ACADEMIA JOURNAL OF BIOLOGY 2019, 41(3): 101–105
DOI: 10.15625/2615-9023/v41n3.10866

MINING CYTOCHROME P450 GENES THROUGH NEXT GENERATION SEQUENCING AND METAGENOMIC ANALYSIS FROM BINH CHAU HOT SPRING

Nguyen Van Tung[1], Nguyen Huy Hoang[1], Nguyen Kim Thoa[2,*]

[1]Institute of Genome Research, VAST, Vietnam
[2]Institute of Biotechnology, VAST, Vietnam

Received 7 November 2018, accepted 10 May 2019

ABSTRACT

Cytochrome P450s (CYPs) are one of the largest distributed enzymes, which catalyze more than 20 different reactions. At present, there has been an increasing realization of the power of P450 biocatalysts for the industrial synthesis of pharmaceuticals, agrochemicals, bulk chemicals, food ingredient, etc. On the other hand, the conditions of industrial processes at high temperature, high-pressure or in chemical solvent require the enzymes, which catalyze the bioconversion, have a specific properties such as thermostability, chemical tolerance or barophilicity. Up to date, the number of thermostable P450s is limited. Nowadays, DNA-metagenome technique gives us a chance to catch novel genes and unique interesting enzymes from microbial community in certain ecology. In this paper, metagenomic DNA extracted from water samples from Binh Chau hot spring was sequenced using Illumina HiSeq platform and was analysed to mining putative genes encoding cytochrome P450. The sequencing generated 9.4 Gb of reads containing 156,093 putative ORFs, of these, 106,903 genes were annotated in NCBI non-redundant protein sequence database. Among all the ORFs were annotated, 68 putative ORFs encoding cytochrome P450 were found belong to 36 specific groups of cytochrome P450 protein family. Of these, the melting temperature (Tm) from thirty-six completed ORFs was predicted for a better understanding of thermodynamic stability.

Keywords: Binh Chau hot spring, cytochrome P450, metagenomic, thermostable enzyme, HiSeq Illumina.

Citation: Nguyen Van Tung, Nguyen Huy Hoang, Nguyen Kim Thoa, 2019. Mining cytochrome P450 genes through next generation sequencing and metagenomic analysis from Binh Chau hot spring. *Academia Journal of Biology*, 41(3): 101–105. https://doi.org/10.15625/2615-9023/v41n3.10866.

[*]*Corresponding author email*: nkthoa@ibt.ac.vn

INTRODUCTION

Cytochrome P450s (CYPs) are one of the largest gene super families, which distribute widely in all living organisms, including bacteria, fungi, plants and animals (Werck-Reichhart & Feyereisen, 2000). They catalyze a large number of mono-oxygenation reactions including aromatic and aliphatic hydroxylation, N-oxidation, etc. The physiological role of P450 enzymes includes biosynthesis of endogenous compounds such as steroids, hormones, and secondary metabolites, fatty acid oxidation, and xenobiotic and especially drug metabolism (Cryle et al., 2003; Guengerich, 2001). Bacterial cytochrome P450 participate in the oxidative biodegradation of natural and man-made chemicals, which can be useful for bioremediation of environment and diversification of natural products through oxygenation, often hydroxylation or epoxidation (Urlacher & Schmid, 2002). Cytochrome P450 have some disadvantages such as limited stability, low levels of activity, and requirement for redox partner(s) (Urlacher et al., 2004; Chefson & Auclair, 2006). The discovery of thermophilic P450s may solve the first problem, limited stability (Mandai et al., 2009).

Thermophilic organisms grow optimally between 50 and 80°C. Enzymes of thermophilic organisms (thermophilic enzymes) show thermostability properties which fall between those of hyperthermophilic and mesophilic enzymes. These thermophilic enzymes are usually optimally active between 60 and 80°C. Intrinsically stable and active at high temperatures, thermophilic and hyperthermophilic enzymes offer major biotechnological advantages such as easier to purify by heat treatment, and performing enzymatic reactions at high temperatures allows higher substrate concentrations, lower viscosity, fewer risks of microbial contaminations, and often higher reaction rates (Vieille & Zeikus, 2001).

This study present the analysis of a large data set generated by Illumina platform sequencing of metagenomic DNA extracted from water samples collected from Binh Chau hot spring in Vung Tau, Viet Nam.

MATERIALS AND METHODS

Metagenome sequencing and assembly

Water was collected from Binh Chau hot spring in Vung Tau, Viet Nam. Sampling was performed directly in a borehole with temperature of 82^{o}C and pH = 7.5 using an ultrafiltration filter has pore size around 0,01 µm. Metagenomic DNA from water samples then was extracted by applying PowerWater® DNA isolation kit and stored at -20^{o}C.

