Protein Engineering, Design & Selection vol. 19 no. 12 pp. 555–562, 2006
Published online November 2, 2006 doi:10.1093/protein/gzl044

Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of α-amylase-related proteins

Mark R.Stam, Etienne G.J.Danchin, Corinne Rancurel, Pedro M.Coutinho and Bernard Henrissat[1]

Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités Aix-Marseille I & II, Case 932, 163 Avenue de Luminy, 13288 Marseille cedex 9, France

[1]To whom correspondence should be addressed.
E-mail: Bernard.Henrissat@afmb.univ-mrs.fr

Family GH13, also known as the α-amylase family, is the largest sequence-based family of glycoside hydrolases and groups together a number of different enzyme activities and substrate specificities acting on α-glycosidic bonds. This polyspecificity results in the fact that the simple membership of this family cannot be used for the prediction of gene function based on sequence alone. In order to establish robust groups that show an improved correlation between sequence and enzymatic specificity, we have performed a large-scale analysis of 1691 family GH13 sequences by combining clustering, similarity search and phylogenetic methods. About 80% of the sequences could be reliably classified into 35 subfamilies. Most subfamilies appear monofunctional (i.e. contain enzymes with the same substrate and the same product). The close examination of the other, apparently polyspecific, subfamilies revealed that they actually group together enzymes with strongly related (or even sometimes virtually identical) activities. Overall our subfamily assignment allows to set the limits for genomic function prediction on this large family of biologically and industrially important enzymes.

Keywords: α-amylase/functional prediction/glycoside hydrolase family GH13/phylogenetic analysis/subfamily classification

Introduction

Starch is the major carbohydrate storage product of terrestrial plants and makes up an important part of the food consumed worldwide. As a direct consequence, the agricultural production of starch-rich plants is massive and exceeded 2.3 billion tons in 2002 just for maize, wheat, potatoes, cassava, rice, barley, oats and millet (FAOSTAT data, 2005; http://faostat.fao.org last accessed December 2005). Besides its direct use as food, starch is also used as a raw material in many industrial applications such as high-fructose corn syrups, glues, sizing agent for the paper industry, ethanol production etc. (van der Maarel *et al*., 2002). Starch is made of amylose, which is a linear polymer of glucose residues linked by α-1,4-glycosidic bonds, and of amylopectin, which is an α-1,4-linked D-glucan with varying proportions of α-1,6-linked branches. Because of its widespread occurrence as a storage product, many enzymes for starch hydrolysis (glycosidases) or modification (transglycosidases) are spread throughout the whole biodiversity. The same is true for enzymes acting on glycogen, the animal and bacterial equivalent of plant starch. Interestingly starch-degrading enzymes are found in just a very few of the numerous families of glycosidases and tranglycosidases (termed GH for glycoside hydrolases). For a review of the classification of glycosidases in families [see Henrissat, 1991; Henrissat and Bairoch, 1993; Bourne and Henrissat, 2001 and the Carbohydrate-Active enZyme (CAZy) database at http://www.cazy.org/CAZY]. In this classification, the majority of the enzymes acting on starch, glycogen, and related oligo- and polysaccharides, are found within family GH13, which represents the largest family of glycoside hydrolases (data from CAZy, May 2006). This family belongs to clan GH-H which contains also families GH70 and GH77. A clan is a hierarchical level higher than the family in the CAZy classification, where families from the same clan are believed to share a common ancestor and catalytic machinery (Davies and Henrissat, 1995; Henrissat and Bairoch, 1996; Henrissat and Davies, 1997; Stam *et al*., 2005). The GH13 family, also known as the α-amylase family, has been identified very early (Nakajima *et al*., 1986; MacGregor, 1988; Svensson, 1988) and groups together enzymes sharing sometimes only very limited sequence similarity. As a consequence, the α-amylase family has been the subject of numerous analyses in order to derive relationships between the sequence and the properties of the enzymes (for example Jespersen *et al*., 1993; Janecek *et al*., 1997; Kuriki and Imanaka, 1999; MacGregor *et al*., 2001). Fuelled by the importance of α-amylases and related enzymes, many crystallographic studies have been performed on GH13 family enzymes, and 50 different members had a known 3-D structure in May 2006 (see for instance Buisson *et al*., 1987; Matsuura *et al*., 1980; Boel *et al*., 1990; Watanabe *et al*., 1991; Burk *et al*., 1993; Kadziola *et al*., 1994; Machius *et al*., 1995). Structurally, the GH13 enzymes are characterized by a conserved structural core composed of three domains often designated as domains A, B and C (Ramasubbu *et al*., 1996): domain A folds as a $(\beta/\alpha)_8$-barrel (Brayer *et al*., 1995; Brzozowski and Davies, 1997; Feese *et al*., 2000; Kanai *et al*., 2001; Abad *et al*., 2002), and domain B is a loop of variable length inserted between strand β3 and helix α3 of the $(\beta/\alpha)_8$-barrel (Janecek, 1997). The active site is found in a cleft between domains A and B where a triad of catalytic residues performs catalysis (Brzozowski and Davies, 1997). Domain C is a C-terminal extension characterized by a Greek key structure (Ramasubbu *et al*., 1996; Janecek, 1997). In addition to this conserved core, some members of family GH13 bear a variable number of supplemental N- or C-terminal extensions such as starch-binding modules (families CBM26, CBM41, CBM34, CBM20 in CAZy) and other modules of still unknown function (Jespersen *et al*., 1991; Janecek, 1997). The conservation of a similar 3-D structure for the catalytic domain of family GH13 is logically accompanied by a conservation of the catalytic residues (Jespersen *et al*., 1991, 1991). From this conserved

