Fold recognition analysis of glycosyltransferase families: further members of structural superfamilies

Octávio L. Franco1,2,3 and Daniel J. Rigden2
2Embrapa Genetic Resources and Biotechnology, Cenargen/Embrapa, Brasília-DF, Brazil; and 3Universidade Católica de Brasília, Pós-Graduação em Ciências Genômicas e Biotecnologia, SGAN Quadra 916, Módulo B, Av. W5 Norte70, 790-160 Brasília-DF, Brazil

Received on April 24, 2003; revised on June 27, 2003; accepted on June 27, 2003

Glycosyltransferases (GTs) are diverse enzymes organized into 65 families. X-ray crystallography and in silico studies have shown many of these to belong to two structural superfamilies: GT-A and GT-B. Through application of fold recognition and iterated sequence searches, we demonstrate that families 60, 62, and 64 may also be grouped into the GT-A fold superfamily. Analysis of conserved acidic residues suggests that catalytic sites are better conserved in superfamily GT-B than in GT-A. Although 26% and 29% of GT families may now be confidently placed in superfamilies GT-A and GT-B, respectively, the remaining 45% of families bear no discernible resemblance to either superfamily, which, given the sensitivity of modern fold recognition methods, suggests the existence of novel structural scaffolds associated with GT activity. Furthermore, bioinformatics studies indicate the apparent ease with which mechanism—inverting or retaining—may change during evolution.

Key words: evolutionary relationships/fold recognition/ glycosyltransferases/MurG/SpsA

Introduction

Glycosyltransferases (GTs; EC 2.4.±.±) constitute a large group of enzymes that are involved in the biosynthesis of oligosaccharides and polysaccharides and that act through the transfer of sugar moieties from activated donor molecules to specific acceptor molecules, forming glycosidic bonds. Particularly abundant are a group of enzymes, present in both prokaryotes and eukaryotes, that utilize an activated nucleotide sugar as a donor and plays significant roles in important biological processes (Verbert and Cacan, 1999).

As of June 2003, the carbohydrate active enzymes (CAZY) classification contains 65 GT families, defined on the basis of sequence similarity (Coutinho and Henrissat, 1999; Coutinho et al., 2003). With the ongoing determinations of GTases structures (for review, see Unligil and Rini, 2000), and computational analyses (Wabl and Grishin, 2001; Breton et al., 2002) a picture has emerged of two GT superfamilies, each containing various families, which do not necessarily share significant sequence similarity. The most studied family within superfamily GT-A is family 2, which contains the inverting glycosyltransferase SpsA from Bacillus subtilis (Figure 1) (Charnock and Davies, 1999; Tarbouriech et al., 2001). This enzyme acts in spore coat formation, and its homologues include cellulose synthase and numerous proteins involved in bacterial cell surface glyclosylation (as reviewed by Unligil and Rini, 2000).

The structure of SpsA is a single domain consisting of parallel β-strands flanked on either side by α-helices (Figure 1) (Charnock and Davies, 1999). Within GT-B, family 28 has received most attention, particularly the MurG protein. This enzyme is an N-acetylglicosaminytransferase involved in the intracellular phase of bacterial peptidoglycan biosynthesis that catalyzes the transfer of N-acetyl-D-glucosamine (GlcNAc) from UDP-GlcNAc to the C4 hydroxyl group of a lipid-linked N-acetylmuramoyl pentapeptide (Ikeda et al., 1990). The structure of MurG contains two domains, each of the Rossmann fold α/β open sheet structure separated by a deep cleft in which the substrates bind (Figure 1) (Ha et al., 2000, Hu et al., 2003).

As of June 2003, the carbohydrate active enzymes (CAZY) classification contains 65 GT families, defined on the basis of sequence similarity (Coutinho and Henrissat, 1999; Coutinho et al., 2003). With the ongoing determinations of GTases structures (for review, see Unligil and Rini, 2000), and computational analyses (Wabl and Grishin, 2001; Breton et al., 2002) a picture has emerged of two GT superfamilies, each containing various families, which do not necessarily share significant sequence similarity. The most studied family within superfamily GT-A is family 2, which contains the inverting glycosyltransferase SpsA from Bacillus subtilis (Figure 1) (Charnock and Davies, 1999; Tarbouriech et al., 2001). This enzyme acts in spore coat formation, and its homologues include cellulose synthase and numerous proteins involved in bacterial cell surface glyclosylation (as reviewed by Unligil and Rini, 2000).

