Menara Perkebunan 2022, 90(1), 40-50
p-ISSN: 0125-9318/ e-ISSN: 1858-3768
DOI: http://dx.doi.org/10.22302/iribb.jur.mp.v90i1.486
Accreditation Number: 164/E/KPT/2021

The in silico study of the COBRA gene family in sugarcane related to potential biomass content

Kajian in silico dari famili gen COBRA pada tanaman tebu yang terkait dengan potensi kandungan biomassa

Riza Arief PUTRANTO[1,2*)], Galuh Wening PERMATASARI[2)] & Rizka Tamania SAPTARI[2)]

[1)] PT Riset Perkebunan Nusantara, Jalan Salak No.1A, Bogor 16128
[2)] Pusat Penelitian Bioteknologi dan Bioindustri Indonesia, Jalan Taman Kencana No.1, Bogor 16128

Received 15 March 2022 / Accepted 20 April 2022

Abstrak

*Tanaman tebu (*Saccharum *sp.) berpotensi sebagai sumber bahan bakar nabati dan biomaterial karena kandungan selulosanya yang tinggi. Selulosa merupakan komponen utama penyusun dinding sel tanaman, sebagai rantai lurus yang tersusun dalam gugusan polisakarida, yang disebut mikrofibril selulosa. Sebuah gen bernama COBRA telah diketahui berperan dalam menentukan arah mikrofibril dan kristalisasi selulosa. Gen COBRA pada spesies* Saccharum *spp. belum banyak dipelajari. Oleh karena itu, kajian* in silico *dilakukan untuk mempelajari gen COBRA pada* Saccharum *sp. Melalui metode perbandingan genomika, gen COBRA dari* Arabidopsis *sp. (AtCOBLs) dibandingkan dengan gen COBRA dari* Saccharum *sp. (SoCOBLs). Domain conserve pada gen kemudian diidentifikasi dan sistem klaster disusun dalam sebuah pohon filogeni. Setelah itu, dibuat model untuk menganalisis struktur dari protein SoCOBL. Dari hasil analisis, sebelas perancah genom* Saccharum *sp. berhasil diidentifikasi. Kemudian, identifikasi daerah lestari menghasilkan sembilan protein SoCOBL. Pohon filogeni menggambarkan dua klaster utama: I dan II, yang membedakan famili SoCOBLs tersebut berdasarkan sekuens protein, motif domain, dan karakteristik asam amino. Karakteristik asam amino menyebabkan variasi pada struktur protein-protein SoCOBL. Secara umum, gen COBRA telah teridentifikasi pada* Saccharum *sp., meskipun fungsi dan ekspresi spesifiknya pada jaringan masih belum diketahui. Diperkirakan gen tersebut berperan sebagai pengatur arah mikrofibril dan proses sintesis selulosa. Oleh karena itu, perlu adanya analisis lebih lanjut pada level in vitro dan in vivo.*

[Kata kunci: selulosa, genomika komparatif, Saccharum *sp.]*

Abstract

Sugarcane (*Saccharum* sp.) is potential as a biofuel and biomaterial source for its high cellulose content. Cellulose is the main constituent of the plant cell wall, as a linear chain arranged in a polysaccharide bundle, called cellulose microfibril. A gene named COBRA has been revealed to play role in the orientation of microfibril and cellulose crystallization. The COBRA gene in the *Saccharum* sp. is under-explored. Therefore, the in silico study was conducted to explore the COBRA gene in *Saccharum* sp. By comparative genomics methods, the COBRA genes from *Arabidopsis* sp. (AtCOBLs) were compared to the *Saccharum* sp. (SoCOBLs). The conserved domain was then identified and the cluster system was constructed under a phylogenetic tree. Furthermore, each SoCOBLs protein was modelled to analyze its structure. According to the analysis, eleven of *Saccharum* sp. genomic scaffolds were successfully identified. Moreover, conserved domain identification resulted in nine SoCOBLs proteins. The phylogenetic tree showed two main clusters: I and II, differentiating those COBLs families based on the protein sequence, domain motif and amino acid properties. It leads to the variation of SoCOBLs protein structure as the results of the amino acid properties. Overall, the COBRA gene has been identified genomically in *Saccharum* sp. Yet, the function and tissue-specific expression are still unclear. It was predicted to act as the regulator of microfibril orientation and the cellulose synthesis process. Hence, further analyses by in vitro and in vivo are indispensable.

[Keywords: cellulose, comparative genomic, *Saccharum* sp.]

Corresponding author: rizaputranto@iribb.org

Introduction

Sugarcane (*Saccharum* sp.), a C4 plant, is potential as a material for biofuels and biomaterials through the abundant lignocellulosic biomasses (Kasirajan *et al*., 2018). The waste of sugarcane industry, called sugarcane bagasse (SCB) contains high cellulose (40 – 50 %) and hemicellulose (25 – 35 %) (Fan *et al*., 2018; Khoo *et al*., 2018). Cellulose is the most common natural fiber with high biocompatibility, high mechanical strength, and good thermal stability (Geng *et al*., 2014). It has diverse applications such as in the fuels, paper, and textile industries (Gupta *et al*., 2016), even in the drugs industries (Wsoo *et al*., 2020). Recent developments are towards the extraction of cellulosic fibers, pure cellulose, cellulose nanofibers, and cellulose nanocrystals from SCB which have diverse applications (Mahmud & Anannya, 2021).

In the plant cells, cellulose is a main component found in the primary and secondary cell walls (Thomas *et al.*, 2013; Meents *et al*., 2018). The structure of cellulose consists of the linear chain beta-1,4-D-glucan (Synytsya & Novak, 2014), building the strong, fibrous and non-soluble characteristics of cell wall. Also, it supports the stability of cell wall's structure suggesting that cellulose is a high strength biomaterial. The chain of cellulose arranged in a polysaccharide bundle called microfibril (Brigham 2018). The building block of cellulose structure is glucose. The process of glucose synthesis in cells is complex. It locates on the cell membrane. It involves the cellulose synthase enzyme to reach out the cell membrane. The glucose-UDP is one of the key-mediator, used by the cellulose synthase enzyme to transport the glucose across the cell membrane or cell wall (Endler *et al*., 2010; Zhang *et al.,* 2021).

