MINISTRY OF EDUCATION
AND TRAINING

VIETNAM ACADEMY OF SCIENCE
AND TECHNOLOGY

GRADUATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

Dao Trong Khoa

CONSTRUCTION OF THE METAGENOMIC DNA DATA OF BACTERIA IN GOAT RUMEN AND STUDY ON THE PROPERTIES OF ENDO-XYLANASE

SUMMARY OF DISSERTATION ON BIOLOGY

Major: Biochemistry

Code: 9 42 01 16

Ha Noi – 2024

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

Supervisors:

1. Supervisors 1: ***Prof. Dr. Truong Nam Hai***
Institute of Biotechnology

Người hướng dẫn 2: ***Assoc. Prof. Dr. Do Thi Huyen***
Institute of Biotechnology

Referee 1: Assoc. Prof. Dr. Pham The Hai
Hanoi University of Science – Vietnam National University

Referee 2: Prof. Dr. Le Mai Huong
Institute of Natural Products Chemistry, Vietnam Academy of Science and Technology

Referee 3: Assoc. Prof. Dr. Truong Quoc Phong
School of Chemistry and Life sciences, Hanoi University of Science and Technology

The dissertation is examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at……………………….. (time, date……)

The dissertation can be found at:

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

INTRODUCTION

1.1. The urgency of the thesis

Lignocellulose, one of the abundant renewable energy sources on Earth, is mostly burned to release waste and smokes seriously affecting the quality of the living environment as well as people's health. Therefore, taking advantage of this surplus source of raw materials to convert them into biofuels not only reduces environmental pollution but also contributes to solving the national energy demand. However, lignocellulose is a solid biomass that is difficult to degrade and convert. The idea of decomposing lignocellulose by biological methods in an environmental-friendly way is highly appreciated and gradually got into application. The search for a source of lignocellulases with strong activity has been one of the key research directions of many scientists around the world. Bacteria residing in lignocellulose-rich ecosystems have been identified as potential sources for gene exploitation in general and lignocellulose-decomposing genes in particular because of their diversity and abundance. However, in reality, 99% of microorganisms cannot be isolated and cultured. To overcome this limitation, Metagenomics techniques allow direct and comprehensive research and evaluation of all microorganisms in the sample without culturing. The mini-ecosystem of the rumen of goats raised in Vietnam is one of the very potential systems that has not been studied in deep. Therefore, this study was conducted to sequencing the DNA of multiple bacterial genomes in the rumen of goats (conventional sequencing to create a small, common data and deep sequencing to create a big, rounder data, thus evaluating the ability to exploit the genes of both datas) and find a new approach to effectively exploit lignocellulose-degrading enzymes, including pre-treatment enzymes, cellulose, hemicellulose and lignindegrading enzymes. Therefore, the thesis have been carried out: "***Construction of the metagenomic DNA data of bacteria in goat rumen and study on the properties of endo-xylanase".***

1.2. Reseach goals:

- Construct the metagenomic DNA data of bacteria in goat rumen;
- Express and characterize an endo-xylanase in the collection of lignocellulose-degrading genes that have been mined from the metagenomic DNA data of bacteria in the goat rumen.

1.3. Reseach subject:

To achieve the reseach goals, the following main research subjects have been carried out:

1. Sequencing metagenomic DNA of bacteria in the goat rumen at normal capacity (8-10 Gbs) and large capacity (deep sequencing, 45-50 Gbs); constructing the metagenomic DNA data and assessing the diversity of bacteria in goat rumen;
2. Mining genes and establishing a HMM-based tool to functionally annotate genes encoding enzymes/proteins involved in lignocellulose metabolism.
3. Selecting an endo-xylanase gene in the metagenomic DNA data of bacteria in the goat rumen, expressing and identifying the characteristics of the recombinant protein.

CHAPTER 1. OVERVIEW

1.1. Lignocellulose

Lignocellulose is an important component of plant cell wall, which accounts for the largest proportion of biomass. Lignocellulose is made up of three main components, all of which are large molecular polymers: cellulose, hemicellulose, and lignin. Lignocellulosic biomass is one of three main biomass sources that can be used to produce biofuels, a new source of energy, overcoming the disadvantages of fossil energy sources. The composition of lignocellulose when decomposed, in addition to providing energy, also has applications in many other socio-economic sectors such as the food industry, medicine, immunology, etc....

1.2. Xylanase

Xylanase is one of the most important xylanolytic enzymes, with the role of cleaving the xylan backbone, facilitating the activity of other enzymes. The most important GH families with xylanase activity are GH 5, 7, 8, 10, 11 and 43, according to the CAZy database. Xylanases are widespread in nature, originating from many classes of organisms, of which xylanases from bacteria and fungi have been widely studied and applied in many industries.

1.3. Metagenomics techniques for effective mining of potential genes

Metagenomics is a technique for studying multi-genome DNA directly without culturing, in which the latest direction is by whole-genome sequencing thanks to advances in sequencing technology. Sequence information is analyzed and processed by many softwares to identify classification and function. Many new methods have been developed to support the analysis and annotation of gene function effectively, in which the method using HMM model is a method with the highest sensitivity and accuracy in representing homologous sequences in sequence families.

