JOURNAL OF TROPICAL LIFE SCIENCE
2022, Vol. 12, No. 3, 377 – 387
http://dx.doi.org/10.11594/jtls.12.03.11

Research Article

In Silico Analysis and 3D Structure Prediction of Putative UDP-Glycosyltransferase 76G1 Protein in *Stevia rebaudiana* MS007

Nor Iswani Mokthar [1], Muhammad Amirul Husni Samsulrizal [2], Afiqah Ramatullah Khan [2], Zarina Zainuddin [3], Tamil Chelvan Meenakshi Sundram [2], Nik Yusnoraini Yusof [4], Nurul Hidayah Samsulrizal [2*]

[1] Department of Biological Sciences and Biotechnology, The National University of Malaysia 43600, Bangi, Selangor, Malaysia
[2] Department of Plant Science, Kulliyyah of Science, International Islamic University of Malaysia, Bandar Indera Mahkota, 25200, Kuantan, Pahang, Malaysia
[3] Plant Productivity and Sustainable Resource Unit, Kuliyyah of Science, International Islamic University Malaysia, Bandar Indera Mahkota, 25200, Kuantan, Pahang, Malaysia
[4] Institute for Research in Molecular Medicine (INFORMM), Universiti Sains Malaysia, Kubang Kerian, 15200 Kota Bharu, Kelantan, Malaysia

Article history:
Submission February 2022
Revised March 2022
Accepted July 2022

**Corresponding author:*
E-mail:
hidayahsamsulrizal@iium.edu.my

ABSTRACT

Stevia rebaudiana is a plant of the Asteraceae family that is used as a natural sweetener. Stevia has been shown to be safe for human consumption and has been utilised as a sweetener substitute for diabetic and obese people. In this study, the structure and gene content involved in the synthesis of putative UDP-glycosyltransferase 76G1 (UGT76G1) protein in *S. rebaudiana* MS007 were analysed using an *in-silico* method. Homologous search using BlastP revealed the highest percentage of identity, score, and E-value for UDP-glycosyltransferase 76G1-like of *Helianthus annuus* (ID: XP_021973845.1). The presence of IPR002213 UDP-glucuronosyl/UDP-glucosyltransferase entry, which is available at locations 89bp to 246 bp, was also verified by the protein family search using InterPro. MEGA-X software was used to construct a molecular phylogeny study, revealing that this protein belongs to the Asteraceae family. To predict the primary, secondary, and tertiary protein structures of the putative UGT76G1 protein, the ProtParam, ExPasy, PSIPRED, and Phyre2 programmes were implemented. The putative UGT76G1 protein's tertiary structure prediction was given a score of 100.0% confidence by the single highest scoring template and a coverage of 98%, with the dimension of the model being (Å) of X: 52.453, Y: 61.270, and Z: 48.102. The UGT76G1 model fulfilled the quality standards and was approved for further analysis after validation performed by PROCHECK, VERIFY3D, and ERRAT. Thus, the findings of this work have contributed to a better knowledge of putative UDP-glycosyltransferase 76G1 features and target recognition processes, which will lead to better information on protein-protein interaction in *S. rebaudiana* MS007.

Keywords: Phylogenetic, Stevia rebaudiana, UGT76G1, 3D Structure Prediction

Introduction

Stevia rebaudiana Bertoni, also known as Stevia, is a perennial herbaceous plant belonging to the Asteraceae family [1]. *S. rebaudiana* can be utilised as a food and medicine due to its significant level of non-caloric sweetening flavouring ingredient [2-6]. The two major steviol glycosides, stevioside and rebaudioside A, are responsible for the sweet flavor of stevia [2]. Stevioside is a glycoside with a glucosyl and sophorosyl residue linked to the steviol aglycon. Meanwhile, rebaudioside A is like stevioside but has glucosyl-(1-3)-sophorosyl residues instead of sophorosyl residues

How to cite:
Mokthar NI, Samsulrizal MAH, Khan AR, et al. (2022) *In Silico* Analysis and 3D Structure Prediction of Putative UDP-Glycosyltransferase 76G1 Protein in *Stevia rebaudiana* MS007. Journal of Tropical Life Science 12 (3): 377 – 387. doi: 10.11594/jtls.12.03.11.

[4]. Both stevioside and rebaudioside are safe for human intake, with stevioside being harmless, non-mutagenic, and incapable of causing cancer, and rebaudioside showing no toxicity effects [5]. Different species and varieties of stevia have varying sweetening chemical capabilities, with *S. rebaudiana* being the sweetest of all [4].

The biosynthesis of steviol glycosides is similar to the gibberellic acid pathway (Figure 1) at the beginning stages of synthesis [2]. Two sta-ges are involved in steviol glycosides biosynthesis: upstream and downstream. Methylerythritol 4-phosphate (MEP) is the upstream pathway in which geranylgeranyl diphosphate (GGDP) is synthesized meanwhile the downstream pathway involves the steviol glycosides biosynthesis from GGDP [7]. In the downstream pathway, uridine diphosphate-dependent glycosyltransferase (UGTs) converts steviol into various glycosides [8, 9]. The steviol glycosides biosynthesis pathway involves the functional genomics of three UGTs, UGT74G1, UGT76G1, and UGT85C2 [8]. The addition of C-13-glucose to steviol is catalysed by UGT85C2, resulting in the production of steviolmonoside. Next, at the C-2' position, 13-O-glucose is added, forming steviolbioside [8]. Then, the C-19 carboxyl of steviolbioside is glycosylated by UGT74G1, which results in the production of stevioside [9, 10, 11]. Finally, UGT76G1 glucosylates the C-3' of the C-13-O-glucose to form rebaudioside A [12].

