The huge diversity of polysaccharides correlates with the diversity of enzymes dedicated to their synthesis and degradation. Henrissat [1], working initially with the glycohydrolases, has developed a logical scheme of classifying this diversity into families. On the basis of hydrophobic cluster analysis and sequence homology, he classified glycohydrolases into families (currently 61, for updates see web page http://expasy/hcuge/cgi-bin/lists?glycosid.txt). Each family consists of enzymes that can be seen to have evolved from a common ancestor on the basis of sequence similarity. However, as three-dimensional structures have been determined for these enzymes, it has become both possible and useful to cluster families together into superfamilies or clans, revealing their evolutionary history and frequently allowing prediction of the mechanism for new enzymes.

In this paper we consider a protein family to comprise proteins with significant sequence

Superfamilies: the 4/7 superfamily of $\beta$-barrel glycosidases and the right-handed parallel $\beta$-helix superfamily

R. Pickersgill\(^1\), G. Harris, L. Lo Leggio\(^2\), O. Mayans\(^3\) and J. Jenkins

Institute of Food Research, Reading Laboratory, Earley Gate, Whiteknights Road, Reading RG6 6BZ, U.K.

The huge diversity of polysaccharides correlates with the diversity of enzymes dedicated to their synthesis and degradation. Henrissat [1], working initially with the glycohydrolases, has developed a logical scheme of classifying this diversity into families. On the basis of hydrophobic cluster analysis and sequence homology, he classified glycohydrolases into families (currently 61, for updates see web page http://expasy/hcuge/cgi-bin/lists?glycosid.txt). Each family consists of enzymes that can be seen to have evolved from a common ancestor on the basis of sequence similarity. However, as three-dimensional structures have been determined for these enzymes, it has become both possible and useful to cluster families together into superfamilies or clans, revealing their evolutionary history and frequently allowing prediction of the mechanism for new enzymes.

In this paper we consider a protein family to comprise proteins with significant sequence
homology and a protein superfamily to comprise proteins whose homology is only revealed by detailed structural comparison or by comparing sequences and structure. The term protein fold defines a protein topology which could have evolved by either convergent or divergent evolution.

We focus on two protein superfamilies each of which have probably diverged from a common ancestor rather than converged to a common architecture. Evolution in these two superfamilies has apparently followed different routes and has not conserved the same features. In the 4/7-superfamily of βα-barrel glycosidases (clan GH-A), as in many other superfamilies such as the serine proteases, the active site is the most conserved region, while in the parallel β-helix proteins it is the structural framework which is best conserved while the active sites show extensive divergence. However, the conservation may partially reflect the criteria used to select the members of the superfamilies. Thus we can speculate that glycosidase active sites and mechanisms can also rapidly diverge while retaining structural similarities.

The 4/7-superfamily of βα-barrel glycosidases (clan GH-A)

Glycosyl hydrolases cleave glycosidic bonds with either inversion or retention of the configuration at the anomeric carbon (C1). Most have two essential carboxylates, one acting as the nucleophile and the other as the proton donor [2]. The typical separation of the acid/base and nucleophile is about 5 Å in retaining enzymes, but is around 10 Å in inverting enzymes where a water molecule is accommodated between the nucleophile and the substrate.

Xylanases cleave β-1,4-linked xylose units with retention of anomeric configuration [3] and are classified into families 10 and 11 of the glycosyl hydrolases. Both families have essential glutamates at their active sites [4,5]. However, despite sharing a common mechanism these enzymes have clearly evolved independently since the catalytic core of the family 10 xylanases comprises an 8-fold βα-barrel [5-8] (Figure 1a) while the family 11 xylanases are smaller and of predominantly β-architecture [4,9]. The structures of two barley β-glucanases, (1-3)-β-glucanase (GHS) and (1-3,1-4)-β-glucanase (GHR) belonging to family 17 [10] had strikingly similar architecture and active site geometry to the family 10 xylanase A from Pseudomonas fluorescens subsp. cellulosa (XynA) [11]. Superimposition of the βα-barrel structures revealed that the nucleophile, Glu-246 on strand 7, of XynA was close to the previously identified nucleophiles Glu-231 of GHS and Glu-232 of GHR and gave the identity of the acid/base in the β-glucanases, with Glu-92 in GHS and Glu-93 in GHR corresponding to Glu-127, on strand 4, in XynA. The structure of E. coli β-galactosidase [12], a member of family 2, and also a βα-barrel, had glutamates at the ends of strands β4 and β7 and the remnants of an ancestral sequence around these glutamates. The conserved βα-barrel architecture and conserved retaining mechanism involving glutamates at the ends of β-strands 4 and 7 led to the name 4/7-superfamily [13]. Further analysis of limited sequence conservation near the active sites and secondary structure predictions led to the suggestion that family 1 and 5 glycosyl-hydrolases also belong to the 4/7-superfamily of βα-barrels.

