH of β-xylosidase in gut of C. gestroi***

| Species | Amount of ORF | GH |
|---|---|---|
| *Bacteroides eggerthii* | 1 | 43 |
| *Clostridium hathewayi* | 1 | 51 |
| *Coprococcus eutactus* | 1 | 43 |
| *Dysgonomonas gadei* | 3 | 43, 51, 116 |
| *Enterobacter mori* | 1 | 43, 116 |
| *Lactococcus raffinolactis* | 2 | 43, 116 |
| *Leeuwenhoekiella blandensis* | 1 | 1 |
| *Mahella australiensis* | 1 | 3 |
| *Marvinbryantia formatexigens* | 2 | 1, 43, 116 |
| *Mitsuokella multacida* | 3 | 1, 51 |
| *Paenibacillus mucilaginosus* | 1 | 1 |
| *Treponema azotonutricium* | 1 | 51 |
| *Treponema primitia* | 5 | 1, 3, 5, 10, 39, 43, 51, 116 |

### 3.2. BUILD METHOD TO FIND OUT GENE CODING β-XYLOSIDASE FROM DNA METAGENOME DATA OF MICROORGANISM IN GUT OF *C. gestroi*

In this study we continue to improve and develop a new method to select genes from the DNA metagenome data. The constructing probes is based on the theory that the amino acid sequences of an enzyme have the same conservation sequences and the same active sites. Probe would help to select quickly the sequence of genes from the DNA metagenome data of microorganisms in the gut of *C. gestroi*. Genes have to encode the target enzyme exactly and perform well after they were expressed. Probe was built in steps:

#### 3.2.1. Identify GH families of CAZY containing β-xylosidase

We clasify the β-xylosidase into each GH, with detailed specification of spatial structure, amino acid composition and proton receptor during enzyme activity. Based on the high conservation of folding in the spatial structure of the enzymes will be classify into large “clan” groups. The results indicate that the enzyme belongs to the four clans: GH-A, GH-D, GH-F, GH-O and 11 families: GH1, 3, 30, 31, 39, 43, 51, 52 54, 116, 120. Details of each family are summarized in Table 3.3.

***Table 3.3. GH families contain the β-xylosidase according to CAZY***

| *Mã E.C* | *GH* | *Clan* | 3D Structure Status | Catalytic Nucleophile/Base | Catalytic Proton Donor | Mechanism |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| 3.2.1.37 | 1 | GH-A | $(\beta/\alpha)_8$ | Glu | Glu | Retaining |
| 3.2.1.8 | 3 | | | Asp | Glu | Retaining |
| 3.2.1.8 | 30 | GH-A | $(\beta/\alpha)_8$ | Glu | Glu | Retaining |
| 3.2.1.8 | 31 | GH-D | $(\beta/\alpha)_8$ barrel | Asp | Asp | Retaining |
| 3.2.1.8 | 39 | GH-A | $(\beta/\alpha)_8$ | Glu | Glu | Retaining |
| 3.2.1.37 | 43 | GH-F | 5-fold β-propeller | Asp | Glu | Inverting |
| 3.2.1.37 | 51 | GH-A | $(\beta/\alpha)_8$ | Glu | Glu | Retaining |
| 3.2.1.37 | 52 | GH-O | | Asp | Glu | Retaining |
| 3.2.1.37 | 54 | | | | | Retaining |
| 3.2.1.37 | 116 | GH-O | | Asp | Glu | Retaining |
| 3.2.1.37 | 120 | | | Asp | Glu | Retaining |

### 3.2.2. Searching the amino acid sequences of the β-xylosidase has been studied experimentally

We search for genes/amino acid sequences of β-xylosidase from CAZY and NCBI that have three criteria: (1) The gene coding β-xylosidase have been studied experimentally and available information about optimal expression, temperature and pH performance, to ensure that this sequence is indeed the correct activity of β-xylosidase; (2) select only the genes/amino acids sequences from the bacteria, to ensure the similarity and coincidence of amino acid residues; (3) Sequences should not differ too much in length, so that, after constructing the probe, the best reference value for the coverage and homology of the probe will be found with the target sequences. Detailed results are summarized in Table 3.4.

***Table 3.4. Summary of data has been studied on β-xylosidase***

| Order | Code in GENBANK | Bacteria | Amount of amino acids | Optimal pH | Optimal temperature ($^oC$) |
|---|---|---|---|---|---|
| **GH1** | | | | | |
| 1 | CAD20872.1 | *bacterium enrichment culture clone P1-6* | 464 | 6 | 40 |
| | | **GH3** | | | |
| 1 | CAD48309.1 | *C. stercorarium* | 715 | | 50 |
| **GH30** | | | | | |
| 1 | ABX45137.1.1 | *Bifidobacterium breve* | 448 | 6 | 45 |
| **GH43** | | | | | |
| 1 | CAA29235.1.1 | *Bacillus pumilus* | 535 | 7 | 40 |
| 2 | AAC97375.1 | *Bacillus pumilus PLS* | 535 | 6 | 45 |
| 3 | AAC27699.1 | *bacterium Bacillus sp. KK-1* | 533 | | 55 |
| 4 | BAA02527.1 | *Clostridium stercorarium* | 473 | 7 | 65 |
| 5 | BAC879411 | *Clostridium stercorarium* | 497 | 3.5 | 80 |
| 6 | AFZ7887.1 | *Enterobacter sp. enrichment culture clone nf1B6* | 536 | 6 | 40 |
| 7 | AAT9862.1 | *Geobacillus stearothermophilus T-6 (XynB3)* | 535 | 6.5 | 60 |
| 8 | ABC750041 | *Geobacillus thermoleovorans IT-08* | 511 | 5 | 70 |
| 9 | ADV16404.1 | *Paenibacillus woosongensis* | 477 | 6-7 | 30-45 |
| 10 | AEF2882.1 | *Thermobifida fusca TM51* | 550 | 4.5 | 50 |
| 11 | BAF982351 | *Vibrio sp. XY-214* | 535 | 7 | 36 |
| **GH52** | | | | | |
| 1 | BAA74507.1 | *Aeromonas caviae* | 729 | 8.7 | 60 |

| | | | | | |
|---|---|---|---|---|---|
| 2 | AGE344791 | *Geobacillus stearothermophilus* | 705 | 5.5 | 70 |
| 3 | ABI49956.1 | *Geobacillus stearothermophilus* | 705 | 6.3 | 65 |
| **GH120** | | | | | |
| 1 | ABM68042.1 | *Thermoanaerobacterium saccharolyticum JW/SL-YS48* | 636 | 6 | 65 |

Of the GH families, only the GH43 has 11 sequence which have large enough to compare and find similarities in the next steps.

