The sequence-based classifications of carbohydrate-active enzymes

Sorting the diverse

Gideon J. Davies (University of York) **and Michael L. Sinnott** (University of Huddersfield)

Carbohydrates offer a structural and chemical diversity unrivalled in Nature: two glucose residues can be joined together in 30 different ways, and, with six different sugars, the number of possible isomers exceeds 1012 [1]. This huge diversity is reflected in the diverse roles for carbohydrates in Nature. Mono-, di-, oligo- and poly-saccharides and glycoconjugates play myriad roles in biology, in addition to well-known ones such as energy storage (starch, glycogen) and maintenance of structure (cellulose, chitin, alginate). The diversity of what is sometimes called the 'glycome' also provides for a subtle means of cellular communication in higher organisms: carbohydrates are the language of the cell. Sugar-mediated interactions not only are important for the communication of healthy cells, but also play crucial roles in disease, viral invasion and bacterial attack and malignancy. Sharon [2] has termed the challenge of carbohydrates as "the last frontier of molecular and cell biology". There is thus considerable interest in the enzymes whose job it is to modify and cleave carbohydrates [GHs (glycoside hydrolases) and lyases] and those involved in their biosynthesis, GTs (glycosyltransferases). Typically, these enzymes make up approx. 1–2% of the genome of any organism [3]. Thus, at the time of writing, there are around 70000 ORFs (open reading frames) known which potentially encode GHs or GTs. A major goal for the scientific community is to extract useful information on the enzymes encoded by these ORFs from sequence alone. This is an enormous challenge, one complicated by the modular nature of the enzymes themselves [4].

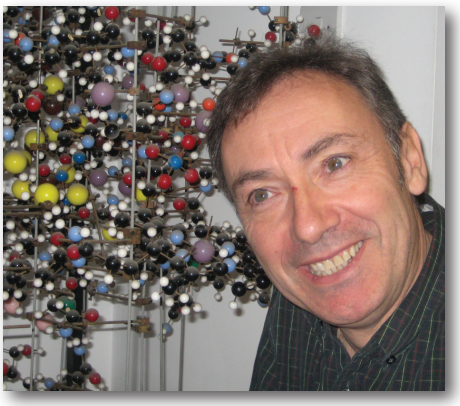


Figure 1. Bernard Henrissat

In the 1990s, Bernard Henrissat (Figure 1) initiated a sequence-based classification of carbohydrate-active enzymes that now underpins all functional, structural and mechanistic consideration of these proteins. His first classic *Biochemical Journal* paper, 'A classification of glycosyl hydrolases based on amino acid sequence similarities' [5] was built largely on the unusual and challenging technique of HCA (hydrophobic cluster analysis) [6] (described below), and defined the first 35 sequence-based families of GHs (termed families GH1–GH35), the enzymes involved in the hydrolysis of the glycosidic bond in di-, oligo- and poly-saccharides and glycoconjugates. The second classic *Biochemical Journal* paper appeared in 1993 [7], when a further 181 GH sequences were analysed, and the number of GH families rose to 45. There was a similar expansion in 1996 [8]. Subsequently, Bernard Henrissat and Pedro Coutinho have established a website with a continuously updated classification database (http://www.cazy.org); at the beginning of 2008, there were 112 GH families containing almost 40000 ORFs (Figure 2). Approx. 3000 pages are downloaded from the CAZy server daily, emphasizing the central position of this sequence classification in carbohydrate research today.

Carbohydrates offer a diversity that far surpasses that available with proteins or nucleic acids. Henrissat was very quick to realize that the wealth of different substrates was more than matched by the plethora of enzymes responsible for their degradation. For example, even a comparatively simple substrate such as cellulose, a regular polysaccharide of β-1,4-linked glucose, requires a complex enzymatic consortium for its complete degradation. Henrissat's first paper on GH classification was inspired by this earlier work on cellulases [9]. In this initial study, Henrissat used HCA to define six distinct families of cellulases, termed families A–F. HCA itself was an unusual, perhaps confusing, technique, derided by some as "French Impressionism", but, in the hands of an expert, it proved to be an amazingly powerful tool for comparing sequences and hence for helping place distantly related enzymes into families. HCA is based on the principle of the 'helical wheel', that one face of an α-helix is predominantly hydrophobic, and so, when the linear amino acid sequence is redrawn in two dimensions, with a helical

pitch, secondary structural elements display characteristic clusters. α-Helices appear aslong horizontal clusters and β-strands as shorter, vertical, arrays [10]. The skilled user was thus able to compare sequences even when the similarity was so low as to escape detection by then conventional means. HCA, augmented at that time (and probably now surpassed) by powerful sequence similarity detection tools, formed the basis for Henrissat's original classification papers.

