Biochem. J. (1991) 280, 309–316 (Printed in Great Britain)

A classification of glycosyl hydrolases based on amino acid sequence similarities

Bernard HENRISSAT
Centre de Recherches sur les Macromolécules Végétales, C.N.R.S., B.P. 53X, F-38041 Grenoble, France

The amino acid sequences of 301 glycosyl hydrolases and related enzymes have been compared. A total of 291 sequences corresponding to 39 EC entries could be classified into 35 families. Only ten sequences (< 5% of the sample) could not be assigned to any family. With the sequences available for this analysis, 18 families were found to be monospecific (containing only one EC number) and 17 were found to be polyspecific (containing at least two EC numbers). Implications on the folding characteristics and mechanism of action of these enzymes and on the evolution of carbohydrate metabolism are discussed. With the steady increase in sequence and structural data, it is suggested that the enzyme classification system should perhaps be revised.

INTRODUCTION

The extensive variety of stereochemistry of carbohydrates is paralleled by a large multiplicity of the enzymes involved in their metabolism. Glycosyl hydrolases (EC 3.2.1.*x*) are key enzymes of carbohydrate metabolism that are found in the three major kingdoms (archaebacteria, eubacteria and eukaryotes). Heritable deficiencies in glycosyl hydrolases, for example lactose intolerance (Auricchio *et al.*, 1963) or mucopolysaccharidosis (Neufeld *et al.*, 1975), are among the most frequent genetically based syndromes in man.

The IUB Enzyme Nomenclature (1984) is based on the type of reaction that enzymes catalyse and on their substrate-specificity. For glycosyl hydrolases (EC 3.2.1.*x*), the first three digits indicate enzymes hydrolysing *O*-glycosyl linkages whereas the last number indicates the substrate and sometimes reflects the molecular mechanism. This classification is very useful, especially to avoid ambiguities and the proliferation of trivial names, and provides a unique classification. However, at least in the case of glycosyl hydrolases, such a classification does not necessarily reflect (and was not intended to) the structural features of the enzymes. In fact, a classification based primarily on the substrate cannot take into account evolutionary events such as (i) divergence (which can result in specificity, and sometimes reaction-type, changes) or (ii) convergent evolution, which may force polypeptides with different folds to catalyse the same reaction on the same substrate. Another problem with the EC classification is that it is not appropriate for enzymes showing broad specificity (i.e. that act on several substrates).

It is now clear that there is a direct relationship between sequence and folding similarities (Chothia & Lesk, 1986). We and others (Henrissat *et al.*, 1989; Henrissat, 1990, 1991; Svensson, 1988; MacGregor & Svensson, 1989; Raimbaud *et al.*, 1989) have begun a systematic comparison of the primary sequences of glycosyl hydrolases and have found (i) families containing several EC entries and (ii) enzymes with similar substrate-specificities that belong to non-related families. A classification that would better reflect sequence (hence structural) similarity would certainly prove useful, especially with the rapidly growing number of glycosyl hydrolase genes that are being sequenced and with the concomitant (although slower) increase in the number of three-dimensional structures being solved. The purpose of the present paper is to present such a classification for glycosyl hydrolases. This classification has become possible because of two favourable factors: (*a*) a rapid increase in the number of available sequences and (*b*) availability of new sequence comparison methods that can deal with low sequence identity (Gaboriaud *et al.*, 1987; Lemesle-Varloot *et al.*, 1990). This classification is necessarily incomplete since many of the enzymes that are listed in the IUB nomenclature are not yet sequenced. Additions to that classification can therefore be expected in the future.

METHODS

A library of glycosyl hydrolase sequence files was constructed with the sequences having an EC 3.2.1.X pointer in the SWISS-PROT Database (Release 15, August 1990, 16241 entries). Because they are structurally related to α-amylases (Hofmann *et al.*, 1989; Farber & Petsko, 1990), cyclodextrin glucanotransferases (EC 2.4.1.19) were also added. After removal of variants, too short and too partial sequences and influenza-virus haemaglutinin/neuraminidases, 70 sequences of sequence fragments not present in this release of the database were manually added to yield a total of 301 entries. The sequences of multiple-domain enzymes were edited (SEQIN, PC/Gene, Genofit, Switzerland) to isolate their catalytic domain(s) in order to facilitate sequence comparison. Comparisons involved DIAG [a laboratory-written dot-matrix program similar to DIAGON (Staden, 1982)] or Hydrophobic Cluster Analysis (Gaboriaud *et al.*, 1987; Lemesle-Varloot *et al.*, 1990) or FSTPSCAN (PC/Gene, Genofit, Switzerland) or a combination of these methods. When necessary, sequence alignments were performed with either MULTALIN (Corpet, 1988), MSA (Lipman *et al.*, 1989) or PALIGN (PC/Gene, Genofit, Switzerland). In the case of low sequence identity (< 25%), significance of sequence similarity was assessed with PCOMPARE (Needleman & Wunsch, 1970) on protein domains of meaningful length (at least 100 residues), sometimes after editing the sequences to eliminate

Abbreviations used: AAMY, α-amylase; AGAL, α-galactosidase; AFUC, α-fucosidase; AGAR, agarase; AGLU, α-glucosidase; AMAN, α-mannosidase; AMG, amyloglucosidase; APU, amylase/pullulanase; ARAF, α-arabinofuranosidase; BAMY, β-amylase; BGAL, β-galactosidase; BGLR, β-glucuronidase; BGLU, β-glucosidase; BMAN, β-mannanase; BXYL, β-xylosidase; CBH, cellobiohydrolase; CDGT, cyclodextrin glucanotransferase; CHI, chitinase; EBGAL, evolved β-galactosidase; EG, endoglucanase; endoNAG, endo-*N*-acetyl-β-glucosaminidase; G4-AMY, maltotetraose-forming amylase; G6-AMY, maltohexaose-forming amylase; GDX, glucodextranase; GLRB, glucosylceramidase; HYAL, hyaluronidase; IAMY, isoamylase; INV, invertase; LAM, laminarinase; LEV, levanase; LIC, lichenase; LPH, lactase/phlorizin hydrolase; NAAGAL, *N*-acetyl-α-galactosaminidase; NABGLU, *N*-acetyl-β-glucosaminidase; NEUR, neuraminidase; NPUL, neopullulanase; PBGAL, 6-phospho-β-galactosidase; PBGLU, 6-phospho-β-glucosidase; PGLR, polygalacturonase; PUL, pullulanase; SGSP, spore-germination specific protein; SI, sucrase/isomaltase; TREH, trehalase; XLAM, exo-laminarinase; XPGLR, exo-polygalacturonase; XYN, xylanase; ORF, open reading frame.

