Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space

William Helbert[a,1], Laurent Poulet[a], Sophie Drouillard[a], Sophie Mathieu[a], Mélanie Loiodice[a], Marie Couturier[a], Vincent Lombard[b,c], Nicolas Terrapon[b,c], Jeremy Turchetto[b,c], Renaud Vincentelli[b,c], and Bernard Henrissat[b,c,d,1]

[a]Centre de Recherches sur les Macromolécules Végétales, CNRS, Grenoble Alpes Université, BP53, 38000 Grenoble Cedex 9, France; [b]CNRS, UMR 7257, Université Aix-Marseille, 13288 Marseille, France; [c]Institut National de la Recherche Agronomique, USC 1408 Architecture et Fonction des Macromolécules Biologiques, 13288 Marseille, France; and [d]Department of Biological Sciences, King Abdulaziz University, 21589 Jeddah, Saudi Arabia

Edited by Stephen G. Withers, University of British Columbia, Vancouver, Canada, and accepted by Editorial Board Member David Baker February 6, 2019 (received for review September 13, 2018)

Over the last two decades, the number of gene/protein sequences gleaned from sequencing projects of individual genomes and environmental DNA has grown exponentially. Only a tiny fraction of these predicted proteins has been experimentally characterized, and the function of most proteins remains hypothetical or only predicted based on sequence similarity. Despite the development of postgenomic methods, such as transcriptomics, proteomics, and metabolomics, the assignment of function to protein sequences remains one of the main challenges in modern biology. As in all classes of proteins, the growing number of predicted carbohydrate-active enzymes (CAZymes) has not been accompanied by a systematic and accurate attribution of function. Taking advantage of the CAZy database, which groups CAZymes into families and subfamilies based on amino acid similarities, we recombinantly produced 564 proteins selected from subfamilies without any biochemically characterized representatives, from distant relatives of characterized enzymes and from nonclassified proteins that show little similarity with known CAZymes. Screening these proteins for activity on a wide collection of carbohydrate substrates led to the discovery of 13 CAZyme families (two of which were also discovered by others during the course of our work), revealed three previously unknown substrate specificities, and assigned a function to 25 subfamilies.

CAZymes | screening | polysaccharides

The last 20 years have witnessed the sequencing of the genomes of isolated unicellular and pluricellular organisms as well as microbial communities from various environments, such as ocean (1, 2), soil (3), and the digestive tract of animals (4) and humans (5, 6). The current challenge is not to obtain even more sequence data, but rather to infer the function of the myriads of already identified proteins (7). Postgenomic approaches, such as transcriptomics, proteomics, and metabolomics, can reveal useful relationships between genes or proteins but do not directly assign function or substrate specificity to hypothetical proteins or enzymes. Therefore, despite the development of faster, cheaper, and miniaturized experimental methods, ascribing a function to a gene product remains the main challenge of biology in the postgenomic era (8).

Reliable functional predictions are based on experimentally determined knowledge and on a suitable estimate of the divergence beyond which precise function cannot be readily extrapolated (9). Inspection of sequence databases show that they are heavily polluted by erroneous functional predictions owing to the lack of universal similarity thresholds that can ensure robust propagation of protein function (8, 10, 11). This problem has become particularly acute with the emergence of bioinformatic methods that can detect extremely remote sequence similarities (12–14).

The enzymes that assemble and deconstruct glycans have been classified into sequence-based families starting in 1991 (15–20). The functional diversity (specificity) of these enzymes is enormous and reflects the wide diversity of glycan structures found in nature. The database of carbohydrate-active enzymes (CAZymes), CAZy (www.cazy.org), compiles the various families of glycoside hydrolases (GHs), polysaccharide lyases (PLs), glycosyltransferases, and several other categories of enzymes that act on carbohydrates (21). In the classification system that underlies the CAZy database, families are defined by sequences that cluster around at least one biochemically characterized member (21). Interestingly, the sequence-based CAZyme families often group together enzymes of differing substrate specificity (15), showing that the acquisition of novel substrate specificity is commonplace among CAZymes. However, as observed in general protein databases, a similarity search conducted against the entries in the CAZy database essentially yields uncharacterized or unreliably named gene products and thus fails to produce reliable functional inference. In addition, as in all protein databases, the number of entries in CAZy is increasing exponentially, but the number of biochemically characterized enzymes is growing much more slowly (21).

Significance

In the context of genome and metagenome sequencing, the assignment of function to sequences is a serious issue. In the case of enzymes of strong substrate specificity, such as those involved in the breakdown of polysaccharides (e.g., glycoside hydrolases and polysaccharide lyases), assignments become unreliable when sequence similarity to experimentally studied enzymes is low. To better explore the sequence-to-function relationships of these enzymes, we successfully applied a strategy based on a rational bioinformatic selection of enzyme targets, synthetic gene synthesis, and screening of recombinant proteins on a wide diversity of carbohydrate substrates. Seventy-nine of our 564 targets exhibited enzymatic activity, including three activities that have not been described previously, and 13 novel enzyme families could be defined.

