Protein family review

Higher plant glycosyltransferases

Joe Ross, Yi Li, Eng-Kiat Lim and Dianna J Bowles

Address: Department of Biology, University of York, York, YO10 5DD, UK.

Correspondence: Joe Ross. E-mail: jr23@york.ac.uk

Published: 7 February 2001

Genome ***Biology*** 2001, **2(2)**:reviews3004.1–3004.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/2/reviews/3004

© BioMed Central Ltd (Print ISSN 1465-6906; Online ISSN 1465-6914)

Summary

Uridine diphosphate (UDP) glycosyltransferases (UGTs) mediate the transfer of glycosyl residues from activated nucleotide sugars to acceptor molecules (aglycones), thus regulating properties of the acceptors such as their bioactivity, solubility and transport within the cell and throughout the organism. A superfamily of over 100 genes encoding UGTs, each containing a 42 amino acid consensus sequence, has been identified in the model plant *Arabidopsis thaliana*. A phylogenetic analysis of the conserved amino acids encoded by these *Arabidopsis* genes reveals the presence of 14 distinct groups of UGTs in this organism. Genes encoding UGTs have also been identified in several other higher plant species. Very little is yet known about the regulation of plant UGT genes or the localization of the enzymes they encode at the cellular and subcellular levels. The substrate specificities of these UGTs are now beginning to be established and will provide a foundation for further analysis of this large enzyme superfamily as well as a platform for future biotechnological applications.

Gene organization and evolutionary history

Classification

Glycosyltransferases (EC 2.4.x.y) catalyze the transfer of sugars to a wide range of acceptor molecules. The enzymes can be classified into families on the basis of sequence similarity, catalytic specificity and the existence of consensus sequences [1-3]. This review concerns only one family of UDP glycosyltransferases in higher plants: that defined by the presence of a carboxy-terminal consensus sequence, the UDP-glycosyltransferases signature, that is thought to be involved in binding of the protein to the UDP moiety of the sugar nucleotide (Figure 1) [3,4]. This consensus sequence can be identified in open reading frames from animal, plant, yeast and bacterial genomes, and it probably defines a single multigene superfamily. The term 'UGT' will be used throughout this review to refer specifically to those glycosyltransferases containing this consensus sequence. The nomenclature of the UGT superfamily is co-ordinated by a group of scientists who were invited by the relevant international nomenclature committees to help with the systematic naming of these and other carbohydrate-handling enzymes [3]. Figure 2 summarizes the system currently used by this group; it is based primarily on amino acid sequence identity. All *Arabidopsis thaliana* UGT genes discussed in this review have been named using this nomenclature system. A broader classification system for all NDP-sugar hexosyltransferases (EC 2.4.1.x) has also been described, and this groups all known glycosyltransferases into 47 distinct families [1,5]. The UGT genes discussed in this review fall into family 1 within the latter classification system. The evolutionary relationship of the different glycosyltransferase families is not yet clear although recent clues have emerged as an increasing number of three-dimensional structures are determined (see the Characteristic structural features section).

Evolution

The sequencing of the model plant species *A. thaliana* has recently been completed [6]. Using the UGT amino acid

[FW]-[2X]-Q-[2X]-[LIVMYA]-[LIMV]-[4-6X]-[LVGAC]-[LVFYA]-[LIVMF]-[STAGCM]-[HNQ]-[STAGC]-G-[2X]-[STAG]-[3X]-[STAGL]-[LIVMFA]-[4X]-[PQR]-[LIVMT]-[3X]-[PA]-[3X]-[DES]-[QEHN]

Figure 1
Amino acid consensus sequence of UDP-glycosyltransferases taken from the PROSITE database of protein families and domains, which was used to identify the 107 *A. thaliana* UGT genes [7]. Letters in brackets denote alternative amino acids at a particular position; X denotes any amino acid.

consensus sequence shown in Figure 1 as a search tool, we have screened its genome and identified a very large glycosyltransferase superfamily containing 107 putative UGT genes and 10 UGT pseudogenes [7]. Analysis of this superfamily has allowed the first characterization of higher plant UGTs at the genomic level to be performed. Information available on other plant UGTs can now begin to be integrated into the results from this genomic analysis.