ancestral scaffold, a large variety of enzymes with varying substrate and product specificity has evolved resulting in the present occurrence in family GH13 of enzymes, with at least 26 different Enzyme Classification (EC) numbers from different enzyme classes: glycoside hydrolases (EC 3.2.1.X, the most abundant), enzymes transferring carbohydrates (EC: 2.4.1.X) and even isomerases (EC 5.4.99.15 and EC 5.4.99.16). These apparently different enzyme categories, however, use the same double displacement catalytic mechanism which proceeds through the build-up and subsequent breakdown of a glycosyl-enzyme intermediate (Davies and Wilson, 1999; Uitdehaag *et al*., 1999) and differ only by the nature of the final acceptor (water for the hydrolases and hydroxyl groups of the substrate for the 'transferases' which are in fact transglycosidases). The EC numbers that describe each enzyme activity are in general very useful, especially to avoid ambiguities and the proliferation of trivial names. However, at least in the case of glycoside hydrolases, and in particular in the case of family GH13, these numbers rarely reflect the common structural features of the enzymes and they are not appropriate for enzymes showing broad specificity (i.e. that act on several substrates). Other problems are that some EC numbers such as EC 3.2.1.98 (maltohexaose-producing α-amylase) are only particular cases of broader enzyme categories such as α-amylase (EC 3.2.1.1) and the distinction depends on the biochemical tests employed (or not) during characterization. Also, some different EC numbers such as EC 3.2.1.10 (oligo-1,6-glucosidase) and EC 3.2.1.70 (glucan 1,6-α-glucosidase) describe basically the same activity. Furthermore, the practical limitations in characterizing the many possible enzyme activities found among the members of this large family lead to biochemical characterizations with limited sets of substrates resulting in biased activity descriptions and annotations (Green and Karp, 2005). The families of glycosidases based on amino acid sequence similarity (Henrissat, 1991) relieved partly these limitations by providing a unified classification system that correlated with the structure and the molecular mechanism of the enzymes (Henrissat and Davies, 1997).

Our continuous updates of CAZy show that family GH13 grew exponentially from 40 entries in 1991 to 2700 in May 2006, e.g. doubled in size approximately every 3 years. A noteworthy fact about the current deluge of sequences that are released by genome sequencing centers and consortia is that virtually all of the novel members of family GH13 are just uncharacterized ORFs with varying degrees of annotations mostly based on unsupervised automatic procedures such as best BLAST hit scores (Rost and Valencia, 1996), which contribute to the creation and subsequent propagation of mis-annotation in public databases (Devos and Valencia, 2001). The increased use of hidden Markov model (HMM)-based annotation methods (Bateman and Haft, 2002; Brown *et al*., 2005) alleviates some of the problems due to the exclusive use of BLAST but presently relies on HMM models of varying quality and annotation. These models suffer from the already mentioned mis-annotations, insufficient biochemical coverage on carbohydrate-active enzymes and pollution with remote similarities (M.R. Stam, E.G.J. Danchin, P.M. Coutinho, B. Henrissat, unpublished data). A reliable tool for substrate specificity prediction is highly desirable and our day-to-day inspection of the current annotations released by genome sequencing centres shows that the situation is particularly critical in the field of glycoside hydrolases and glycosyltransferases, essentially due to the modular structure and the varying substrate specificity within sequence-based families (Coutinho and Henrissat, 1999). This situation is progressively worsened by the decreasing number of novel enzymatic characterization reports in modern scientific literature, perhaps reflecting the fact that the quest for increased impact factors renders journals reluctant to publish such characterizations. Because the situation is unlikely to change, it is becoming important to make the best possible use of the existing and future experimental data. In its field, the CAZy classification effort represents the beginning of a solution, because it usually restrains the number of possible activities for a new sequence assigned to a family, especially when the number of experimentally characterized members is significant. However, the problem remains for large families such as family GH13 that group enzymes of different substrate specificities or even different enzymatic activities overall catalyzed chemical reactions (e.g. hydrolase, transferase, isomerase). To address these problems, and to make progress towards improved annotation of carbohydrate-active enzymes in genomic sequences, we have classified family GH13 into subfamilies following the accepted idea that sequences sharing high similarity should share more biochemical properties than those more distantly related. The difficulties we had to overcome for this work were with the sheer size of the GH13 family, the varying modular structure of its members and the variety of EC numbers present.

Materials and methods

The sequences of catalytic modules of family GH13 members were extracted from the CAZy database. These sequences are the result of a 10-year manual annotation effort where the boundaries of the different catalytic modules were identified using a combination of information resulting from (i) 3-D structure analyses, (ii) deletion studies, (iii) hydrophobic cluster analysis (Gaboriaud *et al*., 1987), (iv) BLAST and PSI-BLAST analysis (Altschul *et al*., 1997) and (v) multiple sequence alignments. A total of 1691 complete catalytic modules sequences were extracted out of a total of 2100 family members available on 26 July 2005, the difference being attributable to fragmentary and other incomplete sequences. The advantage of analyzing exclusively complete and isolated catalytic module sequences is that the background noise due to the remaining component of the coding sequences, which include signal peptides, variable modules such as carbohydrate-binding modules (CBMs) and linker peptides, is eliminated. Moreover, additional modules associated with GH13 such as CBMs can have a different evolutionary history compared to that of the catalytic modules and can potentially produce inconsistencies in phylogenetic reconstructions (Machovic *et al*., 2005).

The extracted sequences corresponding to catalytic modules (GH13), comprising domains A, B and C, were subjected to a multiple sequence alignment using MUSCLE version 3.52 (Edgar, 2004), a program that reliably aligns large sets of protein sequences. The aligned sequences were clustered using the SECATOR algorithm (Wicker *et al*., 2001) as implemented in CLUSPACK (http://www-bio3d-igbmc.u-strasbg.fr/~wicker/programs.html). The underlying algorithm relies on BIONJ (Gascuel, 1997) to build a tree from the multiple sequence alignment and subsequently collapses the branches from subtrees after identification of the nodes joining different

subtrees (Wicker *et al*., 2001). The resulting clusters of aligned sequences were considered as seeds for the creation of subfamilies. Many of the clusters contained too many sequences to make relevant subfamilies. A supplementary step was necessary to remove sequences sharing insufficient similarity with the remainder of the sequences of the cluster. Therefore an automated analysis was followed by a comparison of each sequence from each cluster against the library of amino acid sequences of GH13 catalytic modules using gapped BLASTP and default parameters (Altschul *et al*., 1997). The following criteria were used to identify sufficiently distinct subfamilies:

(i) sequences belonging to the same subfamily share higher sequence similarity than with the remainder of the family (Figure 1) and therefore appear at the top of the BLAST report;

(ii) to ensure sufficient discriminative power, a significant shift in the BLAST *E*-value should be observed between the subfamily members and the remainder of the family (Figure 1);

(iii) when the BLAST results were not consistent with a cluster identified by CLUSPACK, another round of CLUSPACK was performed using only the sequences of the cluster in order to obtain smaller clusters.

(iv) a subfamily should contain at least five sequences from different organisms.