The superfamily therefore includes enzymes with no detectable sequence homology and various activities which have nevertheless evolved from a common ancestral protein. Subsequent structures of members of families 1 and 5 have confirmed the validity of this superfamily. The superfamily, or clan GH-A, has also been extended to include the enzymes in families 26, 30, 35, 39, 42 and 53 [14]. Although there are many proteins which have the 8-fold βα-barrel structure, the βα-barrels of the 4/7 superfamily structures are elliptical in cross section with the major axis running between β-strands 1 and 5. Lisa Holm (personal communication) points out that if structural similarity alone were used, the family 10 xylanases (retaining) would lie closer to β-amylase (inverting) than to the family 17 enzymes (retaining). Family 18 and 20 glycosyl hydrolases which have an acid/base at the end of β4 but probably use substrate assisted catalysis may also be related [15-17]. There are structural similarities, including a rare non-prolyl cis-peptide at the end of β8, between families 1, 2, 5, 17 and 18 suggesting an evolutionary relationship [18]. This highlights the importance of the criteria used to define the superfamily.

β-Bulges are distortions in the normal hydrogen bonding pattern in a β-sheet, such that two residues on the bulged strand are opposite a single residue on the adjacent strand, putting the usual alternation of side-chain direction out of register in the bulged strand and introducing a slight bend in the β-sheet [19,20]. β-Bulges are
known to play important biological roles in other proteins by positioning crucial residues [20]. A detailed comparison of the structural features of XynA and the homologous family 10 xylanase/glucanase β-1,4-glycanase from Cellulomonas fimi (Cex) [7], highlighted the importance of two conserved β-bulge distortions occurring on β-strands 4 and 7 involving the catalytic glutamic residues [11], namely a double β-bulge on β-strand 4, giving rise to a crankshaft-like structural feature (in XynA, Val-124 and Val-125 followed by an Asn-126 and the acid base Glu-127) and a single β-bulge on β-strand 7 (in XynA, Thr-245 and the nucleophile Glu-246). The crankshaft feature arises from the need to place two adjacent hydrophobic (valine) residues in the middle of β-strand 4, with the following two polar residues thrown outwards. Both this double β-bulge and the single β-bulge on β-strand 7, serve to orient the catalytically important residues in the active site. The importance of these β-bulges is shown by their structural conservation in all glycosyl hydrolase family 10 structures. The crankshaft feature on strand 4 and β-bulge on strand 7 are also conserved in family 1, 2, 5 and 17 glycosyl hydrolases. Figures 1b and 1c illustrate the conservation of the β-bulge distortions on strands 4 and 7 respectively for families 5, 10 and 17 of the 4/7 superfamily. From this analysis, the conservation of the distortions of the β-strands 4 and 7 can be taken as a structural signature of the 4/7 superfamily together with the shape of the βx-barrel.

Coiled protein folds
The structures of pectate lyase PelC from Erwinia chrysanthemi [21] and the alkaline protease from Pseudomonas aeruginosa [22] overthrew the general belief that purely parallel β-structures would be unstable and inspired great interest in how these coiled folds were organized and stabi-
lized [23–25]. Within a year three further parallel \( \beta \)-helix structures were published, pectate lyases from \textit{Bacillus subtilis} (BsPel) [26] and PelE from \textit{Erwinia chrysanthemi} [27] and a fragment of the P22 tailspike protein (TSP) [28], as well as the alkaline protease from \textit{Seratia marcescens} [29].

