Article

New families in the classification of glycosyl hydrolases based on amino acid sequence similarities

HENRISSAT, Bernard, BAIROCH, Amos Marc

Abstract

301 glycosyl hydrolases and related enzymes corresponding to 39 EC entries of the I.U.B. classification system have been classified into 35 families on the basis of amino-acid-sequence similarities [Henrissat (1991) Biochem. J. 280, 309-316]. Approximately half of the families were found to be monospecific (containing only one EC number), whereas the other half were found to be polyspecific (containing at least two EC numbers). A > 60% increase in sequence data for glycosyl hydrolases (181 additional enzymes or enzyme domains sequences have since become available) allowed us to update the classification not only by the addition of more members to already identified families, but also by the finding of ten new families. On the basis of a comparison of 482 sequences corresponding to 52 EC entries, 45 families, out of which 22 are polyspecific, can now be defined. This classification has been implemented in the SWISS-PROT protein sequence data bank.

Reference

HENRISSAT, Bernard, BAIROCH, Amos Marc. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical journal*, 1993, vol. 293 (Pt 3), p. 781-8

PMID : 8352747

New families in the classification of glycosyl hydrolases based on amino acid sequence similarities

Bernard HENRISSAT*‡ and Amos BAIROCH†

*Centre de Recherches sur les Macromolécules Végétales, C.N.R.S., BP 53X, F-38041 Grenoble, France, and †Medical Biochemistry Department, Centre Médical Universitaire, CH-1211 Geneva 4, Switzerland

301 glycosyl hydrolases and related enzymes corresponding to 39 EC entries of the I.U.B. classification system have been classified into 35 families on the basis of amino-acid-sequence similarities [Henrissat (1991) Biochem. J. **280**, 309–316]. Approximately half of the families were found to be monospecific (containing only one EC number), whereas the other half were found to be polyspecific (containing at least two EC numbers). A > 60% increase in sequence data for glycosyl hydrolases (181 additional enzymes or enzyme domains sequences have since become available) allowed us to update the classification not only by the addition of more members to already identified families, but also by the finding of ten new families. On the basis of a comparison of 482 sequences corresponding to 52 EC entries, 45 families, out of which 22 are polyspecific, can now be defined. This classification has been implemented in the SWISS-PROT protein sequence data bank.

INTRODUCTION

A new classification system for glycosyl hydrolases has recently been proposed (Henrissat, 1991) to complement the I.U.B. (1984) enzyme nomenclature of this class of enzymes. The proposed classification system is based on amino acid sequence similarities and intends to better reflect the structural features of these enzymes than their sole substrate specificity. The comparison of 301 sequences of glycosyl hydrolases and related enzymes allowed their grouping into 35 families (each containing at least two sequences), out of which 17 were polyspecific (containing at least two EC numbers). Only 10 sequences could not be classified (i.e. had no counterpart) and were expected to form new families when new sequences would become available. This classification was necessarily incomplete, since it was based on a large, but limited, number of sequences. About 180 additional glycosyl hydrolase sequences have since become available and provide the opportunity to test and update the classification by the addition of new members to a number of existing families, as well as by the finding of ten new families.

METHODS

Sequence comparisons were conducted as described previously (Henrissat, 1991). A total of 179 new (or newly available) sequences of glycosyl hydrolases (or related enzymes) have been extracted from the SWISS-PROT protein sequence database (Release 24, December 1992) or entered manually from the literature. Some of these sequences were edited to separate their multiple constitutive catalytic domains, generating a total of 181 proteins (or protein domains), which have been compared to the previous classification. If significant similarities were found with members of one of the 35 described families, the sequence under study was assigned to that family and added to the total set. If no sequence similarity was detected, comparison was extended to each unclassified sequence. Significant similarities between new and previously unclassified sequences (or between at least two new sequences showing no detectable similarities with established families) led to the definition of new families.

RESULTS

Out of the 181 newly analysed sequences (Table 1), 159 (87%) could be classified in the families defined in the previous paper (Henrissat, 1991). Five sequences were found to show similarity to previously unclassified sequences and thus defined four new families (36, 37 38 and 45; Table 1); 12 sequences displaying pairwise similarities, but none with any of the identified families, allowed the definition of families 39–44 (Table 1). Only five sequences could not be classified (i.e. did not exhibit significant similarity to any of the families nor with the unclassified sequences) hence are likely to form new families when related sequences appear. In addition, the previously unclassified cellodextrinase of *Ruminococcus flavefaciens* FD1 has been recently shown to be incorrect, re-sequenced and found to belong to family 5 (Wang and Thomson, 1992). Similarly, the β-D-xylosidase of *Bacillus pumilus*, which could not be classified earlier for lack of similarity, has recently been corrected (Xu et al., 1991). The corrected sequence is significantly similar to that of a newly available bifunctional enzyme displaying both β-D-xylosidase and α-L-arabinofuranosidase activity (Utt et al., 1991) and allowed definition of family 43 (Table 1).

As of January 1993, the classification is based on the comparison of 482 sequences and comprises 45 families with at least two members. Almost half of the families (22) are polyspecific

Abbreviations used: AAMY, α-amylase; AGAL, α-galactosidase; AGAR, agarase; AGLU, α-glucosidase; AIDU, α-L-iduronase; AMAN, α-mannosidase; AMG, amyloglucosidase; ARAF, α-L-arabinofuranosidase; BAMY, β-amylase; BGAL, β-galactosidase; BGLU, β-glucosidase; BMAN, β-mannanase; BXYL, β-xylosidase; CBH, cellobiohydrolase; CDX, cyclodextrinase; CED, cellodextrinase; CDGT, cyclodextrin glucanotransferase; CHI, chitinase; CHITO, chitosanase; DEX, dextranase; EG, endoglucanase; endoNAG, endo *N*-acetyl-β-glucosaminidase; EXG, exo-1,3-β-glucanase; FRU, exo-β-fructosidase; G5-AMY, maltopentaose-forming amylase; G6-AMY, maltohexaose-forming amylase; IAMY, isoamylase; INU, inulinase; INV, invertase; LAM, laminarinase; LIC, lichenase; LPH, lactase phlorizin hydrolase; LVS, levansucrase; LYS, lysozyme; NABGLU, *N*-acetyl-β-glucosaminidase; NEUR, neuraminidase; OGLU, oligo-1,6-α-glucosidase; PBGLU, 6-phospho-β-glucosidase; PGLR, polygalacturonase; PUL, pullulanase; SI, sucrase-isomaltase; TREH, trehalase; XYN, xylanase.

