ZIBELINE INTERNATIONAL™ PUBLISHING

ISSN: 2682-7786 (Online)
CODEN: BDAIDR

Big Data in Agriculture (BDA)

DOI: http://doi.org/10.26480/bda.02.2021.51.55

RESEARCH ARTICLE

IN SILICO CHARACTERIZATION OF PUTATIVE ZAT10 GENE IN *Stevia rebaudiana* ACCESSION MS007

Nurul Hidayah Samsulrizal*, Nur Farhana Mustafa, Zabirah Abdul Rahim & Siti Noor Eliana Mohamad Nazar

Department of Plant Science, Kulliyyah of Science, International Islamic University Malaysia, Jalan Sultan Ahmad Shah, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia.
****Corresponding author e-mail: hidayahsamsulrizal@iium.edu.my***

ARTICLE DETAILS

Article History:

Received 18 April 2021
Accepted 21 May 2021
Available online 18 June 2021

ABSTRACT

Stevia rebaudiana is among recognized medicinal plants used as an artificial sweetener in food and drinks as it contains very low sugar content. It is recognized that stevia contains component steviol glycoside which contains stevioside and rebaudioside A has 50 to 100 more sweetness than sucrose. Due to this, stevia extraction of steviol glycoside is very demanded in the food and pharmaceutical industries. Research shows that stevioside levels are the highest at the time of flower bud formation and lower at the time preceding and following flower bud formation. Hence, this study aims to identify and characterize the putative gene that may involve in the flowering of local *S. rebaudiana* accession MS007 from the analysis of bioinformatics tools. The outcome of this study will help further in the study of the manipulation of the flowering process to increase the outcome yield of steviol glycoside. This study involves characterization of putative zinc finger *ZAT10* gene by homology search by BLAST, protein domain search using InterPro, physicochemical analysis using ProtParam and TMHMM, multiple sequence alignment using MUSCLE, and phylogenetic tree construction using MEGA. The bioinformatics analysis using these programs and software manages to identify the protein domain of *ZAT10* which shows the function of metal ion binding and involve in transcriptional regulation. The analysis of the phylogenetic tree also shows that ZAT10 protein may have a recent common ancestor in its own Asteraceae families such as *Lactuca sativa* and *Heliantus annus*.

KEYWORDS

Stevia rebaudiana, zinc finger ZAT10 protein, flowering, *in silico*.

1. INTRODUCTION

Stevia rebaudiana Bertoni is from the Asteraceae family and native to Northeast Paraguay. This branched bushy shrub plant has unique chemical components that have been studied and used as a substitute to regular synthetic and commercialize sugar such as aspartame, advantame, and saccharin. With the increases in sugar intake in the food industries, it will cause many health issues such as diabetes (Alizadeh et al., 2014). Research on stevia had shown that there are chemical components in the plant called steviol glycoside in the leaves that excreted sweet taste with the benefit of having approximately zero calories (Gasmalla et al., 2014). Steviol glycoside can provide a natural alternative sweetener that has a lower risk to the sensitive insulin response especially for insulin-resistant people like diabetic patients (Samsurizal et al., 2019). It also contains antioxidant properties.

Steviol glycoside produces compounds such as Stevioside and Rebaudioside A which give components that give it a sweet taste 50 to 100 times sweeter than sucrose (Gasmalla et al., 2014). The studies of *S. rebaudiana* sweet diterpene compounds has gained popularity in food and pharmaceutical industries as sugar substitute. There are numerous factors involved in the accumulation of steviol glycoside in stevia. It has been proved that the content of the steviol glycoside content will decrease after flowering stage and at the highest concentration at the formation of flower budding stage (Ceunen et al., 2012). A study reported that stevia that undergoes short day conditions will express early flowering and cause the stagnation of steviol glycoside accumulation (Ceunen et al., 2012).

