FUNCTIONAL METAGENOMIC APPROACH IN IDENTIFY CELLULOSE-DEGRADING ENZYMES FROM MALAYSIAN PALM OIL MILL EFFLUENT

BY

FARAH FADWA BEN BELGACEM

A thesis submitted in fulfilment of the requirement for the degree of Doctor of Philosophy of Engineering (Biotechnology Engineering)

Kulliyah of Engineering
International Islamic University Malaysia

OCTOBER 2020

ABSTRACT

Major source of enzymes is microbes and during cultivation step, more than 85% of them resist cultivation and their genetic patrimony is loss. This work sought to bioprospect cellulose-degrading enzymes from microbiota of the palm oil mill effluent (POME). To get access to this great microbial diversity and to discover the biocatalysts behind, this research has adopted metagenomic approach which technically escapes the cultivation step and screens the microbial DNA for desired enzymatic activity. Metagenomics consists of the creation and screening of metagenomic DNA libraries. *In vivo* identification of cellulose-degrading enzymes was carried out with high-throughput screening and *in silico* identification of genes encoded the enzymes was performed with algorithm-based methods while the results validation was executed by recombinant enzymes expression, purification and characterisation. Culture-enrichment strategy based on natural selection principle was used in the early stage to enhance the screening hit rate. Metagenomic DNA was extracted from enriched and non-enriched sample to construct 109,824 fosmids (4.49 Gb) metagenomic DNA library. In this library, pCC1FOS fosmid and *E. coli* EPI300T1^{R} are the vector and surrogate host of the cloning system, respectively. A high throughput functional metagenomic screen was developed and applied to search for cellulose-degrading enzymes within the library clones using 4-methylumbelliferyl-ß-D-glucopyranoside (MUGlc) and 4-methylumbelliferyl-ß-D-cellobioside (MUC) fluorogenic substrates. The screens were normalised using robust z-score and highest rated clones (100) were then selected. Their fosmids were isolated and sequenced with Hiseq (Illumina) of next-generation sequencing strategy. For quality control of the reads, SolexaQA and FastQC tools were used. Poor quality bases were removed with DynamicTrim algorithm, all bases with Qphred less than 20 were trimmed, and the LengthSort algorithm was used to remove sequences less than 50 bp. *de Bruijn* graph of *de novo* assembly algorithm has organised the reads on *k*-mers to build contigs, and Velvet optimiser has selected the optimum *k*-mers. These contigs were the input of SSPACE algorithm used to locate and orient contigs. Codon DNA sequences (CDS) were identified with PRODIGAL software. The genes identification was carried out following Blastp and SmartBLAST. Seventeen of bioprospected putative cellulose-degrading enzymes were cloned into pBAD-TOPO plasmid and expressed in TOP10 *E.coli* cloning. Enzymes were purified with HisTrap HP column with the aid of a FPLC system. Two putative glucanases and two putative ß-glucosidases were then biochemically characterised for optimal pH and temperature in the presence of substrates MUGlc and MUC, *p*NPG and *p*NPC and CMC as well. In NGS-data analysis step, 4900 contigs and 3540 scaffolds were constructed. 42,247 CDS were detected and 96 potential cellulose-degrading enzymes were identified which evinces the richness of POME metagenome on biocatalysts. The protein sequences of 15 cellulose-degrading enzymes are 100% similar to protein sequences available in protein databases while 40 enzymes show (80-99%) similarities, 24 enzymes (60-79%), 14 enzymes (40-59%) and 3 enzymes show less than 40% similarity, this reflects the qualification of functional metagenomics to bioprospect untapped enzymes. The potential types of enzymes are 19.20% glucanases, 31.32% glucosidases and 46.48% glucoside hydrolases with cellulose-degrading enzyme conserved domains. For the 17 expressed enzymes, three different glycoside hydrolase

families, (enzyme 1, 2, 10 and 21) from GH3, (enzyme 6, 12 and 20) from GH5, (enzyme 3) from GH8 and other glycoside hydrolase families. Enzyme 3 is probably an example of untapped enzyme; it was active toward MUC and MUGlc. The optimum catalysis activity by enzyme 3 occurred at 50 °C and pH 4. For enzymes 4, 11 and 13, no enzymatic activity was detected due to low expression level. This research was very challenging but rewarding. It lays the foundation of diverse and untapped biocatalysts discovery. The bioprospected enzymes found in this metagenomic DNA library can be produced and optimised to be used in different industrial applications. In addition, the NGS-analysed-data can be usd to study the diversity of POME.

