PERGAMON

Phytochemistry 62 (2003) 399–413

PHYTOCHEMISTRY

www.elsevier.com/locate/phytochem

On the origin of family 1 plant glycosyltransferases

Suzanne Paquette[a,b], Birger Lindberg Møller[b], Søren Bak[b,*]

[a]Department of Biological Structure, HSBG-514, Box 357420, University of Washington Medical School, Seattle, WA 98145-9420, USA

[b]Plant Biochemistry Laboratory, Department of Plant Biology, Royal Veterinary and Agricultural University, Thorvaldsensvej 40, and Center of Molecular Plant Physiology (PlaCe), DK-1871 Frederiksberg C, Copenhagen, Denmark

Received 29 August 2002; received in revised form 8 October 2002

Dedicated to Meinhart H. Zenk on the occasion of his 70th birthday

Abstract

The phylogeny of highly divergent multigene families is often difficult to validate but can be substantiated by inclusion of data outside of the phylogeny, such as signature motifs, intron splice site conservation, unique substitutions of conserved residues, similar gene functions, and out groups. The Family 1 Glycosyltransferases (UGTs) comprises such a highly divergent, polyphyletic multigene family. Phylogenetic comparisons of UGTs from plants, animals, fungi, bacteria, and viruses reveal that plant UGTs represent three distinct clades. The majority of the plant sequences appears to be monophyletic and have diverged after the bifurcation of the animal/fungi/plant kingdoms. The two minor clades contain the sterol and lipid glycosyltransferases and each show more homology to non-plant sequences. The lipid glycosyltransferase clade is homologous to bacterial lipid glycosyltransferases and reflects the bacterial origin of chloroplasts. The fully sequenced *Arabidopsis thaliana* genome contains 120 UGTs including 8 apparent pseudogenes. The phylogeny of plant glycosyltransferases is substantiated with complete phylogenetic analysis of the *A. thaliana* UGT multigene family, including intron-exon organization and chromosomal localization.

Keywords: Phylogenetic analysis; Gene organization analysis; UDPG-glycosyltransferases; UGT; Molecular evolution; *Arabidopsis thaliana*; Brassicaceae; Thale cress

1. Introduction

As whole genome sequences of relevant species become available, genetic analysis is entering the post-genome era. Data mining within whole genome sequences is greatly facilitated by phylogenetic and bioinformatic analyses that define and characterize entire multigene families, and profits from the availability of the sequences of all genes putatively involved. Multigene families can be researched within specific organisms and across phyla in complete groups to reduce the amount of bias caused by the availability of a limited number of gene representatives that may represent specific subgroups, such as highly transcribed genes.

Glycosyltransferases are a highly divergent, polyphyletic, multigene family (Mackenzie et al., 1997). They are responsible for glycosylation reactions, i.e. the conjugation of a glycose residue from an activated sugar donor to a receptor molecule. Glycosylation can result in the formation of poly-glycosides, di-saccharides, and various mono-glycosides of non-carbohydrate moieties such as proteins, lipids, steroids, and other small molecules. The glycosyltransferase multigene family is categorized into 54 numbered families according to sequence similarity, signature motifs, stereochemistry of the glucoside linkage formed, and known target specificity (Campbell et al., 1997; http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html). Of these 54 families, the family 1 contains the UDPG-glycosyltransferases (UGTs) (Mackenzie et al., 1997; Li et al., 2001; Ross et al., 2001). UGTs have been identified in plants, animals, fungi, and bacteria, and also viruses (Campell et al., 1997, http://www.unisa.edu.au/pharm_medsci/Gluc_trans/table21.htm). These glycosyltransferases are characterized by utilization of UDP-activated sugar moieties as the donor molecule, and contain a conserved UGT-defining sequence motif near the C-terminus (Mackenzie et al., 1997). This UGT-defining motif is often the

* Corresponding author. Tel.: +45-3528-3346; fax: +45-3528-3333.

E-mail address: bak@kvl.dk (S. Bak).

0031-9422/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
PII: S0031-9422(02)00558-7

only significant region of similarity in sequence alignments within and across phyla. Glycosylation reactions serve to convert reactive and toxic aglycones into more stable and non-reactive storage forms. In addition, attachment of the hydrophilic glucose moiety to hydrophobic aglycones dictates increased water solubility. While UDP-glucose and UDP-glucuronic acid are considered the most typical donor molecules for the family 1 glycosyltransferases, examples of UDP-rhamnose, UDP-xylose, and UDP-galactose also exist.

Glycosylation by a UGT constitutes a prominent modification process and is often the last step in the biosynthesis of natural products in plants (Jones and Vogt, 2001). The glycosylation reaction is not restricted to endogenous substrates; it is also a key step in general detoxification mechanisms for exogenous substrates (xenobiotics) in higher plants (Sandermann, 1992), thereby allowing plants to cope with environmental challenges. The biological function of the glycosylation step in plants is therefore to facilitate storage, and intra- and intercellular transport. Glycosylation also serves as a regulatory step in homeostasis of plant growth regulators, as seen for auxins, gibberellins and brassinolides.

In this paper we show that the plant UGTs cover three distinct clades when compared to non-plant UGTs in bootstrapped Neighbor-Join trees. One of the clades is vastly expanded and specific to plants, and is monophyletic. The two minor clades representing sterol and lipid UGTs are more related to non-plant clades than to other plant-specific clades. The plant lipid glucosyltransferases form a distinct clade with the bacterial lipid glucosyltransferases and reflect the endosymbiotic origin of the chloroplast. The phylogeny of the plant specific UGT clade has previously been examined in *Arabidopsis* (Ross et al., 2001; Li et al., 2001). However, with the completion of the *A. thaliana* genome (TAGI, 2000), it is now possible to examine the *Arabidopsis* UGT gene family as a whole including the sterol and lipid UGT clades. *Arabidopsis* has a large UDP-glycosyltransferase family, containing 112 full-length genes and 8 apparent pseudogenes. Our analysis is supplemented by an extended analysis of intron splice site position and phase as it relates to the phylogeny of a bootstrap Neighbor-Join tree and the chromosomal location of all the *Arabidopsis* UGT genes.

2. Results

2.1. Optimizing multiple alignment of divergent UGT sequences

The phylogeny of large multigene families can be difficult to authenticate, because many of these families contain divergent members that complicate the validation of multiple sequence alignments and phylogenetic trees. Often, extensive sequence diversity makes multiple alignments and phylogenetic trees appear less accurate and ambiguous, due to a large numbers of gaps, long branches, and low bootstrap or parsimony values (Brocchieri, 2001). However, these apparent inaccuracies and weaknesses do not necessarily mean that the given method or program provides data of poor significance. It is possible to assess and underscore the accuracies and inaccuracies of a given phylogenetic tree by examining data outside of the phylogeny itself, such as intron splice site conservation, unique substitutions of conserved residues, and gene function similarities to validate the results obtained.

The availability of complete or nearly complete genomes for an increasing number of species renders it possible to carry out thorough and detailed large-scale analysis of multigene families, and to predict phylogenetic origins. Likewise, benefits and drawbacks of the different analytical procedures used to carry out sequence alignments and to predict phylogenetic relationships can now be assessed. Large data sets are important to overcome the limitations of programs used to form multiple alignments. With respect to the *Arabidopsis* P450s, a complete data set resulted in an improved overall sequence alignment of the family, and enabled the construction of more accurate phylogenetic trees (Paquette et al., 2000 vs. Werck-Reichhart et al., 2002). In order to achieve optimal alignments of the very divergent data sets we have empirically exhausted the use of different available protein weight matrix series, namely PAM, BLOSUM, and GONNET, as well as altered the default values of the ClustalW and ClustalX software (Thompson et al., 1994, 1997) (data not shown). Accordingly, to align the more inhomogeneous UGTs for the multi-organism tree (Fig. 1), we have used the GONNET matrix series instead of the BLOSUM matrix series that was used with the relatively more homogenous *Arabidopsis* sequences (Fig. 4). The GONNET matrices are derived using a similar procedure as the Dayhoff PAM matrices, but are based on a far larger data set and generally considered to be more sensitive than the PAM matrices (Gonnet et al., 1992; Benner et al., 1994; Bork and Gibson, 1996). Generally speaking, the GONNET series are thought to perform better than other matrices for aligning large data sets of highly variant sequences that in addition have undergone changes at different evolutionary rates (Benner et al., 1994; Bork and Gibson, 1996). Empirically, we found the GONNET matrices performed better on the data set of Fig. 1 than the BLOSUM series (data not shown). In addition, we have used an increased delay divergent score of 38% for the multi-organism UGT alignment as opposed to a score of 35% for the *Arabidopsis* UGT multiple alignment; an increase in the delay divergent score assures that the relatively more divergent subclades have a higher chance of aligning, and produces better alignments of the gaps and weakly conserved positions. We have lowered the Gap Extension Penalty (GEP) in the multiple-alignment setting for the multi-organism tree to 0.15 to facilitate introduction of longer gaps, rather than increasing the number of gaps. For the *Arabidopsis* UGT tree, a lower GEP value was used in the pairwise parameters, but a GEP of 0.20 was used in the multiple-alignment settings. For the tree construction we have chosen the Neighbor Joining method as opposed to e.g. parsimony method as the Neighbor Joining method is generally better at resolving long branches due to the "long branches attracts" bias

