Plant Omics Journal POJ

POJ 5(3):200-210 (2012) ISSN:1836-3644

In silico motif diversity analysis of the glycon preferentiality of plant secondary metabolic glycosyltransferases

Ritesh Kumar, Rajender S. Sangwan , Smrati Mishra, Farzana Sabir and Neelam S. Sangwan*

Metabolic and Structural Biology Department, Central Institute of Medicinal and Aromatic Plants (CSIR-CIMAP), P.O.-CIMAP, Lucknow-226015, India

***Corresponding author: sangwan.neelam@gmail.com**

Abstract

Glycosyltransferase are the class of enzymes which specifically glycosylate various natural and artificial substrate aglycons into their glycosidic linked compounds with enhanced water solubility and transport. In several instances, glycosylation is the last step in the biosynthesis of a number of secondary plant products involving flavonoids, terpenoids, steroidal alkaloids, and saponin biosynthetic pathway. The conjugation reactions catalyzed by UGTs may therefore are critical in regulating the levels of several secondary metabolites including signaling and hormonal compounds. In this work we have analyzed genes from the databases for the presence of GT's in diverse plant families. Considerable degree of homology was seen in alignment of all available GT sequences in dicot plants as revealed by ClustalW2 and other phylogenetic tree constructing tools. Also, a highly conserved motif in their C-terminus, named the PSPG box (Plant secondary product glycosyl transferase) was found in all the sequences through motif discovery tools e.g. MEME. The motif discovery tool identified two other distinct motifs in GT sequences, however interestingly *P. patens* and a putative GT sequence from *A. thaliana* was found to be deficient in motif 3 at N terminal of the sequence. A wide range of gene sequences were analyzed in a systematic manner to determine the structure, function and evolution of PSPG box motif found at the C-terminal.

Keywords: Conserved motif, glycosyltransferase, PSPG box, secondary metabolites.
Abbreviations: Multiple sequence alignment MSA; Plant secondary product glycosyltransferase PSPG; glycosyltransferase GT; uridine diphosphate glucose, UDPG; UDP-dependent glycosyltranferases (UGTs); CAZy (Carbohydrate-active enzymes); nucleotide diphosphate (NDP).

Introduction

Glycosyltransferases, members of a multigene superfamily in plants, are ubiquitous group of enzymes that catalyze glycosylation reactions. Glycosylation is a very widespread conjugative modification of plant secondary metabolites that involves transfer of a single or multiple sugar units from activated sugars (e.g. uridine diphosphate glucose, UDPG) to a range of phytochemicals leading to the forming glycosidic bond(s). Glycosylation reactions are integral to several specific plant functions like the regulation of hormone homeostasis (Bowles et al., 2005), detoxification of xenobiotics (Loutre et al., 2003), biosynthesis and storage of secondary compounds. GTs is found in organisms in all the three kingdoms of life forms (plants, animals and microbes). In mammals, glycosylation plays an important role in drug detoxification, while in plants glycosylation often constitutes the last step in the biosynthesis of numerous plant natural products of chemical classes- terpenoids, phenylpropanoids, cyanogenic glycosides and alkaloids (Masao et al., 2000). However, in certain cases, it also constitutes an intermediary step e.g. secologanin biosynthesis in *Catharanthus roseus*. Plant UDP-dependent glycosyltranferases (UGTs) catalyzes glycosylation of various secondary metabolites, and xenobiotics alters properties of acceptor aglycones in terms of their hydrophilicity, stability, chemical interactivity/binding with other molecules including binding with macromolecules, intracellular localization etc. (Bowles et al., 2005; Kristensen et al., 2005; Kramer et al., 2003). Thus, glycosylation plays an important role in maintaining cellular homeostasis of glycosylated and non-glycosylated molecules, buffering the impact of xenobiotic challenges to the plant (Loutre et al., 2003), regulation of plant growth and development, defence response to stresses (Jones and Vogt, 2001; Hou et al., 2004).

UGTs roughly correspond to CAZy (Carbohydrate-active enzymes) family 1, a classification scheme that now describes more than 91 distinct families of CAZy GTs at the CAZY database (www.cazy.org). This classification is based on the nature of substrates accepted by the enzymes and the score of their sequence identity (Campbell et al., 1997; Coutinho et al., 2003). Plants, in contrast with the other kingdoms, have notably many more UGT genes in their genome. *Arabidopsis thaliana*, for example, contains about 120 UGT encoding genes (Claire et al., 2005; Sarah et al., 2009). Phylogenetic analyses have divided these 120 UGTs into three clades - two minor clades having sterol and lipid GTs as members and a major monophyletic clade of plant secondary metabolism UGTs characterized by the presence of a highly conserved motif called plant secondary product glycosyltransferase (PSPG) box, designated so by Hughes and Hughes (1994) for plant enzymes. The PSPG box represents nucleotide diphosphate (NDP) sugar binding site (Vogt, 2002) and spans as 44 amino acid residues towards C-terminal end as typical of all UGTs. UGTs transfer uridine-

diphosphate (UDP) activated glucose to low molecular weight acceptor substrates. The activated sugar form can also be UDP-galactose, UDP-rhamnose, UDP-xylose, UDP-glucuronic acid (Fig. 1) (Merken and Beecher, 2000). These single or multiple glycosylation at the same site in series or at multiple sites on the same acceptor molecule can occur at -OH, -COOH, -NH2, -SH, and at C-C groups (Bowels et al., 2006; Breton et al., 2006). Plant family 1 UDP-dependent glycosyltransferases (UGTs) catalyze the glycosylation of a plethora of bioactive natural products which are immensely important for the pharmaceutical and medicinal purposes. Family 1 UGTs have shown that there is greater variability within the N-terminal regions of these proteins than in their C-terminal regions, and have indicated that amino acid residues in the N-terminal half of the proteins were responsible for acceptor binding, whereas those in the C-terminal half were involved mainly in interactions with donor substrates (Bowles et al., 2005, 2006). In present study, we have analyzed the features of C-terminal region located PSPG box of the secondary metabolism GTs among distinct members of plant kingdom and their evolutionary relatedness based on the complete as well as PSPG-specific amino acid sequences. Further, we have identified other highly conserved/semi-conserved motifs of the UGTs and analyzed them in the perspectives of aglycone substrate acceptability and modulation of the catalytic glycosylation. The analysis could be particularly relevant to design of novel biocatalysts for the production of therapeutic or otherwise useful glycosylated products of terpenoids, steroids, flavonoids etc.

