Bioinformatics Analysis to Construct *Cellulose-binding Module* Synthetic Gene and Design Primer

Febriana Dwi Wahyuni [1], Asrul Muhamad Fuad [2]
[1]Faculty of Health Sciences, University of Esa Unggul, Indonesia.
[2]Research Center for Biotechnology, LIPI, Indonesia.
email: febriana@esaunggul.ac.id

Abstract

Cellulose-binding module (CBM) is a protein domain commonly found in various types of cellulase enzymes. The function of this CBM can be used for the binding process and the immobilization of a protein in the cellulose matrix. CBM can be obtained from several organisms, one of them is Trichoderma reesei. To get a gene, it does not have to be isolated from the original organism. Gene sequences can be obtained synthetically through bioinformatics analysis in accordance with the same gene sequences as those at Gene Bank. Bioinformatics analysis can be used to find new gene sequences or existing genes. This study aims to get a cbmsyn synthetic gene quickly and efficiently without reducing protein activity, which can then be ligated with other genes so that it functions as an immobilized enzyme. From the results of bioinformatics analysis, obtained DNA sequences measuring around 498 pb with 166 amino acid protein lengths. The sequence was modified by adding several restriction sites, namely BamHI, AfeI, and ScaI. The DNA sequences obtained were optimized with the Pichia pastoris codon.

Keywords: *bioinformatics, synthetic gene, cellulose-binding module, primer design.*

1. INTRODUCTION

Cellulose-binding Module (CBM) is a protein domain found in many types of cellulase enzymes. CBM generally functions in the process of binding cellulase enzymes to the substrate, namely cellulose so that it can facilitate the work of the cellulose enzyme (Wang et al. 2001). This CBM function can be used for the binding process and immobilization of a protein in the cellulose matrix.

CBM was initially classified as Cellulose Binding Domain (CBD) based on the initial discovery of several modules that bind cellulose (Ito et al. 2004). CBM can be obtained from various organisms, one of which is Trichoderma reesei. Trichoderma is a cellulase enzyme producing microorganism that has often been used in industry. To obtain a CBM protein coding gene, it does not have to be isolated from the original organism. Gene sequences can be obtained synthetically through bioinformatics applications.

Bioinformatics is the science of collecting and analyzing biological data such as the genetic code (Saraswati, 2017). Bioinformatics is an alternative in exploring sequences of genes or enzymes because there are various biological information presented in an online database (Wahyuni, 2018). The advantage of this method is that it is economical and can be a preliminary study before a real experiment is conducted.

This bioinformatics analysis aims to create synthetic CBM genes (cbmsyn) with web-based programs from T. reesei organisms at GeneBank. In addition, a specific primary design for cbmsyn gene amplification will also be made.

2. RESEARCH METHOD

Analysis of structure, function, and expression of cbmsyn genes

The analysis was done using several bioinformatics databases such as http://www. ncbi.nlm.gov, Pubmed, http://www.uniprot. org, and http://www.ensembl.org. The CBM gene sequences obtained were then optimized with the Pichia pastoris codon using the DNAWorks 3.2 program.

Received 14 February 2019 | Received in revised form 15 March 2019 | Accepted 29 March 2019 | Published online 1 April 2019

CBM protein sequence analysis

Sequence analysis of CBM proteins (protein domains, psycho-chemical characteristics, amino acid scale profiles, signal peptide predictions, target peptide predictions) using the site http://www.expasy.org, Prosite, Protparam, Pro-Scale, Psipred, SignalP, TargetP and PeptideCutter.

Analysis of 3D structure of CBM proteins.

The analysis performed using the Bank Data protein site, http://bioinf.cs.ucl.ac.uk/ PSIPRED menu and 3D protein using the website http://www.pdb.org Swiss menu PYMOL model and software.

Primary Design Analysis

The analysis conducted using Primary 3 software from gene sequences which have high similarity (in the dominant cds base region) with a unique base for primary design. To avoid the hairpin, select the sequence of results from the primary Perlprimer software design.