The pectate lyase parallel \( \beta \)-helix and the alkaline protease parallel \( \beta \)-roll are both characterized by the polypeptide chain coiling around and around to form a large right-handed superhelix [30–32]. The parallel \( \beta \)-helix contains three \( \beta \)-strands per turn and consecutive turns stack one on another to form the superhelix (Figure 2). A single turn comprises a strand of PB1, a turn T1, a strand of PB2, a short turn T2 and PB3. Turn T3 links to the next turn of the superhelix. PB1 and PB2 form an antiparallel \( \beta \)-sandwich with PB3 approximately perpendicular to PB2. The similar 'L' shaped parallel \( \beta \)-helix allows this nomenclature to be applied to the P22 tailspike protein so that the PB1 is equivalent to \( \beta \)-sheet C [28], PB2 equivalent to A and PB3 equivalent to B. There is no sequence repetition in the right-handed parallel \( \beta \)-helix.

**Figure 2**

The structures of pectate lyases: PelC and PelE from \textit{Erwinia chrysanthemi}, BsPel from \textit{Bacillus subtilis} and pectin lyase A (PnlA) from \textit{Aspergillus niger}. 
structures. The structure of UDP-N-acetylglucosamine acyltransferase [33] comprises a left-handed parallel $\beta$-helix. The left-handed cross-over connections between the $\beta$-strands of this structure challenged an almost 20-year-old dogma that such connections are rare [34]. Richardson [34] argues that the inherent right-handed twist of extended polypeptide and $\alpha$-helical segments naturally folds into right-handed coils as the ends of these segments are brought together. However in the left-handed parallel $\beta$-helix the sheets are extremely flat and the connections between adjacent $\beta$-strands are long. The cross section of the left-handed parallel $\beta$-helix is much closer to an equilateral triangle than 'L' shaped and the sequence much more repetitive. Several other structures have been determined which have coiled folds. These include a leucine-rich repeat $\beta/\alpha$-horseshoe structure [35], a leucine-rich repeat variant which forms an $\alpha/3_1$-coil [36] and the right-handed spiral folds including enoyl-coenzyme A hydratase [37] and 4-chlorobenzoyl coenzyme A dehalogenase [38].

Three new right-handed parallel $\beta$-helix structures have also been recently reported, P69 pertactin from Bordetella pertussis [39], rhamnogalacturonase A from Aspergillus aculeatus (RGase) [40] and pectin lyase A from Aspergillus niger (PnlA) [41]. At the same time, the structure of tailspike protein has been refined to higher resolution, characterized as an endohamnosidase and the binding of substrates analysed crystallographically [42].

Below we consider the relationships between these coiled protein folds and conclude that the structural similarities between the right-handed parallel $\beta$-helix proteins are so striking that it is likely these proteins diverged from a common ancestor.

A comparison of right-handed parallel $\beta$-helix proteins

The structures of the extracellular pectate lyase family, PelC [21], BsPel [26], PelE [27] and PnlA [41], were superimposed using MYNFIT [43] with a cut-off value of 3.0 Å. The sequence identity, after structural alignment, is between 13.6 and 31.8%. There are 152 equivalent $\alpha$-carbons across the four proteins and the root mean square displacement (rmsd) values vary from 0.75 Å for PelE and BsPel to 0.98 Å for PelE and PelC. The parallel $\beta$-helix domains of these lyases superimpose well. In particular, the conformation of the parallel $\beta$-sheets 2 and 3 and the short T2 turn is highly conserved. This region, far from the active site, contains the only truly conserved sequence motif, $v$WiDH, used to recognize enzymes of the extracellular pectate lyase family [44]. The major differences are in the long T3 loops: the T3 loops of neither RGase A nor TSP have detectable sequence similarity with those of the lyase and both have 12 coils of parallel $\beta$-helix, in comparison to lyase's seven. The superimposition of TSP, for which co-ordinates are available, can therefore be done using a number of registers. The fit with the $\alpha$-helices capping the N-terminus of the parallel $\beta$-helix displaced by one turn, but spatially close, gives 67 equivalences and a rmsd of 0.91–1.04 Å. The lyase $\alpha$-carbon co-ordinates show rmsds of 0.62–0.74 Å for these 67 residues, revealing that these are the structurally most conserved residues of the lyases which are in the PB2–PB3 region and have relatively well conserved sequences.

