Tracing Determinants of Dual Substrate Specificity in Glycoside Hydrolase Family 5*

Received for publication, March 14, 2012, and in revised form, May 18, 2012. Published, JBC Papers in Press, May 29, 2012, DOI 10.1074/jbc.M112.362640

Zhiwei Chen‡‡, Gregory D. Friedland‡‡, Jose H. Pereira‡‡, Sonia A. Revenco‡, Rosa Chan, Joshua I. Park‡‡, Michael P. Thelen‡‡, Paul D. Adams‡‡‡, Adam P. Arkin‡‡‡, Jay D. Keasling‡‡‡, Harvey W. Blanch‡‡‡, Blake A. Simmons‡‡, Kenneth L. Sale‡, Dylan Chivian‡‡, and Swapnil R. Chhabra‡‡

From the ‡‡Joint BioEnergy Institute, Emeryville, California 94608, the ‡§Sandia National Laboratories, Livermore, California 94555, the ‡¶Lawrence Berkeley National Laboratory, Berkeley, California 94720, the Departments of ‡§Nutritional Sciences and Toxicology, ‡¶Bioengineering, and ‡¶‡‡Chemical and Biomolecular Engineering, University of California, Berkeley, California 94720, and the ‡¶‡‡Lawrence Livermore National Laboratory, Livermore, California 94550

Enzymes are commonly viewed as highly specific for their natural substrates; however, this view obscures the fact that many have the ability to perform multiple activities (1, 2). This “promiscuity” typically involves the same chemistry applied to different substrates or, alternatively, can use different catalytic machinery within the same active site (3, 4). One well studied example is the serum paraoxonase PON1, which can hydrolyze lactones, thiolactones, carbonates, esters, and phosphothiesteres, using one set of active site residues for some functions and other residues for different functions (4, 5). Enzyme multispecificity is essential, in many cases, to organismal survival and has also been argued to be a byproduct of divergent evolution from unspecialized ancestor enzymes, potentially explaining why secondary functions of one enzyme are often primary functions in other members of the same family or superfamily (3, 4). The relative prevalence of enzyme promiscuity is an open question, but it has been suggested to be a fundamental characteristic of enzymes in general (2).

Multispecific enzymes can be found in numerous different protein families including the glycoside hydrolases (GHs), a large class of enzymes, which catalyze the hydrolysis of plant polysaccharides (6). GHs are categorized in the Carbohydrate-Active Enzymes (6) (CAZY) database into more than 100 sequence-based families including endo-, exo-, and side chain-acting hydrolases specific to glucose-, xylose-, mannose-, galactose-, and arabinose-containing polysaccharides, among others. An important application of GHs is in the hydrolysis of cellulose and hemicellulose into fermentable sugars for subsequent conversion to biofuels or commodity chemicals. Members of a given CAZY family share structural features and conserved catalytic residues but may or may not exhibit identical substrate specificity. For example, members of the

* This work was supported by the Office of Science, Office of Biological and Environmental Research, of the United States Department of Energy under Contract DE-AC02-05CH11231.
† This article contains supplemental Tables S1–S5 and Figs. S1–S3.
‡ Both authors contributed equally to this work.
§ Present address: Dept. of Pediatrics, Memorial Sloan-Kettering Cancer Center, New York, NY 10065.
¶ To whom correspondence may be addressed: Joint BioEnergy Institute, 5885 Hollis St., Emeryville, CA 94608. Tel.: 510-643-3722; Fax: 510-486-6219; E-mail: dcchivian@lbl.gov.
‡¶3 To whom correspondence may be addressed: Intrexon Corporation, 201 Industrial Rd., San Carlos, CA 94070. E-mail: schhabra@intrexon.com.
§§ The abbreviations used are: GH, glycoside hydrolase; GH5, glycoside hydrolase family 5; CAZy, Carbohydrate-Active Enzymes database; MSA, multiple sequence alignment; MBTH, 3-methyl-2-benzothiazolinonehydrazone; CMC, carboxymethyl cellulose.
A Substrate Specificity Governing Motif for GH5 A4 Subfamily

GH1 family are active on a number of sugar types linked through the β-1,4 bond, including β-galactose, β-mannose, and β-glucose. In contrast, all existing members of GH64 are β-1,3-endoglucanases.

