

Insight into Ligand Diversity and Novel Biological Roles for Family 32 Carbohydrate-Binding Modules

D. Wade Abbott,*¹ José María Eirín-López,*^{†1} and Alisdair B. Boraston*

*Department of Biochemistry and Microbiology, University of Victoria, Victoria, Canada; and [†]Departamento de Biología Celular y Molecular, Universidade da Coruña, A Coruña, Spain

Family 32 carbohydrate-binding modules (CBM32s) are found in a diverse group of microorganisms, including archaea, eubacteria, and fungi. Significantly, many members of this family belong to plant and animal pathogens where they are likely to play a key role in enzyme toxin targeting and function. Indeed, ligand targets have been shown to range from insoluble plant cell wall polysaccharides to complex eukaryotic glycans. Besides a potential direct involvement in microbial pathogenesis, CBM32s also represent an important family for the study of CBM evolution due to the wide variety of complex protein architectures that they are associated with. This complexity ranges from independent lectin-like proteins through to large multimodular enzyme toxins where they can be present in multiple copies (multimodularity). Presented here is a rigorous analysis of the evolutionary relationships between available polypeptide sequences for family 32 CBMs within the carbohydrate active enzyme database. This approach is especially helpful for determining the roles of CBM32s that are present in multiple copies within an enzyme as each module tends to cluster into groups that are associated with distinct enzyme classes. For enzymes that contain multiple copies of CBM32s, however, there are differential clustering patterns as modules can either cluster together or in very distant sections of the tree. These data suggest that enzymes containing multiple copies possess complex mechanisms of ligand recognition. By applying this well-developed approach to the specific analysis of CBM relatedness, we have generated here a new platform for the prediction of CBM binding specificity and highlight significant new targets for biochemical and structural characterization.

Introduction

Carbohydrate-binding modules (CBMs) are critical components of microbial carbohydrate active enzymes (CAZymes). Primarily, CBMs function in nature by concentrating the parent enzyme at their dedicated substrates thereby enhancing its catalytic efficacy (Boraston et al. 2004). This process helps the enzyme to degrade refractory substrates such as insoluble structural polysaccharides and to target complex substrates resident within eukaryotic extracellular glycans. Originally, CBMs were classified as cellulose-binding domains because the first documented examples interacted preferentially with crystalline cellulose (Gilkes et al. 1988; Tomme et al. 1988; Van Tilbeurgh et al. 1989). More recently, however, the roles of CBMs in more exotic processes such as animal and plant pathogenesis (Boraston et al. Forthcoming) and the metabolism of more complex organisms, including plants and animals, have emerged (Machovic and Janecek 2006a). Elucidating the potential function(s) of these modular components in such diverse biological roles is now under intensive investigation.

Although, the classical definition of a CBM includes the constraint that they "... are found within the primary structures of carbohydrate-active enzymes..." (Boraston et al. Forthcoming), several exceptions to this paradigm now have been reported (Charnock et al. 2002; Flint et al. 2004; Vaaje-Kolstad et al. 2005; Abbott et al. 2007). For example, *Yersinia enterocolitica* contains a family 32 carbohydrate-binding module (CBM32) (YeCBM32) that is an independent periplasmic protein that interacts with polygalacturonic acid (Abbott et al. 2007). This interaction appears to retain polygalacturonic acid within the periplasm

where it becomes a substrate for resident depolymerases. The oligogalacturonide products of these reactions are the preferential ligands for TogB (Abbott and Boraston 2007), the specificity determinant of the intracellular TogMNAB transporter. Reconstruction of the pathway indicates, therefore, that ligand and substrate selectivities of these proteins support a funneling mechanism initiated by YeCBM32 to direct pectin breakdown products into the cell and prevent their escape back into the environment.

Although the most closely related proteins to YeCBM32 (CBM32s from *Yersinia intermedia*, *Yersinia pestis*, *Yersinia pseudotuberculosis*, *Erwinia carotovora*, and *Vibrio vulnificus*) retain a conserved role in pectin utilization, the majority of CBM32s are appended to a wide variety of enzymes and in many cases can be detected in multiple copies within the same enzyme. This observation underlines a diverse functional repertoire for this family, and many CBM32s are likely to be directly involved in pathogenesis. For example, CBM32s are found in *Clostridium perfringens*, *Burkholderia cepacia*, *Bacteroides* sp., and *Streptococcus pneumoniae*; however, in most cases direct roles for these proteins in virulence remain to be determined. To date, CBM32s have been shown to bind a variety of carbohydrate ligands displaying a signature galactose-configured moiety (i.e., axial C4). These include galactose, LacNAc (β -D-galactosyl-1,4- β -D-N-acetylglucosamine), type II blood group antigen H-trisaccharide, and polygalacturonic acid (Newstead et al. 2005; Ficko-Blean and Boraston 2006; Abbott et al. 2007). However, the full spectrum of biological ligands is likely to be much larger as these proteins are found in a milieu of biochemically uncharacterized modular enzymatic architectures with diverse predicted activities.

In order to characterize the relatedness of family 32 CBMs and provide a framework for predicting the diverse structure-function relationships of these proteins, we have performed a rigorous analysis on sequences available within the CAZy database (Coutinho and Henrissat

¹ These authors contributed equally to this work.

Key words: carbohydrate-binding module, family 32, evolution, multimodularity, structural characterization.

E-mail: wabbott@uvic.ca.

Mol. Biol. Evol. 25(1):155–167. 2008

doi:10.1093/molbev/msm243

Advance Access publication November 20, 2007

1999). These data provide a research platform for predicting the ligand specificities of CBM32s based upon a functional classification and also highlight CBM32s with distinctive biological roles. To the best of our knowledge, this approach has only been reported for starch-binding CBMs from family 20 and 21, which display a stringently conserved pattern of ligand binding (Machovic et al. 2005; Machovic and Janecek 2006a, 2006b). In addition to presenting the overall phylogeny of CBM family 32, we have focused on the multimodularity present within *C. perfringens* and *Saccharophagus degradans* proteins and the independent CBM32s from *Myxococcus xanthus* and *Y. enterocolitica*. When possible, we have included the structural information garnered by crystallographic studies and highlighted strategic avenues for future biochemical and structural studies.

Materials and Methods

Module Boundary Determination

A total of 180 polypeptide sequences for carbohydrate-binding modules belonging to different species have been used in our analyses (Supplementary Materials online), including 4 outgroup sequences from eukaryotes. Complete protein sequences were retrieved from the CAZy Database (<http://www.cazy.org/>). Domain architecture was determined using a complementary analysis of IntroProScan (Zdobnov and Apweiler 2001) and PSI-Blast (Altschul et al. 1997) platforms. In some cases, the selection of domain boundaries was fine-tuned by direct sequence alignments with ClustalW (Thompson et al. 1994).

Evolutionary Analysis of CBM32 Modules

Nucleotide coding sequences for the modules analyzed in the focused examples were aligned on the basis of their translated amino acid sequences using the BioEdit (Hall 1999) and MEGA ver. 3.1 (Kumar et al. 2004) programs with default parameters. The alignment for the complete set of CBM sequences consisted of 245 amino acid sites, and a bar chart representation was used in order to represent the frequency of each residue at every position of the alignment, using the LogoBar program (Perez-Bercoff et al. 2006).

All molecular evolutionary analyses in the present work were carried out using the program MEGA ver. 3.1 (Kumar et al. 2004). The extent of nucleotide and amino acid variation between sequences was estimated by means of the uncorrected differences (p -distance) as this distance is known to give better results than more complicated methods when the number of sequences is large and the number of positions used is relatively small (Nei and Kumar 2000).

The numbers of synonymous (p_S) and nonsynonymous (p_N) nucleotide differences per site were computed using the modified Nei–Gojobori method (Zhang et al. 1998), providing in both cases the transition/transversion ratio (R). Although saturation in synonymous sites is possible, p_S values have been used in the present work in order to ascertain the nature of the selective process operating on different modules, as well as to perform Z -test of selection

locally from comparisons between modules within species. In this background, saturation in synonymous sites could be discarded.

Distances were estimated using the pairwise-deletion option and standard errors were calculated by the bootstrap method with 1,000 replicates. The presence and nature of selection was tested in CBM modules by using the codon-based Z -test for selection, establishing the alternative hypotheses as either $H_1: p_N < p_S$ or $H_1: p_N > p_S$ and the null hypothesis as $H_0: p_N = p_S$ (Nei and Kumar 2000). The Z -statistic and the probability that the null hypothesis is rejected were obtained, indicating the significance level as $**P$ ($P < 0.001$) and $*P$ ($P < 0.05$).