On the other hand, the long stay conditioning on the stevia will increase the accumulation of steviol glycoside content by prolonging its vegetative period (Ceunen et al., 2012). The genes activation involving the flowering of stevia play a crucial role in determining the accumulation of steviol glycoside. Thus, identifying and knowing the function of the genes involved in flowering stages in stevia will give more understanding in the process of accumulation of steviol glycoside. The *ZAT10* gene is one of the genes in stevia that has not been well defined to provide a complete understanding of its functions. Hence, the goal of this research is to recognize and characterize the putative *ZAT10* gene and to understand how the steviol glycoside content accumulation can be manipulated and increased.

2. MATERIAL AND METHODS

2.1 Translating Nucleotide

The sequence of putative zinc finger protein *ZAT10* of Stevia rebaudiana

Quick Response Code	Access this article online	
	Website: www.bigdatainagriculture.com	**DOI:** 10.26480/bda.02.2021.51.55

Cite the Article: Nurul Hidayah Samsulrizal, Nur Farhana Mustafa, Zabirah Abdul Rahim & Siti Noor Eliana Mohamad Nazar (2021). In Silico Characterization of Putative Zat10 Gene in *Stevia Rebaudiana* Accession Ms007. Big Data in Agriculture, 3(2): 51-55.

MS007 was obtained from (Samsulrizal et al., 2020). Then, the putative *ZAT10* gene sequence was translated using ExPaSy software [https://web.expasy.org/translate/] and the longest open reading frame was used for the next steps of analysis (Ceunen et al., 2012).

2.2 Homology Search

Homology search was done using the BLAST program. BLASTP was used to analyze the *ZAT10* amino acid sequence of stevia to find the homology search at [https://blast.ncbi.nlm.nih.gov]. First, using the reference sequence library RefSeq, the putative *ZAT10* amino acid sequence of stevia has been blasted and 20 sequences were chosen from various organisms with the highest identity percentage and E-value.

2.3 Domain Search

The putative ZAT10 sequence was analyzed by the InterPro database [http://www.ebi.ac.uk/interpro/] which is a platform for the domain search (Mitchell et al., 2019). InterPro database integrates with other sources of database such as Protein family (pfam) and Simple Modular Architecture Research Tool (SMART) to predict and integrate the sequence by representing protein domain, families, and functional sites (Hunter et al., 2009).

2.4 Physicochemical Properties of ZAT10 Protein

ProtParam [https://web.expasy.org/protparam/] was used to study the molecular weight, theoretical pl, and amino acid composition of the protein sequence. Then, Transmembrane Helices of Hidden Markov Model (TMHMM) [http://www.cbs.dtu.dk/services/TMHMM-2.0/] was used to give statistics and list of the location of the predicted transmembrane helices and predicted location of the intervening loop regions (Gasteiger et al., 2005).

2.5 Multiple Sequence Alignment

MUSCLE is one of the programs that can manage to align multiple sequences of protein or nucleotide. By using the MUSCLE program offered in the MEGA software can make the alignment easier and faster. A total of 21 protein sequences retrieved from BLAST were inserted into the MEGA in FASTA format and produced alignments from the sequences that can be differentiated by different colors (Kumar et al., 2018).

2.6 Phylogenetic Analysis

MEGA [https://www.megasoftware.net/] software was used to construct phylogenetic trees (Hall, 2013). From 21 aligned protein sequences, the phylogeny was chosen in the MEGA to construct a phylogenetic tree and the Maximum Likelihood was chosen as the preferred method with 1000 replication bootstrap value (Tamura et al., 2011).

3. Results and Discussion

The aims of this study are to identify and characterize the *ZAT10* gene in *S. rebaudiana* accession MS007 by using bioinformatics approaches. In order to identify and characterize the gene, the transcriptome database from the previous study was used for further analysis (Samsulrizal et al., 2020). The *ZAT10* gene that was selected from the previous transcriptome database was identified as Cluster-25047.0. The sequence is then translated into protein sequence using ExPaSy and the longest open reading frame was selected.

