KURSOR Journal
Research on Computing and its Aplications

Vol. 7, No. 1, January 2013 **ISSN 0216 – 0544**

DESIGN OF POTENTIAL CELLULASE PRIMER USING MULTIPLE SEQUENCE ALIGNMENT METHOD

[a]Bahrul Ulum, [b]Wisnu Ananta Kusuma, [c]Joni Prasetyo

[a,b]Department of Computer Science, Bogor Agricultural University, Bogor, Indonesia
[a]Department of Informatics Engineering, Al-Kamal Institute of Science and Technology, Jakarta
[c]Renewable Energy Division, BPPT, Serpong, Indonesia
E-mail: aabahrul@gmail.com

Abstrak

Selulase mempunyai peranan utama dalam pemanfaatan limbah biomassa yang mengandung lignin, hemicellulose, dan cellulose (lignocellulose). Limbah biomassa ini sangat banyak terdapat di lingkungan dan sampai saat ini masih belum dimanfaatkan secara maksimal, dikarenakan banyak mikroorganisme dari alam yang memproduksi enzim selulase dengan jumlah terbatas (aktifitasnya rendah). Dalam rangka meningkatkan produktivitas mikroorganisme untuk menghasilkan selulase, salah satu cara yang dapat diterapkan adalah merancang primer sekuens gen penyandi selulase yang dirangkum dari beberapa mikroorganisme penghasil selulase. Dalam penelitian ini, kami melakukan penyejajaran sekuen DNA penyandi selulase untuk mencari potensial primer untuk meningkatkan produktivitas enzim selulase dengan teknik Multiple Sequence Alignment (MSA). Metode yang digunakan adalah metode progresif (Progressive Alignment Algorithms). Hasil penelitian menunjukan bahwa pada tahap penyejajaran, didapatkan tiga daerah konservatif (conserved regions). Sedangkan pada tahap perancangan dengan beberapa parameter yang telah ditentukan didapatkan 46 pasang primer dari lima sekuen gen penyandi selulase yang didapat dari National Center for Biotechnology Information (NCBI).

Kata kunci: Selulase, *Multiple Sequence Alignment*, Perancangan Primer.

Abstract

The role of cellulase is very important in degrading cellulose which is abundant in the environment, such as in biomass waste that is containing lignin, hemicellulose, and cellulose. Biomass waste is abundant in the environment and is still not fully utilized, because many of the natural microorganisms that produce cellulase enzymesproduce the enzyme in a limited amount (have low activity). In order to improve the productivity of microorganisms in producing cellulase, one of the ways that can be applied is to design primer sequences of genes encoding cellulase summarized from several cellulase-producing microorganisms. In this research,we perform alignment of DNA sequences coding of cellulase to look for potential primer in order to increase the productivity of cellulase enzymes by Multiple Sequence Alignment (MSA) method. The method used is progressive (Progressive Alignment Algorithms). The results showed that in the alignment phase, three conserved regions were obtained. However, in the planning phase by using some predetermined parameters 46 pairs of primer sequences were obtained from five genes encoding cellulase taken from NCBI.

Keywords: Cellulase, Multiple Sequence Alignment, Primer design.

INTRODUCTION

The use of cellulase in Indonesia has been increasing, because cellulase is used for bioconversion of lignocellulosic materials into energy resources from renewable raw materials, due to the depletion of fossil fuel reserves available [1-2]. Moreover, cellulase is highly required in a variety of industries, especially industries with substantial use of cellulase such as the textile, pulp and paper, detergents, pharmacy, agriculture and food [3-4]. Various applications of cellulase make it potential to be produced in Indonesia. Right now, cellulases is generally imported. Cellulase can be produced by a group of bacteria, molds and yeasts. Microbe that is most commonly used is *Trichoderma reesei* [5].