### *3.2.3. Group the sequences to identify the similarities by ClustalW – PBIL*

ClustalW - PBIL is a tool that allows to compare multiple amino acid sequences to determine the level of amino acid preservation and similarities between sequences. The results of comparing the sequences will give the highest order of conservation (Symbol: *Prim.cons*).

```
Sequence1     --------------------KITNPVLKGFNPDPSICRAGEDYYMAVSTFEWFPGVQIY
Sequence2     --------------------KIINPVLKGFNPDPSICRVGEDYYMAVSTFEWFPGVQIY
Sequence3     --------------------KIINPVLKGFNPDPSICRVGEDYYIAVSTFEWFPGVQIH
Sequence7     -------------------SKIKNPILTGFHPDPSICRVGDDYYIAVSTFEWFPGVRIY
Sequence6     --------------------EITNPILTGFNPDPSLCRQGEDYYIATSTFEWFPGVRIY
Sequence11    ------------------TTTIQNPILKGFNPDPSIVRVGDDYYIATSTFEWFPGIQLH
Sequence10    ---TSPQVTSSPSREEPRAGTIRNPVLTGFYPDPSILRVGDDYYMATSTFEWYPGVTLH
Sequence8     --------------------EYSNPVIKGFYPDPSICRVGSDYYLVTSSFQYFPGVPIF
Sequence4     -----------------------RKQRFNPYLPSWEYIPDAEPYVFNGRVYIYGSHDRFN
Sequence9     -----------------------TKQGLNPYLPSWEYVPDGEPHVFNDRVYVYGSHDRFN
Sequence5     QPDNKHYKSAVRKWGDLGNGFYRNPVLNSDYSDPDVIRVGGDFYMVCSEFHYMGMPVLH
                                        :  :.   .. .    .  . :    : :     :
Prim.cons.    QP2222222222222234MKI2NPVLKGFNPDPSICRVGEDYYIA2STFEWFPGV2I2
Sequence1     -HSKDLIHWRLAARPLQKTSQLDMKGNPDSGGV-WAPCLSYADGQFWLIYSDIKVVDG-
Sequence2     -HSKDLVHWRLAARPLQKTSQLDMKGNPDSGGV-WAPCLSYADGQFWLIYSDIKVVDG-
Sequence3     -HSKDLVNWRLIAHPLQRVSQLDMKGNPDSGGV-WAPCLSYSEGKFWLIYTDVKVVDG-
Sequence7     -HSKDLKNWRLVARPLNRLSQLNMIGNPDSGGV-WAPHLSYSDGKFWLIYTDVKVVEG-
Sequence6     -HSRDLKNWTLVSTPLDRVSMLDMKGNPDSGGI-WAPCLSYADGKFWLLYTDVKIVDS-
Sequence11    -HSRDLINWRLVGHALTRTSQLNMMGMDNSEGV-YAPALTYSDGTFWLCFSNVHSCRGG
Sequence10    -HSRDLVHWRPLGGALTETRLLDLAGRRDGAGV-WAPALSYRDGLFFLVFTNVASYSG-
Sequence8     -HSTNLINWNKIGYCLIRPSQLMLNNATNRSGI-FAPTLRYHEGIFYLITTNVTLKKN-
Sequence4     GHAFCLNDYVCWSAPVDDLSEWRYEG-VIYRK---TDDPLNPDGRMCLYAPDVTLGPD-
Sequence9     GHAFCLNDYVCWSAPVADLADWRYEG-VIYKK---TDDPLNPDGSMCLYAPDVTVGPD-
Sequence5     --SKDLVNWTIIGRVYDSLKHDPKYDNMEGYAKGSAPAIRYHNGRFYVYF-CTPDEPG-
                :  * .:   .            .         :      :* : :         .
Prim.cons.    GHSKDLVNWRL22RPL2R2SQLDMKGNPDSGGVSWAPCLSY2DGKFWLIYTDVKVVDGG
Sequence1     PFKDGHNYLVTADAVDGEWS-DPVRLNSSG-DPSLFHDPSGKKYVLNMLWDHREKHHSF
Sequence2     PFKDGHNYLVTASEVDGDWS-EPIRLNSSG-DPSLFHDQSGKKYVLNMLWDHREKHHSF
Sequence3     AWKDCHNYLVTCETVDGDWG-EPIQLNSSG-DASLFHDVDGRKYLLNMLWDHRIGRHSF
Sequence7     QWKDGHNYLVTCDTIDGAWS-DPIYLNSSG-DPSLFHDEDGRKYLVNMYWDHRVDHHPF
Sequence6     PWKNGRNFLVTAPSIEGPWS-EPIPMGNGG-DPSLFHDDDGRKYYLYRPWGPRHHSNPH
Sequence11    NWMATPSYVVTADSIEGPWS-EPVPIGNYG-DPSLFHDDDGKKYMLNMIWGGRAKTNFF
Sequence10    NFWDAPNYVTTAPDITGPWS-DPVPLHSLG-DPSLFHDDDGRSWLLSTSMDWRPGRDAF
Sequence8     -------FIVMSEDLQGEWS-EPIWIDGWGGDPSLFFDNGKVYITGTND-NARGEE--
Sequence4     ----GRYYLYYVLDKVPVVS-VAVCDTPAG-KYEFYGYVR---YADGTRLGEREGD--
Sequence9     ----GRYYLYYVLDKVPIVS-VAVCDSPAG-EYEFYGYVR---YSDGTRLGERQGD--
Sequence5     ------LFMSTATDPAGPWSPLHEVVRVAGWDPCPFWDDG------NAYLGHSTVG-A
                     ::          .         * .   :             .
Prim.cons.    PWKDGHNYLVTADDIDGPWSPEP2PLNSSG2DPSLFHDDDG2KYLLNMLWGHREGHHSF
Sequence1    AGIALQEYSVSEKKLVGERKVIFKGTPIKLTEAPHLYYINDVYYLLTAEGGTRYEHAATI
Sequence2    AGIALQEYSVAEKKLIGRRKVIFKGTPIKLTEAPHLYHIGDNYYLLTAEGGTRYEHAATI
Sequence3    GGIVMQEFSVREKRLIHQPKIIFEGTDIKLTEAPHLYHIGDYYYLLTAEGGTRYEHAATI
Sequence7    YGIVLQEYSVEQKKLVGEPKIIFKGTDLRITEGPHLYKINGYYYLLTAEGGTRYNHAATI
Sequence6    NTIVMQEFDPQTGTLSPERKTLFTGTPLCYTEGAHLYRHAGWYYLMVAEXGTSYEHAVVV
Sequence11   GGIIMQEFDADEGKLVGAPKTVFEGTELGCTEGPQLLKKDDYYYLITAEGGTERNHAVTV
Sequence10   GGIVAQEFSVRDMKLVGEPVIIFTGTEAGVTEAPHIYKRDGWYYLVTAEGGTQWEHQVTV
Sequence8    LGIYQAEIDLKKGSIIGERKLIWKGTGGSYPEAPHLYKVNGWYYLLIAEGGTEYGHMVTV
Sequence4    ----WPQFDPAVLTE-GERTYLYTG----FCPKGDKSRKGAMATVLGPDMLTVVEEPVII
```

```
Sequence9    ----EPQFDPAVLTE-GEFTYLYTG----FCAIGDKSRKGAMATVLGRDMLTIVEEPVFV
Sequence5    GPIIIHKMSPDGTKLLDDGVIVYVG---KIAEGTKIYKRNGYYYLIIPEGGVKTG-WQTV
                   : .            :: *         .         ::  :  .       :
Prim.cons.   GGIV2QEFSVAEKKLVGERKIIF2GTP2KLTEAPHLYKI2GYYYLLTAEGGTRYEHAVTV
Sequence1     ARSSRIDGPYEVHPDNPILTAFHAPSHPLQKCGHASIVQTHTNEWYLAHLTGRPIH-SK
Sequence2     ARSSHIEGPYEVHPDNPILSAFHAPEHPLQKCGHASIVQTHTNEWYLAHLTGVPIQ-NK
Sequence3     ARSKHIEGPYEIHPDNPILTSWHEPRNPLQKCGHASIVETHTGEWYLAHLTGRPIH-AD
Sequence7     ARSTSLYGPYEVHPDNPLLTSWPYPRNPLQKAGHASIVHTHTDEWFLVHLTGRPLP-EG
Sequence6     LRAKTIDGPYELHPDVTMMTSWHLPENPLQKSGHGSLLQTHTGEWYMAYLTSRPLR-PG
Sequence11    -RSKHIWGPYEVHPENPILTSRFQEHAELSRAGHGFLVETQTGEWYMSHLCGRRIPNPE
Sequence10    ARSRSVTGPYEVDPAGPALTSRHVPEAPLQKAGHASMVETQHGEWYFAHLTGRPMP-P-
Sequence8     ARSKYPFGPFESCPFNPILTHRS--NHPLQAIGHADIVQYHDGSWWAVFHGTRPISYPP
Sequence4     VPSE----PYSRGSG—EGHEFFEAPSNPIRKKGDTYYFIYSSVVMHELCYATSKHPTKG
Sequence9     APSE----PYSKGSG--EGHEFFEAPSPIRKRGDTYYLIYSSVVMHELCYATSPFPTKG
Sequence5     LRSKSIYGPYERKVV-----LQTG-NTNINGPHQGALVELEGESWFMHFQDTGIL----
                :     *:.                  :    .   .
Prim.cons.    ARSK2I3GPYEVHPDNPILTS2H2PE2PLQK2GHASIV2THTGEWYLAHLTGRPPTPPG
Sequence1     ESIFQQ-RGWCPLGRETAIQKLEWK-DGWPYVVGGKEGLLEVEAPA-MSVKEFSPTYHI
Sequence2     ESIFQQ-RGWCPLGRETAIQKLEWK-DGWPYVVGGKEGTLEVEAPK-IEEKDFAPTYHT
Sequence3     DSVIHQ-RGYCPLGRETAIQKLEWK-DGWPYVAGGKEGRVEVEAPR-ISETKFPSTYQE
Sequence7     QPLLEH-RGYCPLGRETAIQRLEWK-DGWPYVVGGNGPSLEIDGPS-VEEVSWEKDYDE
Sequence6     PLLASGGRGYCPLGRETGIARIEWR-DGWPYVEGGKHAQLTVKGPQ-VAEQPAAVQGSW
Sequence11    QFMPKYDNGFSILGRESALQKAHWQ-DDWPYIATGKTPVVEVEAPN-LPLHPWPES-PA
Sequence10    -------SGRCVLGRETALQKIEWSSDGWPRVRN-AEPLLEVPGPRGLAPHPWPQP-SE
Sequence8     ---------KHHLGRETCLAPIKWTDDGWPIIGYNGRIDIKMDAGYLPVKEKNIGDEII
Sequence4     ----FKYGGVIVSNCDLHIDSYKPAEKPMYYGGNNHGSIVEINGEWYIFYHRHTNG---
Sequence9     ----FTYQGVIVSNNDLHIDSYKPADKPMYYGGNNHGGAVEIQGQWYIFYHRHTNG---
Sequence5     -------------GRVCHLQPVTWV-DNWPLMGCDGEPVTVYRKPRAGKEYERTFP-QT
                            .    :       .
Prim.cons.    ES22QQ4RGYCPLGRETAIQKLEWKDDGWPYVGGGKE22LEVE2PRYI4EHP222TY3E
Sequence1     VDEFKDSSLNRHFQTLRIPFTDQIGSVTENPHHLRLYGQE-LTSKFTQAFVARRWQSFY
Sequence2     VDEFKESTLNRHFQTLRIPFTDQIGSLTEKPRHLRLFGRE-LTSKFTQAFVARRWQSFY
Sequence3     VDQFADRTLNMNFQTLRIPFTKELGSLTERPNHLRLYGRE-LTSTFTQAFVARRWQSLR
Sequence7     KDDFDGDTLNHHFQTLRIPLGEDIATLKARPGHLRLYGRE-LTSRFTQAFVARRWQHFH
Sequence6     RDDFDGSTLDPELQTLRIPFDDTLGSLTARPGYLRLYGND-LNSTFTQSTVARRWQHFI
Sequence11    RDEFIDPTLS-LISTLREPVSEKWLSLSERPGFLRLKGRH-LYSRYEQSMVARRFQAHN
Sequence10    TDHFDDPTPRPEWSTLRRPFDSSWVSLTERPGYLRIRGGQ-PAGLHEPSLVARRLQHRA
Sequence8     EDDFNSDIFSTDWNFIQNPRLEHYSS-KG RPSWLKMRGTEKTLNDINSPTFIRRQEHF
Sequence4     -TSFSRQGCMEKIKILEDGSIPQVEMTSCGSADEPLPGRG-YPAYIACNLFCGEESVYT
Sequence9     -TAFSRQGCMEPISFREDGTIPQVEMTSCGPNGGPLAGRG-YPAYLACNLFCKDEELYT
Sequence5     SDEFDGPELGLQWQWNHNPVNERWS-LSKRPGYLTLEAMY-ESLLKARNTLTQKLIGEK
                 *          .  .          .  .    : .          .
Prim.cons.    VDEFDD2TLN3HFQTLRIPFTEQ2GSLTERPGHLRLYGRESLTS5FTQAFVARRWQHF2
Sequence1     FEAETAVSFFPKNFQQAAGLVNYYNTENWTALQVTYDD-LGRILELSVCENLAFSQPLI
Sequence2     FEAETAVSFFPENFQQSAGLVNYYNTENWTALQVTYDE-LGRTLELSVCQNLAFSQPLA
Sequence3     FTAETAVDFNPETFQQAAGLVNYYNTENWTALQVTYDE-LGRILDVTICDNFTFSQPIK
Sequence7     FVAETKVSFRPTTFQQSAGLVNYYNTQNWTTLQITWHE-KGRILELMTCDHLVVDQPLR
Sequence6     FRAETRMQFSPVHFQQSAGLTCYYNSKNWSYCFVDYEE-QGRTIKVIQLDHNVPSWPLH
Sequence11    ATVETKLEFKPNTPYEMAGLCAYYARNGHYFLKMTAND-GERVLQVVGNINDVYGEYSN
Sequence10    CIFEACLEFKPEDFRQMAGITAYYNTRQWHYLRINRDD-GGVFAGVLTSDRGIIREVGR
Sequence8     NVS-TLLEFKPNQDNEEAGLTVYMNEKHHYEIAL-TKKNGRINVVLKKTVGDIQVVVN-
Sequence4     ---LTGAWMNN-QFPKITQDGKDGDEEPGYIANMKDSATAGFKYFDCKGIKSVKIKVR-
Sequence9     GFGASGAWMDS-RFPKITQDGKDGDEEMGYIANM-TDSATAGFKYFDCHGIRRMTIQV-
Sequence5     GTATTELNTENLKNQQRAGLAFLGGTQENWIG-VREGE---SYIKAVTAGIRYHGPEIE
                                : :          .      :
Prim.cons.    FTAETA22FKP2TFQQSAGLVNYYNTENW22LQVTYDET-GGR2L3V33CDNLV2SQPL
Sequence1     KK-IIIPDEIPYVYLKVTVQRE-TYTYSYSFDQQEWEKIDVPLESTHLSDDFIRGGG--
Sequence2     HK-IIIPDEVTYVYLKVTVRKE-TYKYSYSFDQKEWKEIDVPLESIHLSDDFIRGGG--
Sequence3     DK-IVIPRDIQYVYLRVNVEWE-TYYYSYSFNKKDWYKIEIPLESKKLSDDYVRGGG--
Sequence7     GREIVVPDDIEYVYLRVTVQAT-TYKYSYSFDGMNWIDLPVTFESYKLSDDYIKSRA--
Sequence6     EQPIPVPEQAESVWLRVDVDRL-VYRYSYSFDGETWHAVPVTYEAWKLSDDYIGGRG--
Sequence11    DV---VIGDADTVYMRLELKTQ-WYQYSYSLDGVDWYEIGPALNSTPLSDEGGPDIF--
Sequence10    RIS-V-TGWPKVFLRAEIDRN—LRFAV--SSDGSTWADMGVRLDMSILSDEYAEERFGN
```

```
Sequence8     YFS-------NTIIFSIQANPE-EYKFSFVDPNTGQTYLLGTGLTTLLSTEVAGGFT--
Sequence4     CR------------G--------VFEVKTSWNGEVLGKIPVEFSN-IWTEFSASIP---
Sequence9     CR------------G--------AFEIKTAWNGPVLGTIPVEFSN-VWKPYSTELV---
Sequence5     TP---------NVWFRAEIDLNGITQFYFSTDNENFIQLGGPCRL-----EAG------
                                                      :
Prim.cons.    2-SI2VPDDI2YVYLRV2VDRE-TYKYSYSFDGE5W3KIPVPLESTKLSD2YIRG2GGN
Sequence1     ----FFTGAFVGMQCQ-DTSGERLPADFKYFRYEETTE------
Sequence2     ----FFTGAFVGMQCQ-DTSGERLPADFHYFRYEETDE------
Sequence3     ----FFTGAFVGMQCQ-DTSGAHQHADFDYFMYKEL--------
Sequence7     ----AFTGAFVGMHCR-DGSGQNNYADFDYFLYKEL--------
Sequence6     ----FFTGAFVGLHCE-DISGDGCHADFDYFTYEPA--------
Sequence11    ----RFTGSFAALFVA-DITGQKRHADFDYFEYLEHSEQXENCE
Sequence10    PIMWGFTGAFLGLWAH-DMTGAGLPADFDFCTYRPQSPS-----
Sequence8     -------GVYFGLYATGNGKVCTAPAFFDWFKYIPEI-------
Sequence4     -----IPDGIHALYFTYRGSGSASLKSFTLCTD-----------
Sequence9     -----IPDGIQALYFTYTGMGSASLASFTLE-------------
Sequence5     ----FWKGARIALFSY-NTVMDGGRADFNWFRYEFE--------
                     .   .:              *
Prim.cons.       PIMWFFTGAFVGL2C2YD2SG5GLPADFDYF2YEE3SE2XENCE
```

***Figure 3.2. Results of comparison of the 11 sequences and constructing the β-xylosidase probe for the GH43 family. Note: The level of conservation of amino acid radicals is marked from the "*" to the ":" "." and is not marked. The undercuts in Sequence Prim.Cons will be used to make probes***

ClustalW-PBIL will calculate and return to results in minutes depending on the number of sequences after entering data. In order to be able to construct the probe, the greater the number of β-xylosidase coding sequences have, the higher the probability of the probe is. Based on amino acid data belonging to different GH families, only GH43 family has a sufficient number of sequences (11 sequences) to be able to compare similar regions. The enzymes collected in the GH43 family which have activities in optimally pH from the acid (pH = 3.5) to neutral (pH = 7) and temperatures between 30°C and 80°C. The length of sequences are not much different, from 473 amino acids to 550 amino acids (Table 3.4). The results of the analysis after comparing the 11 sequences of GH43 with the ClustalW-PBIL software are shown in Figure 3.2.