At that time, the field needed a useful, predictive, manner of classifying enzymes and their sequences. Showing characteristic insight, Henrissat appreciated that other enzyme classifications were inadequate for this task. The International Union of Biochemistry and Molecular Biology Enzyme Commission (EC) numbers (for example EC 3.2.1.x for GHs), neither have enough scope to reflect all known GH specificities, nor reflect structural and mechanistic features. Furthermore, EC numbers cannot cope with the broad overlapping specificities as frequently observed in this field. Yet, the challenge in the early 1990s, and one that is even greater today with whole genome sequencing, was how to handle the vast amount of sequence data, and more importantly how to harness it to provide useful analytical output. The presence both of divergent evolution from a common ancestor to acquire new specificities, and convergent evolution towards similar enzyme mechanisms, means there is frequently no correlation between the EC number of an enzyme class and the sequences of the enzymes that perform these reactions. For example, enzymes classified as EC 3.2.1.4, endo-β-1,4-glucanases, populate 14 sequence-distinct GH families. In contrast, within a GH family, there may be a range of different specificities and hence EC numbers. Family GH1, for example, contains enzymes with 18 different EC numbers.

Genome sequencing, which is now remarkably rapid, provides information on vast numbers of gene sequences, but the present consensus is that the number of three-dimensional protein folds is limited to perhaps as little as a few thousand

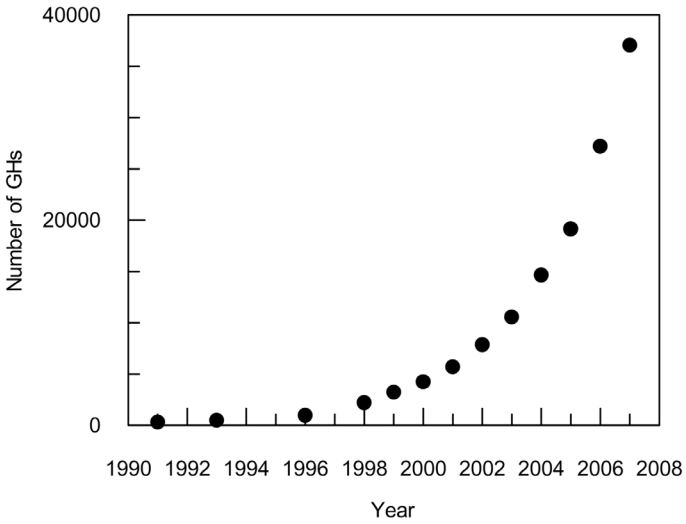


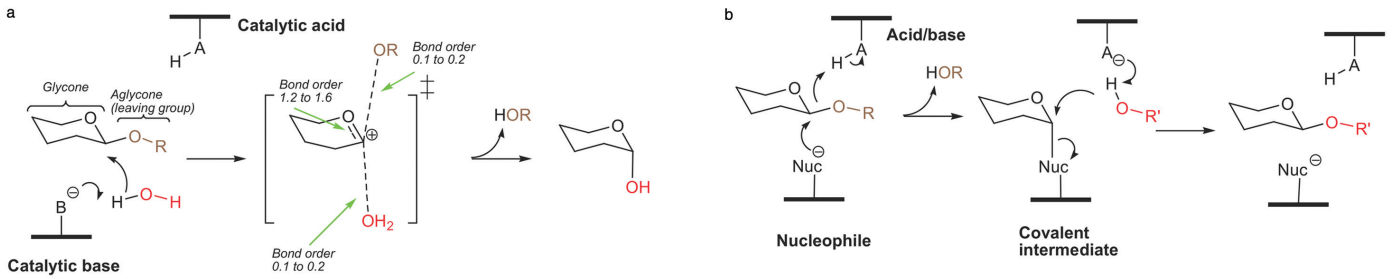
Figure 2. The year-on-year growth in GH ORFs highlights the challenge for a sequence-based classification

(reviewed in [11]). Determination of the fold of an individual protein by X-ray crystallography can now be reasonably fast (using recombinant proteins incorporating selenomethionine residues, powerful synchrotron radiation sources and cryo-crystallography), but it is only quick once suitable crystals are available. At the mechanistic level, complete delineation of a catalytic mechanism can take much more time. The beauty of Henrissat's classification is that because sequence and three-dimensional fold are related, a classification based on sequence similarity alone provides a significant level of structural and mechanistic insight. As we shall see, members of the same family have the same protein fold and closely related catalytic mechanisms, so, as Ronald Reagan is alleged to have said of redwood trees when he was Governor of California, "When you've seen one, you've seen 'em all".