To yield cellulase enzyme, we need to perform a primer design. Primer is a strand of nucleic acid that functions as starting point to synthesize Deoxyribo Nucleic Acid (DNA). Primer designing can be done based on know DNA sequence or on protein sequence. If neither the targeted protein or the DNA sequence know to have closest relationship. One of the ways is to use multiple sequence alignment technique [6].

The multiple alignment of biological sequences has become an essential tool in computational molecular biology. It is used to find conserved regions and motifs in protein families, to detect the homology between new sequences and groups of sequences having an already known function and in a preliminary phase of protein structure prediction [7]. Multiple alignment is also extensively used in molecular evolutionary analysis [8].

The related research of multiple sequence alignment is still focusing on protein sequences [9]. Multiple sequence alignment is also applied on the study of in silico production of hyaluronidase leech (*Hirudo medicinalis*) in genetic engineering [10]. A wide range of algorithms has been investigated in multiple sequence alignment. This algorithm is generally classified into three categories according to their properties exact algorithms [11] progressive algorithms [12-13] and iterative algorithms [14].

One of the three alignment algorithms, progressive algorithm has the advantage of speed and simple sides. Moreover, the algorithm is quite sensitive as alignment algorithm. Progressive alignment algorithms is an approach algorithms in finding sequence globally in multiple sequence alignment.

In this paper we propose to cellulase encoding DNA sequence via multiple sequence alignment technique with progressive alignment algorithm. Furthermore, the results of the alignment will be designed in order to get its potential primer.

SEQUENCE ALIGNMENT

The procedure of this research in general is divided into three parts, which are data preparation, MSA with progressive alignment algorithm and primer design cellulase. More details can be seen in Figure 1.

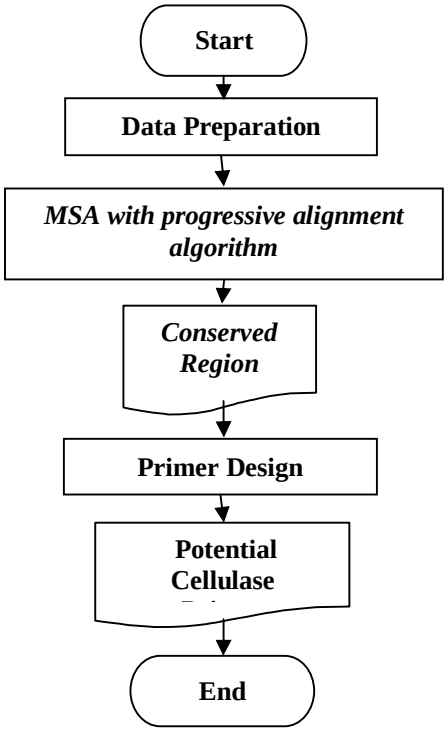

Figure 1. Flowchart of analysis procedure

Data Preparation

In this step, the data is retrieved from National Center for Biotechnology Information (NCBI) GenBank and the data format is set in fasta format.

Progressive Alignment Algorithm

In the second phase DNA data will be aligned using the progressive alignment algorithm. This algorithm is used to find the global sequence alignment of multiple sequences.

Progressive alignment methods use the dynamic programming method to build a multiple sequence alignment starting with the most related sequences and then progressively adding less related sequences or groups of sequences to the initial alignment [15-16]. Relationships among the sequences are modeled by an evolutionary tree where the outer branches or leaves are the sequences. The tree is based on pairwise comparisons of the sequences using one of the phylogenetic methods [17]. Progressive alignment algorithm is as follows:

1. Make a "Guide Tree" based on mutual similarity scores.
2. Start from pairwise alignment with most inner pair.
3. Following the "Guide Tree", add sequences step by step.
4. Perform sequence-profile (or profile-profile) alignment.
5. go to (3) unless all sequence has been aligned.

