In silico identification of three putative *SWEET* genes in *Metroxylon sagu*

R A Putranto*, I Martiansyah and D A Sari

Indonesian Research Institute for Biotechnology and Bioindustry, Jalan Taman Kencana No. 1, Bogor 16128, West Java, Indonesia

*E-mail: rizaputranto@iribb.org

Abstract. Recent studies have identified Sugars Will Eventually be Exported Transporters (SWEET), a novel type of sugar transporters in diverse plant species. This gene family selectively transports different kinds of sugar substrates, including sucrose, fructose and glucose. In this paper, three *Metroxylon sagu SWEET* genes (*MsSWEET-X*, *MsSWEET-Y* and *MsSWEET-Z*), predicted to be involved in starch accumulation, were identified from the NCBI EST database. A comparative analysis was carried out against *Arabidopsis thaliana* TAIR and *Elaeis guineensis* NCBI genome databases resulting in amino acid residues similarity of three *MsSWEET* genes of 21.32 to 76.25 %. One full-length coding sequence (CDS) of 229 amino acids from *MsSWEET-X* has been annotated as opposed to the partial CDSs from the other two. Three types of putative protein domains (Calreticulin, Glycosyl hydrolases and Triose-phosphate transporter) were predicted for *MsSWEET-X*, *MsSWEET-Y* and *MsSWEET-Z*, respectively. Multiple Alignment sing Fast Fourier Transform (MAFFT) has identified three conserved amino acid motifs (Motif-A, Motif-B and Motif-C) among three compared species. Phylogenetic analysis using Maximum-Likelihood Estimation has revealed two genes *AtCRT3* and *MsSWEET-X* at the upstream of initial tree branches (0.17 and 0.12 length) showing their early evolutionary orthology. By contrast, *MsSWEET-Y* gene was predicted to be the latest homolog of *SWEET16* and *SWEET17* undergoing speciation events from both *Arabidopsis* and oil palm. Taken together, these results showed that even though the oil palm and sago palm shared the common ancestry of monocotyledonous family, their *SWEET* genes were divergent. The gene *MsSWEET-X* was highly close to its homolog in *Arabidopsis*.

Keywords: *in silico*, MAFFT, maximum-likelihood estimation, *Metroxylon sagu*, phylogenetic analysis, *SWEET*.

1. Introduction

The Sugars Will Eventually be Exported Transporters (SWEET), a novel type of sugar transporters have been identified in diverse plant species, including thale cress, tomato, wheat, rubber tree and oil palm [1–4]. Various copy numbers were found ranging from 15 to 29 *SWEET* genes for each plant species. This gene family selectively transports different kinds of sugar substrates, including sucrose, fructose and glucose across the cell membrane regulating physiological aspects in plant species [2,5]. SWEET proteins were generally encoded by seven helices transmembrane domains consisting of a typical tandem repeat of three transmembrane domains, *SWEET*-type TMD helices [5,6].

Considered as a potential staple crop due to its high starch content, sago palm (*Metroxylon sagu*) incorporates the mechanism of sugar transport resulting in starch storage that has not yet been elucidated. The *M. sagu SWEET* gene family is potentially one of the gene families which have an indispensable role in the process of sugar storage. In addition, the genome information for this species is equally limited. An Expressed Sequence Tags (ESTs) library from the leaves of sago palm was published [7]. The database showed that the majority of 372 transcripts was potentially contributed in primary metabolism such as sugar biosynthesis and storage as well as stress tolerance.

This paper aimed to *in silico* identify and characterize the *M. sagu* SWEET gene family by performing comparative analysis from ESTs database against *A. thaliana* and *E. guineensis* genomic databases. Multiple sequence comparisons using MAFFT and BLAST using public platform Galaxy were respectively carried out to identify orthologous sequences between those three species. Phylogenetic analysis was constructed using PhyML to group evolutionary relationship of *SWEET* genes. TMHMM analysis was performed to predict the capability of transmembrane activity for SWEET protein.