In this study, we probed the extent and mechanisms of multisubstrate specificity in a highly diverse GH family using phylogenetic analysis and biochemical characterization. GH family 5 (GH5) comprises enzymes with a wide range of activities critical for the deconstruction of lignocellulose including endo-β-1,4-glucanase, endo-β-1,4-mannanase, endo-β-1,3-glucanase, endo-β-1,6-galactanase, lichenase, xylouglucan-specific endo-β-1,4-glucanases, and endo-β-1,4-xylanase (6). The number of substrates catalyzed by GH5 enzymes suggests that this family may contain single domain enzymes with multiple specificities. To test this hypothesis, we built a phylogenetic tree for this highly diverse family from a multiple sequence alignment (MSA) built using both sequence and structure information. This combined sequence and structure approach allowed the resulting alignment to contain genes with low pairwise sequence identity and a variety of functions, something that would not have been possible with standard sequence-only MSA-building methods. We analyzed sequence patterns from the resulting subfamilies to identify glucan and mannan specificity-determining residues and, through extensive biochemical characterization, validated a conserved motif that enables dual substrate specificity within a single catalytic domain. Subsequently, we applied this motif to enhance the catalytic efficiency of a GH5 enzyme for glucan and mannan hydrolysis by 175 and 1,600%, respectively. The conserved motif allowed us to discover new enzymes active on both glucan and mannan substrates, and its presence in over 70 members of GH5 strongly suggests the widespread prevalence of dual specificity in this family. Finally, extending the aforementioned phylogenetic analyses to other CAZy families, GH1 and GH43, further indicates that single domain multispecific GH enzymes may be more common than is currently characterized. As such, single domain multispecific GUs would be expected to reduce the complexity of designing enzyme mixtures, as well as microbial hosts for consolidated bioprocessing of lignocellulose.

EXPERIMENTAL PROCEDURES

Creation of Structure-based Sequence Alignments—To build a high quality sequence alignment in this diverse protein family, we used a combination of structural and sequence information. First, we performed pairwise structural alignments with 3Dhit (7) of 22 GH5 family structures (chain A of Protein Data Bank (PDB) IDs 2JEP, 3VDH, 1BQC, 2WHL, 7A3H, 2OSX, 1H1N, 1QNR, 1TVN, 1R9H, 1UUQ, 1EDG, 2C0H, 2CKS, 1WKY, 1H4P, 2PC8, 1CEO, 2ZUM, 1VJZ, 1EGZ, and 1ECE) to the Cel5A_Tma structure (PDB ID 3MMW, chain A). These 23 structures were selected based on their resolution and to remove redundancy at 90% sequence identity. For each of these structures, we used BLAST on GH5 sequences (after removing short sequence fragments) from the CAZy database to find sequences between 25 and 90% sequence identity with the sequence of the structure, and the resulting sequences were aligned with MUSCLE (8). These 23 MSAs were then combined into one MSA by aligning equivalent positions in the individual MSAs using the pairwise structural alignments to 3MMW. Redundant sequences were filtered out at 90% sequence identity, preferentially keeping sequences with structures, experimental characterization, and longer lengths, in this order of priority. We did not filter explicitly for active site residue identity; however, 94% of the sequences in the alignment contained both catalytic glutamates, and we do not expect removal of the small number of sequences not containing both glutamates to significantly alter the tree. Filtering for required inclusion of other active site residues is possible but was not performed here as some of these active site residues were of interest in finding the specificity-determining motif.

Creation of the Phylogenetic Tree—To test this hypothesis, we built a phylogenetic tree for this highly diverse family from a multiple sequence alignment (MSA) built using both sequence and structure information. This combined sequence and structure approach allowed the resulting alignment to contain genes with low pairwise sequence identity and a variety of functions, something that would not have been possible with standard sequence-only MSA-building methods. We analyzed sequence patterns from the resulting subfamilies to identify glucan and mannan specificity-determining residues and, through extensive biochemical characterization, validated a conserved motif that enables dual substrate specificity within a single catalytic domain. Subsequently, we applied this motif to enhance the catalytic efficiency of a GH5 enzyme for glucan and mannan hydrolysis by 175 and 1,600%, respectively. The conserved motif allowed us to discover new enzymes active on both glucan and mannan substrates, and its presence in over 70 members of GH5 strongly suggests the widespread prevalence of dual specificity in this family. Finally, extending the aforementioned phylogenetic analyses to other CAZy families, GH1 and GH43, further indicates that single domain multispecific GH enzymes may be more common than is currently characterized. As such, single domain multispecific GUs would be expected to reduce the complexity of designing enzyme mixtures, as well as microbial hosts for consolidated bioprocessing of lignocellulose.
Determination of Conserved and Nonconserved Active Site Positions—We used the following equation to calculate a BLOSUM-weighted profile difference score for alignment position $i$

$$\text{score}_i = \sum \sum \text{BLOSUM}_{aa1,aa2}(p_{aa1,i} - p_{aa2,i})^2$$  

(Eq. 1)

where $p_{aa,i}$ is the probability of amino acid aa occurring at position $i$ in the alignment and BLOSUM$_{aa1,aa2}$ is the BLOSUM substitution matrix value for amino acids aa1 and aa2. The resulting profile difference scores for the active site profiles of subfamily A4 versus each of the large primarily endoglucanase or mannanase subfamilies (A1, A2, A5/6, A7, and A8) are summarized in supplemental Table S3. Positions with averaged profile difference scores greater than 0.05 are classified as non-conserved (Asn-20, Glu-23, Pro-53, His-95, His-96, Phe-201, and Glu-287), and those below this cutoff are classified as conserved (Asn-20, Glu-23, Pro-53, His-95, His-96, Phe-201, and Glu-287). Higher cutoff values up to 20% of the maximum score would yield identical results.