3.1 Homology Search Using BLAST

Pearson stated that sequence similarity searching to identify homologous sequences is the first and most informative step in the analysis of newly determined sequences (Pearson, 2013). As modern sequence databases such as BLAST are very comprehensive, many metagenomics sequence samples may have similarity with the protein in the database. Homologous searching is very important and effective so that sequence with significant similarity may be inferred to share a common ancestor (Samsulrizal et al., 2020). Expectation value (E-value) indicates the similarity quality score of expected hits. E-value is used as an initial filter for BLAST search results. The smaller E-value showed a better match and the highest E-value in this search was 10^{-30}. The homology search by BLASTP gives 100 sequences hit with the putative *ZAT10* gene. In Table 1.0, the BLAST result shows the lowest score was 204 and the percentage identity was 51.42% from *Quercus lobata*.

The highest score and identity percentage gathered from the homology search are *Helianthus annuus* species with 306 total score and 72.69% percentage identity. Percentage identity is important to evaluate the top hit by BLAST. Percentage identity is a quantitative measurement that shows the similarity of the sequences. High percentage identity such as percentage identity from *Helianthus annuus* and *Lactuca sativa* species are expected to be more closely related to the putative ZAT10 protein from stevia than other species with lower percentage identity. Hence, percentage identity can be said to reflect the degree of relatedness of the species (Zhang et al., 2008).

Table 1: Top hit from homology search using BLASTP in NCBI database

Description	Total Score	E-Value	Percentage Identification (%)
Zinc finger protein ZAT10-like [*Heliantus annuus]*	306	2e-103	72.69
Zinc finger ZAT10-like [*Lactuca sativa]*	277	1e-91	67.24
Zinc finger protein ZAT10-like [*Heliantus annuus]*	275	9e-91	67.84
Zinc finger protein ZAT10-like [*Cynara cardunculus var. scolymus*]	261	7e-85	61.18
Zinc finger ZAT10-like [*Lactuca sativa]*	257	3e-83	58.78
Zinc finger protein ZAT10-like [*Heliantus annuus]*	252	2e-81	61.04
Zinc finger protein ZAT10-like [*Cynara cardunculus var. scolymus*	239	1e-76	61.80
Zinc finger protein ZAT10-like [*Heliantus annuus]*	475	7e-72	59.01
Zinc finger protein ZAT10-like [*Nicotiana tomentosiformis]*	214	9e-67	54.59
Zinc finger protein ZAT10-like [*Vitis riparia*]	213	1e-66	59.11
PREDICTED: zinc finger protein ZAT10 [*Vitis vinifera*]	211	3e-66	58.67

3.2 Domain Analysis

InterPro consists of multiple databases that are used to predict specific sequence sites describing the same functional domain, family, and site (Hunter et al., 2009). There are many benefits to using different sources such as highlighted potentially incorrect hits from one source as compared to another data source as it may be an outlier that may be due to false positives (Hunter et al., 2009). Protein may consist of one or more domains. A domain refers to a distinct and structural unit in protein that is responsible for a particular function and interaction which may define the overall protein function. There are varieties of domains and similar domains can be found in protein with different functions.

In this study, the domain database InterPro was used as shown in Table 2.0. Zinc finger C2H2 domains are the only domain found using the PROSITE database. This domain has been identified to have a nucleic acid-binding protein structure which is identified in transcription factor TFIIIA in Xenopus. This domain has 22 amino acid residues based on the putative ZAT10 amino acid sequence. Two cysteine and two histidine residues are found at both extremities of the domain and they involve in the tetrahedral coordination of a zinc atom. As many classes of zinc finger domain are characterized based on their position and number of the histidine and

cysteine residues that involve coordination of the zinc atom. It also states that some zinc finger class members have zinc-dependent DNA or RNA binding property.