Table 1. Sequence similarity-based classification of glycosyl hydrolases

Family	Enzyme	Source	EC number	Reference or SWISS-PROT entry[a]
1	BGAL	*Sulfolobus solfataricus*	3.2.1.23	BGAL$SULSO
	BGAL S	*Sulfolobus solfataricus*	3.2.1.23	Cubellis *et al.*, 1990
	BGLU	*Agrobacterium* sp.	3.2.1.21	BGLS$AGRSP
	BGLU A	*Bacillus polymyxa*	3.2.1.21	Gonzales-Candelas *et al.*, 1990
	BGLU B	*Bacillus polymyxa*	3.2.1.21	Gonzales-Candelas *et al.*, 1990
	BGLU A	*Caldocellum saccharolyticum*	3.2.1.21	BGLS$CALSA
	PBGAL	*Lactobacillus casei*	3.2.1.85	LACG$LACCA
	PBGAL	*Staphylococcus aureus*	3.2.1.85	LACG$STAAU
	PBGAL	*Streptococcus lactis*	3.2.1.85	LACG$LACLA
	PBGLU	*Escherichia coli*	3.2.1.86	BGLB$ECOLI
	LPH[b]	human	3.2.1.62/108	LPH$HUMAN
	LPH[b]	rabbit	3.2.1.62/108	LPH$RABIT
2	BGAL	*Clostridium acetobutylicum*	3.2.1.23	Hancock *et al.*, 1991
	BGAL Z	*Escherichia coli*	3.2.1.23	BGAL$ECOLI
	EBGAL	*Escherichia coli*	3.2.1.23	BGA2$ECOLI
	BGAL Z	*Klebsiella pneumoniae*	3.2.1.23	BGAL$KLEPN
	BGAL	*Lactobacillus bulgaricus*	3.2.1.23	BGAL$KLULA
	BGLR	*Escherichia coli*	3.2.1.31	BGLR$ECOLI
	BGLR	human	3.2.1.31	BGLR$HUMAN
	BGLR	mouse	3.2.1.31	BGLR$MOUSE
	BGLR	rat	3.2.1.31	BGLR$RAT
3	BGLU A[c]	*Butyrivibrio fibrisolvens*	3.2.1.21	BGLS$BUTFI
	BGLU B	*Clostridium thermocellum*	3.2.1.21	BGLS$CLOTM
	BGLU[c]	*Ruminococcus albus*	3.2.1.21	GBLS$RUMAL
	BGLU	*Schizophyllum commune*	3.2.1.21	Moranelli *et al.*, 1986
	BGLU 1	*Saccharomycopsis fibuligera*	3.2.1.21	Machida *et al.*, 1988
	BGLU 2	*Saccharomycopsis fibuligera*	3.2.1.21	Machida *et al.*, 1988
	BGLU	*Hansenula anomala*	3.2.1.21	BGLS$HANAN
	BGLU	*Kluyveromyces fragilis*	3.2.1.21	BGLS$KLUFR
4	PBGLU	*Escherichia coli*	3.2.1.86	CELF$ECOLI
	AGAL A	*Escherichia coli*	3.2.1.22	AGAL$ECOLI
5	EG	*Bacillus* sp. strain 1139	3.2.1.4	GUN$BACS1
	EG	*Bacillus* sp. strain KSM-635	3.2.1.4	Ozaki *et al.*, 1990
	EG A	*Bacillus* sp. strain N-4	3.2.1.4	GUN1$BACS4
	EG B	*Bacillus* sp. strain N-4	3.2.1.4	GUN2$BACS4
	EG C	*Bacillus* sp. strain N-4	3.2.1.4	Fukumori *et al.*, 1989
	EG B	*Bacillus lautus*	3.2.1.4	Joergensen & Hansen 1990
	EG	*Bacillus polymyxa*	3.2.1.4	Baird *et al.*, 1990b
	EG	*Bacillus subtilis*	3.2.1.4	GUN1$BACSU; GUN2$BACSU
	EG	*Bacillus subtilis* BSE616	3.2.1.4	Park *et al.*, 1991
	EG	*Bacteroides ruminicola*	3.2.1.4	Matsushita *et al.*, 1990
	EG 1	*Butyrivibrio fibrisolvens*	3.2.1.4	Berger *et al.* 1989
	EG/CBH[d]	*Caldocellum saccharolyticum*	3.2.1.4	CELB$CALSA
	EG	*Clostridium acetobutylicum*	3.2.1.4	GUN$CLOAB
	EG A	*Clostridium cellulolyticum*	3.2.1.4	Faure *et al.*, 1989
	EG B	*Clostridium thermocellum*	3.2.1.4	GUNB$CLOTM
	EG C	*Clostridium thermocellum*	3.2.1.4	GUNC$CLOTM
	EG E	*Clostridium thermocellum*	3.2.1.4	GUNE$CLOTM
	EG H[d]	*Clostridium thermocellum*	3.2.1.4	GUNH$CLOTM
	EG C307	*Clostridium* sp. F1	3.2.1.4	Sakka *et al.*, 1991