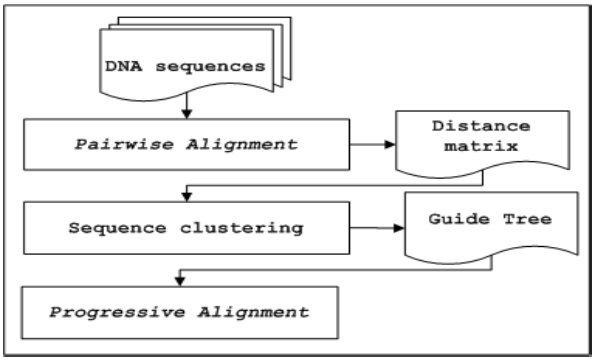

Figure 2. Multiple sequences alignment by progressive alignment algorithm

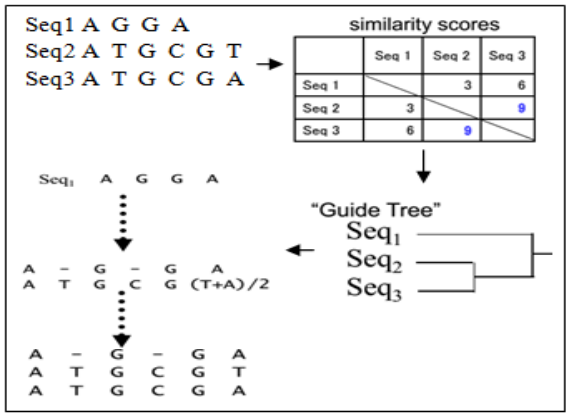

Figure 3. Process progressive alignment algorithm

Primer Design

During last phase, the cellulase encoding DNA, that has been aligned via progressive alignment algorithm will be used to design a potential primer with a predetermined criteria / parameters. The parameters used in this research are as follow: primer length is 20 bp (base pair), primer composition must contain bases G and C with minimum percentage of 45% and maximum percentage of 55% and also melting Temperature (Tm) of minimum = 50°C and maximum of = 60°C. Theoretically (Tm) can be calculated using the formula [2 (A + T) + 4 (C + G)] [18]. The sequence does not form stable hairpins, does not self dimerize, does not cross dimerize with other primers in the reaction, and has a GC clamp at the 3' end of the primer [19]. Primer design aimed to obtain the balance between specificity and efficiency of amplification.

RESULT AND DISCUSSION

In this research, we used DNA data coding for cellulase from eukaryotes organisms. The data is retrieved from NCBI GenBank and the data format is set in fasta format. More details can be seen in table 1.

The search of conserved region of the gene encoding cellulase was done by multiple sequence alignment using progressive alignment algorithm shown in Figure 2. Multiple Sequence Alignment was use to find the highest level of similarity between DNA sequences and homology between related DNA sequences coding for cellulases shown in Figure 4. In simple manual execution of the algorithm that show the step by step of execution describe in figure 3.

Table 1. Cellulase data encoding

ORGANISM	ACCESION NUMBER	LENGTH
Solanum lycopersicum	NM_001247953	1895 bp
Triticum aestivum	AY091512	2196 bp
Arabidopsis thaliana	NM_124350	2360 bp
Solanum lycopersicum	NM_001247933	1717 bp
Solanum lycopersicum	NM_001247938	1780 bp