A study reported that seventeen *S. rebaudiana* accessions were utilized to analyze the morphology of all accessions in preparation for advanced breeding development with fourteen accessions obtained, including MS007 from across Malaysia and three accessions from Paraguay [13]. The MS007 variety was found to be among the tallest and showed good qualities in terms of plant height, number of leaves and leaf size and was remarked as having a compact habit, i.e., dwarf-plants and obovate-shaped leaves after all the plants were exposed to the same environmental and climatic conditions [14]. To date, increased production of steviol glycosides while maintaining the sweetener's safety for human use is one of the challenges in developing *S. rebaudiana.* However, Mirzaei & Shakoory-Moghadam (2022) stated that, stevia use is advised for diabetic patients as a blood sugar stabilizer [15]. Nevertheless, there have been no investigations on the associated genes in the biosynthesis of steviol glycosides from *S. rebaudiana* accession MS007 that we are aware of. Noteworthily, it is essential to provide a complete mechanistic understanding of the associated genes in the steviol glycosides production. Hence, this study aims to utilize sequence- and structure-based bioinformatics approach to characterize the protein structure and function of UDP-glycosyltransferase 76G1 in *S. rebaudiana* accession MS007.

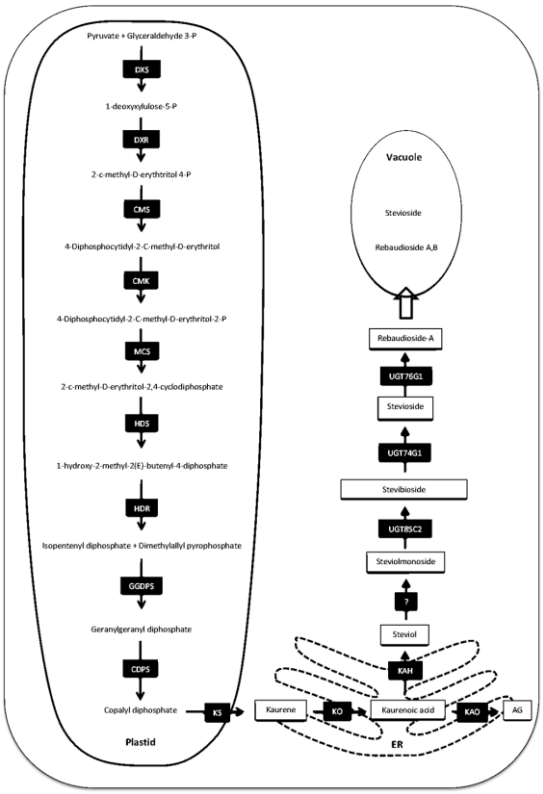


Figure 1. An illustration of the genes involved in production of SGs (taken from Samsulrizal *et al*., 2019)

Material and Methods

Protein translation

The transcriptomic data of the *S. rebaudiana* accessions MS007 was sequenced using Illumina technology, and the *de novo* transcriptome was assembled using Trinity RNA-Seq v2.0.6 [16]. Then, the Cluster-31069.44602 UGT76G1 was selected as the longest gene available. The nucleotide sequence of Cluster-31069.44602 UGT76G1 was inserted into the query box of ExPASy Translate (web.expasy.org/translate/). ExPASy Translate is a sequence analysis tool that helps to perform protein translation from a nucleotide sequence in six reading frames [17]. The longest reading frame that started with amino acid methi-

onine (M) was selected and saved in FASTA file format.

Homolog search

The protein sequence of Cluster-31069.44602 UGT76G1 of *Stevia rebaudiana* MS007 was analysed via homology search by using BlastP at National Center for Biotechnology Information (NCBI) website (https://blast.ncbi.nlm.nih.gov/Blast.cgi). All the parameters were maintained as default i.e., non-redundant database (*nr*).

Protein domains and families

InterPro database was used to find protein domains and families involved (www.ebi.ac.uk /interpro/search/sequence/). This database offers protein family classifications and can predict the presence of conserved domains and sites [18]. Pfam database (pfam.xfam.org/search/sequence) was also used to determine accurate protein families and domains which are represented by two multiple sequence alignments and two profile-Hidden Markov Models (profile-HMMs) [19].

Physicochemical properties of UGT76G1 protein

The translated protein sequence of UDP-glycosyltransferase 76G1 (UGT76G1) was inserted into the query box of the ProtParam-ExPasy Tool (web.expasy.org/protparam/). The molecular weight and theoretical pI were calculated using this program [20].

Constructing phylogenetic tree

Multiple sequence alignment of 15 selected protein sequences, including UGT76G1 protein, was done at www.ebi.ac.uk/Tools/msa/muscle/. Based on high and low consensus qualities, these 15 sequences were chosen. Molecular Evolutionary Genetics Analysis (MEGA X) software was downloaded at www.megasoftware.net/. Protein sequences were imported into MEGA X software to build the phylogenetic tree to identify the evolutionary relationships between the sequences [21]. To analyse the concluded evolutionary trees, the phylogenetic tree was constructed using the Maximum Likelihood method and the JTT model [22] with bootstrapping value set at 1000 replications.

PSIPRED Workbench

UGT76G1 amino acid sequence was uploaded to PSIPRED Workbench (bioinf.cs.ucl.ac.uk/psipred/). Based on position-specific scoring matrices, PSIPRED 4.0 (Predict Secondary Structure) was utilised to predict secondary protein structure [23]. PSIPRED was selected for "Popular Analyses". DeepMetaPSICOV 1.0 (Structural Contact Prediction) was chosen for "Contact Analysis" but features for "Fold Recognition", "Structure Modelling" and "Function Prediction" were not selected. In addition, DomPred (Protein Domain Prediction) was used to select for "Domain Prediction".

Phyre2 Protein Fold Recognition Server

Phyre2 Protein Fold Recognition Server is applied to estimate the 3D structure of proteins [24]. Therefore, this program was used to predict UGT76G1 tertiary protein structure by uploading the amino acid sequence at www.sbg.bio.ic.ac.uk/phyre2/ and the normal modelling mode was selected.

Validation of UGT76G1 structure model

The three different methods, PROCHECK [25], ERRAT [26] and VERIFY 3D [27] were used to assess the quality of the structural model, UGT76G1 model was interpreted based on the geometric quality of the backbone conformation, the residue interaction and contacts and the energy profile of the structure. Ramachandran Plot (services.mbi.ucla.edu/SAVES/Ramachandran/), ERRAT (services.mbi.ucla.edu/ERRAT/), Verify 3D (services.mbi.ucla.edu/Verify_3D/), and PROVE were used in model refinement analysis.