A gene named COBRA coded glycosylphosphatidylinositol (GPI) protein, playing role in the orientation of microfibril and cellulose crystallization. COBRA-like genes (COBLs) consisted of signal peptide and cellulose binding motif (CBM) in the N terminus of the gene, and a short cysteine-rich (CCVS) motif and GPI-anchoring motif at the C-terminus (Roudier *et al*., 2002). The COBRA family have been identified in some plants such as *Arabidopsis thaliana* (Roudier *et al*., 2002), *Zea mays* (Brady *et al*., 2007), *Oryza sativa* (Dai *et al*., 2011), *Solanum lycopersicum* (Cao *et al*., 2012), *Gossypium raimondii* (Niu *et al*., 2015), *Hevea brasiliensis* (Putranto *et al*., 2017), and *Saccharum spp*. (Kasirajan *et al*., 2018).

The information related to the classification and structure of COBRA genes in *Saccharum* spp. is still lacking. Therefore, this study aimed to identify COBRA gene in *Saccharum* sp. by using genomic comparative approach to analyze the cluster classification, amino acid properties, as well as the protein structure prediction of COBRA in *Saccharum* sp. The genomic comparative study has been successfully identified the presence of putative Protease Inhibitor genes in *Hevea brasiliensis* (HbPI) (Martiansyah *et al.,* 2017) and putative *SWEET* genes in *Metroxylon sagu* (Putranto *et al.*, 2020). Therefore, the present study might provide initial information about COBRA genes structure in *Saccharum* sp. for further studies on its function.

Materials and Methods

Collection of genomic data and comparative genomics with Arabidopsis

The genome of sugarcane (R570) was obtained from Dr Angelique d'Hont (Head of Research Team: Structures an Evolution of Genome) (https://sugarcane-genome.cirad.fr/content/download) (Garsmeur *et al.,* 2018). Thereafter, twelve sequences of Arabidopsis (*Arabidopsis thaliana*) COBRA genes (*AtCOBLs*) encoding GPI-anchored proteins (Supplementary File 1) were collected from the TAIR database (https://www.arabidopsis.org/servlets/Search?type=general&search_action=detail&method=1&show_obsolete=F&name=COBRA&sub_type=gene&SEARCH_EXACT=4&SEARCH_CONTAINS=1). All genomic data were then analyzed using the Southern France Galaxy bioinformatics platform (http://galaxy.southern.fr/galaxy/). The Arabidopsis COBRA genes were used as reference to identify COBRA gene family in sugarcane (*SoCOBL*) by carried out an NCBI MEGA-BLAST + tbalstn of Arabidopsis COBRA genes against sugarcane genome, with the default value cutoff of 0.001 using BLOSUM62 scoring matrix. The results were then manually sorted based on ≥ 50% of sequence similarity with minimum length of 150 bp, and ≤ 6.79e-54 of e-value. The analysis parameter was chosen based on the sequence length and the availability of filtered sample.

Conserve domain analysis and annotation of sugarcane genomic scaffolds

The COBRA conserve domain analysis of the sugarcane genomic scaffolds was carried out based on method by Putranto *et al.* (2015) as described: each selection scaffold from the previous analysis containing *SoCOBL* was screened in NCBI Conserve Domain Database Search (CDD) (www.ncbi.nml.nih.gov/Structure/cdd/cdd.shtml) to find the conserve protein domain named "COBRA" with the accession number of pfam04833 and c104787. The length of COBRA domain was between 70 to 180 amino acids residues (Roudier *et al.*, 2002). Thereafter, the scaffolds were manually annotated using Geneious Prime software v2020.2.2 (Biomatters Ltd, USA), including the non-translated region (5' and 3'-UTR), coding region or sequence (CDS) of *SoCOBL*, and the COBRA protein domain. Geneious Prime software provides manual

annotation features and with the user-friendly interface and interactive visualizations to generate publication-quality images (Kearse *et al.*, 2012).

C*haracterization and classification of SoCOBL proteins*

Characterization of SoCOBL proteins was conducted in Geneious Prime software by performing multiple sequence alignment on the amino acid sequences of the COBRA domain from SoCOBL proteins, and predicting the physicochemical properties of the proteins, such as molecular weight, isoelectric point, molecular charge, and hydrophobicity. Afterwards, phylogenetic analysis was performed to classify the orthology and paralogy of SoCOBLs. Classification of SoCOBLs was performed in Geneious Prime software by performing multiple sequence alignment on 9 amino acid sequences of *SoCOBLs* and 11 gene sequences of *AtSOBLs*. The Neighbor-Joining algorithm with 1000 bootstrap replicates were used for reconstructing the phylogenetic trees. The bootstrap method is important for reliability of the phylogenetics tree construction. Bootstrap entails resampling with replacement from one's molecular data to generate fictional datasets of the same size, known as bootstrap replicates. A forest of B bootstrap trees is estimated using the B replicates (one per replicate). Finally, a branch of the original tree's bootstrap value (BP) represents its frequency of recurrence in the forest (Mariadassou *et al.*, 2019).

SoCOBL protein modelling

Homology protein modelling was carried out to confirm the phylogenetic characters of SoCOBLs, using Phyre2 bioinformatics webtools (http://www.sbg.bio.ic.ac.uk/phyre2). Phyre2 was used because of the user-friendly interface to predict and analyze protein structure. Other than that, the samples in Phyre2 are automatically run in a server and within 30 minutes to 2 hours after submission the structure prediction will be released (Kelley *et al.*, 2015). Amino acid sequences of SoCOBLs were annotated using Hhblits for gathering homologous sequences in the database. Hhblits were used because of its sensitivity and high-quality of multiple sequence alignment performance. The matches sequences were aligned towards pre-built HMM databases, obtained from protein sequence database (Remmert *et al.*, 2011). Afterwards, the prediction of secondary structure was carried out with PSIPRED to generate the crude backbone-only models. The PSIPRED software were used because of its capability to produce the highest accuracy up to 76.5% (Q3 score), and PSIPRED is the first rank software for secondary structure prediction methods (Orengo *et al.*, 1999). Analysis was then followed by loop modeling resulting the final 3D model of the protein. The overall workflow of the study is described in Figure 1.

