BIOCHEMICAL

JOURNAL

LETTERS

The DAG family of glycosyl hydrolases combines two previously identified protein families

Glycosyl hydrolases are important metabolic enzymes which catalyse the hydrolysis of glycosidic linkages in carbohydrates. They are widely distributed among plants, animals and micro-organisms, and a large number of genes encoding these enzymes have been cloned and sequenced. Henrissat and co-workers have previously used comparative amino acid sequence analysis to classify a range of distinct families of glycosyl hydrolases [1,2]. Here we focus on two of the families defined by Henrissat and co-workers, families 27 and 36. We demonstrate that they comprise two subclusters of a single superfamily, show that an isomaltodextranase is also a member of this superfamily, and present a phylogenetic analysis of this new amalgamated superfamily, which we have designated the DAG [(isomalto)dextranase/ α -galactosidase] family.

Family 27 and family 36 have previously been shown to consist of α -galactosidases from eukaryotes and prokaryotes, respectively [1,2]. Comparison of representative members of these two families with the protein sequence databases, using the program BLAST [3], identified a number of new potential homologues, including several α -galactosidases, two α -N-acetyl-galactosaminidases [4], and an isomaltodextranase from the bacterium *Arthrobacter globiformis* [5] (see Table 1). Homology between these enzymes was established by determining binary comparison scores, expressed in standard deviations (S.D.), by using the RDF2 program [6] with 500 random shuffles. For example, comparison of the *Aspergillus niger* α -galactosidase (a member of former family 27) with a *Pediococcus pentosaceus* α -galactosidase (a member of former family 36) yielded a score of 10 S.D., whereas comparison of the *A. globiformis* isomaltodextranase (previously believed to be unique [5]) with the

Cyamposis tetragonoloba α -galactosidase (a member of former family 27) gave an optimized score of 13 S.D. These and other scores obtained are of sufficient magnitude to establish evolutionary relatedness [7] between all of the fully sequenced members of this family (Table 1).

Phylogenetic analysis revealed that this family consists of two primary clusters, with isomaltodextranase falling on a separate branch (Figure 1). The first cluster includes eukaryotic α -

Figure 1 Phylogenetic tree for the proteins comprising the DAG family

The TREE program [8] was used for construction of the phylogenetic tree and calculation of branch lengths. Relative evolutionary distances are provided adjacent to each branch. The abbreviations used are presented in Table 1.

Table 1 Current members of the DAG family

Abbreviations: pir, Protein Identification Resource database; sp, SwissProt database; gp, GenBank database.

Abbreviation	Protein	Organism	Length (no. of amino acids)	Accession no.
Cluster 1				
NAG Gga	α -N-Acetylglactosaminidase	<i>Gallus gallus</i>	405	pir S45522
NAG Hsa	α -N-Acetylglactosaminidase	<i>Homo sapiens</i>	412	sp P17050
GAL Hsa	α -Galactosidase	<i>Homo sapiens</i>	429	sp P06280
GAL Cte	α -Galactosidase	<i>Cyamposis tetragonoloba</i>	411	sp P14749
GAL Sce	α -Galactosidase	<i>Saccharomyces cerevisiae</i>	472	sp P04824
GAL Ani	α -Galactosidase	<i>Aspergillus niger</i>	545	sp P28351
Cluster 2				
GAL Eco	α -Galactosidase	<i>Escherichia coli</i>	708	sp P16551
GAL Ppe1	α -Galactosidase	<i>Pediococcus pentosaceus</i> (3)	734	gp L32093
GAL Ppe2	α -Galactosidase	<i>Pediococcus pentosaceus</i> (5)	720	gp L32093
GAL Smu	α -Galactosidase	<i>Streptococcus mutans</i>	720	sp P27756
Branch 3				
IMD Agl	Isomaltodextranase	<i>Arthrobacter globiformis</i> T6	641	gp D30761

NAG Gga (53)	LGYKINIDDCWAAKQDAE...GRLVPDPERFPRG
NAG Hsa (70)	MGYTYLINIDDCWIG.GRDAS...GRLMPDPKRFPHG
GAL Hsa (84)	AGYEYLCIDDCWMAPOQDSE...GRLQADPQRFPHG
GAL Cte (90)	LGYQYINLDDCWAEINRDSE...GNMVPNAAAFPSG
GAL Sce (64)	MGYKYIILDDCWSS.GRDS...GFLVADEQKFPNG
GAL Ani (76)	AGYNRLNIDDCWMAYQRSDN...GSLQWNTTKFPFHG
GAL Eco (321)	LQVERFIIDDGWFKGRNDDRAALQDWYTDQKYPNG
GAL Ppe1 (360)	LQIEMFVLDDGWFVNRNGENGLQDWQVDHRKFPQG
GAL Ppe2 (352)	LQIQLVLDDGWFVNRNGENGLQDWQVDRIKFPNG
GAL Smu (352)	LQIELFVLDDGWFVNRNDDNSLQDWQVFNEEKLGGS
IMD Agl (99)	FQYDI.ACTDGWIEGSSRTTGN.QYITSYNSWQHD
Consensus	LGY DDCW RD G D KFP G

Figure 2 Multiple sequence alignment of the region of highest sequence conservation within the members of the DAG family

The numbers in parentheses refer to the first residue exhibited in each protein. Residues conserved in the majority of proteins are displayed in the consensus sequence, and fully conserved residues are highlighted in **bold**. A dot indicates a gap in the aligned sequences, generated to optimize identities.

galactosidases and α -N-acetylgalactosaminidases, and the second contains prokaryotic α -galactosidases. Multiple sequence alignment of the DAG family revealed one region that showed greatest conservation when all members of the family were considered. This region is shown in Figure 2: four residues in the short alignment portrayed are fully conserved (two glycines, a tryptophan and an aspartic acid). Moreover, approximately half of the residues presented are conserved in the majority of the proteins. It seems likely that this region is of functional significance, perhaps as part of the active site of these enzymes. Supporting this notion, residues adjacent to this region have been demonstrated to be functionally important in the α -galactosidases from *C. tetragonoloba* [9] and *Homo sapiens* [10].