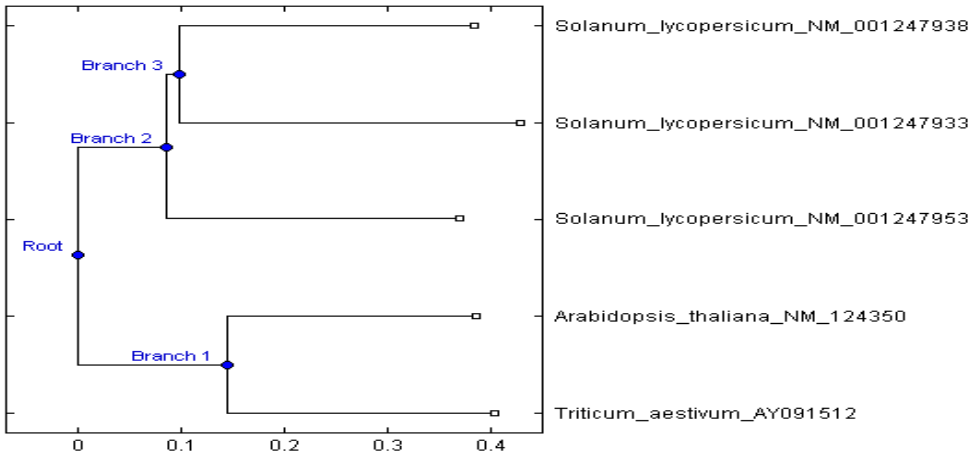

Figure 4. Phylogenetic tree from five cellulase sequence

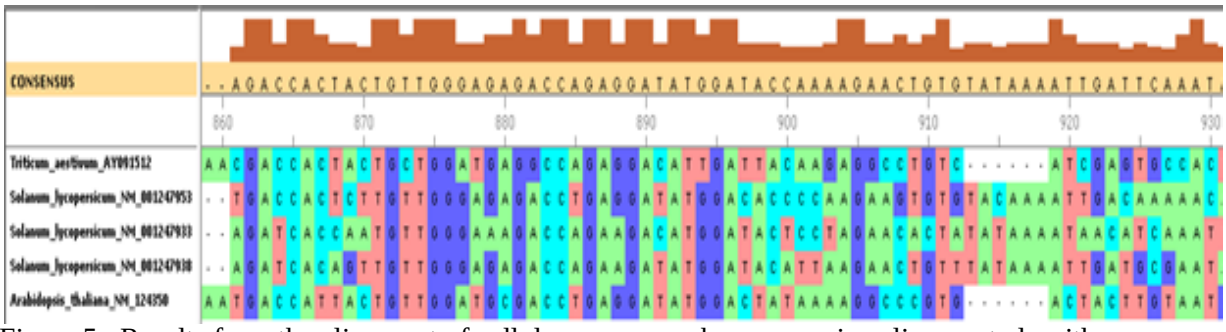

Figure 5. Results from the alignment of cellulase sequence by progressive alignment algorithm

Table 2. Conserved region.

REGION	POSITION	SEQUENCES
1	862-1175	GACCACTCTTGTTGGGAGAGACCTGAGGATATGGACACCCCAAGAAGTGTGTA CAAAATTGACAAAAACACTCCTGGGACTGAAGTTGCTGCTGAAACTGCTGCTG CTCTCGCTGCTGCTTCCTTAGTCTTTAGGAAATGCAACCCATCTTACTCCAAGAT ACTAATCAAAAGGGCCATCAGGGTGTTTGCCTTTGCTGATAAGTATAGAGGTTC ATACAGCAATGGTCTGAGAAAAGTAGTGTGCCCATACTACTGCTCAGTTTCGG GATATGAGGATGAGCTGTTGTGGGGTGCTGCTTGGTTACATAGAGC
2	1567-1765	TGTGGTGGAGTTGTTATTACACCAAAGAGGCTTCGAAATGTAGCCAAAAAACA GGTGGACTATTTGTTAGGAGACAATCCACTAAAAATGTCATACATGGTGGGAT ATGGAGCAAGGTATCCACAGAGGATTCATCACAGGGGATCCTCATTACCCTCA GTCGCGAACCATCCAGCAAAGATACAATGCAGGGATGGTT
3	1790-1915	CACCAAACCCGAACGTACTAGTAGGGGCTGTGGTAGGTGGTCCTGATGAGCAT GATCGTTTCCCAGACGAGCGTTCAGATTACGAGCAATCTGAACCTGCCACTTAC ATTAATGCTCCACTTGTTG

From the analysis result of the alignment of DNA cellulase data encoding, the species who has the score or the highest similarity is *Solanum lycopersicum* NM-001247953. The template used to design potential cellulase primer is the high similary region (conserved region) within *Solanum lycopersicum* NM-001247953 sequence. After that we will be able to know position the potential conserved region primer. Conserved region was derived from the one with the highest similarity within the whole region, as shown in the histogram (Figure 5).