CHAPTER 2. MATERIALS AND METHODS

2.1. Materials

2.1.1. Objective materials

✓ ***Research objects***: Rumen samples of 3 Co goats and 2 Bach Thao goats were collected in Ninh Binh province (GPS coordinates 20.269002 105.893267), 2 Co goats and 3 Bach Thao goats were collected in Thanh Hoa province (GPS coordinates 19.897450 105.795899). The goats selected were goats that ate grass, leaves and branches on the mountains during the day, and were fed various agricultural by-products at night, without feeding bran.

✓ ***Bacterial strains:*** *E. coli* strain DH10B from Invitrogen (USA) was used as the recipient in the gene cloning experiment; *E. coli* strains BL21(DE3), Rosetta1, JM109, SoluBL21 (BL21 Soluble), Origami were used as the recipient for gene expression.

✓ ***Plasmid:*** pET22b plasmid was used as expression vector (Thermo Scientific, USA).

2.1.2. Chemicals

Renown chemicals were purchased from famous companies Merck, Sigma... Quality kits were purchased from Qiagen, Fermentas, Amersharm....

2.1.3. Instruments

Instruments originated from famous companies such as Shimadzu (Japan), BioTek (USA), Bio-rad (USA). Sorvall (USA), Amersham

Pharmacia (USA), Applied Biosystems (USA), GE-Healthcare (Sweden), Implen (Germany), Precisa (Switzerland).

2.2. Research methods

2.2.1. Molecular biology methods

- Extraction, purification of metagenomic DNA;
- Construction of the expression vector containing *exl* gene;
- Transfromation of plasmid into *E. coli*;
- Extraction of plasmid DNA from *E. coli;*
- Agarose gel electrophoresis;
- DNA purification from agarose gel.

2.2.2. Biochemistry/protein methods

- Recombinant protein expression in *E. coli;*
- SDS-PAGE;
- Protein purification by His-tag affinity chromatography;
- Determination of the purity of recombinant protein by Quantity One software;
- Protein quantitative by Bradford;
- Xylanase activity determination;
- Identification of the effects of temperature, pH, metal ion and some chemicals on the enzyme activity;
- Identification the thermal stability of the enzyme;
- Identification the kinetics parameters of the enzyme.

2.2.3. Bioinfomatic methods

- Construction of DNA metagenome into contigs, gene annotation;
- Pfam analysis of sequences;
- Inspection of the conservative domains and prediction of the spartial structure of sequences;
- Prediction of the alkaline/acidic enzymes;
- Prediction of the thermal stability of enzymes based on sequences;
- ORF taxonomy classification;
- Codon optimization and gene synthesis;

- Data processing.

CHAPTER 3. RESULTS AND DISCUSSION

3.1. Deep sequencing, metagenomic DNA data construction and assessment of bacterial diversity in the goat rumens

3.1.1. Extraction of bacterial metagenomic DNA

Before extraction of metagenomic DNA, bacteria from 10 rumens were fractionated. The metagenomic DNAs with large-sized DNA were successfully extracted and purified to get high quality meeting for high throughput sequencing. The results of testing DNA quality and DNA concentration using the nanodrop machine showed that the DNA concentration reached from 53.5 to 137 ng/µL and the A260/280 index reached from 1.921 to 2.028. All 10 metagenome DNA samples from goats rumen bacteria were used as templates for amplifying the 16S rRNA gene. The obtained results ensured that the DNA samples did not contain polymerase inhibitors, so the DNA samples were ready for sequencing.

3.1.2. Sequencing, quality assessment of the metagenomic DNA data and gene functional annotation

The results showed that both data sets had good quality expressed by Q30 above 90%. The total capacity of the metagenome sequencing data of bacteria in the goat rumen was 392.63 million reads. After filtering, 324 million clean data reads were obtained, equivalent to 48.66 Gbs. The reads with Q30 quality accounted for 94.59% and the clean read rate reached 82.61%. After assembling into contigs, the number of contigs of the two data sets was quite large. The metagenome DNA data set of bacteria in the goat rumen was assembled into 3,411,867 contigs with a total length of 3,164 Mbs. Of which, 50% of the sequences were larger than 1,162 bps, the average length of the contigs was 927 bps and the largest contig was 295,214 bps. The contigs cover approximately 64.22% of the reads.

3.1.3. Assessment of diversity in DNA metagenome samples

3.1.3.1. Assessment of diversity based on the 8,6 Gbs data

From the 8.6 Gb sequencing data, 164,644 genes were identified, of which 99.8% were of bacterial origin. Of these, 39,579 ORFs were identified and classified, while 99.8% ORFs belonged to bacteria. The most abundant bacterial phylum was Bacteroidetes (63.5%), followed by Firmicutes (22.6%), Proteobacteria (7.5%), and Synergistetes (3.1%). At the genus level, *Prevotella* (35.3%) and *Bacteroides* (16%) belonging to the order Bacteroidales were the most abundant.