Table 2: Domain research findings through the Interpro database

Accession	Domain	Domain function
IPR041057	ZZHX, C2H2 finger domain	ZHX_Znf_C2H2 ▪ Associated with roles in transcriptional regulation ▪ Able to form both homo- and heterodimers via the region containing homeodomain ▪ ZHX1 is a transcriptional repressor which is ubiquitously expressed. It interacts with nuclear factor Y subunit A (NFYA) and DNA methyl transferase 3B (DNMT3B) for its repression activity
IPR003604	Matrin/U1-C-like, C2H2-type zinc finger	▪ Nucleic acid binding ▪ Zinc ion binding ▪ Binding DNA, RNA, protein and/or lipid substrates

3.3 Physicocemical Properties of ZAT10 Protein

ProtParam has been used to study the physicochemical parameter of a protein sequence. This parameter includes the theoretical pl, amino acid composition, molecular weight, atomic acid composition, estimated half-life, instability index, extinction coefficient, grand average of hydropathicity (GRAVY) and aliphatic index (Gasteiger et al., 2005). Table 3 shows the result of ProtParam analysis on every parameter it provides. By studying various chemical and physical parameters of the protein sequence from ProtParam, it enables the primary analysis of protein. Protein is very important to provide various ranges of function to living organisms. Protein is built up from a group of amino acids that are linked to each other by peptide bonds and form structure differs from one to another. This structure then indicates the levels of organization of a protein molecule.

The *ZAT10* gene sequence has a molecular weight of 2378.27 which is calculated by the addition of average isotopic masses of amino acids and the average isotopic mass of one water molecule. For the instability index, it was used to determine whether the protein would be stable in a test tube. If the index is low than 40, then the protein is in stable condition but if the index value is higher than 40, the protein is considered unstable. The instability index shows that the protein is unstable at 65.17. TMHMM is termed to Trans Membrane Hidden Markov Model where it can give predictions of the topology of protein which shows the overall topology of the protein (Krogh et al., 2001). This program can predict transmembrane helices and discriminate between soluble and membrane proteins with a high degree of accuracy. The TMHMM analysis shows the putative ZAT10 protein contains three color lines which the posterior probabilities of inside, outside of the transmembrane helix.

The pink line indicates the residue located outside of the cell, the blue line indicates that either residue is in the cytoplasm or not and the bold pink line indicates the residue is inside the cell. Based on the analysis the probability of the protein residue shows to be located outside of the cell. It also analyzed that the protein does not have transmembrane as the number of predicted transmembrane is zero. Transmembrane helices and helical bundles form a major building block of the protein membrane. Since there is no transmembrane predicted to be present in the putative ZAT10 protein of stevia thus, the protein could be unstable.

Table 3: The analysis result of physico-analysis properties of putative ZAT10 protein

Number of amino acids	**219**
Molecular weight	2378.27
Theoretical pI	9.16
Total number of negatively charged residue (Asp + Glu)	19
Total number of positively charged residue (Arg +Lys)	25
Atomic composition (formula = C1015H1608N316O326S7)	Carbon = 1015, Hydrogen = 1608, Nitrogen = 316, Oxygen = 326, Sulfur = 7
Total number of atoms	3272
Extinction coefficients	-7700 M^{-1} cm^{-1} assuming all Cys residue are reduce -7450 M^{-1} cm^{-1} assuming all Cys residue are reduce
Estimated half life	- 30 hours in mammalian reticulocytes, in vitro. - >20 hours in yeast, in vivo. - >10 hours in *Escherichia coli*, in vivo
Instability index	65.17 (protein is unstable)
Alipathic index	61.46
Grad average of hydropathicity (GRAVY)	-0.674

3.4 Constructing the Phylogenetic Tree

Multiple sequence alignment is vital for molecular phylogenetic assessment, structure prediction, and residue identification. One of the algorithms that were used was the progressive method by first estimating a tree and then constructing a pairwise alignment of the subs-trees found on each internal node (Edgar, 2004). However, misalignment will happen in progressive methods that need further refinement. MUSCLE uses kmer distance and Kimura distance for a pair of sequences where kmer is for unaligned pairs and Kimura is for aligned pair sequences (Edgar, 2004). There are 21 of ZAT10 protein sequences aligned where 20 sequences are from the top hit of BLASTP results and one of our candidate sequences. Figure 1 shows the alignment of 21 protein sequences. Many sequences align and show the same amino acid sequences which can be analyzed in terms of the conserved region of the protein.