Results and Discussion

A transcriptomic analysis study of *S. rebaudiana* accession MS007 was performed at International Islamic University Malaysia by Samsul rizal and the team since 2019. From the study, the UGT76G1 protein was found as a protein that is mainly involved in steviol glycosides biosynthesis. The longest open reading frame of UGT76G1 with a cumulative length of 283 amino acids was used in this study.

Homology search

The list with significant sequence homology to the UGT76G1 sequence is shown in Table 1. The results were sorted according to the Expect value (E-value). The lower the e-value, the more significant the score. Based on the results, the most similar alignment for the UGT76G1 protein se-

Table 1. BlastP analysis of UGT76G1 sequence for *S. rebaudiana* MS007

Identical genus & species to UGT76G1 sequence of *S. rebaudiana* MS007	Accession No.	Identity (%)	E-value
UDP-glycosyltransferase 76G1-like [*Helianthus annuus*]	XP_021973844.1	77.39	4e-161
UDP-glycosyltransferase 76G1-like [*Helianthus annuus*]	XP_021973845.1	77.66	9e-156
putative UDP-glucuronosyl/UDP-glucosyltransferase [*Helianthus annuus*]	KAF5800049.1	77.66	6e-155
UDP-glycosyltransferase 76G1 isoform X1 [*Helianthus annuus*]	XP_021981266.1	77.03	5e-154
UDP-glycosyltransferase 76G1 [*Helianthus annuus*]	XP_021973866.1	74.82	1e-146
UDP-glycosyltransferase 76G1 isoform X2 [*Helianthus annuus*]	XP_035833363.1	77.65	3e-144
hypothetical protein E3N88_03269 [*Mikania micrantha*]	KAD7480133.1	72.79	5e-144

```
Query  20   MAEQQKIKKSPHVLLIPFPLQGHINPFIQFGKRLISKGVKTTLVTTIhtlnstlnhsntt  79
            MAEQ K+ KSPHVLL P+P QGHINP IQFGKRL+SKGVKTTLVTTI+ LN+ L+H   T
Sbjct  1    MAEQHKMNKSPHVLLFPYPSQGHINPIIQFGKRLLSKGVKTTLVTTIYILNTKLSHKENT  60

Query  80   ttsIEIQAISDGCDEGGFMSA--GESYLETFKQVGSKSLADLIKKLQSEGTTIDAIIYDS  137
            T+ I+I+AISDG DEGG  SA   E+YL+TFK+VGSKSLADLIKKLQSEG T+DAIIYDS
Sbjct  61   TS-IKIKAISDGFDEGGSASAESSETYLKTFKEVGSKSLADLIKKLQSEGNTVDAIIYDS  119

Query  138  MTEWVLDVAIEFGIDGGSFFTQACVVNSLYYHVHKGLISLPCGSTVSVPGLPELKHWETP  197
               W LDVA+EFGIDGGSF TQAC VNS+YYHV+KGLISLP G+ V+VPGLP L+ WETP
Sbjct  120  FMTWALDVAMEFGIDGGSFLTQACSVNSIYYHVYKGLISLPLGAPVTVPGLPLLESWETP  179

Query  198  SFVHNYGPYPGWSKTVFSQFDNIDQARWVFTNSFYELEAQVIEWMRKKWNLKVIGPTLPS  257
            SFVHNYGPYPGWS+ VF+QF NIDQARWVFTNSFY+LE +VIEWMRK WNLKVIGPTLPS
Sbjct  180  SFVHNYGPYPGWSEIVFNQFANIDQARWVFTNSFYQLEDKVIEWMRKMWNLKVIGPTLPS  239

Query  258  MYLDKRLEDDKDYGFNLYKANHNECMNWLNNKPKESVVYVSFGSSAKLEPEHMEEMAWGL  317
            MYLDKRL DDKDYGFNL+K NHN+CMNWLN KPK SVVY+SFGS+AKL+ E MEE+A GL
Sbjct  240  MYLDKRLHDDKDYGFNLFKVNHNKCMNWLNEKPKRSVVYISFGSNAKLQHEQMEEIAQGL  299

Query  318  IDSNMNFLWVVRAEEEEKLPKEFVHHKLSGKGMVVAWCRQLDVLAHESVGCFVTHCGFNS  377
             DS++NFLWVVR EEE KLPK+F+    +GKG+VVAWCRQLDVLAHESVGCFVTHCGFNS
Sbjct  300  SDSHVNFLWVVRVEEESKLPKDFMDVTENGKGLVVAWCRQLDVLAHESVGCFVTHCGFNS  359

Query  378  TLEAISLGVPVVAMPQWTDQITNAKFIDEIWGVGVRVKADENGIVRRENLASCIKTIMED  437
            TLEAISLGVPVV MPQWTDQ TNAK +DE WGVGVRVKADENGIVRR NL SCIK IME+
Sbjct  360  TLEAISLGVPVVGMPQWTDQTTNAKLLDETWGVGVRVKADENGIVRRGNLVSCIKKIMEE  419

Query  438  ERGVIVQKKTIKWRDLAKLAVDKGGSSEKDIDEFVSELLRE  478
            E+GV+ +   +KWR+LAK AVD+GGSS+KDI EFV++L  E
Sbjct  420  EKGVLARMNAVKWRELAKAAVDEGGSSDKDIHEFVNDLKHE  460
```

Figure 2. The results of the BlastP alignment between the subject sequences and UGT76G1

quence of *S. rebaudiana* MS007) is UDP-glycosyltransferase 76G1-like from *Helianthus annuus* with a high percentage identity of 77.66%. The length of the subject sequence is 455 amino acids.