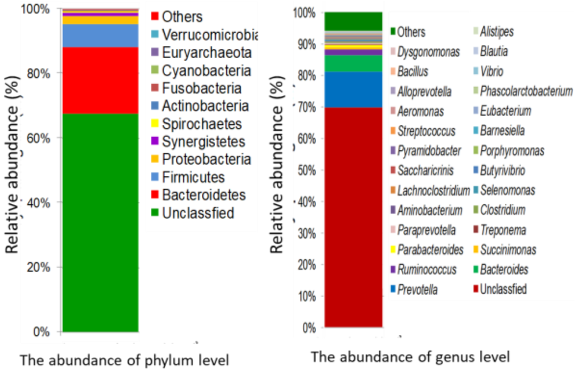

Fig 3. 3. Distribution diagram of taxonomic diversity at phylum and genus levels of bacteria in goat rumen mined from 8.6 Gb data

3.1.3.2. Assessment of diversity based on the deep sequencing data

The results of deep DNA sequencing of the goat rumen bacterial metagenome yielded 48.66 Gb of data. Compared to the 8.6 Gb sequencing data, the classification results were quite similar when the proportion of bacterial genes possessed 99.8%. The Bacteroidetes phylum accounted for the largest proportion with 45.29% of the total number of genes, followed by the Firmicutes phylum with 30.38%. At the genus level, 49.93% of the genes remained unclassified. The most abundant genus was *Prevotella*, accounting for 25.79% of the total number of genes.

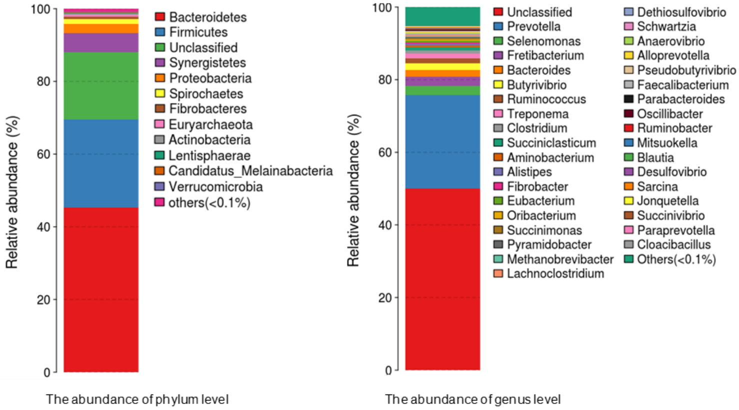

Fig 3. 4. Distribution diagram of taxonomic diversity at phylum and genus levels of bacteria in goat rumen mined from deep sequencing data

3.2. Gene mining and HMM tool establishment for gene annotation; mining genes encoding proteins/enzymes involved in lignocellulose hydrolysis

3.2.1. Mining genes encoding proteins/enzymes involved in lignocellulose hydrolysis based on KEGG

3.2.1.1. Mining genes from the 8.6 Gbs sequencing data

From the DNA sequencing data, 821 ORFs containing carbohydrate esterase (CE) and polysaccharide lyase (PL) domains involved in the pretreatment process in lignocellulose metabolism specifically lignin, 816 ORFs encoding 11 glycoside hydrolase (GH) families with cellulase activity, 2252 ORFs carrying 22 GH families with hemicellulase activity were mined.

3.2.1.2. Mining genes from the deep sequencing 48,6 Gbs data

From the results of deep DNA sequencing of goat rumen bacterial metagenome, 48.66 Gb of data were obtained, 5,367,270 genes with a total length of 2,828,583,591 bp were identified. Among the above genes, 4,385,296 genes had their functions estimated based on comparing the corresponding protein sequences with the Nr, Swissprot, KEGG, eggNOG databases.

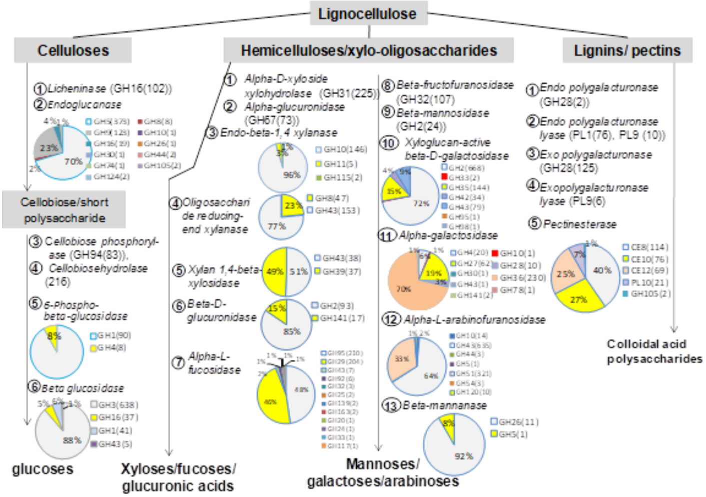

Fig 3.5. Overview of GH/CE/PL families involved in bacterial lignocellulose degradation in goat rumen

Specifically, with the KEGG database, 2,809,791 genes had their functions estimated, of which 317,154 genes (11.3%) were identified as related to carbohydrate metabolism.