The Neighbor-Joining tree-building method (Saitou and Nei 1987) was used to reconstruct the phylogenetic trees. In order to assess that our results are not dependent on this choice, phylogenetic inference analyses were completed by the reconstruction of a maximum parsimony tree (Rzhetsky and Nei 1992a) using the close-neighbor-interchange search method with search level 1 and with 10 replications for the random addition trees option. We combined the bootstrap (Felsenstein 1985; Efron et al. 1996) and the interior-branch test methods (Rzhetsky and Nei 1992a; Sitnikova 1995) in order to test the reliability of the obtained topologies, producing the bootstrap probability (BP) and the confidence probability (CP) values for each internal branch, assuming $BP > 80\%$ and $CP \geq 95\%$ as statistically significant (Sitnikova et al. 1995; Rzhetsky and Nei 1992b). The CBM module associated with the galactose oxidase enzyme in 3 species of fungi was used as outgroup in the reconstruction.

The analysis of the nucleotide variation across coding regions was performed using a sliding-window approach, by estimating the total (π) and the synonymous (π_S) nucleotide diversity (average number of nucleotide differences per site between 2 sequences) with a window length of 20 bp and a step size of 5 bp (for π) and a window length of 5 bp and a step size of 1 bp (for π_S). The codon usage bias in genes encoding CBM32-containing molecules was estimated as the effective number of codons (Wright 1990), where the highest value (61) indicates that all synonymous codons are used equally (no bias) and the lowest (20) that only a preferred codon is used in each synonymous class (extreme bias). Both analyses were conducted with the program DnaSP ver. 4.10 (Rozas et al. 2003).

Results and Discussion

Overall Evolutionary Relationships in CBM Family 32

The phylogenetic tree presented in figure 1 depicts the evolutionary relationships among family 32 CBMs deposited within the CAZy database (Coutinho and Henrissat 1999). At the time of preparation of this paper, there were 202 available polypeptide sequences, which included representatives from archaeobacteria, eubacteria, and eukaryotes (the later ones being used to root the tree). These sequences contain the entire primary structure for an enzyme containing at least 1 copy of a CBM32. Redundant sequences within the same species from closely related strains have been filtered in order to simplify the analysis.

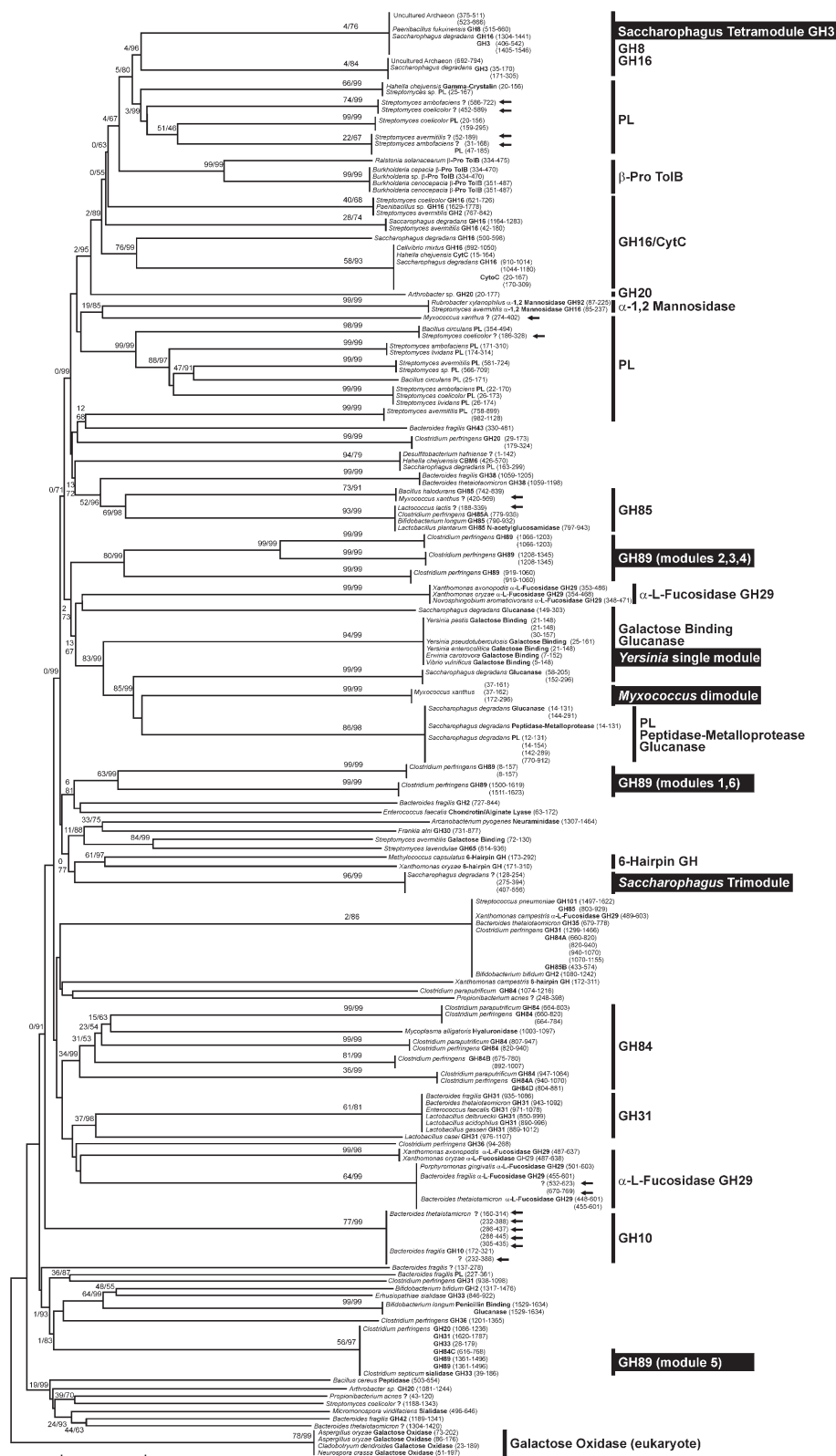


FIG. 1.—Phylogenetic relationships of CBM32s. The reconstruction was carried out by calculating the evolutionary amino acid *p*-distances from the sequences of all the organisms analyzed. CBM-associated enzymes are indicated on the right-hand side near the species names and the modules are indicated by the interval they occupy in the CBM sequence (within parentheses). Numbers for interior nodes indicate BP and CP values. Confidence values were based on 1,000 replications and are only shown when the value is greater than 50%. CBMs whose specific enzyme is not known are indicated by an interrogation symbol (?). Black arrows highlight those sequences whose specific enzyme could be inferred from the groups defined by the present topology.

The complete list of sequences, including their descriptions of modular boundaries is shown in Supplementary Materials online.

Overall, CBM32 sequences display a clustering pattern that parallels the activities of the catalytic module in which the module is appended to. These enzyme groups, which include family 8, 10, 16, 29, 31, 84, 85, and 89 glycoside hydrolases (GHs); pectate lyases (both pectate lyase and β -helix and β -propeller TolB folds); an unclassified metalloprotease; α -1,2 mannosidases; and the galactose oxidases, are indicated in the right margin. Although many of these enzymes remain to be characterized biochemically, their predicted activities span a wide range of substrates, including chitin, α - and β -glucans, mannose, fucose, and hexosamines. In addition to the catalytically active proteins, the galacturonic acid-binding modules (i.e., *Yersinia* sp. single modules) cluster into a grouping that is independent of a catalytic module. Such heterogeneity within a CBM family is in contrast to other previously reported phylogenetic analyses of the starch-binding CBMs from family 20 and 21. These closely related families fall into a clan that also includes more distantly related starch-binding CBM families 25, 26, 34, 41, and 45 (Machovic et al. 2005; Machovic and Janecek 2006a). Although within CBM20s and CBM21s there is some variation in the modular position of the starch-binding domain within the context of the entire protein sequence, these CBMs are all predicted to bind α -1,4 glucans and in most cases are associated with an enzymatic module involved in the biosynthesis or utilization of starch (Machovic and Janecek 2006a, 2006b). In this regard, we believe that the CBM32 phylogeny presented here may be useful for both predicting potential target ligands and understanding some of the biological processes involved in selecting for highly modular CAZymes.