Amplification of the CBMSyn gene with Polymerase Chain Reaction (PCR)

Amplification of the cbmsyn gene using cbmF 'specific primers (5'-GTT ACTCCTATCGATTCTAGAAGCGCTGA TTACAAGGACGATGAT-3') and cbmR '(5'-GATGAGTTTTGTTCTAGAGACA AACATTGTGAGTAGTAATCGTTAGA GTA-3') with a total volume of 25 µl containing 1 µl DNA, 0.2 µl DNA polymerase enzyme, 0.625 µl primers cmbf' and cbmR', 0.5 µl dNTPs, 5 µl Buffers and 17.05 ddH2O. PCR amplification was carried out in 30 cycles. One cycle consists of three stages, namely denaturation (denaturation), attachment (annealing), and elongation (extension). The PCR condition consisted of initial denaturation at 980C for 3 minutes.

The next stage of the PCR cycle was carried out 30 times starting with denaturation at 980C for 30 seconds, annealing at a temperature of 44.5-57.50C for 30 seconds, initial polymerization at a temperature of 720C for 3 minutes. The PCR cycle was followed by final polymerization at 720C for 5 minutes and the temperature dropped to 40C. The PCR results were then visualized by 1% agarose gel electrophoresis.

Electrophoresis of PCR Results

The results of PCR were migrated into 1.5% agarose gel at 100 Volt 40 minutes. 1 kb DNA marker is used as a marker. Gel staining using ethidium bromide (10µg / mL) for 10 minutes, then put into distilled water for 5 minutes to wash ethidium bromide which is still attached to the gel. Gels containing DNA fragments are visualized using the UV Trans Illuminator and documented using the Digibox Camera Documentation System Gel.

Purification of the cbmsyn gene from agarose gel

Purification of the cbmsyn gene from agarose gel was carried out by DNA Fragments Extraction Kit / PCR from Geneaid. Purification of DNA from agarose gel consists of four main stages, namely gel dissociation, DNA binding, washing, and DNA elution.

3. RESULT AND DISCUSSION

CBM gene sequence analysis

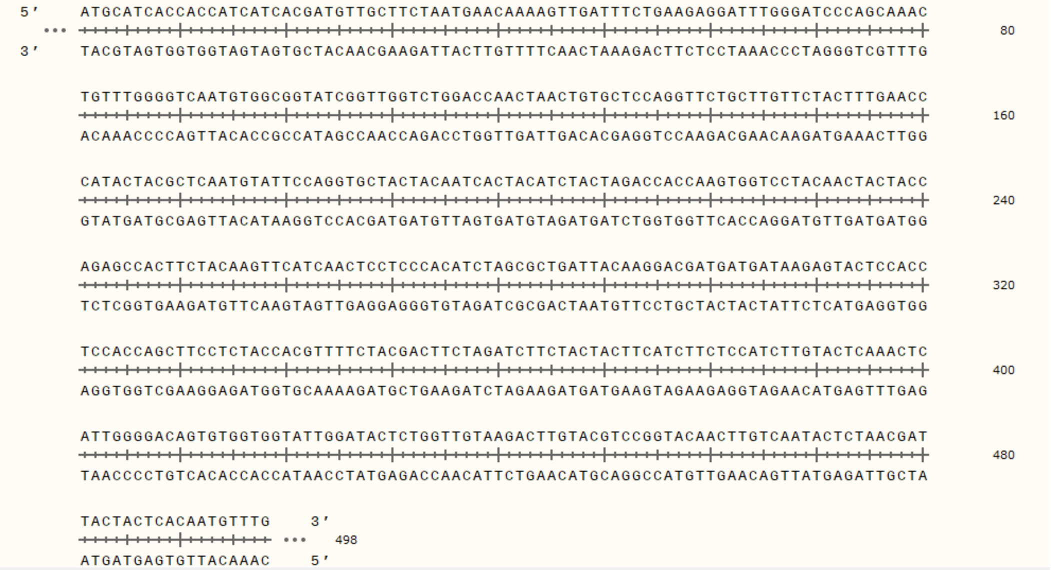
Based on the search results of the CBM gene sequences in the geneBank through NCBI with access number M15665, cbm gene sequences were obtained from Trichoderma reesei (Hypocrea jecorina) organisms consisting of 498 base pairs (figure 1.)