	EG Z	*Erwinia chrysanthemi*	3.2.1.4	GUNZ$ERWCH
	EG 3	*Fibrobacter succinogenes*	3.2.1.4	GUN3$FIBSU
	EG I	*Robillarda* sp. Y-20	3.2.1.4	Yoshigi *et al.*, 1990
	EG I	*Ruminococcus albus*	3.2.1.4	GUN1$RUMAL
	EG A	*Ruminococcus albus*	3.2.1.4	Poole *et al.*, 1990
	EG B	*Ruminococcus albus*	3.2.1.4	Poole *et al.*, 1990
	EG I[(e)]	*Schizophyllum commune*	3.2.1.4	Saloheimo *et al.*, 1988
	EG III	*Trichoderma reesei*	3.2.1.4	GUN3$TRIRE
	EG	*Xanthomonas campestris*	3.2.1.4	Gough *et al.*, 1990
	BMAN A	*Caldocellum saccharolyticum*	3.2.1.78	Lthi *et al.*, 1991
6	EG A	*Cellulomonas fimi*	3.2.1.4	GUNA$CELFI
	EG A	*Microbispora bispora*	3.2.1.4	Yablonsky *et al.*, 1989
	EG A	*Streptomyces* sp. (KSM-9)	3.2.1.4	GUN1$STRSP
	CBH II	*Trichoderma reesei*	3.2.1.91	GUX2$TRIRE
7	EG I	*Trichoderma reesei*	3.2.1.4	GUN1$TRIRE
	CBH I	*Humicola grisea*	3.2.1.91	GUX1$HUMGR
	CBH I	*Phanerochaete chrysosporium*	3.2.1.91	GUX1$PHACH
	CBH I	*Trichoderma reesei*	3.2.1.91	GUX1$TRIRE
	CBH I	*Trichoderma viride*	3.2.1.91	Cheng *et al.*, 1990
8	EG	*Cellulomonas uda*	3.2.1.4	Nakamura *et al.*, 1986
	EG A	*Clostridium thermocellum*	3.2.1.4	GUNA$CLOTM
	EG Y	*Erwinia chrysanthemi*	3.2.1.4	Giuseppi, 1988
	LIC	*Bacillus circulans* WL-12	3.2.1.73	Bueno *et al.*, 1990
9	EG 1	*Butyrivibrio fibrisolvens*	3.2.1.4	Berger *et al.*, 1990
	EG B	*Cellulomonas fimi*	3.2.1.4	Meinke *et al.*, 1991
	EG C	*Cellulomonas fimi*	3.2.1.4	GUNC$CELFI
	EG Z	*Clostridium stercorarium*	3.2.1.4	Jauris *et al.*, 1990
	EG D	*Clostridium thermocellum*	3.2.1.4	GUND$CLOTM
	SGSP	*Dictyostelium discoideum*	n.d.	Giorda *et al.*, 1990
	EG A	*Fibrobacter succinogenes*	3.2.1.4	Cavicchioli *et al.*, 1991
	EG	*Persea americana*	3.2.1.4	GUN$PERAE
	EG 1	*Persea americana*	3.2.1.4	Cass *et al.*, 1990
	EG 2	*Persea americana*	3.2.1.4	Cass *et al.*, 1990
	EG	*Phaseolus vulgaris*	3.2.1.4	Tucker & Milligan, 1991
	EG	*Pseudomonas fluorescens*	3.2.1.4	GUN$PSEFL
10	XYN A	*Bacillus* sp. strain C-125	3.2.1.8	XYNA$BACS5
	XYN A	*Butyrivibrio fibrisolvens*	3.2.1.8	Mannarelli *et al.*, 1990
	EG/CBH[(f)]	*Caldocellum saccharolyticum*	3.2.1.91	CELB$CALSA
	XYN A	*Caldocellum saccharolyticum*	3.2.1.8	Lüthi *et al.*, 1990
	ORF4	*Caldocellum saccharolyticum*	n.d.	Lüthi *et al.*, 1990
	CBH	*Cellulomonas fimi*	3.2.1.91	GUX$CELFI
	XYN Z	*Clostridium thermocellum*	3.2.1.8	XYNZ$CLOTM
	XYN	*Cryptococcus albidus*	3.2.1.8	XYNA$CRYAL
	XYN A	*Pseudomonas fluorescens*	3.2.1.8	XYNA$PSEFL
	XYN B	*Pseudomonas fluorescens*	3.2.1.8	Kellett *et al.*, 1990
	XYN	*Thermoascus aurantiacus*	3.2.1.8	Srinivasa *et al.*, 1990
11	XYN A	*Bacillus circulans*	3.2.1.8	XYNA$BACCI
	XYN A	*Bacillus pumilus*	3.2.1.8	XYNA$BACPU
	XYN A	*Bacillus subtilis*	3.2.1.8	Paice *et al.*, 1986
	XYN B	*Clostridium acetobutylicum*	3.2.1.8	XYNA$CLOAC
12	EG	*Aspergillus aculeatus*	3.2.1.4	Ooi *et al.*, 1990
	EG S	*Erwinia carotovora*	3.2.1.4	GUNS$ERWCA