As revealed by the branch lengths within the CBM family tree (fig. 1), there is significant amino acid variation among members of the family 32 at all levels, with the peculiarity that the extent of this protein variation is roughly the same within and between species, as well as within and between enzyme groups. Surprisingly, a large variation is evident between different modules associated with a particular enzyme in a given species, resulting in an interspersed distribution of CBM sequences across the phylogeny. This property is evidenced in the modules from the *C. perfringens* isozymes from both family GH84 (GH84A-E) and GH85 (GH85A-B) enzymes, which contain some of the most divergent CBM32 sequences in the analysis.

In some cases, a noticeable difference in relatedness is also evident for modules associated with homologous enzymes from different species. For example, there are 2 divergent pectate lyase enzyme clusters, which appear to have been diverging at a fast rate for a long time period rather than suggesting the possibility of a polyphyletic origin. However, it is possible that the differentiation within these 2 groups has been influenced by unique selective constraints acting on the different modules associated with the catalytic module. For example, pectate lyases operate by a β -elimination mechanism to depolymerize pectic substrates found within primary plant cell walls (Marin-Rodriguez et al. 2002). Although pectin primarily comprises α -1,4 linked galacturonic acid backbone, it is a

heterogeneous polysaccharide that also contains rhamnose, arabinose, and galactose moieties, and there are various modifications that can be presented including C6 methoxyl esters and C2/C3 acetylations (Willats et al. 2001). Selection on these 2 clusters of CBM32s therefore may be based upon the degree of polymerization and asymmetry of the galacturonic acid residues not only within the pectic backbone but also toward regions of carbohydrate heterogeneity.

One of the greatest challenges in CBM research is determining ligand specificity. This is a critical first step toward characterizing the molecular determinants of protein-carbohydrate complex formation. In this respect, the phylogenetic analysis presented here is a novel approach and potentially invaluable tool for streamlining the experimental investigation of uncharacterized CBM32s. In several cases, CBM32s that are components of enzymes with unknown activities cluster with CBMs appended to predicted or characterized enzymes (fig. 1—highlighted with black arrows). For example, several lactic acid bacteria have 1 CBM32 member defined within the database. These CBMs cluster into 2 very different enzyme groupings. The GH31 (α -glucosidase) group consists of *Lactobacillus delbrueckii*, *Lactobacillus acidophilus*, *Lactobacillus gasseri*, and *Lactobacillus casei*, whereas the GH85 (mannosyl glycoprotein endo- β -*N*-acetylglucosaminidase) group contains *Lactobacillus plantarum* (Lp_0182) and the uncharacterized protein from *Lactococcus lactis* subsp. *lactis* (YpcC). Sequence comparisons between these 2 closely related proteins reveal that the *L. lactis* CBM32 is an independent acting protein, which aligns very well with the C-terminal portion of the *L. plantarum* enzyme (42% identity—not shown). This observation provides an informed avenue for investigating the mechanisms of carbohydrate recognition and utilization by these currently uncharacterized proteins.

In addition to predicting the specificity of CBM32s appended to unknown catalytic modules, this approach is also helpful for classifying enzyme architectures that contain more than 1 CBM. This concept can be clearly demonstrated for both “homogeneous” and “heterogeneous” clustering patterns. Homogeneous clustering occurs when 2 or more CBM32s from the same enzyme display the most similarity to each other. For example, the GH3 from *S. degradans* contains 4 tandemly arranged CBM32s that cluster very closely within the phylogeny, an observation that may indicate overlapping ligand specificities. Heterogeneous clustering on the other hand refers to the occurrence of a module from a multimodular CBM32 protein displaying the most similarity to a CBM32 from a different enzyme grouping. This is the case of the GH89 α -*N*-acetylglucosaminidase from *C. perfringens*, which has 6 appended CBM32s that cluster into 3 distinct groups. An additional complication comes from species-specific clustering patterns (i.e., modules GH20, GH31, GH33, and GH84C from *C. perfringens*) or by an enzyme-specific pattern (i.e., Fucosidase GH29 in *Porphyromonas*, *Bacteroides fragilis* and *Bacteroides thetaiotaomicron*). The high variation presented between modules even within the same species and in the same enzyme class masks the phylogenetic relationships in some cases. In general terms, however, it is possible to define the different lineages across the phylogeny. The differences in amino acid composition of these modules

will likely translate into differences in ligand-binding specificities and/or affinities. In the following section, we will explore the phenomenon of multimodularity in greater detail as it relates to homogeneous (GH3) and heterogeneous (GH89) dispersion of CBMs.

Analysis of the GH89 Multimodular Enzyme from *C. perfringens*

CpGH89 is a 239.5-kDa protein that has a complicated modular architecture consisting of an α -*N*-glucosaminidase catalytic module (215–911), 6 CBM32 modules (1: 8–157, 2: 919–1060, 3: 1066–1203, 4: 1208–1345, 5: 1361–1496, 6: 1511–1623), and 3 C-terminal modules possibly involved in protein–protein complex formation (fig. 2A). Although the biological role of this enzyme in virulence has yet to be determined, the human homolog is involved in the lysosomal degradation of heparan sulfate and mutations within its gene have been causally linked to Sanfilippo B syndrome (Weber et al. 1996).

To determine the relatedness between the 6 modules of GH89, the sequences at both the amino acid and nucleotide levels were analyzed. The variation between protein sequences is very high although the numbers of synonymous and nonsynonymous substitutions per site are equivalent. Indeed, only 5 out of a total 163 positions are conserved in every sequence (table 1; fig. 2B). This divergence in amino acid composition has 2 implications. Firstly, there is plasticity in amino acid selection involved in the formation of secondary structure elements, as variation is abundant throughout the β -structures. Secondly, there is poor conservation of functional residues implicated in ligand recognition, suggesting that different modules interact with different ligands—or perhaps similar ligands through different mechanisms. For example, when GH89 module 5 is compared with the remaining GH89 modules, only a single arginine residue predicted to be involved in galactose binding (Ficko-Blean and Boraston 2006) is strictly conserved with module 2 (fig. 2B).

Closer analysis of the CpGH89 CBM32 modules reveals the presence of 3 distinct modular clusters: 2, 3, and 4; 1 and 6; and module 5 (fig. 1). This scenario is an example of heterogeneous multimodularity, where CBM32s that cluster differently within the phylogeny are present within the same enzyme. The cluster containing modules 2, 3, and 4 has an average amino acid difference per site that ranges from 0.34 to 0.71 (meaning that 34–71% of the time an amino acid is conserved), with modules 3 and 4 displaying the highest similarity (0.34 amino acid differences per amino acid site). In the overall phylogeny, these modules group with the GH85 CBMs and the GH29 fucosidase CBMs from *Xanthomonas* sp. The module 1 and 6 clusters have an amino acid difference of 0.68 per site and group with the CBM32s from the *B. fragilis* GH2 and *Enterococcus faecalis* chondroitin lyase. Module 5 displays high similarity to several other modules from different *C. perfringens* enzymes including the single modules from GH33 and GH84C, and module 3 from both GH20 and GH31, which suggests that a mechanism based on the horizontal transfer of genetic material (very common in bacterial evolution) could contribute to a certain extent to the

homologies detected between modules from different CBMs in this species. The structure of the GH84C CBM32 has been solved enabling the comparison of the secondary structures between these modules (Ficko-Blean and Boraston 2006). Within these modules, amino acid variation is most common within the loop regions of the β -sandwich (fig. 3A). A noticeable insertion appears within loop 6 for GH31 module 3, which is distal to the predicted binding site, and loop 7 of GH84C and GH31 module 3 that encodes residues shown to be exclusive determinants of LacNAc (Ficko-Blean and Boraston 2006). Importantly, amino acids required for galactose binding are stringently conserved among this group (Ficko-Blean and Boraston 2006) (fig. 3A and B), suggesting that this cluster of modules may be important for the recognition of galacto-configured carbohydrates. From an evolutionary perspective, accommodating the binding site within these loop regions may allow for selective and rapid transformation of protein structure leading to variance in ligand specificity without compromising the core architecture of the protein (analogous to antibodies). This is potentially a powerful mechanism to increase the diversity of protein–carbohydrate interactions through subtle changes in amino acid composition.

Another key observation about these enzyme architectures is their modular positioning. The related sequences are found in a different order along the enzyme coordinate. Within CpGH84C, CpGH31, CpGH89, and CpGH20, the module is located C-terminal of the catalytic domain at various distances, and for CpGH33, the CBM32 is N-terminal (fig. 3C). Although this observation may simply underline the inherent flexibility of the connecting regions between modules, its complete functional significance awaits structural characterization that will define the modular positions in 3-dimensional space.

