Glycoside Hydrolases and Glycosyltransferases. Families, Modules, and Implications for Genomics[1]

Bernard Henrissat* and Gideon J. Davies

Architecture et Fonction des Macromolécules Biologiques, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 6098, 31 Chemin Joseph Aiguier, 13402 Marseille cedex 20, France (B.H.); and Department of Chemistry, Structural Biology Laboratory, University of York, Heslington, York YO10 5DD, United Kingdom (G.J.D.)

One of the first insights into the modular nature of carbohydrate-active enzymes was provided by the dissection of a plant cell wall-degrading enzyme into two functional modules (van Tilbeurgh et al., 1986). The general architecture deduced for this protein featured two independent globular modules: a cellulase catalytic domain, responsible for the hydrolysis reaction itself, and a cellulose-binding module, devoid of catalytic activity but promoting adsorption of the enzyme onto insoluble crystalline cellulose. Similar observations were made for other polysaccharide-degrading enzymes such as plant chitinases (Lucas et al., 1985; Shinshi et al., 1990; Lerner and Raikhel, 1992). In the early 1990s it was shown that this modular structure could be deduced from sequence examination alone (Gilkes et al., 1991). It is now clear that the two major classes of carbohydrate-active enzymes, glycoside hydrolases and glycosyltransferases, frequently display a modular structure (Figs. 1 and 2). In the genomic era, this modularity is of particular importance for correct open reading frame (ORF) annotation and functional prediction.

A classification system of the catalytic domains of glycoside hydrolases and transglycosylases into families based on amino acid similarities was introduced a decade ago (Henrissat, 1991) and updated regularly (Henrissat and Bairoch, 1993, 1996). In marked contrast to the International Union of Biochemistry and Molecular Biology enzyme nomenclature, the new classification scheme was designed to integrate both structural and mechanistic features of these enzymes. It is striking that the system based on sequence similarities (hence also reflecting similar structural features) often grouped enzymes of different substrate specificity in a single "poly-specific" family. This classification system was later extended to glycosyltransferases (Campbell et al., 1997). Over the years the number of families of glycoside hydrolases and glycosyltransferases has grown steadily and currently there are 82 and 47 families, respectively. These families, as well as others featuring polysaccharide lyases and carbohydrate esterases, are available on the continuously updated carbohydrate active enzymes (CAZy) web server at http://afmb.cnrs-mrs.fr/~pedro/CAZY/db.html.

It soon became clear that ancillary, non-catalytic modules were frequently borne by polysaccharide-degrading enzymes (Svensson et al., 1989; Gilkes et al., 1991; Raikhel et al., 1993). The first function reported for these modules was the binding of insoluble polysaccharides such as cellulose, chitin, and starch. Warren and his colleagues showed that, like the families of catalytic domains, the polysaccharide-binding modules also formed a number of distinct families (Coutinho et al., 1993; Tomme et al., 1995; Warren, 1996). Today, 24 families of carbohydrate-binding modules are known and characterized, but the role of many families of ancillary modules that could be detected by careful sequence comparisons remains unknown (Coutinho and Henrissat, 1999a). We have already detected over 60 such modules of unknown function (termed "X" modules) by systematic sequence analysis of a number of carbohydrate-active enzymes (P.M. Coutinho and B. Henrissat, unpublished data). A further complication is that with the present deluge of sequence data, modular enzymes with more than one catalytic domain are discovered. Figures 1 and 2 show a few examples of modular glycoside hydrolases and glycosyltransferases, many of which have particular relevance to plant science.

The classifications of carbohydrate-active enzymes and their associated modules were shown to be of major importance for "pregenomic" applications. Three-dimensional structure is conserved within the families (Davies and Henrissat, 1995; Henrissat and Davies, 1997). This means that once the structure has been established for any family member it may direct and inform strategies for investigation of other members, including their structure solution by molecular replacement and the homology modeling of related sequences. Family-specific sequences have been used to design degenerate oligodeoxyribonucleotide

[1] This work was funded in part by the European Commission (grant no. BIO4–97–2303). The Marseille and York laboratories are supported by the Centre National de la Recherche Scientifique, the Biotechnology and Biological Science Research Council, and the Wellcome Trust. G.J.D. is a Royal Society University Research Fellow.