Chothia and Lesk argue that as structures diverge in sequence, the elements of secondary structure move to accommodate changes in the hydrophobic core of the protein [45]. The comparisons above suggest that the parallel $\beta$-helix proteins are unusual in that the main chain co-ordinates align better than would be expected given the low level of sequence identity. We argue below that this arises because the parallel $\beta$-helix fold is stacked and because of the unusual turns between nearly perpendicular $\beta$-sheets.

Side-chain stacking in parallel $\beta$-helix folds

Stacking is perhaps the most striking single characteristic of the pectate lyase fold (Figure 3). Not only are the residues in the $\beta$-sheets stacked but there are other repetitive features such as stacked $\alpha_1$-turns at T2 and even short stacks of residues in $\alpha_R$ conformation just before PB1. Petersen et al. [40] make the useful distinction between residues being aligned (a characteristic property of a $\beta$-sheet) and stacked (meaning that similar residue types have similar side chain conformation in successive turns). The overall effect of stacking is to give structural regularity and help maintain the relative disposition of secondary structure elements. The regularity is suggestive of a modular structure which allows new turns of the helix to be easily incorporated into the parallel $\beta$-helix fold. The observation that
each turn interacts principally with the turns on either side strengthens this suggestion. The ‘like-on-like’ stacks have obvious implications for evolution allowing duplication of a single turn of the superhelix. If DNA sequences are similar then duplication is likely, thus an initial duplication could be followed by subsequent duplications providing that structure and function are conserved. If the binding of a polysaccharide to a parallel β-helix protein is roughly parallel to the helix axis, such duplications might be selected because they would allow the binding of longer substrates. The most obvious amino acid stacks are the offset stacks of the aromatics, the asparagine ladder and aliphatic stacks. Favourable packing of aromatics is achieved by the twisting of the right-handed parallel β-helix which couples the main chain conformation to the side-chain packing. The aromatic stacks may therefore be favoured by the twist of the sheets and have not been observed in the flat left-handed parallel β-helices. Aromatic stacking, often involving tyrosines, is also seen on the outside of the helix. Aliphatic stacks are mostly found inside the parallel β-helix and are probably stabilized by hydrophobic interactions and the ‘cupped stacking’ type of arrangement [33]. The aliphatic stacks tend to be in contact with other unstacked aliphatic side chains so that some segregation of aromatic and aliphatic residues occurs within the right-handed parallel β-helices. In general TSP shows less evidence of stacking than the pectate and pectin lyases. However, the folding of TSP is best characterized by mutagenesis, which has revealed that mutations which convert a broken aliphatic stack into a continuous stack [46] act as suppressors of temperature-sensitive folding mutants.

*α₁* turns at the junction between perpendicular parallel β-sheets

Another obvious characteristic of the right-handed parallel β-helical proteins is that extended regions of the chain often change

Figure 3

The regularity, side-chain stacking and 80° T2 turn of the right-handed parallel β-helix proteins, illustrated using pectin lyase A

The PB2–T2–PB3 region is the most highly conserved in the right-handed parallel β-helix proteins.
direction sharply by about 80° with a single residue in \( \alpha_t \)-conformation interrupting the residues in \( \beta \)-conformation. This offers the possibility, not always exploited, of forming a continuous network of hydrogen bonds through the turn. This turn has been described as a distorted \( \gamma \beta_k \) turn [21] after the classification of Wilmot and Thornton [47]. Chothia and Murzin [25] viewed the turn as a bend in a single continuous \( \beta \)-sheet similar to the \( \beta \)-bulge [19,20], while Pickersgill et al. [26] identify the occurrence of the \( \alpha_t \)-bounded \( \beta \)-strand as a new supersecondary structure motif. This important element of supersecondary structure repeated in consecutive turns of the superhelix allows stacking through the turns and is rare because approximately perpendicular \( \beta \)-sheets such as PB2 and PB3 are uncommon. It is arguably different from an aperiodic \( \beta \)-breaker [48] in that it is a motif repeated in consecutive turns of the superhelix. Stacked \( \alpha \)-bounded \( \beta \)-strands are likely to confer stability both thermodynamically and through evolution.