The results given by the clustering method were compared with those from an independent bootstrap-supported phylogenetic analysis. Starting with the multiple sequence alignment determined earlier, a new phylogenetic tree was created by the minimum evolution method (Kidd and Sgaramella-Zonta, 1971) with 100 bootstrap replicates using MEGA version 3.1 (Kumar *et al*., 2004). Because of the large number of sequences in the alignment, the 'complete deletion' option of MEGA was selected. This option removes all columns that contain gaps from the multiple sequence alignment.

We used TreeDyn (http://www.treedyn.org/) to analyze the resulting tree. TreeDyn allows the annotation of any leaf of the tree with external information. Here the different leaves were annotated with pertinent information for the subsequent interpretation: (i) EC number of biochemically characterized enzymes, (ii) the taxonomic group of the organisms present and (iii) the subfamilies identified in the clustering process.

The enzyme activities (EC numbers) reported for members of each subfamily were identified and checked for consistency in the context of related sequences. As routinely performed in CAZy, in order to eliminate self-propagating errors and for most sequences originating from genome sequencing efforts, all predicted activities were discarded. Biochemical activities were extracted and cross-checked using the literature and electronic data from different sources: (i) sequence and structure databases: GenBank (Benson *et al*., 2005), UniProt (Bairoch *et al*., 2005), PDB (Berman *et al*., 2000); (ii) biochemical databases: EMP (Selkov *et al*., 1996), PMD (Kawabata *et al*., 1999), and occasionally BRENDA (Schomburg *et al*., 2002); (iii) literature: PubMed (http://www.pubmed.org). Activities exhibited by only a limited number of elements in a subfamily were systematically checked to ensure reliability and support by accessible online resources.

Results and discussion

The results of our functional classification effort are presented under the form of an annotated phylogenetic tree of the GH13 family (Figure 2). Each subfamily is represented by a coloured subtree. In order to adopt a general naming system that can be extended to other families of glycoside hydrolases, we chose to designate the subfamilies with Arabic numerals following the family number, by order of creation. For instance subfamily 5 of family GH13 is designated GH13_5. Table I presents a summary of the different subfamilies created in family GH13. For each of the 35 subfamilies we report the identified EC numbers, the associated activities and the taxonomic group to which the sequences belong, according to the NCBI taxonomy (Benson *et al*., 2005).

We have cross-checked the subfamilies created by the clustering method with the phylogenetic tree generated by MEGA (Figure 2). All the subfamilies correspond to a subtree

Sequences producing significant alignments:	(bits)	Value
GH13_28 a-amylase (Amy) *Bacillus subtilis* SUH4...	883	0.0
GH13_28 a-amylase (AmyE;AmyA;BSU03040) *Bacillus* ...	879	0.0
GH13_28 a-amylase *Bacillus subtilis* HA401 [(1...	876	0.0
GH13_28 a-amylase *Bacillus subtilis* W168 PY79 ...	876	0.0
GH13_28 a-amylase *Bacillus subtilis* [(1-27)SI...	875	0.0
GH13_28 a-amylase *Bacillus subtilis* 233 [(1-2...	869	0.0
GH13_28 a-amylase *Bacillus subtilis* 2633 / natto...	868	0.0
GH13_28 a-amylase *Bacillus subtilis* (EC 3.2.1.1)...	811	0.0
GH13_28 a-amylase *Lactobacillus amylovorus* (EC ...	619	e-176
GH13_28 a-amylase *Lactobacillus plantarum* A6 (E...	615	e-175
GH13_28 a-amylase *Lactobacillus manihotivorans* ...	595	e-169
GH13_28 a-amylase *Streptococcus bovis* 148 (EC 3...	548	e-155
GH13_28 a-amylase (AmyB) *Bifidobacterium adole*...	358	7e-98
GH13_28 AmyA *Butyrivibrio fibrisolvens* 16/4 [...	312	6e-84
GH13_28 a-amylase *Butyrivibrio fibrisolvens* H1...	309	4e-83
GH13_28 a-amylase (amyA) *Clostridium acetobutyl*...	291	8e-78
GH13_28 CAP0098 *Clostridium acetobutylicum* ATC...	271	2e-71
GH13_15 Tfu_0985 *Thermobifida fusca* YX [(1-33...	136	4e-31
GH13 a-amylase *Thermomonospora curvata* (EC 3.2....	136	6e-31
GH13_32 a-amylase (Aml) *Streptomyces lividans* (...	130	3e-29
GH13_32 SCO7020 or AmlB *Streptomyces coelicolor*...	123	4e-27

Fig. 1. Example of a BLAST report obtained starting from a sequence from subfamily GH13_28 Starting with the sequence of the α-amylase from *Bacillus subtilis* SUH4-2 (Cho *et al*., 2000), sequences of the sub-family GH13_28 (framed in black box) are retrieved first with a slow and progressive increase of the *E*-value. The regularity of the progression is interrupted by a large difference in the *E*-value when members of other subfamilies are retrieved.