After collected and extracted, the metagenomic DNA was sequenced using Illumina HiSeq platform. The raw sequence data was analyzed using a standard bioinformatics approach. Raw paired-end reads were assessed and subjected to quality control using FastQC and Trimmomatic (Bolger et al., 2014). The reads containing more than 10% ambiguous bases or containing adapter sequences or the reads containing more than 50% low quality (Q < 15) bases were removed. Preprocessed reads were assembled using SOAPdenovo2 (Luo et al., 2012) by default parameter. Using the Bowtie2 tool (Langmead & Salzberg, 2012), high quality reads were mapped to their own contigs, thus, single or incorrect paired-end reads were filtered from correct reads. During assembly, K-mer was used for statistical analysis. After the assembly process, only contigs no less than 500 bp were kept for further analysis.

Gene prediction and functional annotation

MetaGeneMark (Zhu et al., 2010) version 2.10 was used to predict open reading frames (ORFs) based on assembly results. All the predicted ORFs were compared with NCBI non-redundant protein sequence database (Pruitt et al., 2005) using an E-value cutoff 1e-5, retrieving proteins with the highest sequence similarity with the given genes along with their protein functional annotations. The sequence of putative ORFs encoding cytochrome P450s were blasted

against D.R Nelson-bacteria database (Nelson, 2009) to determine specific groups of cytochrome P450 protein family.

Melting temperature prediction

Melting temperature of complete putative genes encoding cytochrome P450 were predicted directly from protein sequences using the Tm Index program (available at http://tm.life.nthu.edu.tw/) (Ku et al., 2009).

RESULTS AND DISCUSSION

Metagenome sequencing and assembly

Illumina sequencing of the metagenomic DNA extracted from Binh Chau hot spring yielded 9.4 Gb of sequence reads. Of these, 0.04% were ambiguous sequences, 0.41% were adapter sequences and 6.97% were low quality reads, resulting in > 9.4 Gb of useful reads (accounting for 92.58% of all data).

Assembly of these useful reads yielded 51,346 contigs (the cut-off value was set at 500 bp) accounting for 65.12% of the total reads (table 1).

Table 1. SOAPdenovo2 assembly metrics of Binh Chau hot spring

Parameter	Metric
Total number of useful reads	93,999,534
Number of reads in contigs	61,212,496
Number of contigs	51,346
Largest contig (bp)	1,767,609
Shortest contig (bp)	500
Average contig length (bp)	3,351
N50 contig length (bp)	9,791
N90 contig length (bp)	1,059

Gene prediction and functional annotation

MetaGeneMark predicted 156,093 putative ORFs. The length distribution of putative genes was shown in figure 1.

All the predicted ORFs were compared with NCBI non-redundant protein sequence database. In total, 106,903 genes were annotated in NCBI non-redundant protein sequence database. Among all the ORFs were annotated, 68 putative ORFs encoding cytochrome P450 were found. Of the identified putative ORFs encoding cytochrome P450, 36 (52.94%) were completed, whereas 32 (47.06%) were fragments of ORFs.

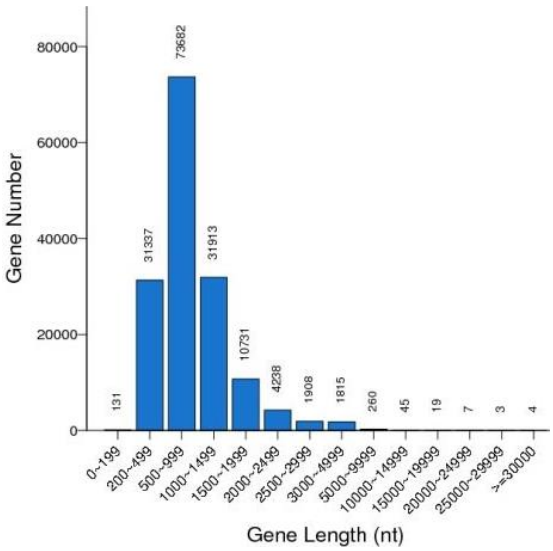

Figure 1. Length distribution of genes obtained from analyzing DNA metagenome Binh Chau hot spring database

Putative cytochrome P450 genes

The sequences of 68 putative ORFs encoding cytochrome P450 were blasted against D.R Nelson-bacteria database. The results showed putative ORFs encoding cytochrome P450 belong to 36 specific groups of cytochrome P450 protein family. Of these, the sample was dominated by five groups CYP253, CYP107, CYP197, CYP205, CYP110 (table 2).

Proteins of both hyperthermophilic and mesophilic microorganisms generally constitute from the same 20 amino acids; however, the extent of thermal tolerance of any given protein is an inherent property of its amino acid sequence. Thermodynamic stability is defined by the protein's free energy of stabilization and melting temperature (Tm, the temperature at which 50% of the protein is unfolded) (Vieille & Zeikus, 2001). In this study, the melting temperature of 36 complete ORFs encoding

cytochrome P450 were predicted using the Tm Index program. Of these, 13 ORFs (34.21%) had Tm higher than 65°C, 23 ORFs (60.53%) had Tm lower 65°C but higher 55°C and 2 ORFs (5.26%) had Tm lower than 55°C (Fig. 2).