ملخص البحث

الميكروبات هي المصدر الاساسي للانزيمات، وخلال مرحلة الاحياء، أكثر من 85 بالمئة من هذه الميكروبات تقاوم العيش في المختبر وبالتالي نفقد المعلومة الجينية الخاصة بها. سعى هذا البحث الى ايجاد انزيمات مفككة للسيليلوز من مجموع الميكروبات المتواجدة في مخلفات مصانع زيت النخيل. للتوصل الى هذا التنوع الميكروبي الهائل واكتشاف المحفزات الحيوية، تبنت هذه الدراسة منهج الميتاجينوميك الذي يتخلى عن خطوة الاحياء ويفحص الحمض النووي للانزيمات. يتضمن منهج الميتاجينوم انشاء وفحص مكتبات الحمض النووي. الكشف عن الانزيمات المفككة للسيليلوز *In vivo* تم بواسطة الفحص عالي المردودية، والكشف عن الجينات الأصلية لهذه الانزيمات *In silico* تم بوسائل لوغارتمية، كما أن التؤكد من النتائج تم عن طريق انتاج الانزيمات، تنقينها، ثم دراسة خصائصها. عملية تغذية الوسط المعتمدة على الانتقاء الطبيعي، استعملت في بادئ الأمر لتدعيم نتائج الفحص. استخلص الحمض النووي منقوص الريبوزوم من كل من العينات المدعمة والغير مدعمة لانشاء مكتبة ميتاجينومية عددها 109824 فوسميد (4.49جيجابايت) ، في هذه المكتبة pCC1FOS هو الفوسميد الحامل و$EPI300T1^R$ هي الخلية المضيفة لعملية الاستنساخ. الفحص الميتاجينومي العالي المردودية للبحث عن هذه الانزيمات، أستخدم كل من MUC و MUGlc كمادتين فاعليتين مفسفرتين. نتائج الفحص عدلت عن طريق حساب Robust z-score. من خلال هذه النتائج تم اختيار اعلى مئة خلية في الترتيب. في المرحلة الثانية، استخلصت فوسميدات هذه الخلايا المئة وتم قراءة حمضها النووي منقوص الريبوزوم بجهاز Hiseq (Illumina) الذي يعتبر من فئة الجيل الثاني. استخدم SolexaQA و FastQC كوسيلتين لترتيب البيانات. التسلسلات ذات الجودة الأقل من اللازمة يتم التخلص منها بواسطة وسيلة DynamicTrim algorithm والسلاسل مع اقل من Qphred20 تم التخلص منها مع السلاسل القصيرة أقل من 50bp عن طريق LengthSort algorithm. De Bruijn graph هو اللوغاريثم المستعمل لاعادة تسلسل الحمض النووي المنقوص الريبوزوم (de *novo* assembly) الى Kmers ومنها بناء Contigs. أما Velvet Optimiser عمل على اختيار ال Kmer الأمثل. SSPACE لوغاريتم