A natural question is why does the turn include the \( \alpha_t \)-conformation rather than the more common \( \alpha_h \)? Possible answers are that the observed \( \alpha_t \)-turn is more stable or that it occurred by chance and has persisted. What is clear is that the \( \alpha_t \)-linked turn is different. PB2 and PB3 could be connected by an \( \alpha_t \)-turn but this would turn the protein inside out. The least disruptive way of swapping from \( \alpha_l \) to \( \alpha_h \) may be to move the position of the turn residue. Moving the position of the turn can be seen to occur in the half turns discussed by Efimov [49], where the turn \( \beta \beta \alpha_t \beta \) becomes \( \beta \alpha_t \beta \beta \) resulting in a flip of one peptide. However this change is unlikely to have occurred simultaneously in multiple turns during evolution as would be needed to preserve the stacking. Also the \( \beta \alpha_t \beta \beta \) turn observed in \( \beta \)-arches [49] may not be compatible with the perpendicular packing of two parallel \( \beta \)-sheets.

If \( \alpha_t \)-turns give better packing and stacking in the context of parallel \( \beta \)-sheets it could be argued that they have arisen by convergent evolution. However, they can only occur as seen in the right-handed parallel \( \beta \)-helices if the two \( \beta \)-sheets, PB2 and PB3, approach a corner with the pleating and hydrogen bonding in the correct register. An example of reversed hydrogen bonding occurs in the spiral folds such as enoyl-coenzyme A hydratase [37] or 4-chlorobenzoyl coenzyme A dehalogenase [38]. The left-handed parallel \( \beta \)-structures again contain many examples of turns with \( \alpha_t \)-conformation but these turns are flatter with out-turned carboxyls and the internal angle of the turn is more acute. There is nothing to suggest that the left- and right-handed parallel \( \beta \)-helix proteins are related. As the \( \alpha_t \)-turn occurs in both left and right-handed parallel \( \beta \)-helix proteins any preference, as suggested by the current sample of structures, must be based on local interactions. It is clear that the \( \alpha_t \)-turn does not determine the overall hand of the parallel \( \beta \)-helix.

Both the lyases and tailspike protein have stacks of \( \alpha_h \) residues, with side chains pointing inwards, at the start of PB1. We consider it significant that alignment of the T2 region also aligns these stacks on the opposite side of the molecules and note that the structure of RGase A is also apparently similar in this region.

**The right-handed parallel \( \beta \)-helix superfamily**

A number of structural similarities leads us to the conclusion that the pectic lyases, rhamnogalacturonase A, tailspike endorhamnosidase and possibly all right-handed parallel \( \beta \)-helices have evolved from a common ancestor. First and foremost is the conservation in the region of the T2 turn. The parallel \( \beta \)-sheets 2 and 3 have the same pattern of hydrogen bonds and pleating in all structures. It is unlikely that the same hydrogen bonding direction and register of the hydrogen bonds arose by chance, since there are four possible arrangements of two sheets each with two possible hydrogen bonding directions. The hydrogen bonding directions of the two sheets, once selected, are locked into the structure over evolutionary time. It is the residue in \( \alpha_t \)-conformation that ensures consistent register but clearly the overall fold is a more collective property of the sequence. Precisely why the \( \alpha_t \)-turn is so highly conserved is not known. Presumably any variation is unfavourable for the folding or stability of the protein and requires simultaneous substitutions in a number of consecutive turns which is unlikely. It is the locking in of the hydrogen bonding direction of the two sheets and the T2 turn that makes the detection of very remote homologies such as the relationship between pectate lyase, rhamnogalacturonase A and tailspike protein possible. Note that although the conformation of the parallel \( \beta \)-sheets is locked in, the number of turns in the structures is not and the modular construction
means that additional turns may be quite easily inserted accounting for the much longer parallel \( \beta \)-helix of the tailspike protein and RGase A. It seems likely that oligomerization as seen in the tailspike protein could also occur. The \( \alpha \)-linked parallel \( \beta \)-sheets and their stacking may account for the puzzling fact that the pectic lyases, tailspike protein and rhamnogalacturonase A have no identifiable sequence similarity, no obvious similarity in the side chains packed in their hydrophobic cores, very few if any conserved residues at their active sites but have very many close similarities in their main chain conformation.