Author contributions: W.H. and B.H. designed research; L.P., S.M., M.L., M.C., and J.T. performed research; M.L. and R.V. contributed new reagents/analytic tools; W.H., S.D., V.L., and N.T. analyzed data; and W.H., N.T., and B.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. S.G.W. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: william.helbert@cermav.cnrs.fr or Bernard.Henrissat@afmb.univ-mrs.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1815791116/-/DCSupplemental.

Published online March 8, 2019.

For sequence-based functional predictions, the occurrence of enzymes of differing specificity in a given CAZy family results in a broad functional categorization, such as "putative glycoside hydrolase," but does not provide a reliable prediction of the actual substrate of the enzyme. Furthermore, there are even examples of proteins that have evolved from CAZymes to acquire novel functions unrelated to their CAZyme ancestor (22). Multiple studies have shown that the breakdown of large multifunctional GH families into subfamilies yields a much narrower set of substrate specificities in each subfamily and offers a clear improvement in functional prediction for those subfamilies that have at least one characterized member (23–26). Conversely, subfamilies with no characterized members can guide enzymology investigations toward unexplored areas of the families.

The steady increase in the number of CAZyme families over the last 20 years and the accumulation of unassigned sequences with similarity too low for reliable assignment to a family suggests that many other CAZyme families remain to be discovered. The most direct route to ascribing a function to putative CAZymes involves demonstrating the actual cleavage of an oligosaccharide or polysaccharide substrate by the protein of interest. In this context, we assayed the degradation of a collection of substrates with a set of enzymes already classified into CAZyme families but assigned to subfamilies with no biochemically characterized member and with a set of highly remote GH and PL homologs too distant to allow their classification into any current CAZy subfamily or family. The strategy was based on a rational bioinformatic selection of targets, automatic gene synthesis, and screening of recombinant proteins on a wide diversity of carbohydrate substrates. This approach increased the number of biochemically characterized subfamilies and led to the discovery of several new enzyme families and of previously unreported substrate specificities.

Results

We selected 564 nucleotide sequences encoding potential glycan-cleaving enzymes from several families of the GH and PL classes of CAZymes. These sequences composed three broad sets of gene products. The first set (142 GHs and 13 PLs, approximately 28% of the investigated sequences) comprised sequences assigned to subfamilies of large GH and PL families with no characterized members. The second set (203 GHxx_dist and 19 PLxx_dist, approximately 39% of the investigated sequences) comprised sequences that fell outside of established subfamilies or were only distantly related to a particular family. The last set (187 candidates, approximately 33% of the investigated sequences) comprised protein sequences that could not be assigned to a family owing to insufficient similarity (<20% identity) with known GHs or PLs. These sequences were typically extracted from the nonclassified category of putative GHs and PLs (www.cazy.org/GH0.html and www.cazy.org/PL0.html). The sequences were not edited to preserve their native specificity; all possible noncatalytic modules were left intact. The two first sets (subfamilies and distantly related proteins) included several eukaryotic sequences, whereas all sequences of candidate GH or PL proteins (the third set) were of prokaryotic origin. The complete list of sequences selected for this work, along with their source organism and family (and subfamily where possible), are given in *SI Appendix*, Table S1). All genes were codon-optimized for *Escherichia coli* expression, synthesized, and cloned in expression vectors encoding an N-terminal His tag for protein purification.

Expression assays were conducted at microplate scale using an autoinducible medium. Automated purification using nickel-affinity chromatography revealed that approximately 60% of the recombinant proteins were obtained in a soluble state (Fig. 1*A*). We observed a significantly reduced number of soluble proteins of eukaryotic origin (9 of 33 soluble proteins) compared with bacterial and archaeal targets (323 of 506 proteins; hypergeometric test $P < 4.10^{-5}$). No other significant correlation was found between