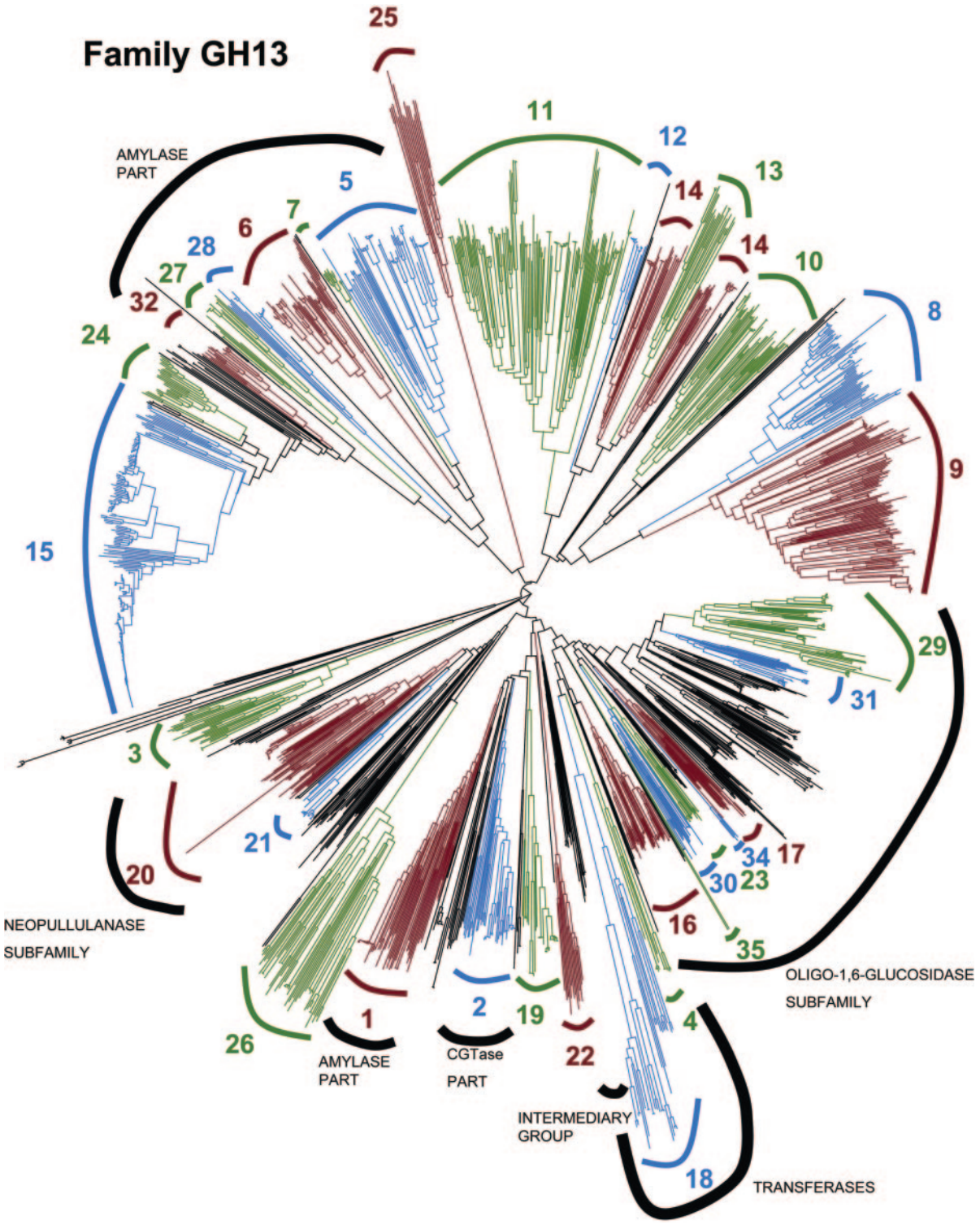

Fig. 2. Phylogenetic tree of family GH13. Sequences classified into subfamilies 1–35 are shown in color. The sequences that were not included into subfamilies appear in black. The external black arcs cover subfamilies previously made by Janecek (Janecek *et al*., 2003) and Oslancava (Oslancova and Janecek, 2002).

of the phylogenetic tree. There is only one exception, namely, subfamily GH13_13, which is found among the branches that compose subfamily GH13_14. This apparent contradiction is explained by the parameters that were selected for computation by MEGA: because of the huge number of sequences to compute, we used the 'complete deletion' option with MEGA, with the consequence that gaps in the alignment were not taken into account. BLAST results (data not shown) and a multiple sequence alignment (Figure 3) confirmed that the corresponding subfamilies are closely related, but distinct, and that the main difference between them are three gaps (Figure 3). The distinctiveness of the two subfamilies was verified by building a new phylogenetic tree with the sequences from subfamilies GH13_13 and GH13_14, using subfamily GH13_12 as an external group, and taking into account sequence gaps.

Table I. Composition of the 35 subfamilies within glycosidase family GH13

Subfamily	EC number[a]	Reported enzyme activities	Taxonomical range[b]
GH13_1	3.2.1.1	α-Amylase	Fungi (48)
GH13_2	3.2.1.1	α-Amylase	Bacteria (38), Archaea (4)
	2.4.1.19	Cyclodextrin glucanotransferase	
	3.2.1.133	Maltogenic α-amylase	
GH13_3	ND	Unknown activity	Bacteria (33), Archaea (1)
GH13_4	2.4.1.4	Amylosucrase	Bacteria (11)
	3.2.1.-	Sucrose hydrolase	
GH13_5	3.2.1.1	α-Amylase	Bacteria (52), Eukaryota (1)
GH13_6	3.2.1.1	α-Amylase	Viridiplantae (44)
GH13_7	3.2.1.1	α-Amylase	Euryarchaeota (9)
GH13_8	2.4.1.8	1,4-α-Glucan branching enzyme	Eukaryota (58), Bacteria (1)
GH13_9	2.4.1.8	1,4-α-Glucan branching enzyme	Bacteria (130), Eukaryota (2)
GH13_10	3.2.1.141	4-α-(1,4-α-Glucano)trehalose -trehalohydrolase	Bacteria (38) Archaea (5)
GH13_11	3.2.1.68	Isoamylase	Bacteria (100), Archaea (6) and Eukaryota (13)
GH13_12	3.2.1.41	Pullulanase	Firmicutes (12)
GH13_13	3.2.1.41	Pullulanase	Bacteria (16), Eukaryota (7)
GH13_14	3.2.1.41	Pullulanase	Bacteria (40)
GH13_15	3.2.1.1	α-Amylase	Metazoa (300), Bacteria (3)
GH13_16	5.4.99.16	Maltose α-glucosyltransferase	Bacteria (38), Archaea (1)
GH13_17	3.2.1.20	α-Glucosidase	Metazoa (18)
GH13_18	2.4.1.7	Sucrose phosphorylase	Bacteria (31)
GH13_19	3.2.1.1	α-Amylase	Bacteria (27)
	3.2.1.98	Maltohexaose-forming α-amylase	
	3.2.1.-	Maltopentaose-forming α-amylase	
GH13_20	3.2.1.54	Cyclomaltodextrinase	Bacteria (56)
	3.2.1.133	Maltogenic α-amylase	
	3.2.1.135	Neopullulanase	
GH13_21	3.2.1.20	α-Glucosidase	Proteobacteria (22) Deinococcus-Thermus (1)
GH13_22	2.4.1.183	α-1,3-Glucan synthase	Fungi (14)
GH13_23	ND	Unknown activity	Proteobacteria (15)
GH13_24	3.2.1.1	α-Amylase	Metazoa (24)
GH13_25	3.2.1.33	Amylo-α-1,6-glucosidase	Eukaryota (15)
GH13_26	5.4.99.15	(1,4)-α-Glucan 1-α-glucosylmutase	Bacteria (43), Archaea (6)
GH13_27	3.2.1.1	α-Amylase	Proteobacteria (19)
GH13_28	3.2.1.1	α-Amylase	Firmicutes (16), Actinobacteria (1)
GH13_29	3.2.1.93	α-Phosphotrehalase	Bacteria (69)
GH13_30	3.2.1.20	α-Glucosidase	Actinobacteria (18)
GH13_31	3.2.1.70	Glucan 1,6-α-glucosidase	Bacteria (23)
	3.2.1.10	Oligo-1,6-glucosidase	
GH13_32	3.2.1.1	α-Amylase	Bacteria (13)
GH13_33	5.4.99.16	Trehalose synthase	Bacteria (6)
GH13_34	NA	Amino acid transporter	Eukaryota (23)
GH13_35	NA	Amino acid transporter	Eukaryota (7)