‡ To whom correspondence should be addressed.

Table 1 Additions to the classification of glycosyl hydrolases

Abbreviation used: n.d., not done. Notes: [a]This sequence was previously unclassified for lack of detectable sequence similarity. It has since been found to be incorrect, resequenced and now falls in family 5 (Wang and Thomason, 1992). [b] C-terminal domain of this multi-domain protein. [c]N-terminal domain of this multi-domain protein. [d]These sequences are not new but were previously unclassified. [e]Full sequence data is now available for this enzyme and confirms its assignment to family 18. [f]β-galactosidases of family 42 display sequence similarity with a segment of $\sim$ 100 residues of β-galactosidases of family 2, but no detectable sequence similarity with the remaining $\sim$ 800 residues. It is difficult to consider the local similarity extending on $\sim$ 100 residues as fortuitous, and this is perhaps indicative of a multidomain structure for β-galactosidases. In the absence of three-dimensional structural data, it is felt safer to consider families 2 and 42 as distinct; although the two families might share some structural features, and perhaps constitute a superfamily, a clear division into two groups will remain. [g]This sequence was previously unclassified for lack of detectable sequence similarity. It has since been found to be incorrect, resequenced (Xu et al., 1991) and now falls in family 43.

Family ⇓	Enzyme	Source	EC number	SWISS-PROT accession number or reference	PROSITE accession number
1	BGLU A	*Clostridium thermocellum*	3.2.1.21	P26208	PDOC00495
1	BGLU	*Erwinia chrysanthemi*	3.2.1.21	P26206	
1	BGLU	*Manihot esculenta*	3.2.1.21	Hughes et al. (1992)	
1	BGLU B	*Microbispora bispora*	3.2.1.21	Wright et al. (1992)	
1	BGLU 1	*Trifolium repens*	3.2.1.21	P26204	
1	BGLU 2	*Trifolium repens*	3.2.1.21	P26205	
1	LPH	Rat	3.2.1.62/108	Duluc et al. (1991)	
1	PBGLU 2	*Escherichia coli*	3.2.1.86	P24240	
2	BGAL	*Clostridium thermosulfurogenes*	3.2.1.23	P26257	PDOC00531
2	BGAL Z	*Streptococcus thermophilus*	3.2.1.23	P23989	
2	BGAL	*Kluyveromyces lactis*	3.2.1.23	Poch et al. (1992)	
2	BGAL Z	*Lactobacillus delbruekii*	3.2.1.23	P20043	
2	BGAL L+M	*Leuconostoc lactis*	3.2.1.23	Davis et al. (1992)	
3	BGLU	*Agrobacterium tumefaciens*	3.2.1.21	P27034	PDOC00621
3	BGLU 3	*Aspergillus wentii*	3.2.1.21	P29090	
3	BGLU	*Trichoderma reesei*	3.2.1.21	Barnett et al. (1991)	
3	CED D	*Pseudomonas fluorescens*	3.2.1.74	Rixon et al. (1992)	
5	EG	*Bacillus* sp. strain KSM-64	3.2.1.4	Sumitomo et al. (1992)	PDOC00565
5	EG C	*Bacillus lautus*	3.2.1.4	Hansen (1992)	
5	EG A	*Butyrivibrio fibrisolvens*	3.2.1.4	P22541	
5	EG D	*Clostridium cellulolyticum*	3.2.1.4	P25472	
5	EG B	*Clostridium cellulovorans*	3.2.1.4	P28621	
5	EG D	*Clostridium cellulovorans*	3.2.1.4	P28623	
5	EG A	*Clostridium josui*	3.2.1.4	Fujino and Ohmiya (1992)	
5	EG 1	*Cryptococcus flavus*	3.2.1.4	Cui et al. (1992)	
5	EG A	*Prevotella (Bacteroides) ruminicola*	3.2.1.4	Vercoe and Gregg (1992)	
5	EG	*Pseudomonas solanacearum*	3.2.1.4	P17974	
5	EG A	*Ruminococcus flavefaciens* 17	3.2.1.4	Cunningham et al. (1991)	
5	EG A[a]	*Ruminococcus flavefaciens* FD1	3.2.1.4	Wang and Thomson (1992)	
5	EG	*Streptomyces lividans* 66	3.2.1.4	P27035	
5	EG 5	*Thermomonospora fusca*	3.2.1.4	Lao et al. (1991)	
5	EXG	*Candida albicans*	3.2.1.58	Cutfield et al. (1992)	
5	EXG	*Saccharomyces cerevisiae*	3.2.1.58	P23776	
5	CED C	*Pseudomonas fluorescens*	3.2.1.74	P27033	
6	EG A	*Streptomyces halstedii*	3.2.1.4	Fernandez-Abalos et al. (1992)	PDOC00563
6	EG 2	*Thermomonospora fusca*	3.2.1.4	P26222	
7	CBH I-1	*Phanerochaete chrysosporium*	3.2.1.91	Covert et al. (1992)	n.d.
7	CBH I-2	*Phanerochaete chrysosporium*	3.2.1.91	Covert et al. (1992)	
8	EG	*Bacillus* sp. KSM-330	3.2.1.4	P29019	PDOC00640

Table 1 (cont.)