2. Materials and methods

2.1. In silico comparative analysis of the SWEET gene family in sago palm, thale cress and oil palm

The *M. sagu* ESTs with accession numbers of JK731189-JK731342 and JK731189-JK731600 were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/). The ESTs consist of 372 tentative unique genes (TUGs) sequences [7]. The ESTs were annotated and stored in the Southern France Galaxy bioinformatics platform (http://galaxy.-southgreen.fr/galaxy/). Thirty-three nucleotide sequences and protein residues of *Arabidopsis thaliana SWEET* (*AtSWEET*) gene family were downloaded from PLAZA 4.0 Dicots Database (https://bioinformatics.psb.ugent.be/ plaza/versions/plaza_v4_-dicots/). In addition, nineteen nucleotide sequences and protein residues of *Elaeis guineensis SWEET* (*EgSWEET*) gene family were downloaded from PLAZA 4.0 Monocots Database (https://bioinformatics.psb.ugent.be/plaza/-versions/plaza_v4_monocots/).

An NCBI MEGA-BLAST + tblastn of *AtSWEETs* and *EgSWEETs* against the *M. sagu* ESTs were carried out at the Galaxy platform using the modified method of Piyatrakul et al. [8] and Putranto et al. [9]. The expectation value cutoff was set to 0.001 using the scoring matrix of BLOSUM62. The results of tblastn were sorted to ensure no duplicated genes noted. The selection was based on the parameters consisting of: (1) hits sharing >50% of sequence similarity with minimum 150 bp length, (2) one unique gene for each location in the EST, (3) one EST hosting more than one gene in the different location, and (4) low E-value.

2.2. Manual annotation and conserved domain analysis of M. sagu ESTs

Manual annotations of *M. sagu* genes were carried out on selected EST encoding potential SWEET domains using Geneious v5.3.6 (Biomatters Ltd, USA). The annotation included the non-translated region (5'- and 3'-UTR), coding sequences (CDS) and protein or domain motifs [10,11]. The putative CDS region for each *M. sagu SWEET* genes was verified using BLASTn in the NCBI (https:// blast.ncbi.nlm.nih.gov/Blast.cgi). The analysis of the conserved domain in the CDS region followed the protocol of Putranto et al. [9]. Each EST containing *SWEET* transcript was screened using the NCBI Conserved Domain Database Search (CDD) (www. ncbi.nlm.nih.gov/-Structure/cdd/cdd.shtml). Multiple sequence alignments between MsSWEET proteins were carried out using Fast Fourier Transform (MAFFT).

2.3. Phylogenetic analysis of MsSWEET against two species models and transmembrane domain analysis

Phylogenetic analysis was performed to determine the orthology and paralogy of a putative *MsSWEET* gene family. The comparative analysis was done on MsSWEET protein residues against 23 AtSWEET and 19 EgSWEET full-length protein residues. The phylogenetic tree was assembled using the BioNJ in the PhyML online software (http://www.atgc-montpellier.fr/phyml) [12]. Model of amino acid substitution used WAG. Bootstrap for the consensus tree was made as many as 6,506 replicates.

The transmembrane domain encoded by MsSWEET proteins was predicted using online server analysis TMHMM V2.0 (www.cbs.dtu.dk/services/TMHMM/). The method used a hidden Markov model (HMM) approach to model various regions of a model protein [13].

3. Results and discussion

3.1. In silico comparative analysis of MsSWEET against AtSWEET and EgSWEET gene families

The identification of several members of *MsSWEET* genes has been successfully performed using *in silico* comparative analysis on the EST database of Wee and Roslan [7] against two databases of model species. Three putative transcripts encoding SWEET proteins (MsSWEET-X, MsSWEET-Y and MsSWEET-Z) were identified from the EST accession numbers JK31566.1, JK31234.1 and JK731407.1, respectively (Table 1). The length of the transcripts ranged from 999 to 1071 bp. The sequence's similarity of *MsSWEET* genes against *AtSWEET* and *EgSWEET* genes were ranged from 21.32 to 76.25% covering 232 to 921 amino acid residues. One full-length coding sequence (CDS) of 229 amino acids from *MsSWEET-X* has been annotated as opposed to the partial CDS from the other two. The partial protein sequences were 57 and 136 amino acid residues. Three types of putative protein domains (Calreticulin, Glycosyl hydrolases and Triose-phosphate transporter) were predicted for *MsSWEET-X*, *MsSWEET-Y* and *MsSWEET-Z*, respectively. The length of the amino acid domains were 141, 57 and 129 bp.