Results

Multiple sequence alignment, and phylogenetic tree construction and analysis

All 40 GT amino acid sequences (Table 1) were subjected to multiple sequence alignment (MSA) using ClustalW2 to comparative localization of their PSPG box motif of 44 amino acid residues near the C-terminal of the sequence. The PSPG box consensus sequence of plant glycosyltransferases, obtained through MSAs shown in Fig. 2a. Phylogenetic tree of the sampled GT sequences was constructed based on the full length of amino acid sequences (Fig. 3) as well as on the basis of PSPG box amino acid sequences for all the 40 GTs from different plants (Fig. 4). The phylogenetic tree based on whole amino acid sequences (Fig. 3) formed three main clusters from the root, one major cluster, two minor clusters, and one the GT from the set (i.e. from *Fragaria x ananassa*, ABB92749.1) appeared to have evolved independently from the phylogenetic root. Major cluster comprised, 35 GTs from 35 different plants whilst 2 GTs from 2 different plants formed a minor cluster. The phylogenetic tree based on amino acid sequences of only PSPG box (Fig. 4) also formed three clusters from the phylogram root, one major cluster with 32 GTs from 32 different plants, and two minor clusters, one of three GTs and other of five GTs. GTs from the two representatives from Poaceae, *T. aestivum* (ACB47884.1) and *Secale cereale* (ACR43490.1), clad together, in both the phylogenetic trees (Fig. 3, 4), but in case of the phylogenetic tree based on the whole of amino acid sequences of GTs (Fig. 3) appeared to have differentiated last from their common ancestor in comparison to the phylogenetic tree based on amino acid sequences of only PSPG box (Fig. 4), in evolutionary time period. This also suggests that the amino acid sequence divergence started earlier in PSPG box to get nearly fixed but amino acid sequences in the rest of the protein still remained subject to variation. It may also indicate a functional freezing of the PSPG box able to afford limited variation further. This means that the amino acid diversification in PSPG box started later in this case or still continued whilst the rest of sequence of the GT declined to vary significantly.

GTs from *R. communis* (XP_002518722.1) and *V. vinifera* (CAN65903.1) that clad together evolved earliest on the whole amino acid sequences of GT (Fig. 3), but later in terms of the amino acid sequences of PSPG box (Fig. 4), to the extent that they clad separately-*V. vinifera* GT (CAN65903.1) was laid individually, while *R. communis* GT (XP_002518722.1) clad with *L. japonicum* GT (BAI63589.1) in one sub-cluster of the main cluster. This might imply that the amino acid diversification in the PSPG box started later, than the other part of the GT, and the amino acid diversification in other part of GT except PSPG box is more significant than the amino acid sequence diversification in the PSPG box, because based on amino acid sequence similarity of whole GT they clad with each other, but on the basis of amino acid sequence of only PSPG box they were evolved from the common ancestor but they do not clad, rather were present in different sub-cluster. GT of *F. x ananassa* (ABB92749.1) was also located individually and evolved from the root of the phylogenetic tree based on the whole amino acid sequences of GT (Fig. 3). This means that the amino acid sequence diversification of whole GT is more significant than the amino acid diversification of only PSPG box, and the amino acid diversification in PSPG box started later than the other part of the whole GT amino acid sequence, in evolutionary time period. Six pair of plants (*Glycyrrhiza echinata* (BAC78438.1) and *Medicago trancatula* (ACT34898.1), *Vingo mungo* (BAA36410.1) and *Cicer arietinum* (CAB88666.1), *Triticum aestivum* (ACB47884.1) and *S. sereal* (ACR43490.1), *Forsythia x intermedia* (BAI65914.1) and *Sesamum indicum* (BAF96582.1), *Nicotiana tobaccum* (AAK28303.1) and *Withania sominifera* (ACD44747.1), and *Eustoma grandiflorum* (BAF49308.1) and *Catharanthus roseus* (BAD 29722.1)) were present together in both the phylogenetic trees (Fig. 3 and 4) with the specificity that both the members of each pair clad with each other, but evolved at different time. Five pair of organisms (*Allium cepa* (AAP88405.1) and *Zea mays* (NP_001148090), *Sorghum bicolor* (XP_002456026) and *Oryza sativa* Indica group (EAY74811.1), *Scutellaria baicalensis* (BAA83484.1) and *Perilla frutescens* (BAG31952.1), *Puraria montana* var. lobata (ACJ72160.1) and *Lotus japonica* (BAI83589.1), and *Populus trichocarpa* (XP_002298737) and *Lycium barbarum* (BAG 80549.1)) were present only in the phylogentic tree based on the whole GT amino acid sequences (Fig. 3), not in the phylogenetic tree based on the amino acid sequences of only PSPG box, with the specificity that both the members of each pair clad to each other. Eight pair of plants (*Glycine max* (ABB85236.1) and *Picea sitchensis* (ABR17572.1), *Arabidopsis thaliana* (NP_181215.1) and *Stevia rebaudiana* (AAR06917.1), *Citrus sinensis* (ACS87992.1) and *Zea mays* (NP_001148090), *Avena strigosa* (ACD03255.1) and *Oryza sativa* Indica group (EAY74811.1), *R. communis* (XP_002518722) and *L. japonicum* (BAI63589.1), *Phytalacca americana* (BAG71127.1) and *L. barbarum* (BAG80549.1), *Dianthus caryophyllus* (BAD52006.1) and *Beta vulgaris* (AAS94329.1), and *Antrrhinum majus* (BAG31950.1) and *P. frutescens* (BAG31952.1)) were present only in the phylogenetic tree based on the amino acid sequences of only PSPG box (Fig. 4), not in the phylogentic tree based on the whole GT amino acid sequences (Fig. 3), with the specificity that both the member of each pair clad to

each other. Two pairs of GTs- *A. thaliana* (NP_181215.1) and *C. sinensis* (ACS87992.1), and *P. americana* (BAG71127.1) and *B. vulgaris* (AAS94329.1) were in the one clade in the phylogenetic tree based on the whole GT amino acid sequences (Fig. 3), but not so in the phylogenetic tree based on the amino acid sequences of only PSPG box (Fig. 4). *A. thaliana* (NP_181215.1) clades with *S. rebaudiana* (AAR06917.1) instead of *C. sinensis* (ACS87991.1), *C. sinensis* (ACS87991.1) clades with *Z. mays* (NP_00114090) instead of *A. thaliana* (NP_181215.1), similarly *P. americana* (BAG71127.1) clades with *Lycium barbarum* (BAG80549.1) instead of *B. vulgaris* (AAS94329.1), and *B. vulgaris* (AAS94329.1) clades with *D. caryophyllus* (BAD52006.1) in the phylogenetic tree based on the amino acid sequences of their PSPG boxes. *Forsythia x intermedia* and *Sesamum indicum*, however, exhibited maximal proximity based on their sequences of whole protein as well as PSPG motif. This may specifies that though PSPG box (located at C-terminal) is an important motif of GTs to govern donor NDP-sugar specificity, nevertheless conservations and divergences elsewhere (towards N terminal region) of the protein significantly influenced the cladding of the plant GTs in the phylogenetic tree and in their evolutionary time period. Functionally, part of these sequences may comprise of conserved/semi-conserved motifs/amino acid residues affecting the catalytic and kinetic properties of glycosylation including specificity of the sugar acceptor substrate.