Based on the level of preservable amino acid sites in the polypeptide chain, when comparing the 11 sequences showed that: seven high conservable amino acid residues (red) is the same sequences, 32 medium conservavle amino acid residues is similar in most sequences (green) and 30 lower conservable positions (blue) is similar in some sequences. At the remaining amino acid sites, the similarities were compared, including: 71 amino acid positions were the same from 8 to 10 sequences (orange), 102 positions were the same from 6 to 7 sequences (pink). In other places, the larger the amino acids will be retained (black). In the case the number of amino acid types is the same, the final sequence will be the number that represents the most repetitive amino acid number (Example “2” meaning that there are two amino acids with numbers the largest equal amount when comparing between 11 sequences). Amino acid sites are less than 50% sequences will be marked purple. This result shows that the amino acid sequences of β-xylosidase belong to the GH43 in bacteria which are highly preserved. The high conservation of amino acids in β-xylosidase sequences suggests that the selection of sequences from bacteria only ensures reliable probe data.

### 3.2.4. Build probe and reference values

Based on the results of comparison between the 11 sequences of the GH43 we will select the preserved sites (red, green, blue), the location of the majority of amino acids are the same (orange and pink), larger proportion (black) and location of equal amino acid (number) to ensure the length of the probe. Amino acid sites are less than 50% sequence will be removed. This probe will be compared against 11 sequences to determine coverage and homology by BLASTP.

The GH43 family's probe was constructed consisting of 464 amino acids. It contains all 7 red amino acids, 32 green positions, 30 blue positions, 71 orange amino acid positions, 102 pink positions, 165 black positions and 57 numbers (Figure 3.2). The probe sequence obtained for the GH43 family is shown in Fig 3.3.

```
I2NPVLKGFNPDPSICRVGEDYYIA2STFEWFPGV2I2HSKDLVNWRL22RPL2R2SQ
LDMKGNPDSGGVWAPCLSY2DGKFWLIYTDVKVVDGPWKDGHNYLVTADDIDGPWSPE
P2PLNSSGFDPSLFHDDDG2KYLLNMLWGHREGHHSFGGIV2QEFSVAEKKLVGERKI
IF2GTP2KLTEAPHLYKI2GYYYLLTAEGGTRYEHAVTVARSK2I3GPYEVHPDNPIL
TS2H2PE2PLQK2GHASIV2THTGEWYLAHLTGRPIPTPPRGYCPLGRETAIQKLEWK
DGWPYVGGGKE22LEVE2PRYI4EHP222TY3EVDEFDD2TLN3HFQTLRIPFTEQ2G
SLTERPGHLRLYGRESLTS5FTQAFVARRWQHF2FTAETA22FKP2TFQQSAGLVNYY
NTENW22LQVTYDETGGR2L3V33CDNLV2SQPL3DDI2YVYLRV2VDREYKYSYSFD
GE5W3KIPVPLESTKLSD2YIRG2GGFFTGAFVGL2C2YD2SG5GLPADFDYF2YEE3
```

***Fig 3.3. Sequence probe belong to GH43 of β- xylosidase***

The comparison of the level of homology between the probe and each sequence using by BLASTP indicates that sequences encoding β-xylosidase (GH43) must have a maximum score of more than 17, minimum coverage and homogeneity of 60% and 30%, respectively (Table 3.5).

***Table 3.5. Comparison of homology with sequences of GH43***

| Sequence | Maxscore | Total score | coverage (%) | E-value | Ident (%) |
|---|---|---|---|---|---|
| Sequence 2 | 608 | 608 | 100 | 0 | 72 |
| Sequence 1 | 603 | 603 | 99 | 0 | 71 |
| Sequence 3 | 596 | 596 | 99 | 0 | 71 |
| Sequence 7 | 547 | 547 | 99 | 0 | 66 |
| Sequence 6 | 444 | 444 | 99 | 5E-155 | 56 |
| Sequence 10 | 358 | 377 | 98 | 9E-122 | 51 |
| Sequence 11 | 351 | 351 | 99 | 4E-119 | 49 |
| Sequence 8 | 231 | 248 | 99 | 2E-73 | 37 |
| Sequence 5 | 101 | 134 | 83 | 3E-27 | 30 |
| Sequence 4 | 25.8 | 76.2 | 61 | 0.003 | 33 |
| Sequence 9 | 17.3 | 113 | 60 | 1.2 | 46 |

However, the results showed that this probe was most suitable for sequences 1, 2, 3, 6, 7, 8, 10, 11 (maximum score over 200, homology and coverage over 37% and 99% respectively) and does not represent 4, 5, 9 sequences. Further exploration of sequence 4, 5 and 9 we continue to use the Swiss model to validate the activity of amino acids 4, 5 and 9 sequence again. The results show that the spatial structure of all three sequences is highly homologous to the structure of β-xylosidase/α-arabinofuranosidase. This result may be assumed that sequence 4, 5 and 9 may be due to the simultaneous performance of two functions, but the alpha-arabinofuranosidase function is more active.

***Table 3.6. Estimates of the tertiary structure of 4, 5, 9 sequence by the Swiss model***

| Sequence | Template | Active enzyme | Coverage | Ident | Method | Structure | Ligand |
|---|---|---|---|---|---|---|---|
| Sequence 4 | 4nov.1.A | Xylosidase/arabinofuranosidase Xsa43E | 0.6 | 30.28 | X-ray, 1.3Å | Monomer | 1 x CA |
| Sequence 5 | 3c2u.1.A | Xylosidase/arabinofuranosidase | 0.94 | 27.08 | X-ray, 1.3Å | homo- | 4 x B3P |
| Sequence 9 | 5glk.1.A | β-xylosidase/α-arabinofuranosidase | 0.56 | 36.84 | X-ray, 1.7Å | Monomer | None |

To exploit maximize the genes coding for β-xylosidase, using probes with maxscore, homology and minimum coverage of 17, 30% and 60%. However, in order to ensure successfully experiement, maxscore, homology and minimum coverage are 200, 37 and 99%, respectively.

### 3.2.5. Exploitation of the β-xylosidase by probe from DNA metagenome data

BGI's anotation noted 46 sequence-encoding β-xylosidases (GH43). When using the probe, if we use a maxscore of 17, minimum coverage of 60% and indentity of 30% we will select 25 sequences, of which 20 sequence coincides with the BGI's prediction. However, there are still 5 sequences different BGI's prediction. If based on a maxscore of more than 200, a minimum homology of 37% and a minimum coverage of 99% are recommended above, there is only one sequence suitable reference index (Table 3.7).

***Table 3.7. Compare the number of sequences which are selected by probes with the BGI's anotations***

| BGI predictions | The number of sequences which are selected by probe |
|---|---|

| | Code gene | Maxscore | Total score | Coverage (%) | E value | Ident (%) |
|---|---|---|---|---|---|---|
| GL0104795 | GL0104795 | 382 | 382 | 99 | 1e-130 | 48 |
| GL0020896 | GL0020896 | 285 | 285 | 62 | 3e-96 | 54 |
| GL0117445 | GL0117445 | 285 | 285 | 62 | 3e-96 | 54 |
| GL0076106 | GL0076106 | 243 | 259 | 85 | 2e-77 | 38 |
| GL0006405 | GL0006405 | 227 | 271 | 68 | 7e-75 | 57 |
| GL0044986 | GL0044986 | 225 | 256 | 87 | 7e-72 | 37 |
| GL0077826 | GL0077826 | 216 | 265 | 67 | 9e-70 | 48 |
| GL0112518 | GL0112518 | 216 | 265 | 67 | 9e-70 | 56 |
| GL0001262 | GL0001262 | 214 | 266 | 82 | 8e-67 | 35 |
| GL0095948 | GL0095948 | 213 | 213 | 63 | 3e-53 | 39 |
| GL0015489 | GL0015489 | 201 | 201 | 67 | 2e-49 | 42 |
| GL0088906 | GL0088906 | 154 | 193 | 62 | 2e-48 | 53 |
| GL0016592 | GL0016592 | 130 | 160 | 64 | 8e-40 | 49 |
| GL0021333 | GL0021333 | 99.8 | 117 | 77 | 2e-27 | 30 |
| GL0090776 | GL0090776 | 54.7 | 119 | 63 | 4e-12 | 37 |
| GL0083296 | GL0083296 | 43.9 | 58.1 | 67 | 4e-09 | 35 |
| GL0091901 | GL0091901 | 40.4 | 54.3 | 64 | 3e-08 | 35 |
| GL0079004 | GL0079004 | 37.0 | 85.1 | 65 | 5e-07 | 36 |
| GL0050001 | GL0050001 | 35.8 | 112 | 67 | 8e-07 | 36 |
| GL0106540 | GL0106540 | 34.7 | 72.0 | 66 | 2e-06 | 35 |
| | GL0119754 | 93.2 | 156 | 61 | 3e-26 | 38 |
| | GL0039878 | 55.8 | 103 | 69 | 2e-13 | 36 |
| | GL0122352 | 27.3 | 102 | 69 | 4e-04 | 32 |
| | GL0068837 | 24.6 | 115 | 66 | 0.002 | 35 |
| | GL0042431 | 22.7 | 84.3 | 61 | 0.011 | 31 |

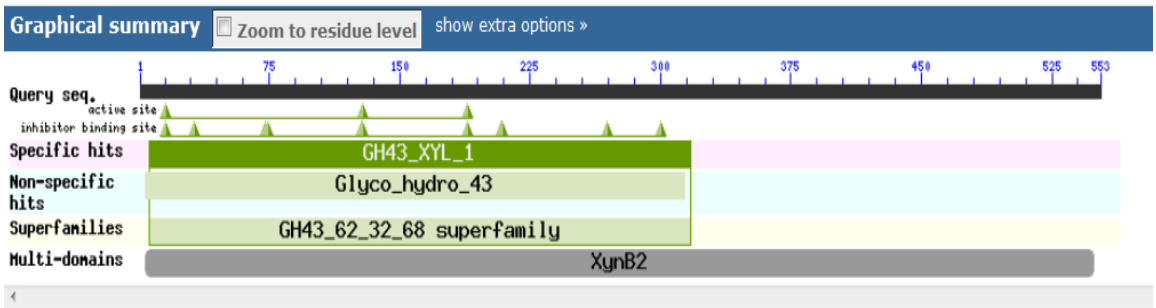


***Figure 3.4. The result of prediction for specific hit and activity site of sequence were selected by the GH43 probe.*** *Note: GH43_XYL:glycosyl family 43 capable of decomposing beta-D-xyloside structure; XynB2: beta-xylosidase*

Results show that: All of 25 sequences which are selected by probes contain specific regions for β-xylosidase (specific hit) and active sites (Figure 3.4). Meanwhile, many of the gene sequences encoding by β-xylosidase predicted by BGI do not encode for this enzyme. Thus the using probes will search less gene sequences than BGI's prediction, but more accurate.