Family ⇓	Enzyme	Source	EC number	SWISS-PROT accession number or reference	PROSITE accession number
9	EG C	*Clostridium cellulovorans*	3.2.1.4	P28622	PDOC00511
9	EG F	*Clostridium thermocellum*	3.2.1.4	P26224	
9	EG 4	*Thermomonospora fusca*	3.2.1.4	P26221	
10	XYN A	*Aspergillus kawachii*	3.2.1.8/4	Ito et al. (1992a)	PDOC00510
10	XYN B	*Butyrivibrio fibrisolvens* H17c	3.2.1.8	P26223	
10	XYN	*Penicillium chrysogenum*	3.2.1.8	Haas et al. (1992)	
10	XYN A[(b)]	*Ruminococcus flavefaciens*	3.2.1.8	P29126	
10	XYN A	*Streptomyces lividans*	3.2.1.8	P26514	
10	XYN I	*Streptomyces thermoviolaceus*	3.2.1.8	Tsujibo et al. (1992)	
11	XYN A	*Aspergillus niger*	3.2.1.8	Maat et al. (1992)	PDOC00622
11	XYN C	*Aspergillus kawachii*	3.2.1.8	Ito et al. (1992b)	
11	XYN A	*Aspergillus tubigensis*	3.2.1.8	de Graaf et al. (1992)	
11	XYN	*Chainia* sp.	3.2.1.8	Bastawde et al. (1991)	
11	XYN A[(b)]	*Neocallimastix patriciarum*	3.2.1.8	P29127	
11	XYN A[(c)]	*Neocallimastix patriciarum*	3.2.1.8	P29127	
11	XYN 2	*Nocardiopsis dassonvillei*	3.2.1.8	Tsujibo et al. (1991)	
11	XYN A[(c)]	*Ruminococcus flavefaciens*	3.2.1.8	P29126	
11	XYN	*Schizophyllum commune*	3.2.1.8	Shareck et al. (1991)	
11	XYN B	*Streptomyces lividans*	3.2.1.8	P26515	
11	XYN C	*Streptomyces lividans*	3.2.1.8	P26220	
11	XYN	*Streptomyces* sp. No. 36a	3.2.1.8	Shareck et al. (1991)	
11	XYN II	*Streptomyces thermoviolaceus*	3.2.1.8	Tsujibo et al. (1992)	
11	XYN	*Trichoderma harzanium*	3.2.1.8	Yaguchi et al. (1992b)	
11	XYN 1	*Trichoderma reesei*	3.2.1.8	Törrönen et al. (1992)	
11	XYN 2	*Trichoderma reesei*	3.2.1.8	Törrönen et al. (1992)	
11	XYN	*Trichoderma viride*	3.2.1.8	Yaguchi et al. (1992a)	
13	AAMY	*Aeromonas hydrophila*	3.2.1.1	P22630	n.d.
13	AAMY	*Alteromonas haloplanctis*	3.2.1.1	Feller et al. (1992)	
13	AAMY	*Bacillus megaterium*	3.2.1.1	P20845	
13	AAMY	*Butyrivibrio fibrisolvens*	3.2.1.1	Rumbak et al. (1991)	
13	AAMY	*Escherichia coli*	3.2.1.1	P25718	
13	AAMY 2	*Escherichia coli*	3.2.1.1	P26612	
13	AAMY	Rice	3.2.1.1	P27935	
13	AAMY-c	Rice	3.2.1.1	P27940	
13	AAMY 2	*Salmonella typhimurium*	3.2.1.1	P26613	
13	AAMY	*Streptomyces violaceus*	3.2.1.1	P22998	
13	AAMY	*Schwanniomyces occidentalis*	3.2.1.1	P19269	
13	AAMY	*Vigna radiata*	3.2.1.1	Koizuka et al. (1990)	
13	AAMY	*Xanthomonas campestris*	3.2.1.1	Hu et al. (1992)	
13	AGLU	*Candida albicans*	3.2.1.20	Geber et al. (1992)	
13	AGLU Z	*Escherichia coli*	3.2.1.20	P21517	
13	CDGT	*Bacillus circulans*	2.4.1.19	Nitschke et al. (1990)	
13	CDGT	*Bacillus ohbensis*	2.4.1.19	P27036	
13	G5-AMY	*Pseudomonas* sp. KO-8940	3.2.1.-	Shida et al. (1992)	
13	G6-AMY	*Bacillus* sp. H-167	3.2.1.98	Shirokizawa et al. (1990)	
13	OGLU	*Bacillus cereus*	3.2.1.10	P21332	
13	OGLU	*Bacillus* sp.	3.2.1.10	P29093	

Table 1 (cont.)