The Tri- and Tetramodule CBM32 Proteins from *S. degradans*

Saccharophagus degradans is a glycophilic marine bacterium. This organism displays a remarkable ability to degrade at least 10 stereochemically unique complex carbohydrates from algal, plant, and invertebrate sources (Taylor et al. 2006). Included in its impressive arsenal of CAZymes are 128 predicted GHs, 36 polysaccharide lyases, 15 carbohydrate esterases, and 133 CBMs, including 25 members from family 32.

Similar to *C. perfringens*, CBM32 multimodularity within *S. degradans* enzymes is commonly observed. Examples include a tetramodular CBM32 appended to a GH3 (SdGH3) catalytic module, an independent trimodular protein, and 6 other enzymes with 2 CBM32 copies (Supplementary Materials online). In order to characterize the relatedness of 2 distinctly different polypeptides containing family 32 CBMs, we analyzed the phylogenies of the independent trimodule and SdGH3.

The independent trimodule contains modular boundaries at amino acid positions 128–254 (module 1), 275–394 (module 2), and 407–556 (module 3). It is difficult to predict the function of this protein for 3 reasons. 1) They lack a catalytic appendage; 2) they cluster within a section of the tree that contains CBMs appended to

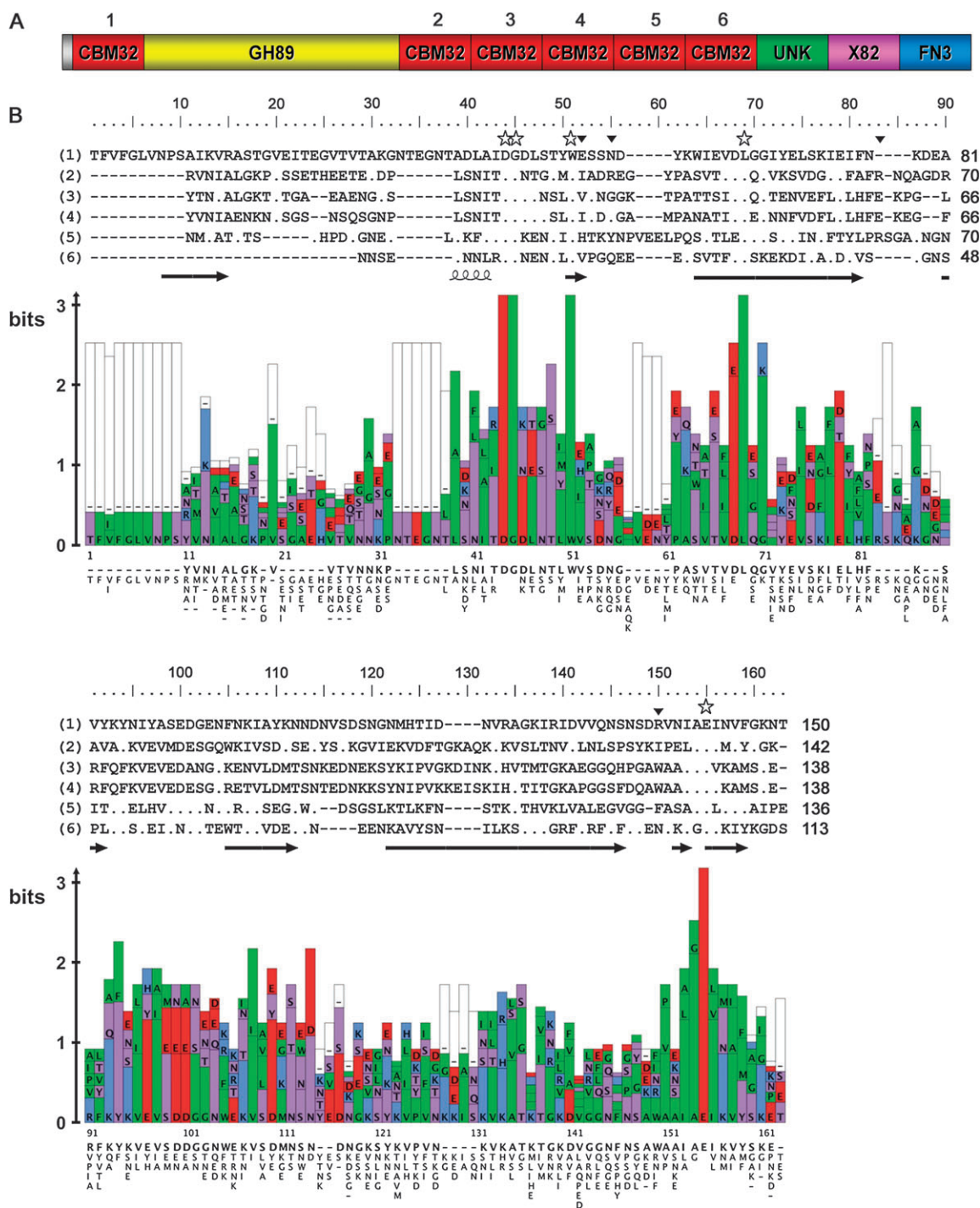


FIG. 2.—Structure of the GH99 from *Clostridium perfringens*. (A) Modular architecture of the multimodular enzyme containing 6 CBM32s. (B) Amino acid alignment of the 6 CBMs associated with the GH99 enzyme in *C. perfringens*. Alignment positions are indicated above the sequences; dots and dashes indicate matching respect to the reference sequence and indels, respectively. White stars above the alignment indicate those positions conserved across modules and black triangles indicate the amino acids involved in galactose binding for module 5 (Ficko-Blean and Boraston 2006). The frequency of each residue at different positions is proportionally represented in the logos bar chart graph below the alignment. Secondary structures were determined from the 3-dimensional structure of GH89_5 structural alignment with the CBM32 from GH84C (PDB ID: 2J1A).

catalytic modules displaying a wide variety of predicted activities (fig. 1); and 3) these sequences are very distantly related to the structurally characterized CBM32s that compromises the reliability of binding site residue identification.

The SdGH3 also has 3 contiguously arranged CBM32s at its N-terminus: module 1 (35–170), 2 (171–305), and 3 (406–542). Module 4 (1405–1546) is separated from this cluster by the catalytic module housed within the protein core (fig. 4A). The functional implications for this

Table 1
Average Numbers of Amino Acid (p_{AA} , upper diagonal) and Nucleotide (p_{NT} , lower diagonal) Differences per Site and Average Nonsynonymous (p_N , upper diagonal boldface) and Synonymous (p_S , lower diagonal boldface) Differences per Site between Modules Associated with the GH89 Enzyme in *Clostridium perfringens*

	1	2	3	4	5	6
1	—	0.82/0.58	0.87/0.60	0.82/0.59	0.74/0.56	0.68/0.45
2	0.56/ 0.50	—	0.71/ 0.50	0.69/ 0.51	0.82/ 0.62	0.75/ 0.53
3	0.58/ 0.50	0.50/ 0.48	—	0.34/ 0.20	0.87/ 0.61	0.85/ 0.62
4	0.57/ 0.51	0.53/ 0.58	0.25/ 0.43	—	0.88/ 0.60	0.84/ 0.61
5	0.55/ 0.52	0.59/ 0.49	0.57/ 0.43	0.60/ 0.56	—	0.81/ 0.60
6	0.47/ 0.55	0.54/ 0.58	0.61/ 0.56	0.59/ 0.51	0.58/ 0.51	—

modular architecture are difficult to determine without appropriate structural information and complementary biochemical analysis of the contributions of each CBM32 to binding; however, it seems possible that the C-terminal module may assist in anchoring and perhaps orientating the enzymatic module at the substrate surface.

Although all 4 of these modules cluster together within the overall family topology, there are clearly 2 subgroup-

ings: modules 1 and 2 sharing 48% identity and modules 3 and 4 sharing 37% identity. The predicted substrate for the family 3 GH is cellulose (β -linked glucan), one of the main structural polysaccharides found within plant cell walls. Although cellulose adopts a relatively simple structure (β -1,4 linked glucose within a linear, fibrous superstructure), the presence of 4 modules within the enzyme with potentially 2 distinctively tailored binding specificities suggests that the enzyme may be targeted to different substructures within the polysaccharide. Although it is tempting to speculate about this, characterization of the CBM32s with SdGH3 is required before appropriate conclusions can be drawn.