Motif discovery

Three sets of data were taken for this investigation. As most of the GTs amino acid sequences are reported in *A. thaliana*, the first set sampled for analysis comprised of 14 *A. thaliana* GTs representing amino acid diversity. The second set as represented in supplementary Table 2. consisted of members of four GTs from four different plants representing highest level of diversity in terms of their amino acid chain length viz. *P. patens* subsp. *patens* (265 amino acids), *Z. mays* (525 amino acids), *T. aestivum* (496 amino acids), *S. cereale* (496 amino acids). In the third set, all the 40 sequences retrieved from NCBI, were considered for motif analysis (Table 2, supplementary Table 3). Motif discovery operation was performed, separately, through motif discovery tool MEME and Glam2 on these three datasets. For the first set, all, except one putative glycosyl transferase showed the existence of three motifs, motif 1 near the C-terminal, motif 2 in middle toward motif 1, motif 3 toward the N-terminal (supplementary Fig 1, Fig. 5, 6). Exceptional member in the set, GT AAD17393.1 lacked motif 3, near N-terminal. For second set of four GTs, all except the smallest (265 amino acids) GTs (*P. patens subsp. patens* (XP_001765134.1) possessed three motifs, motif 1 near the C-terminal, motif 2 in middle toward motif 1, motif 3 toward the N-terminal (supplementary Fig 2, Fig 5, 6). The small size *P. patens subsp. patens* GT (XP_001765134.1) lacked the motif 3 near N-terminal. Two GTs-glucuronosyl transferase of *T. aestivum* (ACB47884.1) and glucosyl transfearse of *S. cereale* (ACR43490.1) which were of the same size (496 amino acids) matched well for the motif 1, 2, and 3 in term of their positions and size. Cyto-O-transferase of *Z. mays* (NP_001148090.1), the biggest in size exhibited little shift in the entire three motif toward N terminal. *P. patens* (XP_001765134.1) was the smallest in size and lacked motif 3, had a considerable shift in motif 1, 2 towards N terminal. For third set, all GTs from all organism except UDP-glucoronosyl/UDP-glucosyl transferase protein of *T. aestivum* (ACB47884.1), UDP-glucosyl transferases of *S. cereale* (ACR43490.1), and *P. patens* (XP_001765134.1) reflected consistent presence of the three motifs (Fig 5, 6, 7). Putative UDP-glucose of *C. arietinum* (CAB88666.1) had the same three motifs but all shifted toward N-terminal side of the gene sequence. UDP-glucoronosyl/UDP-glucosyl transferase protein of *T. aestivum* (ACB47884.1), UDP-glucosyl transferase of *S. cereale* (ACR43490.1), both of same amino acid length, possessed an additional motif 1 near N-terminal besides the above three motifs (Fig. 5).

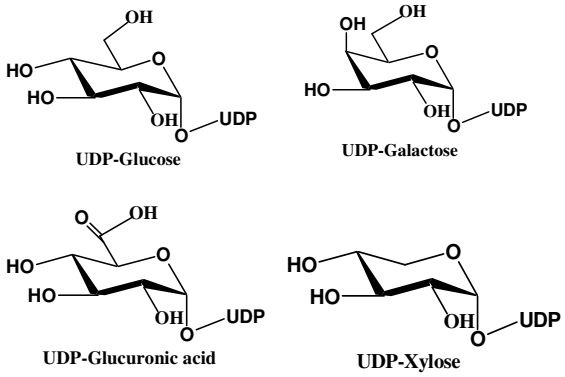

Fig 1. Major sugar donors utilized by the plant glycosyltransferases.

Discussion

In plants, enzymes of GT class are known to recognize a great diversity of substrates including hormones, secondary metabolites and xenobiotics such as pesticides and herbicides (Bowles et al., 2006). The sugar donor is generally UDP-glucose, although UDP-rhamnose, UDP-galactose and UDP-xylose have also been identified as activated sugars for the transfer reactions (Bowles et al., 2005; He et al., 2006). There

Table 1. List of GTs accessions utilized in the analysis.

S.N.	Accession No.	Plants
1	ABB85236.1	*Glycine max*
2	BAC78438.1	*Glycyrrhiza echinata*
3	ACT34898.1	*Medicago truncatula*
4	BAA36410.1	*Vigna mungo*
5	XP_002518722.1	*Ricinus communis*
6	XP_002298737.1	*Populus trichocarpa*
7	ACJ72160.1	*Pueraria montana var. lobata*
8	CAB88666.1	*Cicer arietinum*
9	ABB92749.1	*Fragaria x ananassa*
10	XP_002456026.1	*Sorghum bicolor*
11	BAF75890.1	*Dianthus caryophyllus*
12	BAG71127.1	*Phytolacca americana*
13	CAN_65903.1	*Vitis vinifera*
14	BAG80549.1	*Lycium barbarum*
15	NP_181215.1	*Arabidopsis thaliana*
16	EAY74811.1	*Oryza sativa Indica Group*
17	ACD03255.1	*Avena strigosa*
18	AAK28303.1	*Nicotiana tabacum*
19	BAG31950.1	*Antirrhinum majus*
20	BAF49308.1	*Eustoma grandiflorum*

Table 1. Continued.

21	BAA83484.1	*Scutellaria baicalensis*
22	AAS55083.1	*Rhodiola sachalinensis*
23	ACO44747.1	*Withania somnifera*
24	BAD29722.1	*Catharanthus roseus*
25	BAG31952.1	*Perilla frutescens*
26	BAI63589.1	*Lotus japonicum*
27	AAS94329.1	*Beta vulgaris*
28	CAA59450.1	*Solanum lycopersicum*
29	CAB56231.1	*Dorotheanthus bellidiformis*
30	ACB47884.1	*Triticum aestivum*
31	ACR43490.1	*Secale cereale*
32	BAF96582.1	*Sesamum indicum*
33	NP_001148090.1	*Zea mays*
34	AAP88405.1	*Allium cepa*
35	XP_001765134.1	*Physcomitrella patens subsp. patens*
36	ABR17572.1	*Picea sitchensis*
37	BAI65915.1	*Anthriscus sylvestris*
38	BAI65914.1	*Forsythia x intermedia*
39	ACS87992.1	*Citrus sinensis*
40	AAR06917.1	*Stevia rebaudiana*