Family ⇓ Enzyme	Source	EC number	SWISS-PROT accession number or reference	PROSITE accession number
13 OGLU	*Bacillus thermoglucosidasus*	3.2.1.10	P29094	
13 OGLU	*Escherichia coli*	3.2.1.10	P28904	
13 PUL	*Bacillus* KSM-1876	3.2.1.41	Igarashi et al. (1992)	
13 IAMY	*Pseudomonas* sp. strain smp1	3.2.1.68	P26501	
13 CDX	*Clostridium thermohydrosulfuricum*	3.2.1.54	Podkovyrov and Zeikus (1992)	
14 BAMY	*Arabidopsis thaliana*	3.2.1.2	P25853	PDOC00414
14 BAMY	*Clostridium thermosulfurogenes*	3.2.1.2	P19584	
14 BAMY	Rye	3.2.1.2	Rorat et al. (1991)	
15 AMG	*Aspergillus oryzae*	3.2.1.3	Hata et al. (1991)	n.d.
15 AMG	*Aspergillus awamori*	3.2.1.3	P23176	
15 AMG	*Aspergillus shirousami*	3.2.1.3	P22832	
16 LIC	*Bacillus licheniformis*	3.2.1.73	P27051	n.d.
16 LIC B	*Clostridium thermocellum*	3.2.1.73	Schimming et al. (1992)	
16 LAM	*Clostridium thermocellum*	3.2.1.39	Zverlov et al. (1991)	
16 AGAR[(d)]	*Streptomyces coelicolor*	3.2.1.81	P07883	
17 LAM A	*Lycopersicon esculentum*	3.2.1.39	van Kan et al. (1992)	PDOC00507
17 LAM B	*Lycopersicon esculentum*	3.2.1.39	van Kan et al. (1992)	
17 LAM	*Nicotiana plumbaginifolia*	3.2.1.39	Castresana et al. (1991)	
17 LAM	*Nicotiana tabacum*	3.2.1.39	P23432	
17 LAM	*Nicotiana tabacum*	3.2.1.39	P23433	
17 LAM	*Nicotiana tabacum*	3.2.1.39	P23547	
17 LAM	*Nicotiana tabacum*	3.2.1.39	P23546	
17 LAM	*Nicotiana plumbaginifolia*	3.2.1.39	P23431	
17 LAM	*Phaseolus vulgaris*	3.2.1.39	P23535	
17 LIC	Rice	3.2.1.73	Simmons et al. (1992)	
18 Toxin α-chain	*Kluyveromyces lactis*	3.2.1.14	P09805	n.d.
18 CHI 1	*Aphanocladium album*	3.2.1.14	Blaiseau and Lafay (1992)	
18 CHI[(e)]	*Hevea brasiliensis*	3.2.1.14	P23472	
18 endoNAG	*Flavobacterium* sp.	3.2.1.96	P80036	
18 endoNAG F1	*Flavobacterium meningosepticum*	3.2.1.96	Tarentino et al. (1992)	
18 CHI D	*Bacillus circulans*	3.2.1.14	P27050	
18 CHI	*Brugia malayi*	3.2.1.14	P29030	
18 CHI A	*Nicotiana tabacum*	3.2.1.14	P29060	
18 CHI B	*Nicotiana tabacum*	3.2.1.14	P29061	
18 CHI A	*Phaseolus angularis*	3.2.1.14	P29024	
18 CHI	*Rhizopus niveus*	3.2.1.14	P29025	
18 CHI I	*Rhizopus oligosporus*	3.2.1.14	P29026	
18 CHI II	*Rhizopus oligosporus*	3.2.1.14	P29027	
18 CHI-1	*Saccharomyces cerevisiae*	3.2.1.14	P29028	
18 CHI-2	*Saccharomyces cerevisiae*	3.2.1.14	P29029	
18 CHI-63	*Streptomyces plicatus*	3.2.1.14	P11220	
18 CHI 1	Sugar beet	3.2.1.14	Mikkelsen et al. (1992)	
19 CHI B4	*Brassica napus*	3.2.1.14	Rasmussen et al. (1992)	PDOC00620
19 CHI	*Dioscorea japonica*	3.2.1.14	P80052	

Table 1 (cont.)

Family ⇓ Enzyme	Source	EC number	SWISS-PROT accession number or reference	PROSITE accession number
19 CHI-26	*Hordeum vulgare*	3.2.1.14	Leah et al. (1991)	
19 CHI A	Maize	3.2.1.14	P29022	
19 CHI B	Maize	3.2.1.14	P29023	
19 CHI	*Nicotiana tabacum*	3.2.1.14	P24091	
19 CHI	*Nicotiana tabacum*	3.2.1.14	P29059	
19 CHI	*Petunia hybrida*	3.2.1.14	P29021	
19 CHI-4	*Phaseolus vulgaris*	3.2.1.14	P27054	
19 CHI B	*Populus trichocarpa*	3.2.1.14	P29031	
19 CHI C	*Populus trichocarpa*	3.2.1.14	P29032	
19 CHI-1	Rice	3.2.1.14	P24626	
19 CHI-2	Rice	3.2.1.14	P25765	
19 CHI 4	Sugar beet	3.2.1.14	Mikkelsen et al. (1992)	
19 CHI	*Urtica dioica*	3.2.1.14	P11218	
20 NABGLU	Mouse (β-chain)	3.2.1.52	P20060	n.d.
22 LYS C	*Chrysolophus amherstiae*	3.2.1.52	P22910	PDOC00119
22 LYS C	*Pavo cristatus*	3.2.1.52	P19849	
22 LYS C	*Pseudocheirus peregrinus*	3.2.1.52	P21776	
22 LYS CP	Mouse	3.2.1.52	P17897	
23 LYS G	Chicken	3.2.1.17	P27042	
25 LYS CH	*Clostridium acetobutylicum*	3.2.1.17	Croux and Garcia (1991)	n.d.
26 EG	*Bacteroides ruminicola*	3.2.1.4	Matsushita et al. (1991)	n.d.
28 PGLR	*Aspergillus tubigensis*	3.2.1.15	P19805	PDOC00502
28 PGLR	*Cochliobolus carbonum*	3.2.1.15	P26215	
28 PGLR	Maize	3.2.1.15	P26216	
28 PGLR	*Oenothera organensis*	3.2.1.15	P24548	
31 AGLU	*Candida tsukubaensis*	3.2.1.20	P29064	PDOC00120
31 AMG	*Schwanniomyces occidentalis*	3.2.1.3	P22861	
31 SI	Rat	3.2.1.48/10	P23739	
32 INV	Carrot	3.2.1.26	P26792	PDOC00532
32 INV	*Escherichia coli*	3.2.1.26	P16553	
32 INV 1	Tomato	3.2.1.26	P29000	
32 INV A	*Vigna radiata*	3.2.1.26	P29001	
32 LVS	*Bacillus amyloliquefaciens*	2.4.1.10	Tang et al. (1990)	
32 INU 1	*Kluyveromyces marxianus*	3.2.1.7	P28999	
32 FRU A	*Streptococcus mutans*	3.2.1.80	Burne and Penders (1992)	
33 NEUR	*Salmonella typhimurium*	3.2.1.18	Hoyer et al. (1992)	n.d.
New families:				
36 AGAL[(d)]	*Escherichia coli*	3.2.1.22	P16551	n.d.
36 AGAL	*Streptococcus mutans*	3.2.1.22	Aduse-Opoku et al. (1991)	

Table 1 (cont.)