Results and Discussion

Comparative genomics of sugarcane and Arabidopsis COBRA gene

The cellulose-rich bagasse from the waste of sugarcane industry is one of the most potential materials for producing biofuels or biomaterials in commercial quantities. COBRA gene family encoding a plant-specific glycosyl-phosphatidylinositol (GPI) anchored protein had been proven to be a key regulator in the cellulose crystallinity status in plant cells through binding cellulose microfibrils (Liu *et al.*, 2013). Transcriptome analysis reported the presence of COBRA-like protein in sugarcane (Kasirajan *et al.*, 2018). Therefore, identification and structural analysis of COBRA gene in sugarcane might give valuable information in understanding cellulose synthesis in this bioenergy crops.

The genomics sample from sugarcane R570 consists of 42.359 genes. Based on the result of tblastn against Arabidopsis COBRA genes, there were eleven of sugarcane genomic scaffolds potentially encoding GPI proteins, named *SoCOB, SoCOBL-A1, SoCOBL-A2, SoCOBL-B1, SoCOBL-B2, SoCOBL-B3, SoCOBL-C, SoCOBL-D, SoCOBL-E, SoCOBL-F* and *SoCOBL-G* (Table 1). Length of the scaffolds ranged from 970 – 3736 bp, both in partial or in full length. The Arabidopsis COBRA consisted of *AtCOB, AtCOBL-02, AtCOBL-04, AtCOBL-05, AtCOBL-06, AtCOBL-07, AtCOBL-08, AtCOBL-09,* and *AtCOBL-10*. Sequence similarity of SoCOBLs and AtCOBLs was 72%, covering 172 – 638 amino acid residues. SoCOBL-A1 and SoCOBL-A2 had 55% of similarity with AtCOBL-02. SoCOBL-B1, SoCOBL-B2, and SoCOBL-B3 had 65% of similarity with AtCOBL-04. SoCOBL-C had 52% of similarity with AtCOBL-05, whereas SoCOBL-D against AtCOBL-06, and SoCOBL-E against AtCOBL-07 had 51% of sequence similarities. SoCOBL-F had 52% and 50% of similarities against AtCOBL-08 and AtCOBL-09 respectively, whereas SoCOBL-G had 58% of similarity against AtCOBL-10.

Similarity searching is an effective and reliable strategy for identifying homologs. Inferring homology needs both sequence and structure similarities, and with reliable statistical estimates (Pearson, 2013). Alignment was conducted to find sequence similarity. According to Pearson (2013), two sequences are considerably homologous if they are more than 30% identical over the entire length. The 30% of similarity score in > 100 residues are mostly statistically significant. Based on the analysis, sequence similarity of SoCOBLs and AtCOBLs was 72%, covering 172 – 638 amino acid-

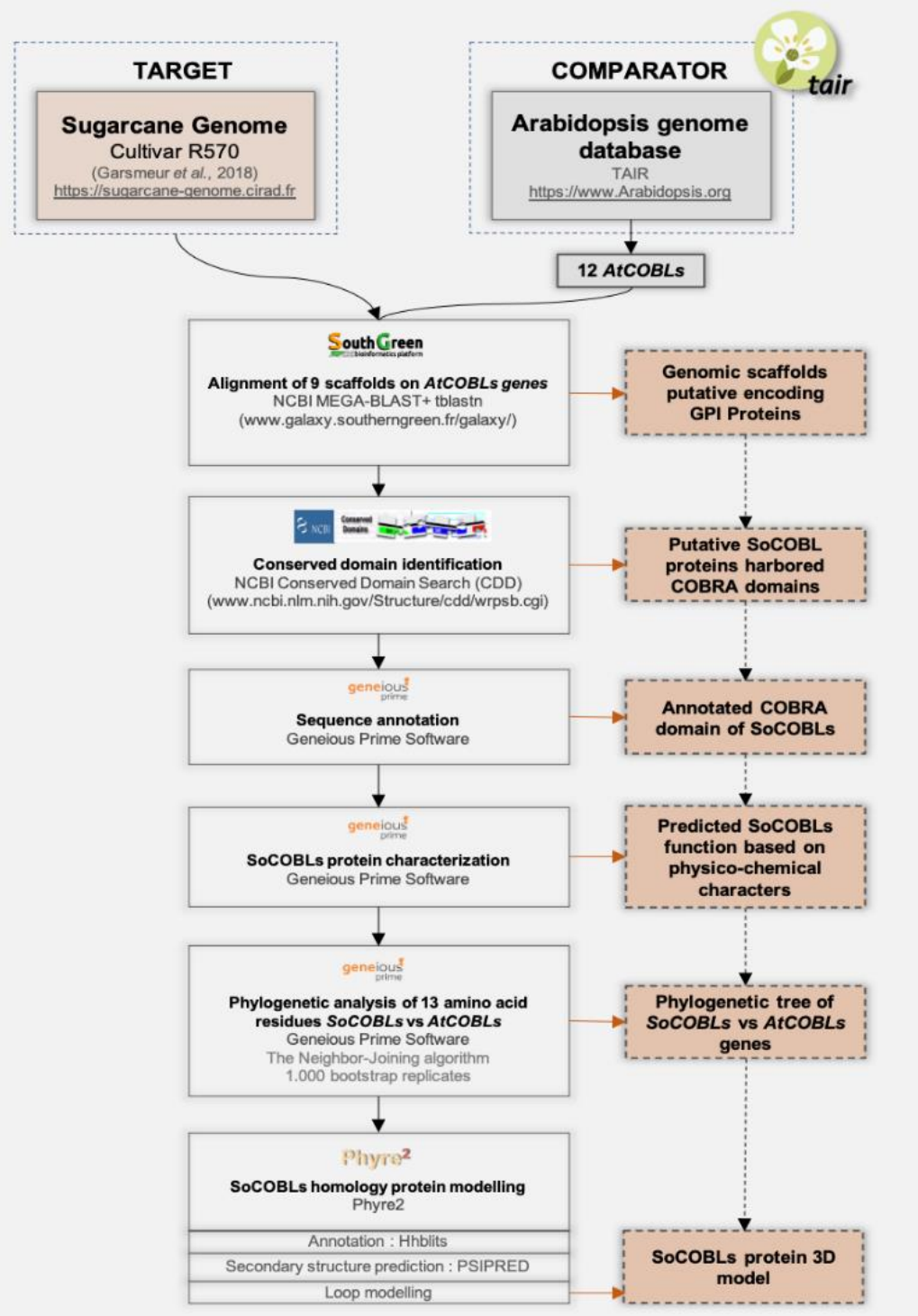

Figure 1. The workflow of genomic comparative study of sugarcane COBRA gene family

Gambar 1. Tahap kerja studi komparatif genomik famili gen COBRA pada tanaman tebu

residues so it has high possibility of homology. Moreover, according to Pearson (2013) the cut off value of 0.001 that we used in this study is reliable to infer homology in protein alignment.