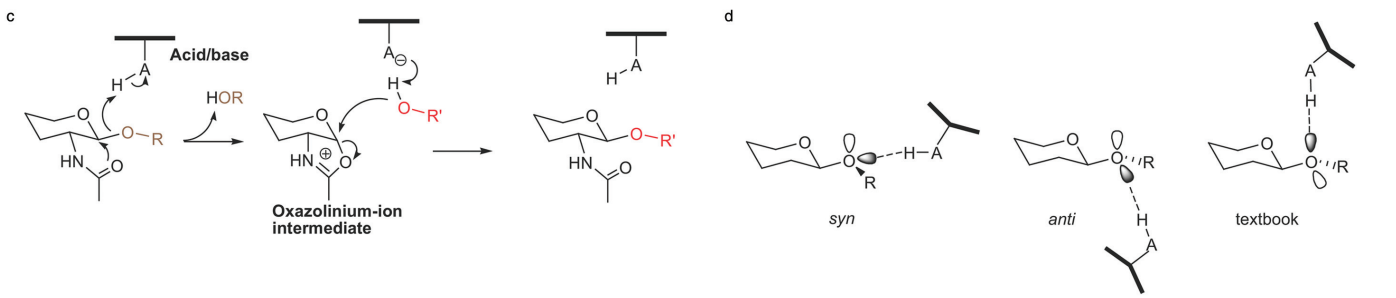
Mechanistic insight from sequence comparison

One of the key benefits of the Henrissat classification is the insight it gives into catalytic mechanism. Although it is in principle possible for a GH to liberate either ring or straight-chain forms of the sugar as initial products, in practice, the ring form of the product is always the same as that of the substrate. The reducing sugar product either has the same configuration at the anomeric carbon or its opposite: hence, catalysis occurs with retention or inversion of the anomeric configuration (GH mechanisms have been reviewed several times, e.g. [12–14]). The only obvious exceptions are family GH23 (which contains both lytic transglycosidases which yield 1,6-anhydromurein residues with retention and goose-type lysozymes which act with inversion) and the chemically unusual NAD^+-dependent GHs exemplified by family GH4 (e.g. [15]). Thus, with very few exceptions, all members of the same GH family have the same reaction stereochemistry, as first reported by Withers and co-workers [16].

Figure 3. Types of GH. Although a six-membered pyranose is shown, the sugar ring ('glycone') can be either six-membered ('pyranose') or five-membered ('furanose'). Leaving groups are shown in gold, and the ultimate attacking nucleophile is in red.



(**a**) Single-displacement (single transition state), inverting. The transition state is the structure of highest energy encountered as the reaction proceeds, and is indicated by ‡. The broken lines represent partial bonds. The glycones in transition states of all glycosyl transfers carry a partial positive charge, originally on the reaction centre, but stabilized by electron release from a lone pair on the ring oxygen. These transition-states are highly 'dissociative' as reflected in the 'bond orders'. (**b**) Double-displacement, retaining, with an enzyme nucleophile. The acid–base catalyst is shown, but not the two transition states, whose glycones are similar to those of the single displacement transition state. If the intermediate is more readily intercepted with water, the enzyme is a GH, if by another alcohol, it is a transglycosidase.



(**c**) Variant double-displacement mechanism using the *trans*-acetamido group of the substrate as nucleophile. The mechanism has so far been established only with pyranosidases. (**d**) *Syn* and *anti* protonation. An illustration of the power of the Henrissat classification is that it correlates well with mechanistic grouping discovered subsequently. Here, GH family membership correlates perfectly with whether the enzyme uses *syn* or *anti* trajectories for protonation of the leaving group [24] (illustrated here with an equatorial pyranoside). No enzymes, thus far, protonate from 'above' as drawn (albeit for convenience) in most reviews and textbooks.