The structure of SpsA is a single domain consisting of parallel β-strands flanked on either side by α-helices (Figure 1) (Charnock and Davies, 1999). Within GT-B, family 28 has received most attention, particularly the MurG protein. This enzyme is an N-acetylglicosaminytransferase involved in the intracellular phase of bacterial peptidoglycan biosynthesis that catalyzes the transfer of N-acetyl-D-glucosamine (GlcNAc) from UDP-GlcNAc to the C4 hydroxyl group of a lipid-linked N-acetylmuramoyl pentapeptide (Ikeda et al., 1990). The structure of MurG contains two domains, each of the Rossmann fold α/β open sheet structure separated by a deep cleft in which the substrates bind (Figure 1) (Ha et al., 2000, Hu et al., 2003).

Here we present the results of an investigation by fold recognition of GT families of still unknown catalytic domain architecture by which families 60, 62, and 64 were grouped to the GT-A fold superfamily. The failure to assign folds to catalytic domains in the remaining families, despite the application of a battery of modern fold recognition methods, suggests that, contrary to the general supposition,
other folds are likely to be associated with GT activity. Surprisingly, just a single iteration of PSI-BLAST is required to demonstrate homology between inverting and retaining GT families.

Results and discussion

Fold recognition analysis and evolutionary relationships by PSI-BLAST

Representative sequences (see Table I) from GT families for which the catalytic domain structure was not known were submitted for analysis at the Meta-Server (Bujnicki et al., 2001a). Although this provides a convenient access to many different methods, we focused on the results of two consensus fold recognition methods—Pcons2 (Lundstrom et al., 2001) and Shotgun on 3 (Fischer, 2003)—that distinguish true and false positives more effectively than individual methods. For comparison we also monitored the results of the 3D-PSSM method, one of the best-performing individual methods (Kelley et al., 2000). Results are also summarized in Table I. To estimate confidence in the results we compared scores with the worst (highest-scoring) false positives produced by each individual method (http://bioinfo.pl/LiveBench; Bujnicki et al., 2001b). Currently, the worst three false positives of the Pcons2 method score 2.42, 1.87 and 1.13. For Shotgun on 3 the figures are 31.4, 22.6, and 21.5, whereas for 3D-PSSM, where the most confident hits have the lowest values, they are 0.037, 0.28, and 0.30. The fact that the top-scoring structures for the GT families presented in Table I were also GT enzymes, along with the unanimity of the predictions, provided further support for the correctness of the assignment, particularly in view of the current perception that most GT families have one of the two known GT-associated folds.

Fold recognition analyses showed that families 12, 16, 21, 25, 27, 45, 54, 55, 60, 62, and 64 matched the structure of SpsA (PDB: 1QG8), a GT from family 2, which belongs to superfAMILY GT-A. Relations between 2 and 12, 16, 21, 25, 27, 45, 54, and 55 have been previously noted (Breton et al., 2002; CAZY home page, http://afmb.cnrs-mrs.fr/CAZY). Thus in this study we have added three more families to superfAMILY GT-A (Table I). No further families could be allocated to GT-B other than those already assigned (Wrabl and Grishin, 2001).