In addition to the proteins listed in Table 1, searches of the translated nucleotide sequence databases identified five partially sequenced unidentified open reading frames (URFs) encoding homologues of established DAG family members. The products of these URFs, from *Oryza sativa* (D24643), *Arabidopsis thaliana* (T41617), *Caenorhabditis elegans* (M92935), *Homo sapiens* (Z15307) and the fungus *Cochliobolus carbonum* (M55979) (GenBank accession numbers provided), exhibited comparison scores with established DAG family members in excess of 10 S.D. The first three of these sequences were found to cluster in the phylogenetic tree with the other eukaryotic members of the DAG family, but the last two sequences clustered with the bacterial α -galactosidases of the DAG family. It remains to be established whether this clustering is artifactual, due to their incomplete sequences, or if the last two sequences represent true eukaryotic members of the bacterial α -galactosidase cluster.

The results summarized here established the occurrence of a diverse superfamily of glycosyl hydrolases of both eukaryotic and prokaryotic origins. They set the groundwork for the eventual re-evaluation and consolidation of the previously assigned families of these enzymes as an increasing number of these sequences become available.

I.T.P. was the recipient of a C. J. Martin Fellowship from the National Health and Medical Research Council (Australia). This work was supported by Public Health Service grants 5R01AI21702 and 5R01AI14176 from the National Institute of Allergy and Infectious Diseases.

BOHAN H. DAGNALL, IAN T. PAULSEN and MILTON H. SAIER, Jr.*

Department of Biology, University of California at San Diego, La Jolla, CA 92093-0116, U.S.A.

* To whom correspondence should be addressed.

- Henrissat, B. (1991) *Biochem. J.* **280**, 309–316
- Henrissat, B. and Bairoch, A. (1993) *Biochem. J.* **293**, 781–788
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410
- Wang, A. M., Bishop, D. F. and Desnick, R. J. (1990) *J. Biol. Chem.* **265**, 21859–21866
- Iwai, A., Ito, H., Mizuno, T., Mori, H., Matsui, H., Honma, M., Okada, G. and Chiba, S. (1994) *J. Bacteriol.* **176**, 7730–7734
- Pearson, W. R. and Lipman, D. J. (1988). *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444–2448
- Saier, M. H., Jr. (1994) *Microbiol. Rev.* **58**, 71–93
- Feng, D.-F. and Doolittle, R. F. (1990) *Methods Enzymol.* **183**, 375–387
- Zhu, A., Wang, Z. K. and Goldstein, J. (1995) *Biochim. Biophys. Acta* **1247**, 260–264
- Ishii, S., Kase, R., Fujita, S., Sugimoto, M., Tomita, K., Semba, T. and Suzuki, Y. (1994) *Biochim. Biophys. Acta* **1204**, 265–270

Received 3 April 1995

Families, superfamilies and subfamilies of glycosyl hydrolases

A classification of glycosyl hydrolases based on amino acid sequence similarities was proposed by us a few years ago [1]. With the fast-growing number of glycosyl hydrolase sequences becoming available, the idea was to provide a classification that would (i) reflect the structural features of these enzymes better than their sole substrate specificity, (ii) help to reveal the evolutionary relationships between these enzymes, and (iii) provide a convenient tool to derive mechanistic information [1,2]. In the classification based on sequence similarities, the enzymes are grouped in families. A family was defined when at least two sequences displayed significant similarities to each other and no significant similarity to other families. Because the folding of proteins is better conserved than their sequences, it was predicted that some of the families could have similar folds [2].

The relatedness of some of the families was indeed confirmed (i) when new sequences became available and which proved to be distantly related to more than one family, (ii) when the sensitivity of sequence comparison methods was increased, and (iii) when structural determinations demonstrated the resemblance of the three-dimensional folds between members of different families. Thus, with the increase in sequence and structure data, and with the improvement of sequence comparison methods, the following groupings have been recently proposed:

Families grouped	References
11–12	[3]
33–34	[4]
7–16	[5]
19–22–23–24	[6]
1–2–5–10–17–30–35–39–42	[7,8]

In the preceding Letter, Dagnall et al. [9] have found that recently available sequences show that families 27 and 36 are another example of distantly related families of glycosyl hydrolases.

We have assessed the significance of pairwise comparisons of members of families 27 and 36, with the BESTFIT program from the University of Wisconsin Genetics Computer Group. The alignment scores, expressed in S.D. above the average of 50 random runs, are reported in Table 1. In our hands, the scores obtained within a family (> 20 S.D.) were consistently larger than those calculated between the two families ($\leq$ 5 S.D. (Table