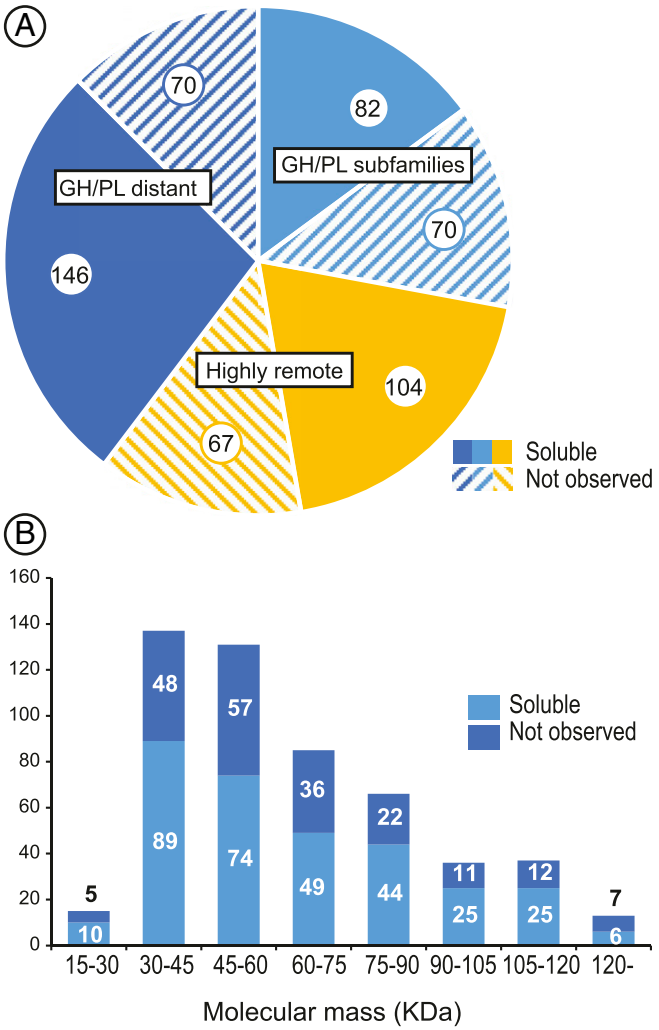

Fig. 1. Overexpression results. (*A*) Results presented according to three broad classes: (*i*) proteins from uncharacterized subfamilies within known families (GH/PL subfamilies), (*ii*) proteins classified into a CAZy family but only distantly related to characterized members (GH/PL distant), and (*iii*) remote homologs whose similarity is too low for inclusion in existing CAZy families (highly remote). "Soluble" refers to overexpressed proteins purified by nickel-affinity chromatography and detected using gel electrophoresis; "not observed," to proteins that did not bind to any affinity column (e.g., inclusion bodies, misfolded proteins). (*B*) Absolute frequency of overexpressed enzymes according to their molecular mass.

solubility and a given taxonomic group (phylum, order, or family rank) or with CAZy families (at the subfamily or family level, with and without inclusion of the distant relatives in the families). Upscaling the cultures to 50-mL flasks to generate sufficient amounts for the screening experiments did not reveal any major shift in expression yield. Similarly, we did not observe any change in the molecular mass of the proteins on expression yield, except for the largest proteins (>120 kDa), which are more likely than small proteins to be multimodular (Fig. 1*B*). To increase the number of soluble targets of the study, we tested the effect of solubilizing tags (27). Thus, 24 genes coding for insoluble proteins were cloned in four of the most popular fusion partners: DsbC, thioredoxin, maltose-binding protein, and CpB (NZYTech). These experiments did not improve the yield of soluble proteins, suggesting that our initial strategy was efficient.

Screening was conducted in 96-well microplates, in which the enzymes were incubated with a set of substrates distributed in the

microwells. The enzyme activity was revealed using a colorimetric reducing assay and size exclusion chromatography. Because some of the substrates were rare, expensive, or difficult to purify, we took advantage of the CAZy classification to divide the set of substrates into subsets to streamline the screening procedure and minimize loss of substrates (*SI Appendix*, Table S2). Members of a given family act on substrates whose glycosidic bonds have the same orientation regardless of the stereochemistry naming conventions (28). For instance, family GH39 contains both β-D-xylosidases and α-L-iduronidases, the substrates of which have an equatorial glycosidic bond (29). Therefore, the screening of enzymes classified into CAZyme families known to act on axially or equatorially linked glycosides was conducted on two sublibraries of substrates containing axial or equatorial glycosidic bonds, respectively (*SI Appendix*, Table S2). In a similar vein, enzymes classified into PL families were screened against a set of substrates containing only hexuronides. All proteins that did not cleave a substrate in their initially assigned sublibrary and all proteins from the most distant hypothetical sugar-cleaving enzyme category were tested on all available substrates.

For the uncharacterized GH/PL subfamily set, a function was ascribed to 38 proteins classified into 25 distinct GH and PL subfamilies. The sequences selected for screening belonged to a small number of well-defined subfamilies with no characterized representative, found mostly in the GH5 and GH43 families (Table 1). The activities observed for the newly characterized subfamilies from the GH5 (e.g., β-mannanase, β-D-glucopyranosidase, β-D-galactofuranosidase) and GH43 (e.g., β-D-galactofuranosidase, α-L-arabinofuranosidase) families were coherent with previously characterized subfamilies from the same families. The glucuronan lyase and heparin lyase activities identified in the PL7_4 and PL15_2 subfamilies, respectively, represent newly described substrate specificities in the corresponding families, which previously included only alginate lyases. These new specificities demonstrate the polyspecificity of these poorly explored PL families.

In the second set, comprising the distant relatives of established families of GHs and PLs (GHxx_dist and PLxx_dist), the success rate of substrate attribution was 23%, only one-half of that obtained with the set of proteins from well-defined subfamilies. Interestingly, however, in several cases, the enzyme activities ascribed to this distant relatives set corresponded to a new substrate specificity for the corresponding family (Table 2).