* Corresponding author; e-mail bernie@afmb.cnrs-mrs.fr; fax 33–4–91164536,

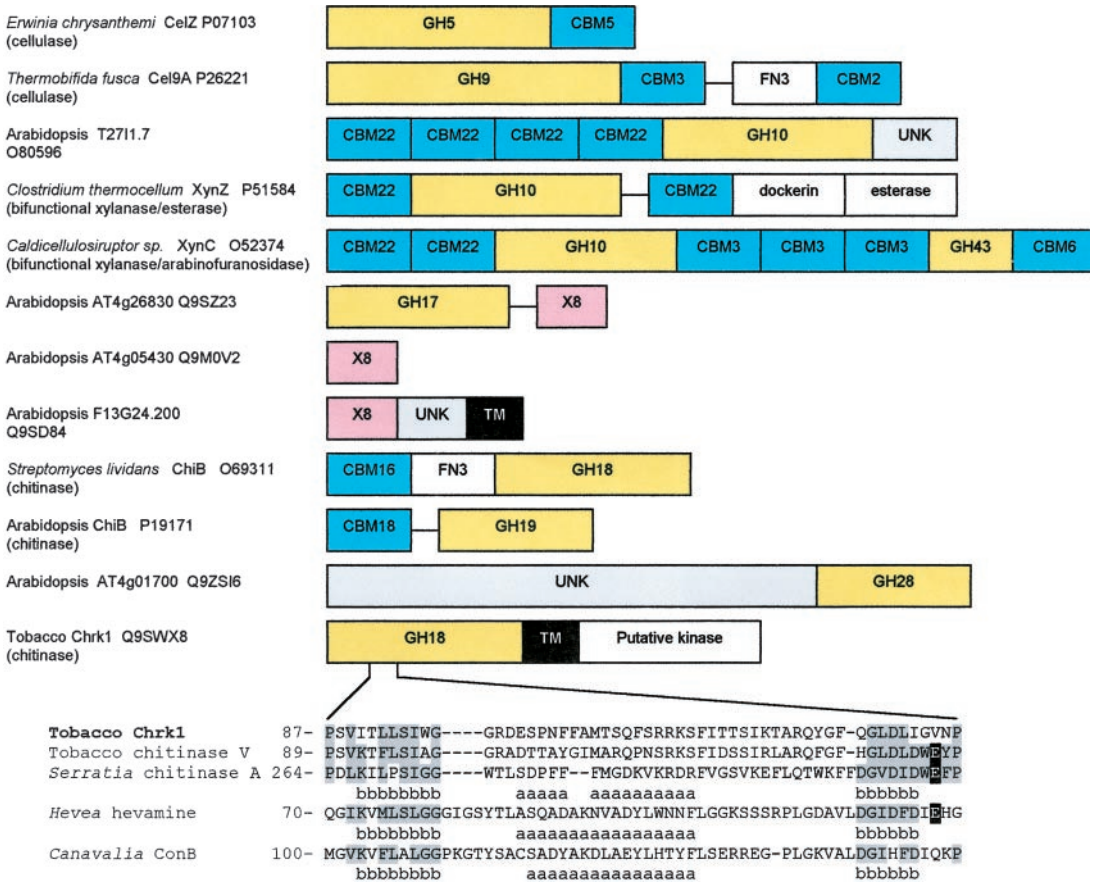

Figure 1. Top, Examples of modular glycoside hydrolases and related proteins. The yellow boxes represent the catalytic domain with the glycoside hydrolase family number indicated after GH. The function of the protein is indicated in parentheses where it was experimentally determined. Carbohydrate-binding modules are shown in blue with the family number appearing after CBM, gray boxes labeled UNK represent regions of unknown function, black boxes labeled TM represent transmembrane segments, other modules are indicated by their function (esterase) or name (dockerin; FN3, fibronectin type III-like), and pink boxes labeled X8 represent a newly identified module family found in plants (see text). When two consecutive modules are separated by a clearly identifiable linker peptide, the peptide is indicated by a horizontal line. Bottom, Multiple sequence alignment of Chrk1 with selected family GH18 members: chitinase V of tobacco (*Nicotiana tabacum*; Q43591); *Serratia marcescens* chitinase A (P07254); *Hevea brasiliensis* chitinase (hevamine; P23472); and concanavalin B of *Canavalia ensiformis* (P49347). Similarities are outlined in gray; the secondary structure (b for strand, a for helix) found in the three-dimensional structures of the chitinases from *S. marcescens*, *H. brasiliensis*, and concanavalin B are indicated under each sequence. The catalytic residue of chitinases is noted in white on a black background.

probes to isolate cDNA coding for other members of the family (Sheppard et al., 1994). This approach has found widespread application for functional cloning. Many families were shown to be poly-specific. This is an example of divergent evolution to acquire new substrate specificity (nature's protein engineering). In contrast, many enzymes displaying identical substrate specificity are found in different families displaying totally unrelated three-dimensional folds (Davies and Henrissat, 1995). At the catalytic level, for the enzymes performing reactions at sugar anomeric carbon, the reaction can proceed either with net retention or inversion of the anomeric configuration. Mechanism is dictated by the location of functional residues within the three-dimensional structure and hence by the sequence. Once the stereochemical mechanism is established for one member of a family, it may be safely extended to other members of that family (Gebler et al., 1992), i.e. catalytic mechanism is conserved within each family.

Furthermore, because the catalytic residues are conserved within a family once they have been identified in both position and function for one member of a family, they can easily be inferred for all members of the family. The absence of a catalytic residue in a member of unknown function (if not a sequence error) generally indicates an interesting alteration of the molecular mechanism or a lack of catalytic activity. An example of this is the plant enzyme myrosinase involved in the hydrolysis of anionic thio-glycosides named glucosinolates. The crystal structure of myrosinase showed that one of the two catalytic glutamates of the otherwise highly homologous β-glucosidases of family GH1 is absent and is