The higher the histogram in the region, the higher the similarity. In this research, the result gives 3 conserved region primer (see Table 2).

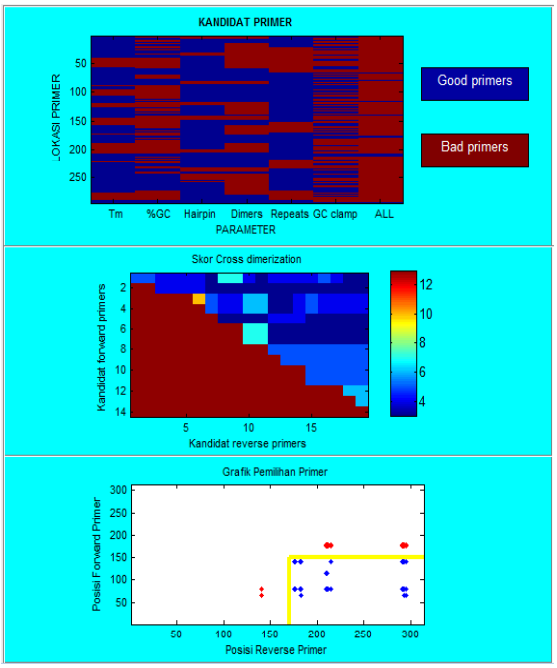

Figure 6. Result primer design from region 1

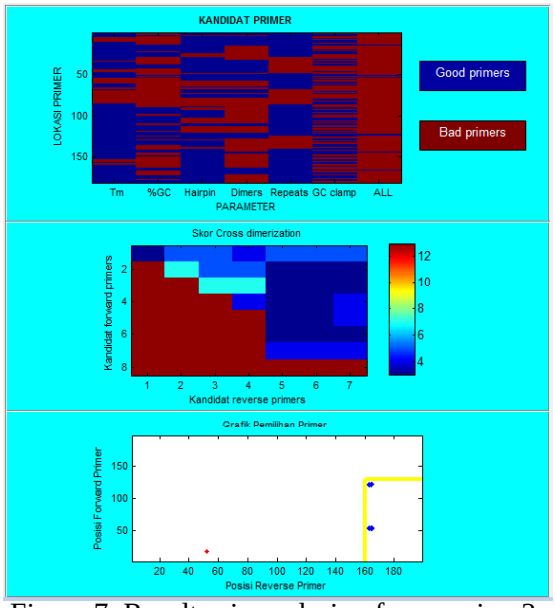

Figure 7. Result primer design from region 2

Based on Figure 6, we are able to known that after the selection the potential location of a good primer from conserved region 1 is positioned between 150 to 210 within the 313 bp sequence. The range of amplification primer selection is 5 '-> 3' = 0-169 and 3 '-> 5' = 170-313. The result of the primer designing from conserved region 1 is 28 primer bases, as shown in table 3.

Based on Figure 7, it can be seen that the location for good potensial primer of conserved region 2 after the selection is between position 50 and 130 from the 198 bp length sequence. The amplification range of primer selection is 5 '-> 3' = 0-150 and 3 '-> 5' = 160 – 198. The result of the primer design from conserve 2 sequence is 13 primer bases, can be seen on Table 4.