### 3.2.6. Investigation of the tertiary structure of the β-xylosidase exploited by the probe

To determine the accuracy of the selected sequence by the probe, we continue to check the spatial structure by the Swiss model tool. Results showed that the majority of selected sequences by probes were estimated to be highly homologous to β-xylosidase. However, some sequences which have a maxscore of less than 100 have other activities of β-xylosidase such as xylanase and arabinofuranosidase (Table 3.8). The maxscore is important to consider the authenticity of genes and proteins when using bioinformatic tools to estimate. This result again helps us confirm that selecting a maxscore of over 200 as the reference value when using the probe will allow for greater accuracy when selecting genes for experimentation.

***Table 3.8. Estimation of the tertiary structure of sequences by the Swiss model***

| Code gene | Maxs core | Template | Active enzyme | coverage | ident | Method | structure | Ligands |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| GL0104795 | 382 | 3c2u.1.A | Xylosidase/ Arabinosidase | 0.95 | 54.77 | X-ray, 1.3Å | homo-tetramer | 4 x B3P |
| GL0020896 | 285 | 3c2u.1.A | beta-D-xylosidase | 0.93 | 52.11 | X-ray, 1.3Å | | 4 x B3P |
| GL0117445 | 285 | 2exj.1.A | beta-D-xylosidase | 0.96 | 55.34 | X-ray, 2.2Å | | 4 x XYS-XYS, 4 x CA, 3 x MES |
| GL0076106 | 243 | 2exk.1.A | beta-D-xylosidase | 0.94 | 37.55 | X-ray, 2.2Å | | 4 x XYS-XYS, 4 x CA, 3 x MES |
| GL0006405 | 227 | 2exk.1.A | beta-D-xylosidase | 0.98 | 59.53 | X-ray, 2.2Å | | 4 x XYS-XYS, 4 x CA, 3 x MES |
| GL0044986 | 225 | 2exj.1.A | beta-D-xylosidase | 0.94 | 40.29 | X-ray, 2.2Å | | 4 x XYS-XYS, 4 x CA, 3 x MES |
| GL0077826 | 216 | 1yrz.1.A | beta-D-xylosidase | 0.96 | 57.85 | X-ray, 2.0Å | monomer | No |
| GL0112518 | 216 | 1yrz.1.A | beta-D-xylosidase | 0.93 | 58.86 | X-ray, 2.0Å | monomer | No |
| GL0001262 | 214 | 2exk.1.A | beta-D-xylosidase | 0.94 | 36.73 | X-ray, 2.2Å | homo-tetramer | 4 x XYS-XYS, 4 x CA, 3 x MES |
| GL0095948 | 173 | 1yi7.1.A | beta-D-xylosidase | 0.93 | 46.52 | X-ray, 1.9Å | homo-tetramer | 4 x CA |
| GL0015489 | 161 | 2exh.1.A | beta-D-xylosidase | 0.98 | 46.56 | X-ray, 1.9Å | homo-tetramer | 4 x CA, 3 x MES |
| GL0088906 | 154 | 1yrz.1.A | beta-D-xylosidase | 0.96 | 64.1 | X-ray, 2.0Å | monomer | No |
| GL0016592 | 130 | 1yi7.1.A | beta-D-xylosidase | 0.98 | 48.9 | X-ray, 1.9Å | homo-tetramer | 4 x CA |
| GL0021333 | 99.8 | 1yrz.1.A | beta-D-xylosidase | 0.96 | 31.65 | X-ray, 2.0Å | monomer | No |
| GL0090776 | 54.7 | 2exj.1.A | Endo-1,4-beta-xylanase | 0.42 | 24.22 | X-ray, 2.2Å | homo-tetramer | 4 x XYS-XYS, 4 x CA, 3 x MES |
| GL0083296 | 43.9 | 3c7g.1.A | Endo-1,4-beta-xylanase | 0.93 | 33.88 | X-ray, 2.0Å | monomer | 4 x XYP, 1 x CA |
| GL0091901 | 40.4 | 3c7g.1.A | Endo-1,4-beta-xylanase | 0.91 | 31.94 | X-ray, 2.0Å | monomer | 4 x XYP, 1 x CA |
| GL0079004 | 37 | 1yrz.1.A | xylan beta-1,4-xylosidase | 0.94 | 18.62 | X-ray, 2.0Å | monomer | No |
| GL0050001 | 35.8 | 3qee.1.A | β-xylosidase/α-L-arabinofuranosidase | 0.99 | 56.51 | X-ray, 1.6Å | monomer | 1 x CA |
| GL0106540 | 34.7 | 4nov.1.A | β-xylosidase/α-arabinofuranosidase | 0.93 | 55.74 | X-ray, 1.3Å | monomer | 1 x CA |
| GL0119754 | 93.2 | 2exh.1.A | xylan beta-1,4-xylosidase | 0.81 | 42.77 | X-ray, 1.9Å | homo-tetramer | 4 x CA, 3 x MES |
| GL0039878 | 55.8 | 1yrz.1.A | xylan beta-1,4-xylosidase | 0.94 | 31.09 | X-ray, 2.0Å | monomer | No |
| GL0122352 | 27.3 | 3kst.1.A | Endo-1,4-beta-xylanase | 0.9 | 36 | X-ray, 1.7Å | monomer | 1 x CA |
| GL0068837 | 24.6 | 2exk.1.A | Endo-1,4-beta-xylanase | 0.89 | 25.68 | X-ray, 2.2Å | homo-tetramer | 4 x XYS-XYS, 4 x CA, 3 x MES |
| GL0042431 | 22.7 | 3c7f.1.A | Endo-1,4-beta-xylanase | 0.82 | 32.08 | X-ray, 1.5Å | monomer | 1 x CA |

*Note:B3P:2-[3-[[1,3-dihydroxy-2-(hydroxymethyl)propan-2-yl]amino]propylamino]-2-(hydroxymethyl) propane-1,3-diol; XYS: xylopyranose; CA: $Ca^{++}$; MES: 2-morpholin-4-ium-4-ylethanesulfonate*

According to Table 3.7 if we based on a maxscore of more than 200, a minimum homology of 37%, and a minimum coverage of 99% there is only one sequence meet the requirements of the reference index. So in order to have more choices when put into practice, we still use GH43 probe with a coverage of 60% and homology 0f 30% reference index, maxscore of 200 and we filtered 11 gene codes from the DNA metagenome data of microorganisms in gut of *C. gestroi* including: GL0104795, GL0020896, GL0117445, GL0076106, GL0006405, GL0044986, GL0077826, GL0112518, GL0001262, GL0095948, GL0015489.

**3.2.7. Predict the structure and function of the gene coding β-xylosidase by bioinformatic tools**

*Table 3.9. Coverage and homology with probes of genes coding β-xylosidase*

| Order | Code gene | Coverage (%) | Ident (%) | Completed gene |
|---|---|---|---|---|
| 1 | GL0104795 | 99 | 48 | X |
| 2 | GL0020896 | 62 | 54 | |
| 3 | GL0117445 | 62 | 54 | |
| 4 | GL0076106 | 85 | 38 | X |
| 5 | GL0006405 | 68 | 57 | |
| 6 | GL0044986 | 87 | 37 | |
| 7 | GL0077826 | 67 | 48 | |
| 8 | GL0112518 | 67 | 56 | X |
| 9 | GL0001262 | 82 | 35 | X |
| 10 | GL0095948 | 63 | 39 | |
| 11 | GL0015489 | 67 | 42 | |

Selection of the gene for the experiment will prioritize the sequences which were chosen by the probe and the initial estimate of the completed ORF. Out of 11 genetic codes, there are completed ORF including: GL0104795, GL0076106, GL0001262 and GL0112518 (Table 3.9).

If the best criterion for the selection of genes in experiment was a maximum score of over 200, homology of 37% and and coverage of 99%, there is only the GL0104795. However, in addition to selecting the genes coding for the β-xylosidase, we are particularly interested in the ability of the enzyme which can act in optimally alkaline and high temperature. Thus, four complete genes will continue to be checked by Alcapred and TBI's tools.

*Table 3.10. Results of the prediction about optimum pH and temperatures of β-xylosidase*

| Order | Code gene | Amount of amino | Optimum pH | Optimum temperatures |
|---|---|---|---|---|
| 1 | GL0104795 | 561 | 0.217171 | 55℃~65℃ |
| 2 | GL0076106 | 553 | 0.871792 | 55℃~65℃ |
| 3 | GL0001262 | 555 | 0.509051 | 55℃~65℃ |
| 4 | GL0112518 | 359 | 0.984522 | 55℃~65℃ |

Using TBI's software predicts all four enzymes are capable of optimal temperature between 55℃ and 65℃. Using Alcapred's software, the results showed that GL0104795; GL0001262; GL0076106 and GL0112518 with a predicted index of 0.217171; 0.509051; 0.871792 and 0.984522 (Table 3.10). According to the predicted results of this software, the index is the closer the value of 1, the active enzyme in optimum alkaline environment will be higher. Thus, the GL0112518 gene is predicted with the highest alkaline tolerance will be prioritized for inclusion in the experiment.

In summary, using the probe of GH43 for β-xylosidase with a length of 464 amino acids, the reference index for homology, coverage and maxscore is over 60%, 30% and 200, respectively. We selected 11 β-xylosidase gene from the DNA metagenome data. Only GL0112518 is a complete gene, which is predicted to perform optimally at high temperatures and alkaline pH.

### 3.2.8. Some detailed predictions of GL0112518 encoding β-xylosidase (*Xbx14*)

#### *3.2.8.1. Results predict the function of Xbx14 by BLASTP*

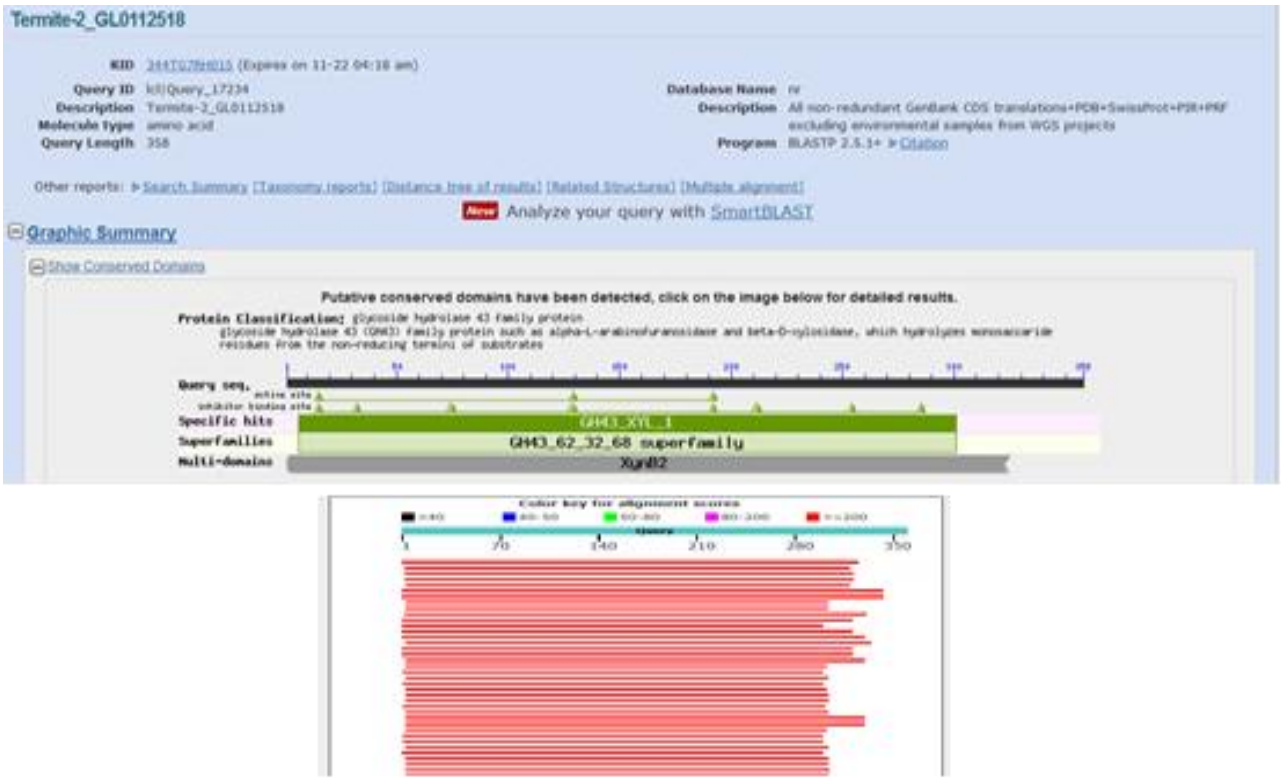


***Figure 3.6. Predict the function of Xbx14 with BLASTP***

The GL0112528 gene has a length of 1077 bp. The BLASTP predicted that it encodes for *Xbx14* of the GH43 family, as originally predicted by the probes (Figure 3.6).