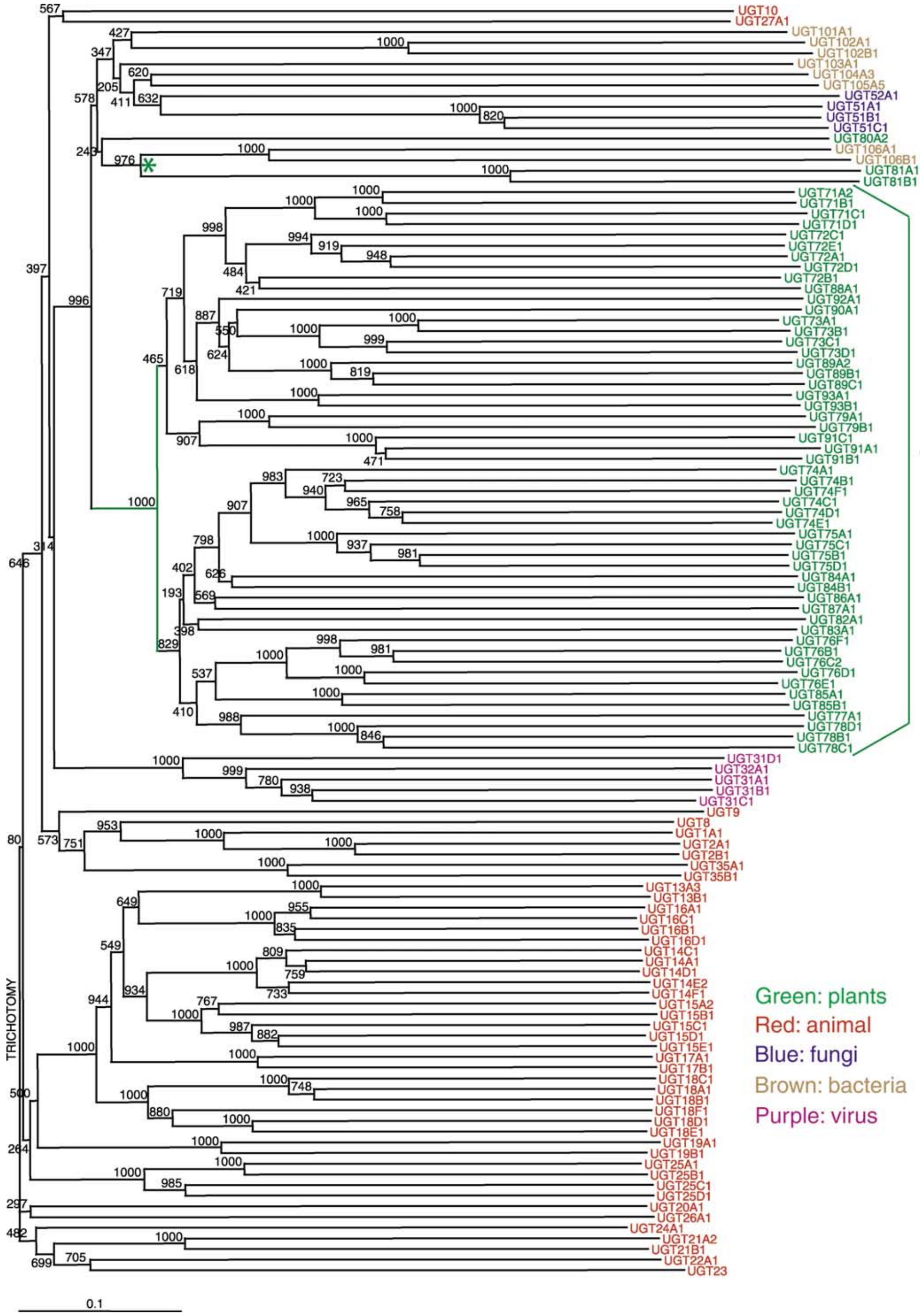


Fig. 1. Multi-organism Neighbor-Join tree containing 119 family 1 glycosyltransferases, including the UGT81 family. A single representative for each subfamily was chosen. *Arabidopsis* sequences took preference over other available plant UGT sequences. The different phyla are color-coded: plant:green; animal:red; fungi:blue; bacteria:brown; and virus:purple. The lipid glycosyltransferase clade is marked with a green asterisk (*). A GONNET protein weight matrix series was used to generate the multiple alignment and the tree was analyzed with 1000 bootstrap trials.

observed when analyzing a divergent data set with sequences under different evolutionary rates (Bruno et al., 2000).

2.2. Family phylogeny

A multi-organism glycosyltransferase family 1 Neighbor-Join tree has been constructed containing UGT sequences representing bacteria, fungi, animals, plants, and viruses (Fig. 1). A single member of each UGT subfamily has been included in the tree. A complete listing of the species chosen to represent the individual subfamilies is given in Table 1 together with GenBank accession numbers. This highly divergent set of sequences represents membrane anchored as well as soluble proteins. Nevertheless, all of these sequences contain the UGT signature motif as outlined by the UGT Nomenclature Committee (Mackenzie et al., 1997) (Fig. 2). The fungal UGT families 51 and 52 differ from all other families by having a unique ~750 amino acid extension at the N terminus. Despite this vast diversity in sequence length and composition, it was possible to align the sequences using ClustalW. To study the effect of the inclusion of these extended UGT51 and UGT52 sequences in the alignment and the tree construction, additional analyses were performed in which these extended N-termini were deleted. This showed that, given the alignment parameters selected for our analysis (see Experimentals), the alignment remained essentially unchanged whether full-length or truncated sequences were used. A criterion for the optimal alignment of the UGT sequences is a proper alignment of the PROSITE UGT consensus sequence. Fig. 2 exemplifies the capability of ClustalW to align these divergent sequences using the alignment parameter chosen (see Materials and methods); for simplicity only the PROSITE consensus is shown representing a single member of each family. The full alignment behind the multi-organism tree (Fig. 1) presenting a single member of each subfamily member is available at our website at: http://www.biobase.dk/P450/Figure1_alignment.pdf.

The plant UGTs represent three distinct clades: a large plant-specific clade and two minor clades represented by UGT80A2 and UGT81A1/81B1. Sequences within the two minor clades show a higher similarity to non-plant UGT sequences than to any sequence of the other plant UGT families (Fig. 1). Despite the low levels of amino acid sequence similarity observed between the three clades, structural studies indicate a common evolutionary origin. Accordingly, the three clades are discussed as belonging to a single supergene family in this paper. The plant UGT81 family represented by UGT81A1 and UGT81B1 from *Arabidopsis* forms a strong (976/1000) clade with UGT106A1 from *Bacillus subtilis* and UGT106B1 from *Staphylococcus aureus*. The node of this clade (Fig. 1, green asterisk) represents