We have performed a detailed analysis of the amino acid sequences of the open reading frames of 88 *A. thaliana* UGT genes using neighbor-joining and parsimony-based analysis methods with statistical confidence measurements by bootstrap analysis [7], resulting in an unrooted phylogenetic tree consisting of 12 well-defined major evolutionary groups. An updated but less detailed analysis using 107 *A. thaliana* UGT genes (including the 88 analyzed in [7]) is shown in Figure 3. Further refinement of more closely related sequences has been shown in the equivalent analysis of 88 *Arabidopsis* UGTs [7]. Bootstrap analysis of this expanded tree shows that the superfamily is likely to contain 14 distinct groups that evolved from an equivalent number of ancestral UGT genes.

Using programs capable of detecting more distantly related sequences, such as PSI-BLAST, two additional *A. thaliana* genes have recently been identified that contain amino acid sequences similar to the UGT consensus sequence. These genes encode proteins 100 residues longer than any of the previously identified *A. thaliana* UGT genes and each contains 13 introns (see the Gene organization section of this article for details of the intron organization of UGT genes). One of these genes has been previously identified as a UDP-glucose sterol β-D-glucosyltransferase [8].

No comparable analysis has yet been carried out for other plant species. Given that so many UGT sequences can be found in the comparatively small genome of *A. thaliana*, however, it is probable that large numbers will also be detected in species throughout the plant kingdom. Similarly, large numbers can be detected in species of the animal kingdom - such as the 60-gene UGT superfamily in *Caenorhabditis elegans*. A complete list of UGTs currently annotated can be found at the UDP Glucuronosyltransferase home page [4].

Gene organization

The *A. thaliana* UGT genes are scattered throughout the genome but do show clustering into groups of two to seven genes; clustered genes show a high degree of amino acid sequence similarity. The genes encoding the *A. thaliana* UGTs contain up to two introns, but over half (58/107) contain no introns (Figure 4). An analysis of the intron-containing UGT genes suggests that a minimum of ten independent intron-insertion events and either one or two intron-loss events have occurred during *A. thaliana* UGT evolution (Figure 3) [7].

Characteristic structural features

Sequence features

The amino acid sequences encoded by the UGT genes containing the consensus sequence shown in Figure 1, which vary in length from 435 to 507 amino acids, have all been found to possess nine conserved regions, including the UGT-defining consensus sequence (Figure 4) [7]. The level of similarity between these UGT amino acid sequences varies from over 95% to lower than 30% identity. The amino-terminal regions are more variable than the carboxy-terminal regions, supporting the suggestion that the domain involved in the recognition and binding of the diverse aglycone substrates is

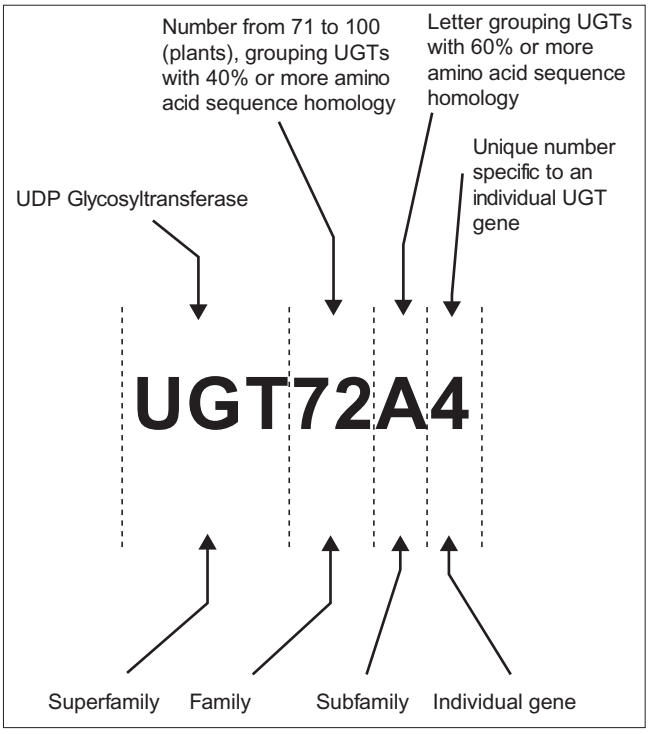

Figure 2
Summary of the current UGT superfamily nomenclature system. The diagram illustrates the system currently used to name plant UDP glycosyltransferases. Further details of this nomenclature system can be found in [3] and on the UDP Glucuronosyltransferase home page [4].