The sequence identity is the number of identical bases between the query and the subject sequences. For this alignment, the sequence identity shows 77%, meanwhile the positive value shows 86% representing the number of residues that either share the same chemical properties or are identical to each other (Figure 2). In the alignment, there is a + symbol which specifies the differences between the amino acids of the query sequences and subject sequences. The residues have

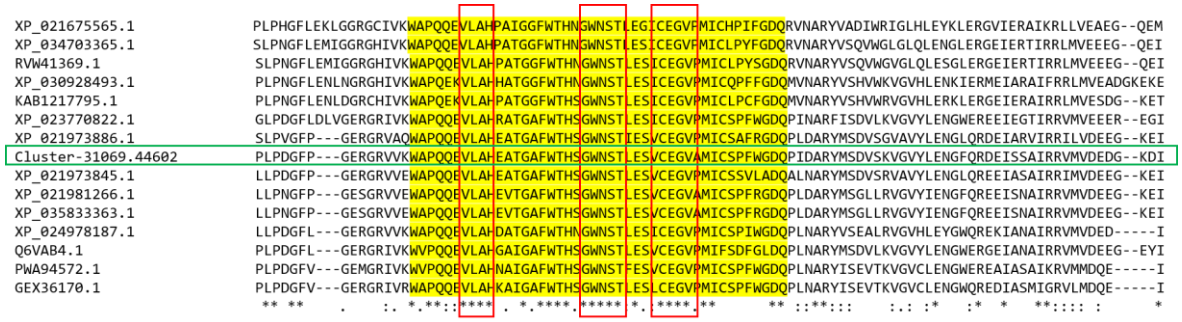


Figure 3. The highlighted region shown is PSPG motif consensus sequence of UGT76G1 protein of *Stevia rebaudiana* and other 14 protein sequences

similar chemical properties.

Protein domains and families

The crucial step in recognising the biological function of selected protein is by performing analysis of protein domains and families. In this study, the InterPro family revealed the presence of entry IPR002213 for UGT76G1 accession MS007 protein sequence which is available at positions 89 to 246. The UGT76G1 protein appeared to have a molecular function that is involved in transferring a glycosyl group from a UDP-sugar to a small hydrophobic molecule. The protein, however, was not in either the biological process or cellular component classification groups.

Multiple Sequence Alignment

Prediction of protein structure and function, phylogeny inference and sequence analysis activities require multiple sequence alignments (MSA) [28]. MSA is required to compare homologous sequences and is a prerequisite for further analyses [29].

In this study, MUSCLE [30] was used to perform multiple sequence alignment involving 15 protein sequences. From the result obtained, the asterisk (*) sign shows conserved amino acids, while the colon (:) sign signifies amino acids conservation with similar properties, whereas conservation between amino acid groups with weakly comparable properties is represented by the period sign (.) [31]. The highlighted region in Figure 3 is the plant secondary product glycosyltransferase (PSPG) motif. Highly conserved amino acids are marked with one (identity > 50%) or two (identity > 80%) asterisks below the amino acid letters [32]. Plant UGTs with the PSPG motif are soluble enzymes and play a role in bioactive natural product synthesis, plant hormone and cell homeostasis control, and xenobiotic detoxification [33, 34].

Maximum Likelihood Estimate of Substitution Matrix

	A	T/U	C	G
A	-	*4.73*	*9.73*	**13.91**
T/U	*7.14*	-	**7.60**	*11.48*
C	*7.14*	**3.70**	-	*11.48*
G	**8.66**	*4.73*	*9.73*	-

Figure 4. The maximum likelihood estimation of substitution matrix of UGT76G1 *S. rebaudiana* nucleotides sequences

The red boxes in Figure 3 represent the highly conserved sequences namely VLAH, GWSNT and CEGV that exist in all the 15 aligned sequences. The probability of substitution (r) from one base (row) to another base is indicated in each entry (column). The substitution pattern and rates were estimated using The Tamura-Nei (1993) model [35]. The rates of different transitional substitutions are bolded. Meanwhile, the trans versional substitutions are italicised (Figure 4).

When assessing instantaneous *r*, the relative values must take into consideration. In a simple way, the total of the *r* values has been set to 100. The nucleotides frequencies are A = 21.59%, T/U = 14.30%, C = 29.41% and G = 34.70%. Automated tree topology was generated to estimate ML values. The maximum Log likelihood was -4855.946. Fifteen nucleotide sequences were involved in this analysis, and $1^{st}+2^{nd}+3^{rd}$+Noncoding codon positions were enclosed. By selecting the complete deletion option, all positions with gaps and missing data were removed. There was a total of 757 positions in the final dataset. MEGA X was used to perform evolutionary analyses, resulting in the estimated values of the substitution matrix of UGT76G1 nucleotide sequences.

Based on Figure 4, the highest value of the estimation rate for transitional substitution between nucleotide A and G is 13.91. In addition, the estimation rate for two transversion substitutions

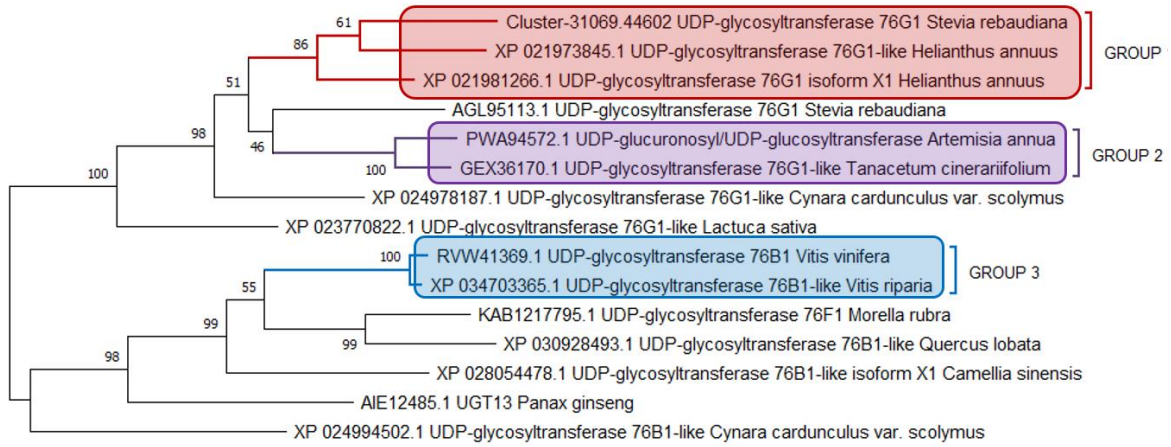


Figure 5. Phylogenetic tree of UGT76G1 using Maximum Likelihood method. The branch lengths are the same as developmental distances utilised in the JTT model with amino acid substitution per site to estimate the evolution process of the phylogenetic tree.

between nucleotide C and T/U showed the lowest value which is 3.70. The substitution of the nucleotide will be easier to perform when the estimation value is lower, meanwhile if the value is higher than the substitution, it is harder to do due to the long distances between nucleotides [36].