When a function could not be attributed to the GHxx_dist and PLxx_dist sequences using the sublibraries corresponding to known substrates of the cognate family, the proteins were screened on all substrates. By doing so, we found that a very distant relative

Table 1. Assignment of function to 25 subfamilies

CAZy subfamily	GenBank accession no.	Substrate	Organism
GH5_13	ZP_02065960.1	pNP-β-D-galactofuranoside	*Bacteroides ovatus* ATCC 8483
GH5_13	WP_018627464.1	pNP-α-L-arabinofuranoside	*Niabella aurantiaca* DSM 17617
GH5_18	ACU71175.1	pNP-β-D-mannopyranoside	*Catenulispora acidiphila* DSM 44928
GH5_35	ACT02895.1	Arabinoxylan	*Paenibacillus* sp. JDR-2
GH5_40	SCG47572.1	Konjac glucomannan	*Micromonospora rifamycinica* DSM 44983
GH5_41	ABD80383.1	β-mannan	*Saccharophagus degradans* 2–40
GH5_43	ADI04784.1	pNP-β-D-glucopyranoside	*Streptomyces bingchenggensis* BCW-1
GH5_45	SDT09889.1	pNP-α-L-arabinofuranoside (weak)	*Azotobacter vinelandii* DJ
GH5_45	ACO76963.1	pNP-β-D-glucopyranoside	*Pseudomonas oryzae* KCTC 32247
GH13_38	WP_029428030.1	pNP-α-D-maltopyranoside	*Bacteroides cellulosilyticus* WH2
GH13_38	ABD79820.1	pNP-α-D-maltopyranoside	*Saccharophagus degradans* 2–40
GH30_6	WP_028726386.1	pNP-β-D-cellobioside	*Parabacteroides gordonii* DSM 23371
GH43_2	ACU61943.1	pNP-α-L-arabinofuranoside	*Chitinophaga pinensis* DSM 2588
GH43_2	SDS19757.1	pNP-α-L-arabinofuranoside	*Mucilaginibacter mallensis* MP1X4
GH43_3	WP_007211145.1	pNP-β-D-galactofuranoside	*Bacteroides cellulosilyticus* WH2
GH43_8	EIY66405.1	pNP-β-D-galactofuranoside	*Bacteroides salyersiae* CL02T12C01
GH43_9	AMX03466.1	pNP-α-L-arabinofuranoside (weak)	*Microbulbifer thermotolerans* DAU221
GH43_17	ADQ05609.1	pNP-α-L-arabinofuranoside	*Caldicellulosiruptor owensensis* OL
GH43_18	WP_029328006.1	pNP-α-L-arabinofuranoside	*Bacteroides cellulosilyticus* WH2
GH43_18	WP_029427512.1	pNP-α-L-arabinofuranoside (weak)	*Bacteroides cellulosilyticus* WH2
GH43_18	WP_018628786.1	pNP-α-L-arabinofuranoside (weak)	*Niabella aurantiaca* DSM 17617
GH43_18	AHF90946.1	pNP-α-L-arabinofuranoside (weak)	*Opitutaceae bacterium* TAV5
GH43_20	SCF26596.1	pNP-α-L-arabinofuranoside	*Micromonospora echinospora* DSM 43816
GH43_20	CBG71495.1	pNP-α-L-arabinofuranoside	*Streptomyces scabiei* 87.22
GH43_23	ADO69162.1	pNP-α-L-arabinofuranoside (weak)	*Stigmatella aurantiaca* DW4/3–1
GH43_30	SCG78792.1	pNP-β-D-galactofuranoside	*Stackebrandtia nassauensis* DSM 44728
GH43_30	ADD39925.1	pNP-β-D-galactofuranoside (weak)	*Micromonospora siamensis* DSM 45097
GH43_31	AFL85801.1	pNP-β-D-galactofuranoside	*Belliella baltica DSM* 15883
GH43_32	ACB77177.1	pNP-β-D-galactofuranoside (weak)	*Opitutus terrae* PB90-1
GH43_32	SDH69004.1	pNP-β-D-galactofuranoside (weak)	*Leifsonia* sp. 197AMF
GH43_34	WP_044096317.1	pNP-α-L-arabinofuranoside	*Bacteroides cellulosilyticus* WH2
GH43_34	ZP_02066340.1	pNP-β-D-galactofuranoside	*Bacteroides ovatus* ATCC 8483
GH43_34	ACS99115.1	pNP-β-D-galactofuranoside	*Paenibacillus* sp. JDR-2
GH43_37	ADJ47124.1	pNP-β-D-galactofuranoside (weak)	*Amycolatopsis mediterranei* U32
PL7_4	ACU70527.1	β-glucuronan	*Catenulispora acidiphila* DSM 44928
PL14_2	AAC96919.1	Alginate	*Paramecium bursaria* chlorella virus 1
PL15_2	ALJ58962.1	Heparan sulfate	*Bacteroides cellulosilyticus* WH2

Enzyme activities (substrate specificities) were established using colorimetric and/or chromatography assays. The substrates used as well as the organism of origin of the protein are indicated. "Weak" indicates limited cleavage.