The trimodule and GH3 tetramodule from *S. degradans* are excellent examples of homogeneous multimodularity, which implies that all the CBM32 components within a protein show higher similarity to each other than to those found within other proteins (table 2). This is in direct contrast to the heterogeneous multimodularity observed within the more complex CpGH89 enzyme. Whether amino acid homogeneity at the structural level translates into overlapping ligand-binding profiles at the functional level remains to be determined. The CBMs from GH3 tetramodule would

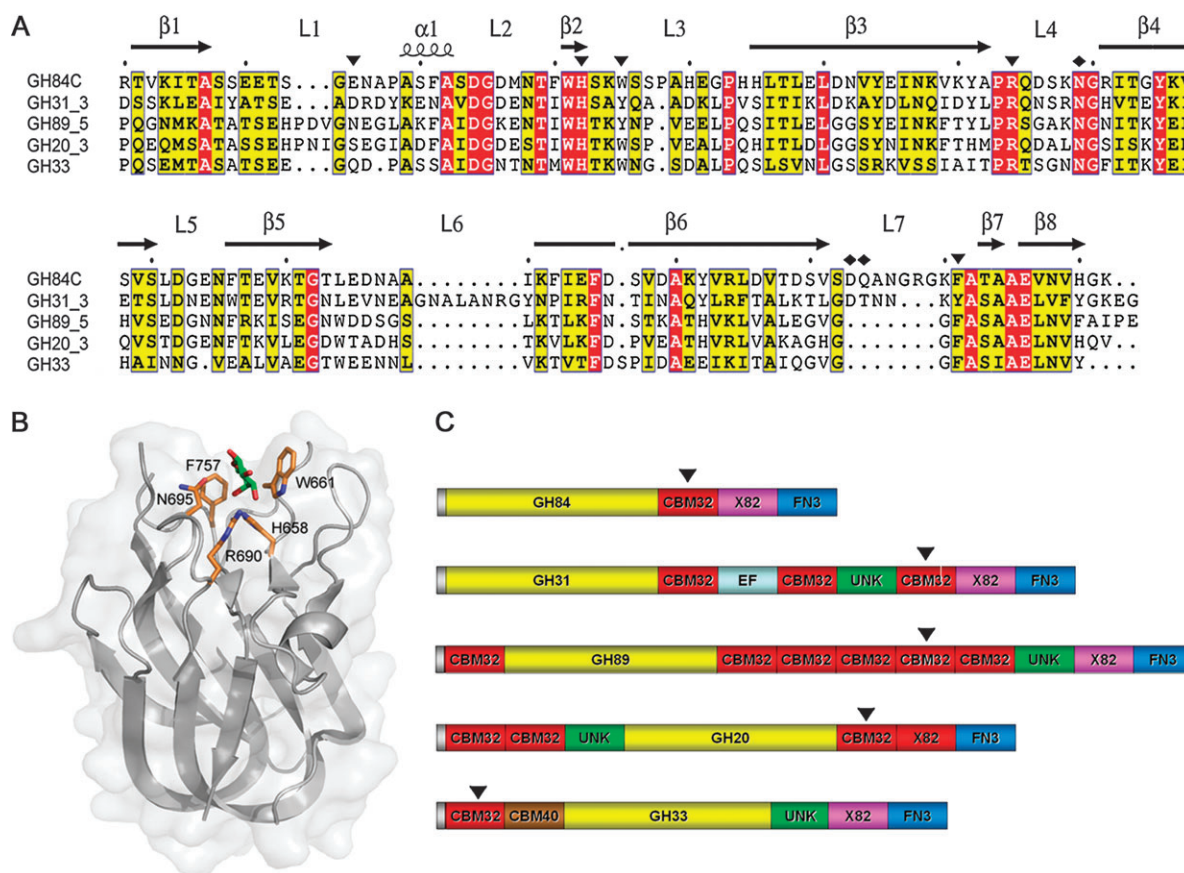


FIG. 3.—Structural analysis of closely related CBM32s from different enzymes in *Clostridium perfringens*. (A) Structural alignment of homologous CBM32s using ESPRIPT (Gouet et al. 2003) and the coordinates from GH84C CBM32 (PDB ID: 2J1A). Secondary structure elements are indicated above the amino acid sequence. Black triangles indicate the amino acids involved in galactose binding and the black diamonds *N*-acetylglucosamine binding. (B) Three-dimensional structure of GH84C CBM32 indicating the conserved amino acids involved in binding galactose. Amino acids are depicted in orange and galactose in green. (C) Modular structures of enzymes from *C. perfringens*. The module displaying homology is indicated with a black triangle.

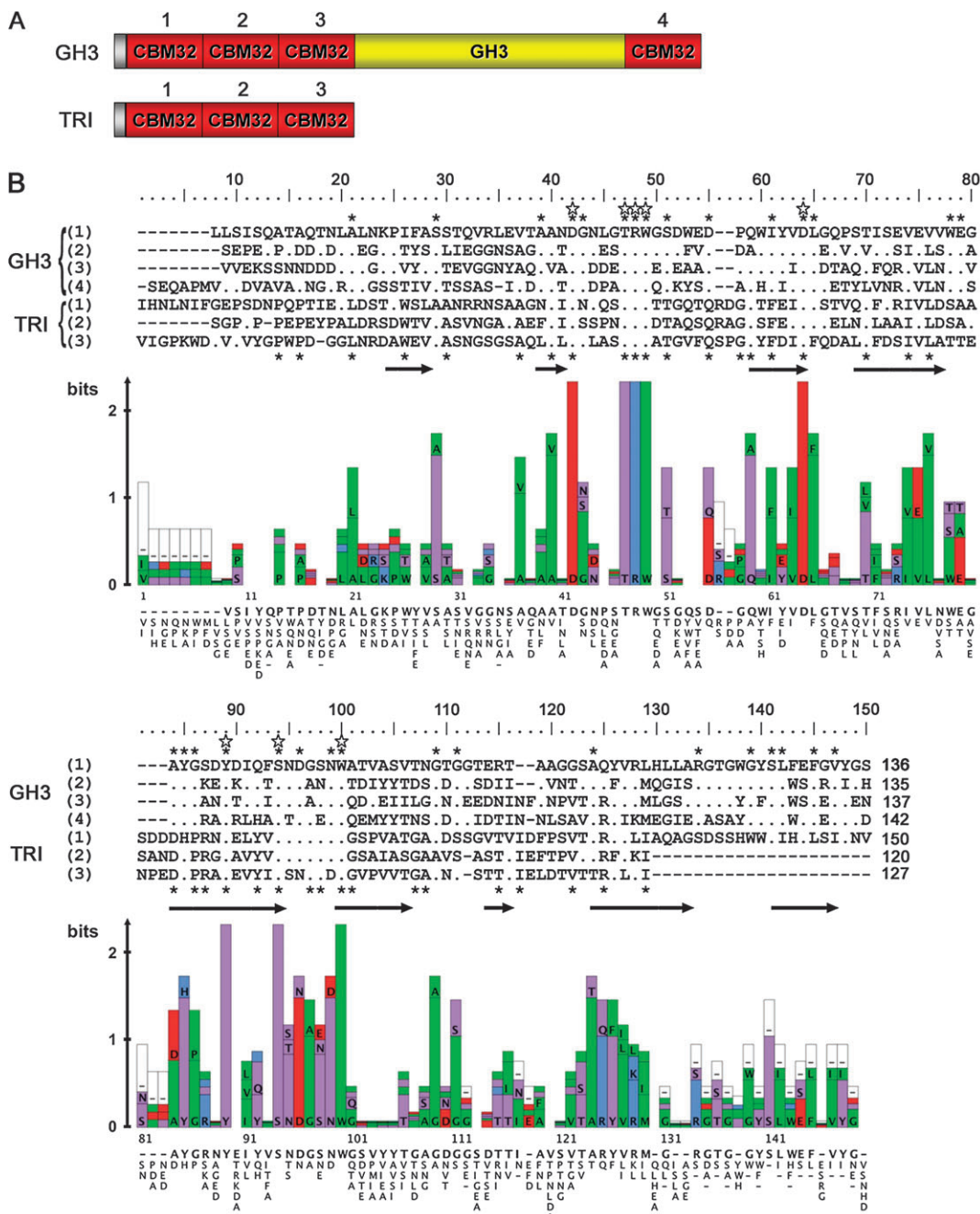


FIG. 4.—Structure of the GH3 and trimodular protein from *Saccharophagus degradans*. (A) Modular architecture of the proteins. (B) Alignment of the sequences constituting the tetramodule and trimodule CBMs in *S. degradans* represented as in figure 2. Asterisks above and below the alignment indicate those positions with conserved residues in the tetramodule and the trimodule, respectively. White stars above the alignment indicate conserved positions across all sequences. Predicted secondary structure (Cuff et al. 1998) is indicated below the sequences.

make excellent targets for future structural studies as they represent some of the most distantly related sequences within the tree to any of the current CBM32s with a known 3-dimensional structure.