### Table I. Families of glycosyltransferases mentioned in the text and belonging to the GT-A superfAMILY

<table>
<thead>
<tr>
<th>Family</th>
<th>Activities</th>
<th>3D NO. in CAZY</th>
<th>Mechanism</th>
<th>3D PSSMa</th>
<th>ShotGun on 3b</th>
<th>Pcons2c</th>
<th>Species</th>
<th>SwissProt Accession</th>
<th>Residues</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>02</td>
<td>Cellulose synthase (EC 2.4.1.12) Chitin synthase (EC 2.4.1.16)</td>
<td>H7L H7Q</td>
<td>Inverting</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Bacillus subtilis</td>
<td>P39621</td>
<td>256</td>
<td>Charnock and Davies, (1999)</td>
</tr>
<tr>
<td></td>
<td>Dolichyl-phosphate β-D-mannosyl transferase (EC 2.4.1.83)</td>
<td>1QG8</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Tarbouriech et al. (2001)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dolichyl-phosphate β-glucosyltransferase (EC 2.4.1.117)</td>
<td>1QGS</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N-acetylglucosaminyltransferase (EC 2.4.1--)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N-acetylgalactosaminyltransferase (EC 2.4.1--).</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hyaluronan synthase (EC 2.4.1--)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chitin oligosaccharide synthase (EC 2.4.1--)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>β-1,3-Glucan synthase (EC 2.4.1.34)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>UDP-GlcNAc:hydroxyproline polypeptide GlcNAc-transferase (EC 2.4.1--)</td>
<td>None</td>
<td>Unknown</td>
<td>0.25 1DBG</td>
<td>5.62 1QG8</td>
<td>1.28 1QG8</td>
<td>Yersinia pestis</td>
<td>Q8ZCN0 445</td>
<td>373</td>
<td>Parkhill et al. (2001)</td>
</tr>
<tr>
<td>62</td>
<td>α1-2-Mannosyltransferase (EC 2.4.1--);</td>
<td>None</td>
<td>Retaining</td>
<td>0.05 1QG8</td>
<td>35.49 1QG8</td>
<td>2.24 1QG8</td>
<td>Ajeellomyces capsulatus</td>
<td>Q9P426 373</td>
<td>373</td>
<td>Colonna-Romano et al. (1998)</td>
</tr>
<tr>
<td></td>
<td>α1,6-Mannosyltransferase (EC 2.4.1--)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>Heparan synthase (EC 2.4.1--)</td>
<td>None</td>
<td>Retaining</td>
<td>0.25 1QG8</td>
<td>21.41 1QG8</td>
<td>2.26 1QG8</td>
<td>Arabidopsis thaliana</td>
<td>Q9C975 329</td>
<td>329</td>
<td>Theologis et al. (2000)</td>
</tr>
</tbody>
</table>

Classified by CAZY, Coutinho and Henrissat, 1999. Experimentally determined structures are available for family 2, structural predictions are presented for the other families.

*a*Score by the 3D-PSSM fold recognition method (Kelley et al., 2000).

*b*Score by the Shotgun on 3 consensus fold recognition method (Fischer, 2003).

*c*Score by the Pcons2 consensus fold recognition method (Lundstrom et al., 2001).
A striking conclusion is that even after comprehensive fold recognition, no folds could be assigned to 45% of GT families, containing 39% of known GT sequences. There are two alternative explanations. First, a large number of extremely divergent GT families, in fact possessing known GT folds, are present in the CAZY database, the catalytic domain folds of which are not identifiable, even after the application of advanced fold recognition tools (Fischer and Rychlewski, 2003). However, given the success of fold recognition when applied to carbohydrate active enzymes (Rigden and Franco, 2002; Rigden, 2002) it is perhaps more likely to suppose the existence of one or more different folds, not currently associated with GT activity. Given the importance of GTs in general (Unligil and Rini, 2000) these would be very significant targets for structural determination.

Because fold recognition can produce significant results for structural analogs as well as distant structural homologs, we sought further evidence regarding evolutionary relationships with sensitive sequence comparisons carried out using PSI-BLAST. The results (Figure 2) provide evidence that most of the families recently identified as members of GT-A (Breton et al., 2002) (Table I) share a common evolutionary origin, although in four cases (families 16, 54, 55, and 60) no relationships could be established (Figure 2). None of the illustrated families leaves the network when only results obtained at the more conservative E-value threshold of 0.001 are considered.

Another important question relates to the evolution of the enzymatic mechanism—what ease can inverting enzymes evolve into retaining ones, and vice versa? As shown in Figure 2, only a single round of PSI-BLAST was sufficient to demonstrate a relationship between family 2 (inverting) and family 45 (retaining). Similarly, within superfamily GT-B, just two iterations of PSI-BLAST at 0.001 are necessary to demonstrate a relationship between retaining family 5 and inverting family 19 (Wrabl and Grishin, 2001; data not shown). These data also suggest that the transition from inverting to retaining mechanism, or the reverse, occurs relatively easily, so that multiple transitions have likely happened in each superfamily. These close relationships between retaining and inverting GTs have also been seen in other studies (Campbell et al., 1998; Breton et al., 1998).