13	AAMY	*Aspergillus oryzae*	3.2.1.1	AMYA$ASPOR
	AAMY	*Bacillus amyloliquefaciens*	3.2.1.1	AMY$BACAM
	AAMY	*Bacillus circulans*	3.2.1.1	AMY$BACCI
	AAMY	*Bacillus licheniformis*	3.2.1.1	AMY$BACLI
	AAMY	*Bacillus stearothermophilus*	3.2.1.1	AMY$BACST
	AAMY	*Bacillus subtilis*	3.2.1.1	AMY$BACSU
	AAMY	*Bacillus* sp.	3.2.1.1	AMYR$BACSP
	AAMY	*Clostridium thermosulfurogenes*	3.2.1.1	Bahl *et al.*, 1991
	AAMY	*Dictyoglomus thermophilum*	3.2.1.1	AMY1$DICTH
	AAMY	*Drosophila melanogaster*	3.2.1.1	AMY$DROME
	AAMY	*Hordeum vulgare*	3.2.1.1	AMY1$HORVU
	AAMY	human (pancreatic & salivary)	3.2.1.1	AMYP$HUMAN; AMYS$HUMAN
	AAMY	mouse (pancreatic & salivary)	3.2.1.1	AMYP$MOUSE;AMYS$MOUSE
	AAMY	pig (pancreatic)	3.2.1.1	AMYP$PIG
	AAMY	rat (pancreatic)	3.2.1.1	AMYP$RAT
	AAMY	rice	3.2.1.1	AMY$ORYSA
	AAMY	*Saccharomycopsis fibuligera*	3.2.1.1	Itoh *et al.*, 1987
	AAMY	*Streptococcus hygroscopicus*	3.2.1.1	AMY$STRHY
	AAMY	*Streptomyces limosus*	3.2.1.1	AMY$STRLM
	AAMY	*Tribolium castaneum*	3.2.1.1	AMY$TRICA
	AAMY	*Triticum aestivum*	3.2.1.1	AMY3$WHEAT
	AAMY	*Vigna mungo*	3.2.1.1	AMYA$VIGMU
	AGLU	*Aedes aegypti*	3.2.1.20	MALT$AEDAE
	AGLU	*Drosophila melanogaster*	3.2.1.20	MAL1$DROME to MAL3$DROME
	AGLU	*Saccharomyces carlsbergensis*	3.2.1.20	MALT$SACCA
	APU	*Clostridium thermohydrosulfuricum*	3.2.1.1/41	APU$CLOTF
	PUL	*Bacillus stearothermophilus*	3.2.1.41	Kuriki *et al.*, 1990
	PUL	*Klebsiella aerogene*	3.2.1.41	PULA$KLEAE
	PUL A	*Klebsiella pneumoniae*	3.2.1.41	PULA$KLEPN
	CDGT	*Bacillus licheniformis*	2.4.1.19	CDGT$BACLI
	CDGT	*Bacillus macerans*	2.4.1.19	CDGT$BACMA
	CDGT	*Bacillus* sp. 1011	2.4.1.19	CDGT$BACS1
	CDGT	*Bacillus* sp. 38-2	2.4.1.19	CDGT$BACS3
	CDGT	*Klebsiella pneumoniae*	2.4.1.19	CDGT$KLEPN
	G4-AMY	*Pseudomonas saccharophila*	3.2.1.60	Zhou *et al.*, 1989
	G4-AMY	*Pseudomonas stutzeri*	3.2.1.60	AMTT$PSEST
	G6-AMY	*Bacillus* sp. 707	3.2.1. -	Tsukamoto *et al.*, 1988
	IAMY	*Pseudomonas amylodermosa*	3.2.1.68	ISOA$PSEAY
	GDX	*Streptococcus mutans*	3.2.1.70	Russell & Ferretti, 1990
	NPUL	*Bacillus stearothermophilus*	3.2.1. -	Kuriki & Imanaka, 1989
14	BAMY	*Bacillus circulans*	3.2.1.2	AMYB$BACCI
	BAMY	*Bacillus polymyxa*	3.2.1.2	Kawazu *et al.*, 1987; Rhodes *et al.*, 1987
	BAMY	*Glycine max*	3.2.1.2	AMYB$SOYBN
	BAMY	*Hordeum vulgare*	3.2.1.2	AMYB$HORVU
	BAMY	*Ipomoea batatas*	3.2.1.2	AMYB$IPOBA
15	AMG	*Aspergillus niger*	3.2.1.3	AMYG$ASPNG
	AMG	*Rhizopus oryzae*	3.2.1.3	AMYG$RHIOR
	AMG	*Saccharomyces cerevisiae*	3.2.1.3	AMYG$YEAST
	AMG	*Saccharomyces distaticus*	3.2.1.3	AMYG$SACDI
	AMG	*Saccharomycopsis fibuligera*	3.2.1.3	AMYG$SACFI
16	LIC A	*Bacillus amyloliquefaciens*	3.2.1.73	GUB$BACAM

	LIC M	*Bacillus macerans*	3.2.1.73	Borriss *et al.*, 1990
	LIC	*Bacillus subtilis*	3.2.1.73	GUB$BACSU
	LIC	*Fibrobacter succinogenes*	3.2.1.73	Teather & Erfle, 1990
	LAM A	*Bacillus circulans*	3.2.1.39	Yahata *et al.*, 1990
17	LAM	*Hordeum vulgare*	3.2.1.39	E13B$HORVU
	LAM	*Nicotiana tabacum*	3.2.1.39	E13B$TOBAC
	XLAM	*Saccharomyces cerevisiae*	3.2.1.58	BGL2$YEAST
	LIC	*Hordeum vulgare*	3.2.1.73	GUB2$HORVU
	LIC	*Nicotiana plumbaginifolia*	3.2.1.73	GUB$NICPL
18	CHI A	*Arabidopsis thaliana*	3.2.1.14	Samac *et al.*, 1990
	CHI A	*Bacillus circulans*	3.2.1.14	Watanabe *et al.*, 1990
	CHI	*Cucumis sativis*	3.2.1.14	CHIT$CUCSA
	CHI A	*Serratia marcescens*	3.2.1.14	CHIA$SERMA
	CHI B	*Serratia marcescens*	3.2.1.14	CHIB$SERMA
	CHI	*Streptomyces erythaeus*	3.2.1.14	CHIT$SACER
	CHI[(g)]	*Hevea brasiliensis*	3.2.1.14/17	Rozeboom *et al.*, 1990
	CHI	*Parthenocissus quinquifolia*	3.2.1.14/17	Bernasconi *et al.*, 1987
	endoNAG H	*Streptomyces plicatus*	3.2.1.96	EBAG$STRPL
19	CHI B	*Arabidopsis thaliana*	3.2.1.14	Samac *et al.*, 1990
	CHI	*Coix lachryma-jobi*	3.2.1.14	IAMY$COILA
	CHI	*Hordeum vulgare*	3.2.1.14	CHIT$HORVU
	CHI	*Nicotiana tabacum*	3.2.1.14	CHIP$TOBAC; CHIQ$TOBAC CHIT$TOBAC
	CHI	*Phaseolus vulgaris*	3.2.1.14	CHIT$PHAVU
	CHI 6	*Populus* sp.	3.2.1.14	CHI6$POPSP
	CHI 8	*Populus* sp.	3.2.1.14	CHI8$POPSP
	CHI	*Solanum tuberosum*	3.2.1.14	CHIT$SOLTU
20	NABGLU	*Vibrio harvey*	3.2.1.30	CHB$VIBHA
	NABGLU	*Dictyostelium discoideum* (α-chain)	3.2.1.52	HEXA$DICDI
	NABGLU	human (α-chain)	3.2.1.52	HEXA$HUMAN
	NABGLU	human (β-chain)	3.2.1.52	HEXB$HUMAN
21	LYS B	example: *Bacillus subtilis*	3.2.1.17	LYB$BACSU
22	LYS C	example: chicken	3.2.1.17	LYC$CHICK
23	LYS G	example: goose	3.2.1.17	LYG$ANSAN
24	LYS	example: bacteriophage T4	3.2.1.17	LYCV$BPT4
25	LYS CH	*Chalaropsis* sp.	3.2.1.17	LYCH$CHASP
	LYS CH	*Streptomyces globisporus*	3.2.1.17	Lichenstein *et al.*, 1990
26	BMAN	*Bacillus* sp.	3.2.1.78	MANB$BACSP
	EG H[(l)]	*Clostridium thermocellum*	n.d.	Yagüe *et al.*, 1990
27	AGAL	*Cyamopsis tetragonoloba*	3.2.1.22	AGAL$CYATE
	AGAL	human	3.2.1.22	AGAL$HUMAN
	AGAL	*Saccharomyces cerevisiae*	3.2.1.22	AGAL$YEAST
	NAAGAL	human	3.2.1.49	NAGA$HUMAN
28	PGLR	*Aspergillus niger*	3.2.1.15	Bussink *et al.*, 1990; Ruttkowski *et al.*, 1990
	PGLR	*Erwinia carotovora*	3.2.1.15	Saarilahti *et al.*, 1990