Structural Modeling—The structural model of Cel5A_Tma in complex with the disaccharide mannan-based substrate has been published previously (10). To create the Cel5A_Tma complex with the disaccharide mannan-based substrate, the glucan-based substrate configuration was altered at OH-C2 by comparison with other mannan-based complex co-crystals. Hydrogens were added using UCSF Chimera (11), and His-95 and Asn-20 dihedrals were optimized for hydrogen bonding with the ligand (resulting in heavy atom root mean square deviation values of 0.43 and 0.32 Å, respectively); other rotatable hydrogen dihedrals were positioned by inspection to assess possible hydrogen bond geometries (supplemental Table S4). The subfamily A7 co-crystal structures described in Results containing an asparagine distant in primary sequence that occupies similar three-dimensional coordinates as Asn-20 are Man5_Tfu (Thermomonospora fusca, PDB ID 3MAN (12)) and Man5A_Bag (B. agaradhaerens, PDB ID 2WHL (13)). The subfamily A8 co-crystal structures containing the aspartate in similar three-dimensional coordinate space as Asn-20 are Man5A_Sly (Solanum lycopersicum, PDB ID 1RH9 (14)) and Man5A_Hje (Hypocrea jecorina, PDB ID 1QNR (15)). The model for Cel5B_Dtu was created with Phyre2 (16).

Chemicals and Reagents—All chemicals and enzymes were analytical grade from Sigma or EMD Chemicals. BugBuster protein extraction reagent, Popculture reagent, rLysozyme solution, Benzonase nuclease HC (purity >90%), and proteinase inhibitor mixture V (EDTA-free) were from Novagen and Calbiochem (EMD Biosciences). The Champion pET101 directional TOPO expression kit was from Invitrogen. Nickel-nitrilotriacetic acid spin columns were from Qiagen. Zeba spin desalting columns (2 ml, 70,000 molecular weight cut off) were from Pierce (Thermo Fisher Scientific). The bicinchoninic acid kit (BCA1-1KT) was from Sigma-Aldrich. Luria-Bertani (LB) medium was from EMD Chemicals, and 2xYT medium was from Sigma-Aldrich.

Gene Synthesis, Cloning, and Mutagenesis—Genes were codon-optimized according to the codon usage in Escherichia coli and synthesized by GenScript USA, Inc. All the genes were amplified and cloned by the pCDF-2 Ek/LIC vector kit (Novagen, EMD Biosciences) except that cel5a_pbr was cloned into pET101 vector (Invitrogen). Cloning primers are listed in supplemental Table S5a. Construct for Cel5A_Tma, pCDF2-cel5a_Tma, has been described before (10). All the constructs were confirmed by DNA sequencing (Quintara Biosciences). Site-directed mutagenesis was conducted by using the QuickChange Lightning site-directed mutagenesis kit according to the instructions of manufacturer (Agilent Technologies). All mutagenic primers are listed in supplemental Table S5b. The mutant plasmids were extracted by the QIAprep spin miniprep kit (Qiagen) and confirmed by DNA sequencing (Quintara Biosciences).

Protein Expression and Purification—All the constructs were transformed into BL21 (DE3) (Novagen, EMD Biosciences) for protein expression. Single colonies were inoculated into 5 ml of LB autoinduction medium (Overtight Express autoinduction system 1, Novagen, EMD Biosciences) containing appropriate antibiotics (100 µg/ml carbenicillin for pET101 constructs and 100 µg/ml streptomycin for the others) and incubated at 30 °C for 24 h. Induced cultures were harvested and preserved at −80 °C until use. Protein extraction, purification, buffer exchange, and concentration determination were as described before (data not shown).

Reducing Sugar Assays—The dinitrosalicylic acid method in a microplate format (17), without adding phenol and sulfite, was used for most of the enzyme assays, whereas 3-methyl-2-benzothiazolinonehydrazide (MBTH) was used for the kinetic assays. The MBTH method was used as described by Anthon and Barrett (18) with the following modifications: 40 µl of sample was mixed with 80 µl of Reagent A (0.25 M sodium hydroxide, 0.075% (w/v) MBTH and 0.025% (w/v) dithiothreitol) and then heated at 80 °C for 15 min. After cooling the samples down to room temperature, 80 µl of Reagent B (0.5% (w/v) FeNH4(SO4)2·12H2O, 0.5% (w/v) sulfamic acid and 0.25 M hydrochloric acid) was added. These mixtures were incubated at room temperature for 30 min. Samples were assayed for absorbance at 620 nm. The linear range of the MBTH method is 0.05–1 mM of reducing sugars (d-glucose for endoglucanases or d-mannose for mannanases).