The gene has been modified by adding some restriction sites like BamHI, AfeI, and ScaI. The DNA sequence was then optimized with the Pichia pastoris codon using the DNA Works 3.2 program.

Analysis of 3D structure of CBM proteins

To determine the presentation of amino acids, molecular weight, isoelectric point (pI) and other physico-chemical properties of the protein Endoglucanase EG-I, an analysis was performed using Postparam Expasy.

Received 14 February 2019 | Received in revised form 15 March 2019 | Accepted 29 March 2019 | Published online 1 April 2019



```
5'   ATGCATCACCACCATCATCACGATGTTGCTTCTAATGAACAAAAGTTGATTTCTGAAGAGGATTTGGGATCCCAGCAAAC
 ...                                                                                  80
3'   TACGTAGTGGTGGTAGTAGTGCTACAACGAAGATTACTTGTTTTCAACTAAAGACTTCTCCTAAACCCTAGGGTCGTTTG

     TGTTTGGGGTCAATGTGGCGGTATCGGTTGGTCTGGACCAACTAACTGTGCTCCAGGTTCTGCTTGTTCTACTTTGAACC
                                                                                     160
     ACAAACCCCAGTTACACCGCCATAGCCAACCAGACCTGGTTGATTGACACGAGGTCCAAGACGAACAAGATGAAACTTGG

     CATACTACGCTCAATGTATTCCAGGTGCTACTACAATCACTACATCTACTAGACCACCAAGTGGTCCTACAACTACTACC
                                                                                     240
     GTATGATGCGAGTTACATAAGGTCCACGATGATGTTAGTGATGTAGATGATCTGGTGGTTCACCAGGATGTTGATGATGG

     AGAGCCACTTCTACAAGTTCATCAACTCCTCCCACATCTAGCGCTGATTACAAGGACGATGATGATAAGAGTACTCCACC
                                                                                     320
     TCTCGGTGAAGATGTTCAAGTAGTTGAGGAGGGTGTAGATCGCGACTAATGTTCCTGCTACTACTATTCTCATGAGGTGG

     TCCACCAGCTTCCTCTACCACGTTTTCTACGACTTCTAGATCTTCTACTACTTCATCTTCTCCATCTTGTACTCAAACTC
                                                                                     400
     AGGTGGTCGAAGGAGATGGTGCAAAAGATGCTGAAGATCTAGAAGATGATGAAGTAGAAGAGGTAGAACATGAGTTTGAG

     ATTGGGGACAGTGTGGTGGTATTGGATACTCTGGTTGTAAGACTTGTACGTCCGGTACAACTTGTCAATACTCTAACGAT
                                                                                     480
     TAACCCCTGTCACACCACCATAACCTATGAGACCAACATTCTGAACATGCAGGCCATGTTGAACAGTTATGAGATTGCTA

     TACTACTCACAATGTTTG    3'
                       ...   498
     ATGATGAGTGTTACAAAC    5'
```

Figure 1.Modified cbmsyn gene sequences by loading several restriction sites (GGATCC: BamHI; AGCGCT: AfeI ’AGTACT: ScaI).