ND, not determined; NA, not applicable.
[a]Experimentally determined.
[b]The number of sequences for each taxon (data from 26 July 2005) is given in brackets.

Thirty-five subfamilies have been identified from the sample of 1691 complete GH13 catalytic domains analyzed (Table I). A total of 1358 (80%) sequences could be assigned to a subfamily. The remaining 333 (20%) sequences were not included because of insufficient statistical support. Typically the latter sequences belong to: (i) sequences left unclustered by the SECATOR procedure; (ii) insufficiently populated small clusters often lacking biochemically characterized members; and (iii) groups that are too populated for the identification of long branches indicative of distinctiveness. It is likely that some sequences in the first two categories will integrate new subfamilies when more closely related members appear.

The (apparently) monospecific subfamilies

The largest subfamilies of family GH13 are subfamilies GH13_15, GH13_9 and GH13_11 which count 303, 132 and 119 members, respectively. Interestingly, only one activity (identified by a single EC number: EC 3.2.1.1, EC 2.4.1.8 and 3.2.1.68, respectively) is observed in each of these families. This feature is observed in fact for 26 of the 35 subfamilies identified, covering 68% of the analyzed sequences. The division into subfamilies coinciding with single activities suggests that the acquisition of these specificities preceded speciation. For example subfamilies GH13_8 and GH13_9 group enzymes with an α-1,4-glucan branching activity (EC 2.4.1.8) and belong to the same subtree. The division into two subfamilies follows the taxonomy: subfamily GH13_8 groups sequences from Eukaryota while subfamily GH13_9 groups sequences from Bacteria.

Subfamily GH13_21 counts only a single biochemically characterized member, namely, an α-glucosidase (Peist *et al*., 1996). However, this enzyme is also highly active on γ-cyclodextrin, which is more coherent with the close relatedness of this subfamily to subfamily GH13_20 (cyclomaltodextrinases) and its more distant relationship to subfamilies GH13_30 and GH13_17 (α-glucosidases). This example illustrates the fact that even when an experimental characterization is available, not all possible activities have been

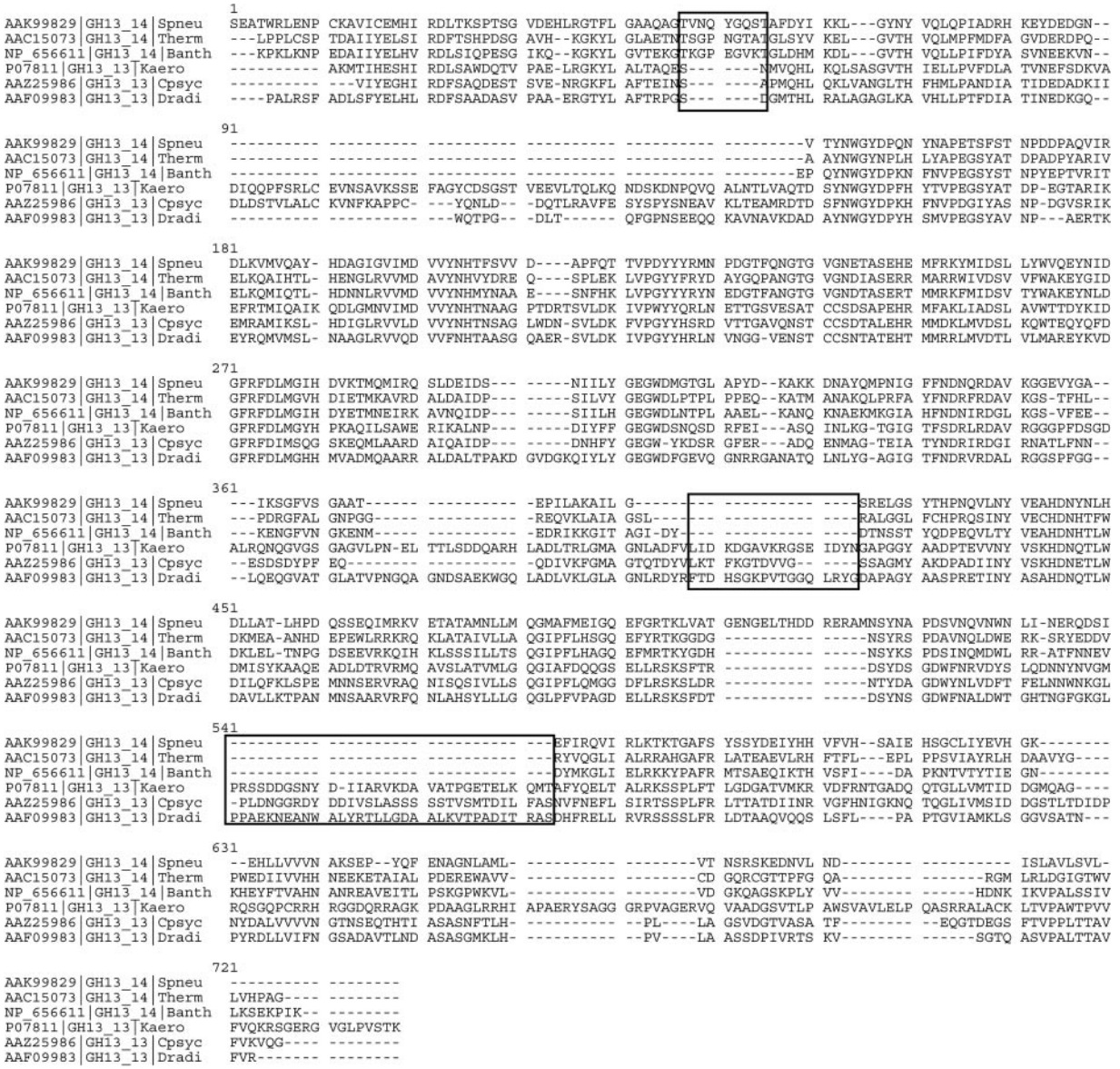

Fig. 3. Multiple sequence alignment of characterized enzymes from subfamilies GH13_13 and GH13_14. The black boxes highlight the gaps present in one subfamily but not in the other.

tested. More characterizations are therefore needed to assign reliably an EC number to subfamily GH13_21.