Enzyme Assays—Enzyme assays for Cel5A_Tma and its mutants were performed at the respective optimal conditions for the two activities, 70 °C and pH 5.00 for endoglucanase activity and 90 °C and pH 5.50 for mannanase activity (data not shown), both in 50 mM sodium citrate buffer. For the other enzymes and their mutants, mesophilic enzymes were assayed at 37 °C, whereas thermophilic enzymes were assayed at 60 °C. 50 mM sodium citrate buffer (pH 5.50) was used for these enzyme reactions. The enzyme reactions contained 0.5% (w/v) carboxymethyl cellulose (CMC, average molecular mass ~90 kDa, Aldrich) and lotus bean gum (Sigma) as substrates for endoglucanase and mannanase activity assays, respectively. d-Glucose and d-mannose (0–5 mM) were used as standards for reducing sugars, as described above, when assaying endoglucanase and mannanase, respectively. The optimal temperatures of Cel5B_Dtu on CMC and lotus bean gum were analyzed from 50–100 °C with 5 °C intervals. 50 mM sodium citrate buffers (pH 3.00–6.50 with 0.50-unit intervals) were used to survey the optimal pHs for endoglucanase and mannanase activities of
A Substrate Specificity Governing Motif for GH5 A4 Subfamily

Cel5B_Dtu. One unit of endoglucanase or mannanase activity is defined as the amount of enzyme required for producing 1 μmol of reducing sugars per minute.

Kinetic Assays—All the specific activities and kinetic assays for Cel5B_Dtu were performed under the optimal conditions (pH 5.00 and 70 °C for endoglucanase activity; pH 5.50 and 75 °C for mannanase activity). CMC and carob galactomannan (low viscosity, Megazyme) instead of lotus bean gum were used as substrates in the kinetic assays. Initial velocities under a wide range of substrate concentrations ([S], 0.2–40 mg × ml⁻¹) were obtained for the calculation of $k_{cat}$ and $K_m$ by the Lineweaver-Burk Plot.

RESULTS

Building a High Quality Phylogenetic Tree for GH5—The CAZy database provides a wealth of data about GHs including lists of genes, activities, and structures within the sequence-based families. However, relationships between members of a given family as revealed by phylogenetic trees are not generally available. To begin our search for single domain multispecific members in GH5, we constructed a phylogenetic tree using available sequence and structure information from this family. Such a tree allows placing the genes into their evolutionary context and identification of subfamilies and sequence patterns between subfamilies with different functions. Phylogenetic tree building relies on the creation first of an MSA containing the sequences of interest. Although there are numerous available tools for building MSAs, their construction for sequence and functionally diverse families is not trivial. Standard MSA tools do not work well when there is sequence identity between members of less than 25%. For example, MSAs have been built with sequence-only approaches (19–23) that covered part of GH5, limiting the overall size and sequence diversity of its constituent genes. Incorporating the complementary information from experimentally determined protein structures can significantly help in the building of alignments and trees for sequence-diverse families (24–27). Given that this combined structure and sequence-based tree building approach has not previously been used on GH families, we chose to draw on the large number of structures in various GH families to build high quality sequence alignments and phylogenetic trees.

Our approach uses the relatively large number of crystal structures in GH5 (more than 30) to combine the low sequence identity parts of the family into a larger MSA. To do this, we created MSAs containing sequences with greater than 25% sequence identity to enzymes with experimentally determined crystal structures and then combined these MSAs using structure alignment methods. We used the resulting GH5 alignment containing 681 sequences to build a phylogenetic tree using FastTree 2 (9) and annotated it with the experimentally characterized activities obtained from CAZy (Fig. 1). In contrast to this phylogenetic analysis, previous studies of GH5 subfamily classifications focused on one subfamily at a time and were limited to sequence identity-based metrics (e.g. Refs. 28 and 29). In this work, we used the tree to classify subfamilies using their distance from the root, the length of the ancestral branch split, and the bootstrap support (Fig. 1).