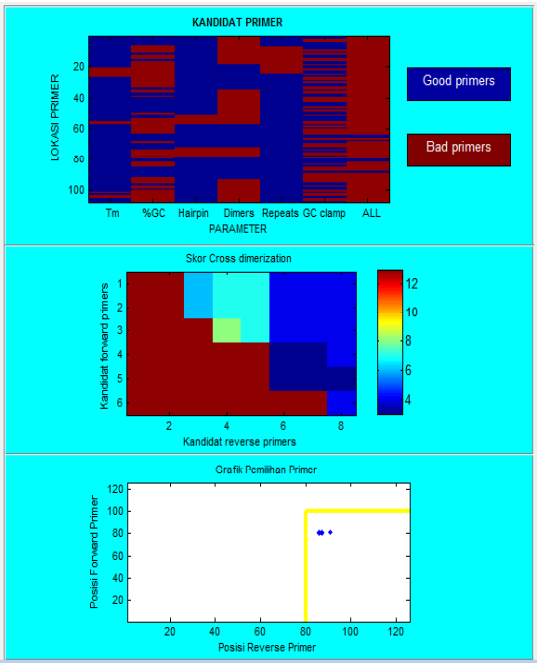

Figure 8. Result primer design from region 3

Based on Figure 8, it is known after the selection the potential location of a good primer for conserved region 3 is positioned between 60 to 90 of 125 bp sequence lenght. The amplification range of primer selection is 5 '-> 3' = 0 – 120 and 3 '-> 5' = 80 – 125. The result of the designed obtained from the conserved region 3 is 5 primer pairs, as shown in Table 5.

CONCLUSION

The main result of this study is that Multiple Sequence Alignment (MSA) using Progressive Alignment Algorithm can generate conserved region of the gene encoding the cellulase used in the primer design. The template used is areas that have a high similarity (conserved region) from *Solanum lycopersicum* NM-001247953 sequence. There are 3 conserved regions that are managed to be obtained, from wich produced 46 primer pairs. The obtained conserved regions are respectively 28 primer bases from conserved region 1, 13 primer bases from conserved region 2, and 5 primer bases from conserved region 3.

Table 3. Primer from Conserved Region 1.

FORWARD / REVERSE PRIMER	START	END	%GC	Tm
GCAACCCATCTTACTCCAAG	140	159	50	54.22
AAAGGCAAACACCCTGATGG	195	176	50	57.22
GCAACCCATCTTACTCCAAG	140	159	50	54.22
CAAAGGCAAACACCCTGATG	196	177	50	55.61
GCAACCCATCTTACTCCAAG	140	159	50	54.22
ATCAGCAAAGGCAAACACCC	201	182	50	57.71
GCAACCCATCTTACTCCAAG	140	159	50	54.22
TATCAGCAAAGGCAAACACC	202	183	45	54.77
GCAACCCATCTTACTCCAAG	140	159	50	54.22
TCTCAGACCATTGCTGTATG	234	215	45	53.03
CTGCTTCCTTAGTCTTTAGG	116	135	45	50.7
ACCATTGCTGTATGAACCTC	228	209	45	53.69
GCTGCTTCCTTAGTCTTTAG	115	134	45	51.26
ACCATTGCTGTATGAACCTC	228	209	45	53.69
CTGCTTCCTTAGTCTTTAGG	116	135	45	50.7
AGACCATTGCTGTATGAACC	230	211	45	53.69
GCTGCTTCCTTAGTCTTTAG	115	134	45	51.26
AGACCATTGCTGTATGAACC	230	211	45	53.69
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
AAAGGCAAACACCCTGATGG	195	176	50	57.22
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
CAAAGGCAAACACCCTGATG	196	177	50	55.61
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
ATCAGCAAAGGCAAACACCC	201	182	50	57.71
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
TATCAGCAAAGGCAAACACC	202	183	45	54.77
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
ACCATTGCTGTATGAACCTC	228	209	45	53.69
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
ACCATTGCTGTATGAACCTC	230	211	45	53.69
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
CAGACCATTGCTGTATGAAC	231	212	45	52.16
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
TCTCAGACCATTGCTGTATG	234	215	45	53.03
AAACACTCCTGGGACTGAAG	66	85	50	55.56
TATCAGCAAAGGCAAACACC	202	183	45	54.77
GCAACCCATCTTACTCCAAG	140	159	50	54.22
TATGTAACCAAGCAGCACCC	310	291	50	56.25
GCAACCCATCTTACTCCAAG	140	159	50	54.22
CTATGTAACCAAGCAGCACC	311	292	50	54.44
GCAACCCATCTTACTCCAAG	140	159	50	54.22
GCTCTATGTAACCAAGCAGC	314	295	50	54.42
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
TATGTAACCAAGCAGCACCC	310	291	50	56.25
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
CTATGTAACCAAGCAGCACC	311	292	50	54.44
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
TCTATGTAACCAAGCAGCAC	312	293	50	53.54
ACTGAAGTTGCTGCTGAAAC	79	98	45	54.81
GCTCTATGTAACCAAGCAGC	314	295	50	54.42
AAACACTCCTGGGACTGAAG	66	85	50	55.56
GCTCTATGTAACCAAGCAGC	312	293	50	53.54
AAACACTCCTGGGACTGAAG	66	85	50	55.56
GCTCTATGTAACCAAGCAGC	314	295	50	54.42