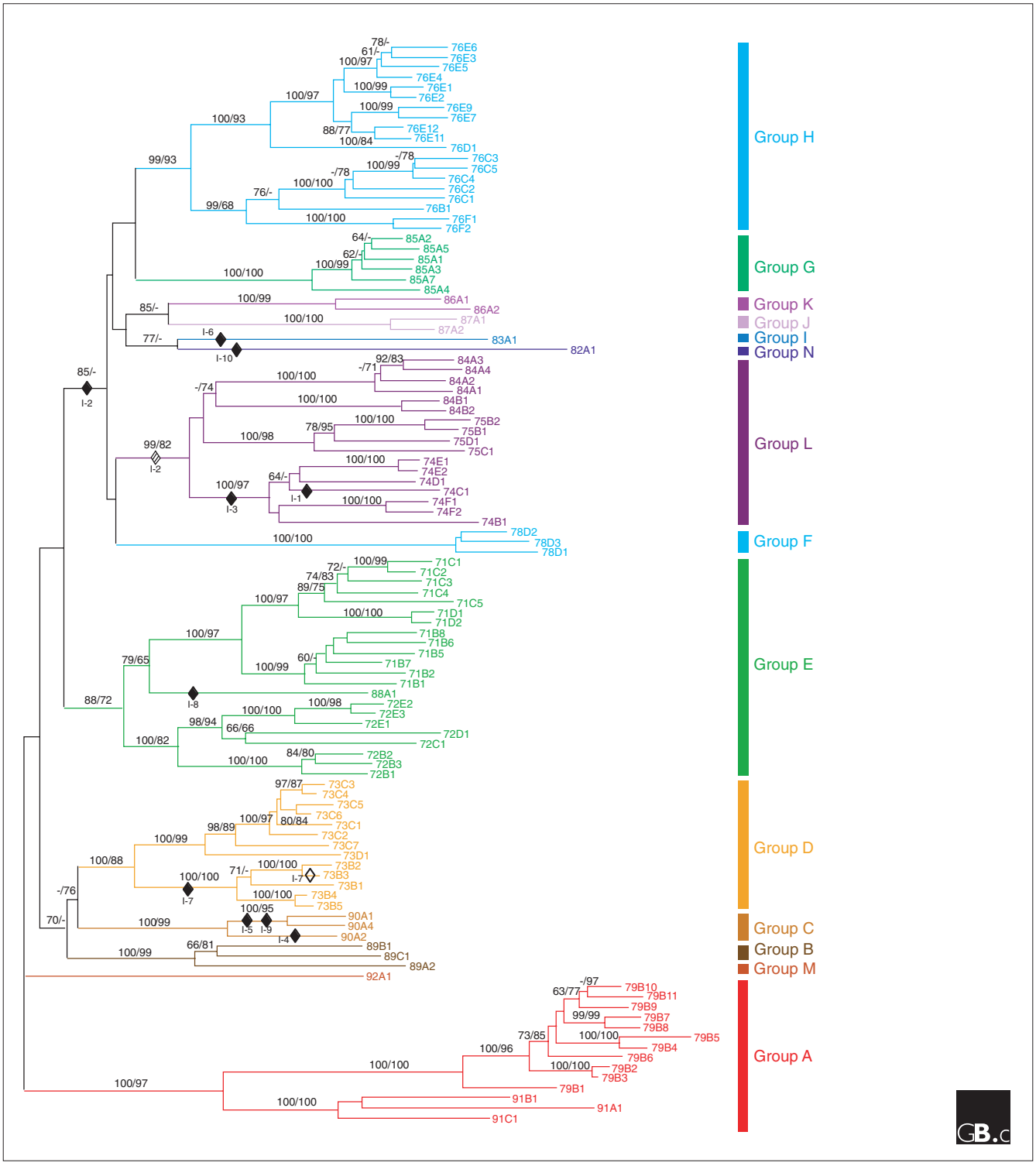

Figure 3
Phylogenetic analysis of the *Arabidopsis* UGT superfamily. Neighbor-joining and parsimony-based analysis of nine conserved amino acid sequences shown in Figure 4 was performed as described previously [7]. Bootstrap values over 60% are indicated above the nodes, with the number on the left indicating neighbor-joining and that on the right indicating parsimony. Dashes indicate bootstrap values under 60%. Further refinement of more closely related sequences has been shown in the equivalent analysis of 88 *Arabidopsis* UGTs [7]. Hypothetical intron gains and losses are indicated by diamonds with the intron number (I) shown (see Figure 4). Postulated intron gains are indicated by filled diamonds, intron losses by unfilled diamonds and the questionable intron loss by a striped diamond.