is considerable information available on the existence and diversity of glycosides, the effect of glycosylation on the activity of the acceptor molecules, and its consequences in relation to cellular homeostasis. In this context, glycosylation is known to provide access to membrane-bound transporters. Glycosides and glucose esters of small molecules, including hormones, secondary metabolites and xenobiotics, have been shown to accumulate in the vacuolar lumen (Bowles et al, 2005). Transporters for some of these compounds have been identified in the vacuolar membrane, and there is evidence to suggest that different mechanisms function for glucosides of endogenous metabolites compared to those of xenobiotics. In contrast to the diversity of sugar donors in plants, the mammalian UGT1 and UGT2 subset invariably use UDP-glucuronic acid. Known acceptors for these glucuronosyl transferases include endogenous substrates such as steroids, bilirubin and bile acids and exogenous xenobiotic substrates such as dietary flavonoids, and drugs such as morphine and naproxen (Radominska-Pandya et al., 1999, King et al., 2000, Tukey and Strassburg, 2000, Miners et al., 2004). To investigate the evolutionary pattern and relationship between the UDP-GT family proteins from distinct organisms of plant kingdom, phylogenetic tree was constructed based on both the whole GT amino acid sequences and also based on amino acid sequences of only PSPG box. The phylogenetic tree obtained is not identical. Some similarity in term of cladding of organism lies in both
the phylogenetic tree, however both are not identical. This specifies that though the PSPG box is an important motif of GTs as concerns the donor specificity, however, the overall sequence similarities/differences could lie at the N terminal region of the sequence and which plays significant role in the cladding of the organisms in the phylogenetic tree and their evolution time in the evolutionary time period. The phylogenetic tree based on the whole GT sequences divides UDP-GT in three groups (A, B, and C), most of the UDP-GTs are coming under group A. The GT of *R. communis* and *V. vinifera* are forming group B, the GT of *P. trichocrapa* and *L. barbarum* are forming group C, and the GT of *F x ananassa* is coming individually from root. All the GTs have conserved PSPG motif near C-terminal. This divergence of group is because of differences in amino acid sequences in other region of GT than PSPG box. The GT of *P. patens*, the smallest GT of 265 amino acid length, comes in group A, showing closeness with *Anthriscus sylvestris* which is of length 485 amino acid. The GT of *P. patens* lacks approximately 230 to 270 amino acid sequences at N-terminal. The GT from *T. aestivum* and *S. sereale* are most conserved and both of them are most distantly related from the GT of *Fragaria x ananassa*. Although the PSPG motif in all the sequences are conserved, but the region other than PSPG box are not conserved, and most of the diversification lies at N-terminal region which is responsible for the sugar acceptor pocket (Offen et al., 2006; Shao et al., 2005). Because the most of the GTs taken have unique sugar donor pocket and the knowledge of sugar acceptor molecule of most of GTs are not available with primary structure, it can be hypothesized that the diversification in the sugar acceptor pocket played crucial role in the cladding of the organisms.

In our present study we have taken GTs from 40 distinct plants of plant kingdom (Table 1). Not only that, 14 distinct GT sequences from *A. thaliana* only (supplementary Table 1), because this is the plant in which the primary structure of most of GTs are available in literature and also 4 distinct GTs (supplementary Table 2) in term of amino acid length, sugar donor specificity and plant origin were taken with objective to investigate the presence of PSPG motif. The PSPG motif is conserved region near C-terminal of all GTs. PSPG motif observed for the set of four distinct sequences was of 39 amino acid, instead of 44 amino acid as already reported in literature (Hughes and Hughes, 1994) and also obtained in the present analysis of set of input sequences. This shortening of the motif length is because of the non-conservative substitution of amino acids at position 40, 41, and 42 among 4 distinct sequences, whereas position 43 showed the conservative substitution and the position 44 is highly conserved. The amino acid residues of position -1 to -11 are strongly conserved among four sequences except position -2 and -3 which are showing semi-conservative substitution and position -4 and -5, showing conservative substitution, becomes the another cause of shortening of PSPG box in this set sequences (Fig 2b). The PSPG box obtained from 14 *A.*

Table 2. Results obtained from MEME tool for Set I, II, and III datasets represent data sets obtained from *A. thaliana* GT sequences, four extremely diverse GT sequences, and all forty diverse GT sequences from 40 different plants respectively.

Parameters	Set I	Set II	Set III
Type of Sequence	Protein	Protein	Protein
Number of sequences	14	4	40
Shortest sequence (amino acid residues)	460	265	265
Longest Sequence (amino acid residues)	496	525	525
Average sequence length (amino acid residues)	487.8	445.5	479.2
Total dataset size (amino acid residues)	7317	1782	19168

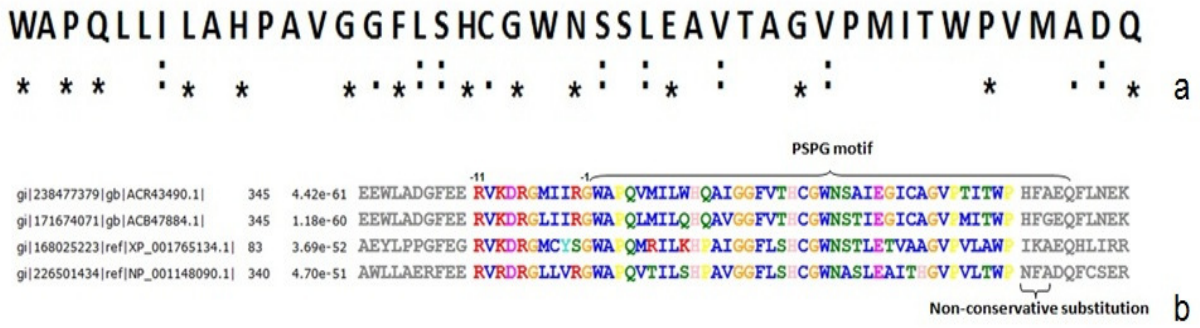

Fig 2. a.Consensus sequence of PSPG box of plant glycosyltransferases obtained through MSA of set III sequences, identical amino acids are indicated by star (*), highly conserved amino acids substitution are indicated by two asterisks (:), and semi-conserved amino acids substitution are indicated by one asterisk below the letters of amino acids; b analysis of PSPG box derived from sequences of set II, as shown in supplementary Table 2.