Table 1. Identification of three putative MsSWEET genes using the *in silico* comparative analysis.

Gene name	Wee's ESTs			Galaxy tblastn				Protein sequence		Domain protein		
	Seq ID	Length (bp)	Species	Ref ID	Similarity (%)	Length (aa)	E-value	CDS	Length (aa)	Name	Accession	Length (aa)
MsSWEET-X	JKT731566.1	1,071	*Ath*	AT1G08450.1	76.25	424	1.66E-84	Full-length	229	Calreticulin	cl02828	141
MsSWEET-Y	JKT731234.1	999	*Ath*	AT5G63840.1	47.368	921	8.75E-12	Partial	57	GH31	cl25582	57
MsSWEET-Z	JKT731407.1	1,055	*Egu*	XP_010922652.1	21.324	232	0.05	Partial	136	TPT	cl26744	129

3.2. Phylogenetic analysis of MsSWEET with AtSWEET and EgSWEET protein residues

The phylogenetic analysis of MsSWEET with AtSWEET and EgSWEET protein residues were carried out. Distinctive evolutionary developments of three putative MsSWEET proteins have been identified (Figure 1). The evolutionary relationship starts with AtCRT3 and MsSWEET-X at the upstream of the *SWEET* gene family with 0.17 and 0.12 value of genetic distance. Furthermore, the MsSWEET-Z had a close orthologous relationship with AtVAL1 with a branch value of 0.11. In general, the *AtSWEET* and *EgSWEET* gene family shared close homolog structures in which both of these families were found in the massive cluster. The protein MsSWEET-Y was relatively close with AtSWEET16 (a branch value of 0.38) and AtSWEET17 (a branch value of 0.56). Regarding the phylogenetic tree structure, of the three compared species, the protein MsSWEET-Y was the latest developed in the *SWEET* gene family resulted from the previous speciation of *A. thaliana*. This result confirmed a potential distinctive function of MsSWEET-Y compared with the other MsSWEET proteins.

The matching of *M. sagu*, *A. thaliana* and *E. guineensis SWEET* genes were carried out using tblastn of SouthGreen Galaxy. The identification of protein domain was carried out using the NCBI CDD.