3.2.2. Analysis of bacterial diversity carrying lignocellulose hydrolysis

genes

3.2.2.1. Diversity of bacteria carrying lignocellulose-degrading genes mined from the 8.6 Gb data

Of the 816 ORFs encoding cellulase genes, 2252 ORFs encoding hemicellulase genes, and 821 ORFs encoding preprocessing genes that were identified, only 221 cellulase genes, 544 hemicellulase genes, and 226 preprocessing genes could be classified, accounting for about 24-27%, while the remaining unclassified fraction was very large. Overall, the majority of the above genes belonged to the Bacteroidetes phylum, specifically 854 ORFs out of a total of 991 ORFs, accounting for 86.2%. The second largest phylum was the Firmicutes phylum with 94 ORFs (9.5%).

3.2.2.2. Diversity of bacteria carrying lignocellulose-degrading genes mined from the deep sequencing databse

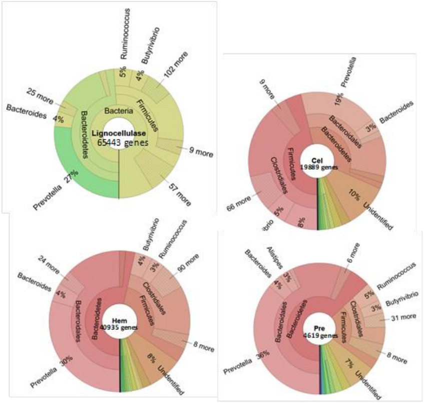

Fig 3.6. Taxonomic diversity of bacteria carrying lignocellulase genes in the rumen of Vietnamese goats annotated by KEGG and classified by MEGAN

All 65,554 genes encoding 30 enzymes/proteins involved in lignocellulose degradation in the goat rumen were subjected into MEGAN software. The results showed that 65,443 genes were classified into taxa (99.85%). Within the genus taxa, the largest genus was *Prevotella*, which contributed 27% of the genes involved in lignocellulose degradation, followed by *Ruminococcus* (5%) and *Bacteroides* (4%). Notably, *Prevotella* contributed significantly to hemicellulose degradation and lignocellulose pretreatment, with this genus contributing 30% of the hemicellulose metabolism genes and 36% of the lignocellulose pretreatment genes.

3.2.2.3. The role of the genus Prevotella in lignocellulose digestion

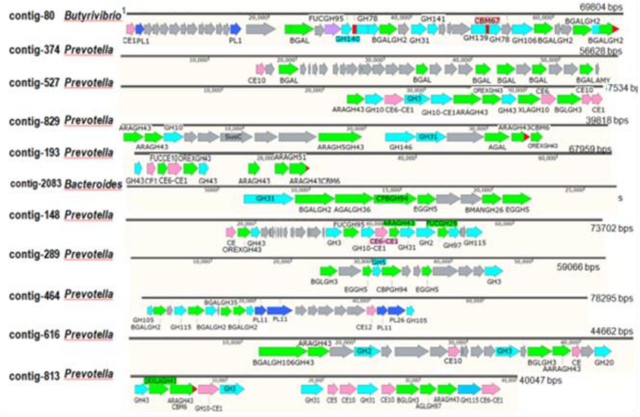

Fig 3.8. Cellulose/hemicellulose-degrading gene loci in potential contigs constructed from goat rumen bacterial metagenome deep sequencing data.

8,900 complete lignocellulase genes were located in 8,364 contigs, of which 7,848 contigs carried only one gene per contig. Of the 22 contigs carrying at least four genes per contig, 18 belonged to the genus *Prevotella*, 2 to the genus *Bacteroides*, 1 to the genus *Clostridium*, and 1 to the genus *Butyrivibrio*. Most of the gene clusters were involved in hemicellulose degradation and were specific for certain substrates. In addition, all genes within a cluster are arranged in the same orientation. In addition to the main enzymes with hemicellulase activity, many genes encoding enzymes belonging to different GHs, which may support the main function of the locus, and some genes of unknown function have also been identified.

3.2.3. Development of a new tool for efficient mining of proteins/enzymes involved in lignocellullose degradation

Based on the HMM model built to mine for 29 different enzymes/accessory domains involved in lignocellulose metabolism, the tool supported mining with the gene groups that were effectively mined in the dataset being galactanase, glucuronyl esterase, hydrogen peroxide oxidoreductase (HPOXRE catalase), xyloglucanase, laccase, CBM (1-84), cellobiohydrolase, beta-glucuronidase, beta-xylosidase, beta-mannosidase GH2, lichenase, alpha-glucuronidase (GH76N) and xylanase GH44.