thaliana GT sequences and 40 diverse secondary metabolic plant GT sequences are similar in term of their length and position specific conservation of amino acids in PSPG motif. We got the three motifs in the GT sequences (Fig 5, 6, 7, supplementary Fig 1, 2a). Motif 1 which was located near C-terminal from approximate position 340th onwards was identified as the PSPG motif based on the amino acid sequence similarity. This shows that the PSPG is an essential part of plant secondary metabolite GTs which is localized near C-terminal of GTs. The MEME tool showed the length of PSPG motif of 50 amino acid but the literature (Hughes and Hughes, 1994; Offen et al., 2006; Shao et al., 2005) support the PSPG motif of length 44 amino acid which is present in the motif 1 showed by the MEME tool which start from position 4 and end at position 47. Exception was existed in the form of *P. patens* GT in which PSPG motif was present in the middle of the GT sequence from 100th amino acid onwards, as the size of GT were 265 amino acids. Another exception was the presence of motif 1 (PSPG motif) twice in GT sequence of *S. cereale* and *T. aestivum*, one at C-terminal and another at N-terminal as per MEME results, but when the sequence was analyzed only one PSPG motif was found near C-terminal from 340th amino acid onwards. Besides two motifs (motif 2, 3) were also found in all GT sequences except *P. patens* GT which lacks motif 3, and motif 2 was present at N-terminal from 10th amino acid residue onwards, the GT of *P. patens* lacks approximately 230 to 270 amino acid sequences at N-terminal. All other GTs has the motif 2 in centre near motif 1 from 260th position onwards. Motif 3 was present near N-terminal from 100th position onwards. In case of *C. arietinum* all motifs are shifted towards N-terminal 20-40 amino acids, this might be because of the shorter length of GT sequences by 40-60 amino acid residues at N terminal. The e value for motif 1, 2 and 3 was 2.7e-1633, 2.6e-1134, and 3.4e-609 respectively. Motif 1 is highly conserved; motif 3 is least conserved and the motif 2 is with intermediate conservation. The acceptor pocket is formed by less conserved regions. The N-terminal domain is less conserved than the C-terminal domain (Offen et al., 2006; Shao et al., 2005). So it can be implied that the motif 3 which is present near N-terminal might have some role in forming acceptor pocket with substrate specificity. How the presence of motif 2 help in functioning of motif 1 and 3 is still matter of speculation. There is no report available so far regarding the presence of motif 2, and 3 in literature. The length of motif 2 and 3, as showed by MEME tool, was 50 amino acids and 34 amino acids respectively, however it requires experimental evidences for functional validation. UDP-sugar is the most commonly used donor for family 1 UGTs, but as for the types of monosaccharides, different UGTs use different monosaccharides. UDP-glucose (UDPGlc) is the most common sugar donor, whilst UDP rhamnose (UDP-Rha), UDP-galactose (UDP-Gal), UDP xylose (UDP-Xyl) and UDP-glucuronic acid (UDP GlcUA) have also been used for some UGTs (Bowles et al., 2005). In case of *W. somnifera*, the GTs utilized only UDP-glucose. UDP galactose could not serve as the sugar donor (Sharma et al., 2007). This specificity was consistent with the recent demonstration that the last amino acid of the PSPG motif in glycosyltransferases controlled relative specificity for UDP glucose or UDP galactose. A glutamine (Q) in glucosyltransferases and histidine (H) in galactosyltr-ansferases is critical to such specificity (Kubo et al., 2004). The presence of glutamine as the last amino acid in UGT prosite motif in SGTL1 corroborates the above functionality.The plant secondary product glycosyltransferase (PSPG) motif is a modification of UGT prosite. The membrane bound plant SGTs differ significantly in the PSPG

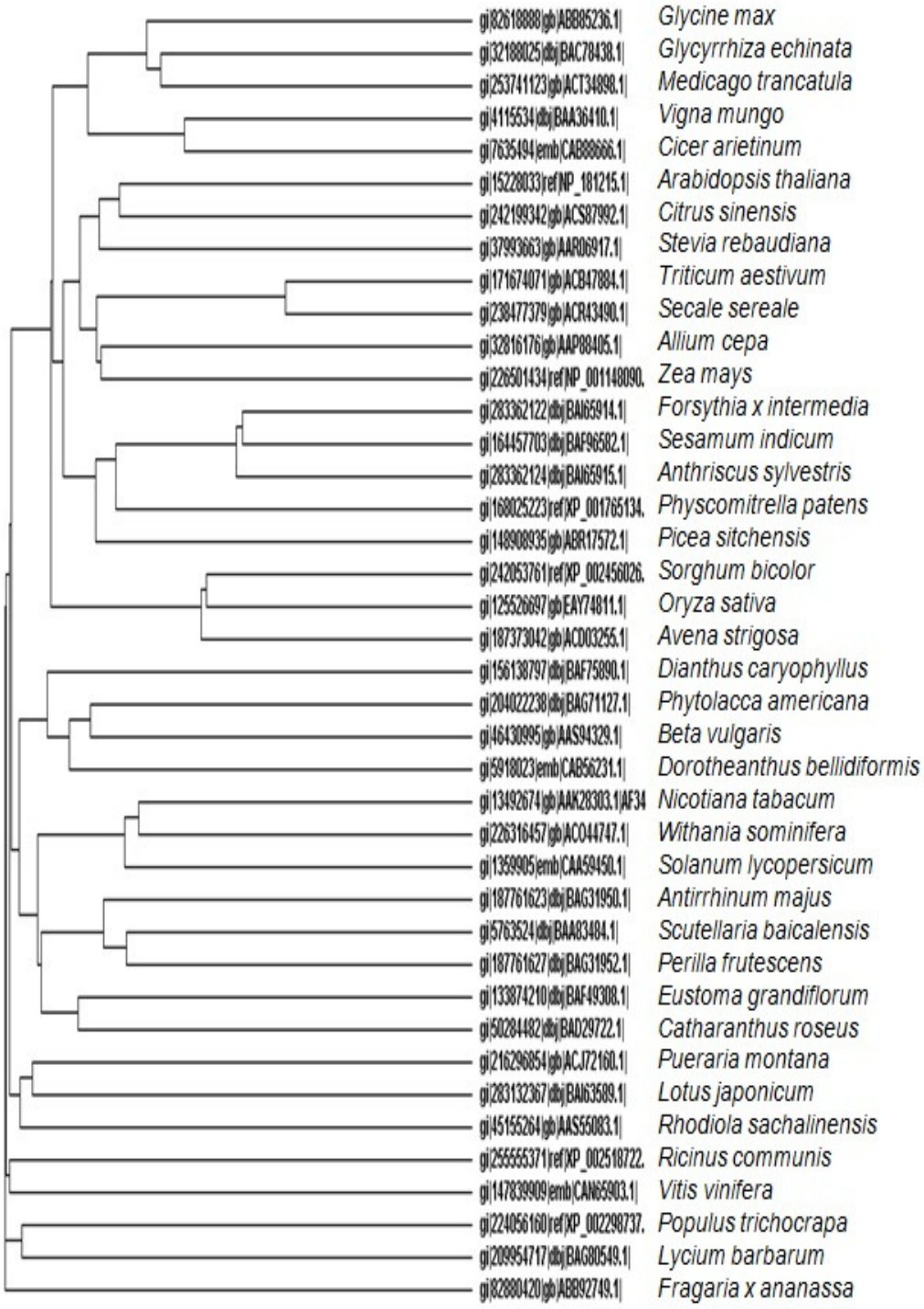

Fig 3. Phylogenetic tree based on the amino acid sequences of complete secondary metabolic GT proteins of 40 diverse plants. The phylogenetic tree was generated from the sequence alignment using the neighbour-joining method.