Comparison of the functional assignments between the subfamilies in this tree shows phylogenetic correspondence with the divisions of different sugar specificities (Fig. 1). Three large subfamilies appear to contain predominantly β-1,4-glucan-specific enzymes (A1, A2, and A5/6); two are predominantly β-1,4-mannan-specific (A7 and A8); and one is predominantly β-1,3-glucan-specific (A9). In terms of substrate specificity, subfamily A4 appears to be the most diverse in GH5 (supplemental Fig. S1a) in that it contains a variety of β-1,4-linked glucan-, mannan-, and xylan-specific enzymes (supplemental Table S1). Notably, several members of subfamily A4 have previously been reported to act on more than one substrate. For instance, GH5 proteins from Prevotella ruminicola (AAC36862.1) and Clostridium cellulovorans (AAA23231.1) have been reported to act on glucon as well as xylan substrates, although detailed biochemical characterization or structural information for these enzymes is not available (30, 31). The most thoroughly characterized GH5 enzyme from subfamily A4 is the thermostable enzyme, Cel5A_Tma (AAD36816.1), from Thermotoga maritima (32). Cel5A_Tma can degrade both galactomannan (71 units/mg) and CMC (616 units/mg) at rates comparable with those of its single substrate-specific counterparts Man5_Tma from GH5 (83 units/mg on galactomannan) and Cel74_Tma from family GH74 (121 units/mg on CMC). Functional genomics studies on T. maritima have revealed recruitment of this enzyme on mannan- and glucon-based growth substrates (33).

Discovery of a Specificity-determining Sequence Motif in Cel5A_Tma—To dissect the determinants of substrate specificity in subfamily A4, we used the comprehensive phylogenetic tree of family GH5 to examine the amino acid profiles of active site residues (see “Experimental Procedures”) in the A4 subfamily, the mainly mannanase subfamilies (A7 and A8), and the predominantly endoglucanase subfamilies (A1, A2, and A5/6) (Fig. 2a). We categorized these positions as either conserved or variable based upon the extent of amino acid diversity (see “Experimental Procedures”) between the subfamily alignments (Fig. 2b, green circles). For example, the catalytic glutamates at positions 136 and 253 (using sequence numbering from Cel5A_Tma) are conserved among all members of GH5; in contrast, position 96 is variable, having mainly histidines in subfamily A4, but relatively few histidines in the other GH5 subfamilies (Fig. 2a).

These analyses resulted in the identification of seven positions (20, 23, 53, 95, 96, 201, and 287; Cel5A_Tma numbering) that varied between the subfamilies, suggesting their involvement in substrate specificity. To evaluate the role of these seven residues in substrate specificity, we generated alanine substitutions at these positions in Cel5A_Tma and assayed the purified enzymes for endoglucanase and mannanase activities (Fig. 2b). Of the seven variable positions, alanine mutations at five positions (N20A, E23A, P53A, H96A, and E287A) resulted in reduced activity on mannan, one mutation (H95A) had reduced activity on glucon, and one mutation (F201A) did not have a large impact for either substrate. For each of the seven positions conserved between the subfamilies (30, 135, 136, 196, 198, 253, and 286), mutation to alanine eliminated activity on both substrates, con-
sistent with the expected role of these positions either in catalysis or in nonspecific sugar substrate binding.

Application of the Motif to Predict Multispecificity—Next, we examined whether the pattern of amino acids at the six specificity-altering residues found in Cel5A_Tma could be generalized by assessing the presence of the pattern across various enzymes in the GH5 A4 subfamily. We reasoned that dual mannanase and glucanase activity might broadly occur within the A4 subfamily despite the lack of experimental characterization of mannanase activity in the A4 subfamily other than Cel5A_Tma (6), perhaps because previous studies did not test for mannanase activity in addition to the more typical glucanase assay. To this end, we searched for the six-residue pattern (allowing either aspartate or glutamate at positions 23 and 287) in the 143 genes in subfamily A4. We identified more than 70 sequences containing the motif (supplemental Fig. S1). Based on the presence of the motif, we predicted that these enzymes may have both endoglucanase and mannanase activities (Fig. 1, pink branches).

To test our prediction of broad multispecificity in subfamily A4, we assayed 10 additional enzymes, selected to broadly cover the phylogenetic diversity in A4 and to either match or differ from the six-residue pattern (Fig. 3a and supplemental Table S2). Of these enzymes, all exhibited endoglucanase activity, and six also had detectable mannanase activity. Of the six characterized dual specificity enzymes, four had the same pattern at the six residues as Cel5A_Tma, whereas two (Cel5A_Umi and Cel5B_Dtu) did not match the pattern, differing at only a single position. Of the four characterized single specificity enzymes, each differed at one position or more from the motif. We further confirmed the specificity determination of the six-residue pattern in other enzymes from the A4 subfamily by characterizing the endoglucanase and mannanase activities of alanine mutants in two dual specificity enzymes from subfamily A4 with low sequence identity to Cel5A_Tma: Cel5C_Cth (29% sequence identity) and Cel5A_Eec (25% sequence identity). The specificity changes resulting from mutations in both Cel5C_Cth and Cel5A_Eec were consistent with the specificity changes resulting from the corresponding mutations in Cel5A_Tma (Fig. 3b), with the exception of the P72A variant of Cel5C_Cth.