3.3. Selection, expression and characterization of endo-xylanase

3.3.1. Selection of endo-xylanase gene for expression

3.3.1.1. The diversity of bacteria carrying the enzyme endo-xylanase

From the 48.6 Gb deep sequencing results, based on the gene functional annotation results with the databases, 3400 genes were identified to encode for endo-1,4-beta-xylanases. Of these, 3213 genes were classified to taxonomic units belonging to 3 kingdoms, 19 phyla, 33 classes, 48 orders, 67 families, 120 genera and 9 species, with only 187 genes having unidentified taxonomic units. At the genus level, 30% of the endo-xylanase genes originated from *Prevotella*, followed by *Ruminococcus* (19%) and *Butyrivibrio* (12%).

At the genus level, 30% of the endo-xylanase genes originated from *Prevotella*, followed by *Ruminococcus* (19%) and *Butyrivibrio* (12%).

3.3.1.2. The diversity of endo-xylanase structure

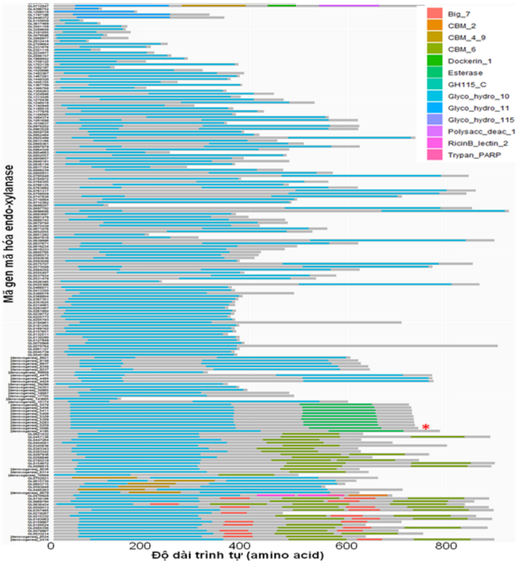

Figure 3.10. Summary of domain structures of endo-xylanase enzymes carrying the GH active site

From 8.6 Gbs conventional metagenomic sequencing data and 48.6 Gbs deep sequencing of bacteria in goat rumen, 739 complete genes encoding endo-xylanases annotated by KEGG were isolated and analyzed for functional domain structure, of which 185 sequences contained glycosyl hydrolase active domains. There were 180 amino acid sequences deduced from 180 genes homologous to endo-xylanases in GenBank, of which a total of 108 sequences originated from *Prevotella*. Of these, 10 genes had sequences for the CE1 sub-active domain, 20 genes had additional CBM6 domain sequences. The endo-xylanases encoded by the above 10 genes had T_m higher than 65°C and were acidic enzymes.

3.3.1.3. Selection of xylanase sequences for expression

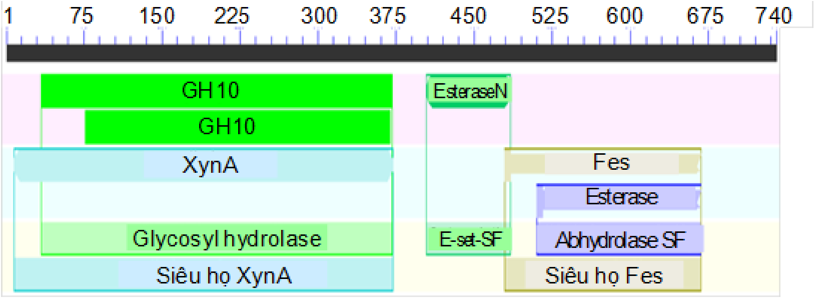

Figure 3. 13. Predicted conserved regions on the denovogenes_5086 sequence

GH10: Glycosyl hydrolase family 10; XynA: Endo-1,4-beta-xylanase; E-set-SF: Superfamily including the N-terminal domain of esterases, EsteraseN: N-terminal domain of esterases; Fes: Enterochelin esterase and related enzymes; SF: superfamily

The protein sequence from the denovogenes_5086 gene had the highest similarity of 96% to endo-1,4-beta-xylanase of *Prevotella* sp. (accession number MBR7030185.1) with 99% coverage. When searching for conserved regions on the gene sequence, the denovogenes_5086 sequence contained two conserved regions: Glyco_hydro_10 (residues 34-371) and the E-set_Esterase region (residues 404-485), in addition to a Fes region identified as enterochelin esterase and related enzymes (residues 479-668). The gene [denovogenes]_5086 had the lowest acid score of 0.363 out of a total score of 1.

3.3.2. Expression of endo-xylanase

3.3.2.1. Codon optimization of exl gene

The gene [desnovogenes]_5086 (*exl)* coding for endo-xylanase was selected for expression in *E. coli*. The *exl* gene sequence was codon optimized for expression in *E. coli* before being artificially synthesized and inserted into the pET22b(+) vector using a pair of restriction enzymes *Nco*I/*Xho*I.