Table 2. Five domination groups of cytochrome P450 in the sample

Groups of cytochrome P450	Total number of ORFs
CYP253	7
CYP107	5
CYP197	5
CYP205	4
CYP110	3

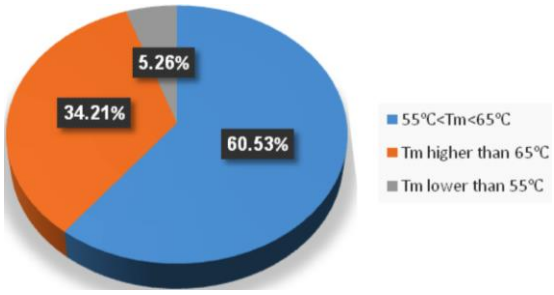

Figure 2. Tm distribution pie chart of complete ORFs encoding cytochrome P450

Enzymes characterized from thermophiles are usually optimally active at temperatures close to the host organism's optimal growth temperature. That explained why most putative ORFs encoding cytochrome P450 were identified in the data had melting temperature higher than 55°C (94.74%).

Vietnam has over 300 hot springs where is diverse with thermophilic microorganisms. However, the number of thermophilic microorganism study is limited despite the potential for great industrial applications. Most of the study performed isolation microorganism from hot spring. It has been estimated that less than 1% of the microorganisms in the natural environment can be cultured in the laboratory. It is increasingly recognized that a huge number of natural products exists in unculturable microbes with chemical, biological, and functional activities for potential uses in various industrial and biomedical applications (Handelsman, 2004).

To date, only a few thermostable P450s have been described in the literature and most of them are of archaeal origin. All of thermostable P450s have been reported was isolate from thermal microoganisms. The first identified and the best studied thermostable P450 CYP119 was isolated from the acidothermophilic archaeon *Sulfolobus sulfataricus* (Wright et al., 1996; Koo et al., 2000, 2002). The best characterized thermostable P450 of bacterial origin is CYP175A1 from *Thermus thermophilus,* the first β-carotene hydroxylase of the P450 superfamily (Blasco et al., 2004). The total denaturation temperature of CYP175A1 is 88°C, while the total denaturation temperature of thermophilic P450 ranges from 47–61°C.

CONCLUSION

In this study, the metagenomic DNA extracted from water samples collected from Binh Chau hot spring in Vung Tau, Viet Nam was sequenced and analysis using bioinformatics pipeline to mining putative ORFs edcoding cytochrome P450. Among all the ORFs were annotated, 68 putative ORFs encoding cytochrome P450 were found belong to 36 specific groups of cytochrome P450 protein family. The melting temperature of 36 complete ORFs was predicted. These genes are strong candidates for future studies.

Acknowledgements: This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number: 106-NN.02-2014.60.

REFERENCES

Blasco F. et al., 2004. CYP175A1 from Thermus thermophilus HB27, the first beta-carotene hydroxylase of the P450 superfamily. *Appl. Microbiol. Biotechnol.*, 64: 671–674.

Bolger A. M. et al., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30: 2114–2120.

Chefson A. and Auclair K., 2006. Progress towards the easier use of P450 enzymes. *Mol. Biosyst.*, 2: 462–469.

Cryle M. J. et al., 2003. Reactions Catalyzed by Bacterial Cytochromes P450. *Aust. J. Chem.*, 56: 749–762.

Guengerich F. P., 2001. Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem. Res. Toxicol.*, 14: 611–650.

Handelsman J., 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev. MMBR*, 68: 669–685.

Koo L. S. et al., 2002. Enhanced electron transfer and lauric acid hydroxylation by site-directed mutagenesis of CYP119. *J. Am. Chem. Soc.*, 124: 5684–5691.

Koo L. S., Tschirret-Guth R. A., Straub W. E., Moënne-Loccoz P., Loehr T. M., & De Montellano P. R. O., 2000. The active site of the thermophilic CYP119 from Sulfolobus solfataricus. *Journal of Biological Chemistry*, 275(19): 14112–14123.

Ku T. et al., 2009. Predicting melting temperature directly from protein sequences. *Comput. Biol. Chem.*, 33: 445–450.

Langmead B. and Salzberg S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9: 357–359.

Luo R. et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1: 18.

Mandai T. et al., 2009. Construction and engineering of a thermostable self-sufficient cytochrome P450. *Biochem. Biophys. Res. Commun.*, 384: 61–65.

Nelson D. R., 2009. The cytochrome p450 homepage. *Hum. Genomics*, 4: 59–65.

Pruitt K. D. et al., 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 33: D501–D504.

Urlacher V. and Schmid R. D., 2002. Biotransformations using prokaryotic P450 monooxygenases. *Curr. Opin. Biotechnol.*, 13: 557–564.

Urlacher V. B. et al., 2004. Microbial P450 enzymes in biotechnology. *Appl. Microbiol. Biotechnol.*, 64: 317–325.

Vieille C. and Zeikus G. J., 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev. MMBR*, 65: 1–43.

Werck-Reichhart D. and Feyereisen R., 2000. Cytochromes P450: a success story. *Genome Biol.*, 1: REVIEWS3003.

Wright R. L. et al., 1996. Cloning of a potential cytochrome P450 from the archaeon Sulfolobus solfataricus. *FEBS Lett.*, 384: 235–239.

Zhu W. et al., 2010. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.*, 38: e132–e132.