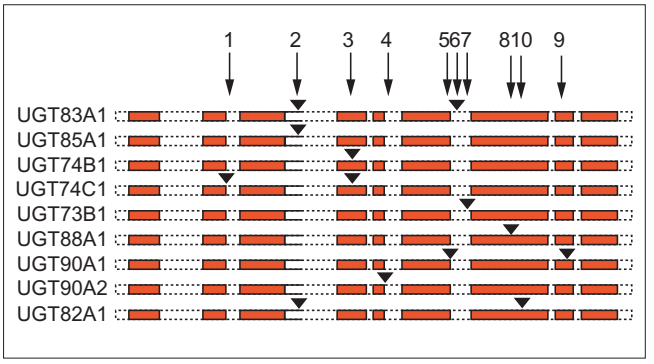

Figure 4
The conserved regions and intron positions of the UGT genes of *A. thaliana*. The nine conserved amino acid regions are shown as red boxes. Segments between these boxes represent regions with a variable number of residues. The positions of introns are indicated by arrows and inverted triangles. Examples of UGT genes containing one or more of the ten introns are shown.

located towards the amino terminus of the protein whereas the carboxy-terminal region encodes a domain involved in binding the nucleotide sugar substrate [9].

Structural features

To date, none of the proteins encoded by the UGT superfamily has been crystallized and their three-dimensional structures are not known. Six glycosyltransferases from other superfamilies have been analyzed structurally, however, and these analyses suggest that, although they were previously thought to be unrelated, they may fall into just two superfamilies [10]. The first of these contains bacteriophage T4 β-glucosyltransferase (BGT) and the *Escherichia coli N*-acetylglucosaminyltransferase MurG, and the second contains *Bacillus subtilis* glycosyltransferase SpsA, bovine β-1,4-galactosyltransferase 1, rabbit *N*-acetylglucosaminyltransferase I and the catalytic fragment of the human glucuronyltransferase I. Interestingly, an approximately 30 amino acid sequence motif in MurG, suggested by the structure to be involved in nucleotide-sugar binding, has been shown to be similar to the UGT consensus sequence described above (Figure 1) [2,11]. Further insight into UGT structure and subsequent structure-function relationship now awaits the resolution of a three-dimensional structure for an enzyme from this superfamily.

Localization and function

Localization

Mammalian UGTs, which transfer glucuronic acid to hydrophobic substrates, are membrane-bound enzymes localized in the endoplasmic reticulum with their catalytic sites facing the lumen. These enzymes contain an amino-terminal leader sequence that is cleaved on cotranslational segregation into the rough endoplasmic reticulum, and a hydrophobic carboxy-terminal halt sequence that anchors the enzyme to the membrane [11]. Our analyses of *A. thaliana* UGTs using TopPred2, SignalP and Psort programs has not identified either of these motifs, supporting the widely held belief that plant UGTs are cytoplasmic enzymes.

Very little information is available from plants regarding the expression of UGT genes. Tomato and tobacco UGTs have been shown to respond rapidly to signals from wounds and pathogen attack [12,13]. There are also now significant data available from the Stanford microarray website on expression [14] of 14 of the 107 *A. thaliana* UGTs. The high level of sequence homology between family members suggests, however, that expression data using either expressed sequence tag (EST) or full length cDNA probes should be treated with caution, as full-length probes may well hybridize to several closely related UGTs and produce misleading expression profiles. No data are yet available to evaluate whether UGT expression is regulated principally at the DNA, RNA or protein level.

Functions

The UGT superfamily in higher plants is thought to encode enzymes that glycosylate a broad array of aglycones, including plant hormones, all major classes of plant secondary metabolites, and xenobiotics such as herbicides [15]. Glycosylation regulates many properties of the aglycones, such as their bioactivity, their solubility and their transport properties within the cell and throughout the plant. In addition to the *A. thaliana* UGT genes, numerous UGT genes have been isolated from a wide range of different plant species and their corresponding gene products either characterized biochemically or defined by genetic analysis [15]. An alignment of these sequences with the *A. thaliana* UGT superfamily and their phylogenetic analysis predicts their position on the *A. thaliana* UGT tree [7], which is shown in Figure 5 along with the available data on substrate specificity of these enzymes.