Table 4. Primer from Conversed Region 2.

FORWARD / REVERSE PRIMER	START	END	%GC	Tm
ACAGAGGATTCATCACAGGG	123	142	50	55.18
ATCTTTGCTGGATGGTTCGC	182	163	50	57.05
ACAGAGGATTCATCACAGGG	123	142	50	55.18
TATCTTTGCTGGATGGTTCG	183	164	45	53.62
CACAGAGGATTCATCACAGG	122	141	50	53.6
ATCTTTGCTGGATGGTTCGC	182	163	50	57.05
ACAGAGGATTCATCACAGGG	123	142	50	55.18
GTATCTTTGCTGGATGGTTC	184	165	45	51.97
CACAGAGGATTCATCACAGG	122	141	50	53.6
TATCTTTGCTGGATGGTTCG	183	164	45	53.62
CCACAGAGGATTCATCACAG	121	140	50	53.6
ATCTTTGCTGGATGGTTCGC	182	163	50	57.05
CCACAGAGGATTCATCACAG	121	140	45	53.6
TATCTTTGCTGGATGGTTCG	183	164	45	53.62
GGTGGACTATTTGTTAGGAG	54	73	45	50.75
ATCTTTGCTGGATGGTTCGC	182	163	50	57.05
GGTGGACTATTTGTTAGGAG	54	73	45	50.75
TATCTTTGCTGGATGGTTCG	183	164	45	53.62
GGTGGACTATTTGTTAGGAG	54	73	45	50.75
GTATCTTTGCTGGATGGTTC	184	165	45	51.97
CAGGTGGACTATTTGTTAGG	52	71	45	51.05
ATCTTTGCTGGATGGTTCGC	182	163	50	57.05
CAGGTGGACTATTTGTTAGG	52	71	45	51.05
TATCTTTGCTGGATGGTTCG	183	164	45	53.62
CAGGTGGACTATTTGTTAGG	52	71	45	51.05
GTATCTTTGCTGGATGGTTC	184	165	45	51.97

Table 5. Primer from Conserverd Region 3

FORWARD / REVERSE PRIMER	START	END	%GC	Tm
TACGAGCAATCTGAACCTGCAA	81	100	50	55.99
GTGGCAGGTTCAGATTGC	105	86	50	57.06
TACGAGCAATCTGAACCTGC	81	100	50	55.99
TAAGTGGCAGGTTCAGATTG	106	87	45	53.6
TTACGAGCAATCTGAACCTG	80	99	45	53.44
AAGTGGCAGGTTCAGATTGC	105	86	50	57.06
TTACGAGCAATCTGAACCTG	80	99	45	53.44
TAAGTGGCAGGTTCAGATTG	106	87	45	53.6
TACGAGCAATCTGAACCTGC	81	100	50	55.99
AATGTAAGTGGCAGGTTCAG	110	91	45	53.9

REFERENCES

[1] T. Anindyawati, "Prospek enzim dan limbah lignoselulosa untuk produksi bioetanol," *BS*, vol. 44, pp. 49-56, 2009.