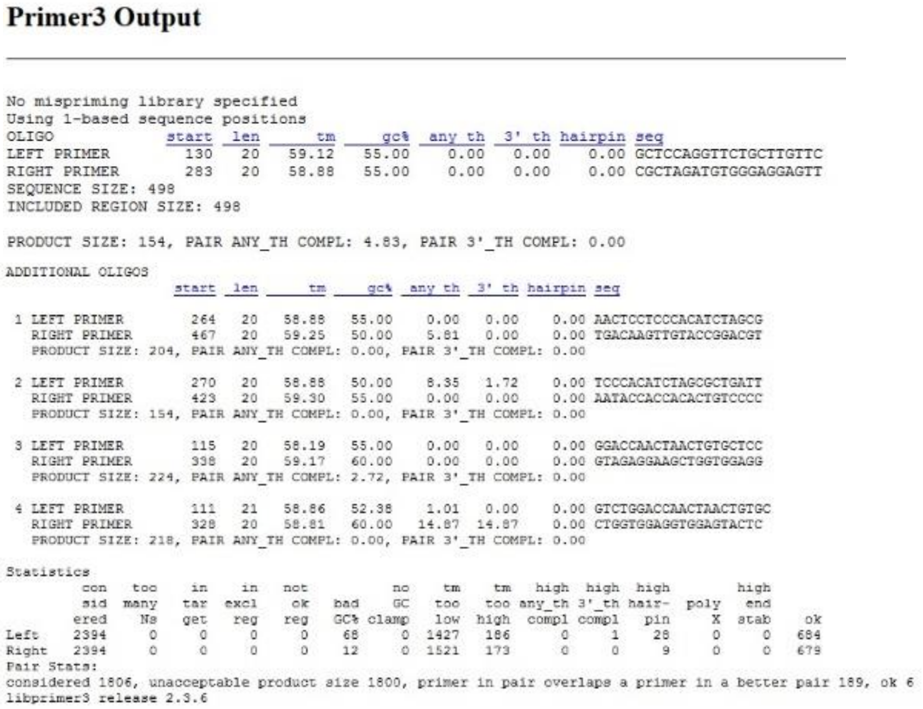
Table 1. Physical and Chemical Properties of the protein Endoglucanase EG-I

Parameter	Protein Endoglucanase EG-1
Berat Molekul	42870.67
pH isoeleltrik	5.17
Komposisi asam amino	ala (A) 128 25.7% Cys (C) 132 26.3 % Gly (G) 89 17.9 % Thr (T) 150 30.1 %
Komposisi atom	Carbon C 1555 Hydrogen H 2614 Nitrogen N 498 Oxygen O 649 Sulfur S 131 $C_{1555}H_{2614}N_{498}O_{649}S_{131}$ Total jumlah atoms: 5447
Index aliphatic	25.70
Grand average of hydropathicity (GRAVY)	0.837

Received 14 February 2019 | Received in revised form 15 March 2019 | Accepted 29 March 2019 | Published online 1 April 2019

Primary Design Analysis

Primary design is part of bioinformatics which is the most important factor in determining unknown DNA sequences (Seprianto, 2018). The primary design using "Primary 3" software was then confirmed by BLAST NCBI to see primary specificity. To see the secondary structure (hairpin loop, dimer) produced by the primer, an analysis was performed using PerlPrimer. The primary design of the sequences was selected and entered into the primary output 3 program, and several primary design alternatives were obtained as follows:



Primer3 Output

```
No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len     tm     gc%  any th  3' th hairpin seq
LEFT PRIMER        130   20  59.12   55.00    0.00   0.00    0.00 GCTCCAGGTTCTGCTTGTTC
RIGHT PRIMER       283   20  58.88   55.00    0.00   0.00    0.00 CGCTAGATGTGGGAGGAGTT
SEQUENCE SIZE: 498
INCLUDED REGION SIZE: 498

PRODUCT SIZE: 154, PAIR ANY_TH COMPL: 4.83, PAIR 3'_TH COMPL: 0.00

ADDITIONAL OLIGOS
                   start  len     tm     gc%  any th  3' th hairpin seq

 1 LEFT PRIMER       264   20  58.88   55.00    0.00   0.00    0.00 AACTCCTCCCACATCTAGCG
   RIGHT PRIMER      467   20  59.25   50.00    5.81   0.00    0.00 TGACAAGTTGTACCGGACGT
   PRODUCT SIZE: 204, PAIR ANY_TH COMPL: 0.00, PAIR 3'_TH COMPL: 0.00

 2 LEFT PRIMER       270   20  58.88   50.00    8.35   1.72    0.00 TCCCACATCTAGCGCTGATT
   RIGHT PRIMER      423   20  59.30   55.00    0.00   0.00    0.00 AATACCACCACACTGTCCCC
   PRODUCT SIZE: 154, PAIR ANY_TH COMPL: 0.00, PAIR 3'_TH COMPL: 0.00

 3 LEFT PRIMER       115   20  58.19   55.00    0.00   0.00    0.00 GGACCAACTAACTGTGCTCC
   RIGHT PRIMER      338   20  59.17   60.00    0.00   0.00    0.00 GTAGAGGAAGCTGGTGGAGG
   PRODUCT SIZE: 224, PAIR ANY_TH COMPL: 2.72, PAIR 3'_TH COMPL: 0.00

 4 LEFT PRIMER       111   21  58.86   52.38    1.01   0.00    0.00 GTCTGGACCAACTAACTGTGC
   RIGHT PRIMER      328   20  58.81   60.00   14.87  14.87    0.00 CTGGTGGAGGTGGAGTACTC
   PRODUCT SIZE: 218, PAIR ANY_TH COMPL: 0.00, PAIR 3'_TH COMPL: 0.00

Statistics
         con   too    in    in   not          no    tm    tm  high  high  high        high
         sid  many   tar  excl    ok   bad    GC   too   too any_th 3'_th hair-  poly   end
         ered   Ns   get   reg   reg   GC% clamp   low  high compl compl   pin     X  stab     ok
Left     2394    0     0     0     0    68     0  1427   186     0     1    28     0     0    684
Right    2394    0     0     0     0    12     0  1521   173     0     0     9     0     0    679
Pair Stats:
considered 1806, unacceptable product size 1800, primer in pair overlaps a primer in a better pair 189, ok 6
libprimer3 release 2.3.6
```

Figure 2. Primary design through the primary 3 output program

From the results of the primary analysis, the one that meets the requirements is primary 2 which has a product size of 154 bp starting at base 270. Forward Primer1 has a length of 20 bases, Tm 58.88 ° C, and% GC 50%. Reverse primer 1 has a length of 20 bases, Tm 59.30 ° C, and% GC 55%. This primer does not form a hairpin loop, dimer, or palindrome after being analyzed using software http://sg.idtdna.com / analyzer / Applications / OligoAnalyzer /

Site Analysis Restriction

To find out the restriction sites found in the cbmsyn gene, the analysis was performed using Snapgene software. The purpose of this restriction site is to ensure that the gene can be cut with one of the desired endonuclease restriction enzymes. The results of the analysis are presented in Figure 3.

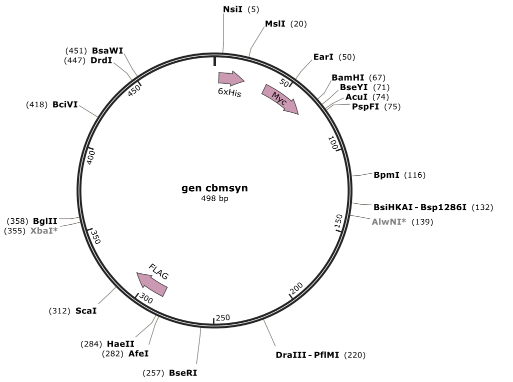


Figure 3. Restriction Enzyme Mapping on the cbmsyn gene with Snapgene software

Received 14 February 2019 | Received in revised form 15 March 2019 | Accepted 29 March 2019 | Published online 1 April 2019