Table 2. Activity of enzymes distantly related to the described GH or PL (GH/PLxx_dist) families

Distant CAZy family	GenBank accession no.	Substrate	Organism
GH2_dist	WP_029427454.1	pNP-β-D-xylopyranoside (new)	*Bacteroides cellulosilyticus* WH2
GH2_dist	WP_029428707.1	Tamarind gum (new)	*Bacteroides cellulosilyticus* WH2
GH2_dist	WP_029428765.1	pNP-β-D-glucuronide	*Bacteroides cellulosilyticus* WH2
GH2_dist	WP_018628801.1	pNP-β-D-glucuronide	*Niabella aurantiaca* DSM 17617
GH3_dist	AJG33435.1	pNP-β-D-*N*-acetyl-glucopyranoside	*Rickettsia rickettsii* str. R
GH5_dist	ZP_06241352.1	pNP-β-D-mannopyranoside	*Victivallis vadensis* ATCC BAA-548
GH10_dist	EMS72420.1	pNP-β-D-xylopyranoside (weak)	*Clostridium termitidis* CT1112
GH16_dist	ZP_02063674.1	pNP-β-D-glucopyranoside (new)	*Bacteroides ovatus* ATCC 8483
GH20_dist	AEV99795.1	pNP-β-D-NAc-6Sulf-glucopyranoside	*Niastella koreensis* GR20-10
GH20_dist	AHF94523.1	pNP-β-D-NAc-glucopyranoside	Opitutaceae bacterium TAV5
GH31_dist	EIY61740.1	pNP-α-D-galactopyranoside	*Bacteroides salyersiae* CL02T12C01
GH36_dist	EIY66649.1	pNP-α-D-galactopyranoside	*Bacteroides salyersiae* CL02T12C01
GH36_dist	ACS99969.1	pNP-α-D-galactopyranoside	*Paenibacillus* sp. JDR-2
GH36_dist	ACS99975.1	pNP-α-D-galactopyranoside	*Paenibacillus* sp. JDR-2
GH36_dist	ZP_06242255.1	pNP-α-D-galactopyranoside	*Victivallis vadensis* ATCC BAA-548
GH42_dist	EIY59668.1	pNP-α-D-mannopyranoside	*Bacteroides salyersiae* CL02T12C01
GH49_dist	EDY96541.1	*Chaetomorpha* sp. CWP (new)	*Bacteroides plebeius* DSM 17135
GH49_dist	EDY96565.1	*Chaetomorpha* sp. CWP (new)	*Bacteroides plebeius* DSM 17135
GH51_dist	WP_084555785.1	Lichenan (new)	*Alkaliflexus imshenetskii* DSM 15055
GH76_dist	ADO68190.1	pNP-α-D-maltoside (new)	*Stigmatella aurantiaca* DW4/3–1
GH106_dist	WP_018627535.1	pNP-α-L-rhamnopyranoside	*Niabella aurantiaca* DSM 17617
GH106_dist	ACT02314.1	pNP-α-L-rhamnopyranoside	*Paenibacillus* sp. JDR-2
GH117_dist	WP_010134686.1	pNP-β-D-galactofuranoside	Flavobacteriaceae bacterium S85

This set encompasses enzymes that fall outside of established subfamilies or that are only distantly related to biochemically characterized enzymes. "New" designates novel specificity in the family. CWP, cell wall polysaccharide.

of family PL9 (GenBank accession no. AEI51087.1) is not a PL, but rather a GH able to cleave the main chain of the exopolysaccharide (EPS) secreted by the ubiquitous cyanobacterium *Nostoc commune*. Therefore, this enzyme and its orthologs define a new GH family, GH160 (Table 3), which may share structural similarity with PL9 lyases. This is the first report of an enzyme able to degrade the EPS of *Nostoc* spp.

The probability of ascribing a function to the most distant hypothetical sugar-cleaving enzymes (third set) was not expected to be very high; however, we validated GH or PL activities for approximately 18% (19 enzymes) of the 104 soluble proteins screened in this category (Table 3). These enzymes show extremely high divergence from enzymes grouped in known CAZyme families, and thus were identified as the first representatives of six new GH families and seven new PL families. Using chromatographic and NMR methods, we performed a thorough analysis of the reaction products of the four most original enzyme activities (three of which were not previously reported in any CAZy family) that we discovered during the course of our work. *SI Appendix*, Figs. S1–S4, respectively report the characterization of the end products of gellan lyase on gellan (founding member of PL33), of an enzyme able to cleave the polysaccharide secreted by *Nostoc* spp. (a founding member of GH160), of a galactanase activity (previously unreported in GH147), and of an endo-acting sulfated-arabinan hydrolase (previously unreported in GH49). In some cases, multiple representatives of the new families were characterized. The newly established PL family (PL33) was clearly polyspecific and grouped together gellan lyase, chondroitin sulfate lyase, and hyaluronan lyase. Two of our 13 new families (GH147 and GH148) were reported by others during the course of our work (30, 31). Although this decreases the number of newly described families from 13 to 11, it confirms that our approach is able to uncover families that were discovered using other approaches. Interestingly, our work revealed enzyme activities in families GH147 and GH148 different from those reported elsewhere, again demonstrating that our approach is valid for enzyme discovery. The characteristics of the new families reported here are summarized in *SI Appendix*.