The conserved region of protein sequences indicated the functional element as it shows a conserved part shared by common ancestors (Stojanovic et al., 1999). However, not all conserved regions are functional. In Figure 1, the amino acid EEEYLALCLMLLAR in position 53 to 66 is conserved in all of 21 protein sequences is one of the six conserved regions found. Along all the multiple sequence alignment of 21 protein sequences, there are five more conserved regions that can be seen from the MUSLCE analysis as shown in the black frame. The different color refers to different physicochemical properties of the protein (Edgar, 2004). The conserved residues denoted with an asterisk *. Thus, the sequence alignment in the analysis had been used to build a good phylogenetic tree by cutting out the N-terminal and C-terminal of the sequences (Edgar, 2004).

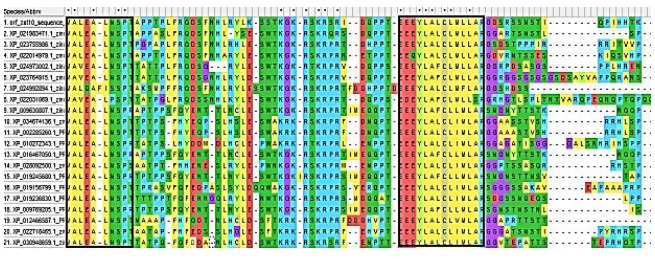

Figure 1: Multiple sequence alignment analysis using MUSCLE from MEGA software

The phylogenetic tree represents a hierarchical organization for biological data. This tree also represents the classification and clustering visualization of the protein sequences after the multiple sequence alignment. The base called root is known as the ancestor and the ends are called the leaves are where the nodes have no children. The maximum likelihood method was described as the probability of an evolutionary event on the tree causing the preferred phylogenetic tree is one with the highest likelihood (Pavlopoulos et al., 2010). Another method used is the Neighbour-Joining (NJ) method. NJ method used the distance matrix and clustering based on the minimum evolution description. The tree has to be with the least total branch length which is more preferable. The NJ method is less expensive, faster and can be applied to larger data sets (Pavlopoulos et al., 2010).

The phylogenetic tree was constructed using the Maximum Likelihood method using MEGA software wherein this phylogenetic tree the highest log-likelihood is -6432.72. A bootstrap role is to assess the accuracy of statistical estimates. The higher the value, from 1 to 100, shows higher the accuracy of the estimation (Efron et al., 1996). Bootstrap value is used in the phylogenetic tree which is used for assessing the accuracy of statistical estimation. For this tree construction, 1000 bootstrap replications were used. The bootstrap value was placed at the nodes of the tree branch and nodes received a high bootstrap value indicating strong support from the nodes. Soltis and Soltis state, the phylogenetic tree rule on the bootstrap replication is given a set of non-contradictory nodes, each with rejection probability below 50% (Soltis and Soltis, 2003). Figure 2 shows the phylogenetic tree construction by the Maximum Likelihood method. It was divided into 3 subgroups based on the bootstrapping value that was over 70% with 1000 replication of the tree.

From the phylogenetic tree, group 1 shows zinc finger protein from stevia has the most recent common ancestor with *Lactuca sativa* with bootstrap confidence at 53%. *Lactuca sativa* is from the same family which is Asteraceae which explains they may come from common ancestors. Next, zinc finger protein from stevia and *Lactuca sativa* also has a similar ancestor with species *Heliantus annus* with bootstrap confidence of 66%. All four of them form a monophyletic group of common ancestors. Based on the phylogenetic tree, zinc finger protein from stevia and *Lactuca sativa* are more closely related compared to *Heliantus annus* because both come from the same nodes which indicate they shared the most recent common ancestor. All these species show that they may share a recent common ancestor with a higher bootstrap value of 90%.