The Independent CBM32s from *Y. enterocolitica* and *M. xanthus*

The galacturonic acid-binding proteins introduce another novel mechanism in CBM32 function. These pro-

teins localize within the periplasm of several gram-negative enteric animal pathogens and operate within a pectin utilization pathway as an independent protein (Rodionov et al. 2004; Abbott et al. 2007). In this regard, the galacturonic acid-binding protein group is more functionally representative of a lectin rather than a traditional CBM. Recently, the CBM32 from *Y. enterocolitica* was structurally and biochemically characterized (Abbott et al. 2007). This protein interacts with polygalacturonic acid, a central component of pectin, and operates to retain

Table 2
Average Numbers of Amino Acid (p_{AA}) and Nucleotide (p_{NT}) Differences per Site and Average Synonymous (p_S) and Nonsynonymous (p_N) Differences per Site within CBMs from *Saccharophagus degradans* (trimodule/tetramodule), *Myxococcus xanthus*, and *Yersinia* species

	p_{AA} (SE)	p_{NT} (SE)	p_S (SE)	p_N (SE)	R	Z-test	P
Trimodule	0.58 ± 0.03	0.52 ± 0.02	0.80 ± 0.03	0.42 ± 0.03	0.6	8.18	0.00**
Tetramodule	0.68 ± 0.02	0.48 ± 0.02	0.63 ± 0.02	0.44 ± 0.03	0.8	5.78	0.00**
<i>Myxococcus</i>	0.36 ± 0.03	0.24 ± 0.02	0.31 ± 0.04	0.22 ± 0.01	0.5	2.30	0.01*
<i>Yersinia</i>	0.10 ± 0.02	0.14 ± 0.01	0.38 ± 0.03	0.06 ± 0.01	1.4	8.47	0.00**

The average transition/transversion ratio used in the estimation of p_S and p_N is denoted as R . SE indicates standard errors calculated by the bootstrap method with 1,000 replicates. $p_S = p_N$ was defined as the null hypothesis (H_0) and $p_S > p_N$ as the alternative hypothesis (H_1) in all comparisons (1-tailed tests). The significance level is indicated as * ($P < 0.05$) and ** ($P < 0.001$).

oligogalacturonide substrates within the periplasm where they can be degraded by enzymes, generating suitable products for intracellular transport through the TogMNAB transporter (Hugouvieux-Cotte-Pattat et al. 2001; Hugouvieux-Cotte-Pattat and Reverchon 2001; Abbott et al. 2007). Although a YeCBM32–oligogalacturonic acid complex is currently lacking, mutagenesis, affinity gel electrophoresis, and sequence analysis determined that the binding site is a pocket lined with basic amino acids that interact with internal residues of a polygalacturonic acid chain (unpublished results) (fig. 5).

The evolution of these proteins is of particular interest toward understanding the overall phylogeny of CBM32s as this group is nestled between 2 of the 3 modular clusters from CpGH89 (modules [2–4] and [1, 6]) (fig. 1). As described above, the cognate ligands for these clusters are currently unknown so direct comparisons of ligand-binding profiles is not possible. Potential similarities in ligand specificity between the CBM32s of CpGH89, an enzyme predicted to be active upon the α -1,4 linked gluco-configured monosaccharides of heparan sulphate (*N*-acetylglucosamine and glucuronic acid), and the *Yersinia* sp. modules, which bind α -1,4 linked galacto-configured (O4 axial) polysaccharides, is not readily clear. Possibly, these proteins may utilize a similar mechanism of binding directed toward the common equatorial C5 carboxylate groups. Experiments have been initiated to help illuminate this relationship.

In proximity to the galacturonic acid-binding cluster, there is another class of independent CBMs belonging to the bacterium *M. xanthus* (table 2; fig. 5A). Within these organisms, there are 2 closely related CBM32-containing proteins (52% identity) that also operate independently of a catalytic module: one of which consists of a single CBM32 (MXAN_0542) and one that contains 2 tandem copies with 84% identity to each other (MXAN_4914). Fused to the C-terminus of these proteins is a large unknown region that displays similarity to a predicted lipoprotein also from *M. xanthus*, suggesting a possible role in extracellular localization (fig. 5B). When the CBM32s from *Yersinia* sp. and *M. xanthus* are compared, there are 24 fixed residues out of 130 alignment positions. As in previous examples of multimodularity, *Myxococcus* modules show a moderate degree of protein and nucleotide variation (table 2); however, it is considerably lower in comparison than the *C. perfringens* GH89 (table 1) and *S. degradans* tri- and tetramodules (table 2). Comparison between silent and

nonsilent variation within the *Y. enterocolitica* and *M. xanthus* proteins indicates the presence of purifying selection maintaining the β -sandwich scaffold (fig. 5A). By analyzing the nucleotide diversity across the modules, it seems that there is a trend toward higher nucleotide (total and silent) variation in regions corresponding to β -structures, which is actually coincident with many positions showing fixed amino acid residues in all sequences (fig. 5A and C). In this regard, the effect of negative selection would be evident in these regions, allowing for high numbers of synonymous substitutions without altering the amino acid residues composing the β -sandwich. In YeCBM32, the codon usage bias is an average of 47.4 ± 1.4 , which contrasts with the extreme bias in the case of the modules of *Myxococcus* (32.5 ± 0.9). Compared with previous estimations, and by taking into account the presence of different modules, it is evident a trend toward higher codon bias values in multimodular CBMs compared with single-module CBMs. The higher amount of codon bias within multimodular CAZymes could be representative of a higher level of polypeptide complexity. Thus, the different amino acids will be encoded by preferred codons in each module, increasing the efficiency and accuracy of translation (Akashi 1994). However, and given that the likelihood of a species-specific codon bias is high considering the ecology of these organisms, further studies would be needed in order to assess if such differences are due to the biased usage of preferred codons in multimodular CBMs (compared with single-module CBMs) or, if otherwise, these are the result of differences in the species-specific codon bias levels between these organisms.

Structural alignments between YeCBM32 and the 3 closely related CBM32s from *Myxococcus* reveal that there is a high level of conservation of the basic amino acids predicted to be involved in ligand binding (Abbott et al. 2007) (fig. 5A and D). It is believed that these residues create a basic pocket designed to interact with the internal sugars of the negatively charged polygalacturonic acid (fig. 5E). In this regard, the noticeable absence of genes within the *M. xanthus* genome encoding classical pectin utilization enzymes, such as polysaccharide lyases (families 1–4, 9–11) and the polygalacturonases (family 28 GHs), is surprising. Biochemical characterization of the *M. xanthus* CBM32s, therefore, would make an interesting study with implications for the role of the conserved basic binding pocket in carbohydrate recognition and possibly extracellular adherence.