Superimposition of the T2 turn region of tailspike and BsPel brings a number of amino acids in \( \alpha \)-conformation at the beginning of PB1 into alignment. Clearly the structural constraints are less severe for the transition region from the T3 loop into PB1 than from PB2 into PB3 but it is significant that there is similarity here since it is this region that forms the substrate binding clefts (Figure 2). There is also functional similarity in these enzymes in that they all bind \( \alpha \)-galactose-containing polymers. As RGase A is homologous to polygalacturonase and Biely et al. [50] have established that these are inverting glycosidases, it is likely that both RGase A and TSP are inverting glycosidases and the active site residues identified by Steinbacher et al. [42] may be common to RGase A. There are further characteristics that suggest a common ancestral right-handed parallel \( \beta \)-helix protein including the \( \alpha \)-helix which caps the amino-terminal end of all the parallel \( \beta \)-helices and other similarities such as the N-terminal location of the long T3 loops. RGase A and tailspike have the same number of turns suggesting that they are closely related.

This work was funded by the B.B.S.R.C. of the U.K., M.A.F.F. and E.U. AIR2 projects CT94-1345 and CT92-1272.


Received 22 December 1997

Structure, function and protein engineering of starch-degrading enzymes
T. E. Gottschalk*, H.-P. Fierobe*, E. Mirgorodskaya†, A. J. Clarke*, D. Tull†, B. W. Sigurskjold‡,
T. Christensen§, N. Payrell, T. P. Frandsen*, N. Juge*, K. A. McGuire*, S. Cottaz‡, P. Roepstorff‡, H. Driguez‡,
G. Williamson† and B. Svensson*†
*Department of Chemistry, Carlsberg Laboratory, Gamle Carlsberg Vej 10, DK-2500 Copenhagen, Valby, Denmark,
†Institute of Molecular Biology, Odense University, Campusvej 55, DK-5230 Odense M, Denmark,
‡Department of Microbiology, University of Guelph, Ontario N1G 2W1, Canada,
§August Krogh Institute, University of Copenhagen,
Universitetsparken 13, DK-2100 Copenhagen Ø, Denmark,
†Centre de Recherches sur les Macromolécules Végétales, 38041 Grenoble 9 Cedex, France,
and *Institute of Food Research, Norwich Research Park, Norwich NR4 7UA, U.K.

Introduction
Information on protein–substrate interactions at atomic resolution and on the mechanism of action in polysaccharide hydrolases is rapidly increasing [1–5]. A total of 62 structural families of glycosyl hydrolases has been reported [6,7]. However, amylolytic or starch degrading and related enzymes currently are represented in only five families; the very large α-amylase family consisting of retaining enzymes with broad variation in specificity [8]; two small families, 14 and 15, which contain the inverting exo-acting enzymes of narrow specificity, the β-amylases [9] and the glucoamylases [10,11], respectively; the retaining α-glucosidases of family 31 [12]; and the small family 57 of extremophile α-amylases [13,14]. Certain members of families 13, 14 and 15 have an extra domain typically as a C-terminal extension of the catalytic domain, which specifically adsorbs on to starch granules, and binds larger soluble polysaccharides and β-cyclodextrin [15,16]. Although the starch-degrading and related enzymes contain additional domains, a function has only been identified for the catalytic