Using the Motif to Engineer Enhanced Activity—In addition to using the six-residue pattern to predict dual specificity, we applied the pattern to engineer enhanced activity. We postulated that the activity of Cel5B_Dtu could be improved by mutating the aspartate at position 14 in Cel5B_Dtu (corresponding to Asn-20 in Cel5A_Tma) to asparagine to fully
match the six-amino acid pattern. Homology modeling of Cel5B_Dtu (data not shown) suggested that D14N could allow three hydrogen bonds to the mannan substrate, whereas an aspartate might be limited to two hydrogen bonds. Mutation of D14N in Cel5B_Dtu resulted in enhanced hydrolysis for both substrates; we found a 70% increase in specific endoglucanase activity and a 300% increase in specific mannanase activity (Table 1). Kinetic analysis revealed that this single amino acid substitution decreased the $K_m$ for galactomannan by 1,500%, accompanied by a 5.2% increase in $k_{cat}$; the $K_m$ for the glucan substrate was reduced by 50%, whereas the $k_{cat}$ was increased by 35%. Notably, improvement in catalytic efficiencies ($k_{cat}/K_m$) attributed to this single mutation for endoglucanase and mannanase activities were 175 and 1,600%, respectively.

**DISCUSSION**

The comprehensive GH5 phylogenetic tree described here led to the identification of an active site motif describing dual specificity for glucan- and mannan-based substrates in the large and diverse A4 subfamily of GH5. However, a sequence motif alone cannot fully determine the substrate specificity of a sequence-distant group of enzymes given the importance of subtle sub-Angstrom level interactions in the active site. It is interesting then that this motif managed to capture the endoglucanase and mannanase specificity pattern for almost all mutations at positions in the Cel5A_Tma active site.

**FIGURE 2.** Determination of positions affecting specificity in glycoside hydrolase family 5. (a) Sequence profiles of the active site positions in the GH subfamily containing Cel5A_Tma (A4), of two predominantly mannanase subfamilies (A7 and A8), and of three predominantly endoglucanase subfamilies (A1, A2, and A5/6) (created with WebLogo (43)). (b) Experimental measurements of the relative specific endoglucanase and mannanase activities of alanine mutants at positions in the Cel5A_Tma active site. Residues that are variable on average between A4 and the other subfamilies are labeled with a green circle (see “Experimental Procedures” for details). Data in panel b are means from three independent experiments; error bars show S.D.

**FIGURE 3.** Characterization of additional GH5 A4 subfamily genes for dual specificity on glucan and mannan. (a) Experimental characterization of the endoglucanase (orange) and mannanase (blue) activities of Cel5A_Tma and 10 other genes from GH5 subfamily A4. These genes were selected to broadly cover the A4 subfamily tree and to contain diversity at the specificity-determining positions. Sequence identity to Cel5A_Tma of each gene is depicted with a black line on the plot, and the amino acid identities of the six specificity-determining positions are shown at right. The pattern of specificity changes in Cel5C_Cth and Cel5A_Eec from subfamily A4 in comparison with the corresponding mutations in Cel5A_Tma of N20A, E23A, P53A, H96A, E287A, and H95A, respectively (Fig. 2b). Cel5C_Cth and Cel5A_Eec are 29 and 25% identical to Cel5A_Tma, respectively, and closely match specificity patterns observed for Cel5A_Tma except the P72A mutation in Cel5C_Cth. Data in a and b are means from three independent experiments; error bars show S.D.
A Substrate Specificity Governing Motif for GH5 A4 Subfamily

TABLE 1
Specific activity and kinetics of Cel5B_Dtu and mutant D14N
CMC, carboxymethyl cellulose; S.A., specific activity; CGM, carob galactomannan.

| Substrate | Activity parameters | WT | D14N | Improvement % |
|-----------|---------------------|----|------|---------------|
| CMC       | S.A. (units × mg⁻¹ protein) | 28.89 ± 0.96 | 50.03 ± 0.97 | 73.17 |
|           | $k_{cat}$ (s⁻¹)     | 408.19 | 550.66 | 34.90 |
|           | $K_m$ (mg × ml⁻¹)   | 24.02 | 11.76 | 104.25 |
|           | $k_{cat}/K_m$ (ml × mg⁻¹ × s⁻¹) | 17.00 | 46.81 | 175.35 |
| CGM       | S.A. (units × mg⁻¹ protein) | 2.11 ± 0.03 | 8.83 ± 0.36 | 318.48 |
|           | $k_{cat}$ (s⁻¹)     | 68.25 | 71.82 | 5.24 |
|           | $K_m$ (mg × ml⁻¹)   | 11.57 | 0.72 | 1506.94 |
|           | $k_{cat}/K_m$ (ml × mg⁻¹ × s⁻¹) | 5.90 | 99.89 | 1593.05 |