Discussion

The selection of our targets was based on exploration of the uncharacterized branches of CAZyme family trees, that is, uncharacterized subfamilies, distant relatives of families (GH/PLxx_dist), or highly divergent proteins (GH/PL_nc). Thus, for the first time, a function was attributed to representatives of 25 well-defined subfamilies of the 48 subfamilies initially targeted. A variety of substrate activities have been previously described in the large GH5 and GH43 families (25, 26), which facilitated our investigation due to the expectation that the uncharacterized subfamilies would share a common activity with previously studied ones. This was particularly true in the case of family GH43, for which 14 of the 18 targeted subfamilies displayed α-L-arabinofuranosidase or β-D-galactofuranosidase activity, as has been observed in many previously described GH43 subfamilies. None of the GH43 targets that we produced exhibited activity against sugar beet arabinan or larchwood arabinogalactan, and the GH43 enzyme activity was recorded only on synthetic para-nitrophenyl (pNP)-glycoside substrates. Previous work has shown that the actual substrate of arabinofuranosidases can arise from the sequential action of other specific enzymes during action on complex glycans, such as arabinoxylan, arabinan, and arabinogalactan (30, 32, 33); however, such partially degraded substrates are often not readily available. Thus, it is possible that some differences may emerge between GH43 subfamilies when assaying the enzymes against complex substrates, as discussed by Mewis et al. (26). In only 7 of the 17 targeted GH5 subfamilies could the function be assigned, most likely due to the large number of eukaryotic targets selected in this family, resulting in a low yield of soluble proteins (16 of 50 soluble proteins in GH5 targets, compared with 66 of 102 soluble proteins in other families; hypergeometric test $P < 10^{-4}$). Seven different substrates—pNP-β-D-galactofuranoside, pNP-α-L-arabinofuranoside, pNP-β-D-mannopyranoside, arabinoxylan, konjac glucomannan, β-mannan, and pNP-β-D-glucopyranoside—were

Table 3. Substrate specificity of new CAZy families

New family	GenBank accession no.	Substrate	Activity	Organism
GH147	WP_029428318.1	β-galactan	Endo-β-(1,4)-galactanase	*Bacteroides cellulosilyticus* WH2
GH147	EFI37897.1	β-galactan	Endo-β-(1,4)-galactanase	*Bacteroides* sp. 3_1_23
GH148	AGN79260.1	Konjac glucomannan	Endo-β-(1,4)-glucosidase	*Pseudomonas putida* H8234
GH148	ACR13278.1	Konjac glucomannan	Endo-β-(1,4)-glucosidase	*Teredinibacter turnerae* T7901
GH157	WP_029429093.1	CM-curdlan	Endo-β-glycosidase	*Bacteroides cellulosilyticus* WH2
GH158	ZP_06243608.1	CM-curdlan	Endo-β-glycosidase	*Victivallis vadensis* ATCC BAA-548
GH159	WP_007210837.1	pNP-β-D-galactofuranoside	β-D-galactosidase	*Bacteroides cellulosilyticus* WH2
GH160	AEI51087.1	EPS *Nostoc commune* (new)	Endo-β-(1,4)-galactosidase	*Runella slithyformis* DSM 19594
PL30	WP_029426181.1	Hyaluronan	Endo-hyaluronan lyase	*Bacteroides cellulosilyticus* WH2
PL31	ABD82242.1	β-glucuronan	Endo-β-(1,4)-glucuronan lyase	*Saccharophagus degradans* 2-40
PL31	AGF62897.1	β-glucuronan	Endo-β-(1,4)-glucuronan lyase	*Streptomyces hygroscopicus* subsp. *jinggangensis* TL01
PL32	EIY62149.1	β-mannuronan	Endo-mannuronan lyase	*Bacteroides salyersiae* CL02T12C01
PL33	ALJ61728.1	Hyaluronan	Endo-hyaluronan	*Bacteroides cellulosilyticus* WH2
PL33	AHF90976.1	Gellan (new)	Endo-gellan lyase	Opitutaceae bacterium TAV5
PL33	AHF90672.1	Chondroitin sulfate	Endo-chondroitin sulfate lyase	Opitutaceae bacterium TAV5
PL33	AHF90411.1	Gellan (new)	Endo-gellan lyase	Opitutaceae bacterium TAV5
PL34	AHF91913.1	Alginate	Endo-alginate lyase	Opitutaceae bacterium TAV5
PL35	ZP_06241351.1	Chondroitin	Endo-chondroitin lyase	*Victivallis vadensis* ATCC BAA-548
PL36	WP_084332190.1	β-mannuronan	Endo-mannuronan lyase	*Flavobacterium denitrificans* DSM 15936

The substrate and the modality of substrate degradation are specified. "New" designates novel specificity not reported previously. Note that families GH147 and 148 were reported by other groups during the course of our work (30, 31). CM, carboxymethyl.

needed to characterize the seven GH5 subfamilies, in agreement with the high polyspecificity already reported for the GH5 family (25).