Phylogenetic tree analysis

The phylogenetic tree visually illustrates evolutionary relationships between different biological species [37]. This study implemented the Maximum Likelihood method and JTT matrix-based model to derive the evolutionary history. The evolutionary history of the taxa studied is des-cribed by the bootstrap consensus tree inferred from 1,000 replications [36]. This analysis involved 15 amino acid sequences selected from the homology search result, and the complete deletion option removed all positions with gaps and missing data. There was a total of 269 positions in the final dataset. Next to the branches are the tree percentages in which the related taxa are grouped together in the bootstrap test (1000 replications). An initial tree is first constructed quickly but not optimally. Then other topology variations are formed using the NNI (nearest neighbour interchange) approach to look for more accurate topologies that match the data. The tree shows the highest log likelihood of -4286.06 (Figure 5).

The resulting phylogenetic trees can then be divided into three main groups, which refer to the bootstrap value for each clade. All organisms in groups I and II are members of the Asteraceae family, specifically *S. rebaudiana* MS007 (refer to UGT76G1), *H. annuus*, *T. cinerariifolium*, and *A. annua*, which are known for their single-seeded achene fruits and composite flower heads. This indicates that our UG76G1 from *S. rebaudiana* MS007 is quite similar to *H. annus*, with a bootstrap value of greater than 50%, according to the consensus tree (Figure 5). This is in contrast to group 3, where all members of this group come from the Vitaceae family i.e. *V. vinifera* and *V. riparia*.

Physicochemical properties of UGT76G1

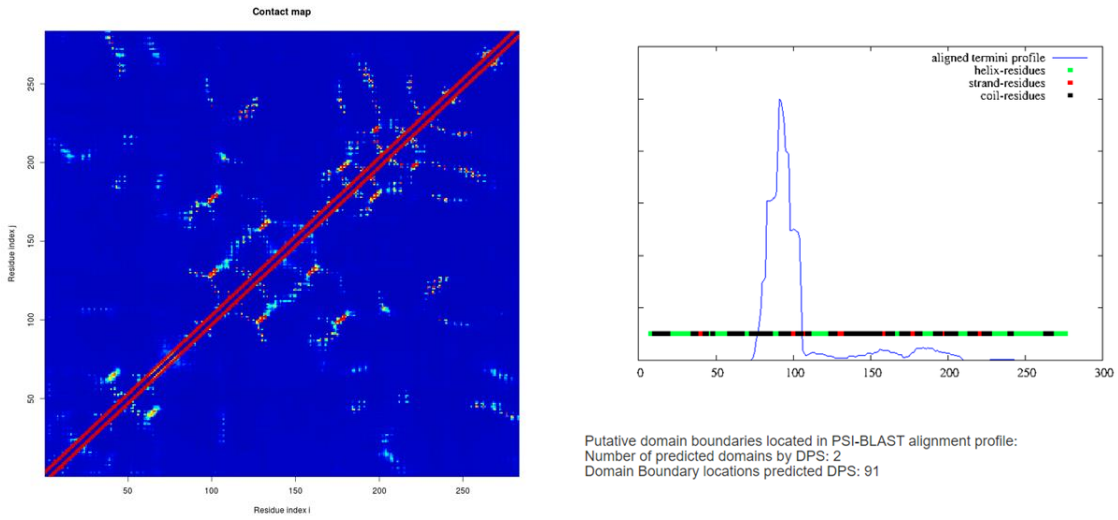
According to physicochemical characteristics in Table 2, the value of the isoelectric point (pI) for UGT76G1 MS007 protein is 5.17. These specifications are needed, in particular, for experimental handling methods, primarily for protein isolation and purification, in order to determine the state of the protein sequence [20]. The highest extinction coefficient (EC) of UGT76G1 MS007 is 54555 M^{-1} cm^{-1} and based on instability index (II), UGT76G1 is predicted to be stable inside a test tube. Table 2 includes a more comprehensive overview of UGT76G1 MS007 protein parameters, such as aliphatic index, molecular weight, and grand average of hydropathy. This information can be used to predict the properties of a protein and are able to help in empirical research [20].

Secondary and tertiary structure prediction

PSIPRED had been used to predict the secondary structure of UGT76G1 MS007, where sequence plot, PSIPRED cartoons, DeepMetaP

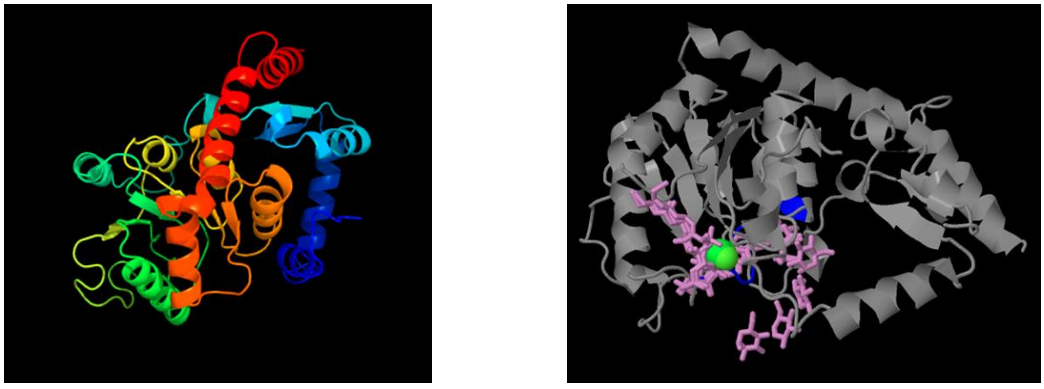
Table 2. Physicochemical characteristics of UGT76G1 *S.rebaudiana* MS007 by ExPasy ProtParam.