Family ⇓ Enzyme	Source	EC number	SWISS-PROT accession number or reference	PROSITE accession number
37 TREH[d]	*Escherichia coli*	3.2.1.28	P13482	n.d.
37 TREH	Rabbit	3.2.1.28	P19813	
38 AMAN[d]	*Saccharomyces cerevisiae*	3.2.1.24	P22855	n.d.
38 AMAN	Mouse	3.2.1.114	P27046	
38 AMAN	Rat	3.2.1.24	P21139	
39 BXYL B	*Caldocellum saccharolyticum*	3.2.1.37	P23552	n.d.
39 AIDU	Dog	3.2.1.76	Stoltzfus et al. (1992)	
39 BXYL	*Thermoanaerobacter* B6A	3.2.1.37	Lee et al. (1992)	
40 BGLU RolB	*Agrobacterium rhizogenes*	3.2.1.-	P20402	n.d.
40 BGLU RolB	*Nicotiana glauca*	3.2.1.-	P09178	
41 BGLU RolC	*Agrobacterium rhizogenes*	3.2.1.-	P20403	n.d.
41 BGLU RolC	*Nicotiana glauca*	3.2.1.-	P07051	
42 BGAL[f]	*Bacillus stearothermophilus*	3.2.1.23	P19668	n.d.
42 BGAL[f]	Thermophilic anaerobe NA10	3.2.1.23	Saito et al. (1992)	
43 BXYL[g]	*Bacillus pumilus*	3.2.1.37	Xu et al. (1991)	n.d.
43 BXYL/ARAF	*Butyrivibrio fibrisolvens*	3.2.1.37/55	Utt et al. (1991)	
44 EG A	*Bacillus lautus*	3.2.1.4	Hansen et al. (1992)	n.d.
44 BMAN/EG[b]	*Caldocellum saccharolyticum*	3.2.1.4	Gibbs et al. (1992)	
45 EG B[d]	*Pseudomonas fluorescens*	3.2.1.4	P18126	n.d.
45 EG V	*Humicola insolens*	3.2.1.4.	Rasmussen et al. (1991)	
Unclassified:				
LAM	*Oerskovia xanthineolytica*	3.2.1.39	Shen et al. (1991)	
DEX	*Arthrobacter* sp.	3.2.1.11	Okushima et al. (1991)	
AMAN	*Dictyostelium discoideum*	3.2.1.24	Schatzle et al. (1992)	
LYS	Bacteriophage SF6	3.2.1.17	P21270	
CHITO	*Bacillus circulans*	3.2.1.-	Ando et al. (1992)	

(contain at least two EC numbers). Only seven sequences (< 1.5% of the sample) have no counterpart and are presently left unclassified. The complete classification is available from the authors on request.

DISCUSSION

The complementarity of this classification system with that of I.U.B. has been outlined in our previous work (Henrissat, 1991). An advantage of this classification is that a **protein** or a **gene translation** or even a **domain** can be classified before even knowing its enzymic activity. In fact, more and more glycosyl hydrolases have been found to consist of several catalytic domains. The present classification is unambiguous, because each catalytic domain can be classified. In addition, the classification based on amino acid similarities can also specify the location (for instance N- or C-terminal) of each of the domains of a multiple domain glycosyl hydrolase.

Although it is likely that most unclassified sequences will form new families when related sequences become available, it is safer not to count them as families until significant sequence similarities are demonstrated (see above examples of the cellodextrinase of *Ruminococcus flavefaciens* FD1 or the β-D-xylosidase of *Bacillus pumilus*). It is also possible that some of the presently unclassified sequences will fall into established families if they are found significantly related to new sequences that are also significantly related to enzymes already classified.

We have implemented the SWISS-PROT protein sequence

data bank (Bairoch & Boeckmann, 1992) with the present classification of glycosyl hydrolases. Each relevant entry now contains, in the comment section (CC lines), the following type of information:

```
CC   -!- SIMILARITY: BELONGS TO FAMILY xx OF GLYCOSYL HYDROLASES.
```

Similarly, the PROSITE dictionary of sites and patterns in proteins (Bairoch, 1992) currently includes signature patterns specific for 18 different families of glycosyl hydrolases. The relevant PROSITE documentation accession numbers for the families shown in Table 1 are indicated. It is planned to develop signature patterns for most if not all the families of glycosyl hydrolases.

Cellulases and xylanases probably represent the first types of glycosyl hydrolases whose classification was greatly clarified by sequence similarity grouping. On the basis of the comparison of 21 sequences, six families termed A–F had been identified (Henrissat et al., 1989). With 67 sequences analysed, three families (G–I) were later added to the classification (Gilkes et al., 1991b). Family I contained only one entry that has recently been re-sequenced and placed in family A (Wang and Thomson, 1992). There are now more than 120 sequences of cellulases, xylanases and related enzymes in the present classification. Because the classification (with letters) of cellulases is older than the present classification (with numbers) and is being widely used, the correspondence between the two classifications is shown in Table 2.

The three-dimensional fold being better conserved than the sequence of proteins, it is expected that the same fold will be found for each member of a family. The validity of this premise can be indirectly verified by determining whether all members of a given family share the same general catalytic mechanism. Gebler et al. (1992b) have examined the stereochemistry of hydrolysis of 16 cellulases and xylanases belonging to six families of the present classification and found that the representatives of a given family indeed displayed the same stereoselectivity. In a similar vein, the hydrolysis patterns of a series of chromophoric glycosides derived from D-glucose, cellobiose, higher cellodextrins, lactose, D-xylose and β-(1,4)-xylobiose by 15 cellulolytic enzymes allowed their grouping in six families coinciding with the classification based on sequence similarities (Claeyssens and Henrissat, 1992). This study also showed that the low-molecular-mass substrates did not discriminate exo- (EC 3.2.1.91) from endo- (EC 3.2.1.4) cellulases. On the other hand, because (i) 8–15% sequence identity can be found in structurally related proteins as well as in unrelated proteins (Chothia, 1992) and (ii) it is still impossible to predict the three-dimensional fold of a protein from its sequence alone, it cannot be excluded that some of the families of the present classification have related folds. In other words, proteins belonging to two different families do not necessarily have different folds. Comparison of the three-dimensional structures of a large number of glycosyl hydrolases from different families would provide an answer, but is presently impossible, for lack of structural information, for the vast majority of these enzymes. It should be noted, however, that the three-dimensional structures of the catalytic domain of α-amylases and cyclodextrin glucanotransferases (CDGTases), both belonging to family 13, have been found to share a superimposable $(\beta/\alpha)_8$ barrel structure (Farber and Petsko, 1990).