The assignment of function to distant GH and PL (GH/PLxx_dist) proteins was more challenging but was also a source of discovery. Seven of the 23 GH/PLxx_dist proteins characterized were active on a substrate that had not been previously reported in the corresponding family. For example, the endo-β(1,4)-glucanase activity of a GH2_dist protein (GenBank accession no. WP_029428707.1), revealed by the degradation of tamarin gum (xyloglucan), had not been previously observed in family GH2. Similarly, another GH2_dist protein (GenBank accession no. WP_029427454.1) displayed a β-D-xylosidase activity not previously reported in family GH2. Interestingly, family GH2 was created in 1991 and has been the subject of numerous biochemical investigations. Therefore, our results demonstrate that polyspecificity remains underestimated even for such well-established GH families, with a direct impact on functional inference from sequence data only. Even more unexpected was the finding that the two distant relatives of the GH49 family (GenBank accession nos. EDY96541.1 and EDY96565.1) can cleave a cell wall polysaccharide from the green algae *Chaeotomorpha* spp. and *Cladophora* spp., whose backbone is composed of sulfated arabinan (34), a structure highly dissimilar to dextran and pullulan, previously known as the sole substrates of family GH49 enzymes. The results of NMR analysis of the reaction products of GenBank EDY96541.1 are presented in *SI Appendix*, Fig. S4.

The rationale for selecting the most distant hypothetical sugar-cleaving enzymes category was to explore the frontiers of the CAZy families so divergent that bioinformatic methods failed to predict putative functions. Functional screening of the proteins of this category led to assignment of the function of enzymes that are the founding members of 13 new families, 2 of which were described by others during the course of our work. From the establishment of the first 35 GH families in 1991 (15) to the 156 families described to date (for a continuously updated classification, see www.cazy.org), an average of approximately 5 new GH families are created each year. The number of PL families is lower because this class of enzymes is specific to polyuronic acid substrates; starting with 9 PL families in 1999 and reaching 29 to date, the number of PL families has grown at a rate of approximately 1 new family per year. Therefore, an average of six new GH and PL families are described each year. Here we have identified roughly twice the number of new families reported worldwide per year. Our substrate screening strategy for the proteins having very low homology with known enzymes has proven to be efficient for identifying novel candidate GHs and PLs. In a virtuous circle, the novel families now define new frontiers to be explored. This method can now be extended to new sets of hypothetical sugar-cleaving enzymes.

We have explored a portion of the large amount of sequence data rationally grouped and classified in the CAZy database. To continue exploring the diversity of sugar-cleaving enzymes, the production of several thousands of recombinant GHs and PLs is now technically possible (35) and is limited only by the cost of gene synthesis which, fortunately, continues to decrease. Thus, the main bottleneck for functional assignment likely is not protein production, but rather the availability of a large and diverse array of substrates. Although this was not a major problem for the screening of enzymes classified in subfamilies of the GH5 and GH43 families, the assignment of function to distantly related enzymes (GHxx_dist and PLxx_dist) and the most distant hypothetical sugar-cleaving enzymes depended directly on the diversity of substrates in the screening library. Thus, the discovery of the first gellan lyase, the first *N. commune* EPS hydrolase, and the first cladophoran hydrolases was possible only because the respective substrates were present in our glycan library. Significantly, the function of more than 243 soluble proteins produced during this work could not be identified, presumably due to of the lack of suitable substrates, representing a large and untapped potential for subsequent discoveries.

Conclusion

We have shown that it is possible to ascribe the function of putative enzymes distantly related to experimentally characterized GHs and PLs through a systematic exploration of the sequence space coupled with a screening procedure against a collection of diverse carbohydrate substrates. The effectiveness of this strategy is illustrated by the description of 11 new families, the discovery of three new substrate specificities, and the assignment of function to

26 subfamilies, starting from a set of 564 bioinformatically selected proteins. A similar approach conducted on thousands of targets would not only generate more discoveries, but also enable a more reliable, knowledge-based functional prediction for gene products from genomic or metagenomic sequencing projects. Given the decreasing cost of recombinant protein production, the main remaining bottleneck is the availability of a substrate library that parallels the diversity of the glycan structures found in nature.

Materials and Methods

Bioinformatics: Selection of Targets. The daily updates of the CAZy database rely on the careful analysis of newly released protein sequences from GenBank by comparing them with previously analyzed/stored sequences (21, 36). To obtain accurate annotation, our procedures make use of sequence libraries of varying levels of granularity: subfamilies, families, and remote relatives. In this work, targets were drawn from three categories: "uncharacterized subfamilies," "distant members within families," and "hypothetical sugar-cleaving enzymes." Details of the selection process are provided in *SI Appendix, Materials and Methods*.

Screening Experiments. For this study, *E. coli* codon optimization, gene synthesis, and cloning of the 539 targets was outsourced to NZYTech. High-throughput expression and purification assays were conducted following the protocol described by Saez and Vincentelli (27). The soluble proteins were screened against the collection of substrates according to the method developed by Fer et al. (37). All positive hits were produced at least twice, and the most interesting enzymes were fully biochemically characterized. The protocol is described in detail in *SI Appendix, Materials and Methods*.