Type	Value
Number of amino acids	283
Molecular weight	32249.87
Theoretical pI	5.17
Formula	$C_{1463}H_{2251}N_{373}O_{427}S_{11}$
Total number of atoms	4525
Extinction coefficient (EC)	54555
Extinction coefficient	54430
Instability index (II)	39.96 (Unstable)
Aliphatic index (AI)	87.42
Grand average of hydrophaticy (GRAVY)	-0.208



a b

Figure 6. Secondary Structure Prediction of UGT76G1 MS007. (a) DeepMetaPSICOV Contact Map and (b) DOMPred results of UGT76G1 *S. rebaudiana* MS007.



a b

Figure 7. Tertiary structure prediction of UGT76G1 *S. rebaudiana* MS007 (a) The model with has a confidence level of 100.0% and a coverage of 98% based on the single highest scoring template. (b) Structural model prediction with the predicted binding site (blue) and other residues (grey).

SICOV contact map and DomPred results were obtained. Figure 6 (a) represents the DeepMetaP-SICOV contact map of UGT76G1 which shows a directly proportional relationship between the x-axis and the y-axis of the graph and Figure 6 (b) shows the aligned termini profile of UGT76G1 at its peak at the scale of approximately 90.

The tertiary structure prediction was done by

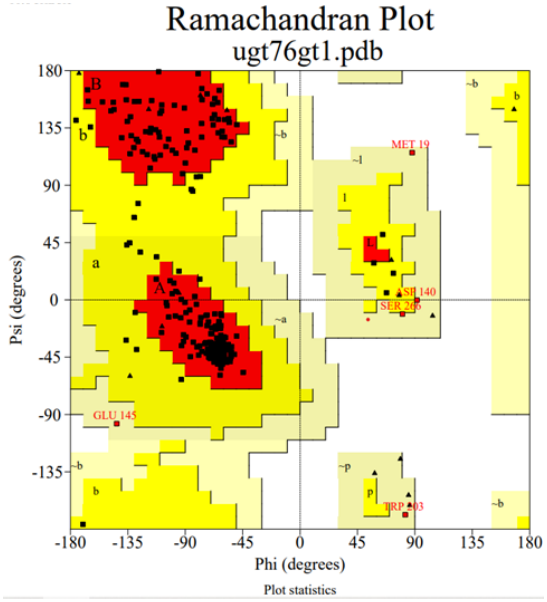


Figure 8. The Ramachandran plot reveals that there were 89.8% of residues in the most liked region, 8.1% of residues in the permitted region, and 1.8% of residues were present in the generously area.

performing template-based homology modelling or fold recognition using Phyre2 Protein Fold Recognition Server [38]. The model was based on template 6O86 from Protein Data Bank (PDB) with the title as crystal structure of SeMet UDP-dependent glucosyltransferase (UGT) from *S. rebaudiana* in complex with UDP [39]. The template structure is a glycosyltransferase molecule containing chain A with 458 of sequence length. Figures 7 (a) and (b) demonstrate the UGT76G1 MS007 tertiary structure prediction outcomes. The position where uridine diphosphate (UDP)-dependent glucosyltransferases would bind to the UGT76G1 enzyme was predicted in Figure 7 (b). The structure of UGT76G1 *S. rebaudiana* MS007 was predicted in this analysis.

Model assessment

The structural model assessment was verified through the SAVES server (PROCHECK, ERRAT, PROVE, and Verify3D) (saves.mbi.ucla.edu/). The PROCHECK program will generate the Ramachandran plot by calculating the residue-by-residue stereochemical quality of the structure of UGT76G1 *S. rebaudiana* MS007 (Figure 8).

The structural model plot demonstrated that no residue was situated in a disallowed area; 89.8% of the residues were positioned in the most liked region and the remaining residues were placed in the additional and generously allocated region (Figure 8). A plot result of greater than 85% suggests that the projected model is of acceptable quality.

The assessment of the proposed model structure using ERRAT yielded a score of 85.24% (Figure 9). ERRAT analyses data from highly refined protein structures to discover local faults within the geometry of a protein structure in order to determine the non-bonded atomic interactions quality factor in general [40]. The high-quality model was indicated by the high ERRAT score is more than 50%.

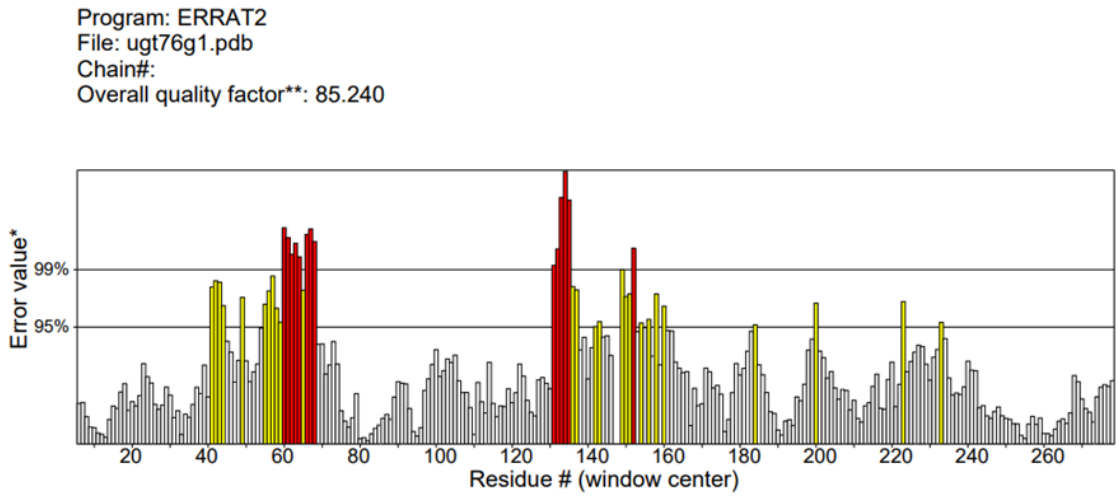


Figure 9. The ERRAT tool results reveal the acceptable model with the overall quality factor of the modelled protein is 85.24% depending on multiple sorts of atoms