عمل على ترتيب ال Contigs وتوجيهها. سلسلة الكودون للحمض النووي منقوص الريبوزوم تم تحديدها عن طريق برنامج PRODIGAL. التعرف على الجينات بواسطة Blastp و SmartBLAST. سبعة عشر من الانزيمات المحتملة استنسخت في البلاسميد pBAD-TOPO للخلية البكتيرية TOP10 E.coli. تم تصفية الانزيمات بتقنية FPLC بواسطة HisTrap HP column. أنزيمان من نوع جلوكانيز وأنزيمان من نوع البيتاجلوكوزيديز تم تحديد خصائصهما البيوكيميائية من ال pH الأمثل ودرجة الحرارة المثلى في حضور كل من المواد المتفاعلة المفسفرة MUC و MUGlc والمواد المتفاعلة الملونة pNPG و pNPC والمادة المتفاعلة CMC أيضا. من نتائج قراءة التسلسل بواسطة الجيل الجديد، شكلنا 4900 contigs و 3540 scaffolds. تم اكتشاف 42247 CDS منها 96 أنزيم مفكك للسيليلوز, مما يؤكد وفرة الانزيمات في عينة البحث POME. أضحت نتائج التسلسل البروتييني أن السلاسل البروتينية ل 15 أنزيم مطابقة 100% لبروتينات متواجدة في قاعدة البيانات، 40 أنزيم مطابق بنسب تتراوح بين (80-99%)، 24 أنزيم مطابقة بنسب تتراوح بين (60-79%)، 14 أنزيم مطابق بنسبب (40-59%) وثلاث أنزيمات مطابقة بأقل من 40%، وهذا مايؤكد أيضا كفائة الميتاجينوميك في اكتشاف الانزيمات الغير معروفة مسبقا. الانزيمات المكتشفة هي 12.20% جلوكانيز، 31.32% جلوكوزايديز، و 46.48% من مختلف أنواع الانزيمات المفككة للسيليلوز. من هذه الانزيمات 17 انزيما غير مكتشف تم تجربتها في المختبر، تنتمي الانزيمات 1، 2، 10 و 21 لعائلة GH3، وتنتمي الانزيمات 6، 12 و 20 لعائلة GH5 وينتمي الانزيم 3 لعائلة GH8. الانزيم 3 يحتمل أن يكون مثالا عن الانزيمات الغير مكتشفة سابقا حيث أنه أبدى فاعليته اتجاه المادتين الفعالتين MUC و MUGlc الفاعلية المفككة المثلى لهذا الانزيم 3 كان في درجة حرارة 50 درجة مئوية وpH. أما فيما يخص الانزيمات 4 و11 و13 فلم يتم رصد تفاعل كيميائي نظرا لمستوى الانتاج الضئيل. هذا البحث يعد صعبا ولكن مجزيا. فهذا يعد حجر أساس لاكتشاف العديد من الانزيمات المكتشفة والغير مكتشفه مسبقا. الانزيمات المكتشفه في هذه المكتبة الميتاجينومية يمكن تصنيعها وتحسينها لاستخدمها في صناعات مختلفة. كما أن البيانات المتحصل عليها من قراءة التسلسل للحمض النووي منقوص الريبوزوم تعتبر مادة غنية لدراسه التنوع في عينة POME.

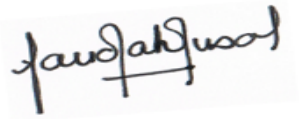
APPROVAL PAGE

The thesis of Farah Fadwa Ben Belgacem has been approved by the following:

_____
Hamzah Mohd. Salleh
Supervisor

_____
Ibrahim Ali Noorbatcha
Co-Supervisor

_____
Md. Zahangir Alam
Co-Supervisor



_____
Faridah Yusof
Internal Examiner

_____
Muhammad Mukram bin Mohamed Mackeen
External Examiner

_____
Amir Feisal Merican
External Examiner

_____
Fouad Mahmoud Mohamed Rawash
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Farah Fadwa Ben Belgacem

Signature ……. Date…….

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

BIOPROSPECTING OF CELLULOSE-DEGRADING ENZYMES FROM MALAYSIAN PALM OIL MILL EFFLUENT BY FUNCTIONAL METAGENOMIC APPROACH

I declare that the copyright holder of this thesis jointly owned by the student and IIUM.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Farah Fadwa Ben Belgacem

…… ……

Signature Date

ACKNOWLEDGEMENTS

Firstly, I thank Allah, the almighty, for giving me the strength to carry on this research and for blessing me with many great people who have been my greatest support in both my personal and professional life.

I would like to seize this opportunity to express my deepest regards and gratitude to whom I dedicate this work, my dear father Mohammed Ben Belgacem, my lovely mother and my mother-in-law Zoura et Hanou, my beloved husband Oualid Abdelkader Bellag and my little angle Mohamed Khaled Alwalid, my pretty sisters Fairouz, Imen and Nesrine, my brothers Abdelkader and Akram, and to all my family members and friends who granted me the gift of their unwavering belief in my ability to accomplish this goal: thank you for your support and patience.