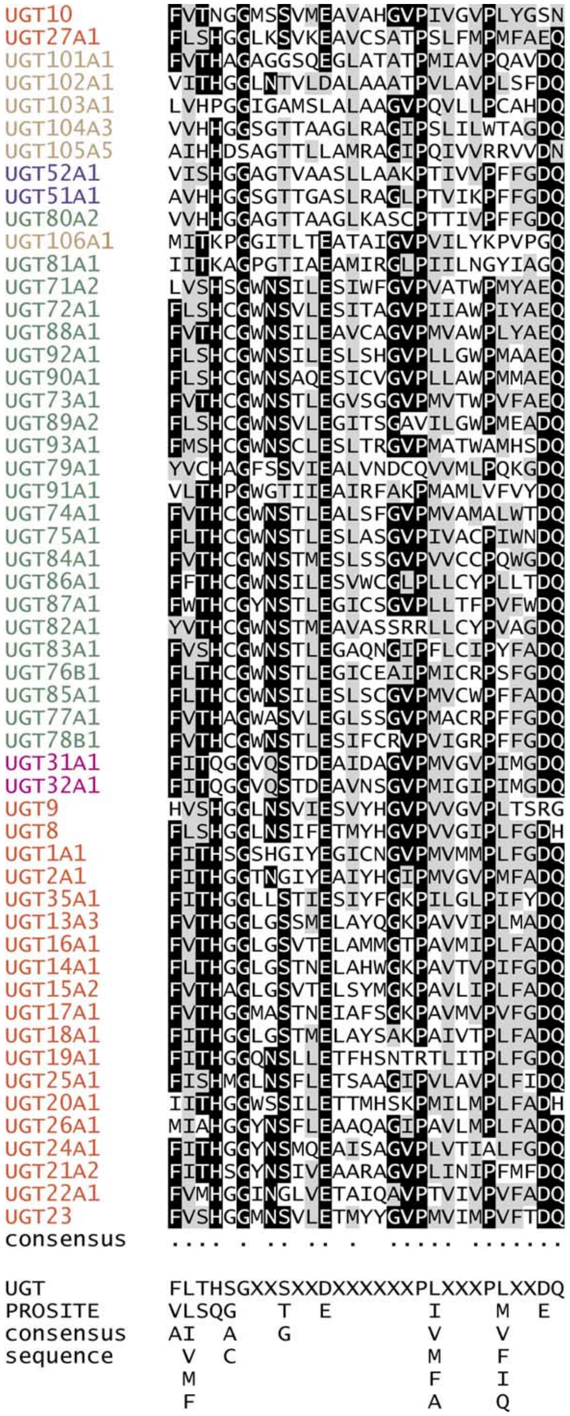

UGT10 FVTNGGMSSVMEAVAHGVPIVGVPLYGSN
UGT27A1 FLSHGGLKSVKEAVCSATPSLFMPMFAEQ
UGT101A1 FVTHAGAGGSQEGLATATPMIAVPQAVDQ
UGT102A1 VITHGGLNTVLDALAAATPVLAVPLSFDQ
UGT103A1 LVHPGGIGAMSLALAAGVPQVLLPCAHDQ
UGT104A3 VVHHGGSGTTAAGLRAGIPSLILWTAGDQ
UGT105A5 AIHHDSAGTTLLAMRAGIPQIVVRRVVDN
UGT52A1 VISHGGAGTVAASLLAAKPTIVVPFFGDQ
UGT51A1 AVHHGGSGTTGASLRAGLPTVIKPFFGDQ
UGT80A2 VVHHGGAGTTAAGLKASCPTTIVPFFGDQ
UGT106A1 MITKPGGITLTEATAIGVPVILYKPVPGQ
UGT81A1 IITKAGPGTIAEAMIRGLPIILNGYIAGQ
UGT71A2 LVSHSGWNSILESIWFGVPVATWPMYAEQ
UGT72A1 FLSHCGWNSVLESITAGVPIIAWPIYAEQ
UGT88A1 FVTHCGWNSILEAVCAGVPMVAWPLYAEQ
UGT92A1 FLSHCGWNSILESLSHGVPLLGWPMAAEQ
UGT90A1 FLSHCGWNSAQESICVGVPLLAWPMMAEQ
UGT73A1 FVTHCGWNSTLEGVSGGVPMVTWPVFAEQ
UGT89A2 FLSHCGWNSVLEGITSGAVILGWPMEADQ
UGT93A1 FMSHCGWNSCLESLTRGVPMATWAMHSDQ
UGT79A1 YVCHAGFSSVIEALVNDCQVVMLPQKGDQ
UGT91A1 VLTHPGWGTIIEAIRFAKPMAMLVFVYDQ
UGT74A1 FVTHCGWNSTLEALSFGVPMVAMALWTDQ
UGT75A1 FLTHCGWNSTLESLASGVPIVACPIWNDQ
UGT84A1 FVTHCGWNSTMESLSSGVPVVCCPQWGDQ
UGT86A1 FFTHCGWNSILESVWCGLPLLCYPLLTDQ
UGT87A1 FWTHCGYNSTLEGICSGVPLLTFPVFWDQ
UGT82A1 YVTHCGWNSTMEAVASSRRLLCYPVAGDQ
UGT83A1 FVSHCGWNSTLEGAQNGIPFLCIPYFADQ
UGT76B1 FLTHCGWNSTLEGICEAIPMICRPSFGDQ
UGT85A1 FLTHCGWNSILESLSCGVPMVCWPFFADQ
UGT77A1 FVTHAGWASVLEGLSSGVPMACRPFFGDQ
UGT78B1 FVTHCGWNSTLESIFCRVPVIGRPFFGDQ
UGT31A1 FITQGGVQSTDEAIDAGVPMVGVPIMGDQ
UGT32A1 FITQGGVQSTDEAVNSGVPMIGIPIMGDQ
UGT9 HVSHGGLNSVIESVYHGVPVVGVPLTSRG
UGT8 FLSHGGLNSIFETMYHGVPVVGIPLFGDH
UGT1A1 FITHSGSHGIYEGICNGVPMVMMPLFGDQ
UGT2A1 FITHGGTNGIYEAIYHGIPMVGVPMFADQ
UGT35A1 FITHGGLLSTIESIYFGKPILGLPIFYDQ
UGT13A3 FVTHGGLGSSMELAYQGKPAVVIPLMADQ
UGT16A1 FVTHGGLGSVTELAMMGTPAVMIPLFADQ
UGT14A1 FLTHGGLGSTNELAHWGKPAVTVPIFGDQ
UGT15A2 FVTHAGLGSVTELSYMGKPAVLIPLFADQ
UGT17A1 FVTHGGMASTNEIAFSGKPAVMVPVFGDQ
UGT18A1 FITHGGLGSTMELAYSAKPAIVTPLFADQ
UGT19A1 FITHGGQNSLLETFHSNTRTLITPLFGDQ
UGT25A1 FISHMGLNSFLETSAAGIPVLAVPLFIDQ
UGT20A1 IITHGGWSSILETTMHSKPMILMPLFADH
UGT26A1 MIAHGGYNSFLEAAQAGIPAVLMPLFADQ
UGT24A1 FITHGGYNSMQEAISAGVPLVTIALFGDQ
UGT21A2 FITHSGYNSIVEAARAGVPLINIPFMFDQ
UGT22A1 FVMHGGINGLVETAIQAVPTVIVPVFADQ
UGT23 FVSHGGMNSVLETMYYGVPMVIMPVFTDQ
consensus ....
UGT FLTHSGXXSXXDXXXXXXPLXXXPLXXDQ
PROSITE VLSQG T E I M E
consensus AI A G V V
sequence V C M F
M F I
F A Q

Fig. 2. Multi-organism alignment of UGT C-terminal PROSITE consensus sequence from 55 family 1 glycosyltransferases. For simplicity, a single sequence of each UGT family was chosen and only the conserved UGT PROSITE consensus sequence is displayed. The PROSITE UGT consensus sequence is shown and the different phyla are colored as in Fig. 1. Residues similar in all sequences are marked with a (.) beneath the alignment. Residues that are identical in at least half of the sequences are printed with a black background, residues that are similar in at least half of the sequences are printed with a gray background, and residues with less than 50% similarity are printed with a white background.

Table 1
Non-*Arabidopsis* UGTs

UGT	Organism	GenBank Accession No.
UGT1A2	*Rattus norvegicus*	M34007
UGT2A1	*R. norvegicus*	X57565
UGT2B1	*R. norvegicus*	M13506
UGT8	*R. norvegicus*	L21698
UGT9	*Caenhorabiditis elegans*	CAA85328
UGT10	*C. elegans*	AAA83572
UGT13A3	*C. elegans*	AAC48056
UGT13B1	*C. elegans*	CAA99950
UGT14A1	*C. elegans*	AAF99874
UGT14C1	*C. elegans*	AAB42343
UGT14D1	*C. elegans*	AAB42345
UGT14E2	*C. elegans*	AAB42344
UGT14F1	*C. elegans*	AAB42350
UGT15A2	*C. elegans*	CAA94904
UGT15B1	*C. elegans*	AAC48241
UGT15C1	*C. elegans*	AAC48235
UGT15D1	*C. elegans*	AAC48238
UGT15E1	*C. elegans*	AAB71324
UGT16A1	*C. elegans*	CAA92791
UGT16B1	*C. elegans*	CAB01674
UGT16C1	*C. elegans*	CAB62783
UGT16D1	*C. elegans*	AAK29790
UGT17A1	*C. elegans*	CAB02883
UGT17B1	*C. elegans*	CAB02884
UGT18A1	*C. elegans*	CAA99957
UGT18B1	*C. elegans*	CAA99955
UGT18C1	*C. elegans*	CAA99959
UGT18D1	*C. elegans*	CAA94871
UGT18E1	*C. elegans*	CAA99954
UGT18F1	*C. elegans*	AAD14733
UGT19A1	*C. elegans*	CAB01584
UGT19B1	*C. elegans*	CAB01585
UGT20A1	*C. elegans*	CAA94365
UGT21A2	*C. elegans*	AAA81769
UGT21B1	*C. elegans*	AAK52183
UGT22A1	*C. elegans*	CAA94866
UGT23	*C. elegans*	CAA90296
UGT24A1	*C. elegans*	CAA84336
UGT25A1	*C. elegans*	CAA94378
UGT25B1	*C. elegans*	AAG24186
UGT25C1	*C. elegans*	AAB66005
UGT25D1	*C. elegans*	AAC02618
UGT26A1	*C. elegans*	CAB54183
UGT27A1	*C. elegans*	AAB65381
UGT31A1	*Spodoptera littoralis nucleopolyhedrovirus*	X84701
UGT31B1	*Mamestra brassicae nucleopolyhedrovirus*	U41999
UGT31C1	*Lymantria dispar nucleopolyhedrovirus*	U04321
UGT31D1	*Choristoneura fumiferana nucleopolyhedrovirus*	U10441
UGT32A1	*Choristoneura fumiferana DEF nucleopolyhedrovirus*	U10476
UGT35A1	*Drosophila melanogaster*	AF116555
UGT35B1	*D. melanogaster*	AF116554
UGT51A1	*Saccharomyces cervisiae*	AAB67475
UGT51B1	*Pichia pastoris*	AF091397
UGT51C1	*Candida albicans*	AF091398
UGT52A1	*Dictyostelium disciodeum*	AF098916
UGT71A2	*Manihot esculentum*	S41950
UGT72A1	*M. esculentum*	S41951
UGT73A1	*Nicotiana tabacum*	U32644
UGT74A1	*Zea mays*	L34847
UGT75A1	*N. tabacum*	AB000623
UGT77A1	*Z. mays*	X13500

(continued on next page)

Table 1 (*continued*)

UGT	Organism	GenBank Accession No.
UGT78B1	*Gentiana triflora*	D85186
UGT78C1	*Solanum melongena*	X77369
UGT79A1	*Petunia hybrida*	Q43716
UGT85B1	*Sorghum bicolor*	AAF17077
UGT93A1	*Phaseolus vulgaris*	AF116858
UGT93B1	*Z. mays*	AF318075
UGT101A1	*Streptomyces antibioticus*	CAA80301
UGT102A1	*Pantoea agglomerans*	AAA64979
UGT102B1	*Pantoea ananatis*	BAA14125
UGT103A1	*Pseudomonas aeruginosa*	AAA62129
UGT104A3	*Mycobacterium avium*	AAC71702
UGT105A5	*Amycolatopsis orientalis*	AAB49292
UGT106A1	*Bacillus subtillus*	P54166
UGT106B1	*Staphylococcus aureus*	CAA74741

lipid glycosyltransferases. UGT81A1 is the monogalactosyldiacylglycerol (MGDG) synthase and catalyzes the transfer of galactose from UDP-galactose to 1,2-diacylglycerol. MGDG is the major glycolipid constituent of the chloroplast (Shimojima et al., 1997). Bacterial UGT106A1 and 106B1 also glycosylate 1,2-diacylglycerol, but use UDP-glucose instead of UDP-galactose as glycosyl. In addition to 1,2-diacylglycerol, UGT81A1 as well as UGT106B1 utilize other lipids as sugar acceptors (Jorasch et al., 1998, 2000).