Table 2 Correspondence between the present classification and that of cellulases

Family in the present classification	Corresponding cellulase family	Number of cellulase/xylanase sequences used for the grouping	Reference
5	A	21	Henrissat et al. (1989)
6	B	21	Henrissat et al. (1989)
7	C	21	Henrissat et al. (1989)
8	D	21	Henrissat et al. (1989)
9	E	21	Henrissat et al. (1989)
10	F	21	Henrissat et al. (1989)
11	G	67	Gilkes et al. (1991a,b)
12	H	67	Gilkes et al. (1991a,b)
26	I	73	Henrissat (1991)
44	J	123	The present work
45	K	123	The present work

During our sequence comparisons, we have examined in detail the location of Asp and Glu residues in each family. These residues are commonly found to be catalytic in glycosyl hydrolases, either as proton donors in their protonated form or as nucleophile or oxocarbonium stabilizing agents in their charged form (Sinnott, 1990). Examination of their conservation constitutes a useful way of predicting the catalytic residues of glycosyl hydrolases (Zvelebil and Sternberg, 1988; Henrissat et al., 1989; Baird et al., 1990). Another example is given by the recent work of Gebler et al. (1992a) who elegantly showed that Glu^{537}, not Glu^{461} is the nucleophile in the active site of β-galactosidase (lacZ) from *Escherichia coli* which belongs to family 2. Since β-glucuronidases belong to the same family (family 2), one can predict which Glu is the nucleophile in β-glucuronidases.

Predictions of catalytic residues based on the conservation of Asp and Glu residues have sometimes been verified experimentally (Baird et al., 1990; Py et al., 1991) thus demonstrating the usefulness of the approach. The method is straightforward, but sensitive to sequencing inaccuracies: for instance, in family 6, only four Asp and Glu residues were found to be conserved (Henrissat et al., 1989), suggesting that the catalytic amino acid(s) of this family of cellulases should be one of these. The three-dimensional structure of one member of this family has since been later solved (Rouvinen et al., 1990) and showed that the catalytic Asp was none of the four candidates and was not conserved in one of the proteins. Because this observation conflicted with the notion that residues in the active centre are better conserved than those in more remote parts of the protein, the gene sequence coding for the 'anomalous' protein was carefully re-sequenced in the critical area. Results showed that a sequencing error in the original work had produced a local reading frameshift and that the corrected gene sequence restored the missing catalytic Asp residue in the protein (Gilkes et al., 1991a). Possible sequencing inaccuracies around catalytic residues have also recently been reported for a β-galactosidase (Gebler et al., 1992a) and for a few cellulases (Henrissat, 1993). It is inevitable that a certain amount of error occurs during the sequencing and/or during sequence data handling/processing, and the present classification could perhaps also help in their detection.

We thank Dr. M. Claeyssens for critically reading the manuscript and Dr. Rémi Spilliaert for suggesting that *Streptomyces coelicolor* agarase could be elated to family 16. This work was supported by a CM2AO research grant from Organibio.

REFERENCES

Aduse-Opoku, J., Tao, L., Ferretti, J. J. and Russell, R. R. B. (1991) J. Gen. Microbiol. **137**, 757–764

Ando, A., Noguchi, K., Yanagi, M., Shinoyama, H., Kagawa, Y., Hirata, H., Yabuki, M. and Fujii, T. (1992) J. Gen. Appl. Microbiol. **38**, 135–144

Baird, S. D., Hefford, M. A., Johnson, D. A., Sung, W. L., Yaguchi, M. and Seligy, V. L. (1990) Biochem. Biophys. Res. Commun. **169**, 1035–1039