[2] B. Joshi, R.M. Bhatt, Sharma D, Joshi J, Malla R, Lakshmaiah, and Sreerama, "Lignocellulosic ethanol production: Current practices and recent developments," *Biotechnology and Molecular Biology Review*, vol. 6, pp. 172-182, 2011.

[3] Kuhad RC, Gupta R, and Singh A, "Microbial Cellulases and Their Industrial Applications," *Enzyme Research*, vol. 2011, pp. 1-10, 2011.

[4] Mojsov K, "Microbial cellulases and their applications in textile processing," *International Journal of Marketing and Technology*, vol. 2, pp. 12-29, 2012.

[5] R.K. Sukumaran, R.R. Singhania, and A. Pandey, "Microbial cellulases – Production, applications and challenges,"

Journal of Scientific & Industrial Research, vol. 64, pp. 832-844, 2005.

[6] P.V. Lakshmi, A.A. Rao, and G.R. Sridhar, "An Efficient Progressive Alignment Algorithm for Multiple Sequence Alignment", *International Journal of Computer Science and Network Security*, vol. 8, pp. 301-305, 2008.

[7] Gambin A, and Otto R, "Contextual Multiple Sequence Alignment," *Journal of Biomedicine and Biotechnology*, vol. 2, pp. 124–131, 2005.

[8] T Kampke, M Kieninger, and N Mecklenburg, "Efficient primer design algorithms," *Bioinformatics*, vol.17, pp. 214–225, 2001.

[9] R.C. Edgar, and S. Batzoglou, "Multiple sequence alignment," *Science Direct*, vol. 16, pp.1-6, 2006.

[10] Djamil, "Studi in Silico Produksi Hyaluronidase Lintah (Hirudo Medicinalis) Secara Rekayasa Genetika," M. Eng. Thesis, University of Indonesia, Depok, 2005.

[11] J. Stoye, V. Moulton, and A.W. Dress, "DCA: An efficient iplementation of the divide and conquer approach to simultaneous Multiple sequence Alignment," *Computer Applications in the Biosciences*, vol.13, pp. 625-626, 1997.

[12] A. Loytynoja and N. Goldman, "An algorithm for progressive multiple alignment of sequences with insertions", in *Proceedings of The National Academy of Sciences of The United States of America*,vol. 102, pp. 10557–10562, 2005.

[13] D.W. Mount, "Using Progressive Methods for Global Multiple Sequence Alignment," *Cold Spring Harb Protoc*, vol. 4, pp. 1-6, 2009.

[14] D. Lupyan, A.L. Macias, and A.R. Ortiz, "A new progressive-iterative algorithm for multiple structure alignment," *Bioinformatics*, vol. 21, pp. 3255-3263, 2005.

[15] J.D. Thompson, D.G. Higgins, and T.J. Gibson, "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", *Nucleic Acids Res*, vol. 22, pp.4673–4680, 1994.

[16] D.G. Higgins, J.D. Thompson, and T.J. Gibson, "Using CLUSTAL for multiple sequence alignments," *Methods Enzymol*, vol. 266, pp. 383–402, 1996.

[17] D.W. Mount, "Distance methods for phylogenetic prediction," *Cold Spring Harb Protoc*, vol.4, pp. 1-6, 2008.

[18] R.B Wallace., J. Shaffer, R.F.Murphy, J. Bonner, T. Hirose, and K. Itakura, "Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch," *Nucleic Acids Res*, vol.6, pp.3543-3557, 1979.

[19] C.W. Dieffenbach, T.M. Lowe, and G.S. Dveksler, "General concepts for PCR primer design," *Cold Spring Harb Protoc*, vol. 3, pp.s30–s37, 1993.