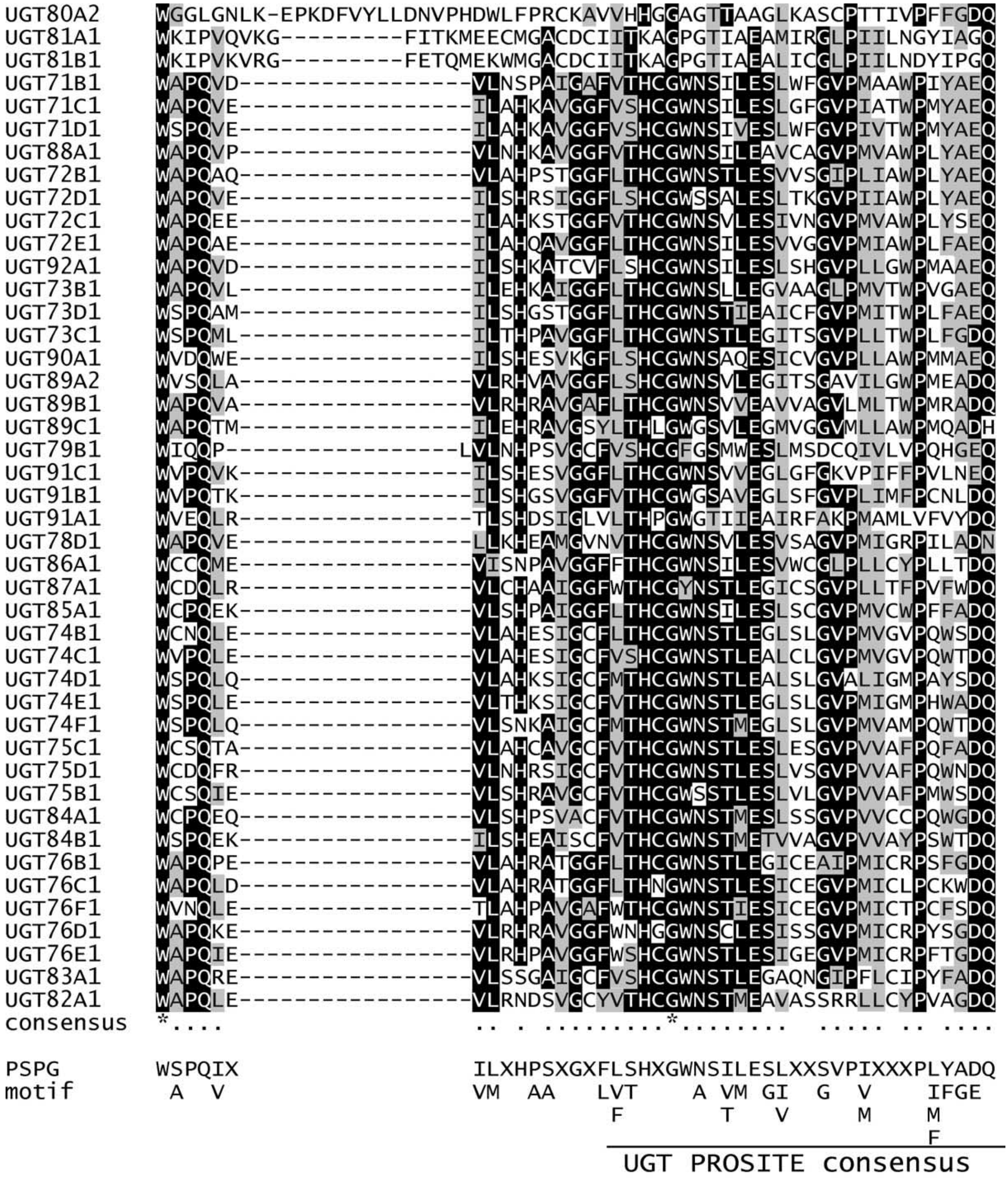
In the multi-organism tree presented in Fig. 1, UGT80A2 does not group with a significant bootstrap value to any other sequence group. UGT80A2 is a sterol-glucosyltransferase involved in the synthesis of sterolglucosides, characteristic lipids of the eukaryotic membrane (Warnecke et al., 1999). The fungal UGT51 and UGT52 families are also sterol-glucosyltransferases (Warnecke et al., 1999) and in the multiorganism tree they form a very strong cluster of 1000/1000 iterations. Members of these three UGT families catalyze the very same reaction, but the plant homolog does not appear to have a close phylogenetic relation to the fungal sterol glucosyltransferases, although some sequence similarities are apparent. UGT80A2 does not contain the large N-terminal extension found in the fungal sequences, although it is ~160 amino acids longer than most other plant UGTs. Compared to fungal UGT51A1, UGT80A1 is 557 amino acids shorter. The 722 N-terminal amino acids of UGT51A1 are not required for catalytic activity in vitro (Warnecke et al., 1999) and no specific biological functions has yet been assigned to these extended N-terminus of the sterol UGTs.

2.3. The plant specific UGT clade

The major part of the plant UGTs form a single strong clade that is specific to plants (Fig. 1). Of the 21 known UGT families in this clade, only the UGT77 and UGT93 families are not represented in the *Arabidopsis* genome. The members of this plant specific clade are characterized by a highly conserved consensus sequence denoted the PSPG (**P**utative **S**econdary **P**lant **G**lycosyltransferase) motif. The PSPG consensus was originally defined by Hughes and Hughes (1994) as a signature motif for a plant UGT involved in glycosylation of secondary metabolites. The PSPG motif is a modification of the PROSITE glycosyltransferase signature sequence shown in Fig. 2 and is extended by 15 amino acids N-terminally to the PROSITE consensus (Fig. 3). The sequence composition of the UGT80 and 81 families differs significantly from the PSPG motif by incorporation of additional residues within the PSPG motif (Fig. 3).

To further validate the phylogeny of the plant UGTs, we compared a bootstrap tree based on amino acid sequences of the entire set of *Arabidopsis* UGTs with the intron-exon organization of the genes (Fig. 4). Of the 120 UGTs in the *Arabidopsis* genome, 112 are full-length. The remaining eight are either truncated, or contain in frame stop codons or frame shifts, and are considered putative pseudogenes. The multiple alignment used for intron mapping has a consensus length of 709 amino acids, and only UGTs that could be assembled into full-length sequences were included in the alignment and tree. Introns were considered conserved if the position was within 1 codon of the majority of that intron's recorded positions and the phase was identical. Introns positioned between two codons are phase 0, introns positioned after the first base in the codon are phase 1, and introns positioned after the second base in the codon are phase 2.

Generally, the UGTs containing the PSPG motif are characterized by having a single or very few introns, if any, as opposed to the UGT80 and UGT81 gene families, which have 13 and 7 or 5 introns, respectively (Fig. 4). The lack of two introns in the *UGT81B2* gene most likely represents a case of intron loss as these two introns are found in UGT81B1 and UGT81A1. Over half of the *Arabidopsis* UGTs lack introns (58/112). Of



```
UGT80A2   WGGLGNLK-EPKDFVYLLDNVPHDWLFPRCKAVVHHGGAGTTAAGLKASCPTTIVPFFGDQ
UGT81A1   WKIPVQVKG---------FITKMEECMGACDCIITKAGPGTIAEAMIRGLPIILNGYIAGQ
UGT81B1   WKIPVKVRG---------FETQMEKWMGACDCIITKAGPGTIAEALICGLPIILNDYIPGQ
UGT71B1   WAPQVD-----------------VLNSPAIGAFVTHCGWNSILESLWFGVPMAAWPIYAEQ
UGT71C1   WAPQVE-----------------ILAHKAVGGFVSHCGWNSILESLGFGVPIATWPMYAEQ
UGT71D1   WSPQVE-----------------ILAHKAVGGFVSHCGWNSIVESLWFGVPIVTWPMYAEQ
UGT88A1   WAPQVP-----------------VLNHKAVGGFVTHCGWNSILEAVCAGVPMVAWPLYAEQ
UGT72B1   WAPQAQ-----------------VLAHPSTGGFLTHCGWNSTLESVVSGIPLIAWPLYAEQ
UGT72D1   WAPQVE-----------------ILSHRSIGGFLSHCGWSSALESLTKGVPIIAWPLYAEQ
UGT72C1   WAPQEE-----------------ILAHKSTGGFVTHCGWNSVLESIVNGVPMVAWPLYSEQ
UGT72E1   WAPQAE-----------------ILAHQAVGGFLTHCGWNSILESVVGGVPMIAWPLFAEQ
UGT92A1   WAPQVD-----------------ILSHKATCVFLSHCGWNSILESLSHGVPLLGWPMAAEQ
UGT73B1   WAPQVL-----------------ILEHKAIGGFLTHCGWNSLLEGVAAGLPMVTWPVGAEQ
UGT73D1   WSPQAM-----------------ILSHGSTGGFLTHCGWNSTIEAICFGVPMITWPLFAEQ
UGT73C1   WSPQML-----------------ILTHPAVGGFLTHCGWNSTLEGITSGVPLLTWPLFGDQ
UGT90A1   WVDQWE-----------------ILSHESVKGFLSHCGWNSAQESICVGVPLLAWPMMAEQ
UGT89A2   WVSQLA-----------------VLRHVAVGGFLSHCGWNSVLEGITSGAVILGWPMEADQ
UGT89B1   WAPQVA-----------------VLRHRAVGAFLTHCGWNSVVEAVVAGVLMLTWPMRADQ
UGT89C1   WAPQTM-----------------ILEHRAVGSYLTHLGWGSVLEGMVGGVMLLAWPMQADH
UGT79B1   WIQQP-----------------LVLNHPSVGCFVSHCGFGSMWESLMSDCQIVLVPQHGEQ
UGT91C1   WVPQVK-----------------ILSHESVGGFLTHCGWNSVVEGLGFGKVPIFFPVLNEQ
UGT91B1   WVPQTK-----------------ILSHGSVGGFVTHCGWGSAVEGLSFGVPLIMFPCNLDQ
UGT91A1   WVEQLR-----------------TLSHDSIGLVLTHPGWGTIIEAIRFAKPMAMLVFVYDQ
UGT78D1   WAPQVE-----------------LLKHEAMGVNVTHCGWNSVLESVSAGVPMIGRPILADN
UGT86A1   WCCQME-----------------VISNPAVGGFFTHCGWNSILESVWCGLPLLCYPLLTDQ
UGT87A1   WCDQLR-----------------VLCHAAIGGFWTHCGYNSTLEGICSGVPLLTFPVFWDQ
UGT85A1   WCPQEK-----------------VLSHPAIGGFLTHCGWNSILESLSCGVPMVCWPFFADQ
UGT74B1   WCNQLE-----------------VLAHESIGCFLTHCGWNSTLEGLSLGVPMVGVPQWSDQ
UGT74C1   WVPQLE-----------------VLAHESIGCFVSHCGWNSTLEALCLGVPMVGVPQWTDQ
UGT74D1   WSPQLQ-----------------VLAHKSIGCFMTHCGWNSTLEALSLGVALIGMPAYSDQ
UGT74E1   WSPQLE-----------------VLTHKSIGCFVTHCGWNSTLEGLSLGVPMIGMPHWADQ
UGT74F1   WSPQLQ-----------------VLSNKAIGCFMTHCGWNSTMEGLSLGVPMVAMPQWTDQ
UGT75C1   WCSQTA-----------------VLAHCAVGCFVTHCGWNSTLESLESGVPVVAFPQFADQ
UGT75D1   WCDQFR-----------------VLNHRSIGCFVTHCGWNSTLESLVSGVPVVAFPQWNDQ
UGT75B1   WCSQIE-----------------VLSHRAVGCFVTHCGWSSTLESLVLGVPVVAFPMWSDQ
UGT84A1   WCPQEQ-----------------VLSHPSVACFVTHCGWNSTMESLSSGVPVVCCPQWGDQ
UGT84B1   WSPQEK-----------------ILSHEAISCFVTHCGWNSTMETVVAGVPVVAYPSWTDQ
UGT76B1   WAPQPE-----------------VLAHRATGGFLTHCGWNSTLEGICEAIPMICRPSFGDQ
UGT76C1   WAPQLD-----------------VLAHRATGGFLTHNGWNSTLESICEGVPMICLPCKWDQ
UGT76F1   WVNQLE-----------------TLAHPAVGAFWTHCGWNSTIESICEGVPMICTPCFSDQ
UGT76D1   WAPQKE-----------------VLRHRAVGGFWNHGGWNSCLESISSGVPMICRPYSGDQ
UGT76E1   WAPQIE-----------------VLRHPAVGGFWSHCGWNSTLESIGEGVPMICRPFTGDQ
UGT83A1   WAPQRE-----------------VLSSGAIGCFVSHCGWNSTLEGAQNGIPFLCIPYFADQ
UGT82A1   WAPQLE-----------------VLRNDSVGCYVTHCGWNSTMEAVASSRRLLCYPVAGDQ
consensus *....                  .. . .........*........  ..... .. ....

PSPG      WSPQIX                 ILXHPSXGXFLSHXGWNSILESLXXSVPIXXXPLYADQ
motif      A  V                  VM  AA  LVT   A VM GI  G  V    IFGE
                                          F      T  V     M    M
                                                               F
                                          ____________________________
                                           UGT PROSITE consensus
```