The task of comprehensively assaying UGT substrate specificity is a formidable one and much work remains to be done. Nevertheless, the identification of substrate specificity of higher plant UGTs is beginning to allow some conclusions to be drawn and some interesting relationships between different UGTs to be detected. For example: enzymes that catalyze the formation of salicylic glucose ester and indole-3-acetic acid glucose ester share the highest sequence homology to Group L from *Arabidopsis*, which contains enzymes that produce hydroxycinnamoyl glucose ester [15-17]; three UGTs known to be involved in the 3-*O*-glucosylation of anthocyanidin in both monocotyledons and dicotyledons are all clustered with the *Arabidopsis* Group F [18]; and two highly homologous sequences encoding enzymes that glycosylate the plant hormone zeatin are distinct from all the major UGT groups of *Arabidopsis,* suggesting the possible presence of *Arabidopsis* zeatin glycosyltransferases that have not been identified in the *A. thaliana* UGT superfamily [19].

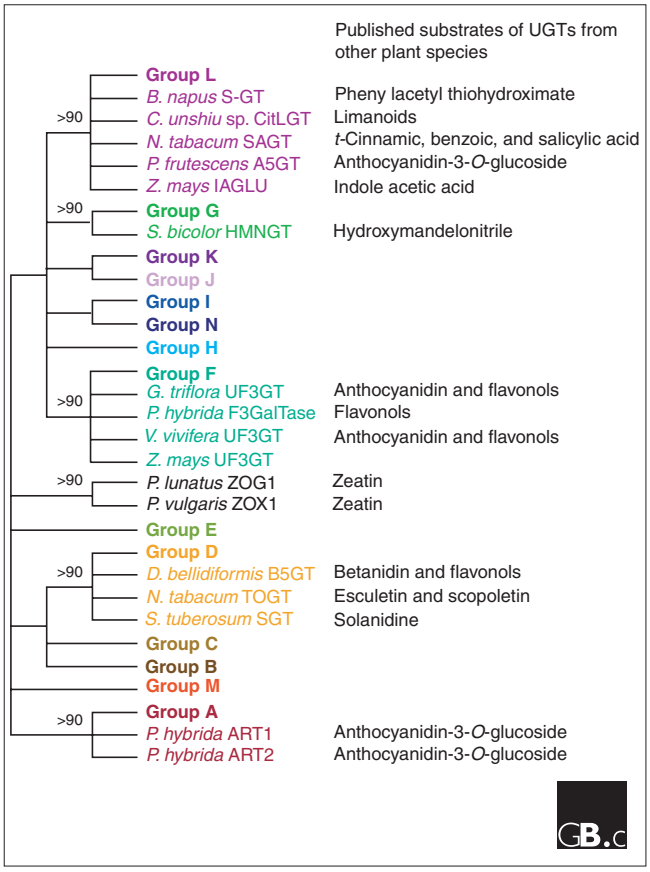

Figure 5
The relationship of the groups of *A. thaliana* UGTs with other published plant UGTs. A simplified version of the *A. thaliana* UGT phylogenetic tree is shown with other plant UGTs added. The bootstrap values, which give the degree of confidence in the branching pattern presented, are 60-90% unless otherwise stated. The published substrate specificities for UGTs other than *A. thaliana* are listed to the right of the figure. Full species names referred to in the figure are as follows: *Brassica napus, Citrus unshiu, Nicotiana tabacum, Perilla frutescens, Zea mays, Sorghum bicolor, Gentiana triflora, Perilla hybrida, Vitis vinifera, Phaseolus lanatus, Phaseolus vulgaris, Dorotheanthus bellidiformis, Solanum tuberosum.*

These data, taken together, provide a useful foundation for starting to understand the structure-activity relationships of the UGT family. It will be interesting to compare the catalytic specificity *in vitro* with the consequences of changing the level of individual enzymes *in vivo*. A broad specificity of recombinant enzymes *in vitro* may not provide insight into the activity *in planta*, because substrate availability will also be relevant in the cellular context.