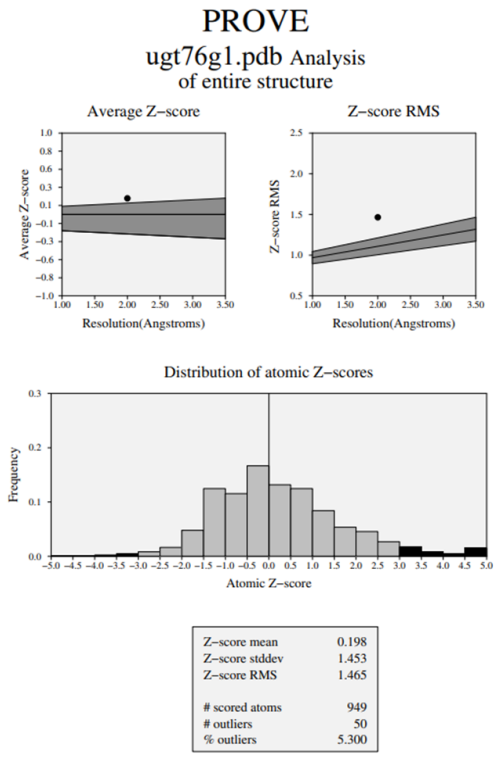


Figure 10. The average Z-score was 0.198 and the Z-score RMS was 1.465

The validation using Protein Volume Evaluation (PROVE), an online server, confirmed the whole structure of the displayed protein. The quality of protein is assessed using volume Z-scores, which represent deviations from the conventional values of atomic volumes. The Z-score is an analytical measurement for evaluating the data's capacity to see the nearest overlay based on possible outcomes. Figure 10 shows that the model protein's normal Z-score was 0.198 and the Z-score RMS was 1.465.

VERIFY 3D analysis revealed that 80.65% of the residues had an average 3D-1D score of more than 0.2. Only 19.35% of the residues were found to be non-complementary to the 3D-1D profile. The VERIFY 3D engages statistical method to determine the compatibility of an atomic (3D) model with its own amino acid (1D) by classifying structures depending on positions and surroundings then evaluating the results in comparison to those of more advanced structures [38]. When the score of VERIFY 3D is greater than 80%, the model is of good quality. Overall, the predicted model structure was found to be good, stable, and sufficient, based on the four assessment program scores.

Conclusion

The putative UDP-glycosyltransferase 76G1 (UGT76G1) protein in *S. rebaudiana* variety MS007 was effectively discovered and described through an *in-silico* approach. The protein sequence and domain analysis strongly imply that it belongs to UDP-sugar with a glycosyl group from UDP glycosyltransferases (UGT) superfamily. The phylogenetic analysis between UGT76G1 and other homologous proteins found that UDP-glycosyltransferase 76G1 from *S. rebaudiana* variety MS007 came from the same common ancestor which is the Asteraceae family. Analysis of the gene sequence and 3D model indicates that it is 98% similar compared to its template (PDB ID: 6O86) and based on structure verification, the UGT76G1 model variety MS007 was found to be acceptable, stable, and adequate. This work effectively addressed the knowledge gap of the previously unannotated UGT76G1 protein in *S. rebaudiana* MS007 by applying *in silico* sequence- and structure-based strategies. As a result, utilizing this information to exploit the structure in order to manufacture artificial sweeteners to meet customer demand would be extremely beneficial.

Acknowledgment

We would like to express our gratitude to International Islamic University Malaysia for providing financial assistance under the Research Acculturation Grant Scheme (IRAGS18-036-0037).

References

1. Yadav AK, Singh S, Dhyani D, Ahuja PS (2011) A review on the improvement of stevia [*Stevia rebaudiana* (Bertoni)]. Canadian Journal of Plant Science 91(1):1-27. DOI: 10.4141/CJPS10086.
2. Samsulrizal NH, Zainuddin Z, Noh AL, Sundram TC (2019) A review of approaches in steviol glycosides synthesis. International Journal of Life Sciences and Biotechnology 2(3):145-57. DOI: 10.38001/ijlsb.577338.
3. Razali A, Samsulrizal NH, Zainuddin Z (2020) Identification of genes involved in flowering in *Stevia rebaudiana* using expressed sequence tags (ESTs). Asia-Pacific Journal of Molecular Biology and Biotechnology 28(2):105-12. DOI: 10.35118/apjmbb.2020.028.2.09.
4. Goyal SK, Samsher, Goyal RK (2010) Stevia (*Stevia rebaudiana*) a bio-sweetener: a review. International journal of food sciences and nutrition 61(1):1-0. DOI: 10.3109/09637480903193049.
5. Lemus-Mondaca R, Vega-Gálvez A, Zura-Bravo L, Ah-Hen K (2012) *Stevia rebaudiana* Bertoni, source of a high-potency natural sweetener: A comprehensive review on the biochemical, nutritional and functional aspects.