Nevertheless, the high similarity score might not always reflect evolutionary relationship. Therefore, examining conserved domain of high-scoring alignments improved accuracy in inferring homology (Pearson, 2013). In this study, the eleven sequences of SoCOBL proteins were then analyzed for its conserved COBRA domain, resulting in nine SoCOBL proteins harbored COBRA domains ranged from 143 – 182 amino acid residues. Meanwhile, COBRA domain was not found in SoCOBL-D and SoCOBL-G.

Table 1. Identification of sugarcane putative genes encoding COBRA proteins using in silico comparative analysis. The matching of sugarcane and Arabidopsis COBRA genes were carried out using tblastn of Galaxy. The identification of protein domain was carried out using NCBI Conserved Domains Database (CDD) Search. All putative genes were annotated in Geneious.

Tabel 1. Identifikasi gen tebu yang berpotensi mengkode protein COBRA menggunakan analisis konparatif in silico. Pencocokan antara gen COBRA pada tebu dan Arabidopsis dilakukan dengan menggunakan tblastn dari Galaxy. Identifikasi domain protein dilakukan dengan menggunakan NCBI Conserved Domains Database (CDD) Search. Semua gen putatif dianotasi pada Geneious.

Gene name *Nama gen*	**Sugarcane R570 genome** *Genom tebu R570*		**Arabidopsis ID** *ID Arabidopsis*		**Galaxy tblastn*** *Galaxy tblastn*			**Gene structure** *Struktur gen*		**Protein sequence** *Sekuens protein*		**COBRA domain** *Domain COBRA*		
	Accession *Aksesi*	**Length (bp)** *Panjang (pasang basa)*	**Gene name** *Nama gen*	**Accession** *Aksesi*	**Identity (%)** *"Identity"*	**Length (aa)** *Panjang*	**E-value** Nilai-e	**Exon nb** *Ekson*	**Intron nb** *Intron*	**Annotated** *Teranotasi*	**Length (aa)** *Panjang*	**CDD Search** *Pencarian pada CDD*	**CDD Accession** *Aksesi CDD*	**Length (aa)** *Panjang*
SoCOB	Sh_247J17_g000060	3736	*AtCOB*	AT5G60920.1	72.093	172	2.99e-72	6	5	Full-length	467	YES	pfam04833	176
SoCOBL-A1	Sh_207G10_g000100	3969	*AtCOBL-02*	AT3G29810.1	55.000	380	3.68e-134	3	2	Partial	306	YES	pfam04833	154
SoCOBL-A2	Sh_207H02_g000160	3875			55.000	380	2.19e-134	3	2	Partial	306	YES	pfam04833	154
SoCOBL-B1	Sh_207G10_g000130	2201			65.502	229	5.34e-102	4	3	Partial	454	YES	pfam04833	172
SoCOBL-B2	Sh_207H02_g000150	1907	*AtCOBL-04*	AT5G15630.1	65.502	229	5.34e-102	1	0	Full-length	166	YES	pfam04833	143
SoCOBL-B3	Sh_213P05_g000100	2508			65.066	229	6.25e-102	3	2	Partial	364	YES	pfam04833	163
SoCOBL-C	Sh_247J17_g000020	3125	*AtCOBL-05*	AT5G60950.1	52.147	163	6.65e-47	4	3	Full-length	416	YES	pfam04833	166
SoCOBL-D	Sh_237O23_g000050	970	*AtCOBL-06*	AT1G09790.1	51.807	166	6.79e-54	2	1	Partial	447	NO	-	-
SoCOBL-E	Sh_230C02_g000070	2106	*AtCOBL-07*	AT4G16120.1	51.236	607	00.00	1	0	Full-length	701	YES	pfam04833	182
SoCOBL-F	Sh_235O12_g000060	3878	*AtCOBL-08*	AT3G16860.1	52.188	617	00.00	1	0	Full-length	669	YES	pfam04833	182
			AtCOBL-09	AT5G49270.1	50.470	638	00.00							
SoCOBL-G	Sh_227E04_g000040	2161	*AtCOBL-10*	AT3G20580.1	58.774	473	00.00	1	0	Full-length	681	NO	-	-

Properties of SoCOBLs based on the COBRA domain

Characterization of the nine SoCOBLs was carried out based on its COBRA domain by performing sequence alignment. The *in silico* analysis showed an average molecular weight of COBRA domain was 18.15 kDa with length of 196 amino acids (Table 2). Sequence alignment was performed on the nine COBRA domains to identify the differences of each SoCOBL proteins based on its COBRA domain (Figure 2). The COBRA domains (196 amino acids) from nine SoCOBL proteins had 54.4% of similarity, indicating the possibility of the nine SoCOBL proteins having similar function. Moreover, 49.9% of the SoCOBLs sequences (745 amino acids) are hydrophobic, 35% are polar uncharged, and 17% are charged (Table 2), indicating potential function of SoCOBL as a transmembrane protein.