Fig. 3. PSPG sequence of 44 *Arabidopsis* UGTs. For simplicity, only a single sequence from each subfamily is shown and only the PSPG consensus is displayed. The PSPG consensus sequence taken from Hughes and Hughes (1994) as well as the PROSITE UGT consensus are displayed. Residues conserved in all sequences are marked with a (*), and similar residues in all sequences are marked with a (.) beneath the alignment. Residues that are identical in at least half of the sequences are printed with a black background, residues that are similar in at least half of the sequences are printed with a gray background, and residues with less than 50% similarity are printed with a white background.

those that contain introns, the vast majority (44/54) have only one. We have identified 31 unique intron positions in the 112 UGTs analyzed, and 21 of these are derived from the UGT80A and the UGT81 family. Sixteen are phase 0, nine are phase 1, and six are phase 2. Of these 31 intron positions only one, denoted A, is conserved across families (Fig. 4).

The UGT76 family and the UGT78, 82, 83, 85, 86 and 87 families all share a phase 1 intron at approximately position 279 on the consensus sequence (Fig. 4, marked A; http://www.biobase.dk/P450/Figure4_alignment.pdf). Due to sequence dissimilarity, the position of the phase 1 intron of the UGT78 family is difficult to unequivocally confirm. However, based on realignments of the

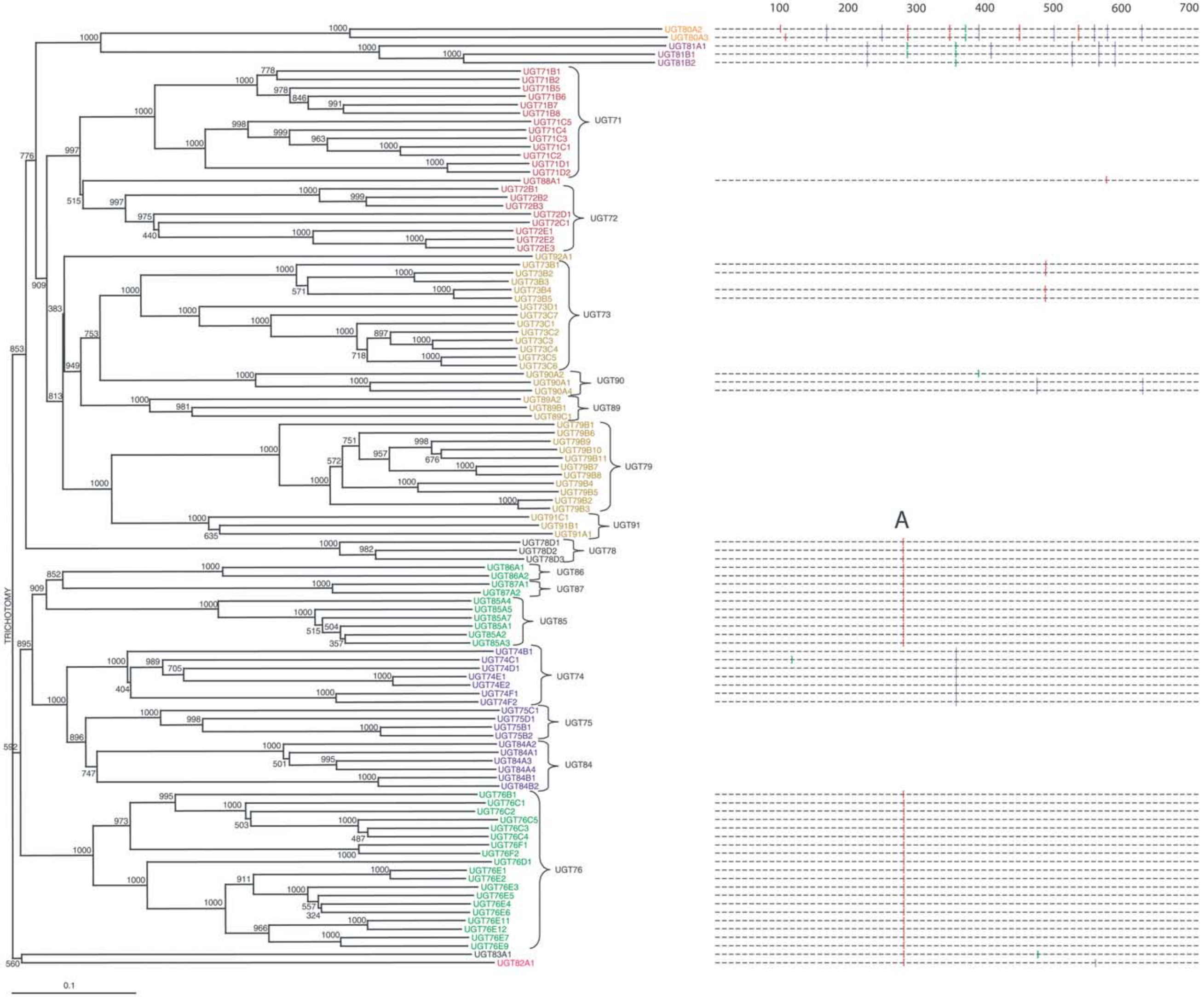


Fig. 4. Neighbor-Join bootstrap tree and corresponding intron map of the 112 *Arabidopsis* UGTs. Subclades supported by strong bootstrap values are assigned a particular color. Phase 0 (|), phase 1 ([ ), and phase 2 ( ]) introns are indicated in their consensus position on the multiple alignment used to form the phylogenetic tree. Phase 0 introns are positioned between two codons, phase 1 introns are positioned after the first base of the codon, and phase 2 introns are positioned after the second base of the codon. A BLOSUM protein weight matrix series was used to align the sequences and the tree was analyzed with 1000 bootstrap trials.

region in question using ClustalX it appears most probable that the intron in the UGT78 family is the conserved A intron. In addition to the conserved A intron, UGT82A1 and UGT83A1 contain a phase 0 and a phase 2 intron, respectively.

UGT89C1 appears to have a large unique gap of 39 amino acids in the sequence when compared to the two other UGT89s, UGT89A2 and UGT89B1. This gap does not appear to be the result of poor annotation i.e. lack of proper identification of intron exon junctions, as the UGT89 gene family contains no introns. The gap in UGT89C1 is confirmed by the EST AV551176, which spans this region. The UGT89 family contains one putative pseudogene UGT89A1P that appears to have a truncated C-terminus that lacks 30 amino acids. Neither of the full-length genes contains this gap and no EST corresponds to UGT89A1P. Such anomalous gaps are difficult to explain and may be the result of improper sequencing of the genome or that these genes have significant different sequences in these regions. However, it is more likely that they are pseudogenes. The presence of an EST for UGT89C1 would then argue that in this case, it is an expressed pseudogene.

The UGT71, 72, and 88 families form a strong (997/1000) clade (Fig. 4), and all but the single-family member UGT88A1 lack introns. UGT88A1 contains a single unique phase 1 splice site in the C-terminus, and might represent a case of intron gain in the UGT88 family or intron loss in the families UGT71 and UGT72 prior to duplication events.

Another strong clade (1000/1000) consists of the UGT74, 75, and 84 families. All the members of the UGT74 family share a phase 0 intron between consensus positions 353 and 354 (Fig. 4). UGT74C1 has an additional phase 2 splice site further towards the N-terminus, and it is the only 2-intron member of this clade. The UGT75 and 84 families both lack introns. This could indicate the conserved intron in the UGT74 family has been lost prior to proliferation of the UGT75 and UGT84 families. Alternatively, the intron in the UGT74 family has been gained prior to proliferation of the UGT74 family.

The UGT90A subfamily contains four members, three of which are shown in the Neighbor-Join tree. The fourth, UGT90A3P, is a pseudogene containing a truncated N-terminus and numerous in-frame stop codons. UGT90A1 and 90A4 both share two phase 0 introns, but UGT90A2 lacks these introns and instead has a phase 2 intron at an earlier consensus position.