Bairoch, A. (1992) Nucleic Acids Res. **20**, 2013–2018
Bairoch, A. and Boeckmann, B. (1992) Nucleic Acids Res. **20**, 2019–2022
Barnett, C. C., Berka, R. M. and Fowler, T. (1991) Bio/Technology **9**, 562–567
Bastawde, K. B., Tabatabai, L. B., Meagher, M. M., Srinivasan, M. C., Vartak, H. G., Rele, M. V. and Reilly, P. J. (1991) ACS Symp. Ser. **460**, 417–425
Blaiseau, P. L. and Lafay, J. F. (1992) Gene **120**, 243–248
Burne, R. A. and Penders, J. E. (1992) Infect. Immun. **60**, 4621–4632
Castresana, C., de Carvalho, F., Gheysen, G., Habets, M., Inzé, D. and Van Montagu, M. (1991) Plant Cell **2**, 1131–1143
Chothia, C. (1992) Nature (London) **357**, 543–544
Claeyssens, M. and Henrissat, B. (1992) Protein Sci. **1**, 1293–1297
Covert, S. F., Wymelenberg, A. V. and Cullen, D. (1992) Appl. Environ. Microbiol. **58**, 2168–2175
Croux, C. and Garcia, J. L. (1991) Gene **104**, 25–31
Cui, Z., Mochizuki, D., Matsuno, Y., Nakamura, T., Liu, Y., Hatano, T., Fukui, S. and Miyakawa, T. (1992) Biosci. Biotechnol. Biochem. **56**, 1230–1235
Cunningham, C., McPherson, C. A., Martin, J., Harris, W. J. and Flint, H. J. (1991) Mol. Gen. Genet. **228**, 320–323
Cutfield, S., Brooke, G., Sullivan, P. and Cutfield, J. (1992) J. Mol. Biol. **225**, 217–218
Davis, S., Stevens, H., van Riel, M., Simons, G. and de Vos, W. M. (1992) J. Bacteriol. **174**, 4475–4481
de Graaf, L. H., van den Broeck, H. C., van Ooijen, A. J. J. and Visser, J. (1992) Xylans and Xylanases (Visser, J., Beldman, G., Kusters-van Someren, M. A. and Voragen, A. G. J., eds.), pp. 235–246, Elsevier Science Publishers, Amsterdam
Duluc, I., Boukamel, R., Mantei, N., Semenza, G., Raul, F. and Freund, J. N. (1991) Gene **103**, 275–276
Farber, G. K. and Petsko, G. A. (1990) Trends Biochem. Sci. **15**, 228–234
Feller, G., Lonhienne, T., Deroanne, C., Libioulle, C., Van Beeumen, J. and Gerday, C. (1992) J. Biol. Chem. **267**, 5217–5221
Fernandez-Abalos, J. M., Sanchez, P., Coll, P. M., Villanueva, J. R., Pérez, P. and Santamaria, R. I. (1992) J. Bacteriol. **174**, 6368–6376
Fujino, T. and Ohmiya, K. (1992) J. Ferment. Bioeng. **73**, 308–313
Geber, A., Williamson, P. R., Rex, J. H., Sweeney, E. C. and Bennett, J. E. (1992) J. Bacteriol. **174**, 6992–6996
Gebler, J. C., Aebersold, R. and Withers, S. G. (1992a) J. Biol. Chem. **267**, 11126–11130
Gebler, J. C., Gilkes, N. R., Claeyssens, M., Wilson, D. B., Béguin, P., Wakarchuk, W. W., Kilburn, D. G., Miller, R. C., Warren, R. A. J. and Withers, S. G. (1992b) J. Biol. Chem. **267**, 12559–12561
Gibbs, M. D., Saul, D. J., Lüthi, E. and Bergquist, P. L. (1992) Appl. Environ. Microbiol. **58**, 3864–3867
Gilkes, N. R., Claeyssens, M., Aebersold, R., Henrissat, B., Meinke, A., Morrison, H. D., Kilburn, D. G., Warren, R. A. J. and Miller, R. C. (1991a) Eur. J. Biochem. **202**, 367–377
Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C. and Warren, R. A. J. (1991b) Microbiol. Rev. **55**, 303–315
Haas, H., Herfurth, E., Stöffler, G. and Redl, B. (1992) Biochim. Biophys. Acta **1117**, 279–286
Hansen, C. K. (1992) Ph.D. Thesis, Technical University of Denmark, Lyngby, Denmark
Hansen, C. K., Diderichsen, B. and Jorgensen, P. L. (1992) J. Bacteriol. **174**, 3522–3531
Hata, Y., Kitamoto, K., Gomi, K., Kumagai, C., Tamura, G. and Hara, S. (1991) Agric. Biol. Chem. **55**, 941–949
Henrissat, B. (1991) Biochem. J. **280**, 309–316
Henrissat, B. (1993) Gene **125**, 199–204
Henrissat, B., Claeyssens, M., Tomme, P., Lemesle, L. and Mornon, J. P. (1989) Gene **81**, 83–95
Hoyer, L. L., Hamilton, A. C., Steenbergen, S. M. and Vimr, E. R. (1992) Mol. Microbiol. **6**, 873–884
Hu, N.-T., Hung, M.-N., Huang, A.-M., Tsai, H.-F., Yang, B.-Y., Chow, T.-Y. and Tseng, Y.-H. (1992) J. Gen. Microbiol. **138**, 1647–1655
Hughes, M. A., Brown, K., Pancoro, A., Murray, B. S., Oxtoby, E. and Hughes, J. (1992) Arch. Biochem. Biophys. **295**, 273–279
I.U.B. (1984) Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry on the Nomenclature and Classification of Enzyme-Catalysed Reactions, Academic Press, London and New York
Igarashi, K., Ara, K., Saeki, K., Ozaki, K., Kawai, S. and Ito, S. (1992) Biosci. Biotechnol. Biochem. **56**, 514–516
Ito, K., Ikemasu, T. and Ishikawa, T. (1992a) Biosci. Biotechnol. Biochem. **56**, 906–912
Ito, K., Iwashita, K. and Iwano, K. (1992b) Biosci. Biotechnol. Biochem. **56**, 1338–1340
Koizuka, N., Tanaka, Y. and Morohashi, Y. (1990) Plant Physiol. **94**, 1488–1491
Lao, G., Ghangas, G. S., Jung, E. D. and Wilson, D. B. (1991) J. Bacteriol. **173**, 3397–3407
Leah, R., Tommerup, H., Svendsen, I. and Mundy, J. (1991) J. Biol. Chem. **266**, 1564–1573
Lee, Y.-E., Lowe, S. E. and Zeikus, J. G. (1992) Xylans and Xylanases (Visser, J., Beldman, G., Kusters-van Someren, M. A. and Voragen, A. G. J. eds.), pp. 275–288, Elsevier Science Publishers, Amsterdam