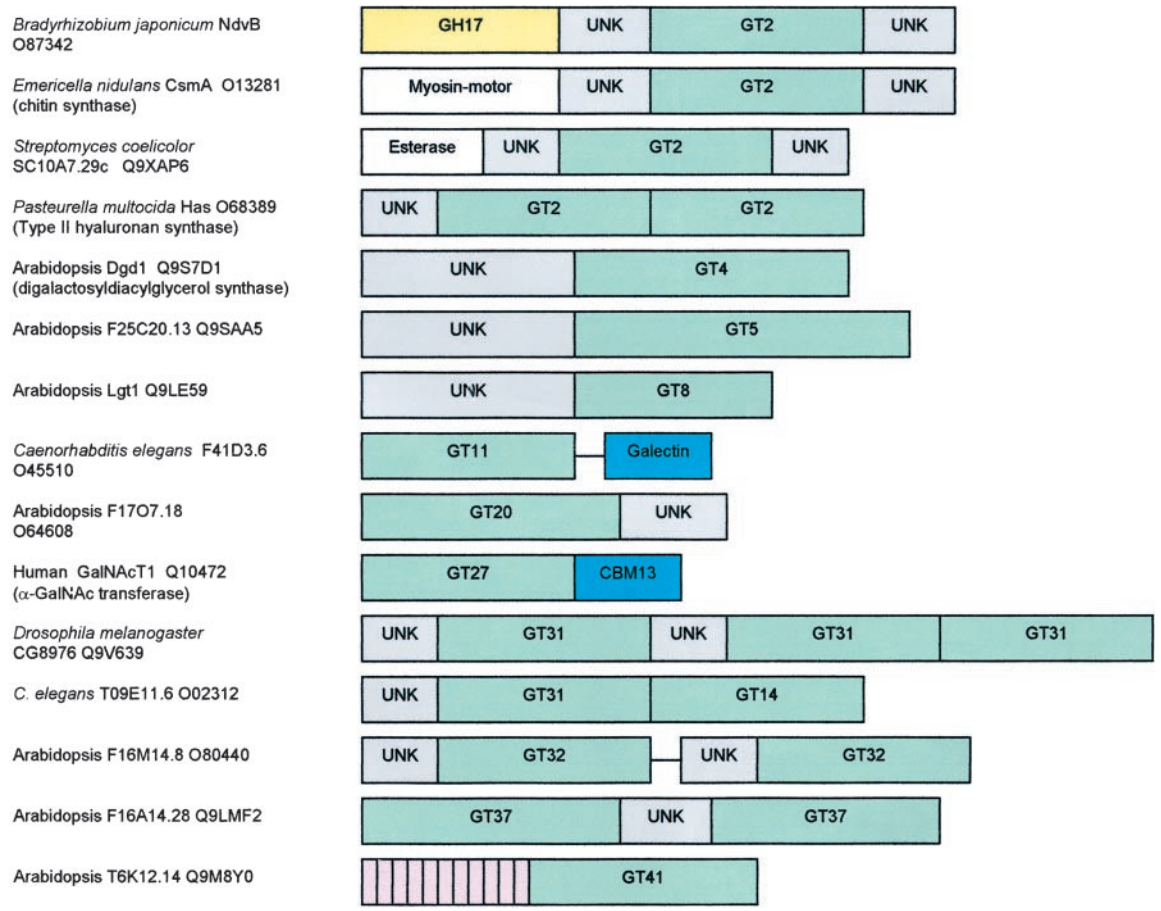

Figure 2. Examples of modular glycosyltransferases and related proteins. Pale-green boxes represent the catalytic domain with the glycosyltransferase family number indicated after GT, carbohydrate-binding modules are shown in blue with the family number appearing after CBM, gray boxes labeled UNK represent regions of unknown function, the yellow box represent a module belonging to glycoside hydrolase family GH17, and the pink boxes on the last line represent tetratricopeptide repeats. Other modules are indicated by their putative function (myosin motor and esterase). Several chitin synthases bear an N-terminal myosin motor protein and this strongly suggests that chitin synthesis may be guided by association with cytoskeletal structures (Fujiwara et al., 1997).

instead replaced by a Gln (Burmeister et al., 1997). Recent crystallographic work shows that myrosinase has evolved to use ascorbate to replace the missing Glu (Burmeister et al., 2000). Another example allows us to predict a putative plant chito-oligosaccharide-signaling receptor. GenBank accession number AF088885 encodes a 739-amino-acid protein from tobacco, Chrk1. This ORF shows significant similarities with family GH-18 chitinases. The similarity is, however, restricted to the first 345 residues of Chrk1. Furthermore, the C-terminal 390 residues of this protein bear strong similarities to a large number of protein kinases, the best scores being with a number of plant Ser/Thr kinases. The two domains of Chrk1 are separated by a central, most likely membrane-spanning, region (Figs. 1 and 3). Contrary to "classical" retaining glycosidases where two catalytic residues perform the catalytic reaction, family GH-18 chitinases use only one catalytic amino acid together with "anchimeric" assistance from the substrate (Terwisscha van Scheltinga et al., 1995). A close inspection of the alignment around the catalytic region of family 18 chitinases (Fig. 1) shows that Chrk1 lacks this catalytic amino acid, as does concanavalin B from *C. ensiformis* (Hennig et al., 1995). No enzymatic activity has been detected for concanavalin B and we therefore conclude that the N-terminal domain of Chrk1 has also lost its catalytic function. It may instead act as a carbohydrate-binding protein separated from a protein kinase signaling domain via a transmembrane helix. Although the similarity with chitinases makes it tempting to suggest the receptor may bind chito-oligosaccharides or their derivatives but the precise molecule that is recognized cannot be inferred from sequence analysis alone. Our predictions are consistent with the emerging picture of the modular structure of plant receptors, with a recruitment of different extracellular domains, which are fused onto intracellular protein kinases via a membrane-spanning region. In this respect, Chrk1 has a