The key point of a sequence, and hence structural, classification is that the active-centre residues that define a given catalytic mechanism are essentially conserved in a given GH family. Koshland's application of Ingold's rules about nucleophilic substitution at a saturated carbon to GH action had, by 1953 [17], led to a clear understanding that inverting GHs catalysed a single displacement and retaining GHs a double displacement (Figures 3a and 3b). The nucleophile in simple retaining GHs is normally an enzyme carboxylate (aspartate or glutamate), although in certain families (GH18, GH20, GH56, GH84, GH85 and possibly GH102–GH104), natural selection has exploited a *trans*-acetamido group of the substrate for this role using 'neighbouring-group participation' (e.g. [18,19]) (Figure 3c). Retaining sialidases (GH33, GH34 and GH83), acting on substrates bearing a carboxy group next to the reaction centre, instead use the electrically neutral phenolic hydroxy group of an enzyme tyrosine [20]. Stable glycosyl-enzyme intermediates, which permit direct and unambiguous identification of the enzyme nucleophile, can be obtained by the now classic Zechel and Withers inactivators [14] or enzymes with a mutated acid–base catalyst; in both cases, the glycosyl-enzyme intermediate (Figure 3b) is formed, but may not turn over. Change in both glycone and acid–base catalyst can be employed simultaneously, as they were to permit the examination of the covalent intermediate in lysozyme action to redefine the textbook mechanism [21]. What is especially noteworthy about the

Henrissat classification is that, within a retaining GH family, the enzyme nucleophile is strictly conserved and hence predictable for all subsequent members of the family.

The acid catalyst in inverting families, and the acid–base catalysis in retaining families, can usually be identified with confidence from the differential effects of its mutation on various substrates [22,23]. The conservation of the acid catalyst within an inverting family, or the acid–base catalyst within a retaining family is again almost perfect. That said, one cannot support Reagan's redwood analogy completely: all mechanistic taxa should be determined for more than one member of a GH family, given the fact any exceptions are frequently extremely interesting, and that people sometimes make mistakes. The handful of exceptions, enzymes in which nucleophile or acid–base are not conserved, reflect enzymes acting on natural substrates whose aglycones do not require protonation, or where the protein completely lacks GH activity as the scaffold has been co-opted by evolution for another function, as observed for GH family members now acting as lectins and proteinaceous enzyme inhibitors. Given their potentially lesser roles in catalysis, identification of the catalytic base in inverting enzymes, and of 'helper' auxiliary residues is less easily divined from sequence alone. Furthermore, for many GH families, identification of the base is ambiguous even after three-dimensional structure determination!

Perhaps the most spectacular example of the success of Henrissat's classification is its ability to correlate a mechanistic grouping or 'taxon' which was only recognized after the classification was established. For schematic diagrams, GH mechanisms were traditionally depicted with the proton from the acid catalyst approaching the aglycone oxygen atom from above, but Heightman and Vasella [24] designed basic inhibitors in which the axis of the nitrogen lone pair was fixed relative to the glycone, and showed that proton transfer could be *syn* to the C1–O5 bond (of an aldopyranoside) or *anti*, but never from above (Figure 3d). This subtle stereochemical observation is also conserved within a GH family.