GPI-anchored proteins (GPI-APs) encoded by COBRA has known to be incorporated into the cell wall, thus facilitate crystallization of cellulose microfibrils, regulating orientation of cell expansion (Schindelman *et al*., 2001). GPI-APs are a class of membrane proteins containing a soluble protein attached by a conserved GPI anchor (Zurzolo and Simons, 2016). GPI-APs are associated with membrane rafts, the micro domains enriched in sphingolipids and cholesterol (Borner *et al*., 2003). The C-terminus of all GPI-APs contains hydrophobic signal sequence that triggers the addition of the GPI anchor, while the N-terminus might comprise hydrophobic signal peptides (Takahashi *et al*., 2016; Zhou, 2019) PI-APs possess no transmembrane domain. However, analysis of SoCOBLs protein properties revealed high hydrophobicity of the COBRA domain (Table 2). The high percentage of hydrophobic amino acids in SoCOBLs protein could be as the N-terminus and C-terminus hydrophobic signal peptides, or those might be transmembrane proteins, similar to other discounted GPI-APs, PIN3, PIN4, and RLKs (Zhou, 2019). These analyses supporting the hypothesis of putative SoCOBLs are probably as the component of sugarcane cell wall, thus involved in deposition of cell wall cellulose.

Table 2. The properties of COBRA domain of SoCOB proteins. The calculations were carried out using Geneious. The data shown were mean values of 9 COBRA domains from each of SoCOB proteins.

Tabel 2. Karakteristik domain COBRA dari protein-protein SoCOB. Perhitungan dilakukan menggunakan Geneious. Data yang ditampilkan merupakan nilai rerata dari 9 domain COBRA dari setiap protein SoCOB.

COBRA Domain Statistics ***Statistik Domain COBRA***		
Length (amino acids) *Panjang (asam amino)*	196	
Identical sites *Situs identik*	37 (18.9 %)	
Pairwise identity *Identitas pencocokan*	54.40 %	
Pairwise positive (BLSM62) *Pencocokan positif*	64.80 %	
COBRA Domain Properties ***Karakteristik Domain COBRA***		
Molecular weight (kDa) *Berat molekul (kDa)*	18.152	
Isoelectric point *Titik isoelektrik*	9.02	
Extinction coefficient *Koefisien ekstingsi*	35.593	
Amino Acids Group ***Kelompok Asam Amino***	**Number** ***Jumlah***	**%**
Acidic *Asam*	83	5.6
Basic *Basa*	173	11.6
Charged *Bermuatan*	256	17.2
Polar uncharged *Polar tidak bermuatan*	534	35.8
Hydrophobic *Hidrofobik*	745	49.9
GC-rich *Kaya GC*	426	28.6
AT-rich *Kaya AT*	322	21,6

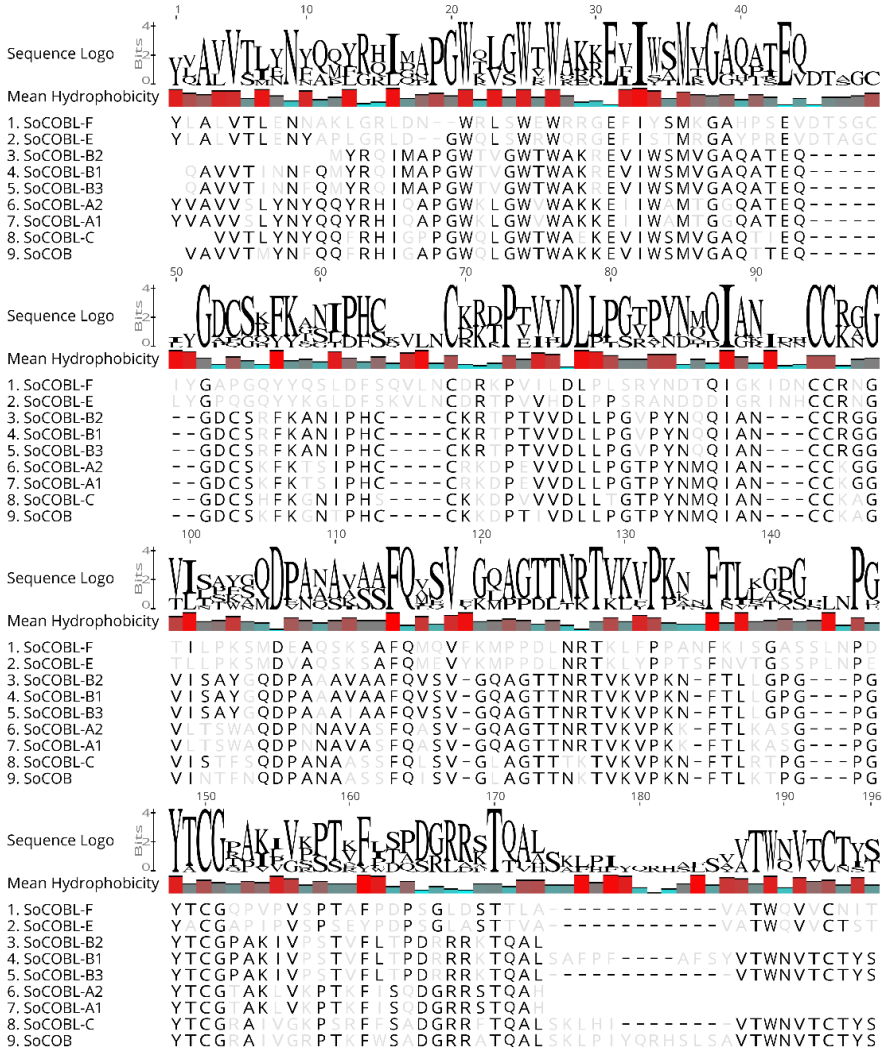

Figure 2. Multiple sequence alignment of COBRA domain from 9 SoCOB proteins. The representation of amino acid dominance was shown in capital logo. The mean hydrophobicity was predicted using Geneious calculations.

Gambar 2. Penjajaran beberapa sekuens domain COBRA dari 9 protein SoCOB. Asam amino yang muncul secara dominan ditunjukkan dengan huruf kapital. Rerata nilai hidrofobisitas diprediksi menggunakan perhitungan Geneious.