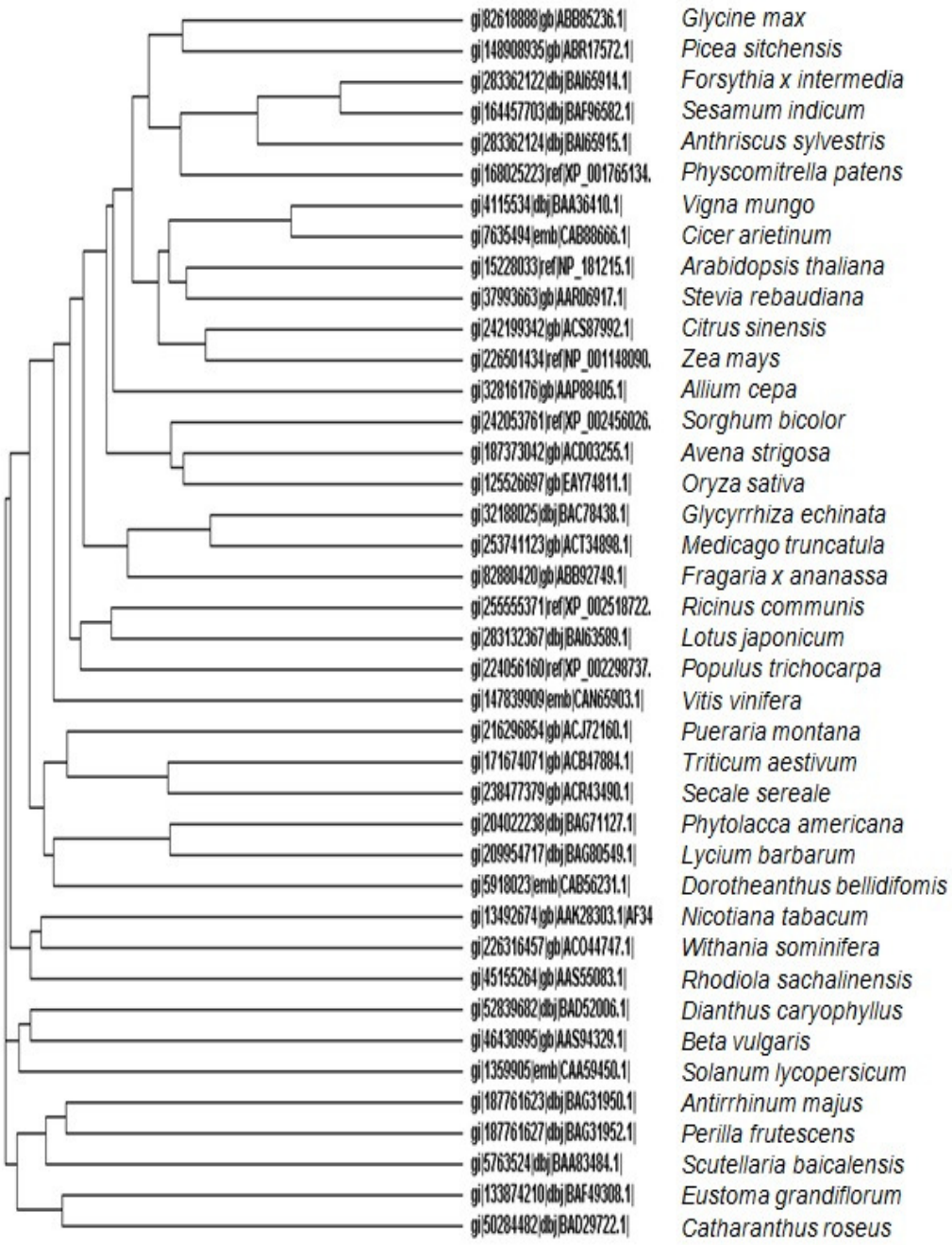

Fig 4. Phylogenetic tree based on the amino acid sequences of PSPG motif of 40 diverse plant secondary metabolic GTs. The phylogenetic tree was generated from sequence alignment using the neighbour-joining method.