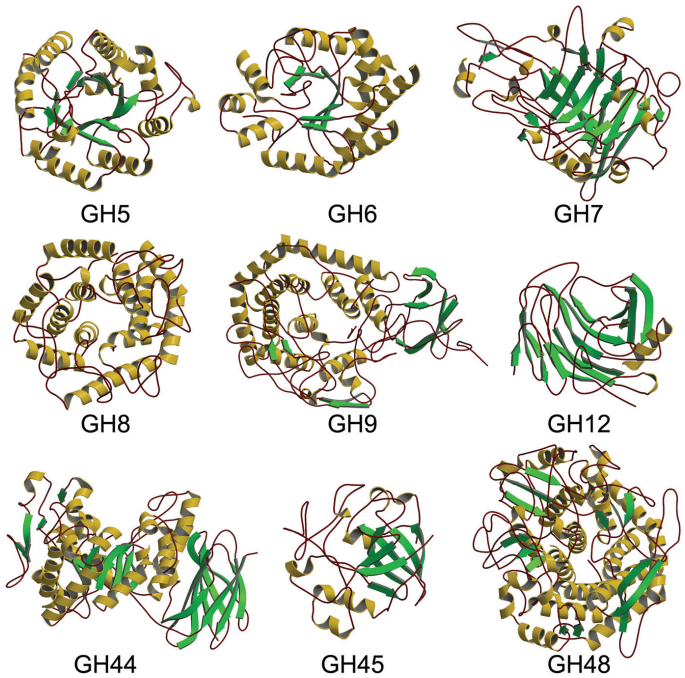


Figure 4. EC number does not correlate with three-dimensional structure, sequence or reaction mechanism. Here are nine of the different enzyme families (GH5–GH9, GH12, GH44, GH45 and GH48) classified as endoglucanases (EC 3.2.1.4). Similarly, three-dimensional structure does not correlate with EC number. Divergent evolution means that several of the GH families are populated with enzymes having many different substrate specificities. The family GH9 and GH44 structures also exemplify one of the challenges of carbohydrate-active enzymes, their modularity (here the enzymes are relatively simple, with a single catalytic domain appended to a β-sheet domain). Multi-modular enzymes cause enormous problems with genome annotations of carbohydrate-active enzymes, leading to frequent mis-annotation based on the homology of just one of many domains. This Figure was drawn with BOBSCRIPT [35].

Families, superfamilies and clans

As discussed, sequence similarity is reflected in structural conservation, meaning that all GH families have a protein fold which is essentially conserved within a family [3,12,13]. Two crucial factors became evident during the 1990s, as the developments in protein structure solution led to an explosion of three-dimensional structures for GHs. One was that there was a plethora of different three-dimensional folds for enzymes catalysing the same reaction; the first three-dimensional structures for cellulases revealed six or seven different protein topologies (Figure 4)! Secondly, it became apparent that many sequence-diverse families displayed the same three-dimensional fold, despite acting on different substrates. These structural relationships suggest that many related families share a common ancestor. The families display similar folds and catalytic apparatus, but their sequence similarities are at, or beyond, the borderline of detection and significance. Although some of these structural similarities had

been predicted, notably in another Henrissat HCA classic published in 1995 [25], many have only been observed subsequently to three-dimensional structure determination.

The whole issue of what to call a grouping of similar three-dimensional folds is a complex and controversial area. For GH families displaying essentially the global three-dimensional topology and with conservation of active-centre chemistry (and hence reaction stereochemistry), Henrissat adopted, with some advice from one of us, the term clan (Authors' note: M.L.S. confesses to suggesting the term 'clan', in place of 'superfamily' to Dr Henrissat, whose mother tongue is French, at the 1991 Wageningen meeting on xylanases, but pleads in mitigation that the film *Braveheart* lay some years in the future) [8,25]. Hence, for example, clan GH-A groups together a large assortment of retaining GHs active on a vast array of different β-D (or the equivalent α-L) glycosides in which the catalytic acid–base is located on strand β-4 and the enzymatic nucleophile on strand β-7 of a $(\beta/\alpha)_8$ barrel [25,26]. Thus far, and with great justification, Henrissat has proved cautious in advocating too many super- (and sub-) groupings of the CAZy families, most crucially because such next-generation classifications have to be 'future proof'. Henrissat would argue, with considerable justification, that too many super- and sub-families are currently being proposed in haste and using too few data. Sadly, many of these rapidly proposed groupings neither stand up to scrutiny nor survive the next release of sequence data. In contrast, the CAZy classification itself has withstood a 100-fold increase in the number of protein sequences since its inception.

Expansion of the sequence classification to other carbohydrate-active enzymes and their component domains

Building upon the success of the GH sequence classification, Henrissat has recently expanded this classification to other classes of carbohydrate-active enzymes and also to their component domains. This latter area is tricky, as it requires the delineation of these complex, and often multi-modular, enzymes [4]. CAZy now defines over 50 families of CBMs (carbohydrate-binding modules) in which at least one member has had its function clarified (reviewed in [27]) and there are many more to be characterized; Henrissat counts over 100 other distinct domains whose function is yet to be reported. CAZy also defines 18 families of polysaccharide lyases (the far smaller number reflecting the requirement for a uronic-acid-containing sugar to facilitate the β-elimination mechanism) and 15 families of carbohydrate esterases. The esterase classification is markedly less predictive than for the other classifications, however, given the astronomical number of 'generic' esterases, with broad substrate specificity, in Nature.