Maat, J., Roza, M., Verbakel, J., Stam, H., Santos da Silva, M. J., Bosse, M., Egmond, M. R., Hagemans, M. L. D., van Gorcom, R. F. M., Hessing, J. G. M., van den Hondel, C. A. M. J. J. and van Rotterdam, C. (1992) Xylans and Xylanases (Visser, J., Beldman, G., Kusters-van Someren, M. A. and Voragen, A. G. J. eds.), pp. 349–360, Elsevier Science Publishers, Amsterdam
Matsushita, O., Russell, J. B. and Wilson, D. B. (1991) J. Bacteriol. **173**, 6919–6926
Mikkelsen, J. D., Berglund, L., Nielsen, K. K., Christiansen, H. and Bojsen, K. (1992) Advances in Chitin and Chitosan, (Brine, C. J., Sandford, P. A. and Zikakis, J. P. eds.), pp. 344–353, Elsevier Applied Science, London
Nitschke, L., Heeger, K., Bender, H. and Schulz, G. E. (1990) Appl. Microbiol. Biotechnol. **33**, 542–546
Okushima, M., Sugino, D., Kouno, Y., Nakano, S., Miyahara, J., Toda, H., Kubo, S. and Matsushiro, A. (1991) Jpn. J. Genet. **66**, 173–187
Poch, O., L'Hôte, H., Dallery, V., Debeaux, F., Fleer, R. and Sodoyer, R. (1992) Gene **118**, 55–63
Podkovyrov, S. M. and Zeikus, J. G. (1992) J. Bacteriol. **174**, 5400–5405
Py, B., Bortoli-German, I., Haiech, J., Chippaux, M. and Barras, F. (1991) Protein Eng. **4**, 325–333
Rasmussen, G., Mikkelsen, J. M., Schülein, M., Patkar, S. A., Hagen, F., Hjort, C. M. and Hastrup, S. (1991) World Pat. WO 91 17,243
Rasmussen, U., Bojsen, K. and Collinge, D. B. (1992) Plant Mol. Biol. **20**, 277–287
Rixon, J. E., Ferreira, L. M. A., Durrant, A. J., Laurie, J. I., Hazlewood, G. P. and Gilbert, H. J. (1992) Biochem. J. **285**, 947–955
Rorat, T., Sadowski, J., Grellet, F., Daussant, J. and Delseny, M. (1991) Theor. Appl. Genet. **83**, 257–263
Rouvinen, J., Bergfors, T., Teeri, T., Knowles, J. K. C. and Jones, T. A. (1990) Science **249**, 380–386
Rumbak, E., Rawlings, D. E., Lindsey, G. G. and Woods, D. R. (1991) J. Bacteriol. **173**, 4203–4211
Saito, T., Kato, K., Suzuki, T., Iijima, S. and Kobayashi, T. (1992) J. Ferment. Bioeng. **73**, 51–53
Schatzle, J., Bush, J. and Cardelli, J. (1992) J. Biol. Chem. **267**, 4000–4007
Schimming, S., Schwarz, W. H. and Staudenbauer, W. L. (1992) Eur. J. Biochem. **204**, 13–19
Shareck, F., Roy, C., Yaguchi, M., Morosoli, R. and Kluepfel, D. (1991) Gene **107**, 75–82
Shen, S. H., Chrétien, P., Bastien, L. and Slilaty, S. N. (1991) J. Biol. Chem. **266**, 1058–1063
Shida, O., Takano, T., Takagi, H., Kadowaki, K. and Kobayashi, S. (1992) Biosci. Biotechnol. Biochem. **56**, 76–80
Shirokizawa, O., Akiba, T. and Horikoshi, K. (1990) FEMS Microbiol. Lett. **70**, 131–136
Simmons, C. R., Litts, J. C., Huang, N. and Rodriguez, R. L. (1992) Plant Mol. Biol. **18**, 33–45
Sinnott, M. L. (1990) Chem. Rev. **90**, 1171–1202
Stoltzfus, L. J., Sosa-Pineda, B., Moskowitz, S. M., Menon, K. P., Dlott, B., Hooper, L., Teplow. D. B., Shull, R. M. and Neufeld, E. F. (1992) J. Biol. Chem. **267**, 6570–6575
Sumitomo, N., Ozaki, K., Kawai, S. and Ito, S. (1992) Biosci. Biotechnol. Biochem. **56**, 872–877
Tang, L. B., Lenstra, R., Borchert, T. V. and Nagarajan, V. (1990) Gene **96**, 89–93
Tarentino, A. L., Quinones, G., Schrader, W. P., Changchien, L. M. and Plummer, T. H. (1992) J. Biol. Chem. **267**, 3868–3872
Törrönen, A., Hodits, R., Gonzalez, R., Mach, R. L., Messner, R., Kalkkinen, N., Harrki, A. and Kubicek, C. P. (1992) Biotechnology in Pulp and Paper Industry (Kuwahara, M. and Shimada, M., eds.), pp. 447–452, Uni Publishers, Tokyo
Tsujibo, H., Sakamoto, T., Miyamoto, K., Hasegawa, T., Fujimoto, M. and Inamori, Y. (1991) Agric. Biol. Chem. **55**, 2173–2174
Tsujibo, H., Miyamoto, K., Kuda, T., Minami, K., Sakamoto, T., Hasegawa, T. and Inamori, Y. (1992) Appl. Environ. Microbiol. **58**, 371–375
Utt, E. A., Eddy, C. K., Keshav, K. F. and Ingram, L. O. (1991) Appl. Environ. Microbiol. **57**, 1227–1234
van Kan, J. A. L., Joosten, M. H. A. J. and Wagemakers, C. A. M. (1992) Plant Mol. Biol. **20**, 513–527
Vercoe, P. E. and Gregg, K. (1992) Mol. Gen. Genet. **233**, 284–292
Wang, W. and Thomson, J. A. (1992) Mol. Gen. Genet. **233**, 492
Wright, R. M., Yablonsky, M. D., Shalita, Z. P., Goyal, A. K. and Eveleigh, D. E. (1992) Appl. Environ. Microbiol. **58**, 3455–3465
Xu, W. Z., Shima, Y., Negoro, S. and Urabe, I. (1991) Eur. J. Biochem. **202**, 1197–1203
Yaguchi, M., Roy, C., Ujiie, M., Watson, D. C. and Wakarchuk, W. (1992a) Xylans and Xylanases (Visser, J., Beldman, G., Kusters-van Someren, M. A. and Voragen, A. G. J., eds.), pp. 149–154, Elsevier Science Publishers, Amsterdam
Yaguchi, M., Roy, C., Watson, D. C., Rollin, F., Tan, L. U. L., Senior, D. J. and Saddler, J. N. (1992b) Xylans and Xylanases (Visser, J., Beldman, G., Kusters-van Someren, M. A. and Voragen, A. G. J., eds.), pp. 435–438, Elsevier Science Publishers, Amsterdam
Zvelebil, M. J. J. and Sternberg, M. J. E. (1988) Prot. Eng. **2**, 127–138
Zverlov, V. V., Laptev, D. M., Tishkov, V. I. and Velikodvorskaja, G. A. (1991) Biochem. Biophys. Res. Commun. **181**, 507–512

Received 30 November 1993/2 February 1993; accepted 18 February 1993