Phylogenic trees of SoCOBLs

Construction of phylogenetic trees of SoCOBL and AtCOBL showed two clusters, named cluster I and cluster II. Cluster I consisted of SoCOBL-F and SoCOBL-E, also grouped with AtCOBL7, AtCOBL8, AtCOBL9, AtCOBL10 and AtCOBL11. By contrast, cluster II consisted of SoCOB, SoCOBL-C, SoCOBL-A2, SoCOBL-A1, SoCOBL-B2, SoCOBL-B3, SoCOBL-B1, also grouped with AtCOB, AtCOBL1, AtCOBL2, AtCOBL4, AtCOBL5, AtCOBL6 (Figure 3). The automatic domain annotation illustrated SoCOBL-E and SoCOBL-F were characterized by long exon in the N-terminal. By contrast, domain motifs in cluster II varied, with only SoCOBL-A1 and SoCOBL-A2 showed similar domain motifs. Nevertheless, SoCOBL-A1, SoCOBL-A2, and SoCOBL-B2 showed the addition of 24 residues in C-terminal at 174 – 197 bp. Moreover, SoCOBL-A1, SoCOBL-A2, SoCOB, and SoCOBL-C were marked by short exon in the N-terminal, compared to SoCOBL-B1, SoCOBL-B2, and SoCOBL-B3 in the same cluster.

Phylogenetic analysis of the genes revealed two clusters of *SoCOBs*, agreed with those COBRA family in *A. thaliana* (Roudier *et al*., 2002), *O. sativa* (Dai *et al*., 2011), *Z. mays* (Brady *et al*., 2007), *S. lycopersicum* (Cao *et al*., 2012), *G. raimondii* (Niu *et al*., 2015), and *H. brasiliensis* (Putranto *et al*., 2017). Moreover, grouped in the same cluster, *SoCOBL-B2* putatively ortholog to *AtCOB*, suggesting the similar role in regulating cellulose microfibrils orientation (Roudier *et al*., 2005). However, domain motif of *SoCOB, SoCOBL-B1, SoCOBL-B2*, and *SoCOBL-B3* were vary.

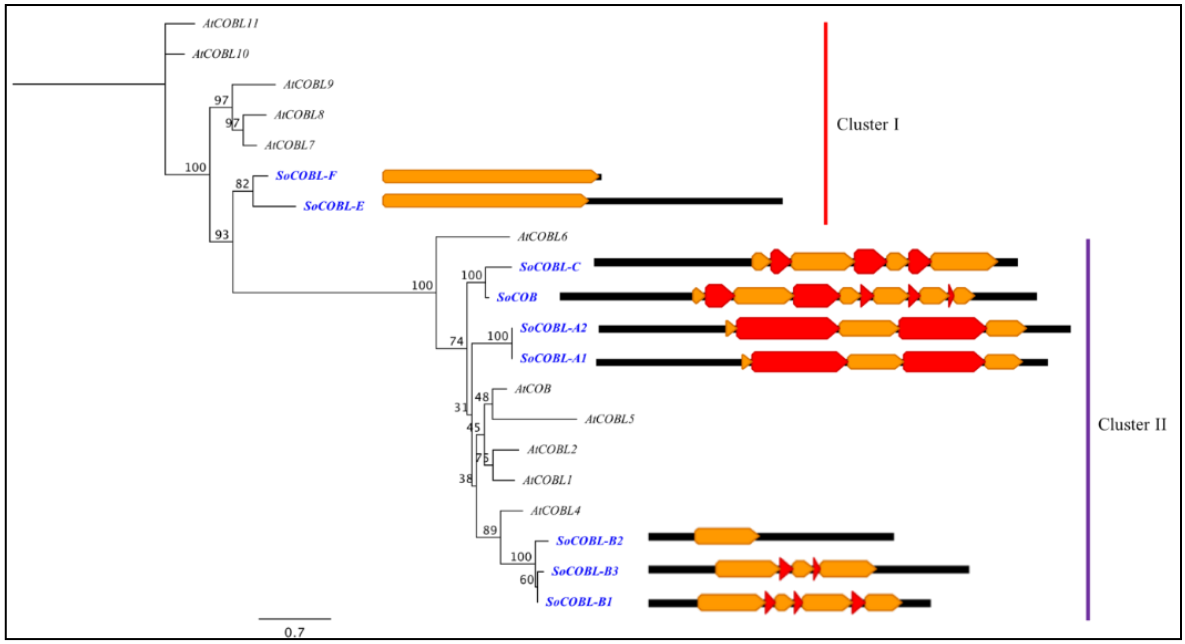

Figure 3. Phyogenetic analysis of 11 and 13 COBRA amino acids from *A. thaliana* and *S. officinarum*. The evolutiorany history was inferred using PhyML tree builder in SouthGreen Galaxy. Tree topology was constructed using NNIs followed by building initial tree using BioNJ. Model for amino acids subsitution was performed using WAG with 4 tree subtitution rate categories. Bootstrap was performed 100 times for branch support. In silico prediction of gene structure for each of 9 SoCOBRA genes was shown in colored bars. The automatic annotation followed by manual validation was carried out in Geneious. The black bar represents the DNA sequence length. The orange bar represents the exon(s) while the red bar represents the intron(s).

Gambar 3. Analisis filogenetik dari 11 dan 13 asam amino COBRA dari A. thaliana *dan* S. officinarum. *Sejarah evolusi disimpulkan menggunakan* PhyML tree builder *pada SouthGreen Galaxy. Topologi pohon filogenetik dikonstruksi menggunakan NNIs dilanjutkan dengan pembuatan pohon inisial menggunakan BioNJ. Pembuatan model dari substitusi asam amino dilakukan menggunakan WAG dengan 4 kategori substitusi.* Bootstrap *dilakukan sebanyak 100 kali untuk mendukung pembuatan cabang pohon filogenetik. Prediksi* in silico *struktur dari setiap gen SoCOBRA ditunjukkan dengan diagram batang berwarna. Anotasi secara otomatis dan dilanjutkan dengan validasi manual dilakukan pada Geneious. Diagram batang berwarna hitam menunjukkan panjang sekuens DNA. Diagram batang oranye menunjukkan ekson sedangkan diagram batang merah menunjukkan intron.*

In Arabidopsis, it has known that COBRA genes were mainly functions in regulating crystallization or deposition of cellulose microfibrils during cell expansion. They are redundant but expressed in different tissues or plant developmental stages. COBRA has known to be expressed during root development while COBL6 and COBL9 were expressed during flower development (Roudier *et al*., 2002). These temporal and tissue specializations may cause variation of domain motifs, although still has similar function. According to the sequence alignment, *AtCOB* was similar with *SoCOB* whereas *AtCOBL9* was similar with *SoCOBL-F*. Meanwhile, the phylogenetic trees grouped *SoCOB* and *SoCOBL-F* in different cluster (Figure 3), indicating possibility of specialization of both genes.