Although there are no reports in the literature for enzymatic activity enhancement for mannan hydrolysis, the largest improvements for glucan and xylan hydrolysis to date are ~80% (35) and ~300% (36), respectively. That the 300% increase in specific mannanase activity and 1,600% improvement in mannanase $k_{cat}/K_m$ observed for the “back-to-motif” mutation in Cel5B_Dtu come from a point mutant is intriguing given the difficulty of enhancing activity in these enzymes through other optimization techniques (37–39). The back-to-motif mutation indicates that this substrate-determining motif could be ancestral to the GH5 family, supporting Jensen’s hypothesis (40) that the spectrum of specificity in the ancestors of an enzyme family can be seen in the descendant families. Similar to back-to-ancestor mutations at nonactive sites, back-to-motif mutations within the active sites may broaden enzyme activity and also make enzymes more evolvable (41).

In addition to endoglucanase and mannanase activity in A4 subfamily enzymes, preliminary work has shown the presence of a third specificity, xylanase, in some enzymes. ⁶ This co-oc-

⁶ Z. Chen, unpublished data.
occurrence of single domain multispecificity and multiple specificities in the subfamily raises the interesting question of whether multispecificity is an “inherent” property of some groups of related enzymes, such as the GH5 A4 subfamily. To investigate the extent of this co-occurrence in CAZy, we extended the aforementioned MSA and phylogenetic analysis to two other well characterized GH families: GH1 and GH43. GH1 contains ~3,500 members (with 232 biochemical characterizations), and GH43 contains ~2,000 members (with 85 biochemical characterizations) (6). Similar to our findings in GH5, we observed the presence of single domain multispecific enzymes within subfamilies bearing different sugar specificities (supplemental Fig. S3, a and b), which suggests that these subfamilies could contain numerous multispecific members. Further analysis in other GH and non-GH families is needed to confirm this observation more generally, but these results support the idea that multispecificity could be an inherent property of some groups of enzymes.

In conclusion, our comprehensive phylogenetic and biochemical analyses of GH5 and subsequent phylogenetic analysis of GH1 and GH43 suggest that multispecific GH enzymes may be more prevalent than have been experimentally characterized. It will be interesting to investigate whether these multiple specificities are utilized in certain conditions by the host organism or whether they are perhaps a latent property of enzymes evolved from a promiscuous ancestor.

Acknowledgments—We thank Morgan Price for writing the scripts used in the sequence alignment trimming and tree rerooting, as well as writing the phylogenetic tree building program FastTree. The codon-optimized Cel5A_Pbr gene was a gift from Dr. Christopher A. Voigt at the University of California, San Francisco. We thank Tanveer Batth and Dr. Christopher J. Petzold in the Technology Division of the Joint BioEnergy Institute for help with mass spectrometry analysis. Tanja Kortemme provided suggestions after critical reading of the manuscript.