Conservation of residues important to enzymatic activity
Current knowledge of GT mechanisms suggests the obligatory involvement of conserved catalytic acidic residues in inverting enzymes. However, as reviewed by Davies and Henrisat (2002), the mechanism of retaining GTs remains obscure. A mechanism analogous to that of retaining glycoside hydrolases would involve a conserved basic residue: Alternative mechanisms might not, although stabilization of the positively charged intermediates might be effectively carried out by acidic residues. It has also often proved difficult, particularly within superfamily GT-B, to locate these catalytic residues, even with the benefit of structural information. For example, based on the structure of the inverting enzyme MurG complexed to UDP-GlcNAc, several acidic residues were mutated, but the loss of none of these abolished catalytic activity (Hu et al., 2003). Thus no aspartate or glutamate has so far been definitively identified as catalytic in the GT-B fold superfamily.

Using the fold recognition alignments totally (100%) and strongly (90%) conserved residues in GT families were identified and their positions compared within superfamilies (Figure 3A, 3B). In this way, tendencies in conserved acidic residue positioning were sought that might help locate the catalytic acidic residues. However, three possible complications had to be borne in mind—inaccuracies in the fold recognition alignments, existence of divergent nonexpressed genome sequences, and presence of noncatalytic proteins within GT families (Unligil and Rini, 2000). It was also necessary to identify conserved acidic residues with known noncatalytic roles.

Within the GT-A family there is a strong candidate for catalytic residue—Asp191 (SpsA numbering; Tarbouriech et al., 2001). A similar residue in the same position was previously described in structures of families GT-7 (Gastinel et al., 1999) and GT-13 (Unligil et al., 2000). In all the families studied for which a relationship with family 2 was observed, a suitable located Asp or Glu (allowing for local errors in alignment) is present (Figure 3A). The presence of a catalytic Glu in this same region has a precedent in the form of family 43 (Pedersen et al., 2000). Several site-directed mutagenesis studies support a key role of these acidic residues in the enzymatic mechanism. When these residues were mutated in glycosyltransferases from Sinorhizobium meliloti (Garinot-Schneider et al. 2000) and from Salmonella enterica (Keenleyside et al., 2001), the enzymatic activity was severely decreased. Most interestingly, similarly placed highly conserved acidic residues are seen in the retaining families and in family 60 whose nature—retaining or inverting—is unknown. This may be seen as support for...
the double displacement mechanism of retention of configuration of mechanism, analogous to that observed in glycoside hydrolases, which would necessarily involve the participation of a basic residue, most likely Asp or Glu (Davies and Henrissat, 2002). However the positively charged intermediates and transition states that would be involved in alternative mechanisms (Davies and Henrissat, 2002) would likely require charge compensation as a means toward their stabilization. Asp and Glu residues could also fulfill that role.

Another conserved Asp (numbered 39 in SpsA), located at the end of the strand β-2 (Figure 1 and 3A), is involved in nucleotide binding, interacting with the uracyl moiety by making a hydrogen bond to N3 of uracil base (Charnock and Davies, 1999). Figure 3A demonstrates that only families 12 and 16 contain fully conserved aspartic acids in the vicinity. The remaining inverting families, with the exception of family 54, have a 90% conserved Asp and/or Glu. In family 54, 75% of sequences contain a conserved Glu residue in this same position. The retaining families 27, 45, and 55 all have conserved acidic residues for binding UDP. In contrast, families 60, 62, and 64 do not show any conservation of acidic residues at this position (Figure 3A), although some of these similarly bind UDP nucleotide compounds. Presumably other specificity-conferring mechanisms function in these families.