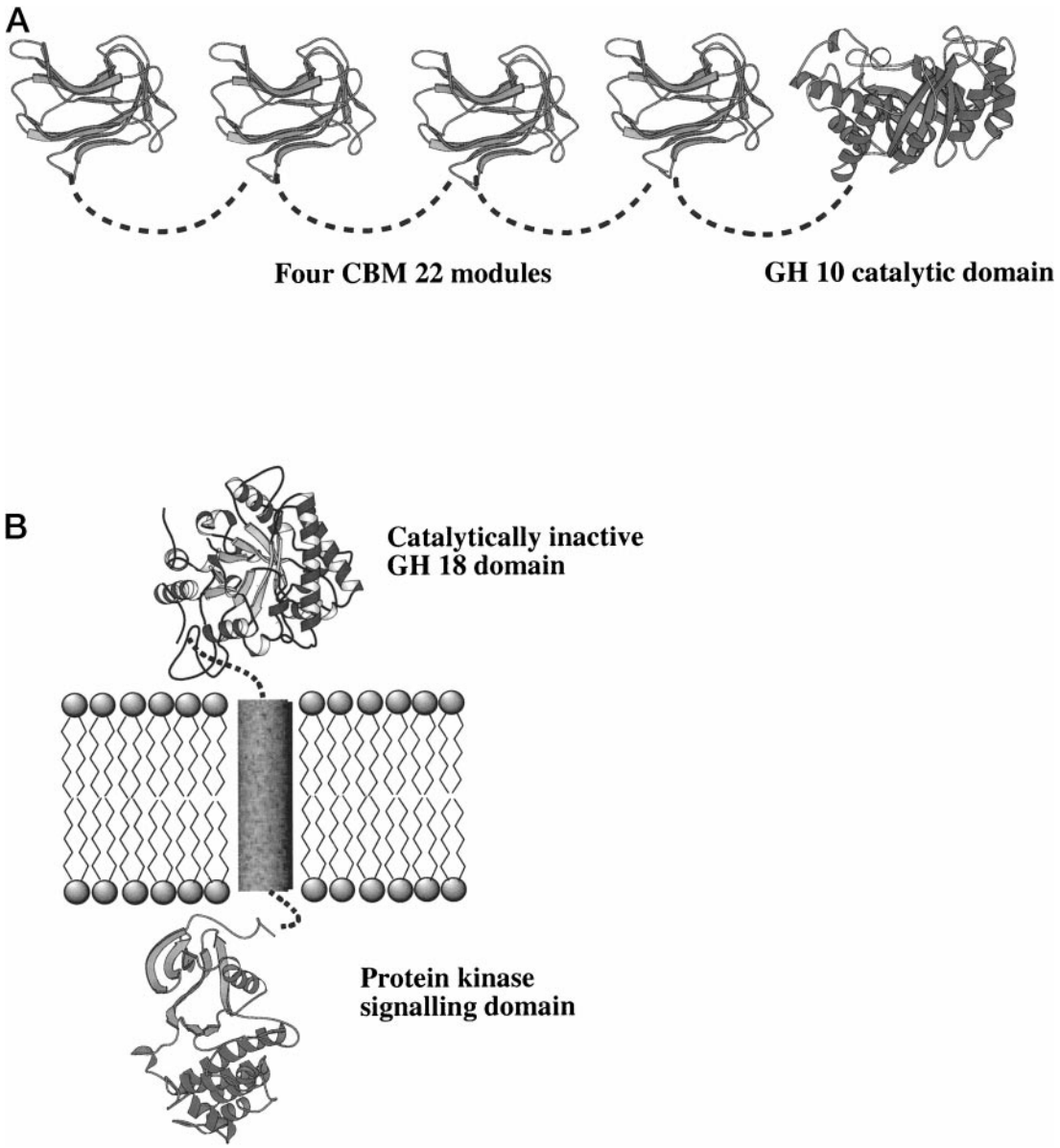

Figure 3. A, Modeled structure of ORF T27I1.7 from Arabidopsis (O80596), which consists of four repeats of a CBM22 module (homologs are implicated in xylan binding; Charnock et al., 2000), together with a family GH10 xylanase catalytic domain (Fig. 1). B, Modeled structure for a putative plant oligosaccharide receptor. ORF Chrk1 from tobacco (Q9SWX8) displays an extracellular domain with homology to family GH-18 chitinases, but lacks the essential catalytic acid residue. This domain is linked via a transmembrane segment to a Ser/Thr kinase domain. This allows us to propose a model for oligosaccharide signaling events in plants. These figures were drawn with the MOLSCRIPT program (Kraulis, 1991) using Protein Data Bank entries with accession numbers 1DYO and 1EOW (A) and 1CTN and 3LCK (B).

modular organization reminiscent of that of the brassinosteroid receptor or the human insulin receptor.

In the genomic era, the families and modular description of carbohydrate-active enzymes have further advantages. The availability of a number of completely sequenced genomes allows us to search and list all the carbohydrate-active enzymes possessed by the organism, as was recently performed for Arabidopsis (Henrissat et al., 2001). Furthermore, one can compare the carbohydrate-active enzymes content of different genomes and derive information on the evolution of carbohydrate metabolism such as the transfer of genes between species (Coutinho and Henrissat, 1999b).

One limitation with the family classifications is that a family can be defined only when one of its members is characterized biochemically. For example, the majority of β-linked polysaccharides in plants are synthesized by glycosyltransferase family 2 (GT-2) enzymes, but several fungal and plant sequences, demonstrably not family GT-2 members and thus