***3.2.8.2. Prediction spatial structure of Xbx14 by PHYRE2***

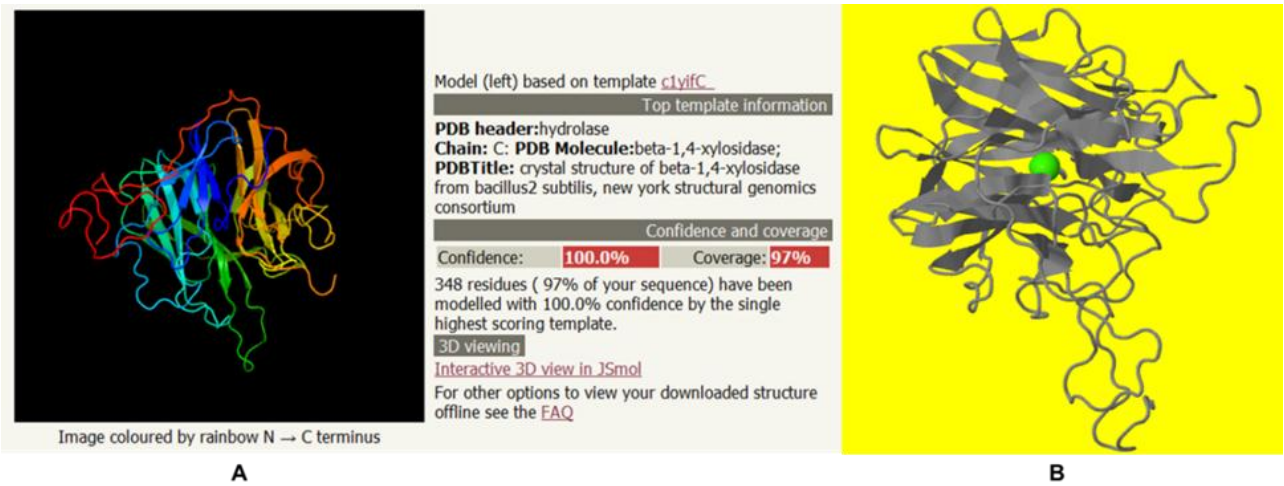


***Figure 3.7. Spatial structure model (A) based on c1ylfC template and activity center (B) of Xbx14***

An analysis second structure of *Xbx14* show that it has 41% of β-folded strands, 1% of α-chains and 62% of unspecified sequences (Figure 3.7A). Spatial modeling indecated that *Xbx14* were spherical folded, with the highest homology and coverage of 97% compared to the beta-xylosidase structure of *Bacillus 2 subtillis* (Fig 3.7B).

***3.2.8.3. Check the source code of GL0112518 with the BLAST - Explorer tool***

Using the BLAST –Explorer tool to check the origin of the GL0112518 gene showed that the gene is within the branch of the *Treponema primitia* (Fig 3.8). This is the largest group of ORFs in the gut of *C. gestroi* (Figure 3.1).

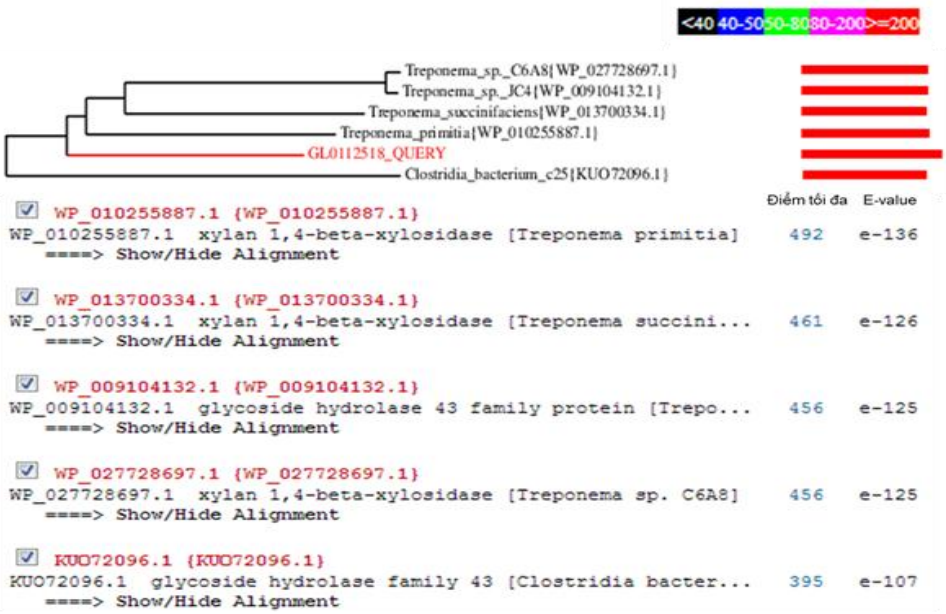


***Figure 3.8. Diagram of origin of GL0112518***

In summary, we predict that the GL0112518 gene which have a length of 1077 nucleotides, coding for β-xylosidase by BLASTP and PHYRE 2. The GL0112518 gene is highly homologous to the gene sequence from the bacterium *Treponema primitia,* a characteristic bacterium found in *C. gestroi* by Explorer-Blast.

## 3.3. EXPRESSION *Xbx14* IN VECTOR PET22b(+)

### 3.3.1. Express *Xbx14* gene.

#### *3.3.1.1. Check Xbx14 gene in the Pet22b (+) vector*

The results of transforming the Pet22Xbx vector into DH10B cells spreaded on LBA plates (Figure 3.9A). Electrophoretic analysis of the five Pet22Xbx plasmid showed that all of them ran higher than the non-gene control vector (Fig 3.9B). This result demonstrates that the Pet22b(+) vector contains DNA fragments which is inserted into the multiple cloning sites, so plasmid will be larger sizes and run slower on the image.

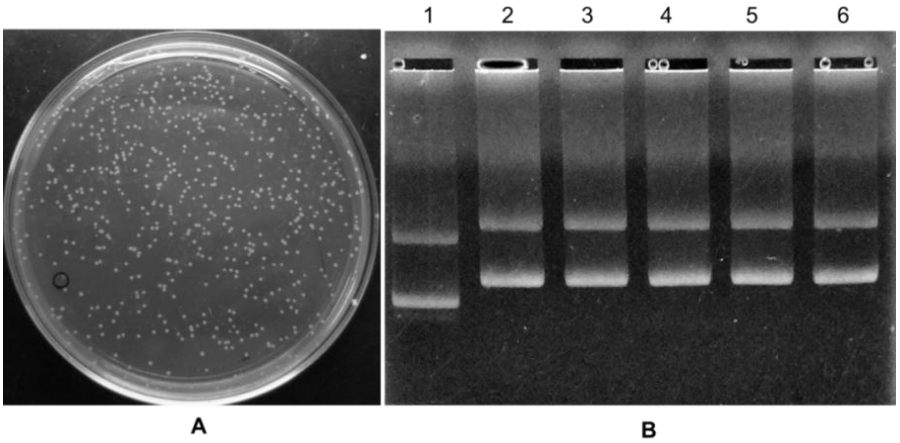


***Figure 3.9. Result of transforming Pet22Xbx vector in E.coli DH10B(A) and electrophysiological analysis of isolation product of Xbx14 gene (B).*** *Line 1: Plasmid Pet22b (+); Line 2-6: Plasmid Pet22Xbx*

The study examined the *Xbx14* gene in the Pet22Xbx vector by PCR, cut-off by restriction enzyme and sequence gene. The cut-off map on the pET22Xbx showed that there is only one point of *Xho*I and *Nco*I and four point of *Hinc*II on the vector (Figure 3.10A). The results of PCR is only one bright, sharp band and the size of the *Xbx14* gene (Figure 3.10B). If pET22*Xbx14* is cut by *Nco*I and *Xho*I, the product are exactly two clear bands of approximately 1.1 kb and 5,5 kb (Fig 3.10C) and three bands of 4.0 kb, 1.8 kb and 0.6 kb when cut by *Hinc*II.This results were completely accurate with theoretical calculations, indicating that the *Xbx14* gene was attached to the pET22b(+) vector. After sequencing *Xbx14* with specific primers were shown that the *Xbx14* sequences gene were attached to the pET22b (+) vector exactly as originally rooted sequences (Appendix 1).

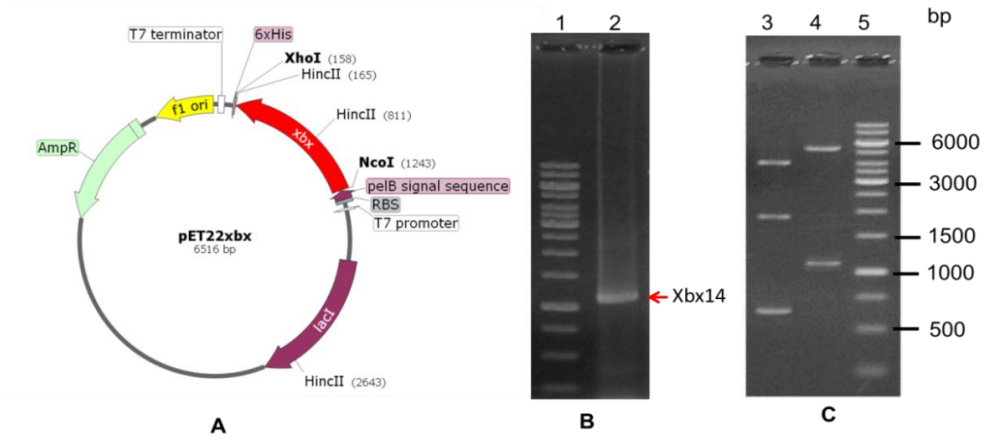


***Figure 3.10. Cutting position diagram of cut-off enzyme HincII, XhoI and NcoI on pET22Xbx vector (A), PCR product electrophoresis (B) and pET22Xbx cutting by NcoI and XhoI, HincII (C).*** *Line 1, 5: Fermentas marker; Line 2: PCR product of Xbx14 gene, Line 3: pET22xbx cut by HincII; line 4: pET22xbx cut by NcoI and XhoI*

***3.3.1.2. Coexpression gene coding Xbx14 and chaperone in E.coli rosetta 1***

Recombinant pET22Xbx vector was transformed into *E. coli Rosetta 1* for expression survey. After optimizing the expression conditions such as temperature, the concentration of the IPTG, the time of induction, the time of sample collection, the results of *Xbx14* expression are quite good, but completely insoluble *Xbx14*. This study continued the co-expression of the *Xbx14* gene in the pET22Xbx vector with the chaperones (GroEL (60 kDa), DnaK (70 kDa) and DnaJ (40 kDa)) in the pG - KJE8 vector. The result was soluble *Xbx14* completely (Fig 3.11).

SDS-PAGE shows that when co-expression with chaperone, soluble *Xbx14* appear with large band, bold and theoretical dimensions. On the other hand, *Xbx14* protein does not appear in the non-inducible IPTG sample. When we compared to expression product with sample without chaperone, the *Xbx14* protein is mostly found in precipitation. This proved that co-expression with chaperone helped *Xbx14* to perform well and insolublely. At the same time the checking crude activity of *Xbx14* showed that the control did not appear yellow color of pNP due to the metabolic catalysis of β-xylosidase such as the recombinant *Xbx14* protein (Figure 3.12).