	PGLR	*Lycopersicon esculentum*	3.2.1.15	PGLR$LYCES
	PGLR	*Pseudomonas solanacearum*	3.2.1.15	Huang & Schell, 1990
	XPGLR	*Erwinia chrysanthemi*	3.2.1.82	PEHX$ERWCH
29	AFUC	*Dictyostelium discoideum*	3.2.1.51	FUCO$DICDI
	AFUC	human	3.2.1.51	FUCO$HUMAN
	AFUC	rat	3.2.1.51	FUCO$RAT
30	GLRB	human	3.2.1.45	GLRB$HUMAN
	GLRB	mouse	3.2.1.45	GLRB$MOUSE
31	AGLU	human	3.2.1.20	LYAG$HUMAN
	SI[h]	human	3.2.1.48/10	SUIS$HUMAN
	SI[h]	rabbit	2.2.1.48/10	SUIS$RABIT
32	INV	*Bacillus subtilis*	3.2.1.26	SCRB$BACSU
	INV	*Saccharomyces cerevisiae*	3.2.1.26	INV1$YEAST to INV7$YEAST
	INV	*Salmonella typhimurium*	3.2.1.26	Gunasekaran *et al.*, 1990
	INV	*Schwanniomyces occidentalis*	3.2.1.26	Klein *et al.*, 1989
	INV	*Streptococcus mutans*	3.2.1.26	SCRB$STRMU
	INV	*Vibrio alginolyticus*	3.2.1.26	SCRB$VIBAN
	INV	*Zymomonas mobilis*	3.2.1.26	Gunasekaran *et al.*, 1990
	LEV	*Bacillus subtilis*	3.2.1.65	SACC$BACSU
33	NEUR	example: *Clostridium perfringens*	3.2.1.18	NANH$CLOPE
34	NEUR	example: influenza virus	3.2.1.18	NRAM$INABA
35[i]	BGAL	human	3.2.1.23	BGAL$HUMAN
	BGAL	mouse	3.2.1.23	Nanba & Suzuki, 1990

[a] References are given only for the sequences not found in the SWISS-PROT databank (Release 15, August 1990).

[b] Pre-pro-LPH has an internal double duplication (Mantei *et al.*, 1988). Each of the four subdomains (I–IV) displays sequence similarity with this family. Mature LPH comprises only domains III and IV (domains I and II are removed by proteolytic cleavage).

[c] These proteins display an inversion of their two-domain structure with respect to the other members of the family (Lemesle-Varloot *et al.*, 1990).

[d] *C*-Terminal part of the protein.

[e] Homology reported with *T. reesei* EG III (Saloheimo *et al.*, 1988), no sequence data.

[f] *N*-Terminal part of the protein.

[g] Homology reported with *C. sativis* and *P. quinquifolia* chitinases (Rozeboom *et al.*, 1990), no sequence data.

[h] SI from human and rabbit are composed of two homologous domains each of which is similar to human AGLU.

[i] A weak nucleotide sequence similarity was reported with *E. coli* β-galactosidase (Nanba & Suzuki, 1990). The amino acid sequences similarity is so weak (not detectable) that the divergence of the polypeptide backbones is expected to be so high that it is safer to consider them as belonging to two different families. This does not rule out the possibility of a common ancestor.

highly variable loops as described by Henrissat *et al.* (1990). Scores greater than 3 standard deviations were considered significant. A family was defined when (i) at least two sequences displayed significant amino acid similarity and (ii) no significant similarity was found with other families.

RESULTS

A total of 301 sequences were compared. Of these, 291 sequences corresponding to 30 EC entries could be classified into 35 families (Table 1). Only ten sequences (< 5% of the sample) could not be assigned to any family (Table 2). With the sequences available so far, 18 families were found to be monospecific (containing only one EC number) and 17 were found to be polyspecific (containing at least two EC numbers). It is possible that some of the monospecific families may turn out to be polyspecific when additional sequence data are available.

DISCUSSION

Structural aspects

Since significant sequence similarity is a strong indication of folding similarities (Chothia & Lesk, 1986), the members of one family most likely share the same folding characteristics, thereby enabling homology modelling if the three-dimensional structure of one member is known. The present classification may also help in choosing the enzyme to crystallize: there is more novelty in the resolution of a new three-dimensional structure than of a three-dimensional structure that could be predicted.