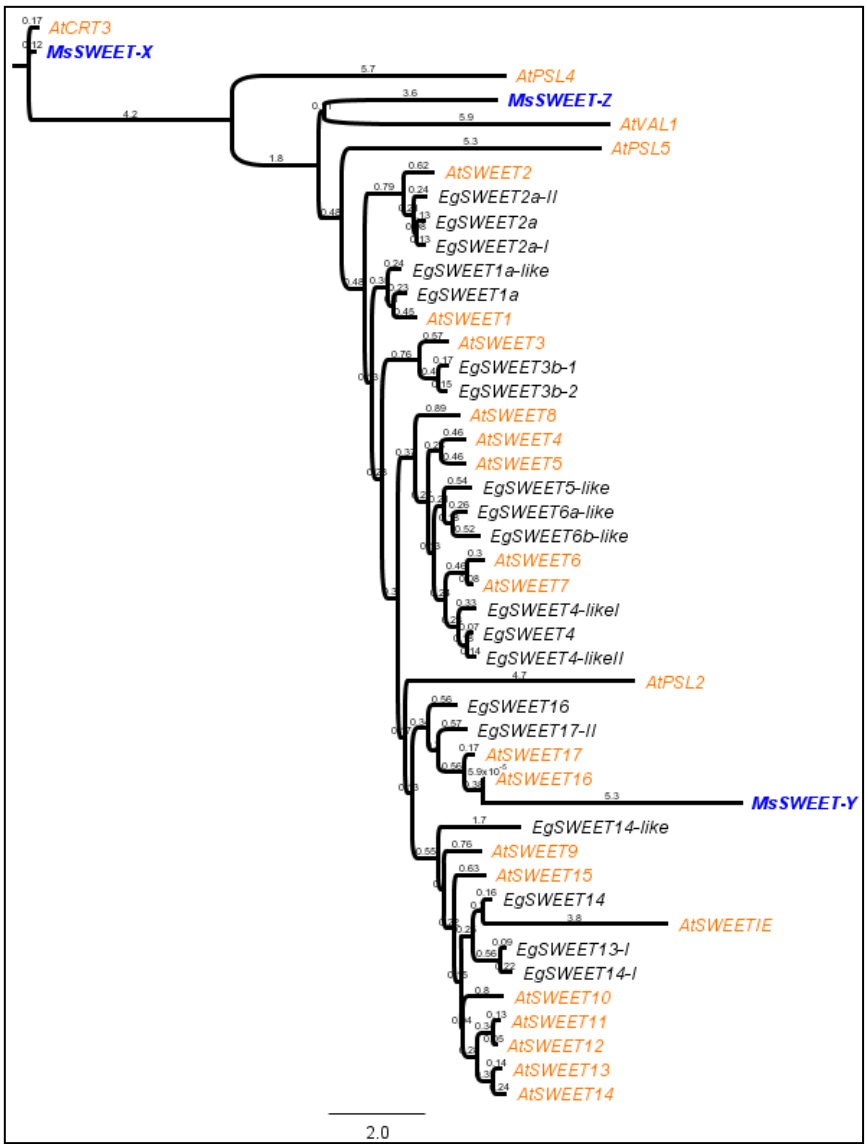

Figure 1. Phylogenetic analysis of 3 MsSWEET against 23 AtSWEET and 19 EgSWEET protein residues. The evolutionary history was inferred using the BioNJ method of PhyML online software. Bootstrap for the consensus tree was made as many as 6,506 repetitions. The bar showed an evolutionary distance of 2.0.

3.3. Conserved motifs and transmembrane domain of putative MsSWEET genes

The multiple sequence alignment using MAFFT has identified three differentiated groups (A, B and C) among the three compared species (Figure 2). Motif-A included 62 amino acid residues dominated by leucine (L) amino acid. Motif-B and Motif-C covered 37 and 27 amino acid residues, respectively. Each of *MsSWEET* gene identified in this work harboured unique sequences differed from *A. thaliana* and *E. guineensis*. In addition, the prediction of the transmembrane domain using TMHMM V2.0 in

three MsSWEET proteins showed that only MsSWEET-Z encodes four predicted transmembrane domains. The locations of predicted transmembrane domains are in 5th–27th, 42nd–61st, 74th–96th and 100th–121st amino acid positions of MsSWEET-Z protein (Figure 3).

MsSWEET-X showed 76.25% orthology to AtCRT3 (AT1G08450.1) encoding calreticulin protein. In *Arabidopsis*, AtCRT3 demonstrated to be involved in the response against the bacterial Pathogen-Associated Molecular Pattern (PAMP) [14]. This result hypothesized a potentially wide range function of the *SWEET* gene family in plants. As for, MsSWEET-X, it has a potentially similar function with AtCRT3.

MsSWEET-Z was identified with a low percentage of similarity (21.32%) to *E. guineensis* protein XP_010922652.1 due to its partial state of the sequence. The comparison to the EST database has its limit as this kind of library was prone to a sequence gap due to the random cloning of expressed CDS. However, reference studies showed that this protein refers to predicted bidirectional sugar transporter SWEET2a and is responsible in sugar transporter. The phylogenetic tree showed a close relationship between *EgSWEET2a* and *AtSWEET2*. The gene *AtSWEET2* encoded a SWEET2 protein mediating both low-affinity uptake and efflux of sugar across the plasma membrane in *Arabidopsis*. The partial similarities of *MsSWEET-Z* to *EgSWEET2a* was confirmed by four transmembrane domains detected using TMHMM V2.0. The transmembrane domain was generally found with helical structure and is responsible for regulating the membrane across activity. This result confirmed the predicted function of *MsSWEET-Z* as one of the key genes responsible for sugar transporter in sago palm.