The (apparently) polyspecific subfamilies

Only five subfamilies (GH13_19, GH13_31, GH13_20, GH13_2 and GH13_4) contain more than one reported activity. However, we have noticed that the activities within each subfamily are closely related. In the case of subfamily GH13_19, there is barely any difference in term of specificity between the α-amylase (EC 3.2.1.1) from *Escherichia coli* K12 (Spiess *et al*., 1997), the maltohexaose-forming α-amylase (EC 3.2.1.98) from *Bacillus halodurans* LBK 34 (Hashim *et al*., 2004) and the maltopentaose-forming α-amylase (no EC number assigned) from alkalophilic Gram positive bacteria DSM 5853 (Candussio *et al*., 1990). In the same way, two apparently 'different' activities are present in subfamily GH13_31, namely, glucan 1,6-α-glucosidase (EC 3.2.1.70) (Whiting *et al*., 1993) and oligo-1,6-glucosidase (EC 3.2.1.10) (Bornke *et al*., 2001). According to the IUBMB Enzyme Nomenclature (Enzyme Nomenclature Committee, 1992), these two activities catalyze the hydrolysis of α-1,6-D-glucosidic linkages. Therefore, the difference is perhaps more semantic than biological. Another example of the assignment of different EC numbers for the same activity is found in subfamily GH13_20, which groups cyclomaltodextrinase (EC 3.2.1.54) from *Paenibacillus sp.* A11 (Kaulpiboon and Pongsawasdi, 2004), maltogenic α-amylase (EC 3.2.1.133) from *Bacillus subtilis* SUH4-2 (Cho *et al*., 2000) and neopullanase (EC 3.2.1.135) from *Thermoactinomyces vulgaris* R-47 (Tonozuka *et al*., 1993). It has been demonstrated that these three enzymes act on the same substrate and generate the same product, therefore they should be classified under the same name and the same EC number (Cheong *et al*., 2002; Lee *et al*., 2002).

Subfamily GH13_2 clusters together two maltogenic α-amylases (EC 3.2.1.133) (Dauter *et al*., 1999) and an acarviosyl transferase (EC 2.4.1.-) (Hemker *et al*., 2001) together with cyclodextrin glucanotransferases (EC 2.4.1.19) (Leemhuis *et al*., 2003). The two maltogenic α-amylases

present also an high catalytic activity on cyclodextrin (Dauter *et al*., 1999) and only one amino acid mutation can change the acarviosyl transferase activity into an enzyme with 4-α-glucanotransferase activity (Leemhuis *et al*., 2004). This example shows the limitations of activity prediction based on subfamily analysis for polyspecific enzymes and for engineered variants.

Finally, two activities, sucrose hydrolase (EC 3.2.1.-) and amylosucrase (EC 2.4.1.4), are found in subfamily GH13_4. These two activities are closer to each other than suggested by their EC numbers, since they operate on the same substrate (sucrose), with the same molecular mechanism and the difference is only with different transglycosylation abilities. It is therefore likely that all members of this subfamily utilize sucrose as the substrate.

In conclusion, the close examination of the polyspecific subfamilies reveals that they actually contain enzymes with strongly related (or even sometimes nearly identical) substrate and/or product specificities, showing that here too subfamily assignment has strong predictive power.

Subfamilies with no associated EC number

Subfamilies GH13_3, GH13_23, GH13_34 and GH13_35 do not contain enzymes with an associated EC number. In fact, subfamilies GH13_34 and GH13_35 contain a particular set of GH13 members, which have lost their catalytic machinery and evolved to a novel function (Broer and Wagner, 2002; Janecek *et al*., 1997). Members from subfamily GH13_34 are known as 4F2 heavy chain proteins, which induce amino acid transport in vertebrates (Estevez *et al*., 1998), whereas subfamily GH13_35 groups cysteine, basic and neutral amino acid transporters and related proteins (Mizoguchi *et al*., 2001). A multiple sequence alignment of subfamily GH13_3 shows that while the catalytic base is conserved, the remainder of the catalytic machinery and other typical conserved motifs of this family are not (data not shown), suggesting that members of this subfamily have probably also acquired a novel, unrelated, function.

In contrast, multiple sequence alignment of subfamily GH13_23 members revealed a conserved catalytic apparatus (data not shown), suggesting that the members of this subfamily have a glycoside hydrolase capability.

Comparison with other efforts of classification within family GH13

Several criteria can be envisioned and used for the definition of subfamilies suitable to derive a better correlation between sequences and enzyme specificity than membership to the broad GH13 family. Here we have created subfamilies based on sequence similarity and phylogenetic reconstruction criteria. Overall sequence differences between catalytic modules reflect functional differences. Earlier a classification of amylases for specificity prediction purposes had been proposed, based on the structure of the small domain B (Janecek *et al*., 1997). Although it is conceivable that domain B has co-evolved to some extent with the remainder of the catalytic domain of the enzymes, it does not cover entirely the active site of family GH13 enzymes and is too short to provide a signal-to-noise ratio sufficient for the classification of hundreds of proteins. Other efforts have attempted to define a limited number of subfamilies based on phylogenetic analyses of partial subsets of family GH13 sequences (Oslancova and Janecek, 2002; Janecek *et al*., 2003). We have mapped the groups resulting from these earlier analyses onto the tree presented in Figure 2. Our results are broadly in agreement with these earlier studies but provide a complete analysis of the entire family GH13 resulting in both finer subdivisions (i.e. more subfamilies) and an improved correlation with activity.