The finding of 35 families is only a consequence of the sample of sequences available to date. It is therefore evident that a number of other families remain to be discovered. The present classification is complementary to that of IUB, and could be simply implemented to provide a better description of glycosyl hydrolases. A possibility could be the addition of another digit to

Table 2. Enzymes that could not be classified

Enzyme	Source	EC no.	Reference or SWISS-PROT entry[a]
EG B	*Pseudomonas fluorescens*	3.2.1.4	Gilbert *et al.* (1990)
EG A	*Ruminococcus flavefaciens*	3.2.1.4	GUX$RUMFL
AGAL[b]	*Escherichia coli*	3.2.1.22	RAFA$ECOLI
AMAN	*Saccharomyces cerevisiae*	3.2.1.24	Yoshihisa & Anraku (1989)
TREH	*Escherichia coli*	3.2.1.28	TREA$ECOLI
BXYL	*Bacillus pumilus*	3.2.1.37	XYNB$BACPU
ARAF C	*Pseudomonas fluorescens*	3.2.1.55	Kellett *et al.* (1990)
AGAR	*Pseudomonas atlantica*	3.2.1.81	AGAR$PSEAT
AGAR	*Streptomyces coelicolor*	3.2.1.81	AGAR$STRCO
HYAL	Bacteriophage H4489A	3.2.1.36	HYAL$BPH4

[a] References are given only for the sequences not found in the SWISS-PROT databank (Release 15, August 1990).

[b] In this sequence, a region spanning approx. 55 residues shows significant similarity to the α-galactosidases in family 27. The size of this region, however, is too small to allow the confident assignment of this sequence in that family.

the present EC number: EC 3.2.1.X.Y where X is the present fourth digit of the EC classification and where Y would represent the sequence-similarity-based family.

It is likely that these concepts are general and that the introduction of structural data to the description of other types of enzymes would also prove useful (Farber & Petsko, 1990).

Molecular mechanism

All glycosyl hydrolases are thought to act by a general acid catalysis mechanism in which two amino acid residues participate in a single-displacement or double-displacement reaction resulting in inversion or retention of configuration respectively at the anomeric carbon atom of the hydrolysed glycoside (Sinnott, 1990). An important outcome of a detailed primary-structure analysis of glycosyl hydrolases is the ability to locate potential active-site residues on the basis of the identification of appropriate invariant amino acid residues (Zvelebil & Sternberg, 1988; Henrissat *et al.*, 1989; Henrissat, 1990). In the case of a family of cellulases (family 5), the number of invariant potentially catalytic amino acid residues was so low that it constituted a direct active-site-residue prediction (Henrissat *et al.*, 1989). Site-directed mutagenesis of the predicted residues supports the prediction (Baird *et al.*, 1990*a*; Py *et al.*, 1991). Examination of invariant potentially catalytic residues is possible for each of the families presented in Table 1. When sequence identity is too high and/or when the number of related sequences is too low, the list of potentially catalytic residues is too long to constitute a useful prediction as such. In a number of cases, however, sequence divergence is high enough for site-directed mutagenesis experiments. Also, if a catalytic residue of a member of one family is known, one can immediately trace the presence of this residue in the other members. For example, Glu-358 of the *Agrobacterium* β-glucosidase (family 1), a retaining enzyme, has been shown to be directly involved in glycosidic bond cleavage by acting as a nucleophile (Withers *et al.*, 1990). This residue is totally invariant in family 1 except LPH domains I and II, which are removed from pre-pro-LPH by proteolytic processing *in vivo*. Sequence comparison predicts that Glu-1273 and Glu-1749 of human LPH domains III and IV are directly involved in catalysis whereas domains I and II are predicted to be non-catalytic.

Evolutionary aspects

The existence of a number of polyspecific families indicates that the acquisition of new specificities by glycosyl hydrolases is a common evolutionary event. The divergence of glycosyl hydrolases to acquire new specificities is not unexpected given the stereochemical resemblance between some of their substrates. This divergence could find its origin in the emergence of new carbohydrate metabolites during evolution. Such a scheme leads to two contradictory evolutionary pressures on one enzyme: (i) optimization of the catalytic efficiency for the original substrate and (ii) divergence to acquire a new specificity. Gene duplication, frequently observed among glycosyl hydrolases, could be a way to cope with such events. One copy would continue dealing with the original substrate, and duplicate copies could constitute templates for constructing enzymes with a specificity directed to a new, but stereochemically similar, substrate. Because they occur in the three major kingdoms, glycosyl hydrolases might thus serve as useful markers, not only for studying evolution of enzyme specificity, but also for studying the evolution of carbohydrate metabolism: if it is not unexpected that enzymes acting on stereochemically similar substrates (for example glucosides and galactosides) can be structurally related, why do enzymes that act on the same substrate sometimes belong to distinct families with no interfamily similarity? An illustration is with the origins of human LPH: why was LPH built after archaebacterial β-galactosidase (family 1) and not after *E. coli* β-galactosidase (LacZ) (family 2), since the latter is present in the human genome under the form of β-glucuronidase (family 2)?

Studying the evolution of the enzymes involved in carbohydrate metabolism will eventually enable evolutionarists to trace the evolution of carbohydrates, and perhaps tell us, for example, which of glucose or galactose appeared first on Earth.

Note added in proof (received 2 September 1991)

Since this paper was submitted, 21 new sequences of glycosyl hydrolases and one of a glycosyltransferase have become available to the author (list available on demand). Of these sequences 18 could be classified in the families defined in this paper. The α-galactosidase of *Streptococcus mutans* (Aduse-Opoku *et al.*, 1991) was found to show sequence similarity to the previously unclassified *rafA* α-galactosidase of *Escherichia coli* and thus defines a 36th family. Three sequences could not be classified (i.e. did not exhibit significant similarity of sequence to any of the families or to the unclassified sequences).

This work was supported by a CM2AO research grant from Organibio. I thank F. Barras for stimulating discussions and for critically reading this manuscript.