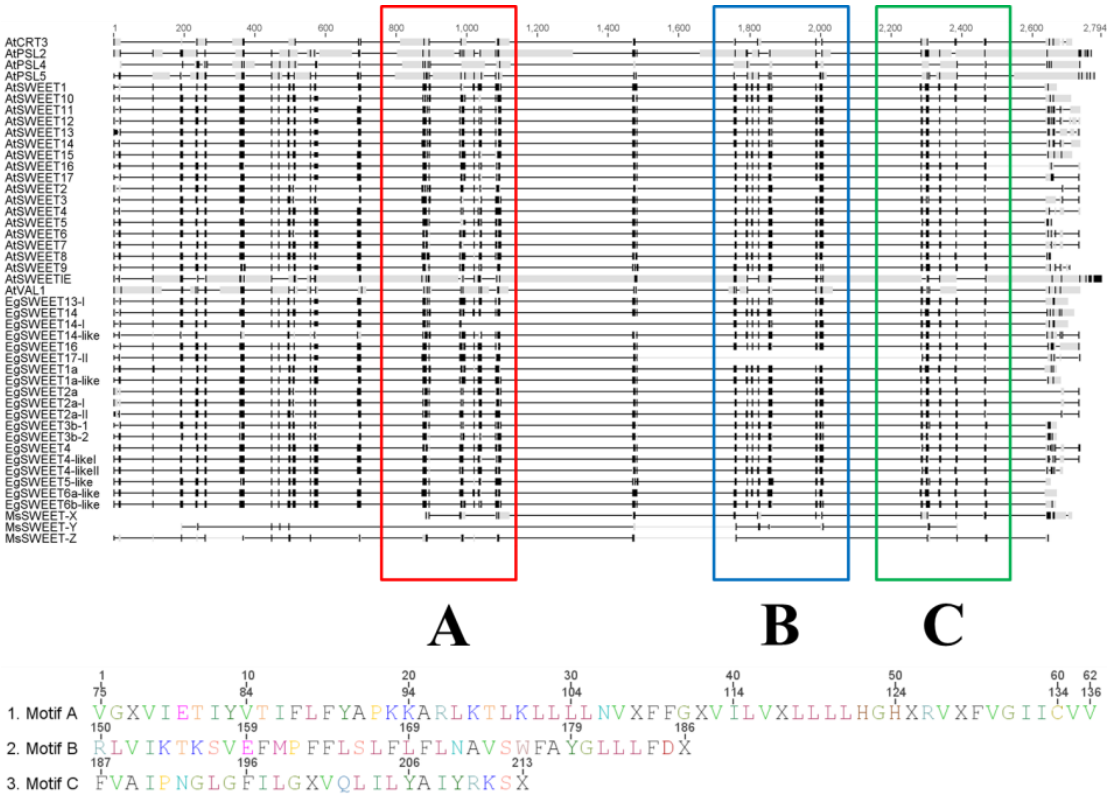

Figure 2. The multiple sequence alignment using MAFFT of the *SWEET* gene family of *M. sagu*, *A. thaliana* and *E. guineensis*. The protein motifs (A, B and C) were identified. The MAFFT was performed on the Galaxy platform with the parameters: matrix BLOSUM 62, maximum numbers of iterations 1,000 and distance method 6 mers.

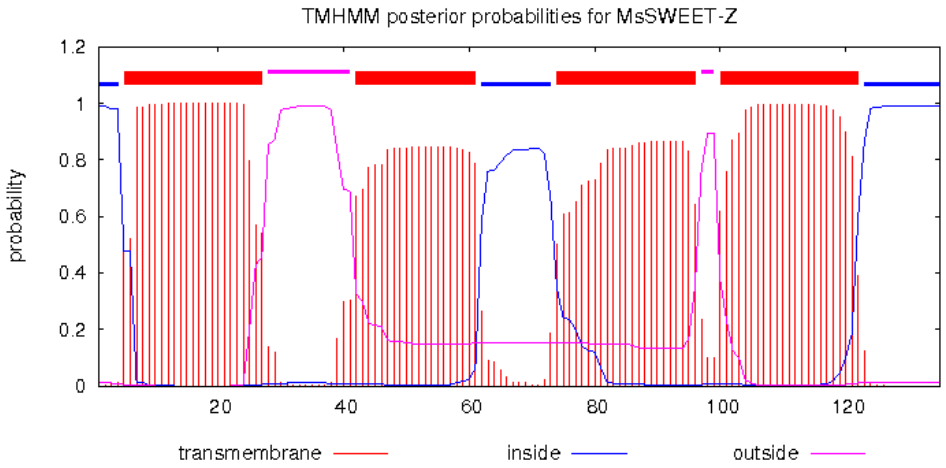

Figure 3. The predicted transmembrane domain encoded by MsSWEET-Z. The analysis was performed using online server TMHMM V2.0.