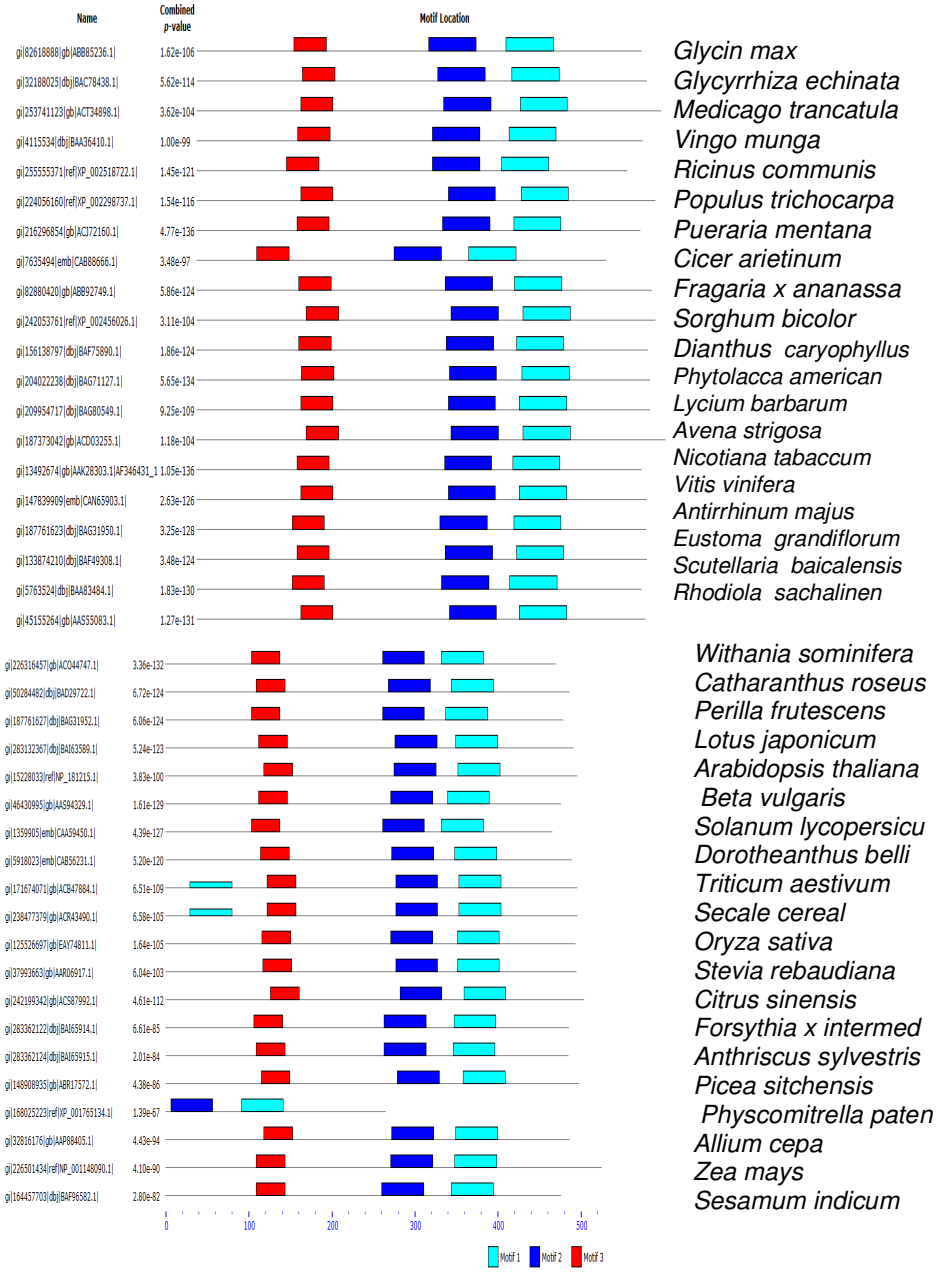

Fig 5. Combined block diagram for all the motifs, constructed using set III of 40 diverse sequences from plants, showing the occurrence and location of the motifs. Putative UDP-glucose of *C. arietinum* (CAB88666.1) has the same three motifs but all motifs are shifted towards the N-terminal side of the sequence. UDP-glucoronosyl/UDP-glucosyl transferase protein of *T. aestivum* (ACB47884.1), and UDP-glucosyl transferase of *S. cereale* (ACR43490.1), both are of same amino acid length, have an extra motif near N-terminal besides the other three motifs. Glycosyl transferase (265aa) of *P. patens subsp. Patens* (XP_001765134.1) did not exhibit box 3, and considerable shift in both the motif towards N-terminal was observed.

motif by incorporation of additional residues within the PSPG motif compared to the cytosolic GTs (Paquette et al., 2003). This additional sequence stretch was found in the SGTL1. The major differences between the SGTL1 and other SGTs are seen in N-terminal part of the protein. Similarity was higher in the middle and C-terminal part. Mutations affected the relative specificities for the sugar donors UDP-galactose and UDP-glucuronic acid, although UDP-glucose was always preferred (He et al., 2006). Curcumin glucosyltransferase (CaUGT2) isolated from cell cultures of *C. roseus* exhibits unique substrate specificity. To identify amino acids involved in substrate recognition and catalytic activity of CaUGT2, a combination of domain swapping and site-directed mutagenesis was carried out.

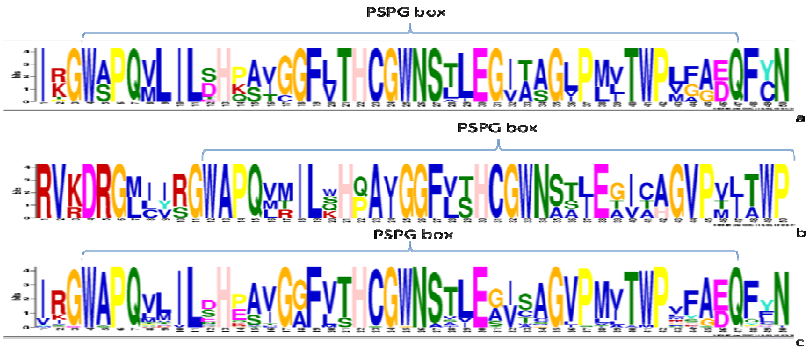

Fig 6. WebLogo of the plant specific PSPG motif constructed from the accessions shown in Table 1. Letter size is proportional to the degree of amino acid conservation. Different colours of letters indicate distinct properties of amino acids such as blue colour is representative of most hydrophobic amino acids (A, C, F, I, L, V, W, and M), green colour is representative of polar, non-charged, non-aliphatic residues (N, Q, S, and T), magenta colour is representative of acidic amino acids (D, E), and red colour is representative of positively charged amino acids (K, R), while the pink, orange, yellow, and turquoise colours represent the residue H, G, P, and Y respectively. a. Motif 1 constructed by using fourteen *A. thaliana* GT sequences as shown in supplementary Table 1; b. Motif 1 constructed by utilizing four most diversified GT sequences as shown in supplementary Table 2; c. Motif 1 constructed by utilizing 40 diversified GT sequences as shown in supplementary Table 3.

Fig 7. WebLogo of the plant GT specific motif constructed from the accessions shown in Table 1. Letter size is proportional to the degree of amino acid conservation. Different colours of letters indicate the distinct properties of amino acids such as blue colour is representative of most hydrophobic amino acids (A, C, F, I, L, V, W, and M), green colour is representative of polar, non-charged, non-aliphatic residues (N, Q, S, and T), magenta colour is representative of acidic amino acids (D,E), and red colour is representative of positively charged amino acids (K, R), while the pink, orange, yellow, and turquoise colour represent the residue H, G, P, and Y respectively. a. Motif 2 of around 50 amino acids found between motif 1 (at C terminal) and motif 3 (at N terminal) located around 240-290 amino acids of the sequence; b- Motif 3 of around 34 amino acids found at N terminal and located around 100-150 amino acids of the sequence.

Exchange of the PSPG-box of CaUGT2 with that of NtGT1b (a phenolic glucosyltransferase from tobacco) led to complete loss of enzyme activity in the resulting recombinant protein. However, replacement of Arg378 of the NtGT1b PSPG-box with cysteine, the corresponding amino acid in CaUGT2, restored the catalytic activity of the chimeric enzyme. Further site-directed mutagenesis revealed that the size of the amino acid side-chain in that particular site is critical to the catalytic activity of CaUGT2 (Masada et al., 2007). Results from the phylogenetic analysis and comparison of substrate recognition patterns among Arabidopsis Family 1 UGTs have shown that there is greater variability within the N-terminal regions of these proteins than in their C-terminal regions, including the PSPG-box, and have indicated that amino acid residues in the N-terminal half of the proteins were responsible for acceptor binding, whereas those in the C-terminal half were involved mainly in interactions with donor substrates. Investigation of the 3D-structures of betanidin 5-O-glucosyltransferase (B5GT) from *D. bellidiformis* and cyanohydrin glucosyltransferase from *Sorghum bicolor* by homology modeling, and of isoflavonoid 3-O-glucosyltransferase from *M. truncatula* and flavonoid 3-O-glucosyltransferase from *V. vinifera* by X-ray crystallography, revealed the role of specific conserved amino acid residues in the PSPG-box that constitute the donor-sugar binding pockets (He et al., 2006). However, the roles of less well conserved amino acids within the motif that may determine the characteristics unique to particular enzymes such as substrate recognition and catalytic potential of the secondary metabolic GTs.

Materials and Methods

NCBI (National Center for Biotechnology Information), websites (http://www.ncbi.nlm.nib.gov/) were used for the retrieval of the amino acid sequences of plants GTs. GTs of the selected 40 different organism (Table 1), based on amino acid sequence diversity, were retrieved, and used for the conserved motif analysis. The MSA of all these diversified 40 GTs were performed by using two MSA software (ClustalW2 and T coffee), and their phylogenetic tree were also constructed by using the same software. Motif discovery tools such as MEME were utilized for computing motif occurrence and analysis.