Protein model of SoCOBLs

The phylogenic character of SoCOBLs was confirmed by the homology protein modelling. Figure 4 illustrated the predicted protein structure of the nine SoCOBLs. Referring to the protein model, SoCOB, SoCOBL-A1, and SoCOBL-A2 had similar protein structure, consisted of a beta turn and two helix structures. This result agrees with the phylogenetic analysis determining the three proteins were in the same cluster. There was no secondary structure predicted in SoCOBL-B3. Furthermore, SoCOBL-E was predicted to have similar protein structure with SoCOBL-F, consisted of a beta turn and a helix structure with also nearly identical folding, conforming the phylogenetic trees, also suggesting possible redundancy.

Altogether, the presence of COBRA gene in sugarcane supporting sugarcane as a potential source of biofuel, due to its high carbohydrate conversion and biomass accumulation. COBRA encoding GPI-anchored proteins regulates deposition of microfibrils cellulose during cell wall expansion (Schindelman *et al*., 2001). Hypothetically, it might be directly related to the length of sugarcane segments which determine the quantity of stored sugar and accumulated biomass.

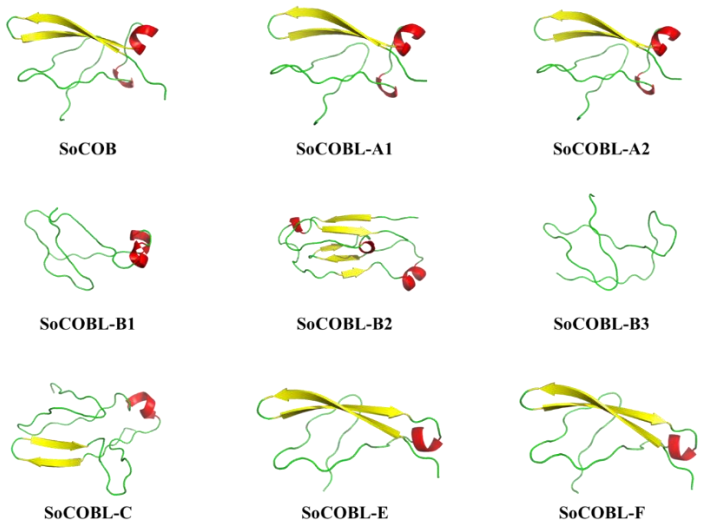

Figure 4. In silico models of 9 SoCOB proteins. The protein modeling was carried out using Phyre2 protocol for each of the SoCOB protein. The original SoCOB amino acid sequences were scanned using Hhblits followed by a PSIPRED analysis using a query of Hidden Markov model and fold library scanning in HMMs database. The final protein model was built using loop modeling. The green color showed a loop. The yellow color showed a beta sheet while the red color showed an alpha helix.

Gambar 4. Model in silico *dari 9 protein SoCOB. Pembuatan model setiap protein SoCOB dilakukan menggunakan protokol dari Phyre2. Sekuens orisinil asam amino SoCOB dicari menggunakan analisis Hhblits dilanjutkan dengan analisis PSIPRED menggunakan* query Hidden Markov model *dan pencarian pustaka pelipatan protein di database HMMs. Model akhir protein dibangun menggunakan* loop modeling. *Warna hijau menunjukkan* loop. *Warna kuning menunjukkan lembaran beta sedangkan warna merah menunjukkan heliks alfa*.

Conclusion

Comparative genomics studies successfully identified the presence of COBRA gene family in sugarcane. The sugarcane COBRA family consisted of two main clusters, cluster I (SoCOBL-F and SoCOBL-E), and cluster II (SoCOB, SoCOBL-C, SoCOBL-A2, SoCOBL-A1, SoCOBL-B2, SoCOBL-B3, and SoCOBL-B1). The exact function and tissue specific expression of sugarcane COBRA genes are still unknown. However, it was predicted to have the same role and redundant with the Arabidopsis COBRA genes, which acts as regulator for microfibril orientation in the synthesis of cellulose. This assumption might be addressed in future studies by using the *in vitro* approach.

Acknowledgements

The author thanks Dr Angelique d'Hont (Head of Research Team: Structures an Evolution of Genome) for providing the sugarcane COBRA sequences.

References

Borner GHH, KS Lilley, TJ Stevens & P Dupree (2003). Identification of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A proteomic and genomic analysis. *Plant Physiol* 132(2), 568–577.

Brady SM, S Song, KS Dhugga, JA Rafalski & PN Benfey (2007). Combining expression and comparative evolutionary analysis the COBRA gene family. *Plant Physiol* 143(1), 172.

Brigham C (2018). Biopolymers: Biodegradable alternatives to traditional plastics. *In*: *Green chemistry: An inclusive approach*. Elsevier Inc. p. 753–770.

Cao Y, X Tang, J Giovannoni, F Xiao & Y Liu (2012). Functional characterization of a tomato COBRA-like gene functioning in fruit development and ripening. *BMC Plant Biol* 12(1), 211.

Dai X, C You, G Chen, X Li, Q Zhang & C Wu (2011). OsBC1L4 encodes a COBRA-like protein that affects cellulose synthesis in rice. *Plant Molec Biol* 75(4–5), 333–345.

Endler A, C Sánchez-Rodríguez & S Persson (2010). Cellulose squeezes through. *Nat Chem Biol* 6(12), 883–884.

Fan M, A Zhang, G Ye, H Zhang & J Xie (2018). Integrating sugarcane molasses into sequential cellulosic biofuel production based on SSF process of high solid loading. *Biotechnol Biofuels* 11, 329.

Garsmeur O, G Droc, R Antonise, J Grimwood, B Potier, K Aitken, J Jenkins, G Martin, C

Charron, C Hervouet, L Costet, et al & A D'Hont (2018). A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nat Com* 9, 2638.