3.3.2.2. Construction of expression vector carrying exl gene

The vector test product has two DNA bands with the correct size of the endo-xylanase gene and the pET22b(+) vector in a straight form. The *exl* sequence was 100% similarity to the optimized sequence. Thus, the endo-xylanase gene has been successfully synthesized and ligated into the pET22b(+) expression vector.

3.3.2.3. Expression of endo-xylanase gene [denovogenes]_5086

(1) Selection of endo-xylanase expression strains

Endo-xylanase protein was successfully expressed in all 5 strains BL21 Soluble, Rosetta1, Rosetta2, JM109, Origami. The amount of soluble protein and activity obtained from a unit of culture of strain Rosetta1 was the highest, so we chose strain Rosetta1 as the strain to express the endo-xylanase gene.

(2) Selection of endo-xylanase expression medium

EXL protein was expressed in LB, modified TB, SOB, PE, YT, SB media. PE medium was chosen for endo-xylanase expression becaused of giving the high yield of the enzyme.

(3) Selection of temperature expression

Endo-xylanase expression was studied at temperatures of 20ºC, 25ºC, 30ºC and 37ºC. The protein was well expressed in soluble form and possessed the highest specific activity in the cell cultured at 20ºC compared to other temperatures where it was expressed in insoluble form or expressed poorly. Therefore, we chose the endo-xylanase expression temperature at 20ºC.

(4) Seclection of IPTG concentration

Different concentrations of IPTG ranging from 0.05 to 1.5 mM were tested for the expression of recombinant EXL. The results showed that the appropriate IPTG concentration was selected as 0.1 mM for further studies.

(5) Selection of sampling time for endo-xylanase expression

The Rosetta1 strain carrying the endo-xylanase gene was studied by comparing the sampling times after induction at 1 hour, 2 hours, 3 hours, 4 hours, 5 hours, 6 hours and 16 hours (overnight). The time chosen to collect samples of endo-xylanase expressing cells was 16 hours after induction.

3.3.3. Purification of recombinant endo-xylanase protein

3.3.3.1. Purification of endo-xylanase by His-tag affinity chromatography

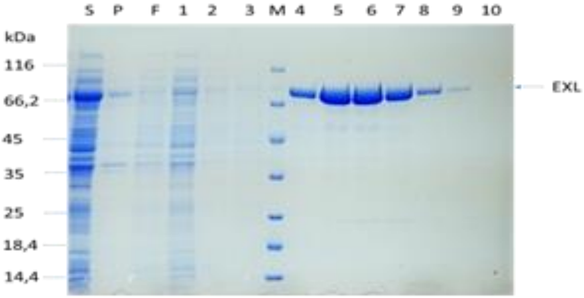

Fig 3.23. Analysis of protein fractions during purification process of the recombinant endo-xylanase in SDS-PAGE.

S, P: total EXL soluble and insoluble protein; F: sample after column passage; 1: sample washed with 50 mM imidazole buffer; 2-3: sample washed with 100 mM imidazole buffer; 4-10: sample fractions after washing with 250 mM imidazole buffer; M: standard protein ladder (Thermo Scientific)

With the appropriate concentration of imidazole competitor, the intracellular proteins that adhered nonspecifically to the substrate were washed away with washing buffer, allowing the recovery of relatively clean xylanase recombinant protein in the fractions collected with the pushing buffer. To obtain high purity protein, the product of the first purification was used for the second purification also using His-tag affinity chromatography column.

The results of software analysis based on the method of comparing the density of the protein band to be analyzed with the entire run showed that the protein samples after purification had only one dark band, the xylanase protein bands all had very high purity, respectively 99.53% - 95.87% - 95.56%, meeting the standard of purified protein with purity above 95%.

3.3.3.2. Desalting protein after chromatography

The purity of the protein sample was calculated as the percentage of purified protein over the total protein in the sample. Using Quantity One software, the purity of the 3 fractions after de-salting was calculated to reach values of 97% - 95% -98% respectively, completely satisfying the requirement that the protein sample after purification should have a purity of over 95%. The recombinant xylanase protein was purified and de-salted 3 times with high repeatability. The activity of the purified protein was recovered at a rate of about 25%.

3.3.4. Study on xylanase enzyme properties and enzyme kinetic parameters

3.3.4.1. Determine the optimal temperature for enzyme activity

The research results showed that xylanase enzyme showed the highest activity when incubated at 40ºC, and showed relatively strong activity when the reaction temperature was between 35ºC - 50ºC (reaching over 90% of maximum activity compared to 40ºC).

3.3.4.2. Determine the optimal pH for enzyme activity

The research results showed that xylanase protein showed the highest activity when incubated at pH 5.5, reaching 25.6 U/mg, and showed relatively strong activity when the reaction pH was in the range of 5-6 (reaching over 90% of maximum activity compared to pH 5.5).

3.3.4.3. Study of thermal stability of enzymes

The research results showed that xylanase enzyme is stable at the optimal temperature for activity of 40ºC, at 50°C, activity begins to decrease after only 1 hour of treatment and completely loses activity after incubation for 24 hours. At high temperature (60ºC), the enzyme completely loses activity after only 1 hour of incubation. Therefore, it can be concluded that xylanase enzyme is not heat stable.