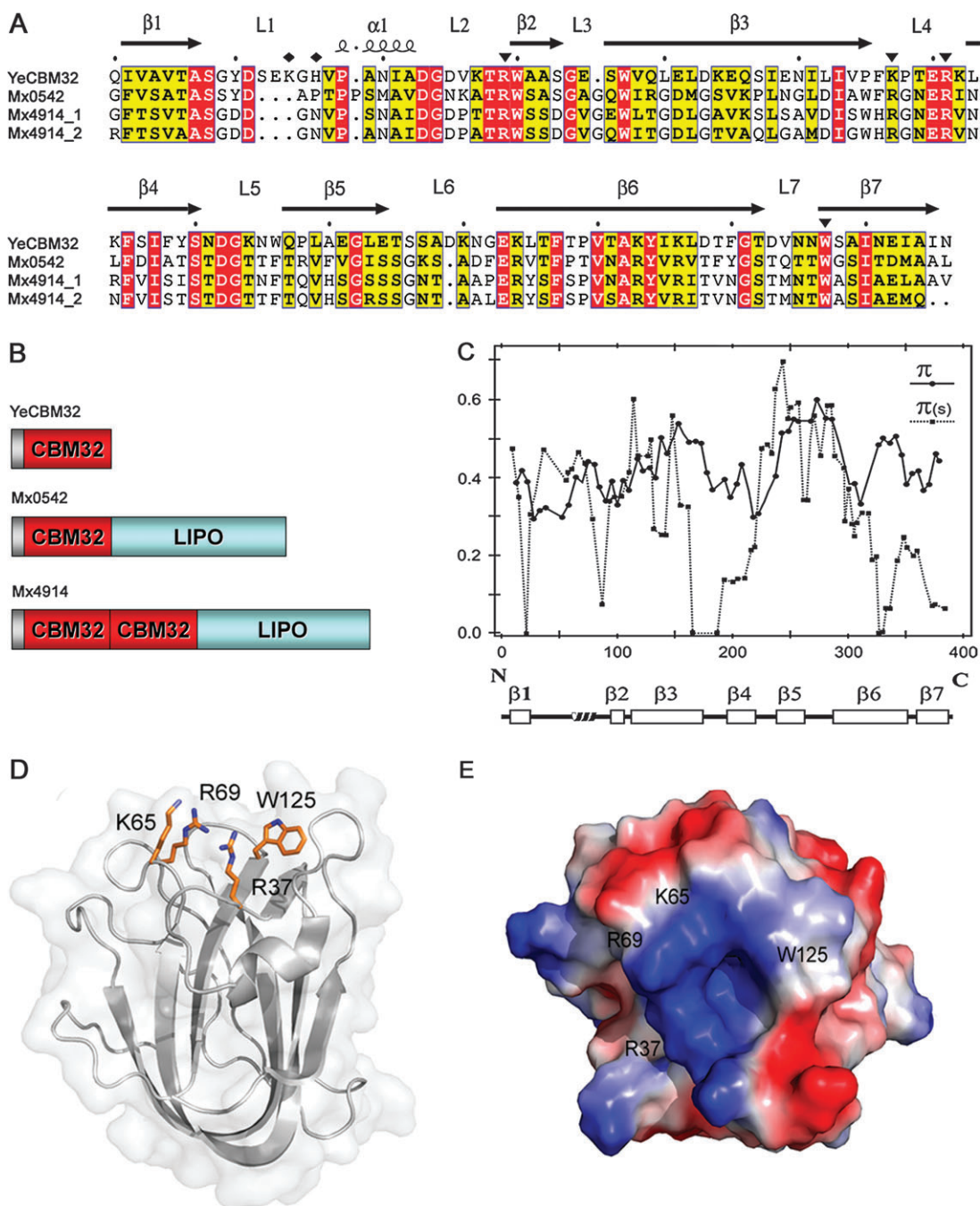


FIG. 5.—Structural analysis of independent CBM32s from *Yersinia enterocolitica* and *Myxococcus xanthus*. (A) Structural alignment between CBMs using ESPRIPT and the coordinates for YeCBM32 (PDB ID: 2JD9). Black triangles indicate amino acids with a major role in polygalacturonic acid binding, and black diamonds indicate amino acids with a minor role. (B) Modular structures of the independent CBMs from *Y. enterocolitica* and *M. xanthus*. (C) Total (π) and synonymous (π_s) nucleotide diversity (expressed as the average number of nucleotide difference per site) across the coding regions of the single module from YeCBM32 and the dimodule from *M. xanthus*. The diversity values were calculated using a sliding-window approach with a window length of 20 bp and a step size of 5 bp (for π), and a window length of 5 bp and a step size of 1 bp (for π_s). The corresponding secondary structure is represented below the graph. (D) Three-dimensional structure of YeCBM32 with the conserved amino acids predicted to be involved in binding shown as orange sticks. (E) Electrostatic potential of the solvent accessible surface of YeCBM32 highlighting the basicity of the conserved binding pocket.

Conclusion

Traditionally, CBMs are viewed as performing relatively simple roles in nature. More recently, however, CBMs have been implicated in many inflammatory forms of microbial pathogenesis and the metabolism of higher or-

ganisms including plants and animals. In addition to this unraveling functional diversity, new complex architectures have emerged, including enzymes with multiple copies of CBM32s and independent CBM32s such as the polygalacturonic acid-binding lectin-like proteins.

The presence of multiple CBMs within 1 enzyme is a phenomenon that invites more intense phylogenetic and biochemical investigation. Toward this end, family 32 provides a repository of information as currently 31 proteins from different organisms have been determined to contain more than 1 CBM32. At first glance, it is tempting to speculate that CBM multimodularity results from a process of intragenic duplication, involving functional domains; however, the process is clearly not that simple. The data presented here suggest that there are at least 2 distinct strategies intended to enhance overall enzyme function: 1) heterogeneous and 2) homogeneous modular arrangement.

Heterogeneous multimodularity occurs when one or more module within an enzyme is more closely related to a CBM found in a different enzyme than the other CBM32s within the same protein. This pattern underlines the likelihood that different CBMs within the same enzyme possess distinct ligand-binding specificities. For example, module 5 from CpGH89 displays a high level of divergence from the other CBMs within this enzyme and is closely related to modules from GH84C, GH33, GH31, and GH20 (fig. 2). The biological importance of heterogeneous CBMs within these enzymes has not been determined; however, there are 2 tantalizing possibilities. Firstly, harnessing CBMs that possess distinct ligand specificities may fine-tune enzyme targeting to a heterogeneous multivalent substrate. Secondly, it may facilitate enzyme residency as nascent ligands are exposed during catalytic turnover that are recognized by ancillary CBMs.

Homogeneous multimodularity refers to CBMs that display the closest similarity to other CBMs within the same protein. It is predicted that such amino acid similarity will translate into overlapping ligand-binding profiles, which increases the propensity of the complete polypeptide to interact more tightly with a target ligand through an "avidity effect." Simply stated, this means that the overall affinity of the CBMs operating in tandem is greater than the individual modules acting independent of one another. This specialized form of multivalency has been previously characterized for a variety of tandem CBMs found in other families, including the xylan-binding CBM2bs from a *Cellulomonas fimi* xylanase (Bolam et al. 2001), the xylo- and cellulose-binding CBM6 triplicate modules from *Clostridium stercoararium* (Boraston et al. 2002), and the starch-binding CBM26s from lactobacillus α -amylases (Guillén et al. 2007). Homogeneous multimodularity within family 32, however, may be a more complex and fine-tuned process. As described above, the presence of 2 subgroups (modules 1, 2 and modules 3, 4) within SdGH3 suggests that the protein may possess tailored CBM32 ligand-binding specificities that work in cooperation to recognize distinct substructures within the target substrate. In essence, this possibility may blur the apparent functional distinction between homogeneous and heterogeneous multimodularity. Nevertheless, both these modular signatures appear to be a powerful evolutionary method to tweak or enhance enzyme activity.

In addition to the contributions of multimodular carbohydrate active proteins, the functions of independent CBMs are now beginning to be understood. Beyond the

galacturonic acid-binding CBM32s, which likely operate to retain oligogalacturonides within the periplasm, independent CBMs have also been documented in families 29 and 33, where they perform novel roles in the degradation of insoluble polysaccharides. Family 29 contains only 1 member: the noncatalytic protein NCP1 from the anaerobic fungus *Piromyces equi*. NCP1 consists of 2 tandem CBMs (CBM29-1 and -2) and is a component of the multisubunit extracellular plant cell wall degradation complex found in the gastrointestinal tract of herbivores (Wong et al. 1995). This protein displays a promiscuous pattern of ligand binding, including xylan, mannan, and cellulose (Charnock et al. 2002; Flint et al. 2005). The tandem positioning of 2 related CBMs enables the protein to interact with carbohydrates in a unique "lamination" mechanism that sandwiches the ligand between the domains (Flint et al. 2004). The CBM33 from *Serratia marcescens* (Cbp21) is an independent acting protein involved in the degradation of chitin. Cbp21 is secreted and interacts with crystalline chitin perturbing its superstructure. This process facilitates destruction of the polysaccharide by chitinase A, B, and C (Vaaje-Kolstad et al. 2005). The abundance of independent proteins within family 32, including the dimodule from *M. xanthus* and the trimodule from *S. degradans*, provides an excellent opportunity to expand our understanding of the operations of independent CBMs and perhaps discover new functions for this interesting class of proteins.

CBM32s represent a complex family of proteins that span a wide range of ligand-binding profiles and functions. In many cases, the presence of multiple CBM32s within an enzyme or the lack of catalytic module complicates their characterization even further. The phylogenetic analysis presented here helps to alleviate these difficulties by providing a novel approach for their classification. By analyzing each CBM32 independently, it positions the module with CBM32s of related sequence that may more accurately represent its biological function. In addition to streamlining several new routes for investigation within family 32, we anticipate that this approach will be helpful for the future classification of members within other CBM families that have diversity in ligand selectivity.

Supplementary Material

Supplementary materials are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

A.B.B. is Canada Research Chair in molecular interactions. This work was funded by the National Sciences and Engineering Research Council of Canada and by a Postdoctoral Marie Curie International Fellowship within the 6th European Community Framework Program (to J.M.E.-L.). We are grateful to L. Ficko-Blean and A. Lammerts van Bueren for their helpful comments and critical review of the manuscript.