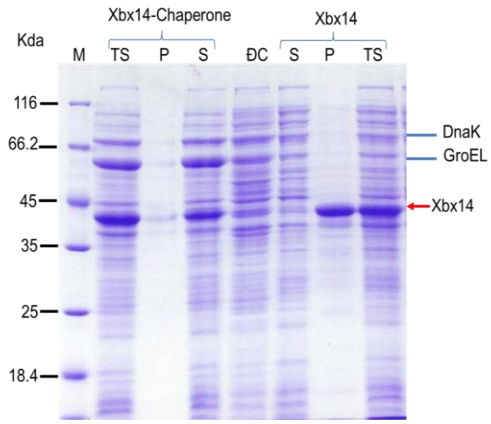


***Figure 3.11. SDS-PAGE of comparation to expresion recombinant Xbx14 with and without chaperone in E.coli Rosetta 1***

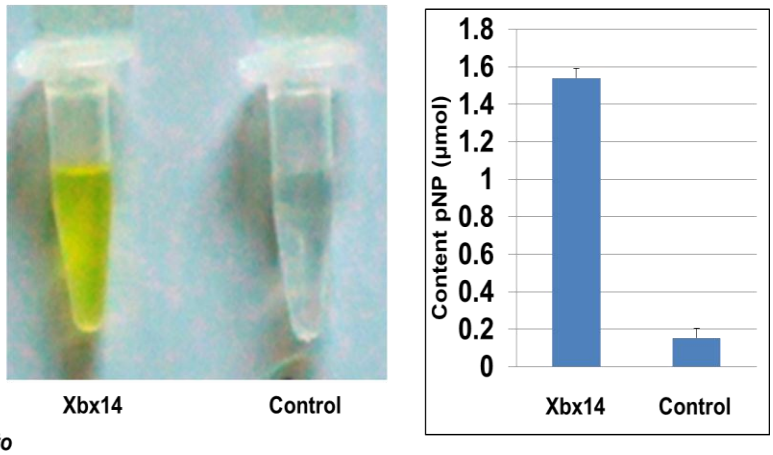


***Figure 3.12. Test activity of rude Xbx14 with specific pNPX***

*Line M:. Protein ladder unstained (Thermo scientific); Line ĐC: Total protein of E.coli Rosetta 1; Line S: Solulbe protein of E.coli Rosetta 1; Line P: insolulbe protein of E.coli Rosetta 1*

***3.3.1.3. Optimal expression conditions of Xbx14 gene***

***a. Influence of temperature on Xbx14 expression in E. coli Rosetta 1***

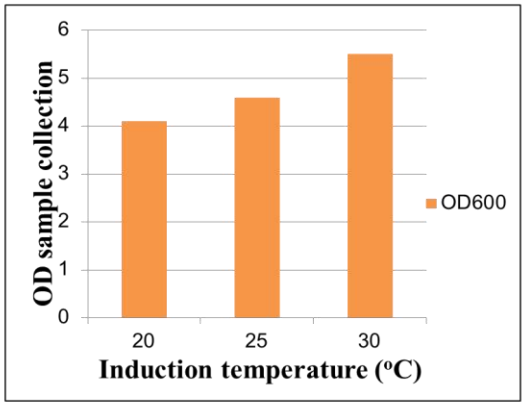


***Figure 3.13. Density of cell inducted IPTG at 20°C, 25°C and 30°C***

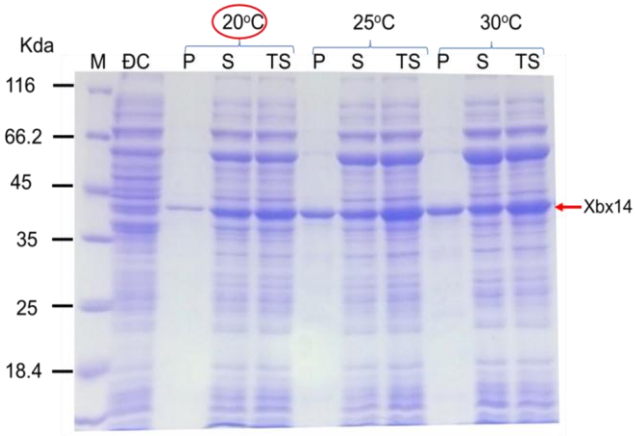


***Figure 3.14. SDS-PAGE recombinant protein coexpressed in E. coli Rosetta 1 after 6 hours of 0.5 mM IPTG induction at 20°C, 25°C, 30°C.***

*Line ĐC: Negative control, Line TS: total protein; Line S: soluble protein; Line P: insoluble protein; Line M: Protein ladder unstained (Thermo scientific).*

we selected three temperatures lower than the growth temperature of *E. coli* at 20°C, 25°C and 30°C to test. On the other hand, according to the supplier's suggestion of pG-KJE8 vector, the appropriate temperature for soluble chaperone expression is less than 25°C. The results showed that the cell density obtained after 5 hours of induction was highest at 30°C and lowest at 20°C (Fig 3.13). Electrophoresis showed that the *Xbx14* protein have both soluble and insoluble forms at 20°C, 25°C, 30°C (Fig 3.14). Although total recombinant *Xbx14* was produced at 30°C was the highest, the amount of soluble *Xbx14* at 20°C was higher. On the other hand, as recommended by the pG-KJE8 vector supplier, the temperature for expression of the chaperone should not exceed 25°C. So we will select a temperature of 20°C to express the *Xbx14* gene in subsequent experiments.

***b. Influence of IPTG concentration on expression of Xbx14***

We selected 9 of IPTG concentrations including: 0 mM; 0.05 mM; 0.1 mM; 0.3 mM; 0.5 mM; 0.7mM; 1mM; 1.2 mM and 1.5 mM to test the effect of IPTG concentration to *Xbx14* gene's expression in *E. coli Rosetta 1* at 20°C. Cells were shaked at 20°C for 5 hours. The collected sample will measure the density of $OD_{600}$. The results showed that the IPTG concentration increase from 0.0 mM to 1.5 mM, the cell biomass reduced (Figure 3.15). The total *Xbx14* were examined by SDS-PAGE (Fig 3.16). At IPTG concentrations of 0.05 mM, the largest recombinant *Xbx14* were collected. Examination of soluble *Xbx14* revealed that at different IPTG concentrations, the recombinant protein content was also proportional to the total *Xbx14* protein content (Fig. 3.16, Fig. 3.17). Thus, to obtain the highest cell biomass as well as the highest soluble *Xbx14* protein, the most appropriate IPTG induction level was 0.05 mM.

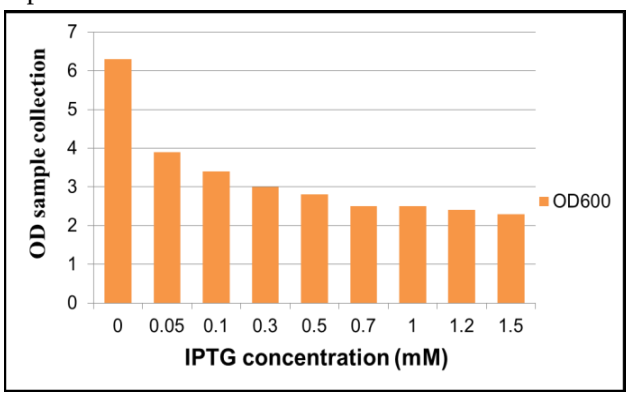


***Figure 3.15. The density of Rosetta 1 E. coli (DE3) cell induced IPTG at concentrations from 0.0 to 1.5 m***

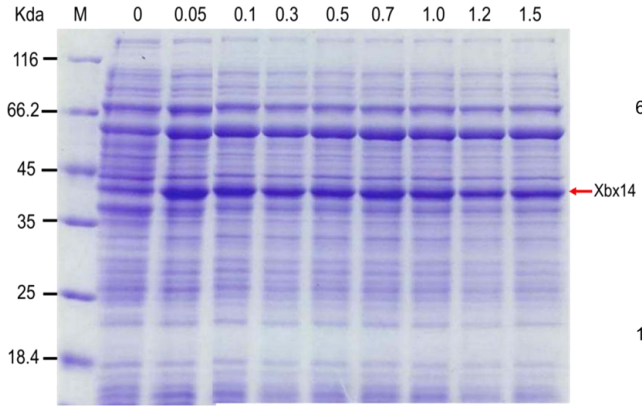


***Figure 3.16. SDS-PAGE total protein expression in E. coli Rosetta 1 inducted with 09 of IPTG concentrations***

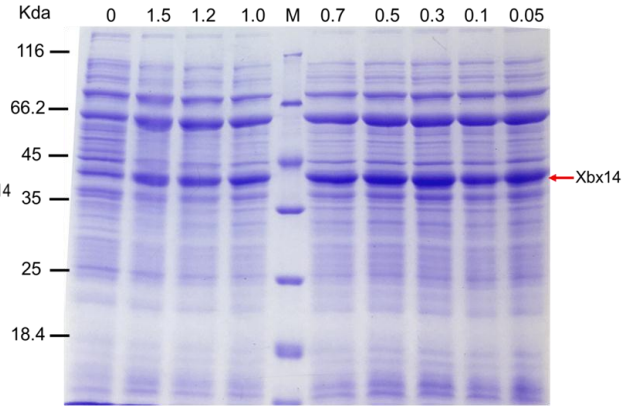


***Figure 3.17. SDS-PAGE soluble protein expression in E. coli Rosetta 1 inducted with 09 of IPTG concentrations***

*Line 0,0 - 1,5: total protein/solulble protein induct from 0,0 to 1,5 mM IPTG. Line M: protein ladder unstained (Thermo scientific)*

***c. Influence of induction time on expression of Xbx14***

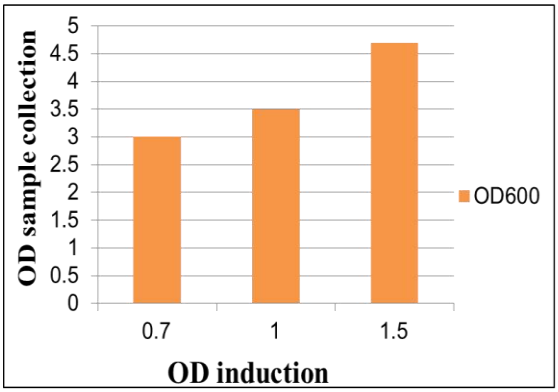


***Figure 3.18. density of E. coli Rosetta 1 (DE3) cell induced at different $OD_{600}$***

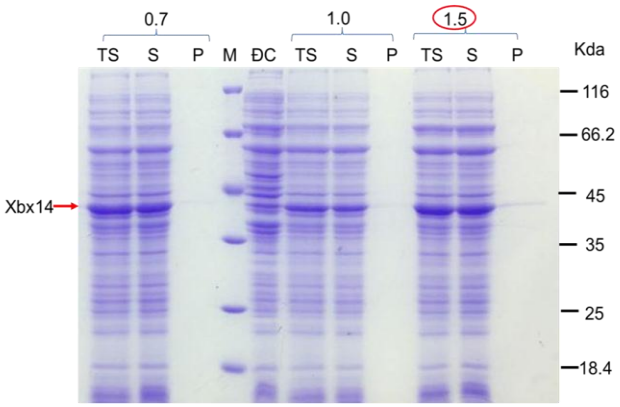


***Figure 3.19. SDS-PAGE soluble protein expression in E. coli Rosetta 1 inducted IPTG at $OD_{600}$ = 0.7; 1.0; 1.5.***

*Line ĐC: Negative control, Line TS: total protein; Line S: soluble protein; Line P: insoluble protein; Line M: Protein ladder unstained (Thermo scientific).*

Cell were cultured at 25°C for induction of chaperone expression and 20°C induction of Xbx14 expression. We chose at three of induction times which is 0.7, 1.0 and 1,5 of the $OD_{600}$. The results indicate when cell is inducted in 1.5 of the $OD_{600}$, the largest biomass cell was obtained (Figure 3.18). The obtained cell sample continued to check the total Xbx14 and soluble Xbx14. Results show that total and soluble Xbx14 protein content were equal at 0.7 and 1.5 of $OD_{600}$. However, cell biomass inducted at $OD_{600}$ = 1.5 is larger (Fig. 3.19). So we will select the touch at 1.5 of the $OD_{600}$ for the next experiment.