Finally, a special thanks to my supervision comity, Prof. Dr. Hamzah Mohd. Salleh, Prof. Dr. Ibrahim Ali Noorbatcha and Prof. Dr. Md. Zahangir Alam for their continuous support, encouragement and leadership and to all the lecturers and researchers in IIUM, and for that, I will be forever grateful. I also would like to thank Prof. S.G. Withers (University of British Columbia, Canada) for kindly providing "chlorocoumarin xylobioside" needed for this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ATA	Amine transaminase
ASRA	Adaptive substituent reordering algorithm
BACs	bacterial artificial chromosomes
BLAST	Basic local alignment search tool
BG	β-Glucosidase
CAZy	Carbohydrate-active enzymes
CBH	Cellobiohydrolase
CMC	Carboxyl methyl cellulose
DBG	*de Bruijn* graph
DI	Digital imaging
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleoside triphosphate
DMSO	Dimethyl sulfoxide
EDTA	Ethylenediaminetetraacetic acid
EtBr	Ethidium bromide
EpPCR	Error-prone PCR
FACS	Fluorescent activated cell sorting
FPLC	Fast protein liquide chromatography
GH	Glycosyl Hydrolase
GFP	Green fluorescent protein
HTS	Hight-Throughput Screening
IVTC	*In vitro* compartmentalization
KCl	Potassium chloride
K_2HPO_4	Potassium hydrogen phosphate
MAD	Median absolute deviation
MEGAWHOP	Megaprimer PCR of whole plasmid
$MgCl_2.6H_2O$	Magnesium chloride hexahydrate
$MgSO_4.7H_2O$	Magnesium sulfate heptahydrate
MUGlc	4-methylumbelliferyl-ß-D-glucopyranoside
MUC	4-methylumbelliferyl-ß-D-cellobioside
NaCl	Sodium chloride
Na_2CO_3	Sodium carbonate
NGS	Next generation sequencing
*p*NPC	4-Nitrophenyl β-D-cellobioside
*p*NPG	4-Nitrophenyl β-D-glucuronide
OD	Optical density
PBD	Protein data bank
PCR	Polymerase chain reaction
PLP	Pyridoxal-5'-phosphate
POME	Palm oil mill effluent
QSAR	Quantitative structure-activity relationship
RBS	Ribosome binding site
RNase	Ribonuclease
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
TAE	Tris-acetate-EDTA

TE	Tris-EDTA
UV	Ultraviolet

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Various industries have utilised enzymes as part of their production due to their high selectivity, efficiency and rapidity in catalytic reactions. By using enzyme in industrial processes, it is possible to eliminate or at least prevent possible hazards caused by extreme parameters (including pressure, temperature, and pH) for chemical reactions to occur. In addition, lower energy consumption is possible as enzyme's optimum parameters are relatively mild. Enzyme usage also allows production of specific products only, therefore it is possible to avoid production of unwanted by-products that may be challenging and expensive to dispose or may even pose negative effect (Binod et al., 2013).

Lignocellulose is the major structural component of plant material; it is a sustainable and renewable resource available for use in biotechnological applications. Thus, cellulose-degrading enzymes are amongst the highly demanded industrial enzymes due to their ap plications in various industries including detergents, fermentation, food, textile, pulp and paper (Hill et al., 2006). Using cellulose-degrading enzymes in enzymatic hydrolysis of cellulosic materials into fermentable sugars for bioethanol production may reduce the process cost greatly as the processes may be improved to take plate at mild conditions (temperature 45 to 50 °C and pH 4 to 6) (Kuhad et al., 2011). Cellulose-degrading enzymes are commonly synthesised by microorganisms such as fungi, bacteria and actinomycetes during their growth on cellulosic materials; these enzymes may be thermophilic, mesophilic, aerobic or

anaerobic. Several genera are more extensively studied compared to others including *Aspergillus*, *Cellulomonas*, *Clostridium*, and *Trichoderma* (Kuhad et al., 2010).

Microbial enzymes are normally obtained through cultivation and subsequent screening of pure strains of microorganisms. It is later found out that only 1-15% of microbial genomes are cultivable under laboratory conditions while more than 85% have never been studied before (Amann, 1995).