The expansion of Henrissat's sequence-based approach, beyond GHs, is perhaps best exemplified by the GT classification, whose magnitude rivals that of the hydrolase work. The CAZy classification of GTs, those enzymes using activated sugar donors to drive glycosidic bond formation, evolved from the pioneering work on GH classification and now (January 2008) defines 90 sequence-based GT families, which contain over 33000 ORFs [28,29]. In marked contrast with GH families, however, the constraint of a nucleotide-binding fold, for the nucleotide-sugar-dependent GTs, means that just two topologies (and small modifications thereof) have been seen for nucleotide-sugar-dependent GTs, termed the GT-A and GT-B folds [29]. Perhaps unfortunately, the terms GT-A and GT-B have been used to describe these two general protein folds, but this does not relate to mechanism. Enzymes with both GT-A and GT-B folds are observed to catalyse with both retaining and inverting mechanisms and with a variety of different chemistries. This does not therefore have the same meaning as the use of GH-A and GH-B for GH clans, which, as we discussed above, reflects a conserved three-dimensional structure, catalytic geometry and reaction stereochemistry. Because GT fold does not correlate with stereochemistry, there is a possibility that extremely large families such as inverting GT2 or retaining GT4 contain a few enzymes that achieve the 'wrong' stereochemistry. That said, there have been very few reclassifications thus far (a notable example is mannosylglycerate synthase which was classified as inverting family GT2 until the enzyme was shown to be retaining; it now forms a new family GT78 whose N-terminal GDP-Man-binding domain is indeed highly similar to that observed for inverting GTs [30]). Henrissat is also quick to rectify any errors when stereochemical data become available.

Thus far, just two classes of lipid-sugar dependent GTs have had their folds determined and, free from the straight-jacket of nucleotide binding, these enzymes have displayed three-dimensional structures both dissimilar to each other and different from the GT-A and GT-B folds. The potential dangers of the indiscriminate proposal of clans, alluded to above, is well exemplified here by the proposal of a GT-C clan for integral membrane GTs before the structure solution of any enzyme of this class [31]. The recent structure solution of the N-glycosylation oligosac-

charyltransferase [32] shows that the likely catalytic domain (with the substantial caveat that function is not demonstrated for the isolated domain) is not that section of the sequence used to define the super grouping. So on the basis of current evidence, 'GT-C' may simply reflect a conserved transmembrane topology to which is appended a GT catalytic domain.

Back to the future

One of us once commented that glycobiologists were a community similar to the Galapagos Islands, which risked evolving with little knowledge of other related communities [33]. It is clear that such worries still persist today, and that, as a field, people are often unaware of important developments on related enzymes simply because they are from different, or distant, organisms. The Henrissat sequence classification, initiated by his two classic *Biochemical Journal* papers in the 1990s, provides the language and framework for communication between the disparate areas of carbohydrate biochemistry and glycobiology. The CAZy sequence classification immediately informs about homologues in different organisms, known three-dimensional structures and catalytic mechanisms in a robust predictive manner. It should also prevent needless repetition of work. Furthermore, there is a wealth of information in the CAZy classification that also needs to be incorporated into expert genome annotation. Thoughtless, worse still computerized, sequence annotations are one of the great banes of modern molecular biology (the rice genome still contains 'brain-specific' and 'muscle-specific' proteins!), and the CAZy classifications could, and should, inform much more insightful annotations across the literature. The great challenge for the modern age is functional dissection. For example, of the ~33000 GT ORFs in CAZy today, approx. 95% encode proteins of undefined function. In these situations, CAZy can provide insight about fold, configuration of donor and product and probable catalytic mechanism. What is far harder to provide, however, is insight into substrate specificity. It is known that as few as one or two amino acid substitutions are all that is required to change the sugar specificity of some GTs. So, with the reservations considered above, it is likely that a robust subfamily dissection of the CAZy families, supported as it must be by structural and functional analysis of representative subfamily members, may be the way forward. This subdivision has already begun with the large α-amylase family GH13 [34]. A careful subfamily analysis of large families has considerable value for improved functional annotation of ORFs that look like carbohydrate-active enzymes in genomes and for a means to classify, compare and relate biological information as one step towards the analysis of 'carbohydrate systems' in whole organisms. CAZy, which has survived a 17 year explosion of sequence data, provides the foundation upon which all in the field build their research. The families have proved useful to diverse biologists, including molecular biologists, mechanistic enzymo logists, structural biologists and now researchers interested in genomes and functional genomics. Given the increasing importance of carbohydrates in cellular biology, and the re-emergence of plant cell wall hydrolysis for biofuel production, it is reassuring that the foundation provided by the sequence-based classification of carbohydrate-active enzymes is solid. ■

This article (doi:10.1042/BJ20080382) was first published in the Biochemical Journal (www.biochemj.org).