***d. Influence of sample collection time on efficiency Xbx14***

Cells after 0.05 mM IPTG induction at 1.5 of $OD_{600}$ will continue to shake at 20°C. Samples will be collected after 5 hours, 8 hours, 15 hours and 20 hours of induction. At the time of sampling after induction, the cell density increased from 5 hours to 15 hours, then decreased at 20 hours (Figure 3.20). The cell is broken by sonic to check the expression and soluble *Xbx14*. Results showed that total and soluble *Xbx14* content were obtained at least after 5 hours induction and higher at 8 hours and 15 hours. Although cell density was reduced due to lack of nutrient media, total and soluble *Xbx14* were stable (Fig. 3.21) after 20 hours of induction. Combined with the final cell density data, we selected time of 15 hours at 20°C to collect sample.

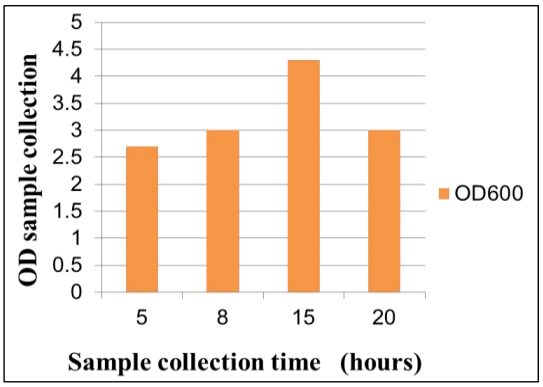


***Figure 3.20. Density of expressing E. coli Rosetta 1 (DE3) cell after 5, 8, 15 and 20 hours of induction***

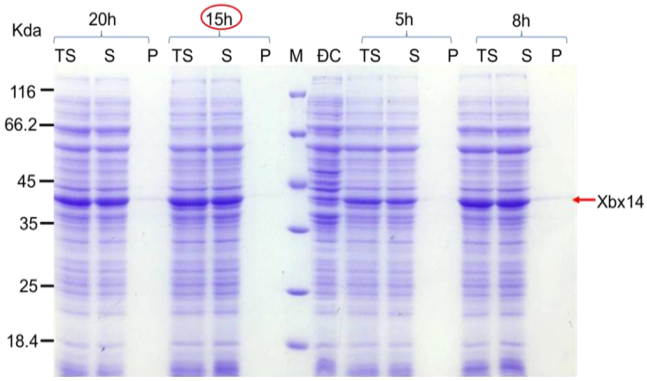


***Figure 3.21. SDS-PAGE protein expression in E. coli Rosetta 1 collected after 5, 8, 15, and 20 hours of IPTG induction.***

*Line ĐC: Negative control, Line TS: total protein Line; S: soluble protein; Line P: insoluble protein; Line M: Protein ladder unstained (Thermo scientific).*

Thus, following the optimal conditions for co-expression of *Xbx14* with chaperones in *Rosetta E. coli*, to obtain large amounts of soluble *Xbx14* need temperature of 20°C, 0.05 mM IPTG, $OD_{600}$ = 1.5 and collect sample after 15 hours induction.

### 3.3.2. Refine *Xbx14*

According to the original design, the GL0112518 gene was attached to the pET22b(+) vector at the multiple cloning sites, behind having a sequence encoding for 6 histidine which is used to refine the affinity chromatographic column. However the *Xbx14* is soluble in water and completely precipitated at very low salt concentrations. This is why we can not use the affinity chromatography column, so we will use $(NH4)_2SO_4$ to purify by precipitation of fractions

#### *3.3.2.1. Purification of Xbx14 by $(NH_4)_2SO_4$*

The result showed that *Xbx14* was precipitated in sections from high salt concentrations to very low salt concentrations. Figure 3.22A shows that *Xbx14* is precipitated at 5% $(NH4)_2SO4$ and less than such as 3% $NH4)_2SO4$ (Figure 3.22B), 1% $NH4)_2SO4$ (Figure 3.22C), 0.2% $NH4)_2SO4$ (Fig. 3.22D) and 0.02% $NH4)_2SO4$ (Fig 3.22E). Almost of larger proteins than *Xbx14* were precipitated, so the amount of contaminated protein is very small. In order to purify the contaminated proteins, the study is based on the isoelectric point of the *Xbx14* protein and dissolves in buffer with different pH. The results indicated that most of the contaminated proteins were dissolved in phosphate buffer pH = 9, while *Xbx14* was still precipitated ( Figure 3.22F).

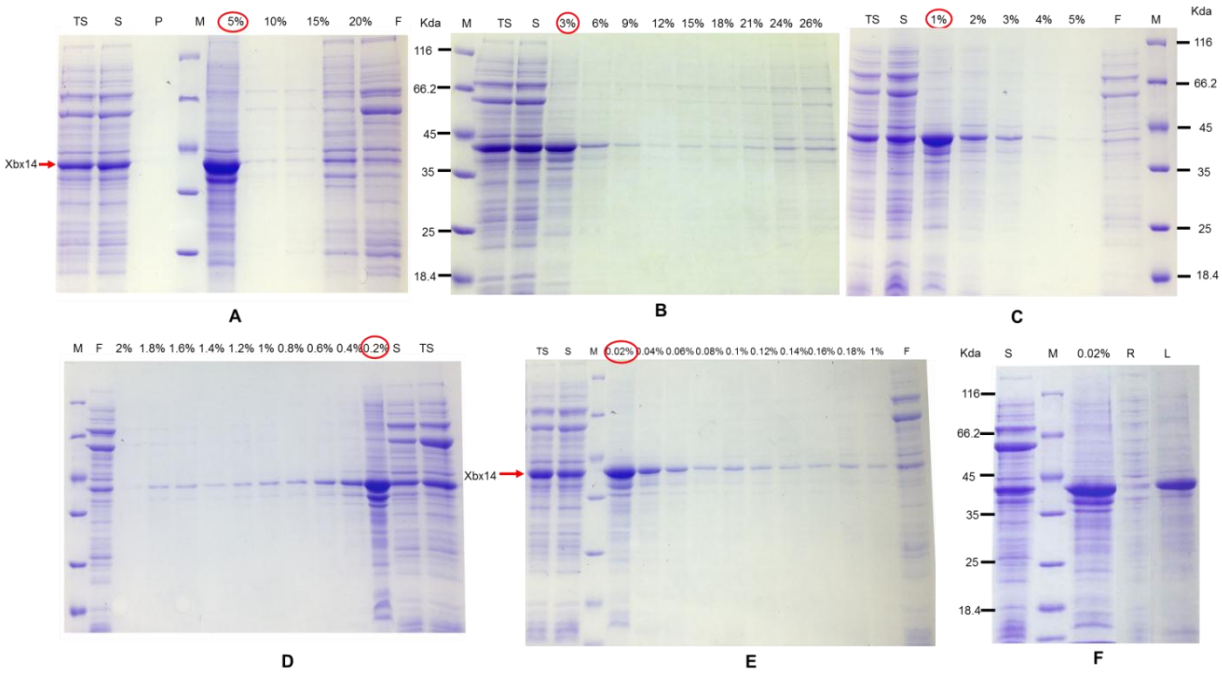


***Figure 3.22. SDS-PAGE of Xbx14 precipitate by $(NH4)_2SO4$: 5% (A), 3% (B), 1% (C) , 0.2% (D), 0.02% (E) and wash with phosphate buffer pH = 9 (F).*** *Line TS : Total protein; Line S: soluble protein; Line 5%, 10% ... corresponding salt concentration $(NH4)_2SO4$ used to precipitate Xbx14; Line F: final fluid after protein precipitation; Line R: fluid after washing the Xbx14 protein precipitate with phosphate buffer (pH = 9); Line L: Xbx14 after washing with phosphate buffer; M: Protein ladder unstained (Thermo scientific).*

#### *3.3.2.2. Check the purity of Xbx14 by Quantity one software*

The relative purity of the *Xbx14* protein was evaluated by Quantity One software. Samples were measured and calculated for each line equivalent to 2 µg and 4 µg protein (Fig 3.23A). A background position (BGR) which has the same color and brightness as the background at the *Xbx14* protein band position are selected. Each band is indicated by a corresponding curve vertex on the graph (Figure 3.23B and 3.22C). The scanned and calculated image of the *Xbx14* protein was compared with the total protein of 19.8%. However, after refining, *Xbx14* accounts for 92.7%.

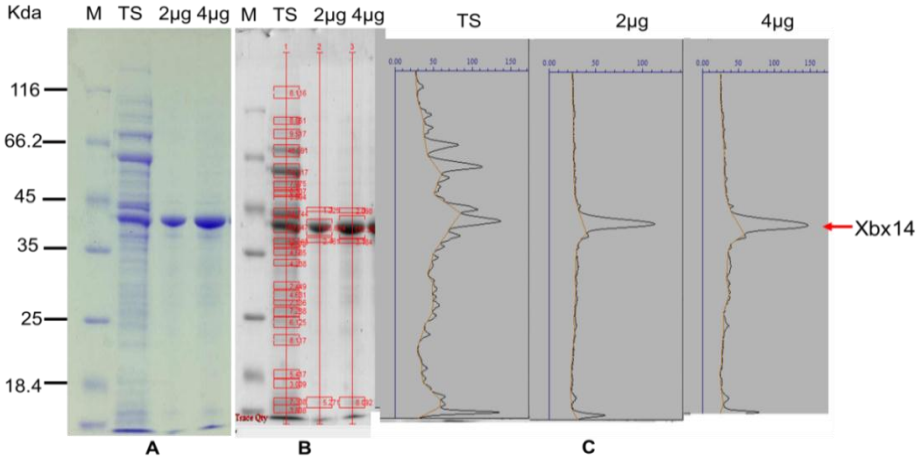


***Figure 3.23. The results for testing cleanness of Xbx14 after purification by $(NH4)_2SO4$ and wash with phosphate buffer (pH = 9).***

### 3.3.3. Study characters of *Xbx14*

#### *3.3.3.1. Determine the activity of Xbx14*

The activity of *Xbx14* was determined by 10 mM pNPX substrate in phosphate buffer (pH = 7) according to the Teng et al's method. The reaction appear yellow of the pNP, and the control did not appear yellow after 1.5 hours of incubation. The activity of the enzyme will be precisely determined by the Elisa method at 405 nm. The results measured and calculated according to the benchmark showing that after 2 hours produced 0.1692 µMol of pNP from 0.253mg of *Xbx14* (Fig 3.24).

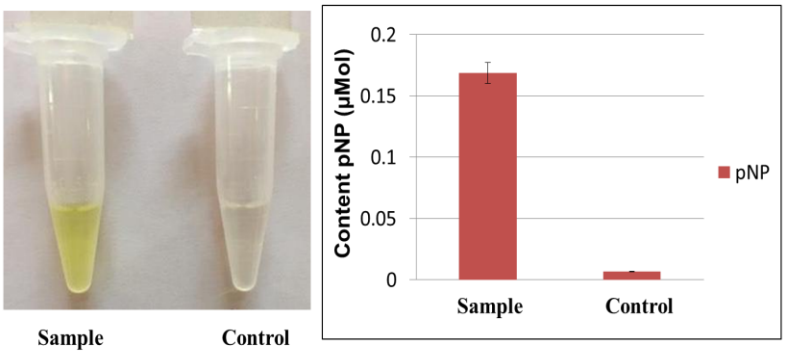


*Figure 3.24. Test activity of Xbx14 with specific pNPX*

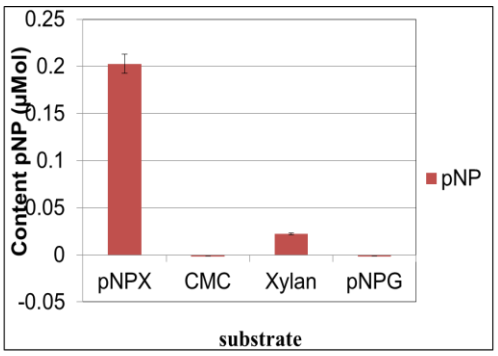


*Fig 3.25. Xbx14 activity on different substrates*

### *3.3.3.2. Determine the substrate specificity of Xbx14*

Some common substrates are CMC, xylan, p-nitrophenyl-β-glucoside (pNPG), 4-nitrophenyl β-D-xylopyranoside (pNPX) are used to test the specificity of *Xbx14*. In which, pNPX is specific substrate of *Xbx14*, CMC is a specific substrate of endoglucanase, xylan is a specific substrate of xylanase and pNPG is a specific substrate of β-glucosidase. The results showed that *Xbx14* perform the highest activity on pNPX substrate, weak activity with xylan and not cleaved CMC and pNPG (Fig 3.25). This proves that *Xbx14* has the specific activity of the β-xylosidase that cleaves double-stranded xylose and cleaves sugar molecules at the ends of the xylan.