Conclusions

The diversity of specificities and activities found in family GH13 shows that this family is old enough to have seen the emergence (and sometime the loss) of many activities. The sequences belonging to subfamilies containing only one EC number represent 68% of the sequences analyzed. This excellent correlation with the subfamilies that we have defined through our phylogenetic analysis suggests that indeed the assignment to a subfamily is a considerable step towards improved functional prediction. However, because not all subfamilies have a biochemical characterized members and because a significant number of sequences are still not included in subfamilies, errors or imprecision are still possible during unsupervised automated genomic annotations.

In addition, the present study points out that there are still some branches of family GH13 that require structural or biochemical characterization and that additional subfamilies will emerge later. Here again, our work points to several limitations: experimental EC number assignments are sometimes ambiguous due to the use of unduly limited sets of substrates. To make things worse, the choice of an EC number appears to occasionally reflect more the opinion of the experimentalist than actual biochemical evidence. Finally, the traditional descriptive EC numbers were not intended nor designed to take into account functional drifts that arise from evolutionary events such as gene loss, convergence or duplication. All these aspects suggest the use of EC numbers in post-genomic approaches (for instance metabolic pathway mapping) with the greatest caution, as they were only designed to provide common names to describe enzyme reactions.

The rigorous approach we developed for the definition of sub-families in family GH13 will be applied in the future to other Carbohydrate-Active enZymes families which in turn will benefit from the improvement of predictability of specificity at a larger scale.

A limitation of the methods we have used is that they are very time-consuming and one cannot repeat this type of an analysis every time the CAZy database is updated. We have therefore developed a series of HMMs based on each of the 35 subfamilies described here, and these allow the rapid assignment of new sequences to the subfamilies defined here. Illustratively, the set of complete sequences of GH13 modules selected at the beginning of our work (1691 sequences as July 2005) has grown to over 2456 in August 2006. Out of the 765 novel full-length sequences added to family GH13 between July 2005 and August 2006, more than 90% could be added to the 35 subfamilies described here. This subfamily assignment will therefore become available and will be updated as an integral part of the data presented for family GH13 in the CAZy database (http://afmb.cnrs-mrs.fr/CAZY/fam/GH13.html).

Acknowledgements

This work was funded in part by the European Commission (STREP FungWall grant, contract: LSHB-CT-2004-511952).

References

Abad,M.C., Binderup,K., Rios-Steiner,J., Arni,R.K., Preiss,J. and Geiger,J.H. (2002) *J. Biol. Chem.*, **277**, 42164–42170.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Bairoch,A. *et al.* (2005) *Nucleic Acids Res.*, **33**, D154–D159.

Bateman,A. and Haft,D.H. (2002) *Brief Bioinform.*, **3**, 236–245.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) *Nucleic Acids Res.*, **33**, D34–D38.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.

Boel,E., Brady,L., Brzozowski,A.M., Derewenda,Z., Dodson,G.G., Jensen,V.J., Petersen,S.B., Swift,H., Thim,L. and Woldike,H.F. (1990) *Biochemistry.*, **29**, 6244–6249.

Bornke,F., Hajirezaei,M. and Sonnewald,U. (2001) *J. Bacteriol.*, **183**, 2425–2430.

Bourne,Y. and Henrissat,B. (2001) *Curr. Opin. Struct. Biol.*, **11**, 593–600.

Brayer,G.D., Luo,Y. and Withers,S.G. (1995) *Protein Sci.*, **4**, 1730–1742.

Broer,S. and Wagner,C.A. (2002) *Cell. Biochem. Biophys.*, **36**, 155–168.

Brown,D., Krishnamurthy,N., Dale,J.M., Christopher,W. and Sjolander,K. (2005) *Pac. Symp. Biocomput.*, pp. 322–333.

Brzozowski,A.M. and Davies,G.J. (1997) *Biochemistry.*, **36**, 10837–10845.

Buisson,G., Duee,E., Haser,R. and Payan,F. (1987) *EMBO J.*, **6**, 3909–3916.

Burk,D., Wang,Y., Dombroski,D., Berghuis,A.M., Evans,S.V., Luo,Y., Withers,S.G. and Brayer,G.D. (1993) *J. Mol. Biol.*, **230**, 1084–1085.

Candussio,A., Schmid,G. and Bock,A. (1990) *Eur. J. Biochem.*, **191**, 177–185.

Cheong,K.A., Kim,T.J., Yoon,J.W., Park,C.S., Lee,T.S., Kim,Y.B., Park,K.H. and Kim,J.W. (2002) *Biotechnol. Appl. Biochem.*, **35**, 27–34.

Cho,H.Y., Kim,Y.W., Kim,T.J., Lee,H.S., Kim,D.Y., Kim,J.W., Lee,Y.W., Leed,S. and Park,K.H. (2000) *Biochim. Biophys. Acta*, **1478**, 333–340.

Coutinho,P.M. and Henrissat,B. (1999) *J. Mol. Microbiol. Biotechnol.*, **1**, 307–308.

Dauter,Z., Dauter,M., Brzozowski,A.M., Christensen,S., Borchert,T.V., Beier,L., Wilson,K.S. and Davies,G.J. (1999) *Biochemistry*, **38**, 8385–8392.

Davies,G. and Henrissat,B. (1995) *Structure*, **3**, 853–859.

Davies,G.J. and Wilson,K.S. (1999) *Nat. Struct. Biol.*, **6**, 406–408.

Devos,D. and Valencia,A. (2001) *Trends Genet.*, **17**, 429–431.

Enzyme Nomenclature Committee (1992), *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, San Diego, CA, USA.

Edgar,R.C. (2004) *Nucleic Acids Res.*, **32**, 1792–1797.

Estevez,R., Camps,M., Rojas,A.M., Testar,X., Deves,R., Hediger,M.A., Zorzano,A. and Palacin,M. (1998) *FASEB J.*, **12**, 1319–1329.

Feese,M.D., Kato,Y., Tamada,T., Kato,M., Komeda,T., Miura,Y., Hirose,M., Hondo,K., Kobayashi,K. and Kuroki,R. (2000) *J. Mol. Biol.*, **301**, 451–464.

Gaboriaud,C., Bissery,V., Benchetrit,T. and Mornon,J.P. (1987) *FEBS Lett.*, **224**, 149–155.

Gascuel,O. (1997) *Mol. Biol. Evol.*, **14**, 685–695.

Green,M.L. and Karp,P.D. (2005) *Nucleic Acids Res.*, **33**, 4035–4039.