Professor Gideon Davies obtained a Ph.D. from the University of Bristol in 1990 under the supervision of Herman Watson and Len Hall. He then moved, via EMBL Hamburg with Keith Wilson, to the University of York to work with Dale Wigley and Guy Dodson on DNA gyrase. Subsequently, and with periods in Uppsala and Grenoble, Gideon was awarded a Royal Society University Research fellowship to pursue his studies of carbohydrate enzymology. Gideon is now one of the University of York's '40th Anniversary' professors and a Royal Society/Wolfson Research Merit Award recipient. His research focuses on the structural enzymology and chemical biology of carbohydrate-active enzymes.
email: davies@ysbl.york.ac.uk

Professor Michael Sinnott obtained a Ph.D. from the University of Bristol under the supervision of Mark Whiting. After a postdoc at Stanford, he joined the Bristol faculty and progressed up the ranks, becoming Reader in 1982. In 1989, he joined the University of Illinois at Chicago (becoming University of Illinois Senior Scholar 1992), but family circumstances dictated a 1996 return to the U.K., and he moved to the Paper Science Department at UMIST (University of Manchester Institute of Science and Technology). Now semi-retired, he continues research in the chemical and biochemical mechanisms of glycosyl transfer through a Visiting Professorship in the School of Applied Sciences, University of Huddersfield. His graduate text, Carbohydrate Chemistry and Biochemistry: Structure and Mechanism, was published in October 2007 by the RSC (Royal Society of Chemistry).
email: m.l.sinnott@hud.ac.uk