### *3.3.3.3. Influence of temperature on activity of Xbx14*

We study optimal activity of *Xbx14* in a temperature range of 20°C, 30°C, 40°C, 50°C, 60°C, 70°C, 80°C and 90°C. Results showed that activity of β-xylosidase increased gradually from 20°C to 60°C and decreased gradually when the temperature increased to 80°C and decreased sharply at 90°C. Thus, the most suitable temperature for the operation of *Xbx14* is 60°C (Fig 3.26), as predicted by the TBI tool. This promises to be a new enzyme with heat resistant.

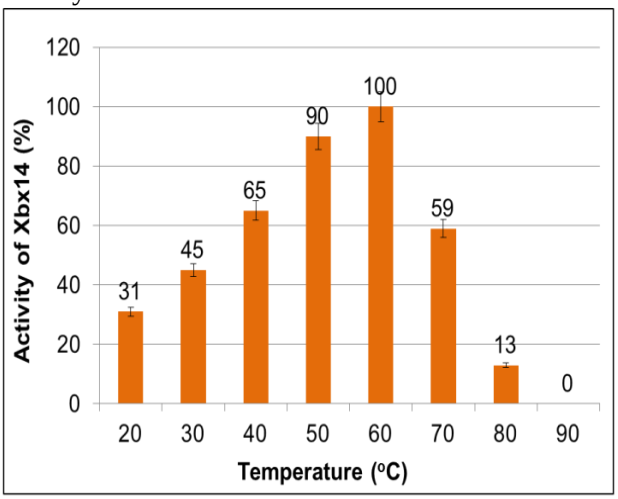


*Figure 3.26 Effect of temperature on activity of Xbx14*

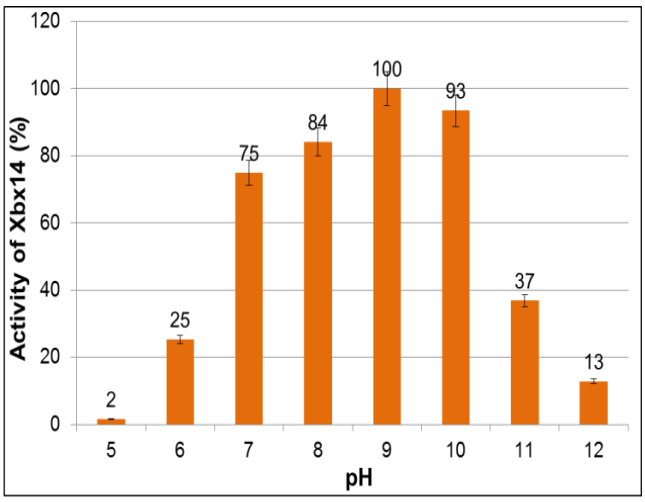


*Figure 3.27. Effect of pH on activity of Xbx14*

### *3.3.3.4. Influence of pH on activity of Xbx14*

*Xbx14* is capable of performing well in the alkaline pH as estimated by Alcapred tool. The activity test showed that the enzyme was inactivated at pH 5, the activity increased gradually from pH6 to pH9 and reached its highest value at pH9. After that the activity decrease lightly at pH10 and decreased deeply at pH12 (Fig 3.27). Thus the optimum pH of *Xbx14* is appropriate in the alkaline environment, consistent with the initial predictions using the Alcapred tool. This enzyme has the potential to be very effective when combined with the treatment of plant biomass by alkaline solution. This is also the enzyme we are looking for in this study.

### *3.3.3.5. Thermal stability of Xbx14*

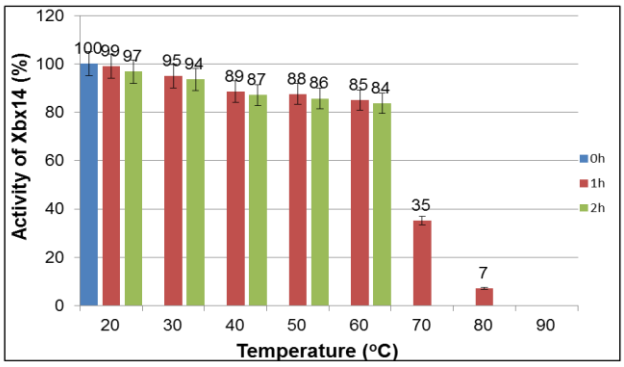


***Figure 3.28. Activity of Xbx14 testing at different temperatures***

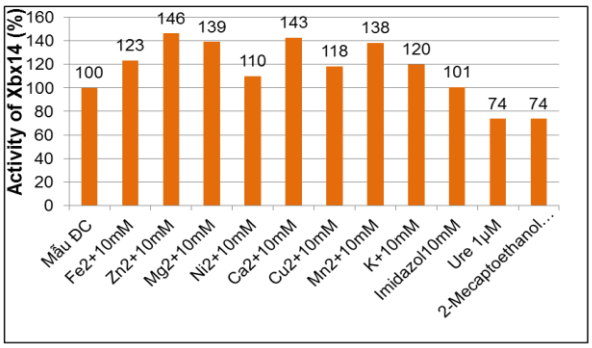


***Figure 3.29. Effect of some metal ions and chemicals on the activity of Xbx14***

*Xbx14* is treated at temperatures at 20°C, 30°C, 40°C, 50°C, 60°C, 70°C, 80°C and 90°C for different periods of 1 hour and 2 hours before incubation with substrate to test thermal stability. Results showed that, *Xbx14* lost almost of activity despite the shortest processing time (1 hour) at temperatures greater than 70°C. The enzyme activity is almost stable after 1 to 2 hours of heat treatment at temperatures between 20°C and 60°C. Thus, the *Xbx14* is quite stable at 20°C to 60°C (Fig 3.28).

***3.3.3.6. . Influence of some metal ions and chemicals on activity of Xbx14***

To evaluate the effect of metal ions on the activity of *Xbx14*, we proceeded according to the method of Lee et al., or Shi et al. The results show that all of metal ions increase the activity of the *Xbx14*. Imidazol did not affect the activity of *Xbx14*, but at 1 µM 2-mercaptoethanol and 1 µM urea reduced the activity of *Xbx14* ( Fig 3.29).

***3.3.3.7. Dynamic characters of Xbx14***

The Km constant of *Xbx14* is determined by the pNPX substrate concentration range from 2.0 to 20 mg/ml in phosphate buffer pH9. Kinetic parameters are determined from Lineweaver-Burke graph. The results showed that with the *Xbx14* concentration of 0.253 mg/ml, the equation of reaction rate dependence on substrate concentration was based on the function y = ax + b with high reliability $R^2$ = 0.9937 (Fig 3.30). Based on the equation, Km and Vmax of *Xbx14* were respectively 13.7 mg/ml and 6.5 µmol/min. Therefore, the specific activity of *Xbx14* is 24.54 U/mg.

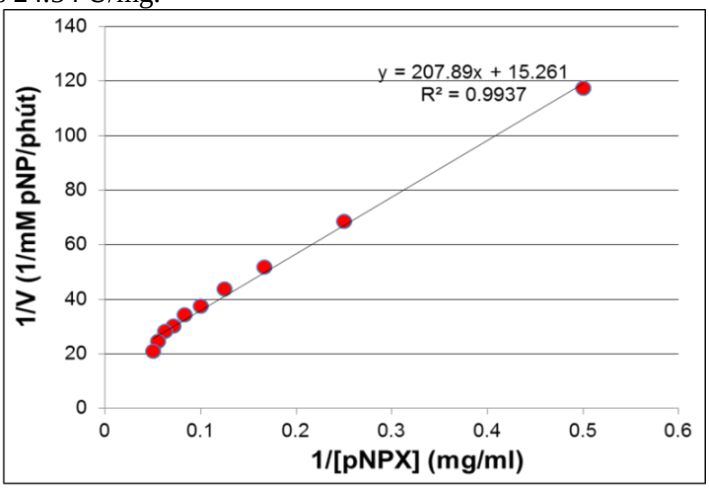


***Figure 3.30. Depend on reaction Xbx14 on substrate concentration pNPX following Linewever-Burk***

In summary, the activity of *Xbx14* shows that this is a novel enzyme, capable of optimal performance at high temperature (60°C) and alkali (pH9). It has with good specific activity of 24.54 U/mg. This enzyme promises to bring many economic benefits when combined with treatment of biomass by alkalis and heat in practice.

## CONCLUSIONS AND RECOMMENDATIONS

### Conclusion

1. This study exploited 20 bacteria species who have been identified to characterize the poor nitrogen and rich lignocellulose diet of *C. gestroi* from available DNA metagenome data. Of the three species involved in nitrogen metabolism and five species involved in hydrolysis of lignocellulose in gut *C. gestroi* without in

other termites. Thus, We classified 578 ORF coding lignocellulase into 52 different GH families according to CAZY classification

2. We developed successfully a method to exploit gene from DNA metagenome data by probes basing on the amino acid sequences which studied detail characters and using support of bioinformatic tools, in the direction of application with important properties of heat resistance and alkalinity resistance. The GL0112518 gene coding for β-xylosidase was found out from the DNA metagenome data of microorganism in gut of the *C. gestroi* by GH43 probe;

3. The GL0112518 gene is expressed succesfully in *E.coli Rosetta* 1. The Xbx14 protein has the β-xylosidase activity, with kinetic parameters of Lineweaver-Burk equation of Km = 3.18mM and Vmax = 6.5 μmol/min and specific activity of 24.54 U/mg. It is a new and highly active enzyme with optimum activity in high temperature and alkaline pH.

**Recommendation**

Use of *Xbx14* in combination with other cellulases and hemicellulases to hydrolyze lignocellulose.

## LIST OF THE AUTHOR'S WORKS


**1. *Minh Giang Nguyen,*** *Thi Huyen Do, Nam Hai Truong* (2015). Using bioinformatic technology in studying metagenomics – A new research approach and application in biology. Tap chi khoa hoc, Padagogy of Ho Chi Minh University, (2), 2015.

**2. *Minh Giang Nguyen,*** *Thi Huyen Do, Nam Hai Truong* (2016). Using bioinformatic tools in exploited gene encoding enzyme to decompose lignocellulose from metagenome of free-living bacteria in gut of the lower termite *Coptotermes gestroi*. Tap chi Cong nghe Sinh hoc 14(1 ): 39-47.

**3. *Minh Giang Nguyen,*** *Thi Huyen Do, Nam Hai Truong* (2016). In silicon mining for alkaline enzymes from metagenomic DNA data of gut microbes of the lower termite *Coptotermes gestroi* in Viet Nam. Tap chi Sinh hoc 38(3): 374-383. DOI: 10.15625/0866-7160/v38n3.7811.

**4**. ***Minh Giang Nguyen,*** *Minh Thu Nguyen, Thi Duyen Nguyen, Thi Huyen Do, Thi Thuong Ho, Bich Ngoc Pham, Hoang Ha Chu, Nam Hai Truong* (2017). Construction of vector pCB301-*Xbxs14*-ELP and expression of Xbxs14 gene coding Xylan 1,4- beta-xylosidase from free living bateria in gut of termite *Coptotermes gestroi* in *Nicotiana benthamiana* by agroinfiltration. Tap chi Sinh hoc 2017, 39(2): 226-235. DOI: 10.15625/0866-7160/v39n2.9843

**5**. ***Minh Giang Nguyen,*** *Thi Huyen Do, Thu Nguyet Phung, Nam Hai Truong* (2017). Probe design for exploiting genes coding xylan 1,4 – beta- xylosidase from DNA metagenome of free-living bacteria in gut of the lower termite *Coptotermes gestroi*. Accepted public

**6. *Nguyen Minh Giang***, *Pham Hai Nhu, Nguyen Minh Thu, Đo Thi Huyen, Đinh Nho Thai, Truong Nam Hai* (2017). Expression of Xbxs14 coding xylan 1,4-β-xylosidase derived frm bacteria in *Coptotermes gestroi* termite gut on *Escherichia coli rosetta 1* (DE3). Tap Chi Cong Nghe Sinh Hoc 15(3):1-7.