ACKNOWLEDGMENTS. This work was supported by the French National Research Agency (Grant ANR-14-CE06-0017) and the French Infrastructure for Integrated Structural Biology (FRISBI) (Grant ANR-10-INSB-05-01). W.H. has received support from the Glyco@Alps Cross-Disciplinary Program (Grant ANR-15-IDEX-02), Labex ARCANE, and Grenoble Graduate School in Chemistry, Biology, and Health (Grant ANR-17-EURE-0003). B.H., N.T., and R.V. have received support from FRISBI (Grant ANR-10-INSB-05-01).

1. Venter JC, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74.
2. Sunagawa S, et al.; Tara Oceans Coordinators (2015) Ocean plankton: Structure and function of the global ocean microbiome. *Science* 348:1261359.
3. Gilbert JA, Jansson JK, Knight R (2014) The Earth microbiome project: Successes and aspirations. *BMC Biol* 12:69.
4. Muegge BD, et al. (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 332:970–974.
5. Methé BA, et al.; Human Microbiome Project Consortium (2012) A framework for human microbiome research. *Nature* 486:215–221.
6. Huttenhower C, et al.; Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214.
7. Hanson AD, Pribat A, Waller JC, de Crécy-Lagard V (2009) "Unknown" proteins and "orphan" enzymes: The missing half of the engineering parts list—and how to find it. *Biochem J* 425:1–11.
8. Roberts RJ (2011) Combrex: Computational bridge to experiments. *Biochem Soc Trans* 39:581–583.
9. Devos D, Valencia A (2000) Practical limits of function prediction. *Proteins* 41:98–107.
10. Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLOS Comput Biol* 5:e1000605.
11. Jones CE, Brown AL, Baumann U (2007) Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics* 8:170.
12. Altschul SF, Koonin EV (1998) Iterated profile searches with PSI-BLAST—A tool for discovery in protein databases. *Trends Biochem Sci* 23:444–447.
13. Eddy SR (2011) Accelerated profile HMM searches. *PLOS Comput Biol* 7:e1002195.
14. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10:845–858.
15. Henrissat B (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* 280:309–316.
16. Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* 293:781–788.
17. Henrissat B, Bairoch A (1996) Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* 316:695–696.
18. Campbell JA, Davies GJ, Bulone V, Henrissat B (1997) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem J* 326:929–939.
19. Lombard V, et al. (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem J* 432:437–444.
20. Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B (2013) Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol Biofuels* 6:41.
21. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42: D490–D495.
22. Coutinho PM, Stam M, Blanc E, Henrissat B (2003) Why are there so many carbohydrate-active enzyme-related genes in plants? *Trends Plant Sci* 8:563–565.
23. Stam MR, Danchin EGJ, Rancurel C, Coutinho PM, Henrissat B (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: Towards improved functional annotations of α-amylase–related proteins. *Protein Eng Des Sel* 19:555–562.
24. St John FJ, González JM, Pozharski E (2010) Consolidation of glycosyl hydrolase family 30: A dual domain 4/7 hydrolase family consisting of two structurally distinct groups. *FEBS Lett* 584:4435–4441.
25. Aspeborg H, Coutinho PM, Wang Y, Brumer H, 3rd, Henrissat B (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol* 12:186.
26. Mewis K, Lenfant N, Lombard V, Henrissat B (2016) Dividing the large glycoside hydrolase family 43 into subfamilies: A motivation for detailed enzyme characterization. *Appl Environ Microbiol* 82:1686–1692.
27. Saez NJ, Vincentelli R (2014) High-throughput expression screening and purification of recombinant proteins in *E. coli*. *Methods Mol Biol* 1091:33–53.
28. Henrissat B, Davies G (1997) Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol* 7:637–644.
29. Henrissat B, et al. (1995) Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc Natl Acad Sci USA* 92: 7090–7094.
30. Luis AS, et al. (2018) Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic *Bacteroides*. *Nat Microbiol* 3:210–219.
31. Angelov A, et al. (2017) A metagenome-derived thermostable β-glucanase with an unusual module architecture which defines the new glycoside hydrolase family GH148. *Sci Rep* 7:17306.
32. Ndeh D, et al. (2017) Metabolism of a complex pectin reveals novel enzymatic adaptations in the human gut microbiota. *Nature* 544:65–70.
33. Cartmell A, et al. (2018) A surface endogalactanase in *Bacteroides thetaiotaomicron* confers keystone status for arabinogalactan degradation. *Nat Microbiol* 3:1314–1326.
34. Arata PX, Quintana I, Raffo MP, Ciancia M (2016) Novel sulfated xylogalactoarabinans from green seaweed *Cladophora falklandica*: Chemical structure and action on the fibrin network. *Carbohydr Polym* 154:139–150.
35. Turchetto J, et al. (2017) High-throughput expression of animal venom toxins in *Escherichia coli* to generate a large library of oxidized disulphide-reticulated peptides for drug discovery. *Microb Cell Fact* 16:6.
36. Cantarel BL, et al. (2009) The carbohydrate-active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res* 37:D233–D238.
37. Fer M, et al. (2012) Medium-throughput profiling method for screening polysaccharide-degrading enzymes in complex bacterial extracts. *J Microbiol Methods* 89:222–229.