Amplification of the cbmsyn gene with Polymerase Chain Reaction (PCR) and Purification

Amplification of the cbmsyn gene is done using CbmF 'and CbmR' specific primers. The length of the cbmsyn gene DNA fragment was 498 pb and after amplification it produced a cbmsyn gene size of 261 pb (Figure 4). The PCR results are then purified with the aim of increasing DNA purity and minimizing the occurrence of contaminants

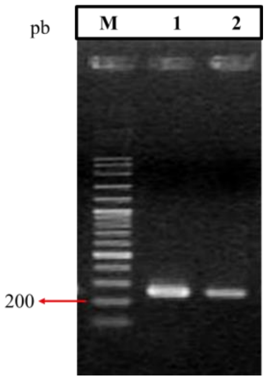


Figure 4. Results of cbm gene PCR with a size of around 261 pb. M = marker 100 bp, 1: result of cbmsyn 1st elute gene amplification purification, 2: cbmsyn gene amplification results of 2nd elute gene.

4. CONCLUSION

From the results of bioinformatics analysis, obtained DNA sequences measuring around 498 pb with 166 amino acid protein lengths. The sequence was modified by adding several restriction sites, namely BamHI, AfeI, and ScaI. The DNA sequences obtained were optimized with the Pichia pastoris codon.

5. REFERENCES

Ito J, Fujita Y, Ueda M, Fukuda H, Kondo A. 2004. Improvement of cellulose-degrading ability of a yeast strain displaying *Trichoderma reesei* endoglucanase II by recombination of cellulose-binding domain. *Biotechnol Prog.* 20: 668-691.

Jhala, M.K., *et al.* (2011). Role of Bioinformatics in Biotechnology. Information Technology Centre, GAU, Anand. Terdapat di http://openmed. nic.in/1383/01/Role_of_ Bioinformatics_ in_ Biotechnology.pdf

Lehtio, Janne. (2001). Functional Studies and engineering of Family I Carbohydrate-binding Modules. Royal Institute of Technolog. Sweden.

Mount, D.W. 2001. *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Narita, V., Arum, A.L., Isnaeni, S., Fawzya N.Y. (2012). Analisis Bioinformatika Berbasis WEB untuk eksplorasi Enzim Kitonase Berdasarkan Kemiripan Sekuens. *Jurnal Al-Azhar Indonesia Seri Sains dan Teknologi*. Vol 1. No 4: 197-203.

Saraswati, H. 2017. Analisa Bioinformatika Gen E1 dan E2 dari Virus Hepatitis C (HCV) Genotipe 1, 2, 3, dan 6 sebagai Kandidat Vaksin *viral-like Particles* (VLP). *Indonesian Journal of Biotechnology and Biodiversity*, 1 (2): 48-57.

Seprianto dan Wahyuni, FD. 2018. Analisis Bioinformatika Gen Potensial Penyandi *Halichondrin*B dari Spons Laut sebagai Kandidat AntiKanker. *Indonesian Journal of Biotechnology and Biodiversity*, 2 (2): 57-66.

Shoseyov O, Shani Z, Ley I. (2006). Carbohydrate Binding Modules: Biochemical Properties and novel applications. *Microbiology and Molecular Biology*. 70(2): 283-295.

Wahyuni FD, Seprianto. 2018. *Candida antarctica* Lipase B Synthetic Gene: A Bioinformatic Analysis. *Bioscience*. 2:20-29.

Wang Y, Slade MB, Gooley AA, Gooley AA, Atwell BJ, & Williams KL. 2001. Cellulose-binding modules from Extracellular matrix proteins of *Dictyostelium discoideum* stalk and sheath. *Eur. J. Biochem.* 268: 4334-4345.

Received 14 February 2019 | Received in revised form 15 March 2019 | Accepted 29 March 2019 | Published online 1 April 2019