Food chemistry 132(3):1121-32. DOI: 10.1016/j.foodchem.2011.11.140.
6. Singh SD, Rao GP (2005) Stevia: The herbal sugar of 21st century. Sugar tech 7(1):17-24. DOI: 10.1007/BF02942413.
7. Kumar H, Kaul K, Bajpai-Gupta S et al. (2012) A comprehensive analysis of fifteen genes of steviol glycosides biosynthesis pathway in *Stevia rebaudiana* (Bertoni). Gene 492(1):276-84. DOI: 10.1016/j.gene.2011.10.015.
8. Richman A, Swanson A, Humphrey T et al. (2005) Functional genomics uncovers three glucosyltransferases involved in the synthesis of the major sweet glucosides of *Stevia rebaudiana*. The Plant Journal 41(1): 56-67. DOI: 10.1111/j.1365-313X.2004.02275.x.
9. Yang YH, Huang SZ, Han YL et al. (2014) Base substitution mutations in uridinediphosphate-dependent glycosyltransferase 76G1 gene of *Stevia rebaudiana* causes the low levels of rebaudioside A: mutations in UGT76G1, a key gene of steviol glycosides synthesis. Plant physiology and biochemistry 80: 220-5. DOI 10.1016/j.plaphy.2014.04.005.
10. Shibata H, Sawa Y, Oka TA et al. (1995) Steviol and steviol-glycoside: glucosyltransferase activities in *Stevia rebaudiana* Bertoni-purification and partial characterization. Archives of biochemistry and biophysics 321(2): 390-6. DOI: 10.1006/abbi.1995.1409.
11. Shibata H, Sonoke S, Ochiai H et al.(1991) Glucosylation of steviol and steviol-glucosides in extracts from *Stevia rebaudiana* Bertoni. Plant physiology 95(1): 152-6. DOI: 10.1104/pp.95.1.152.
12. Brandle JE, Telmer PG. Steviol glycoside biosynthesis (2007) Phytochemistry 68(14): 1855-63. DOI: 10.1016/j.phytochem.2007.02.010.
13. Othman HS, Osman M, Zainuddin Z (2018) Genetic variabilities of *Stevia rebaudiana* Bertoni cultivated in Malaysia as revealed by morphological, chemical and molecular characterisations. AGRIVITA, Journal of AgricuLtural Science 40(2): 267-83. DOI: 10.17503/agrivita.v40i2.1365.
14. Samsulrizal NH, Khadzran KS, Sundram TS et al. (2021) Transcriptome profiling of Stevia rebaudiana MS007 revealed genes involved in flower development. Turkish Journal of Biology 45(3): 314-22. DOI: 10.3906/biy-2103-3.
15. Mirzaei AR, Shakoory-Moghadam V (2022) Bioinformatics analysis and pharmacological effect of Stevia rebaudiana in the prevention of type-2 diabetes. Cellular, Molecular and Biomedical Reports, 2(2), pp.64-73.
16. Samsulrizal NH, Khadzran KS, Shaarani SH et al. (2020) De novo transcriptome dataset of *Stevia rebaudiana* accession MS007. Data in brief 28. DOI: 10.1016/j.dib.2019.104811.
17. Gasteiger E, Gattiker A, Hoogland C et al. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic acids research 31(13): 3784-8. DOI: 10.1093/nar/gkg563.
18. Mitchell AL, Attwood TK, Babbitt PC et al. (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic acids research 47(D1): D351-60. DOI: 10.1093/nar/gky1100.
19. Bateman A, Coin L, Durbin R, Finn RD et al. (2004) The Pfam protein families database. Nucleic acids research 32(suppl_1): D138-41. DOI: 10.1093/nar/gkh121.
20. Gasteiger E, Hoogland C, Gattiker A et al. (2005). Protein identification and analysis tools on the ExPASy server. In: John M. Walker (Ed): The Proteomics Protocols Handbook, 571–607.
21. Hall BG (2013) Building phylogenetic trees from molecular data with MEGA. Molecular biology and evolution 30(5): 1229-35. DOI: 10.1093/molbev/mst012.
22. Tamura K, Peterson D, Peterson N et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular biology and evolution 28(10): 2731-9. DOI: 10.1093/molbev/msr121.
23. Buchan DW, Jones DT. (2019) The PSIPRED protein analysis workbench: 20 years on. Nucleic acids research 47(W1): W402-7. DOI: 10.1093/nar/gkz297.
24. Kelley LA, Mezulis S, Yates CM et al. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nature protocols 10(6): 845-58. DOI: 10.1038/nprot.2015.053.
25. Lovell SC, Davis IW, Arendall III WB et al. (2003) Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins: Structure, Function, and Bioinformatics 50(3): 437-50. DOI: 10.1002/prot.10286.
26. Tomii K, Hirokawa T, Motono C (2005) Protein structure prediction using a variety of profile libraries and 3D verification. Proteins: Structure, Function, and Bioinformatics 61(S7): 114-21. DOI: 10.1002/prot.20727.
27. Bowie JU, Lüthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016): 164-70.
28. Edgar RC, Batzoglou S. (2006) Multiple sequence alignment. Current opinion in structural biology 16(3): 368-73. DOI: 10.1016/j.sbi.2006.04.004.
29. Wallace IM, Blackshields G, Higgins DG (2005) Multiple sequence alignments. Current Opinion in Structural Biology 15(3) 261–266. DOI: 10.1016/j.sbi.2005.04.002.
30. Edgar RC. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32(5):1792-7. DOI: 10.1093/nar/gkh340.
31. Madeira F, Park YM, Lee J et al. (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic acids research 47(W1): W636-41. DOI: 10.1093/nar/gkz268.
32. Wang J, Hou B (2009) Glycosyltransferases: key players involved in the modification of plant secondary metabolites. Frontiers of Biology in China 4(1): 39-46. DOI: 10.1007/s11515-008-0111-1.
33. Osmani SA, Bak S, Møller BL (2009) Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling. Phytochemistry 70(3): 325-47. DOI: 10.1016/j.phytochem.2008.12.009.
34. Rahmatullah-Khan A, Mokhtar NI, Zainuddin Z, Samsulrizal NH (2021) In silico characterization of UGT74G1 protein in Stevia rebaudiana Bertoni Accession MS007. Journal of Tropical Life Science 11(3). DOI: 10.11594/jtls.11.03.09.
35. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular biology and evolution 10(3): 512-26. DOI: 10.1093/oxfordjournals.molbev.a040023.
36. Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis

across computing platforms. Molecular biology and evolution 35(6): 1547. DOI: 10.1093/molbev/msy096.
37. Roy SS, Dasgupta R, Bagchi A (2014) A review on phylogenetic analysis: a journey through modern era. Computational Molecular Bioscience 4(3): 39. DOI: 10.4236/cmb.2014.43005.
38. Kelley LA, Sternberg MJ (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nature protocols 4(3): 363-71. DOI: 10.1038/nprot.2009.2.
39. Lee SG, Salomon E, Yu O, Jez JM (2019) Molecular basis for branched steviol glucoside biosynthesis. Proceedings of the National Academy of Sciences 116(26): 13131-6. DOI: 10.1073/pnas.1902104116.
40. Yusof NY, Firdaus-Raih M, Mahadi NM et la. (2017) In silico analysis and 3D structure prediction of a chitinase from psychrophilic yeast Glaciozyma antarctica PI12. Malaysian Applied Biology, 46, pp.117-123.

This page is intentionally left blank.