potentially forming a separate glycosyltransferase family, are annotated in sequence databanks as potential β-1,3 glucan synthases. The lack of direct experimental evidence for the UDP-glucose glucosyltransferase activity of these proteins has prevented their assignment to a glycosyltransferase family until a very recent report demonstrated this activity unequivocally (Kottom and Limper, 2000). The β-1,3-glucan synthases now form family GT-48.

The plurimodular structure of carbohydrate-active enzymes has major implications for genomic annotations and discovery of gene function. A large number of annotations are incorrect because they reflect a hit with a non-catalytic module only. A vivid example is with the proteins carrying an approximately 100-amino-acid module termed "X8" (in pink in Fig. 1). In a significant number of these proteins, the X8 module is found at the C terminus of a family GH-17 β-1,3-glucanase, suggesting β-1,3-glucan-binding function. This module, however, is also found fused to proteins that are not glycoside hydrolases and even appears in isolation (Fig. 1). However, because the first occurrence of a protein containing this X8 module was in a β-1,3-glucanase, several of the X8 proteins are misleadingly annotated as "β-1,3-glucanase-like" or as displaying "similarity to β-1,3-glucanase," even when the X8 module is not attached to a catalytic entity. This family of modules, present in plant sequences only, must have great significance as no less than 38 copies are found in the Arabidopsis genome. The fact that this domain is found fused to catalytic domains, in isolation, and linked to a transmembrane segment points to a spectrum of different cellular functions. In addition to the problems caused by modularity, further genomic annotation errors occur because of the poly-specific nature of the sequence families. This leads to both over-prediction, such as "putative cellulose synthase" (when it is known that family GT-2 contains a vast spectrum of substrate specificities from the synthesis of cellulose through complex cell surface glycolipid formation), and under-prediction, such as "putative sugar hydrolase."