References

1. Laine, R.A. (1994) A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05×10^{12} structures for a reducing hexasaccharide: the *Isomer Barrier* to development of single-method saccharide sequencing or synthesis systems. Glycobiology **4**, 759–767
2. Sharon, N. (2001) The conquest of the last frontier of molecular and cell biology. Biochimie **83**, 555
3. Davies, G.J., Gloster, T.M. and Henrissat, B. (2005) Recent structural insights into the expanding world of carbohydrate-active enzymes. Curr. Opin. Struct. Biol. **15**, 637–645
4. Henrissat, B. and Davies, G.J. (2000) Glycoside hydrolases and glycosyltransferases: families, modules, and implications for genomics. Plant Physiol. **124**, 1515–1519
5. Henrissat, B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. **280**, 309–316
6. Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.P. (1987) Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett. **224**, 149–155
7. Henrissat, B. and Bairoch, A. (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. **293**, 781–788
8. Henrissat, B. and Bairoch, A. (1996) Updating the sequence-based classification of glycosyl hydrolases. Biochem. J. 316, 695–696
9. Henrissat, B., Claeyssens, M., Tomme, P., Lemesle, L. and Mornon, J.P. (1989) Cellulase families revealed by hydrophobic cluster analysis. Gene **81**, 83–95
10. Woodcock, S., Mornon, J.P. and Henrissat, B. (1992) Detection of secondary structure elements in proteins by hydrophobic cluster analysis. Protein Eng. **5**, 629–635
11. Grant, A., Lee, D. and Orengo, C. (2004) Progress towards mapping the universe of protein folds. Genome Biol. **5**, 107
12. Davies, G. and Henrissat, B. (1995) Structures and mechanisms of glycosyl hydrolases. Structure **3**, 853–859
13. Davies, G.J., Sinnott, M.L. and Withers, S.G. (1997) Glycosyl transfer. In Comprehensive Biological Catalysis (Sinnott, M.L., ed.), pp. 119–209, Academic Press, London
14. Zechel, D.L. and Withers, S.G. (2000) Glycosidase mechanisms: anatomy of a finely tuned catalyst. Acc. Chem. Rev. **33**, 11–18
15. Yip, V.L.Y., Varrot, A., Davies, G.J., Rajan, S.S., Yang, X., Thompson, J., Anderson, W.F. and Withers, S.G. (2004) An unusual mechanism of glycoside hydrolysis involving redox and elimination-steps by a family 4 β-glycosidase from *Thermatoga maritima*. J. Am. Chem. Soc. **126**, 8354–8355
16. Gebler, J., Gilkes, N.R., Claeyssens, M., Wilson, D.B., Beguin, P., Wakarchuk, W.W., Kilburn, D.G., Miller, R.C., Warren, R.A.J. and Withers, S.G. (1992) Stereoselective hydrolysis catalyzed by related β-1,4-glucanases and β-1,4-xylanases. J. Biol. Chem. **267**, 12559–12561
17. Koshland, D.E. (1953) Stereochemistry and the mechanism of enzymatic reactions. Biol. Rev. **28**, 416–436
18. Terwisscha van Scheltinga, A.C., Armand, S., Kalk, K.H., Isogai, A., Henrissat, B. and Dijkstra, B.W. (1995) Stereochemistry of chitin hydrolysis by a plant chitinase/lysozyme and X-ray structure of a complex with allosamidin: evidence for substrate assisted catalysis. Biochemistry **34**, 15619–15623
19. Dennis, R.J., Taylor, E.J., Maculey, M.S., Stubbs, K.A., Turkenburg, J.P., Hart, S.J., Black, G., Vocadlo, D.J. and Davies, G.J. (2006) Structure and mechanism of a bacterial β-glucosaminidase having O-GlcNAcase activity. Nat. Struct. Mol. Biol. **13**, 365–371
20. Watts, A.G., Damager, I., Amaya, M.L., Buschiazzo, A., Alzari, P., Frasch, A.C. and Withers, S.G. (2003) *Trypanosoma cruzi trans*-sialidase operates through a covalent sialyl-enzyme intermediate: tyrosine is the catalytic nucleophile. J. Am. Chem. Soc. **125**, 7532–7533
21. Vocadlo, D.J., Davies, G.J., Laine, R. and Withers, S.G. (2001) Catalysis by hen egg-white lysozyme proceeds via a covalent intermediate. Nature **412**, 835–838
22. Damude, H.G., Withers, S.G., Kilburn, D.G., Miller, Jr, R.C. and Warren, R.A.J. (1995) Site-directed mutagenesis of the putative catalytic residues of endoglucanase CenA from *Cellulomonas fimi*. Biochemistry **34**, 2220–2224
23. MacLeod, A.M., Lindhorst, T., Withers, S.G. and Warren, R.A.J. (1994) The acid /base catalyst in the exoglucanase/xylanase from *Cellulomonas fimi* is glutamic acid 127: evidence from detailed kinetic studies of mutants. Biochemistry **33**, 6371–6376
24. Heightman, T.D. and Vasella, A.T. (1999) Recent insights into inhibition, structure, and mechanism of configuration-retaining glycosidases. Angew. Chem. Int. Ed. **38**, 750–770
25. Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J.P. and Davies, G. (1995) Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. Proc. Natl. Acad. Sci. U.S.A. **92**, 7090–7094
26. Jenkins, J., Leggio, L.L., Harris, G. and Pickersgill, R. (1995) β-Glucosidase, β-galactosidase, family A cellulases. family F xylanases and two barley glucanases form a superfamily of enzymes with 8-fold β/α architecture and with two conserved glutamates near the carboxy-terminal ends of β-strands four and seven. FEBS Lett. **362**, 281–285
27. Boraston, A.B., Bolam, D.N., Gilbert, H.J. and Davies, G.J. (2004) Carbohydrate-binding modules: fine tuning polysaccharide recognition. Biochem. J. **382**, 769–781
28. Campbell, J.A., Davies, G.J., Bulone, V. and Henrissat, B. (1997) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem. J. **326**, 929–939
29. Coutinho, P., Deleury, E., Davies, G.J. and Henrissat, B. (2003) An evolving heirarchical family classification for glycosyltransferases. J. Mol. Biol. **328**, 307–317
30. Flint, J., Taylor, E., Yang, M., Bolam, D.N., Tailford, L.E., Martinez-Flietes, C., Dodson, E.J., Davis, B.G., Gilbert, H.J. and Davies, G.J. (2005) Structural dissection and high-throughput screening of mannosylglycerate synthase. Nat. Struct. Mol. Biol. **12**, 608–614
31. Liu, J. and Mushegian, A. (2003) Three monophyletic superfamilies account for the majority of the known glycosyltransferases. Protein Sci. **12**, 1418–1431
32. Igura, M., Maita, N., Kamishikiryo, J., Yamada, M., Obita, T., Maenaka, K. and Kohda, D. (2008) Structure-guided identification of a new catalytic motif of oligosaccharyltransferase. EMBO J. **27**, 234–243
33. Davies, G.J. and Henrissat, B. (2002) Structural enzymology of carbohydrate-active enzymes: implications for plant glycogenomics. Biochem. Soc. Trans. **30**, 291–297
34. Stam, M.R., Danchin, E.G.J., Rancurel, C., Coutinho, P.M. and Henrissat, B. (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of α-amylase-related proteins. Protein Eng. Des. Sel. **19**, 555–562
35. Esnouf, R.M. (1997) An extensively modified version of MolScript that includes greatly enhanced colouring capabilities. J. Mol. Graphics Modell. **15**, 132–134