The UGT73B and UGT73C subfamilies occupy a well-defined clade (1000/1000). The UGT73C subfamily lacks introns, and all of the UGT73Bs except UGT 73B3 have a single phase 0 splice site. UGT73B3 has no introns.

The alignment used to produce this intronmap, complete with intron positions marked, is available at http://www.biobase.dk/P450/Figure4_alignment.pdf.

2.4. Chromosomal localization

The UGT genes appear to be evenly dispersed throughout the genome and are generally organized in clusters with the reading frames facing the same direction (Fig. 5). A few groups have UGTs with reading frames that oppose one another, e.g. the UGT71C cluster on chromosome 1. Singular UGTs account for 29 of the 120 genes in the genome. It is notable, that most of the putative pseudogenes are located next to a full-length gene of the same subfamily. This tendency of subfamilies with multiple members and putative pseudogenes to be closely clustered on the chromosomes is indicative of recent gene duplication events. The *Arabidopsis* Genome Initiative has found that the overall ratio of genes to pseudogenes is 1/0.03 (TAGI, 2000). In the UGT gene family, the ratio is closer to 1/0.067, approximately twice as high as for the whole organism. No pseudogenes of the UGT80 or UGT81 family were found in the genome, which may indicate that these gene families are more stable than the UGTs with the PSPG motif. UGT80 and UGT81 code for enzymes that catalyze conserved household keeping reactions in contrast to the UGTs with the PSPG motif, which are putatively involved in secondary metabolism and thus subjected to recruitment for novel functions.

3. Discussion

3.1. Plant family type 1 glycosyltransferases

Based on comparison with UGTs from other phyla, we have identified three distinct clades that contain plant UGTs (Fig. 1). The UGT80 and UGT81 families show more sequence homology to non-plant UGT families than to other plant sequences, arguing that they evolved before the radiation of plants from the other phyla. This is in accordance with sterols and lipids being biological molecules that have evolved before the radiation of the plant/animal/fungi kingdoms. The homology between the plant UGT81s and the prokaryotic UGT106 lipid glycosyltransferases reflects the endosymbiotic origin of the chloroplast (Douglas, 1998). The presence of UGT81 homologs in other plant species combined with the absence of a homolog in the eukaryotic world outside the plant kingdom supports this hypothesis. The present form of UGT81 is nuclear encoded, and in agreement with a prokaryotic origin and function in the chloroplast, it contains a chloroplast targeting pre-sequence (data not shown). Accordingly, the UGT81 family represents a case of gene-transfer from an organelle genome to the nucleus.

The apparent lack of a homolog between the plant sterol glycosyltransferases of the UGT80 family and the fungal functional homologs UGT51 and UGT52,

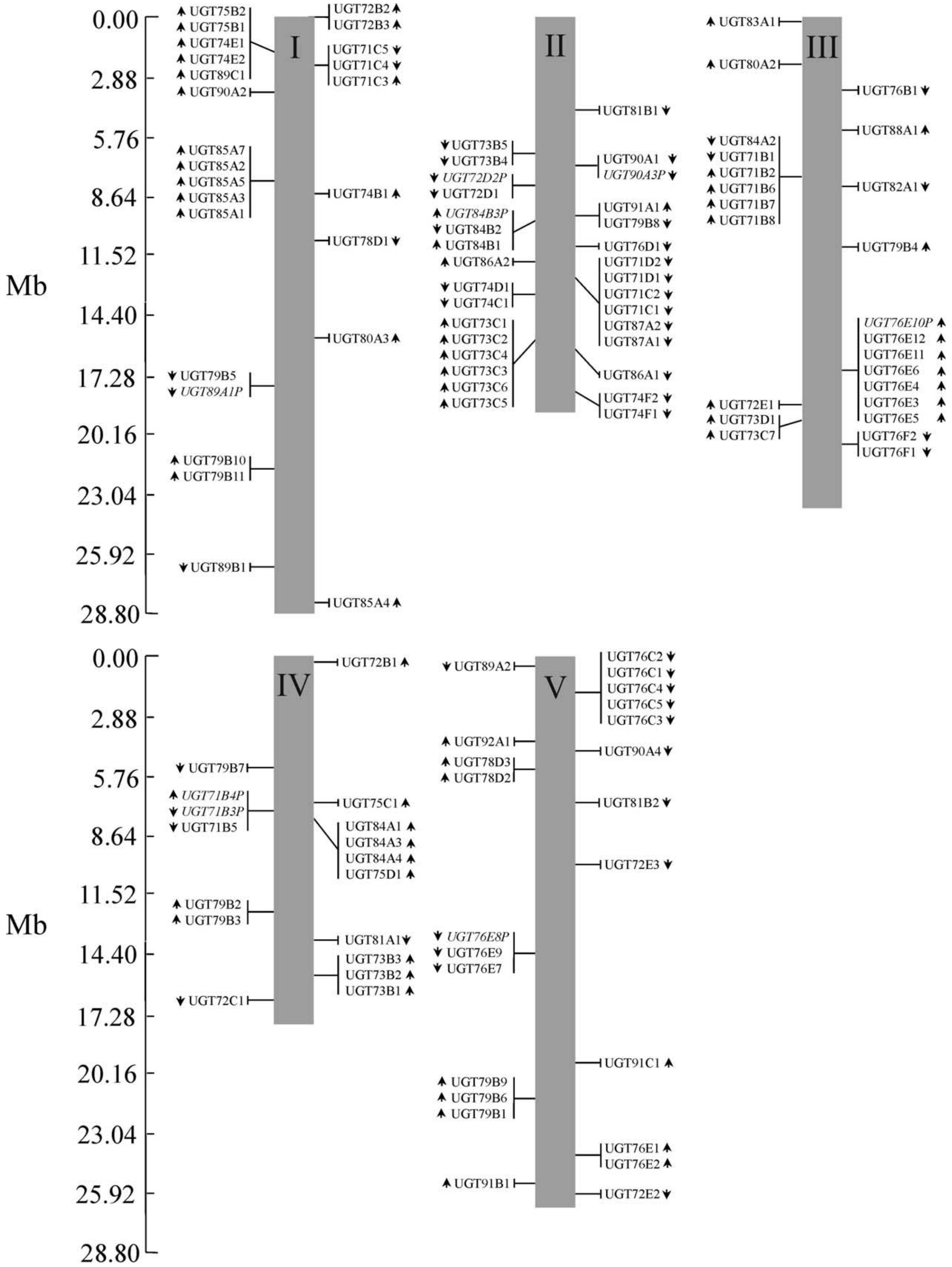


Fig. 5. Chromosome Map of 120 *Arabidopsis* UGTs. Arrows indicate the direction of the coding sequence (up, antisense; down, sense). Gene position and orientation were determined using a BLAST alignment against the complete chromosome sequence available from NCBI. Pseudogene names end with a P and are italicized.

argues that sterol glycosylation might be a case of convergence in these phyla, despite the fact that generally sterol biosynthesis is evolutionary related in eukaryotes. In favor of this argument is the apparent lack of sterol glycosyltransferase from the other phyla that group with the UGT51, UGT52, and UGT80 families.

Plants have a large group of UGTs that contain the PSPG motif (Figs. 1 and 3). These UGTs are thought to be involved in glycosylation of secondary plant metabolites or natural products (Vogt and Jones, 2000; Jones and Vogt, 2001). Sequence alignments, phylogenetic analysis (Fig. 1) and the similar simple intron-exon organization (Fig. 4) suggest that the UGTs with the PSPG motif are monophyletic, i.e. the result of a single ancestral UGT in the plant lineage whose descendants have become numerous. The presence of multiple pseudogenes and the organization of subfamily clusters in the chromosomes (Fig. 5) in this clade is characteristic of recent duplication events and indicates that these genes are in a process of being recruited for new functions (Walsh, 1987, 1995; Lynch and Conery, 2000). The classification of these UGTs as being involved in secondary metabolism should be expanded to include other plant specific pathways, given the recent discovery of UGTs in this clade that are involved in the glycosylation of primary plant metabolites. For example, UGT75B1 has been proposed to be a subunit of the callose synthase complex by facilitating transfer of UDP-glucose from sucrose synthase to callose synthase (Hong et al., 2001). Recombinant UGT84s and UGT72s have in vitro been identified to glucosylate hydroxycinnamates and thus putatively be involved in sinapate and lignin biosynthesis (Milkowski et al., 2000; Lim et al., 2001) Likewise recombinant UGT84B1 catalyzes glycosylation of the plant growth regulator indole-3-acetic acid in vitro (Jackson et al., 2001).

Unlike UGT80A2 and UGT81A1 that are membrane-anchored proteins, the PSPG motif containing UGTs are generally regarded as soluble enzymes. However, upon cellular fractionation they may partly copurify with the microsomal fractions because they are components of multi-enzyme complexes associated with membrane anchored enzymes (e.g. Møller and Conn, 1980; Burbulis and Winkel-Shirley, 1999; Hong et al., 2001; Tattersall et al., 2001). The association of UGTs with specific multi-enzyme complexes is expected to be an important factor in the formation of metabolons that facilitate channeling of reactive or toxic intermediates in biosynthetic pathways.

3.2. Intron-exon organization and phylogeny of the plant specific UGTs

Introns may be considered as evolutionary fossils in a gene family, with intron position, phase, loss, and gain serving as diagnostic tools to validate phylogenies (Long et al., 1995; Stoltzfus et al., 1997). We have previously used intron-exon organization to validate the phylogeny of the cytochrome P450 genes of *Arabidopsis thaliana*, which constitute a large multigene family of 273 members (Werck-Reichhart et al., 2002; Paquette et al., 2000). This method can be applied to other gene families whose sequence diversity makes them challenging to work with, such as the UGT multigene family.