The modularity of carbohydrate-active enzymes is of a significance that goes beyond plant science. With the rapidly growing number of genomes being sequenced, great care must be taken to "dissect" the various modules from single polypeptides or ORFs during sequence comparisons. To aid this process, the modular description of all carbohydrate-active enzymes is currently being undertaken in our laboratories.

ACKNOWLEDGMENT

The help of Dr. Pedro M. Coutinho is gratefully acknowledged.

Received September 8, 2000; accepted September 19, 2000.

LITERATURE CITED

Burmeister WP, Cottaz S, Driguez H, Iori R, Palmieri S, Henrissat B (1997) Structure **5:** 663–675

Burmeister WP, Cottaz S, Rollin P, Vasella A, Henrissat B (2000) J Biol Chem (in press)

Campbell JA, Davies GJ, Bulone V, Henrissat B (1997) Biochem J **326:** 929–939

Charnock SJ, Bolam DN, Turkenburg JP, Gilbert HJ, Ferreira LMA, Davies GJ, Fontes CMGA (2000) Biochemistry **39:** 5013–5021

Coutinho JB, Gilkes NR, Warren RAJ, Kilburn DG, Miller RC Jr (1993) Mol Microbiol **6:** 1243–1252

Coutinho PM, Henrissat B (1999a) *In* HJ Gilbert, G Davies, B Henrissat, B Svensson, eds, Recent Advances in Carbohydrate Bioengineering. The Royal Society of Chemistry, Cambridge, United Kingdom, pp 3–12

Coutinho PM, Henrissat B (1999b) J Mol Microbiol Biotechnol **1:** 307–308

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

Fujiwara M, Horiuchi H, Ohta A, Takagi M (1997) Biochem Biophys Res Commun **236:** 75–78

Gebler J, Gilkes NR, Claeyssens M, Wilson DB, Béguin P, Wakarchuk WW, Kilburn DG, Miller RC Jr, Warren RA, Withers SG (1992) J Biol Chem. **267:** 12559–12561

Gilkes NR, Henrissat B, Kilburn DG, Miller RC Jr, Warren RA (1991) Microbiol Rev (1991) **55:** 303–315

Hennig M, Jansonius JN, Terwisscha van Scheltinga AC, Dijkstra BW, Schlesier B (1995) J Mol Biol **254:** 237–246

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

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

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

Henrissat B, Coutinho PM, Davies GJ (2001) Plant Mol Biol (in press)

Henrissat B, Davies GJ (1997) Curr Opin Struct Biol **7:** 637–644

Kottom TJ, Limper AH (2000) J Biol Chem (in press)

Kraulis PJ (1991) J Appl Cryst **24:** 946–950

Lerner DR, Raikhel NV (1992) J Biol Chem **267:** 11085–11091

Lucas J, Henschen A, Lottspeich F, Vögeli U, Boller T (1985) FEBS Lett **193:** 208–210

Raikhel NV, Lee HI, Broekaert WF (1993) Annu Rev Plant Physiol Plant Mol Biol **44:** 591–615

Sheppard PO, Grant FJ, Oort PJ, Sprecher CA, Foster DC, Hagen FS, Upshall A, McKnight GL, O'Hara PJ (1994) Gene **150:** 163–167

Shinshi H, Neuhaus JM, Ryals J, Meins F Jr (1990) Plant Mol Biol **14:** 357–368

Svensson B, Jespersen H, Sierks MR, MacGregor EA (1989) Biochem J **264:** 309–311

Terwisscha van Scheltinga AC, Armand S, Kalk KH, Isogai A, Henrissat B, Dijkstra BW (1995) Biochemistry **34:** 15619–15623

Tomme P, Warren RAJ, Gilkes NR (1995) Adv Microb Physiol **37:** 1–81

van Tilbeurgh H, Tomme P, Claeyssens M, Bhikhabhai R, Pettersson G (1986) FEBS Lett **204:** 223–227

Warren RAJ (1996) Ann Rev Microbiol **50:** 183–212