It has been suggested that many UGTs may not exhibit high substrate specificity at all, but rather recognize individual hydroxyl groups present on a wide range of different aglycones [15]. Our substrate-specificity data do not seem to support this suggestion, as screening of 36 *Arabidopsis* UGTs revealed only one enzyme capable of glucosylating indole-3-acetic acid [16]. Thus, for at least certain UGTs, reactions may be directed by substrate specificity rather than regiospecificity. A much clearer picture will emerge when substrates of more *Arabidopsis* enzymes have been identified and these data are considered within the context of temporal and spatial expression profiles *in planta*.

Enzyme mechanism

UGTs transfer nucleotide-diphosphate-activated sugars to low-molecular-weight aglycone substrates. In plants, the activated sugar is usually UDP-glucose but other sugars such as UDP-xylose [19] are also found. The conjugation of the sugar can lead to the formation of a range of glycosylated molecules including glucose esters, cyanogenic glucosides, phenolic glucosides and glucosinolates containing a β-thioglucose moiety. Many aglycones, such as the flavonols, can also accept more than one sugar if a number of sites are available for glycosylation. The exact catalytic mechanism used by UGTs is not yet known. As discussed above, the enzymes are generally thought to contain an aglycone-binding amino terminus and a UDP-sugar-binding carboxyl terminus but any conclusions regarding enzymatic mechanism await determination of the crystal structure.

Frontiers

It will be essential to integrate data from *in vitro* and *in vivo* studies to gain a more complete picture of the potential biological roles of UGTs in plants. This is now feasible with current technology: microarray data, details of the catalytic activities of specific recombinant proteins, metabolite profiles of plants over-expressing or lacking individual UGTs, as well as information on the cell- and tissue-specificity of gene expression, can all be accessed and integrated. Similarly, once the three-dimensional structure of one UGT has been accomplished, molecular modeling will provide very rapid insights into the structural relatedness of other superfamily members and how this relatedness is reflected in catalytic activities.

The recent realization that *Arabidopsis*, with such a small genome relative to other species in the plant kingdom, has so many UGTs opens up a whole range of new frontiers, both for the fundamental understanding of UGT functions and for the many strategic applications of the UGT superfamily.

Additional data file

An Excel file containing the accession numbers of the *Arabidopsis* BAC clones that contain UGT genes and the location of each gene in the clone is included online (file added on 17 July 2001).

References

1. Campbell JA, Davies GJ, Bulone V, Henrissat B: **A classification of nucleotide-diphospho-sugar glycosyltransferases based on**

amino acid sequence similarities [published erratum appears in *Biochem J* 1998, **329:**719]. *Biochem J* 1997, **326:**929-939.
Describes a comprehensive classification of all known NDP-sugar hexosyltransferases and presents the use of this system to group the known glycosyltransferases into 26 families. A more recent version of this analysis is available at the Introduction to Glycosyltransferase website [5].