GT-A also contains a motif known as the DxD motif (Wiggins and Munro, 1998; Shibayama et al., 1998; Tarbouriech et al., 2001) (Figure 3), although the arrangement of the two Asp residues varies. In SpsA, this motif adopts the sequence xD98D99. In four representative structures available for the folding superfamily, the first aspartate residue of the motif binds to hydroxyl groups on the ribose moiety, whereas the second aspartic acid binds the divalent metal ion, which could be Mn²⁺ (Tarbouriech et al., 2001). The Mn²⁺ ion is clearly positioned to counter the negative charge that develops on the β-phosphate on cleavage of the donor sugar–phosphate linkage (Cowan, 1998). As shown in Figure 3A, with the exception of families 27 and 62, all other families have an at least 90% conserved acidic residue in the region. In family 27 the figure drops to 88%. Again, the families newly assigned to GT-A provide a surprise; no conserved acidic residue is seen in this region for family 62.

Our degree of understanding of mechanism in GT-B is less advanced. One particular region of sequence conservation between families of the GT-B superfamily (Ha et al., 2000) lies around 250–290 (MurG numbering). Three G-loops (glycine-rich loops located at turns between the carboxyl ends of β-strands and the N-termini of the following α-helices in Rossman fold domains; Baker et al., 1992) have also been the focus of attention (Ha et al., 2000). Later structural determination of the UDP-GlcNAc:MurG complex (Hu et al., 2003) revealed that two G-loops and the 250–290 conserved region are responsible for binding to the UDP-GlcNAc substrate. All inverting families had the acidic residues conserved in this region, with the exception of families 9 and 41. In family 41, 78% of sequences contained a conserved glutamine. Among the residues forming hydrogen bonds to the substrate is Glu269 (Figure 1), for which a role in distinguishing between UDP and TDP has been proposed (Hu et al., 2003). This residue is 90% conserved in family 28 (containing MurG itself) and 100% conserved in family 30 (Figure 3B), whereas an Asp residue may functionally substitute in family 19 and 33. A catalytic role for this residue may be effectively ruled out because its
replacement with Ala in MurG led to only modest loss of activity (Hu et al., 2003). Families 5, 9, and 41 do not have any conservation of acidic residues in this region and must use other mechanisms to confer substrate specificity.

From these analyses it is clear that further work is required to help locate catalytic acidic residues in GT-B and in retaining families of GT-A. The lack of any perceptible trends in positioning of conserved acidic residues in GT-B (Figure 3B), even considering only inverting enzymes, suggests that several catalytic site architectures may well be present in the superfamily.

Materials and methods

Members of GT families 2, 12, 16, 21, 25, 27, 28, 45, 54, 55, 56, 60, 62, and 64 were located in the CAZY database and retrieved using Entrez (www.ncbi.nlm.nih.gov/Entrez). Groups of sequences were aligned with ClustalW (Higgins et al., 1994). Manipulation and limited hand-editing of alignments were performed with Jalview (www.ebi.ac.uk/~michele/jalview). Residues with 90% or 100% of conservation was located using the MView program (Brown et al., 1998). Fold recognition experiments made use of the Structure Prediction META server (Bujnicki et al., 2001a). Particular attention was paid to the results of the two consensus fold recognition analyses, Peons2 (Lundstrom et al., 2001) and Shotgun on 3 consensus prediction (Fischer, 2003), which produces a score based on the results of three independent fold recognition methods, FFAS (Rychlewski et al., 2000), 3D-PSSM (Kelley et al., 2000, and Inbgu (Fischer, 2000).

Iterated sequence database searches were carried out using PSI-BLAST (Altschul et al., 1997) at the NCBI (www.ncbi.nlm.nih.gov/BLAST), using either 0.01 or 0.001 as the E-value cut-off, below which a sequence is included in the next iteration. Appearance of a member of a given GT family in the list of sequences resulting from a search using a different GT family was taken to indicate significant sequence similarity and hence to support a common evolutionary origin for the two families. As input for the fold recognition and iterated database searches we used representative sequences of all families as listed in Table I. When searches with a certain GT family member produced members of another GT family among the significant results, a possible evolutionary origin for the two families was suggested.

Abbreviations

CAZY, carbohydrate active enzymes; GT, glycosyltransferases; GT-A, glycosyltransferase from fold superfamily A; GT-B, glycosyltransferase from fold superfamily B; MurG, GT-B from E. coli; SpsA, GT-A from B. subtilis.

References