All these species in group 1 have zinc finger ZAT10-like protein which may have similar conserved regions and domains due to their closeness to the common ancestor to each other. Thus, it can be concluded that the protein may have played a role in the abiotic tolerance to salt. In group 2 which known as paraphyletic group as it consists of the same family of *Lactuca sativa* and *Heliantus annuus* from one common ancestor to group 1. Both species shared 49% bootstrap value which based on bootstrap confidence value means the nodes are less supported. *Lactuca sativa* has ZAT10 protein while *Heliantus annuus* has ZAT9 protein. ZAT9 protein and ZAT10 protein shared the same C2H2-type 2 domain according to the Uniprot database and it also has the same molecular function as involving in DNA-binding transcription factor activity. Both ZAT10 and ZAT9 proteins have similar zinc finger C2H2 domain.

This means that both proteins may express many similar functional and structural units. Then, the most less recent common ancestors shared in group 3 by zinc finger protein from stevia is *Gossypium raimondii* and *Durio zibethinus* group which come from polyphyletic group to group 1 and group 2. Polyphyletic group means the group 1 and group 3 are not defined by the common ancestor. Both species in group 3 have no similarity with group 1 up besides they both are from the Plantae kingdom. Based on the NCBI database on the predicted zinc finger protein ZAT10-like in *G. raimondii* has domain zinc finger C2H2-type 6 as opposed to the domain of the species in group 1 which is domain zinc finger C2H2-type 2. In the zinc finger C2H2 superfamily, it has the common function as DNA-binding in transcription factor (Jones et al., 2014). Both groups are found to be less related as to no recent common ancestor compared to stevia and *Lactuca sativa* and *Heliantus annus* from group 1 and group 2 which come from the same plant Asteraceae family.

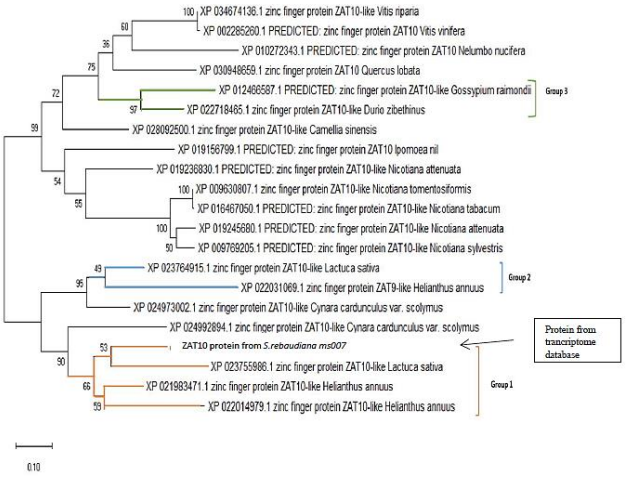

Figure 2: The phylogenetic tree is classified into three groups i.e., Group1, Group 2 and Group3, using the Maximum Likelihood algorithm.

4. Conclusion

This study aims to identify and characterize the *ZAT10* gene in *S. rebaudiana* MS007 by using the transcriptomic data from a previous study. The analysis of transcriptome data manages to study the protein domain of the putative *ZAT10* gene which shows its domain which is zinc finger C2H2 domain has to function as nucleic acid-binding protein structures and some members of this domain demonstrate zinc-dependent DNA or RNA binding property. For the physicochemical analysis, the Protparam program also analyzed putative ZAT10 protein *S. rebaudiana* MS007 is not stable due to high instability index at 65.17 and TMHMM analysis show that putative ZAT10 protein of *S. rebaudiana* is located outside the cell and does not has transmembrane helix which may cause the protein to be unstable. In the MUSCLE sequence alignment, six conserved regions had been identified which shows they could have more than one similar functional and structural unit. The phylogenetic tree shows that putative ZAT10 protein from *S. rebaudiana* MS007 is highly related to *Lactuca sativa* and *Heliantus annus*.