Geng H, Z Yuan, Q Fan, X Dai, Y Zhao, Z Wang & M Qin (2014). Characterisation of cellulose films regenerated from acetone/water coagulants. *Carbohydr Polym* 102, 438 – 444.

Gupta VK, PJM Carrott, R Singh & M Chaudhary (2016). Cellulose: a review as natural, modified and activated carbon adsorbent. *Bioresour Technol* 216, 1066 – 1076.

Kasirajan L, N Hoang, A Furtado, FC Botha & RJ Henry (2018). Transcriptome analysis highlights key differentially expressed genes involved in cellulose and lignin biosynthesis of sugarcane genotypes varying in fiber content. *Sci Rep* 8(1), 11612.

Kearse M, R Moir, A Wilson, S Stones-Havas, M Cheung, S Sturrock, S Buxton, A Cooper, S Markowitz, C Duran, T Thierer, B Ashton, P Mentjes & A Drummond (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2(12), 1647 – 1649.

Kelley LA, S Mezulis, CM Yates, MN Wass & MJE Sternberg (2015). The Phyre2 web portal for protein, prediction and analysis. *Nat Protoc* 10, 845 – 858.

Khoo RZ, WS Chow & H Ismail (2018). Sugarcane bagasse fiber and its cellulose nanocrystals for polymer reinforcement and heavy metal adsorbent: a review. *Cellulose* 25, 4303 – 4330.

Liu L, K Shang-Guan, B Zhang, X Liu, M Yan, L Zhang, Y Shi, M Zhang, Q Qian & J Li (2013). Brittle Culm1, a COBRA-Like protein, functions in cellulose assembly through binding cellulose microfibrils. PLoS Genet 9(8), e1003704.

Mahmud MA & FR Anannya (2021). Sugarcane bagasse – A source of cellulosic fiber for diverse applications. *Heliyon* 7(8), e07771.

Mariadassou M, A Bar-Hen & H Kishino (2019). Tree evaluation and robustness testing. *Encyclopedia of bioinformatics and computational biology.* Academmic Press, 736 – 745.

Martiansyah I, RA Putranto & N Khumaida (2017). Identifikasi famili gen putatif penyandi protease inhibitor dengan pendekatan *in silico* komparatif pada genom *Hevea brasiliensis* Muell. Arg. *Menara Perkebunan* 85(2), 53 – 66.

Meents MJ, Y Watanabe & AL Samuels. The cell biology of secondary cell wall biosynthesis. *Ann Bot* 121(6), 1107 – 1125.

Niu E, X Shang, C Cheng, J Bao, Y Zeng, C Cai, X Du & W Guo (2015). Comprehensive analysis of the COBRA-like (COBL) gene family in Gossypium identifies two COBLs potentially associated with fiber quality. *PLoS ONE* 10(12), e0145725.

Orengo CA, JE Bray, T Hubbard, L LoConte & I Sillitoe (1999). Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins* Suppl 3, 149 – 170.

Pearson WR (2013). An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics* 42(1), 3.1.1 – 3.1.8.

Putranto RA, C Duan, Kuswanhadi, T Chaidamsari, M Rio, P Piyatrakul, E Herlinawati, J Pirrello, F Dessailly, & J Leclercq (2015). Ethylene response factors are controlled by multiple harvesting stresses in *Hevea brasiliensis*. *PLoS ONE* 10(4), e0123618.

Putranto RA, I Martiansyah & RT Saptari (2017). In silico identification and comparative analysis of *Hevea brasiliensis* COBRA gene family. *Proceeding International Conference on Science and Engineering* 1, 39–47.

Putranto RA, I Martiansyah & DA Sari (2020). In silico identification of three putative SWEET genes in *Metroxylon sagu. IOP Conf Ser: Earth Environ Sci* 482, 012026.

Remmert M, A Biegert, A Hauser & J Soding (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9, 173 – 175.

Roudier F, AG Fernandez, M Fujita, R Himmelspach, GHH Borner, G Schindelman, S Song, TI Baskin, P Dupree & GO Wasteneys (2005). COBRA, an Arabidopsis extracellular glycosyl-phosphatidyl inositol-anchored protein, specifically controls highly anisotropic expansion through its involvement in cellulose microfibril orientation. *Plant Cell* 17(6), 1749 – 1763.

Roudier F, G Schindelman, R DeSalle, PN Benfey (2002). The COBRA family of putative GPI-anchored proteins in Arabidopsis. A new fellowship in expansion. *Plant Physiol*. 130(2), 538–548.

Schindelman G, A Morikami, J Jung, TI Baskin, NC Carpita, P Derbyshire, MC McCann & PN

Benfey (2001). COBRA encodes a putative GPI-anchored protein, which is polarly localized and necessary for oriented cell expansion in arabidopsis. *Genes Dev* 15(9), 1115–1127.

Synytsya A & M Novak (2014). Structural analysis of glucans. *Ann Transl Med* 2(2), 17.

Takahashi D, Y Kawamura, M Uemura (2016). Cold acclimation is accompanied by complex responses of glycosylphosphatidylinositol (GPI)-anchored proteins in Arabidopsis. *J Exp Bot* 67(17), 5203–5215.

Thomas LH, VT Forsyth, A Sturcova, CJ Kennedy, RP May, CM Altaner, DC Apperley, TJ Wess & MC Jarvis (2013). *Plant Physiol* 161(1), 465 – 476.

Wsoo MA, S Shahir, SPM Bohari, NHM Nayan & SIA Razak (2020). A review on the properties of electrospun cellulose acetate and its application in drug delivery systems: A new prespective. *Carbohydr Res* 491, 107978.

Zhang W, W Qin, H Li & A Wu (2021). Biosynthesis and transport of nucleotide sugars for plant hemicellulose. *Front Plant Sci* 12, 723128.

Zhou K (2019). Glycosylphosphatidylinositol-anchored proteins in Arabidopsis and one of their common roles in signaling transduction. *Front Plant Sci* 10, 1022.

Zurzolo C & K Simons (2016). Glycosylphosphatidylinositol-anchored proteins: Membrane organization and transport. *Biomembranes* 1858(4), 632 – 639.