2. Kapitonov D, Yu RK: **Conserved domains of glycosyltransferases.** *Glycobiology* 1999, **9:**961-978.
Identifies and aligns three glycosyltransferase conserved domains. The evolutionary relationship of each of these domains is presented along with a potential mechanism for the glycosyltransferase catalytic reaction.
3. Mackenzie P, Owens I, Burchell B, Bock K, Bairoch A, Bélanger A, Fournel-Gigleux S, Green M, Hum D, Iyanagi T, *et al.*: **The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence.** *Pharmacogenetics* 1997, **7:**255-269.
An update of the nomenclature system for UDP glycosyltransferases. Amino acid sequences of proteins from animal, yeast, plant and bacteria are compared to define 33 families.
4. **UDP Glucuronosyltransferase home page**
[http://www.unisa.edu.au/pharm_medsci/Gluc_trans/Gt_ttl.htm]
This is the home page of the committee for naming UDP glucuronosyltransferase. The site has links to relevant databases and information resources.
5. **Introduction to Glycosyltransferase**
[http://afmb.cnrs-mrs.fr/CAZY/GT.html]
Site further describing the classification of glycosyltransferases that use nucleotide diphospho-sugars, nucleotide monophospho-sugars and sugar phosphates (EC 2.4.1.x). Enzymes are grouped into distinct sequence-based families [1].
6. The *Arabidopsis* Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796-815.
Report describing the sequencing of the complete genome of *Arabidopsis thaliana*.
7. Li Y, Baldauf S, Lim E-K, Bowles DJ: **Phylogenetic analysis of the UDP-glycosyltransferase multigene family of *Arabidopsis thaliana*.** *J Biol Chem* 2001, in press.
Phylogenetic analysis of 88 UGT amino acid sequences, defining 12 major groups of glycosyltransferases.
8. Warnecke DC, Baltrusch M, Buck F, Wolter FP, Heinz E: **UDP-glucose:sterol glucosyltransferase: cloning and functional expression in *Escherichia coli*.** *Plant Mol Biol* 1997, **35:**597-603.
A description of the properties of a membrane bound UDP-glucosyltransferase that glucosylates plant sterols.
9. Meech R, MacKenzie PI: **Structure and function of uridine diphosphate glucuronosyltransferases.** *Clin Exp Pharmacol Physiol* 1997, **24:**907-915.
Discusses the concepual division of UGT proteins into two domains, an amino-terminal half containing the aglycone binding site and a carboxy-terminal half believed to contain a UDP-sugar binding site.
10. Unligil UM, Rini JM: **Glycosyltransferase structure and mechanism.** *Curr Opin Struct Biol* 2000, **10:**510-517.
Summary of the six known glycosyltransferase three-dimensional structures and discussion of the grouping of these enzymes into two superfamilies on the basis of their structural similarities.
11. Radominska-Pandya A, Czernik PJ, Little JM, Battaglia E, MacKenzie PI: **Structural and functional studies of UDP-glucuronosyltransferases.** *Drug Metab Rev* 1999, **31:**817-899.
Review describing current information on substrate specificity, structure and topology of UGT1A and 2B family glucuronosyltransferases.
12. O'Donnell PJ, Truesdale MR, Calvert CM, Dorans A, Roberts MR, Bowles DJ: **A novel tomato gene that rapidly responds to wound- and pathogen-related signals.** *Plant J* 1998, **14:**137-142.
Characterization of a wound-induced glucosyltransferase gene.
13. Roberts MR, Warner SAJ, Darby R, Lim EK, Draper J, Bowles DJ: **Differential regulation of a glucosyl transferase gene homologue during defence responses in tobacco.** *J Exp Bot* 1999, **50:**407-410.
Investigation of the expression profile of a glucosyltransferase that is rapidly induced during the defence response.
14. **Stanford Microarray Database**
[http://genomewww4.stanford.edu/MicroArray/SMD/]
Site providing raw and normalized data from microarray experiments as well as their corresponding image files.
15. Vogt T, Jones P: **Glycosyltransferases in plant natural product synthesis: characterization of a supergene family.** *Trends Plant Sci* 2000, **5:**380-386.
A recent review of glycosyltransferases of plant secondary metabolism.
16. Jackson R, Lim E-K, Li Y, Kowalczyk M, Sandberg G, Hoggett J, Ashford DA, Bowles DJ: **Identification and biochemical characterisation of an *Arabidopsis* indole-3-acetic acid glucosyltransferase**. *J Biol Chem* 2001, in press.
This report describes the *in vitro* substrate specificity of a member of the *A. thaliana* UGT superfamily that shows activity to IAA.
17. Lim E-K, Li Y, Parr A, Jackson J, Ashford DA, Bowles DJ: **Identification of glucosyltransferase genes involved in sinapate metabolism and lignin synthesis in *Arabidopsis*.** *J Biol Chem* 2001, in press.
An analysis of the *in vitro* expression of 36 recombinant *A. thaliana* UGTs and the identification of enzymes that produce hydroxycinnamoyl glucose conjugates.
18. Ford CM, Boss PK, Hoj PB: **Cloning and characterization of *Vitis vinifera* UDP-glucose:flavonoid 3-O-glucosyltransferase, a homologue of the enzyme encoded by the maize Bronze-1 locus that may primarily serve to glucosylate anthocyanidins *in vivo*.** *J Biol Chem* 1998, **273:**9224-9233.
Describes the identification of a UGT with activity to anthocyanidin substrates.
19. Martin RC, Mok MC, Mok DW: **A gene encoding the cytokinin enzyme zeatin O-xylosyltransferase of *Phaseolus vulgaris*.** *Plant Physiol* 1999, **120:**553-558.
Description of a zeatin glycosyltransferase that uses UDP-xylose as the nucleotide sugar donor.