Literature Cited

- Abbott DW, Boraston AB. 2007. Specific recognition of saturated and 4,5-unsaturated hexuronate sugars by a periplasmic binding protein involved in pectin catabolism. *J Mol Biol.*
- Abbott DW, Hryniuk S, Boraston AB. 2007. Identification and characterization of a novel periplasmic polygalacturonic acid binding protein from *Yersinia enterocolitica*. *J Mol Biol.*
- Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics.* 136:927–935.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.
- Bolam DN, Xie H, White P, Simpson PJ, Hancock SM, Williamson MP, Gilbert HJ. 2001. Evidence for synergy between family 2b carbohydrate binding modules in *Cellulomonas fimi* xylanase 11A. *Biochemistry.* 40:2468–2477.
- Boraston A, Lammerts van Bueren A, Ficko-Blean E, Abbott DW. Forthcoming. Carbohydrate-protein interactions: carbohydrate-binding modules. *Carbohydr Glycosci.*
- Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. 2004. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J.* 382:769–781.
- Boraston AB, McLean BW, Chen G, Li A, Warren RA, Kilburn DG. 2002. Co-operative binding of triplicate carbohydrate-binding modules from a thermophilic xylanase. *Mol Microbiol.* 43:187–194.
- Chamock SJ, Bolam DN, Nurizzo D, Szabo L, McKie VA, Gilbert HJ, Davies GJ. 2002. Promiscuity in ligand-binding: the three-dimensional structure of a *Piromyces* carbohydrate-binding module, CBM29-2, in complex with cello- and mannohexaose. *Proc Natl Acad Sci USA.* 99:14077–14082.
- Coutinho PM, Henrissat B. 1999. Carbohydrate-active enzymes: an integrated database approach. In: Gilbert GDHJ, Henrissat B, Svensson B, editors. *Recent advances in carbohydrate bioengineering*. Cambridge: The Royal Society of Chemistry. p. 3–12.
- Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. 1998. JPred: a consensus secondary structure prediction server. *Bioinformatics.* 14:892–893.
- Efron B, Halloran E, Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci USA.* 93:13429–13434.
- Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution Int J Org Evolution.* 39:783–791.
- Ficko-Blean E, Boraston AB. 2006. The interaction of carbohydrate-binding module from a clostridium perfringens N-acetyl-beta-hexosaminidase with its carbohydrate receptor. *J Biol Chem.* 281:37748–37757.
- Flint J, Bolam DN, Nurizzo D, Taylor EJ, Williamson MP, Walters C, Davies GJ, Gilbert HJ. 2005. Probing the mechanism of ligand recognition in family 29 carbohydrate-binding modules. *J Biol Chem.* 280:23718–23726.
- Flint J, Nurizzo D, Harding SE, Longman E, Davies GJ, Gilbert HJ, Bolam DN. 2004. Ligand-mediated dimerization of a carbohydrate-binding molecule reveals a novel mechanism for protein-carbohydrate recognition. *J Mol Biol.* 337:417–426.
- Gilkes NR, Warren RA, Miller RC Jr, Kilburn DG. 1988. Precise excision of the cellulose binding domains from two *Cellulomonas fimi* cellulases by a homologous protease and the effect on catalysis. *J Biol Chem.* 263:10401–10407.
- Gouet P, Robert X, Courcelle E. 2003. ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* 31:3320–3323.
- Guillén D, Santiago M, Linares L, Pérez R, Morlon J, Ruiz B, Sánchez S, Rodríguez-Sanoja R. 2007. Alpha-amylase starch binding domains: cooperative effects of binding to starch granules of multiple tandemly arranged domains. *Appl Environ Microbiol.* 73:3833–3837.
- Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser.* 41:95–98.
- Hugouvieux-Cotte-Pattat N, Blot N, Reverchon S. 2001. Identification of TogMNAB, an ABC transporter which mediates the uptake of pectic oligomers in *Erwinia chrysanthemi* 3937. *Mol Microbiol.* 41:1113–1123.
- Hugouvieux-Cotte-Pattat N, Reverchon S. 2001. Two transporters, TogT and TogMNAB, are responsible for oligogalacturonide uptake in *Erwinia chrysanthemi* 3937. *Mol Microbiol.* 41:1125–1132.
- Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform.* 5:150–163.
- Machovic M, Janecek S. 2006a. Starch-binding domains in the post-genome era. *Cell Mol Life Sci.* 63:2710–2724.
- Machovic M, Janecek S. 2006b. The evolution of putative starch-binding domains. *FEBS Lett.* 580:6349–6356.
- Machovic M, Svensson B, MacGregor EA, Janecek S. 2005. A new clan of CBM families based on bioinformatics of starch-binding domains from families CBM20 and CBM21. *FEBS J.* 272:5497–5513.
- Marin-Rodriguez MC, Orchard J, Seymour GB. 2002. Pectate lyases, cell wall degradation and fruit softening. *J Exp Bot.* 53:2115–2119.
- Nei M, Kumar S. 2000. *Molecular evolution and phylogenetics*. New York: Oxford University Press.
- Newstead SL, Watson JN, Bennet AJ, Taylor G. 2005. Galactose recognition by the carbohydrate-binding module of a bacterial sialidase. *Acta Crystallogr Sect D Biol Crystallogr.* 61:1483–1491.
- Perez-Bercoff A, Koch J, Burglin TR. 2006. LogoBar: bar graph visualization of protein logos with gaps. *Bioinformatics.* 22:112–114.
- Rodionov DA, Gelfand MS, Hugouvieux-Cotte-Pattat N. 2004. Comparative genomics of the KdgR regulon in *Erwinia chrysanthemi* 3937 and other gamma-proteobacteria. *Microbiology.* 150:3571–3590.
- Rozas J, Sanchez-DelBarrio JC, Messegueur X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics.* 19:2496–2497.
- Rzhetsky A, Nei M. 1992a. A simple method for estimating and testing minimum-evolution trees. *Mol Biol Evol.* 9:945–967.
- Rzhetsky A, Nei M. 1992b. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *J Mol Evol.* 35:367–375.
- Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.
- Sitnikova T. 1996. Bootstrap method of interior-branch test for phylogenetic trees. *Mol Biol Evol.* 13:605–611.
- Sitnikova T, Rzhetsky A, Nei M. 1995. Interior-branch and bootstrap tests of phylogenetic trees. *Mol Biol Evol.* 12:319–333.
- Taylor LE 2nd, Henrissat B, Coutinho PM, Ekborg NA, Hutcheson SW, Weiner RM. 2006. Complete cellulase system in the marine bacterium *Saccharophagus degradans* strain 2-40T. *J Bacteriol.* 188:3849–3861.

- Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673–4680.
- Tomme P, Van Tilbeurgh H, Pettersson G, Van Damme J, Vandekerckhove J, Knowles J, Teeri T, Claeysens M. 1988. Studies of the cellulolytic system of *Trichoderma reesei* QM 9414. Analysis of domain function in two cellobiohydrolases by limited proteolysis. *Eur J Biochem.* 170:575–581.
- Vaaje-Kolstad G, Horn SJ, van Aalten DM, Synstad B, Eijsink VG. 2005. The non-catalytic chitin-binding protein CBP21 from *Serratia marcescens* is essential for chitin degradation. *J Biol Chem.* 280:28492–28497.
- Van Tilbeurgh H, Loontjens FG, Engelborgs Y, Claeysens M. 1989. Studies of the cellulolytic system of *Trichoderma reesei* QM 9414. Binding of small ligands to the 1,4-beta-glucan cellobiohydrolase II and influence of glucose on their affinity. *Eur J Biochem.* 184:553–559.
- Weber B, Blanch L, Clements PR, Scott HS, Hopwood JJ. 1996. Cloning and expression of the gene involved in Sanfilippo B syndrome (mucopolysaccharidosis III B). *Hum Mol Genet.* 5:771–777.
- Willats WG, McCartney L, Mackie W, Knox JP. 2001. Pectin: cell biology and prospects for functional analysis. *Plant Mol Biol.* 47:9–27.
- Wong MV, Ho YW, Tan SG, Abdullah N, Jalaludin S. 1995. Isozyme and morphological characteristics of the anaerobic fungus *Piromyces mae* isolated from the duodenum, rumen and faeces of sheep. *FEMS Microbiol Lett.* 134:9–14.
- Wright F. 1990. The 'effective number of codons' used in a gene. *Gene.* 87:23–29.
- Zdobnov EM, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 17:847–848.
- Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA.* 95:3708–3713.

Claudia Kappen, Associate Editor

Accepted October 26, 2007