REFERENCES

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

Auricchio, S., Rubino, A., Landolt, M., Semenza, G. & Prader, A. (1963) Lancet **ii**, 324–326

Bahl, H., Buchhardt, G., Spreinat, A., Haeckel, K., Wienecke, A., Schmidt, B. & Antranikian, G. (1991) Appl. Environ. Microbiol. **57**, 1554–1559

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

Baird, S. D., Johnson, D. A. & Seligy, V. L. (1990*b*) J. Bacteriol. **172**, 1576–1586

Berger, E., Jones, W. A., Jones, D. T. & Woods, D. R. (1989) Mol. Gen. Genet. **219**, 193–198

Berger, E., Jones, W. A., Jones, D. T. & Woods, D. R. (1990) Mol. Gen. Genet. **223**, 310–318

Bernasconi, P., Locher, R., Pilet, P. E., Jollès, J. & Jollès, P. (1987) Biochim. Biophys. Acta **915**, 254–260

Borriss, R., Buettner, K. & Maentsaelae, P. (1990) Mol. Gen. Genet. **222**, 278–283

Bueno, A., Vazquez de Aldana, C. R., Correa, J. & del Rey, F. (1990) Nucleic Acids Res. **18**, 4248

Bussink, H. J. D., Kester, H. C. M. & Visser, J. (1990) FEBS Lett. **273**, 127–130
Cass, L. G., Kirven, K. A. & Christoffersen, R. E. (1990) Mol. Gen. Genet. **223**, 76–86
Cavicchioli, R., East, P. D. & Watson, K. (1991) J. Bacteriol. **173**, 3265–3268
Cheng, C., Tsukagoshi, N. & Udaka, S. (1990) Nucleic Acids Res. **18**, 5559
Chothia, C. & Lesk, A. M. (1986) EMBO J. **5**, 823–826
Corpet, F. (1988) Nucleic Acids Res. **16**, 10881–10890
Cubellis, M. V., Rozzo, C., Montecucchi, P. & Rossi, M. (1990) Gene **94**, 89–94
Enzyme Nomenclature (1984) 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
Farber, G. K. & Petsko, G. A. (1990) Trends Biochem. Sci. **15**, 228–234
Faure, E., Belaich, A., Bagnara, C., Gaudin, C. & Belaich, J. P. (1989) Gene **84**, 39–46
Fukumori, F., Kudo, T., Sashihara, N., Nagata, Y., Ito, K. & Horikoshi, K. (1989) Gene **76**, 289–298
Gaboriaud, C., Bissery, V., Benchetrit, T. & Mornon, J. P. (1987) FEBS Lett. **224**, 149–155
Gilbert, H. J., Hall, J., Hazelwood, G. P. & Ferreira, L. M. A. (1990) Mol. Microbiol. **4**, 759–767
Giorda, R., Ohmachi, T., Shaw, D. R. & Ennis, H. L. (1990) Biochemistry **29**, 7264–7269
Giuseppi, A. (1988) Ph.D. Thesis, University of Aix–Marseille I
Gonzales-Candelas, L., Ramon, D. & Polaina, J. (1990) Gene **95**, 31–38
Gough, C. L., Dow, J. M., Keen, J., Henrissat, B. & Daniels, M. J. (1990) Gene **89**, 53–59
Gunasekaran, P., Karunakaran, T., Cami, B., Mukundan, A. G., Preziosi, L. & Baratti, J. (1990) J. Bacteriol. **172**, 6727–6735
Hancock, K. R., Rockman, E., Young, C. A., Pearce, L., Maddox, I. S. & Scott, D. B. (1991) J. Bacteriol. **173**, 3084–3095
Henrissat, B. (1990) Protein Sequences Data Anal. **3**, 523–526
Henrissat, B. (1991) Protein Sequences Data Anal. **4**, 61–62
Henrissat, B., Claeyssens, M., Tomme, P., Lemesle, L. & Mornon, J. P. (1989) Gene **81**, 83–95
Henrissat, B., Saloheimo, M., Lavaitte, S. & Knowles, J. K. C. (1990) Proteins **8**, 251–257
Hofmann, B. E., Bender, H. & Schulz, G. E. (1989) J. Mol. Biol. **209**, 793–800
Huang, J. & Schell, M. A. (1990) J. Bacteriol. **172**, 3879–3887
Itoh, T., Yamashita, I. & Fukui, S. (1987) FEBS Lett. **219**, 339–342
Jauris, S., Rücknagel, K. P., Schwarz, W. H., Kratzsch, P., Bronnenmeier, K. & Staudenbauer, W. L. (1990) Mol. Gen. Genet. **223**, 258–267
Joergensen, P. L. & Hansen, C. K. (1990) Gene **93**, 55–60
Kawazu, T., Nakanishi, Y., Uozumi, N., Sasaki, T., Yamagata, H., Tsukagoshi, N. & Udaka, S. (1987) J. Bacteriol. **169**, 1564–1570
Kellett, L. E., Poole, D. M., Ferreira, L. M. A., Durrant, A. J., Hazelwood, G. P. & Gilbert, H. J. (1990) Biochem. J. **272**, 369–376
Klein, R. D., Poorman, R. A., Favreau, M. A., Shea, M. H., Hatzenbuhler, N. T. & Nulf, S. C. (1989) Curr. Genet. **16**, 145–152
Kuriki, T. & Imanaka, T. (1989) J. Gen. Microbiol. **135**, 1521–1528
Kuriki, T., Park, J. H. & Imanaka, T. (1990) J. Ferment, Bioeng. **69**, 204–210
Lemesle-Varloot, L., Henrissat, B., Gaboriaud, C., Bissery, V., Morgat, A. & Mornon, J. P. (1990) Biochimie **72**, 555–574
Lichenstein, H. S., Hastings, A. E., Langley, K. E., Mendiaz, E. A., Rohde, M. F., Elmore, R. & Zukowski, M. M. (1990) Gene **88**, 81–86
Lipman, D. J., Altschul, S. F. & Kececioglu, J. D. (1989) Proc. Natl. Acad. Sci. U.S.A. **86**, 4412–4415
Lüthi, E., Love, D. R., McAnulty, J., Wallace, C., Caughey, P. A., Saul, D. & Bergquist, P. L. (1990) Appl. Environ. Microbiol. **56**, 1017–1024