3.3.4.4. Study on the effect of some metal ions and chemicals on xylanase activity

The results of xylanase activity test showed that in the presence of Fe^{3+}, Cu^{2+}, Co^{2+}, Zn^{2+} ions and SDS denaturant, the activity of xylanase enzyme decreased sharply, especially the two ions Fe^{3+}, Cu^{2+} almost made the enzyme lose its activity (activity only reached 7% compared to the control sample without added ions). In the environment with the presence of chemicals 2-mercaptoethanol, Tween-20 and Triton X100, the activity of endo-xylanase was also inhibited to only about 40%.

3.3.4.5. Study of substrate specificity of enzyme

Based on the graph, xylanase shows strong activity with the specific substrate xylan, on the contrary, it shows almost no activity with other substrates such as CMC, filter paper, pNPG and pNPX, specifically the catalytic activity for these substrates is negligible. Therefore, it can be concluded that the xylanase enzyme is specific for xylan substrate. In addition, the EXL enzyme has also been tested for activity with some esterase substrates such as gelatin, pectin, tributyrin or skim milk at a concentration of 1%, however, the enzyme did not show activity.

3.3.4.6. Study of enzyme kinetic parameters

The kinetic parameters K_m and V_{max} of the xylanase enzyme were determined to be 14.56 mg/ml and 0.86 µmol/min respectively with a specific activity of 171.56 IU/mg.

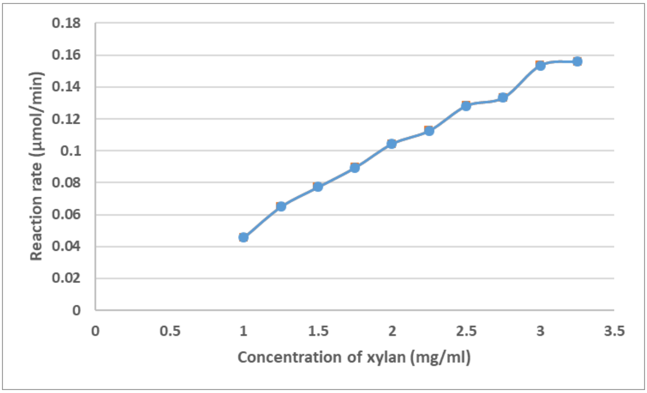

Fig 3. 33. Correlation graph between substrate concentration and reaction rate of recombinant endo xylanase enzyme.

CONCLUSIONS AND FUTURE WORK DIRECTION

CONCLUSIONS

1. The DNA metagenome of bacteria in the goat rumen was sequenced with a conventional size of 8.6 Gbs and a large size (deep sequencing) of 48.66 Gbs. From the deep sequencing data, 3,411,867 contigs were assembled with a coverage of 64.22%, 5,367,270 genes were annotated, of which 4,311,093 genes (80.32%) were classified. Both data predicted genes of bacterial origin to account for 99.8%. The phylum Bacteroidetes was the most abundant phylum with 45.29% of the total genes, followed by the phylum Firmicutes with 30.38%. At the genus level, 49.93% of the genes remained unclassified. The most abundant genus was *Prevotella,* accounting for 25.79% of the total genes.

2. From the 48.66 Gb data, 65,443 genes encoding 30 enzymes/proteins involved in lignocellulose degradation were classified into taxa (accounting for 99.85% of the total lignocellulase genes), including 21,029 cellulase genes, 41,756 hemicellulase genes, and 4,769 lignocellulose pretreatment genes. The genus *Prevotella* played an important role, providing up to 27% of the genes involved in lignocellulose degradation. These genes were identified on PUL structures to enhance lignocellulose hydrolysis. The HMM modelling tool was successfully designed for efficient mining of galactanase, glucuronyl esterase, hydrogen peroxide oxidoreductase, xyloglucanase, laccase, CBM (1-84), cellobiohydrolase, beta-glucuronidase, beta-xylosidase, beta-mannosidase GH2, lichenase, alpha-glucuronidase (GH76N) and xylanase GH44 genes.

3. Among the 739 complete genes encoding endo-xylanases, 108 sequences were identified from Prevotella, all of which belonged to the GH10 family, and 10 enzymes had an additional CE1 auxiliary region with

a T_m higher than 65°C and were acidic enzymes. The denovogenes_5086 sequence encoding the GH10-CE1 endo-xylanase was selected, codon optimized, and successfully expressed in *E. coli*. The enzyme was purified by His-tag affinity chromatography with a purity of over 95% and a recovery of 25%. The purified enzyme had optimal activity at 40°C, pH 5.5. The enzyme was stable and stable at the optimal temperature for up to 24 hours but lost activity immediately after incubation at 60°C for 1 hour. The enzyme acted on a specific substrate, xylan. The enzyme activity decreased when adding Fe^{3+}, Cu^{2+}, Co^{2+}, Zn^{2+}, Ni^{2+}, Ca^{2+}, Na^{+}, Mn^{2+} (1 mM) and commonly used chemicals such as SDS (1%), urea (1 µM), 2-mercaptoethanol (1 µM), EDTA (1 µM), tween 80 (1mM), triton X-100 (1 µM). The kinetic parameters K_m and V_{max} of the xylanase enzyme were determined to be 14.56 mg/ml and 0.86 µmol/min respectively, with a specific activity of 171,56 IU/mg.