Hashim,S.O., Delgado,O., Hatti-Kaul,R., Mulaa,F.J. and Mattiasson,B. (2004) *Biotechnol. Lett.*, **26**, 823–828.

Hemker,M., Stratmann,A., Goeke,K., Schroder,W., Lenz,J., Piepersberg,W. and Pape,H. (2001) *J. Bacteriol.*, **183**, 4484–4492.

Henrissat,B. (1991) *Biochem. J.*, **280**, 309–316.

Henrissat,B. and Bairoch,A. (1993) *Biochem. J.*, **293**, 781–788.

Henrissat,B. and Bairoch,A. (1996) *Biochem. J.*, **316**, 695–696.

Henrissat,B. and Davies,G. (1997) *Curr. Opin. Struct. Biol.*, **7**, 637–644.

Janecek,S. (1997) *Prog. Biophys. Mol. Biol.*, **67**, 67–97.

Janecek,S., Svensson,B. and Henrissat,B. (1997) *J. Mol. Evol.*, **45**, 322–331.

Janecek,S., Svensson,B. and MacGregor,E.A. (2003) *Eur. J. Biochem.*, **270**, 635–645.

Jespersen,H.M., MacGregor,E.A., Sierks,M.R. and Svensson,B. (1991) *Biochem. J.*, **280**, 51–55.

Jespersen,H.M., MacGregor,E.A., Henrissat,B., Sierks,M.R. and Svensson,B. (1993) *J. Protein Chem.*, **12**, 791–805.

Kadziola,A., Abe,J., Svensson,B. and Haser,R. (1994) *J. Mol. Biol.*, **239**, 104–121.

Kanai,R., Haga,K., Yamane,K. and Harata,K. (2001) *J. Biochem. (Tokyo)*, **129**, 593–598.

Kaulpiboon,J. and Pongsawasdi,P. (2004) *J. Biochem. Mol. Biol.*, **37**, 408–415.

Kawabata,T., Ota,M. and Nishikawa,K. (1999) *Nucleic Acids Res.*, **27**, 355–357.

Kidd,K.K. and Sgaramella-Zonta,L.A. (1971) *Am. J. Hum. Genet.*, **23**, 235–252.

Kumar,S., Tamura,K. and Nei,M. (2004) *Brief Bioinform.*, **5**, 150–163.

Kuriki,T. and Imanaka,T. (1999) *J. Biosci. Bioeng.*, **87**, 557–565.

Lee,H.S., Kim,M.S., Cho,H.S., Kim,J.I., Kim,T.J., Choi,J.H., Park,C., Oh,B.H. and Park,K.H. (2002) *J. Biol. Chem.*, **277**, 21891–21897.

Leemhuis,H., Dijkstra,B.W. and Dijkhuizen,L. (2003) *Eur. J. Biochem.*, **270**, 155–162.

Leemhuis,H., Wehmeier,U.F. and Dijkhuizen,L. (2004) *Biochemistry*, **43**, 13204–13213.

MacGregor,E.A. (1988) *J. Protein Chem.*, **7**, 399–415.

MacGregor,E.A., Janecek,S. and Svensson,B. (2001) *Biochim. Biophys. Acta*, **1546**, 1–20.

Machius,M., Wiegand,G. and Huber,R. (1995) *J. Mol. Biol.*, **246**, 545–559.

Machovic,M., Svensson,B., MacGregor,E.A. and Janecek,S. (2005) *FEBS J.*, **272**, 5497–5513.

Matsuura,Y., Kusunoki,M., Harada,W., Tanaka,N., Iga,Y., Yasuoka,N., Toda,H., Narita,K. and Kakudo,M. (1980) *J. Biochem. (Tokyo)*, **87**, 1555–1558.

Mizoguchi,K. *et al.* (2001) *Kidney Int.*, **59**, 1821–1833.

Nakajima,R., Imanaka,T. and Aiba,S. (1986) *Appl. Microbiol. Biotechnol*, **23**, 355–360.

Oslancova,A. and Janecek,A. (2002) *Cell Mol. Life. Sci.*, **59**, 1945–1959.

Peist,R., Schneider-Fresenius,C. and Boos,W. (1996) *J. Biol. Chem.*, **271**, 10681–10689.

Ramasubbu,N., Paloth,V., Luo,Y., Brayer,G.D. and Levine,M.J. (1996) *Acta Crystallogr. D Biol. Crystallogr.*, **52**, 435–446.

Rost,B. and Valencia,A. (1996) *Curr. Opin. Biotechnol.*, **7**, 457–461.

Schomburg,I., Chang,A. and Schomburg,D. (2002) *Nucleic Acids Res.*, **30**, 47–49.

Selkov,E. *et al.* (1996) *Nucleic Acids Res.*, **24**, 26–28.

Spiess,C., Happersberger,H.P., Glocker,M.O., Spiess,E., Rippe,K. and Ehrmann,M. (1997) *J. Biol. Chem.*, **272**, 22125–22133.

Stam,M.R., Blanc,E., Coutinho,P.M. and Henrissat,B. (2005) *Carbohydr. Res.*, **340**, 2728–2734.

Svensson,B. (1988) *FEBS Lett.*, **230**, 72–76.

Tonozuka,T., Ohtsuka,M., Mogi,S., Sakai,H., Ohta,T. and Sakano,Y. (1993) *Biosci. Biotechnol. Biochem.*, **57**, 395–401.

Uitdehaag,J.C., Mosi,R., Kalk,K.H., van der Veen,B.A., Dijkhuizen,L., Withers,S.G. and Dijkstra,B.W. (1999) *Nat. Struct. Biol.*, **6**, 432–436.

van der Maarel,M.J., van der Veen,B., Uitdehaag,J.C., Leemhuis,H. and Dijkhuizen,L. (2002) *J. Biotechnol.*, **94**, 137–155.

Watanabe,K., Kitamura,K., Hata,Y., Katsube,Y. and Suzuki,Y. (1991) *FEBS Lett.*, **290**, 221–223.

Whiting,G.C., Sutcliffe,I.C. and Russell,R.R. (1993) *J. Gen. Microbiol.*, **139**, 2019–2026.

Wicker,N., Perrin,G.R., Thierry,J.C. and Poch,O. (2001) *Mol. Biol. Evol.*, **18**, 1435–1441.

Received June 27, 2006; revised August 31, 2006; accepted September 18, 2006

Edited by Dick Janssen