The plant specific clade has previously been analyzed for intron position and gain and loss, which were shown to correlate well with the predicted gene phylogeny (Li et al., 2001; Ross et al., 2001), but at the time the entire genome sequence of *Arabidopsis* was not available. As has been previously shown (Li et al., 2001), the intron splice site positions of the *Arabidopsis* UGTs follow the phylogeny predicted by trees constructed from primary amino acid sequence data. In *Arabidopsis* an average of 79% of the nuclear encoded genes contain introns, and the average exon size is ~250 bp (TAGI, 2000). Because an average UGT is approximately 500 amino acids, the expected average number of introns would be five. It is evident from Fig. 4 that the number of introns found in the UGT multigene family is much below average, with the UGT80 and UGT81 families being the only exceptions. This may be interpreted to indicate a massive loss of introns in the PSPG containing UGTs. Alternatively, the plant UGTs were always exceptionally low in introns, but the UGT80 and UGT81 families subsequently gained a large number whereas the PSPG motif containing ones never did. The UGT subfamilies in *Arabidopsis* that do contain introns share them strongly. Only a few exceptions to this statement are found in the entire data set (Fig. 4) as exemplified by UGT73B3, UGT74B1, and UGT90A2

In the analysis provided by Ross et al. (2001), 14 groups within the PSPG containing UGTs were suggested that have evolved from an equivalent number of distinct ancestral UGT genes. Our analysis concurs with these 14 groups with two exceptions. Given the conservation of the A intron (Fig. 4) and the high bootstrap value (909/1000) (Fig. 4) we argue that the families UGT78, UGT86, UGT87, UGT85 occupy a single group and not four independent groups. Likewise families UGT73, UGT90, and UGT89 are grouped with a bootstrap value of 949/1000 (Fig. 4) and are characterized by not having the conserved A intron but are either contain no introns or contain scattered introns. In the phylogeny presented by Ross et al. (2001) these three families were listed as independent groups. In fact, our data suggest that the PSPG containing UGTs consist of two major groups; those that group with the conserved intron A, and those that group with the ones that primarily contain no introns. The latter group is particularly supported with a bootstrap value of 909/

1000 (Fig. 4), suggesting a monophyletic origin of the families UGT71, 88, 72, 92, 73, 90, 89, 79, and 91. The A-intron containing UGTs are less confirmed by the bootstrap analysis (Fig. 4). However, taking the more divergent data set used in Fig. 1 for the multi-organism tree, this group consisting of UGT families 74, 75, 84, 86, 87, 82, 83, 76, 85, 77, and 78, is supported with a bootstrap value of 829/1000. The overall similarity of the amino acid sequences and the similar intron exon organization of the genes, suggest that the PSPG motif containing UGT's are monophyletic and that early in the development of the clade either the A intron has been lost or gained prior to the massive proliferation of the clade. The inability to resolve the deep branches and the corresponding weak bootstrap values observed when taking only *Arabidopsis* sequences into account, reflects the general problems observed with resolving long branches (Philippe and Laurent, 1998; Bruno et al., 2000), and has in this study been overcome by the inclusion of out groups that stabilize the otherwise very long branches observed in the smaller and less divergent *Arabidopsis* data set (Fig. 4, Ross et al., 2001; Li et al., 2001).

3.3. UGT function and phylogenetic grouping

The functions of the *Arabidopsis* UGTs are largely unknown. The use of microarrays and large-scale heterologous expression, in combination with phylogenetic analysis and the availability of knock-out mutants, constitute important tools for substrate identification and determination of biological functions. Expression of UGTs has already provided candidates for UGTs involved in glycosylation of a variety of substrates such as IAA, salicylic acid, sinapate and 4-hydrozybenzoates, thiohydroxymates, cyanohydrins, betanin, and flavanols (Lim et al., 2001, 2002; Milkowski et al., 2000, 2001; Jackson et al., 2001; Jones et al., 1999; Vogt, 2002; Vogt and Jones, 2000). Based on the data currently available it appears that in vitro these enzymes exhibit broader substrate specificity than in planta (e.g. Jones et al., 1999). This renders pinpointing of the true and key in vivo substrate ambiguous when solely based on in vitro experimentation. The work of the York group has provided evidence for the conservation of substrate specificity within subclades (Lim et al., 2001, 2002; Jackson et al., 2001), whereas the groups in Halle and Copenhagen argue for the importance of regiospecificity as a determinant of whether a compound may be glycosylated or not by a specific UGT (Vogt and Jones, 2000; Vogt, 2002). A clear case of regiospecificity and convergent evolution is found in the two UGTs from *Dorotheanthus bellidiformis* catalyzing either the 5- or 6- hydroxylation of betanin. These two UGTs are only 19% identical and belong to two phylogenetically distinct UGT families, UGT71 (6-GT) and UGT73 (5-GT), yet they have the same substrate (Vogt, 2002).

3.4. Co-evolution of UGTs and cytochromes P450

Plants synthesize a vast array of more than 200,000 natural products to accommodate biotic and abiotic stresses and to secure communication with other plants. The last step in the biosynthesis of many natural products is often a glycosylation reaction catalyzed by a UGT. Glycosylation facilitates stability, storage, and also intra- and intercellular transport. Typically, a hydroxylation reaction precedes the glycosylation reaction to provide a hydroxyl group amenable for glycosylation. The hydroxylation reactions are often catalyzed by cytochromes P450, another large multigene family consisting of 246 full-length genes in *Arabidopsis* (Werck-Reichhart et al., 2002; http://www.biobase.dk/P450/p450.shtml). As seen for the plant UGTs, plant cytochromes P450 contain a large group of plant-specific genes involved in the biosynthesis of plant natural products, commonly referred to as the A-type P450s (Durst and Nelson, 1995; Werck-Reichhart et al., 2002). In *Arabidopsis,* the full-length A-type cytochromes P450 comprises of 153 members or 62% of the total number of cytochromes P450 present. They are thought to be of a monophyletic origin based on intron position and phase conservation (Paquette et al., 2000; Werck-Reichhart et al., 2002). The explosion of the P450 family is thought to have occurred via gene duplication and conversion starting 430 million years ago when plants began to colonize land (Kahn and Durst, 2001). Similarly the plant-specific UGT population is large and appears to have exploded due to recent duplication events (Fig. 5). Another common feature between the A-type cytochrome P450′s and the PSPG motif containing UGTs is the high number of putative pseudogenes. This possibly reflects that these two multigene families are subject to rather frequent changes to secure recruitments for novel functions. Accordingly, it is tempting to suggest that the PSPG motif containing UGTs have co-evolved with cytochromes P450s to secure stabilization of toxic aglycones generated by the cytochromes P450s. A complete listing of putative Arabidopsis cytochrome P450 and UGT pseudogenes and can be found at our website: *The Arabidopsis P450, cytochrome b5, P450 reductase, and Glycosyltransferase Family 1 Site at PlaCe* (htpp://www.biobase.dk/P450). It is intriguing that many of the natural products are detoxified in herbivores by cytochrome P450-catalyzed hydroxylation reactions, followed by UGT-catalyzed glucoronylation to facilitate secretion. The evolution and explosion in the genome of the plant specific A-type cytochromes P450 and the plant specific UGTs with the PSPG motif reflect the necessity of plants to produce new defense compounds to be competitive in the constant chemical warfare between plants and herbivores as well as to cope with other abiotic and biotic stresses.

4. Experimentals

4.1. Obtaining sequences

All of the UGT sequences used in this analysis are available through GenBank as complete cDNA sequence, complete protein sequence, annotated BAC sequence, or complete chromosomal sequence. A complete list of all non-*Arabidopsis* sequences and their GenBank accession numbers is presented in Table 1. The *Arabidopsis* UGT sequences are publicly available at *The Arabidopsis P450, cytochrome b5, P450 reductase, and Glycosyltransferase Family 1 Site at PlaCe* (http://www.biobase.dk/P450/UGT.shtml). Putative *Arabidopsis* UGTs were identified by BLAST (Altschul et al., 1990) comparison against the TAIR *Arabidopsis* and NCBI databases using the following UGT signature motif as outlined by the UGT Nomenclature Committee (Mackenzie et al., 1997):

[F/V/A]-[L/I/V/M/F]-[T/S]-[H/Q]-[S/G/A/C]-GXX-[S/T/G]-XX-[D/E]-XXXXXXP-[L/I/V/M/F/A]-XX-P-[L/M/V/F/I/Q]-XX-[D/E]-Q

Amino acids enclosed in brackets ([]) and separated by slashes are variations at a single position. *Arabidopsis* sequences with this motif, as well as sequences having sequence identity greater than 50% to known UGTs from other organisms, were annotated using a combination of PUBLISH from the GCG Wisconsin Package version 10.2 and the NetGene2 intron prediction server (Hebsgaard et al., 1996; http://www.cbs.dtu.dk/services/NetGene2/). Conservation of sequences within a family was given precedence over predicted intron splice sites. The resulting amino acid sequences were aligned with known members of the gene family and checked for any variances. GenBank BAC annotations and published cDNA sequences were taken into account when available.

The UGT nomenclature used in this paper follows that previously defined by the UGT Nomenclature Committee (Mackenzie et al., 1997) and used by Ross et al. (2001) and Li et al. (2001). Accordingly, sequences more than 40% identical on the amino acid level are placed in the same family and sequences more than 60% identical in the same subfamily.

4.2. Alignments and phylogenetic trees

Multiple sequence alignments and Neighbor-Join bootstrap trees were initially constructed using ClustalW for GCG 10.2 (Thompson et al., 1994) and refined using ClustalX version 1.81 to realign selected regions within the multiple alignment (Thompson et al. 1997), and visualized using Treeview 1.6.6 (Page, 1996). ClustalW settings for the multi-organism UGT alignment underlying Fig. 1 were as follows: For the initial pairwise alignment used to generate the guided tree file, a Gap Opening Penalty (GOP) and a Gap Extension Penalty (GEP) of 10.00 and 0.10 were selected, and the GONNET protein weight matrix series were used. To generate the subsequent multiple alignment, the GOP was kept at 10.00 while the GEP was increased to 0.15 and the delay divergent sequence score was set to 38% and the GONNET protein weight matrix series were used. For the *Arabidopsis* UGT alignment, the following pairwise parameters were used: GOP 10.00, GEP 0.08 and the BLOSUM protein weight matrixes series. For the subsequent multiple alignment, a GOP of 10.00, GEP of 0.20, a delay divergent sequences of 35%, and the BLOSUM protein weight matrix series were used. To facilitate a better alignment of the multi-organism UGTs, the extensive N-terminals of the UGT51 and UGT52 family members were deleted. The alignments used to construct the trees presented in this paper are available at http://www.biobase.dk/P450/Figure1_alignment.pdf and http://www.biobase.dk/P450/Figure4_alignment.pdf. The Neighbor-Join trees were analyzed with 1000 bootstrap trials. Exclude positions with gaps and correct for multiple substitutions were both set to off.