REFERENCES
1. Nobeli, I., Favia, A. D., and Thornton, J. M. (2009) Protein promiscuity and its implications for biotechnology. Nat. Biotechnol. 27, 157–167
2. Khersonsky, O., and Tawfik, D. S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu. Rev. Biochem. 79, 471–505
3. Peisajovich, S. G., and Tawfik, D. S. (2007) Protein engineers turned evolutionary selection native and complex structures of thermostable β-mannanase from Thermomonospora fusca: substrate specificity in glycosyl hydrolase family 5. Biochemistry 46, 7009–7018
4. Bourgault, R., Oakley, A. J., Bewley, J. D., and Wilke, M. C. (2005) Three-dimensional structure of (1,4)-β-D-mannan mannanohydrolase from tamaro fruit. Protein Sci. 14, 1233–1241
5. Sabini, E., Schubert, H., Mursudov, G., Wilson, K. S., Siika-Aho, M., and Penttilä, M. (2000) The three-dimensional structure of a Thermococcus reesei β-mannanase from glycoside hydrolase family 5. Acta Crystallogr. D. Biol. Crystallogr. 56, 3–13
6. Kelley, L. A., and Sternberg, M. J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. 4, 363–371
7. Xiao, Z., Storms, R., and Tsang, A. (2005) Microplate-based carboxymethylcellulose assay for endoglucanase activity. Anal. Biochem. 342, 176–178
8. Anthon, G. E., and Barrett, D. M. (2002) Determination of reducing sugars with 3-methyl-2-benzothiazolinonehydrazone. Anal. Biochem. 305, 287–289
9. Wang, Y., Wang, X., Tang, R., Yu, S., Zheng, B., and Feng, Y. (2010) A novel thermostable cellulase from Fervidobacterium nodosum. J. Mol. Catal. B. Enzym. 66, 294–301
10. Pereira, J. H., Chen, Z., McAndrew, R. P., Sapra, R., Chhabra, S. R., Sale, K. L., Simmons, B. A., and Adams, P. D. (2010) Biochemical characterization and crystal structure of endoglucanase Cel5A from the hyperthermophilic Thermotoga maritima. J. Struct. Biol. 172, 372–379
11. Petersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera: a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612
12. Hilge, M., Gloor, S. M., Rypniewski, W., Sauer, O., Heightman, T. D., Zimmermann, W., Winterhalter, K., and Piontek, K. (1998) High-resolution native and complex structures of thermostable β-mannanases recognize heterogeneous substrates. Biochemistry 48, 7009–7018
13. Tailford, L. E., Ducros, V. M., Flint, J. E., Roberts, S. M., Morland, C., Zechel, D. L., Smith, N., Bjørnvad, M. E., Borchert, T. V., Wilson, K. S., Davies, G. J., and Gilbert, H. J. (2009) Understanding how diverse β-mannanases recognize heterogeneous substrates. Biochemistry 48, 7009–7018
14. Bourgault, R., Oakley, A. J., Bewley, J. D., and Wilke, M. C. (2005) Three-dimensional structure of (1,4)-β-D-mannan mannanohydrolase from tamaro fruit. Protein Sci. 14, 1233–1241
A Substrate Specificity Governing Motif for GH5 A4 Subfamily

31. Whitehead, T. R. (1993) Analyses of the gene and amino acid sequence of the Prevotella (Bacteroides) ruminicola 23 xylanase reveals unexpected homology with endoglucanases from other genera of bacteria. *Curr. Microbiol.* 27, 27–33

32. Chhabra, S. R., Shockley, K. R., Ward, D. E., and Kelly, R. M. (2002) Regulation of endo-acting glycosyl hydrolases in the hyperthermophilic bacterium *Thermotoga maritima* grown on glucan- and mannan-based polysaccharides. *Appl. Environ. Microbiol.* 68, 545–554

33. Chhabra, S. R., Shockley, K. R., Conners, S. B., Scott, K. L., Wollfinger, R. D., and Kelly, R. M. (2003) Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima*. *J. Biol. Chem.* 278, 7540–7552

34. Wu, T. H., Huang, C. H., Ko, T. P., Lai, H. L., Ma, Y., Chen, C. C., Cheng, Y. S., Liu, J. R., and Guo, R. T. (2011) Diverse substrate recognition mechanism revealed by *Thermotoga maritima* Cel5A structures in complex with celletetraose, cellobiose, and mannotriose. *Biochim. Biophys. Acta* 1814, 1832–1840

35. Liang, C., Fioroni, M., Rodriguez-Ropero, F., Xue, Y., Schwanenberg, U., and Ma, Y. (2011) Directed evolution of a thermophilic endoglucanase (Cel5A) into highly active Cel5A variants with an expanded temperature profile. *J. Biotechnol.* 154, 46–53

36. Rettinger, S., Yu, Y., Wicki, I., Ludwiczek, M., D’Angelo, I., Baturin, S., Okon, M., Strynadka, N. C., Lutz, S., Withers, S. G., and McIntosh, L. P. (2010) Circular permutation of *Bacillus circulans* xylanase: a kinetic and structural study. *Biochemistry* 49, 2464–2474

37. Himmel, M. E., Ding, S. Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W., and Foust, T. D. (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* 315, 804–807

38. Wen, F., Nair, N. U., and Zhao, H. (2009) Protein engineering in designing tailored enzymes and microorganisms for biofuels production. *Curr. Opin. Biotechnol.* 20, 412–419

39. Percival Zhang, Y. H., Himmel, M. E., and Mielenz, J. R. (2006) Outlook for cellulase improvement: screening and selection strategies. *Biotechnol. Adv.* 24, 452–481

40. Jensen, R. A. (1976) Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* 30, 409–425

41. Bershtein, S., Goldin, K., and Tawfik, D. S. (2008) Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol.* 379, 1029–1044

42. Letunic, I., and Bork, P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* 39, W475–W478

43. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190

44. Davies, G. J., Wilson, K. S., and Henriissat, B. (1997) Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochem. J.* 321, 557–559

45. Davies, G. J., Mackenzie, L., Varrot, A., Dauter, M., Brzozowski, A. M., Schülein, M., and Withers, S. G. (1998) Snapshots along an enzymatic reaction coordinate: analysis of a retaining β-glycoside hydrolase. *Biochemistry* 37, 11707–11713