Conclusion

The GTs, involved in plant secondary metabolism, possess important motif playing key role in enzymatic catalysis. The PSPG box consisting of 44 amino acids possess a unique plant secondary product glycosyltransferase signature. The PSPG motif is an important motif for the catalysis involving the donor substrate and is situated at the C terminal of the gene sequence. The aglycon specificity resides putatively at the N terminal end and contributes towards the functionality for the acceptor molecule(s) as substrates. The phylogenetic tree based on 40 GTs sequences differed significantly, however the phylogenetic relationship based on the PSPG motif could reveal a closer relationship. This specifies that though PSPG box is an important motif of GTs as concerns the donor specificity, however, the overall sequence similarities/differences could lie at the N terminal region of the sequence representing sugar acceptor molecule, which plays significant role in the cladding of the resource plants in the phylogenetic tree and signifies their evolution. Besides the PSPG box at C terminal two more motifs are also present in GT sequences, one at N-terminal which might possess the catalytic potential for various aglycon acceptor pocket available for glycosylation. The presence of motif 2 in between the two acceptor and donor pocket could be required for some regulatory and/or catalytic functions and need further study to establish that.

Acknowledgements

Authors are thankful to Director, CIMAP, for providing facilities and encouragement. Financial grant under NWP-09 Network Program of CSIR, New Delhi is gratefully acknowledged. RK is thankful to University Grants Commission, New Delhi for the award of Junior Research Fellowship.

References

Bowles D, Isayenkova J, Lim E K, Poppenberger B (2005). Glycosyltransferases: managers of small molecoles. Curr Opin Plant Biol, 8: 254-263.

Bowles D, Eng-Kiat L, Brigitte P, Fabian EV (2006) Glycosyltransferases of lipophilic small molecules, Annu Rev Plant Bio, 57:567-97.

Breton C, Snajdrova L, Jeanneau C, Koca J, Imberty A (2006) Structure and mechanisms of glycosyltransferases. Glycobiology 16, 29-37.

Campbell JA, Davies GJ, Bulone V, Henrissat B (1997) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem J 326: 929-993.

Coutinho PM, Deleury E, Davies GJ, Henrissat B (2003) An evolving hierarchical family classification for glycosyltransferases. J Mol Biol 328: 307-317.

Claire MM, Mathilde LM, Patrick S (2005) Plant secondary metabolism glycosyltransferases: the emerging functional analysis. Trends Plant Sci 10:542-9.

Hou B, Lim E K, Higgins G S, Bowels D J (2004). N-glucosylation of cytokinins by glucosyltransferases of *Arabidopsis thaliana*. J Biol Chem 279: 47822-47832.

He X-Z, Wang X, Dixon RA (2006) Mutational analysis of the *Medicago* Glycosyltransferase UGT71G1 reveals residues that control regioselectivity for (iso) flavonoid glycosylation. J Biol Chem 281: 34441-47.

Hughes J, Hughes MA (1994) Multiple secondary plant product UDP-glucose glucosyltransferase genes expressed in cassava (*Manihot esculenta* Crantz) cotyledons. DNA Sequence 5: 41-49.

Jones P, Vogt T (2001). Glycosyltransferases in secondary plants metabolism: tranquilizers and stimulant controllers. Planta 213: 164-174.

King CD, Rios GR, Green MD, Tephly TR (2000) UDP-glucuronosyltransferases. Curr Drug Metab 1, 143–161.

Kramer CM, Prata RTN, Willits MG, De Luca V Steffens JC, Graser G (2003) Cloning and region specificity studies of two flavonoid glucosyltransferases from *Allium cepa*. Phytochemistry 64, 1069-1076.

Kristensen C, Morant M, Olsen CE, Ekstrom CT, Galbraith DW, Moller BL, Bak S (2005) Metabolic engineering of dhurrin in transgenic Arabidopsis plants with marginal inadvertent effects on the metabolome and transcriptome. Proc Natl Acad Sci USA 102: 1779-1784.

Kubo A, Arai Y, Nagashima S, Yoshikawa T (2004) Alteration of sugar donor specificity of plant glycosyltransferase by a single point mutation. Arch Biochem Biophys 429 198–203.

Loutre C, Dixon DP, Brazier M, Slater M, Cole DJ, Edwards R (2003) Isolation of a glucosyltransferase from *Arabidopsis thaliana* active in the metabolism of the persistent pollutant 3,4-dichloroaniline. Plant J 34: 485-493.

Masao, H., Kuroda R., Hideyuki, S, Yoshikawa, T (2000) Cloning and expression of UDP-glucose: flavonoid 7-O-glucosyltransferase from hairy root cultures of *Scutellaria baicalensis*. Planta 210, 1006-1013.

Masad S, Terasaka K, Mizukami H (2007) A single amino acid in the PSPG-box plays an important role in the catalytic function of CaUGT2 (Curcumin glucosyltransferase), a Group D Family 1 glucosyltransferase from *Catharanthus roseus*. FEBS Letters 581: 2605–2610.

Merken HM, Beecher GR (2000) Liquid chromatographic method for the separation and quantification of prominent flavonoid aglycones. J Chromatogr A 897: 177-184.

Miners JO, Smith PA, Sorich MJ, Mckinnon RA, Mackenzie PI (2004) Predicting human drug glucuronidation parameters: application of in vitro and *in silico* modelling approaches. Annu Rev Pharmacol Toxicol 44: 1–25.

Offen, W, Martinez-Fleites, C, Yang, M, Kiat-Lim, E, Davis, BG, Tarling, CA, Ford, CM, Bowles, DJ and Davies, GJ(2006) Structure of a flavonoid glucosyltransferase reveals the basis for plant natural product modification. Embo J 25: 1396–1405.

Paquette S, Moller BL, Bak S (2003) On the origin of family 1 plant glycosyltransferases. Phytochemistry 62 399–413.

Radominska-Pandya A, Czernik PJ, Little JM, Battaglia E, Mackenzie PI (1999) Structural and functional studies of UDPglucuronosyltransferases. Drug Metab Rev 31: 817–899.

Sarah AO, Soren B, Birger LM (2009) Substrate specificity of plant UDP-dependent glycosyltransferases predicted from the crystal structures and homology modeling, Phytochemistry 70: 325-47.

Sharma LK, Madina BR, Chaturvedi P, Sangwan R S, Tuli R (2007) Molecular cloning and characterization of one member of 3b-hydroxy sterol glucosyltransferase gene family in *Withania somnifera.* Arch Biochem Biophys 460: 48–55.

Shao H, He X, Achnine L, Blount JW, Dixon RA, Wang X (2005) Crystal structures of a multifunctional triterpene/flavonoid glycosyltransferase from *Medicago truncatula*. Plant Cell. 17: 3141–3154.

Tukey RH, Strassburg CP (2000) Human UDP-glucuronosyltransferases: metabolism, expression and disease. Annu Rev Pharmacol Toxicol 40: 581–616.

Vogt T (2002) Substrate specificity and sequence analysis define a polyphyletic origin of betanidin 5- and 6-O-glucosyltransferase from *Dorotheanthus bellidiformis*. Planta 214: 492-495.