Non-*Arabidopsis* family 1 glycosyltransferase sequences were chosen using the sequence list available at the UDP-glucuronosyltransferase homepage (http://www.unisa.edu.au/pharm_medsci/Gluc_trans/table21.htm) or identified via BLAST searches. When available, one sequence from each subfamily was used. In cases where more than one species was represented in a subfamily, *Arabidopsis* sequences were preferred. All other species were chosen at random and are tabularized in Table 1.

4.3. Intron maps

Arabidopsis UGT intron maps were constructed by determining the intron splice site phase and position using a combination of the NetGene2 prediction server, PUBLISH from GCG Wisconsin package version 10.2, and family sequence similarity as done previously with the *Arabidopsis* P450s (Paquette et al., 2000; Werck-Reichhart et al., 2002). These splice site positions are numbered relative to their position on the consensus sequence produced by the multiple alignment of the *Arabidopsis* UGTs.

4.4. Chromosome maps

Chromosome maps were constructed by BLAST comparison of each UGT sequence against whole chromosome sequence. UGTs were grouped when the distance between them was less than 300kbp. Direction of the reading frame is compared to the chromosome sequence available from GenBank (ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/A_thaliana/).

Acknowledgements

The authors wish to thank Hans Ullitz Møller and Bo Mikkelsen of BioBase, the Danish Biotechnological Database, for advice and for hosting *The Arabidopsis P450, cytochrome b5, P450 reductase, and Glycosyltransferase Family 1 Site at PlaCe* (htpp://www.biobase.dk/P450). We thank Dr. Peter MacKenzie and Dr. Dianna Bowles for help on UGT nomenclature. The Danish Research Foundation supported this work.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Campbell, J.A., Davies, G.J., Bulone, V., Henrissat, B., 1997. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem. J. 326 (Pt 3), 929–939.

Benner, S.A., Cohen, M.A., Gonnet, G.H., 1994. Amino acid substitutions during functionally constrained divergent evolution of protein sequences. Protein Eng. 7, 1323–1332.

Bork, P., Gibson, T.J., 1996. Applying motif and profile searches. Methods Enzymol. 266, 162–184.

Brocchieri, L., 2001. Phylogenetic inferences from molecular sequences: review and critique. Theor. Popul. Biol. 59, 27–40.

Bruno, W.J., Socci, N.D., Halpern, A.L., 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. 17, 189–197.

Burbulis, I.E., Winkel-Shirley, B., 1999. Interactions among enzymes of the *Arabidopsis* flavonoid biosynthetic pathway. Proc. Natl. Acad. Sci. USA 96, 12929–12934.

Durst, F., Nielsen, D.R., 1995. Diversity and evolution of plant P450 and P450-reductases. Drug Metab. Drug Interact. 12, 189–206.

Douglas, S.E., 1998. Plastid evolution: origins, diversity, trends. Curr. Opin. Genet. Dev. 8, 655–661.

Gonnet, G.H., Cohen, M.A., Benner, S.A., 1992. Exhaustive matching of the entire protein sequence databse. Science 256, 1443–1445.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P., Brunak, S., 1996. Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. Nucleic Acids Res. 24, 3439–3452.

Hong, Z., Zhang, Z., Olson, J.M., Verma, D.P., 2001. A novel UDP-glucose transferase is part of the callose synthase complex and interacts with phragmoplastin at the forming cell plate. Plant Cell 13, 769–779.

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

Jackson, R.G., Lim, E.K., Li, Y., Kowalczyk, M., Sandberg, G., Hoggett, J., Ashford, D.A., Bowles, D.J., 2001. Identification and biochemical characterization of an Arabidopsis indole-3-acetic acid glucosyltransferase. J. Biol. Chem. 276, 4350–4356.

Jones, P., Vogt, T., 2001. Glycosyltransferases in secondary plant metabolism: tranquilizers and stimulant controllers. Planta 213, 164–174.

Jones, P.R., Møller, B.L., Hoj, P.B., 1999. The UDP-glucose:*p*-hydroxymandelonitrile-*O*-glucosyltransferase that catalyzes the last step in synthesis of the cyanogenic glucoside dhurrin in *Sorghum bicolor*. Isolation, cloning, heterologous expression, and substrate specificity. Journal of Biol. Chem. 274, 35483–35491.

Jorasch, P., Warnecke, D.C., Lindner, B., Zahringer, U., Heinz, E., 2000. Novel processive and nonprocessive glycosyltransferases from *Staphylococcus aureus* and *Arabidopsis thaliana* synthesize glycoglycerolipids, glycophospholipids, glycosphingolipids and glycosylsterols. Eur. J. Biochem. 267, 3770–3783.

Jorasch, P., Wolter, F.P., Zahringer, U., Heinz, E., 1998. A UDP glucosyltransferase from *Bacillus subtilis* successively transfers up to four glucoseresidues to 1,2-diacylglycerol: expression of ypfP in *Escherichia coli* and structural analysis of its reaction products. Mol. Microbiol. 29, 419–430.

Kahn, R., Durst, F., 2001. Function and evolution of plant cytochrome P450. Recent. Adv. Phytochem. 34, 151–189.

Li, Y., Baldauf, S., Lim, E.K., Bowles, D.J., 2001. Phylogenetic analysis of the UDP-glycosyltransferase multigene family of *Arabidopsis thaliana*. J. Biol. Chem. 276, 4338–4343.

Lim, E.K., Doucet, C.J., Li, Y., Elias, L., Worrall, D., Spencer, S.P., Ross, J., Bowles, D.J., 2002. The activity of *Arabidopsis* glycosyltransferases toward salicylic acid, 4-hydroxybenzoic acid, and other benzoates. J. Biol. Chem. 277, 586–592.

Lim, E.K., Li, Y., Parr, A., Jackson, R., Ashford, D.A., Bowles, D.J., 2001. Identification of glucosyltransferase genes involved in sinapate metabolism and lignin synthesis in *Arabidopsis*. J. Biol. Chem. 276, 4344–4349.

Long, M., Rosenberg, C., Gilbert, W., 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. Pro. Natl. Acad. Sci. USA 92, 12495–12499.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequence of duplicated genes. Science 290, 1151–1155.

Mackenzie, P.I., Owens, I.S., Burchell, B., Bock, K.W., Bairoch, A., Belanger, A., Fournel-Gigleux, S., Green, M., Hum, D.W., Iyanagi, T., Lancet, D., Louisot, P., Magdalou, J., Chowdhury, J.R., Ritter, J.K., Schachter, H., Tephly, T.R., Tipton, K.F., Nebert, D.W., 1997. The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. Pharmacogenetics 7, 255–269.

Milkowski, C., Baumert, A., Strack, D., 2001. Identification of four *Arabidopsis* genes encoding hydroxycinnamate glucosyltransferases. FEBS Letters 486, 183–184.

Milkowski, C., Baumert, A., Strack, D., 2000. Cloning and heterologous expression of a rape cDNA encoding UDP-glucose:sinapate glucosyltransferase. Planta 211, 883–886.

Møller, B.L., Conn, E.E., 1980. The biosynthesis of cyanogenic glucosides in higher plants. Channeling of intermediates in dhurrin biosynthesis by a microsomal system from *Sorghum bicolor* (linn) Moench. J. Biol. Chem. 255, 3049–3056.

Page, R.D., 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12, 357–358.

Paquette, S.M., Bak, S., Feyereisen, R., 2000. Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. DNA Cell Biol. 19, 307–317.

Philippe, H., Laurent, J., 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. 8, 616–623.

Ross, J., Li, Y., Lim, E., Bowles, D.J., 2001. Higher plant glycosyltransferases. Genome Biol. 2, 3004.1–3004.6.

Sandermann Jr., H., 1992. Plant metabolism of xenobiotics. Trends Biochem. Sci. 17, 82–84.

Shimojima, M., Ohta, H., Iwamatsu, A., Masuda, T., Shioi, Y., Takamiya, K., 1997. Cloning of the gene for monogalactosyldiacylglycerol synthase and its evolutionary origin. Proc. Natl. Acad. Sci. USA 94, 333–337.

Stoltzfus, A., Logsdon jr, J.M., Palmer, J.D., Doolittle, W.F., 1997. Intron "sliding" and the diversity of intron positions. Proc. Natl. Acad. Sci. USA 94, 10739–10744.

TAGI (The *Arabidopsis* Genome Initiative), 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

Tattersall, D.B., Bak, S., Jones, P.R., Olsen, C.E., Nielsen, J.K., Hansen, M.L., Høj, P.B., Møller, B.L., 2001. Resistance to an herbivore through engineered cyanogenic glucoside synthesis. Science 293, 1826–1828.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

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.

Vogt, T., Jones, P., 2000. Glycosyltransferases in plant natural product synthesis: characterization of a supergene family. Trends Plant Sci. 5, 380–386.

Walsh, J.B., 1987. Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics 117, 543–557.

Walsh, J.B., 1995. How often do duplicated genes evolve new functions? Genetics 139, 421–428.

Warnecke, D., Erdmann, R., Fahl, A., Hube, B., Muller, F., Zank, T., Zahringer, U., Heinz, E., 1999. Cloning and functional expression of UGT genes encoding sterol glucosyltransferases from *Saccharomyces cerevisiae*, *Candida albicans*, *Pichia pastoris*, and *Dictyostelium discoideum*. J. Biol. Chem. 274, 13048–13059.

Werck-Reichhart, D., Bak, S., Paquette, S., 2002. Cytochromes P450. In: Somerville, C.R., Meyerowitz, E.M. (Eds.), The *Arabidopsis* Book. American Society of Plant Biologists, Rockville, MD, pp. 1–29. (doi/10.1199/tab.0028) http://www.aspb.org/downloads/arabidopsis/werckfinal.pdf.