4. Conclusions

The *in silico* comparative sequence analysis has successfully identified three putative *MsSWEET* genes (*MsSWEET-X*, *MsSWEET-Y* and *MsSWEET-Z*). This work correspondingly showed that even though the oil palm and sago palm shared the common ancestry of the monocotyledonous family, their *SWEET* genes were divergent. The gene *MsSWEET-X* was highly close to its homolog in *Arabidopsis* (*AtCRT3* gene) as opposed to the gene *MsSWEET-Z* predicted to be the real transmembrane protein. Future research should focus on the gene expression of each *MsSWEET* gene in accordance with the sago palm with high producing starch.

5. References

[1] Feng C Y, Han J X, Han X X and Jiang J 2015 Genome-wide identification, phylogeny, and expression analysis of the *SWEET* gene family in tomato *Gene* **573** 261–72

[2] Jia B, Zhu X F, Pu Z J, Duan Y X, Hao L J, Zhang J, Chen L Q, Jeon C O and Xuan Y H 2017 Integrative view of the diversity and evolution of SWEET and SemiSWEET sugar transporters *Front. Plant Sci.* **8** 1–18

[3] Sui J L, Xiao X H, Qi J Y, Fang Y J and Tang C R 2017 The *SWEET* gene family in *Hevea brasiliensis*–its evolution and expression compared with four other plant species *FEBS Open Bio* **7** 1943–59

[4] Gao Y, Wang Z Y, Kumar V, Xu X F, Yuan D P, Zhu X F, Li T Y, Jia B and Xuan Y H 2018 Genome-wide identification of the *SWEET* gene family in wheat *Gene* **642** 284–92

[5] Eom J S, Chen L Q, Sosso D, Julius B T, Lin I W, Qu X Q, Braun D M and Frommer W B 2015 SWEETs, transporters for intracellular and intercellular sugar translocation *Curr. Opin. Plant Biol.* **25** 53–62

[6] Baker F R, Leach K A and Braun D M 2012 SWEET as sugar: new sucrose effluxers in plants *Mol. Plant* **5** 766–8

[7] Wee C C and Roslan H A 2012 Expressed sequence tags (ESTs) from young leaves of *Metroxylon sagu 3 Biotech* **2** 211–8

[8] Piyatrakul P, Yang M, Putranto R A, Pirrello J, Dessailly F, Hu S, Summo M, Theeravatanasuk K, Leclercq J, Kuswanhadi and Montoro P 2014 Sequence and expression analyses of ethylene response factors highly expressed in latex cells from *Hevea brasiliensis PLoS One* **9** e99367

[9] Putranto R A, Duan C, Kuswanhadi, Chaidamsari T, Rio M, Piyatrakul P, Herlinawati E, Pirrello J, Dessailly F, Leclercq J, Bonnot F, Tang C, Hu S and Montoro P 2015 Ethylene response

factors are controlled by multiple harvesting stresses in *Hevea brasiliensis PLoS One* **10** e0123618

[10] Martiansyah I, Putranto R A and Khumaida N 2017 Identification of putative gene family encoding protease inhibitors by *in silico* comparative analysis in *Hevea brasiliensis* Müell.Arg. genome *Menara Perkebunan* **85** 53–66

[11] Putranto R A, Martiansyah I and Saptari R T 2017 *In silico* identification and comparative analysis of *Hevea brasiliensis COBRA* gene family. In: *International Conference on Science and Engineering 2017* Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, 12–13 October 2017

[12] Guindon S, Dufayard J F, Lefort V, Anisimova M, Hordijk W and Gascuel O 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0 *Syst. Biol.* **59** 307–21

[13] Reynolds S M, Käll L, Riffle M E, Bilmes J A and Noble W S 2008 Transmembrane topology and signal peptide prediction using dynamic bayesian networks *PLoS Comput. Biol.* **4** e1000213

[14] Christensen A, Svensson K, Thelin L, Zhang W, Tintor N, Prins D, Funke N, Michalak M, Schulze-Lefert P, Saijo Y, Sommarin M, Widell S and Persson S 2010 Higher plant calreticulins have acquired specialized functions in *Arabidopsis PLoS One* **5** e11342