References

Alizadeh, M., Azizi-lalabadi, M., Hojat-ansari, H. and Kheirouri, S., 2014. Effect of Stevia as a substitute for sugar on physicochemical and sensory properties of fruit-based milk shake. Journal of scientific research and reports, Pp. 1421-1429.

Ceunen, S., Werbrouck, S., Geuns, J.M., 2012. Stimulation of steviol glycoside accumulation in Stevia rebaudiana by red LED light. Journal of plant physiology, 169 (7), Pp.749-752.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32 (5), Pp. 1792-1797.

Efron, B., Halloran, E. and Holmes, S., 1996. Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences, 93 (23), Pp. 13429-13429.

Gasmalla, M.A.A., Yang, R. and Hua, X., 2014. Stevia rebaudiana Bertoni: an alternative sugar replacer and its application in food industry. Food Engineering Reviews, 6 (4), Pp. 150-162.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R.D. and Bairoch, A., 2005. Protein identification and analysis tools on the ExPASy server. (In) John M. Walker (Ed): The Proteomics Protocols Handbook, Pp. 571–607.

Hall, B.G., 2013. Building phylogenetic trees from molecular data with MEGA. Molecular Biology and Evolution, 30 (5), Pp. 1229–1235.

https://doi.org/10.1093/molbev/mst012

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., 2009. InterPro: the integrative protein signature database. Nucleic acids research, 37(suppl_1), Pp. D211-D215.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. and Pesseat, S., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics, 30 (9), Pp.1236-1240.

Krogh, A., Larsson, B., Von Heijne, G. and Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology, 305 (3), Pp. 567-580.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K., 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution, 35 (6), Pp. 1547–1549. https://doi.org/10.1093/molbev/msy096

Mitchell, A.L., Attwood, T.K., Babbitt, P.C., Blum, M., Bork, P., Bridge, A., Brown, S.D., Chang, H., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Baeur, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J. A., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, D., Tosatto, S. C. E., Yong, S., and Finn, R. D., 2019. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research, 47 (D1), D351–D360. https://doi.org/10.1093/nar/gky1100

Pavlopoulos, G.A., Soldatos, T.G., Barbosa-Silva, A. and Schneider, R., 2010. A reference guide for tree analysis and visualization. BioData mining, 3 (1), Pp. 1-24.

Pearson, W.R., 2013. An introduction to sequence similarity ("homology") searching. Current protocols in bioinformatics, 42 (1), Pp. 3-1.

Samsulrizal, N.H., Khadzran, K.S., Shaarani, S.H., Noh, A.L., Sundram, T.C., Naim, M.A. and Zainuddin, Z., 2020. De novo transcriptome dataset of Stevia rebaudiana accession MS007. Data in brief, 28.

Samsulrizal, N.H., Zainuddin, Z., Noh, A.L., Sundram, T.C., 2019. A Review of Approaches in Steviol Glycosides Synthesis. International Journal of Life Sciences and Biotechnology, 2 (3), Pp. 145-157.

Soltis, P.S. and Soltis, D.E., 2003. Applying the bootstrap in phylogeny reconstruction. Statistical Science, Pp. 256-267.

Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., Miller, W. and Hardison, R., 1999. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. Nucleic Acids Research, 27 (19), Pp. 3899-3910.

Tamura, Koichiro, Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S., 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution, 28 (10), Pp. 2731–2739. https://doi.org/10.1093/molbev/msr121

Zhang, Y., Song, G., Vinař, T., Green, E.D., Siepel, A. and Miller, W., 2008, March. Reconstructing the evolutionary history of complex human gene clusters. In Annual International Conference on Research in Computational Molecular Biology, Pp. 29-49. Springer, Berlin, Heidelberg.