FUTURE WORK DIRECTION

Continue to study suitable methods and conditions to determine the esterase activity of EXL enzyme, towards producing enzyme mixtures for application in environmental treatment and production of raw materials - biofuels.

NEW CONTRIBUTIONS OF THE THESIS

1. A 48.66 Gb goat rumen bacterial metagenome DNA data has been constructed and for the first time the role of *Prevotella* in enhancing feed digestion in the goat rumen was deeply analyzed and clarified.
2. For the first time, the thesis has built a hidden Markov model (HMM) tool for functional annotation of the gene group encoding the carbohydrate binding module (CBM) and some enzymes involved in the pretreatment of lignocellulose, cellulase, and hemicellulase.
3. Endo-xylanase EXL encoded from the gene of goat rumen bacteria has been successfully expressed and purified with high activity.

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

1. **Trong-Khoa Dao**, Thi-Huyen Do, Ngoc-Giang Le, Hong-Duong Nguyen, Thi-Quy Nguyen, Thi-Thu-Hong Le, Nam-Hai Truong, 2021, Understanding the role of *Prevotella* genus in the digestion of lignocellulose and other substrates in Vietnamese native goats' rumen by metagenomic deep sequencing, *Animals* 11(11), 3257.

2. Thi-Huyen Do, **Trong-Khoa Dao**, Khanh-Hoang-Viet Nguyen, Ngoc-Giang Le, Thi-Mai-Phuong Nguyen, Tung-Lam Le, Thu-Nguyet Phung, Nico M. van Straalen, Dick Roelofs, Nam-Hai Truong, 2018, Metagenomic analysis of bacterial community structure and diversity of lignocellulolytic bacteria in Vietnamese native goat rumen. *Asian-Australasian journal of animal sciences*, 31(5), 738–747.

3. Thi Huyen Do, Ngoc Giang Le, **Trong Khoa Dao**, Thi Mai Phuong Nguyen, Tung Lam Le, Han Ly Luu, Khanh Hoang Viet Nguyen, Van Lam Nguyen, Lan Anh Le, Thu Nguyet Phung, Nico M. van Straalen, Dick Roelofs, Nam Hai Truong, 2018, Metagenomic insights into lignocellulose-degrading genes through Illumina-based de novo sequencing of the microbiome in Vietnamese native goats rumen. *Journal of General and Applied Microbiology*. 64(3):108-116.

4. Khanh Hoang Viet Nguyen, **Trong Khoa Dao**, Hong Duong Nguyen, Khanh Hai Nguyen, Thi Quy Nguyen, Thuy Tien Nguyen, Thi Mai Phuong Nguyen, Nam Hai Truong, Thi Huyen Do, 2021, Some characters of bacterial cellulases in goats' rumen elucidated by metagenomic DNA analysis and the role of fibronectin 3 module for endoglucanase function. *Animal Bioscience* 34(5): 867-879.

5. **Đào Trọng Khoa,** Đỗ Thị Huyền, Trσιơng Nam Hải, 2018, Nghiên cứu biểu hiện expansin tái tổ hợp trong *Escherichia coli, Báo cáo khoa học Hội nghị Công nghệ sinh học toàn quốc 2018*, 138-142.

6. **Đào Trọng Khoa**, Đỗ Thị Huyền, Trương Nam Hải, 2020, Khai thác gene mã hóa endo-1,4-beta-xylanase từ dữ liệu DNA metagenome vi khuẩn trong dạ cỏ dê bằng mẫu dò, *Tạp chí Công nghệ Sinh học* **19**(3): 519-528.

7. Trương Nam Hải, Đỗ Thị Huyền, **Đào Trọng Khoa,** 2021, Trình tự gene mã hóa expansin có nguồn gốc từ vi khuẩn trong dạ cỏ dê và expansin tái tổ hợp có khả năng làm tăng chuyển hóa xenluloza tinh thể của xenlulaza. Bằng độc quyền giải pháp hữu ích số 2701, số đơn 2-2017-00312, Cục sở hữu trí tuệ.

8. Nguyen Hai Dang, Do Thi Huyen, Nguyen Thi Kien, Ha Thi Thuy Hoa, Le Quynh Giang, **Dao Trong Khoa**, Truong Nam Hai, 2021, Expression of gene coding endoglucanase GH5-4 derived from meagenomic DNA data of bacteria in goats rumen in *Escherichia coli, Academia Journal of Biology*, 43(2): 17–26.