Lüthi, E., Jasmat, N. B., Grayling, R. A., Love, D. R. & Bergquist, P. L. (1991) Appl. Environ. Microbiol. **57**, 694–700
MacGregor, E. A. & Svensson, B. (1989) Biochem. J. **259**, 145–152
Machida, M., Ohtsuki, I., Fukui, S. & Yamashita, I. (1988) Appl. Environ. Microbiol. **54**, 3147–3155
Mannarelli, B. M., Evans, S. & Lee, D. (1990) J. Bacteriol. **172**, 4247–4254
Mantei, N., Villa, M., Enzler, T., Wacker, H., Boll, W., James, P., Hunziker, W. & Semenza, G. (1988) EMBO J. **7**, 2705–2713
Matsushita, O., Russell, J. B. & Wilson, D. B. (1990) J. Bacteriol. **172**, 3620–3630
Meinke, A., Braun, C., Gilkes, N. R., Kilburn, D. G., Miller, R. C. & Warren, R. A. J. (1991) J. Bacteriol. **173**, 308–314
Moranelli, F., Barbier, J. R., Dove, M. J., MacKay, R. M., Seligy, V. L., Yaguchi, M. & Willick, G. E. (1986) Biochem. Int. **12**, 905–912
Nakamura, K., Misawa, N. & Kitamura, K. (1986) J. Biotechnol. **4**, 247–254
Nanba, E. & Suzuki, K. (1990) Biochem. Biophys. Res. Commun. **173**, 141–148
Needleman, S. B. & Wunsch, C. D. (1970) J. Mol. Biol. **48**, 443–453
Neufeld, E. F., Lim, T. W. & Shapiro, L. J. (1975) Annu. Rev. Biochem. **44**, 357–376
Ooi, T., Shinmyo, A., Okada, H., Murao, S., Kawaguchi, T. & Arai, M. (1990) Nucleic Acids Res. **18**, 5884
Ozaki, K., Shikata, S., Kawai, S., Ito, S. & Okamoto, K. (1990) J. Gen. Microbiol. **136**, 1327–1334
Paice, M. G., Bourbonnais, R., Desrochers, M., Jurasek, L. & Yaguchi, M. (1986) Arch. Microbiol. **144**, 201–206
Park, S. H., Kim, H. K. & Pack, M. Y. (1991) Agric. Biol. Chem. **55**, 441–448
Poole, D. M., Hazlewood, G. P., Laurie, J. I., Barker, P. J. & Gilbert, H. J. (1990) Mol. Gen. Genet. **223**, 217–223
Py, B., Bortoli-German, I., Haiech, J., Chippaux, M. & Barras, F. (1991) Protein Eng. **4**, 325–333
Raimbaud, E., Buléon, A., Pérez, S. & Henrissat, B. (1989) Int. J. Biol. Macromol. **11**, 217–225
Rhodes, C., Strasser, J. & Friedberg, G. (1987) Nucleic Acids Res. **15**, 3934
Rozeboom, H. J., Budiani, A., Beintema, J. J. & Dijkstra, B. W. (1990) J. Mol. Biol. **212**, 441–443
Russell, R. R. B. & Ferretti, J. J. (1990) J. Gen. Microbiol. **136**, 803–810
Ruttkowski, E., Labitzke, R., Khanh, N. Q., Löffler, F., Gottschalk, M. & Jany, K. D. (1990) Biochim. Biophys. Acta **1087**, 104–106
Sarrilahti, H. T., Heino, P., Pakkanen, R., Kalkkinen, N., Palva, I. & Palva, E. T. (1990) Mol. Microbiol. **4**, 1037–1044
Sakka, K., Shimanuki, T. & Shimada, K. (1991) Agric. Biol. Chem. **55**, 347–350
Saloheimo, M., Lehtovaara, P., Penttilä, M., Teeri, T. T., Stahlberg, J., Johansson, G., Petterson, G., Claeyssens, M. & Knowles, J. (1988) Gene **63**, 11–21
Samac, D. A., Hironaka, C. M., Yallaly, P. E. & Shah, D. M. (1990) Plant Physiol. **93**, 907–914
Sinnott, M. L. (1990) Chem. Rev. **90**, 1171–1202
Srinivasa, B. R., Vithayathil, P. J., Roy, R. P. & Swaminathan, K. R. (1990) J. Protein Chem. **9**, 337–338
Staden, R. (1982) Nucleic Acids Res. **10**, 2951–2961
Svensson, B. (1988) FEBS Lett. **230**, 72–76
Teather, R. M. & Erfle, J. D. (1990) J. Bacteriol. **172**, 3837–3841
Tsukamoto, A., Kimura, K., Ishii, Y., Takano, T. & Yamane, K. (1988) Biochem. Biophys. Res. Commun. **151**, 25–31
Tucker, M. L. & Milligan, S. B. (1991) Plant Physiol. **95**, 928–933
Watanabe, T., Suzuki, K., Oyanagi, W., Ohnishi, K. & Tanaka, H. (1990) J. Biol. Chem. **265**, 15659–15665
Withers, S. G., Warren, R. A. J., Street, I. P., Rupitz, K., Kempton, J. B. & Aebersold, R. (1990) J. Am. Chem. Soc. **112**, 5887–5889
Yablonsky, M. D., Elliston, K. O. & Eveleigh, D. E. (1989) in Production, Characterization and Application of Cellulose, Hemicellulose and Lignin Enzyme Degrading Systems (Coughlan, M., ed.), pp. 73–83, Elsevier, London
Yagüe, E., Béguin, P. & Aubert, J. P. (1990) Gene **89**, 61–67
Yahata, N., Watanabe, T., Nakamura, Y., Yamamoto, Y., Kamimiya, S. & Tanaka, H. (1990) Gene **86**, 113–117
Yoshigi, N., Taniguchi, H. & Sasaki, T. (1990) J. Biochem. (Tokyo) **108**, 338–392
Yoshihisa, T. & Anraku, Y. (1989) Biochem. Biophys. Res. Commun. **163**, 908–915
Zhou, J., Baba, T., Takano, T., Kobayashi, S. & Arai, Y. (1989) FEBS Lett. **255**, 37–41
Zvelebil, M. J. J. & Sternberg, M. J. E. (1988) Protein Eng. **2**, 127–138

Received 2 April 1991/23 May 1991; accepted 30 May 1991