Techniques of specifically cloning environmental DNA involving the hereditary diagrams of whole microbial consortia what is called metagenome, furnish molecular grouping space that alongside cunning *in vitro* advancement technologies will collaborate synergistically to bring a limit of accessible succession space into biocatalytic application (Lorenz et al., 2002). The new method involves direct DNA isolation from an environmental sample followed by direct cloning for subsequent screening and product expression to allow unbiased genomic representation of the microbes.

Molecular techniques of environmental DNA cloning in synergy with *in vitro* microbiological technologies facilitate the biocatalytic application (Handelsman et al., 1998). This field is mostly similar to genome library development and screening, with the distinction that the cloned DNA does not start from a solitary known microorganism, but instead from the whole consortia in a special environment. Function-based metagenomics has developed as a strong strategy for DNA model approval and protein bioprospecting from natural and hand-engineered biological systems (Mewis et al., 2013).

The success of the cloning may be proven by positive presence of novel proteins through function-based screening where the clones' biological activity is monitored (Kumar et al., 2015). Screening may be performed through two

alternatives; sequence-based and activity or function-based screening. It is recommended to combine both methods to obtain the complete picture of the community. Sequence based screening allows the identification of the sequence of interest while functional screening provides identification of unknown and novel genes that might not be recognisable by only sequence based screening (Riesenfeld et al., 2004). It is a challenge to develop effective and sensitive functional screening where it previously relies on noticeable changes in colony morphology or the appearance of zymogram (Teather and Wood, 1982). High throughput screening is introduced where screens are conducted in 384-well plate format to increase the efficiency and comparability between samples. Cell lysis is also imposed in this method to overcome intracellular accumulation of enzyme activity. Substrate of interest (in this case fluorogenic substrate) may interact with the enzyme released from the cells and allow the reading of enzyme activity using fluorescence-based microplate readers (Taupp et al., 2011).

In this study, palm oil mill effluent (POME) was chosen because it is a rich habitat for microorganisms since it contains nutrients and growth factors. In the presence of significant quantity of cellulosic fruit residues, microorganisms secrete cellulose-degrading enzymes to adapt and survive in POME environment. Malaysia's POME samples were collected and enriched in the laboratory to enhance the screening of cellulose-degrading enzymes. After metagenomic DNA extraction, sequences of around 40 kb were cloned into a fosmid vector and transformed into a bacterial host, which is generally *Escherichia coli.* A high number of multiplications of *E. coli* with the inserted DNA created fosmid libraries of POME metagenomic DNA. The presence of cellulose-degrading enzymes was conducted with microtiter high-throughput screening of the four libraries using fluorogenic substrates in fluorescence-able

microplate reader. Only clones with the desired enzymes were able to metabolise the substrate and merge fluorescence to predict their presence. Due to the limited budget, only 100 high-rated clones were sequenced with next generation sequencing (NGS). Their DNA was sequenced to define the coding DNA sequence of each enzyme. Second cloning was carried out with those sequences into an expression vector. Finally, the enzymes were expressed, purified and characterised to discover more about the cellulose-degrading enzymes.

1.2 PROBLEM STATEMENT AND RESEARCH SIGNIFICANCE

The number of cellulose-degrading enzymes which are currently successfully utilised at the industrial level is considered low compared to the high demand due to the variety of cellulosic composition in plant biomass. In addition, the available enzymes in the market are considered expensive which imposes a real obstacle of the process. These problems indicate the need for further research and development of versatile and low-cost enzymes to grow progressively sustainable and financially competitive generation processes (Adrio & Demain, 2014). As microorganisms are the main source of enzymes, the real problem behind the lack of enzymes versatility and novelty is the weakness of classical methods to benifit the maximum possible of this available source (Kumar et al., 2015; Vavourakis et al., 2018). The evolution of bioinformatic techniques to study the biodiversity of microorganisms revealed the presence of unexpected numbers of unkown species with complicated phylogenetic connections which have been missed by traditional cultivation methods. Named “unculturable microorganisms”, are a prosperous source of biocatalysts if the right method of investigation is followed.